Contact is welcome but reply is not instant. Slides are yours to use though I would appreciate acknowledgement if it is possible to do so.

Daniel E. Geer, Jr., Sc.D.
Geer Risk Services
P.O. Box 390244
Cambridge, Mass. 02139
U.S.A.
+1.617.492.6814
dan@geer.org

When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.

-- William Thomson, Lord Kelvin, 1883

To measure is to know.

-- James Clerk Maxwell, 1831-1879

The revolutionary idea that defines the boundary between modern times and the past is the mastery of risk: the notion that the future is more than a whim of the gods and that men and women are not passive before nature. Until human beings discovered a way across that boundary, the future was the mirror of the past or the murky domain of oracles and soothsayers who held a monopoly over knowledge of anticipated events.

-- Peter Bernstein, 1996

Amateurs study cryptography;
professionals study economics.

-- Allan Schiffman, 2 July 04

Security folks are from Mars.
Business people are from Wharton.

-- Adam Shostack, 19 October 04

Measurement motivates.

-- John Kenneth Galbraith

Outline

- Outline
- Measurement for a reason
- State of the art
- Breaking new ground
- Sustainability is not simple
- Wrapup

An outline in spirit -- this is what we'll cover, but we'll do it over and over, not once per bullet.

Good enough security

Sandhu

1. Good enough is good enough
2. Good enough always beats perfect
3. The really hard part is determining what is good enough

This is why we are here today.

An opening thought... we are here because we need to know how to deal with #3. That's all.

Sandhu R: "Good-Enough Security: Toward a Pragmatic Business-Driven Discipline," IEEE Internet Computing, January/February 2003, v5 n3 p66; as found at http://www.list.gmu.edu/journals/ic/03-sandhu-good.pdf

Goal for today

Self sufficiency, not expertness

Expertness is good, no, wonderful, but we are too young yet at this and, in any case, self sufficiency is half knowing how to invent and half knowing when to do so. We hope for nothing more than that.

Goal for tomorrow

To move from a culture of fear
to a culture of awareness and
then a culture of measurement.

Everyone likely to peruse this tutorial has moved away from fear, uncertainty, and doubt (FUD) and is doubtless trying to bring his/her organization to a culture of awareness. The author hopes that this tutorial helps move those organizations to a culture of measurement.

The 20th century was notable for the creation of a culture of measurement through and through the manufacturing sector. The 21st century seems to be ready to mirror that in the information sector.

Why measure?

- There’s never enough <X> to go around
- To play better, you must keep score
- Discipline is easier with numbers

Each is sufficient and collectively more so; if you are allocating scarcity then you need to know how much scarcity you have. If you want to improve performance, then you need a performance measure. If many people have duties, then you need a way to say what those duties are such that most anyone can tell if the duties are actually being performed.

Trade-offs

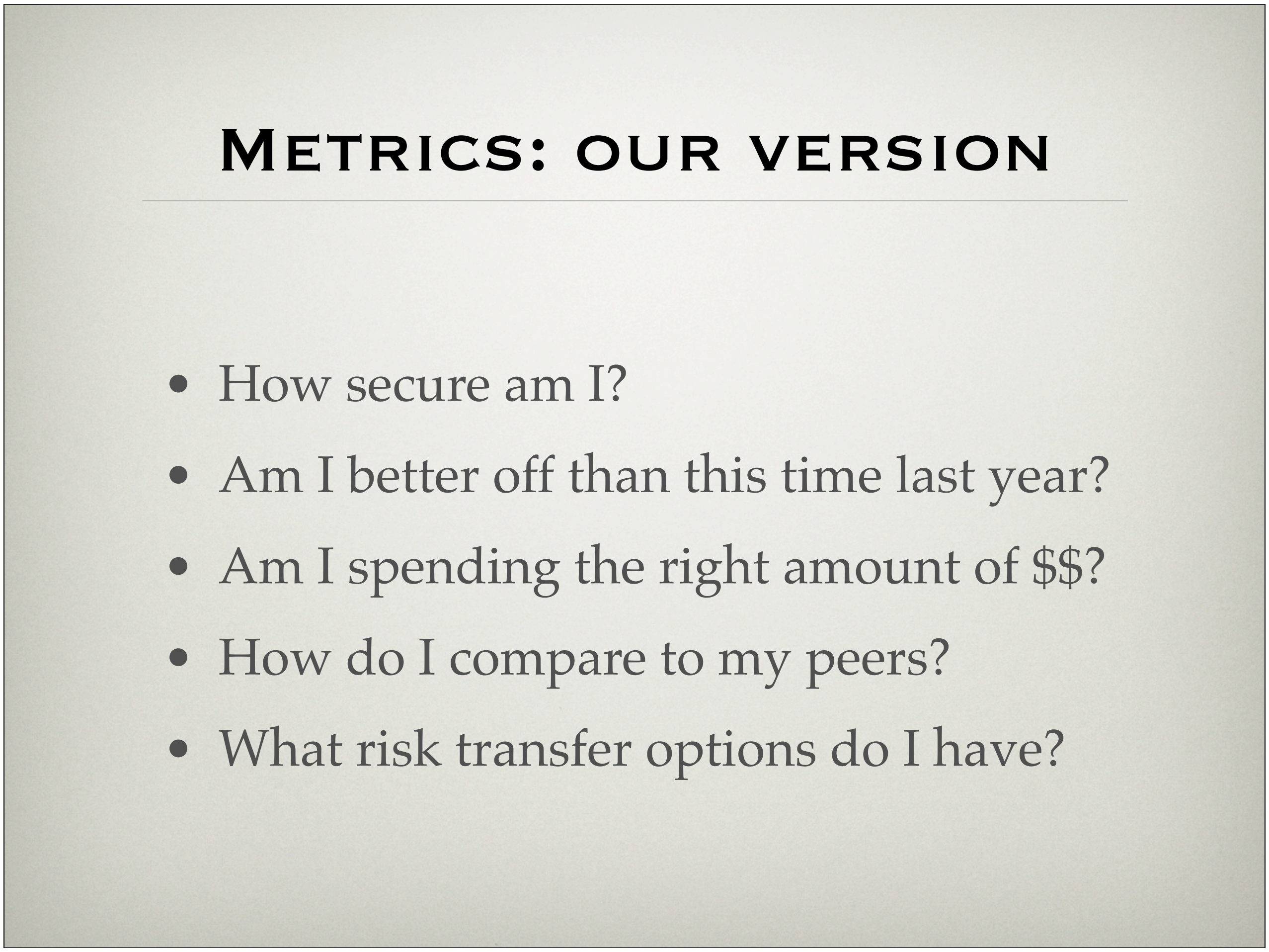
- Security is about tradeoffs; but you know that
- It is easier to make tradeoffs when you have a measure to compare them with
- Even then, it is not necessarily easy

All security is about tradeoffs, and tradeoffs are easier to make in the common language of numbers commonly agreed upon, and that that is so is understandable to tinkers, tailors, soldiers, sailors, rich men, poor men, beggar men, thiefs, doctors, lawyers, and indian chiefs alike.

Metrics

- Other industries have theirs, why not us?
  - Logistics: $/mile, percent full loads
  - Warehouses: $/□, turn-rate
  - Telecom: $/connection, saturation

What can we do here in security? What should we measure or, for that matter, what can we measure?



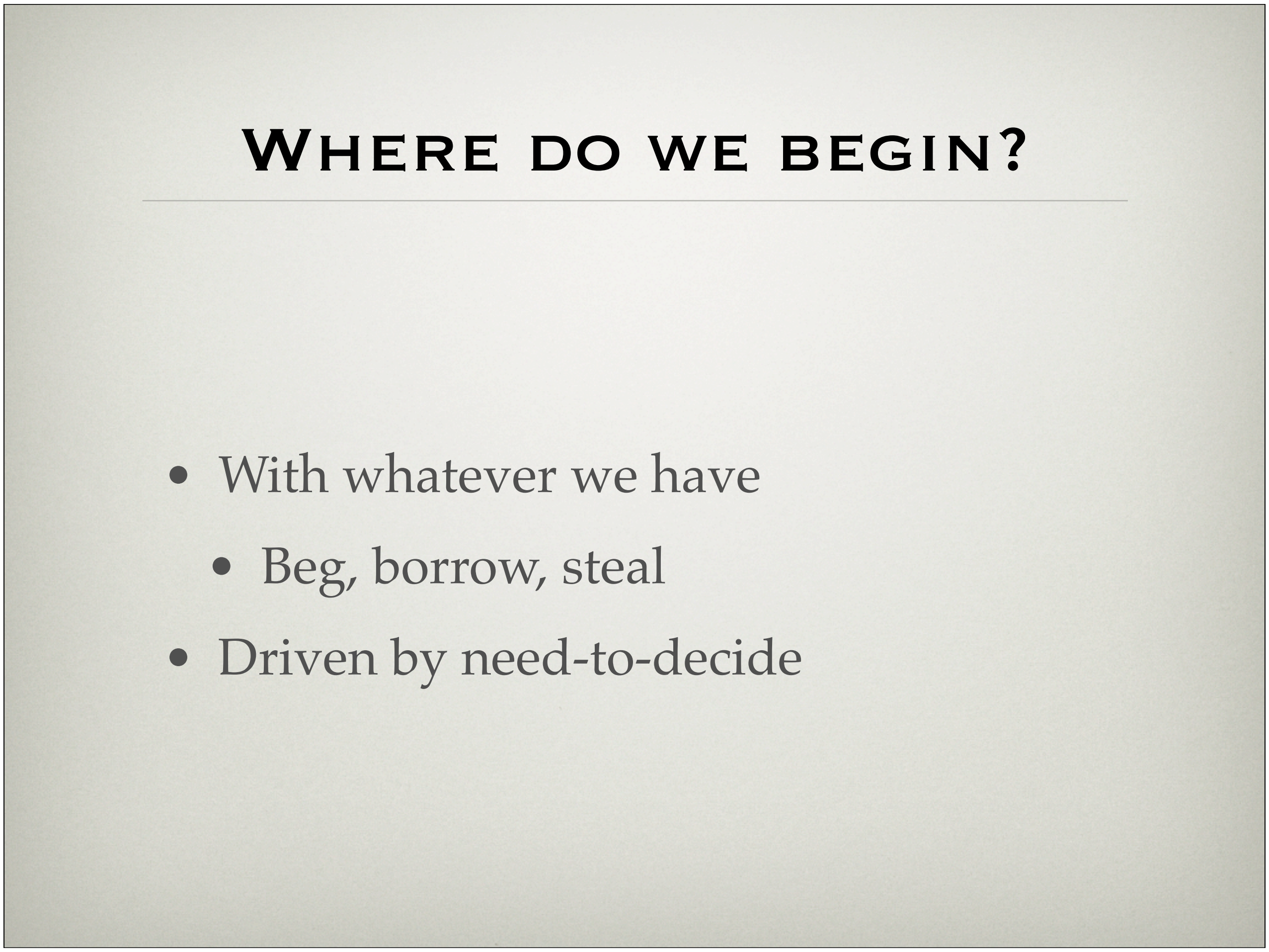
These are precisely the questions that any CFO would want to know and we are not in a good position to answer. The present author was confronted with this list, exactly as it is, by the CISO of a major Wall Street bank with the preface “Are you security people so stupid that you cannot tell me....”

This particular CISO came from management audit and therefore was also saying that were he in any other part of the bank, bond portfolios, dervative pricing, equity trading strategies, etc., he would be able to answer such questions to five digit accuracy. The questions are sound.

Measure what?

- We'll come to that, but...
- Early on: anything you can
- Later: what models tell you to

Start from where you are and go to where you want to be.



No one is without enough resources to begin. To the extent you have any choices, choose (and we will say this over and over) to be decision support.

Theft is here a virtue

- Public Health
- Insurance
- Accelerated Failure Time testing
- Portfolio Management
- *..., etc.*

There are others. We will steal from them and everyone else. We will use the skill sets that have already had their evolutionary morphing and apply them to our field. Why, because WE DO NOT HAVE TIME TO START FROM SCRATCH.

Terms of engagement

Definitions

- Computer security has tended to reinvent terms when perfectly good terms already exist
- Hence, we have to be careful about terms

All fields reach a point at which they begin to have specialized words. This is sometimes good -- the field may have concepts that need the marker of a word to go with them -- and it is sometimes bad -- using words that are unfamiliar when there are perfectly good non-specialist uses, a phenomenon that is generally due to making a guild out of some set of practitioners.

Definitions

- Vulnerability
- Threat-Source
- Threat
- Risk, systematic & unsystematic
- Risk Management

These are some of the terms for which we need agreed upon, common understandings. In most ways, what we agree upon is not as important as that we agree.

Vulnerability

NIST SP 800-30

A flaw or weakness in system security procedures, design, implementation, or internal controls that could be exercised and result in a security breach or a violation of a system's security policy

SP 800-30: Risk Management Guide for Information Technology Systems, July 2002.
http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf

Threat-Source

NIST SP 800-30

Either (1) intent and method targeted at the intentional exploitation of a vulnerability or (2) a situation and method that may accidentally trigger a vulnerability

ibid

Common Threat-Sources

NIST SP 800-30

- Natural - flood, earthquake
- Human
  - unintentional (drop vase)
  - intentional (throw vase)
- Environmental - chemical tank leak

ibid

Threat

NIST SP 800-30

The potential for a threat-source to exercise (accidentally trigger or intentionally exploit) a specific vulnerability

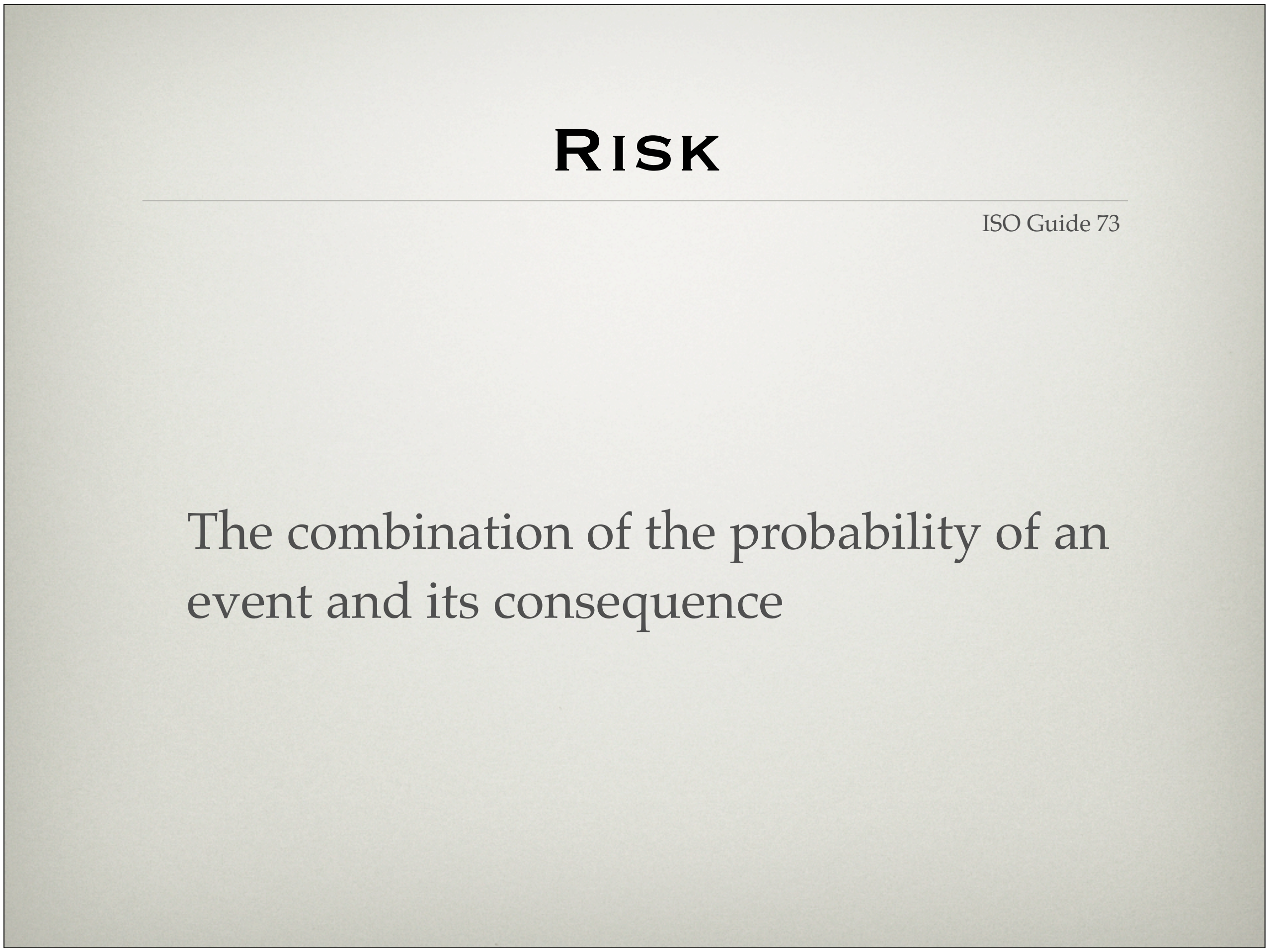
ibid

Threat likelihood

NIST SP 800-30

controls ☞ opponent ☟	are effective	are ineffective
has capability & motivation	Medium	High
does not have	Low	Low

ibid



By quotation; ISO standards are not open source (which is preposterous, but not today's focus).

Risk, systematic

RIMS

Chance of loss that is predictable under relatively stable circumstances

(fire, wind, or flood produce losses that, in the aggregate over time, can be accurately predicted despite short-term fluctuations)

Risk & Insurance Management Society: http://www.rims.org/MGTemplate.cfm?Section=Glossary&template=Magazine/GlossaryDisplay.cfm&GlossaryID=1545

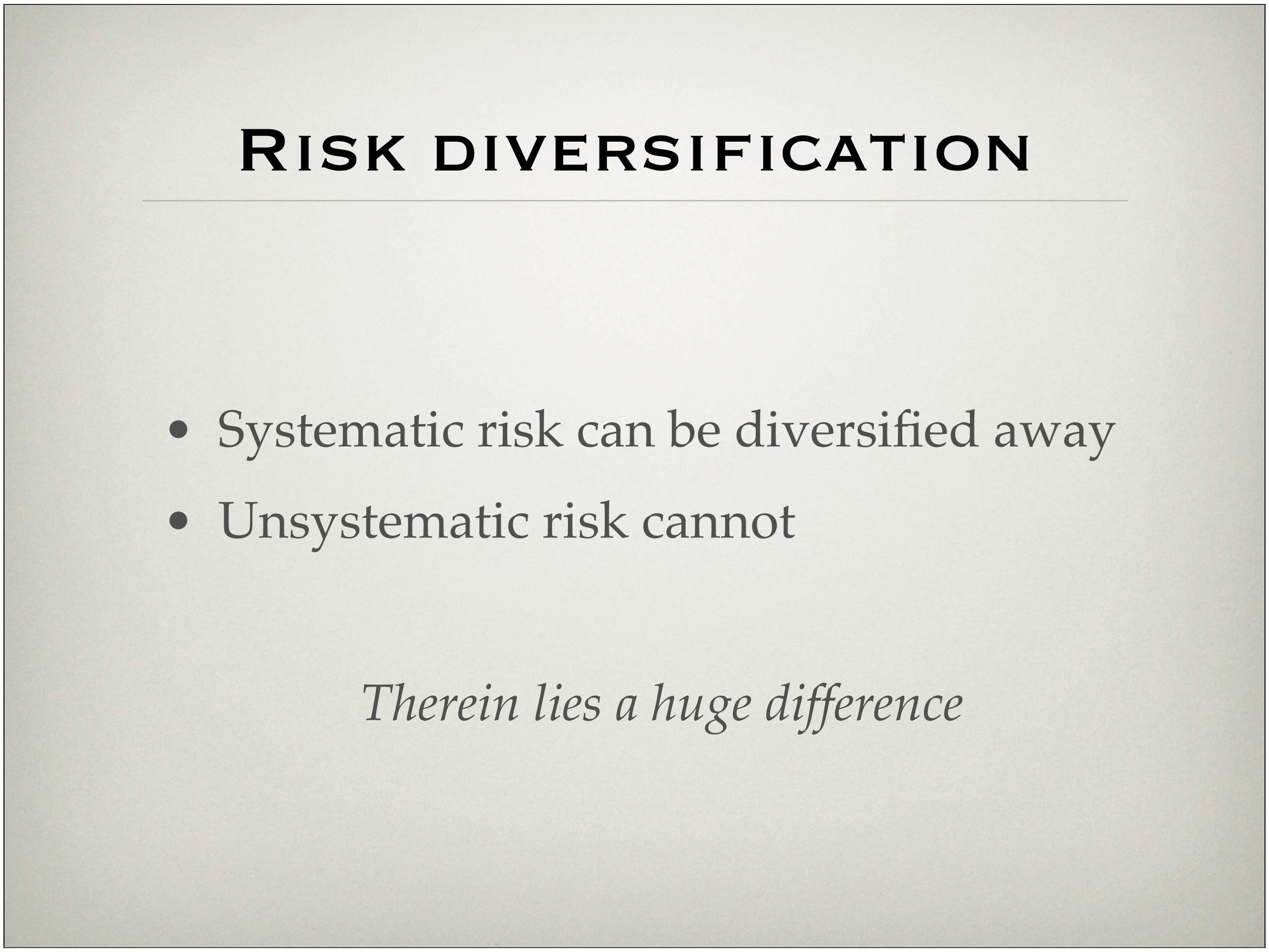
Risk, unsystematic

RIMS

Chance of loss that is unpredictable in the aggregate because it results from forces difficult to predict

(recession, unemployment, epidemics, war-related events, etc.)

Risk & Insurance Management Society: http://www.rims.org/MGTemplate.cfm?Section=Glossary&template=Magazine/GlossaryDisplay.cfm&GlossaryID=1547



http://www.riskglossary.com/link/risk_aversion.htm

(first mentioned in Markowitz H : “Portfolio Selection,” Journal of Finance, v7 p77-91, 1952.)

Diversification against systematic risk

- TCP/IP assumes diverse paths
- Data centers rely on diversified power and bandwidth
  - Lessons learned in Manahattan, 9/11
- Reciprocity / mutual aid agreements
- Subject to measurement and models

Preparation against unsystematic risk

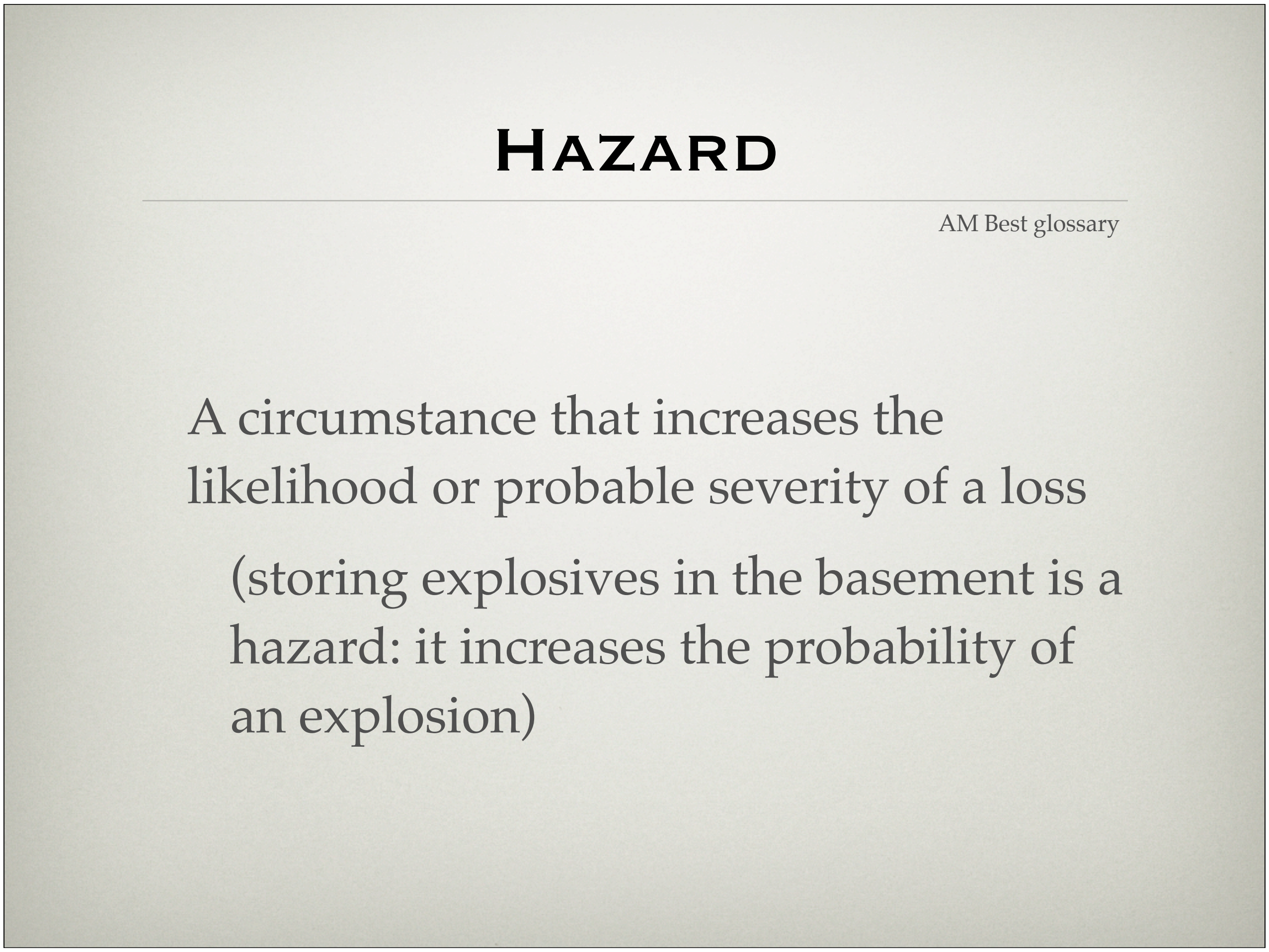
- Disaster recovery / fallback plans
  - Diminished modes of operation
- Unsystematic risks can affect all parties
  - Reciprocity / mutual aid may fail
- Not directly subject to measurement

Risk management, what

Bernstein

The essence of risk management lies in maximizing the areas where we have some control over the outcome, while minimizing the areas where we have absolutely no control over the outcomes and the linkage between effect and cause is hidden from us.

Bernstein P: _Against the Gods_, John Wiley & Sons, 1996.



A.M. Best's glossary is everywhere (except on the company's own website).

Peril

AM Best glossary

The cause of a possible loss

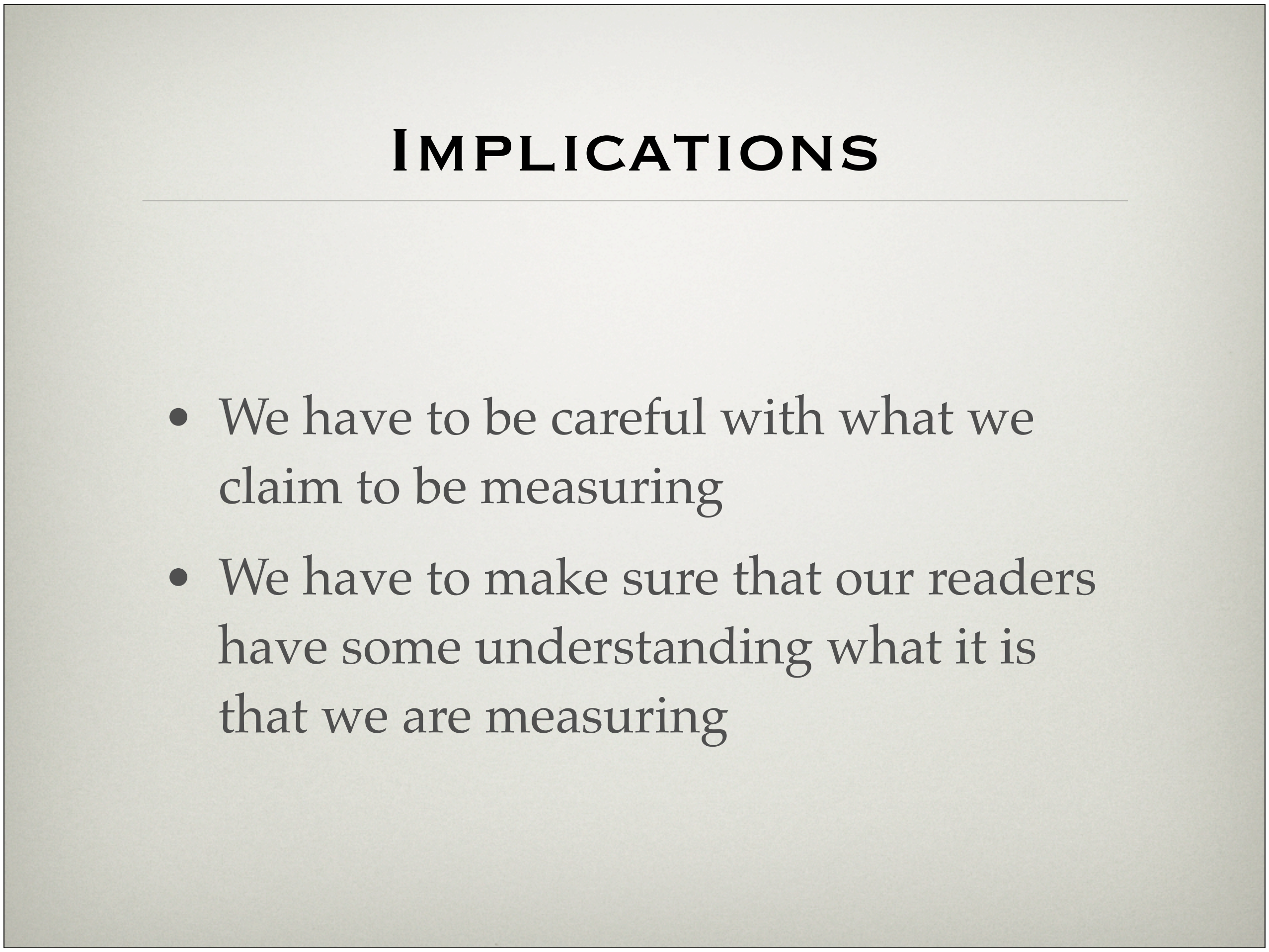
ibid

Why measure? (again)

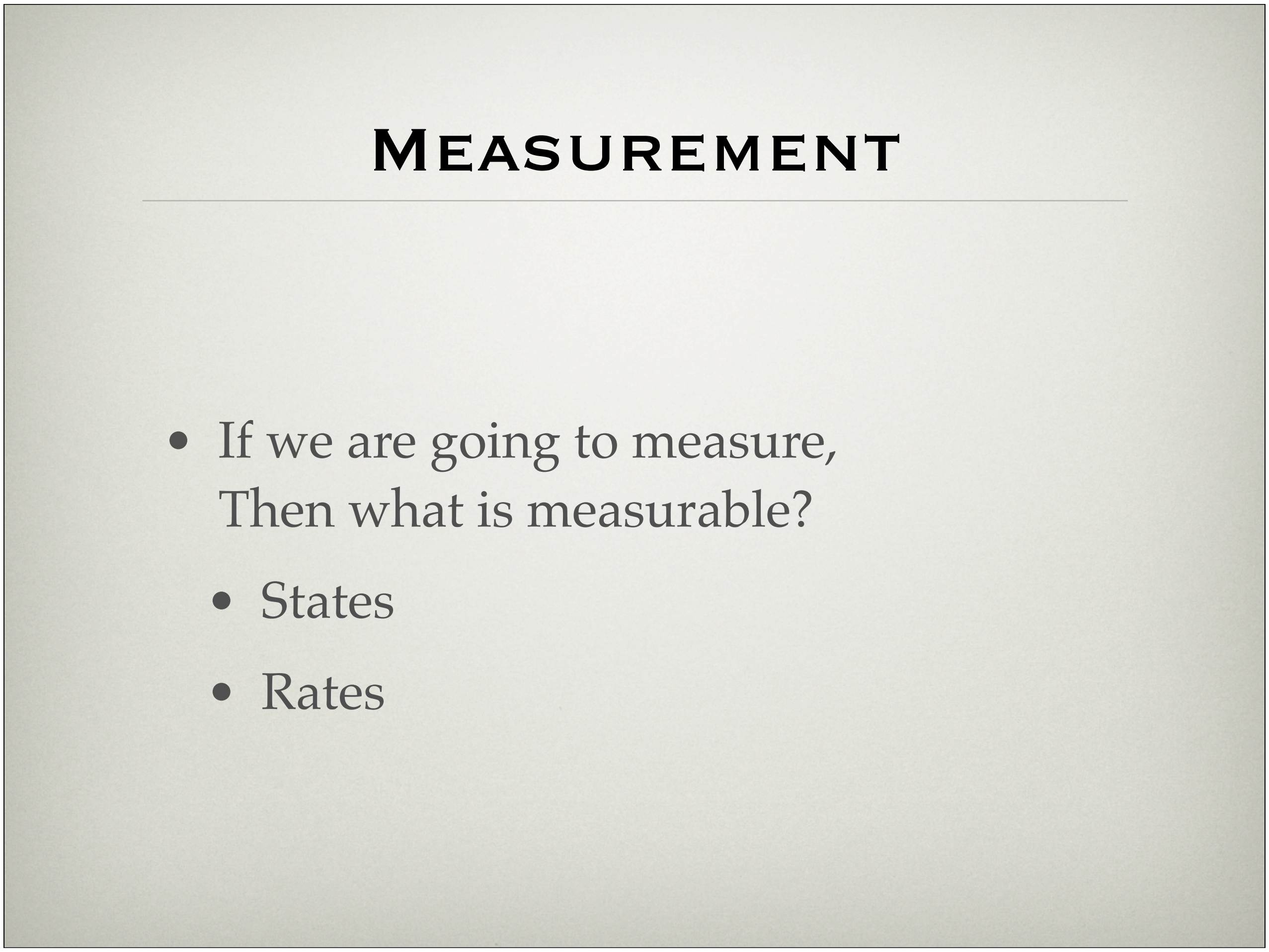
Premise: We measure to support decision making, possibly under fire

Consequent: Knowing which way you err can more vital than suppressing error

So, why are measurements made? To support decision making and, therefore, errors only matter if they bias decisions. As such, suppressing error can be useful or it can be useless. The question is often more the latter, at least insofar as knowing that you are high or low, east or west, etc., is often all that is needed to decide whether to ascend or descent, to go left or to go right.



If we are, however, making decisions then it does pay to be careful what we put into those decisions. As has been said since the 1950s, “garbage in , garbage out” and for measurement-driven decision making that is oh, so true.



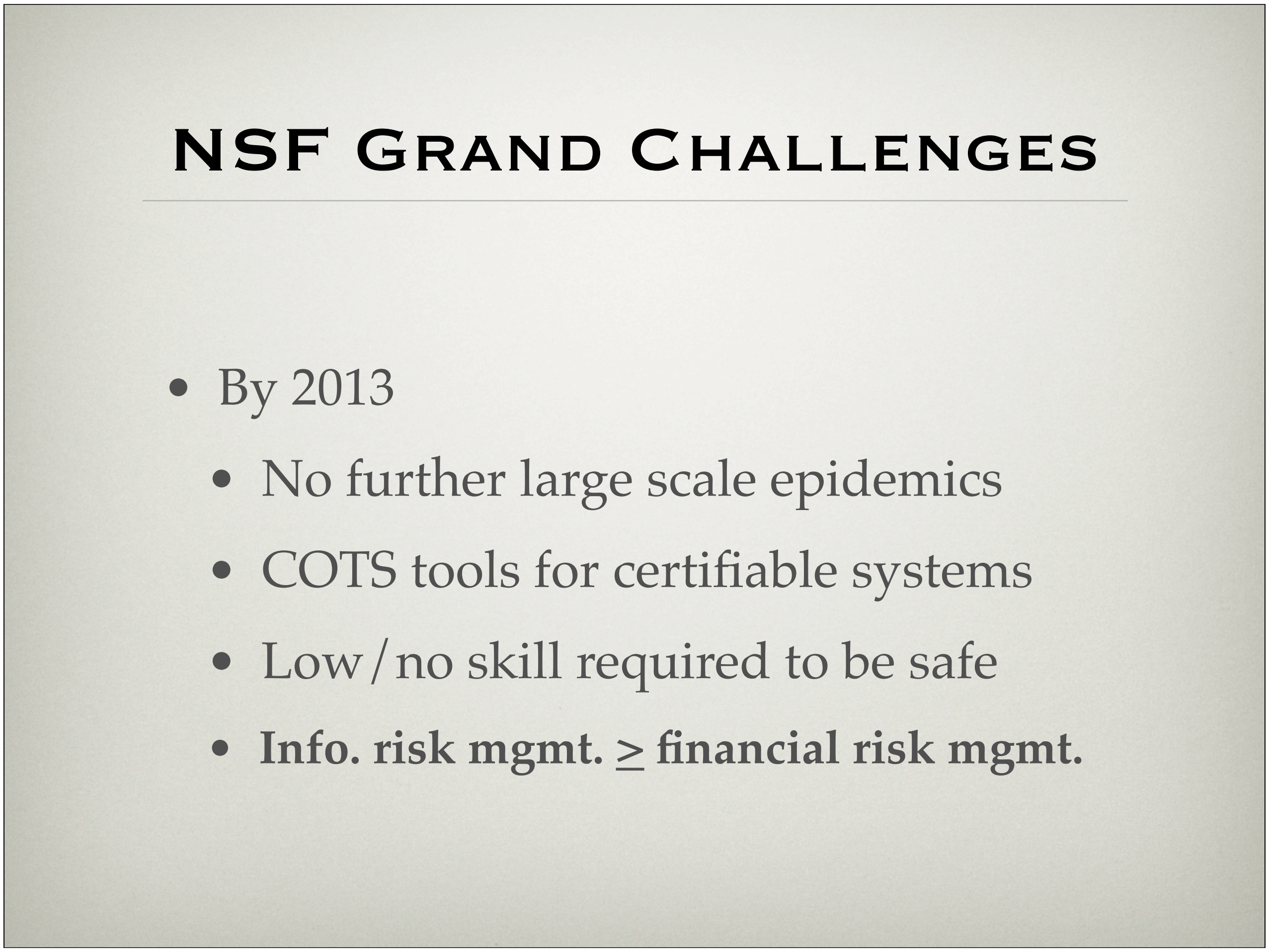
That which can be measured really comes down to two categories of measurements, viz., states and rates. If this reminds you of Heisenberg, well, it should. You can measure the position of something or you can measure the momentum of something. This is not atomic physics, so we can perhaps do both at once.

Risk Management

ISO Guide 73

The process of determining an acceptable level of risk, assessing the current level of risk, taking steps to reduce risk to the acceptable level, and maintaining that level of risk

ibid



In November, 2003, the Computing Research Association held a limited attendance, invitation only retreat in Virginia at the behest of the National Science Foundation. The purpose was to set the ten-year research agenda in information security <http://www.cra.org/Activities/grand.challenges/security/home.html>. Here are the results in lay terms: An end to epidemics, commercial off the shelf (COTS) tools for building certifiable systems, improvements in semantics and user interface such that one need not be an expert to be safe, and information risk management of a quantitative sophistication as good as that of financial risk management.

These are high goals, and at the same time it is horrifying that any of them could take a decade to deliver. On the other hand, if they do take as much as a decade, then starting now is crucial.

See http://www.cra.org/Activities/grand.challenges/security/home.html

Business drivers towards risk management

Jaquith

- Information asset fragility
- Provable security
- Cost pressure
- Accountability

The first of several references from Security Metrics, Andrew Jaquith, Addison-Wesley, 2006, this listing describes why risk management is the only real alternative to fear, uncertainty, and doubt. Which is more, the fragility of information assets, the absence of provable security (hence no natural upper bound on what you could spend), general competition for money with other IT projects, and the rain of new information-related regulations all work together to make a measurement regime for risk management essential.

Risk management, culture

J.Reason

Pathologic	Bureaucratic	Generative
Don't want to know	May not find out	Actively seek
Messengers "shot"	Heard if they arrive	Messengers rewarded
Responsibility shirked	Compartmentalized	Responsibility shared
Failure punished	Local repairs only	Failures beget reforms
Ideas discouraged	Ideas beget problems	Ideas welcomed

Reason J: _Managing the Risks of Organizational Accidents_, Ashgate Publishing Limited, 1997.

Risk management, what

Borge

Risk management means taking deliberate action to shift the odds in your favor – increasing the odds of good outcomes and reducing the odds of bad outcomes.

Borge D: _The Book of Risk_, John Wiley & Sons, 2001.

Dan Borge was in charge of risk management for Bankers Trust when they made more money than anyone else and did so because, explicitly, of their ability to take better chances.

Risk management, why

Borge

The purpose of risk management is
to improve the future,
not to explain the past.

ibid

Security metrics...

... are the servants of risk management

And risk management...

...is about making decisions

Therefore

The only security metrics we are interested in are those that support decision making about risk for the purpose of managing that risk.

This is a real bias. If the metric does not have a role in decision making, leave it to someone else to do, if ever.

As a kind of logic,

without	there can be no
measurement	reproducibility
reproducibility	conclusion
conclusion	control
control	security

Another way of thinking about it, specifically that if you want security then you must control the future, if you want to control the future then you must be able to draw conclusions from what you know, if you want to draw conclusions then the basis for those conclusions must be reproducible, and if you want reproducible bases you have to have a measurement regime.

So, what do good metrics look like?

A good metric must (1/5)

Jaquith

Be consistently measured

- The criteria must be objective.
- The criteria must be repeatable.

Andrew Jaquith as quoted in Berinato S, "A Few Good Metrics," CSO Magazine, July 2005; see http://www.csoonline.com/read/070105/metrics.html

A good metric must (2/5)

Jaquith

Be cheap to gather

- Using automated tools helps
- ...such as scanning software or password crackers.

ibid

A good metric must (3/5)

Jaquith

Contain units of measure

- Time, dollars, or some numerical scale should be included
- ...not just, say, “green,” “yellow” or “red” risks.

ibid

A good metric must (4/5)

Jaquith

Be expressed as a number

- Give the results as a percentage, ratio or some other kind of actual measurement
- ...not just subjective opinions such as "low risk" or "high priority."

ibid

A good metric must (5/5)

Jaquith

Be contextually specific

- Relevant do decision making
- Does not result in an aggravated "Uh, this helps me how?"

ibid

& Answer RM challenges

- How secure am I?
- Am I better than this time last year?
- Am I spending the right amount of $$?
- How do I compare to my peers?
- What risk transfer options do I have?

Those challenges were risk management challenges. How secure am I leads directly to looking at that very fact over the timeline so that you can say whether you made progress in the previous interval. The right amount of dollars, as we shall see, is like Goldilock's porridge -- it can be too hot, too cold, of just right. If you are rather different from your peers either you're crazy or they are; care to know which? And, of course, if someone will take you risk but let you keep your reward, then by all means let them. You can cry all the way to the bank, if you must.

Bad metrics

- Can't be consistent
- Aren't cheap to gather
- Missing numbers, units, or both
- Result in a shrug

A good metric invites use; a bad metric is dismissed.

Kinds of numbers

Types

- Continuous – infinite number of values
  - What is your weight?
- Discrete – countable number of values
  - How many children do you have

(and most raw security data are counted)

Discrete variables are almost always counts of things. Continuous variables are usually measurements. Not a big distinction in terms of decision support but does have some implications in how to handle statistics if you have enough data to do statistics.

Scaling

- Nominal: named, no numeric meaning
- Ordinal: lined up, $a<b<c \Rightarrow a<c$
- Interval: 22-14 = 8 = 42-34
- Ratio: 8/4 = 2 = 4/2

Numbers for comparison purposes come in several flavors called “scales” (scales in the sense of music, not in the sense of having 100,000 computers on your internal network). The four scales are nominal scale, ordinal scale, interval scale, and ratio scale.

Nominal scale

- Classification data, *e.g.*, {M, F}
- No ordering, *e.g.*, M > F meaningless
- Arbitrary labels, *e.g.*, {M, F} or {1, 0}

Standard definitions follow, but this text copied from http://www.stat.sfu.ca/~cschwarz/Stat-301/Handouts/node5.html

In a nominal ("name only") scale, the categories are nothing but categories and the labels are nothing but labels.

Ordinal scale

- Ordered data, differences between values are not comparable, *e.g.*,
  - Political parties on left to right spectrum given labels 0, 1, 2
  - Rank on a scale of 1..5 your degree of satisfaction ("Likert Scale")
- Good enough for decision support

In an ordinal scale, the categories and the labels are still just categories and labels, but now there is an unambiguous sense that there is a natural sequence to them as otherwise arbitrary as they are. That this is a weak condition is not an insult; it is a benefit in that weak conditions bias outcomes less than strong conditions and this weak condition is good enough to produce the decision support we seek.

Likert scale example: http://www.icbl.hw.ac.uk/ltdi/cookbook/info_likert_scale/

Interval scale

- Ordered data, constant scale, but no natural zero
- Differences make sense, but ratios do not, *e.g.*,
  - 30°-20° = 20°-10°, but
  - 20°/10° is not twice as hot

An interval scale as fixed intervals between items of like kind. In the example, we see that thirty degrees is ten degress warmer than twenty degrees just as twenty degrees is tend degrees warmer than ten degrees. That does not mean that twenty degress is twice as hot as ten degress. The intervals are subject to addition and subtraction but not to multiplication or division.

Ratio scale

- Ordered data, constant scale, has a natural zero
- Ratios do matter, *e.g.*,
  - Height, weight, age, length

Whereas an interval scale can support addition and subtraction, now we get multiplication and division. Praise be to the inventor of “zero.”

Table/Progression

	order?	constant scale?	natural zero?
nominal	no		
ordinal	yes	no	
interval	yes	yes	no
ratio	yes	yes	yes

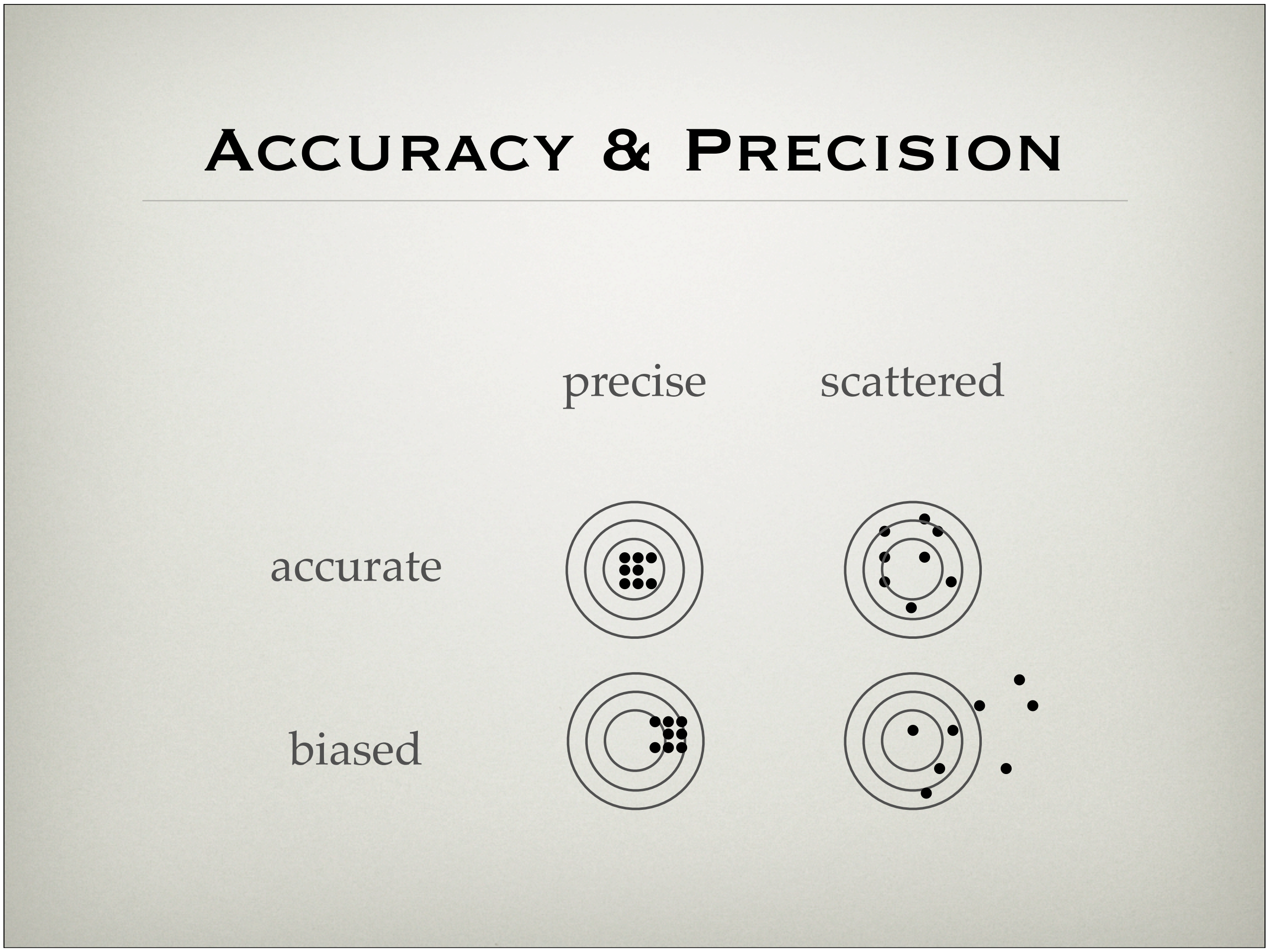
Perhaps easier to visualize using this table.

Accuracy & Precision

- Accuracy is unbiased; errors don’t lead
- Precision is narrowness, even if skewed

Accuracy is the “freedom from mistake or error” while precision is “exactness and thus limited” (per the Merriam Webster Unabridged).

As used here (and not just here) accuracy is free from misleading bias while precision is free from fuzzy lack of clarity. For trend analysis and then decision making, the consistency of which accuracy speaks is the more important.

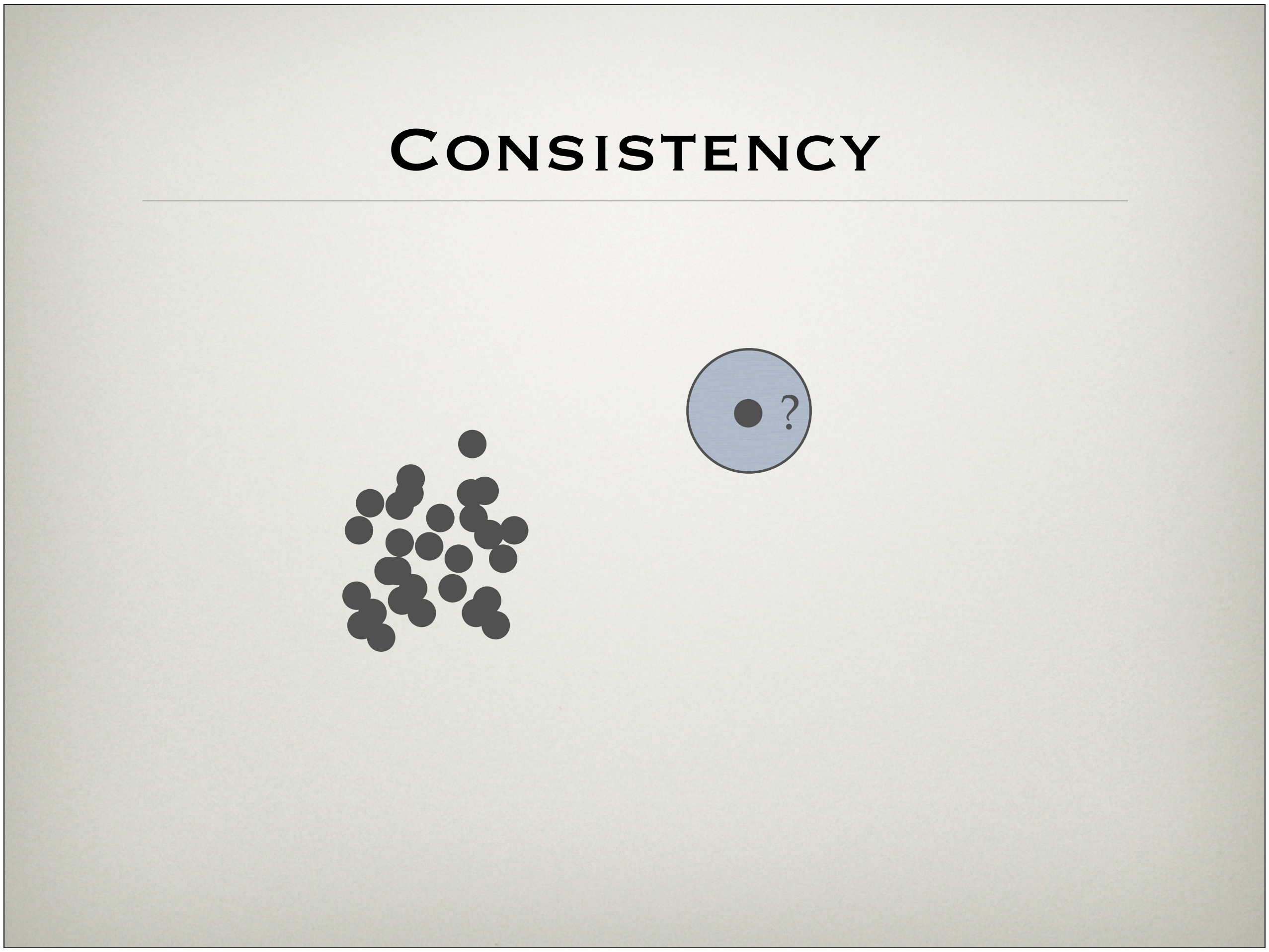


Visual depiction instead.

Consistent

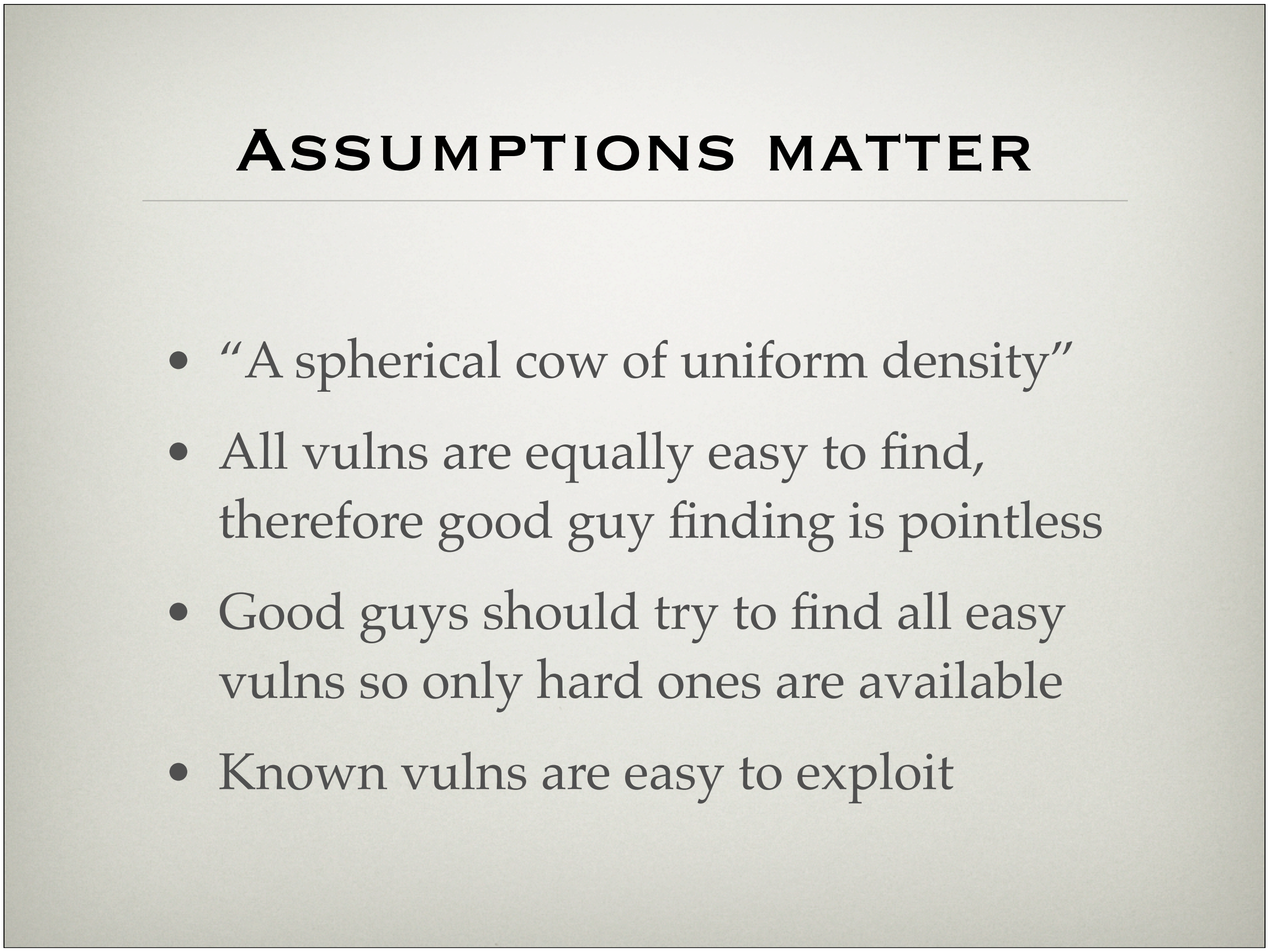
- Precision and accuracy are nice, but for trend analysis, consistent is all you need
- Consistent is a weaker condition of constant bias and constant scatter, improving the estimate as more data becomes available

Put differently, “consistent estimators” converge on the true value. That they converge means precision increases and that they converge to the true value means that bias is not present.



Consistency is a state of measurement where additional measurements enable greater precision. Sampling a random number does not have this characteristic, but otherwise a consistent measure is possible and preferred.

Note that unless there is a consistency to your measure you will not be able to identify the outliers as the sense of outlier -- different in substantial ways to the central tendency -- cannot be identified if there is no consistency to the underlying measure.



In any place where numbers and estimates around them are used and made, there will be assumptions. The first is a classic in the physics lecture -- "Assume a spherical cow of uniform density." But in our world we can make leaps of equal reach; we might assume that all vulnerabilities are equally easy to find which, as a matter of logic, would tell Good Guys to stop trying to find them since they could not reduce supply for Bad Guys as the Bad Guys would just find different vulns equivalently easily. The contrasting assumption, that the difficulty of finding vulns is ordinal, not nominal, would encourage Good Guys to find as many of the easy to find vulns as possible as the remaining vulns would be more costly for the Bad Guys to find. Closer to fact than to assumption, we might still wonder whether or not that a vuln is known means that the vuln is exploitable. That could be an hypothesis and we could then seek data to confirm or deny it.

Normalization

- Technically correct meaning: to make an unknown distribution look like the Normal distribution
- Colloquial meaning: to make it possible to realistically compare two different distributions

Several times in the following material we’ll normalize a number. Normalization means to bring varied measures to a common scale, often a dimensionless relative scale, for the purpose of making the disparate comparable. It is to make “normal” which usually means relative to a well understood base state. In the precise meaning of mathematical statistics, it is to convert a distribution to something resembling a “Normal” distribution. In general work, it is less precisely means to make multiple distributions comparable.

Normalization, cont.

- As Normal as possible: $Z = \dfrac{X - \mu}{\sigma}$
  - Subtract the mean to mutually center data sets
  - Divide by standard deviation to get common scale
- Many other options; use with care

You need not subtract the mean if you want, say, each distribution to start with an initial value at a common time. You may not divide by the standard deviation if you want to highlight, say, rates of change (where dividing each distribution by its own median might be a better idea). As with most general topics, this is as far as we can go without examples.

Testing

Testing

truth / test	+	-
+	a	b
-	c	d

a = true positives
b = false positives
c = false negatives
d = true negatives
accuracy = (a+d)/t

true positives
  a = positive testers who have disease
true negatives
  d = negative testers who are without disease
false positives
  b = positive testers who are without disease
false negatives
  c = negative testers who have disease

And the accuracy of the test is the number right as a fraction of all tested, i.e., (a+d)/t

Testing

test \ truth	+	-	
+	a	b	a+b
-	c	d	c+d
	a+c	b+d	t

(a+c)/t = prevalence
a/(a+c) = sensitivity (recall)
d/(b+d) = specificity
a/(a+b) = predictive value positive (precision)
d/(c+d) = predictive value negative

prevalence
(a+c)/t = fraction of population that has disease
sensitivity
a/(a+c) = what fraction of those with disease test positive
specificity
d/(b+d) = what fraction of those without disease test negative
predictive value positive
a/(a+b) = what fraction of positive tests have disease
predictive value negative
d/(c+d) = what fraction of negative tests are without disease

Interpretation

- Get a negative for highly sensitive test?
  - Likely a true negative ("rule out")
- Get a positive for highly specific test?
  - Likely a true positive ("rule in")

http://www.poems.msu.edu/EBM/Diagnosis/Diagnosis.htm, specifically, http://www.poems.msu.edu/EBM/Diagnosis/SensSpec.htm

Interpretation, cont.

- Predictive value depends on the prevalence of the condition (rows)
- Sensitivity, specificity do not (columns)

∴ We can describe how good the test is without knowing prevalence, but we cannot say what an individual test result predicts without prevalence estimates.

This is important: while the specificity and sensitivity of a test are characteristics of the test independent of the population on which that test is used, the predictive values positive and negative are dependent on those populations. Put differently, a test of constant specificity and constant sensitivity will have a different predictive value when the true rates of disease change. In the table shown before, this is whether you are working with columns or rows. Go back and look.

Choosing tests

- If false negative is serious,
  Then favor sensitivity (& treat false pos)
- If false positive is serious,
  Then favor specificity (& lose false neg)

One might favor sensitivity if the treatment is painless and cheap but the disease is serious; re-imaging a virtual machine when there is <u>any</u> doubt about its integrity, say.

One might favor specificity if the treatment is painful or costly while the disease is mild; complete emergency patch rollout to correct a spelling error, say.

Testing in security

- AVS signature finding
- IDS anomaly identification
- Automated code analyses
- Firewall packet inspection
- Patch management performance
- ... and on and on ...

There are many testing and testing-like activities in security, as listed here in hint form.

Multi-stage testing

- Maximizes cost-effectiveness

  Stage 1 “screen”: dirt cheap, high sensitivity

  Stage 2 “confirm”: expensive, high specificity

- Combination has higher specificity at expense of sensitivity, *e.g.*, policing the blood supply for HIV

A multi-stage test is one where testing is done sequentially. As such, the results of any one stage are conditional on the results of the previous stage. This can have significant economic impact.

As a rule of thumb, you cannot increase sensitivity and specificity at the same time. For a resonably rare disease, non-cases will strongly outnumber cases hence a negative test result is more likely. Working with that, you have a first stage that confirms negative status, i.e., it is highly sensitive resulting in false positives but, in turn, low false negatives. In other words, the first test releases as many as possible (and no more) from further work-up. The second stage wants no false negatives so it is highly specific and, if indeed most subjects were rejected in the first stage, that second stage test can be quite expensive (and definitive).

MS-testing in security

- Many examples
  - Router logs (S1) post-processed by log-analysis tools (S2)
  - Anomaly detection (S1) reviewed by human eyes (S2)
  - SIGINT traffic analysis (S1) to sieve which crypto is worth breaking (S2)

We have many examples of multi-stage testing in security, as outlined here. There are others, and it might pay us to look harder at multi-stage security testing.

Worked example

- Situation
  - 10^6 to screen
  - Prevalence = 1%
- Problem: Identify those 10,000

A million people, lines of code, or whatever to screen and the idea that 1% of them are the problem -- just which 1% is now our problem.

Testing

test \ truth	+	-	
+			
-			
	10,000	990,000	10^6

1% = prevalence
10^6 = population size

This is what we know.

S1: rule-out negatives

truth / test	+	-	
+	*99.99%*	*10%*	
-	*.01%*	*90%*	
	10,000	990,000	10^6

99.99% = sensitivity (our focus here)
90% = specificity
10^6 = test population

We begin with a test that is sensitive, i.e., which misses few true positives at the cost of a meaningful number of false positives, and for which a negative result is not enormously meaningful.

S1: RULE-OUT NEGATIVES

test \ truth	+	-	
+	9,999	99,000	108,999
-	1	891,000	~~891,001~~
	10,000	990,000	10^6

99.99% = sensitivity, and so 1 false negative

90% = specificity, and so 99,000 false positives

.999999 = predictive value negative

.09 = predictive value positive

So, with a sensitivity of 99.99% we get one false negative and we can forget about 89% of the population. At this prevalence, a negative result is .999999 likely to be correct (odds of 1 in 10,000,000 of being wrong).

S2: RULE-IN POSITIVES

test \ truth	+	-	
+	*90%*	*.01%*	
-	*10%*	*99.99%*	
	9,999	99,000	108,999

90%	=	sensitivity (no longer the focus)
99.99%	=	specificity (now our focus)
108,999	=	test population

Now we take just the remainder and use a second that has, for convenience, the reverse sensitivity and specificity.

S2: rule-in positives

test \ truth	+	-	
+	8,999	10	9,009
-	1,000	98,990	~~99,990~~
	9,999	99,000	108,999

90%	=	sensitivity, and so 1,000 false negatives
99.99%	=	specificity, and so 10 false positives
.9989	=	predictive value positive
.99	=	predictive value negative

This test gets us 10 false positives and 1,000 false negatives, but as a matter of management we ignore any negative results

S1|S2: overall

test \ truth	+	-	
+	8,999	10	9,009
-	1,001	989,990	~~990,991~~
	10,000	990,000	10^6

89.99% = sensitivity with 10 false positives
99.999% = specificity with 1,001 false negatives
.9989 = predictive value positive
.99899 = predictive value negative

We now have a compound result in which the predictive value of the compound test is high both for positives and for negatives, which is arguably what we would want though debate may ensue on the downstream cost of a false negative versus a false positive.

Cost effectiveness

S1 @ 30¢/test ⇒ \$0.3M & 99,001 wrong

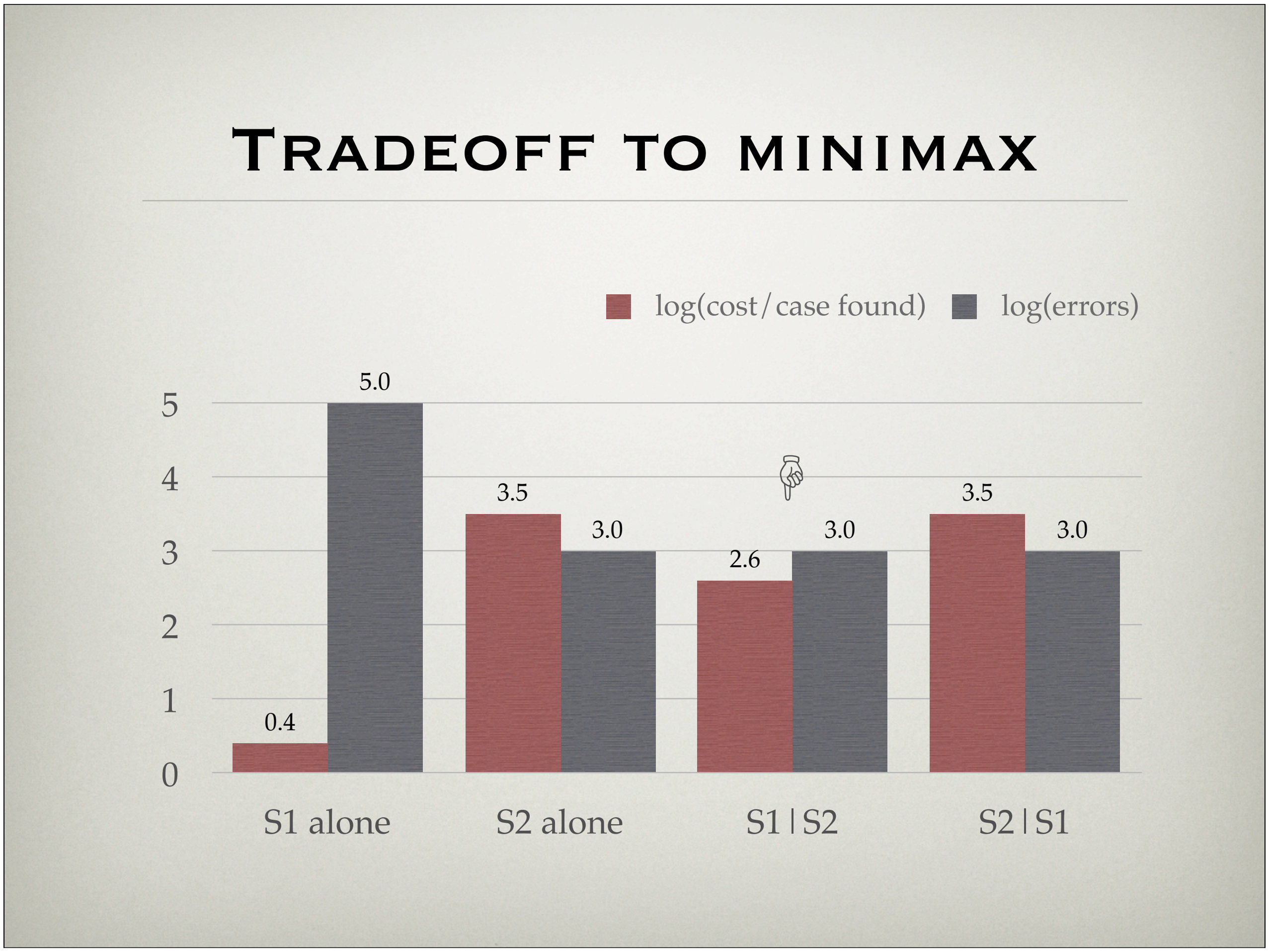
S2 @ \$30/test ⇒ \$30M & 1,099 wrong

☞ **S1**|**S2** ⇒ \$3.6M & 1,011 wrong

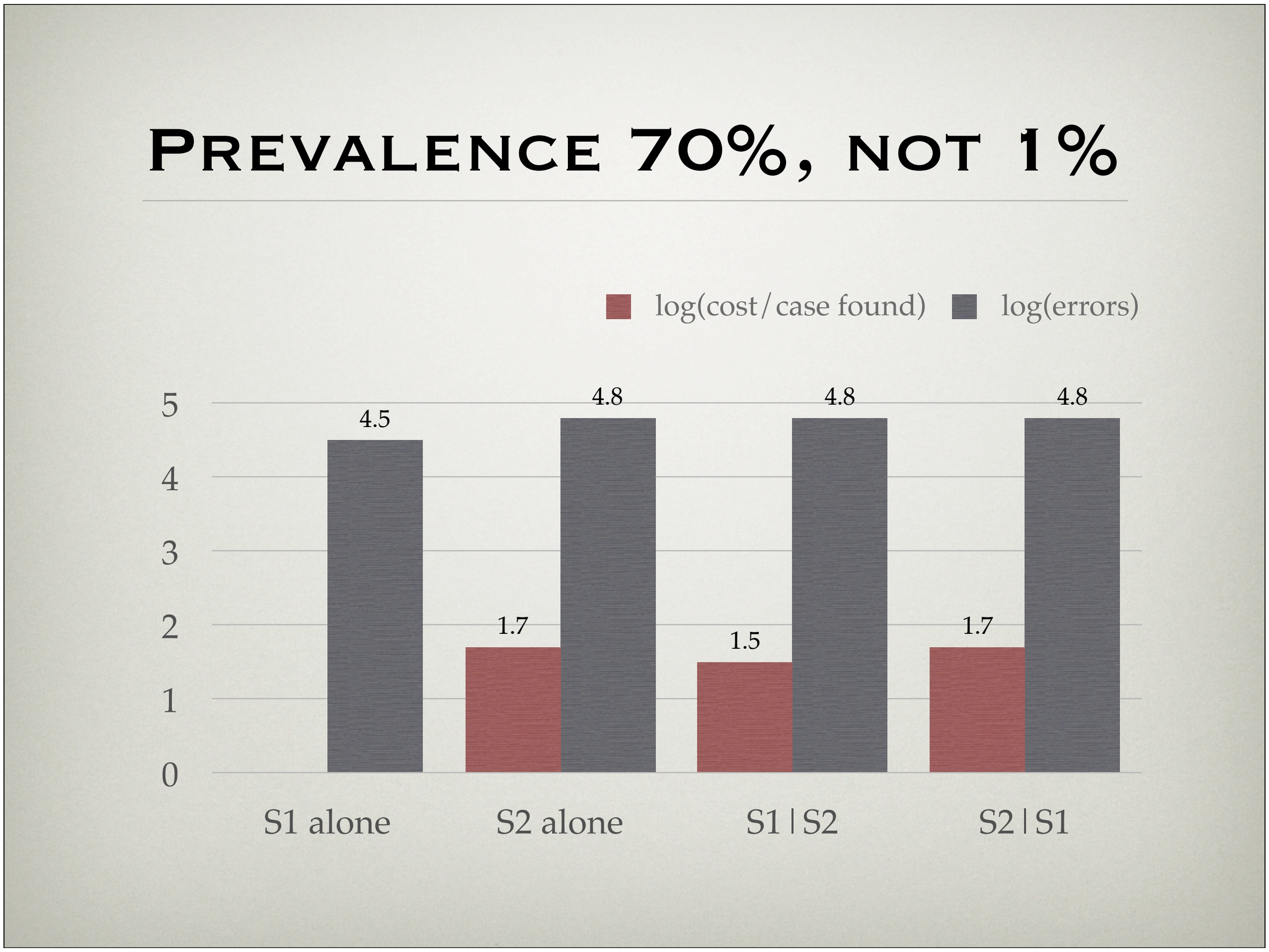
S2|**S1** ⇒ \$30M & 1,011 wrong

If we did S1 alone at a cost of 30¢ per test, we'd spend \$300,000 and have nearly 100,000 incorrect results. Similarly, if we did S2 alone at a cost of \$30 per test, we'd spend \$30,000,000 and have over 1,000 incorrect results.

