Looking For Truth Or At Least Data

Elizabeth D. Zwicky
zwicky@otoh.org

LISA 2009

Important Disclaimers

- All the numbers in this presentation are made up.
- The stories are true.
- I am not a statistician.
- I’m done with the funky transitions now.

Audience

- System Administrators
- Not statisticians
- Mostly collecting data about machines

- Numbers: good
- Believing appearances: bad
- Making stuff up: ??

What Am I Talking About?

- An attitude
- A hobby
- Where science, system administration, and security overlap

Fundamentals

- “That’s interesting. I wonder what I could find out about it?”
- Distinguish between “what appears to be” and “what is”.
- Understand numbers.

Why Might You Care?

- Planning systems and upgrades
- Troubleshooting
- Being good at security
- Just plain fun
- Not falling for pseudo-science

Recognizing Data

- Is this data?
- What is it data about?
- What conclusions can we draw from it?

Is This Data?

- “The CEO says the network is slow.”
- “47 users complained about network slowness yesterday.”
- “Average network latency yesterday was 15 milliseconds.”

Is This Data?

- "I feel like something might be wrong with a core router."
- "Brand A's router has an error rate 200% worse than Brand B."
- "Sites that use Brand A's router report slowness more often."

Is This Data?

- “We didn’t change anything around the time people started complaining about the network.”
- “We changed the routing just before people started complaining about the network.”
- “People are complaining because you changed the routing.”

Not Data

- Hearsay
- Numbers without context
- Conclusions

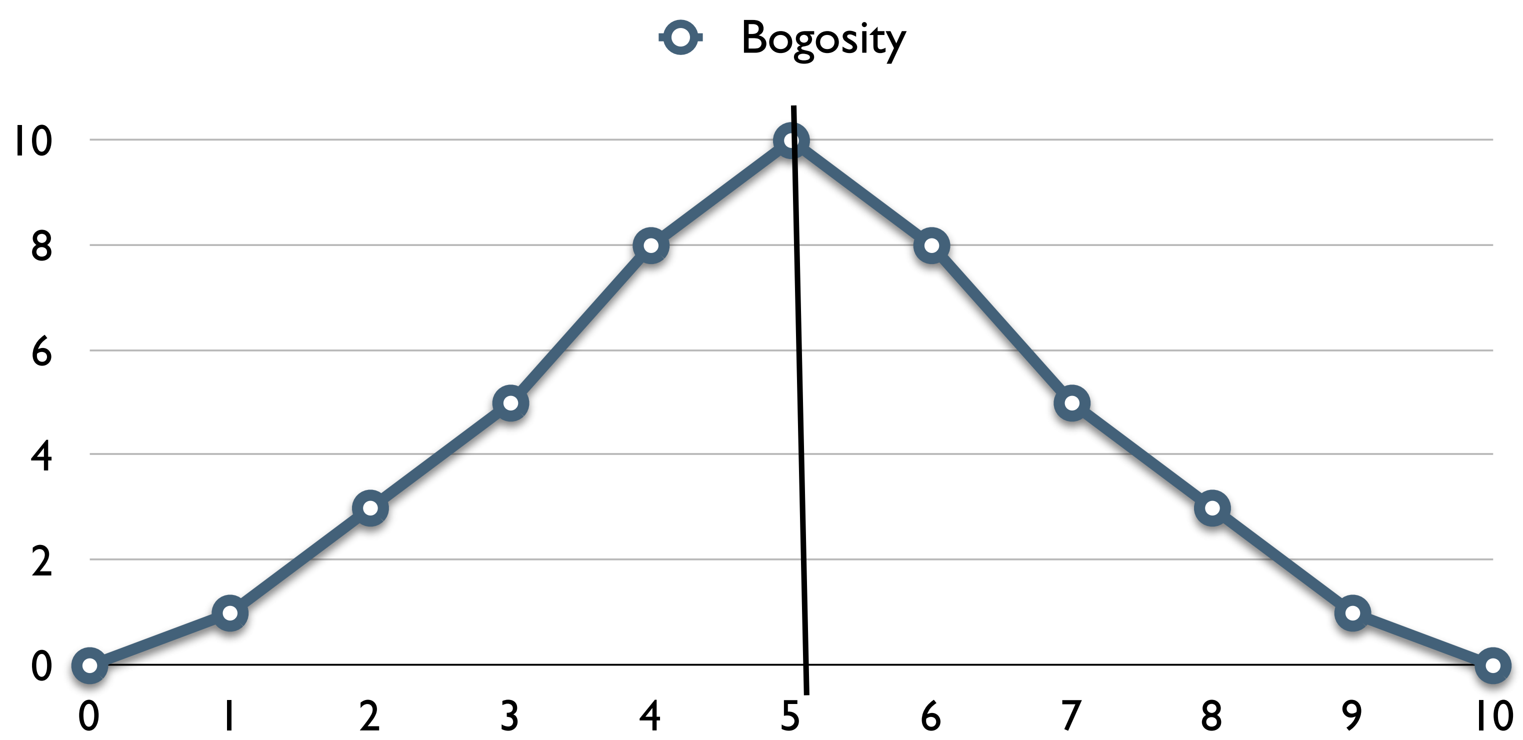
Data

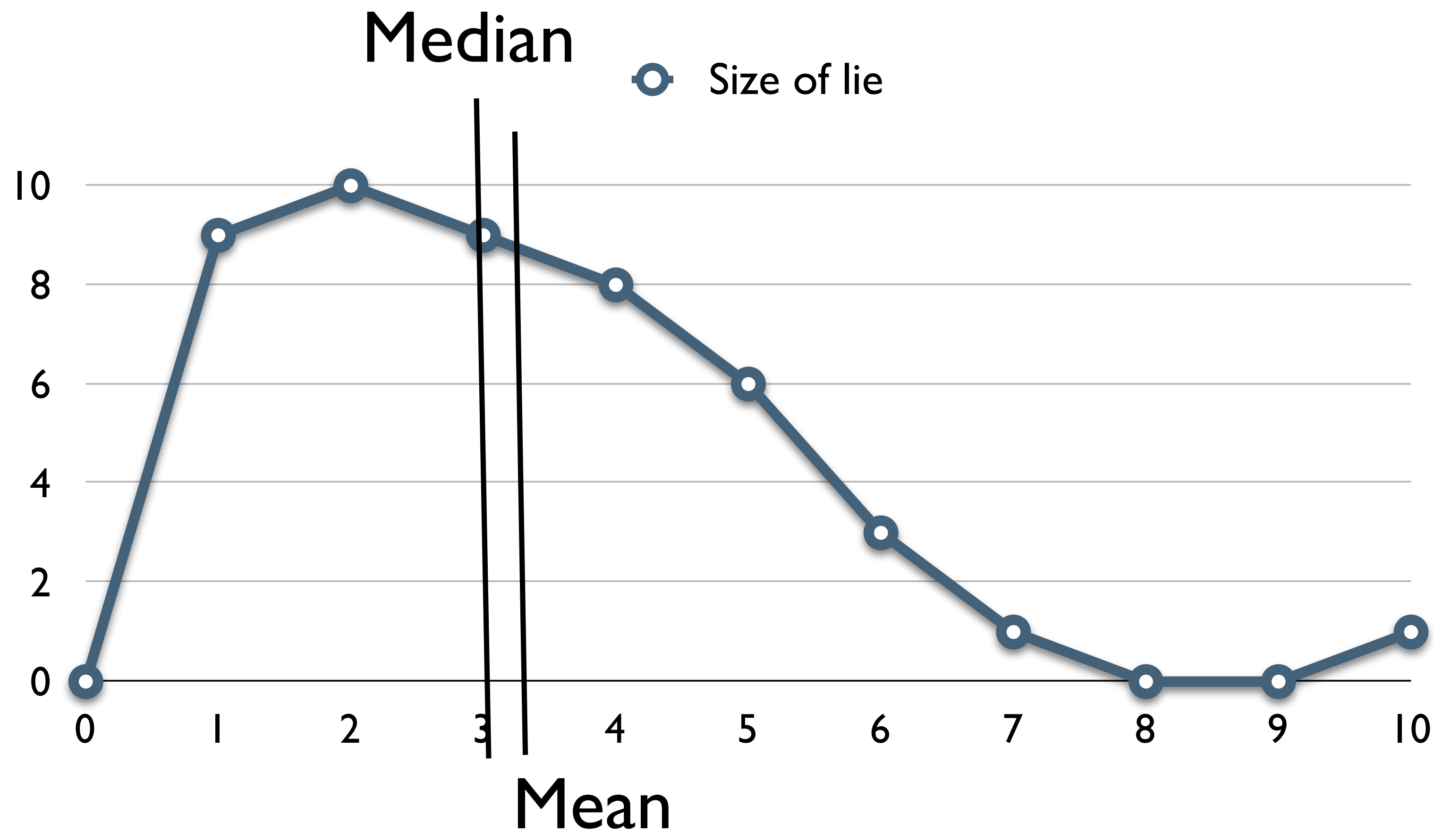
- Observations
- Self-report
- Numbers in context

