Metadata

Steven M. Bellovin
http://www.cs.columbia.edu/~smb

What is "Metadata"?

- No precise definition
- Data "about" the primary data
- Roughly speaking, everything except the conversation itself
- For non-communications, often everything except the "essence" (whatever that is)

A Legal Intuition?

- Metadata (especially for a communication) is a message to or from another party
- Often, this message would not exist or would be drastically different if the same communication took place in a different form
- Example: dialed digits are a message to the phone company—you wouldn't use them when talking to someone in person, but you might say the same thing to the other party.
- Example: image metadata doesn't exist in hard-copy photographs
- The third party doctrine may (or may not) come into play

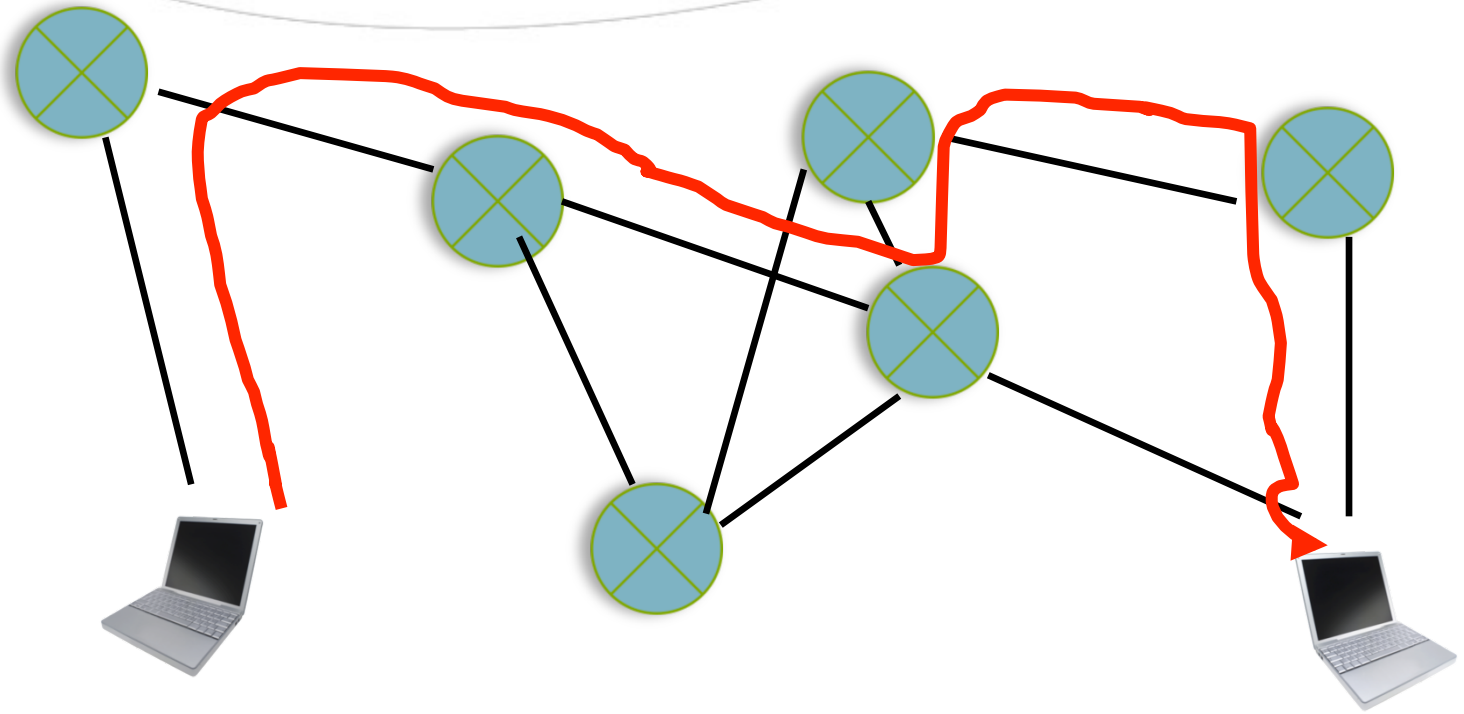
Historical Note

- An old, old technique, both for law enforcement and for intelligence
  - Called "traffic analysis" by the intelligence community
- Who talks to whom is very revealing
  - Military: learn "order of battle"

