It’s All About the Data

200 DATA 3, 4
210 DATA 5, 12
220 DATA 20, 17

Bill Cheswick
Visiting Scholar, University of Pennsylvania
ches@cheswick.com
Old data

- PDP-1
- Dartmouth BASIC
- AMSAT
- Lunar orbiter photos
• This was created and run on the original Dartmouth BASIC.

• The simulator emulates the GE instruction set!

• There is even a simulated TTY font! (Parens, “*” are broken)

• Dykstra is rolling in his grave: GOTOs are considered harmful.

```
LIST
HYPOT  11:15   TUE 06 JUN 2017
100 READ X, Y
110 LET H = SQR(X↑2 + Y↑2]
120 PRINT X, Y, H
130 GOTO 100
200 DATA 3, 4
210 DATA 5, 12
220 DATA 20, 17
999 END

READY

RUN
HYPOT  11:15   TUE 06 JUN 2017
3              4              5
5              12             13
20             17
26.2488

OUT OF DATA IN  100

TIME:  Ø SECS.
```
AMSAT Phase 2 ham satellite: AO-7

- Launched 15 Mov, 1974
- Went silent in 1981
- Back to life in ~2002
- Similar experience with the Meteor M-N1 weather satellite
- International Sun/Earth Explorer, launched in 1978
  - Needs complete ground station software rewrite
Lunar orbiter image recovery project

- ~1,500 tapes
- Needed Ampex FR-900 tape drives. Spare parts on eBay!
- Needed specialized demodulation hardware
- Tapes had four times the dynamic range of the original film images, and twice the resolution.
Preserving data

- Preserving the bits
- Understanding the formats
Preserving bits

- I am not happy with current archival data storage solutions
  - longevity is uncertain
  - I want TB-sized media
- This means periodic copying, or
- relying on cloud providers
Which bits to preserve?

- All of them. Storage is getting cheaper (see below)
- Curation can be a pain. Leave this as technical debt to future generations
- The Eternal Web Site
Eternal Web Site

- Neil Sloane idea from the 1990s
- A web site never goes away
- Curated by a reliable company, perhaps with an affiliated research group
- Funded by cheap monthly payments for updates, the cemetery model for the rest
- Commitment to update formats over time
- I presented this idea to AT&T management
Eternal data preservation

• Mores and politics change
• Duplicate and distribute the data too-widely to be completely extinguished
  • DVDs in the ocean trenches, and orbiting Jupiter. Litter the moon with them.
Forgetting data

- mostly seems to happen accidentally
- (Cygnherd, and special keywords)
  - chessecretkeyword
- freedom to be forgotten?
Understanding old data

- Old formats: the .doc story. They will be figured out, probably easier than archeologist’s tasks.
Data collection

- The Internet Mapping Project
Lumeta

- Spun off from Bell Labs in 2000.
- Gathering data and converting it to information
- Topological mapping, display, and analysis
- Perimeter leaks
- What, exactly, is on the “inside” of a network
The Internet at night
Uses of these pretty pictures

- Hanging at the DOJ, FCC, given to Premier of India
- AT&T used them for FCC net neutrality arguments.
- Posters, tee shirts, etc.
- Have appeared in numerous talks, mostly unattributed.
US military, reached with ICMP ping
Un film par Steve “Hollywood” Branigan...
Serbian propaganda site found

- Dead babies, “Hiroshima”, etc.
- Do I take it down?
- Cheswick needs a personal foreign policy
- At the time, this might have come as a surprise to the US State Department.
The Internet after 9/11
NYC routers after 9/11
Vast data

- Internet topography and traffic (Internet Mapping Project)
- The World Wide Web
- The cell system
  - think about “burners”
- aerial mapping
- licence plate readers
- Video surveillance and face recognition
  - “Super recognizers”
- Cellular proteins
- Genetic information
- Metabalome, virome, immune system
- Brain connections, neural net analysis
Vast data challenges

• All of this is about converting data to information.
• The visual system is well-suited to this, given thoughtful implementation
• Lumeta customers: “What are my top five problems?”
• The ball of yarn problem
• Showing changes over time. (OpenGL and 3D? VR?)
• Internet Mapping Project has about 1TB of Internet scans from 1998 to 2011
Data exfiltration

