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**

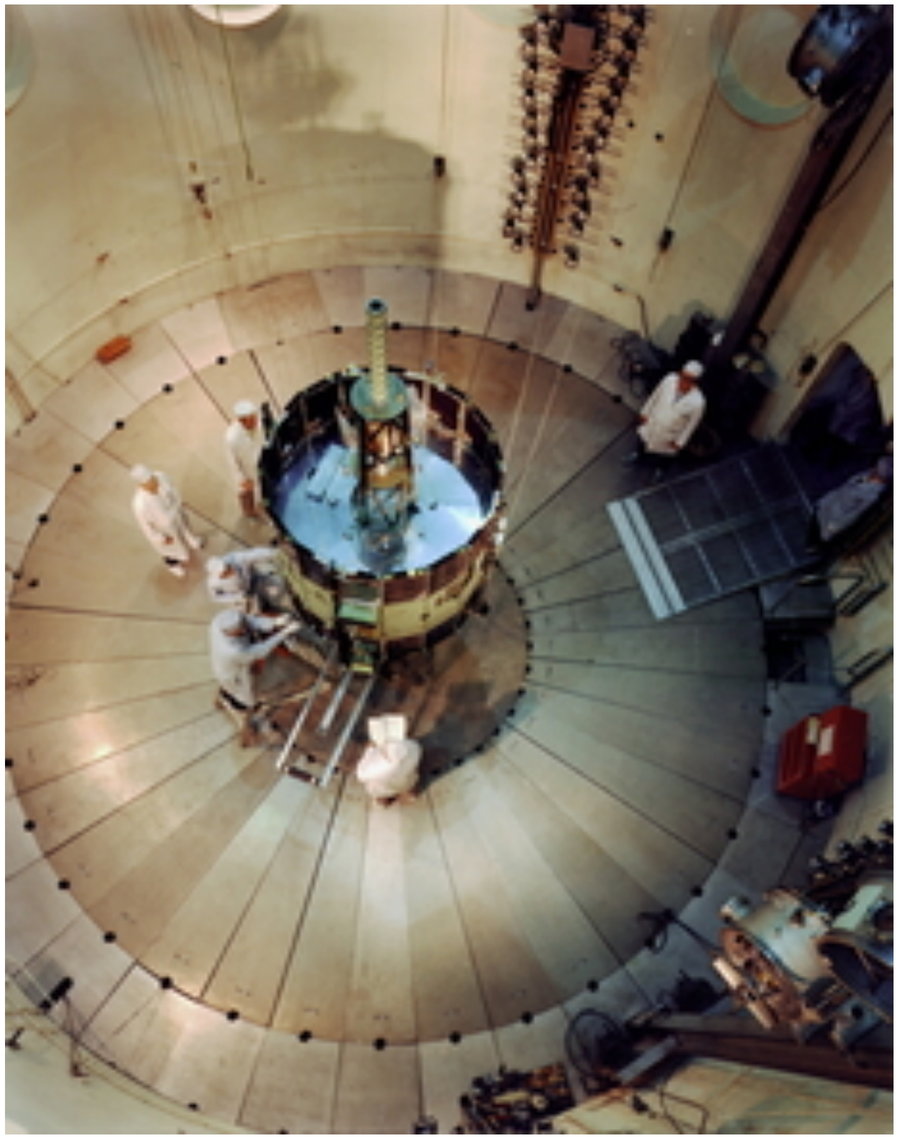

digital EQUIPMENT CORPORATION

- **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 Ø6 JUN 2Ø17

1ØØ READ X, Y
11Ø LET H = SQR[X↑2 + Y↑2]
12Ø PRINT X, Y, H
13Ø GOTO 1ØØ
2ØØ DATA 3, 4
21Ø DATA 5, 12
22Ø DATA 2Ø, 17
999 END

READY

RUN

HYPOT     11:15   TUE Ø6 JUN 2Ø17

 3                4                5
 5                12               13
 2Ø               17
26.2488

OUT OF DATA IN  1ØØ


TIME:    Ø SECS.
```

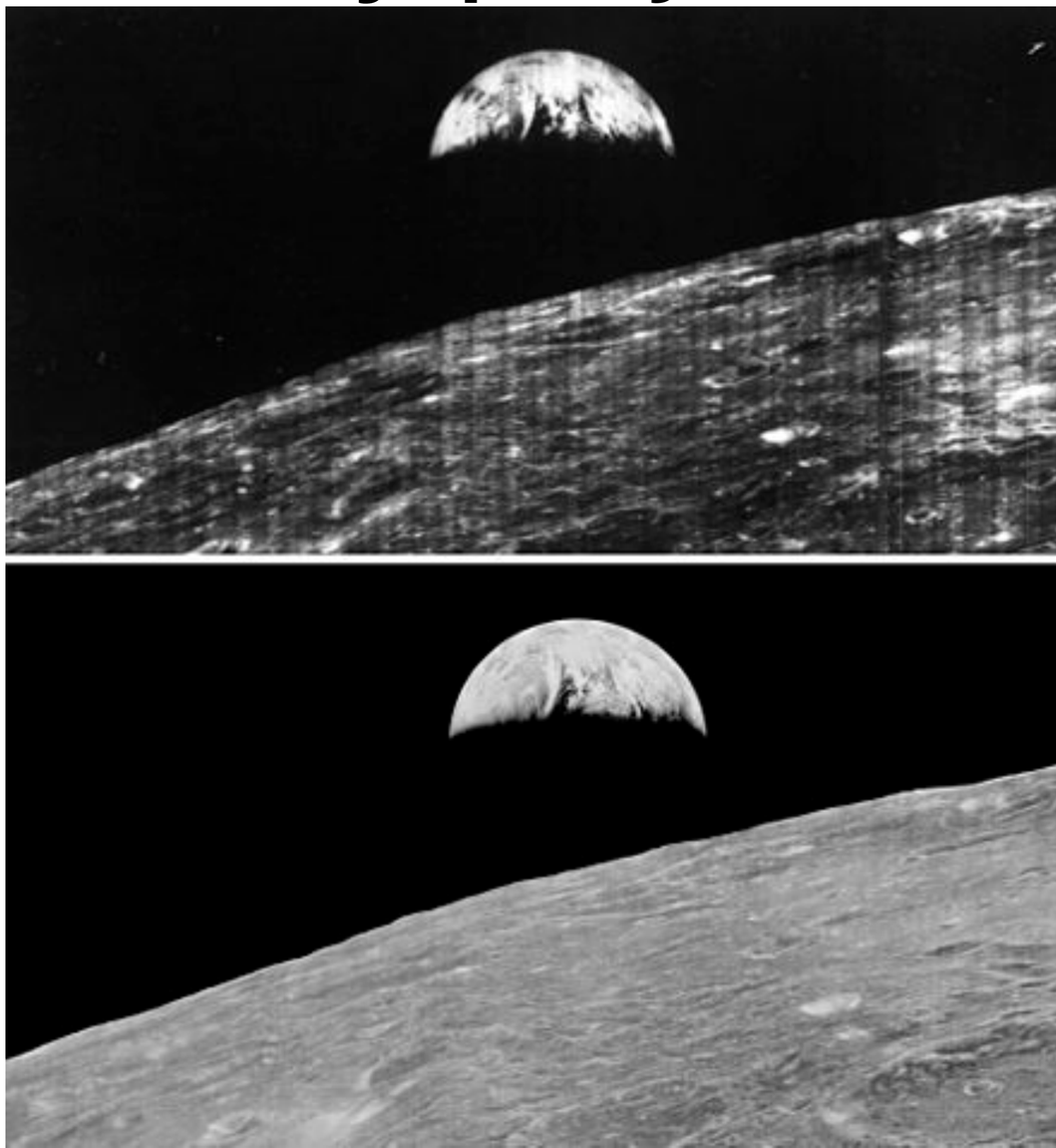
AMSAT Phase 2 ham satellite: AO-7