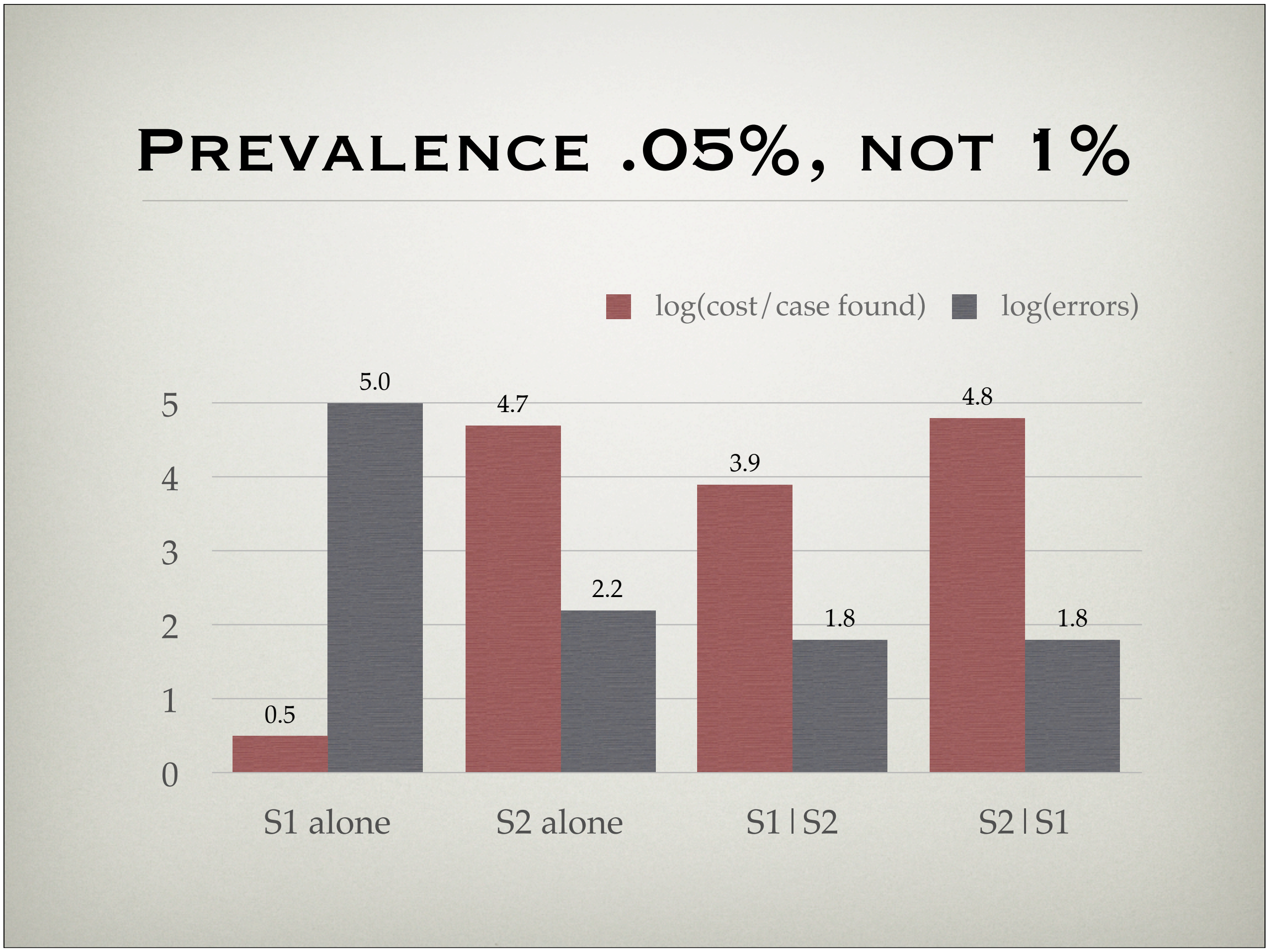
Whether we do S1 or S2 first and the other second, we still get just over 1,000 incorrect results but with S1 first we spend \$3,600,000 rather than \$30,000,000.



As with all of security, you are doing tradeoffs. In this case it is unduly easy to get that minimax solution: the maximal favorable result at minimal cost. However, as you can appreciate, a very large differential between the general cost effects of false positives versus false negatives.



The effect of converting from 1% prevalence to 70% prevalence makes this a harder decision.



The effect of converting from 1% prevalence to .05% prevalence further highlights the choices to be made, and how they are sensitive to the prevalence of the disease.

Calibration?

- So what if your test gives a value, not a binary result?
- How do you decide what is a positive and what is a negative (since, after all, you have to make a decision)?

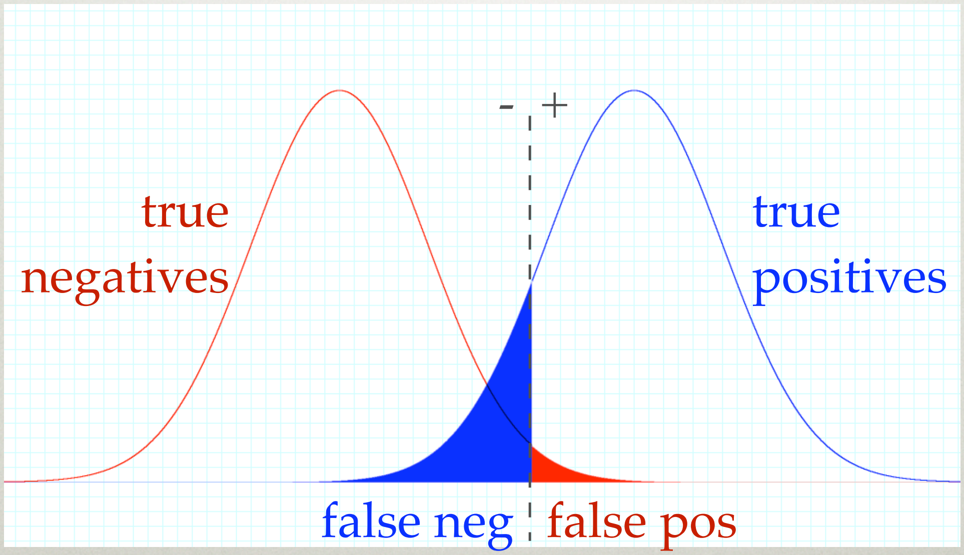
Do I treat or not treat? Rebuild or not rebuild? – Questions like that need binary decisions even if the test I am doing is returning not a "Yes/No" response but rather a value. This leads to a different class of problems.

Calibration: ROC

- "Receiver Operating Characteristic"
  - Analysis of the test itself
  - You adjust the sensitivity *v* specificity
  - Simple idea: what is cutuoff value?
- Works for any kind of data

This is all about setting cutoffs on continuous scales so that above value X you say "Yes" and below X you say "No" – but what is the right value of X?

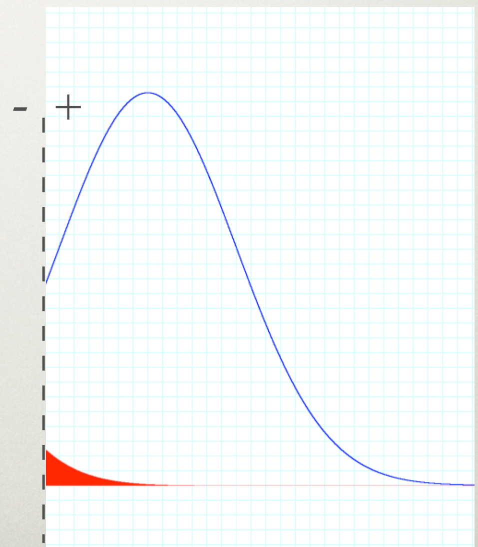
Pick a cutoff value



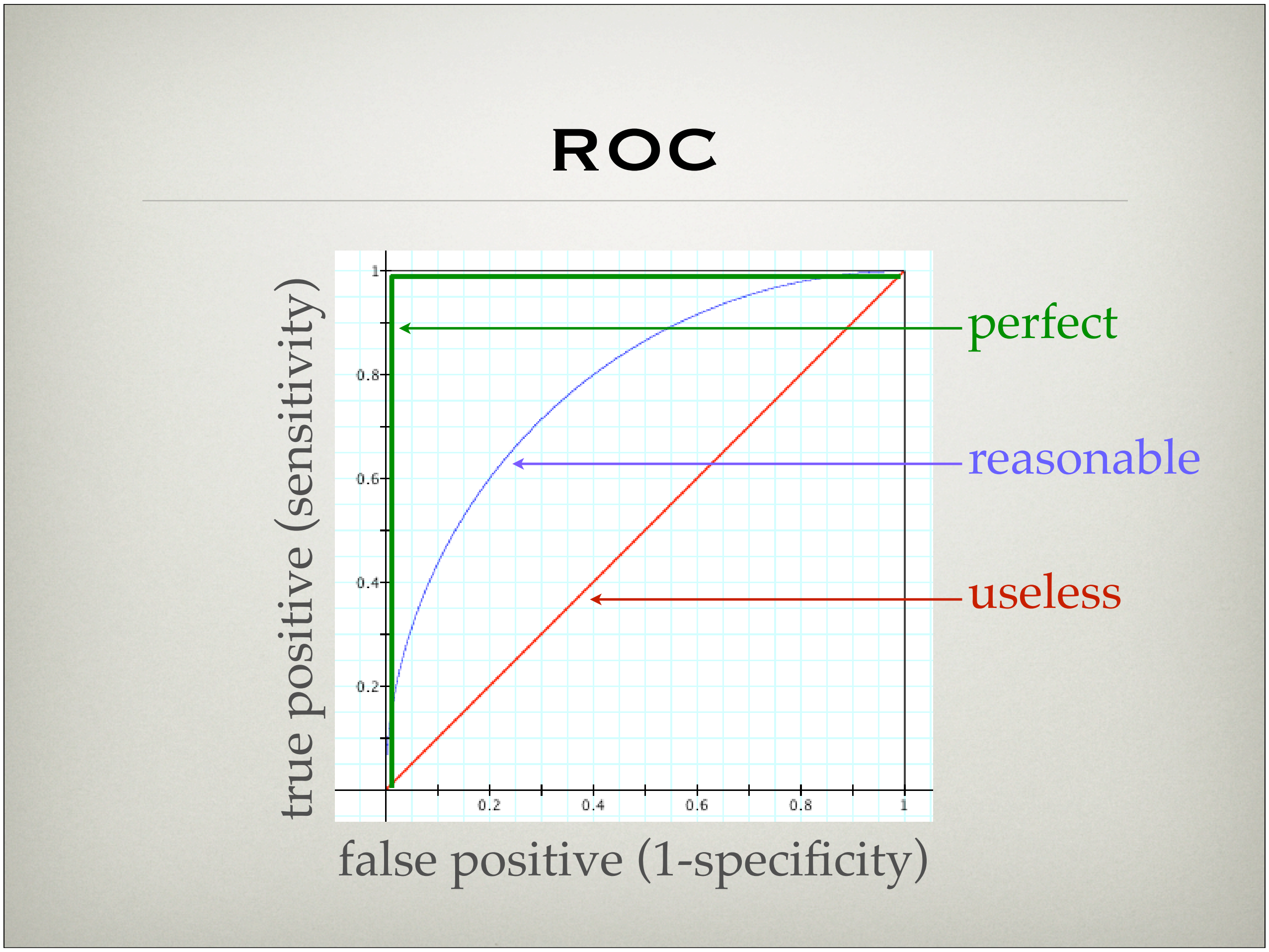
In this example, the true negatives get a test result that is centered around a low value while the true positives get a test result that is centered around a higher value. The problem is that the tails of the distributions overlap, so it is not possible to avoid some false positives or negatives. You have to pick a threshhold value below which your test calls “negative” and above which it calls “posiive.”

Pick a cutoff value

$$ROC = \frac{\text{true positives}}{\text{false positives}}$$



The ROC is the ratio of true positives to false positives as you vary the cutoff point.



Sensitivity, the fraction of true positives that test positive, is traded off against specificity, the fraction of true negatives that test negative. A useless test is one where you cannot increase either the specificity or the sensitivity, and effectively where whether a test returns positive or negative has no meaning. A perfect test has no false positives and no false negatives. A reasonable test is one where

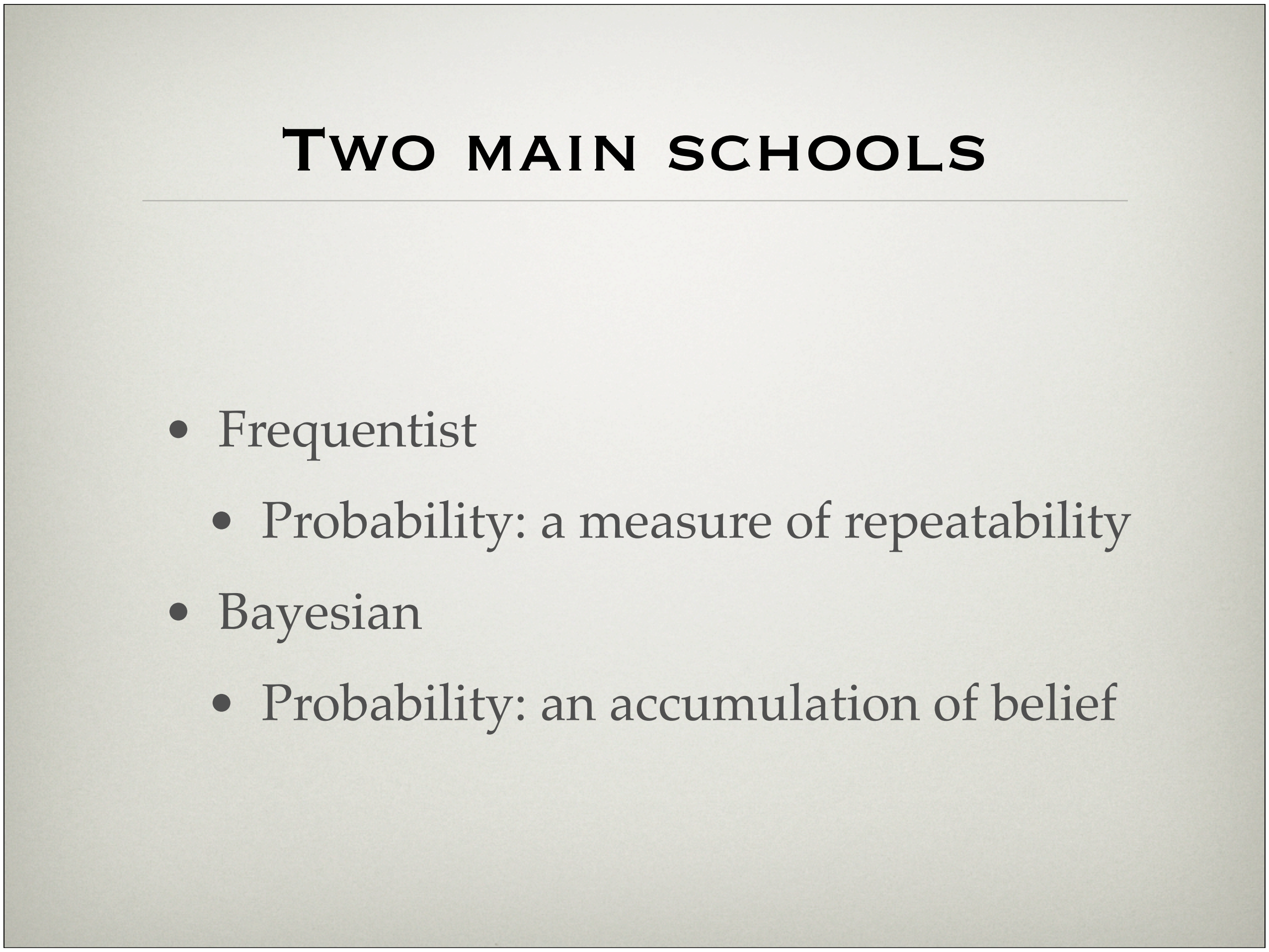
ROC, why bother?

- Allows calibration of a practical test by a definitive but expensive one
- Lets you make tradeoffs based on whether false positives or false negatives are more critical
- Works with any data (Likert scales, say)

For more, see the Wikipedia entry and follow the links, or just search, or just look up Wilcoxon and/or Mann-Whitney testing in a statistics text if you can read those.

Also see http://glue.umd.edu/~acardena/Papers/Oakland06.pdf and ../AAAI06-255.pdf

Probability



These are not just idle differences; they really lead somewhere (or not, since the arguments around them have been going on amongst specialists for decades now).

In any case, the frequentist focus is on frequency, i.e., that a coin flip has equal probability of heads or tails is something that is the observable result of flipping that coin over and over and over. By contrast, the Bayesian viewpoint is that you have come to believe that the heads and the tails are of equal odds and you will make your decisions not on the sceptical repetition of innumerable coin flips but on the belief you have, a belief possibly well bolstered by experience but not derived purely from that experience.

Bayesian

$$\Pr(H|d) = \frac{\Pr(d|H) \times \Pr(H)}{\Pr(d)}$$

$$\Pr(H|d) = \text{belief in H if data d obtains}$$
$$\Pr(d|H) = \text{probability of d if H is true}$$
$$\Pr(H) = \text{prior probability of H}$$
$$\Pr(d) = \text{prior probability of d}$$

The math is not otherwise necessary, but this is Bayes Rule. It allows you to reverse a set of probabilities you might know to get at one you might not know and thus change your belief in the Hypothesis based on the data as observed.

Bayesian's strength

- Makes use of non-repeatable data
- Directly valuable for decision making
- Handles multiple hypotheses easily

$$\Pr(d) = \sum_i \Pr(d|H_i) \times \Pr(H_i)$$

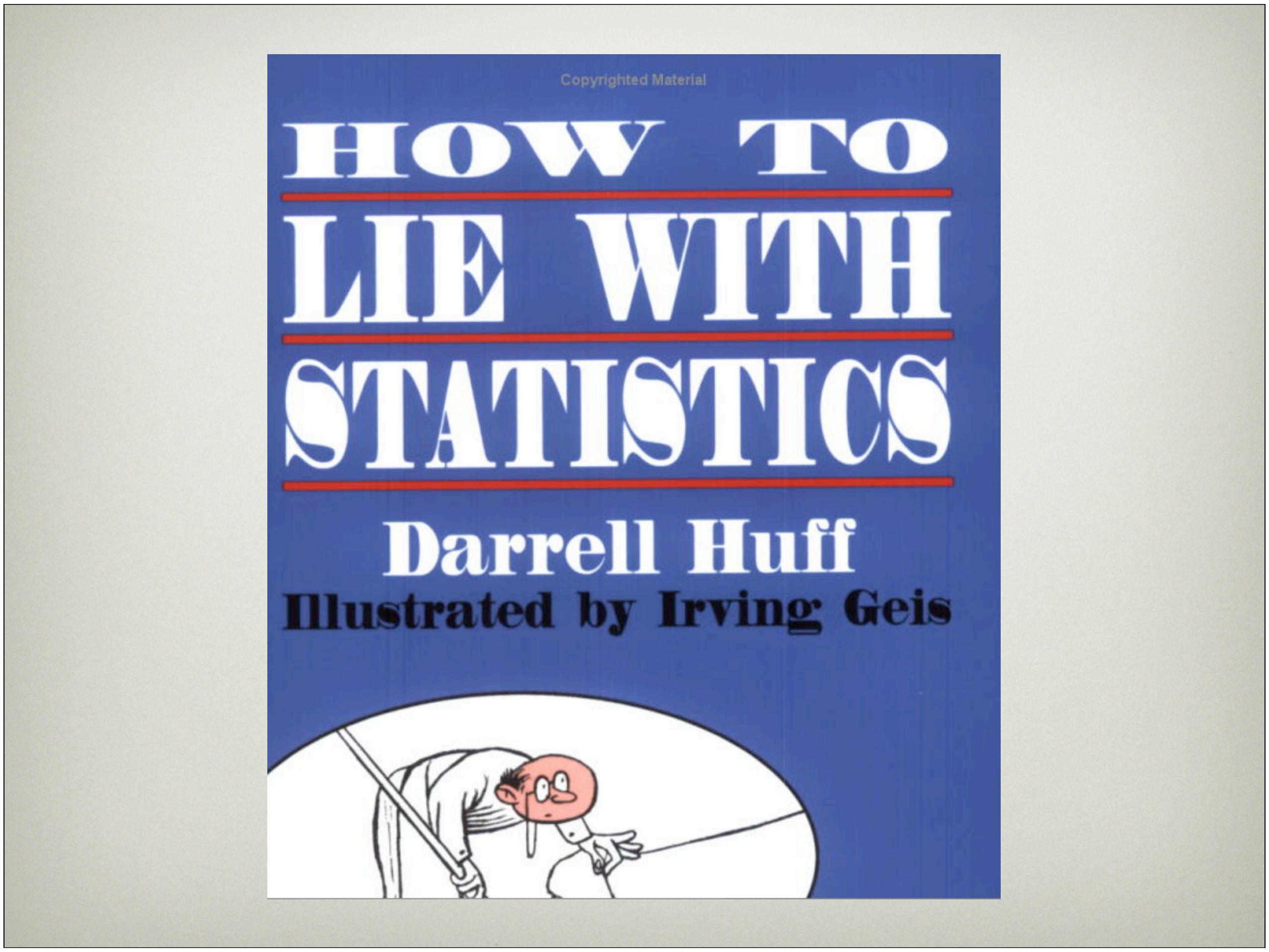
Because of its focus on belief, non-repeatable data that would, by it is non-repeatability, be out of scope for the frequentist can be used by the Bayesian. This is good for real world decision making but not so good for basic scientific research.

Also in Bayesianism's list of advantages is that you can handle multiple hypotheses at the same time.

Out of scope

- No time here to cover statistics and probability or simulation for that matter
- Already in use in security in*, e.g.,* Bayesian spam filters

Bayesian and frequentists methods are implicitly in use all the time in security, but the Bayesians have the upper hand. It is widely acknowledged, for example, that Bayesian spam filters are the ones to use. In some sense, you don't want to have to do repetitive spam experiments -- you get plenty of those as it is -- and you certainly do not want to rely on repeatability when the opponent is trying to make every spam e-mail different from every other (thus requiring you to learn, if learn you will, from non-repeatable data).

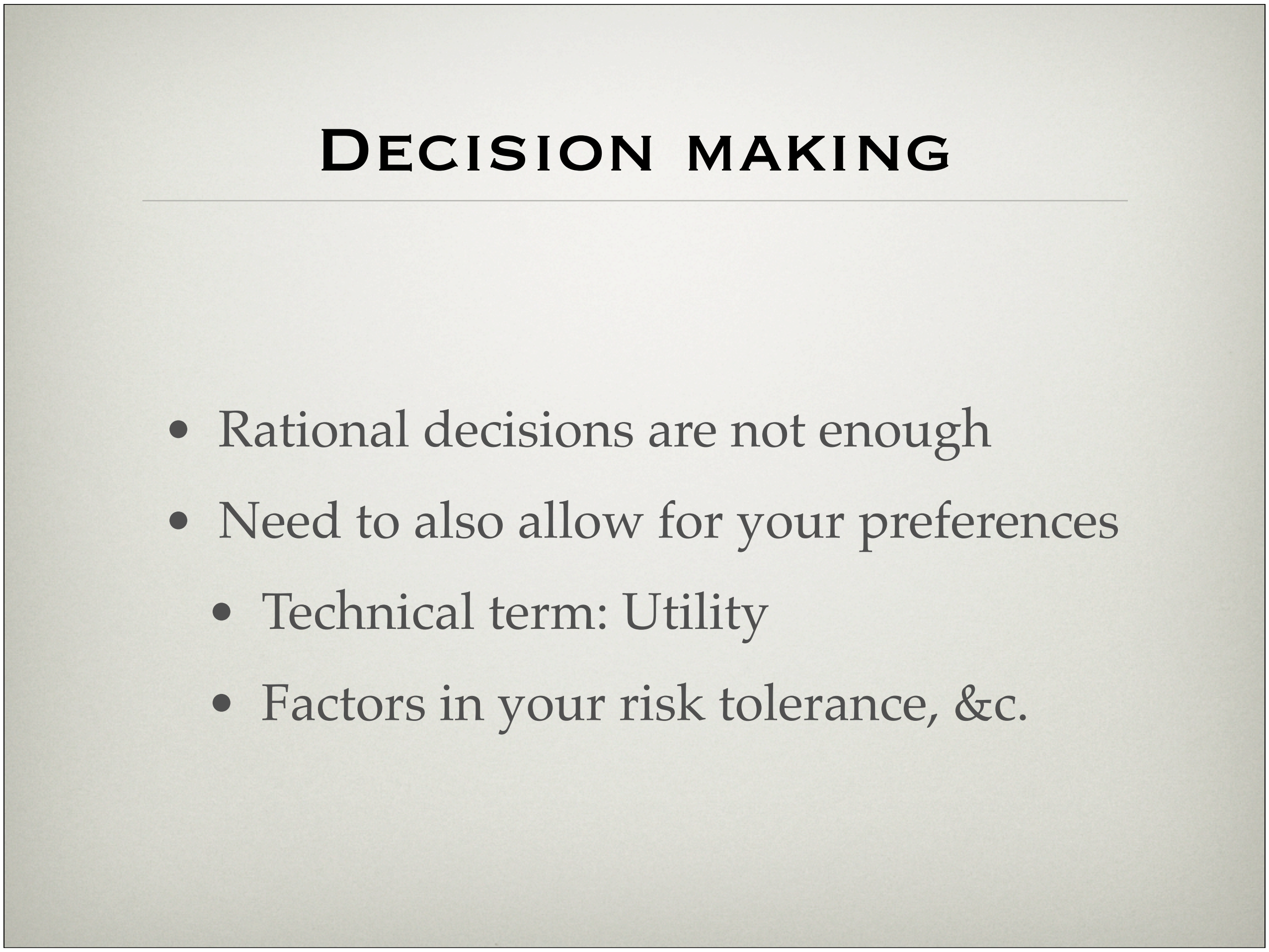


Huff D : How to Lie with Statistics, W. W. Norton & Company, 1952. (reissue edition, September 1993)

I cannot recommend this enough. You might also like:

Paulos JA : A Mathematician Reads the Newspaper, Anchor, March, 1996.

Decision making



Decision making is at its best when it is rational. Or is it?

Actually, it isn't. What is needed is a way to represent your preferences, not just the cold hard facts, since your preferences are the only thing that matters for part of the decision making process. The technical term for this is utility, and we are now talking about "decision analysis."

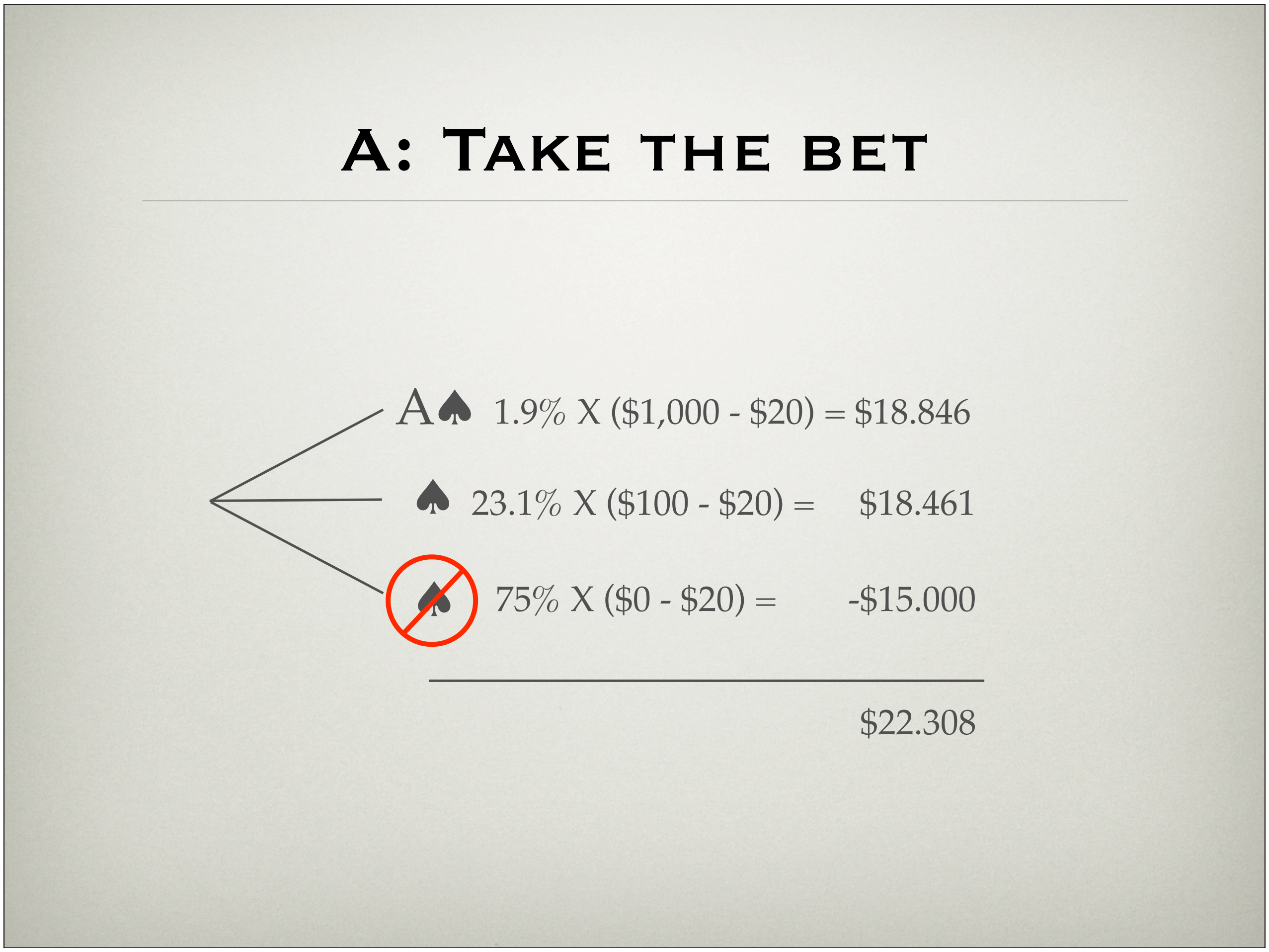
A gamble

- Draw one card from a fair deck
  - $1,000 for A♠
  - $100 for any other ♠
  - -0- otherwise

Q: Play or don't play for $20?

You pay me $20. I put a deck of (fair) cards on the table. You draw one card. If it is the Ace of Spades, I give you $1,000. If it is not the Ace but is still a Spade, I give you $100. Otherwise I give you nothing.

Want to play?



All in all, you should indeed play as factored over every possible outcome you should expect to get $22.308 on average per game. This is, in other words, a purely frequentist view -- you play the game over and over and over and eventually the winnings will average $22.308 per round.

Why is that simple?

- We took the "expected value"

$$Pr(X)*X = E(X)$$

- This works when decision making is rational

This worked simply because we calculated the "expected value" by simply multiplying the probability of any particular outcome by the value of that outcome, discounting that outcome, if you will, by the chance of getting it.

This is precisely what a purely rational decision making process looks like.

Not always so simple

- Flip a fair coin
  - Heads, I give you $50,000
  - Tails, you give me $50,000
- Completely fair, yet few will play

So, let's play a different game. It is completely, completely fair. I flip a coin and one of us gives the other one $50,000. Despite being fair, few will play.

A little harder

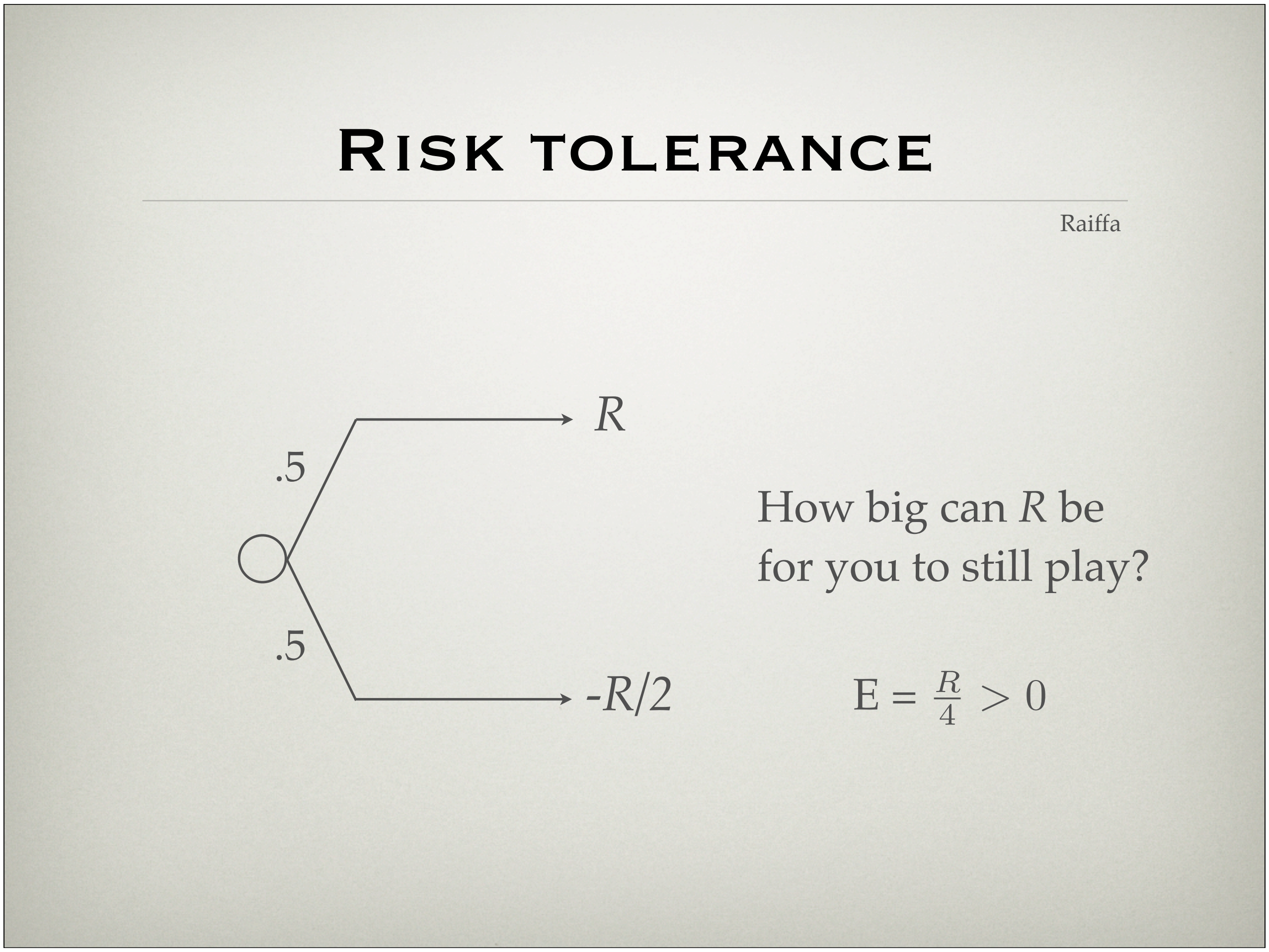
- Flip that fair coin again
  - Heads, I still give you $50,000
  - Tails, you give me $40,000
- Odds in your favor; still few will play

Few will play even when the odds favor them rather strongly -- the expected value of this game is $5,000 positive for you. Nevertheless, except at high-roller tables in Las Vegas, bets of this sort don't come up very often.

Risk aversion

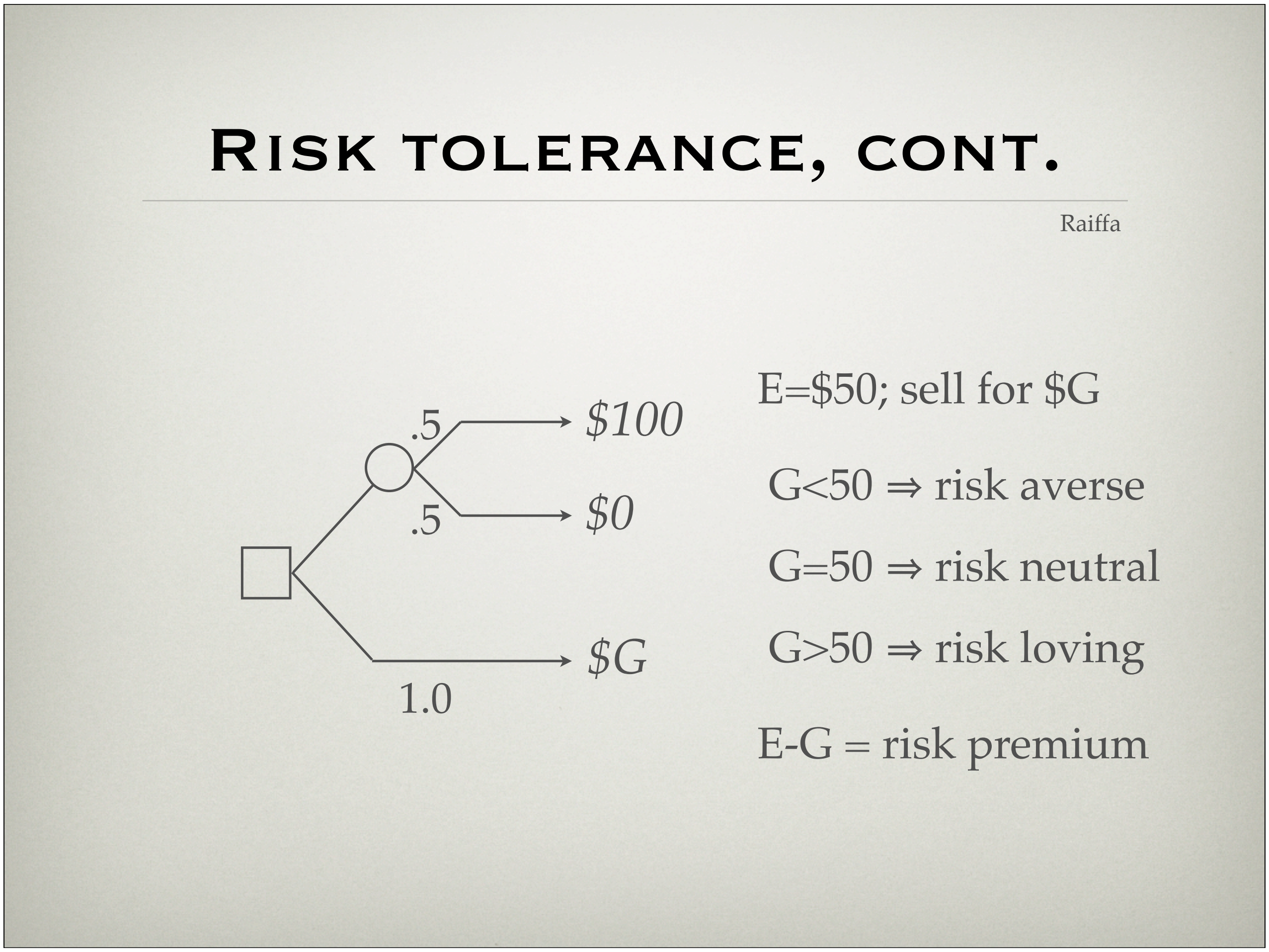
- The issue is risk aversion
- You can be averse to risks you know
- You can be averse to risks you don’t
- You can be blind to one or another risk

The issue is “risk aversion” -- the desire to avoid risk. As it says, that can be risks which you know and understand or it can be otherwise.



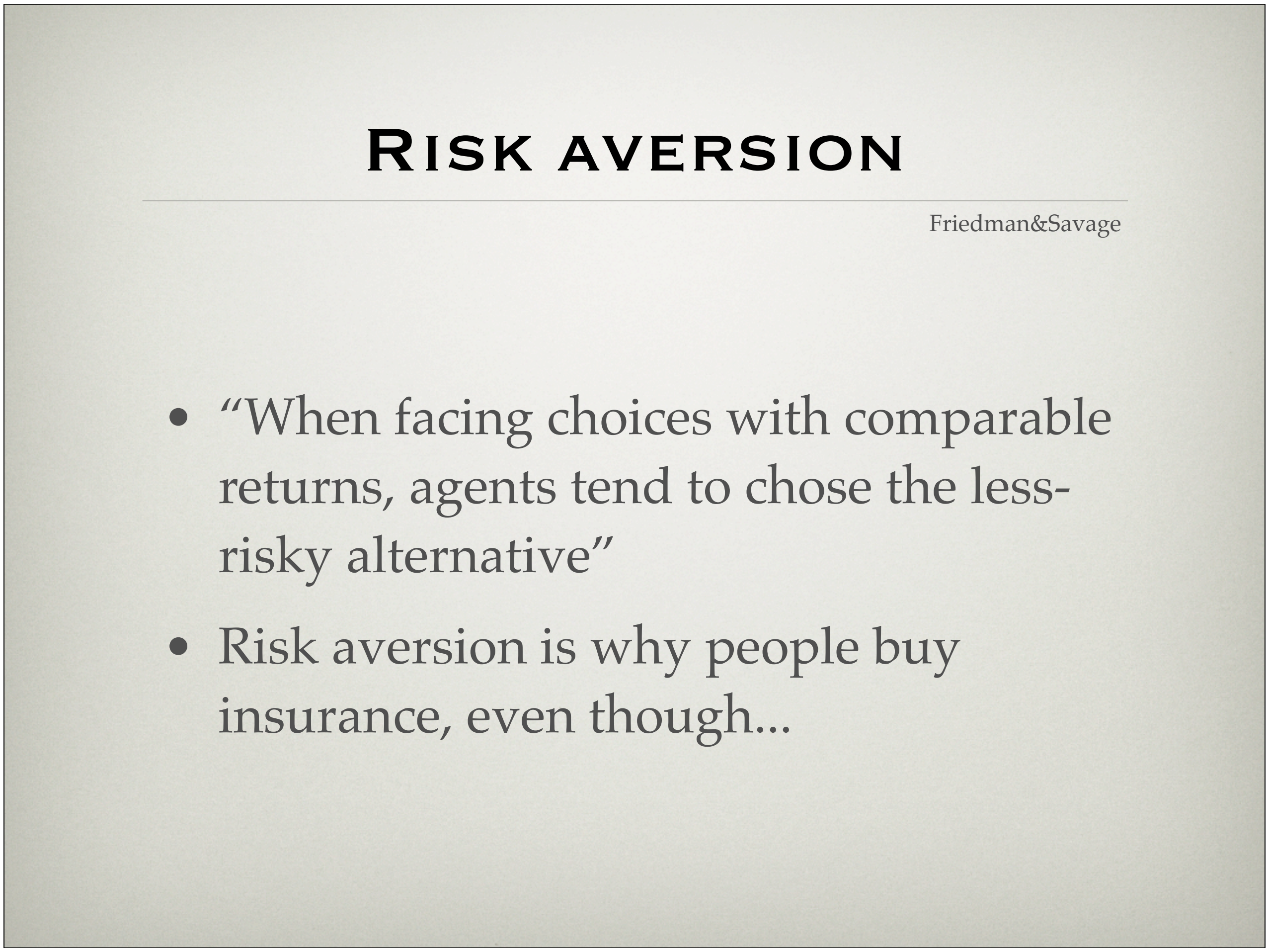
Howard Raiffa used this in his (Harvard Business School) lectures. A fair toss of a fair coin and you get $R for heads and you lose $R/2 for tails; how big an R will you play for? The expected value is positive throughout at $R/4.

Raiffa H: _Decision Analysis_, Addison-Wesley, 1968.



A fair toss of a fair coin (the circle symbol) and you win $100 or you lose nothing. This is certainly in your favor, with an expected value E of $50.

What amount of money would you accept to skip the coin toss? If it is, say, $40 you will note that $40 is less than $50 and your risk aversion actually has a name ("risk premium") and a numeric value ($10). At $50 excactly, you are exactly risk neutral. Over $50 and you prefer the risk to the payoff -- you are risk seeking.



There is nothing wrong with risk aversion; it is perfectly natural and it explains why people buy insurance.

Friedman M & Savage LP: "The Utility Analysis of Choices Involving Risk," Journal of Political Economy, v56 pp279-304, 1948.

Risk aversion, cont.

Baker

Type of gamble	Expected return
Individual health insurance	You lose 40% of each bet
Large group health insurance	You lose 10% of each bet
Roulette	You lose 5% of each bet

If you buy individual health insurance (which, incidentally, it is illegal to sell in Massachusetts), you will generally collect $60 in benefits for every $100 in premiums you pay. Large groups tend to suppress the loss, generally to the tune of losing $10 per $100 of premium. Roulette is a better bet than either.

[In the American version, there are 38 slots of which 2 (0,00) are for the house and the payout is 35-to-1: (35 - 37)/38 x 100 = -5.26% The European version has only on (0) which is for the house, and it, too, pays at 35-to-1: (35 - 36)/37 x 100 = -2.70%]

Baker S: course notes, Univ. of South Carolina, 2001, at http://hspm.sph.sc.edu/COURSES/ECON/RiskA/RiskA.html

Term of art

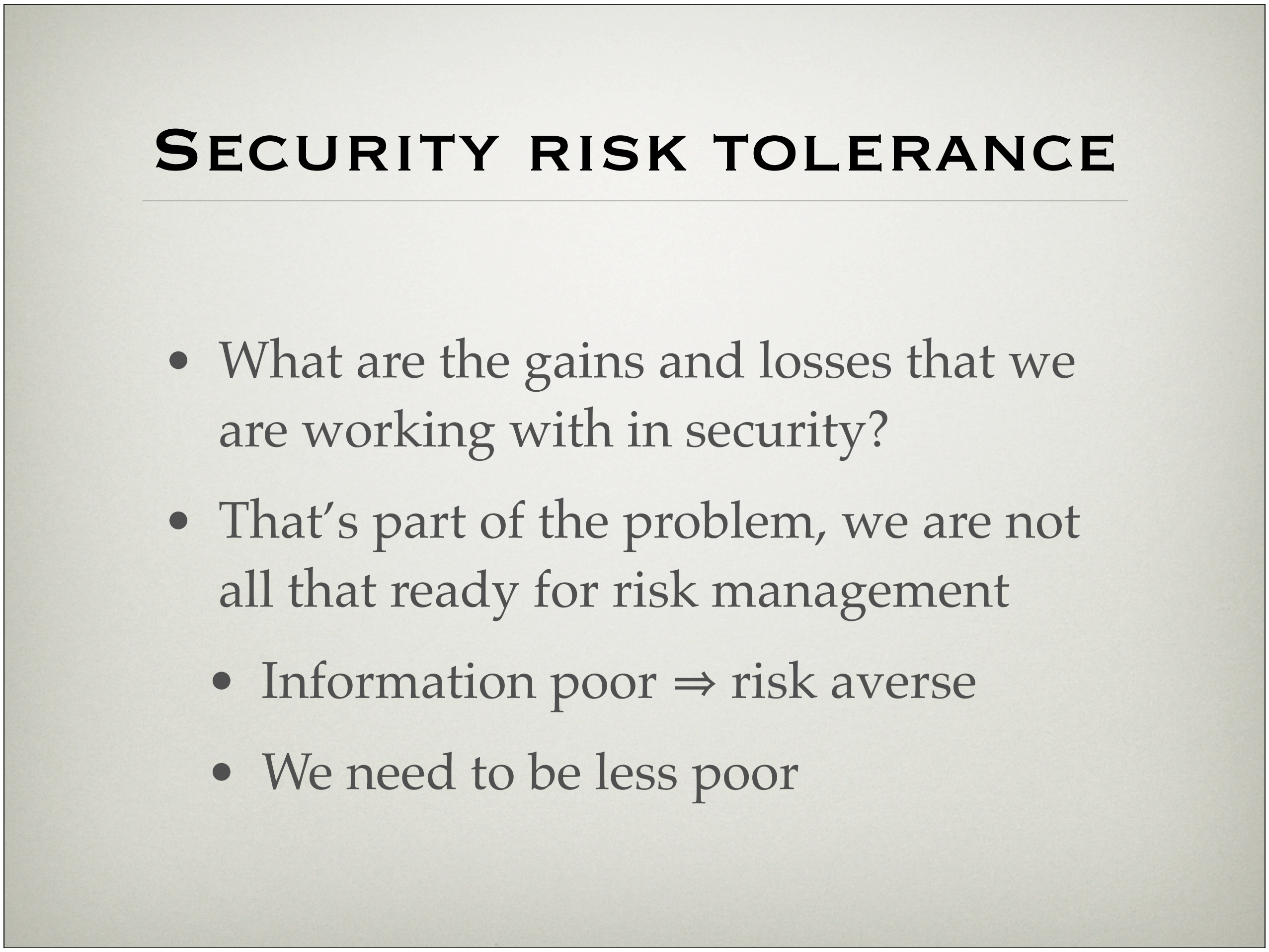
- Read annual report for BC/BS
  - The “Medical Loss Ratio” ≅ 90%
  - 1-*MLR* ≅ 10% ≡ your expected loss
- Unless you lose, BC/BS goes bankrupt
  - Your risk aversion makes the business

If you read the annual report for a health insurer like Blue Cross Blue Shield, find the phrase “medical loss ratio” which is the converse of your loss ratio; 90% for them is 10% for you, but if it wasn’t your loss they wouldn’t still be in business.

Use in security

- Can we understand the risk tolerances of our clients?
- Can we make security decisions based on risk pricing?
- Do we see risk-aversion or -seeking?

So, let's ask, does this have use in security? Is the risk tolerance of our clientele something we can gauge or, to the point, make decisions on or with? Are our consumers risk-averse (as they probably are or they would not hire us)?



And what are those tolerances for risk?

We don't know. In fact, we don't know much. Here's one of the crystalline truths in this entire lecture: Being more information poor makes you more risk averse.

Gets to the heart of it

- Risk aversion is why a General Counsel will say that if you could have lost data you have act as if you did
- Risk aversion is why some keep no records
- What is your reputation worth?

A bank in New York had a Chief Information Security Officer. This CISO wanted to invest in identity management. The system involved cost real money. The CISO got the money by asking what is essentially a risk aversion question: “This investment is worth it if the reputation capital of the firm is at least as much as one basis point of our market cap” (basis point = .01%). No officer of that bank was willing to bet the reputation of the firm as being worth less than .01% of the market value of the firm, and so the CISO got his identity management system. True story.

Mining what we have

Security metrics

- How do we get less information poor?
- What is our starting point?
- How do we measure success?
- What are the minimum assumptions?
- When does the game end?

So let's try this again... How do we, in fact, get to be less information poor given that we are starting from where we are, there are assumptions to be made, success is itself a measurement question and while this may be a life's work for some of us may it please the Court that the game at least end at some point.

"Laws of log analysis"

Ranum

1. Never keep more than you can conceive of possibly looking at
2. The number of times an uninteresting thing happens is an interesting thing
3. Keep everything you possibly can except for where you come into conflict with the First Law

Marcus Ranum's three laws of log analysis according to Marcus Ranum, found variously, e.g., http://seclists.org/lists/firewall-wizards/2004/Oct/0018.html

Ranum built the first firewall, which became DECSeal.

Where to start

- Steal techniques from other fields
- Mine data we have at least
- Make testable hypotheses
- Share data where we can

So, to be less information poor what do we do? We steal from other fields, that's what we do, and we will never again have as many security practitioners trained in other fields as we do today. While they are still present, let's mine their brains, let's examine data we already know how to collect before we tackle data we don't know how to collect and by all means let's put up some hypotheses to prove or disprove.

Virtuous theft, again

- Public Health
- Insurance
- Accelerated Failure Time testing
- Portfolio Management
- Physics

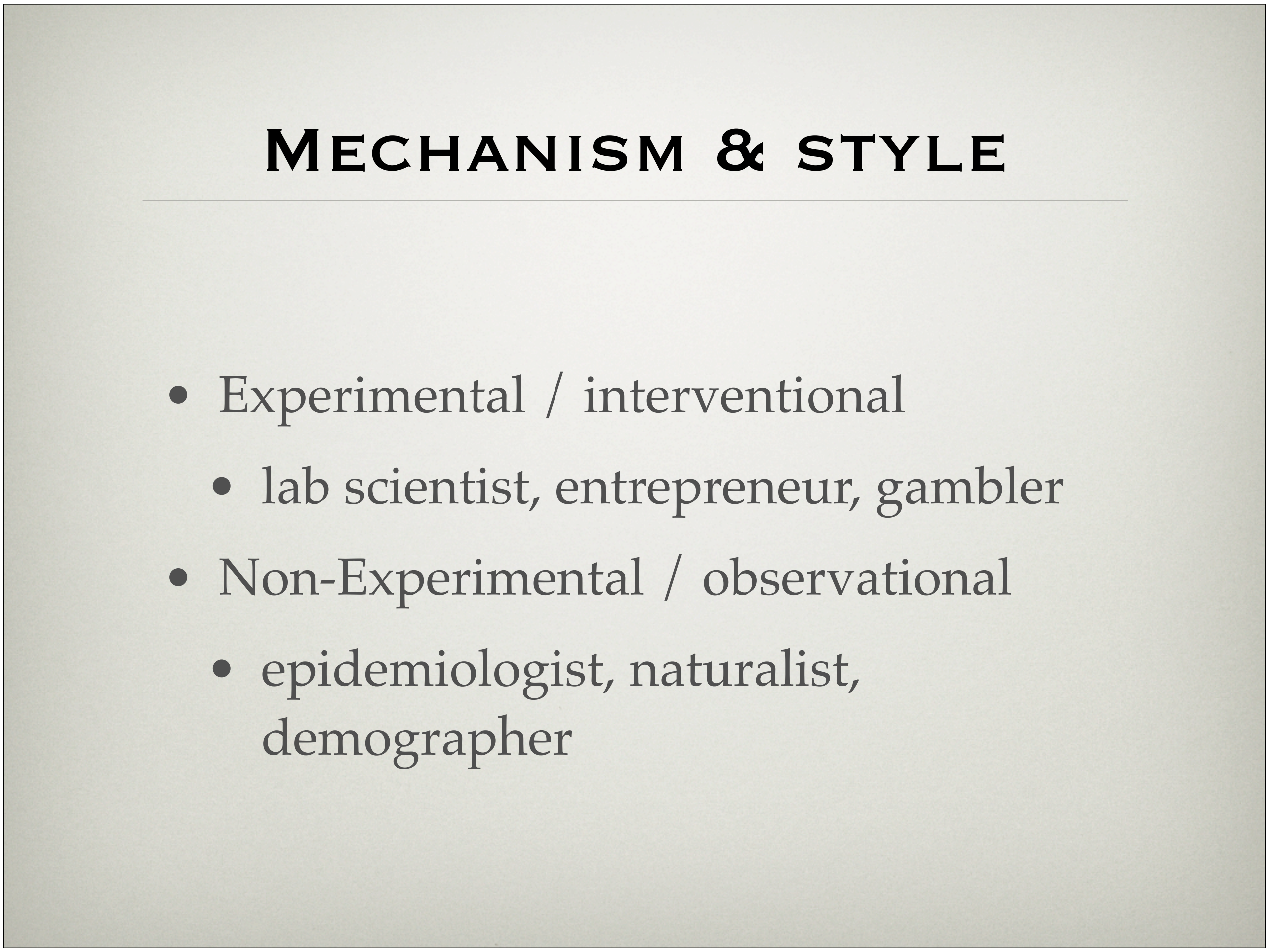
These are just examples of fields from which we can steal.

Public Health

Public Health

- Concern is disease spread, not disease
  - Does not require knowledge of causality if control is possible without it
  - Epidemiology "invented" by tracing cholera's transmission
- Focus on practical intervention, e.g., hygiene

Public health concerns it self with the spread of disease regardless of whether it is understood; indeed early efforts at community hygiene bore fruit before underlying biology could explain why they did. The classic case is On the Mode of Communication of Cholera, London, 1855, where John Snow concluded what it was that transmitted cholera without knowing what cholera was.



A lab scientist or an entrepreneur or a gambler intervenes in their world to see what happens. An epidemiologist, a naturalist, or a demographer doesn't intervene but instead observes.

Things to measure

- Disease processes
    - Incidence
    - Prevalence
- Generally: density, rate, proportion
- Models correct for biases

If it is public health we are measuring then the states and rates include incidence and prevalence. Where models are relevant is when absent models we have bias that is uncorrectable.

Incidence

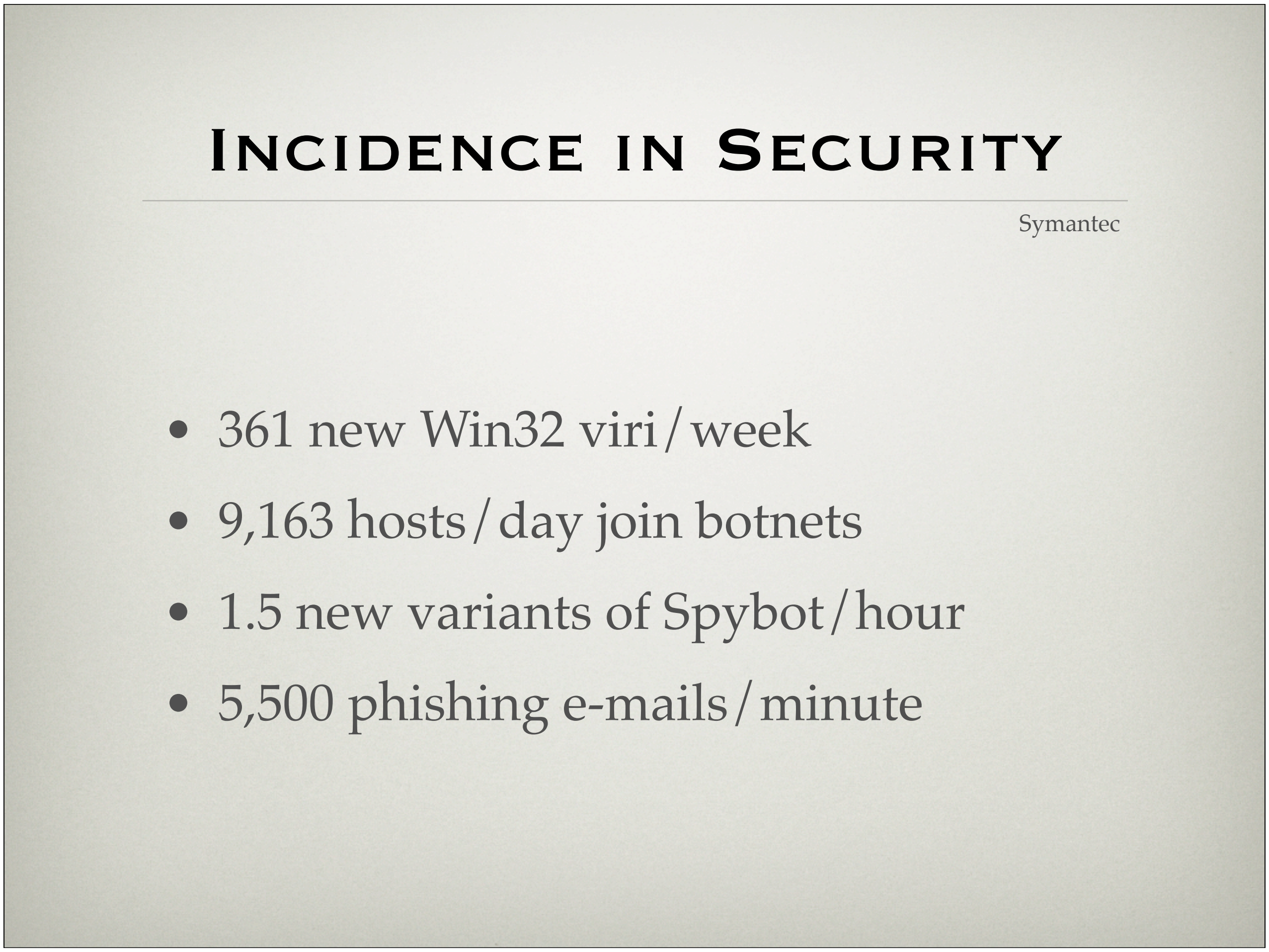
BMJ

The rate at which new cases occur in a population during a specified period:

$$I = \frac{\textit{number of new cases}}{\textit{person-years at risk}}$$

The definition of the British Medical Journal.

http://bmj.bmjjournals.com/epidem/epid.2.html



Symantec Threat Report IX, March, 2006 [requires registration, and, of course, it is not expressed as incidence the way we express it here.]

Prevalence

BMJ

The proportion of a population that are cases at a point in time:

$$P = \frac{\textit{number of cases}}{\textit{size of population}}$$

British Medical Journal, again.

http://bmj.bmjjournals.com/epidem/epid.2.html

Prevalence in Security

Symantec

- 85% of today's e-mail is spam
- 56% of today's spam originated in U.S.
- 1 of 119 last year's e-mails were phishes
- 50% of home computers are unpatched

ibid, except that 85% figure is from Message Labs rather than the Symantec report listed earlier

Relationships

*Prevalence = Incidence * Average Duration*

Bot fraction = Øwned/*t* * delay(patching)

How the arithmetic works, and an example from the security world: The fraction of hosts that are members of botnets is the rate at which hosts are 0wned times the delay in patching. Symantec's number is 30,000/day while Qualys' number is patching delay of 45 days, on average, hence the Bot-prevalence is 30,000 hosts/day incident times 45 days duration hence 1,350,000 hosts in botnets. This is probably a serious underestimate as detection and duration figures alike are likely low, but it illustrates the idea of the relationship between prevalence, incidence, and average duration.

Prevalence sampling

- You cannot have perfect knowledge
  - So you sample
    - You must be aware that you are
- How you sample has impact on what inferences you can draw

As with any sample, you are sampling because perfect knowledge is itself not possible and thus you must attend to how the sampling is done lest the sampling introduce a bias that you did not anticipate and thus do not correct for.

Sampling, 1/3

- Sampling fraction $f = N/M$
  - Sample of size N
  - Population of size M
- If M is unknown, so is f

The simple form of analyzing a sample.

Sampling, 2/3

- For much of what we care about, *M* and therefore *f* are indeed unknown
  - How many computers are there on the Internet?
  - How many privately held exploits are there?

In the security arena, we don't know the population size hence it is difficult to estimate the sampling fraction. The two sub-bullets are examples of how hard it is to know what the sampling fraction is. That is not insurmountable in and of itself, but it must be acknowledged in any analysis.

Sampling, 3/3

- If f is unknown but nevertheless stable, Then trend data can still be valid:

$$\text{trend}_N \propto \text{trend}_M$$

- Also connects incidence to prevalence

Trends that are true in the population will be equally true in the sample if the sampling fraction is stable.

Selection bias

- $N = Mf$, again where
  - N = reported incidence
  - M = true incidence
  - f = fraction of vulns reported
- If $f \neq$ *constant* (or just not predictable), Then $trend_N$ does not track $trend_M$

If, however, the sampling fraction is unstable (or, to be precise, not predictable) means that the trends evident in the sample may or may not track any trend truly present in the population at large.

Security selection bias

Symantec

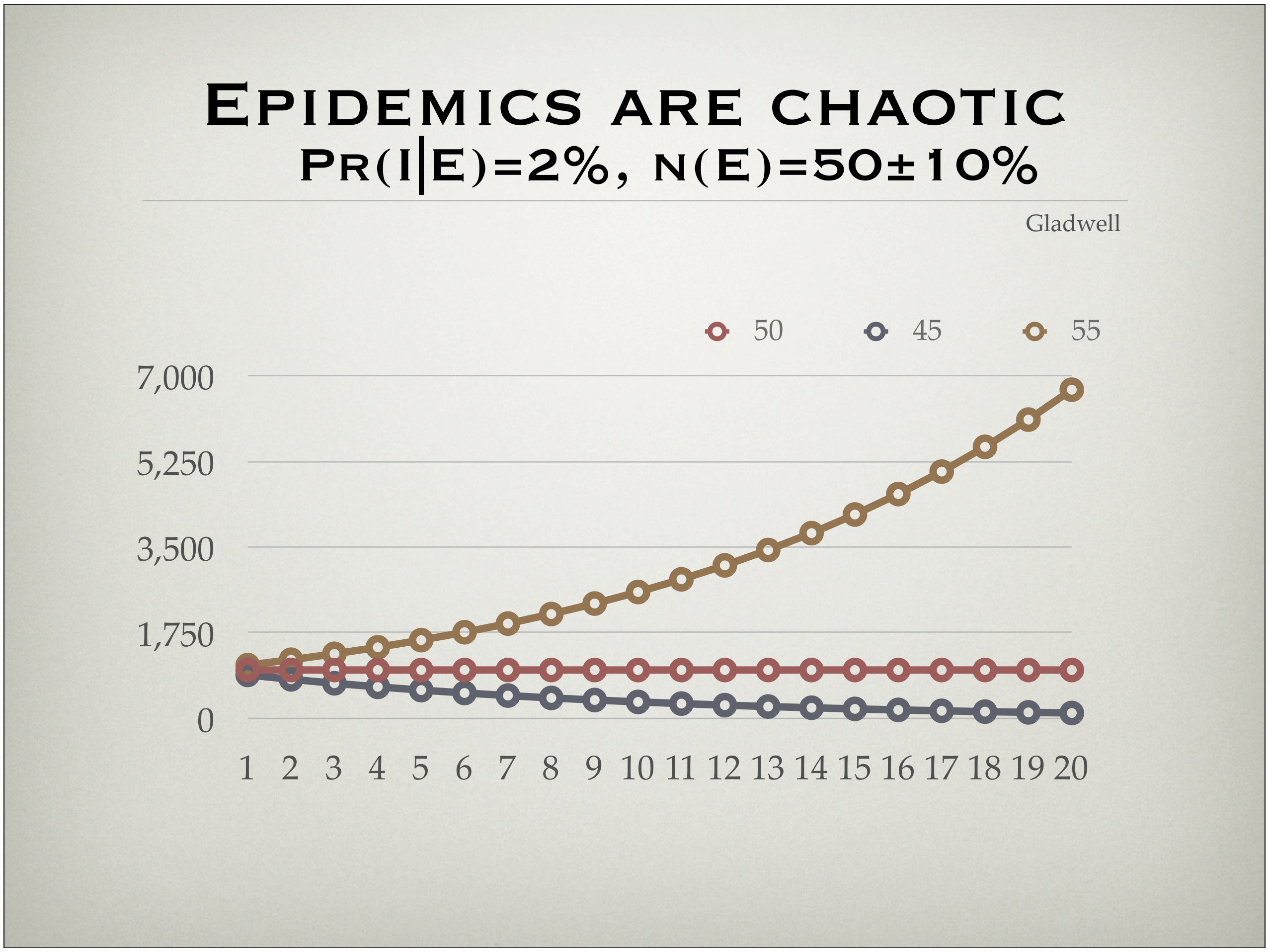
"Symantec speculates that while the number of publicly disclosed vulnerabilities could decrease, the window of exposure to potential threats could increase [if] details about vulnerabilities are held privately for greater periods of time."

