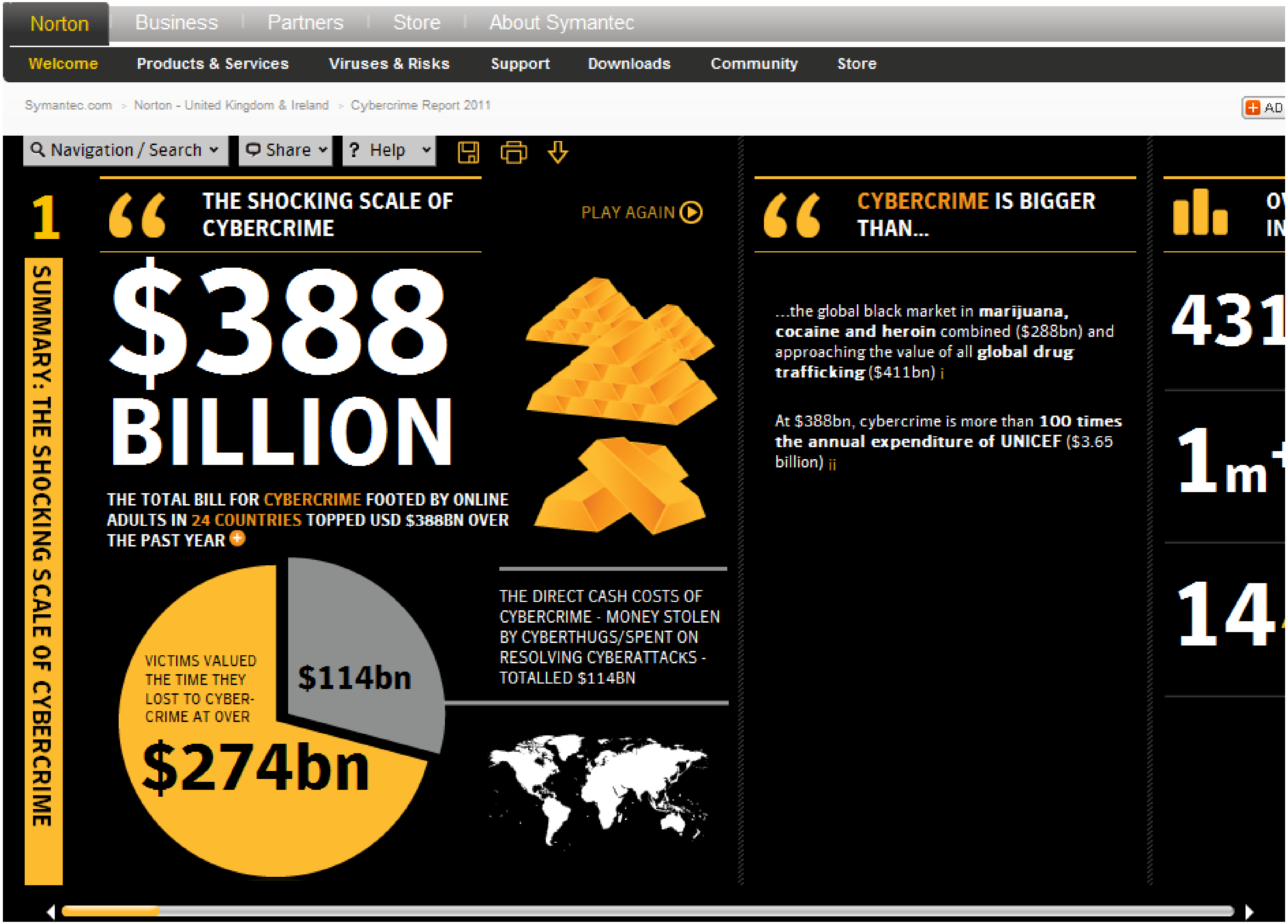
Sex, Lies and Cyber-crime Surveys

Dinei Florêncio and Cormac Herley

Microsoft Research, Redmond

Link to paper from WEIS 2011



THE SHOCKING SCALE OF CYBERCRIME

$388 BILLION

THE TOTAL BILL FOR CYBERCRIME FOOTED BY ONLINE ADULTS IN 24 COUNTRIES TOPPED USD $388BN OVER THE PAST YEAR

VICTIMS VALUED THE TIME THEY LOST TO CYBER-CRIME AT OVER $274bn

$114bn

THE DIRECT CASH COSTS OF CYBERCRIME - MONEY STOLEN BY CYBERTHUGS/SPENT ON RESOLVING CYBERATTACKS - TOTALLED $114BN

CYBERCRIME IS BIGGER THAN...

...the global black market in **marijuana, cocaine and heroin** combined ($288bn) and approaching the value of all **global drug trafficking** ($411bn) i

At $388bn, cybercrime is more than **100 times the annual expenditure of UNICEF** ($3.65 billion) ii

- Cybercrime estimates come from surveys

$$Estimate = \frac{|\boldsymbol{X}|}{|\boldsymbol{R}|} \sum_{i \in \boldsymbol{R}} f[r_i]$$

- Methodology extremely flawed

Conclusion: no faith in any of these estimates

“You should never trust user input”

Writing Secure Code, Howard et al.