Telephony Metadata

- Caller's number and called number
- Start and end time of the call
- Start and end location (for mobile, this includes cell site, antenna "sector", approximate distance)
  - "Location" is controversial
- Device serial number (for mobile)

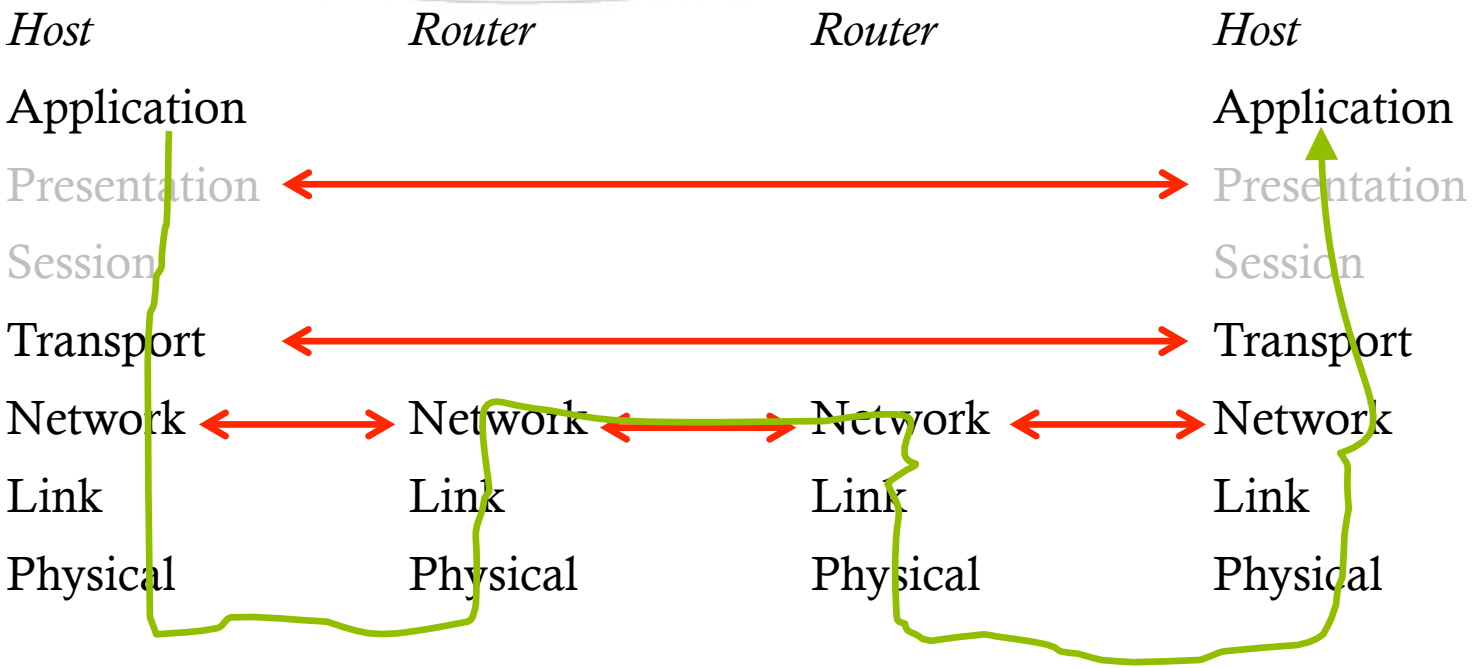
Routing Through the Internet



The Network Stack

7	Application	Email, Web, etc
6	Presentation	
5	Session	
4	Transport	TCP
3	Network	IP
2	Link	WiFi, Ethernet
1	Physical	Radio, fiber, etc.