This is Symantec saying precisely that a sampling fraction instability produces a possibility of misleading inference. If a changing sampling fraction is related to an effect of interest, then trends that are correlated with that effect will be likely misleading.

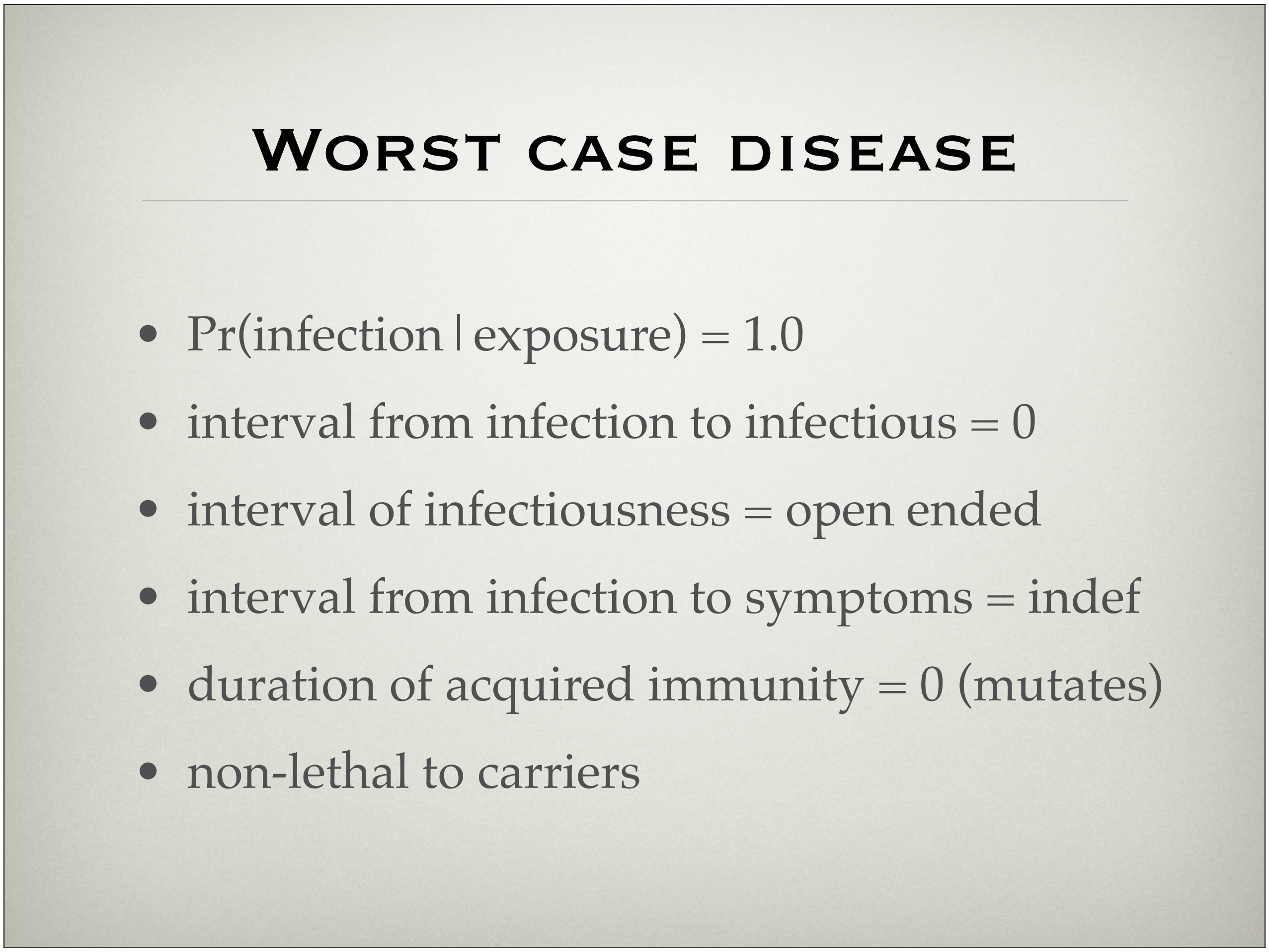
Epidemics

- Characteristics of infectious processes
  - Pr(infection | exposure)
  - interval from infection to infectious
  - duration of infectiousness
  - interval from infection to symptoms
  - duration of acquired immunity

The math for modeling epidemics is well developed, as is the math for accelerated failure time testing, actuarial science, portfolio management, and others. There is no need, and no time, to invent new science before progress can be made. Steal these skills, and do so while the senior practitioners in security still include people with these sort of skills learned elsewhere.



This is simply the example used in Malcolm Gladwell's The Tipping Point, Little Brown, 2000. It illustrates the chaotic nature of epidemics which is to say that small changes in initial conditions produce large changes in downstream values. This example is where the initial number of cases is 1,000, the probability of infection given exposure is 2%, the number of exposure events while infectious is 50 plus or minus 5 (10%), and the downstream shows that in only 20 days at -10% the disease will die out while in only 20 days at +10% the epidemic will be well underway.



If you were designing a pessimal disease, it would be perfectly transmissable (100% chance of getting the disease once exposed and no acquired immunity), no symptomatic sign of infection, and an instantaneous conversion from pre-infection to infectious (or from prey to predator, if you prefer).

The above describes worm propagation, or DDOS zombies, or the stockpiling of unannounced vulnerabilities.

Does the law have an answer for designer disease with pessimal characteristics and self-obscured authors? Is “terrorism” an appropriate model or is it more like mandatory seat belt laws?

Public health strategy

- Immunization efficiency
- Limit contacts with the infected
- Slow rates of transmission

To “steal” from Public Health, then, one might also look at strategy. Within that field, strategies around disease control tend to involve immunization efficiency, the limitation of contacts between infected and susceptible individuals, and taking steps that otherwise slow transmission rates.

Use in security

- Immunization efficiency
  - Patching trades reliability risk for penetration risk
  - Infections peak at one rev off current
  - Either keep up (immunize) or stay behind (diversify over version)

Infections peak at one revision off of current. Current revisions have fewer attacks probably because they are new; older revisions have few attacks probably because they are old. In other words, either keep up or fall behind.

Use in security, cont.

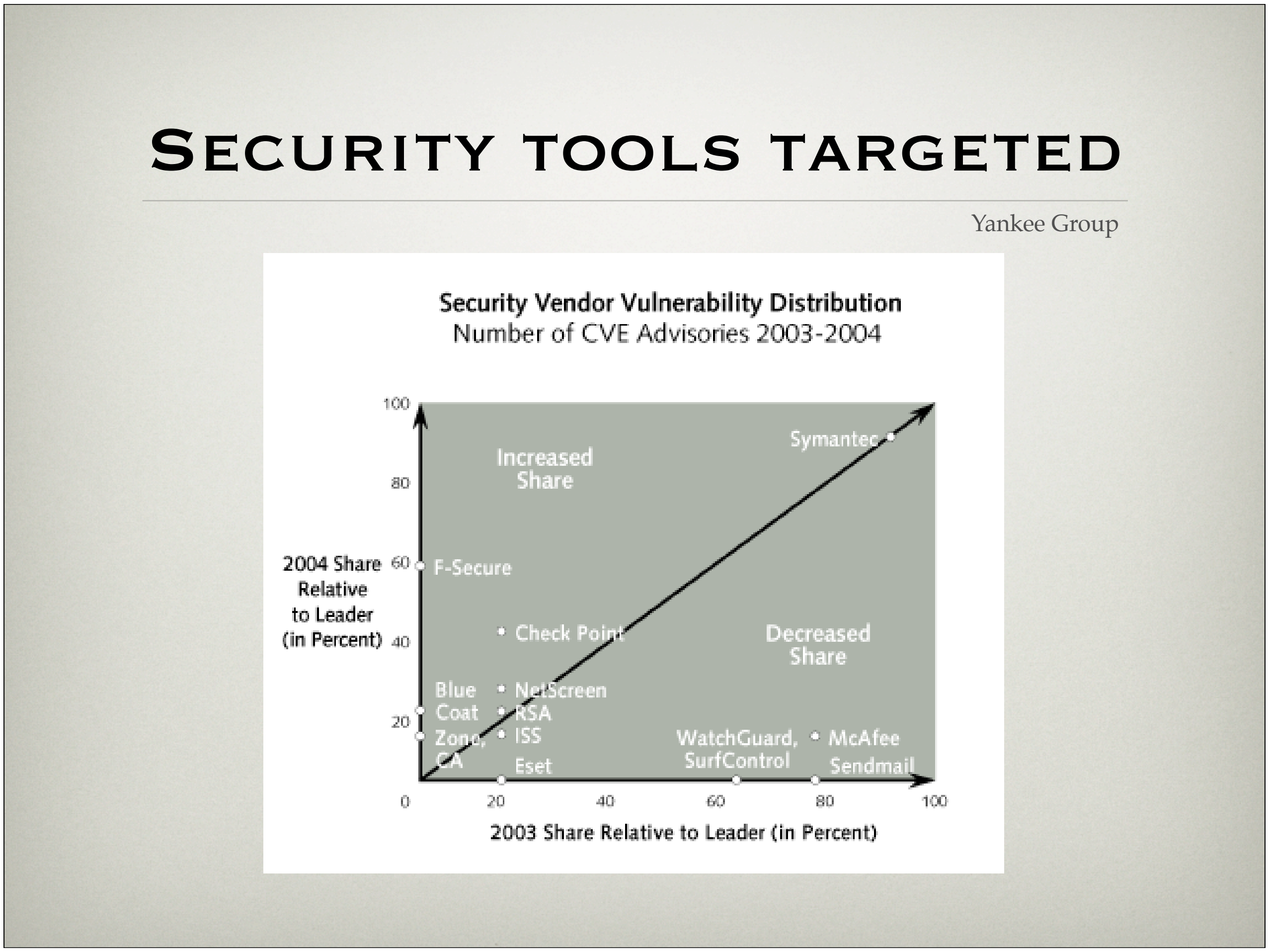
- In any case, measure your own patch latency against
  - Information at risk
  - Organizational features

A direct example of how a measurement tactic would exactly mirror the public health style; information at risk is the analog of susceptability and organization features relate to the level of contact between the infected and the susceptible.

Use in security, cont.

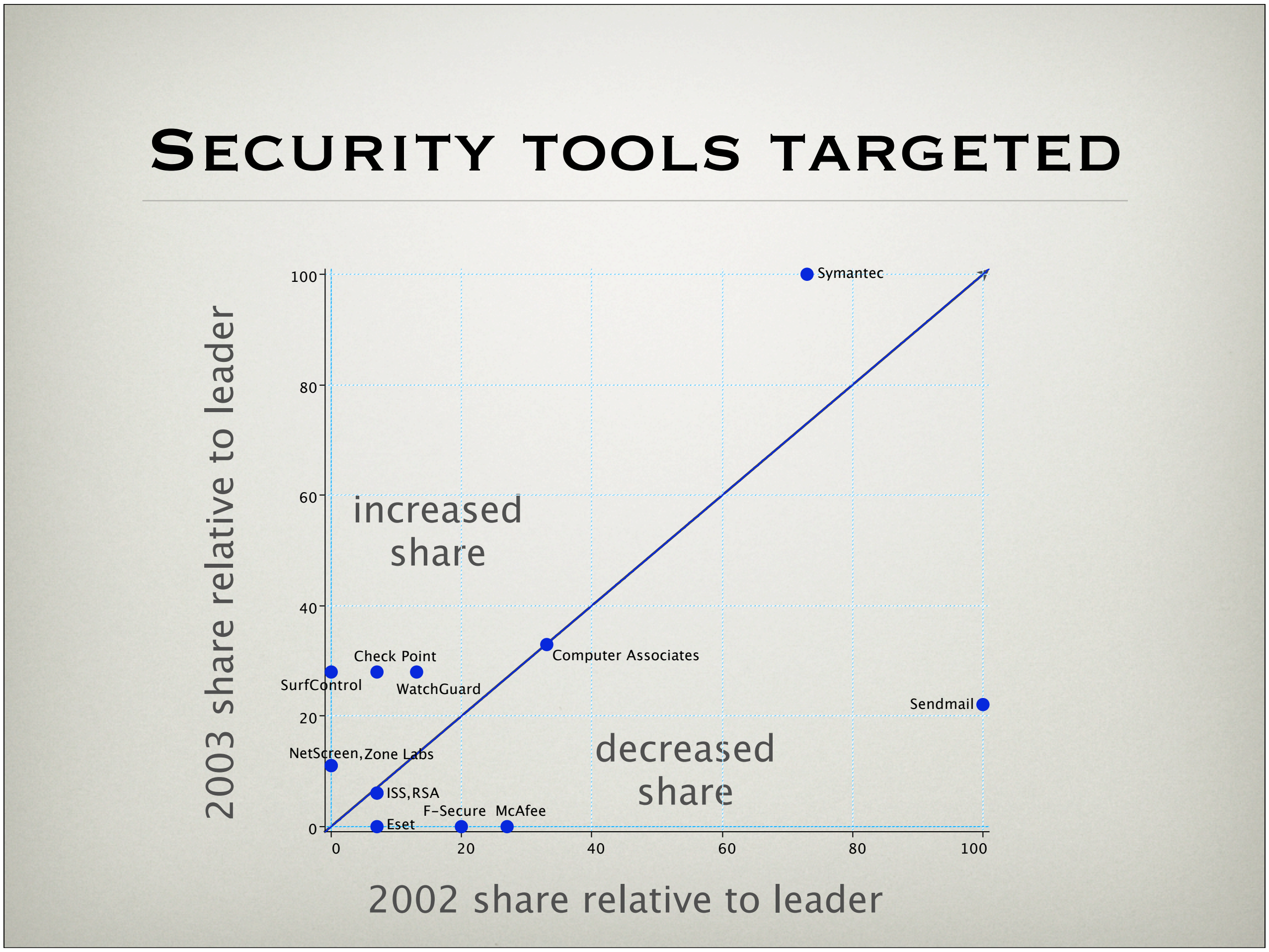
- Limit contacts with the infected
  - Internal role for quarantine
  - Care-givers need especially to be careful
  - Security products are care givers

Infections spread by contact. Having everyone on a flat network or similarly universally reachable catalyzes transmission rates. Internal segmentation -- often a side effect of regulation anyhow -- serves to limit the number of infectible parties the already infected can contact. In the real world of, say, an Ebola outbreak, it is care givers who suffer most and may represent the most significant transmission vector. Ebola, because of its lethality, is not the best example of transmission but it well illustrates that care givers and themselves be vectors.



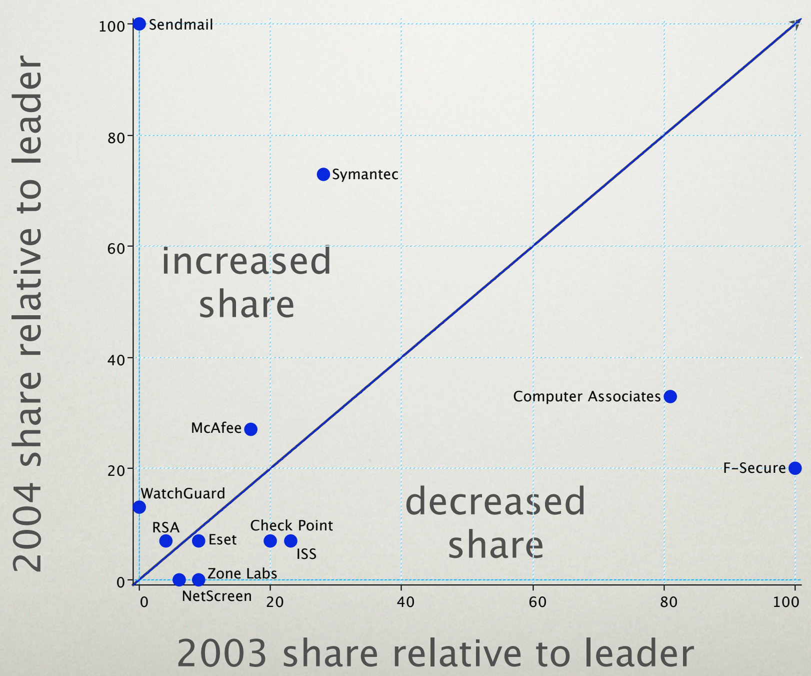
Security products have high privilege and market penetration, hence they, too, are attack targets. Symantec is currently the (unfortunately for them) reigning king of attackable vulnerabilities so this chart tells you how a given vendor's number of vulnerabilities in 2003 (horizontal) and 2004 (vertical) compare to the leader (Symantec). For example, CheckPoint had 20% as many flaws as Symantec in 2003 but 40% as many in 2004, meaning it is being could be targeted more. Note that we say "could" – if there is a non-declining percentage of vulnerabilities that are exploited then these vulnerability counts are forward-looking indicators of future attacks. This display is a analytic method that is valuable in many situations; search for "bivariate scatter plot" to see more.

Jaquith A & Singer J, "Fear and Loathing in Las Vegas: The Hackers Turn Pro," Yankee Group Trend Analysis, May 25, 2005.

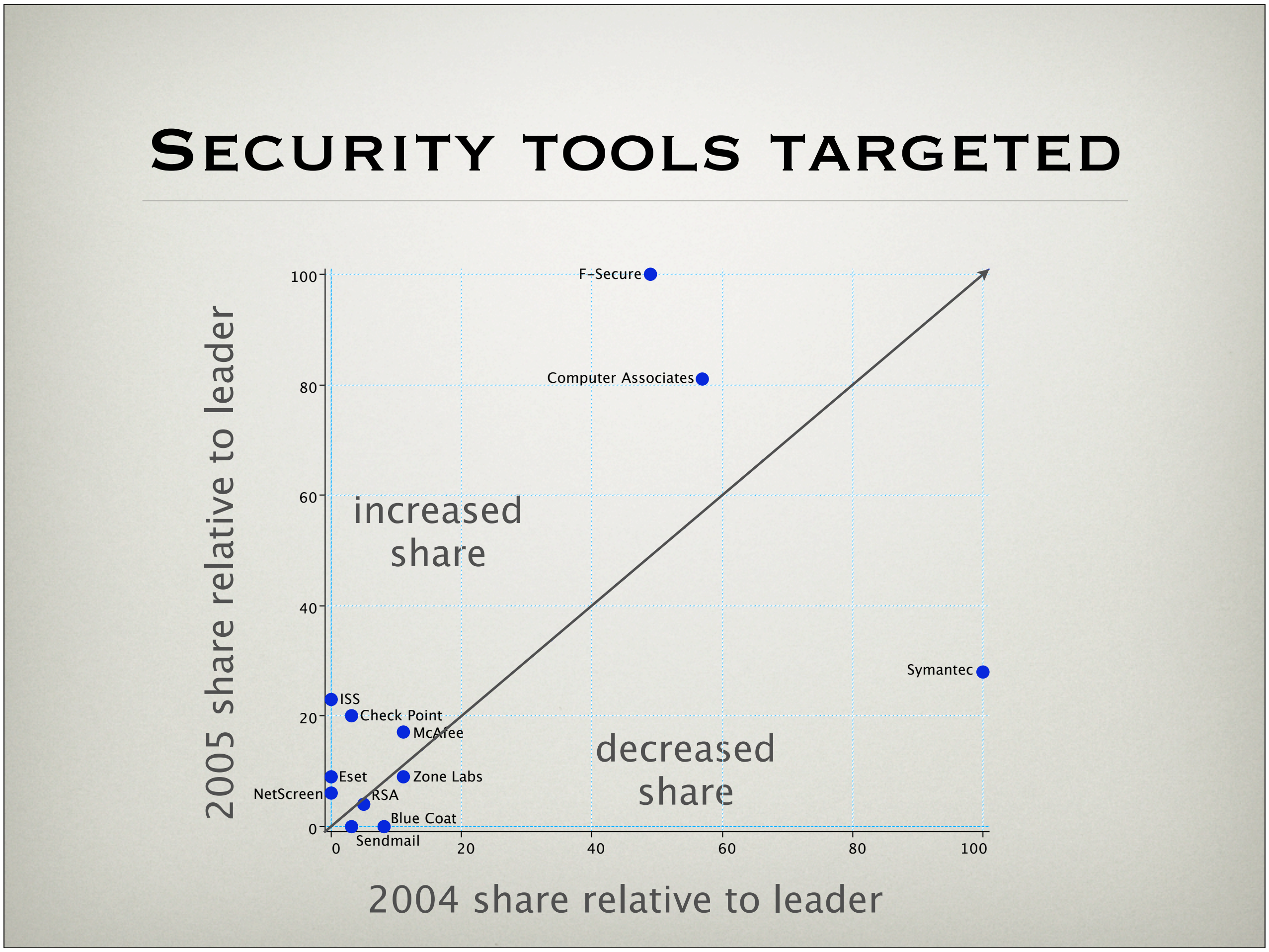


A different analysis using the NVD (national vulnerability database) in its XML form as found at http://nvd.nist.gov/download.cfm

Security tools targeted

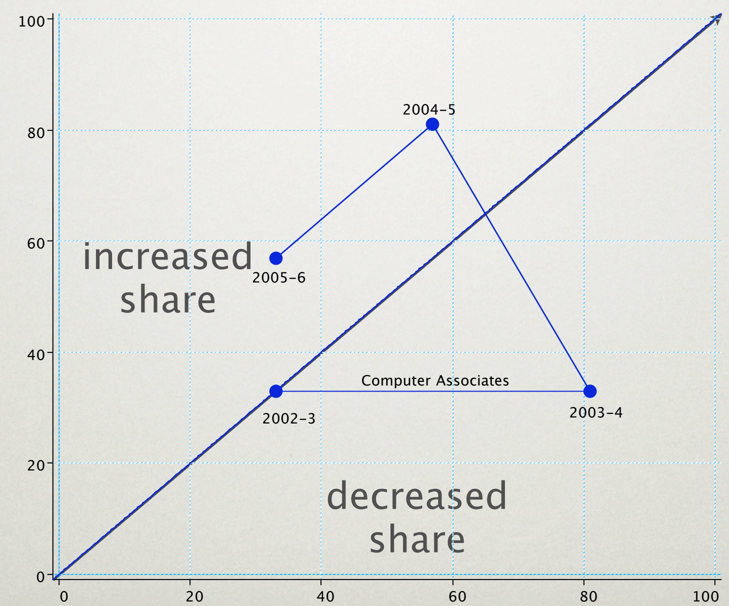


A different analysis using the NVD (national vulnerability database) in its XML form as found at http://nvd.nist.gov/download.cfm



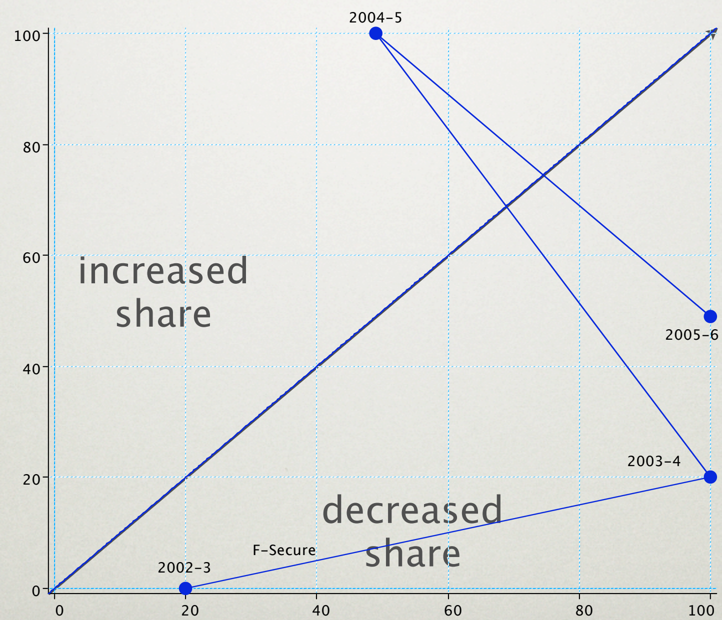
Continuing the analysis using the NVD (national vulnerability database) in its XML form as found at http://nvd.nist.gov/download.cfm

Tracking performance



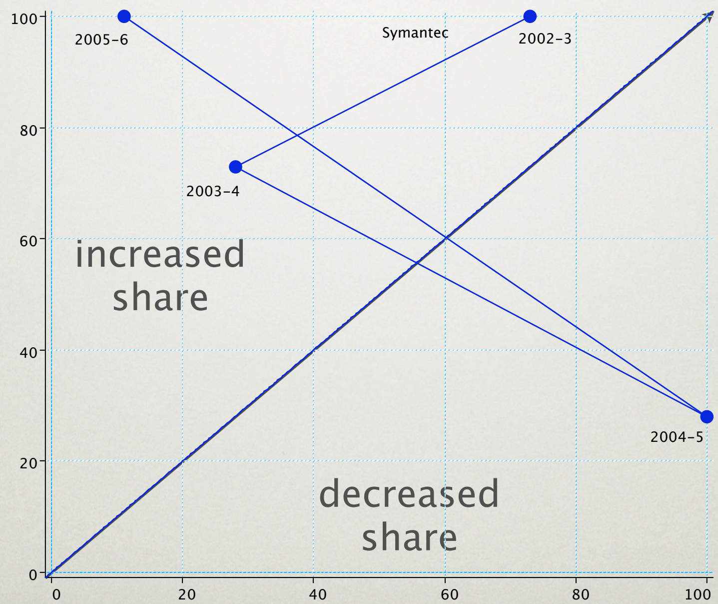
Converting to time-line form

Tracking performance



Converting to time-line form

Tracking performance



Converting to time-line form

Use in security, cont.

- Slow rates of transmission
  - Throttle network demand spikes
  - Egress filtering
    - If you wouldn't let it, why would you let it out?

Third, when an infection occurs, do something about rates of transmission. As shown by a team at the U of New Mexico, a sharp uptick in network transmission demand should be met with a sharp reduction in available bandwidth. See "Technological Networks and the Spread of Computer Viruses," Balthrop, et al., Science, v304 n23 p527-52, April, 2004.

CDC model

- Centers for Disease Control
  - Longitudinal trending to calibrate models and identify excess incidence
  - Away teams for emergencies
  - Mandatory reporting of communicable diseases

This is what the Centers for Disease Control do, what makes them what they are. The first item is the one the others are based on -- check into a hospital with Bubonic Plague and medical privacy notwithstanding, your case will be on the CDC's agenda the same day. The mandatory reporting gets them the very data to base longitudinal trend analysis on so that questions like "How many cases of tuberculosis in Atlanta is too many?"

The away teams are for when, say, a hemorrhagic fever like Ebola shows up.

Mandatory reporting is the lynchpin for public health; can we get it for digital security?

CDC model, cont.

- Share information
  - Cannot tell whether you are a target of choice or chance unless you do
  - Share normal data more than exceptions
  - Include security in DR procedures

If you can, join the Information Sharing and Analysis Center (ISAC) for your sector. Don't expect miracles, but do demand them. Information sharing is a stupendously important thing to do and never easy to get people to do. Corporate general counsels are hard to convince that the short term risk of exposure is worth it since the gain from sharing is diffuse and deferred. If you don't share normal data then questionably abnormal has no comparand. Finally, disaster recovery (DR) plans have to include security as well.

Mandatory reporting

- Can we get this in the digital world?
- Jurisdiction
    - A mess, globalized, define “disease”
- Technical
    - De-identification/anonymization essential?

Mandatory reporting is the lynchpin for public health; can we get it for digital security?

On the jurisdictional side, mandatory reporting in one locale would force events to officially occur in other locales. As to the technical side, no corporate counsel will agree to sharing attack and protection data if he thinks it can be traced back hence de-idenfication may be a technical requirement.

Mandatory reporting

- But we can get it in the enterprise
- Make sure you do

Mandatory reporting, however hard it is to do at sector-wide or national scale, is possible within the enterprise and is essential for all to do.

Quarantine

- Usually managed locally
  - Applies to border control as well
- Requires finding of dangerousness
  - Can be open-ended
- Alternative to vaccination

There is a long history of quarantine powers being reserved to the state, going all the way back to leper colonies two millenia ago. Infection control in hospitals can require quarantine, but in the public health arena everyone has heard of Typhoid Mary.

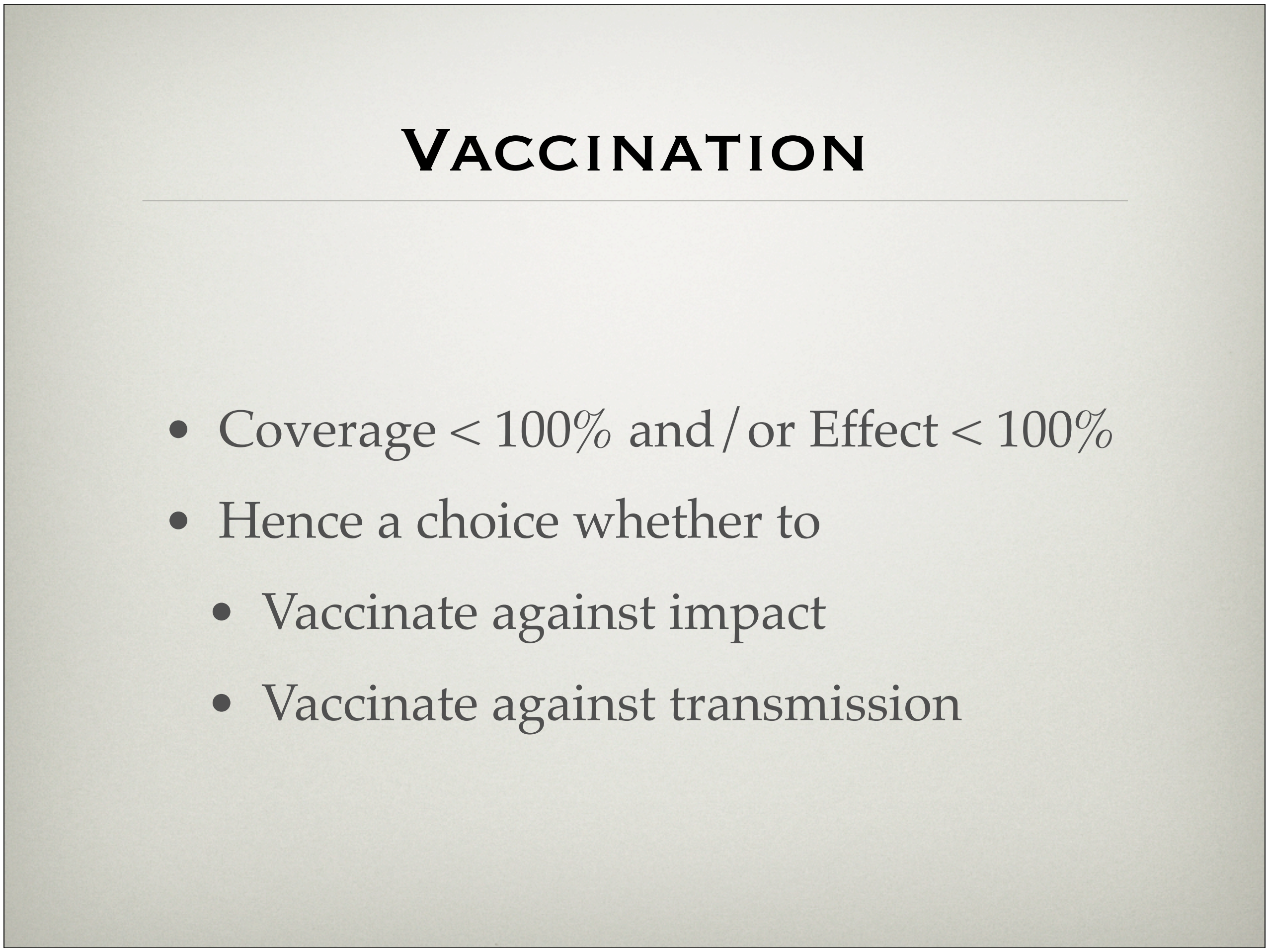
When the (2004) Witty Worm was imminent, U Cal Berkeley and Lawrence Berkeley Labs took different approaches. UCB warned systems administrators to administer a patch. LBL scanned their computers and only those who had taken the patch were allowed on the network. UCB had 800 infections; LBL had 1. Quarantine works if there are diagnostic tests.

Use in security

- "Measure" by scanning for known vulnerabilities
- Isolate them at the switch until they patch
  - One good natural experiment

Scanning for known vulnerabilities is a confirming tactic for assessing susceptibility at the population level.

When Witty broke out, there was a 48-hour warning interval. At U Cal Berkeley, lab heads and system administrators were notified and offered the patch. At the nearly identical Lawrence Berkeley Labs (25% the size of UCB), scanning and isolation at the switch was done. The scorecard? At UCB: 800 infections. At LBL: 1 infection.



Because vaccination (patching) is never fully effective, either because of not getting 100% coverage or because the vaccine is not 100% effective, in the public health situation one is left with a choice of whether to steer the vaccination program by impact reduction or by transmission suppression.

Against impact

- Ordering to <u>minimize harm</u>
  - Worst failures get first protection
- Real world: flu vaccine to old/young
- Security: patch important machines first
  - Worst = side effects like data loss

If vaccination against harm, then you supply the vaccine to those who would suffer most. In health, the sick and the weak get first intervention. In security, the juciest targets (data or control) get first intervention. Scoring this is by relative risk of harm measured before and after or at milestone intervals.

Against transmission

- Ordering to maximize herd immunity
  - Prevent replacement of cases
  - Vaccination failure ≡ Susceptible
- Real world: flu vaccine to nurses/docs
- Security: patch chatty machines first

If vaccinating against transmission, the term of art is "herd immunity" which means what it sounds like -- making the herd immune rather than the individual. In the real world, you vaccinate those most likely to transmit such as care givers themselves. In security, machines with the greatest number of connectable counterparties (perhaps instant messaging servers, say).

So why public health?

- Macro scale effects due to micro scale events
- How many of event X is too many?
- Where are the hot spots?
    - Visualization thus has a role, especially in comparing against baselines

How many of event X is too many and/or “compare and contrast departments by such and such a measure.

Public Health lessons

- Get baseline numbers but be consistent
- Share data where you can
- Keep an eye on anomalies
- At least one of {Quarantine,Immunize}

More ineresting/useful information at various sites on the Internet at large, on finance in particular (through the Financial Services Information Sharing and Analysis Center), and the CDC's "Morbidity and Mortality Weekly Report" which shows what sharing gets the practitioners of the public health discipline.

http://www.usenix.org/events/sec02/staniford.html
http://www.icsi.berkeley.edu/news/2004/nb0419.html
http://www.caida.org
http://www.fsisac.com
http://www.cdc.gov/mmwr

Insurance

Insurance models

1. Annualized Loss Expectancy (ALE)
2. Market pricing of risk transfer
3. Catastrophe Bonds

The insurance world, often said to be the salvation of security, has three main areas of focus.

1. ALE

$$ALE = \sum_{i=1}^{n} I(O_i) \times F_i$$

where:

$$O_1 \cdots O_n = \text{Set of Harmful Outcomes}$$

$$I(O_i) = \text{Impact of } Outcome_i \text{ in dollars}$$

$$F_i = \text{Frequency of } Outcome_i$$

Annualized Loss Expectancy is just a negative expected value summation across all losses (within the fixed time period of one year).

1. ALE

- Pro:

  consistent, unbiased, extracts value from experience, familiar

- Con:

  useless absent actuarial tail; also:

$$1\ event \times \frac{\$10^6}{event} = 10^6\ events \times \frac{\$1}{event}$$

The advantages of ALE are roughly that for any class of events that are widely feared and widely likely, there is an existing body of measured data sufficient to provide consistent, unbiased estimates which are then the basis for financial transactions as needed. However, when the events are rare, or the substrate changes often, this is harder. In the case of digital goods, however, the losses are subject to intentional initiation and automation of technique which does rather change things, as illustrated by the last line of the above.

2. Transfer market

- Re-insurance exists to diversify risks by pooling and then trading them
- For unique risks, auction pricing based more on risk aversion of seller than risk seeking of buyer, *i.e.*, over-priced risk

An insurance company diversifies its client base but it also lays off a portion of its risk -- all with an eye to avoiding a level of retained risk that is a threat to the capital base. When the risks being laid off are numerous and subject to calculations by both buyer and seller, they are tradable commodities. When they are unique, the seller's risk aversion, expressed as a risk premium demanded by the buyer, means that sellers will generally have to accept payment for the transfered risk that is less than their own valuation of it.

3. Catastrophe Bonds

- If no catastrophe,
  Then acts like a high yield bond
- If defined catastrophe does occur,
  Then principal diverted to beneficiaries

Main backstop for Florida hurricane coverage

When the market refuses to sell insurance, such as did occur after the Loma Prieta (California, 1989) earthquake, an alternative is to seek not insurers but investors who buy “catastrophe bonds.” Cat bonds pay a high rate of interest and are for fixed intervals. If, during that interval, no catastrophe occurs, the investors receive their capital back at the end of the bonded period. If the catastrophe does occur, then the capital is not returned, the bond effectively defaults, and the funds are diverted to defined beneficiaries for the mitigation of the catastrophe. Florida hurricane coverage is still available as insurance but cat bonds are the stop-loss backstop to much of the coverage.

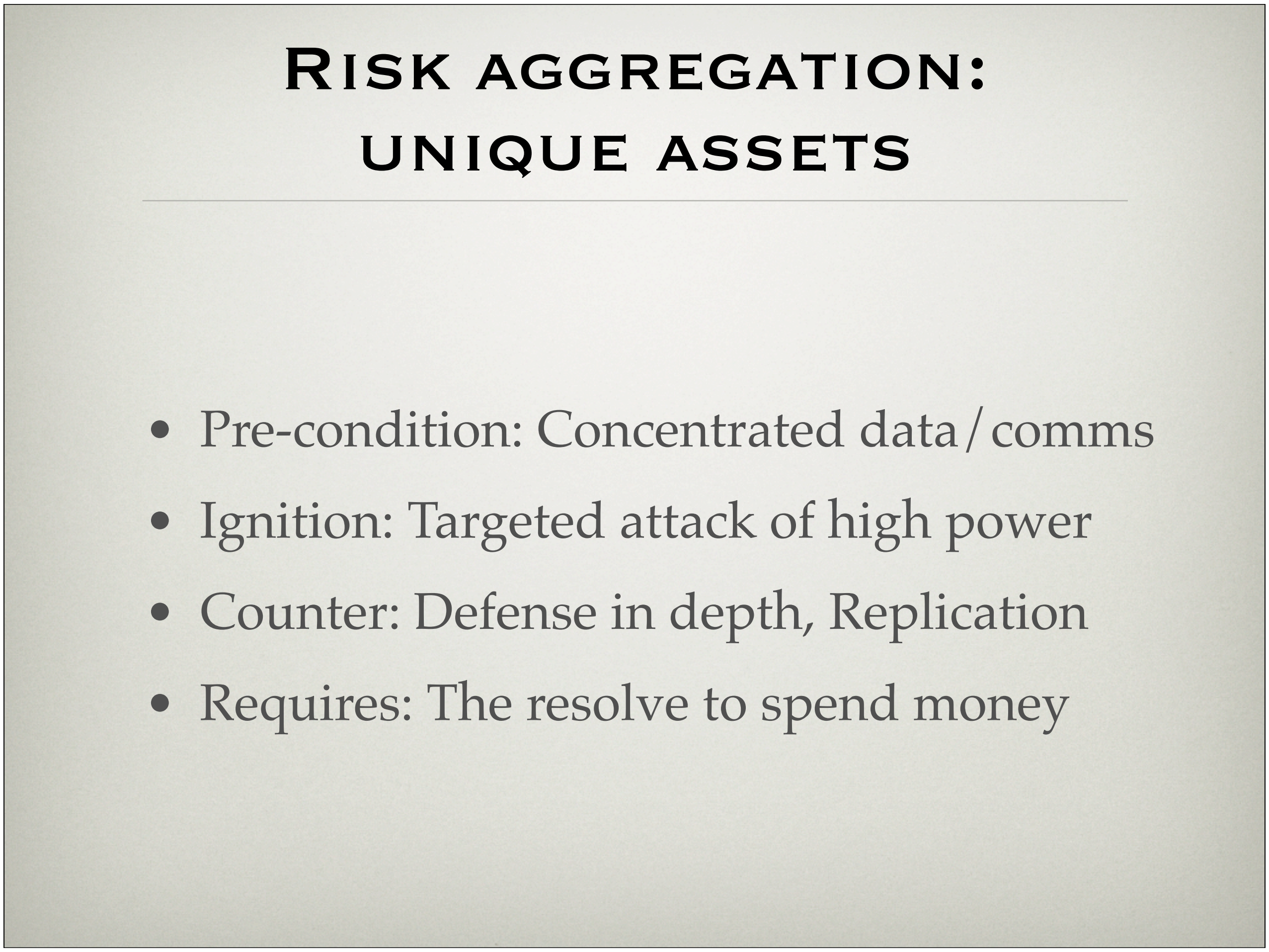
Insurance

- Risk aggregation
  - Consistent with zero loss history
  - Undermines premium pricing
  - Inherent to intentionally unique assets
  - Exacerbates cascade failure (impact on DR)

Insurance has an extremely valuable concept: “risk aggregation.”

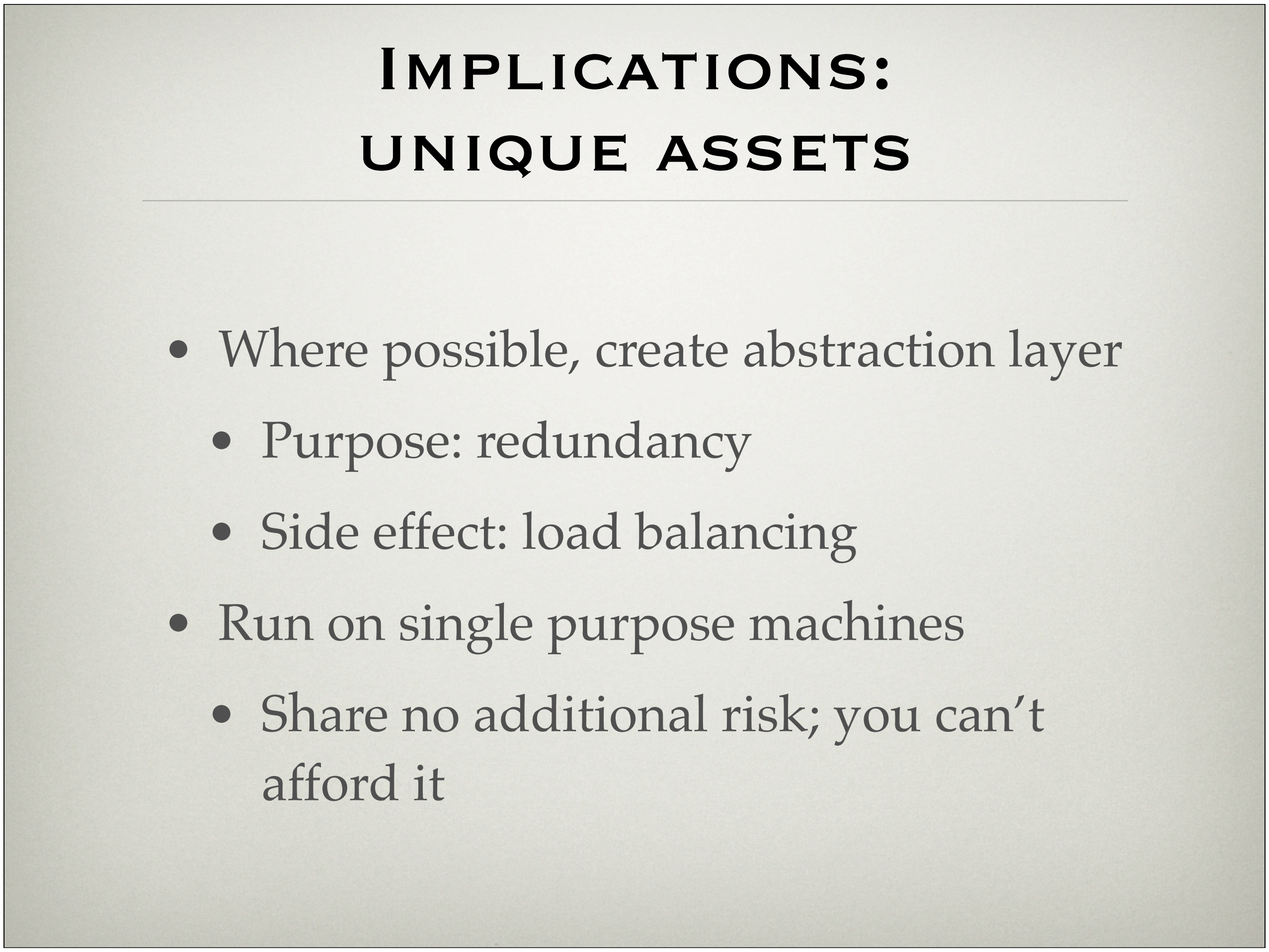
Risk aggregation undermines a portfolio as it makes the appearance (in time) of claims be correlated with events that may not have yet occurred. No writer of homeowners’ insurance wants closely adjacent houses lest if one burns down the other will, too. Worse still, no writer of homeowners’ insurance wants to discover that an earthquake burns down all the houses in an entire county. The problem is, without a long actuarial tail, it is not possible to disambiguate a history of zero insurable losses with an event where all individual risks are globally correlated but which has yet to occur.

The scariest digital risk is loss of an intentionally unique asset. The most uncontrollable digital risk is cascade failure.



For unique assets to be a risk at the national scale, you need the pre-condition of some high concentration of data, communications, or both. The ignition of that risk is a targeted attack of high power up to and including the actions of nation states. The counter to this latent risk is “defense in depth” which may include replication. Defense in depth is ultimately (at the policy level) a referendum on the willingness to spend money.

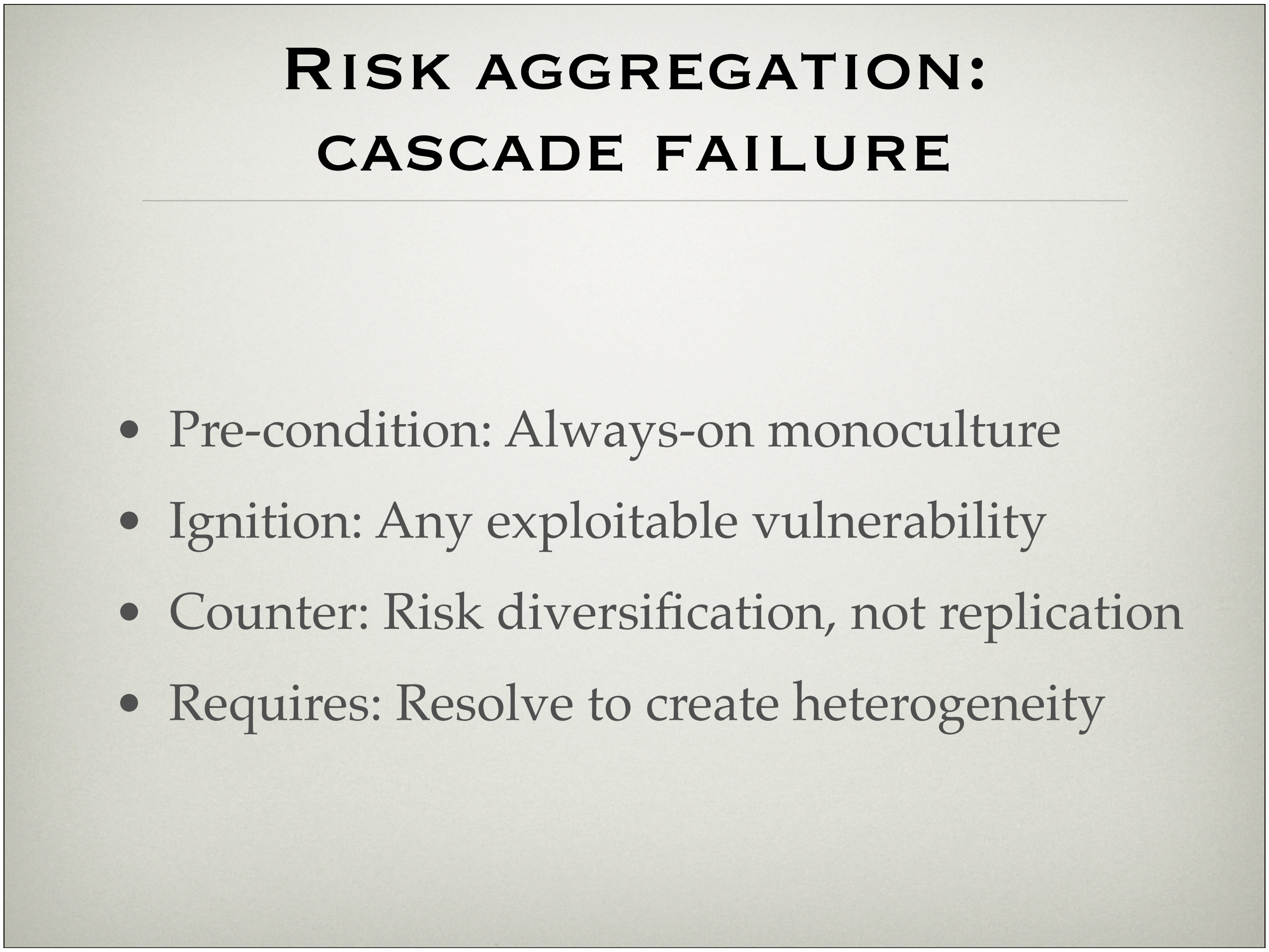
As such, there is nothing more to say at the general level and we lay this branch of the tree aside so as to focus on the other.



For your unique assets, your best bet is redundancy. If the Prob(failure)=m, n-way redundancy changes that to Prob(failure)=m^n, a result which assumes your redundancy does not create monocultural cascade failure possibilities. Note that the Internet's Domain Name Service (DNS) is 13-way redundant and the implementations at each of the 13 root name servers are different, radically so in general.

Also, do not inherit risks you don't need. If the Prob(failure)=m for your service but other things on the current machine could cause machine failure with Prob=x, then the Prob(success) for your service is [1-((1-m)(1-x))]. Taking m=10^-4 and x=10^-2, a combined machine has Prob(failure) of .989901 or close to the value of x, not m.

Take a Kerberos Key Distribution Center (KDC) for example; its security must be paramount so you must run it on a single host running nothing else. However, the absence of the KDC service cascades to all other services relying upon it, so it must be replicated. To avoid creating new failure modes, you run with one master and several slaves so as to trade the diminished operation (no password changes if the master is offline) for avoiding an overall absence of Kerberos service.

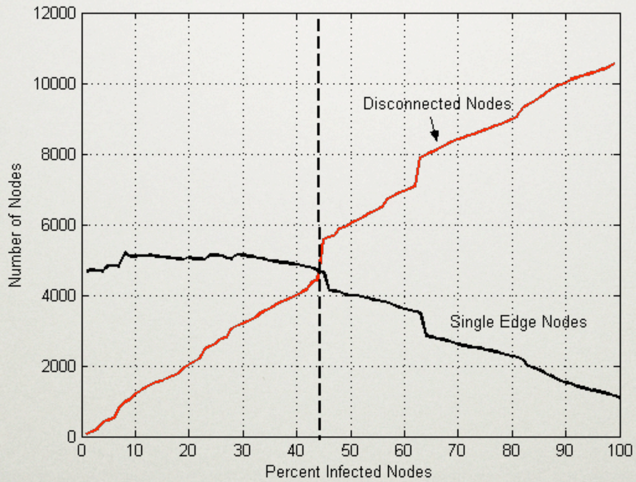


For cascade failure to be a risk at the national scale, you need the pre-condition of an always-on monoculture. The ignition of that risk is an attack on vulnerable entity within the always on monoculture so long as it has a communication path to other like entities. The counter to this latent risk is risk diversification which absolutely does not include replication. Cascade avoidance is ultimately (at the policy level) a referendum on the resolve to treat shared risk as a real cost, per se.

We now follow this branch to see where it leads. Sean Gorman of George Mason University has an upcoming publication that suggests that the risk-cost of homogeneity kicks in at rather low densities (preliminary results indicate 43% for leaf nodes, 17% for core fabric).

Cascade failure threshold

Gorman



This graph is the result of a simulation where a monoculture of hosts is increasingly infected with malware that prevents further communication with that host. The point is the discontinuity at 43%, and the planning implications of that (such as to avoid having more than 43% of any particular platform in the total mix of platforms within a single enterprise).

Source: Gorman SP, Kulkarni R, Schintler L & Stough R, "Is Microsoft a threat to national security? The effect of technology monocultures on critical infrastructure", George Mason University, Infrastructure Mapping Project Working Paper, 2004.

Monoculture as cascade

$$\begin{array}{lrcl} \text{let} & sizeof(enterprise) & = & y \\ \text{and} & Pr(individual\_infection) & = & x \\ \text{hence} & Pr(no\_individual\_infection) & = & 1-x \\ & Pr(no\_group\_infection) & = & (1-x)^y \\ & Pr(group\_infection) & = & 1-(1-x)^y \end{array}$$

we want LD50: $x \mid y$ such that $Pr(group\_infection) = 50\%$

$$\begin{array}{lrcl} \text{which means:} & .50 & = & 1-(1-x)^y \\ & (1-x)^y & = & .50 \\ & (1-x) & = & \sqrt[y]{.50} \\ & x & = & 1-\sqrt[y]{.50} \end{array}$$

This may be akin to beating a dead horse, but you are welcome to work through the math which, in turn, is the basis for the next page.

What we want to know is this: For a given enterprise size (y) how much risk can each desktop separately have before there is a greater than even chance of a cascade failure of the enterprise as a whole.

Notation: Pr(A) is the Probability of A; LD50 is the Dose which will prove Lethal to 50% of the experimental animals.

Cascade triggering

Internet: $n(websites_{total}) \approx 25 \times 10^6$

For $y = 5,000$, $x \approx \frac{1}{7,200}$
$n(websites_{infected}) = x \ \times \ 25 \times 10^6 \approx 3,500$

For $y = 100,000$, $x \approx \frac{1}{144,000}$
$n(websites_{infected}) = x \ \times \ 25 \times 10^6 \approx 175$

This is a fabricated example, but it illustrates how much voltage is on the wire unless there are some resistors and capacitors to damp it out.

If we estimate the total number of websites as twenty-five million and we have the somewhat fanciful idea that every person in the enterprise visits one of them at random, then a cascadable monoculture within the enterprise means that the LD50 for five thousand seats is .00014 so that if there are at least 3,400 infected web sites amongst the twenty-five million the odds favor the enterprise getting an infection. For one hundred thousand seats, if there are at least 175 infected web sites then the odds of infection are at least fifty percent.

Implications: cascade failure

- Perimeter-centric defensive posture
  - Anything that stops propagation: platform diversity, network segmentation
- Decision: ingress or egress filtering?
  - Attacking customers is especially bad

Avoiding cascades is about putting up roadblocks to the easy flow of hostile bits, regardless of the particulars of how those bits are sourced, organized, or targeted. This is where insurance mindsets and public health mindsets are much the same; public wants to stop propagation while insurance wants to limit propagatability.

Head of worldwide operations, NYC investment bank, said “Last year, we stopped 75,000 inbound viruses but I am prouder that we stopped 500 outbound ones.” Parsing that, this individual is saying that in decision analytic terms the “utility” of stopping an outbound virus is 150-to-1 that of stopping an inbound virus. Two orders of magnitude -- sounds about right though maybe three would be better. If that is not convincing, consider active attacks outbound and not just propagating attacks outbound.

Use in security

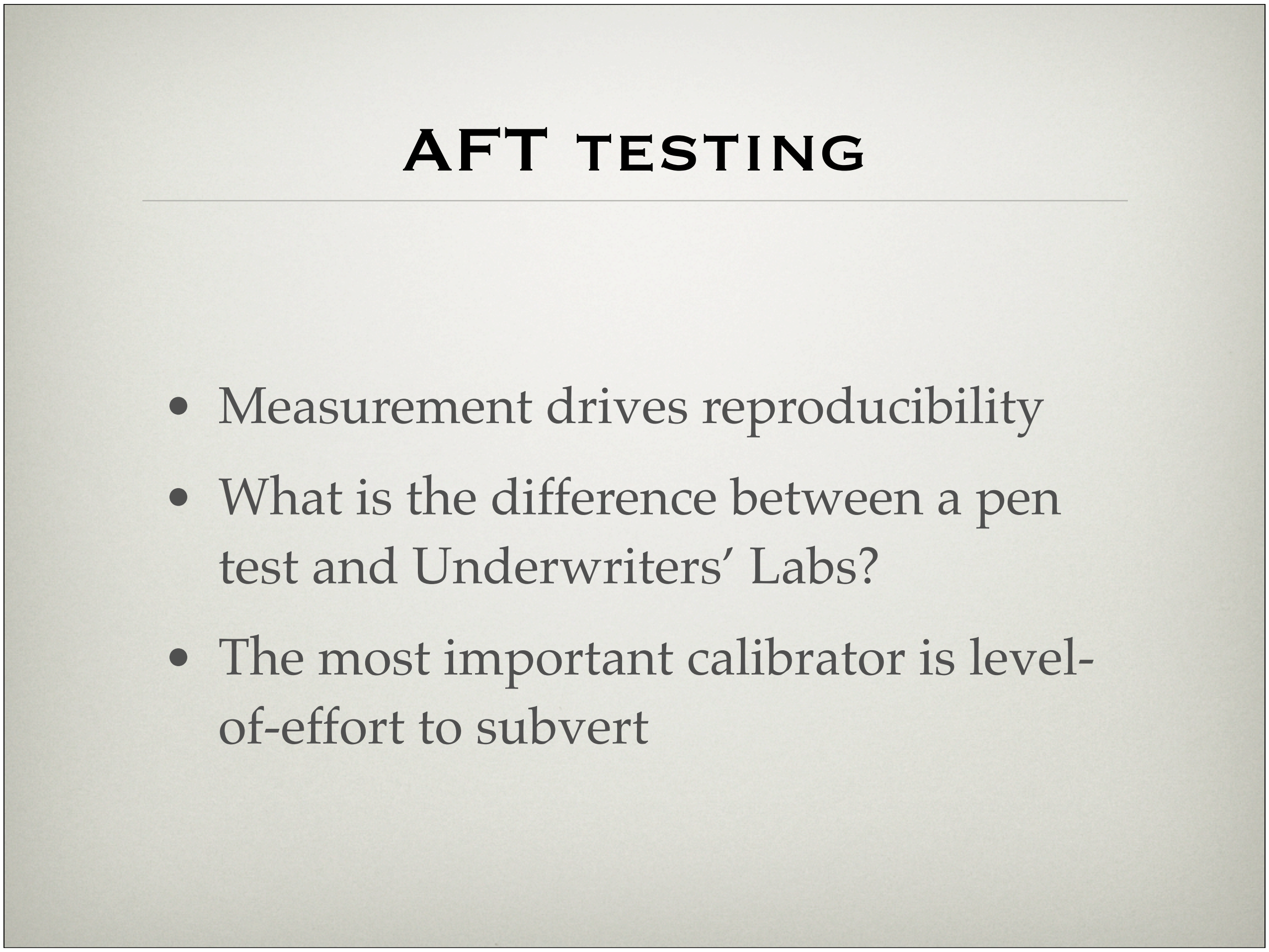
- Today: business continuity policies
- Tomorrow: track evolution of liability
- AIG *netAdvantage*: security & privacy liability, cyber extortion & terrorism, injury to information assets, business interruption, crisis communication

Three papers from http://www.infosecon.net/workshop/ are relevant here.

In one, the authors show that the risk due to platform monoculture is mitigated by introduction of a second platform even if that second platform is itself less secure than the first, i.e., diversity alone results in reduced firm-wide risk. In the the second paper, it is argued on social capital grounds that the public policy consequence of a monoculture must be mandatory sharing of vulnerability and incident data. In the third paper, a full-tilt, academic-grade mathematical economics argument is made for differential insurance premiums for diversity as a counter to risk-correlation.

AIG, a leading insurer, has been first to market with a number of digital security offerings; http://www.aignetadvantage.com/

Accelerated failure time testing



Once again, the need for metrics is clear and only if we measure can we achieve reproducibility.

Note that a penetration test is just like what Underwriters' Laboratories does with, say, a toaster. The question is not whether UL can break a toaster, of course they can. The question is whether putting the handle up and down 5,000 times breaks the toaster or whether it takes 10,000 times.

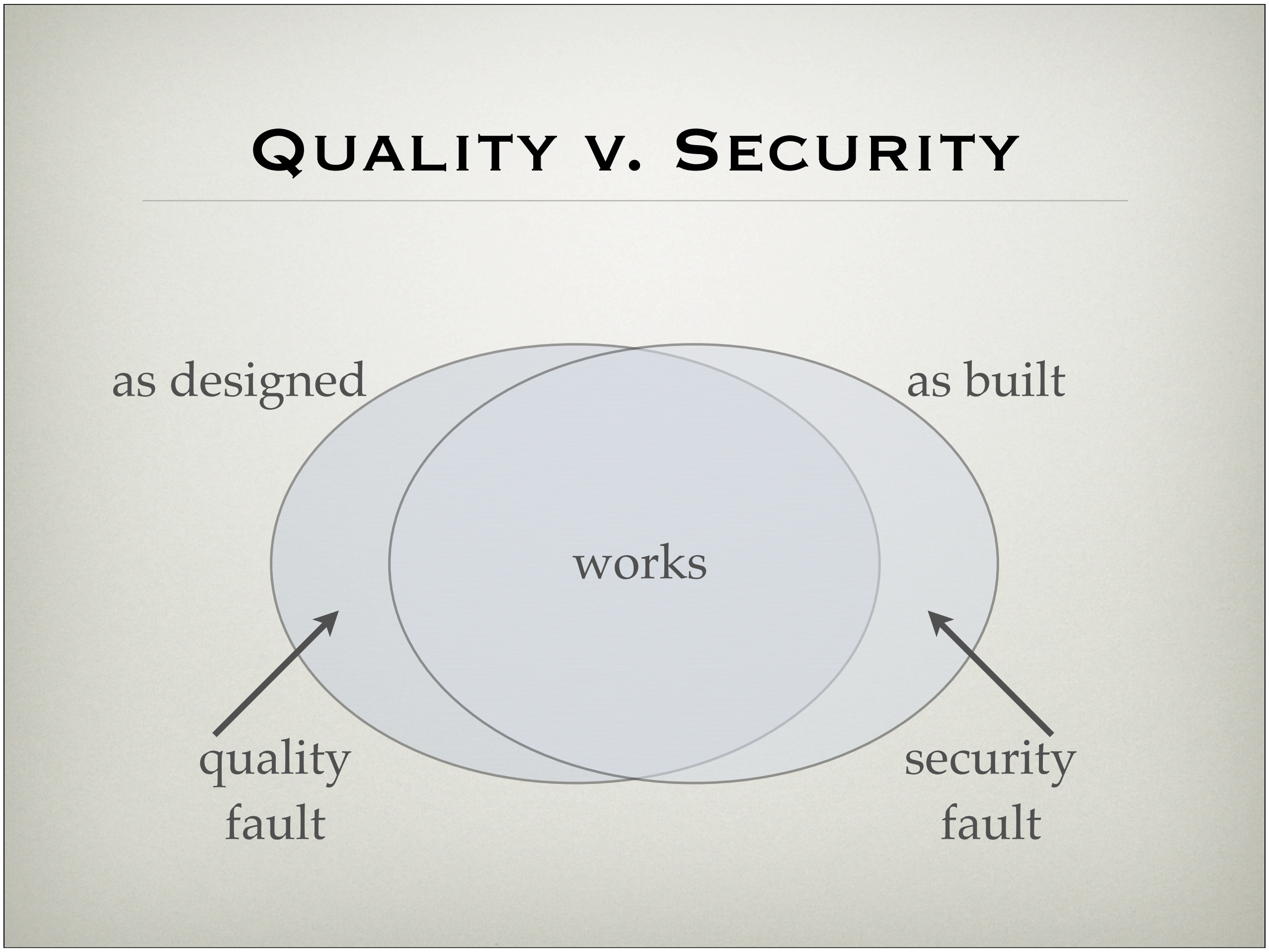
The most important measure, hands down, is the level of effort to penetrate -- what does it take the penetrator to achieve his aim? This allows two important things, relative ordering of like products or like threats -and- a way to assess whether a proposed mitigation is against something that is worth mitigating, e.g., mitigating against takeover by a national laboratory is not a reasonable strategy for a taxi-cab company.

AFT testing

- Tests known, established modes of failure
- Doesn’t exist to test if failure can happen
- Tests what it takes to cause failure
  - Ex: slam a car door until it falls off

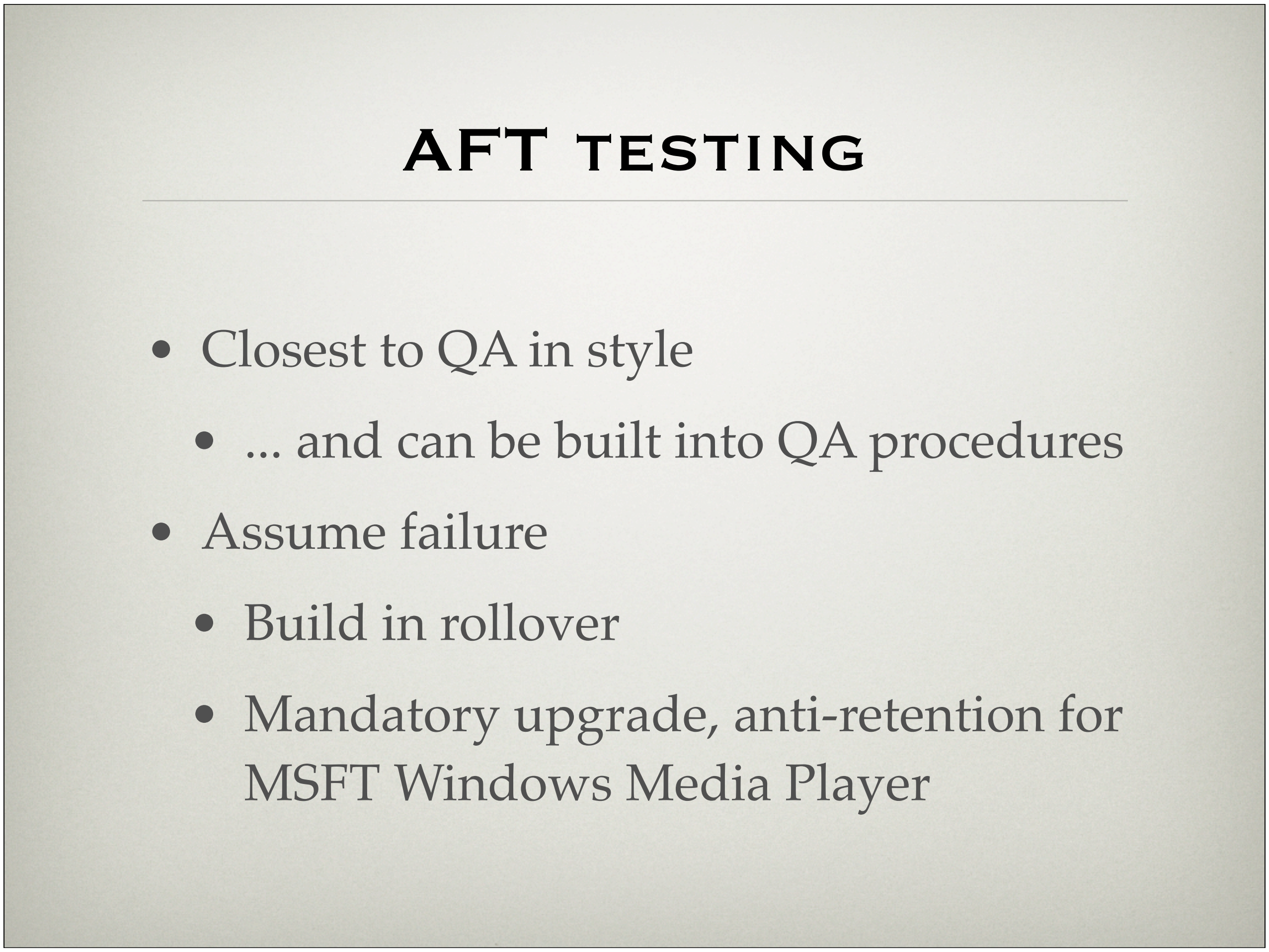
An important point: AFT requires that you know what you are looking for. It does not discover whether a failure can happen given the particular stress but rather how much of that stress does it take to cause the (inevitable) failure.

This can be a quantitative result, like “It takes 7,500 door slams on average to fatigue the hinge post enough to sag the door beyond operational limits.” This can also be a qualitative result, like “Direct contact between swords A and B showed that B sustains damage at a rate faster than A and will thus fail first in actual use.”



The picture (from Doctor Dobbs' Journal) of what makes quality and security so similar but so distinct: Where there is a design requirement but no implementation there is a quality fault. Where there is an implementation but no design requirement, there is a security fault. The strategy for forcing early failure is different if what you are looking for is a design point that was not correctly implemented versus an implementation fact that was not in the original design.

Thompson HH & Whittaker JA, "Testing for Software Security," Dr. Dobbs Journal, v342 p24-34, November, 2002.



This is the closest of all these measures to what a quality assurance engineer would recognize. You load a server until it degrades, you increase transaction rate until you saturate (“TPC”), and so forth. The idea is to assume failure and to prepare for it.

Ignoring any questions of who owns what, for some time the MSFT Windows Media Player has had features of mandatory upgrade (if an upgrade is available, the user must accept that upgrade) and anti-retention (if an upgrade is taken, the previous version must be deleted). Within the corporate environment, similar policies are often in place even if not formalized in contractual language.