- Importance of input validation in security
  - SQL Injection
  - Buffer overflows

$$\frac{|\boldsymbol{X}|}{|\boldsymbol{R}|} \sum_{i \in \boldsymbol{R}} f[r_i] \quad \equiv$$

```
total = 0.0;
for (i=0; i < survey_size; i++){
        total += (double) strcpy(user_input[i]);
}
estimate  = total * population_size/survey_size
```

- Practice unacceptable in writing code ubiquitous in forming estimates

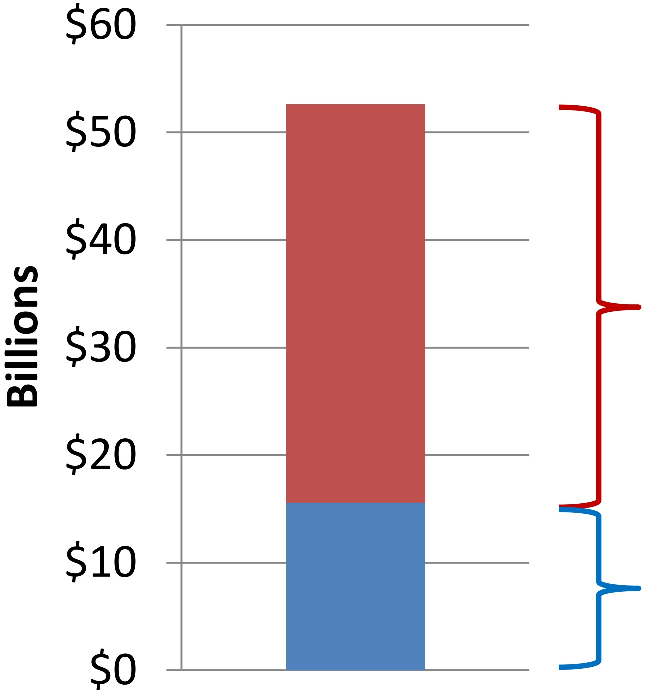
Surveys and errors…

- Single transcription error inflates US household wealth estimate by $1 trillion. (Fed. Reserve Consumer Finance Survey, 1983)

- Two respondents add $37 billion to ID Theft survey. (FTC ID Theft Survey, 2006)
  - Including these two: $52 billion
  - Excluding these two: $15.6 billion

- Men claim between 3x and 9x more lifetime heterosexual partners than women. (Various sex surveys)

FTC ‘06 ID Theft Survey



Portion of estimate
From 2 respondents

Portion of estimate
From 4998 respondents

- Two respondents contribute a factor of $\frac{37}{2}\Big/\frac{15.6}{4998}$ = 5927 more than average

- Two vote at 6000x strength of everyone else.

Sex and Lies

Men Report 3-9x More Female Partners than Women Report Male Partners (survey, after survey after survey).

Morris [Nature ‘93]: “male reports only slightly exceed female reports for lower 90% of the sample, but dramatically exceed them for the upper 2%”

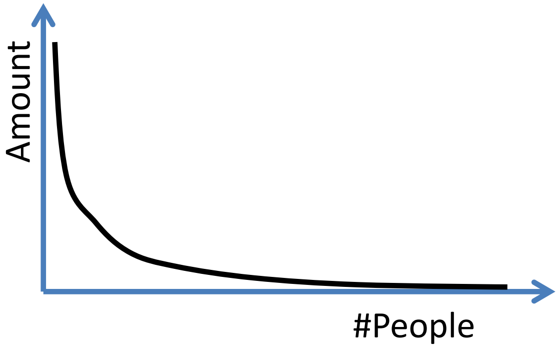
	Male/female reports
All	3.2
Bottom 90%	1.2
Top 10%	10.1

Simple Explanation: some men tell whoppers

- Consider man with 2x overstatement
  - E.g. correct answer = 50, but claims 100
- Takes 16 men at the median to understate by 2x to cancel this single error

What do these Surveys have in common?

- Uneven distribution
  - Wealth, losses, #sex partners



- Small number of outliers explain the problem

- **Bias is to the upside, not downside.**
  - **Errors don’t cancel**

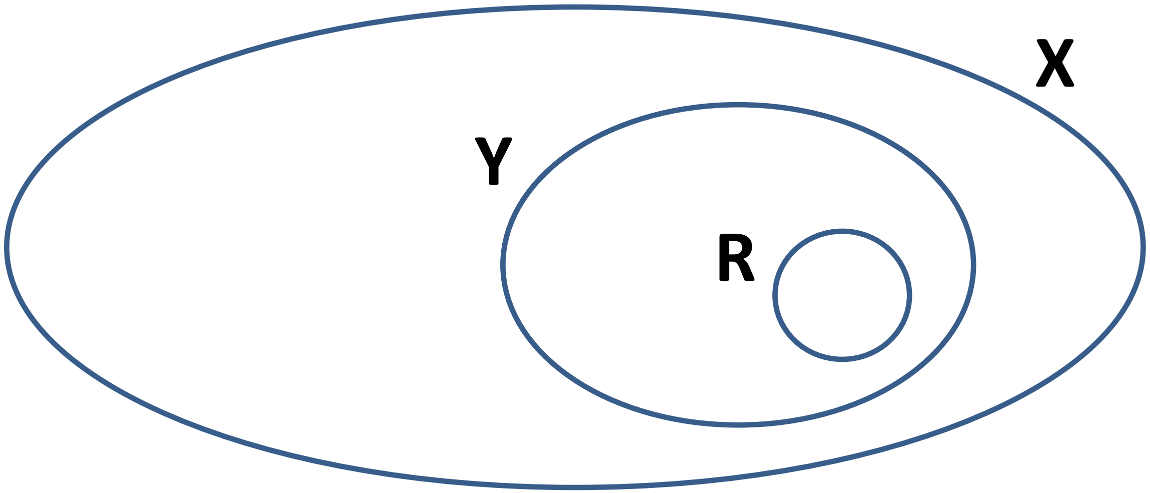
Answers can be wrong by orders of magnitude

How can this be? Aren't election surveys pretty accurate?