- **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

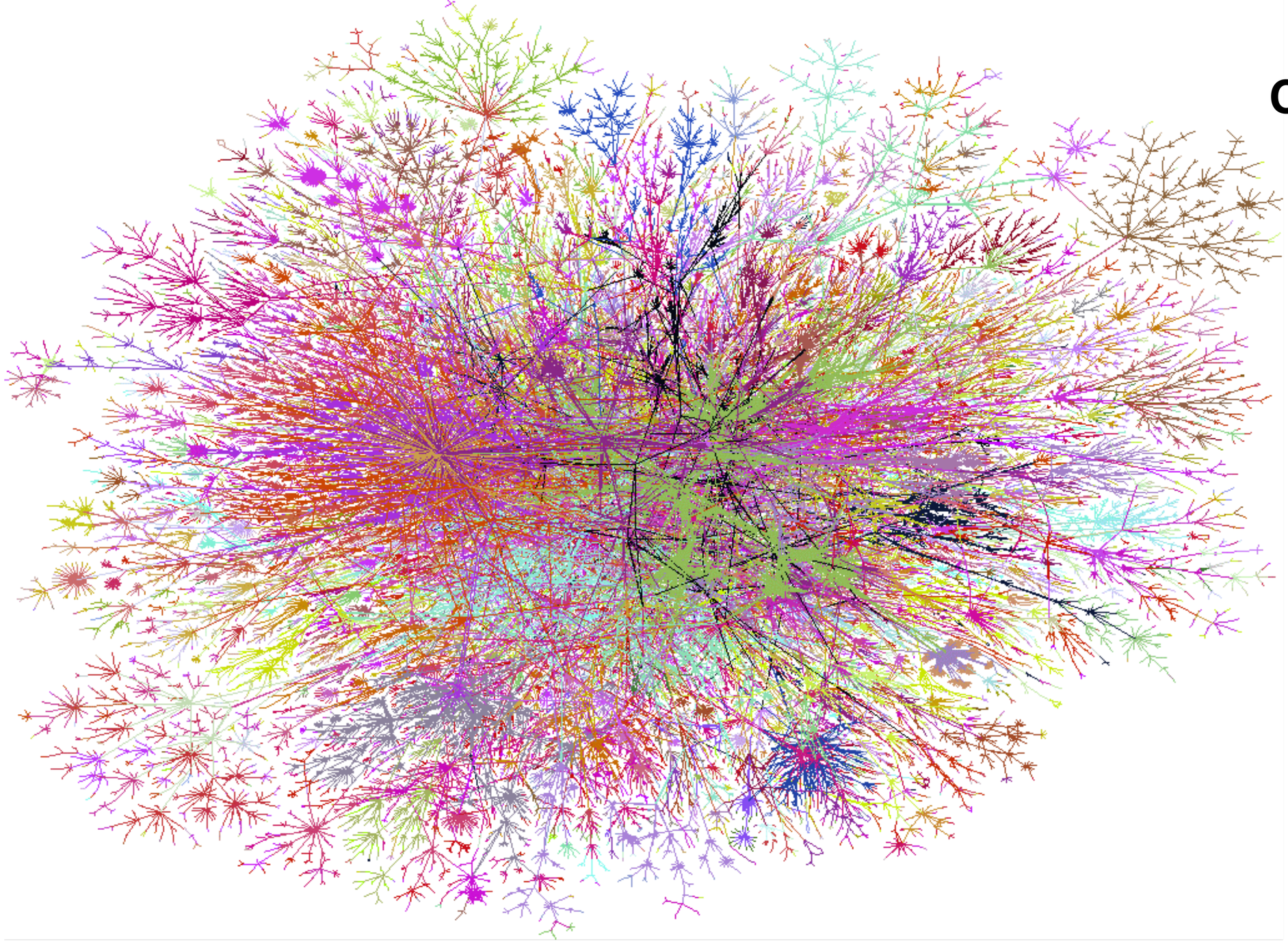
- **Old formats: the .doc story. they will be figured out, probably easier than archeologist's tasks.**

Data collection

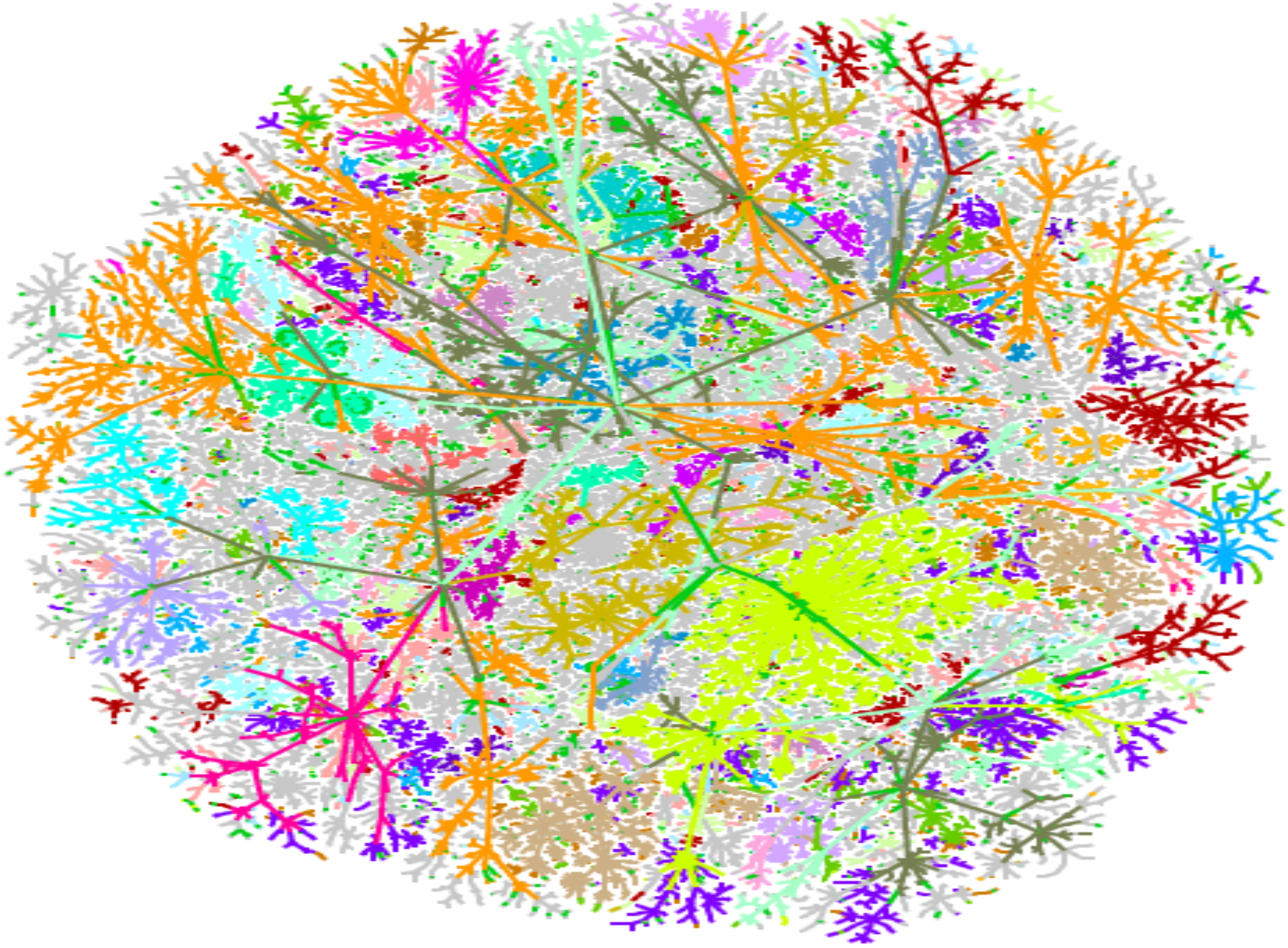
- **The Internet Mapping Project**

Lumeta

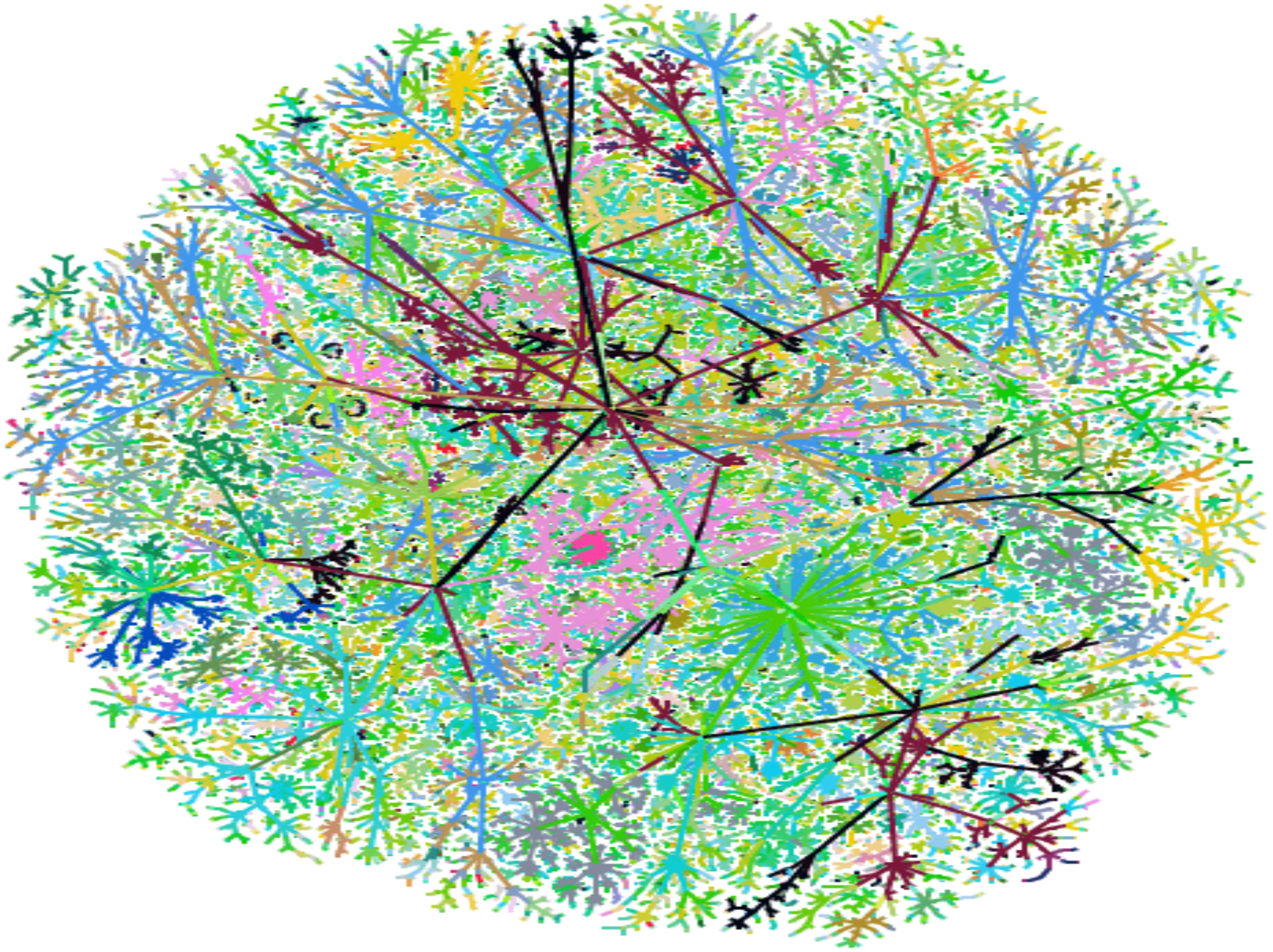
- **Spun off from Bell Labs in 2000.**
- **Gathering data and converting it to information**
- **Topological mapping, display, and analysis**
- **Perimeter leaks**
- **What, exactly, is on the “inside” of a network**



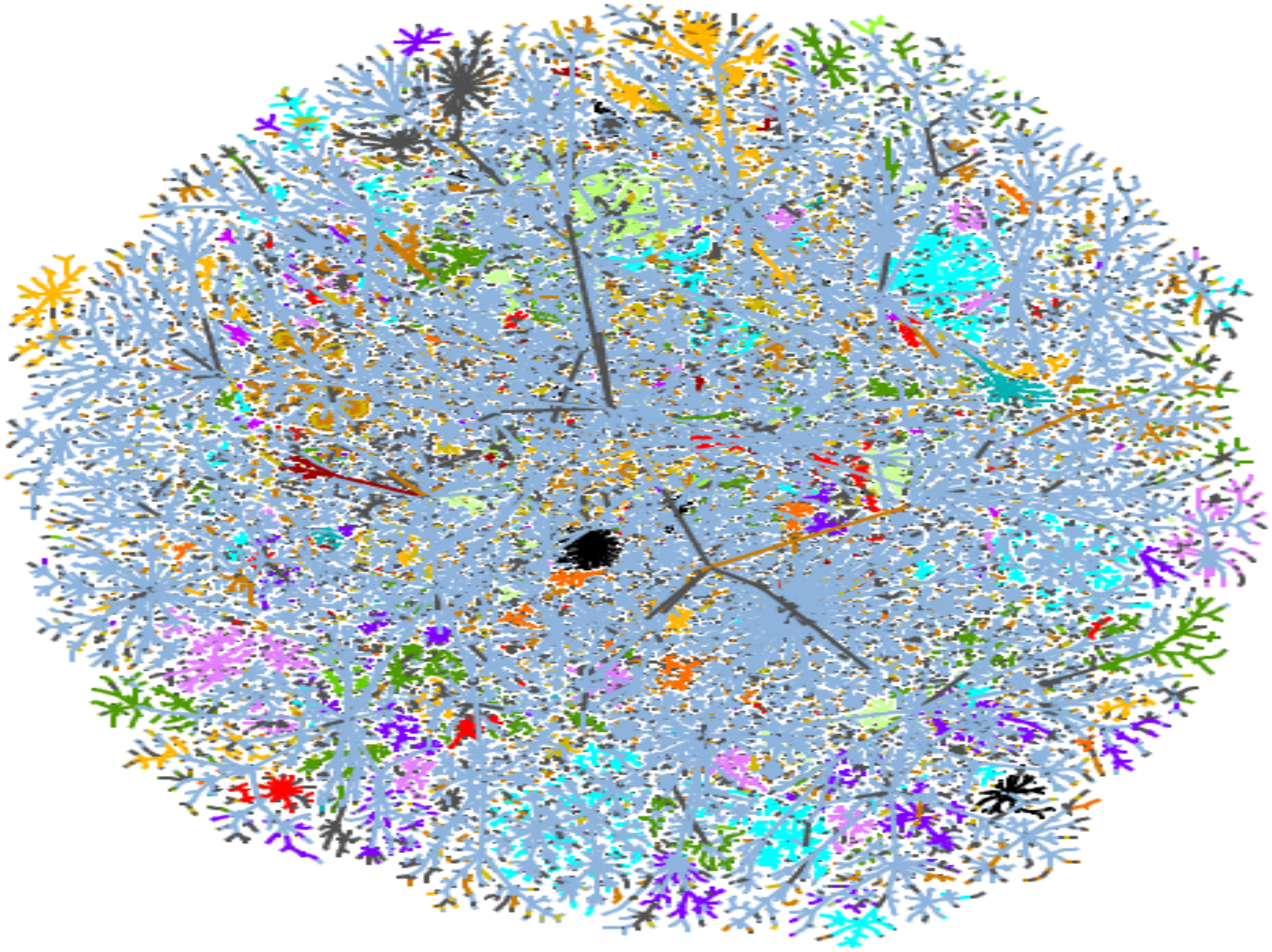
Colored by IP address (!)