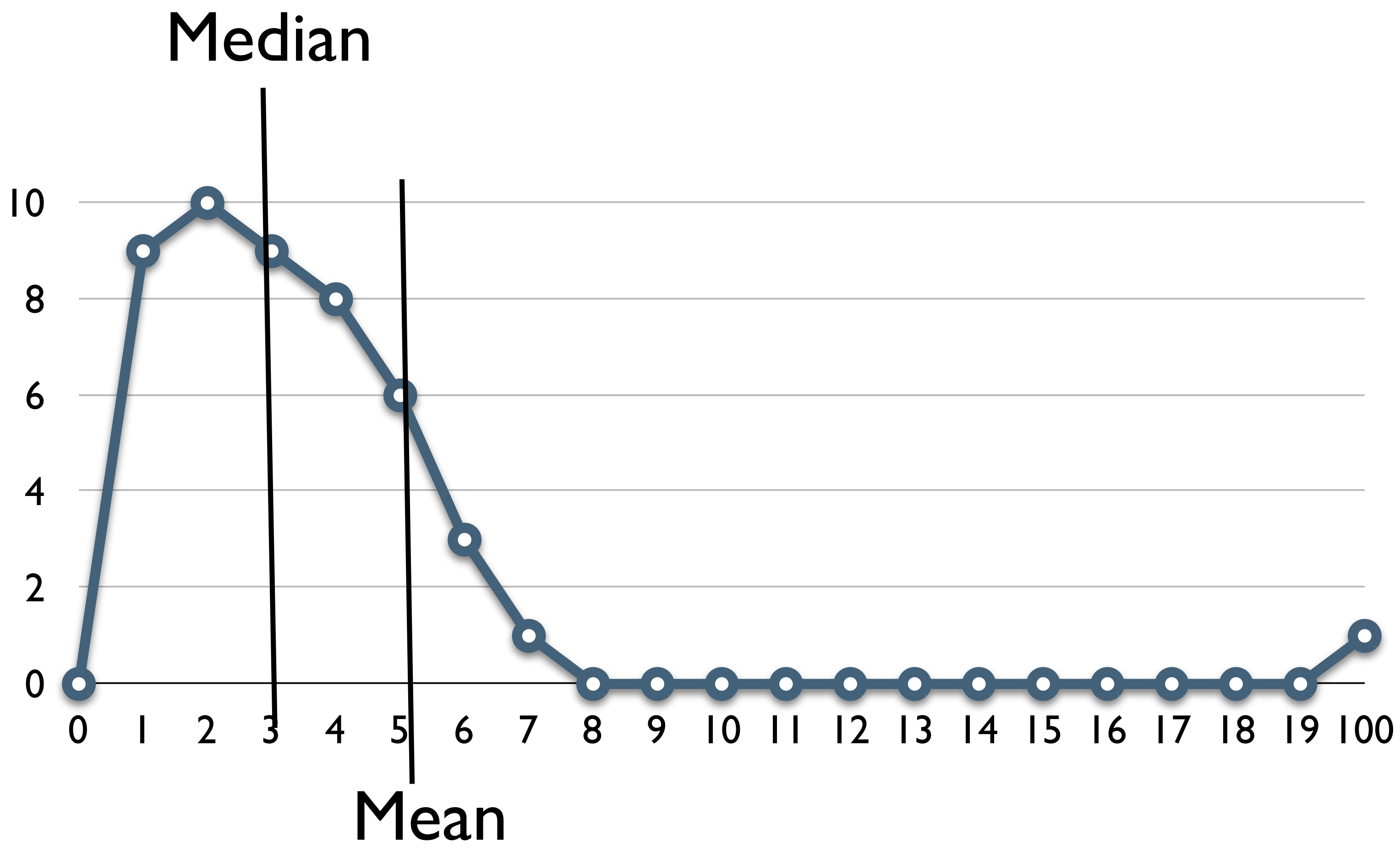
Why Those Numbers Aren’t Data

Basic Statistical Skepticism

- What do you mean “average”?
- Compared to what?
- What do you mean by “correlated”?


Bogosity
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10


Median
Size of lie
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10
Mean


Median
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
100
Mean

Average

- Means are only interesting for symmetrical single-peaked curves.
- Your data probably does not make one of them.
- You probably want median, quartiles, or percentiles.
- If you do want a mean, you want a standard deviation.

What Can You Do?

- Forget the average, look at a picture of the numbers.
- Ask what kind of average it is.
- Ask what the standard deviation is.

Compared to...

- Is 99.9% accuracy good?
- If your false positive rate on network packets is .1%, you get a false alarm every...
- And your false negative rate?

Better and Worse

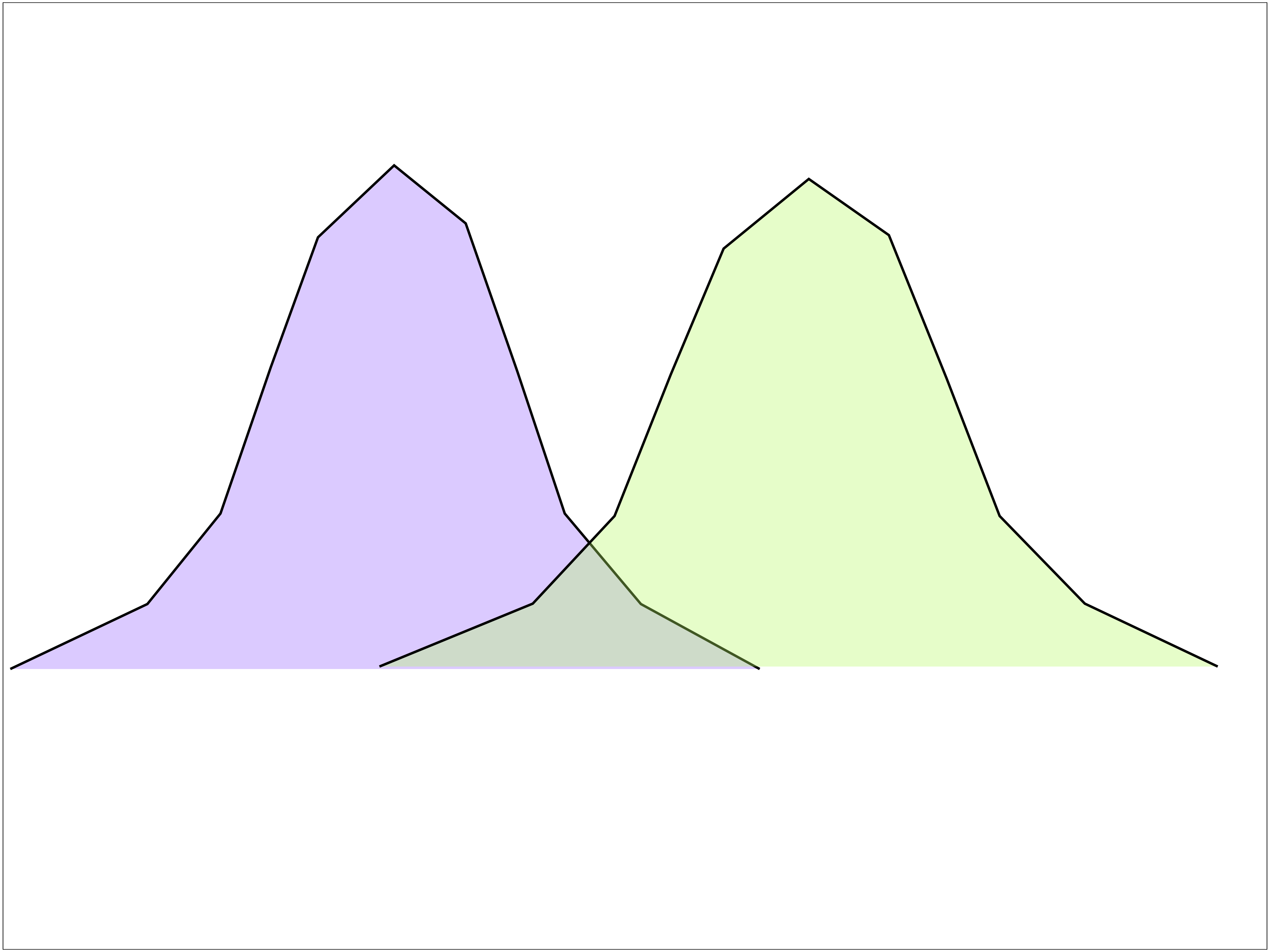
- Is a 200% increase in error rate bad?
- If your initial error rate was 1 in 4, your new error is 3 in 4.
- If your initial error rate was 1 in a million, your new error rate is 3 in a million.