Layers Talk to Peer Layers



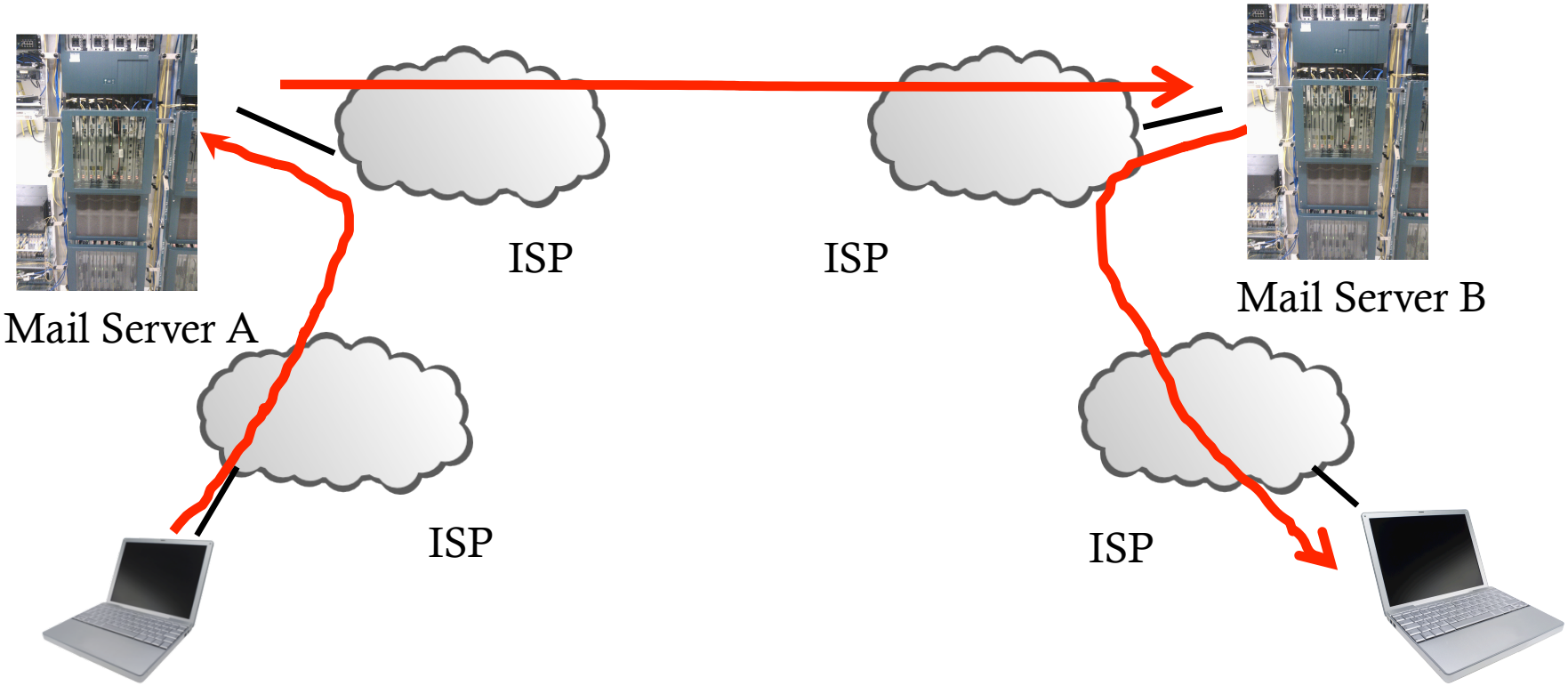
Application and transport are end-to-end; network is hop-by-hop.

Where's the Metadata?

- The source and destination IP (network layer) address are hop-by-hop, and are used by every ISP along the way
  - Close analog to phone numbers in *Smith*
- The TCP (transport layer) "port numbers"—the application service identifier—are end-to-end, and hence are not given to third parties
  - But ISPs sometimes filter on them—does that make them metadata?
- Application data is generally end-to-end—but not always…

What is *Legally* Metadata?

Who owns the mail servers?