- Set membership vs Numeric surveys
- Voting Intentions Survey
  - Voting power is evenly spread
  - No outliers (one vote per respondent)
  - Errors cancel

Sources of errors in Surveys

X: whole population
Y: contacted population
R: responding population
$f[r_i]$ = observed i-th individual response



$$\text{Estimate} = |X|\bar{x} \overset{?}{\approx} \frac{|X|}{|R|}\overline{f[r]} = \frac{|X|}{|R|}\sum_{i \in R} f[r_i]$$

$$\bar{x} - \overline{f[r]} = (\underset{\uparrow}{\bar{x} - \bar{y}}) + (\underset{\uparrow}{\bar{y} - \bar{r}}) + (\underset{\uparrow}{\bar{r} - \overline{f[r]}})$$

Sampling bias   Non-response error   Response bias

Sources of Survey Error

- Sampling bias: $\bar{x} \neq \bar{y}$
  - Contacted pop not representative
  - Inadequate sampling (margin of error)
- Non-response bias: $\bar{y} \neq \bar{r}$
  - Responding pop not representative
- Response bias: $\bar{r} \neq \overline{f[r]}$
  - Responses not accurate
  - Unbounded for non-negative quantities

Toy Example: "How much did you win from gambling last week?"

X = 230 million, 0.1% have avg. winning \$100,

$$|\mathbf{X}|\,\bar{x} = \$23 \text{ million}$$

- Selection bias ($\bar{x} - \bar{y}$): contact 50,000 people, but sample contain twice as many gamblers.
- Non-response Bias ($\bar{y} - \bar{r}$): 10% answer survey, but gamblers 10X more likely to answer the survey.
- Response bias ($\bar{r} - \overline{f[r]}$): Suppose gamblers exaggerate by 5X. One gambler claims \$1 million

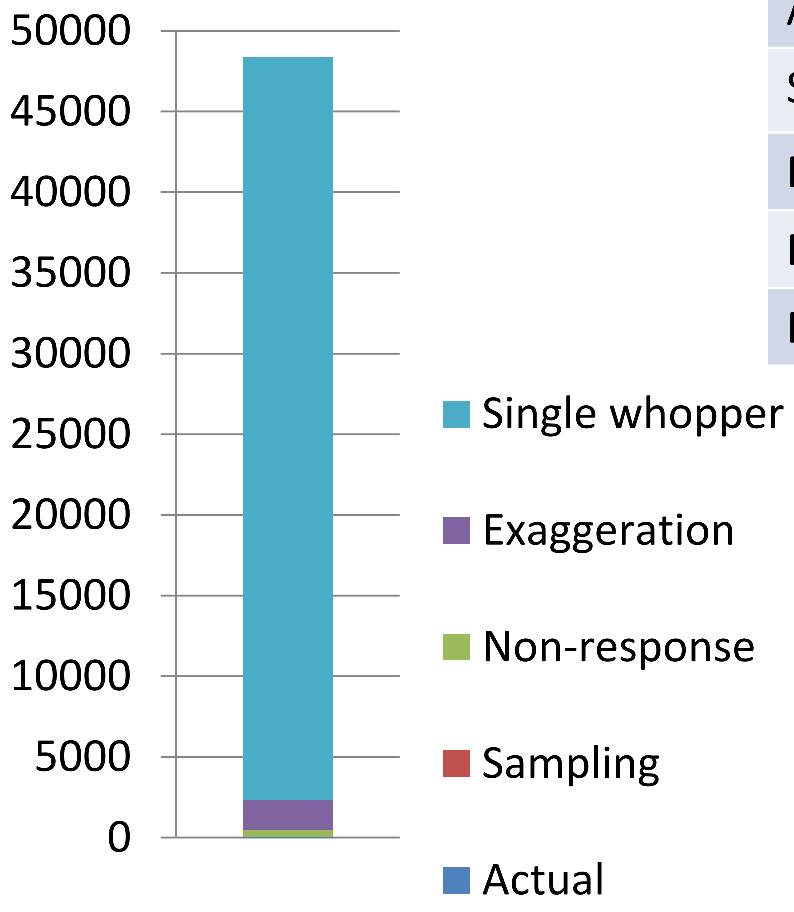
$$\text{Estimate} = \frac{|X|}{|R|}\overline{f[r]} = \frac{230 \times 10^6}{5000}(4 \times 500 + 1 \times 10^6)$$

= \$48 billion

$\pm$0.27% at 95% confidence

Contributions to 2000x Error

	$ millions
Actual Answer	$23
Sampling bias	+$23
Non-response bias	+$414
Response bias [5x exaggeration]	+$1886
Response bias [single whopper]	+$46000



±0.27% at 95% confidence!!

(refers only to Sampling error)

How did we get here?
Why so different from voting surveys?

$$\frac{|\boldsymbol{X}|}{|\boldsymbol{R}|}\sum_{i\,\in\boldsymbol{R}} f[r_i]$$

In contrast to voting surveys:

- Error ($f[r_i]-r_i$) unbounded above
- Errors don’t cancel
- Errors multiplied by $|\mathbf{X}|/|\mathbf{R}|$ = 46,000

Asymmetric Response Error

- When surveyed quantity is non-negative
- Response Error = $(f[r_i] - r_i)$
  - bounded below: $-r_i$
  - unbounded above

- Sex Survey: it takes 2x understatement by 16 men at the median to cancel a liar who adds 50 to his total
- Gambling Survey: 95% of the estimate is coming from single whopper. No cancelation possible.