See: Arbaugh W, Fithen W, & McHugh J, “Windows of Vulnerability: A Case Study Analysis,” IEEE Computer, v33 n12 p52-59, December, 2000; http://computer.org/computer/co2000/rz052abs.htm

Example: LOPHTcrack

- L0PHTcrack breaks passwords
  - Tells you how easy passwords are
  - Tells you who hasn't got the message
  - Tells you gross percentages
  - Permits divisional comparisons

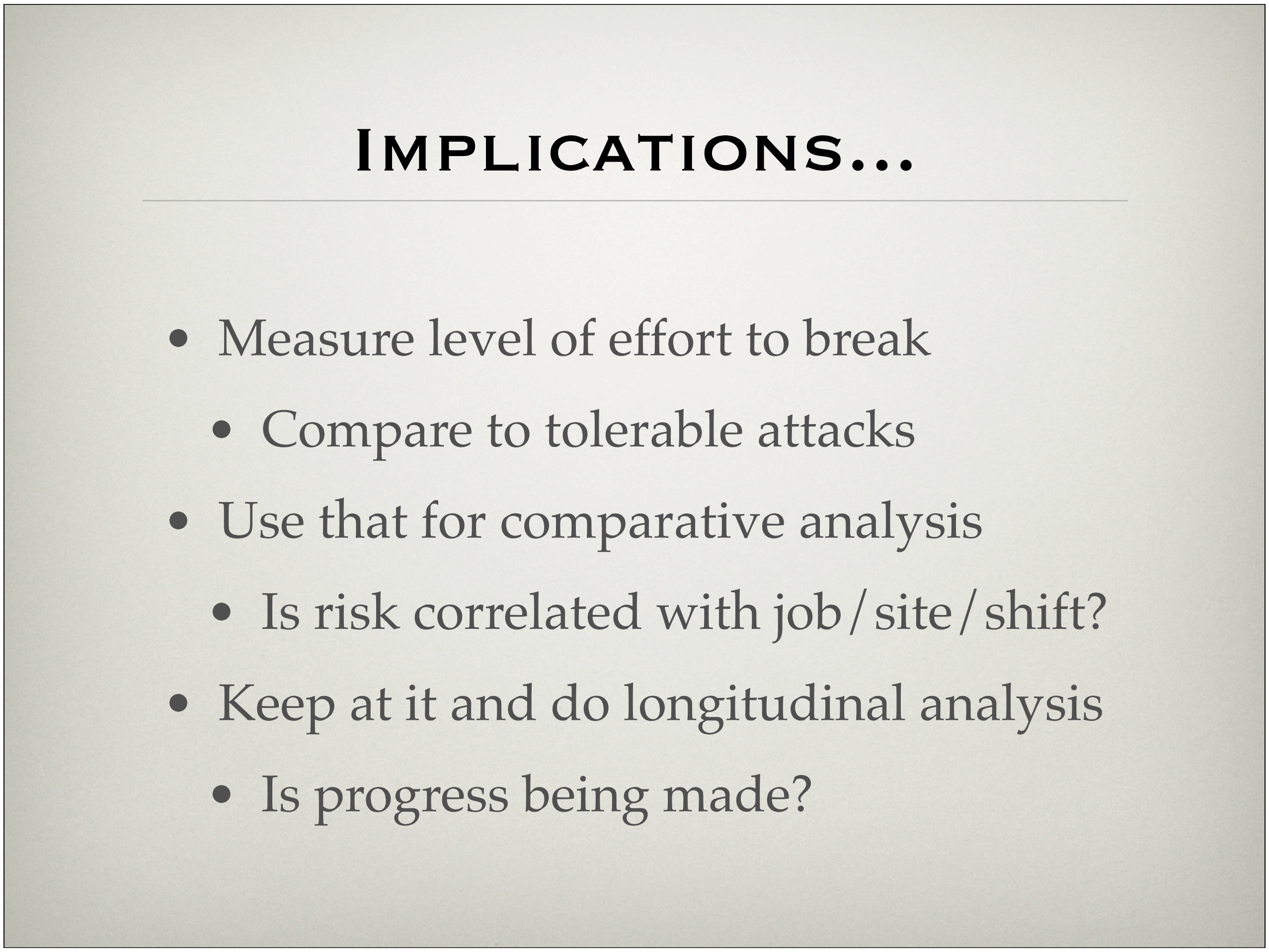
L0PHTcrack, now known as LC5, is available from Symantec (having bought @stake having bought L0PHT Heavy Industries). It is the admitted best such commercially available tool.

Example: LOPHTcrack

- Actual metrics (across ________)
  - Average time to break
  - Percentage breakable in X minutes
  - Quartile analysis (*see later slides)*
- Optional:
  controlled trial of awareness program

These are examples of what you can measure, such as to compare roles (authorization levels) for the average time to break their passwords, to compare departments on the percentage of their passwords that are breakable over a lunch hour, and to look at the spread in results using quartile analysis (which we demonstrate later).

As an option, measure average time to break for two groups and then for one of those groups apply user awareness training then wait a month and then re-measure. You have a "case-control" study of the effectiveness of (back to public health) immunization against poor password choice.



The reason to know level of effort to break is to compare that to what is tolerable risk.

The reason to compare across various lines in the business is to focus attention on remediation. If your internal network is flat and all your firewalls are good but one, then what is your perimeter really? Order your divisions and attend to them in that order.

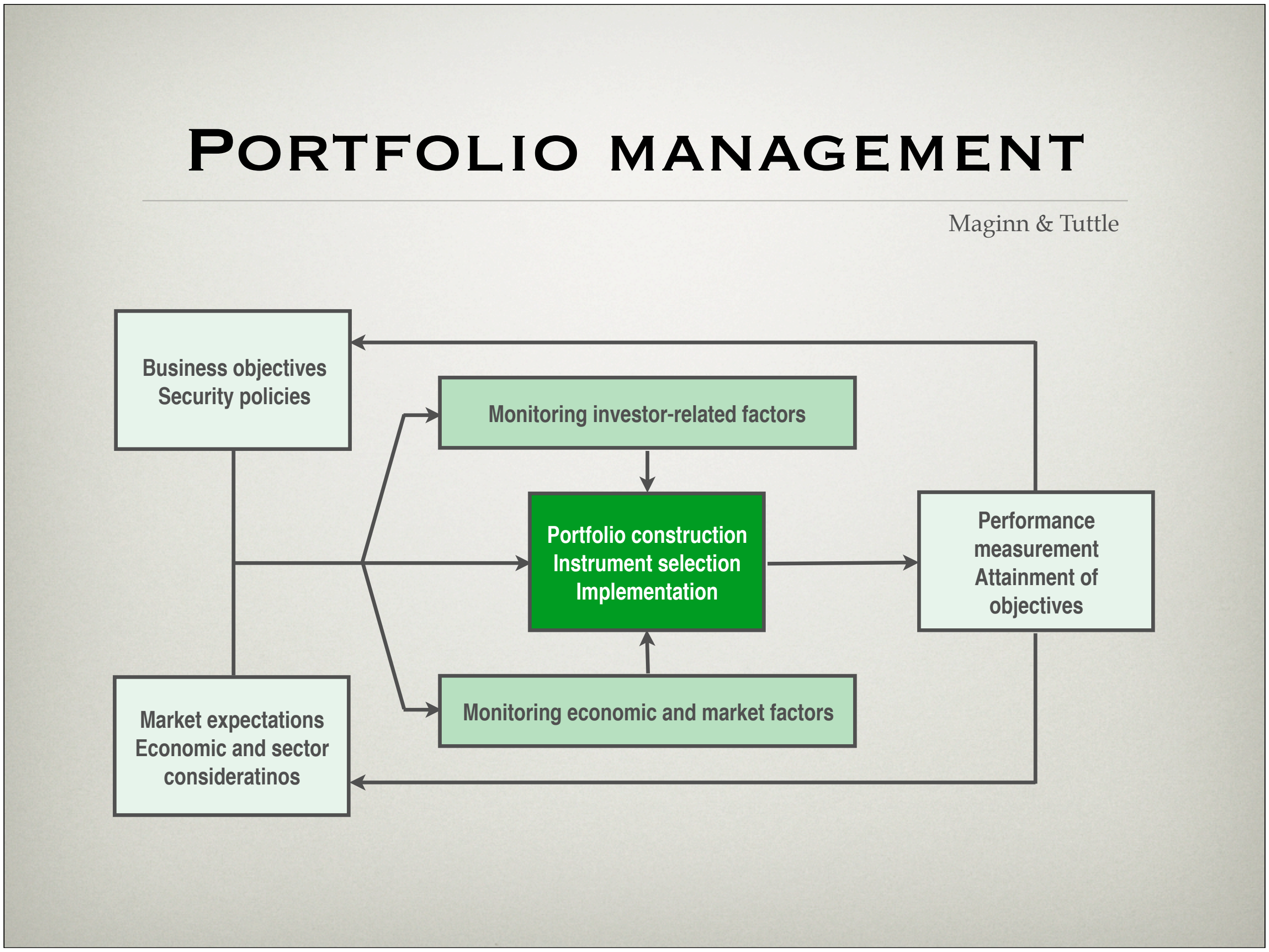
Back to those original questions, “Am I better off than I was this time last year?,” one sees the point of longitudinal (time trend) analysis. Public health touched on that as well with the difference that it is an observational regime around inherited risk whereas accelerated failure time testing is an intentional provocation of

When an event is inevitable eventually, your best effort is to compress that time to event so that you can reduce its chance of actually happening by understanding the risk factors that would make it come sooner. This amounts to removing the (cliche alert) “low hanging fruit” opportunistic attacks rely upon.

Doing such work over and over allows comparisons either across definable sub-groupings of the firm or longitudinally across time. Either way, you get a relative vulnerability ordering and that alone is sufficient for decision support in security operations.

For more on relative vulnerability, see: Cowan C, “Relative Vulnerability: An Empirical Assurance Metric,” Workshop on Measuring Assurance in Cyberspace, June 26, 2003, Monterey, California, at http://www.laas.fr/IFIPWG/Workshops&Meetings/44/W1/10-Cowan.pdf and, later, http://www.homeport.org/~adam/shmoocon/shmoocon-cowan.pdf

Portfolio Management



A classic formulation from a classic text, as first proposed in Jaquith A, "Learning from Wall Street: Risk Management for Applications," Secure Business Quarterly, Q2 2002; see http://www.sbq.com/sbq/app_security/sbq_app_wall_street.pdf

Source: Maginn DL & Tuttle DW, Managing Investment Portfolios, 2nd edition, 1990, Warren Gorham & Lamont.

Portfolios for security

- Hedging-like ideas, such as
  - DHS says "Orange" ⇒ adjust knobs
  - Diversify risk
  - If future uncertain, invest for flexibility
- Hard to find "leading security indicators"

Perhaps illustrating that this is an idea whose time has come, analysts are now touting portfolio management ideas for security management. The first known reference, by the Giga Group, is at http://www.cio.com/analyst/012502_giga.html; there are others, of course.

Portfolio theory

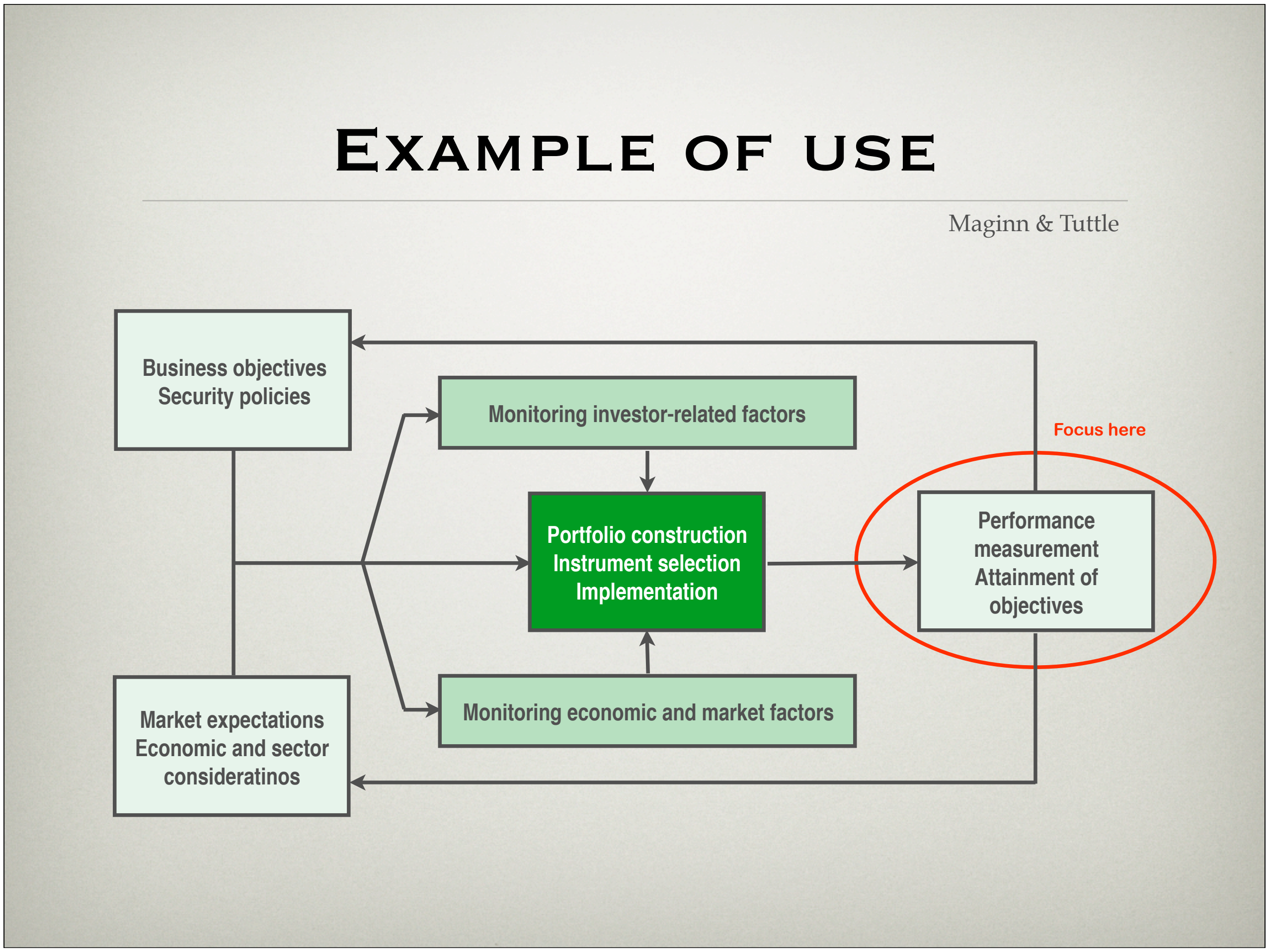
- Risk is a commodity that can be
  - Classified
  - Measured
  - Priced
  - Traded
- Portfolios balance the risk of multiple investments

The point for financial types is to get risk into commodity status, keep it there, and make some money. Risk as understood in finance is not bad so long as it is priced correctly and hedged adroitly. The job of the portfolio manager is to balance the aggregate risk of multiple investments in the portfolio.

Portfolio risk handling

- Security analytics measure portfolio performance
  - Drive return on security investment (ROSI) calculations
  - Feed back into risk quantification

This is harder to get just right than it sounds, but also easier to get started.



Let's work an example using just the performance measurement / attainment of objectives part of this.

...A LITTLE EXAMPLE USING POOLED DATA...

Real life example from published literature over the next several slides.

Geer DE, Jaquith A, & Soo Hoo K : “Information Security -- Why the Future Belongs to the Quants,” IEEE Security & Privacy, v1 n4 p24-32, July/August, 2003.

QA LIT ON LEVEL OF EFFORT

Relative cost to fix issues, by stage	
Design	1
Implementation	6.5
Testing	15
Maintenance	100

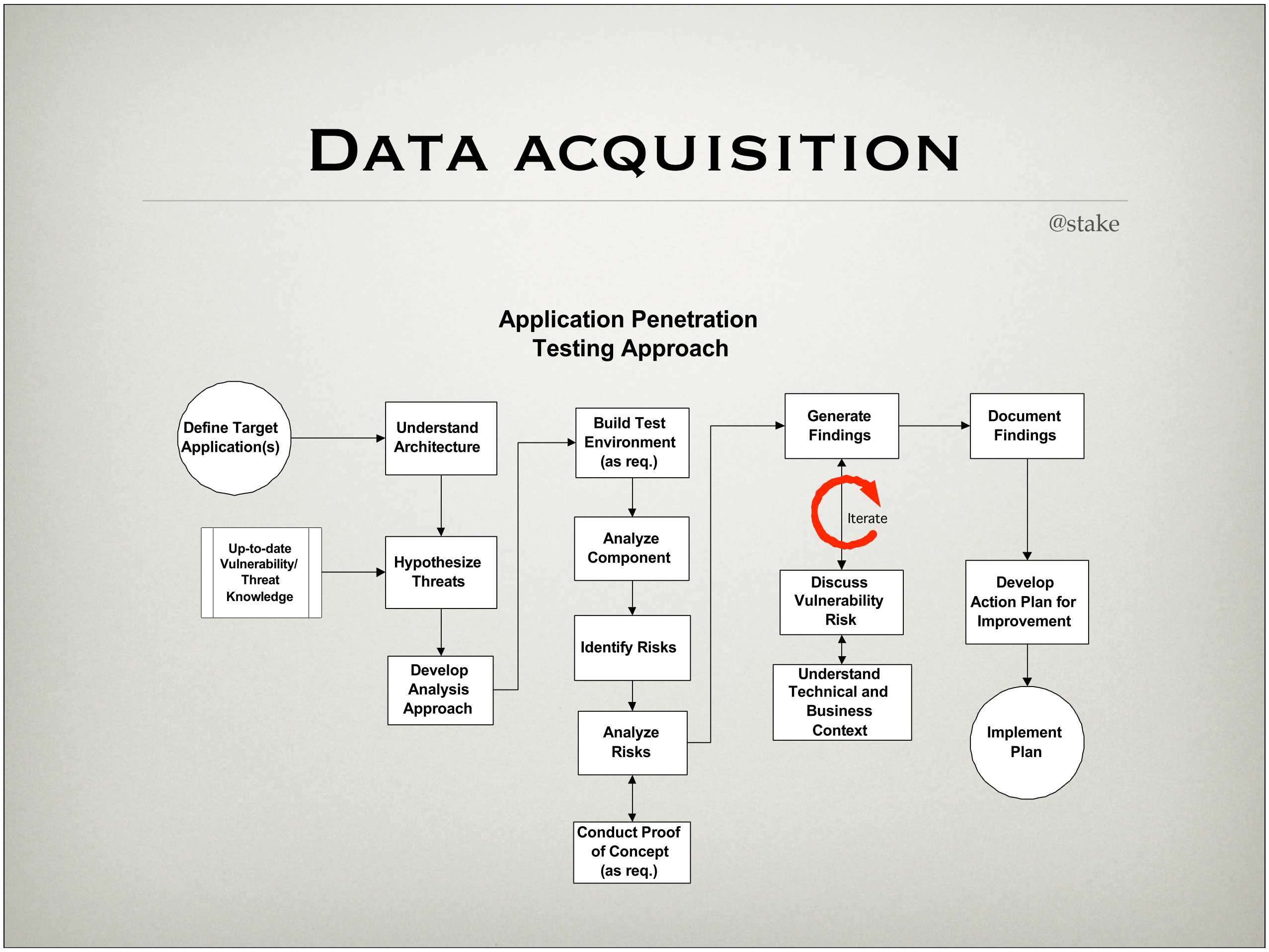
Implementing Software Inspections, IBM Systems Sciences Institute, IBM, 1981

Software development costs, by stage	
Design	15%
Implementation	60%
Testing	25%

Architectures for Software Systems, course Notes, Garlan & Kazman, CS, CMU, 1998

The IBM study (by Barry Boehm) said that $1 in design bought as much as $100 did in field maintenance, and this was for a time when product lifecycles were more relaxed than they are now and location-independent attack was impossible. More can be found in his book, Software Engineering Economics, Prentice Hall, 1981.

The Software Engineering Institute at Carnegie Mellon arrived at the other numbers by measuring practice, not as a proscription for what to do.



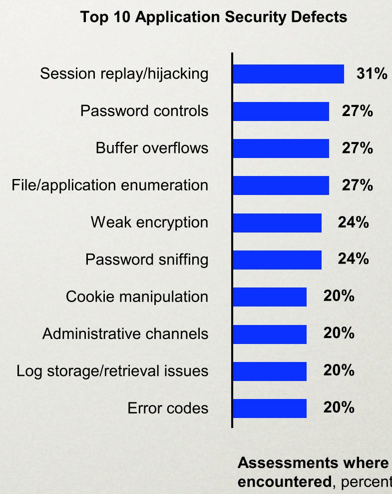
This is the standard methodology of a well known security company to penetration testing applications. It does not matter what the approach is precisely, but it does matter that this approach was used on many, many engagements hence bias of observation can be analyzed away since all data that was collected under this methodology had the same biases.

Security defects common

Security Defects by Category

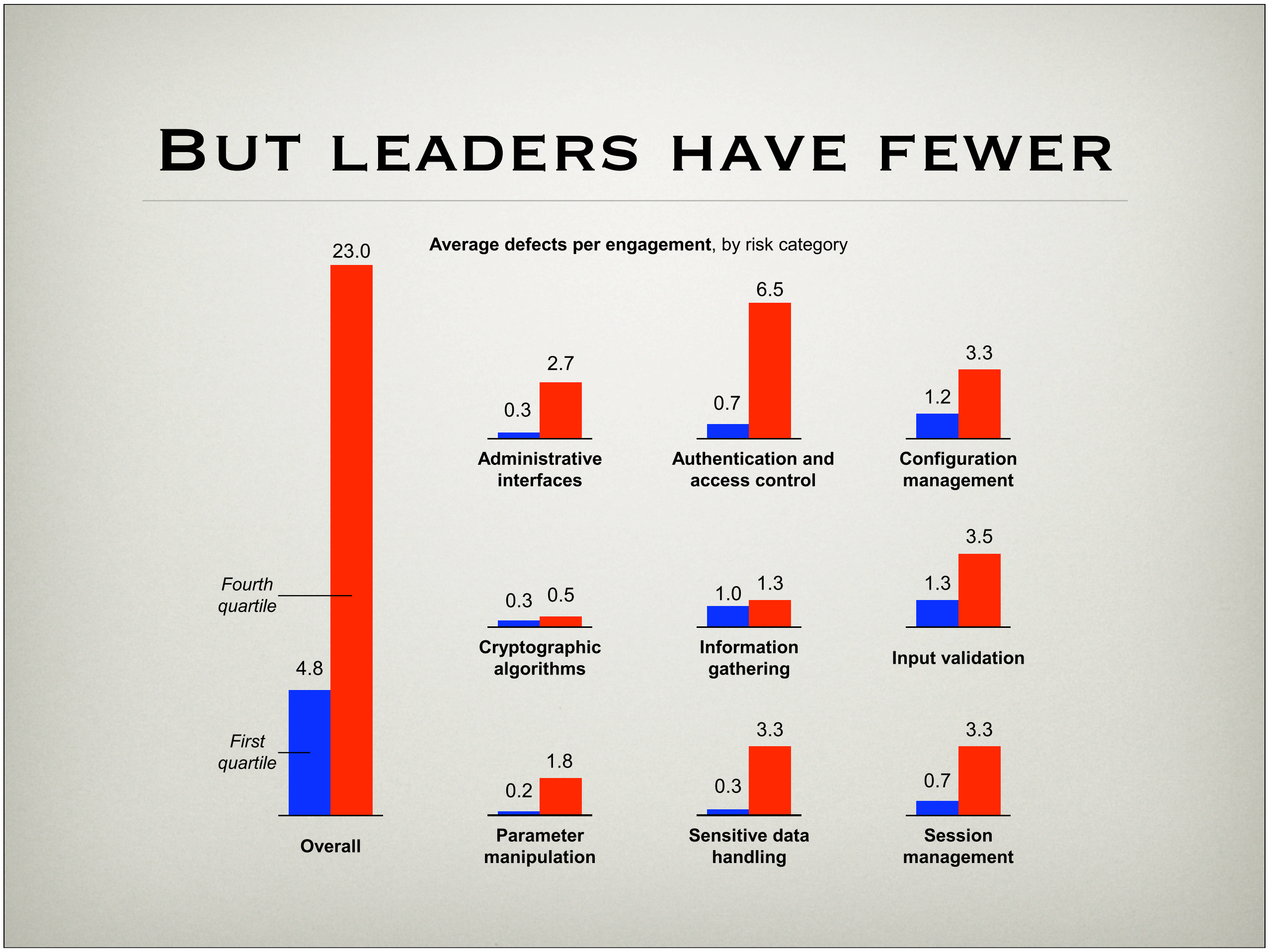
Category	Engagements where observed	Design related	Serious design flaws*
Administrative interfaces	31%	57%	**36%**
Authentication/access control	62%	89%	**64%**
Configuration management	42%	41%	**16%**
Cryptographic algorithms	33%	93%	**61%**
Information gathering	47%	51%	**20%**
Input validation	71%	50%	**32%**
Parameter manipulation	33%	81%	**73%**
Sensitive data handling	33%	70%	**41%**
Session management	40%	94%	**79%**
Total	**45**	**70%**	**47%**

*Scores of 3 or higher for exploit risk *and* business impact



First finding from pooling: Security defects are common. The nine categories have sixty particular defects amongst them (genus & species). The rightmost column is how many engagements had any occurence of each of the sixty, with the result ordered by the percentage thereof.

It also illustrates that when doing measurement you actually do have to make assumptions: If a design says “must be resistant to hostile input” and the implementation is vulnerable to hostile input, then that is treated here as an implementation fault. If, on the other hand, the design is silent on hostile input and consequently the implementation is vulnerable thereto, then that is treated here as a design fault. Oh, and a “serious design fault” is one which produced an above-median risk which, if exploited, also produced an above-median impact.

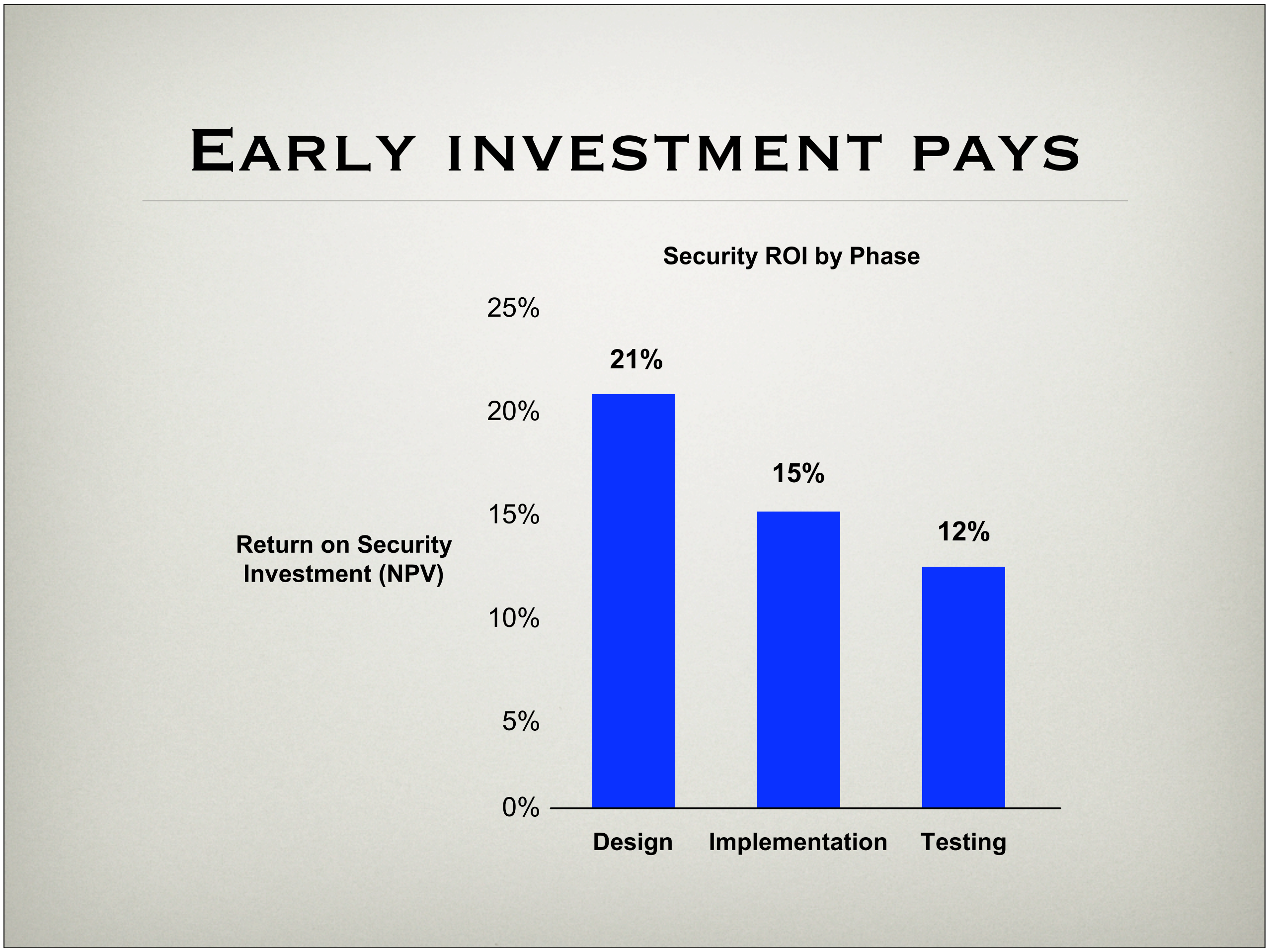


Second finding: There is a real difference between the top quartile and the bottom quartile, which is all about heightening contrasts.

In the first finding, it was a binary decision of whether an engagement found any of a flaw type. In this finding, we count flaws per engagement. Then, for each type, we divide the population into quartiles (four equal sized buckets). By comparing the lowest quartile to the highest, you then get a sense of spread and range for the measure at hand. In this case, it is the count of flaws per engagement by each of the nine categories (genus) of risk. As you can see, incidence and ratios of incidence vary a lot. Crytographic algorithms have a small range of flaw density while the greatest ratio is that of Sensitive data handling. If you assume that your environment is a leader, then the above suggests you might concentrate your efforts on Input validation; if you assume that your environment is a laggard, then instead you might focus on Authentication and access control. And so forth.

Third finding: Leaders not only have fewer out and out flaws, it translates into less business risk.

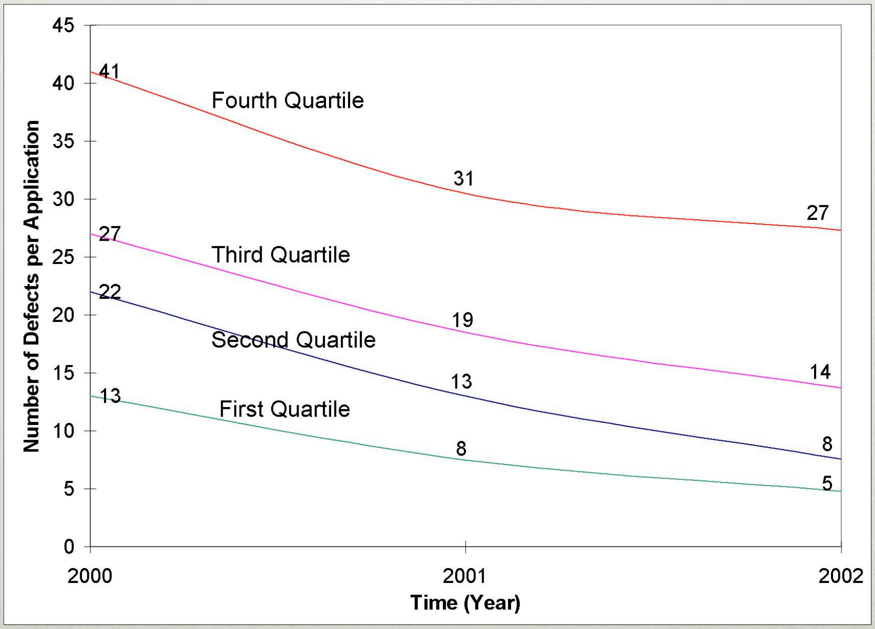
Business risk was assessed on an ordinal scale, 1-5, with 1 as lowest risk and 5 as highest. We invented the scale, using an odd-number of categories (which is recommended) and in parallel did one for ease of exploit and one for business impact. After assignment of a score to each vulnerability on both those scales, we summed up the risks for a composite "business adjusted risk score." While the details do matter, see the paper for them. The point is that we then compared BAR scores, again by quartiles, in the form you see above (which is called a "waterfall" graph). Now we're getting somewhere: If we can price the cost of moving a given system from the bottom quartile to the top, we can say that the cost effectiveness of doing, say, better session management is $XYZ for an 88% reduction in risk versus $ABC for a 20% reduction in risk for better information gathering.



Fourth finding: Early money is better than late money, an is so as measured by returns on security investment.

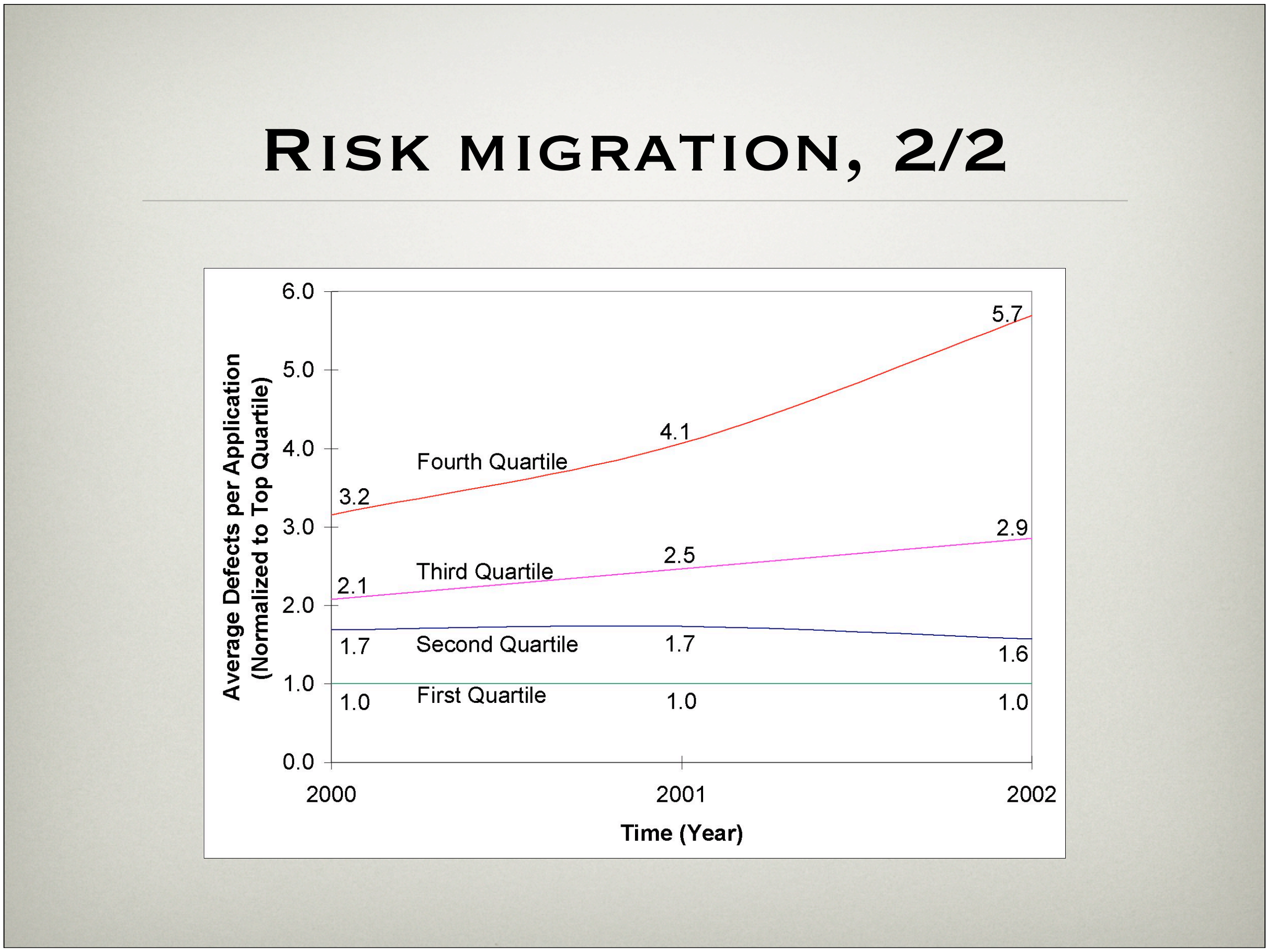
Going back to the 1981 IBM study and the 1998 CMU course notes, we combine the fees charged by the consulting firm for the risks found. Since clients hired the consulting firm at various stages, we can say whether the dollars involved in that hiring were expended at one of the three stages of product lifetime. Looking at BAR reductions but asking how the money would have changed had the work been done at a different stage of product lifetime, we finally come to a net return curve that looks like this. It is likely that 21%/15%/12% are wrong, but the shape is right -- it really does pay in classic economic terms to find your flaws early. Real data.

Risk migration, 1/2



This is the amount of risk per assessment, separated by quartiles, and graphed over time. The news is apparently quite good with advances on all fronts.

That work continued over three years during which the four quartiles of customers all cleaned up their acts, presumably under the influence of the consulting reports they were getting. This is good news. The three years shown are simple what was covered in the referenced paper; the work continues.



This is the same graph, but normalized to the best quartile. Now you can see that although risk is declining for each quartile (the good news), the first quartile is getting better faster than the fourth quartile and thus the ratio between the best and the worst is broadening over time.

The implications of this are arguably profound -- If you are doing a good job at this (systemic) risk reduction, then the fraction of your total risk that is due to your counterparties (the unique risk) is rising. Were we talking about medicine we could doubtless agree that if we were to cure heart disease then cancer would become even more important than it now is; that is what you see here. By normalizing to the best quartile we have removed nearly all measurement artifacts that might affect our inferences; the inferences remaining are perhaps weaker but less likely to be artifacts. Divergence of risk is a solid finding.

Portfolio lessons

- Need broad market measures
- Aim of analysis is to heighten contrasts
- It is possible to price risk
- Define your risk first, then your metrics

To have a portfolio measure of any sort you need a broad measure of the market within which the portfolio lies. Whether this is direct data sharing, implicit data sharing with a common truste third party, or is limited to divisional difference within a single enterprise, you must have an aggregate comparand.

Regardless of the comparand, the point of analysis is to heighten contrasts. This is directly consistent with looking for leverage in a set of candidate financial transactions. The examples given here -- of quartile, waterfall, non-parametric ordinal assignments to categories, etc. -- are just examples. There are many alternatives. Do some exploratory data analysis.

The bottom line is that it is possible to price risk, even if (as was shown) what you are pricing is relative risk reductions against a baseline for which there is no known calibration.

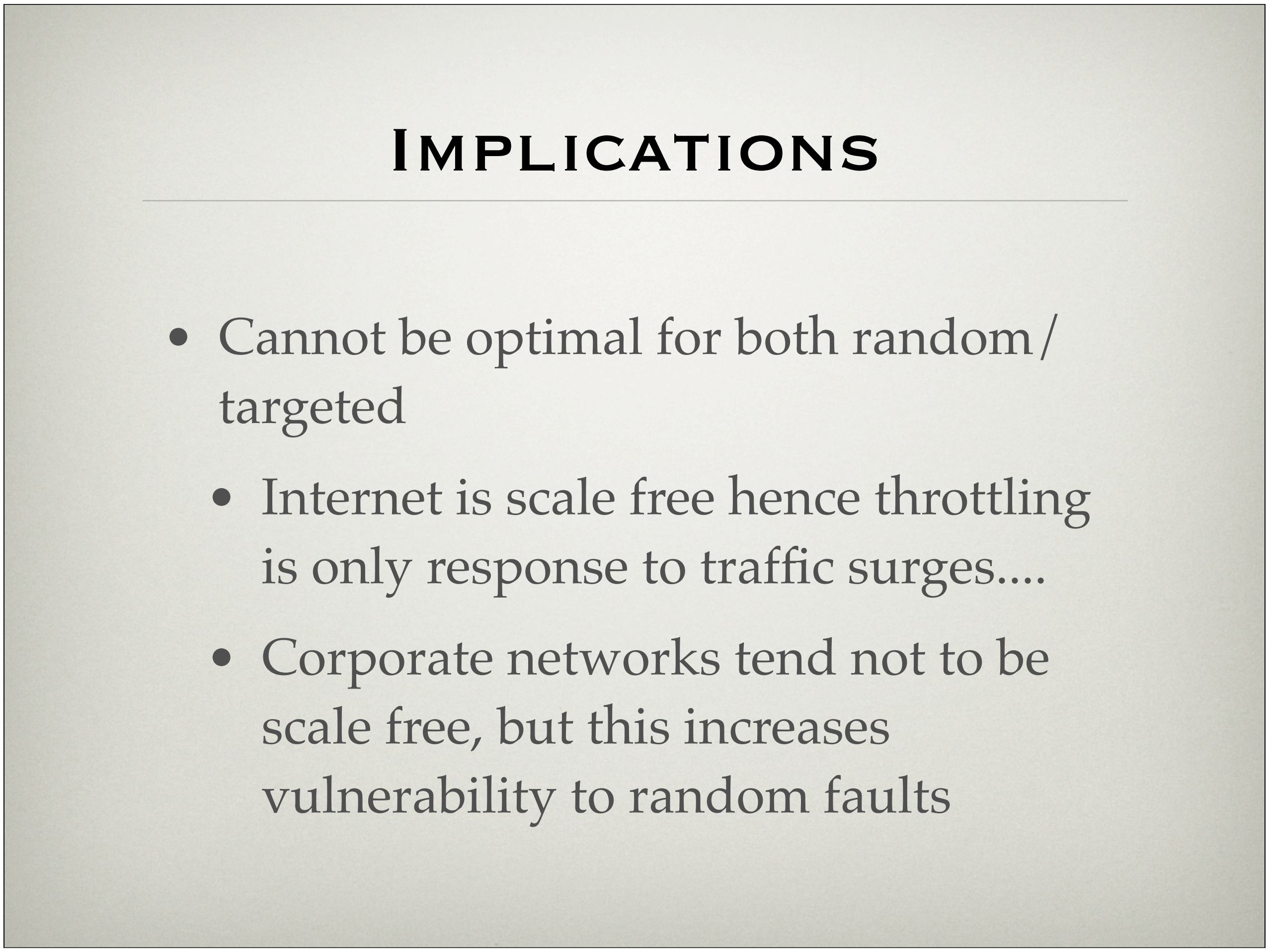
Physics

Physics of nets

- Random connectivity – like it sounds
  - Maximal resistance to targeted faults
- Scale free – looks the same at any scale, like fractals
  - Maximal resistance to random faults

A rather startling result in the physics literature has mathematically shown that a network design has to trade off vulnerability to random faults and vulnerability to targeted faults, that it is not possible to be maximally resistant to random component failure without creating the conditions in which targeted attacks cause outscale connectivity losses just as it is not possible to be maximally resistant to targeted attacks without creating the conditions in which random faults cause outscale connectivity losses.

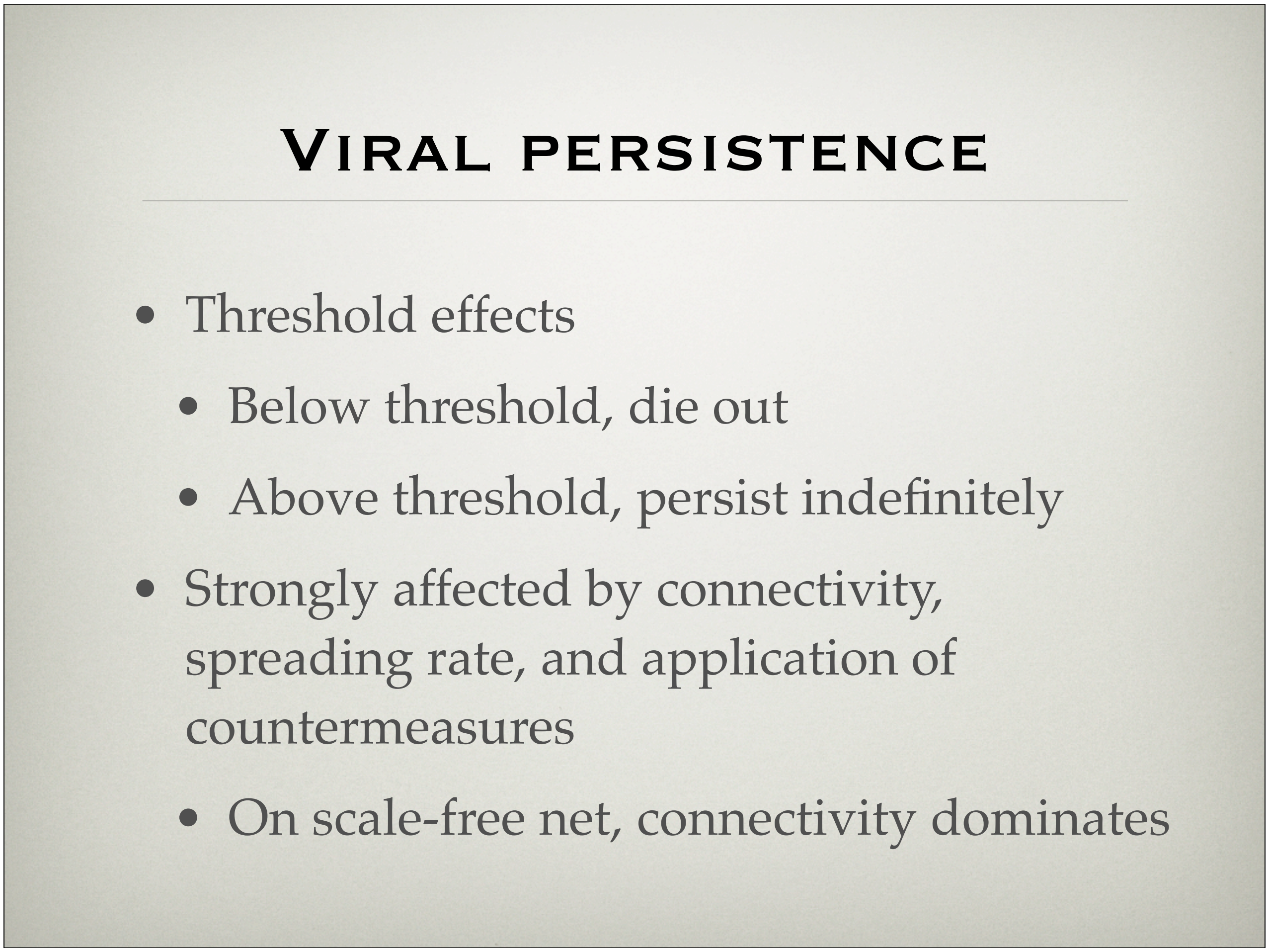
See a short discussion at http://en.wikipedia.org/wiki/Scale-free_network or Albert-Laszlo Barabasi's book, Linked: How Everything is Connected to Everything Else, Morgan Kaufmann, 2004.



This is the impact of the insight in the physics literature on scale-free networks. The claim that the Internet is scale free is in fact true -- measurement of Internet connection patterns is what brought the original authors to the conclusion that the Internet was scale free, not the other way around (that is, the measure was not to confirm theory; rather the theory grew out of measurement).

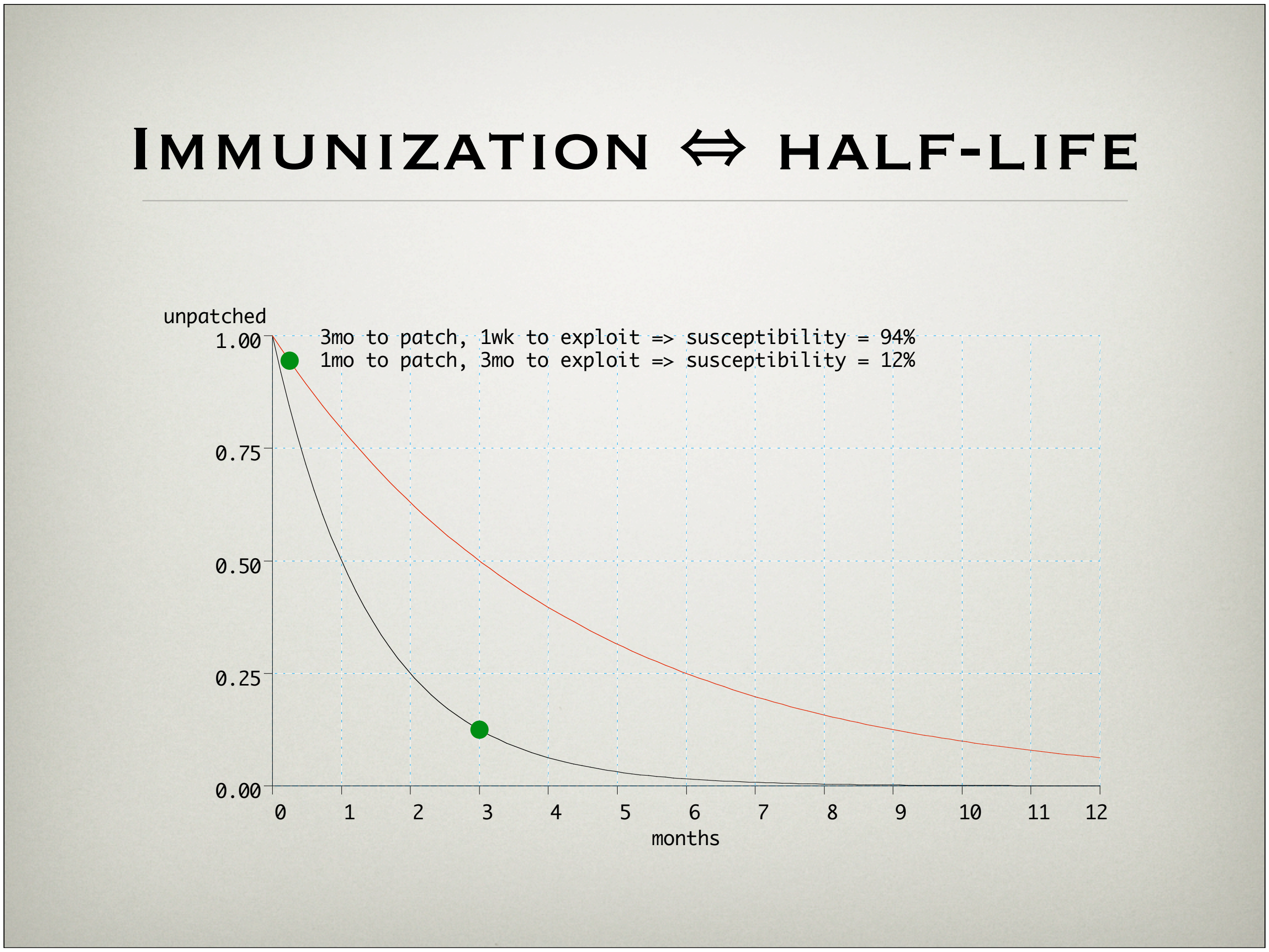
If a network is intentionally scale-free, then targeted faults can have substantial impact. Mitigating that means having some mechanism to throttle demand, and that is the case in many commercial ISPs who will not let traffic volume rise too steeply whether inbound or outbound (relative to their peering points with other ISPs).

Corporate networks tend to be designed a bit more, not accreted by the near-random process that grow scale-free networks. As such, they may well be more resilient to targeted attacks but by the theoretic result, this means that they have a compensating rise in vulnerability to random faults, perhaps explaining the necessity for a network operations center (NOC).



None of this should be surprising, but the physicists develop these ideas with particular rigor. The reason they find their results surprising is that with the scale-free property connectivity so dominates that it becomes likely that old viruses never die, i.e., there is not really a minimum threshold of infection required to sustain a virus' presence in the Internet at large.

Source: Pastor-Satorras R & Vespignani A, "Epidemic spreading in scale-free networks", Phys. Rev. Lett. 86, 3200, 2001.



Posting a patch starts a race wherein the patch is reverse-engineered to produce exploits. The two data points are intended to bracket current reality. In the one case, if patching does have a one-month half-life while the reverse engineering interval is 90 days, then the susceptibility would be 12% at the moment of exploit. By contrast, if patching has a three-month half-life while the reverse engineering interval is one week, then the susceptibility would be 94% at the moment of exploit.

Time-to-exploit is shrinking while the time-to-patch is lengthening (if you factor in the growth of always-on, always-connected home machines) so the question becomes whether “mandatory” is a word we must use and, if so, what would it mean?

Implications

- Like thermodynamics
  - Can’t win, break even, or get out of game
- Perfection is clearly unobtainable
  - Relative vulnerability (ordinal scale) works
- Thresholds exist

With the brusqueness of physics, the point is obviously that security will not be perfect hence relative vulnerability is likely to be the actual measure of choice. As said at the outset, a relative vulnerability focus is admitting that an ordinal scale is all we are going to get or, in brighter language, we are able to get an ordinal scale and with that there are lots of things we can do.

Physics shows us that there are thresholds, e.g., for viral persistence, for connectivity as both a value and a source of risk, and so forth. Having physics to occasionally fall back on is actually reassuring as nothing else has the same rigor, the same swagger, as physics.

Physics: theory v practice

- Does scale free actually happen?
  - Not exactly, but almost
- What does happen?
  - Design optimality for use cases but otherwise scale free

This scale free network model may or may not apply to real networks. It probably does not wherever policy tends to trump free choice of interconnection. However the lesson that optimality tradeoffs around what sort of threat you are resistant to and what are you not is worth repeating. The reference below, which hard reading, adds that you can do better in designing a network for effective bandwidth and resistance to faults but only if you design for that rather than permitting random interconnection. This remains an area of theoretic debate, but there are lessons to be learned now and no doubt lessons to be learned later.

Doyle JC, Alderson DL, Li L, Low S, Roughan M, Shalunov S, Tanaka R, and Willinger W : "The 'robust yet fragile' nature of the Internet," Proceedings National Academy of Sciences, v102 p14497-14502, http://www.pnas.org/cgi/reprint/102/41/14497.pdf

Use in security

- Must choose what to optimize
- Time constants and connectivity matter
- Make models you can test

As the scale-free versus designed network discussion shows, optimizing for one variable not unsurprisingly may well de-optimize another. So long as you are watching all the dials, that is no problem. So watch all the dials.

What physics has told us is mostly about the interplay between connectivity and time, which are both perhaps related to propagation of change whether that change is for the better (as in a patch management system) or for the worse (as in a geometrically propagating worm). Physics also tells us the importance of having a testable theory, a sense of the big picture yet in simple terms. That is hard to do, but it is so powerful when it obtains.

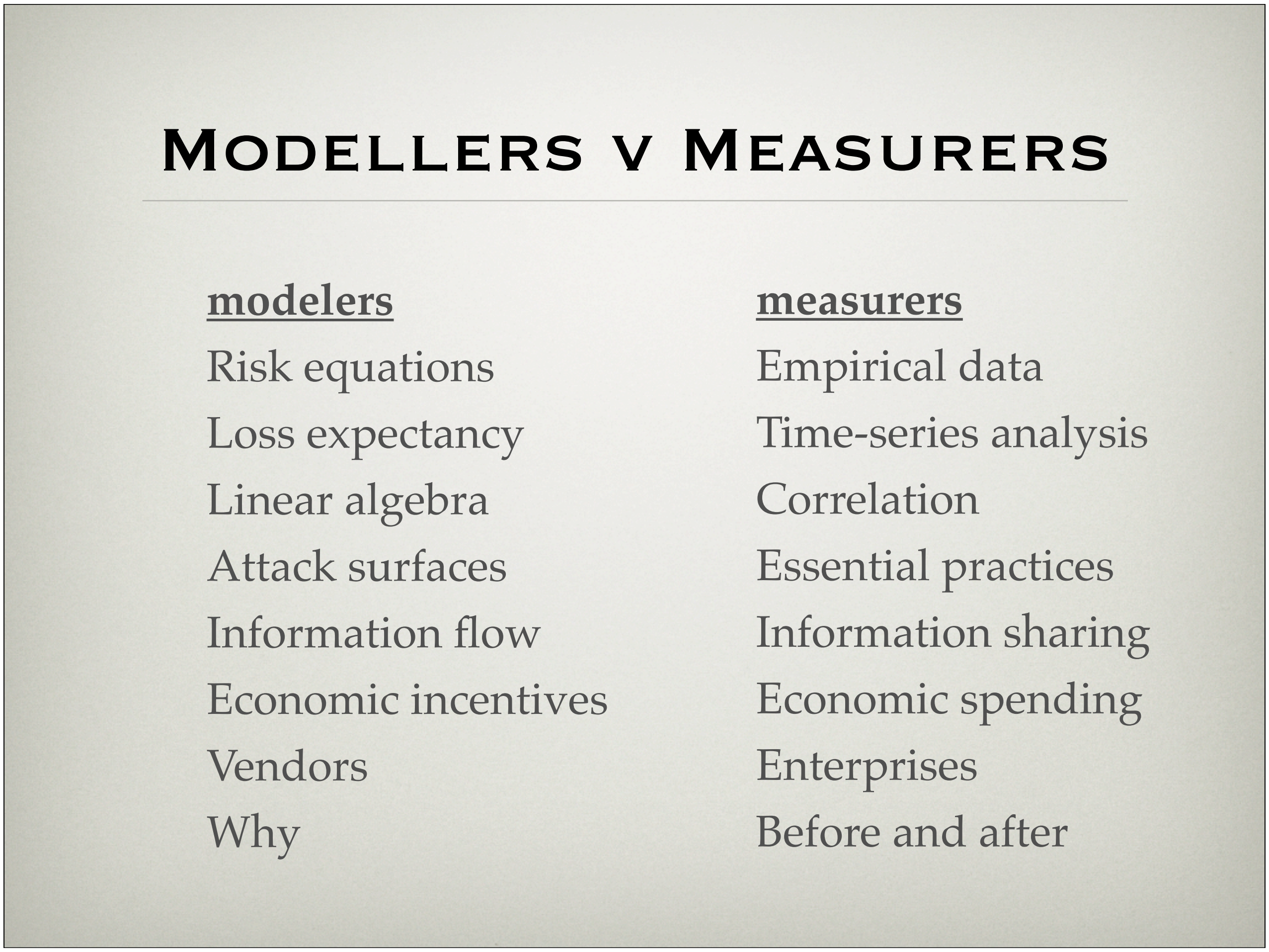
Other

Other areas to mine

- Hurricane models; property & casualty insurance *v.* building codes
- Bio-informatics applied to protocol analysis
- Sensor networks
- “Value at Risk” simulations

Perhaps in the next revision of these notes we will explore all of the above, which is, as well, not a complete list. The field is wide open to you to innovate yourself.

See http://www.wired.com/news/infostructure/0,1377,65191,00.html for more ideas.



As you think on these topics, ask yourself if you are a modeler or a measurer. The discuss@securitymetrics.org mailing list has -- with this result of how to tell.

Yes, this is only the security field's version of Isaiah Berlin's famous essay which is highly recommended; Berlin I : The Hedgehog and the Fox, Simon & Schuster, 1953.

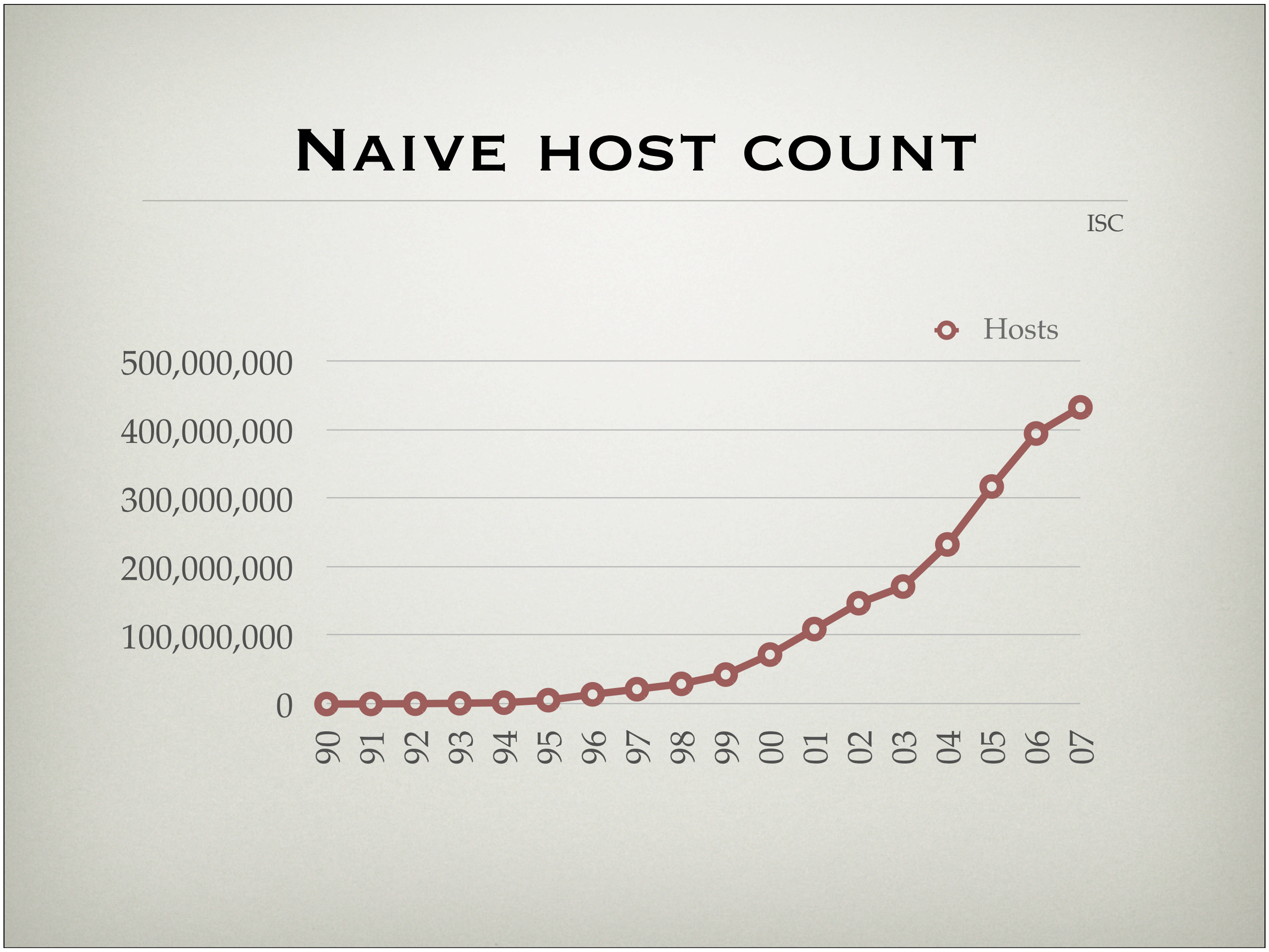
Trend Analysis

Trend

OECD

A long-term movement in an ordered series, say a time series, which may be regarded, together with the oscillation and random component, as generating the observed values

http://stats.oecd.org/glossary/search.asp



Public ISC data; http://www.isc.org/index.pl?/ops/ds/host-count-history.php

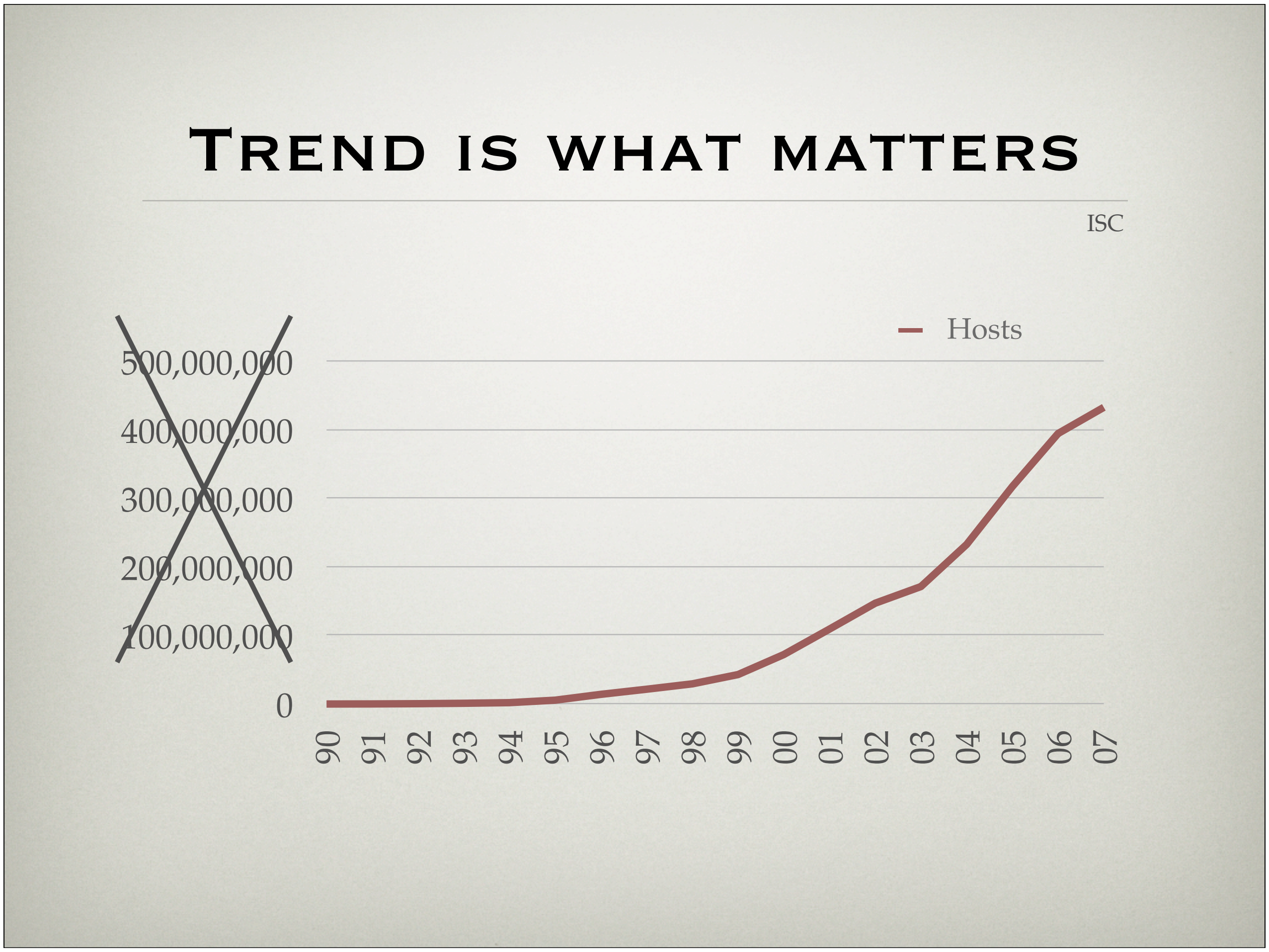
... in thousands

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
236	376	727	1,313	2,217	5,846	14,352	21,819	29,670	43,230	72,398	109,574	147,345	171,638	285,139	353,284	439,286	433,193

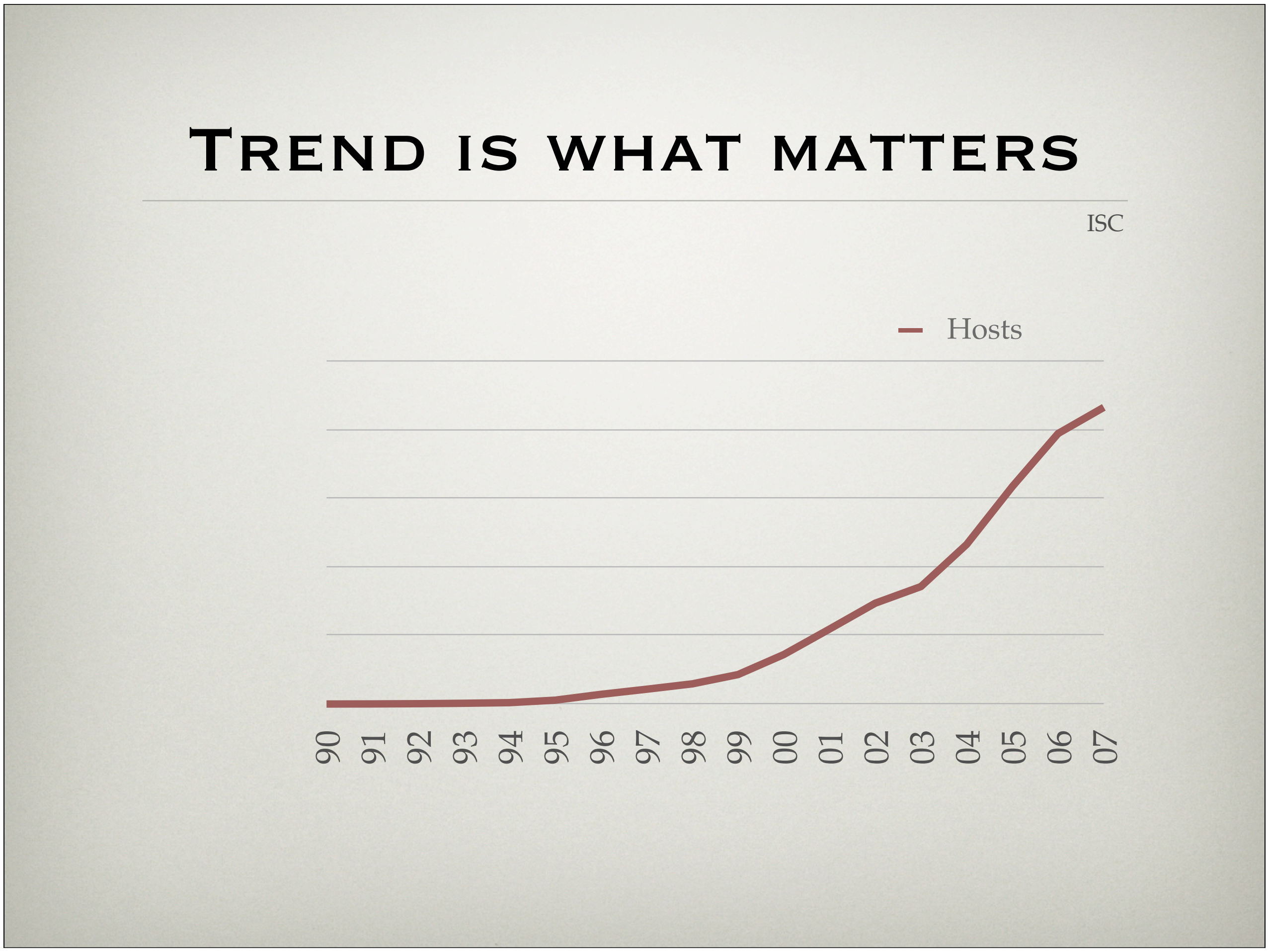
Selection bias here?