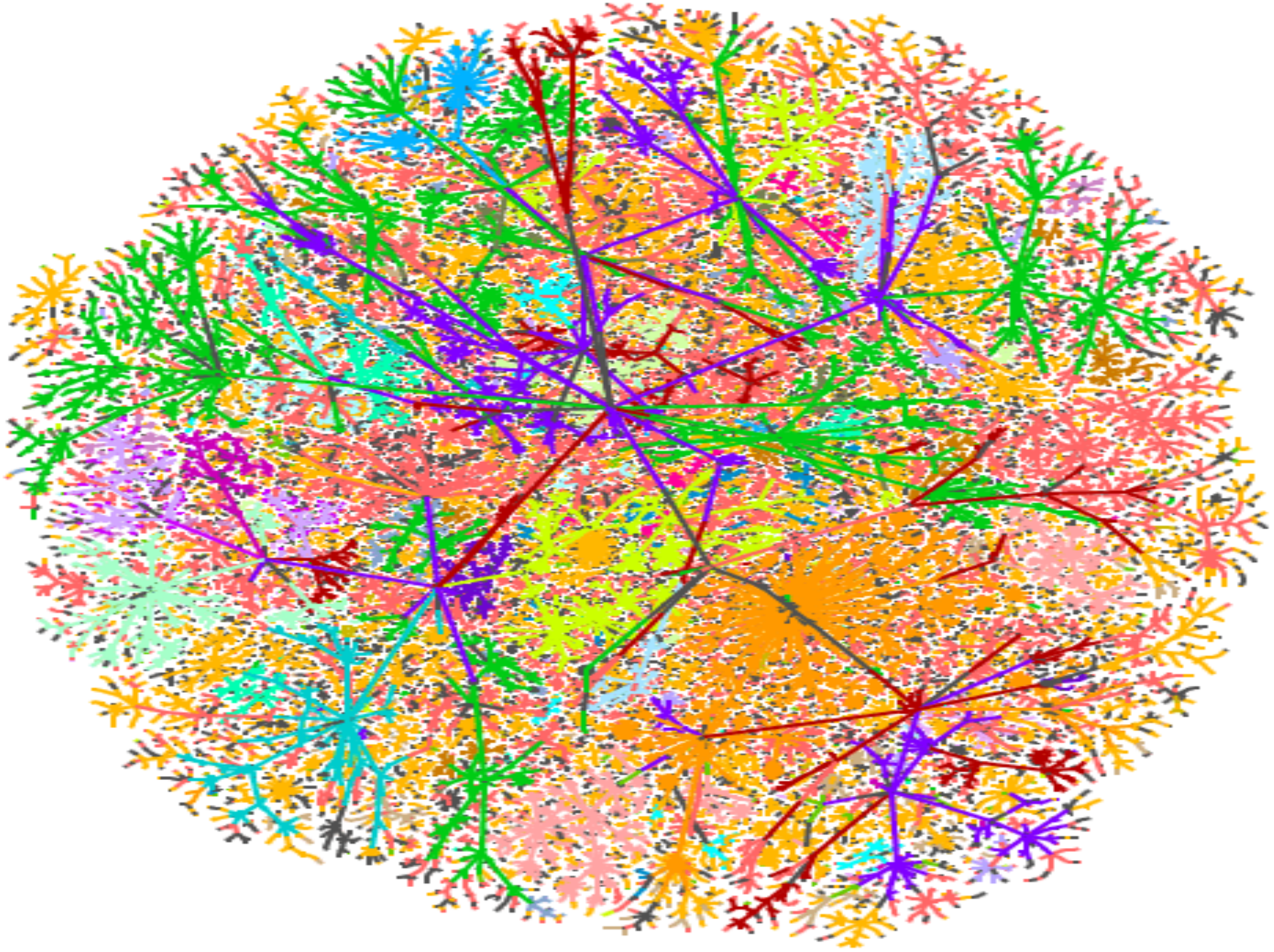
Colored by AS number



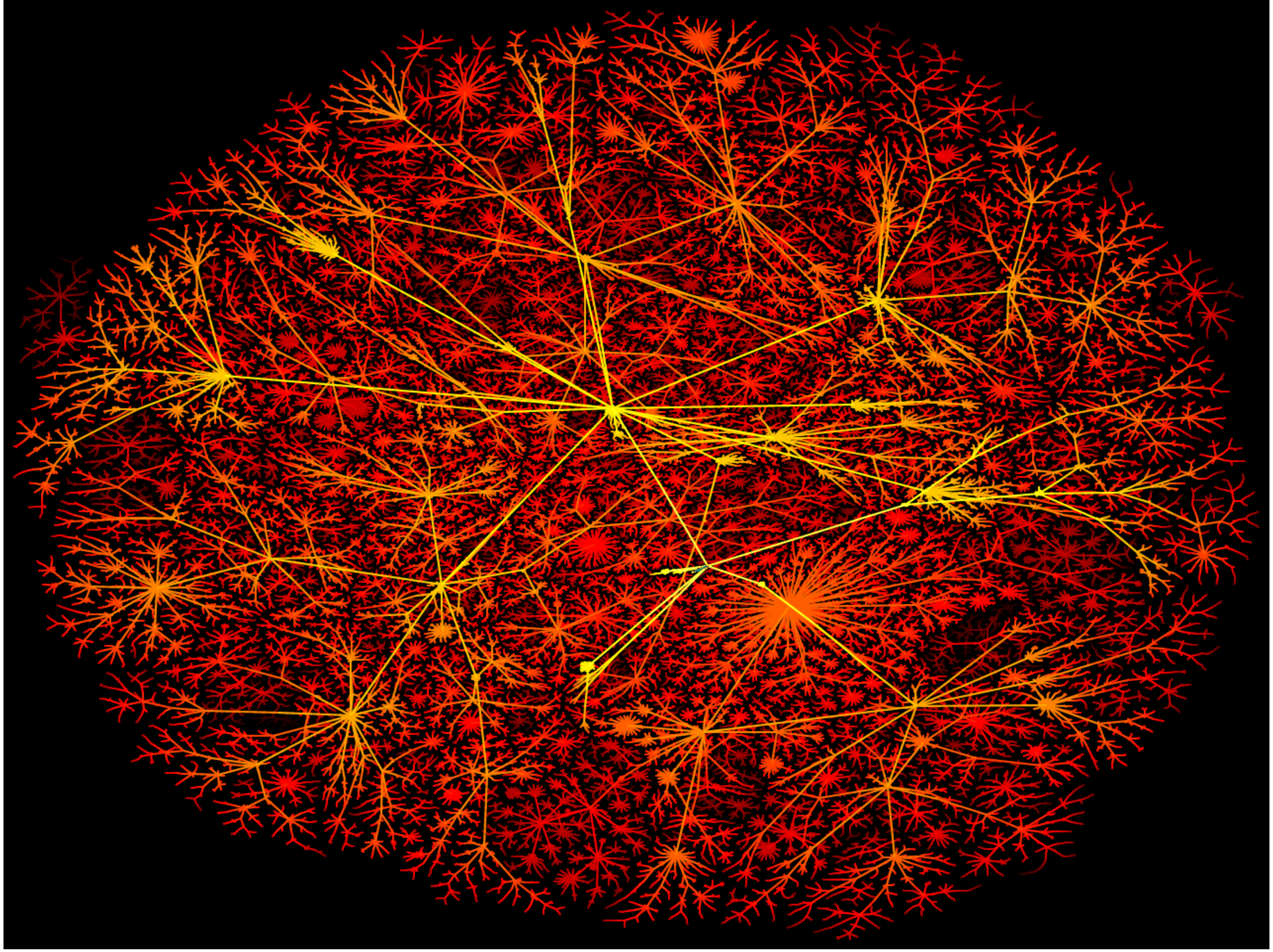
MDST colored by IP address



Colored by country or TLD



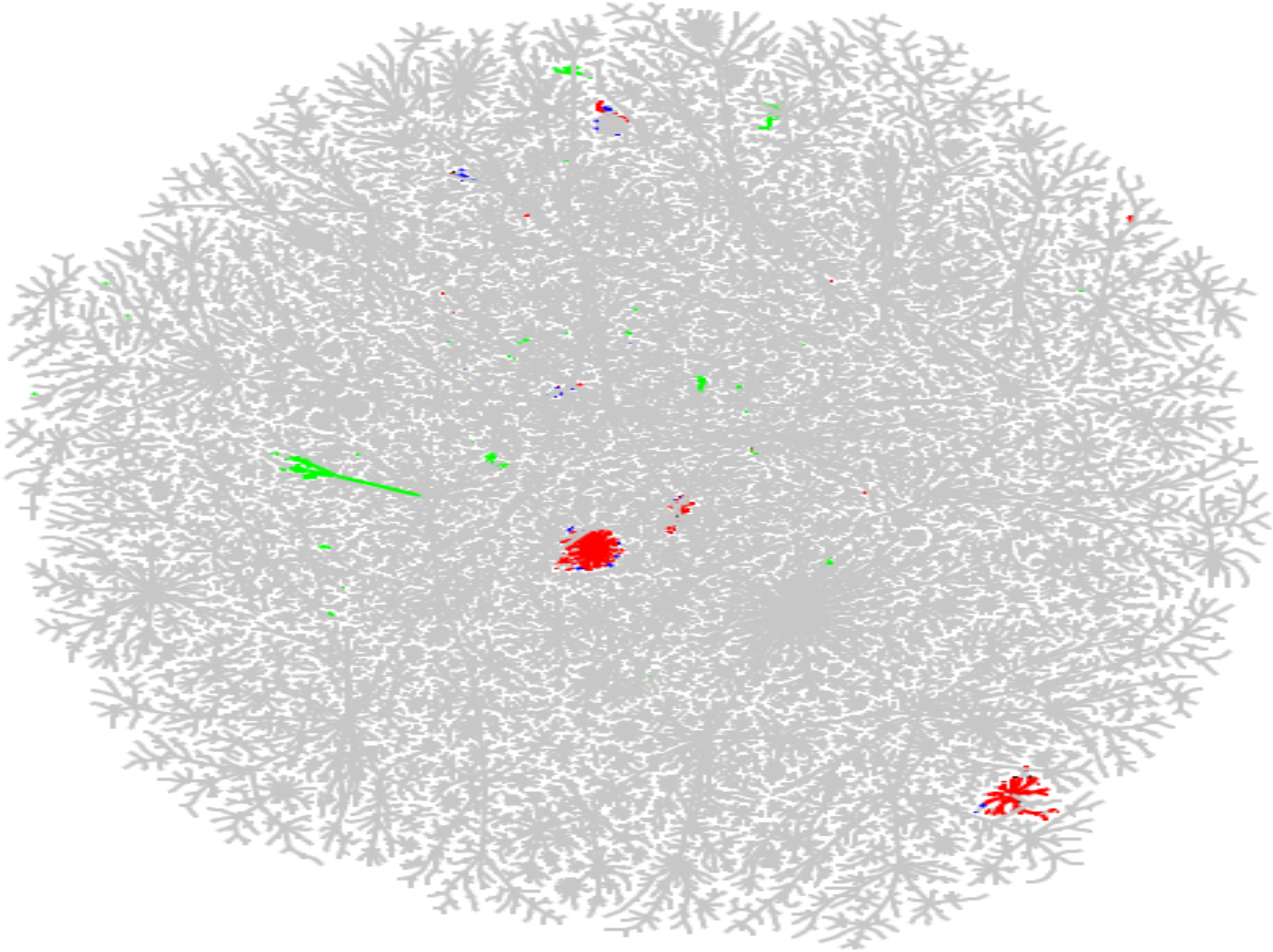
MDST colored by ISP



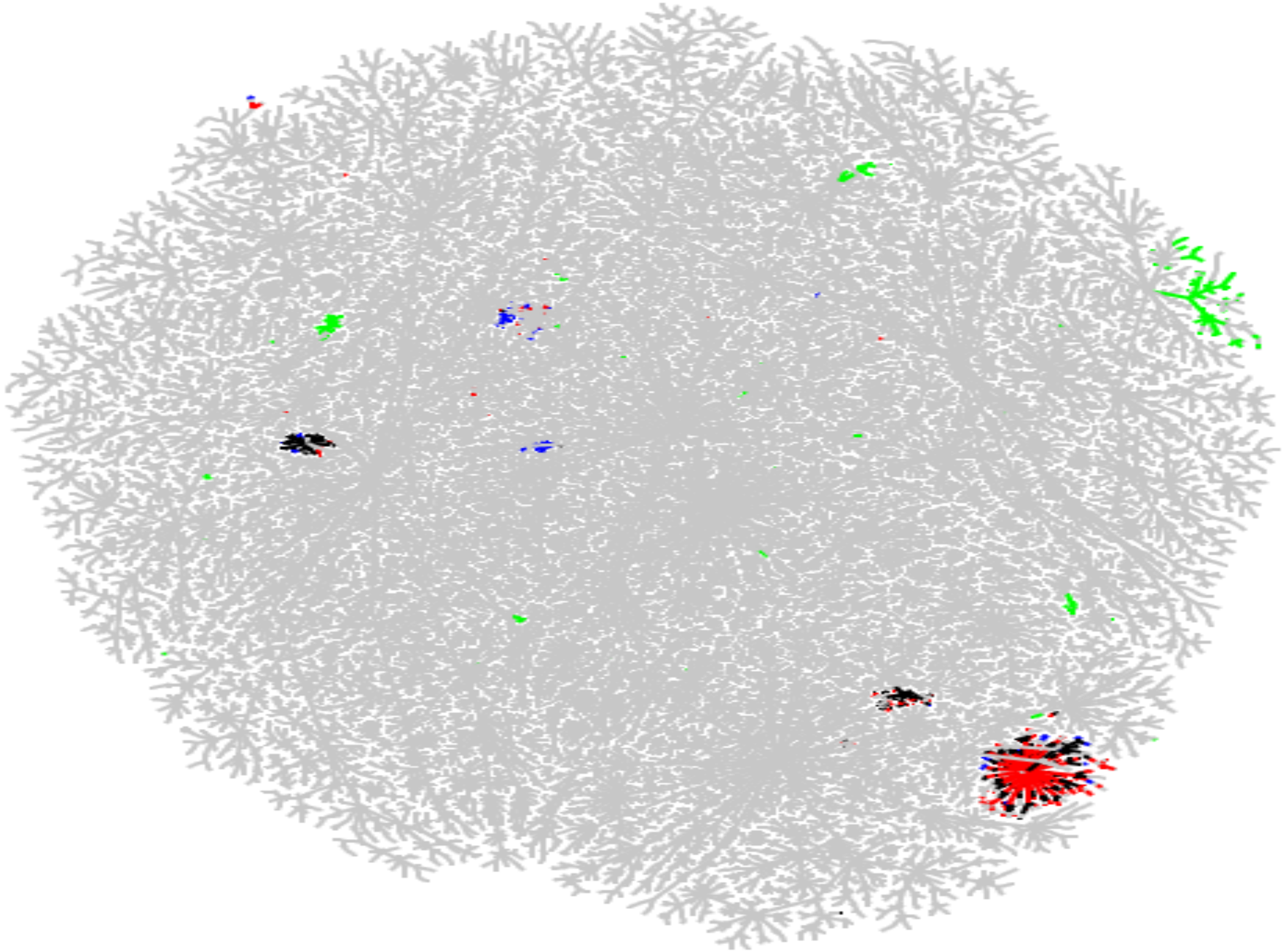
The Internet at night

Uses of these pretty pictures

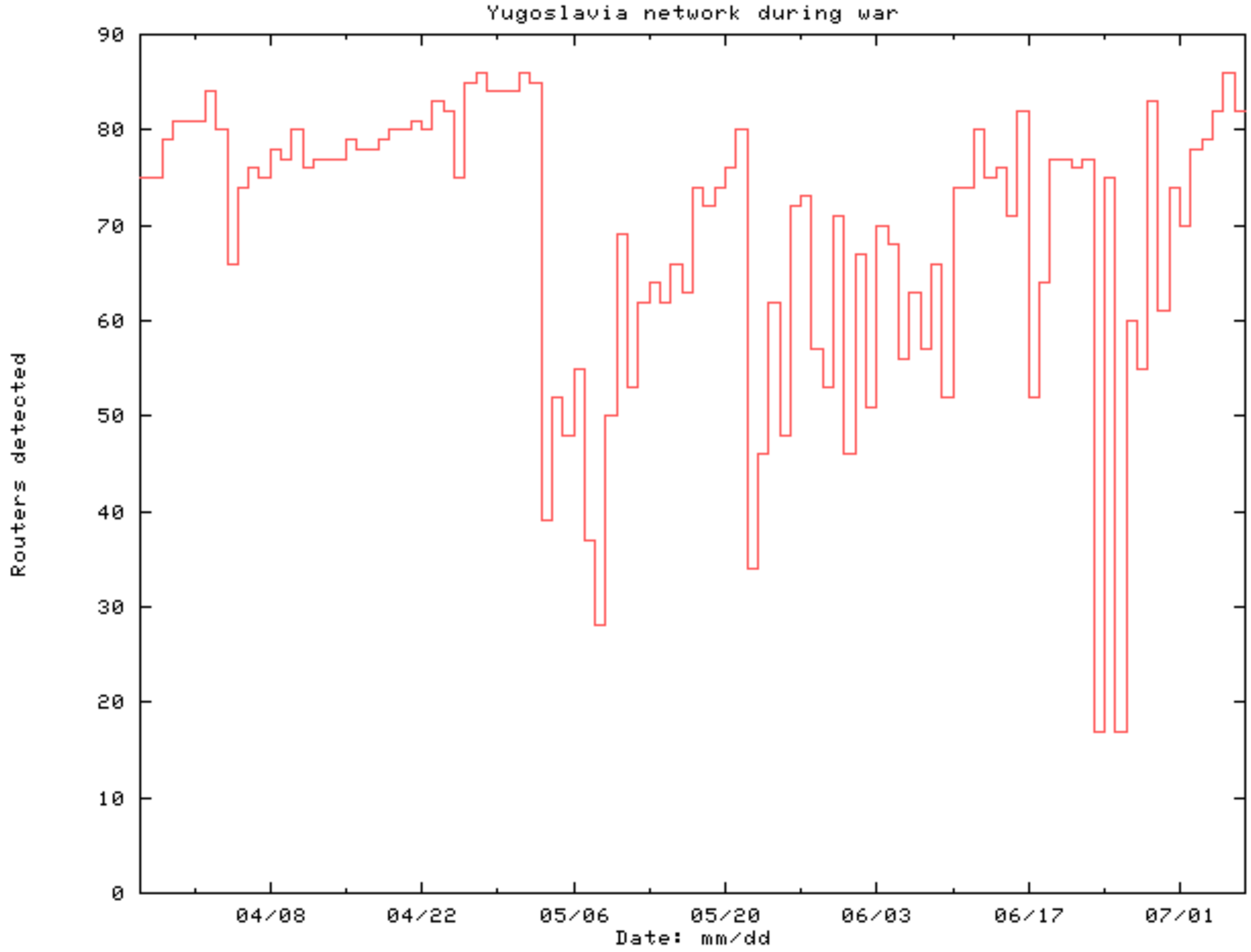
- **Hanging at the DOJ, FCC, given to Premier of India**
