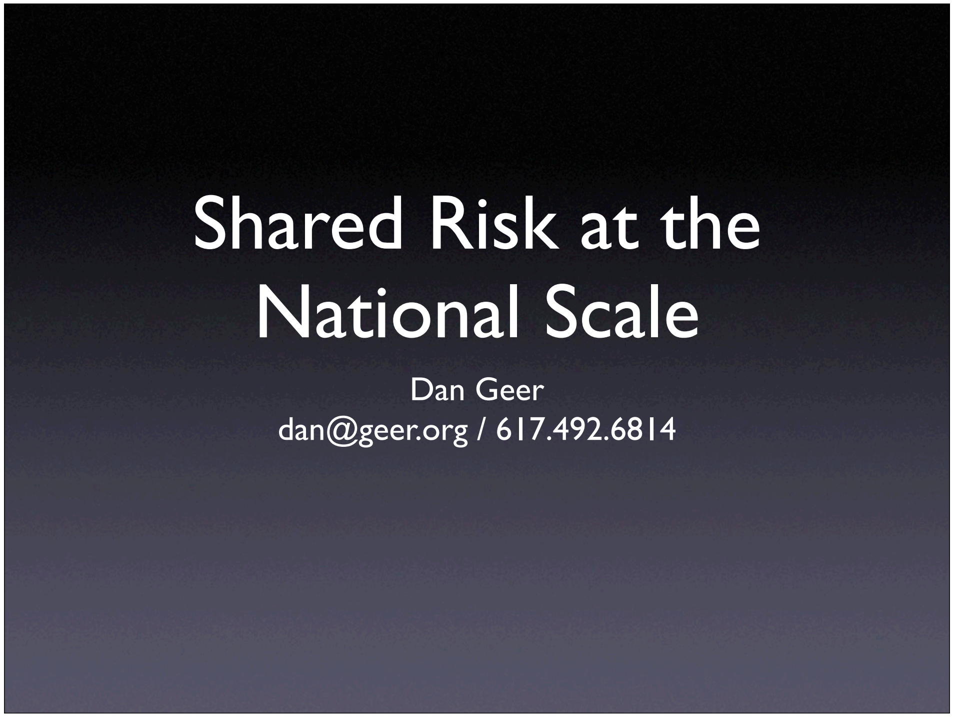


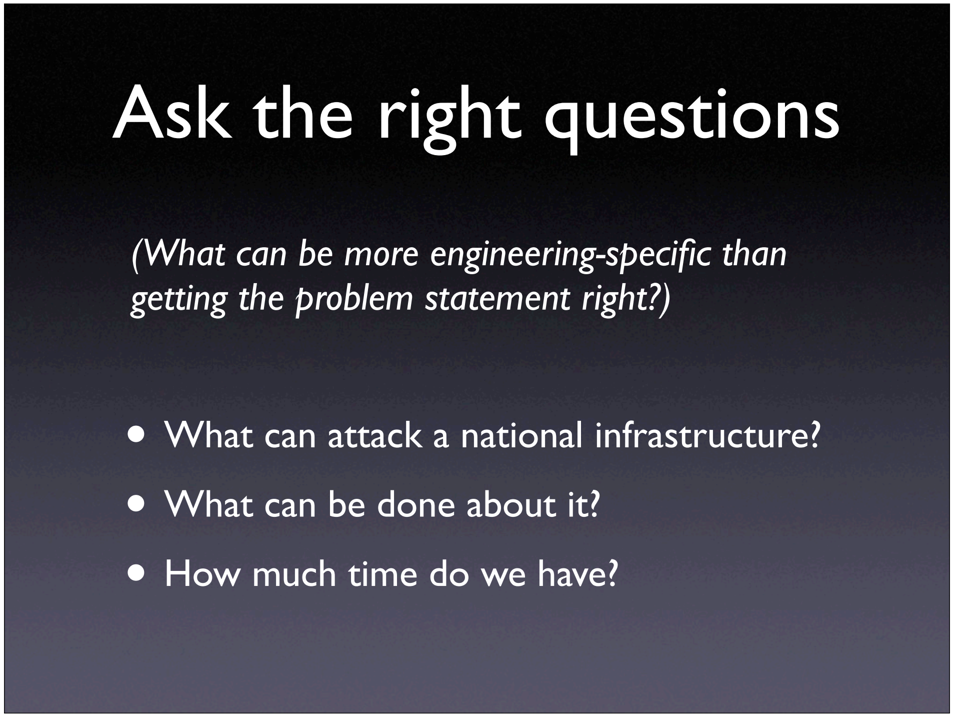
Shared Risk at the National Scale

Dan Geer
dan@geer.org / 617.492.6814

Daniel E. Geer, Jr., Sc.D.

Principal, Geer Risk Services, LLC
P.O. Box 390244
Cambridge, Mass. 02139
Telephone: +1 617 492 6814
Facsimile: +1 617 491 6464
Email: dan@geer.org

VP/Chief Scientist, Verdasys, Inc.
950 Winter St., Suite 2600
Waltham, Mass. 02451
Direct-in: +1 781 902 5629
Corporate: +1 781 788 8180
Facsimile: +1 781 788 8188
Email: geer@verdasys.com

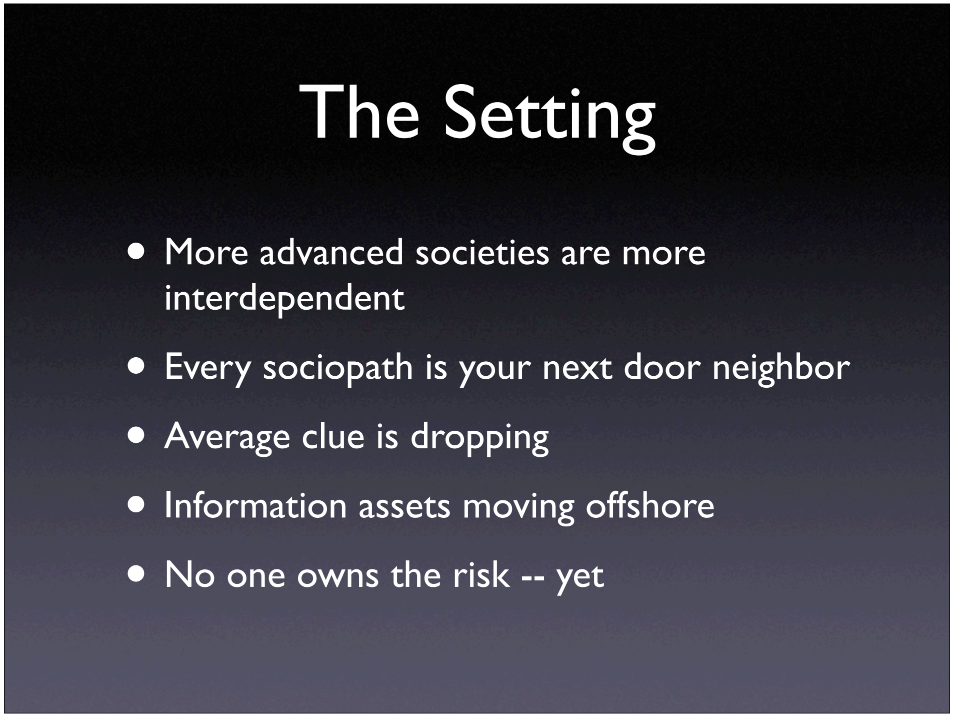


Ask the right questions

(What can be more engineering-specific than getting the problem statement right?)

- What can attack a national infrastructure?
- What can be done about it?
- How much time do we have?

In all of engineering, getting the problem statement right is job 1. Without the right problem statement you get “we solved the wrong problem” or “this is a solution is search of a problem” or worse.



The Setting

- More advanced societies are more interdependent
- Every sociopath is your next door neighbor
- Average clue is dropping
- Information assets moving offshore
- No one owns the risk -- yet

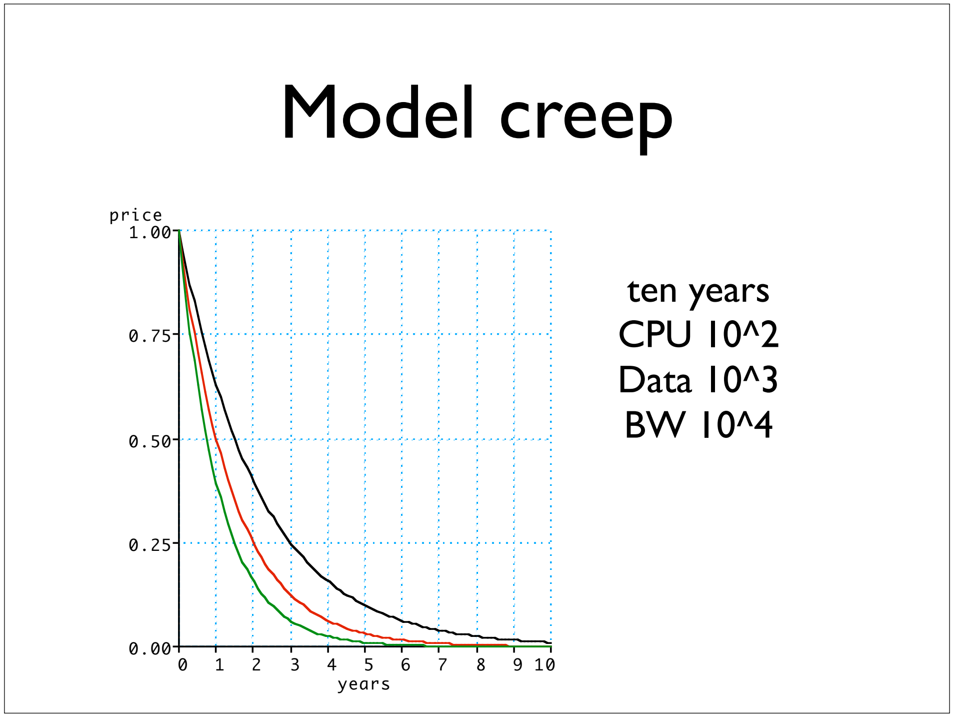
The more advanced the society the more interdependent it is. Which is the cause and which is the effect is a debate for sociology or economics, but it is a tight correlation.

Equidistance and near zero latency is what distinguishes the Internet from the physical world.

Power doubles every 12-18 months and, obviously, skill on the part of the user base does not. Hence the ratio of skill to power falls. This has broad implications.

Information does not want to be free, but it does want to be located to its advantage.

In finance, risk taking and reward are tightly correlated and there is zero ambiguity over who owns what risk; cf. that in the digital security sphere where there is nothing but ambiguity over who owns what risk.