- Open question
  - NAT causes under-estimate
  - Multi-homing causes over-estimate
- Are the above fractions changing?
  - If so, there is selection bias

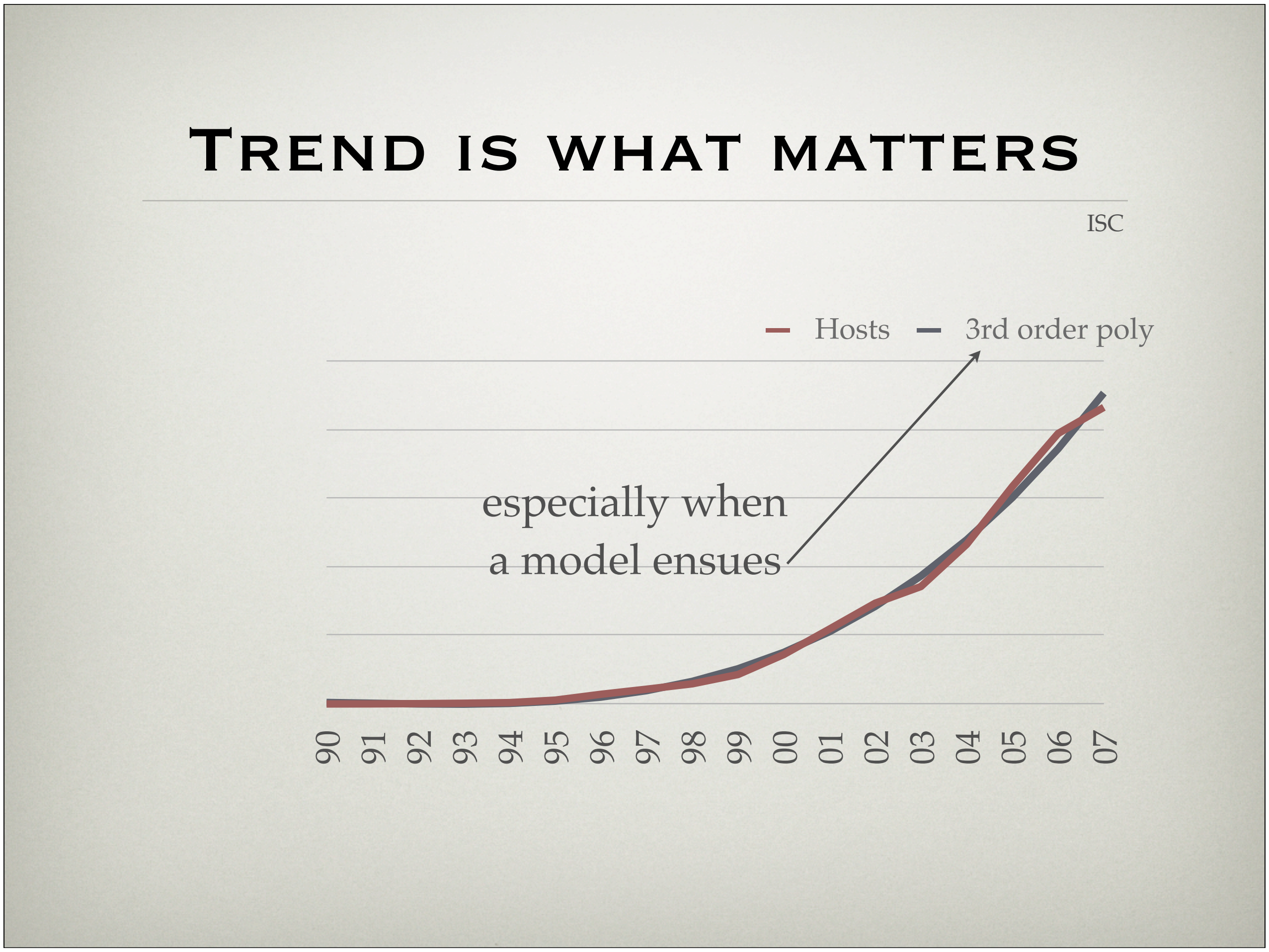
So, does the estimate of total Internet hosts exhibit selection bias? Of course it does: network address translation (NAT) makes a raft of hosts appear as one while multi-homed hosts, having as they do multiple addresses, can cause over-estimate. However, if either or both of these is true it is of no import so long as the fractions are relatively stable.



It is the trend that matters. Ignore the ordinate (Y axis) and look at the shape.



Without an ordinate, it is easier to look just at the shape.

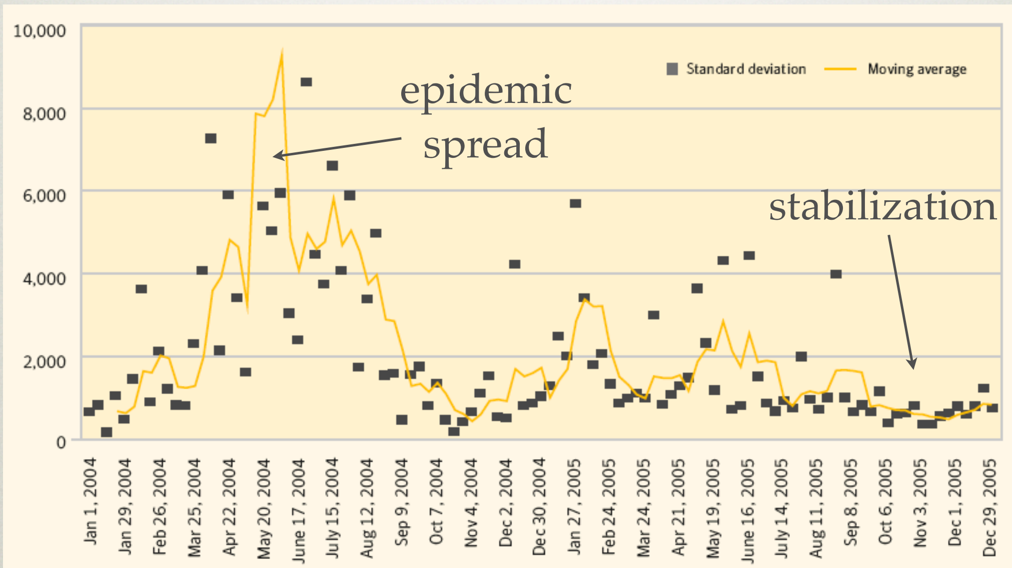


Trend is what matters, especially when a model can be fit to it, as is the case here.

By the way, that is a pretty good fit: R^2 = .9976 (coefficient of correlation between the observed data and the fitted curve).

Cyclic trends, e.g., bots

Symantec

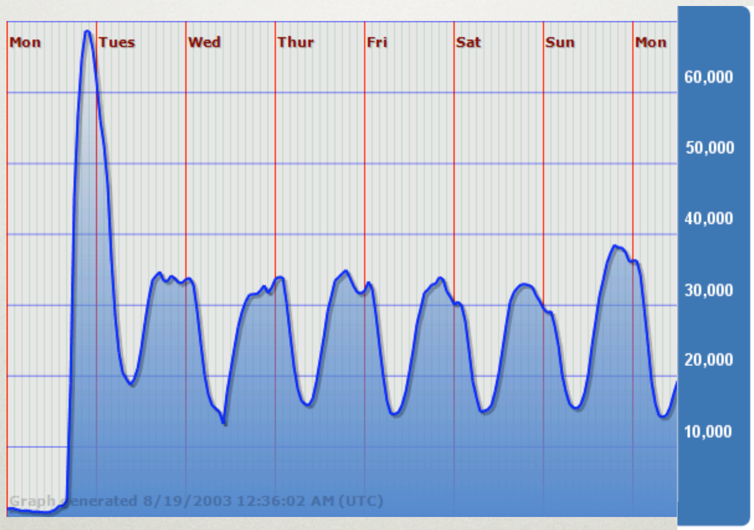


Sometimes trends are cyclic. Symantec views robot network (bot) recruitment as cyclic, here with a period of epidemic spread followed by a period of stabilization until some new attack method appears making possible another cycle of recruitment.

Symantec Threat Report IX, March 2006. Not open source.

HARMONICS, E.G., PEOPLE

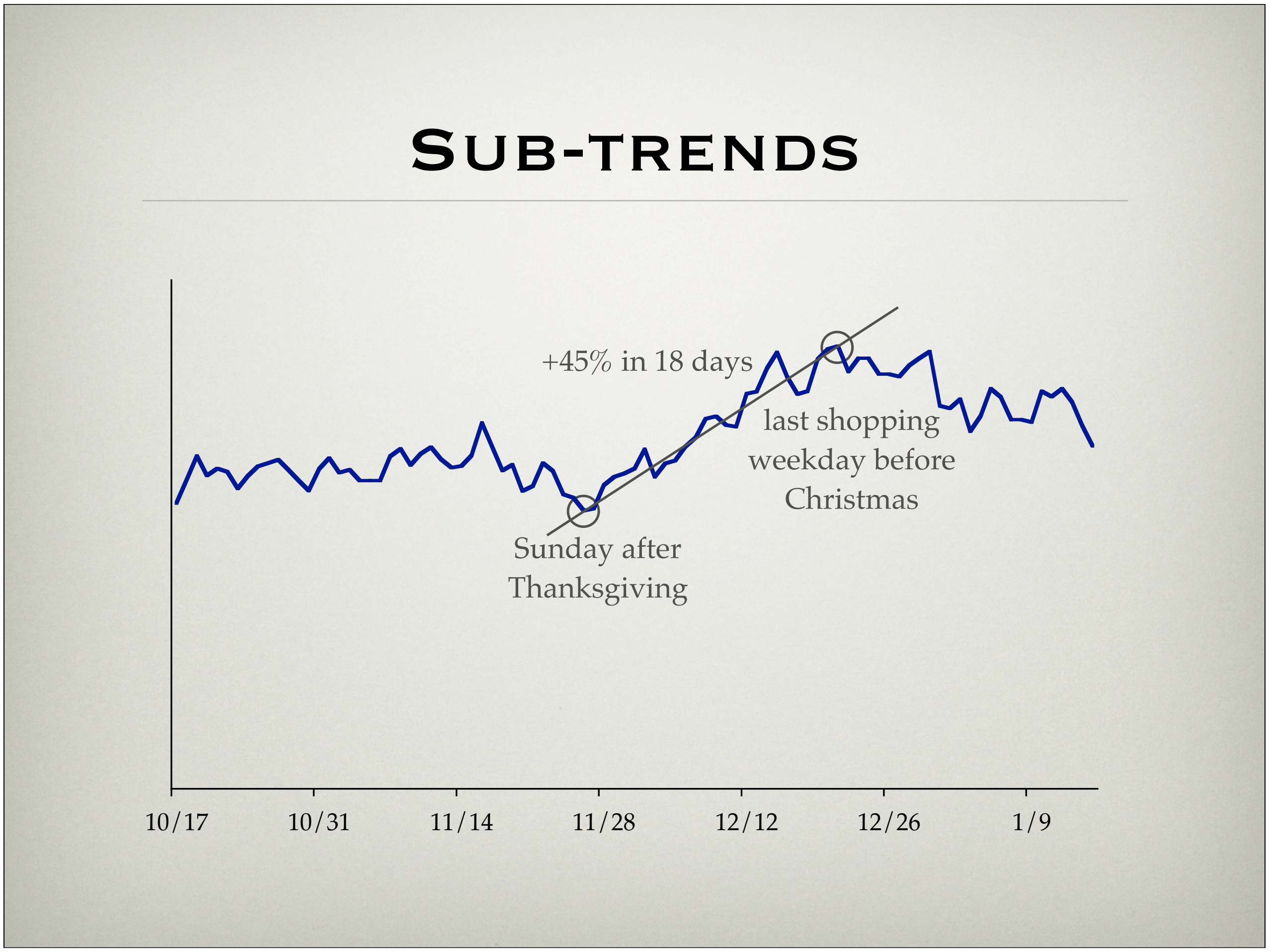
McAfee



Unique Attackers per Hour with Blaster

Trends can also be harmonic; here the attack graph for Blaster as provided by McAfee and clearly showing differences between when people are awake versus when they are asleep.

http://www.hackerwatch.org/checkup/graph.asp
http://www.hackerwatch.org/img/map/worm.png



So, does this show an all-out assault on the Internet Christmas shopper? If not, what does it show?

The point is that trends can occur in smaller-than-full intervals though, of course, you have to be careful not to over-read the data here. This is the NVD workfactor data that we will come to later, but the illustration done with it belongs here.

Trends do matter

- If the street price for meth is declining, Then L.E. is losing at drug control
- If price for stolen data is declining, Then we are losing at data security

Trends come up in everyday life all the time, such as the example above where law enforcement uses street prices for drugs as a calibrator on whether their control efforts are winning or losing.

Since the price of stolen data seems to be falling, we might as well face up to the fact that we are losing our control problem.

Use in security

- Nearly any reproducible metric that has meaning to you can be looked at as a trend
- Trend analysis is a component of decision making, particularly in the case of cost-effectiveness-based decisions

In security, trends are going to often be the best we can do and they are consistent with ordinal scale measurement. As has been said before, if decisions can be made on that basis, trend analysis is good enough and particularly so for cost-effective decision making.

TANSTAAFL

"There Ain't No Such Thing As A Free Lunch" from Heinlein RA : The Moon Is a Harsh Mistress, 1966, which, incidentally, was adopted as a title by economist Milton Friedman.

A central idea

- Cost effectiveness, yes
  - Cost of improvement trend
- Cost benefit, no
  - Cost of intangibles, per se

This is an idea we will now elaborate.

Cost-benefit

$$CB_{ratio} = \frac{Cost_{new\ strategy}}{Benefit_{new\ strategy}}$$

$$CB_{ratio} < 1.0 \Rightarrow \text{favorable}$$

Cost-benefit ratio is, surprise, the ratio of a cost to the benefit it provides. This is valuable if it is less than 1.0, i.e., you get more benefit than your cost was for getting that benefit.

Cost-benefit

- CB asks if you want to spend the money
- Requires pricing benefits in $$
  - How much is a human life worth?
  - High quality timber *v.* wilderness?
  - Cheap housing *v.* code compliance?

Cost-benefit analysis requires pricing the cost and the benefit on a common scale so that you can ask whether you would rather have the money (avoid the cost) or the benefit (incur the cost). This can be hard.

Cost-effectiveness

$$CE_{ratio} = \frac{Cost_{new\ strategy} - Cost_{current\ practice}}{Benefit_{new\ strategy} - Benefit_{current\ practice}}$$

$$CE_{ratio} < 1.0 \Rightarrow \text{favorable}$$

Cost-effectiveness analysis asks how much benefit can you get for how much cost.

Cost-effectiveness

- CE assumes you will spend the money
- CE asks how well you can spend it
  - $10B: safer cars *v.* law enforcement?
  - $1M: 100% uptime *v.* instant recovery?
  - $100: 1 fine dinner *v.* 20 lunches?

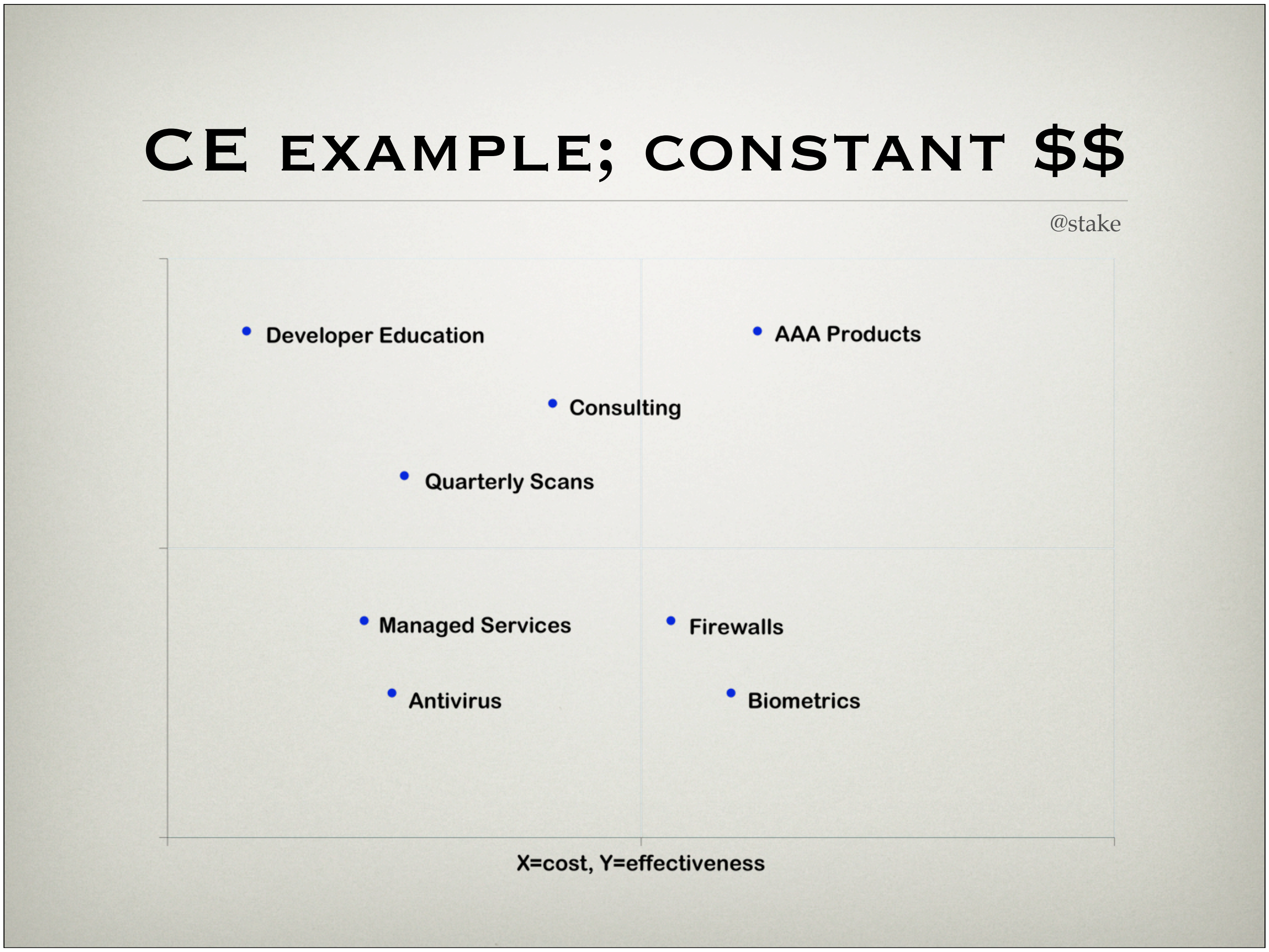
Cost-effectiveness assumes that you will, indeed, spend the money and thus your interest is in how much you can get for your money, not whether you'd rather keep your money in the first place.

CB v. CE in security

- Today's job in measuring security is to enable cost-effective decision making
- We will not answer "What is the value of security?" but rather "How much security can I get for reasonable $$?"

Putting a stake in the ground, it isn't whether one would rather keep the money or get the benefit (CB) but rather what good can you do for the budget dollars you have.

CE is always tractable; CB is only tractable or stable when the conversions of benefits to dollars are stable.



With fictitious data, this is a guess as to how you might look at a set of CE options. In this picture, one would see that the cost effectiveness of Developer Education is very good indeed, while far from good for biometrics. If your budget situation were that you spent no more than $X, rightward parts of this option graph might disappear.

Total cost is a mix

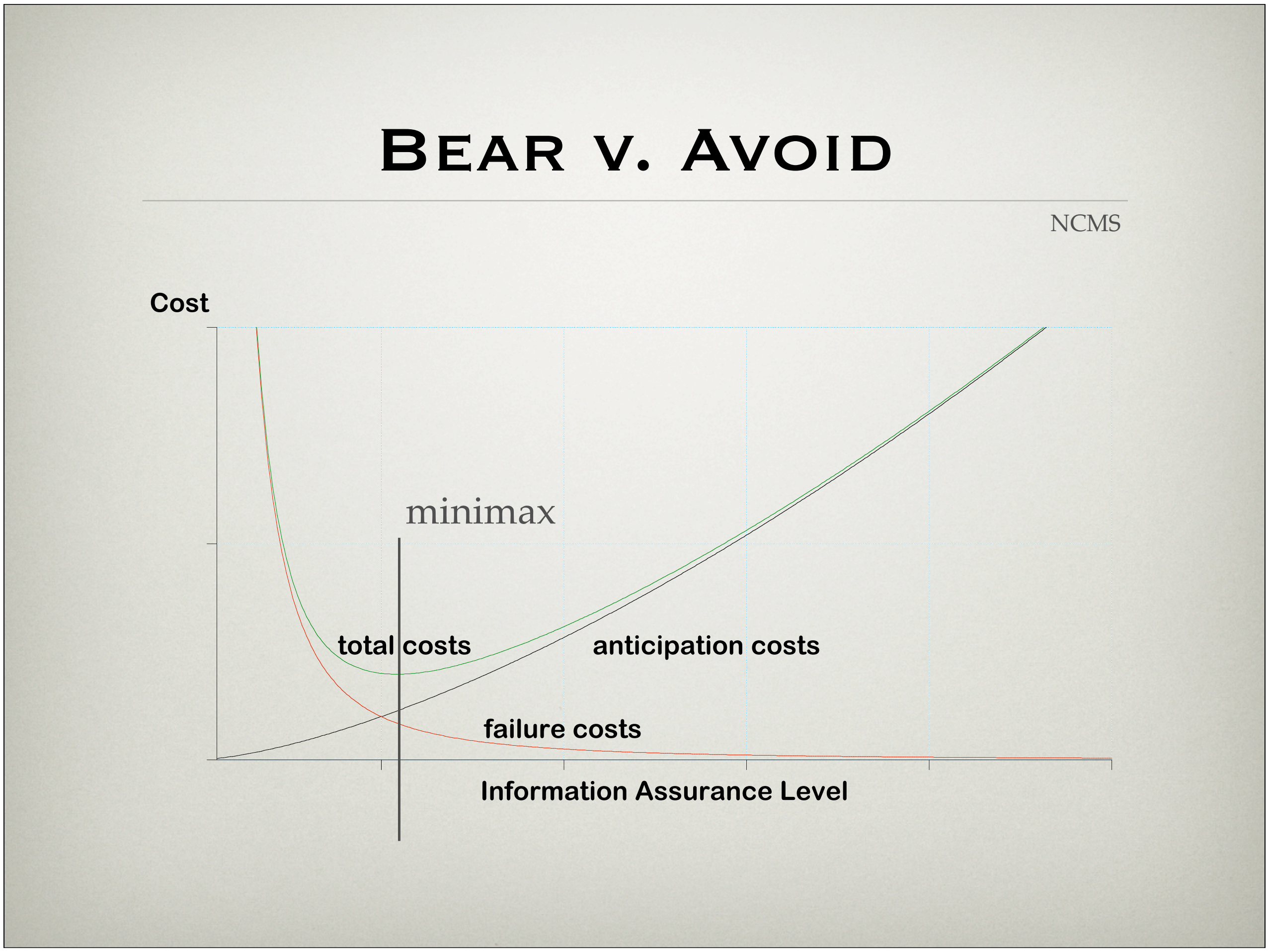
- Anticipation costs – what you spend to avoid trouble, i.e., prevention
- Failure costs – what you spend to clean up from trouble, i.e,. recovery
- Total cost is the sum of anticipation and failure costs

When talking cost, it is good to make sure that you are talking total cost. For security, these costs are of two classes, costs expended to prevent trouble (anticipation) and costs spent to clean up from trouble (failure). Total cost is the sum of both.

Minimax solutions

- You want the maximum advantage for the minimum cost
- You want the most cost-effective strategy for the cost you can endure

The goal you seek is to maximize on variable, the benefit, and to minimize another, the cost. This is what cost-effectiveness seeks to provide. Economists would likely call this optimality and be done with it.



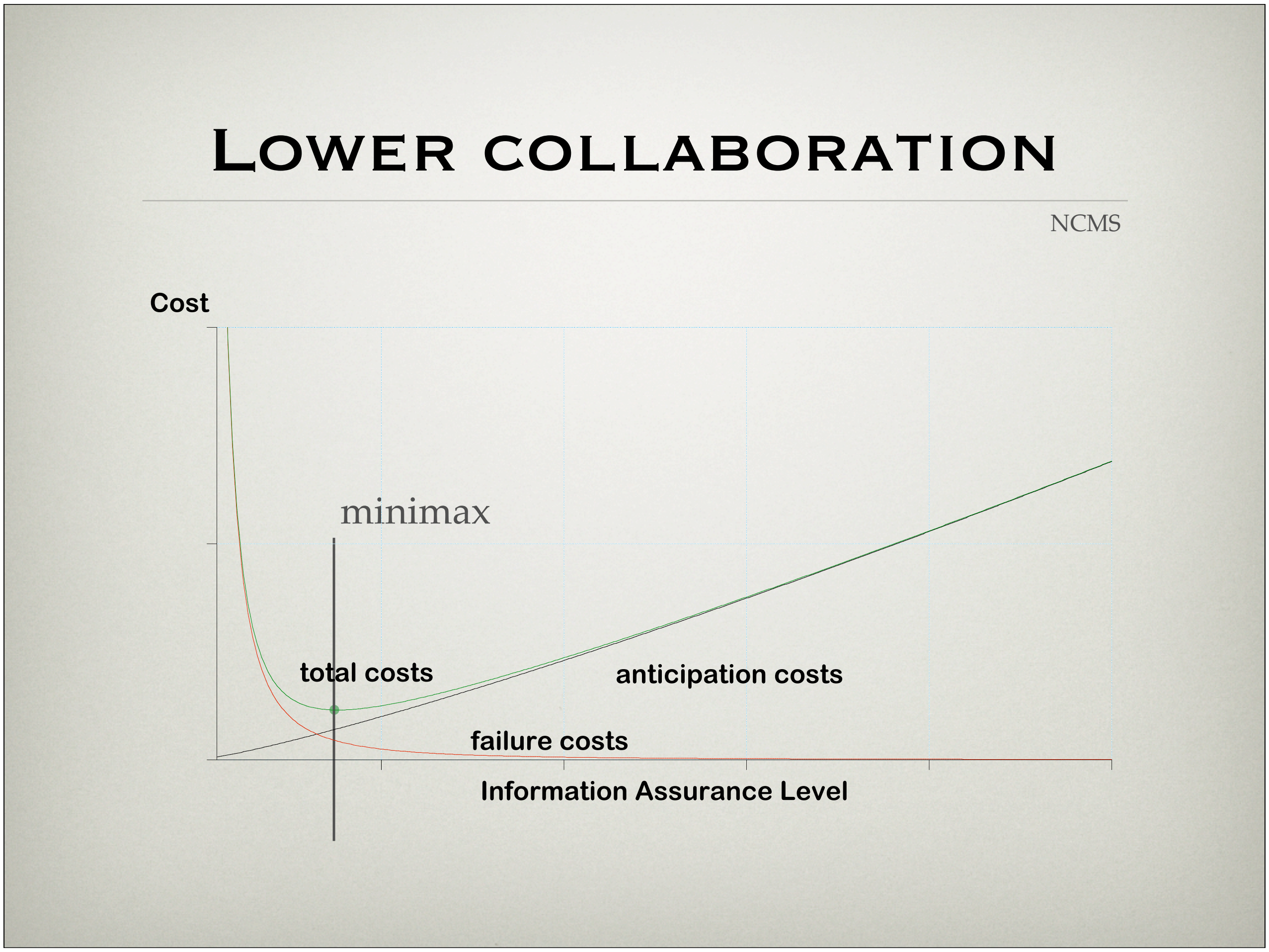
Risk transfer is about trading one risk for another; that can be internal as well as external. This picture does not specify, but it illustrates the tradeoff between anticipation (prevention) costs and failures (mitigation) costs. The total cost is the sum of the two and, as the graph shows, spending nothing on anticipation maximizes failures costs just as spending too much on anticipation minimizes failure costs. The saddle point is your management target.

Source: "Costs of Information Assurance," National Center for Manufacturing Sciences, August, 2002; see http://trust.ncms.org/pdf/CostInfoAssur-NCMS.pdf

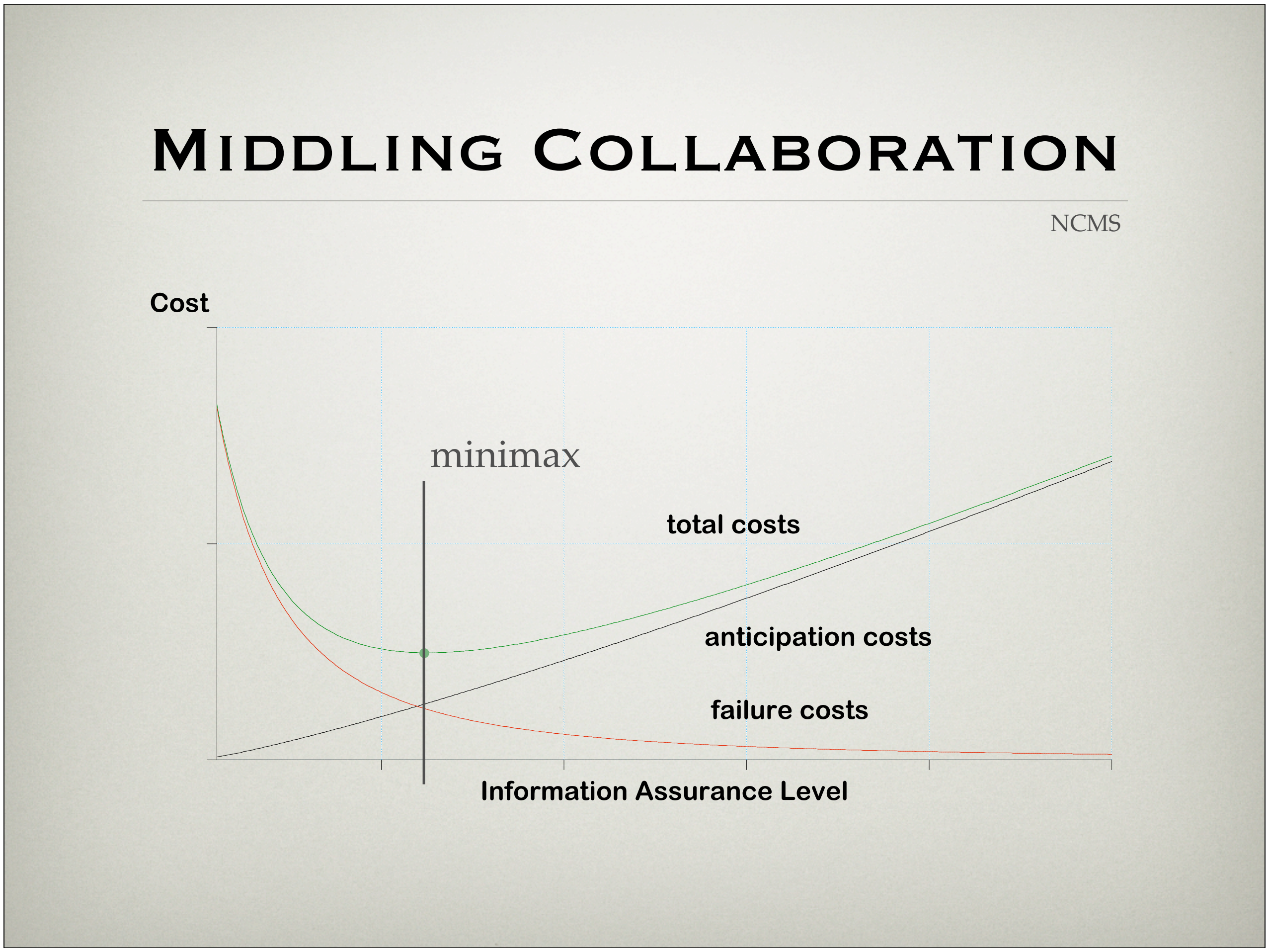
Setting matters

- Little collaboration
  ⇒ low failure cost ⇒ spend little
- High collaboration
  ⇒ high failure cost ⇒ spend more

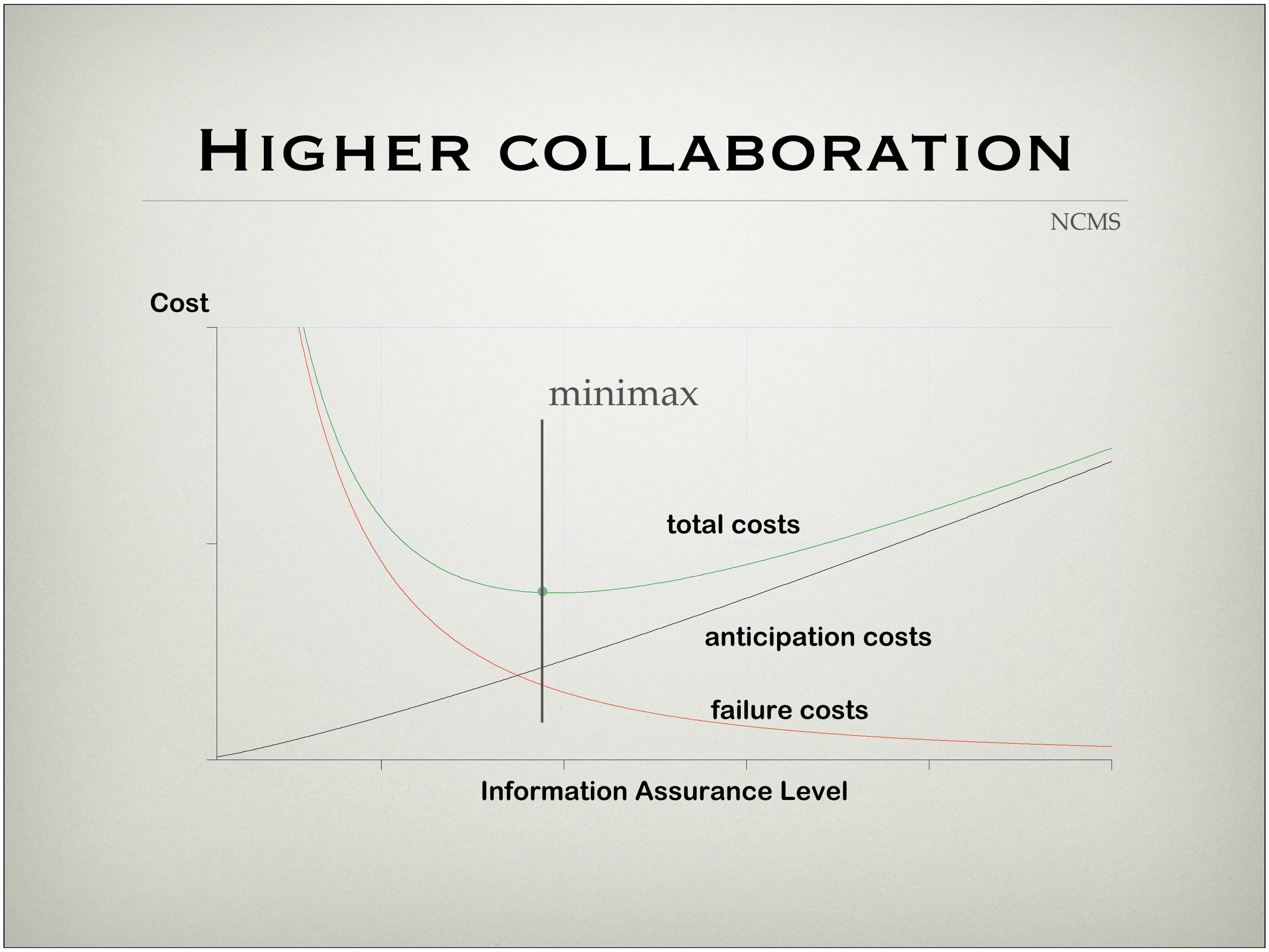
One of the things that NCMS points out well is that the level of collaboration you have with your customers, suppliers, and other counterparties affects the cost of failure should you be unable to have that collaboration. If you have little collaboration, you can be offline, say, at little effect. If you have a high degree of collaboration, the effects of being offline are more profound. Were these true, you might have to adjust your spend up or down to reach optimality.



So at low collaboration, the total cost has its minimax point where anticipation costs are minimal because failure costs are also minimal.



At middling collaboration, the failure costs have risen so the minimax point has moved rightward.



At high collaboration, more money still must be spent on anticipation if the minimax point is to be achieved.

All this is obviously of direct application to nearly any security setting. If you have a loss you are willing to eat (like an insurance deductible), then you can set your protection level accordingly. You can transfer some risk and anticipate other risk while bearing yet another. More to the point, your business people will be able to have a conversation like this.

Decision support

Return on Invesment

- Of course you would like to show this
- Of course it is hard
- This where Cost-Effectiveness comes to your rescue

Let's start with a non-security example...

Attempting true rigor in calculating return on security investment (ROSI) can be a time-sink but it is also a fundamentally valid question in a risk management world.

CE and ROI decisions

NCPA

By spending $227,000 every year for sickle cell screening for unscreened black newborns, we add 961 years collectively to their lives at a cost of $236 for each year of life saved.

"Dying Too Soon: How Cost-Effectiveness Analysis Can Save Lives, National Center for Policy Analysis," Washington, D.C., 1997, available at `http://www.ncpa.org/studies/s204/s204.html`, and, in particular, Table VI at `http://www.ncpa.org/studies/s204/table6.gif`

CE and ROI decisions

NCPA

By spending about $460 million per year on heart transplants, we add about 2,900 years to the lives of heart patients at a cost of $158,000 per year of life saved.

ibid

CE and ROI decisions

NCPA

Equipping school buses with seat belts costs about $53 million per year; but since this effort will save only two children's lives every year, the cost is about $2.8 million per year of life saved.

ibid

CE and ROI decisions

NCPA

We spend $2.8 million every year on radionuclide emission control at elemental phosphorus plants (which refine mined phosphorus before it goes to other uses); but since this effort will saves at most one life alternate years, the cost is $5.4 million per year of life saved.

ibid

CE and ROI decisions

NCPA

While banning asbestos in automatic transmission components costs but $22,000 per year, the cost per year of life saved is $66 million.

ibid

CE and ROI decisions

NCPA

	$/yr in M	+Life Years	$/LY in M
sickle screen	0.227	961.	0.00024
heart txplant	460.	2,900.	0.158
bus seatbelts	53.	2.	2.8
radioactivity	2.8	0.1	5.4
asbestos	0.022	.0003	66.4

ibid, in summary form

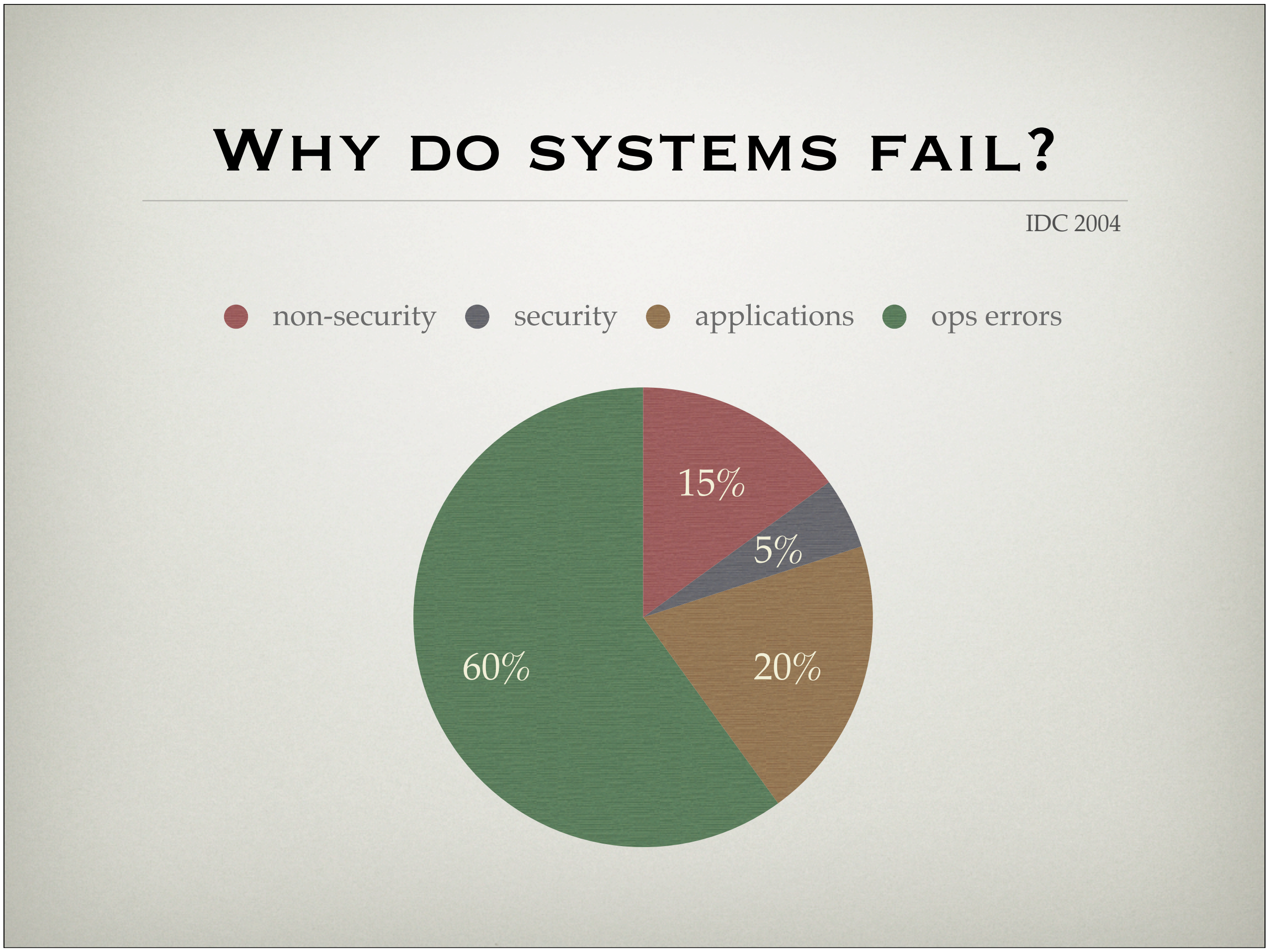
CE example: ROSI

- Use application scanner to manufacture some risk index r_a
- Apply patch, rescan to get r_b
- Determine rollout cost c_r
- Dollars per unit of risk reduction = $\frac{c_r}{(r_a - r_b)}$

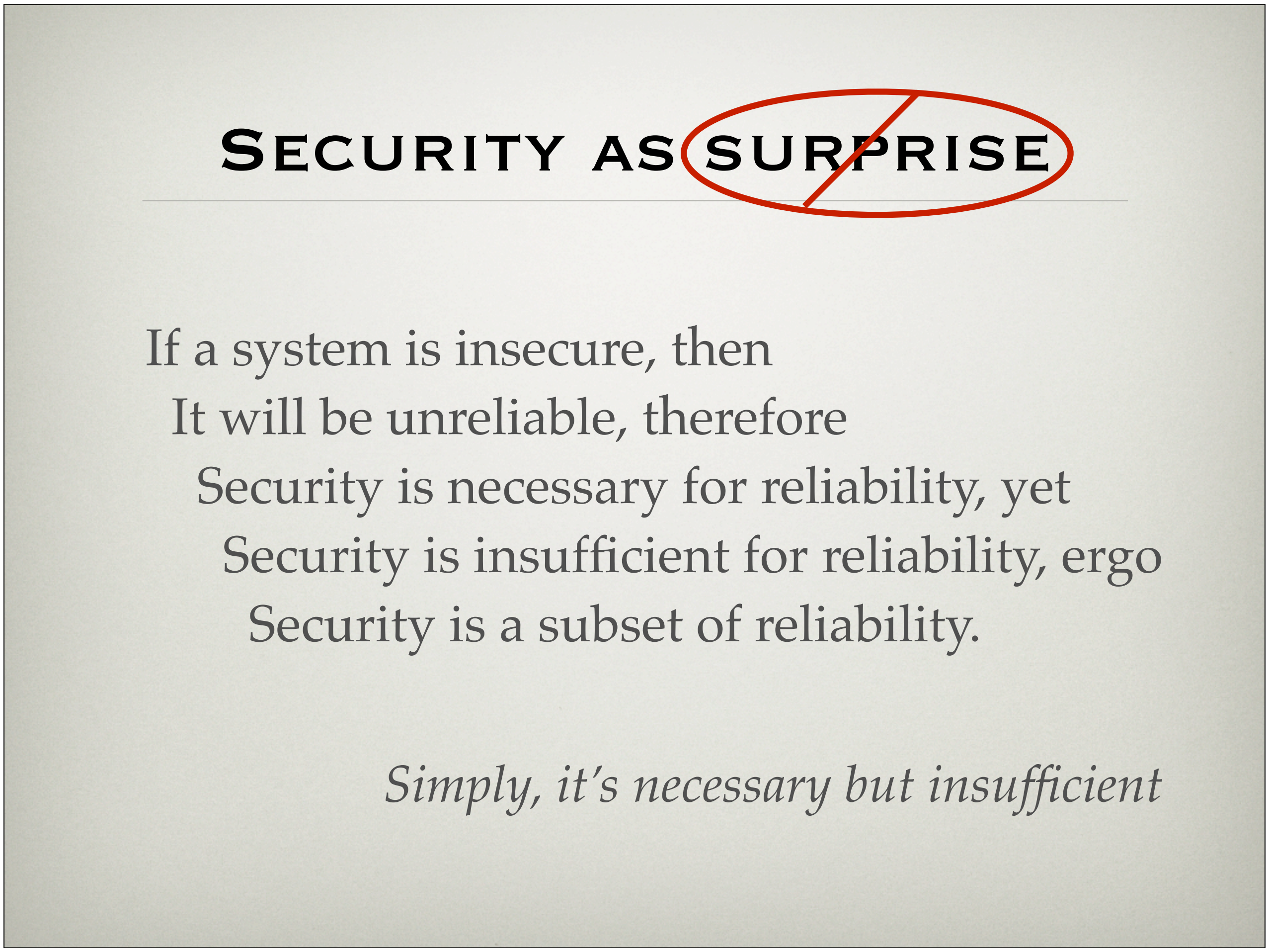
Let's work a simple example.

In this case, we work out a dollar value for each unit of risk reduction. So long as we consistently measure the before and the after, the relative vulnerability of the before and the after can be then used for comparison, as is done here. If you have many different options on what you might do, sample the lot of them, order the results, and just proceed from most cost-effective toward the least.

Thinking more broadly about costs



IDC, non-open source reference.



The more mature the infrastructural entity is the more security is a subset of reliability, per the logic above.

The parallel: that if a system is unregulated then it is unpredictable, therefore regulation necessary is for predictability, yet regulation is insufficient for predictability, therefore regulation is a subset of predictability suggests itself. If as correct as the relation between security and reliability, then the question for the law is how to regulate for predictability without damping out innovation or the motivation to improve. This is hardly a new topic, but the digital physics will stress security as a subset of reliability.

As Whit Diffie (Stanford) has observed, computing would become free were it not for security.

Availability

US Army

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTTR} + \text{MTBF}}$$

High availability can come from

↑ Mean Time Between Failure (MTBF)

↓ Mean Time To Repair (MTTR)

This section drawn from http://www.usace.army.mil/publications/armytm/tm5-698-3/glossary.pdf but see also http://www.weibull.com/SystemRelWeb/availability.htm

MTBF v. MTTR

- MTBF anticipates failure so as to avoid it
- MTTR anticipates failure so as to recover
- Neither is cost effective at the margins
- Sum of the two is the TCO of your strategy

Mean Time Between Failures is the measure of the average time between (in our case) security events.

Mean Time To Repair is the measure of the average time to recover from (in our case) a security event.

Making MTBF infinite is infinitely expensive. Making MTTR zero likewise. Neither is the whole answer separately but together you have a risk management decision that permits actual, sane discussion of the Total Cost of Ownership for the security technology and processes that you do deploy.

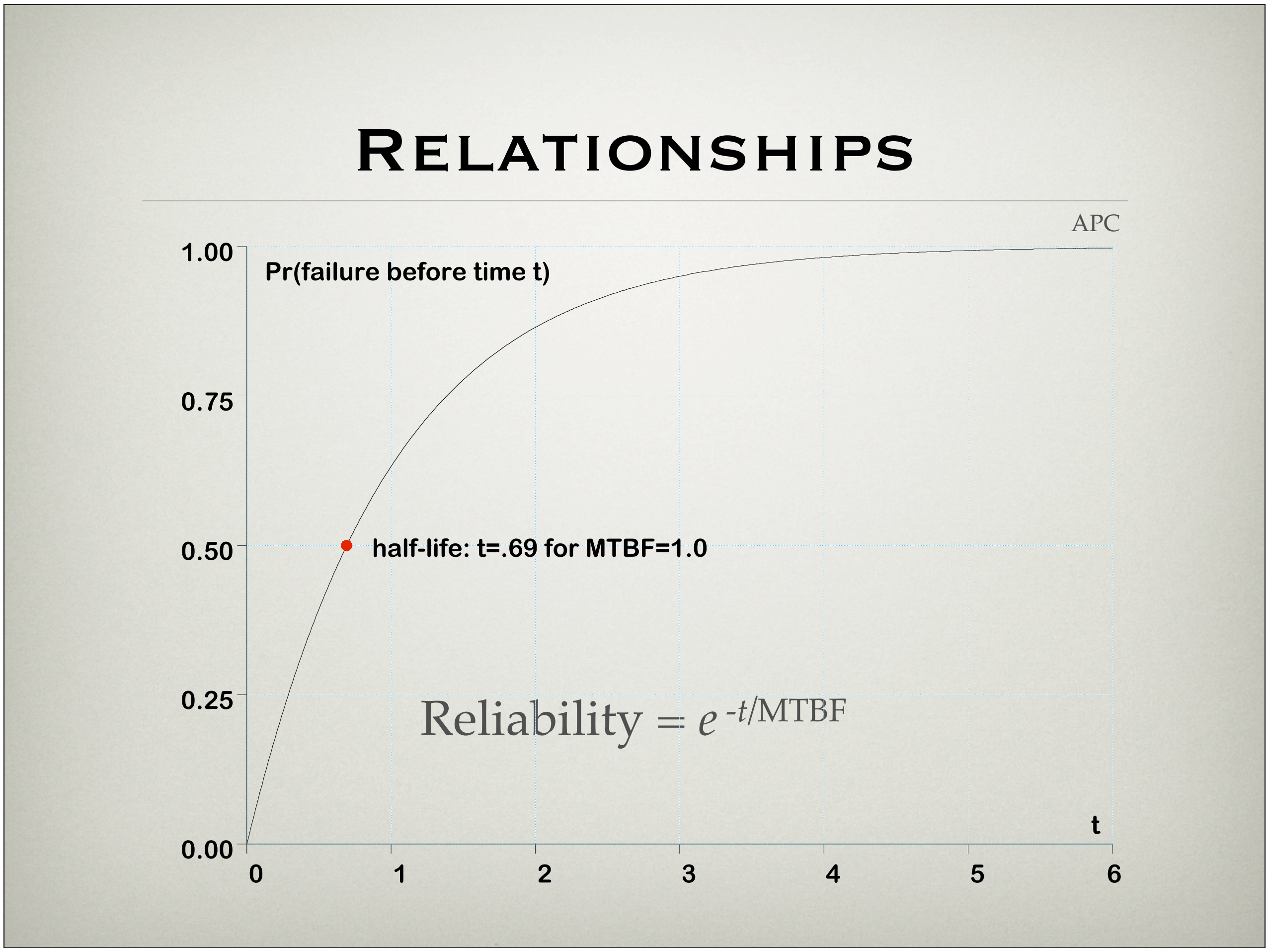
CE & Availability

US Army

$$A = 1 = \begin{cases} \frac{\text{MTBF}}{0+\text{MTBF}} & \text{when MTTR} = 0 \\ \frac{\infty}{\text{MTTR}+\infty} & \text{when MTBF} = \infty \end{cases}$$

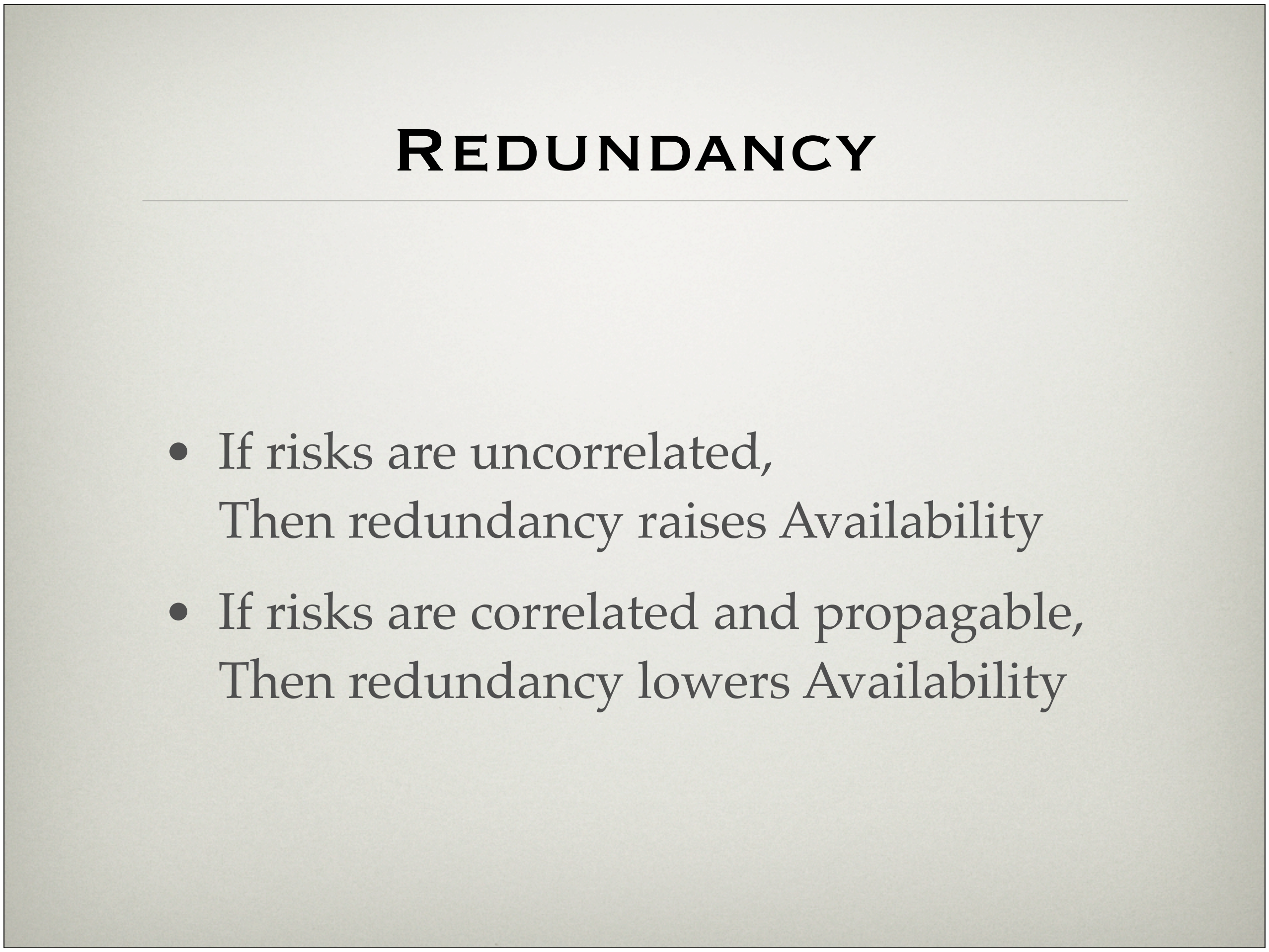
In other words, which is more CE to approach, zero recovery time or infinite uptime?

In some circumstances, your availability improvements are most cost effective when approached through suppressing failures. In others, they may come from shortening repair time. For a space mission, no failures clearly wins. For Google, throwing out a misbehaving board-level Linux blade is the answer.



Note, in case it was not obvious, that “MTBF” is not the same as half-life. This is a side issue, but the exponential curve above says that if MTBF is 1 unit of time that 50% of the component it covers will have died by 0.69 units of time.

“Availability & Reliability Theory,” APC, at https://ilcsupport.desy.de/cdsagenda/askArchive. php?base=agenda&categ=a0533&id=a0533s1t12/moreinfo



An important point: If the risk of failure is uncorrelated across multiple instances then redundancy will raise availability (more units will have to fail to break availability). If, however, the risk of failure is correlated and transmissible, then adding units decreases availability.

2004 Turing lecture

Adi Shamir

- Absolutely secure systems do not exist
- To halve your vulnerability, you have to double your expenditure
- Cryptography is typically bypassed, not penetrated

Adi Shamir, the “S” of “RSA,” received the Turing Award in 2004. His acceptance lecture included three points, as above.

http://www.acm.org/awards/turing_lectures_project/turing/S/s-pp/shamir_1files_files/TextOnly/index.html

Liability

Liability

- No clear answer on this yet, so must speculate
- Amongst those doing a good job on security, the residual risk is that of counterparty risk, *i.e.*, the risk that your business partners will lose your data for you – and are you liable?

Probably not the way to run the railroad, but some management responds only to this.

Legal corroboration

Jeffrey Ritter, Esq.: That which...

...is not documented does not exist.

...was not recorded did not happen.

...has not been audited is vulnerable.

He does not mean a path to invisibility, but rather that these are the pre-conditions for liability. He is advising law firms on just this sort of thing, i.e., that their own handling of co-mingled documents from their clients is dangerous to their clients and themselves unless that handling is done with rigor. (His firm is Waters Edge Consulting, wec-llc.com, co-founded with Karen Worstell, former CISO for Microsoft.)

Liability and quant law

Hand 1947

- Given
  - P = the probability of loss
  - L = the amount of said loss
  - B = the cost of adequate precautions
- Then
  - Liability whenever $B < PL$

Judge Learned Hand says simply that if it is more cost effective to anticipate and thus prevent a failure than it is to bear the risk, then there is liability for not having done so. This is a precedential case for all of U.S. liability case law.

UNITED STATES et al. v. CARROLL TOWING CO., Inc., et al.; Nos. 96, 97, Dockets 20371, 20372; SECOND CIRCUIT COURT OF APPEALS; 159 F.2d 169; January 9, 1947.

Ex: Identity fraud

FTC 2003

$$P = 4.6\%$$

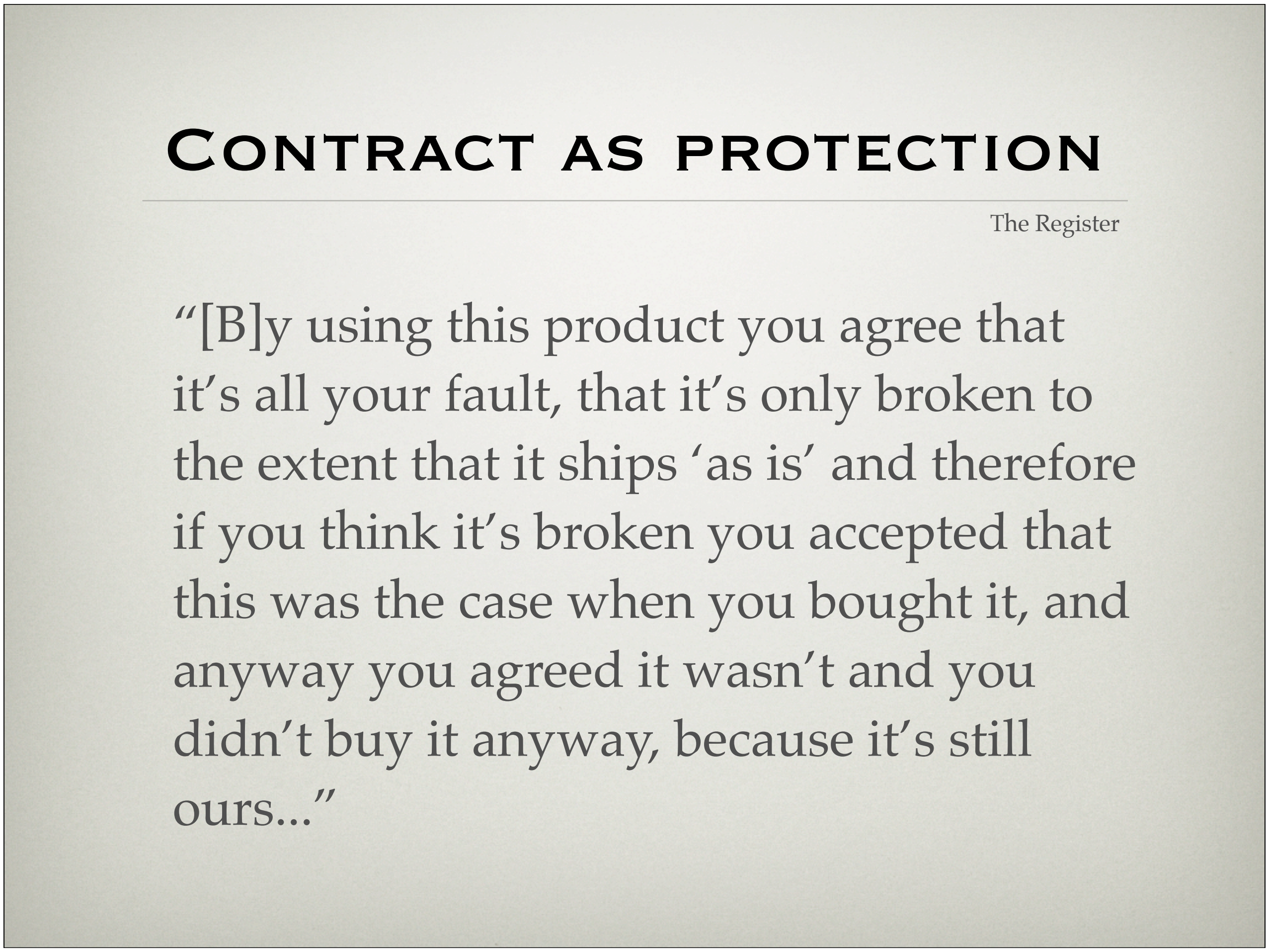
$$L = \frac{3x10^{8}hr*\$5.15/hr + \$5x10^{9}}{10^{7}\ victims} = \$655/v$$

$$P*L = \$30.11 = B_{cutoff}$$

Is $30.11/yr/consumer enough to cure?

Applying Hand's calculus to data from the Federal Trade Commission on identity theft, 4.6% of the population has suffered an identity fraud. In the process, they spent three hundred million hours repairing the troubles caused (priced here at the Federal minimum wage) and they also spent five billion dollars out of pocket. The ten million people involved thus had a $655 loss per incident. Since the probability (P) of a loss, 4.6%, times the loss (L) of $655 imposes a burden (B) of $30.11, the question then is whether it is possible to protect an individual against identity fraud for $30.11 per annum. If it is, then liability is found. If not, not.

Identity Theft Survey Report, Federal Trade Commission, September, 2003, as found at http://www.ftc.gov/os/2003/09/synovatereport.pdf



This is the wonderfully curmudgeonly UK digital publication “The Register” synopsizing the plain english meaning of most software licenses. Liability can, as ever, be removed by contract as this one tries illustratively doing. See http://www.theregister.co.uk/content/4/33082.html

Software metrics

Pricing software

- COnstructive COst MOdel (COCOMO)
  - Organic - small teams work to less than rigid requirements
  - Semi-detached - mixed teams meet requirements of varying rigidity
  - Embedded - tight hardware, software, and operational constraints

One of several methodologies for producing reliable software at attractive costs, the COnstructive COst MOdel (COCOMO) provides differing advice for three tiers of development teams and requirements, as stated here. Material taken from http://www1.jsc.nasa.gov/bu2/COCOMO.html

COCOMO EQUATION 1/2

$$E = \text{person-months} = a(KLOC)^b$$

$$D = \text{chronological months} = c(E)^d$$

$$P = \text{people required} = E/D$$

$$KLOC = \text{thousand lines of code}$$

Defining three outcome variables, E, D, and P, and noting that software cost is going to be dependent on sheer volume measured in thousands of lines of code.

COCOMO EQUATION 2/2

where a, b, c and d are:

Software project	a	b	c	d
Organic	2.4	1.05	2.5	0.38
Semi-detached	3.0	1.12	2.5	0.35
Embedded	3.6	1.20	2.5	0.32

Coefficients derived from practice under a NASA contract to TRW. See original reference to pursue this.

COCOMO PUNCHLINE

$KLOC$		1	1,000	100,000
organic	E	2	3,390	426,787
	D	7	55	345
	P	0.3	62	1,238
semi-detached	E	3	6,873	1,194,322
	D	4	55	335
	P	0.8	125	3,566
embedded	E	4	14,332	3,600,000
	D	4	53	313
	P	1.0	268	11,491

A punchline: as projects grown in size they also grow in complexity and requirements, i.e., the path tends to be diagonal as seen here. The chart contrasts 1,000 lines of code with 1,000,000 and then 100,000,000. In the upper left box, we find 500 LOC / person month of effort. In the lower right box, we are down to 27 LOC / person month of effort. This is a complexity tax imposed by a requirement for constant quality.

As this talk is not about software construction, we will move forward, but there is a lot of work to be done on stealing from the quality assurance literature for the benefit of software security.

Rapid ferment

- Several firms are pushing hard on software security and metrics for them
- Metricon 1.0 and 2.0 both had/have multiple papers on this issue
- Whatever you do, be consistent (per the arguments about trend analysis) and provide relative risk per dollar

Any attempt to list them all will inevitably leave somebody out who will inevitably be aggrieved. Get the digest of Metricon 1.0 now (see below) and the digest of Metricon 2.0 as soon as it is available (in a similar location). This is a very intensely worked area with real competition, and, that means whatever is written down here is going to quickly outdate and, of course, those who are competing will not share all they know.

http://www.securitymetrics.org//content/attach/Welcome_blogentry_010806_1/metricon.notes.PDF

Field estimation

Capture/re-capture

- N = #(population)
- n_1 = #(1st capture; mark & release)
- n_2 = #(2nd capture)
- m_2 = #(2nd capture found to be marked)
- "Lincoln Index:" $\frac{m_2}{n_2} = \frac{n_1}{N} \implies N = \frac{n_1 n_2}{m_2}$

Sometimes, you want to estimate how many frogs there are in a pond. For that you capture some frogs, band them, release them, wait a while, and capture some more frogs. The ratio of frogs captured in the second pass that do or do not have a band tells you what you want to know -- the total number of frogs in the pond, as we shall see. The assumption, and of course there is one, is that your chancing of catching each individual frog is the same.

Use in security

- Select all e-mails in one hour, say
- Record senders of Bad Mail (n_1)
- Repeat in one week (n_2 and m_2)
- Estimate number of violators (N)

e.g., {41,62,6} $\Rightarrow \exists$ 424 Bad Mail senders

In security, we might well use this. We would catch all the e-mail outbound in a one hour capture. We'd band (remember) whomever sent bad e-mail by whatever definition we wished to use. At a later time, we'd do this a second time. People who re-appeared from the first time we'd treat as banded. We could now estimate the number of people who are sending bad e-mail.

In the example, forty-one in pass one plus sixty-two in pass two, of which six are repeats, and now we have an estimate of the population of senders of bad e-mail.

(FOR THE RECORD)

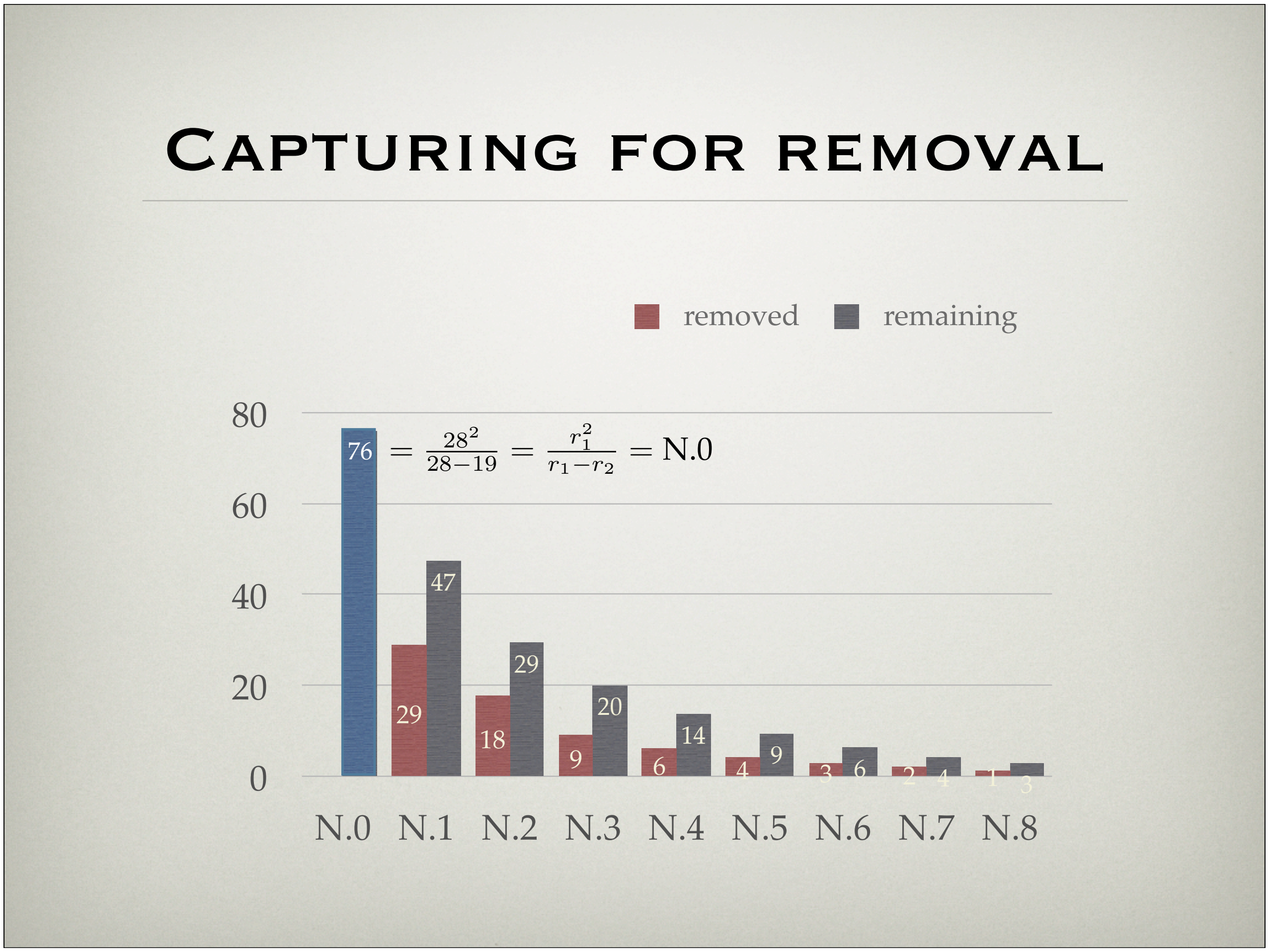
- Better to calculate $N = \frac{(n_1 + 1)(n_2 + 1)}{m_2 + 1} - 1$
- So our example is really
  {41,62,6} ⇒ $\exists$ 377 Bad Mail senders
- Has seen real use in analyzing repeated design reviews by independent teams

Without going into it, the better statistical measure is as seen here which leads to a different though similar estimate of the number of senders of bad e-mail. For further discussion, see several texts. For security purposes, and remembering the Dr. Dobbs illustration, you might especially want to read Vander Wiel SA & Votta LG : Assessing Software Designs Using Capture-Recapture Methods, IEEE Trans on Software Eng, v19 n11 p1045-1054.