- **AT&T used them for FCC net neutrality arguments.**
- **Posters, tee shirts, etc.**
- **Have appeared in numerous talks, mostly unattributed.**



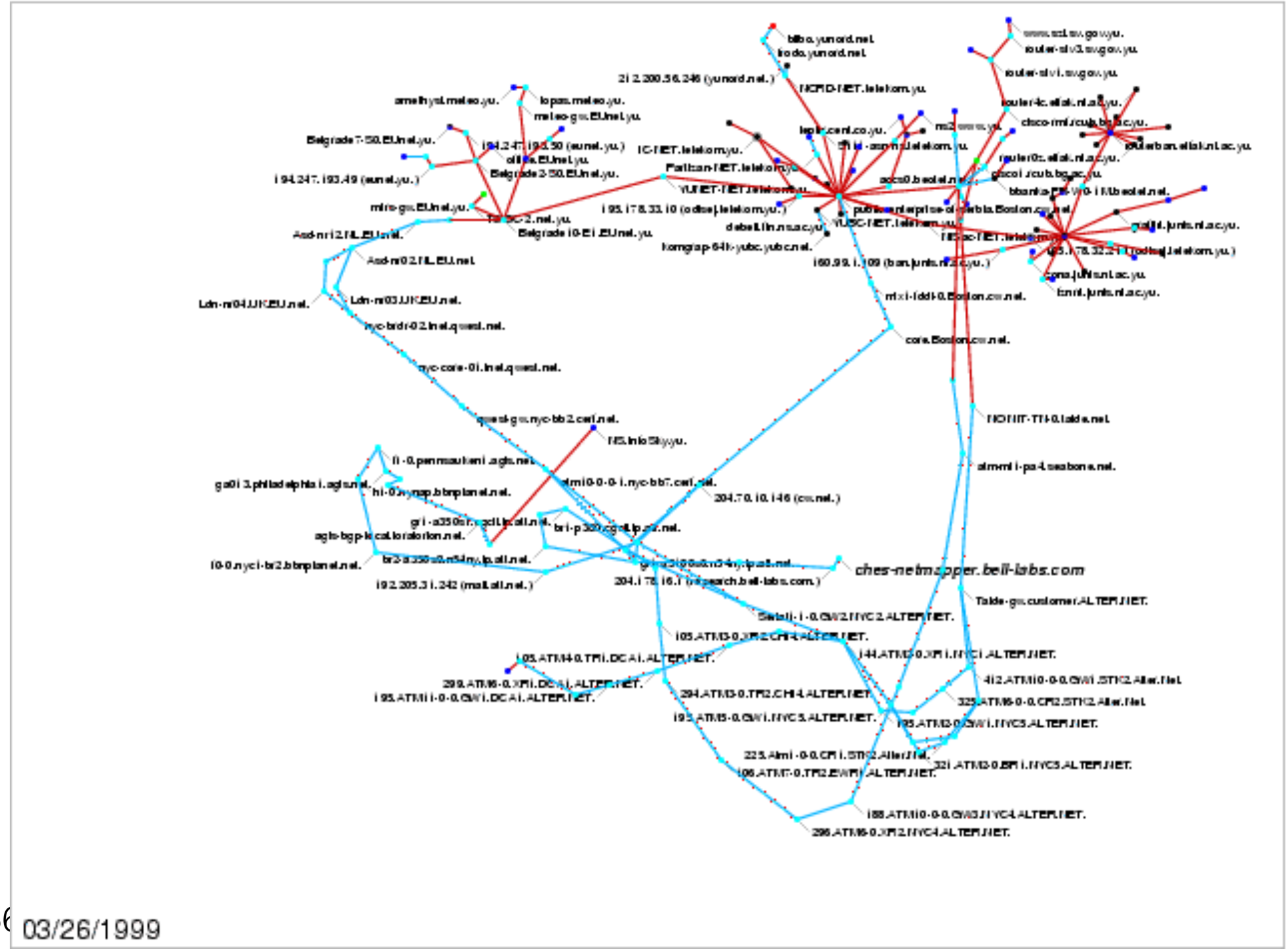
US military, reached with ICMP ping

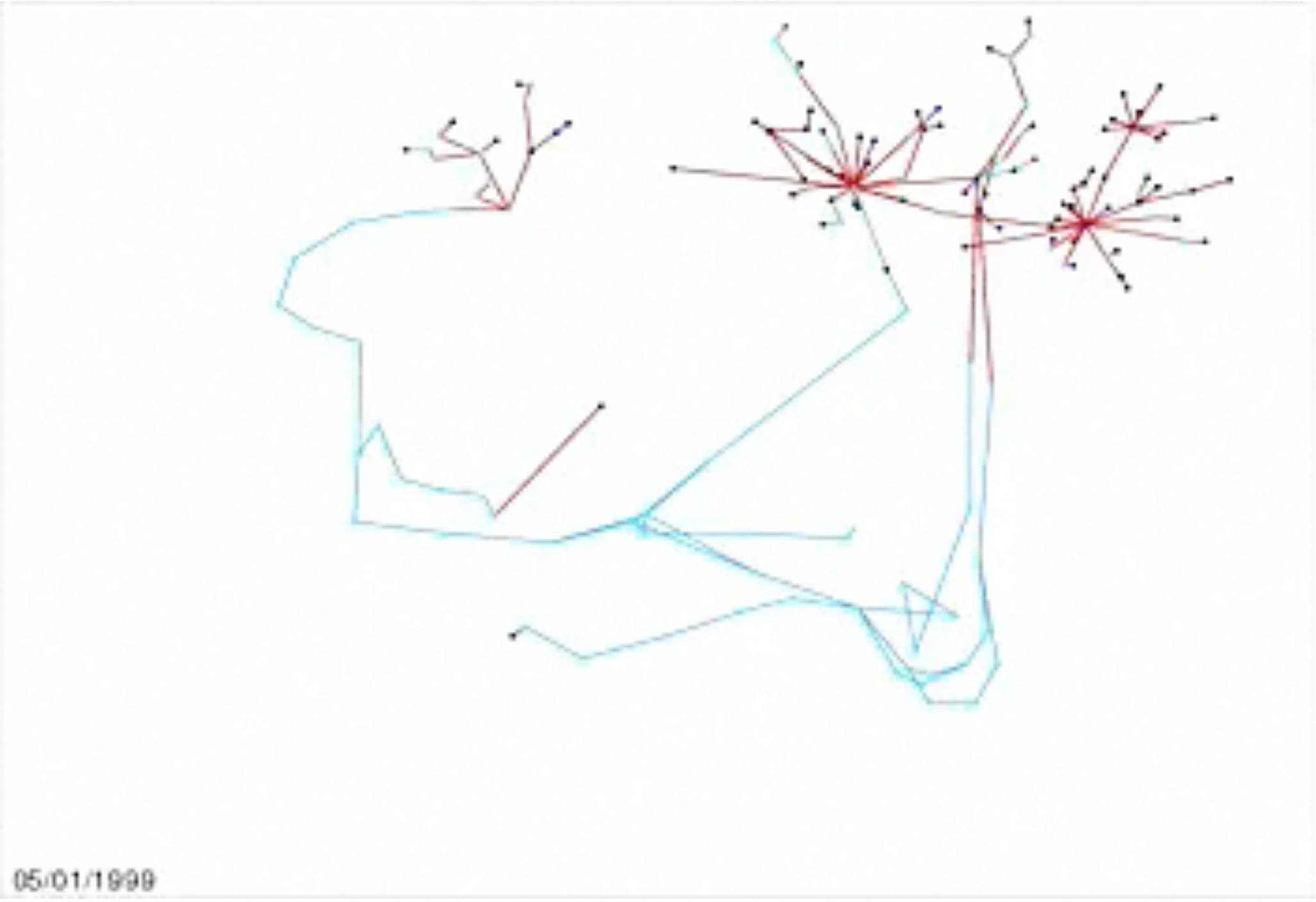


US military, reached with UDP


Yugoslavia network during war
Routers detected
90
80
70
60
50
40
30
20
10
0
04/08
04/22
05/06
05/20
06/03
