Massive Data Sets: Challenges for Statistics

Karen Kafadar

Indiana University

kkafadar@indiana.edu

`http://math.cudenver.edu/~kk`

Acknowledgements:

E.J. Wegman, GMU; D. Marchette, NSWC; G. Davis, UVA;

R.L. Jacobsen, UCB

AFOSR F49620-01-1-0274 (EJW); ARO D07-xxx (KK)

OUTLINE

1. Motivation: Massive data from Genomics, Internet, High-energy physics; Fraud detection literature
2. Guiding principles; Take-home messages
3. Genomics: Microarray data; Multiple testing Multiple testing
4. Internet traffic data (Sample data from GMU)
   IP addresses, ports, message sizes
5. High-energy physics experimental data
   Classification, Tiny signals in massive background
6. Morals: Computing, Analysis, Inference
7. Further research
   New framework for inference from massive data sets?

1. Motivation: Massive Data

- Genomics:

  Measure mRNA expression in response to stimulus from 50K-70K genes/person

  Goal: Identify function of genes

  Size: 0.5–1.0 Million/experiment (50K $\times$ 20 people)

  Relatively easy: Specified hypotheses

- Internet:

  Cybersecurity, computer viruses, network attacks: Scientific computing, financial transactions, business operations, security procedures, ...

  Goal: Detect attacks *before* they force shutdowns

  Size: Thousands of packet transmits per minute

- High-energy Physics (HEP):
  - Colliding beams of electrons (SLAC) or protons (CERN) accelerated at very high energies (MeV/GeV/TeV)
  - Collisions yield short-lived particles that decay into more short-lived particles in any one of 100,000 ways ("events")
  - Most events well-characterized (particles, speeds, lifetimes)
  - Others less well understood (e.g. those with b-mesons)

Goal: Find target events of interest amidst millions of "uninteresting" events

Size: Millions per minute

Common Theme: **Tiny signal in vast sea of noise**

Detect abnormal behavior: disease surveillance; nuclear product mfg; phone/charge card usage; financial transactions; ...

Easier when:

- data streams stratify into smaller data sets
- smaller sets are roughly independent of each other
- sets can be modeled simply and parametrically
- nature of potential abnormality is well-characterized
- residual distribution is well specified $\Rightarrow$ assess *probability* of abnormality
- SPC-type tools are applicable

Genomics: Many hypotheses – but straightforward

Internet: Many vaguely-specified hypotheses (outliers, excess packets/transmissions, signals of potential attacks

HEP: Too many partially-specified hypotheses to characterize — so the best we may be able to do is to reduce background (EDA) – i.e., find a rough measure of "uninteresting" events and remove them (rates 1%, 0.3%, ...)

Features of Internet traffic data:

- Relentless ("streaming")
- Not independent of other systems: thousands of messages from thousands of ports/addresses each *minute*
- Diverse (text, numeric, image)
- Dispersed (geographically)
- Data often not from some convenient mathematical pdf

What data should be collected?

How can anomolies be detected?

Features of HEP data:

- Relentless ("streaming")
- "Events" are *assumed* to be relatively independent of each other (e.g., occurrence of event type A tells us nothing about the probability of event B occurence)
- High-precision measurements — when we can see them
- Mis-identified tracks associated with events
- Data are *not* from a convenient mathematical pdf

An example of an "inconvenient" mathematical pdf (from high-energy physics):