Capturing for removal

- First removal = r_1, second removal = r_2
- If $\frac{r_1}{N_0} = \frac{r_2}{N_0 - r_1}$, then $N_0 = \frac{r_1^2}{r_1 - r_2}$

Sometimes you don't band the frogs and throw them back -- sometimes you want to get the frogs out of the pond. You can still estimate size of the beginning population by, you guessed it, making an assumption: that on any round of catching you catch the same percentage of frogs.



So, with only two numbers, 29 captures on the first round and 18 captures on the second round, we can say that we began with 76 frogs in this pond and that the population will decline further to 20 thence 14 thence 9 thence 6 thence... under repeated removal of 29/76=38% of the frogs at each round.

Use in security

- University screening inbound laptops for malware before Reg Day
- Egress filtering with feedback to senders of Bad Bits ("Don't let me catch you again")

In security, we might say that within a university the returning students and their laptops are the pond and the ones that have malware are the frogs to be removed. Or we might do a capture/re-capture experiment to tell people sending bad e-mail that they should not do it again.

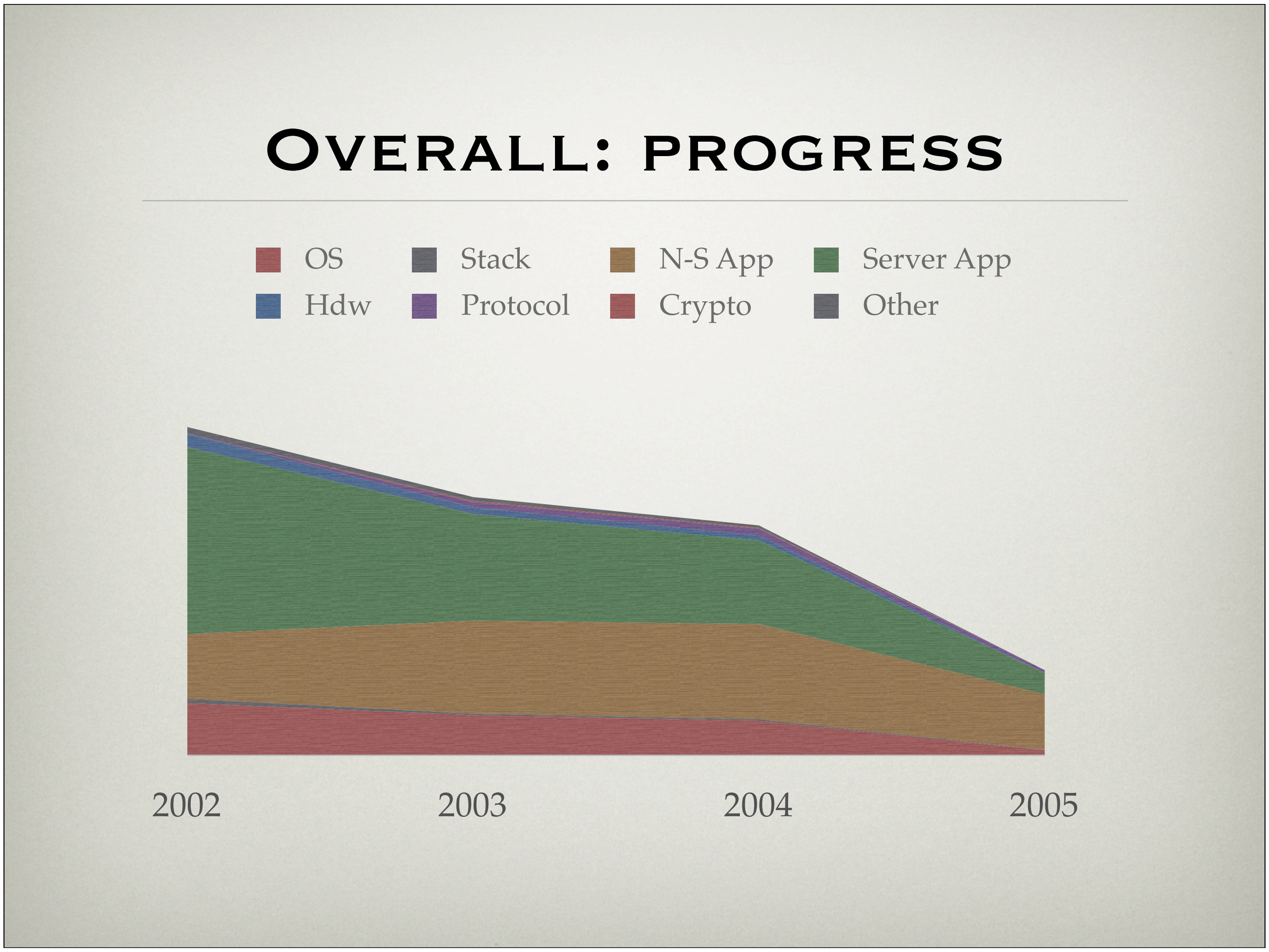
More trend analysis

Remotely exploitable vulns

NIST

Component	2005	2004	2003	2002
OS	19	140	163	213
Net Stack	1	6	6	18
Non-Server App	229	393	384	267
Server App	88	345	440	771
Hardware	0	20	27	54
Protocol	12	28	22	2
Crypto	0	4	5	0
Other	0	10	16	27

This is data right from the National Institute for Standards and Technologies. I don't like it; it doesn't tell you anything; the column order is reverse chronological and the raw counts offer no insight. But let's start with it, as seen at http://icat.nist.gov/icat.cfm?function=statistics



Let's see if there is progress being made by making a stacked area graph and running time in the forward direction. It does indeed look like progress.

NON-UNIFORM ΔN(VULNS)

	CAGR
Hardware	-73.5%
Other	-66.7%
Net Stack	-61.8%
OS	-55.3%
Server App	-51.5%
Non-Server App	-5.0%
Protocol	81.7%
Crypto	-na-

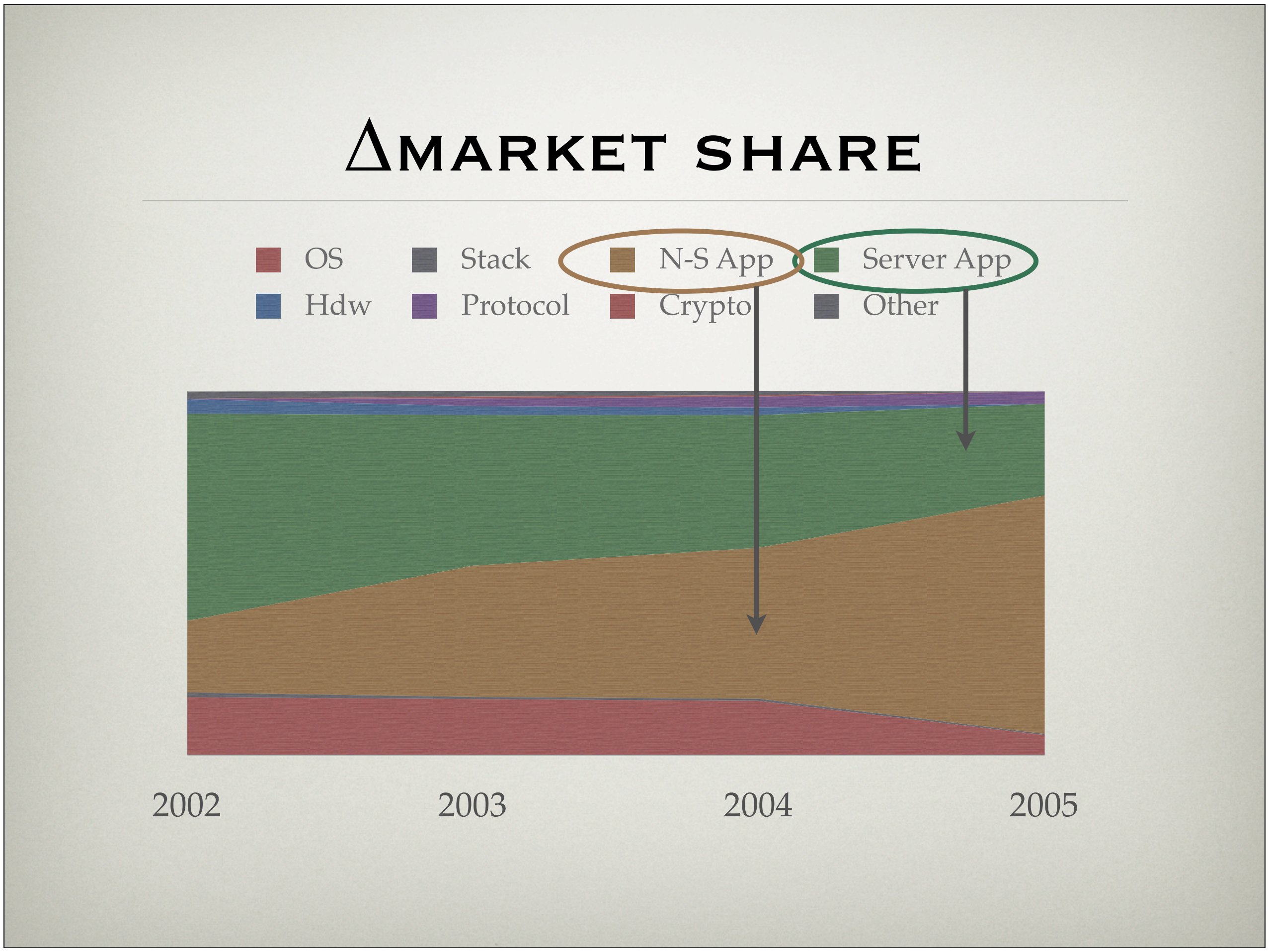
-36%

But the progress is hardly uniform. The compound annual growth rate (CAGR) varies from −73.5% to +81.7%, which is quite a range, and has an overal CAGR of −36%.

Market share

Component	2005	2004	2003	2002
OS	5%	15%	15%	16%
Net Stack	0%	1%	1%	1%
Non-Server App	66%	42%	36%	20%
Server App	25%	36%	41%	57%
Hardware	0%	2%	3%	4%
Protocol	3%	3%	2%	0%
Crypto	0%	0%	0%	0%
Other	0%	1%	2%	2%

It might be more instructive to look at market share rather than pure count. In the format of the original, it looks like this (which is still pretty useless).



But as market share we can now see something worth seeing, that the green Server Application category was once dominant but is in fast decline, its place taken by the brown Non-Server Application category.

This tells you...

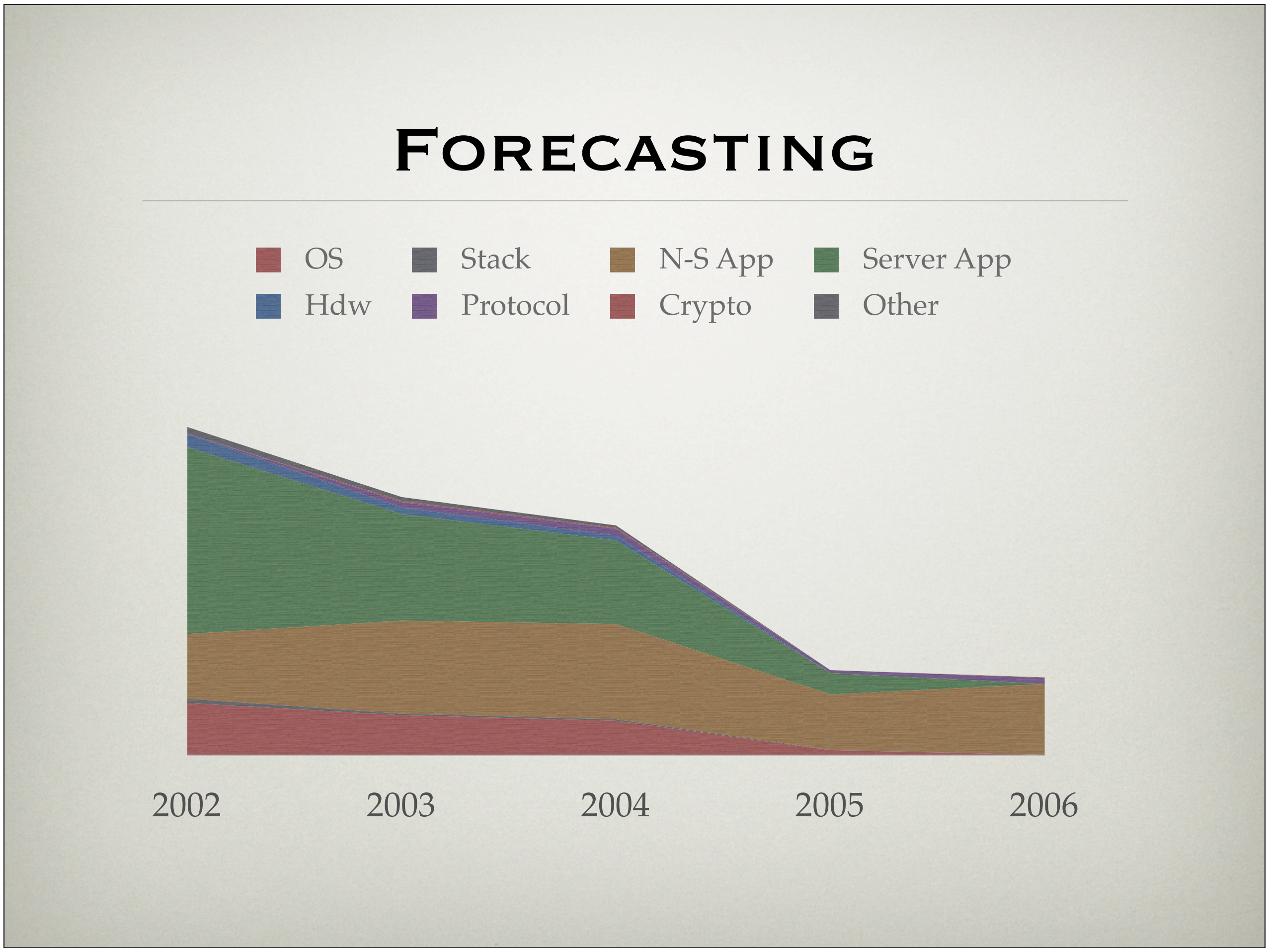
- ...to work on non-server applications
- Market share tells the story
  - NS-apps 20%→66%, CAGR =49%
  - Other 80%→34%, CAGR = -25%

Now we have a metric on which to base a decision; we need to work on these Non-Server Applications as they have a CAGR of 49% in market share terms while everything else has a CAGR of -25% in market share terms.

2006 forecast

	linear regression	
OS	0	
Net Stack	0	
Server App	0	
Hardware	0	
Other	0	
Crypto	2	0.6%
Protocol	25	7.8%
Non-Server App	292	91.5%

If we take the numbers as given and just do a linear regression so that there is a 2006 (plus one year) prediction, we'd expect the year 2006 values to be down to three (from eight) classes with Non-Server Applications now at 91.5%, thus reinforcing the idea that we need to attend to that line item above all others.



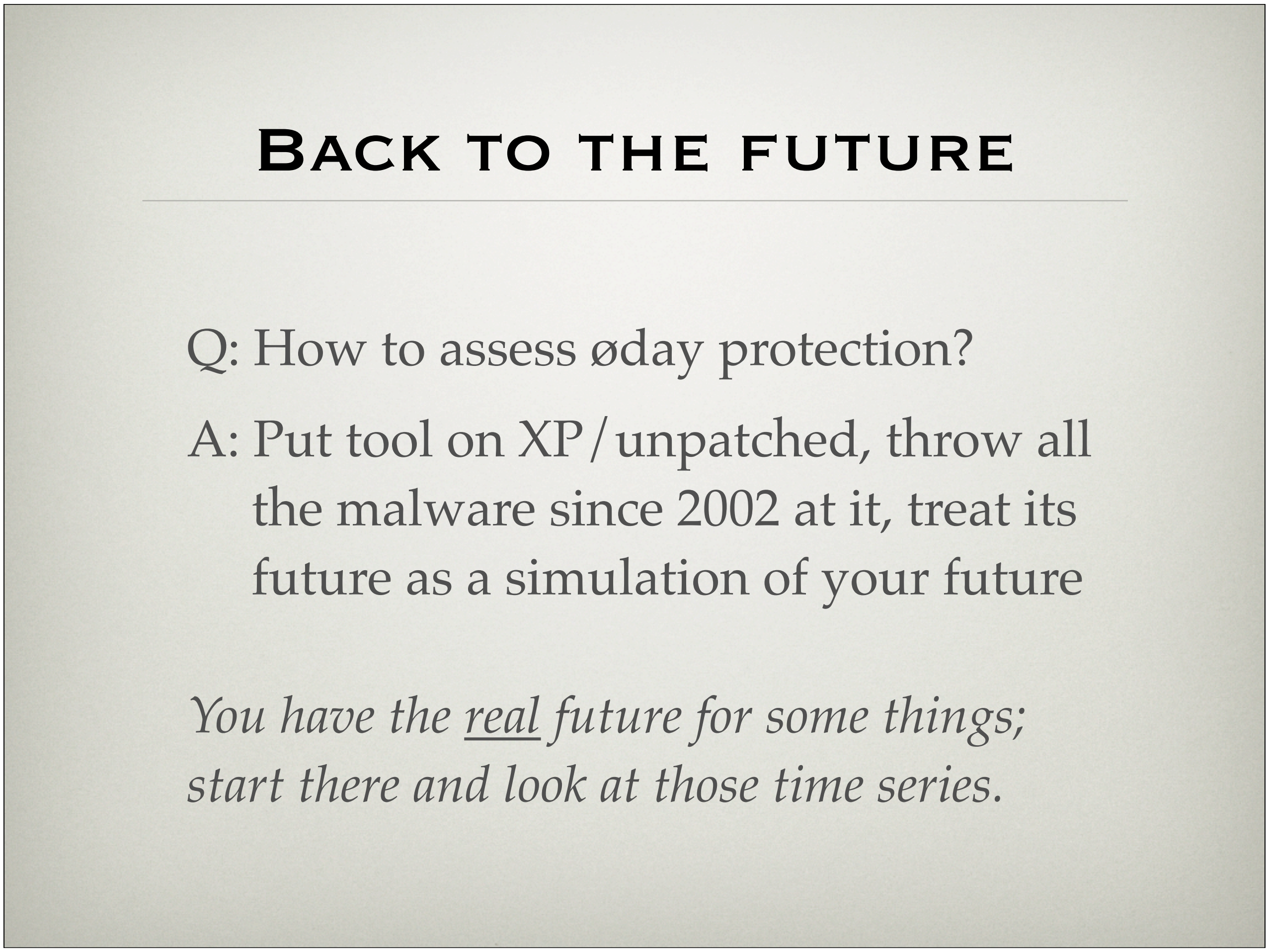
Graphing, in the same style, with the forecast in place gets the point across to almost anyone.

Forecasting is good

- Can help with
  "What if we hadn't done anything?"
- Don't overdo it; stay skeptical

And, in fact, forecasting with an intervention versus non-intervention dichotomy is often a very good decision tool indeed. You can overdo it, of course, but a healthy skepticism is an adequate protection here.

Scepticism is the chastity of the intellect; it is shameful to give
it up too soon, or to the first comer.
-- George Santayana



Sometimes the time series you really want is a deep projection into the future. That is hard to do. As an example of trying, the present author wanted to assess a 0day protection strategy. Of course, one cannot ask for samples of future 0day attacks, so an unpatched Windows XP system, vintage 2002, was taken and the protection installed on it. As 2002 is four years back, all the worms and viruses that have appeared in four years can be said to be a sample of what the future held for XP in 2002 and thus throwing all of those attacks at the unpatched XP system was, in fact, a simulation of repeated 0day attacks and, in turn, an adequate test of whether the installed tool had protective value against 0day malware. As a testing strategy, it worked and worked well.

Visualization

Some guidelines

- Communication, not beauty
- Less is more
- Work on your graphs, etc., yourself
- If it doesn't add anything, leave it out
- Color for focus, not for decoration
- If it isn't labeled, then it doesn't exist

The point is, you are collecting and presenting metrics as decision support and not for art, fame, glory, fun, or self-protection. Or at least let us assume that is the case.

Highly recommended are the books of Edward Tufte, but you probably already know that. See http://www.edwardtufte.com/tufte/ for more; they are seriously good (humbling, really).

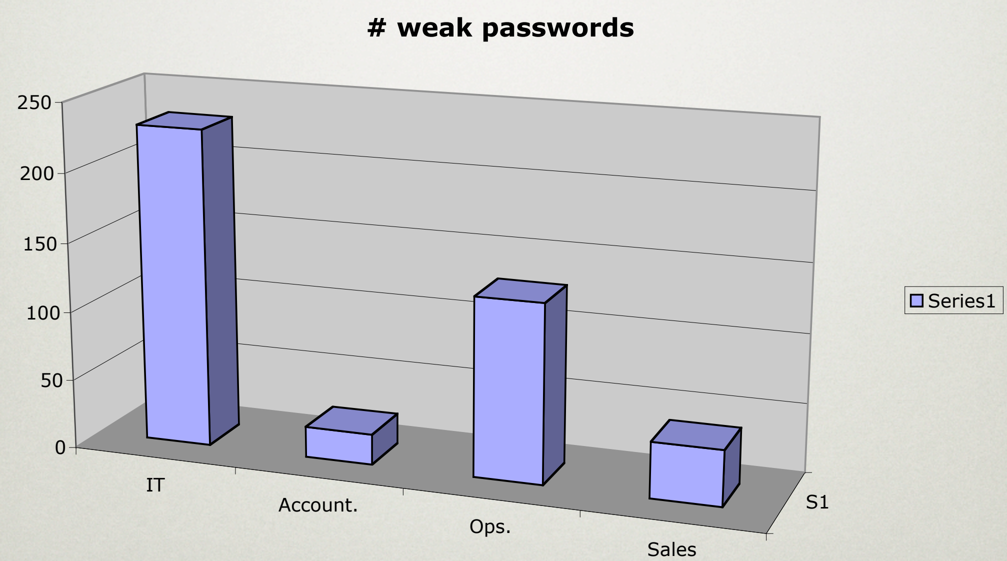
How to show...

basic
counted
data

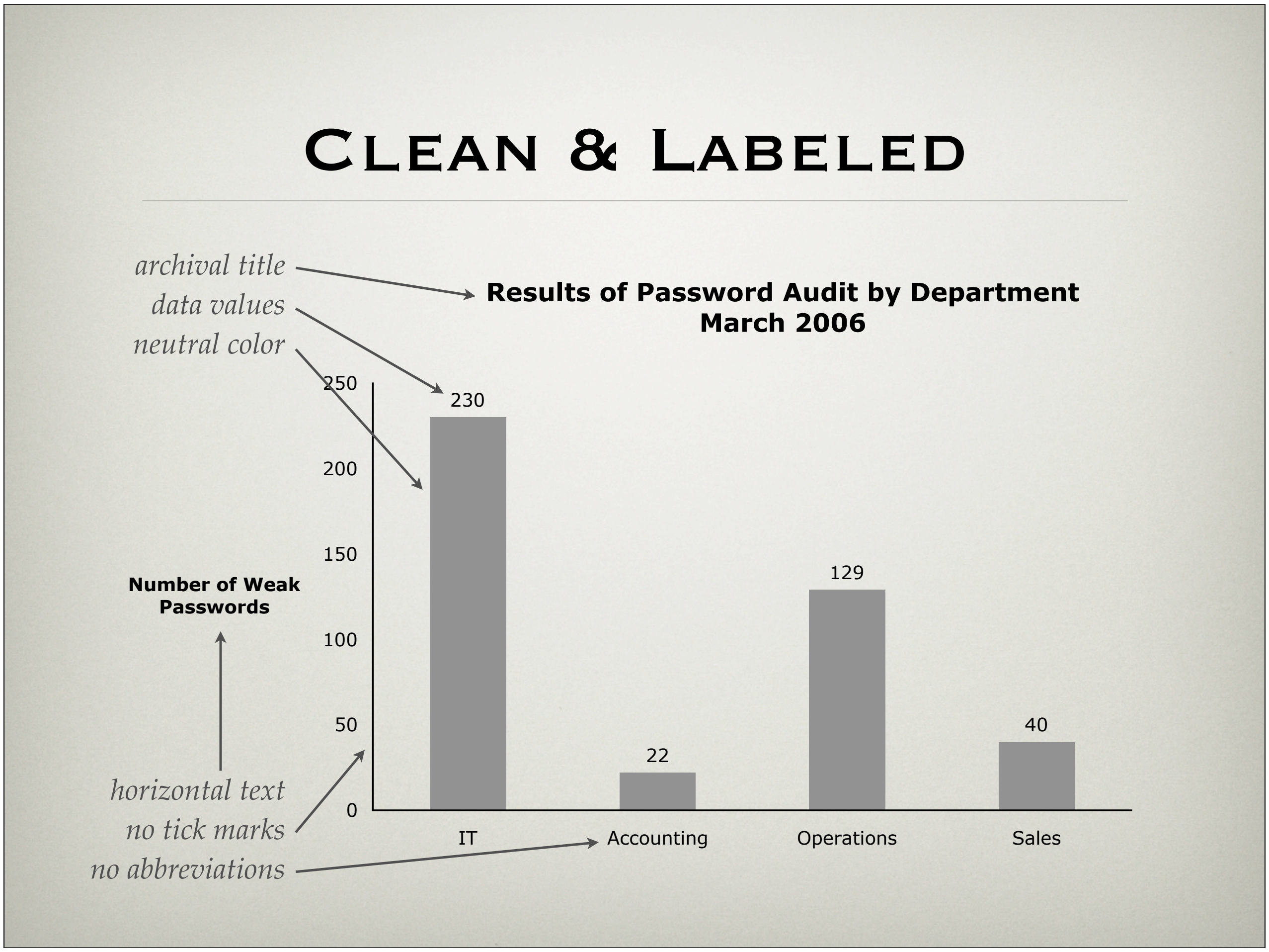
department	weak pwds
IT	230
Account.	22
Ops.	129
Sales	40

Manufactured data, Jaquith, op cit., p.166

How not to show...



Junk. All the doodads and visual effects add *nothing* and should not be present.



Clearer and thus better. Save the visual effects for when you need them.

How to show...

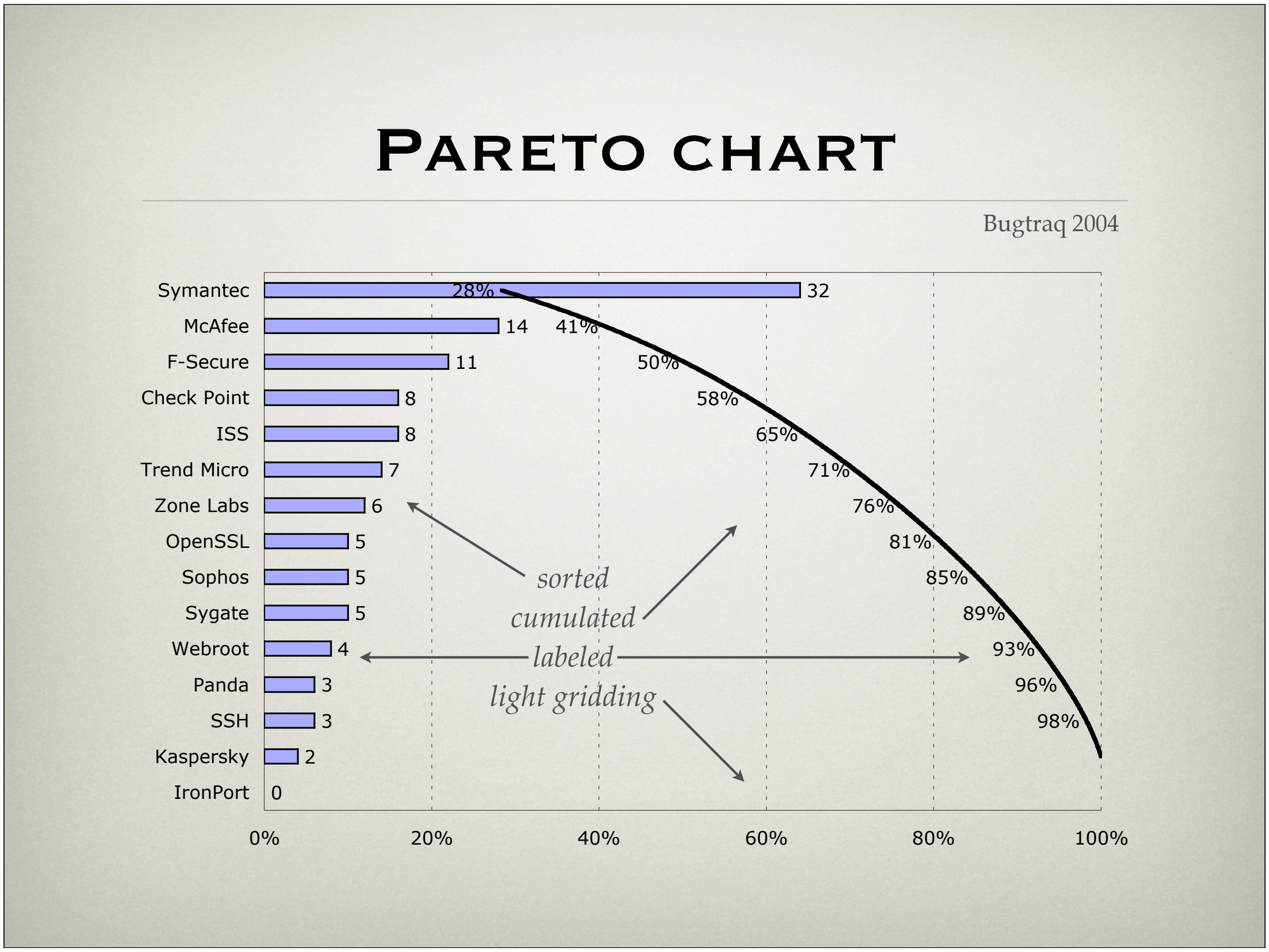
Bugtraq 2004

comparative
counted
data

vendor	vulns
Check Point	8
F-Secure	11
IronPort	0
ISS	8
Kaspersky	2
McAfee	14
OpenSSL	5
Panda	3
Sophos	5
SSH	3
Sygate	5
Symantec	32
Trend Micro	7
Webroot	4
Zone Labs	6

The point is understanding, and by ordering and showing both the point and cumulative curves, presumably those who want to understand, do.

2004 Bugtraq data as reported by Jaquith, op cit., p.194



As with the previous display, inessentials are kept at a minimum, labels are straightforward, and a lot is packed into a small space.

Excel: 3 column worksheet: label, data, cumulative %age. Insert a bar chart. 2click category scale getting "format axis", then "scale" tab, then check "categories in reverse order." 2click secondary (cumulative) bar to get "axis" where you check "secondary axis." 2click primary axis then set "max" to true max and "min" to true min (they will be wrong). Turn on "show data value" for both, but 2click a secondary data value then click "alignment" where you change it to "inside end." Click secondary bar then, from top bar, "Add Trendline..." which should be a polynomial of order 5 (or so). 2click secondary bar going to "Patterns" here you set the Border to "none" and the Area to "none" making the secondary bar chart disappear. 2click the top (primary) scale and turn off tick marks and tick mark labels, making the top scale disappear. 2click the bottom (secondary) scale and turn off tick marks but leave tick labels on. 2click the left (vertical) scale and turn off tick marks but leave tick labels on. Remove gridlines, color, border, and legend from the chart itself. Remove the last (extraneous) "100%" label on the secondary trendline.

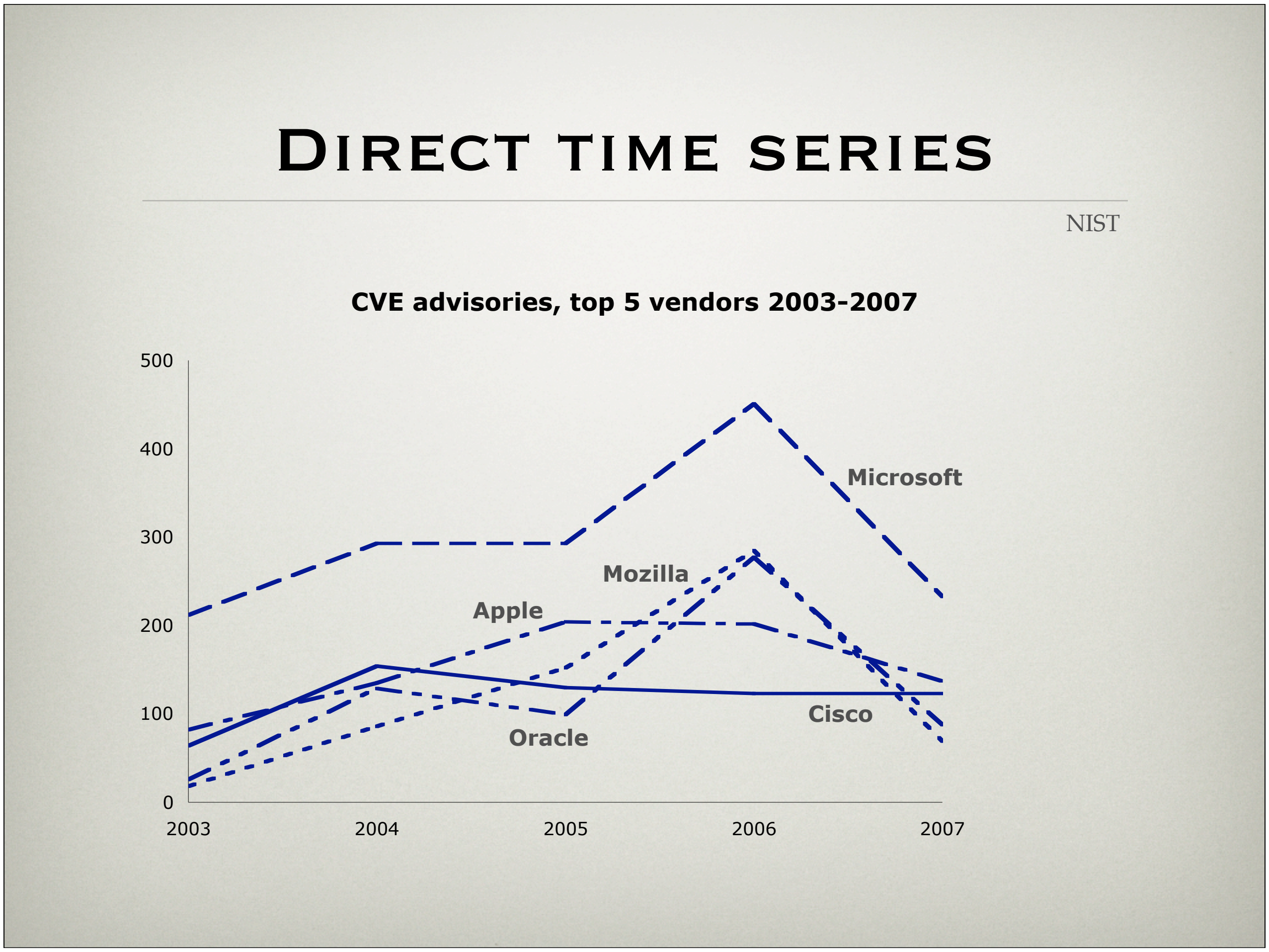
How to show...

NIST

comparative counted time series

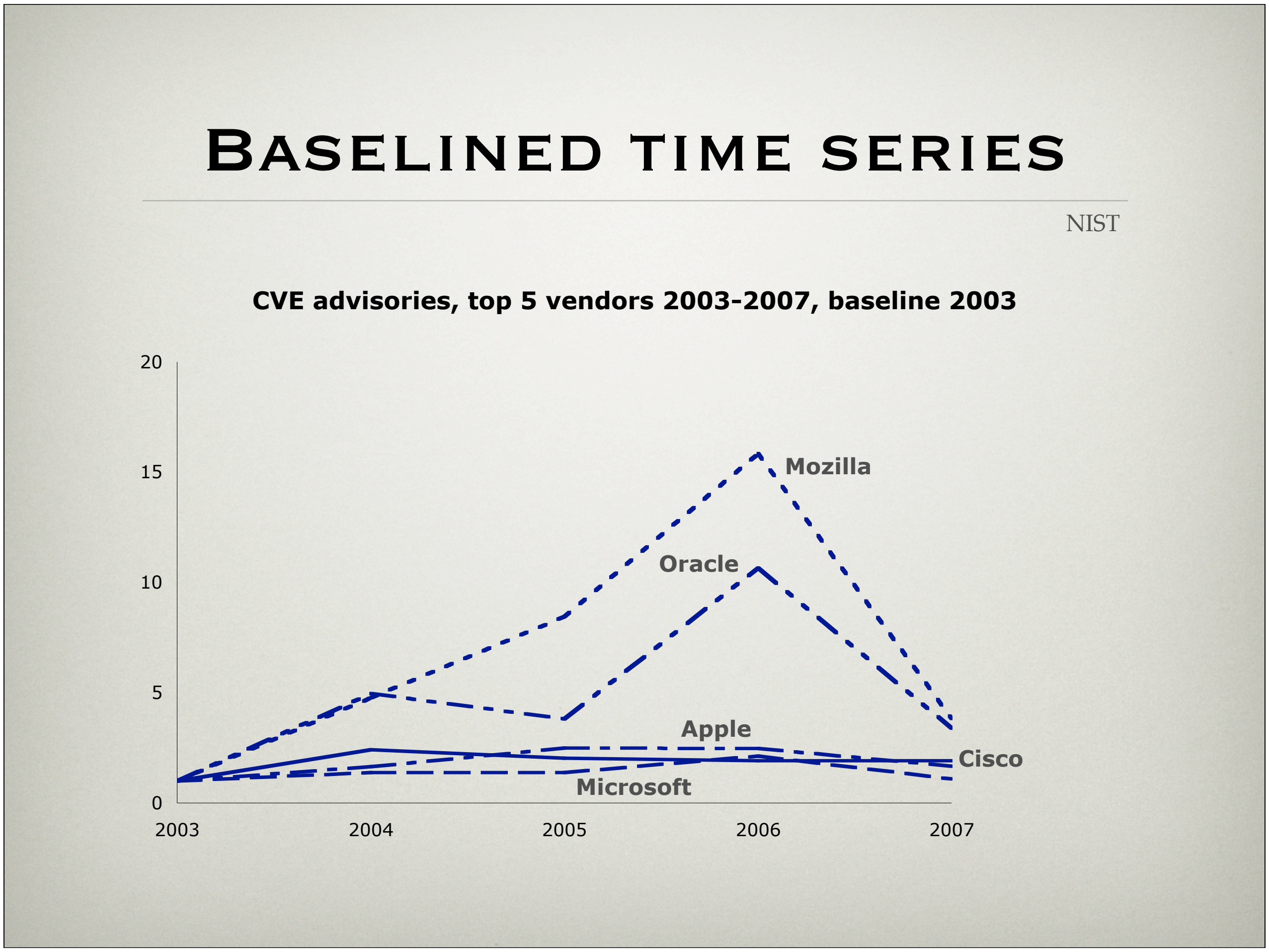
	2003	2004	2005	2006	2007	total
Microsoft	212	293	293	451	233	1482
Apple	82	135	204	202	137	760
Oracle	26	129	99	277	88	619
Mozilla	18	86	152	285	69	610
Cisco	64	154	130	123	123	594

Data as derived from several files available at http://nvd.nist.gov/download.cfm, plus assorted awk scripts...



This is simply the side-by-side plots of how many CVE vulnerabilities were posted against the given vendor by year.

Gives you a sense of scale for “Who is the problem?”



This now normalizes to a baseline, namely the number of vulnerabilities reported in 2003, and from there the multiple of that year that a given year represents.

Gives you a sense of scale for “Who is having a rough time?”

How to show...

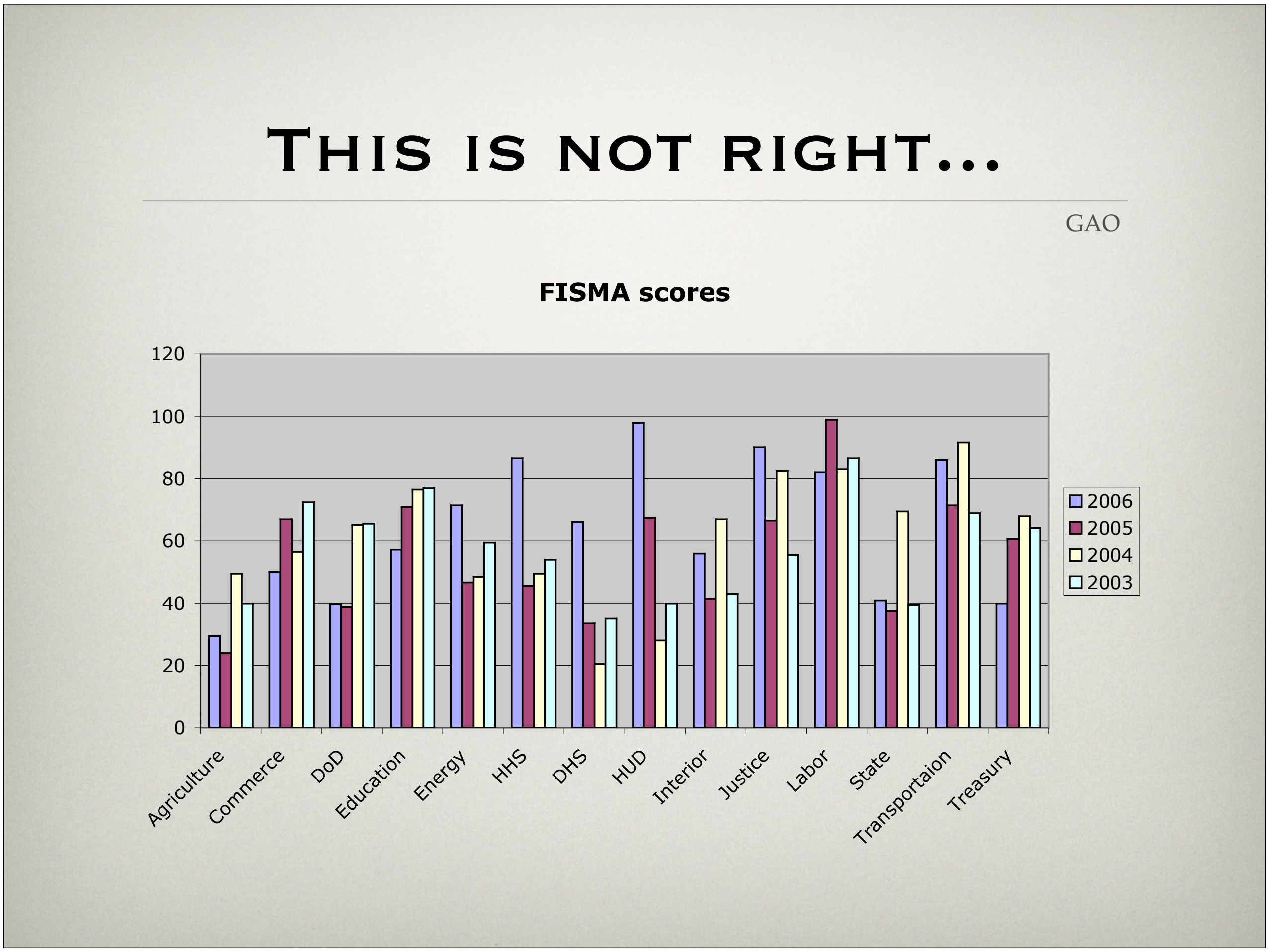
GAO

Agency	2006 Score	2006 Grade	2005 Score	2005 Grade	2004 Score	2004 Grade	2003 Score	2003 Grade
Agriculture	29.5	F	24	F	49.5	F	40	F
AID	99	A+	100	A+	99	A+	70.5	C-
Commerce	50	F	67	D+	56.5	F	72.5	C-
DOD*	39.75	F	38.75	F	65	D	65.5	D
Education	57.25	F	71	C-	76.5	C	77	C+
Energy	71.5	C-	46.75	F	48.5	F	59.5	F
EPA	92	A-	97.5	A+	84	B	74.5	C
GSA	95	A	92.5	A-	79.5	C+	65	D
HHS	86.5	B	45.5	F	49.5	F	54	F
DHS	66	D	33.5	F	20.5	F	34	F
HUD	98	A+	67.5	D+	28	F	40	F
Interior	56	F	41.5	F	67	D+	43	F
Justice	90	A-	66.5	D	82.5	B-	55.5	F
Labor	82	B-	99	A+	83	B-	86.5	B
NASA	60.75	D-	80	B-	60	D-	60.5	D-
NRC	53	F	60.5	D-	88	B+	94.5	A
NSF	99	A+	95	A	77.5	C+	90.5	A-
OPM	99	A+	98	A+	72.5	C-	61.5	D-
SBA	89	B+	77	C+	60	D-	71	C-
SSA	96.5	A	99	A+	86	B	88	B+
State	41	F	37.5	F	69.5	D+	39.5	F
Transportation	86	B	71.5	C-	91.5	A-	69	D+
Treasury*	40	F	60.5	D-	68	D+	64	D
VA*	**	**	46	F	50	F	76.5	C
Government-wide Average	**72.9**	C-	**67.3**	D+	**67.2**	D+	**65**	D

***The Inspector General for these agencies did not provide independent evaluations of their agencies' FISMA reports for FY03. Therefore, the scores are based on self-reported numbers submitted by the agencies.**

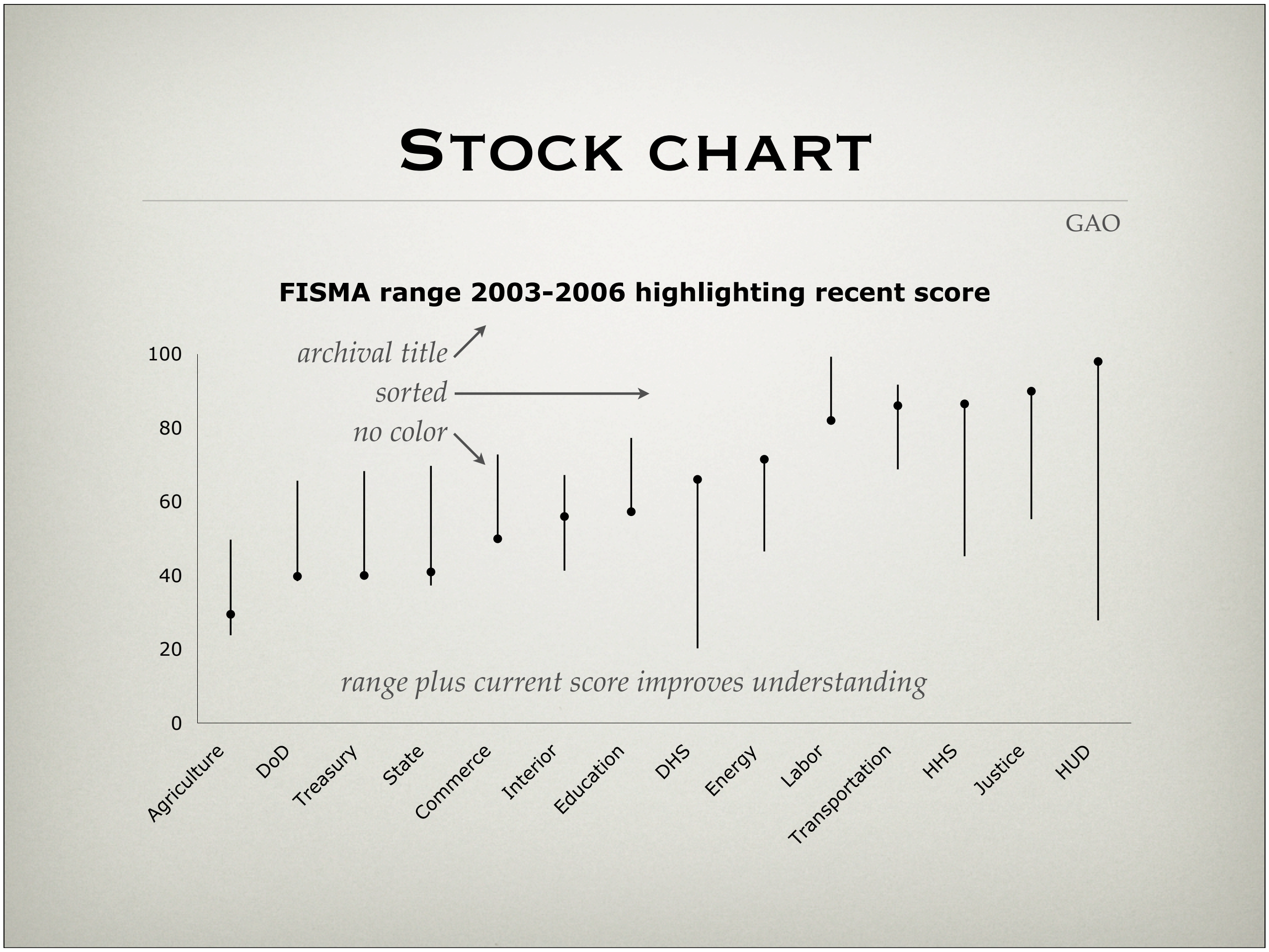
****The Department of Veterans Affairs did not provide its FY06 FISMA report.**

FISMA scores, courtesy of Richard Bejtlich's http://www.taosecurity.com/images/fisma2003-6.jpg



This now normalizes to a baseline, namely the number of vulnerabilities reported in 2003, and from there the multiple of that year that a given year represents.

Gives you a sense of scale for “Who is having a rough time?”



This now normalizes to a baseline, namely the number of vulnerabilities reported in 2003, and from there the multiple of that year that a given year represents.

Gives you a sense of scale for “Who is having a rough time?”

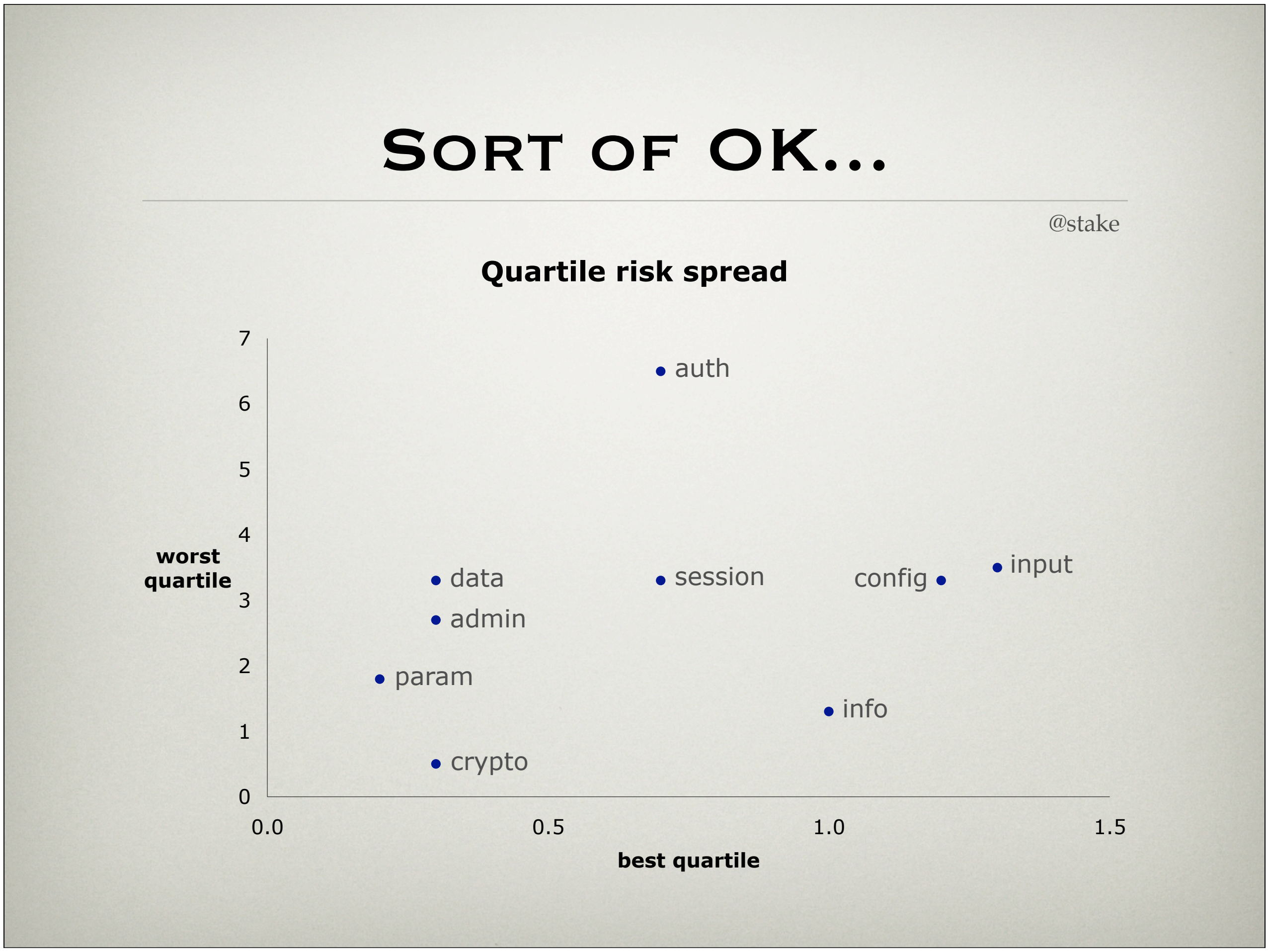
How to show...

@stake

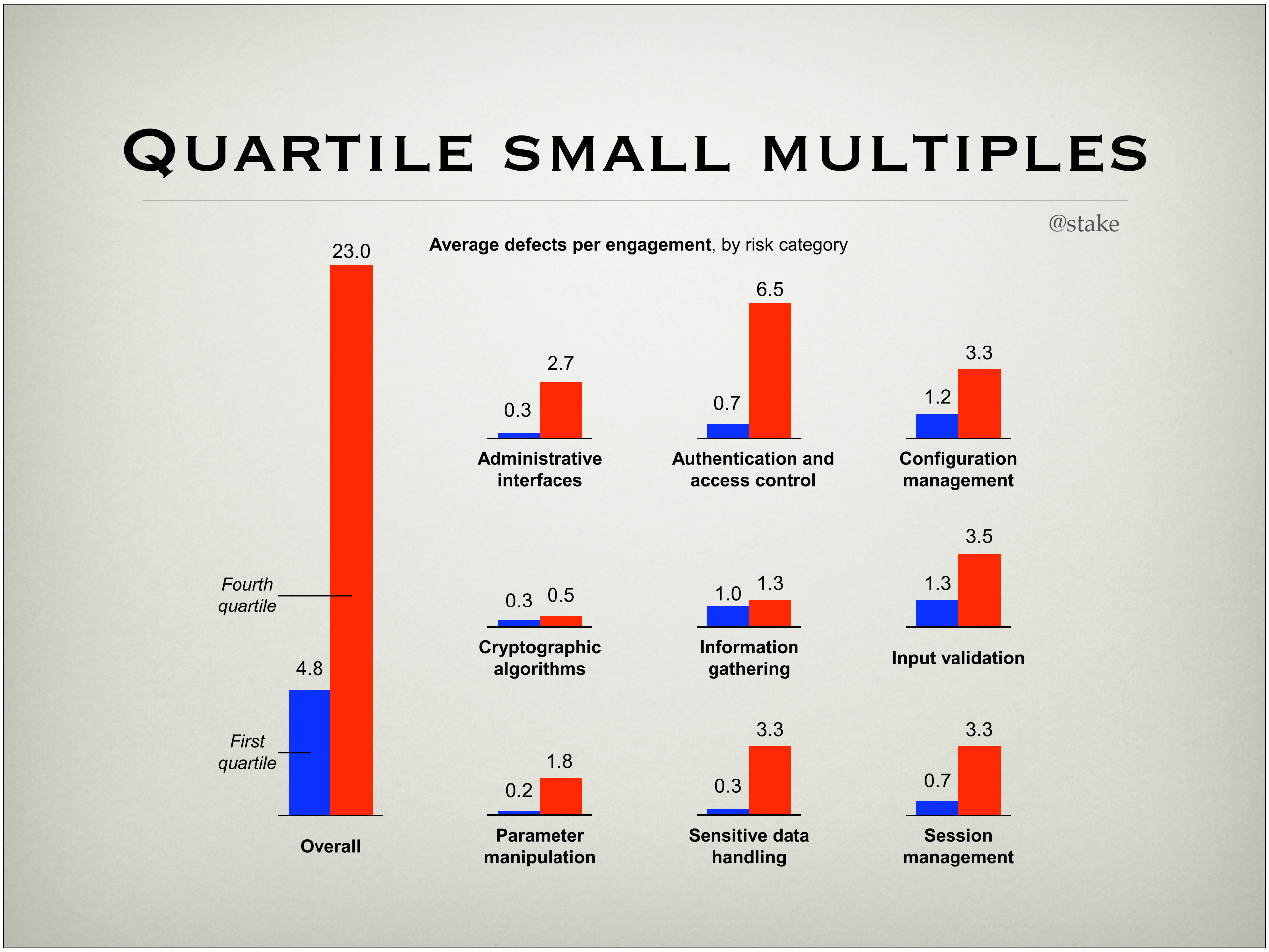
categoric quartiles

	Q1	Q3	Q3	Q4
Authentication & access control	0.7	..	..	6.5
Configuration management	1.2	..	..	3.3
Cryptographic algorithms	0.3	..	..	0.5
Information gathering	1.0	..	..	1.3
Input validation	1.3	..	..	3.5
Parameter manipulation	0.2	..	..	1.8
Sensitive data handlin	0.3	..	..	3.3
Session management	0.7	..	..	3.3

Taken from Geer DE, Jaquith A & Soo Hoo K: "Information Security -- Why the Future Belongs to the Quants," IEEE Security & Privacy, v1 n4 pp24-32, July/August 2003.



Sort of OK in that you can look at this and quickly see that cryptographic algorithms do not much matter in looking for further risk reductions but authentication and access control appear to be a very fruitful place to do further work as the separation between the first and last quartiles is broad there.



The data is discussed elsewhere, but the illustration is how small multiples displays lends itself especially well to quartile analysis.

How to show...

A mix of numbers, trends, and words in a single display that keeps a train of thought moving along?

Sometimes the point is to get a lot of data across without breaking the train of thought for something like “Turning now to Figure 4, you can see” when Figure 4 may well be on the next page.

Not wonderful...

As you can see in Figure 3,

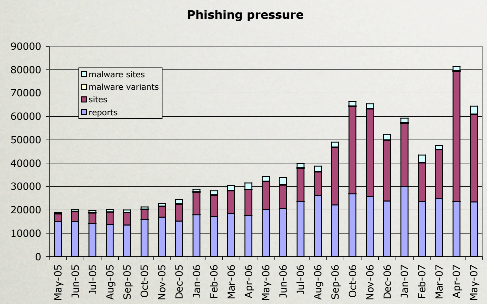
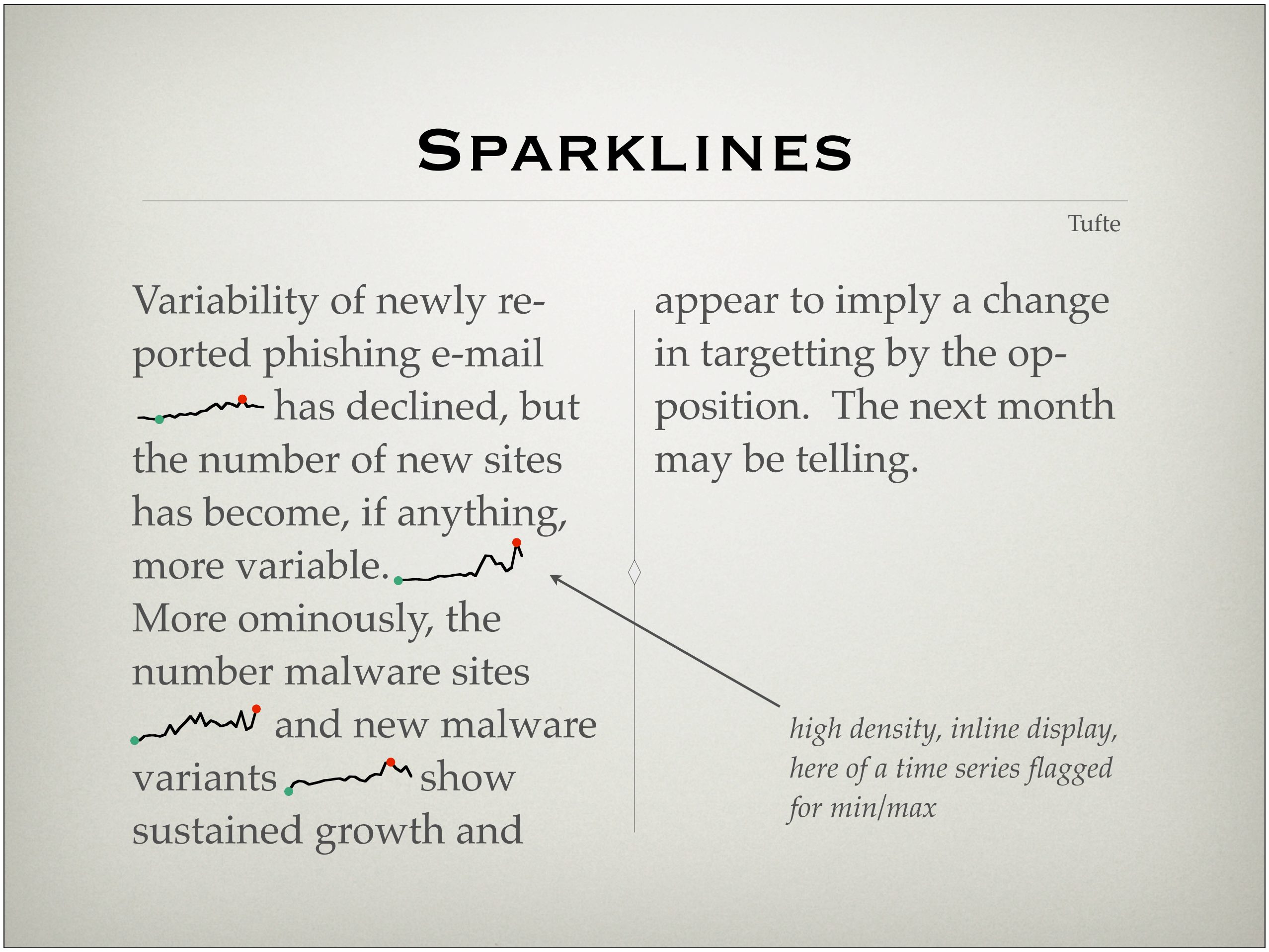


Figure 3

the variability of newly reported phishing e-mail has declined, but the number of new sites has become, if anything, more variable. Less obvious is that the number malware sites and variants show sustained growth and appear to imply a change in target-ting by the opposition. The next month may be telling.

The conventional method, in both academia and, for that matter, The Economist.



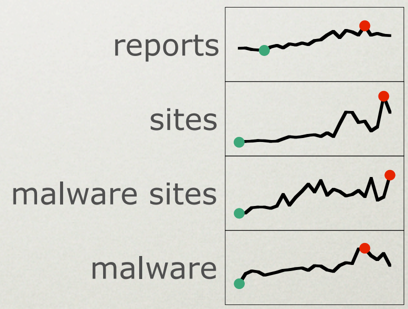
For discussion, see either Edward Tufte's Beautiful Evidence, 2006, or this URL

http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001OR&topic_id=1

Sparklines

Tufte

Variability of newly reported phishing e-mail has declined, but the number of new sites has become, if anything, more variable. More ominously, the number malware sites and new malware variants show sustained growth and appear to imply a change in targetting by the opposition. The next month may be telling. The last eighteen months look like this:



A different way to use them, that's all. Sparklines are also useful in dashboard applications

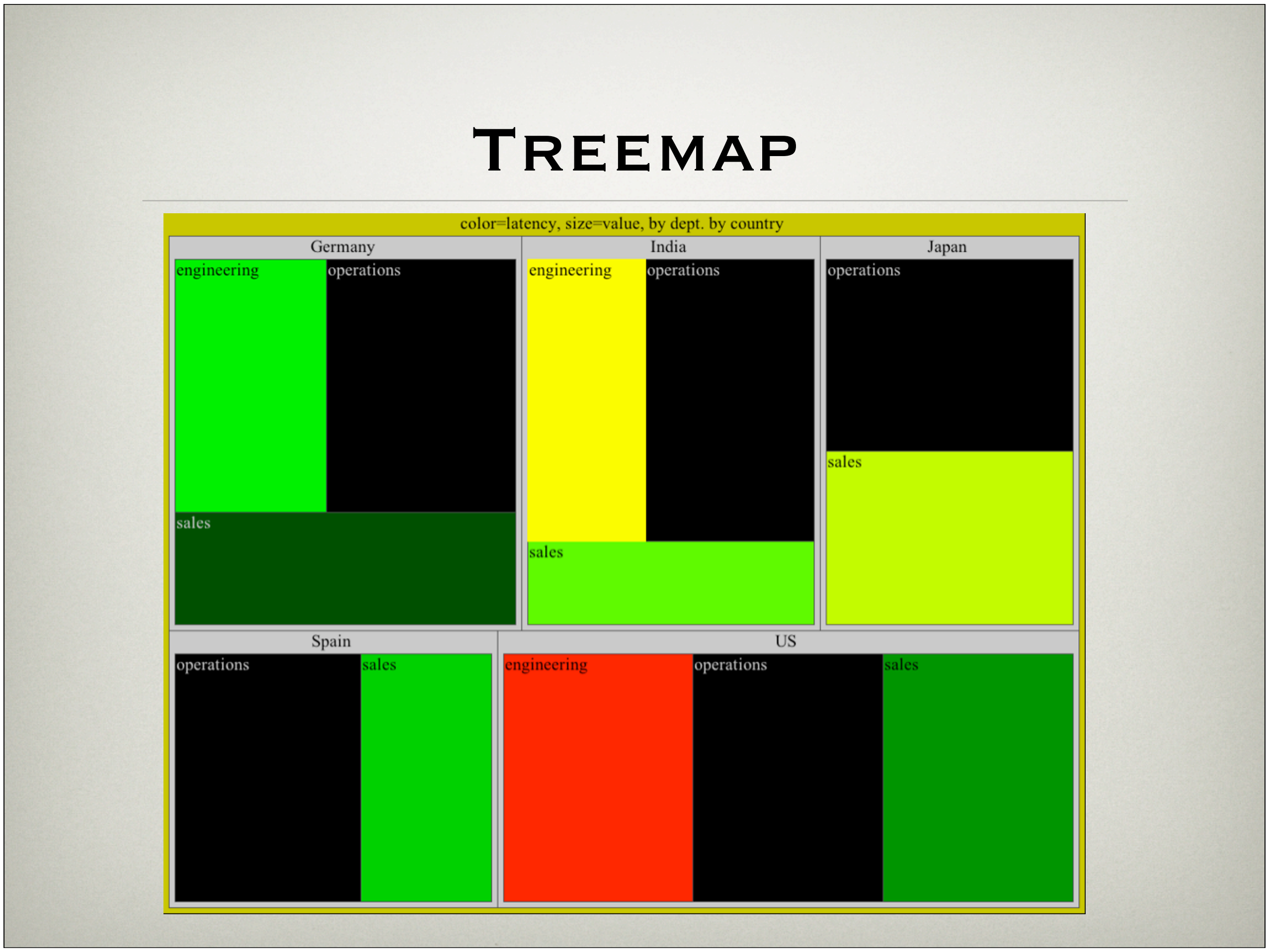
For some Python code to generate these, see http://bitworking.org/news/Sparklines_in_data_URIs_in_Python, though the ready may be amused to learn that these sparklines were done using Microsoft Excel charts (heavily abused) just to see if it could be done.

How to show...