06/17
07/01
Date: mm/dd

Un film par Steve “Hollywood” Branigan...

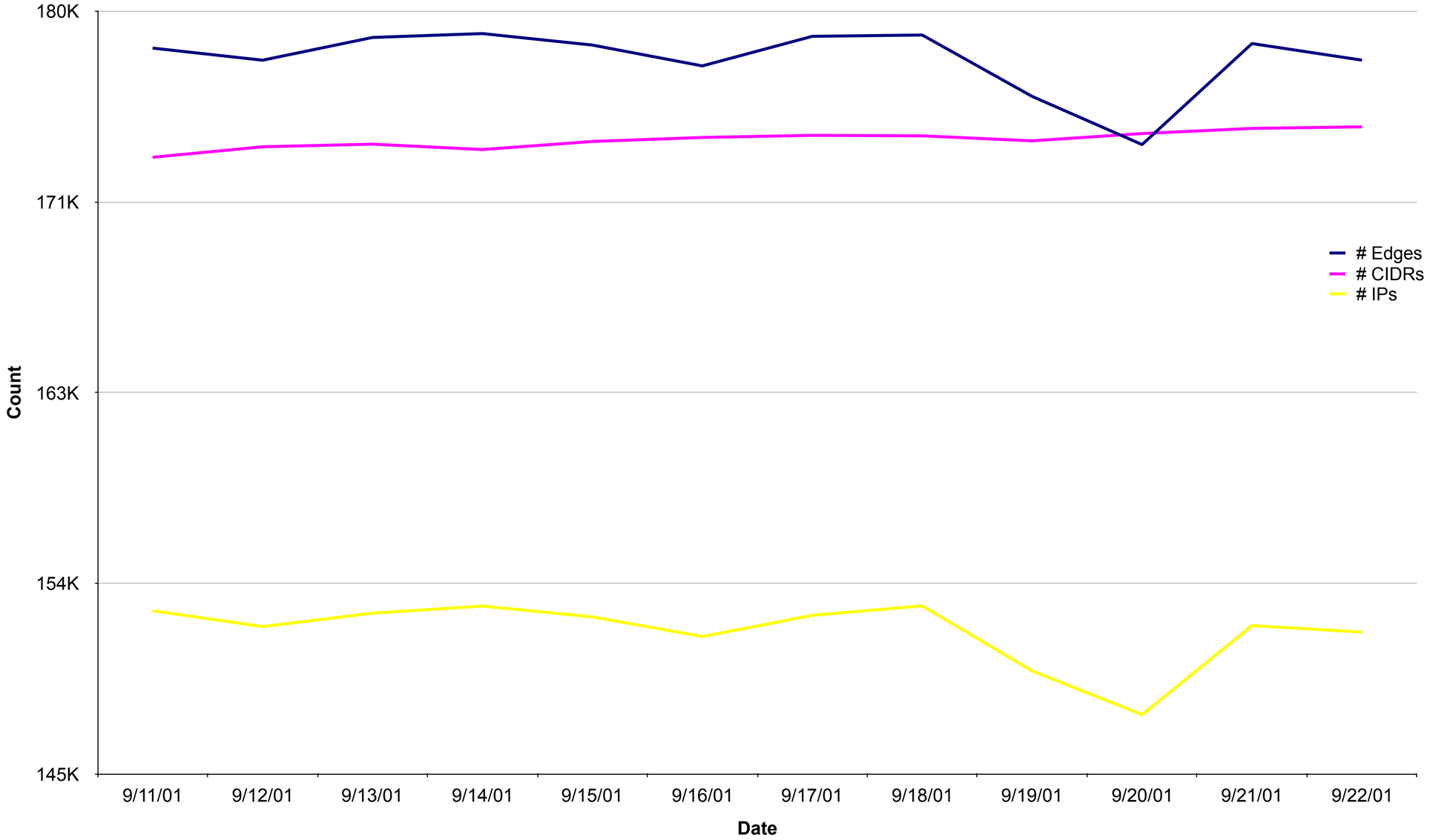

ches-netmapper.bell-labs.com


05/01/1999

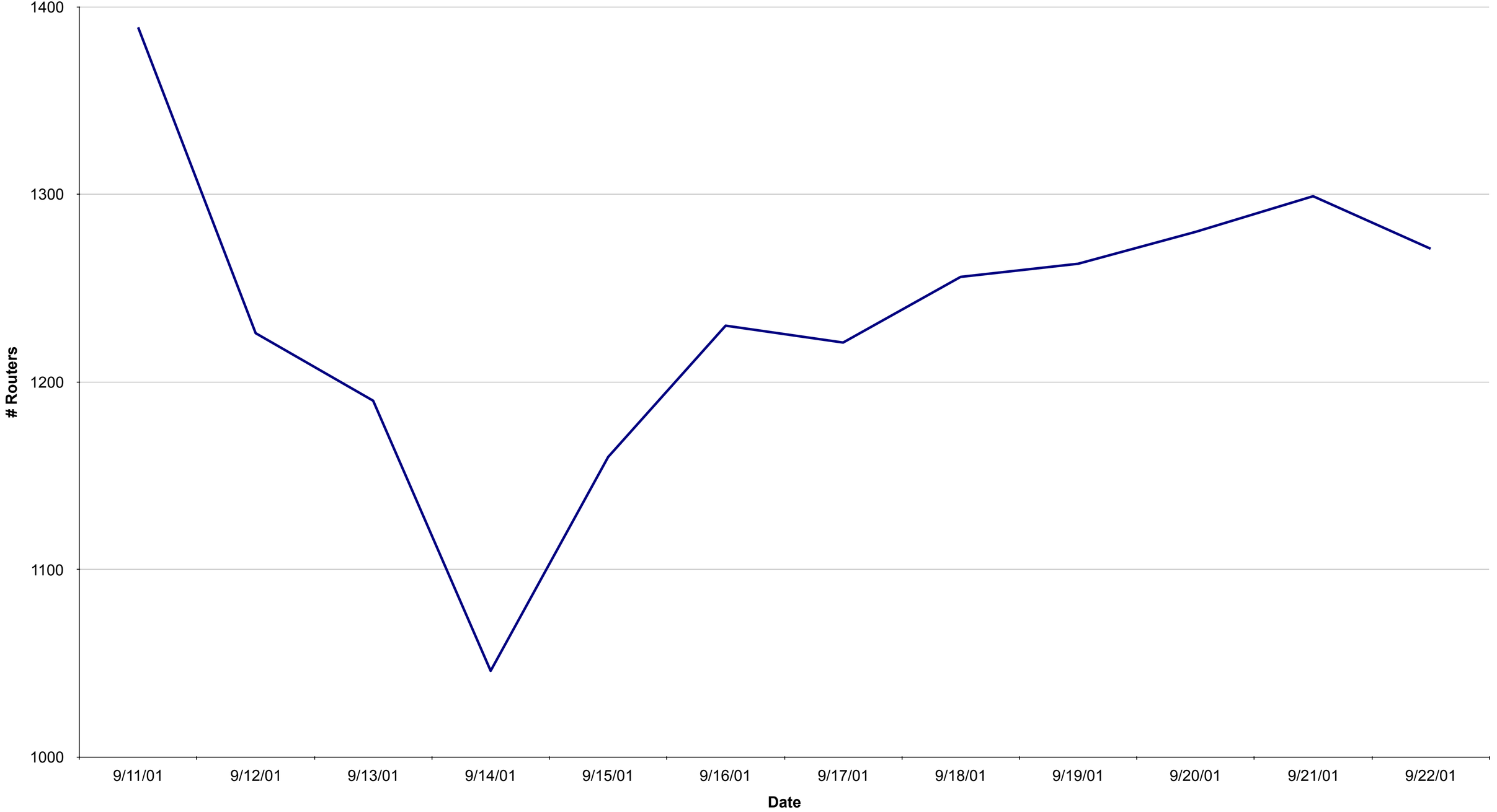
fin

Serbian propaganda site found

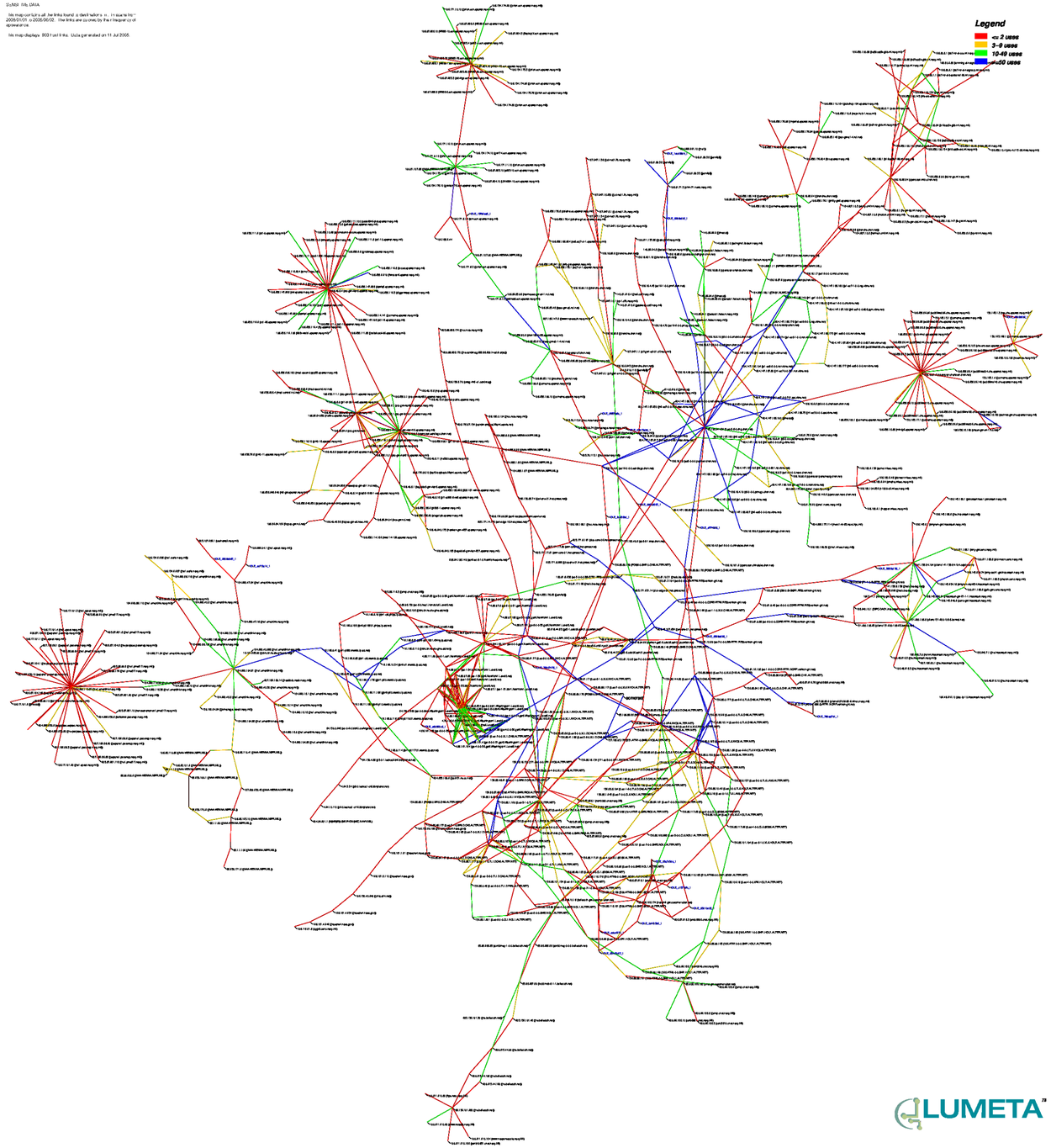
- **Dead babies, “Hiroshima”, etc.**
- **Do I take it down?**
- **Cheswick needs a personal foreign policy**