$$f(q^2 = (p_\ell + p_\mu)^2, \cos\theta_\ell, \cos\theta_\nu, \chi) =$$
$$\frac{3G_F^2|V_c b|^2}{8(4\pi)^4}\frac{\rho_{D*}}{(q^2-m_l^2)^2}q^2 B \times [((1+\cos^2\theta_\ell) + \frac{m_\ell^2}{q^2}\sin^2\theta_\ell)\sin^2\theta_V(|H_+|^2 +$$
$$|H_-|^2) + 4(\sin^2\theta_\ell + \frac{m_\ell^2}{q^2}\cos^2\theta_\ell)\cos^2\theta_V|H_0|^2 - 2\cos\theta_\ell\sin^2\theta_V(|H_+|^2 -$$
$$|H_-|^2) - 2(1-\frac{m_\ell^2}{q^2})\sin^2\theta_\ell\cos 2\chi\sin^2\theta_V Re(H_+H_-^*) + (1 -$$
$$\frac{m_\ell^2}{q^2})\sin 2\theta_\ell\cos\chi\sin 2\theta_V Re(H_+ + H_-)H_0^* -$$
$$\sin\theta_\ell\cos\chi\sin 2\theta_V Re(H_+ - H_-)H_0^* + 4\frac{m_\ell^2}{q^2}\cos^2\theta_V|H_t|^2$$

(Can be simplified by assuming H_-, H_0, H_+ are real)

Graphics and visualization are critical

Fraud Detection Literature

- *General overview*: R.J. Bolton, D.J. Hand (2002), "Statistical fraud detection: A review" (with discussion), *Stat. Sci.*
- *Experimental design*: M. Schonlau, W. DuMouchel, W-H Ju, A.F. Karr, M. Theus, Y. Vardi (2002), "Computer intrusion: Detecting masquerades," *Stat. Sci.*: designed experiments to evaluate algorithms for detecting masquerade user (stratify by user, identify characteristic features of user's "signature")
- *Modeling*: W.S. Cleveland, D.X. Sun (2000), "Internet traffic data," *JASA*: Models for times between web accesses and challenges of long-range dependence and stationarity

- *Telephone calling fraud* (stratify by user): K.C. Cox, S.G. Eick, G.J. Wills (1997), "Visual data mining: Recognizing telephone calling fraud;" C. Cortes, D. Pregibon (2001), "Signature-based methods for data streams," *Data Mining and Knowledge Discovery Data Mining and Knowledge Discovery*:
- *Nuclear product manufacturing*: Spiegelman+Rosenblatt 1984
- *Disease surveillance*: Siegrist et al. 2004, Stroup et al. 1989, Waller and Gotway 2004
- *Bioterrorism*: Hutwagner et al. 2003, many others
- *Visualizing network data*: S. Krasser et al. 2005: "Real-time and forensic network data analysis using animated and coordinated visualization" *IEEE Workshop on Information Assurance, USMA*: PC/time plots

New data types/structures have lead to advances in statistics (EJW, PJH)

- Data from agricultural expts $\Rightarrow$ Design of experiments
- "Large" data sets $\Rightarrow$ Statistical graphics
- No specified probability distribution $\Rightarrow$ Nonparametrics
- Non-Gaussian distributions $\Rightarrow$ Robust methods
- Many-featured data $\Rightarrow$ Multivariate statistics/displays
- Clinical trials $\Rightarrow$ Sequential analysis
- Testing many hypotheses $\Rightarrow$ Multiple comparisons
- Many other examples ...

"Take-Home Messages"

- Prevalence of *streaming* data will increase
- Basically unusuable in raw form; require much pre-processing
- Detecting "exotic" requires characterizing "typical"
- Streaming data offer new challenges for Statistics/CS:

  data value: what data to collect/discard (STAT)

  data warehouse: acquisition, storage, distribution (CS)

  tools, algorithms for pre-processing (STAT, CS)

  new methods for analysis – robust, "sufficient" (STAT)

  new informative visual graphical displays (STAT, CS)

Guiding Principles

Lessons from EDA

"... 'exploratory data analysiss' is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as for those we believe might be there. Except for emphasis on graphs, its tools are secondary to its purposes.... the tool-kit of exploratory data analysis is, and must remain, open-ended."

"Data analysts regard their models as a basis from which to measure deviation, as a convenient bench mark in the wilderness, expecting little truth and relying on less."