Concentration: signs of potential problem

- Gambling Survey
  - Median $500
  - Mean $200,000
  - Mean/Median = 400
- Large ratio of mean/median means that small fraction accounts for majority of estimate

Concentration in Cyber-crime Surveys

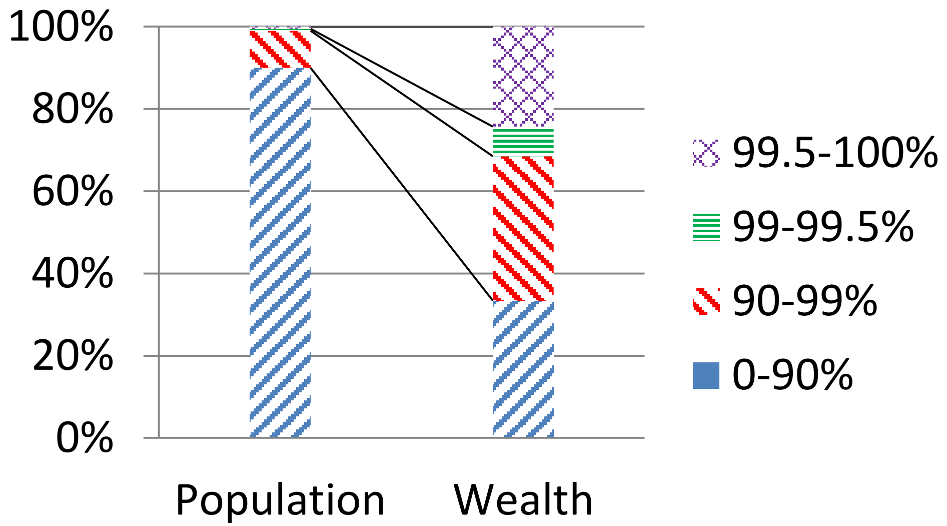
	Mean/Median	Top 1% accounts for*
US Wealth	4.12	33%
FTC ID Theft ‘06	3.75	30%
Gartner Phishing ‘07	4.5	59%
DoJ ID Theft Survey ‘08	9.3	72%
IC3 ID theft 2007	9.7	78%

Eye-popping levels of concentration: tiny minority of respondents Account for majority of estimate.

* Fitting a Pareto (i.e. power law) distribution

Many Phenomena really are Concentrated

- Concentration of US wealth: mean/median = 4.3



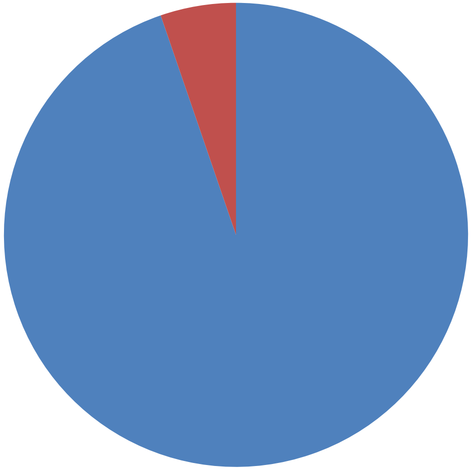
- Problem is: can’t tell difference from true outlier and lie/exaggeration

Doing it Right

- US Bureau of Census Wealth Survey
  - Uniform sample of 3824 households
  - Second sample of 438 high-wealth households
    - Two strata: median $50 million, $300 million
- It takes real effort to sample the upper tail
- Surveying Concentrated phenomena is hard

Rare Concentrated Phenomena

- **|R|**= 5000 person survey of phenomenon that affects 5% => 250
- Top 1% of 250 → 2.5 respondents
- If top 1% accounts for 75% then 1-3 respondents dominating estimate.



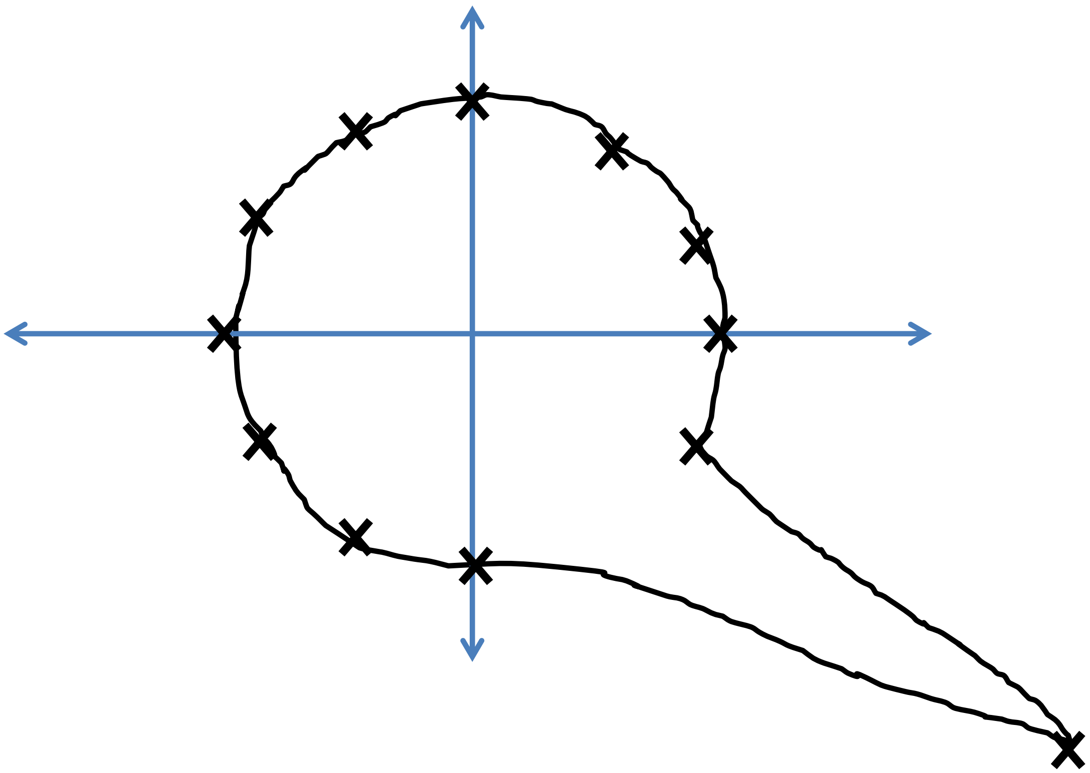
	Survey Size	Percent	# Victims	Estimate (millions)
FTC (all)	4917	3.7%	181	$15,600
FTC (old accts)	4917	1.4%	68	NA
FTC (new accts)	4917	0.8%	39	NA
Javelin ‘03	4000	0.0723%	3	$367