- **At the time, this might have come as a surprise to the US State Department.**



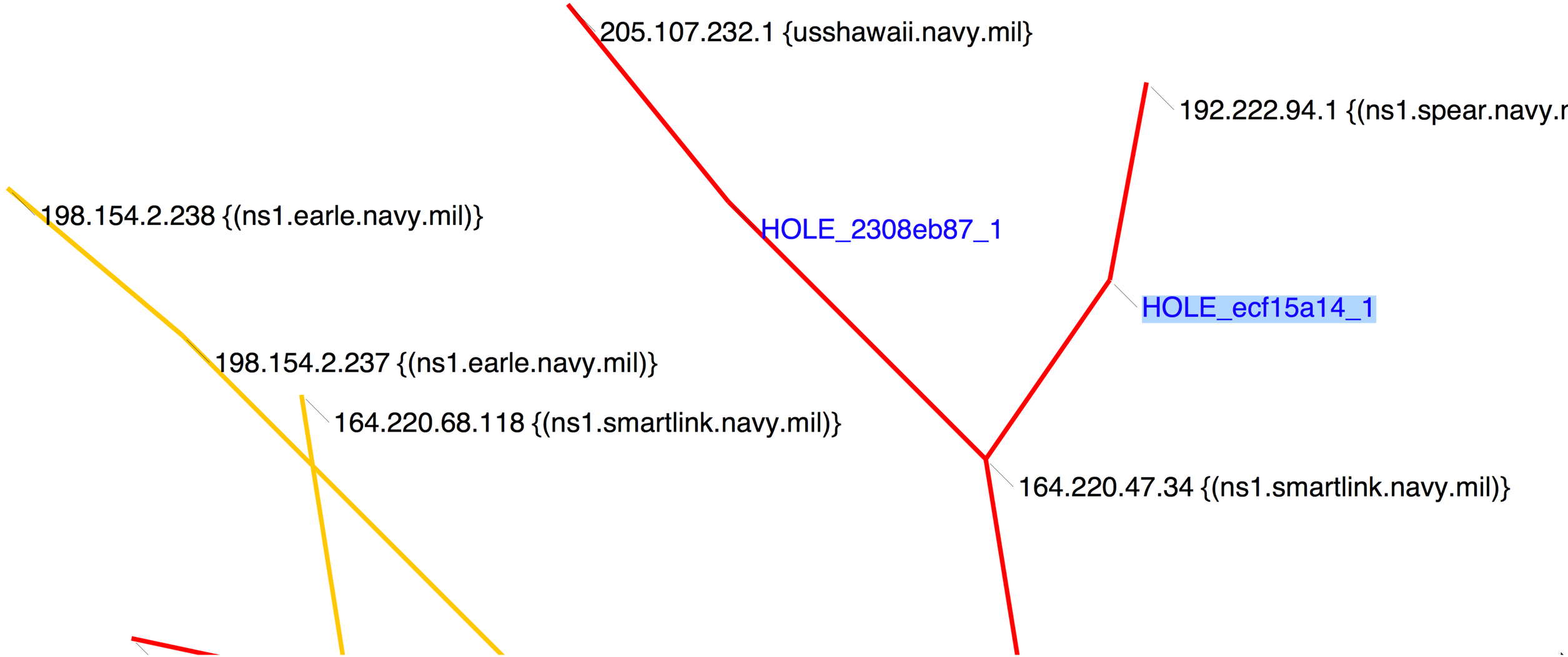
The Internet after 9/11



NYC routers after 9/11


Legend
<= 2 uses
3 - 9 uses
10-49 uses
>=50 uses


198.62.64.162 {fitcpac-gw.nci.net}
128.49.43.56 {fitcpac-gw-s2.nosc.mil}
198.62.64.94 {ivcs-gw.nci.net}
198.253.136.2 {pl606-1.spawar.navy.mil}
198.253.136.20 {siegel-pb.spawar.navy.mil}
128.49.240.175 {hawken-gw-v257.spawar.