– JWT, "Comment" (Parzen), *JASA* 1979, pp.121-122

"Statistics is 'reactive" — very responsive to new problems that arise in chemistry, biology, physics, ..." — P. Hall

"Advances in powerful computing equipment has had a dramatic impact on statistical methods and theory. It has changed forever the way data are analyzed" – P. Hall

"Far better an approximate answer to the ***right*** *question, which is often vague, than an* ***exact*** *answer to the wrong question, which can always be made precise"* – JWT 1962, p.13

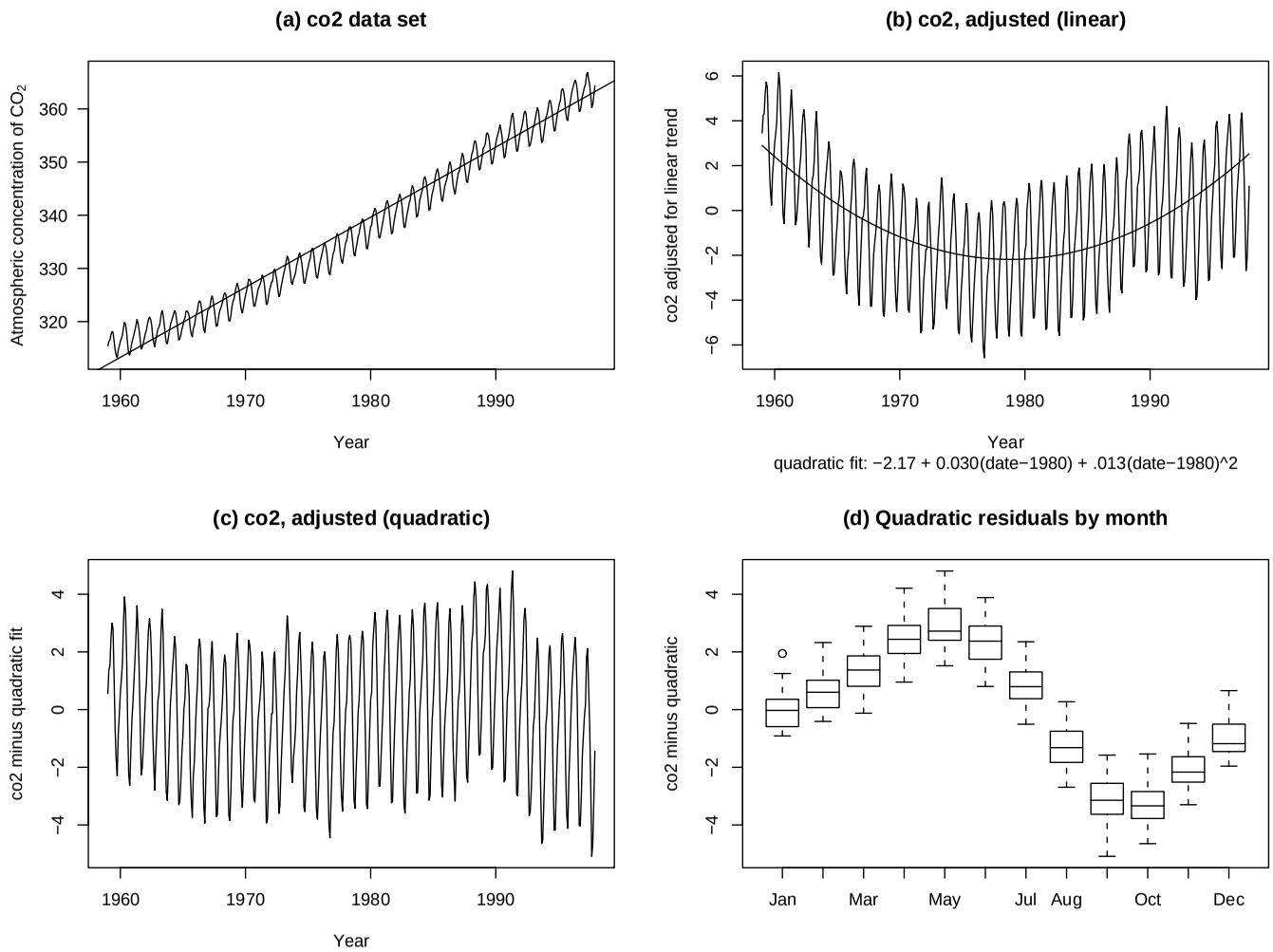
"Better a precise measure of something important than an imprecise measure of something unimportant" – D. Byar