- “Dirty words”, data spills
- Covert channels are a big deal
- eat a thumb drive
- glue in the USB slots?
- traffic analysis: Gordon Welchman at Bletchley. “metadata”
- FBI wiretap jokes

- Chaum networks (TOR) don’t work if they are infiltrated
- leak detection
- How I might spend $100M in black money
Data Spills

- Credit card and banking data
- PCI vs VPNs
- Wikileaks, Snowden, CIA hacking tools, Stuxnet, OPM
- Certified Data Handler?
Leak detection

- A sends a packet to B, with a spoofed return address of D
- B doesn’t care where the test packet comes from
- It has a route for D, and sends a reply from C
- D receives a packet from C, containing info that says that A sent it to B
What Might I Do with $100M of black money
(inspired by Banford’s *Body of Secrets*)

- Problem statement
- Deploy taps
- Find matching packet patterns
- I suspect that traffic analysis techniques are a major remaining World War II secret
Packet size (log)

0 -100 -1,000

0 -10 -100

0 10 100 1,000

0 10 20 30 40 50 60 70 80 90 100 110 120 130 140

Packet size (log)

0 -100 -1,000

0 -10 -100

0 10 100 1,000

0 10 20 30 40 50 60 70 80 90 100 110 120 130 140

43 of about 56
Protecting Data

- Encryption: AES, MD5/SHA, etc.
- End-to-end
- VPN: use elliptic curves
- Safer clients. iOS least unsafe?
- CDH
- n-factor authentication
- 105 and zoompass demos?
Time release data

• The digital fuse
Data Policies

- Government policies
- access vs privacy
- What about bad governments, whatever you may conceive them to be?
- HIPPA v. checklists
- Public data
Public data

- Many state, local, and federal government agencies are making their data available to the public
  - They need systematic data, and good APIs
  - They need courage: the citizens find problems
- The arguments are similar to the open source security arguments
- Ben Wellington found the most-ticketed legal parking spot in NYC
  - see http://iquantny.tumblr.com/
Complicated data: neural nets

- It is mostly about processing data
- “ten hundred thousand fingers have fingers.”
How to wreck a nice beach (Guido)
How to wreck a nice beach (ches)
Neural nets and machine learning

- How do you know how they work?
  - How do you test them?
  - How do you test people? How confident are you in them? How do you hack them?
- What part does each neuron play? These are probably beyond comprehension.
- ISO standard neural nets, approved for a particular use?
- What about updated nets? How do you qualify the training data?
- Do you name nets?
- Open source v. proprietary nets
Real AI?

- Start with a set of rules for thought (this is the hard part)
- Feed in the entire world wide web.
  - Biased? Not a problem, that evaluation is part of AI processing, learning that information has probabilities, points of view, etc.
- Give advice, and ask for comments on the advice?
  - “Don’t trust Fox News.”
- Read the commentary on bias in Google News
- AI victory: makes research suggestions that are intriguing to the experts.
Conclusion

• Data is everywhere, and it is easy to obtain new data. Data needs conversion to information.

• Data analysis can yield fascinating and useful results

• Data mining is often more valuable than gold mining, and you don’t have to go to Alaska.

• These are skills that cross disciplines.

• Learning how to handle it is time well-spent:
  • Screen scraping
  • Statistics
  • Databases (don’t forget simple Unix filters).
  • There are PhD-level problems remaining
Suggestions

• Think how you can gather data:
  • Unusual fields in Internet packets (xprobe2)
  • Low TTL values suggest you may be getting mapped

• Unusual traffic:
  • spoofed packets
  • weird protocols and values

• Log everything, then understand the logs’ contents.

• Look for outliers. I use tail(1) a lot

• grep -v the stuff you understand, and see what’s left
It’s All About the Data

200 DATA 3, 4
210 DATA 5, 12
220 DATA 20, 17

Bill Cheswick
Visiting Scholar, University of Pennsylvania
ches@cheswick.com