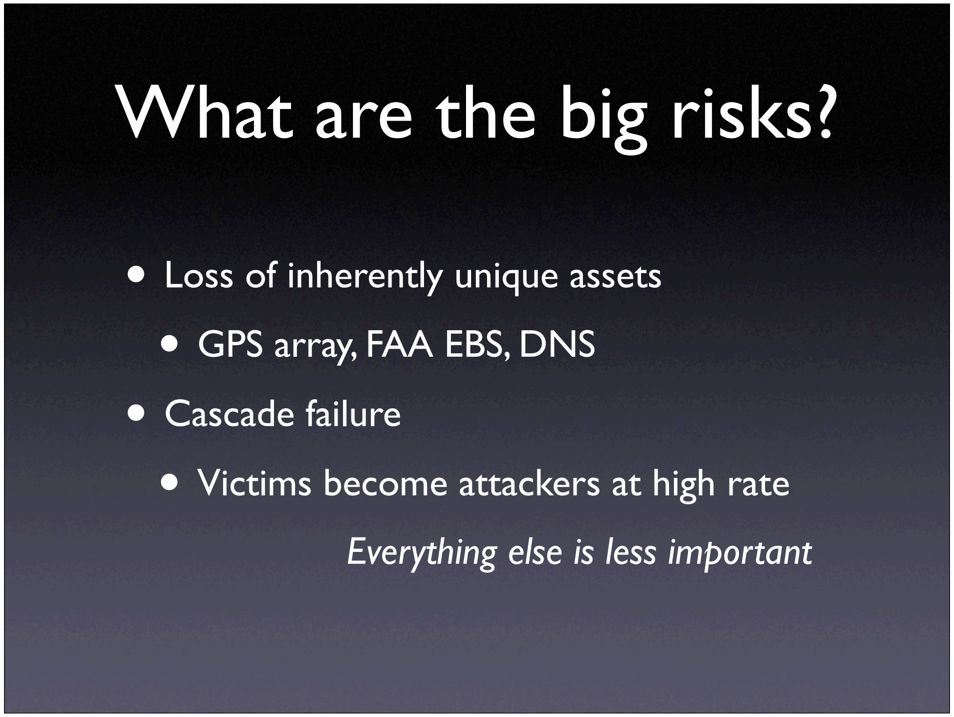


Model creep

ten years
CPU 10^2
Data 10^3
BW 10^4

Black line is “Moore’s Law” whereby $/MHz drops by half every 18 months. It’s unnamed twins 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 charactgeristic of the then optimal computing plant.

[forgive background color reversal needed for clarity]



What are the big risks?

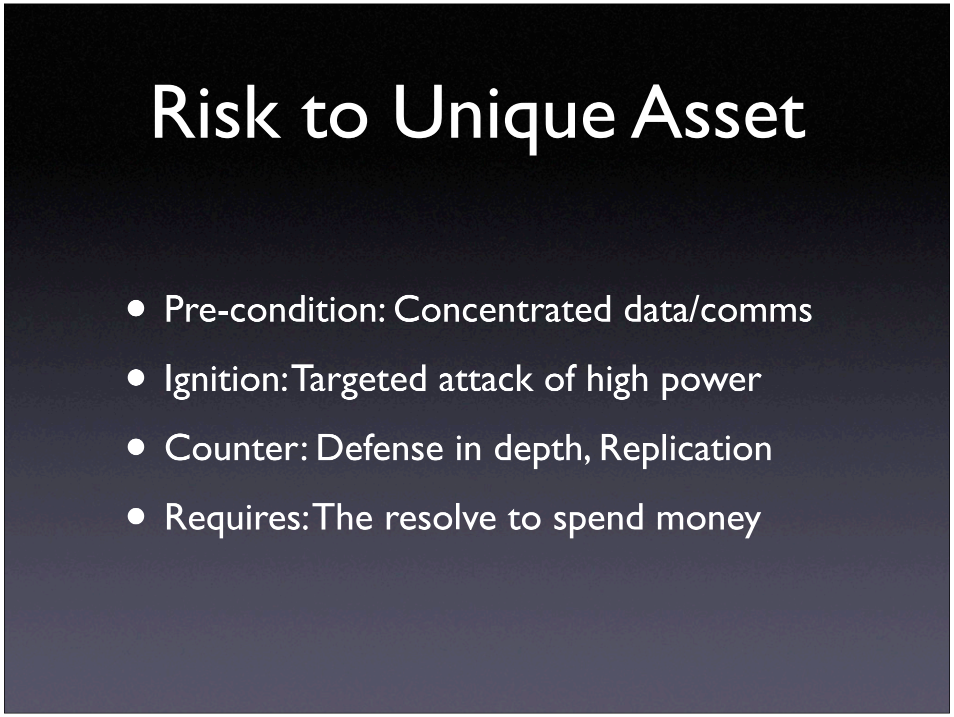
- Loss of inherently unique assets
  - GPS array, FAA EBS, DNS
- Cascade failure
  - Victims become attackers at high rate

Everything else is less important

If having to name the only risks that matter at the national scale, there seem to be two classes and only two classes.

On the one hand, there are entities that are inherently unique by design. For example, the Global Positioning System satellite array (taken as a unit) is one such entity; the Federal Aviation Administrations emergency broadcast system is another, and the Domain Naming System is another. In each case, it is an authoritative data or control source which would be less authoritative if it was surrounded by alternatives. Putting it differently, you only want one red telephone though losing that red telephone is a risk at the national scale.

On the other hand, there entities that are dangers to each other in proportion to their number -- any risk which, like an avalanche, can be initiated by the one but propagated by the many. This “force multiplication” makes any of this class of risks a suitable candidate for national scale.

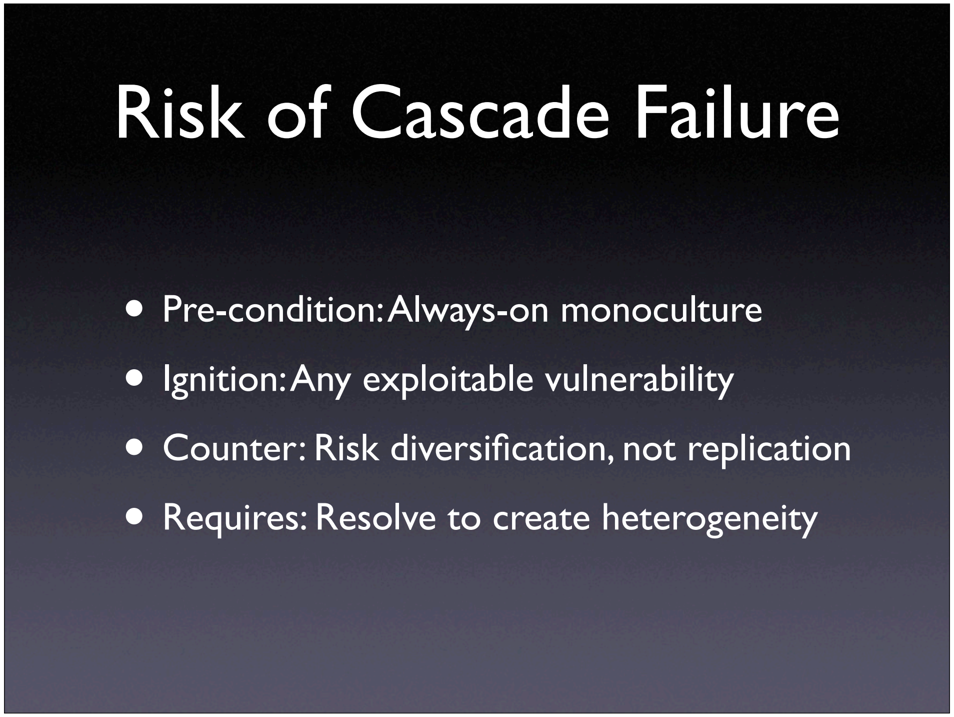


Risk to Unique Asset

- Pre-condition: Concentrated data/comms
- Ignition: Targeted attack of high power
- Counter: Defense in depth, Replication
- Requires: The resolve to spend money

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.

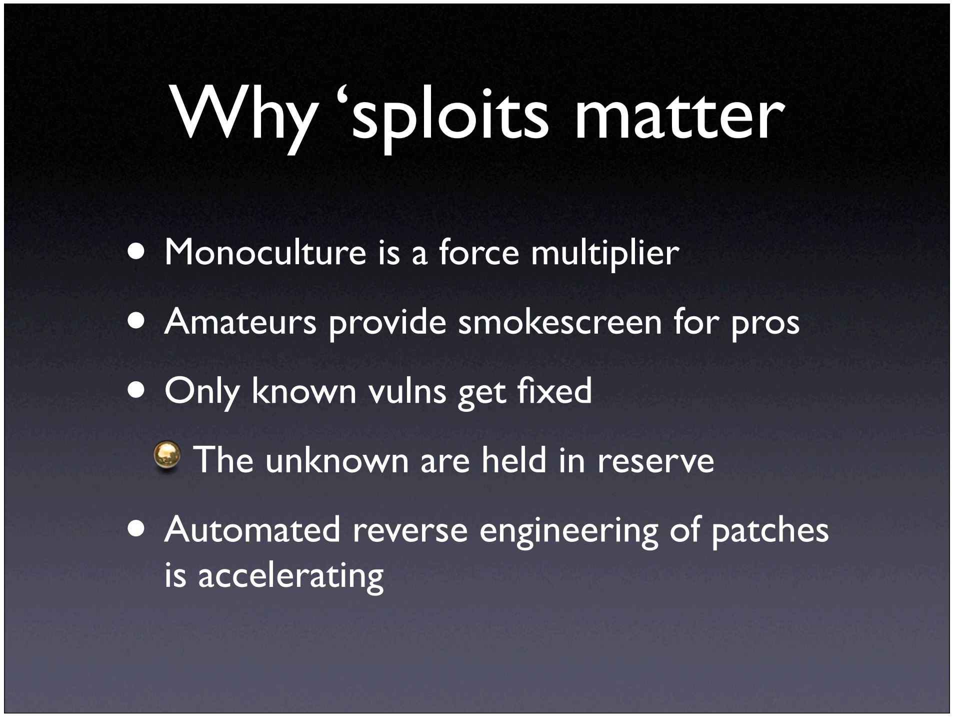


Risk of Cascade Failure

- Pre-condition: Always-on monoculture
- Ignition: Any exploitable vulnerability
- Counter: Risk diversification, not replication
- Requires: Resolve to create heterogeneity

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.