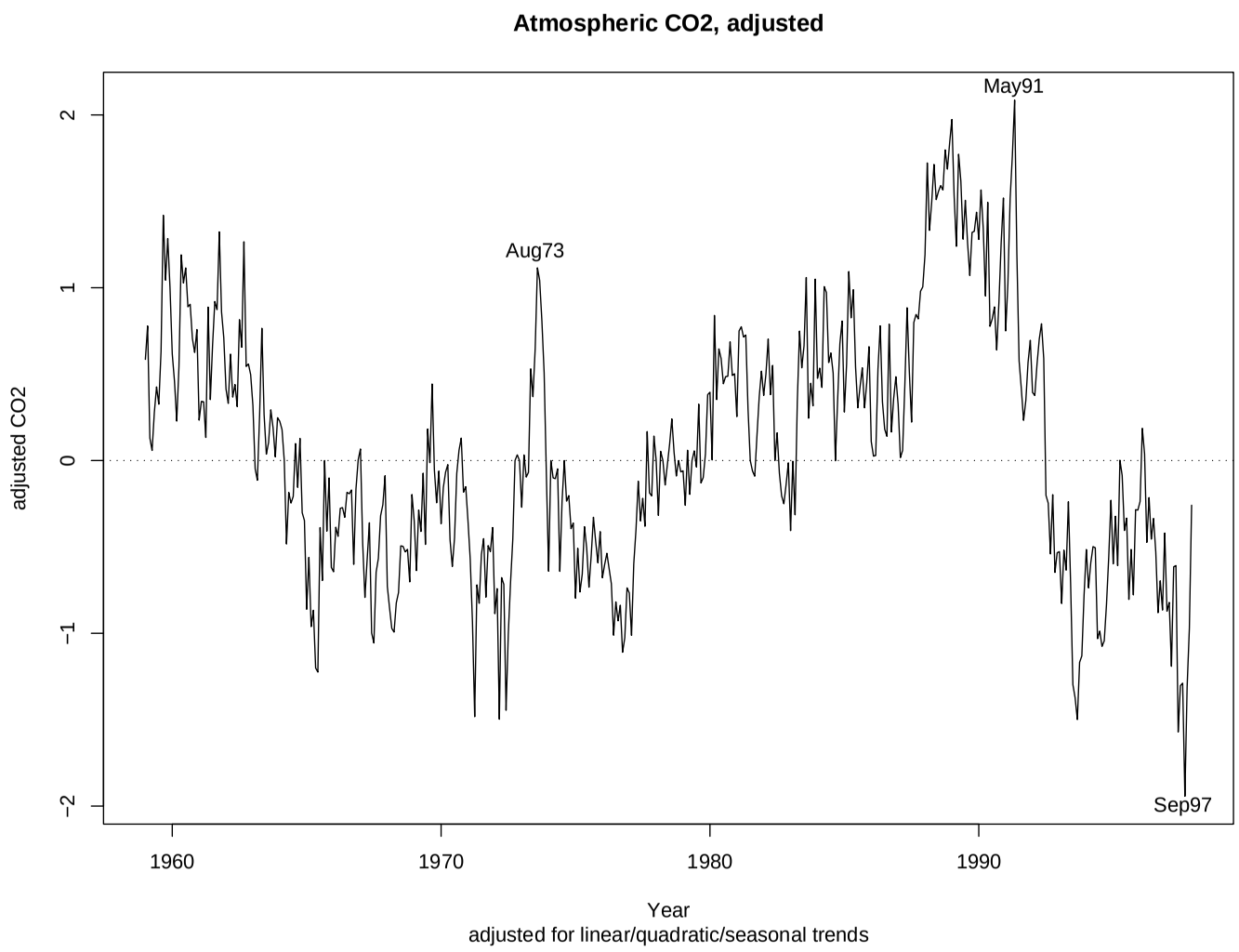
"**The greatest value of a picture** is when it *forces* us to notice **what we never expected to see**" – JWT, *EDA*, p.vi