Nested hierarchy of problems

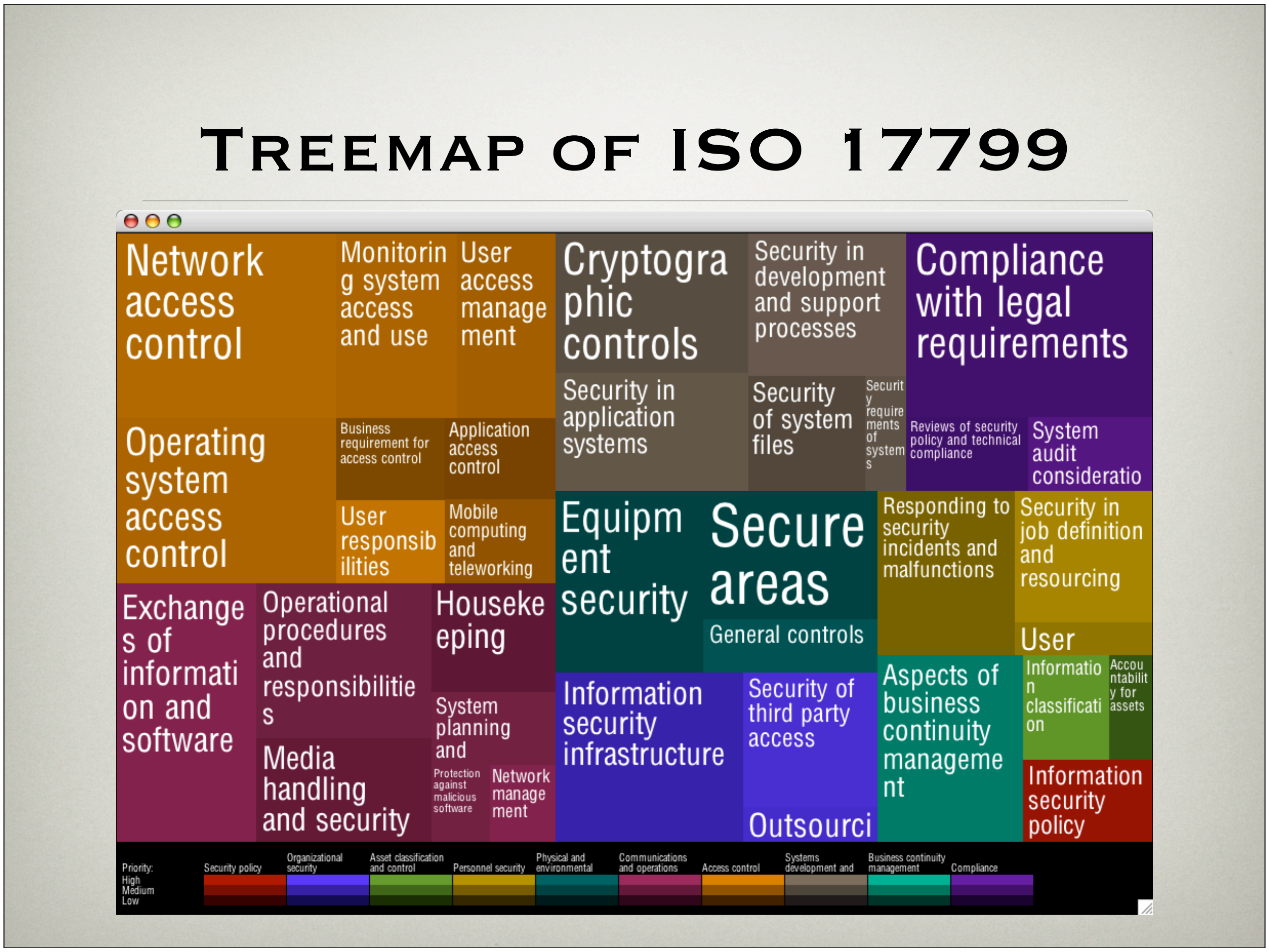
sector	country	patch latency	value at risk
sales	Spain	12	0.7
sales	Japan	22	0.9
sales	India	18	0.5
engineering	US	40	1
engineering	Germany	16	0.8
engineering	India	28	0.7
operations	US	5	1
operations	Germany	5	1
operations	Spain	5	1
operations	Japan	5	1
operations	India	5	1

Not an everyday need, but when hierarchy is the core of how decisions must be made and accountability rendered, it is important for the management and the technical sides of the house to agree on something, and a picture is often the best method to do so.



Ben Schneiderman's "Tree Map" is something you probably want to learn about but which is too much for today's discussion to do justice to. His home page, http://www.cs.umd.edu/~ben/, has several references.

The picture is fictional data to illustrate how patch latency might be described for management purposes. In this case, it is a location by function breakout where size of a block is scaled to number of seats while the color indicates latency -- in this case black is lowest latency (hence no concern) and runs upward through green, yellow, and finally to red where latency is highest. In this fictitious data set, one might conclude that operations has a common and quite effective latency minimization while US Engineering apparently ignores patching, etc.



The ISO 7799 standard, at a view depth of two, showing what is relatively important to what. See http://www.freshcookies.org/jtreemap/iso-example.png.

From numbers to Inference

Trivial example

- Setting:
  keycard required for entry and exit
- Measure:
  simple count of keycard use
- Infer:
  odd numbers imply tailgating

Example due to Jaquith.

"Laws of vulnerabilities"

Qualys

- Multi-year field observation type study
- Some selection bias
- Meta-analysis = "measure of measures"

In a multi-year field observation study where company's exteriors were scanned for vulnerabilities and the results pooled for descriptive purpose, Gerhard Eschelbeck of Qualys came to publish his "Laws of Vulnerabilities" as found at http://www.qualys.com/research/rnd/vulnlaws/

Yes, there is selection bias in that the company's scanned invited the scanning, thus proving that they have an interest in security that is probably in excess of the average interest. Nevertheless, this combining of multiple analyses into one is itself an analysis, a meta-analysis, and it is a measure or measures of some real value.

Half-life

Qualys

The length of time it takes users to patch half of their systems

Eschelbeck noted that patching behavior is like radioactive half-life; each succeeding fixed interval of time has the same fall-off in the number of patched systems.

Half-life, cont.

Qualys

$t_{1/2}$	2004	2005	
external	21	19	-10%
internal	62	48	-23%
	3.0	2.5	

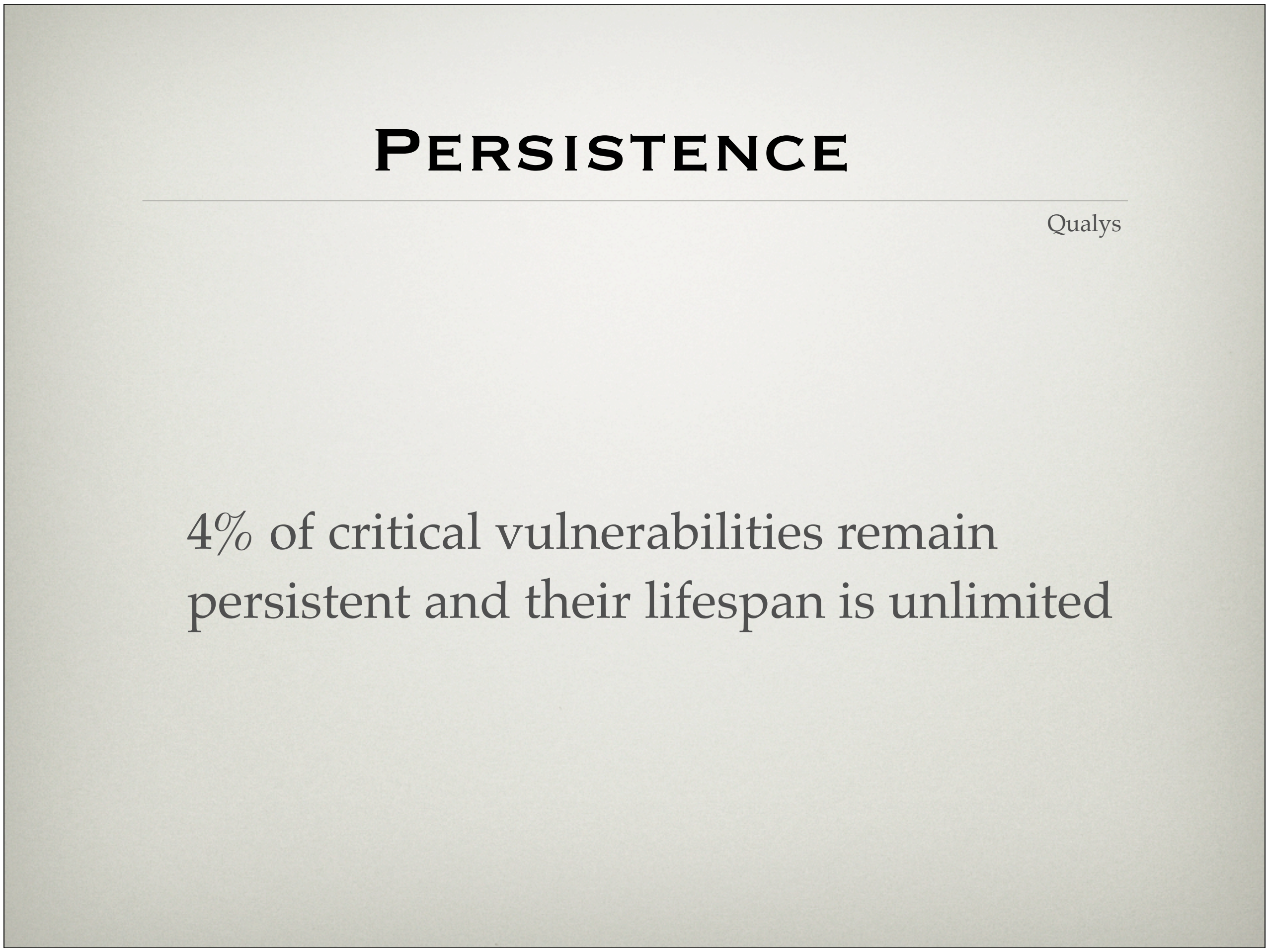
Over one year of observation, here 2004 through 2005, the half-life constant for internal systems changed from 62 days to 48 days, a reduction in patch latency of 23%, a better figure than the 10% reduction seen in external machines. On the other hand, external machines were three times quicker in 2004 in getting patches and still a respectable 2.5 times as fast in 2005.

Prevalence

Qualys

50% of the most prevalent and critical vulnerabilities are replaced by new vulnerabilities on an annual basis

In turn, Eschelbeck makes some observations, the first of which is that half of the problems to really worry about are new each year.



and that one in twenty-five of all vulnerabilities is effectively immortal,

Focus

Qualys

90% of vulnerability exposure is caused by 10% of critical vulnerabilities

and that there is a 90/10 rule for magnitude of danger and count of vehicles for carrying that danger

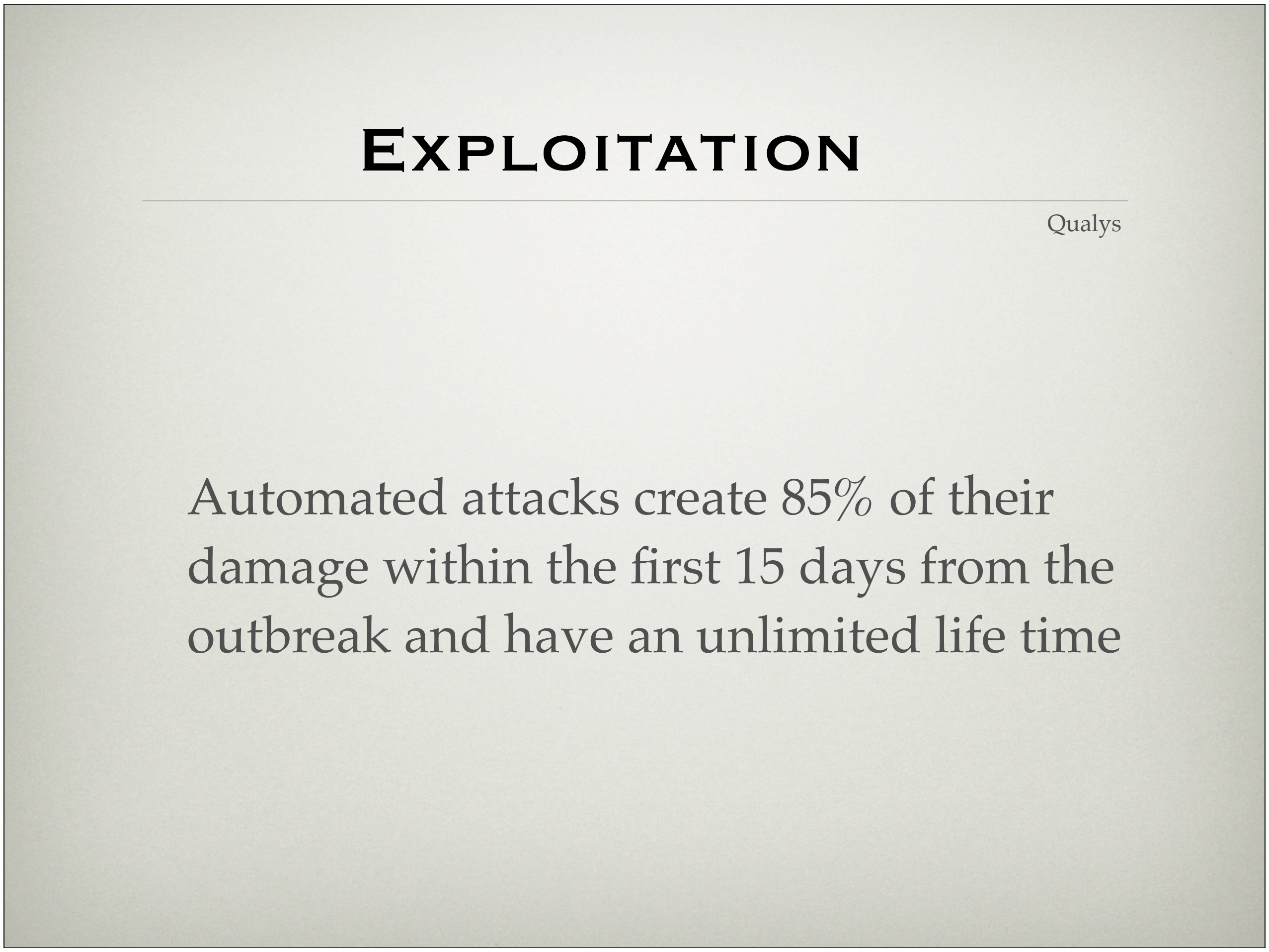
Window of exposure

Qualys

The time-to-exploit cycle is shrinking faster than the remediation cycle;

80% of exploits are available within the first half-life period of critical vulnerabilities

As we all have guessed, the interval between warning and attack shrinks. Eschelbeck's data is that in that first half-life decline from 100% unpatched to 50% patched, 80% of all exploits become available. This means that in patching one has the quick and the dead.



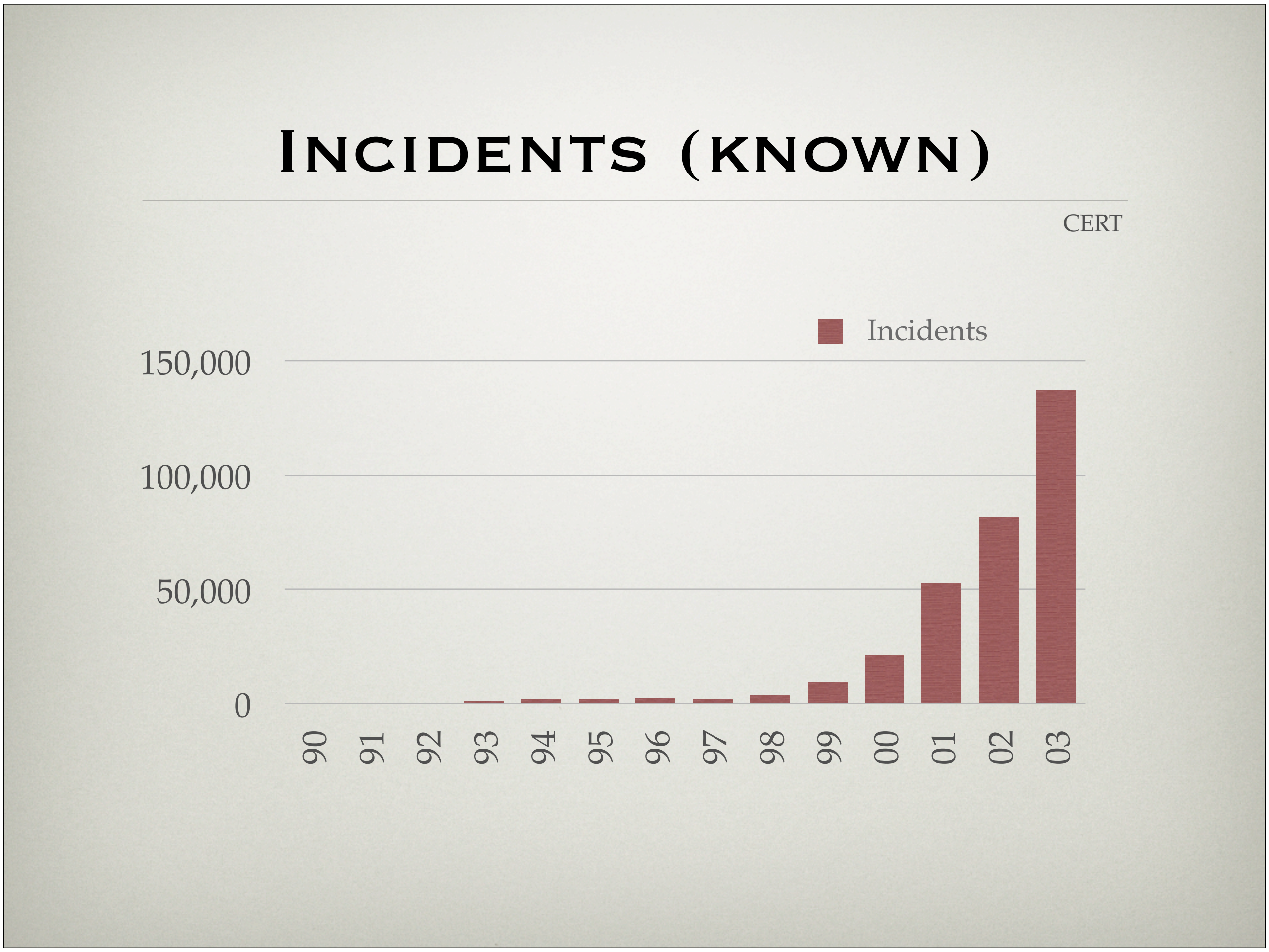
Automation is, of course, what in many ways separates digital security from physical security more than any other. This tends to front-load the damage to the earliest parts of the period of susceptibility, and in Eschelbeck's data that is 85% of damage in days 1-15.

Something to think on

A plane lands in England full of bullet holes. The repair mechanics suggest armoring the plane where the holes are. The pilot notes that the planes which come back do not have the kind of bullet holes that need armoring.

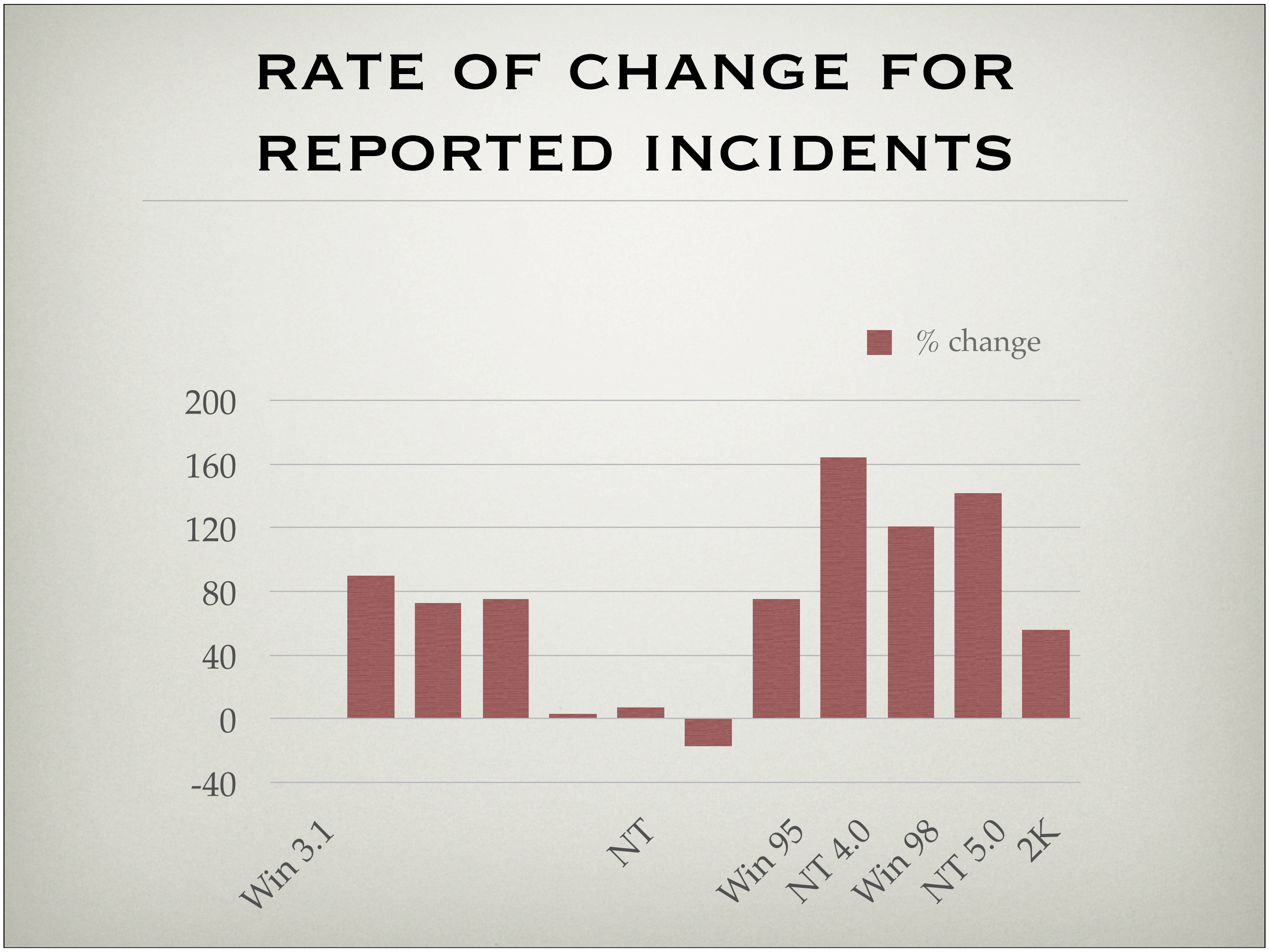
An apocryphal story, but a good one. It reminds us that what we can observe has already been through some filter. In this case, the returning planes were still flyable so the bullet holes they carried were, ipso facto, not the ones to fear most.

Yet more trend analysis



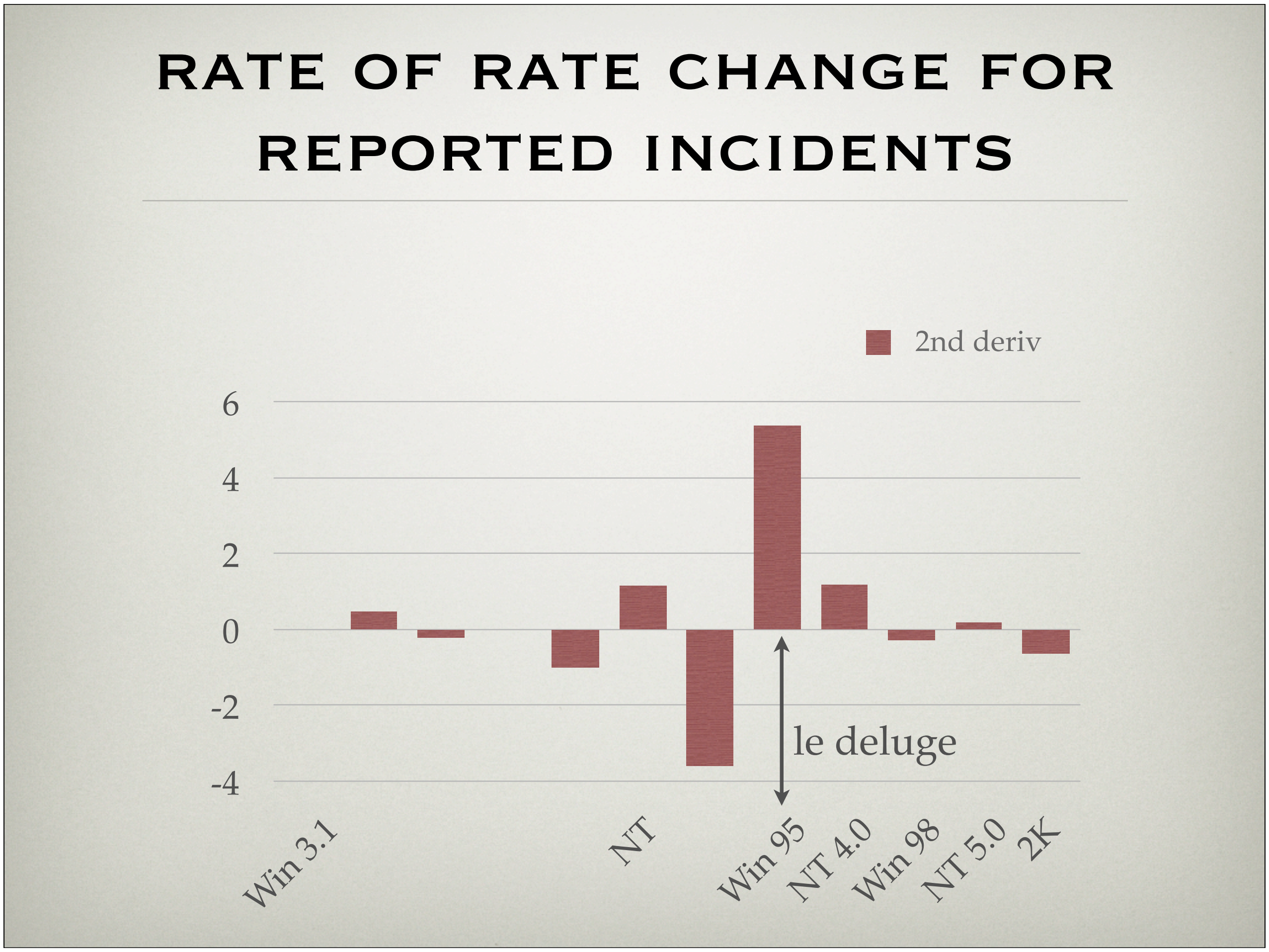
Public CERT data, http://www.cert.org/stats/ for numbers of incidents reported (which they stopped doing in 2003, hence the cutoff).

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
252	406	773	1,334	2,340	2,412	2,573	2,134	3,734	9,859	21,756	52,658	82,094	137,529



The calendar is per IDC, and added to the incidents per CERT, we can now look at the rate of change of incident reports to CERT demarcated by the dates of product release by Microsoft (MSFT has a 94% market share so we can safely ignore all other vendors here).

Win 3.1	Win NT	Win 95	NT 4.0	Win 98	NT 5.0	Win 2K	Win XP
1990	1995	1997	1998	1999	2000	2001	2002



Now we look at the rate of rate of change, i.e., the second derivative. In this case we now see when the problem began: Windows 95 and its introduction of a TCP/IP network stack. This suddenly glued an operating system that had been designed for a single authoritative user on a truly local network to the entire world including a lot of Bad Guys. The rest is history. The realization that the Internet was important caused Gates & Co. to expose their unprotected user base to everyone else. All else follows.

Use in security

- Don't just look at trends
- Look at rate of change of rate of change
  - What is correlated with changes in the rate of change?
  - If $X \propto Y$ and $Y = f(t)$, then pull out the impact on X of changes due only to t

In security, trends are almost always good enough for decision making but sometimes you want trends of trends, as the previous pictures attempted to illustrate. When doing this, you will generally be looking for correlated events that correspond to sharp rate changes. In other words, when X is proportional to Y and Y is a function of time, then try to find what part of the change of X is due to time in and of itself.

Economics

The question is now

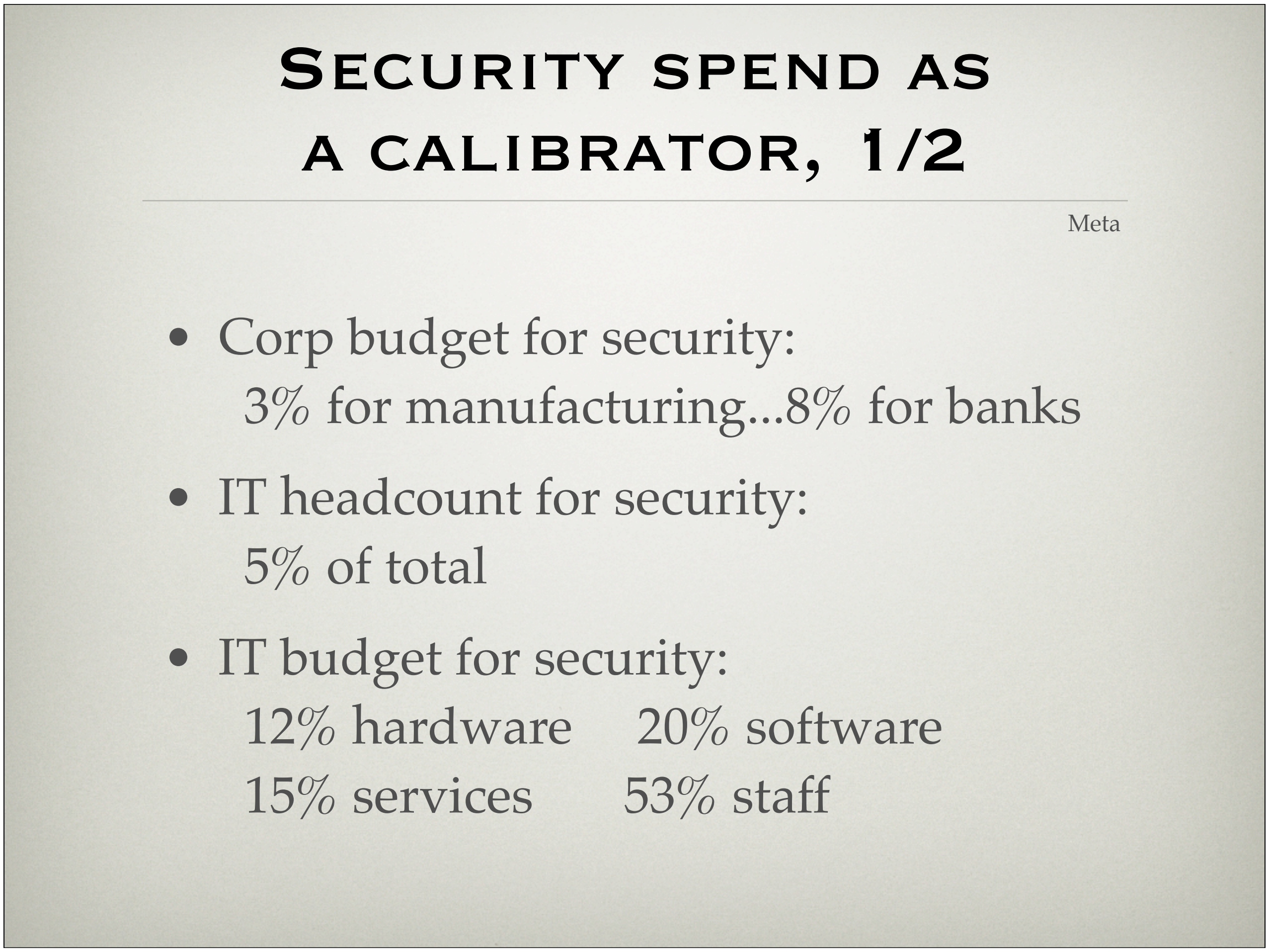
"The next ten years will be a referendum on whether we consume the entire productivity growth of the US economy for increased security spend." [*paraphrase summary*]

Chief US Economist, Morgan Stanley
Op-Ed, NY Times, 23 October 2001

"The Terror Economy," Richard Berner, NY Times, 23 October 01, Page A23, Column 1

ABSTRACT - Op-Ed article by Morgan Stanley economist Richard Berner warns that war against terrorism will impose long-term economic costs in form of higher insurance and security costs, maintenance of larger inventories and new Internet security measures; explains that spending more on defense will erase decade-long 'peace dividend' and crowd out other investments that helped transform budget deficits into surpluses.

Full article available from author; no longer available online. The point is real for us: We cannot be the opponent of wealth creation by withdrawing all the productivity growth our economy provides and using it for our non-productive purposes.



The Meta Group, Diamond report #2856, recommendations on how much of IT budet should be allocated to security spend. Regardless of whether yoiu believe these numbes, once published they are the numbers you have to prove your numbers are better than if you want your numbers to now be the de facto standard.

Security spend as a calibrator, 2/2

Gartner

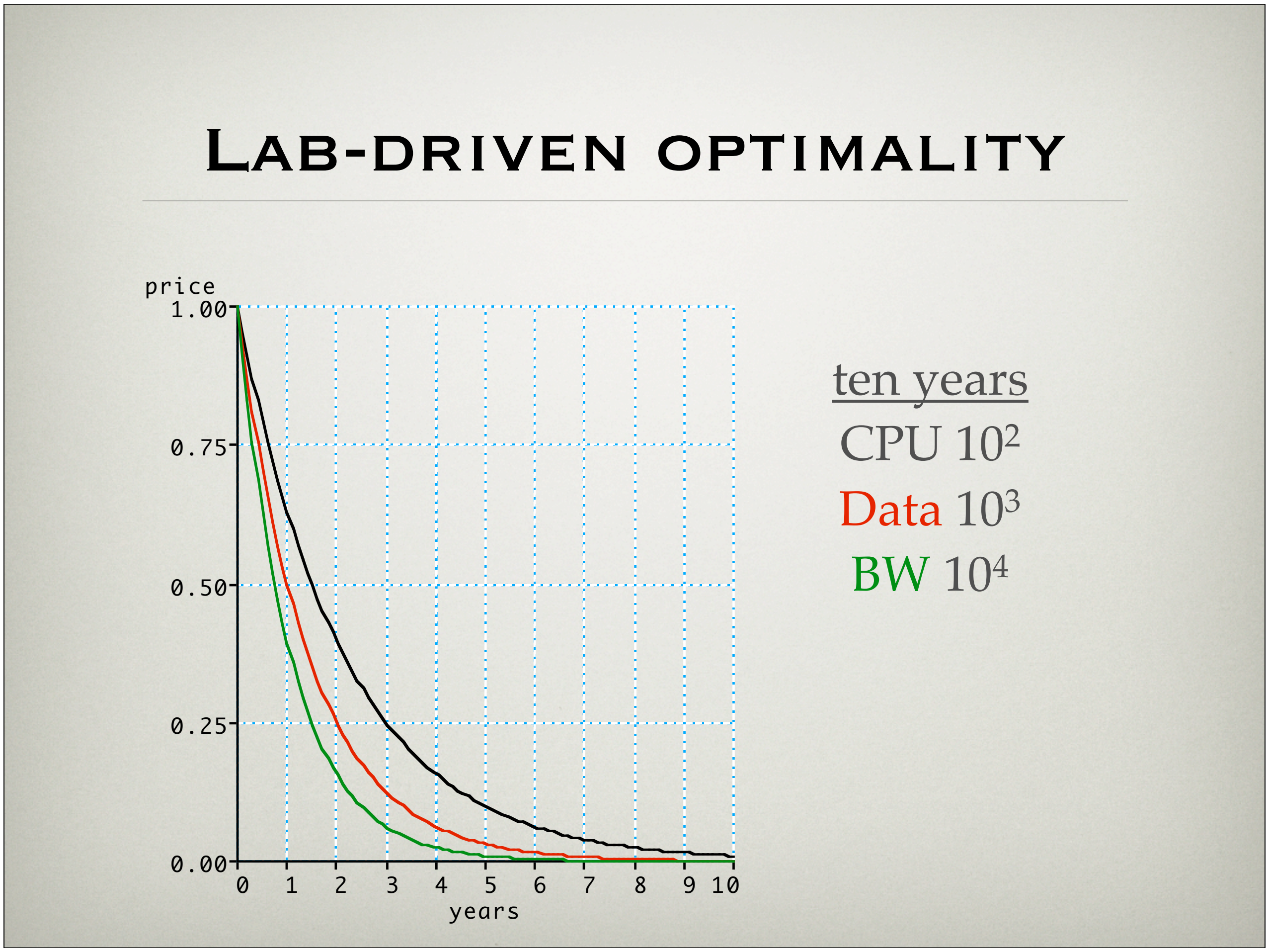
- Corp budget for security:
  3-6%, visibility and size as drivers
- Security software
  small companies greater budget %age
- Security %age spending by sector:
  health > government > education > ...

A similar set of de facto numbers, this time from The Gartner Group, Research #G00126733, recommendations on how much of IT budet should be allocated to security spend.

Security's focus changing

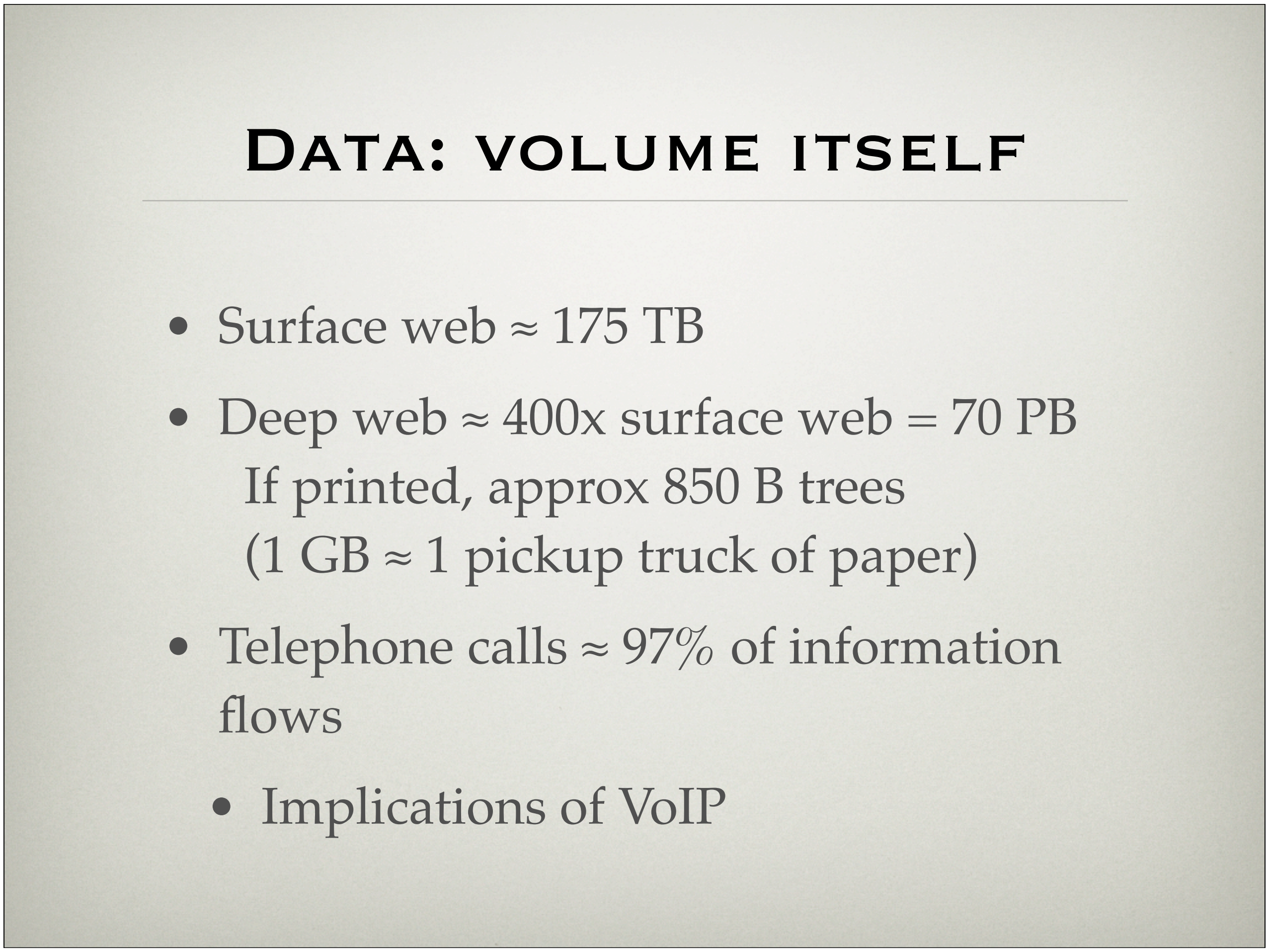
- Long-term trend is towards
  - greater data-richness
  - greater data-mobility
- If true, then implications for planning
- Why is it true?

The long-term prospect for computing is that it changes over time based on predictable pressures; see following slides. The short form is that data takes command. If so, this is where our metrics need to go.



Black line is “Moore’s Law” whereby $/MHz drops by half every 18 months. It’s unnamed siblings are, in red, the price of storage (12 month) and, in green, bandwidth (9 month). Taken over a decade, while CPU will rise by two orders of magnitude, the constant dollar buyer will have 10 times as much data per computer cycle available but that data will be movable to another CPU in only 1/10th the time. This has profound implications for what is the general characteristic of the then optimal computing plant.

And, even if there are wiggles here and there, the general point that there is a drift over time in the optimal computer design stands.



The volume of data as estimated by Lyman & Varian in their annual survey on “how much information?”, specifically that the apparent World Wide Web is 175 terabytes whereas the total spinning data volume is estimated at four hundred times as large, meaning 70 petabytes. Were the “deep web” printed it would consume 850 billion trees. As a rule of thumb, 1 gigabyte of information would, if similarly printed, fill a pickup truck.

If, as Lyman & Varian suggest, 97% of information flows are in telephone form, then a wholesale trend to “Voice over Internet Protocol” will have profound implications to the amount of data exposed to threat and the nature of that threat.

See http://www.sims.berkeley.edu/research/projects/how-much-info/

Data: volume growth

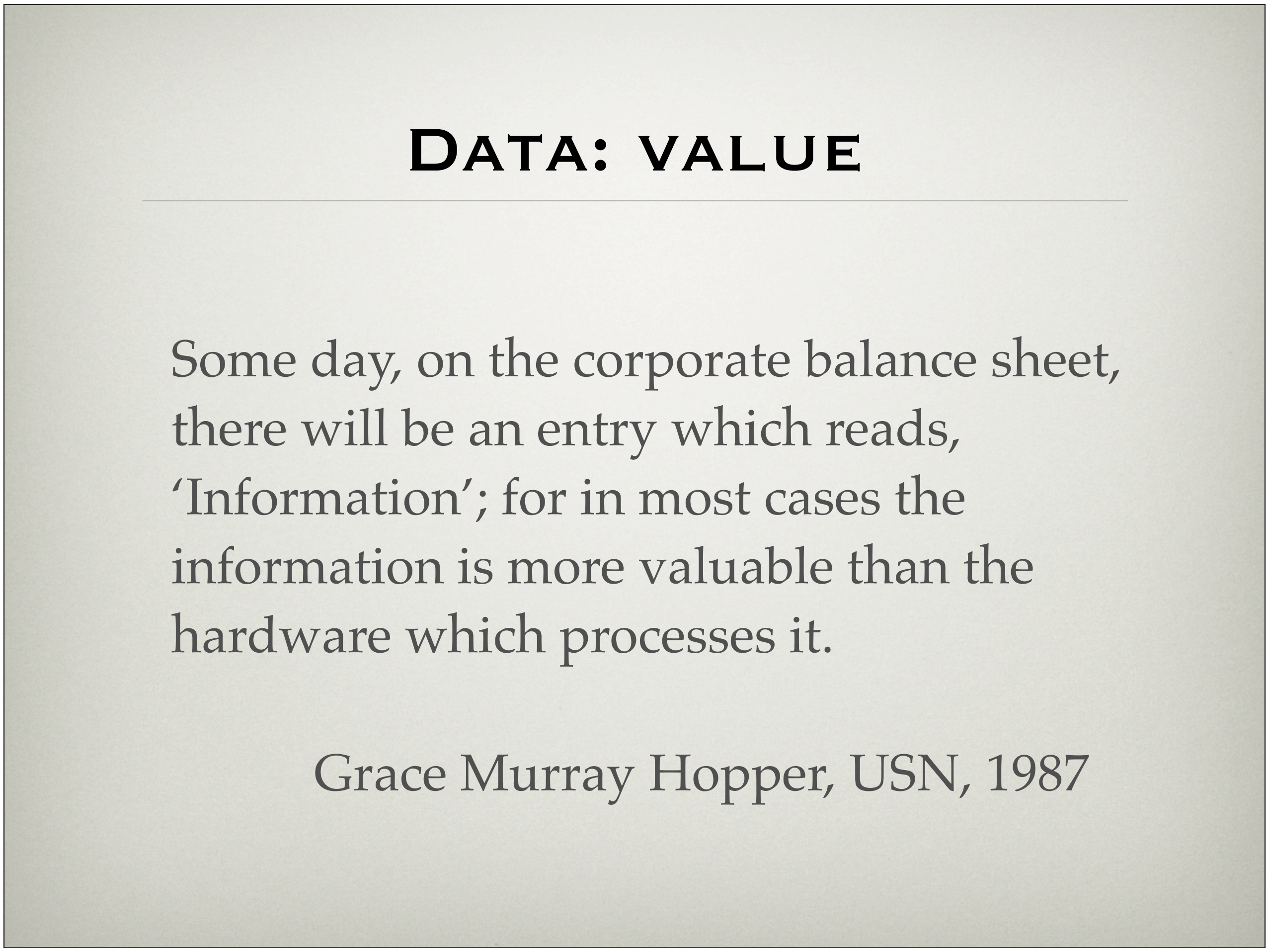
- Corporate IT spending on storage:

  4% in 1999 v. 17% in 2003 (Forrester) for net capacity of +150%/year (Gartner) while bits/$ up 16x in same interval (vide supra)

- Retained volume doubling at ~30 months

The volume of data is substantial, getting more so, and will likely dominate security's rational focus from this point forward.

Forrester and Gartner numbers from research documents (subscription).



Rear Admiral Grace Murray Hopper, USN (Ret), Washington, D.C., 1987.

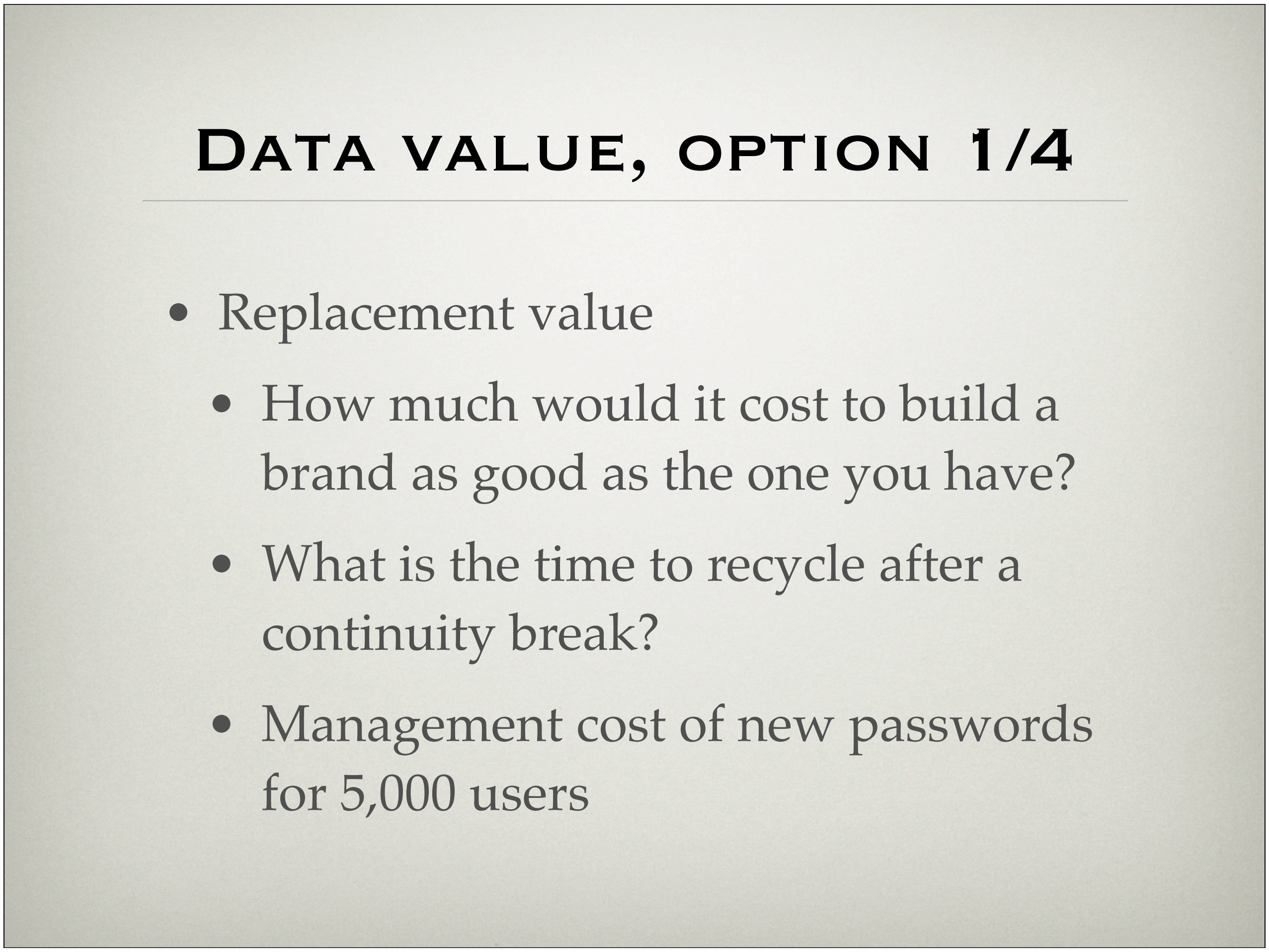
Question for the reader: Is that point now? Has it already occurred? Where do information assets appear on the balance sheet and/or how are they treated when describing shareholder value?

Data: value

The information about the packages we ship is more valuable than the packages themselves.

Fred Smith, Federal Express, ca 1990

Fred Smith, founder of Federal Express.



You ask a management team “How much is your brand worth?” and you get blank stares or wild guesses. Try it a different way, ask “How much would it cost you, knowing what you know today, to build a brand from scratch as good as the one you have now?” This will get an answer that is probably a lower bound for replacement value. If such a value is sufficient basis to make whatever managerial decision around security that is on the table, then that is good enough for the time being.

Similarly, if your business has a “non-interruptibility” requirement, such as continuous monitoring of weather conditions for a period of time before a power plant can be sited, then the re-formulated form of “How much is your information worth?” would be more like “How much incremental cost would you incur if your continuity of measurement were broken and you had to start over?”

A different sense of the value of good passwords or good password protection would be to not ask “How much are your passwords worth?” but rather “If today you had to get all 50,000 people in your firm to pick a new password within 36 hours how much incremental cost would you incur?”

Data value, option 2/4

- Black economy market price
  - AOL screen names: 0.1¢/name
  - Bot-net host rental for spam: $1/wk
  - Deadbeat details: $10/name
  - Financial screenshot: $500
  - Game skin 90 days out: $50,000

A different way to look at the value of information is to ask what the black market pays, if indeed that is a question that can be answered in a way that is sufficiently close to where you are to be valuable via analogy. For example, a thief was paid $100,000 for 92MM AOL screen names.

Computers that are taken over silently are occasionally rented to others, e.g., as spam relays. The rental fee approximates $1/week by some estimates. That tells you at the very least that the supply of machines taken over is great as such a price is obviously slight. That would mean that your data on your machine is, by analogy, very easy to get at by others. If you don't know how easy it is, then you would conservatively assume that breaking into your machine is worth a dollar on the open market.

More directly, a major west coast bank reports that its tellers are routinely offered $500 per screenshot of customer identifying data for customers with over $50,000 of assets. So a clerk making $10/hour can give herself an after-tax raise of $26,000/year for the price of one sheet of paper per week. Not every clerk is immune to this temptation.

The other prices are based on publicly reported events.

Data value, option 3/4

- Future economic value
  - From eureka to FDA filing costs circa $700M, 80% is information
  - Derivative pricing algorithm alone carried on books as $300M
  - Patent losses: CDMA in India & China at $750M/annum

In a pharmaceutical company, the critical period begins with the “Eureka!” moment and closes with the FDA formal filing. In this interval, the pharmaceutical can expect to spend $700,000,000 at the end of which 80% of the value is the information in the can. This is a hard to get figure and was obtained in conversations variously.

A single bank in NYC that is known for its derivative trading carries its apparatus for pricing same as a $300,000,000 asset.

The inability of Qualcomm to effectively patent its CDMA technology in China and India represents an information loss to them of $750,000,000 per year based on current usage rates of the CDMA technology.

Data value, option 4/4

Lindstrom

- Lower bound value
  - At least as much as the total IT budget including depreciation & amortization

A painful observation by Pete Lindstrom, Spire Security.

Data value, option 5?

- Ransom value
  - "Ransomware" fast rising, and faster becoming unbreakable
  - In a sense, this is yet another valuation

"The Crimeware Landscape: Malware, Phishing, Identity Theft and Beyond" A Joint Report of the US Department of Homeland Security, SRI International Identity Theft Technology Council, and the Anti-Phishing Working Group

More at http://www.antiphishing.org/reports/APWG_CrimewareReport.pdf

An obvious conclusion: the economics of security and what it is for point us towards data as a focus rather than infrastructure, as has been the case heretofore. Our metrics actually do have to follow.

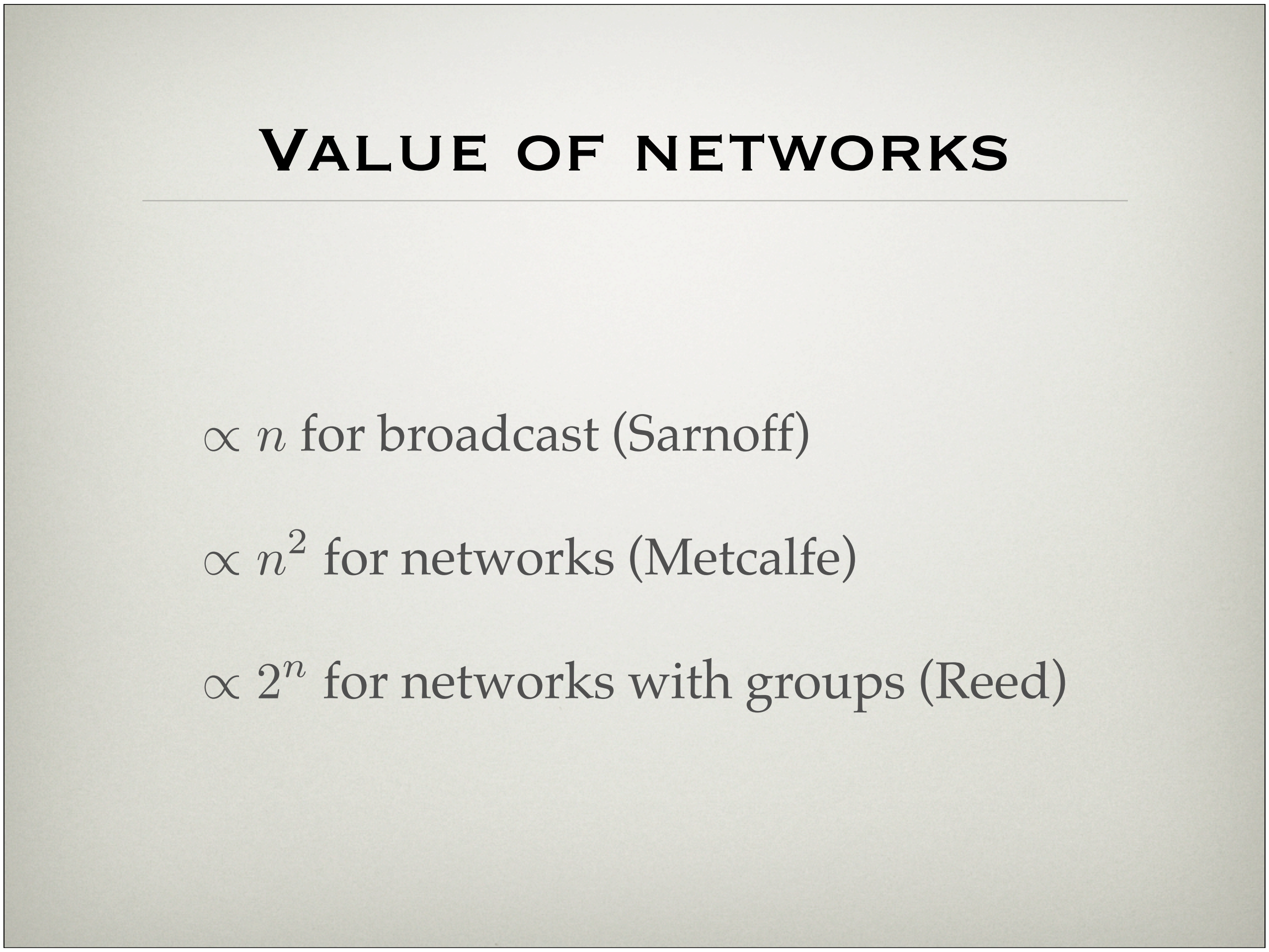
Data value, alternate

- "Hey, just add it up!"

  Info Asset Val ≥ Salary(IT)
  + Capital Expense
  + Salary(Other)
  + Revenue(IT)
  + Intellectual Property

- If inequality sufficient for decision, then all well and good

Of course, you can value data a different way -- the budgetary way. If this gets a decision, all well and good. See value #4 (Lindstrom) for what this is a variation on.



Side issue, but networks have value. You have probably heard of the second. The first is for a broadcast network like radio, which says that the number of listeners is the proportionality constant for the value of the network itself (Sarnoff). The second is for communications networks like the Internet, and it says that the number of potential conversations is the proportionaility constant (Metcalfe). The third is more sociologic; it says that sub-groups and not individuals are what make networks valuable and thus the number of possible groups is the proportionality constant (Reed).

No absolute security

- Security absolute only w.r.t. opponent
- Opponents mutate
- Hence cannot stay absolutely secure even if once achieved

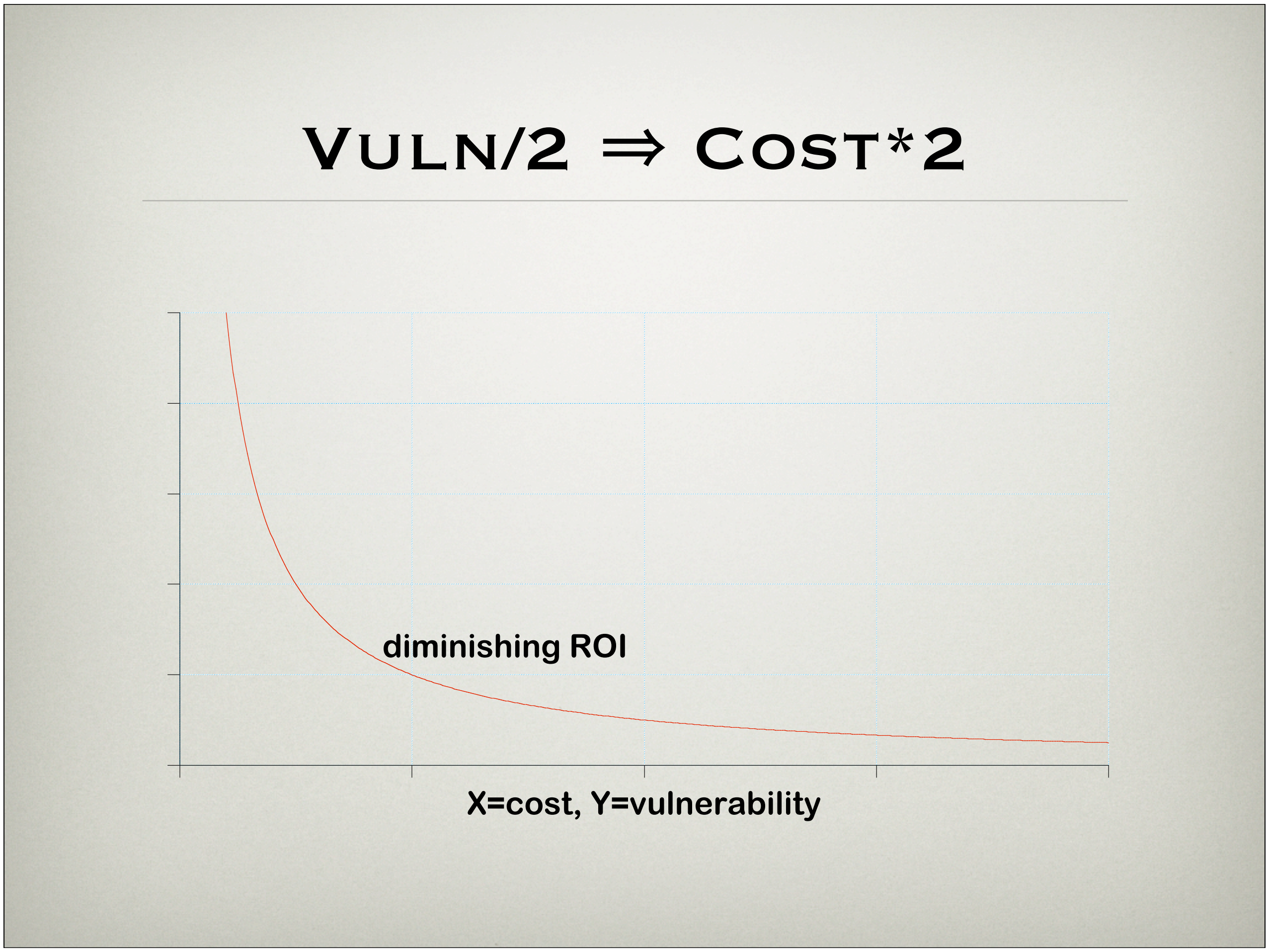
Ian Grigg, in a paper on security economics, reminds us that “secure” has a sub-text of “secure against opponent X” but since opponents change often it is thus proved that you cannot stay absolutely secure even if you might achieve it for an instant; see http://iang.org/papers/pareto-secure.html

2004 Turing lecture

Adi Shamir

- Absolutely secure systems do not exist
- To halve your vulnerability, you have to double your expenditure
- Cryptography is typically bypassed, not penetrated

Repeating Adi Shamir, with emphasis.



This is what Shamir's statement looks like and it in its own way illustrates how a defense in depth strategy is a better strategy; diminishing returns along one axis can be abandoned for robust returns along another axis.

Pareto efficiency

- *Pareto improvement*:
  - someone is better off, and
  - no one is worse off
- *Pareto efficient*:
  - no improvements left

As Grigg points out, the economics literature has a concept of a Pareto improvement (rate) and Pareto efficient (state) that may as well be applied to security.

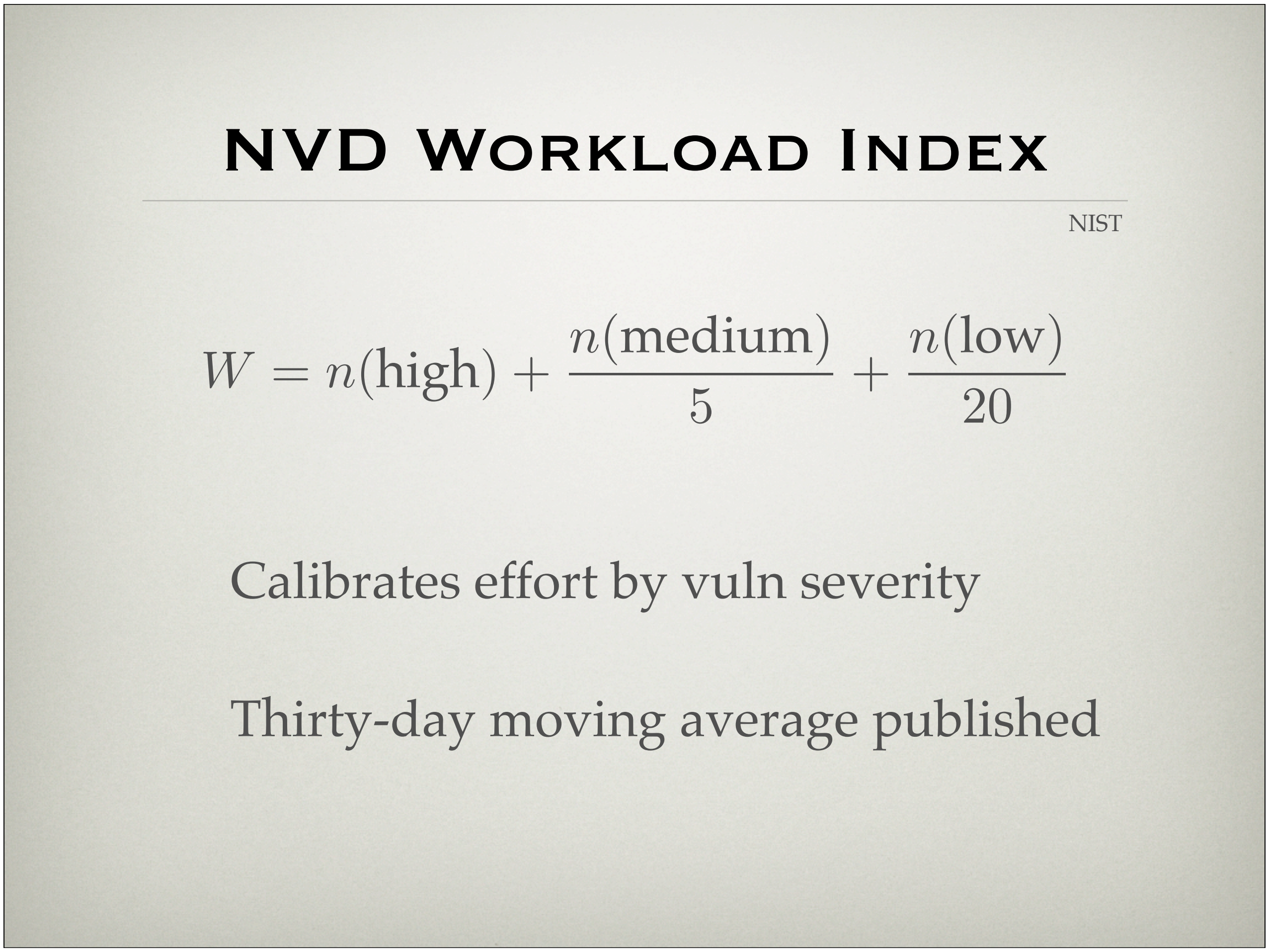
Pareto secure

Grigg

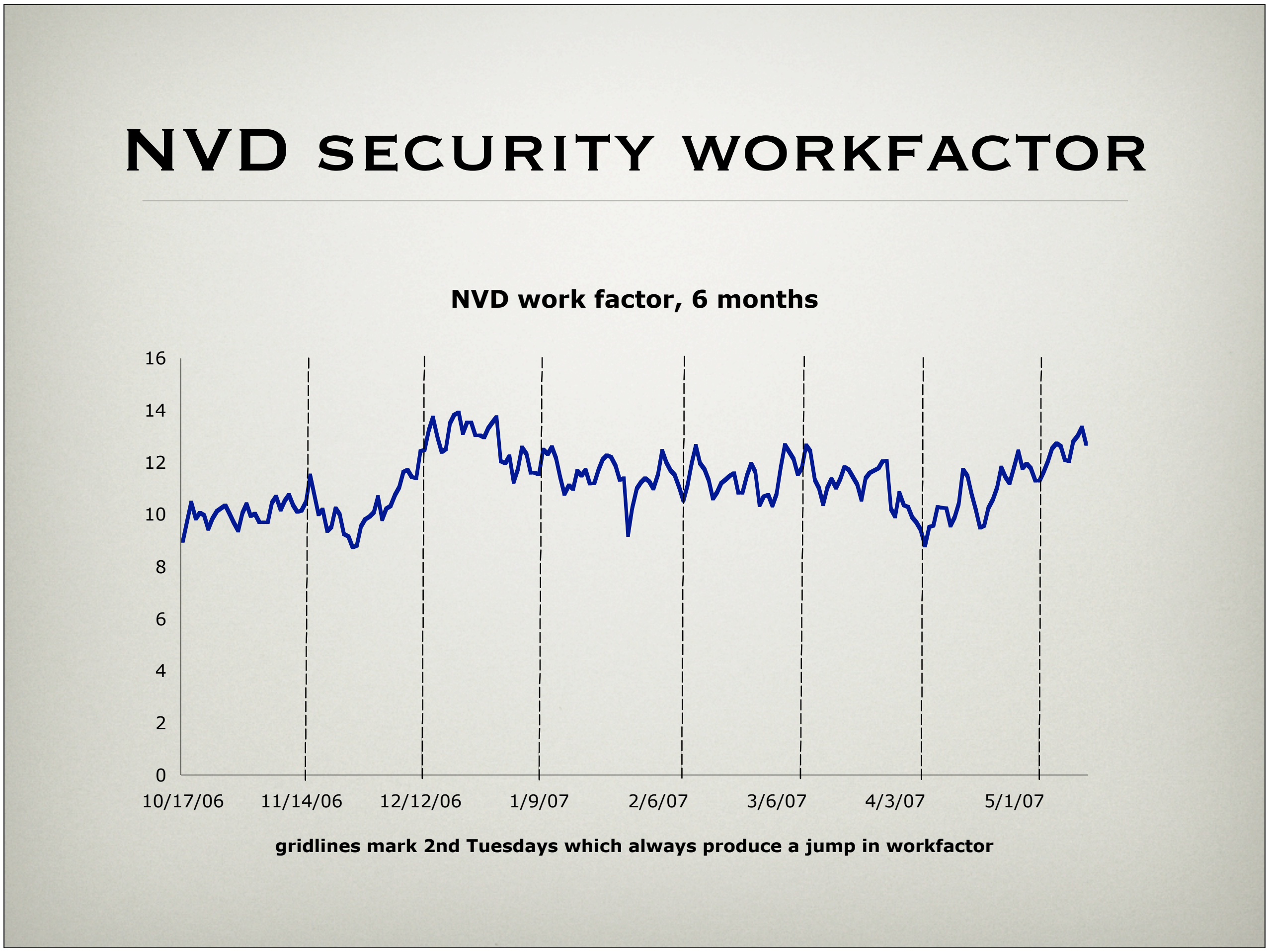
- Proposed change increases security in some area, and
- Does not decrease security in any area
- Another form of CE analysis

Example: AES versus DES

In particular, a Pareto secure state is one where there are no changes that can be made which are unqualifiedly good; see again Grigg I : Pareto-Secure, r1.6, Systemics, 2005, http://iang.org/papers/pareto-secure.html



The National Vulnerability Database as seen at http://nvd.nist.gov/nvd.cfm?workloadindex has a different metric more of interest to operational people than any other but in this case it is a work-load predictor based on the current vulnerability rankings. As you can see, it has a ratio scale for work (1-0.2-0.05) and what they publish is a thirty-day moving average. This can be adapted to other uses or, more likely used as a calibrator to some metric you are yourself using, which is why it is here.