Email: What is the Metadata?

220 yyy.com ESMTP Exim 4.82 Tue, 11 Mar 2014 19:43:03 +0000
HELO xxx.cs.columbia.edu
250 yyy.com Hello xxx.cs.columbia.edu [2001:18d8:ffff:16:12dd:b1ff:feef:8868]
MAIL FROM:<smb@xxx.cs.columbia.edu>
250 OK
RCPT TO:<smb@yyy.com>
250 Accepted
DATA
354 Enter message, ending with "." on a line by itself
From: Barack Obama <president@whitehouse.gov>
To: <smb2132@columbia.edu>
Subject: Test

This is a test
.
250 OK id=1WNSaS-0001z5-1d
QUIT
221 yyy.com closing connection

Web Servers and Metadata

- Server log:

108.178.71.xx - - [03/May/2012:21:12:19 -0400] "GET /1e/ HTTP/1.1" 200 2623

- Many people have constant IP addresses, at home or work
- They do not realize that these are logged
- What can be learned?

Information Leakage

- DNS lookup:

  rrcs-108-178-71-xx.sw.biz.rr.com.

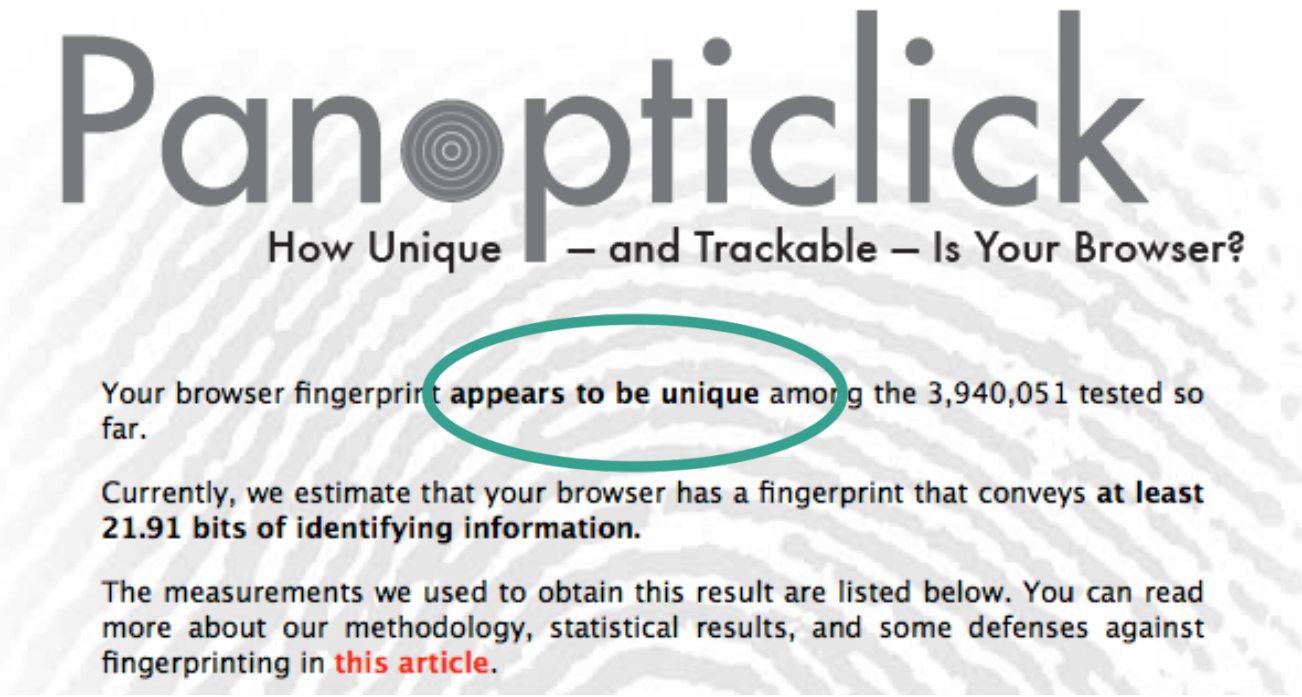
- It's a business – and probably a small one, since Time Warner RoadRunner is a cable ISP and primarily serves consumers
- It's in the southwest
- Investigation with traceroute tells me it's in Austin, TX