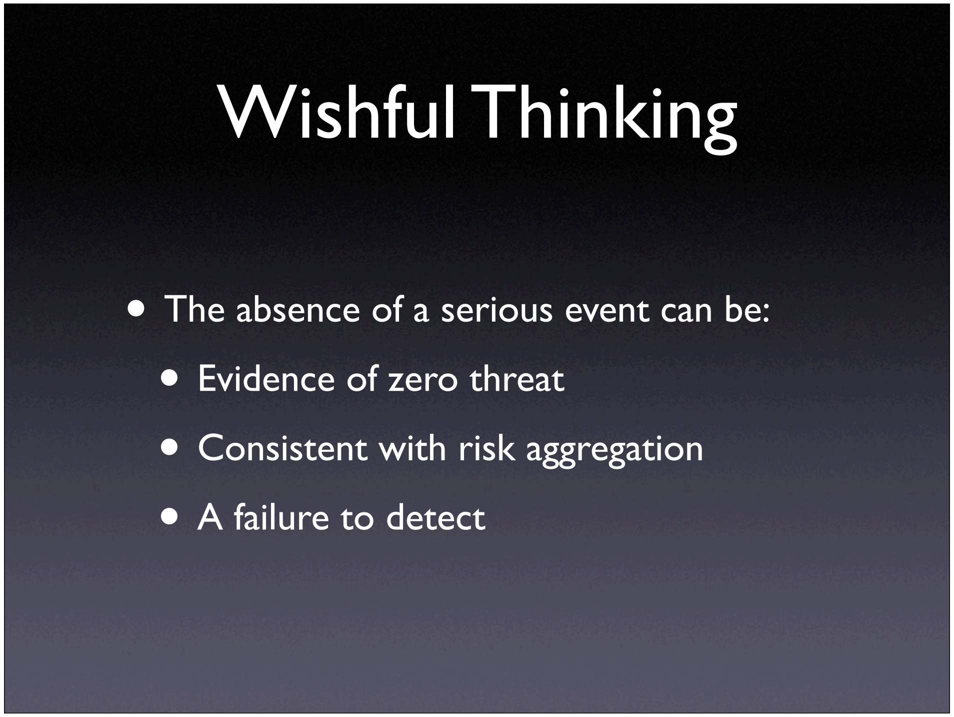
Why ‘sploits matter

- Monoculture is a force multiplier
- Amateurs provide smokescreen for pros
- Only known vulns get fixed
  - The unknown are held in reserve
- Automated reverse engineering of patches is accelerating

So why do exploits matter? Because in a monoculture they are the ignition and their propagation amongst potential instigators of a cascade failure is well documented. Of course, the extent of their existence and propagation is unknowable in and of itself, but it is clear that the testing of exploits by the most expert is sufficiently obscured by the constant rain of amateur attacks. One estimate (by John Quarterman of Internet Perils) is that perhaps 10% of total Internet backbone traffic is low-level scans while another (by Vern Paxson of Lawrence Livermore) is that for a site such as Livermore one can expect perhaps 40% of inbound connections to be attacks.

Because only known vulnerabilities get fixed, the central question is who knows what and when. The conservative assumption for a vulnerability discoverer is that he was not the first to discover the current vulnerability. A similarly conservative assumption is that not all vulnerability discoverers are of good will. Therefore the question is “How many vulnerabilities are known, silently, to persons not of good will?” The corroborating evidence that this number is non-zero lies in observing that all major virus or worm attacks to date have exploited previously known vulnerabilities, never unknown ones. With such evidence, either all vulnerabilities are discovered by persons of good will or there is a reservoir of vulnerabilities being held in reserve.

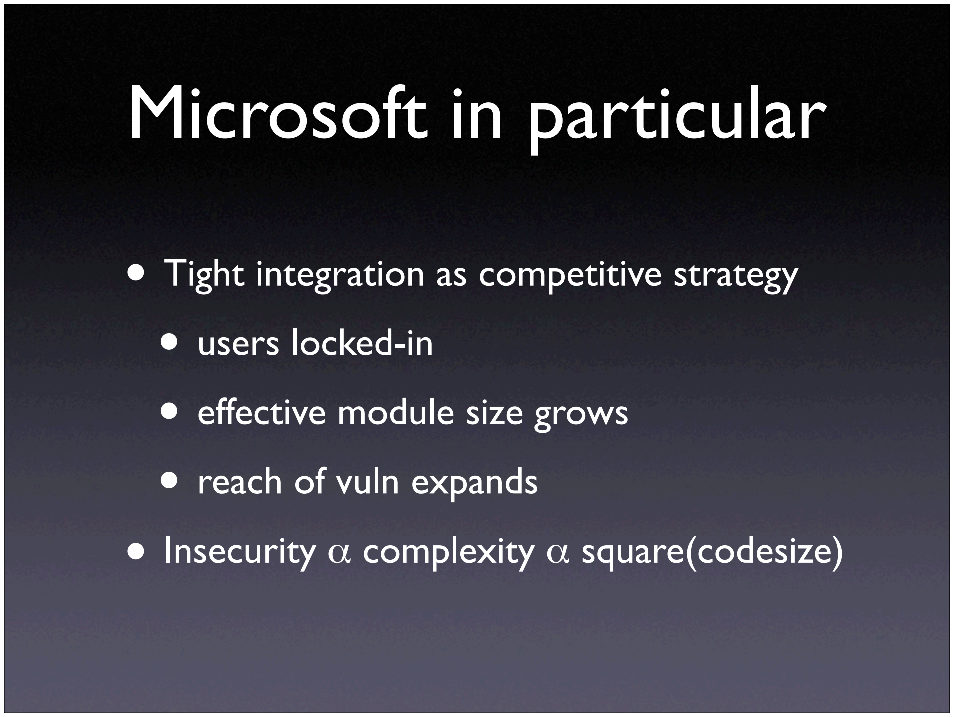
Note that converting patches into vulnerabilities by reverse engineering the patches is not only now the dominant source of exploits but that it is becoming much quicker due to automation. In two years the times have dropped from six months to under a week for principal attacks of public interest. Further declines may no longer matter.



Wishful Thinking

- The absence of a serious event can be:
  - Evidence of zero threat
  - Consistent with risk aggregation
  - A failure to detect

The absence of a major attack event, the situation in which we find ourselves today, is not, as it might seem on first blush, reassuring of low threat. It is consistent with low/no threat to be sure, but it is also consistent with risk aggregation (an insurance term where instead of 1,000 claims occurring at random one instead gets 1,000 claims all at once, the difference between house fires and earthquakes). It is also consistent with a failure to detect, though less likely that "major" and "indetectible" are likely to appear in the same sentence.

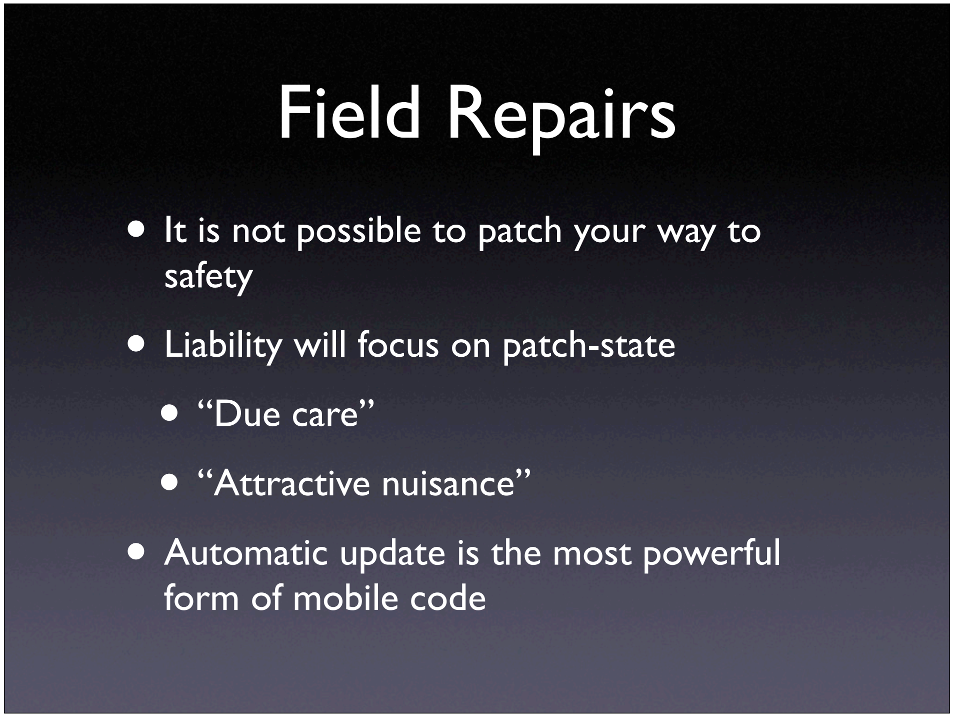


Microsoft in particular

- Tight integration as competitive strategy
  - users locked-in
  - effective module size grows
  - reach of vuln expands
- Insecurity α complexity α square(codesize)

The situation with Microsoft is the critical focus just as when discussing solar power one must speak of Sol.

The quality control literature leads one to expect that as effective code size grows complexity grows as the square of that code size. Similarly, the quality control literature expects total flaws, of which security flaws are a subset, to grow linearly in complexity. Microsoft's competitive strategy is manifestly to achieve user-level lock-in via tight integration of applications. This tight integration, besides violating software engineering wisdom, expands effective module size to that of the tightly integrated whole and thus inevitably creates the platform most likely to have security flaws, and by a wide margin. Coupled with its 94% market share one thus achieves the vulnerable monoculture on which cascade failure depends.

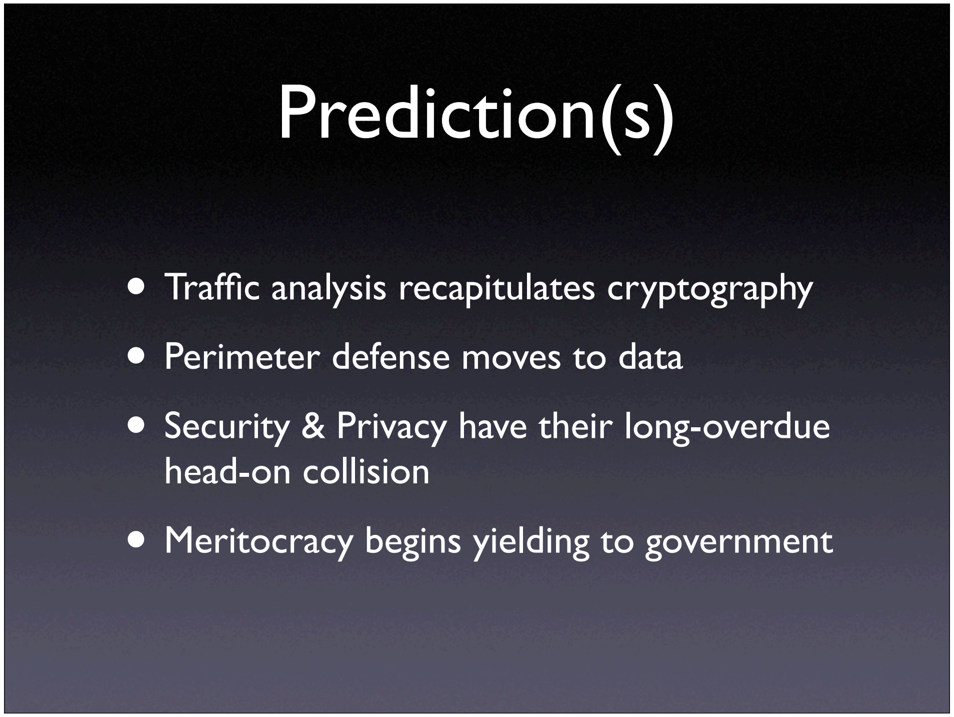


Field Repairs

- It is not possible to patch your way to safety
- Liability will focus on patch-state
  - “Due care”
  - “Attractive nuisance”
- Automatic update is the most powerful form of mobile code

Field repairs, the dominant activity if not strategy of the present time, are at best a damage containment. It is not possible to patch yourself to safety: any significant patch latency preserves the critical mass (of vulnerable entities) while every patch has a non-zero chance of collateral damage. Thus we come to a discussion of formal liability and the high likelihood that in the short term such discussion will focus on patch-state as a proxy for culpability either in the sense of patching being evidence of due care or the lack of patching, particularly within substantial enterprises, being evidence of an attractive nuisance (like an unfenced swimming pool).