198.253.114.132 {nws114-132.spawar.navy.mil}
128.49.240.165 {bayside2-gw-vlan-2
206.24.193.149 {acr2-so
205.107.232.1 {usshawaii.navy.mil}
192.222.94.1 {(ns1.spear.navy.mil)}
HOLE_2308eb87_1
HOLE_ecf15a14_1
198.154.2.238 {(ns1.earle.navy.mil)}
198.154.2.237 {(ns1.earle.navy.mil)}
164.220.68.118 {(ns1.smartlink.navy.mil)}
164.220.47.34 {(ns1.smartlink.navy.mil)}
205.68.89.5 {aiscgw11nns.supship.navy.mil}
164.220.68.110 {(ns1.smartlink.navy.mil)}
164.220.40.6 {(ns1.smartlink.navy.mil)}
y.mil)}
99.20 {(ns1.pms317.navy.mil)}
164.220.47.18 {(ns1.smartlink.navy.mil)}
205.67.231.3 {(ns1.pms317.navy.mil)}
164.220.50.162 {(ns1.smartlink.navy.mil)}
164.220.192.98 {(ns1.smartlink.navy.mil)}
206.37.164.3 {(ns-peopasc.peosnap.navy.mil)}
164.220.68.166 {(ns1.smartlink.navy.mil)}
avy.mil)}
164.220.71.34 {(ns1.smartlink.navy.mil)}
164.220.40.226 {(ns1.smartlink.navy.mil)}
12.124.235.90 {(cbru.br.ns.els-gms.att.net)}
205.66.99.3 {(ns1.pms317.navy.mil)}
205.66.99.172 {(ns1.pms317.navy.mil)}
164.220.194.81 {(ns1.smartlink.navy.mil)}
164.220.113.26 {(ns1.smartlink.navy.mil)}
164.220.192.10 {(ns1.smartlink.navy.mil)}
s1.smartlink.navy.mil)}
164.220.113.58 {(ns1.smartlink.navy.mil)}
207.132.169.14 {(nasbbdc.nasb.navy.mil)}
164.220.40.50 {(ns1.smartlink.navy.mil)}
205.67.231.20 {(ns1.pms317.navy.mil)}
164.220.43.6 {(ns1.smartlink.navy.mil)}
164.220.16.1 {(ns1.smartlink.navy.mil)}
164.220.16.14 {(ns1.smartlink.navy.mil)}
159.120.34.34 {(pxmanage.navair.navy.mil)}
164.220.16.13 {(ns1.smartlink.navy.mil)}
205.67.231.10 {nsnosanantonex1.pms317.navy.mil}
204.222.250.3 {belltower.peosnap.navy.mil}
164.220.194.66 {(ns1.smartlink.navy.mil)}
205.68.112.5 {(smtp.cnet.navy.Mil)}
4.68.97.35 {ge-5-0-0-52.gar2.NewYork1.Level3.ne
4.68.96.150 {so-6-0.hsa1.Honolulu2.Level3.net}
12.123.3.106 {(xbru.br.ns.els-gms.att.net)}
12.122.2.18 {tbr1-cl8.phlpa.ip.att.net}
12.122.9.166 {tbr2-p013601.phlpa.ip.att.net}
4.68.128.174 {(ns2.Level3.net)}
204.255.168.2 {(auth51.n
12.123.3.61 {tbr2-p011601.n54ny.ip.att.net}
12.123.3.110 {(xbru.br.ns.els-gms.att.net)}