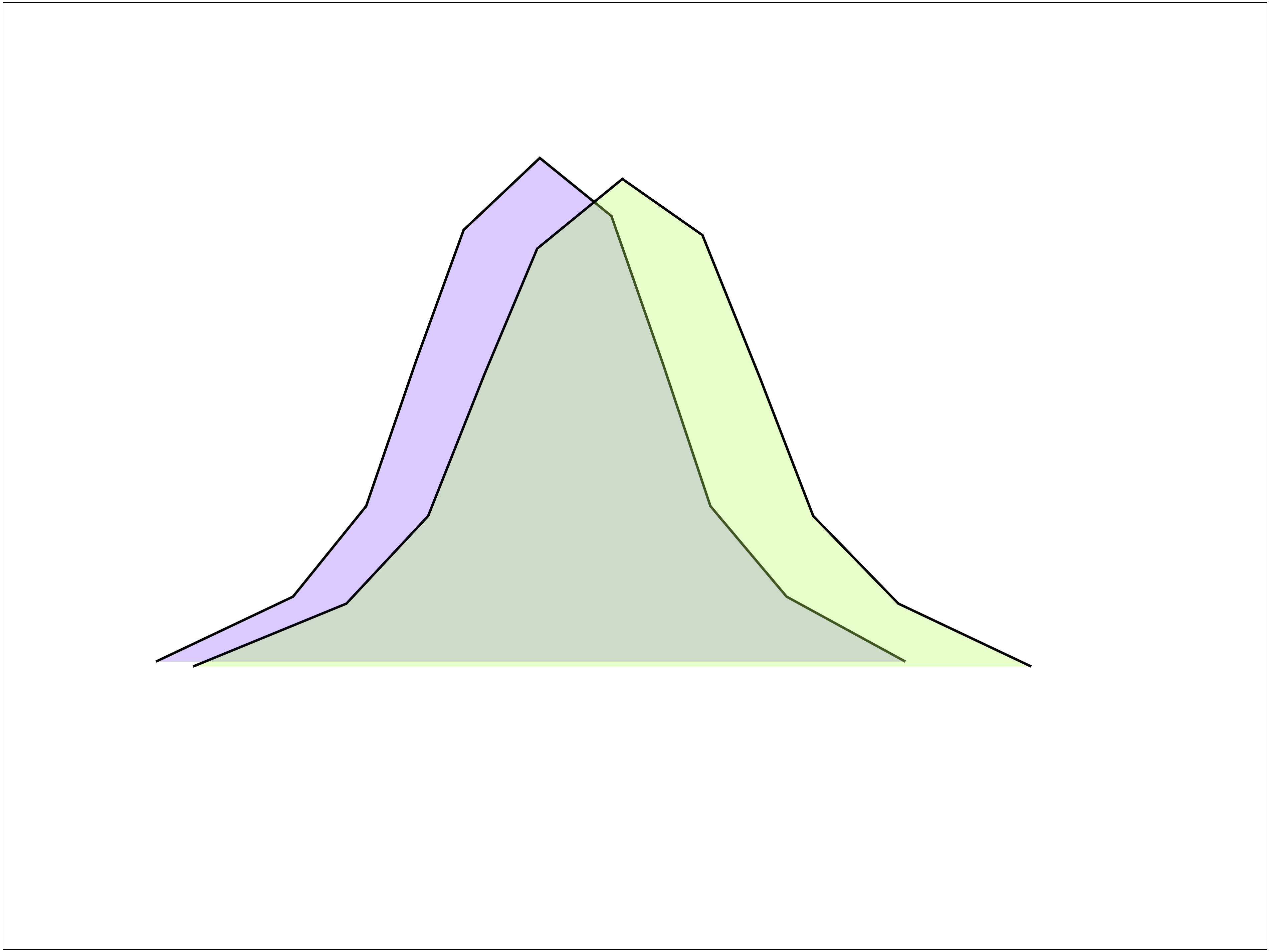
Error Rates Again

- Suppose both routers have the same error rate
- but one of them eats every millionth packet (random error)
- and the other eats every packet of a rare type (systematic error)

Correlations

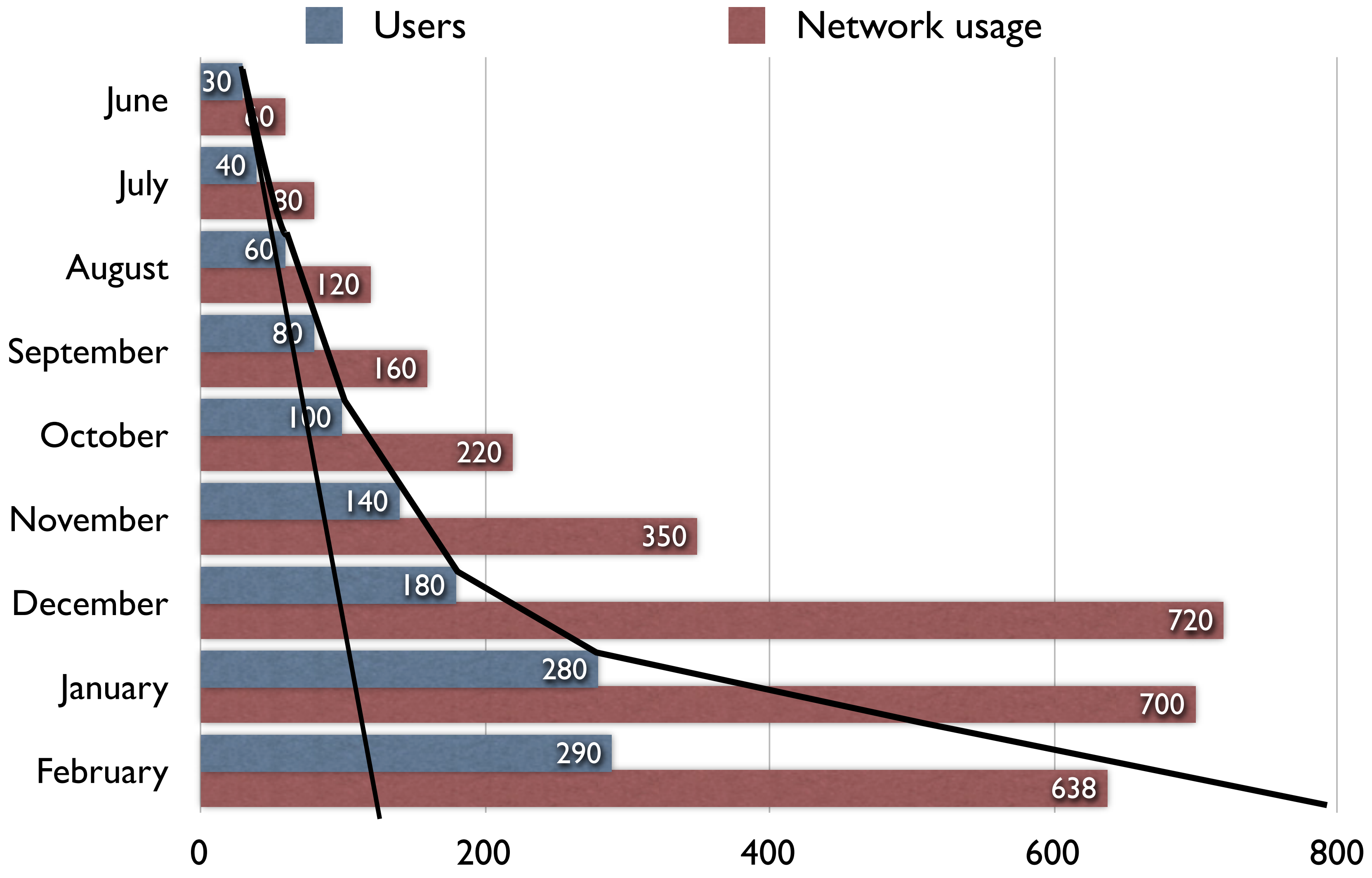
- “Sites that use Brand A routers are more likely to report slowness.”
- Correlation does not imply causation.
- Some correlations are weak.
- If you look at enough correlations, some of them will be “strong”.