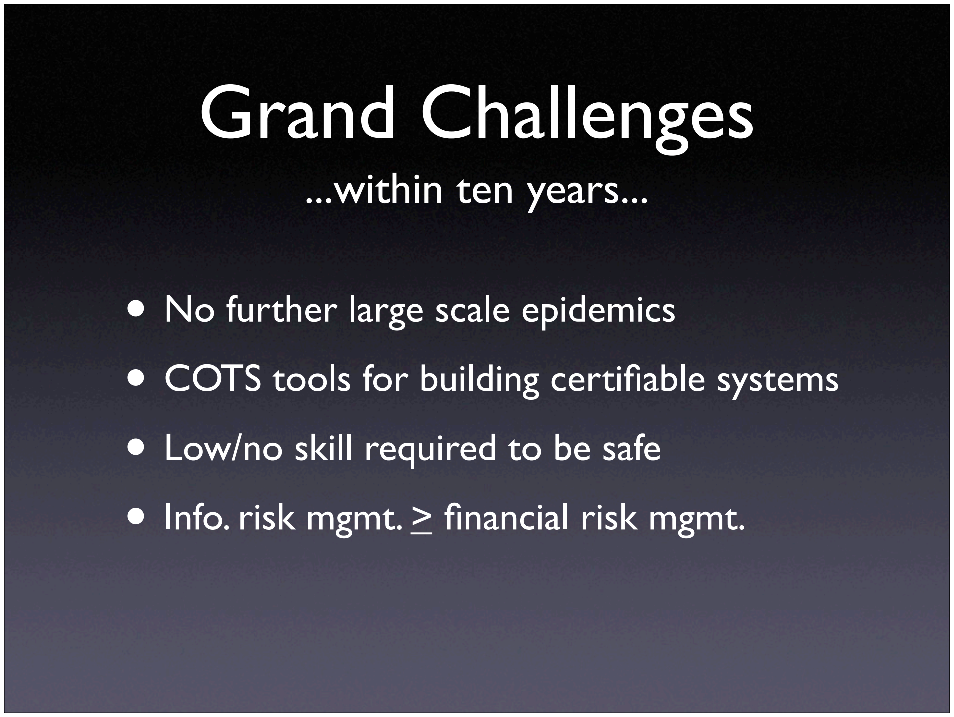
Looking dispassionately at risk, one must also conclude that automatic update is the ultimately powerful form of mobile code. Automatic patching does harden systems but it does so more toward brittleness than toward toughness in that if ever the automatic patching pathway is itself effectively co-opted then the game is largely over at that moment.



Prediction(s)

- Traffic analysis recapitulates cryptography
- Perimeter defense moves to data
- Security & Privacy have their long-overdue head-on collision
- Meritocracy begins yielding to government

No discussion of national level threat can look at the current point in time; it must instead lead its target just as a hunter must his. In that sense, the next ten years (or less) will have the commercial sector catching up to the military in traffic analysis just as the last ten years had that catch-up in cryptography. At the same time, increasing threat will, as it must, lead to shrinking perimeters thus away from a focus on enterprise-scale perimeters and more toward perimeters at the level of individual data objects. Security and privacy are, indeed, interlocking but, much as with twins in the womb, the neoplastic growth of the one will be to the detriment of the other hence the bland happy talk of there being no conflict between the two will be soon shown to be merely that. Finally, the Internet as a creature built by, of, and for the technical and ethical elite being no longer consistent with the facts on the ground, its meritocratic governance will yield to the anti-meritocratic tendencies of government(s).



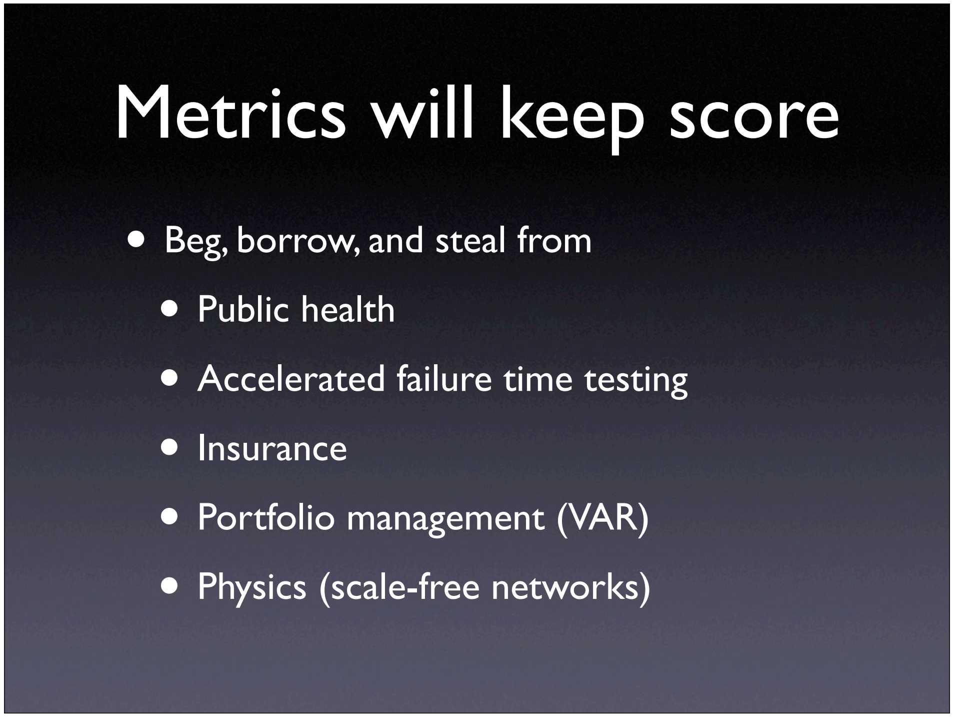
Grand Challenges

...within ten years...

- No further large scale epidemics
- COTS tools for building certifiable systems
- Low/no skill required to be safe
- Info. risk mgmt. ≥ financial risk mgmt.

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.

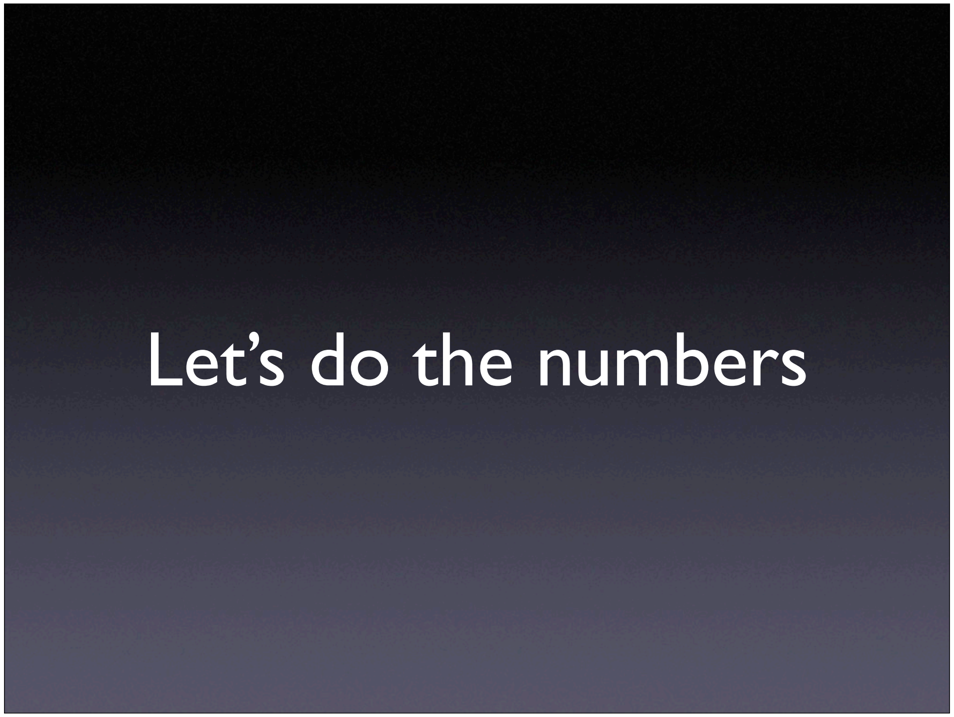


Metrics will keep score

- Beg, borrow, and steal from
  - Public health
  - Accelerated failure time testing
  - Insurance
  - Portfolio management (VAR)
  - Physics (scale-free networks)

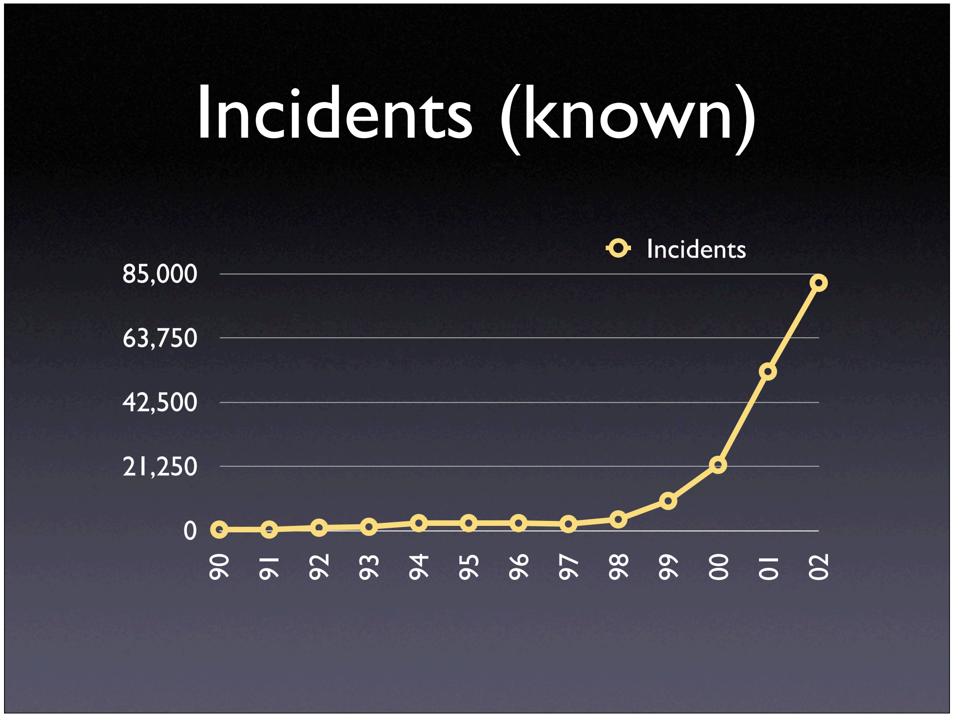
The future belongs to the quants, full stop. As such, and bearing in mind the critical need for security at this time, we must borrow from other fields as we have no time to invent everything from scratch. We are likewise lucky that at this time the field has the maximum of hybrid vigor in that all of its leaders were trained at something else hence our ability to extract from that “something else” is maximal.