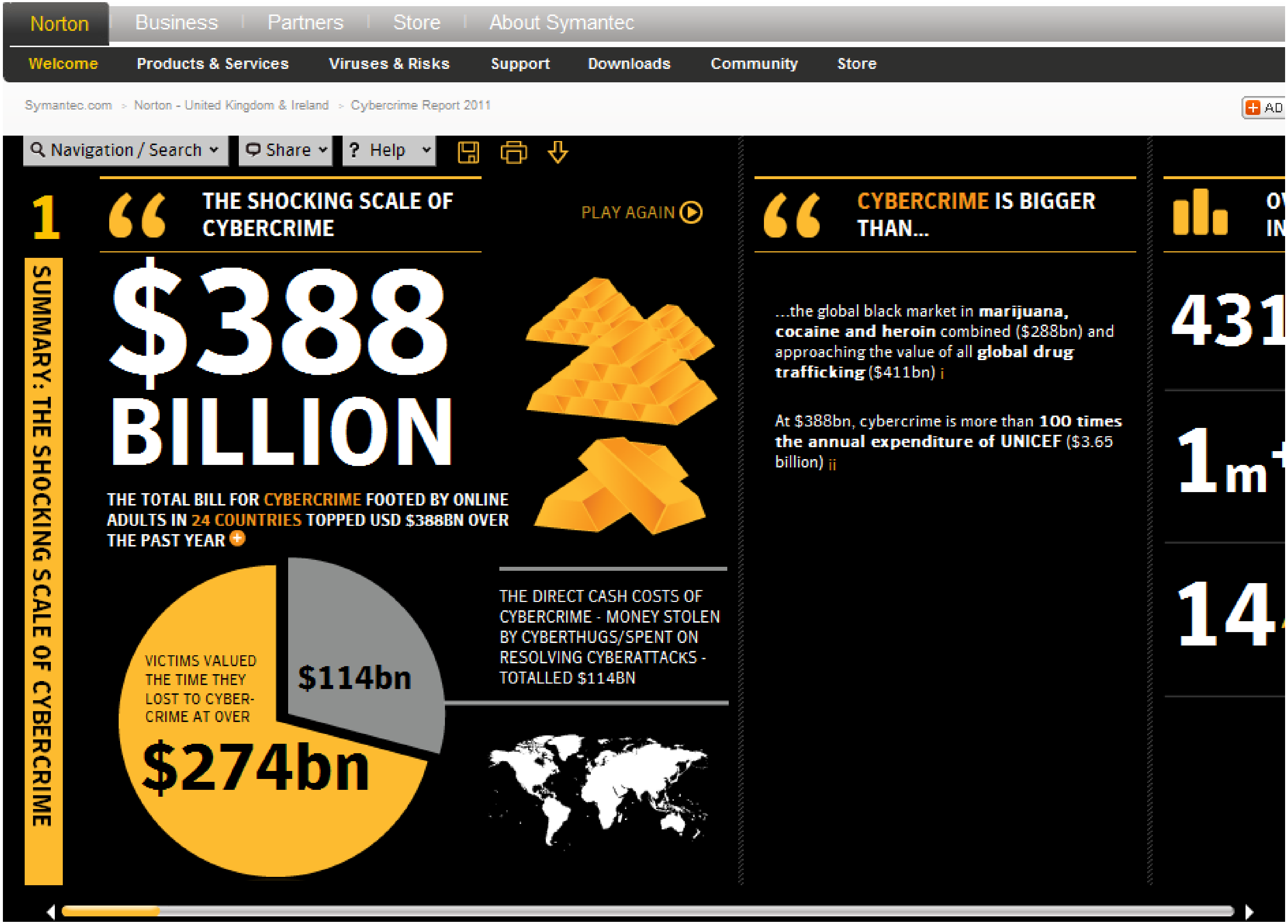
Lessons

Mean/median > 3 => tiny fraction of respondents account for majority of estimate

Lessons: Check Outliers

- Draw the graph of $x^2 + y^2 = 1$:





1 SUMMARY: THE SHOCKING SCALE OF CYBERCRIME

THE SHOCKING SCALE OF CYBERCRIME

$388 BILLION

THE TOTAL BILL FOR CYBERCRIME FOOTED BY ONLINE ADULTS IN 24 COUNTRIES TOPPED USD $388BN OVER THE PAST YEAR

VICTIMS VALUED THE TIME THEY LOST TO CYBER-CRIME AT OVER $274bn

$114bn

THE DIRECT CASH COSTS OF CYBERCRIME - MONEY STOLEN BY CYBERTHUGS/SPENT ON RESOLVING CYBERATTACKS - TOTALLED $114BN

CYBERCRIME IS BIGGER THAN...