Browsers Leak Other Data to Web Sites

```
Connection from 128.59.13.10:57461 at Wed May  2 16:49:39 2012
GET /favicon.ico HTTP/1.1
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.7; rv:12.0)
        Gecko/20100101 Firefox/12.0
Accept: image/png,image/*;q=0.8,*/*;q=0.5
Accept-Language: en-us,en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
```

The EFF says that only one in 200,000 browsers looks like this. (Connect to http://gg1.cs.columbia.edu to see what your browser tells web servers.)

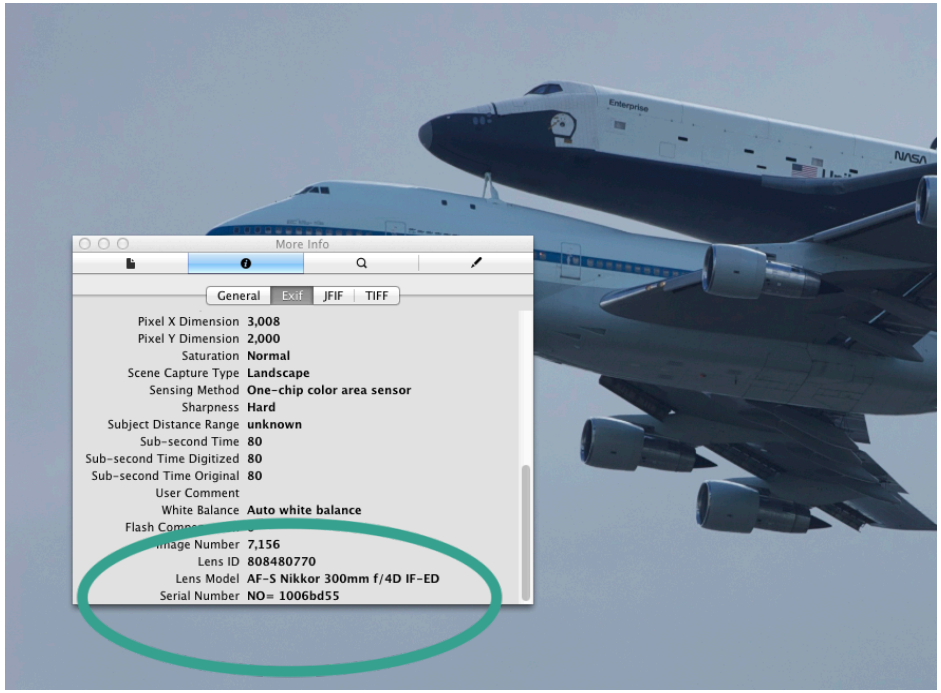
A Second Browser was Unique!



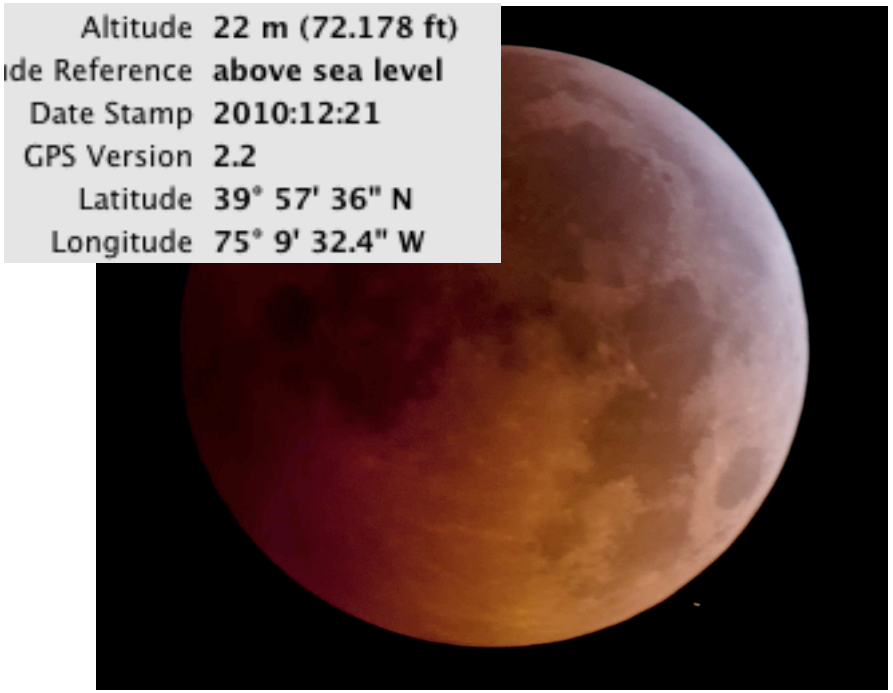
(Via https://panopticlick.eff.org)