For a sense of this, see <http://www.stake.com/research/reports/acrobat/ieee_quant.pdf>, or <http://www.securitymetrics.org>.



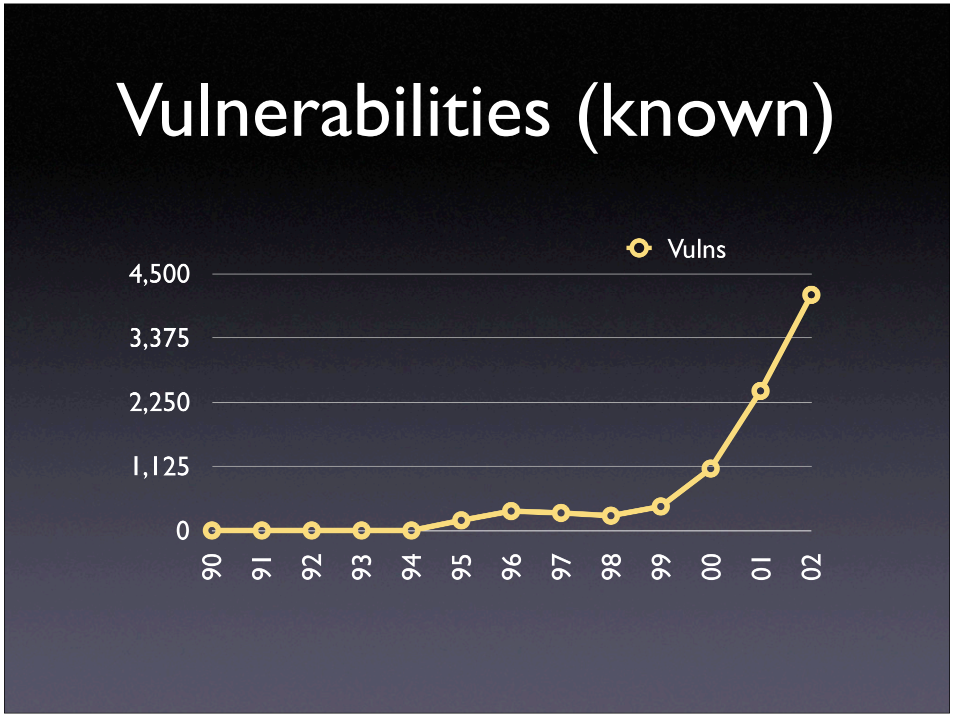
Let’s do the numbers

It is time to illustrate these points.



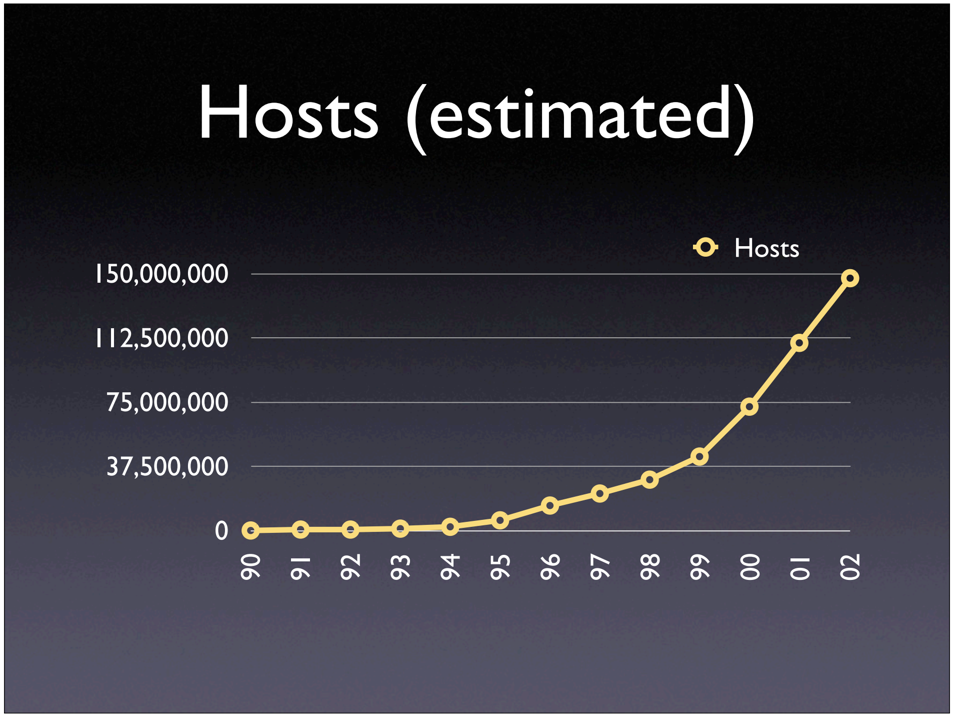
Incidents (known)

Public CERT data



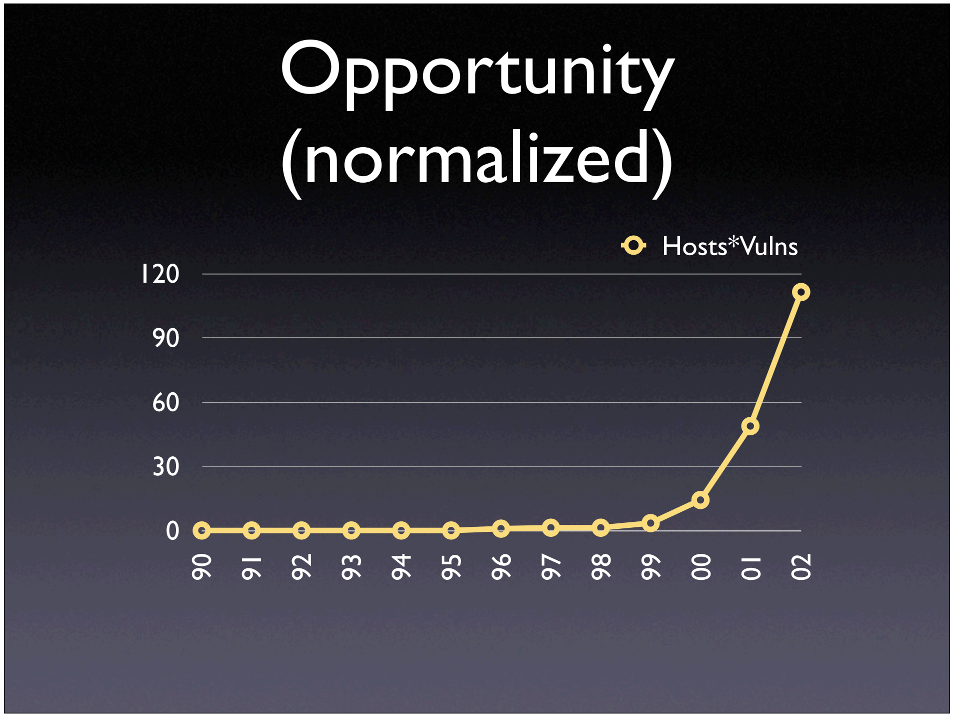
Vulnerabilities (known)

Public CERT data



Hosts (estimated)

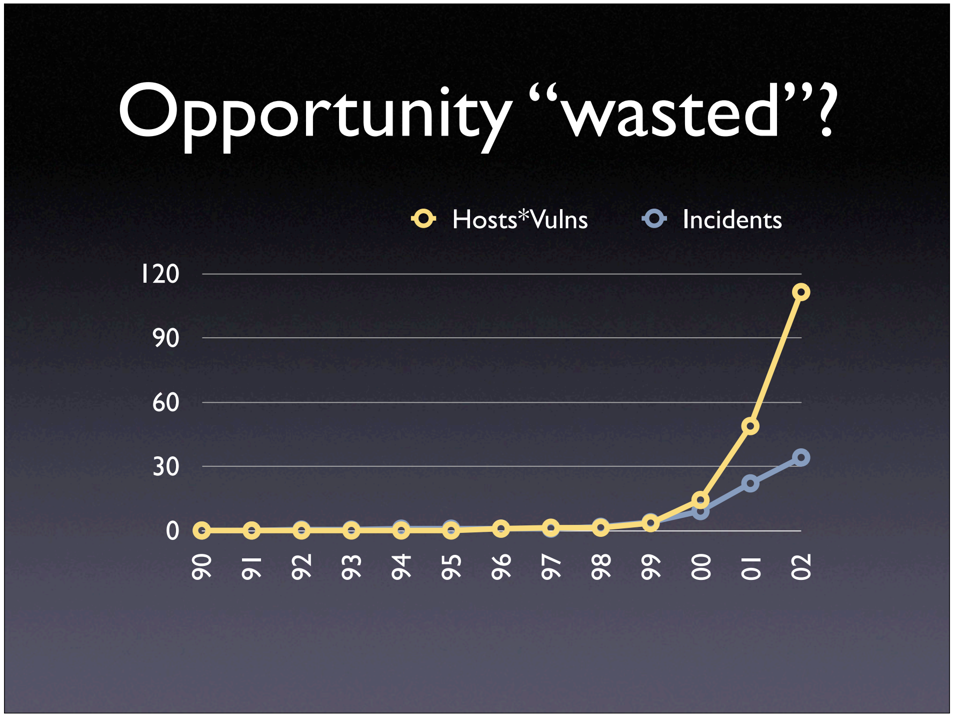
Public ISOC data



Opportunity (normalized)

So how much opportunity is there? Is it well modeled by total number of open holes, i.e., by the product of the number of hosts times the number of vulnerabilities?