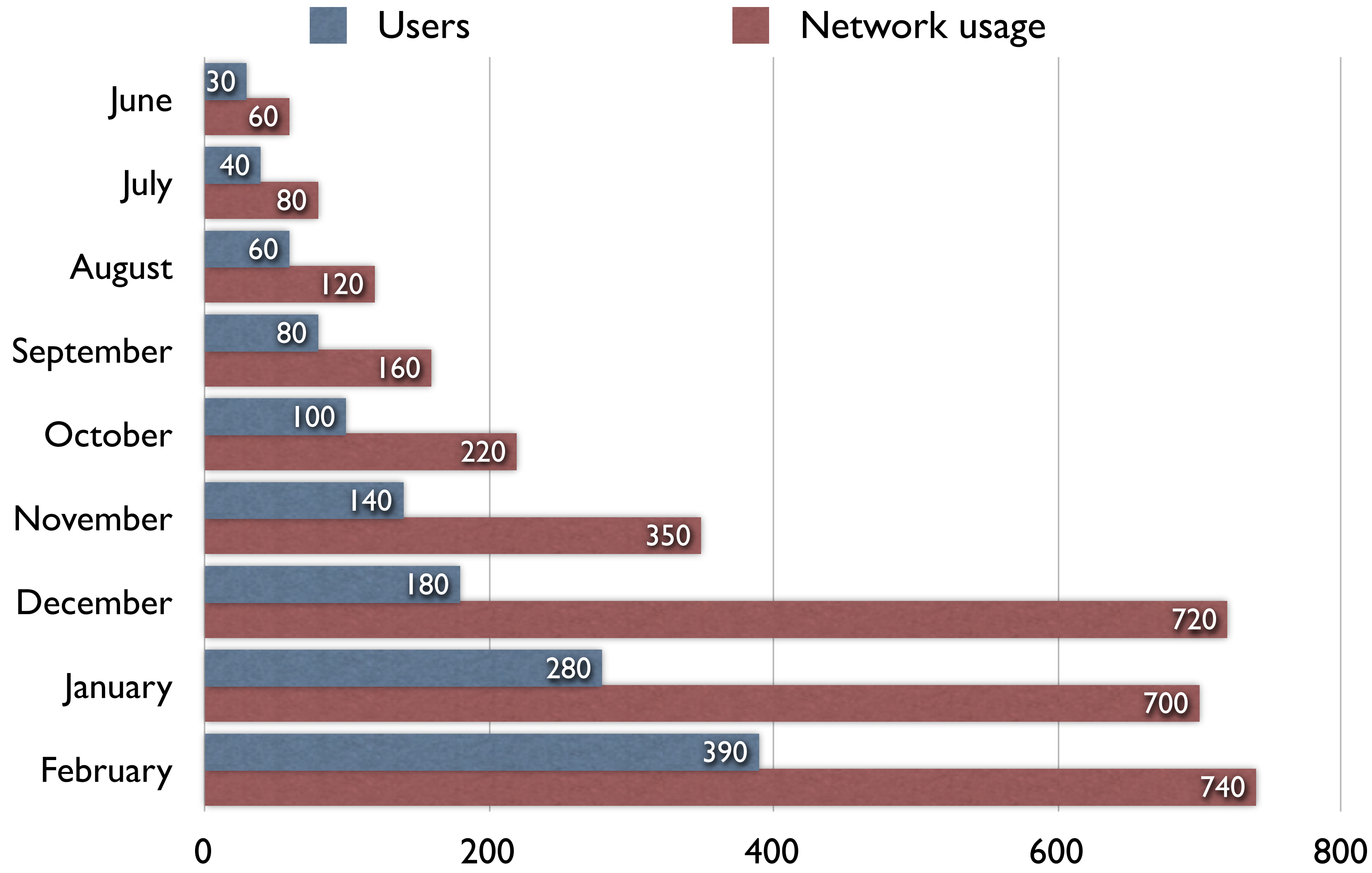




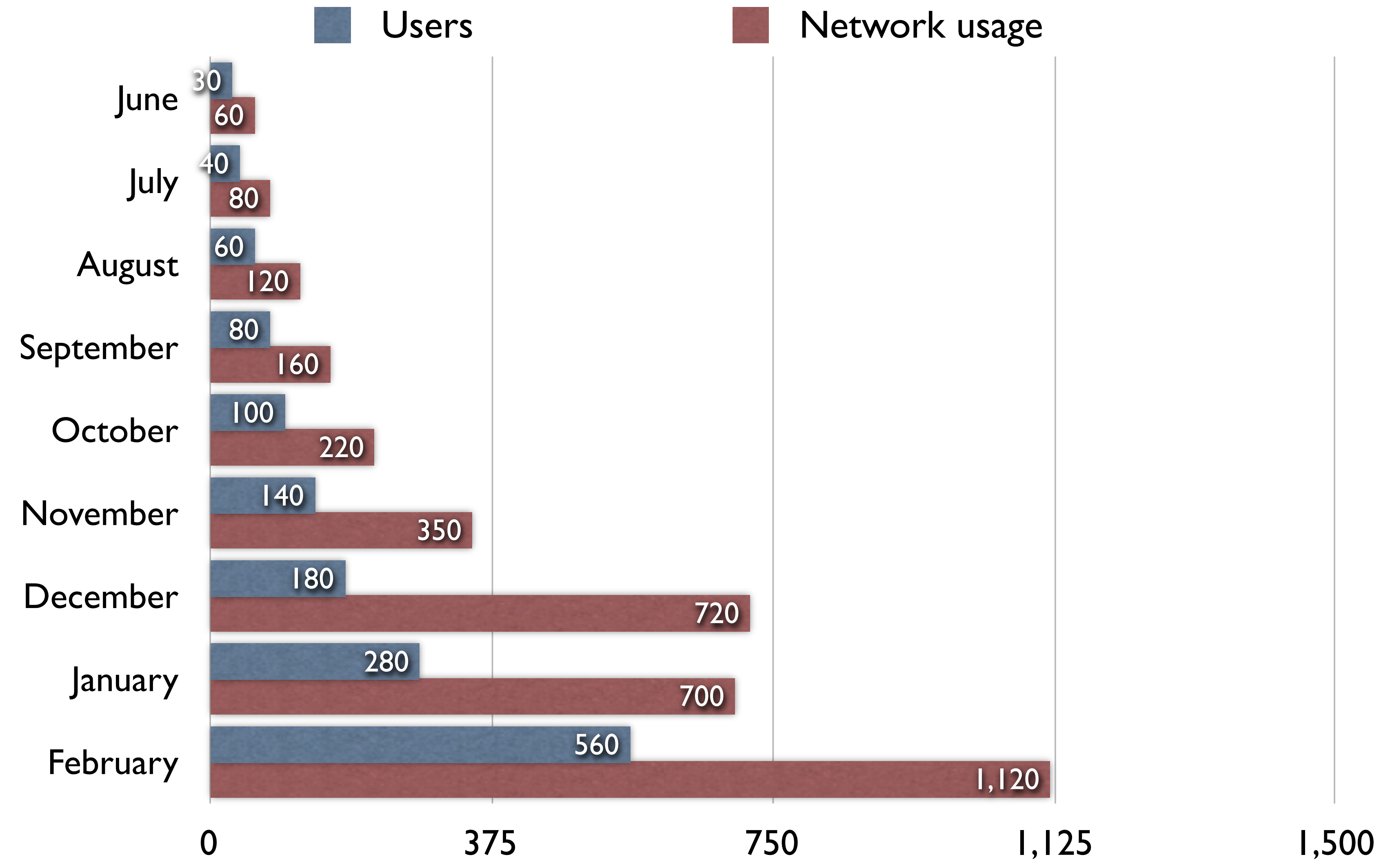
What Is It About?

- “47 users complained about network slowness yesterday”
  - is real data
  - about users
- “Network usage is increasing rapidly”


Users
Network usage
June
July
August
September
October
November
December
January
February
30
60
40
80
60
120
80
160
100
220
140
350
180
720
280
700
290
638
0
200
400
600
800



🟦 Users
🟥 Network usage
June
30
60
July
40
80
August
60
120
September
80
160
October
100
220
November
140
350
December
180
720
January
280
700
February
390
740
0
200
400
600
800


Users
Network usage
June
30
60
July
40
80
August
60
120
September
80
160
October
100
220
November
140
350
December
180
720
January
280
700
February
560
1,120
0
375
750
1,125
1,500

What Is It About?

- Most data is about lots of things
- The users are complaining it’s slow because
  - it’s slower
  - they changed applications
  - they’re unhappy

What conclusions?

- From the data I’ve shown:
    - Either your network will be overprovisioned most of the year, or December is going to be nasty.

What Conclusions?

- Data is a lot easier to find than truth.
- Be very cautious in the conclusions you draw from data.
- Correlation does not imply causation.

Gathering Data

Basic Tools

- A programming language, preferably one that's good with text.
- Some programs for looking at the guts of things.
- Some programs for making data into pictures.

Looking at Guts

- trace, dtrace, truss
- wireshark, tcpdump
- Windows sysInternals

Making Data into Pictures

- Your favorite spreadsheet
- GraphViz
- gnuplot

Basic Knowledge

- Regular expressions
- SQL
- XML
- Basic statistics

Finding Data

- Mine existing sources
- Collect data
- Simulate and/or extrapolate
- Find somebody else with data
- Make stuff up

Mine Existing Data

- How many files have we got? Count them.
- What are people's names like? Look them up.
- Those log files must be good for something

Collect Data

- Add logging
- Save snapshots of changing data
- Use tracing or network sniffing
- Run tests

Simulate and/or Extrapolate

- Set up a test situation
- Find a similar situation
- And then go back to mining or collecting data

Find Somebody Else With Data

- Published sources
- Friends and colleagues
- Get the rawest available data
- Know as much about it as possible

Make Stuff Up

- If all else fails, try guessing
- Get a lot of guesses
- Base guesses on knowns as much as possible
- Play around to see how changing guesses changes outcomes

Backups

- How much data will a given backup scheme backup?
- Mining: pull data from existing backup system.
- Collection: record statistics by day
- Simulation: make up a model of how people behave, see how much data

Educating Users on Security

- Mining: What do people currently look for or read?
- Collection: What do they do with changed content?
- Research: What do we know about naive users and security?

Collecting Data About People

- Human Subjects Boards and ethics
- Random sampling is good
- If you can’t be right,
  - be qualitative instead of quantitative
  - be wrong lots of different ways
  - at least understand why you’re wrong

What Next?

- Maybe fascinating things will just jump out at you.
- Maybe you just need to ask “why”?
- Maybe you’re going to use that data.