12.123.9.41 {gar1-p360.wswdc.ip.att.net}
4.68.111.29 {so-3-1.car1.NewYork1.Level3.net}
12.122.2.85 {tbr1-cl9.wswdc.ip.att.net}
12.122.10.54 {tbr2-cl1.wswdc.ip.att.net}
4.68.97.97 {ae-1-54.bbr2.
12.123.9.45 {gar1-p370.wswdc.ip.att.net}
4.68.97.33 {ae-1
12.122.11.178 {gbr1-p20.wswdc.ip.att.net}
4.68.97.161 {ae-1-56.bbr2.
4.68.97.129 {ae-1-55.bbr1.NewY
12.122.11.182 {gbr2-p20.wswdc.ip.att.net}
12.122.11.166 {gbr2-p10.wswdc.ip.att.net}
64.159.3.2
64.159.0.230 {ae-0-0.bbr2.Washington1.Level3.net}
4.68.128.201 {(ns2.Level3
12.123.9.118 {tbr2-p011901.wswdc.ip.att.net}
4.79.192.6 {VERIZON-INTE.ga


205.107.232.1 {usshawaii.navy.mil}
192.222.94.1 {(ns1.spear.navy.
198.154.2.238 {(ns1.earle.navy.mil)}
HOLE_2308eb87_1
HOLE_ecf15a14_1
198.154.2.237 {(ns1.earle.navy.mil)}
164.220.68.118 {(ns1.smartlink.navy.mil)}
164.220.47.34 {(ns1.smartlink.navy.mil)}

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

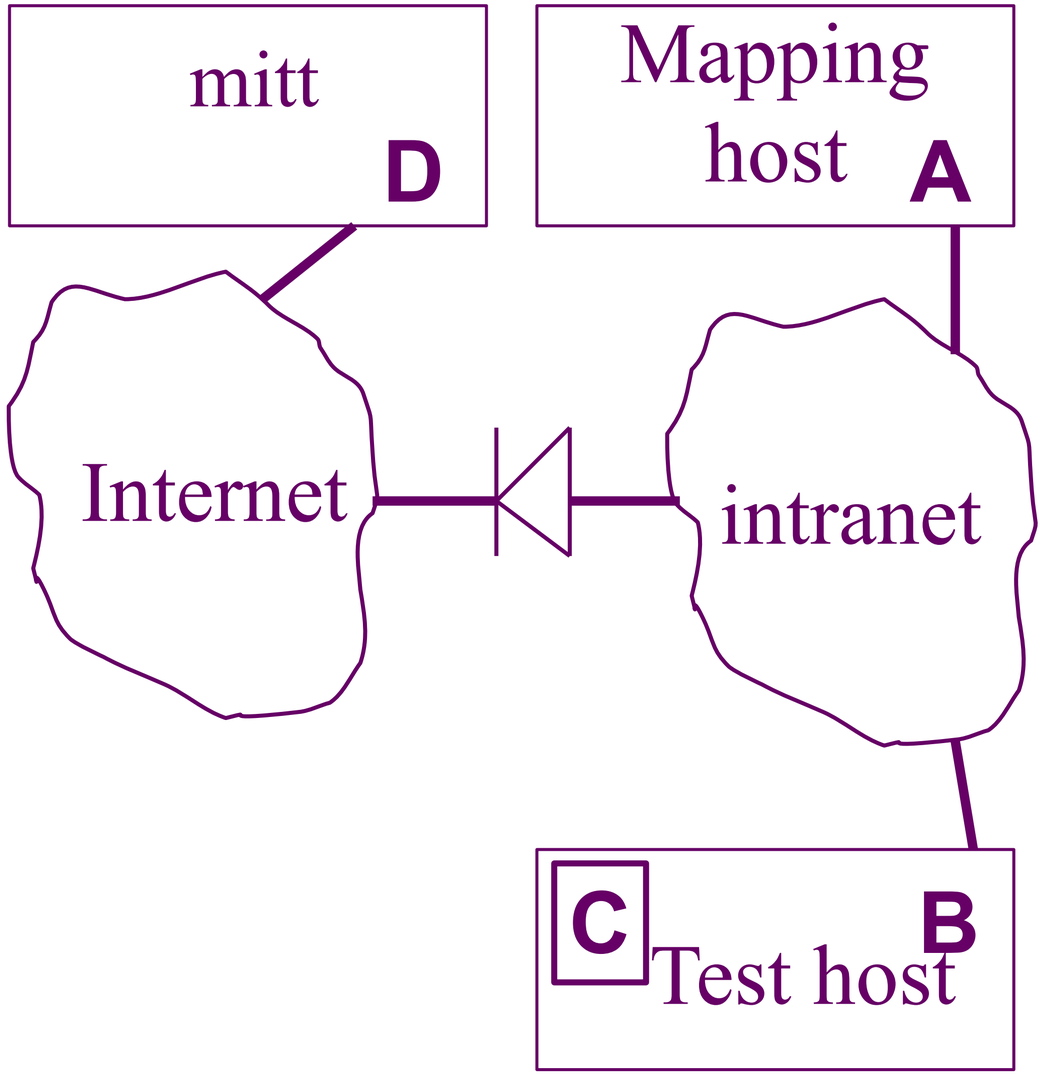
- **“Dirty words”, data spills**
- **Covert channels are a big deal**
- **eat a thumb drive**
- **glue in the USB slots?**
- **traffic analysis: Gordon Welchman at Bletchley. “metadata”**
- **FBI wiretap jokes**
- **Chaum networks (TOR) don’t work if they are infiltrated**
- **leak detection**
- **How 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?**

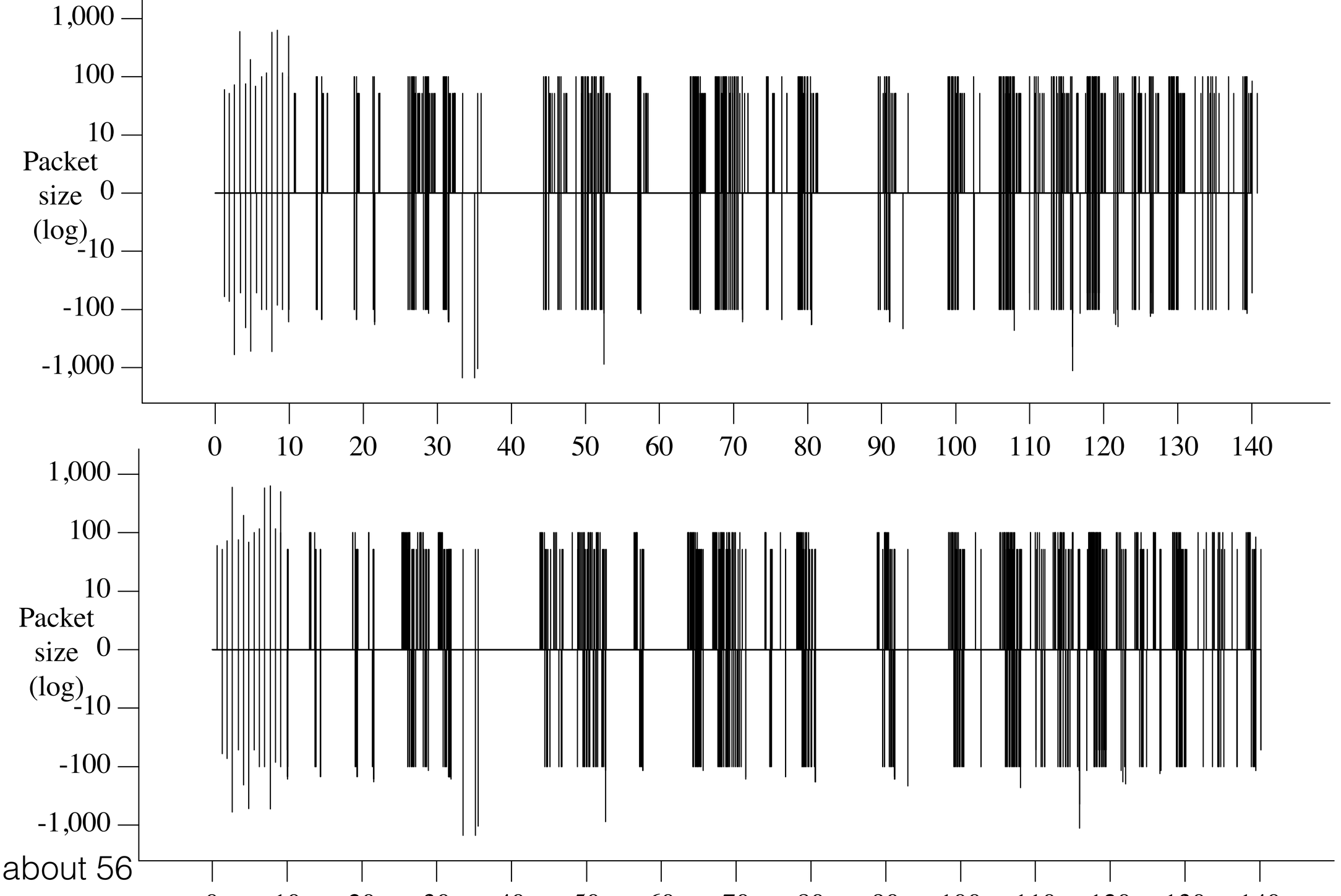
Leak detection

- **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**


1,000
100
10
Packet
size
(log)
0
-10
-100
-1,000
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
1,000
100
10
Packet
size
(log)
0
-10
-100
-1,000

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**

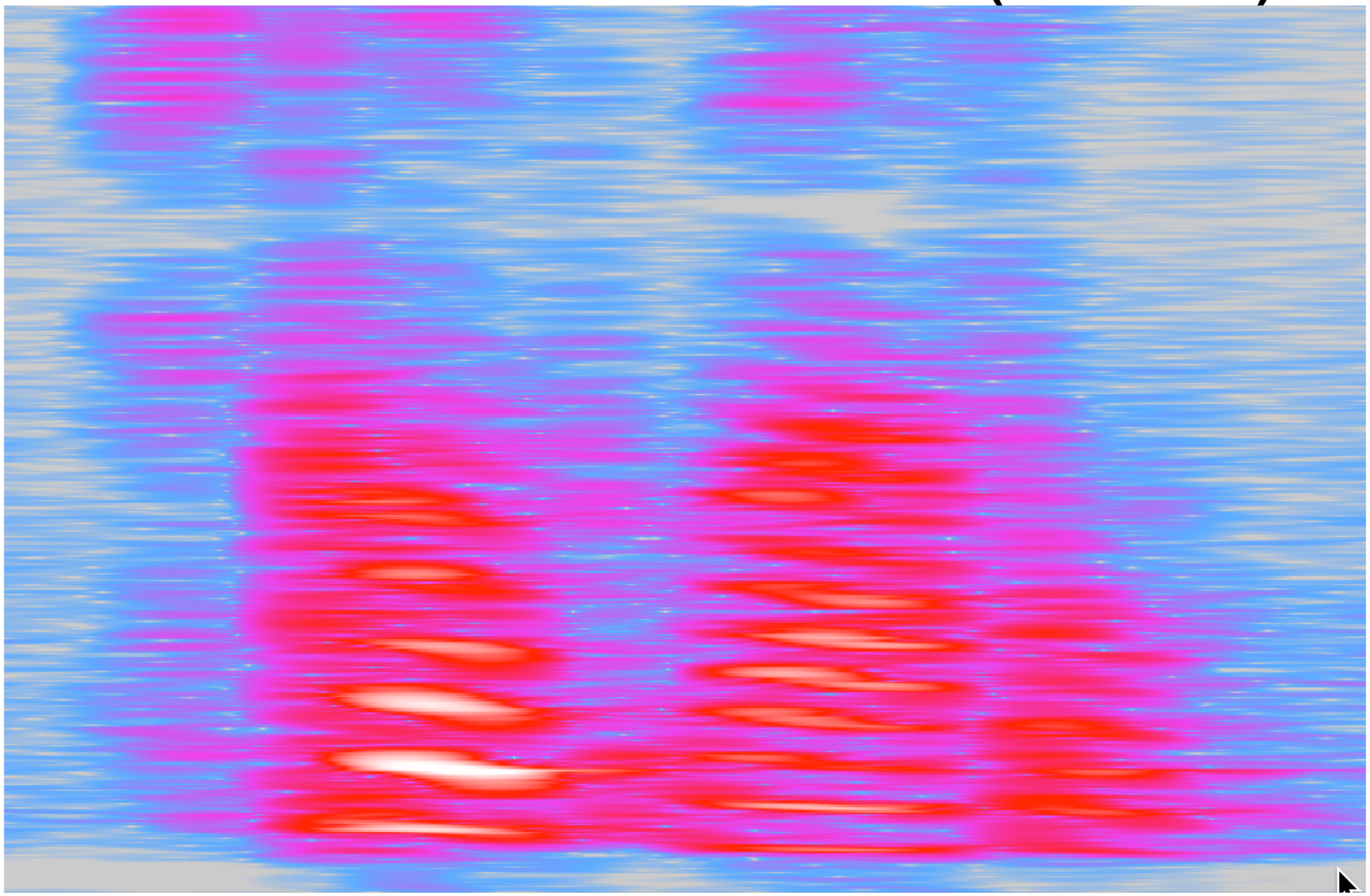
Public data

- **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/**

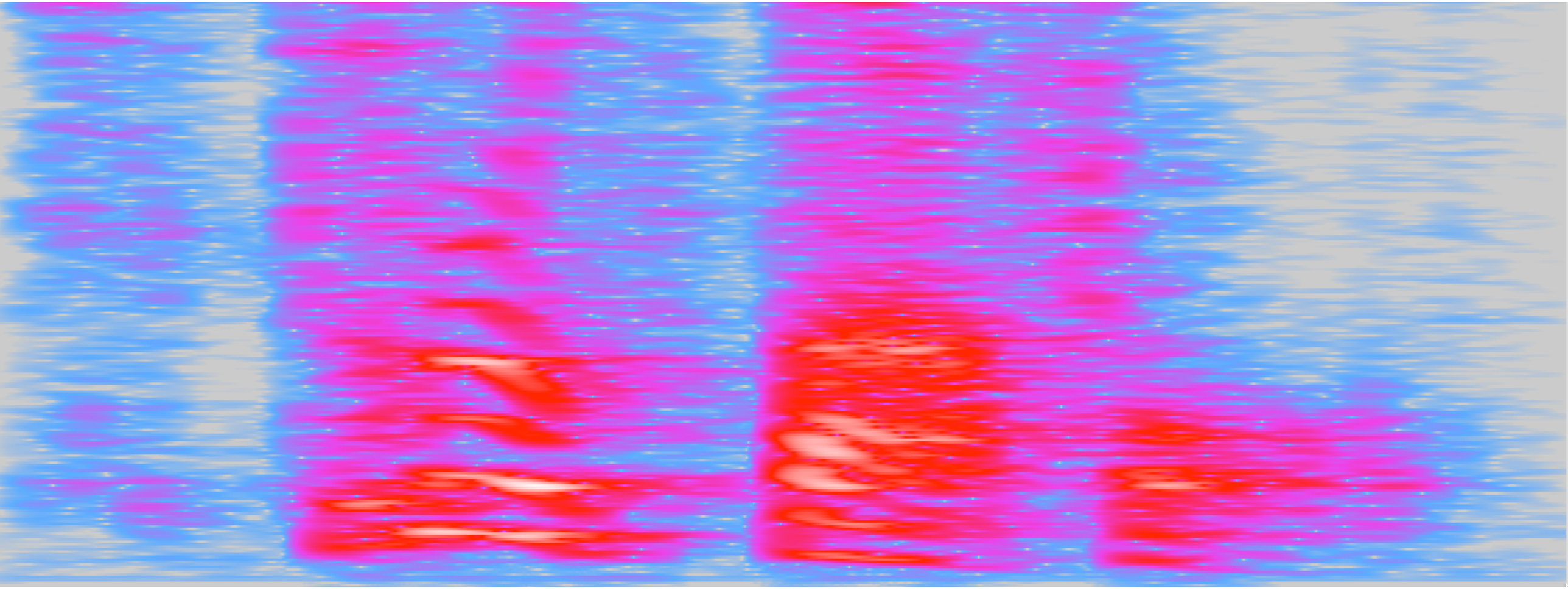
Complicated data: neural nets

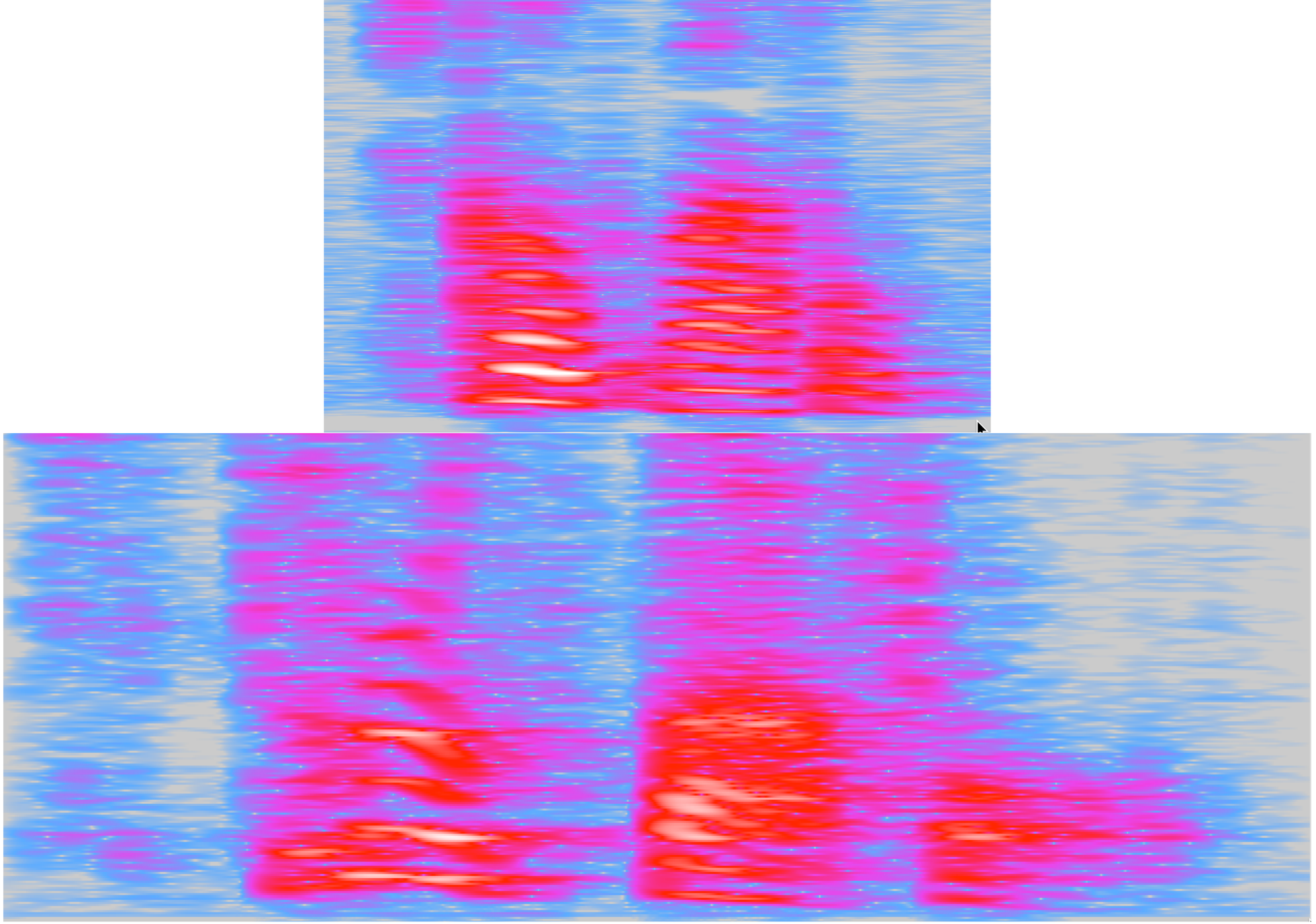
- **It is mostly about processing data**
- **“ten hundred thousand fingers have fingers.”**

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