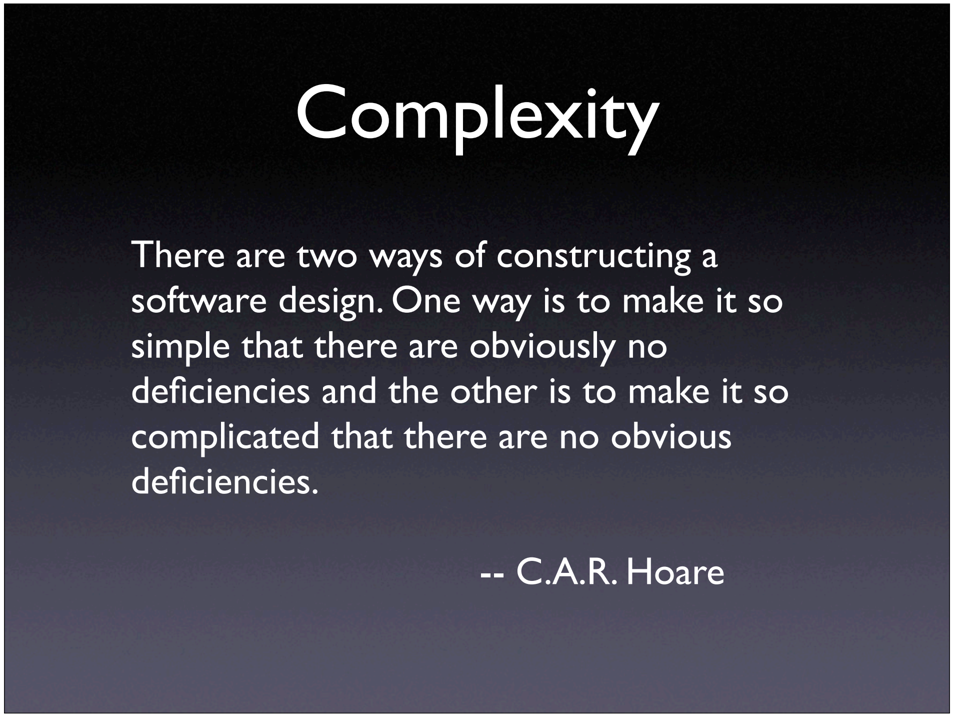
If so, the curve looks like this, and it has taken an amazingly steep turn upward.



Opportunity “wasted”?

If opportunity is proportional to the product of hosts times vulns,...

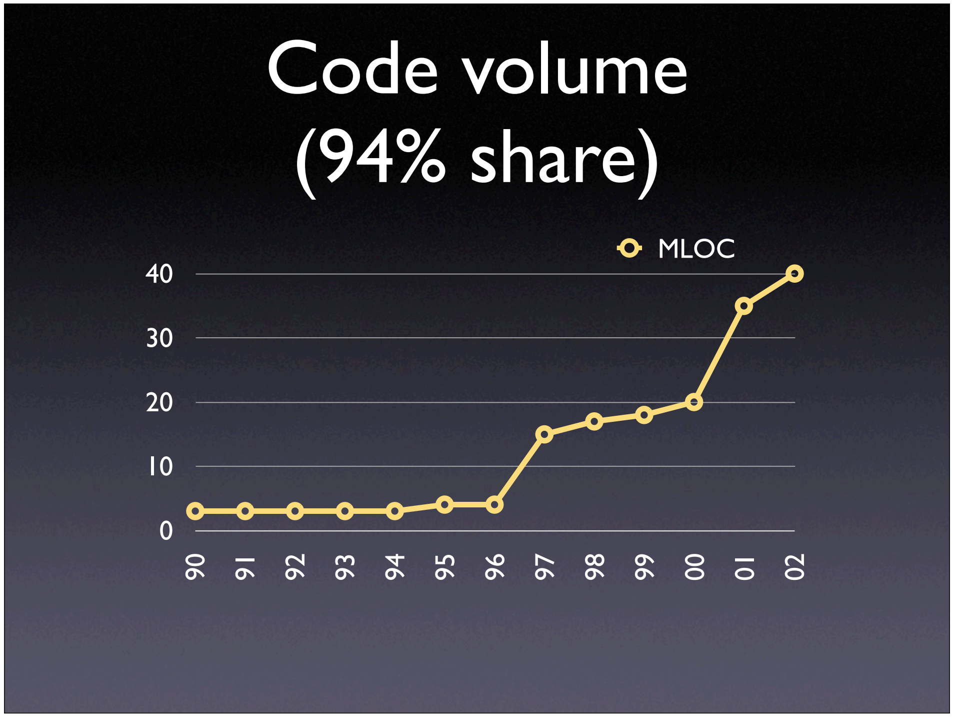
Then either “we” are doing a good job at keeping the crime rate from growing as fast as the opportunity is growing, there is some degree of the attack community holding vulnerabilities in reserve, or there is a growing reservoir of untapped opportunity for attack.



Complexity

> There are two ways of constructing a software design. One way is to make it so simple that there are obviously no deficiencies and the other is to make it so complicated that there are no obvious deficiencies.
>
> -- C.A.R. Hoare

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.

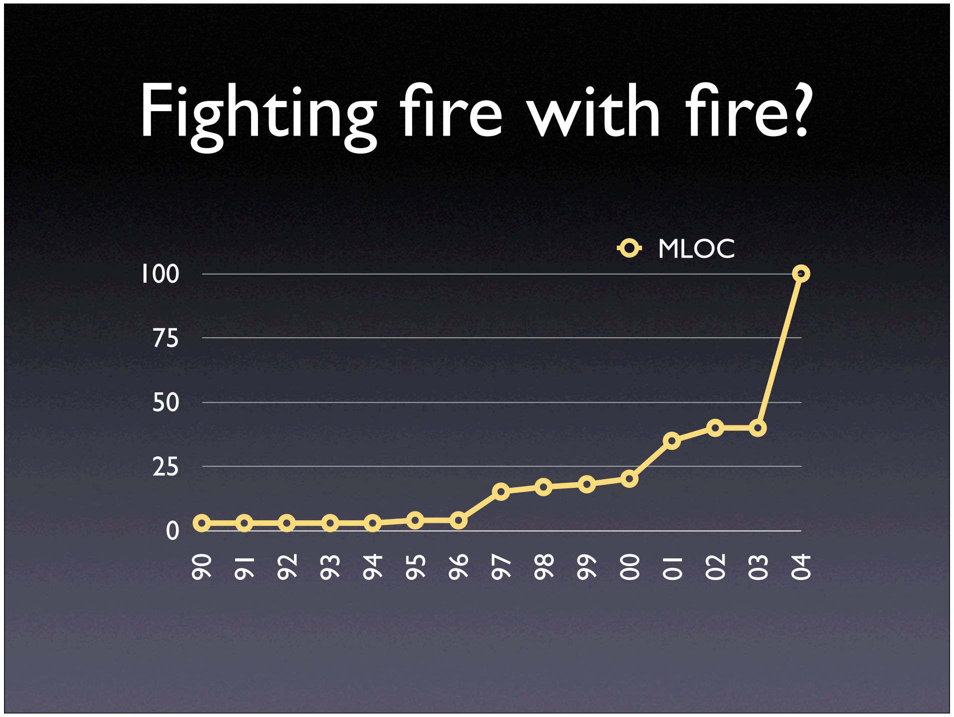


Code volume (94% share)

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

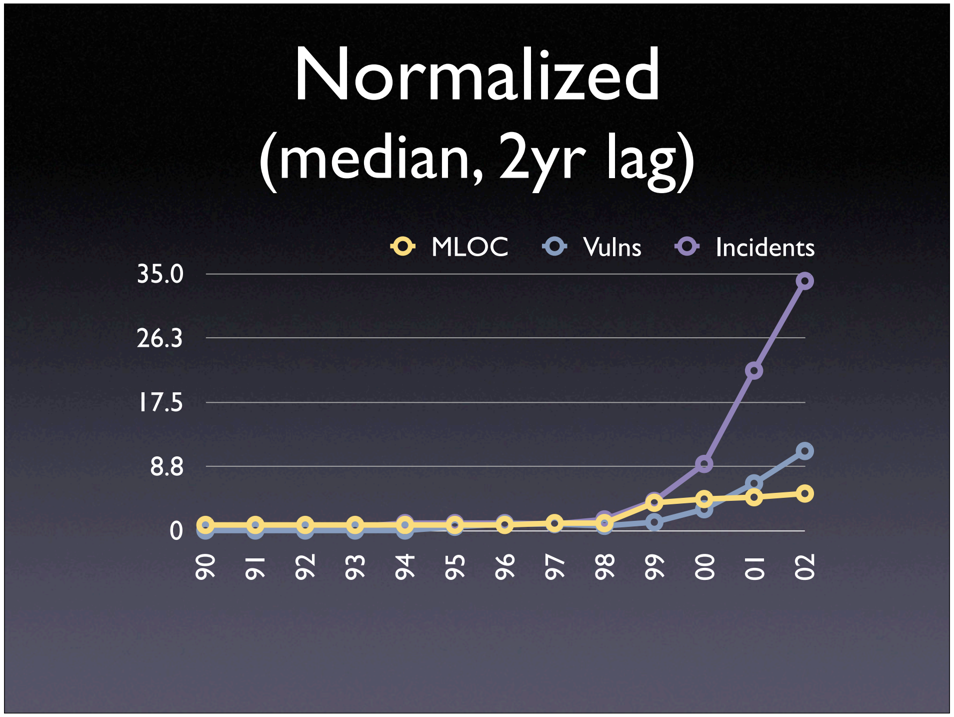


Fighting fire with fire?

Code volume as observed:

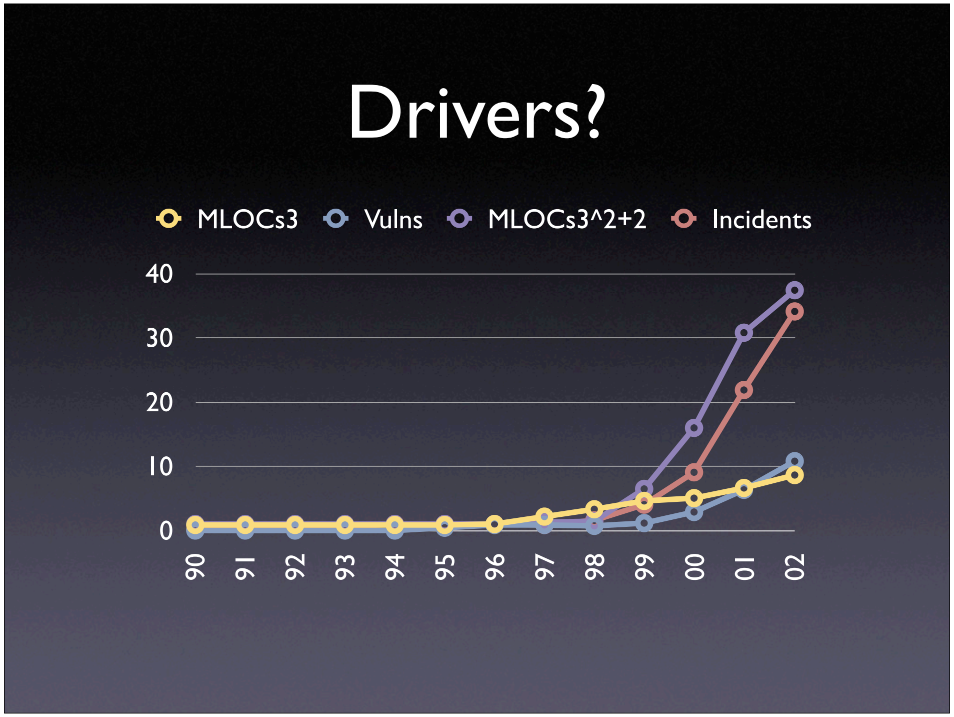
Win 3.1	Win NT	Win 95	NT 4.0	Win 98	NT 5.0	Win 2K	Win XP	Longhorn?
3	4	15	17	18	20	35	40	100?
1990	1995	1997	1998	1999	2000	2001	2002	2004?