Cuckoo's Egg

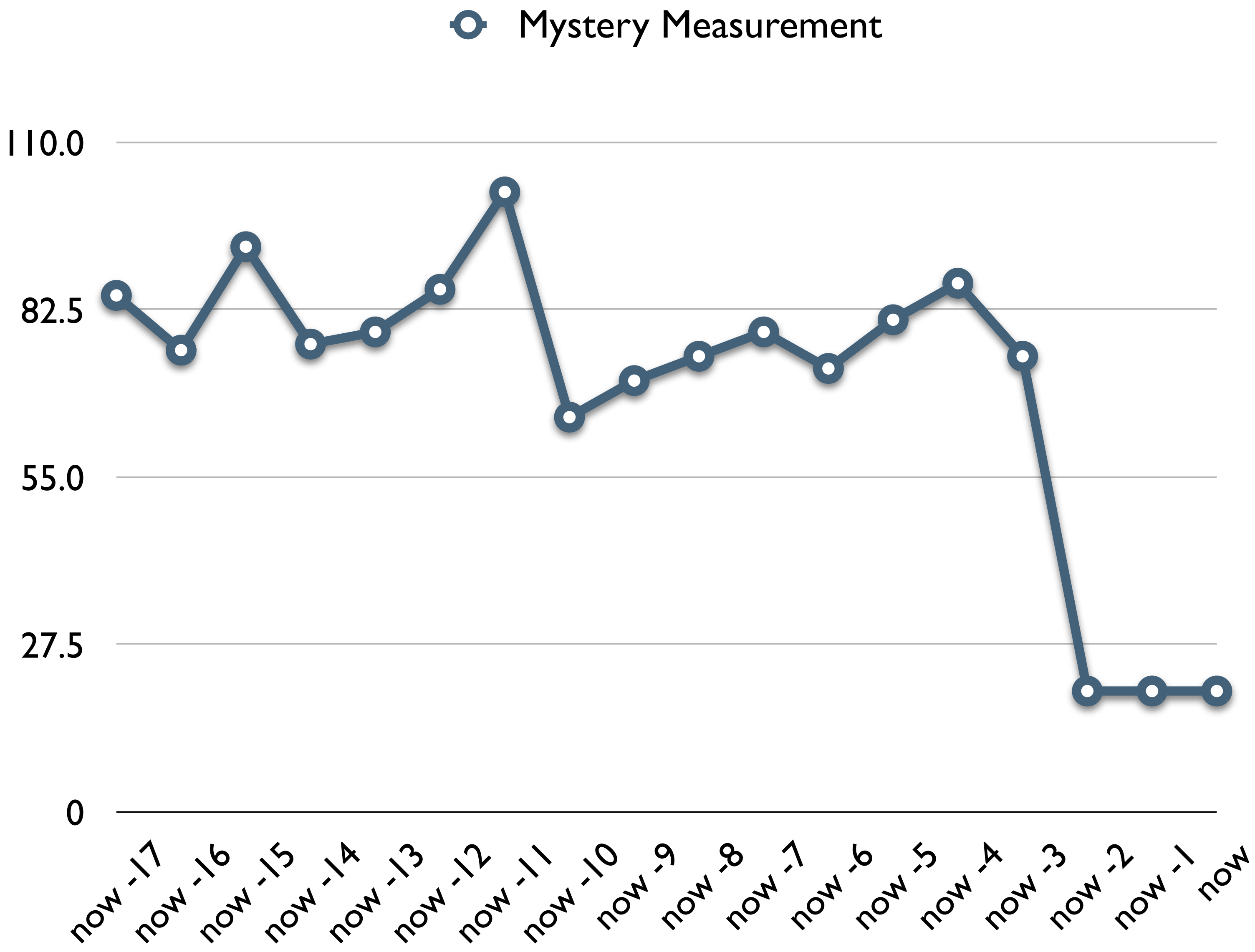
- Cliff Stoll tracks a quarter

Sanity Checking

- Another reason you might be asking "why"?
- Some data collection is wrong
- Some data collection reveals other problems

Analyzing Data

- Let the data lead you
- Know what questions you want to ask
- Humans are good at very specific sorts of pattern recognition


Mystery Measurement
110.0
82.5
55.0
27.5
0
now -17
now -16
now -15
now -14
now -13
now -12
now -11
now -10
now -9
now -8
now -7
now -6
now -5
now -4
now -3
now -2
now -1
now

Humans are Good At

- Noticing abrupt change
- Finding correlation
- Seeing faces

Humans are Bad At

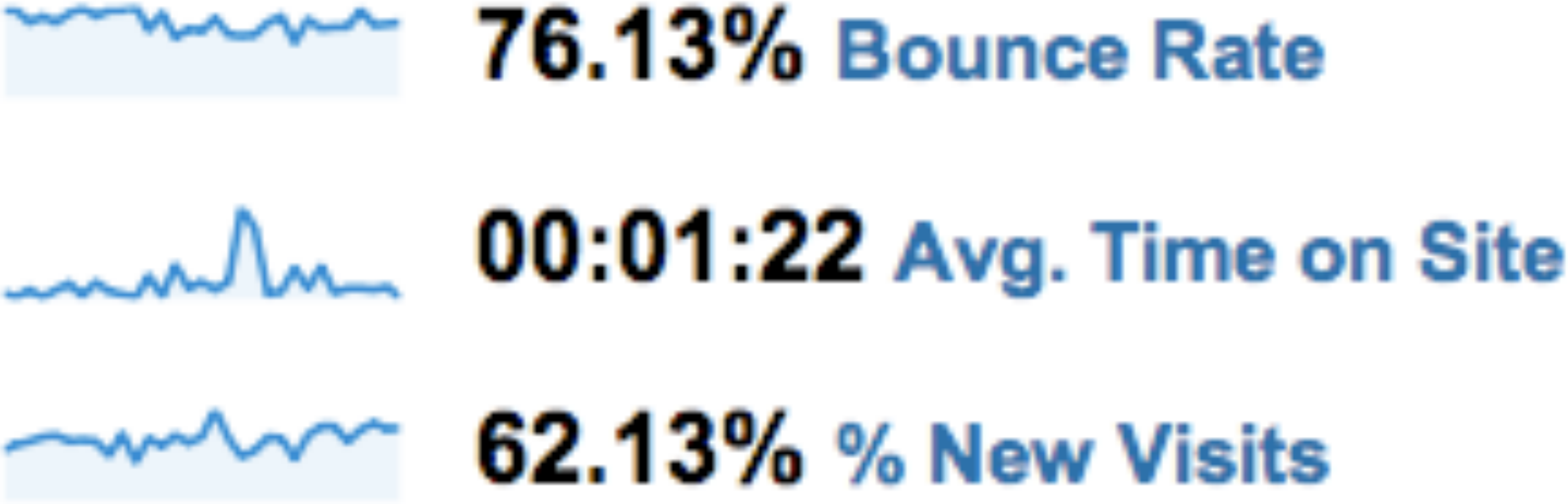
- Evaluating probability
- Finding non-correlation
- Perceiving slow change
- Perceiving correlation with time delay

Displaying Data

- Decide what you want to say
- Display that with only minimal other facts

Not Lying With Graphs

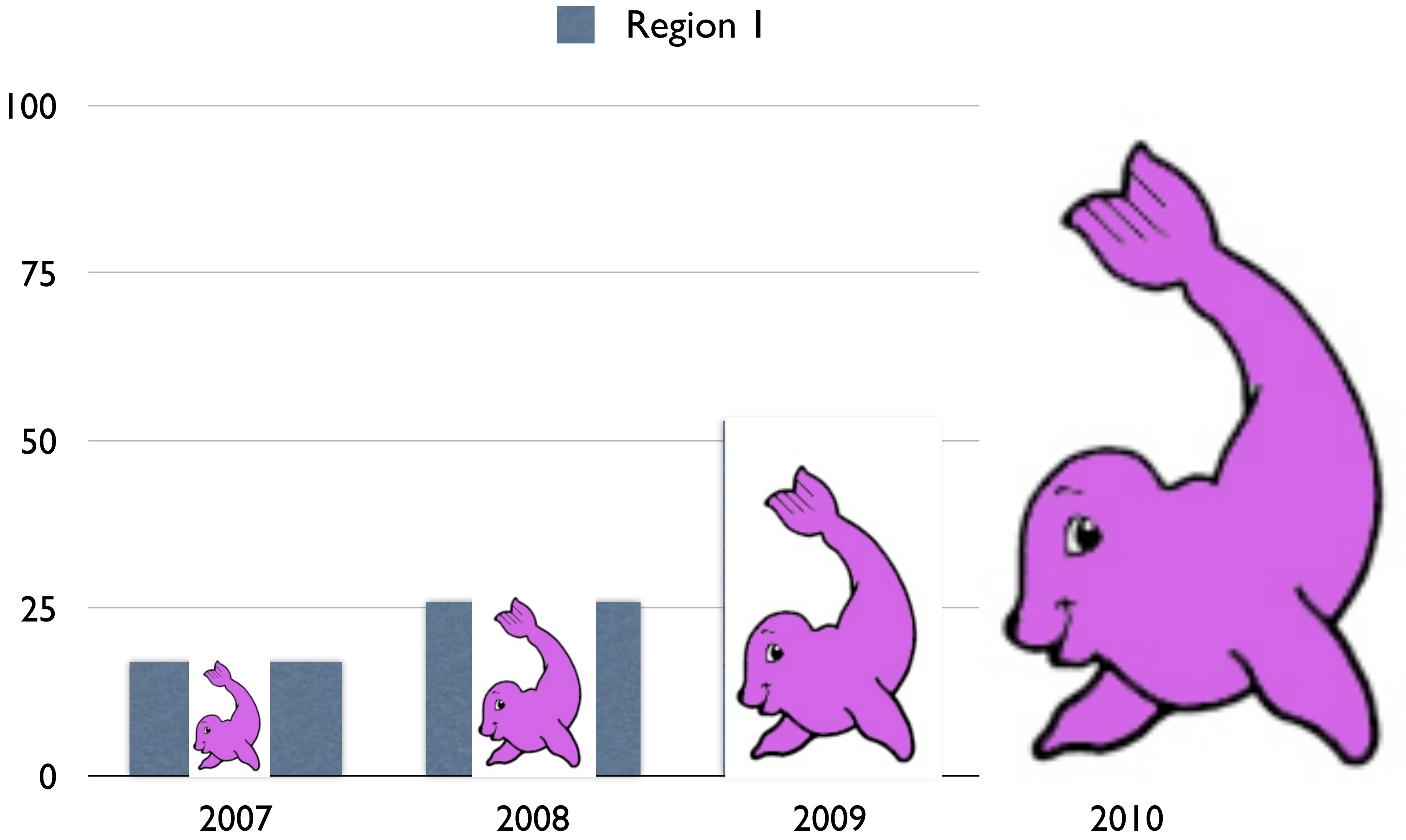
- Up is good, down is bad.
- Humans perceive area, but not well.
- Whenever possible, start at 0.