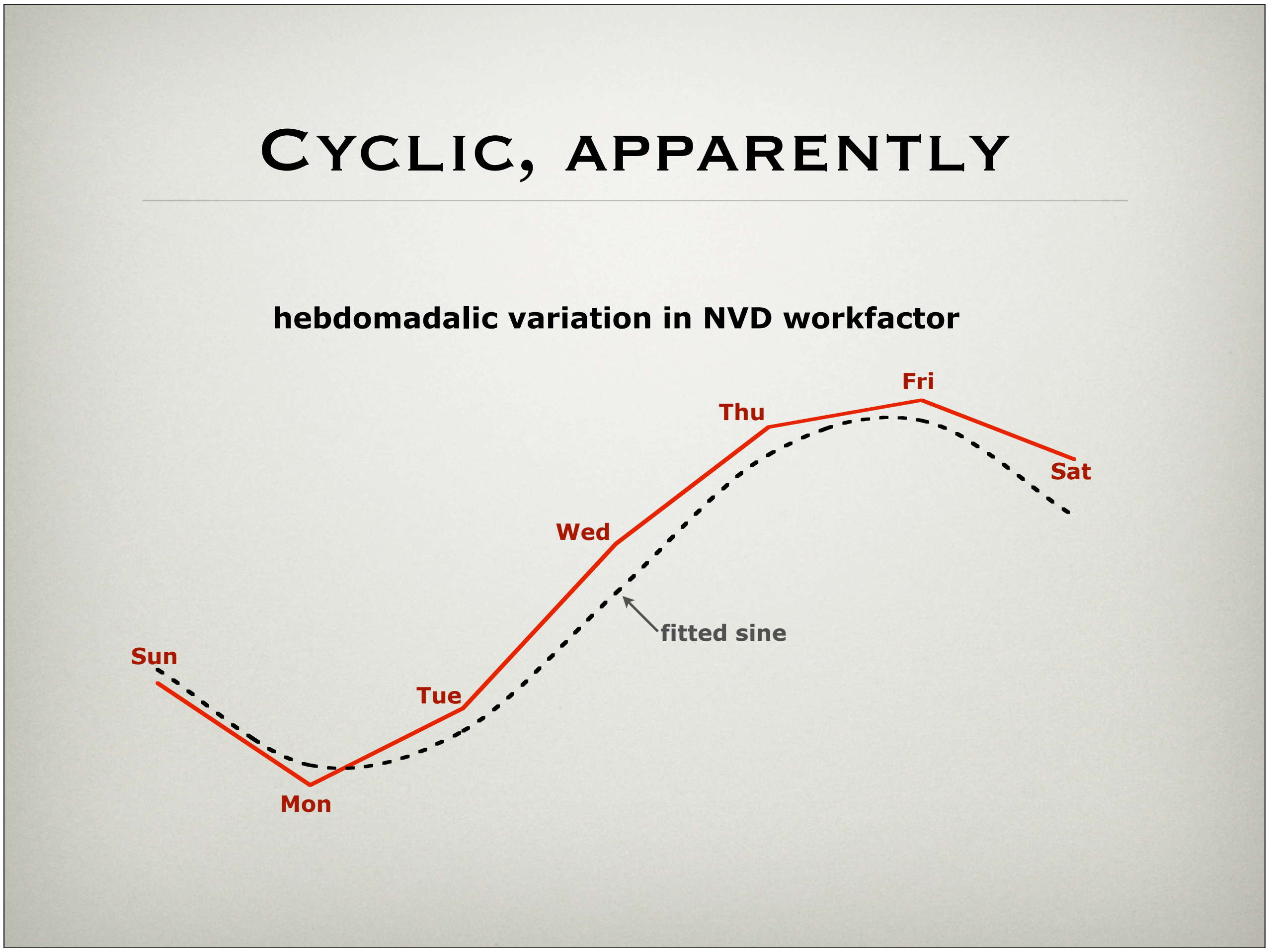


Published everyday at `http://nvd.nist.gov/`, but not otherwise charted. The workfactor number is a composite measure of vulnerabilities and their severities then outstanding.

In this chart, the dotted verticals are Microsoft patch days, the two pyramidal arrows are marking the days of max and min in this window, the blue line is the actual Workfactor Indes, and the red line is a moving 7 day average of the workfactor.

Published everyday at `http://nvd.nist.gov/`, but not otherwise charted. The workfactor number is a composite measure of vulnerabilities and their severities then outstanding.

In this chart, the dotted verticals are Microsoft patch days, the two pyramidal arrows are marking the days of max and min in this window, the blue line is the actual Workfactor Indes, and the red line is a moving 7 day average of the workfactor.



As data accumulates, the curve does reshape from time to time. At the time of this writing, the dotted black line is a fitted sine curve while the red solid line is the mean workfactor by day of the week for the past 100 days.

Fault trees

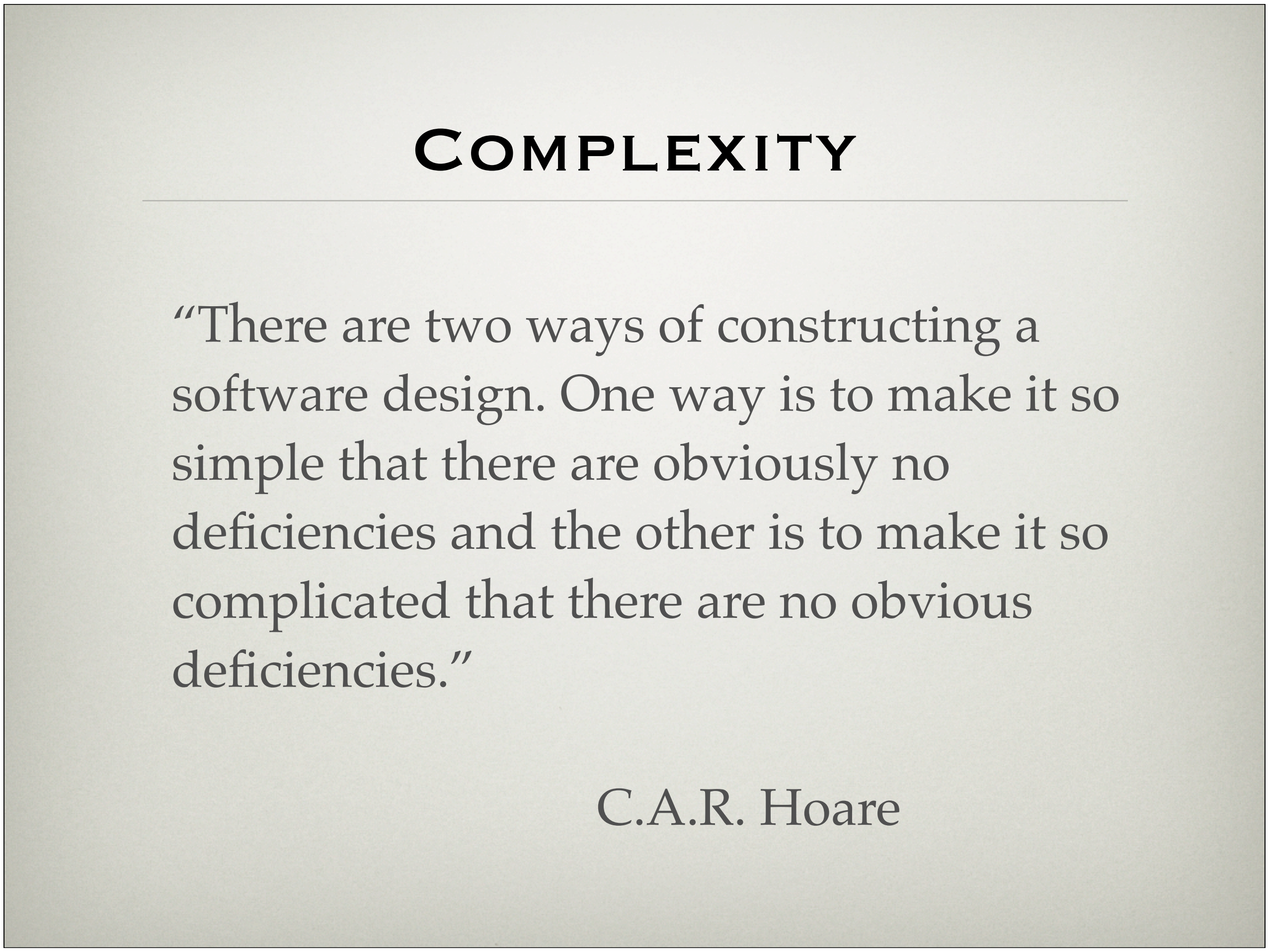
Soo Hoo

Impractical (combinatoric cost)

$$n(scenarios) = 3 \times 4 \times 4 \times 4 = 192$$

In case you have been tempted to try this, don’t bother with fault trees; they suffer from combinatoric explosion in terms of the numbers of scenarios that have to be considered if full exhaustion of the problem space is your goal; just in this picture it is 192 such scenarios. For more discussion (and on much broader scale and more relevant scope) see Soo Hoo KJ: "How Much Is Enough? A Risk-Management Approach to Computer Security," CISAC Working Paper, August 2000; as found at http://cisac.stanford.edu/publications/11900/

Complexity



This sums up the question of complexity. The parallels to current market leading suppliers, competing as they are on feature richness, is obvious and daunting. We mention complexity here as it will come again and again through the course of the day.

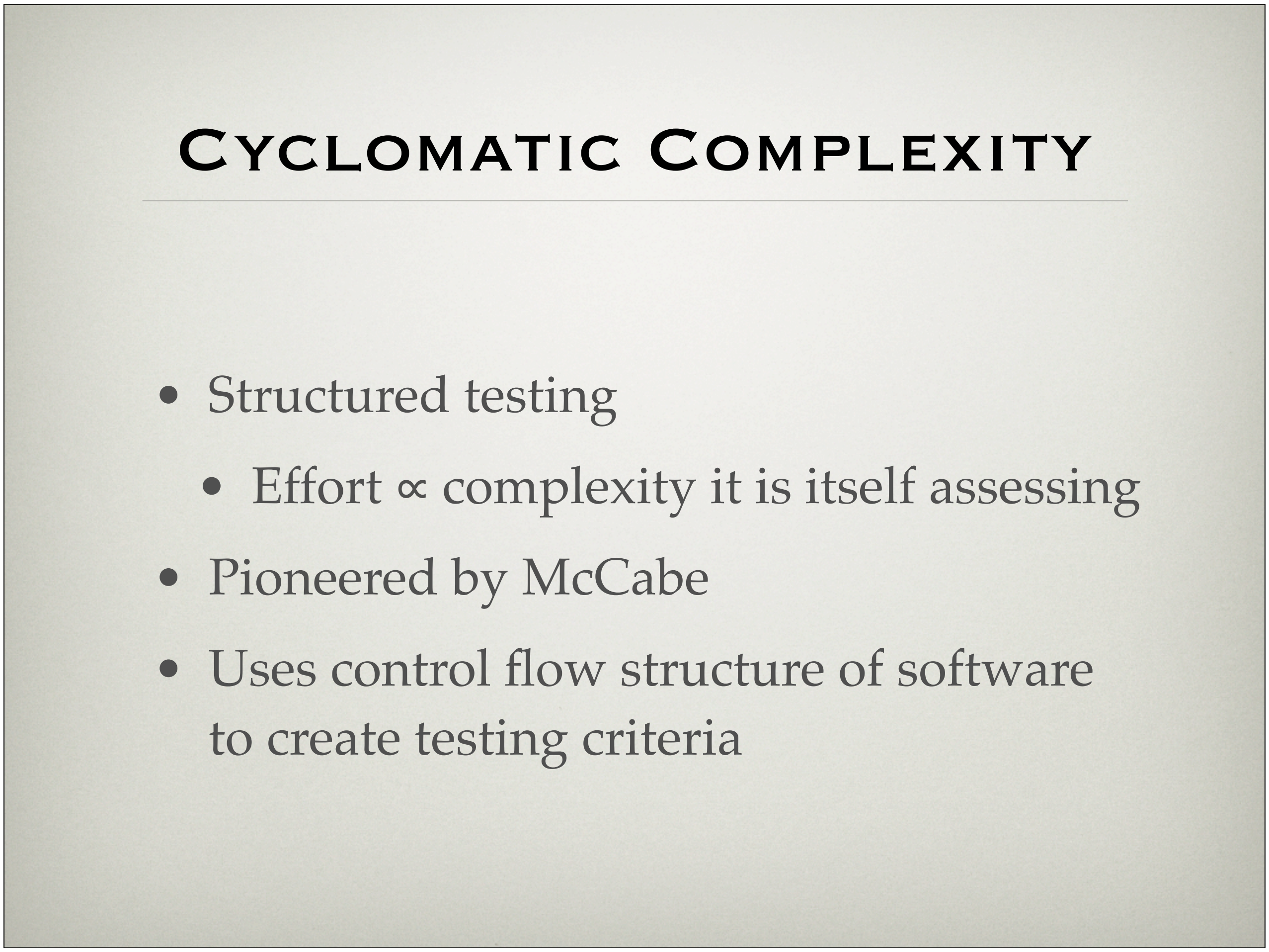
Code complexity

- Lots (lots) of measurements of code complexity out there
- If security faults ⊂ quality faults,

  And quality faults ∝ complexity,

  Then security ∝ {1-complexity}

If you assume that security faults are a subset of quality faults and are thus scattered and inadvertent, and if you accept the quality control literature that suggests complexity is the biggest contributor to quality faults, then we need to look at code complexity if we are to understand security faults in that code. See RELATED CASE STUDIES section in http://hissa.ncsl.nist.gov/HHRFdata/Artifacts/ITLdoc/235/appendix.htm#418907 in particular.

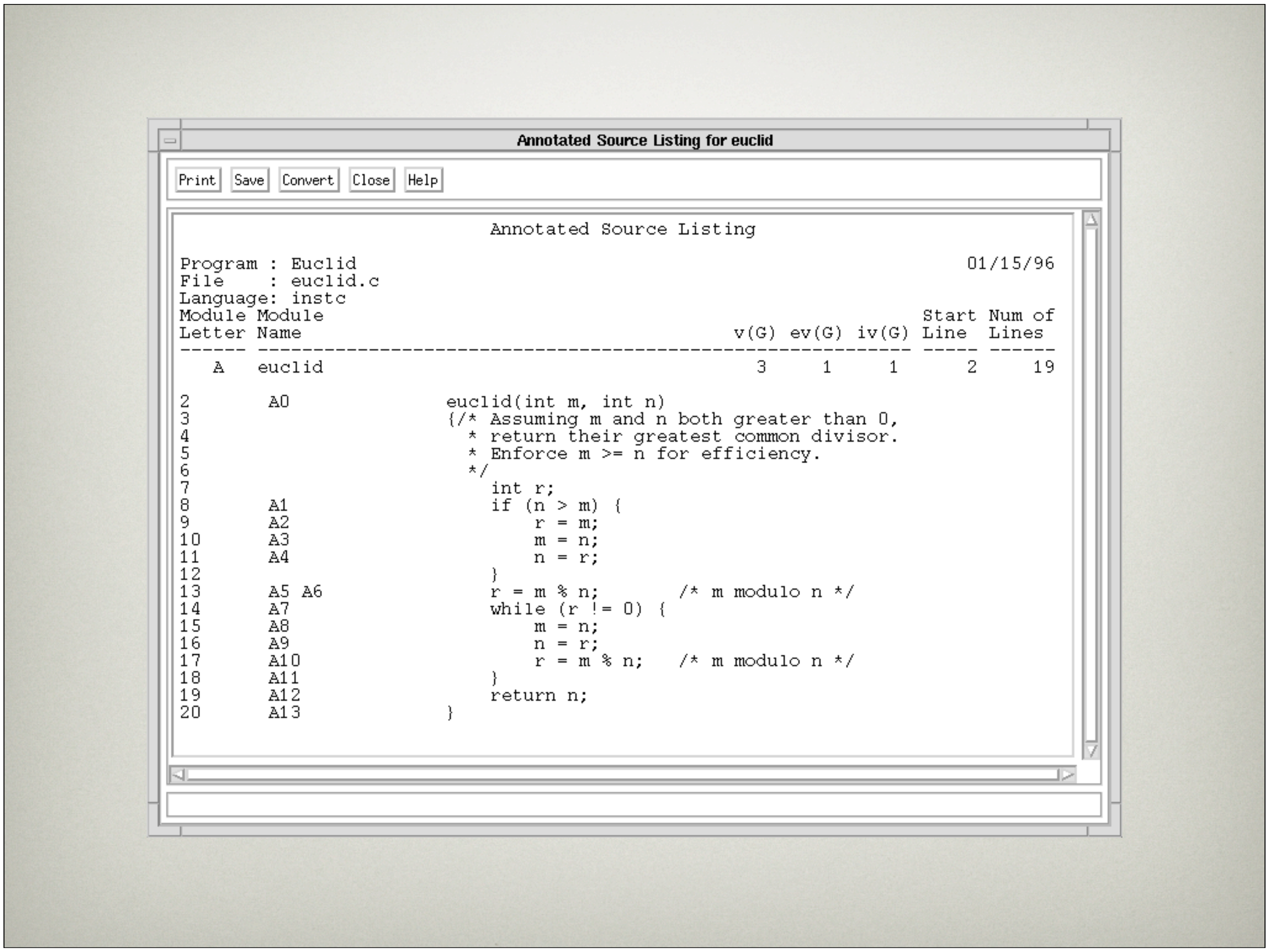


The most widely deployed measure of code complexity is McCabe's "cyclomatic" figure where the idea is to graph the control flow of a body of software and then to create testing criteria that are informed by that structure. Obviously, as that structure becomes more complex the task of testing against it becomes more daunting. In the limit, complexity above a threshold prevents testing from being efficacious thus leaving quality faults undiscovered (and thence delivered to the field).

Cyclomatic complexity

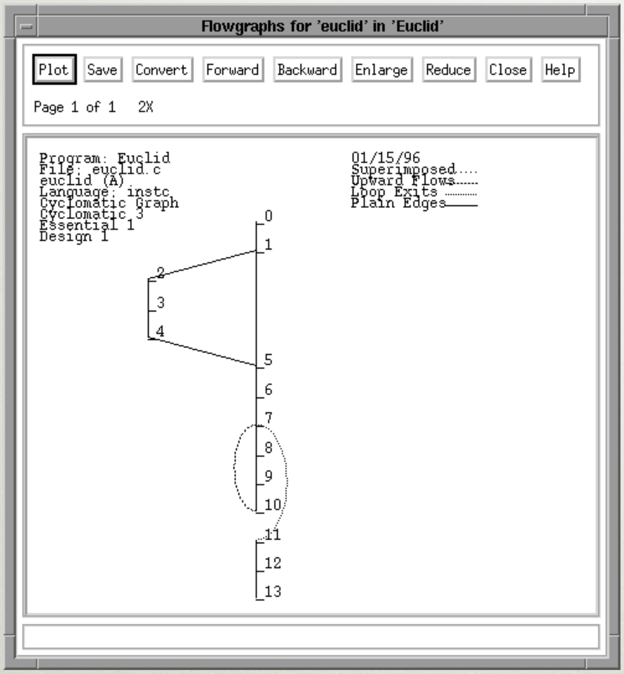
- McCabe often integrated to build env.
  - $v(G) = e - n + 2$, where e and n are the number of edges and nodes in the control flow graph
- Outside scope, except
  - Trouble when $v(G) \geq 10$

McCabe calculations are today often integrated into build environments, i.e., they are often available at zero marginal cost to developers and analysts. The definition of the score is graph-theory at work: the number of edges minus the number of nodes plus two. How to derive this is out of scope here, but as you can see more edges than nodes means more paths through the code and a rising McCabe score. Though opinions vary, a score greater than ten is thought troubling and over fifteen perhaps fatal.



```
Annotated Source Listing for euclid

Print  Save  Convert  Close  Help

                         Annotated Source Listing

Program : Euclid                                              01/15/96
File    : euclid.c
Language: instc
Module Module                                                Start Num of
Letter Name                               v(G) ev(G) iv(G) Line  Lines
------ ----------------------------------------------- ----- ------
   A   euclid                                3     1     1      2     19

2       A0              euclid(int m, int n)
3                       {/* Assuming m and n both greater than 0,
4                         * return their greatest common divisor.
5                         * Enforce m >= n for efficiency.
6                         */
7                           int r;
8       A1                  if (n > m) {
9       A2                      r = m;
10      A3                      m = n;
11      A4                      n = r;
12                          }
13      A5 A6               r = m % n;       /* m modulo n */
14      A7                  while (r != 0) {
15      A8                      m = n;
16      A9                      n = r;
17      A10                     r = m % n;   /* m modulo n */
18      A11                 }
19      A12                 return n;
20      A13             }
```

For the programmers in the audience, a simple set of three illustrations. Here, some simple source.

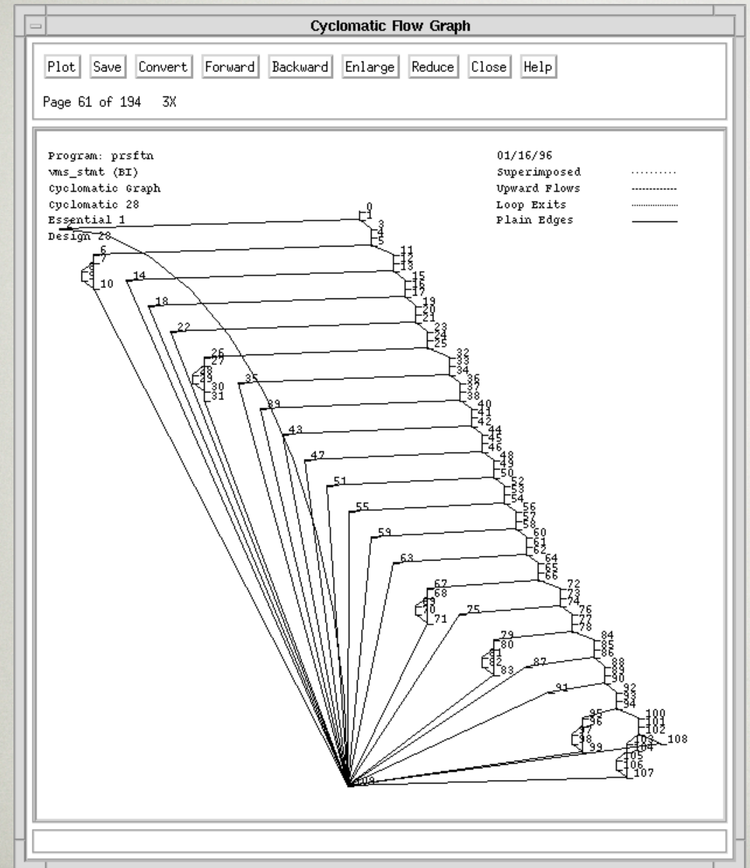


$$e = 15$$

$$n = 14$$

$$v(G) = 3 \leq 10$$

...some simple source with a McCabe score of 3, which is much less than 10.



$$e = 138$$
$$n = 108$$
$$v(G) = 28 > 10$$

A less than simple picture now with a score of 28 which rather higher than 10 and quite likely untestable.

A warning

- Limits on sizes of modules? No
- Limits on complexity thereof? Yes
- Aftermarket tools to assess complexity of <u>binaries</u> are appearing.

Note that most patches increase v(G)

So what should the security metrics person do? Limit sizes of modules? No. Limit complexity of modules? Yes. Note that there are aftermarket tools now appearing for this even if all you have is a binary and not source. Once this is possible, the default good practice shifts to “Why aren’t you looking at this?” and that shift-point is more or less now.

Note that as most patches involve at least one extra node and two extra edges that most patches increase complexity scores. What a surprise.

Use in security

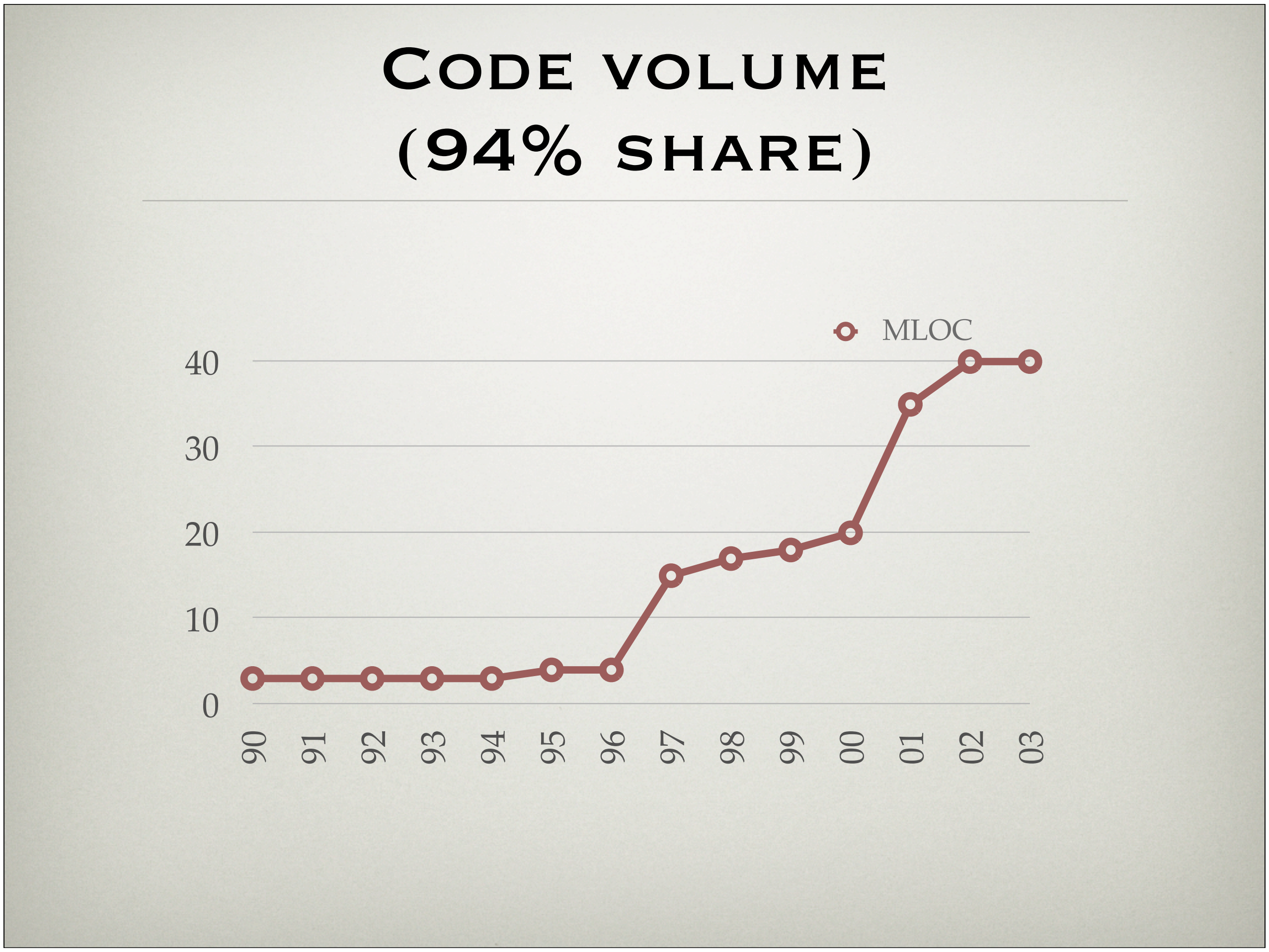
- Hotspotting – look at your outliers
- Trending – recent check-ins different?
- Verify assumptions – c.f. code coverage

How might we use McCabe or other complexity metrics? As stolen directly from http://www.enterpriseintegrationpatterns.com/ramblings/41_metrics.html the anser is to do one or more of (1) looking at your outliers, (2) looking at recent versus historical trends, and/or (3) verify that you are or are not getting testing that actually can be said to have enough coverage to predict field experience before field deployment. The higher your requirement for reliability the higher the need for this strategy.

Naive example

- Naive in that...
  - it uses the only data we have, code volume, and
  - estimates complexity as square of code volume (a venerable metric)

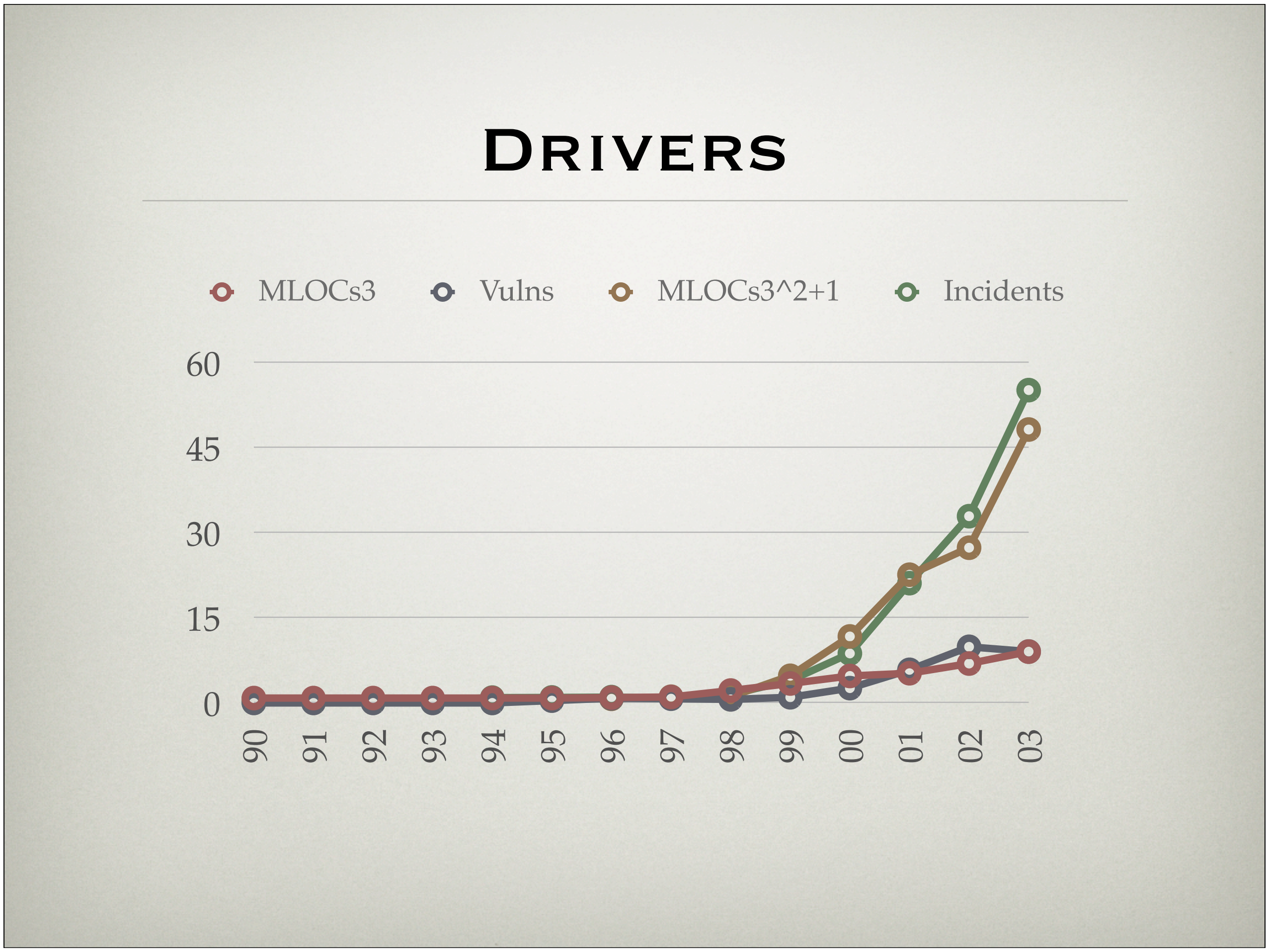
A naive example might be as follows, noting that this is unproven (even if once you’ve seen it you would tend, as does the present author, to suggest that the burden of proof has shifted to those who say that the following isn’t so).



Windows 94% market share per IDC

Code volume as observed:

Win 3.1	Win NT	Win 95	NT 4.0	Win 98	NT 5.0	Win 2K	Win XP
3	4	15	17	18	20	35	40
1990	1995	1997	1998	1999	2000	2001	2002



Each curve is normalized against its own median over this period.

Code volume curve, MLOCs3, is the three year moving average of code volume, perhaps a better estimator of effective code volume in the population at large.

The second code volume curve, MLOCs3^2+1, is the square of the three year moving average of code volume, and then shifted right one year. The argument is this: Security faults are a subset of quality faults and the literature says that quality faults will tend to be a function of code complexity, itself proportional to the square of code volume. As such, the average complexity in the field should be a predictor of the attack-ability in an a priori sense. Shifting it right one year is to permit the attack community time to acquire access and skill to that growing code base complexity. This is not a statement of proven causality -- it is exploratory data analysis.

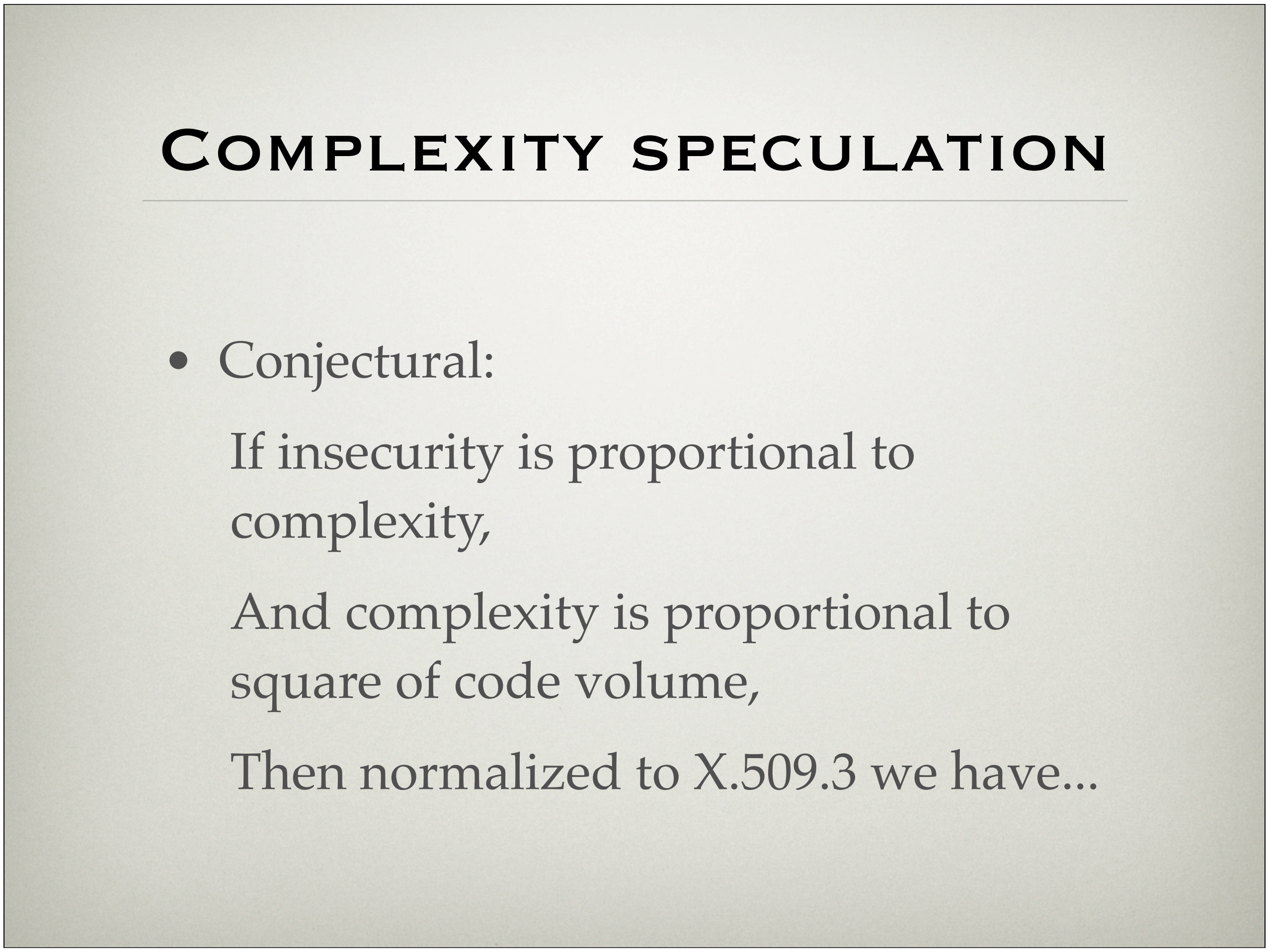
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
S3	.86	.86	.86	.86	.86	.86	.95	1.05	2.19	3.43	4.76	5.24	6.95	9.05
V	0	0	0	0	0	0.41	0.83	0.75	0.63	1.00	2.61	5.84	9.90	9.07
^2	0.73	0.73	0.73	0.73	0.73	0.73	0.73	1.10	1.10	4.79	11.73	22.62	27.38	48.23
I	0.1	0.16	0.31	0.54	0.94	0.97	1.03	0.86	1.50	3.96	8.73	21.13	32.94	55.18

Complexity speculation

- Factual:

X.509r1	20 lines of ASN.1
X.509r3	600 lines of ASN.1
SET	3000 lines of ASN.1

A different speculation. X.509 is the ISO standard for public key infrastructure (PKI) certificate structure. The above is factual.



We might have an hypothesis that security is, as we said earlier, proportional to complexity and, again as we said before, that complexity is proportional to the square of code volume. In that case, and normalizing to X.509c3, we'd have...

Complexity speculation

- Then normalized to X.509.3 we have...

X.509r1	.001 units of complexity
X.509r3	1 unit of complexity
SET	25 units of complexity

...normalizing to X.509c3, we'd have X.509r3 at the normalization point of 1 unit of complexity with X.509r1 at three orders of magnitude below and SET at one and a half orders of magnitude above. That's a fair dynamic range, to say the least.

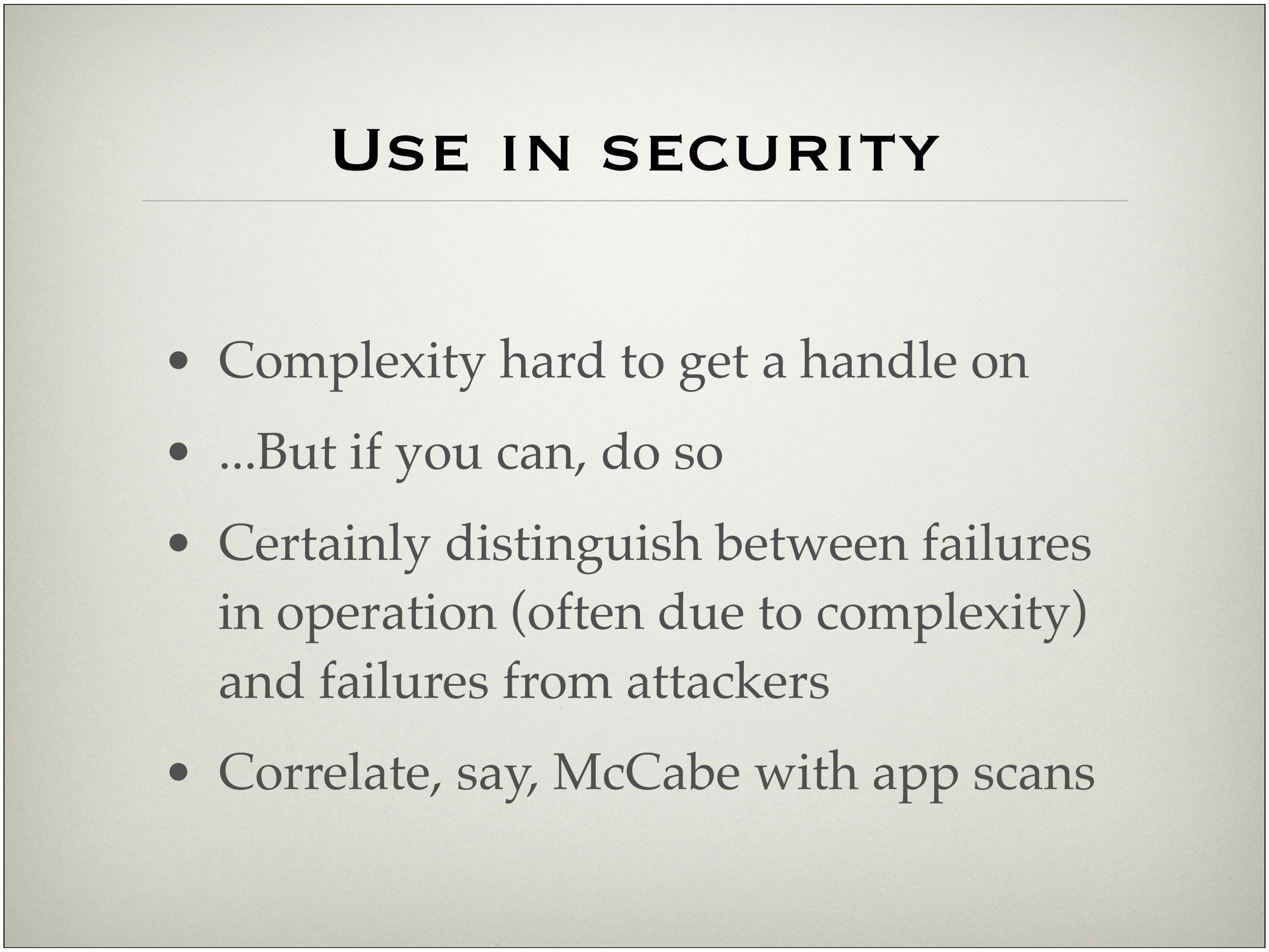
[Secure Electronic Transactions (SET) was a 1996 standard jointly by Mastercard and VISA; see http://www.echeck.org/overview/comparison/set.html]

Complexity speculation

- Back to factual:

X.509r1	insufficient expressiveness thus not much in use
X.509r3	widely used, but too expressive for assured interop
SET	couldn't be cost-effectively implemented thus failed

X.509r1 was too small, SET too large, and therefore by exhaustion X.509r3 must be just right (apologies to Goldilocks). The present author thinks that this is a textbook case of cost in complexity terms putting an upper bound on what could be cost effectively implemented, not that PKI with client-side certificates has exactly taken over Internet-based commerce.



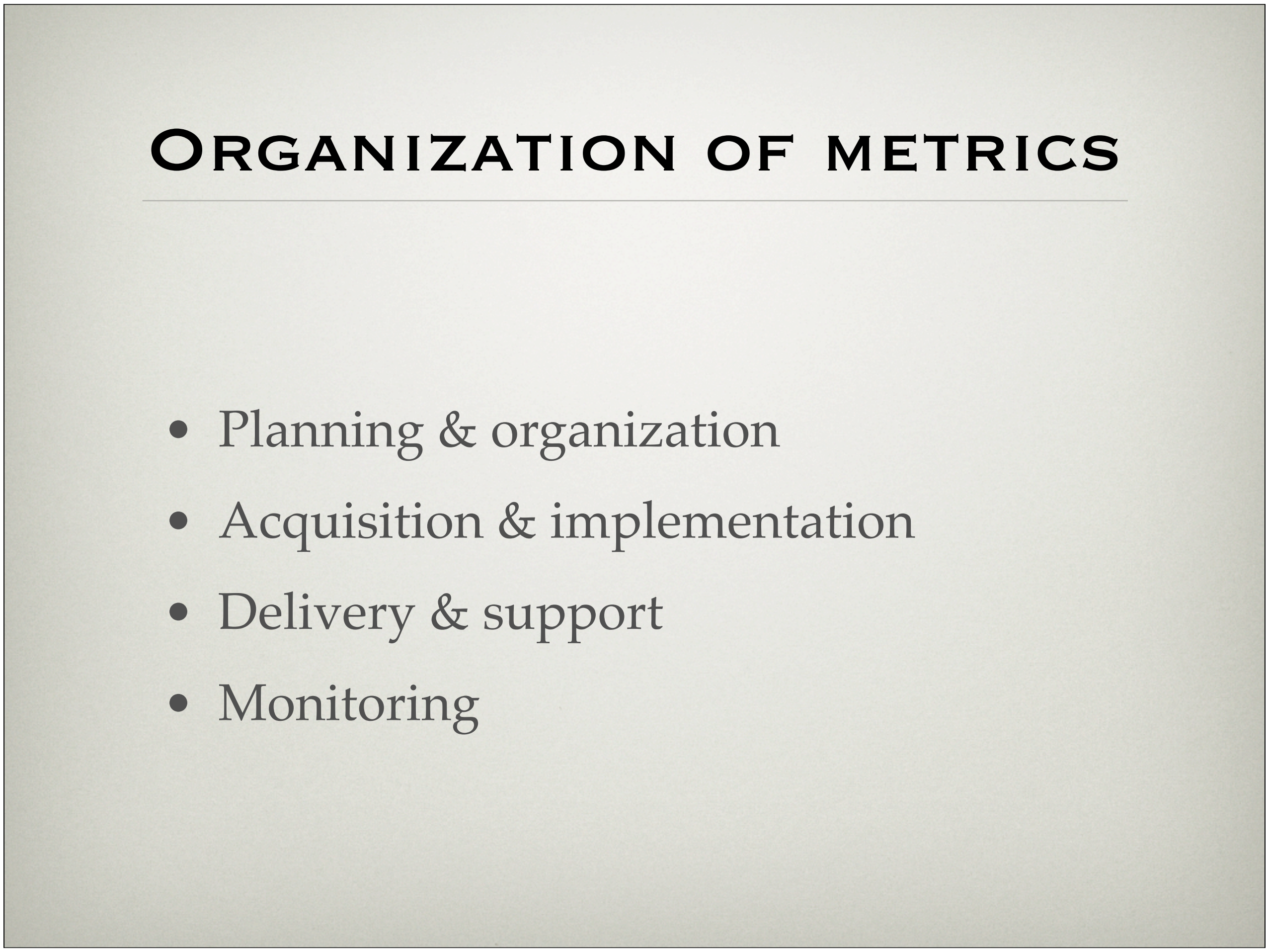
In security, if you can get a handle on it, is absolutely vital in any program of security metrics. At the very least, see if you can find predictive correlations such as McCabe scores and application scanning against the built products those code bases deliver. This requires further discussion than can be done here, sadly.

Metrics Programs

When picking metrics...

- Metric name
- Metric description
- Metric purpose/objective
- Required data sources
- Required logic, algorithms, or formulae
- Frequency of measurement
- Units of measure
- Benchmark or goal
- Visualization
- Publication schedule

Current consensus on the discuss@securitymetrics.org list of how to define a lasting metric.



For a really useful read on aligning COBIT, ISO, ITIL, and NIST, see

Hodgkiss G, Guldentops E, et al., Office of Government Commerce, United Kingdom, "Aligning COBIT, ITIL and ISO 17799 for Business Benefit: Management Summary," 2005, as found at http://www.itgovernance.co.uk/files/ITIL-COBiT-ISO17799JointFramework.pdf

Planning & organization

Assess & manage IT risks

Manage IT human resources

Manage IT investment

Jaquith, op cit., p.100

Planning & organization

Metrics to assess & manage IT risks:

% of critical assets...
...on compliant servers
...reviewed for physical risks
...with cost of compromise estimated
...with documented risk assessment
...with documented risk mitigation plans

Jaquith, op cit., p.100

Planning & organization

Metrics to manage IT human resources:

- % of perf reviews with eval of responsibilities & compliance
- % of position descriptions with clarity
- % of trusted users with background checks

Jaquith, op cit., p.100

Planning & organization

Metrics to manage IT investment:

- Budget allocation for security (operational, new programs, discretionary)

Jaquith, op cit., p.100

Acquisition & implementation

Solution identification

Installation & accreditation

Jaquith, op cit., p.104

Solution identification

Metrics to identify:

% coverage of confidentiality controls
% coverage of integrity controls
# consultations between sec. & dev.
# customer consulatations w/sec.
# sec. team consultations by B.U.s
% new systems with initial sec. consults

Jaquith, op cit., p.105

Installation & accreditation

Metrics to identify:
- % systems with certification
- % systems with risks accepted (sign-off)
- % systems with security costs built-in

Jaquith, op cit., p.108

Delivery & support

Educate & train users

Ensure system security

Identify & allocate costs

Manage data

Manage third-party servers

Jaquith, op cit., p.113

Educate & train users

Metrics to identify:

- # security skills mastered
- % new employees awareness trained
- % security staff with certification
- Fulfillment rate of retraining/external
- Objective raining effectiveness

Jaquith, op cit., p.113

Ensure system security

Metrics to identify:

- # active user IDs assigned to one person
- % users with sysadmin rights
- % assets with role-based assignments
- % systems with segregation of duties
- Cycle time to de-provision users by type

Jaquith, op cit., p.113

Identify & allocate costs

Metrics to identify:

- Cost of security for revenue-generation
- % security costs charged back to B.U.s
- Estimated $ cost from all incidents
- % incidents with no measurable costs

Jaquith, op cit., p.113

Manage data

Metrics to identify:

- Data flow numbers
- Toxicity rate in customer data
- % backup media stored offsite
- % media sanitized prior to disposal
- # data privacy escalations, with costs

Jaquith, op cit., p.113

Manage third-party servers

Metrics to identify:

- Cycle time to grant access
- % third-party applicantes vetted
- # unauthorized transactions, by app
- % 3rd-party agreements with security
- % agreements with external validation

Jaquith, op cit., p.113

Monitoring

Monitor process

Monitor internal controls

Ensure compliance

Jaquith, op cit., p.126

Monitor process

Metrics to identify:

% systems with monitored logs

% external-facing systems with logs

% watched for configuration integrity

Jaquith, op cit., p.126

Monitor internal controls

Metrics to identify:

- % systems reviewed for compliance
- % 3rd parties reviewed for compliance
- % controls working as designed
- % systems with any serious deficiency
- Per system cost of assurance

Jaquith, op cit., p.126

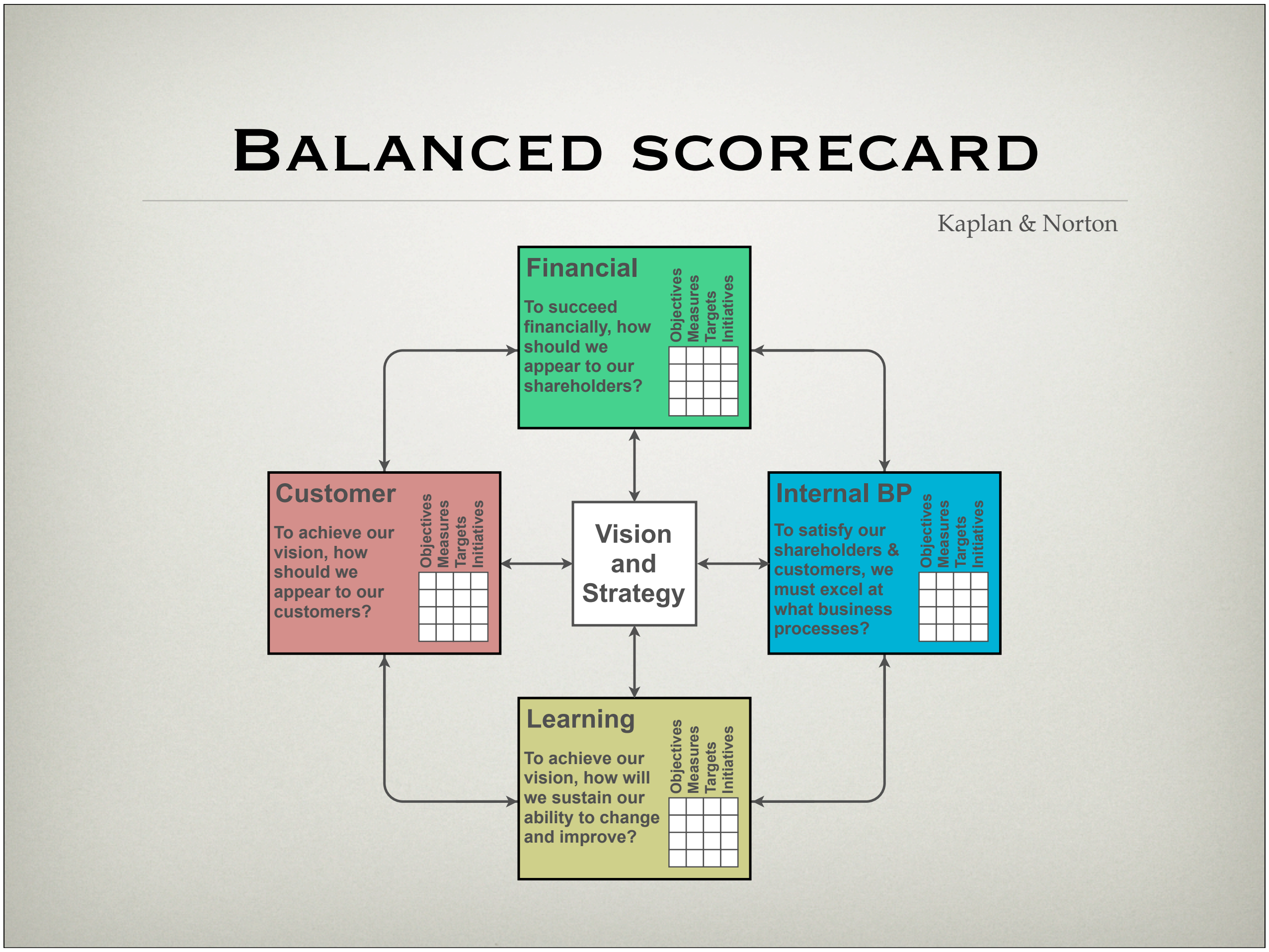
Ensure compliance

Metrics to identify:

- # audits sucessfully completed
- # pending items with cost-to-complete
- # pending customer-related items
- % key requirements externally audited
- Cost of remediations

Jaquith, op cit., p.126

Keeping score



A balanced scorecard is a now-standard part of business planning albeit not for security. It comes from Kaplan RS & Norton DP : "The Balanced Scorecard: Measures That Drive Performance," Harvard Business Review, January-February 1992, and, with a redrawn picture from http://www.balancedscorecard.org/ looks like the above.

So how to?

- What are the security versions of the four corners of a balanced scorecard:
  - Financial *v* Security
  - Internal business processes *v* Security
  - Learning and growth *v* Security
  - Customer *v* Security

What does the present author actually want? A balanced scorecard built with security in mind. And what is a balanced scorecard?

Financial view

- Ensure revenue-generation proceeds
- Preserve integrity of needed records
- Lower risk to the revenue generators
- Security as contributor to reliability

Jaquith, op cit., p270, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

Example metrics

- Cost of security per transaction
- DoS and other attack downtimes
- Data flow per transaction & per source
- Budget correlation with risk measures
- Comparison with like firms

Jaquith, op cit., p270, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

Business process view

- Minimizing unknown unknowns
- Protection of information assets
- Least-privilege and need-to-know
- Verification & accreditation

Jaquith, op cit., p281, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

Example metrics

- %age of critical systems under DR plan
- %age of systems obeying ______ policy
- MTBF & MTTR for security incidents
- Number of security team consultations
- Latency to obey ______ change orders

Jaquith, op cit., p281, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

Learning & growth view

- Awareness and culture generally
- Certification and/or training
- Levels of collaboration, esp. early
- Engagement with outside peers/groups

Jaquith, op cit., p287, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

Example metrics

- %age of job reviews involving security
- %age of security workers with training
- Ratio of b.u. security staff to central staff
- New system timely security consultations
- %age of programs with budgeted security

Jaquith, op cit., p287, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

Customer view

- Decrease security backscatter
- Increase options without increasing risk
- Compliance and the proof thereof
- Security as contributor to reliability

Jaquith, op cit., p273, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

Example metrics

- %age of SLAs with security standards
- %age of tested external-facing apps
- Number of non-employees with access
- %age of data secure-by-default
- %age of customer data outside data center

Jaquith, op cit., p273, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

How to get this to fly

- Cascading scorecards build bridges
- Acceptance *v.* Accountability
- Field test everything
- Show cause and effect
- Counter the idea that security is reactive

Jaquith, op cit., p293, the discuss@securitymetrics.org mailing list, and experience in/with/around banks

There's a lot more...

Maximize metric leverage

Use ratios to heighten contrasts

e.g., an outcome measure per unit of process cost

Ratio of this to that can be more explanatory, and is also consistent with situations where the underlying scale and accuracy of measurement is suspect.

Measures to pursue

Odds & odds ratio: estimate effect size

$$\mathrm{OR}(x,y) = \frac{\mathrm{Odds}(x)}{\mathrm{Odds}(y)} = \frac{\Pr(x)/(\Pr(\bar{x})}{\Pr(y)/\Pr(\bar{y})}$$

Relative risk: esp. for unlikely outcomes

$$\mathrm{RR}(x,y) = \frac{\Pr(x)}{\Pr(y)}$$

These are measures that will soon see use in security modeling. The Odds Ratio has useful properties, and the Relative Risk is trivial to calculate and seems intuitive to most.

AND IN EXAMPLE:

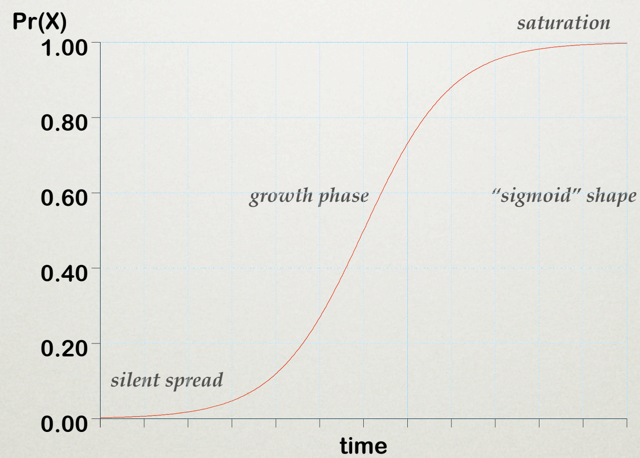
host is	managed	unmanaged
vulnerable	3	7
clean	6	4

$$\mathrm{OR}(vulnerable|unmanged) = \frac{7/4}{3/6} = 3.5$$

This illustrates the Odds Ratio (OR) with artificial but not all that unlikely data.

In this case, the odds of being vulnerable when unmanaged are 7-to-4 (or 1.75-to-1) while the odds of being vulnerable when managed are 3-to-6 (or 2-to-1 against). The ratio of the two then tells you that being unmanaged increases your odds of being vulnerable by a factor of 3.5, which is a clear way to express the value (and has statistical support wrapped around it that will, in time, become commonplace).

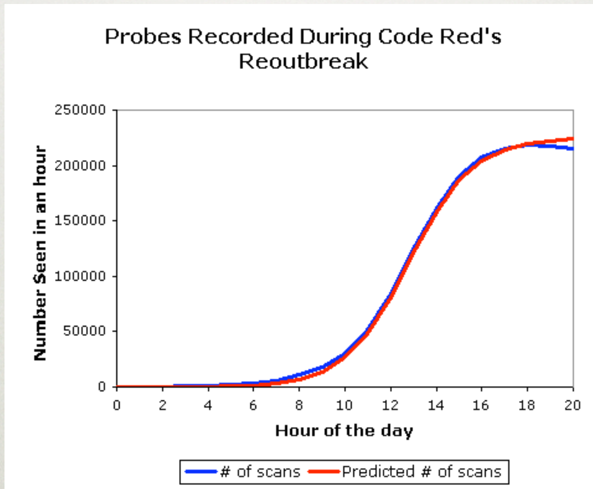
Sigmoid curves...



Sigmoid curves show the change in odds ratio over time when there is a saturation effect, as seen in the next slide.

..ARE DIRECTLY APPLICABLE

CAIDA



And this is exactly how you might use it to look at spread rates of a worm.

This graphic was taken directly as is from http://www.caida.org/publications/papers/2003/sapphire/sapphire.html

Logistic regression

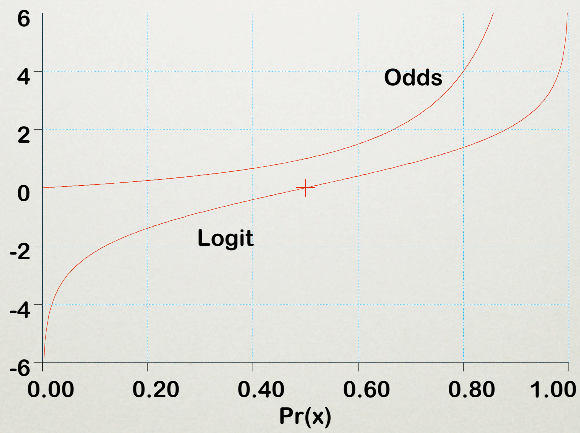
$$\text{logit}(x) = log(\text{Odds}(x))$$
$$= log(\text{Pr}(x)) - log(\text{Pr}(\bar{x}))$$

modelled as

$$= \sum_{i=1}^{n} \beta_i y_i$$

The logit (or “log-odds”) has very useful properties, especially in logistic regression where log(Odds(x))=log(Pr(x))-log(Pr(1-x)) and thus you can get a general linear model where, frankly, log(Pr(1-x)) does not matter. Suitable for most anything with a binary outcome. Widely used in clinical trials, for example.

Logit (logistic) curve



Logistic curves shows the change in odds over the range of probabilities. Obviously, it is linear in the middle but goes to ±infinity at the margins. The point in taking the log of the odds is to make the curve symmetric and, for the central portion where decisions are harder to make, to induce linearity.

Meta-analysis

Combining several measures of the same thing to get a better handle

e.g., different vulnerability scores across different departments to get an enterprise-wide value

$$confidence = \frac{signal}{noise} \times \sqrt{sample\ size}$$

Meta-analysis, in the statistical literature, generally means to combine measurements of the same topic done by multiple researchers so as to arrive at a composite, but well-supported, central truth. Stretching that idea just a little, we want metrics that combine measurements that each contribute something to the understanding of some slippery topic. See, for example, http://www.pitt.edu/~super1/lecture/lec1171/index.htm

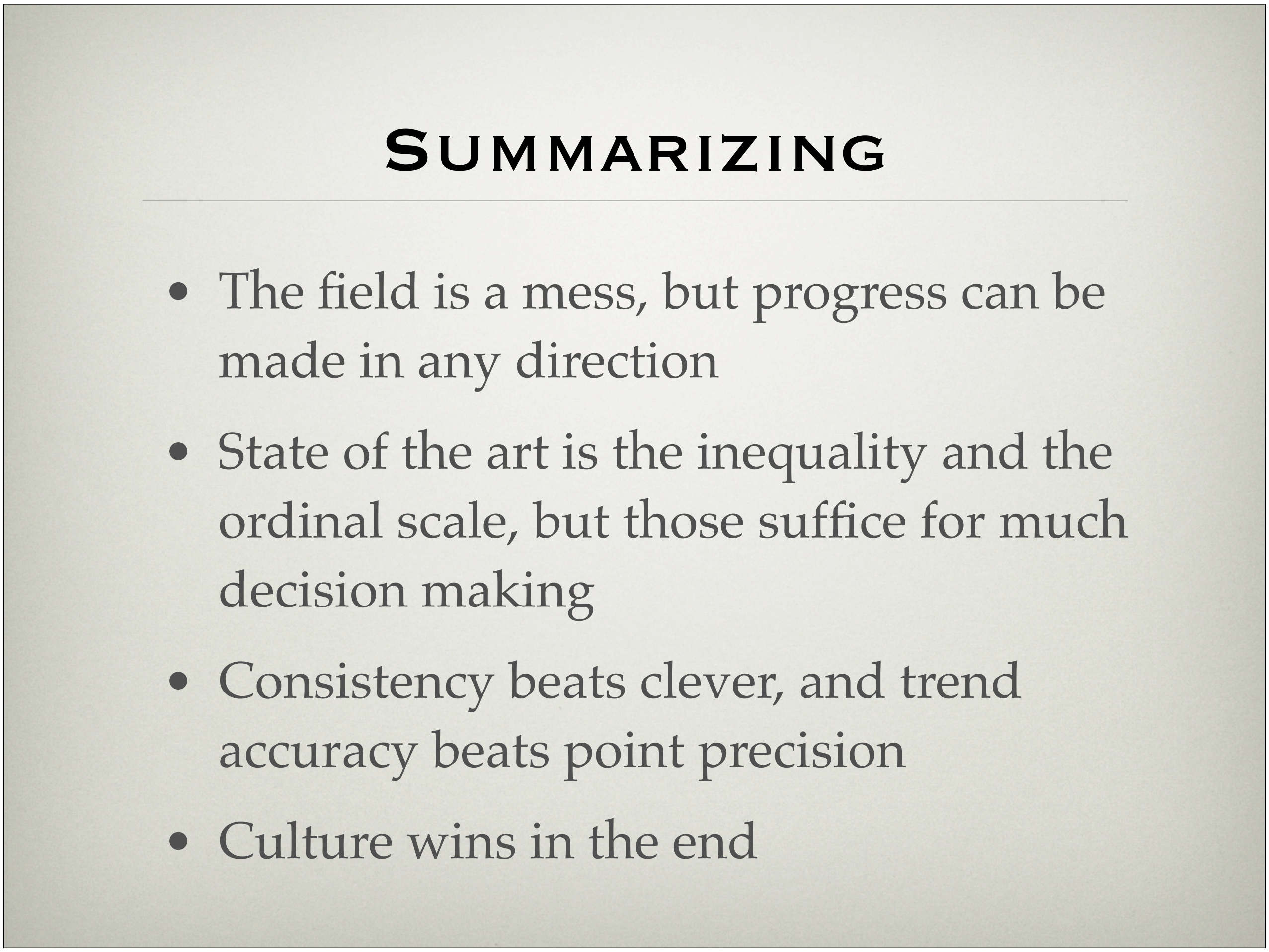
Survival analysis

- Many kinds, all are about predicting end-point failures
- In people, death is unambiguous but in machinery, ambiguity is common
- Nevertheless, like to be an important part of security metrics going forward

The “names” to remember are Weibull, the Cox proportional hazards model and Cox regression, the Kaplan–Meier estimator, and on and on.

The side references to the Wikipedia entries on “hazard ratio” and “survival analysis” are as good a place for the novice to begin as any.

In other words,
we've only just begun

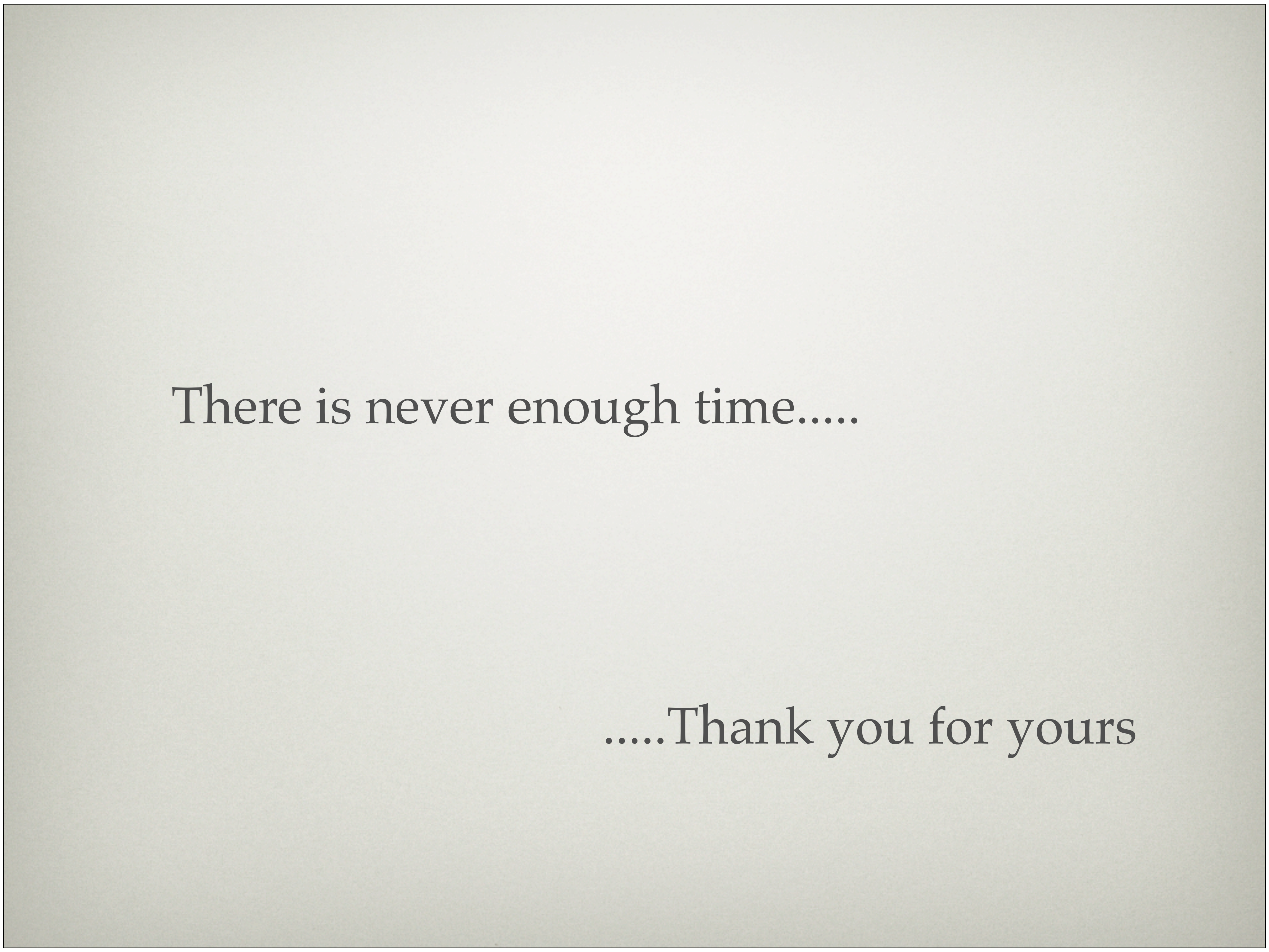


Summarizing is virtually impossible, but simply put at our present state of knowledge, ordinal information is both good enough and almost all we can do. Clever we have not got time for and the clever will be busy shooting themselves in their own feet in any case. Progress can be made if you are not afraid to try.

Go forth and

- Instrument your firm/unit
- Derive metrics from instrument output
- Map metrics to values
- Make decisions

More or less stolen from Betsy Nichols, CTO of ClearPoint Metrics.



It has been entirely my pleasure. Contact is welcome but reply is not instant. Slides are yours to use though I would appreciate acknowledgement if it is possible to do so.

Daniel E. Geer, Jr., Sc.D.
Geer Risk Services
P.O. Box 390244
Cambridge, Mass. 02139
U.S.A.
+1.617.492.6814
dan@geer.org