(How big will Longhorn be?)



Each curve is normalized against its own median over this period. Therefore, overlaying the curves is legitimate.

Code volume curve is shifted right two years to crudely simulate diffusion delay.

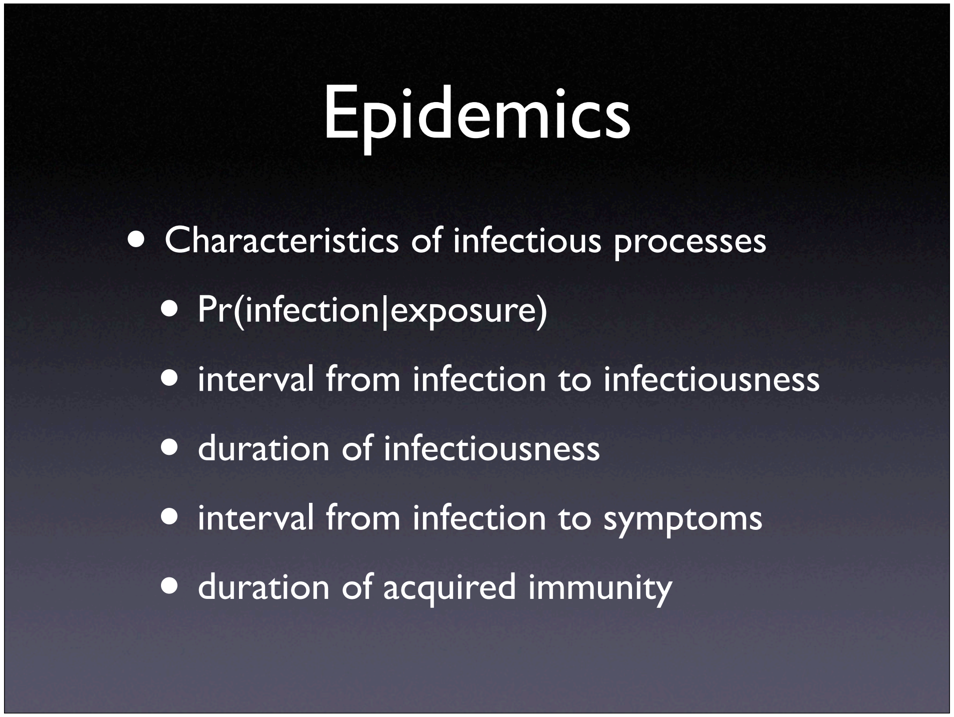


Drivers?

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.

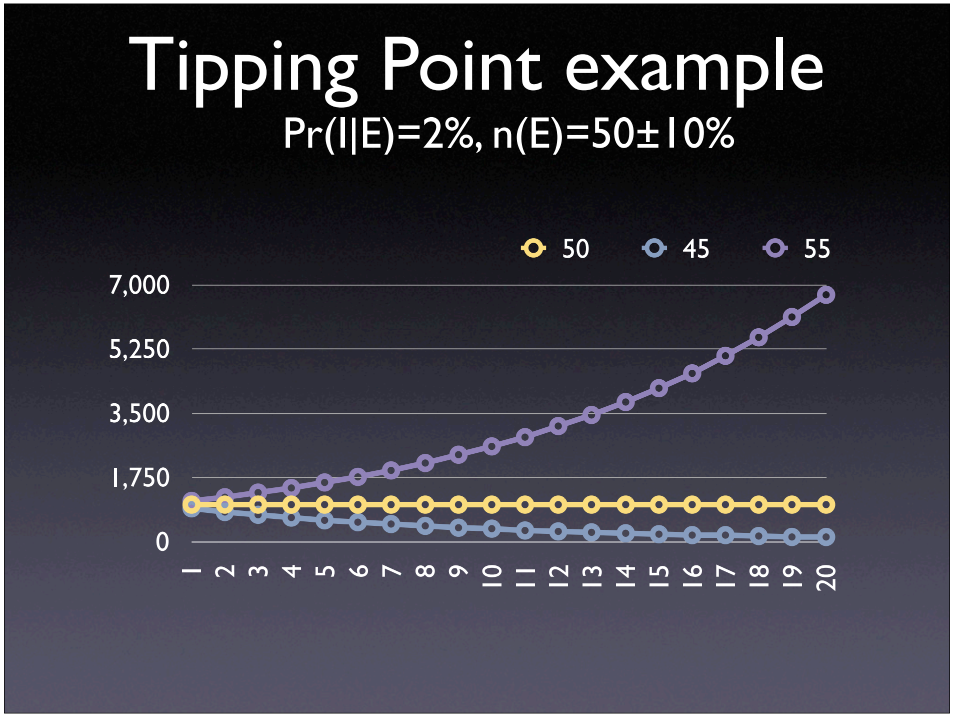
The second code volume curve, MLOCs3^2+2, is the square of the three year moving average of code volume, and then shifted right two years. The argument is this: Security faults are a subset of quality faults and the literature says that quality faults will tend to be a function of code complexity, itself 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 two years 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.



Epidemics

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

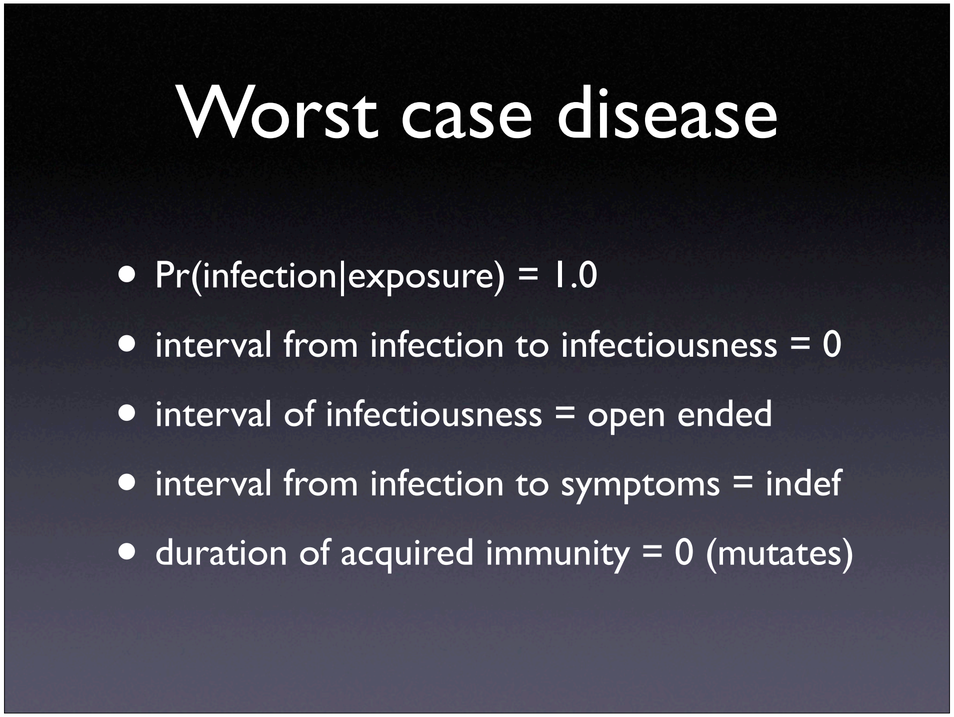
The math for modeling epidemics is well developed, as is the math for accelerated failure time testing, actuarial science, portfolio management, and others. There is no need, and no time, to invent new science before progress can be made. Steal these skills, and do so while the senior practitioners in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



Tipping Point example

Pr(I|E)=2%, n(E)=50±10%

This is simply the example used in Gladwell's The Tipping Point. 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.



Worst case disease

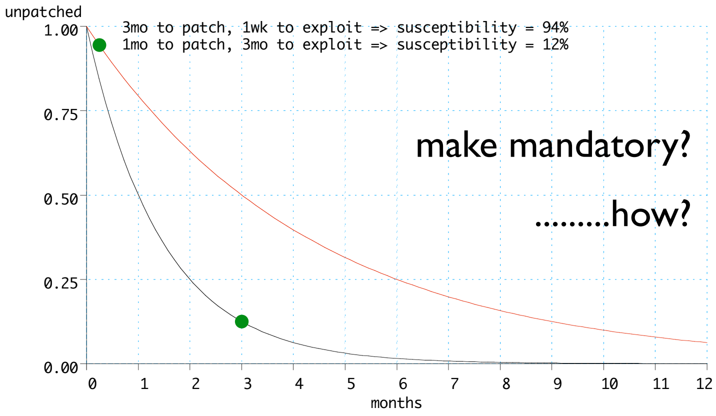
- Pr(infection|exposure) = 1.0
- interval from infection to infectiousness = 0
- interval of infectiousness = open ended
- interval from infection to symptoms = indef
- duration of acquired immunity = 0 (mutates)

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?

Immunization



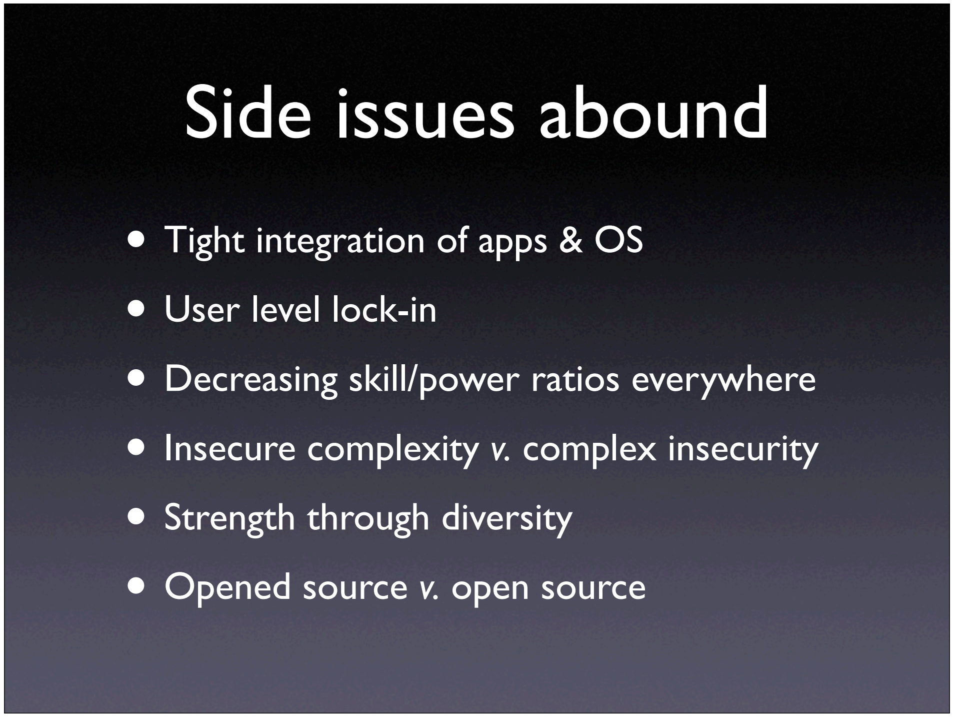
make mandatory?