...the global black market in **marijuana, cocaine and heroin** combined ($288bn) and approaching the value of all **global drug trafficking** ($411bn) i

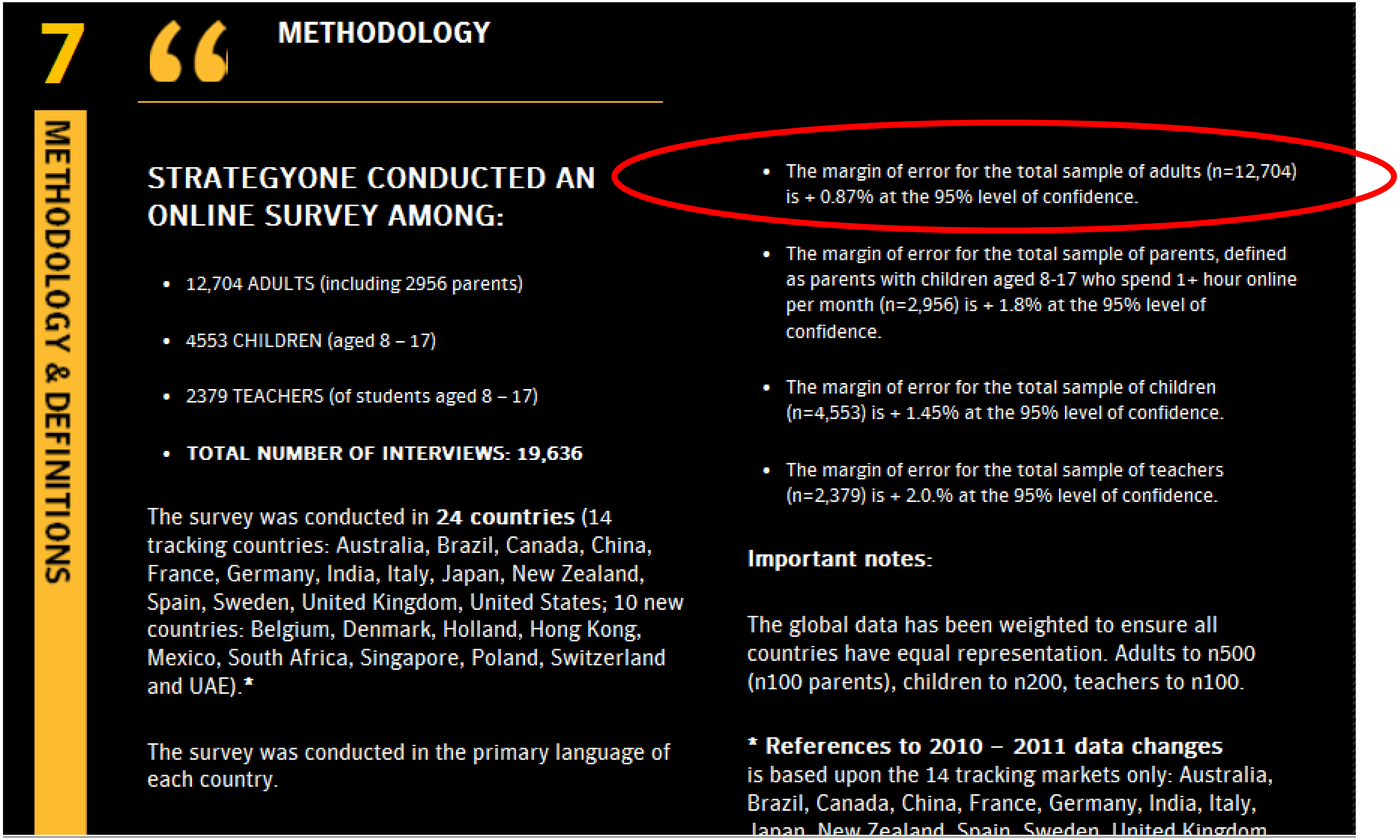
At $388bn, cybercrime is more than **100 times the annual expenditure of UNICEF** ($3.65 billion) ii

O
IN

431

1m

14



7

METHODOLOGY & DEFINITIONS

METHODOLOGY

STRATEGYONE CONDUCTED AN ONLINE SURVEY AMONG:

- 12,704 ADULTS (including 2956 parents)
- 4553 CHILDREN (aged 8 – 17)
- 2379 TEACHERS (of students aged 8 – 17)
- **TOTAL NUMBER OF INTERVIEWS: 19,636**

The survey was conducted in **24 countries** (14 tracking countries: Australia, Brazil, Canada, China, France, Germany, India, Italy, Japan, New Zealand, Spain, Sweden, United Kingdom, United States; 10 new countries: Belgium, Denmark, Holland, Hong Kong, Mexico, South Africa, Singapore, Poland, Switzerland and UAE).*

The survey was conducted in the primary language of each country.