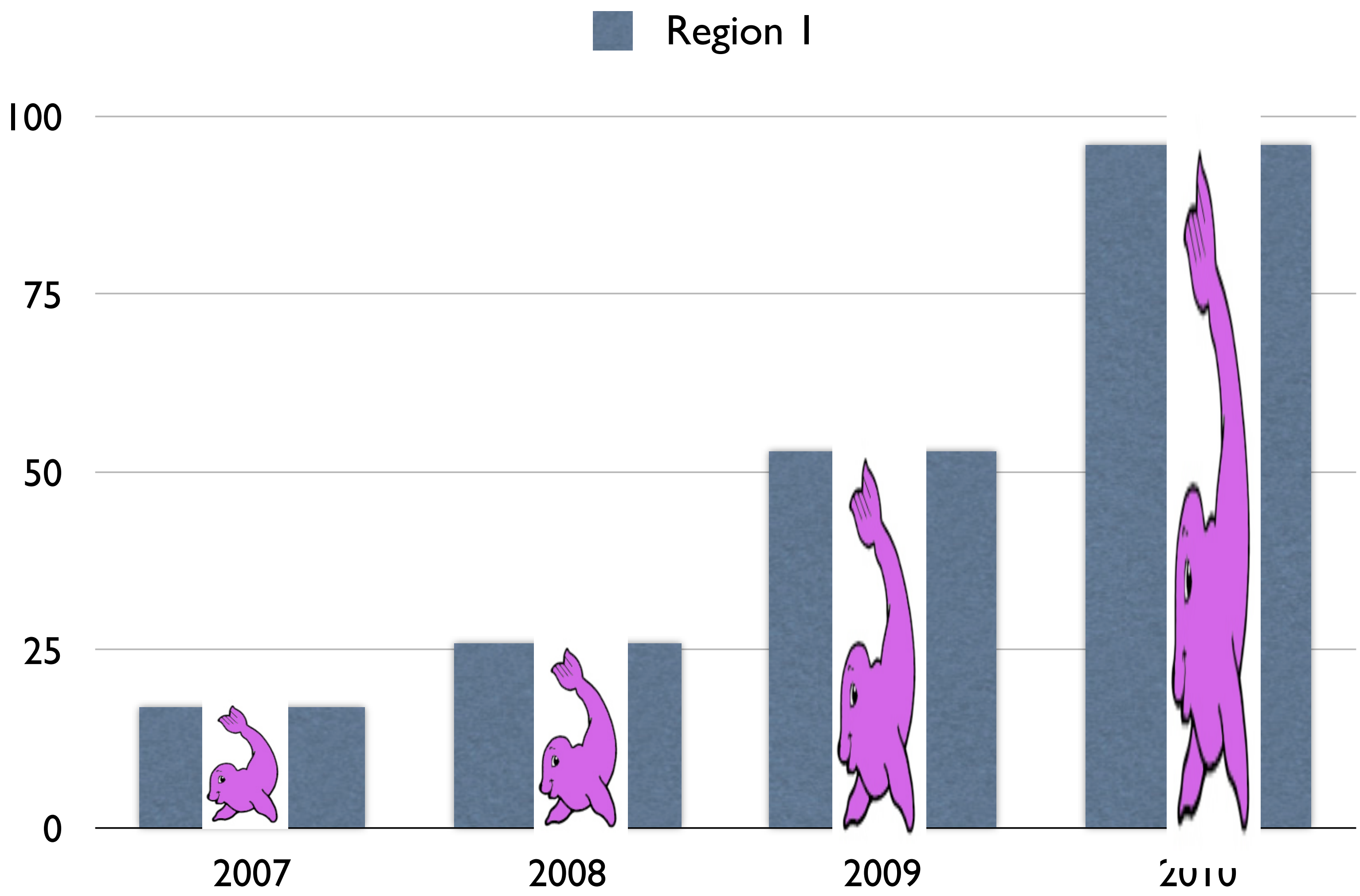


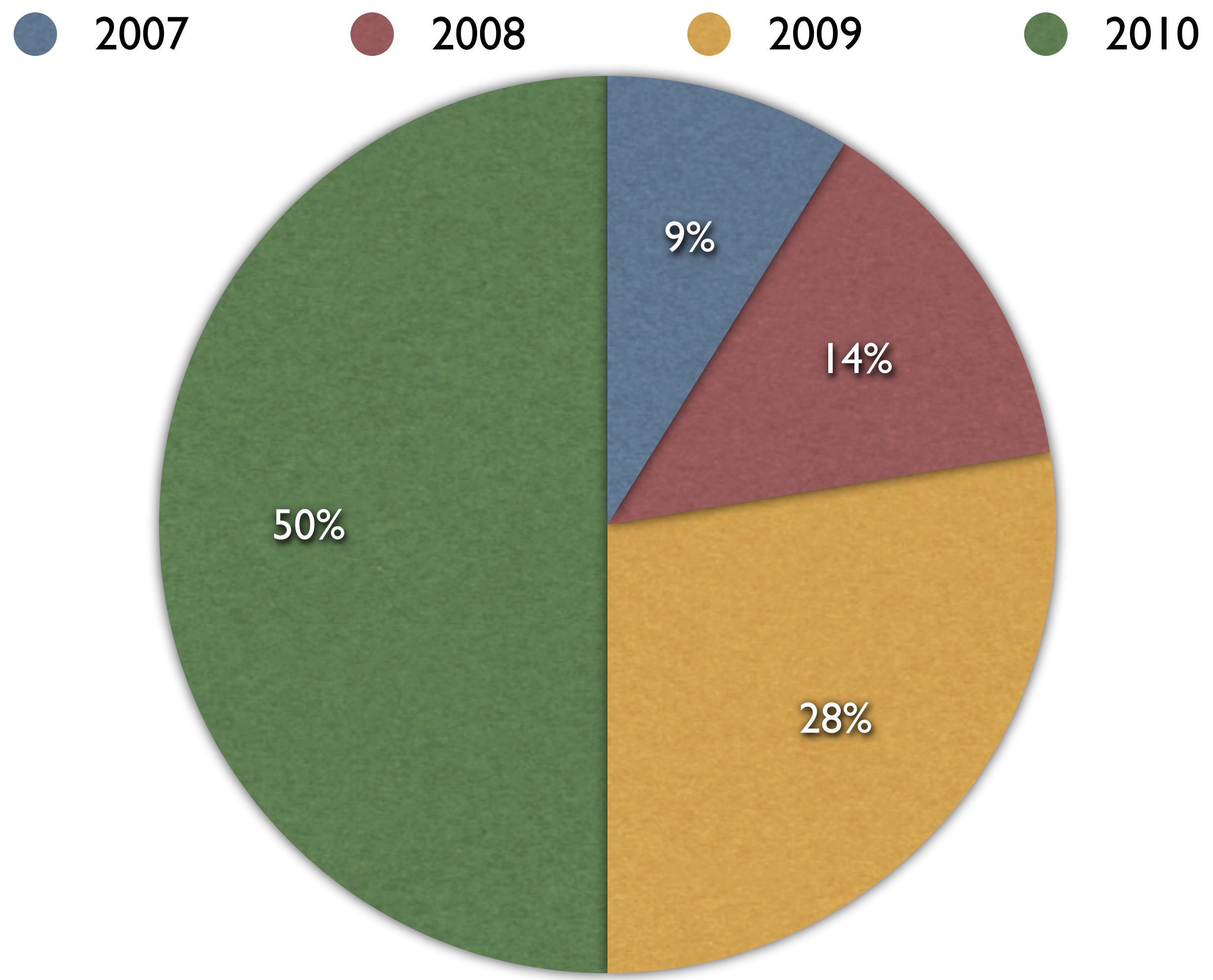
76.13% Bounce Rate

00:01:22 Avg. Time on Site

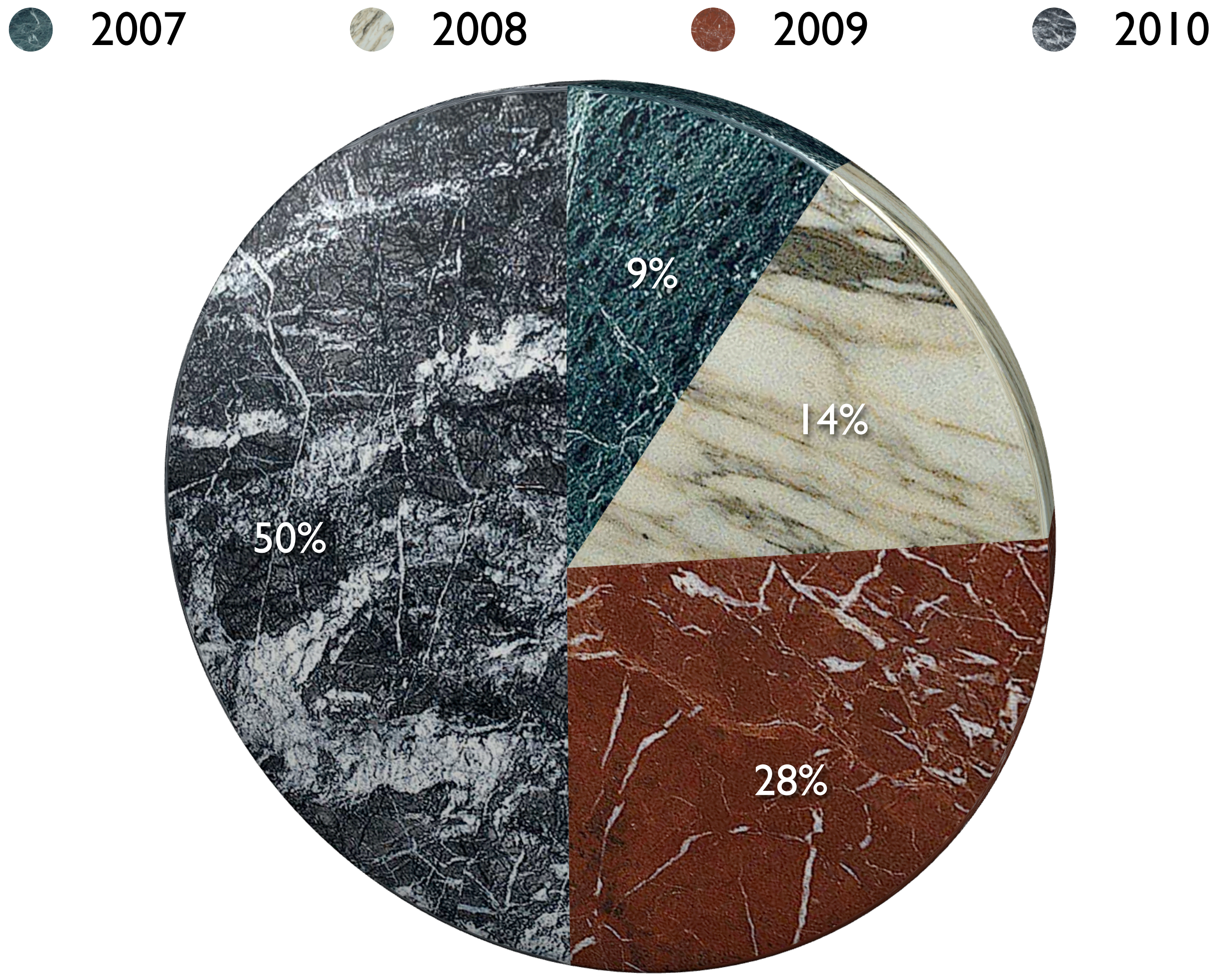
62.13% % New Visits

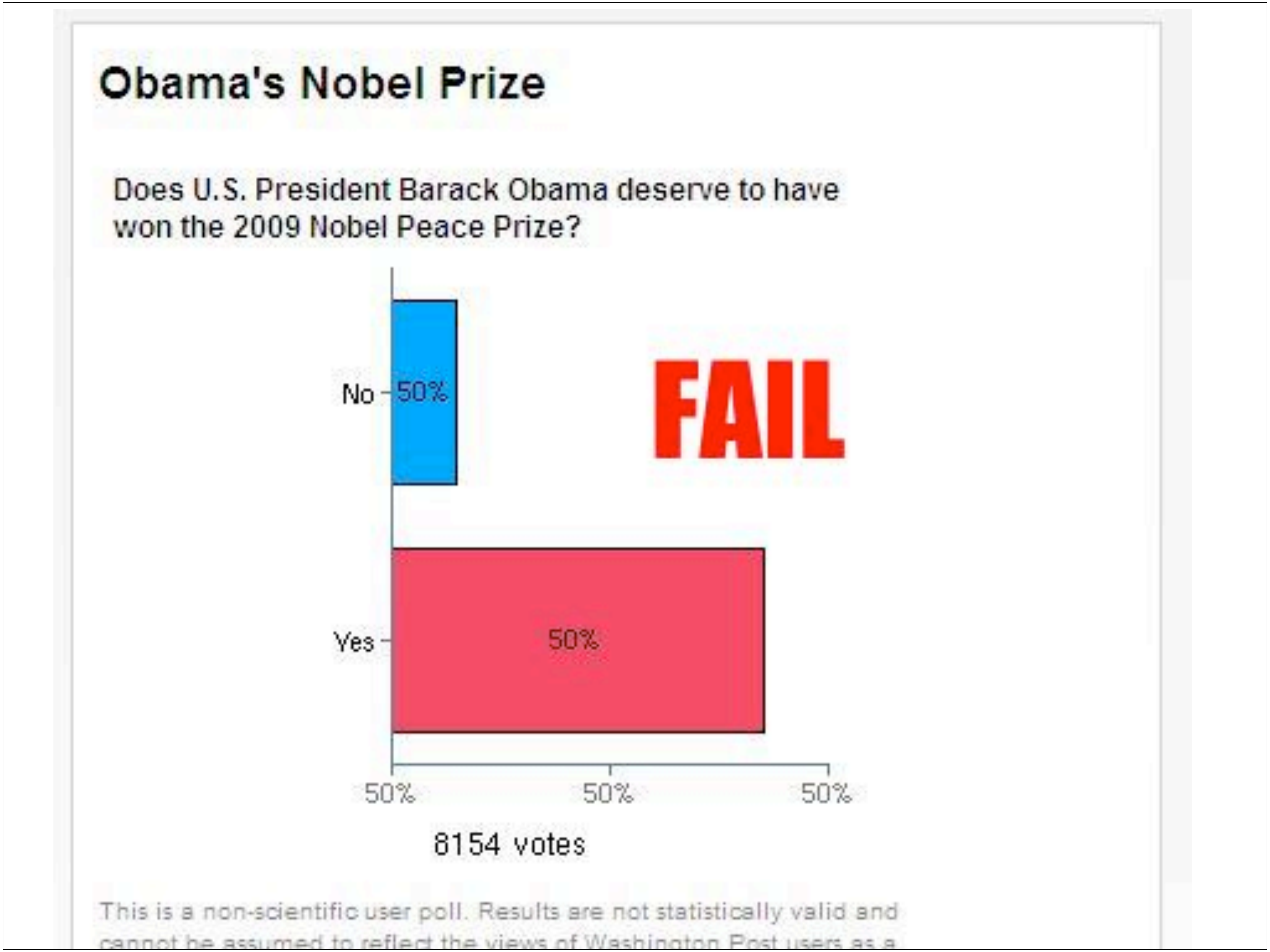

Region 1
100
75
50
25
0
2007
2008
2009
2010


Region 1
100
75
50
25
0
2007
2008
2009
2010



2007
2008
2009
2010
9%
14%
28%
50%


2007
2008
2009
2010
9%
14%
28%
50%



Obama's Nobel Prize