.........how?

Qualys, Inc., has data that implies patching is like radioactive decay in that 50% of the remaining unpatched systems will be patched in each succeeding “half-life.” Qualys’s figure is 30 days.

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.

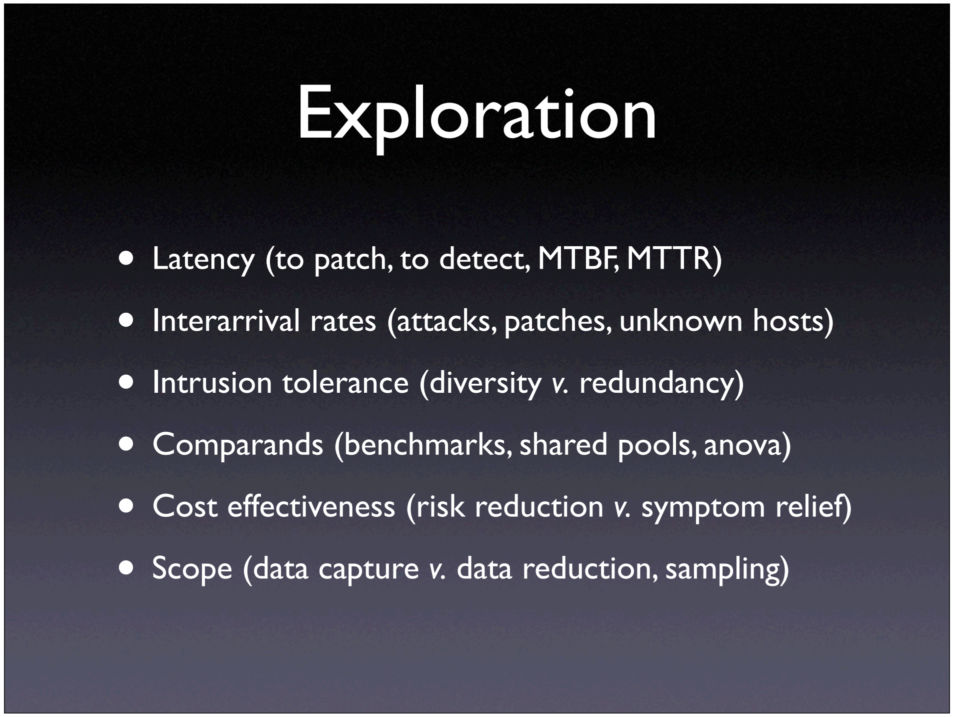
Time-to-exploit is shrinking while the time-to-patch is lengthening (if you factor in the growth of always-on, always-connected home machines) so the question becomes whether “mandatory” is a word we must use and, if so, what would it mean. What does the law say?



Side issues abound

- Tight integration of apps & OS
- User level lock-in
- Decreasing skill/power ratios everywhere
- Insecure complexity *v.* complex insecurity
- Strength through diversity
- Opened source *v.* open source

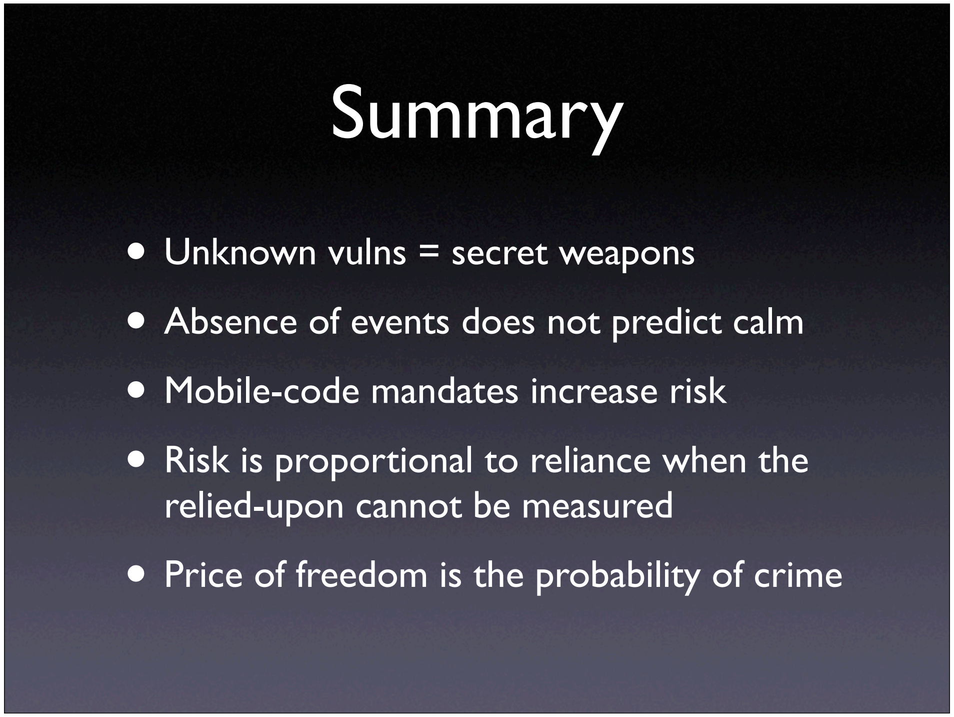
This list is indicative, not exhaustive. It includes the monopolization questions of tying the applications to the operating system thus to insure that a security failure of the one is a security failure of the other, whether user level lock-in plays a role in assessing the locus of liability for security faults, and whether the skill to operate ever more powerfully interconnected computers does not at some point require some a priori proof of capability. It asks to distinguish complexity that is insecure from insecurity that is itself complex. It ponders the question of genetic diversity as a survival advantage in a world where predators have just arisen. It distinguishes the value of public disclosure in the open source tradition to the private disclosure of the entirety of the Windows (94% share) source code pool to potentially hostile nation states. It could go on. The challenge is substantial and historically crucial. What will the law say, and can it say it without adding noise?



Exploration

- Latency (to patch, to detect, MTBF, MTTR)
- Interarrival rates (attacks, patches, unknown hosts)
- Intrusion tolerance (diversity *v.* redundancy)
- Comparands (benchmarks, shared pools, anova)
- Cost effectiveness (risk reduction *v.* symptom relief)
- Scope (data capture *v.* data reduction, sampling)

To go on from here we can’t use words, they don’t say enough. We must use numbers. These are indicative and intended to push you to think of more. Even if the shorthand does not read clearly, the point is this: now that the digital world is essential, statistics based on the realities of digital physics will be, at least, how score is kept. Perhaps we will be fortunate and statistics based on the realities of digital physics will also inform decision making at the highest levels, including the law.



Summary

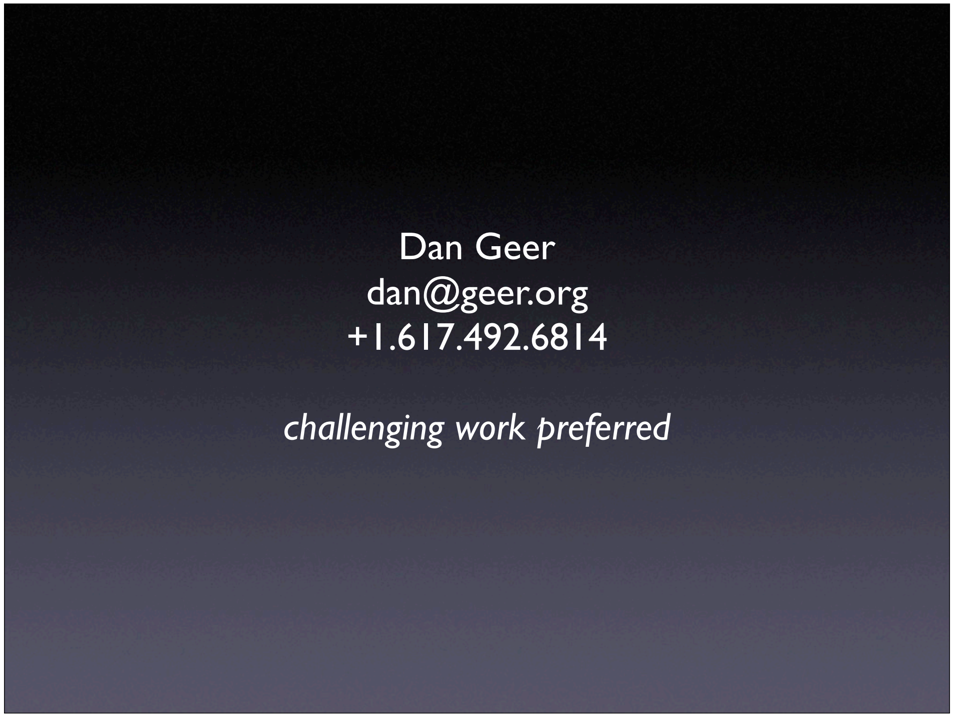
- Unknown vulns = secret weapons
- Absence of events does not predict calm
- Mobile-code mandates increase risk
- Risk is proportional to reliance when the relied-upon cannot be measured
- Price of freedom is the probability of crime

In summary, the pool of selectively known vulnerabilities is the secret weapon of the serious enemy, the absence of a significant catastrophe to date is most assuredly not evidence that the risk is low because in a risk aggregated world significant events make up in their severity what they lack in frequency, that mandates for automatic patching are effectively mandates for more powerful mobile code and are thus risk creating in the larger sense of risk aggregating, that risk is itself proportional to the reliance on places in the entity being relied upon exactly when there are no effective measures, and that the tradeoff between freedom (default permit) and safety (default deny) is real and present.

There is never enough time.....

.....Thank you for yours

It has been entirely my pleasure.



Dan Geer
dan@geer.org
+1.617.492.6814

challenging work preferred

Further contact is welcome particularly if it brings problems of the sort that illustrate the bounds of our knowledge.