A Picture I Took

Let's Look at the Metadata



A Lunar Eclipse Photo with GPS Information



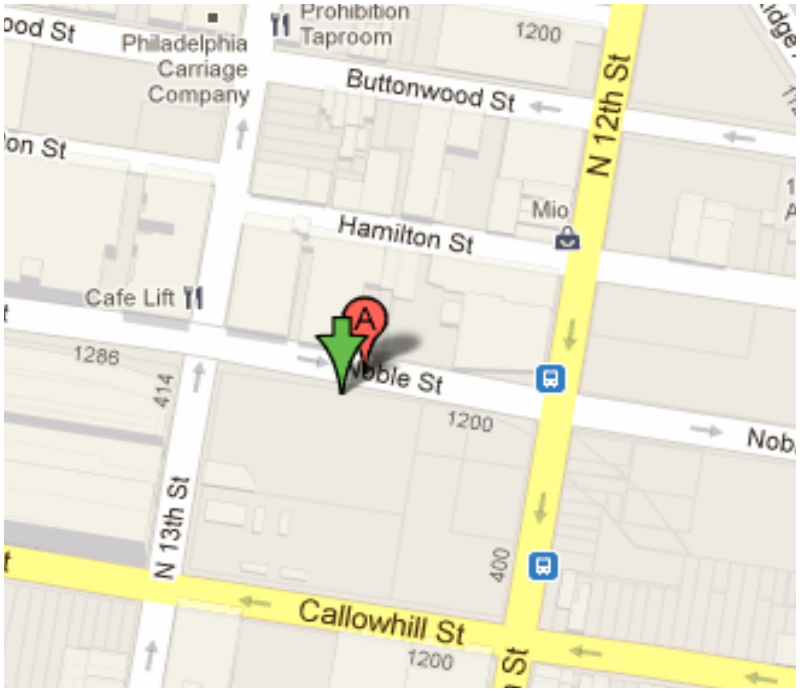


Photo and metadata by Matt Blaze; used by permission

What Can Metadata Do?

- Identify callers who have changed their phone numbers, in a 300 terabyte dataset (Cortes et al, AT&T)
- Identify communities of interest (*Id.*)
- Identify language in encrypted VoIP calls (White et al., UNC)
- Recommend books, movies, etc. (Amazon, TiVo, Netflix, more)

Location Paints a Full Picture

"Disclosed in [GPS] data . . . will be trips the indisputably private nature of which takes little imagination to conjure: trips to the psychiatrist, the plastic surgeon, the abortion clinic, the AIDS treatment center, the strip club, the criminal defense attorney, the by-the-hour motel, the union meeting, the mosque, synagogue or church, the gay bar and on and on"

Quoted in J. Sotomayor's concurrence in *Jones*

How Does This Work?