Does U.S. President Barack Obama deserve to have won the 2009 Nobel Peace Prize?

8154 votes

This is a non-scientific user poll. Results are not statistically valid and cannot be assumed to reflect the views of Washington Post users as a

A Complex Example

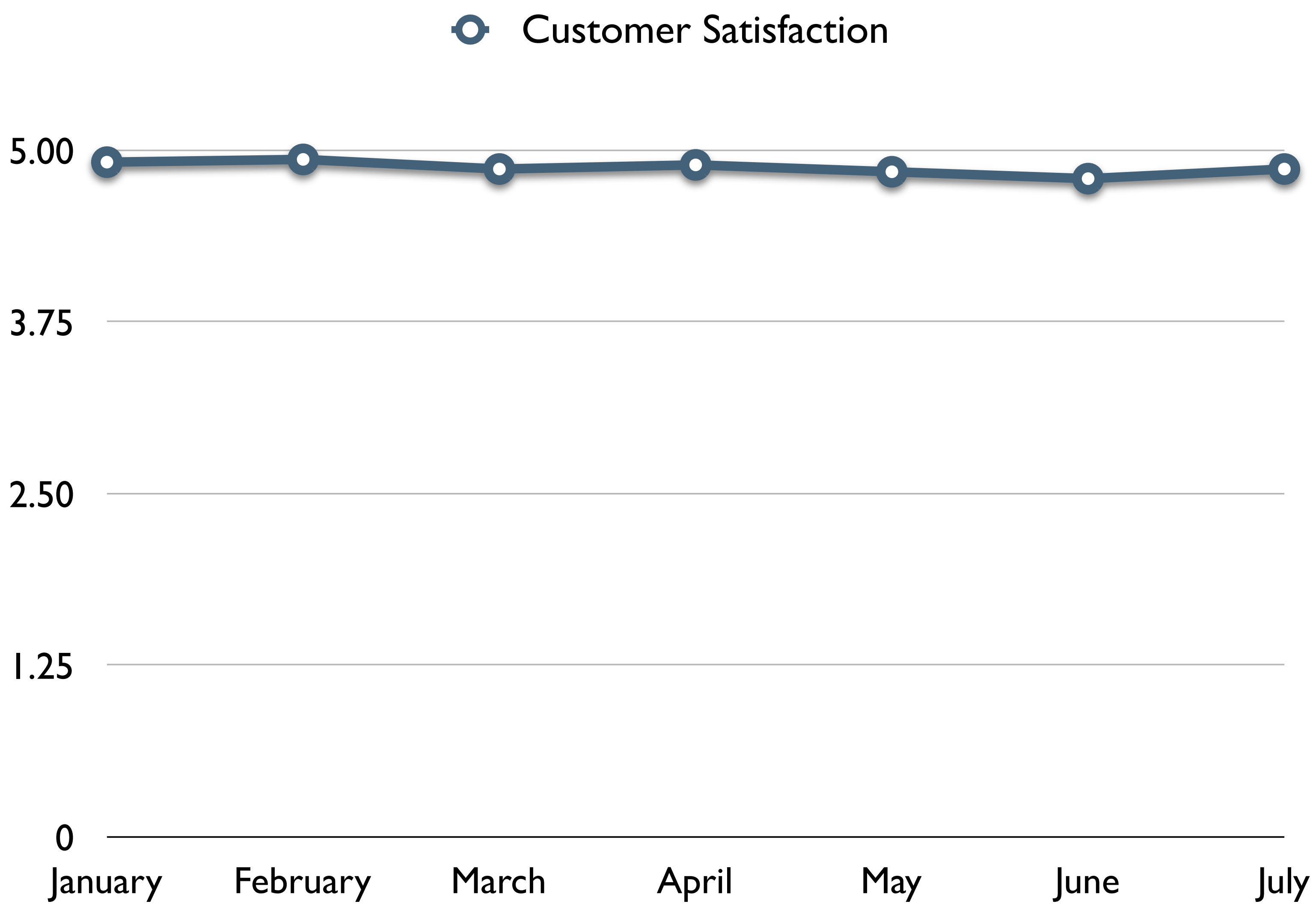
- Help desk performance
- Time to resolve == unhappy customers, unhappy partners
- Customer satisfaction?