- The margin of error for the total sample of adults (n=12,704) is + 0.87% at the 95% level of confidence.
- The margin of error for the total sample of parents, defined as parents with children aged 8-17 who spend 1+ hour online per month (n=2,956) is + 1.8% at the 95% level of confidence.
- The margin of error for the total sample of children (n=4,553) is + 1.45% at the 95% level of confidence.
- The margin of error for the total sample of teachers (n=2,379) is + 2.0.% at the 95% level of confidence.

Important notes:

The global data has been weighted to ensure all countries have equal representation. Adults to n500 (n100 parents), children to n200, teachers to n100.

*** References to 2010 – 2011 data changes** is based upon the 14 tracking markets only: Australia, Brazil, Canada, China, France, Germany, India, Italy, Japan, New Zealand, Spain, Sweden, United Kingdom

Confidence in this estimate?

- Let p = fraction of liars in population.
- Free of bias if:

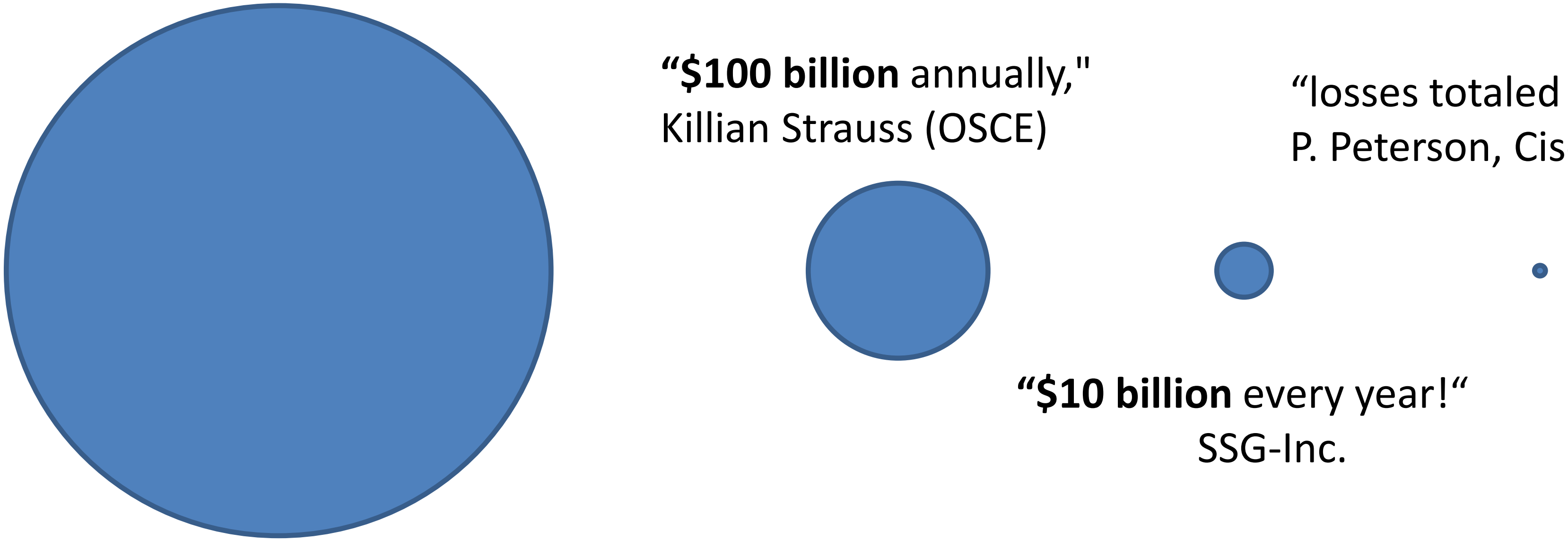
$$(1-p)^{19000} > 0.95$$

- Suppose p = 1%
  - 0.99^19636 ~ 1/10^85 = 0.00000000000000000000000000000000000000000000
    00000000000000000000000000000000000000001

- **Conclusion: you can have no confidence whatever that estimate represents the underlying phenomenon**

So what: How big is cyber-crime?

“Approximately **$1 trillion**,"
E. Amoroso, CSO, AT&T



“$100 billion annually,"
Killian Strauss (OSCE)