- Complex mathematical algorithms find *correlations*
  - Note: *not* causality; linkages aren't always obvious to humans
- "Supervised learning": a human annotates a sample dataset during a "training" phase; the algorithms learn what pattern is formed
- "Unsupervised learning": no labeling necessary; can easily find which records clusters with which

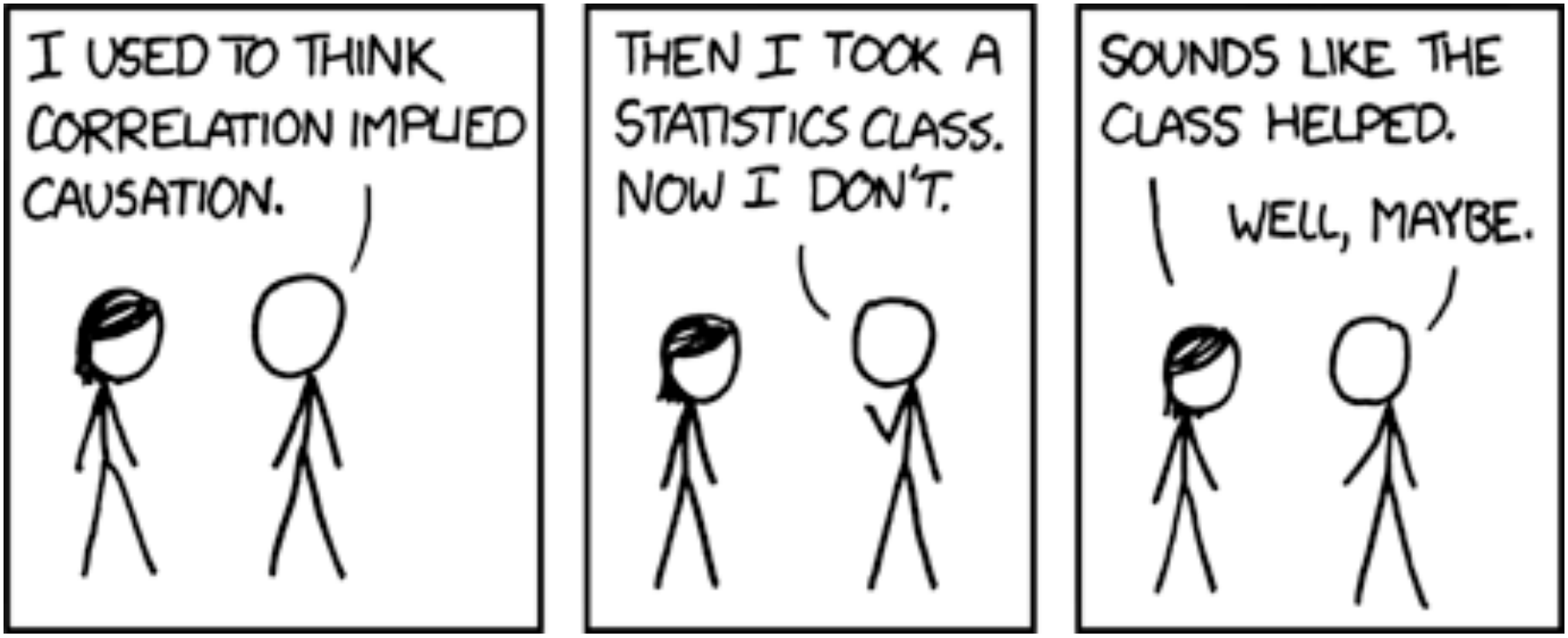
Sample Scenario

- Short, weekly calls from New York to (suspicious country)
- A series of very frequent long calls
- Silence…

Two Explanations

- "Hi, Mom. Yes, I'm studying hard. Yes, my grades are good. Yes, I miss you."
- "Oh, no—how serious do the doctors say it is? What is the prognosis? I'm worried!"
- (plane flight home)

- "Status update: The neighbors aren't suspicious; no contact with law enforcement."
- "Here's the plan. Can you procure…? Is the target guarded?"
- (the plan is about to be executed.)



(https://xkcd.com/552/)

Concerns

- More effort or less effort (aka “expense”)
- More or less certainty
- Knowing or unknowing disclosure
- Prospective or retrospective
- Confirmatory or accusatory