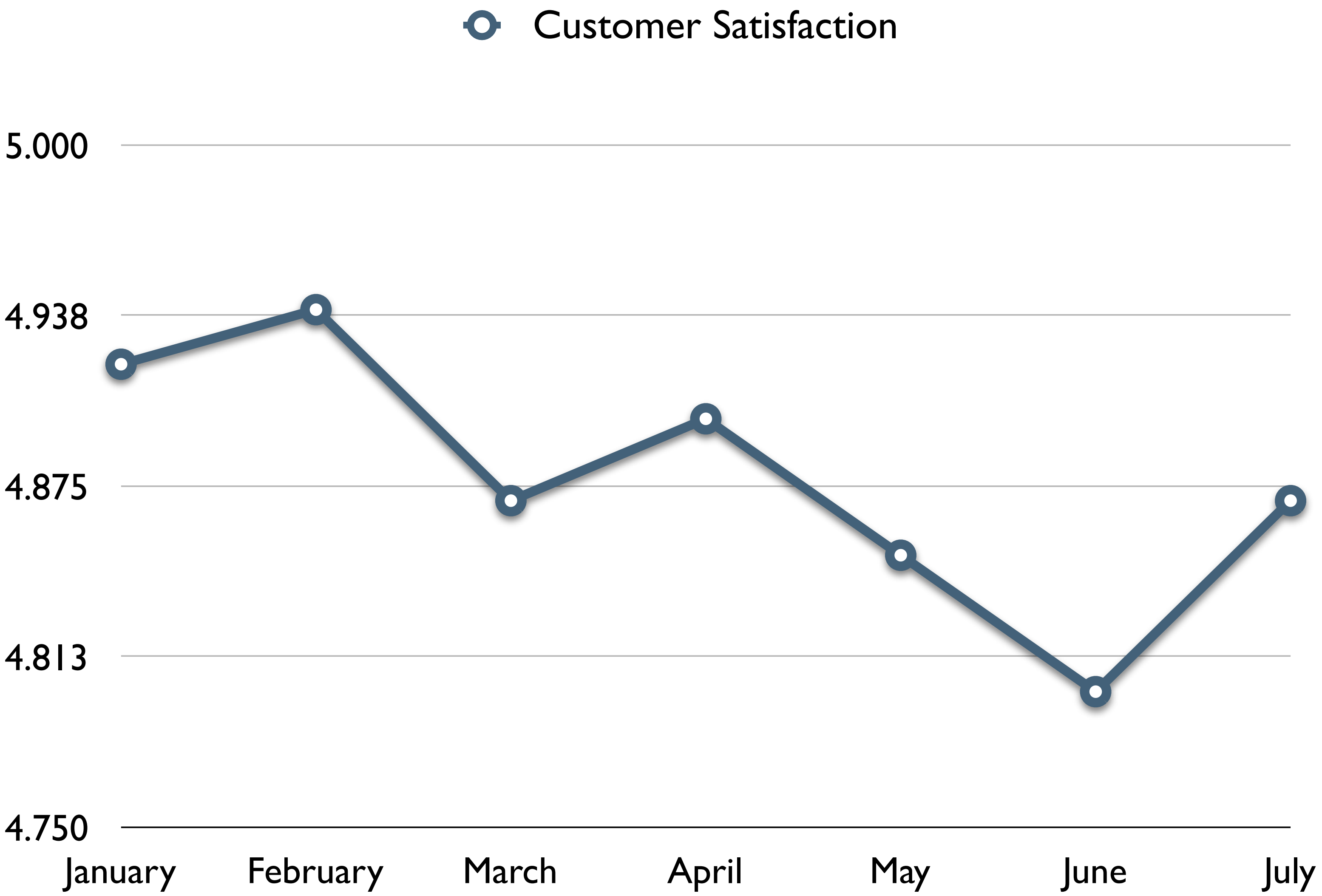
Customer Satisfaction

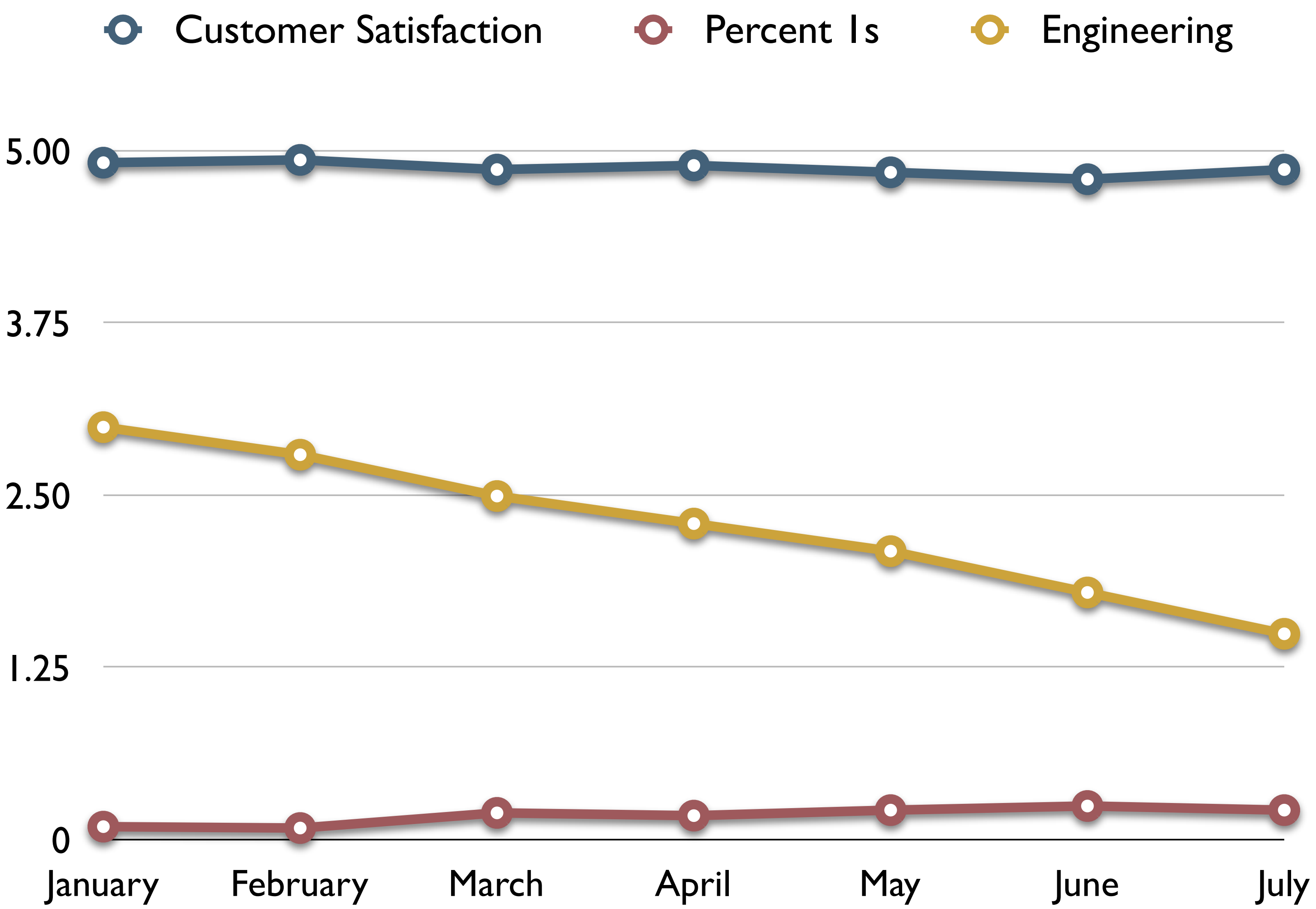
- Self-selected sample
- People who are especially unhappy or happy
- People who follow instructions

The Problem

- Help desk operators say users are unhappy
- Help desk management looks at numbers, says there's no problem


Customer Satisfaction
5.00
3.75
2.50
1.25
0
January
February
March
April
May
June
July


Customer Satisfaction
5.000
4.938
4.875
4.813
4.750
January
February
March
April
May
June
July


Customer Satisfaction
Percent 1s
Engineering
5.00
3.75
2.50
1.25
0
January
February
March
April
May
June
July

Most Relevant Books

- *Automating System Administration with Perl* by David Blank-Edelman
- *Visualizing Data* by Ben Fry
- *Data Crunching* by Greg Wilson

Classics

- *How to Lie With Statistics* by Darrell Huff
- *The Visual Display of Quantitative Information* by Edward Tufte

Background

- *Head First Statistics* by Dawn Griffiths
- *Predictably Irrational* by Dan Ariely
- *The Logic of Failure* by Dietrich Dörner

Blogs about data

- Junk charts: http://junkcharts.typepad.com/junk_charts/
- Chris Jordan: http://www.chrisjordan.com
- Chart Porn: http://chartporn.org/

Blogs that think this way

- Cognitive Daily: http://scienceblogs.com/cognitivedaily/
- Language Log: http://languagelog.ldc.upenn.edu
- Bad Science: http://www.badscience.net/

Elizabeth D. Zwicky
zwicky@otoh.org