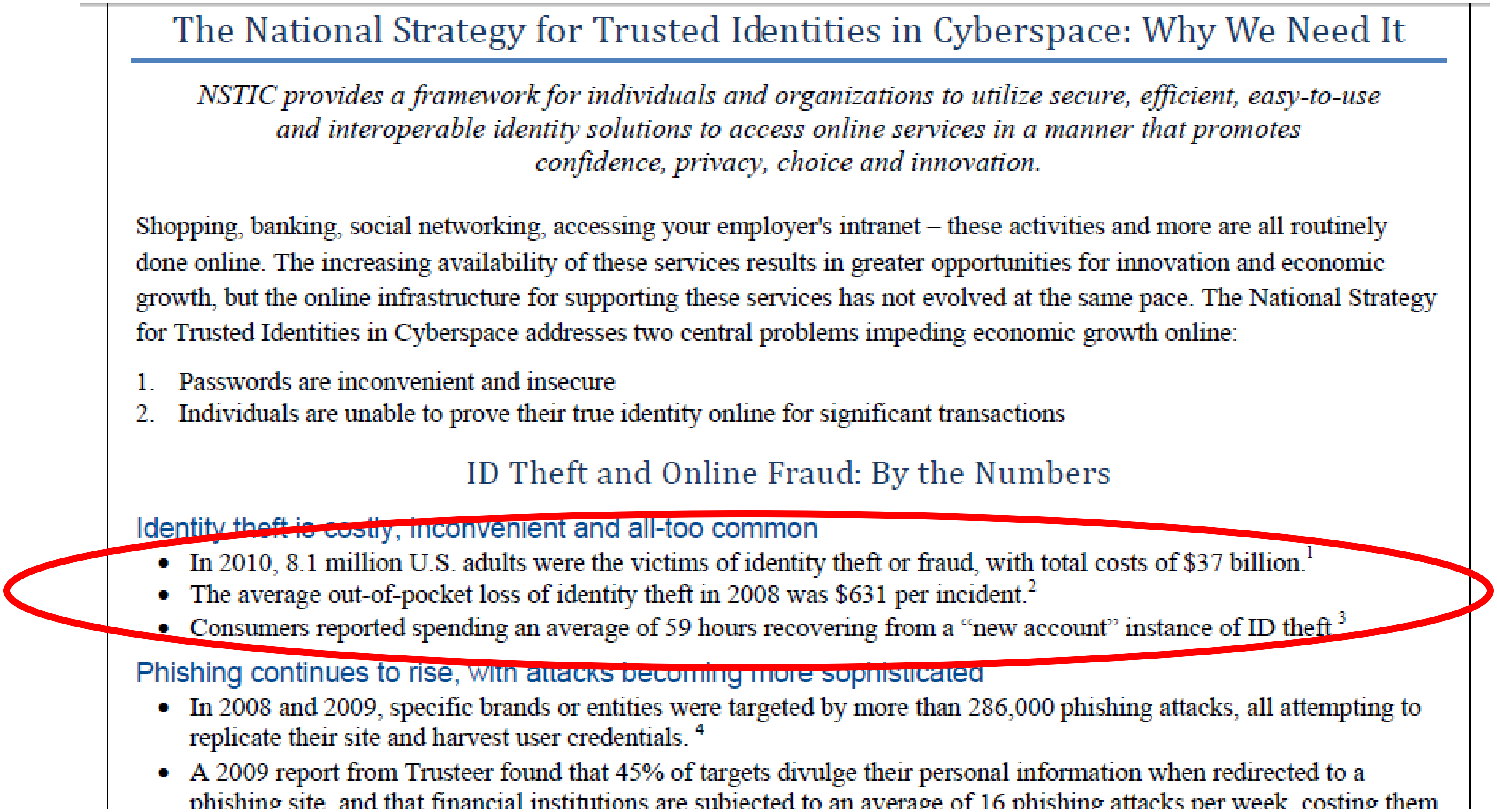
“losses totaled **$560m**,"
P. Peterson, Cisco Fellow

“$10 billion every year!“
SSG-Inc.

Are these people talking about the same thing?

$$\frac{\$1\ trillion}{\$560\ million} = \frac{GDP\ USA}{GDP\ Chad} = \frac{Internet\ Population\ 2011\ (2\ billion)}{Internet\ Population\ 1991\ (1.2\ million)}$$

So what?



The National Strategy for Trusted Identities in Cyberspace: Why We Need It

NSTIC provides a framework for individuals and organizations to utilize secure, efficient, easy-to-use and interoperable identity solutions to access online services in a manner that promotes confidence, privacy, choice and innovation.

Shopping, banking, social networking, accessing your employer's intranet – these activities and more are all routinely done online. The increasing availability of these services results in greater opportunities for innovation and economic growth, but the online infrastructure for supporting these services has not evolved at the same pace. The National Strategy for Trusted Identities in Cyberspace addresses two central problems impeding economic growth online:

1. Passwords are inconvenient and insecure
2. Individuals are unable to prove their true identity online for significant transactions

ID Theft and Online Fraud: By the Numbers

Identity theft is costly, inconvenient and all-too common

- In 2010, 8.1 million U.S. adults were the victims of identity theft or fraud, with total costs of $37 billion.[1]
- The average out-of-pocket loss of identity theft in 2008 was $631 per incident.[2]
- Consumers reported spending an average of 59 hours recovering from a "new account" instance of ID theft.[3]

Phishing continues to rise, with attacks becoming more sophisticated

- In 2008 and 2009, specific brands or entities were targeted by more than 286,000 phishing attacks, all attempting to replicate their site and harvest user credentials. [4]
- A 2009 report from Trusteer found that 45% of targets divulge their personal information when redirected to a phishing site, and that financial institutions are subjected to an average of 16 phishing attacks per week, costing them

- Hard to say how much something will improve if we don't know the size of the problem.

“You should never trust user input”

Writing Secure Code, Howard et al.

- Importance of input validation in security
  - SQL Injection
  - Buffer overflows

$$\frac{|\boldsymbol{X}|}{|\boldsymbol{R}|}\sum_{i\,\in\boldsymbol{R}} f[r_i] \qquad \equiv$$

```
total = 0.0;
for (i=0; i < survey_size; i++){
            total += (double) strcpy(user_input[i]);
}
estimate  = total * population_size/survey_size
```

- Practice unacceptable in writing code ubiquitous in forming estimates

“So you’re saying cyber-crime is no big deal?”

When you don’t know say you don’t know

We don’t have much idea of cyber-crime losses

Conclusions

- Response error can eclipse all other errors
  - Ratio mean/median > 3 is sign of concentration
  - Failure to disclose median => ignore
- This isn’t just a failure to achieve perfection
  - No confidence estimate reflects phenomenon
- Never trust user input