"Rx":

- Start with simplest of models
- Remove the obvious, magnify the residual
- Iterate robustly, note non-conforming pieces
- Control $E(missed)$, not $P\{missed\}$ — by piece
- Make good use of graphical displays
- "Cognostics", "scagnostics" – diagnostics from cognitive/scatterplot displays (Tukey + Tukey 1985)

"Graphs will certainly be increasingly 'drawn' by the computer without being touched by hands. More and more, too, as better procedures of diagnosis and indication are automated, graphs, and other forms of expository output, will, in many instances, cease to be the vehicle through which a man diagnoses or seeks indications, becoming, instead, the vehicle through which the man supervises, and approves or disapproves, the diagnoses and indications already found by the machine." – JWT 1962, p.60


(a) co2 data set
Atmospheric concentration of CO_2
360
350
340
330
320
1960
1970
1980
1990
Year
(b) co2, adjusted (linear)
co2 adjusted for linear trend
6
4
2
0
−2
−4
−6
1960
1970
1980
1990
Year
quadratic fit: −2.17 + 0.030(date−1980) + .013(date−1980)^2
(c) co2, adjusted (quadratic)
co2 minus quadratic fit
4
2
0
−2
−4
1960
1970
1980
1990
Year
(d) Quadratic residuals by month
co2 minus quadratic
4
2
0
−2
−4
Jan
Mar
May
Jul
Aug
Oct
Dec


Atmospheric CO2, adjusted
May91
Aug73
Sep97
adjusted CO2
2
1
0
−1
−2
1960
1970
1980
1990
Year
adjusted for linear/quadratic/seasonal trends

3. Microarray data

- DNA (genetic code: A,C,G,T) in cell **nucleus**
- triplets of nucleotides code for **amino acids**
  (ex: AAA or AAG codes for Lysine)
- Genes = organized strings of nucleotides (in triplets)
- Genes code for proteins (e.g., insulin)
- Proteins made in cell **cytoplasm**
- Cell needs genes into cytoplasm (only coding part of DNA)
- DNA copy (cDNA: AT/CG); Excise introns; T $\rightarrow$ U (mRNA)
- **Gene expression = measure of mRNA concentration**
  — which *may* correlate with protein production

Ex (Kim Kafadar): Ca^{2+} signaling in yeast cells

Ca^{2+} needed for cells to survive stresses (e.g., high salt, alkaline pH, cell wall damage); promotes signaling through calcineurin (protein phosphatase)

What does calcineurin do?

dephosphorylates & activates Crz1p/Tcn1p/Hal8p transcription factor; Crz1p accumulates in cell nucleus, activates gene transcription whose products promote adaptation to stress

What inhibits Crz1p?

Start with Ca^{2+}

- Grow 2 batches of cells: w/Ca^{2+}, w/o Ca^{2+}
- Wait 30 ($\pm$) min; harvest cells, collect mRNA
- Reverse-transcribe mRNA back to cDNA (U $\rightarrow$ T)
- Denature ("un-zip" — single strand)
- Label "no-Ca" cells w/green flourophore (532nm)
- Label "Ca" cells w/red flourophore (635nm)
- Hybridize both to cDNA slide

cDNA slide:

- Single-strand copy of each gene ("probe") printed in defined locations ("spots")
- Ca/no-Ca flouorophore-labeled cDNA mixture placed on slide
- cDNAs in mixture find matching partners
- match $\Rightarrow$ binding energy $\Rightarrow$ radiates (532 if green; 635 if red)
- Laser scanner records fluorescence by pixel
- "Red" if more red-tagged mRNA (Ca batch)
- "Green" if more green-tagged mRNA (no-Ca batch)
- "Yellow" if not much difference

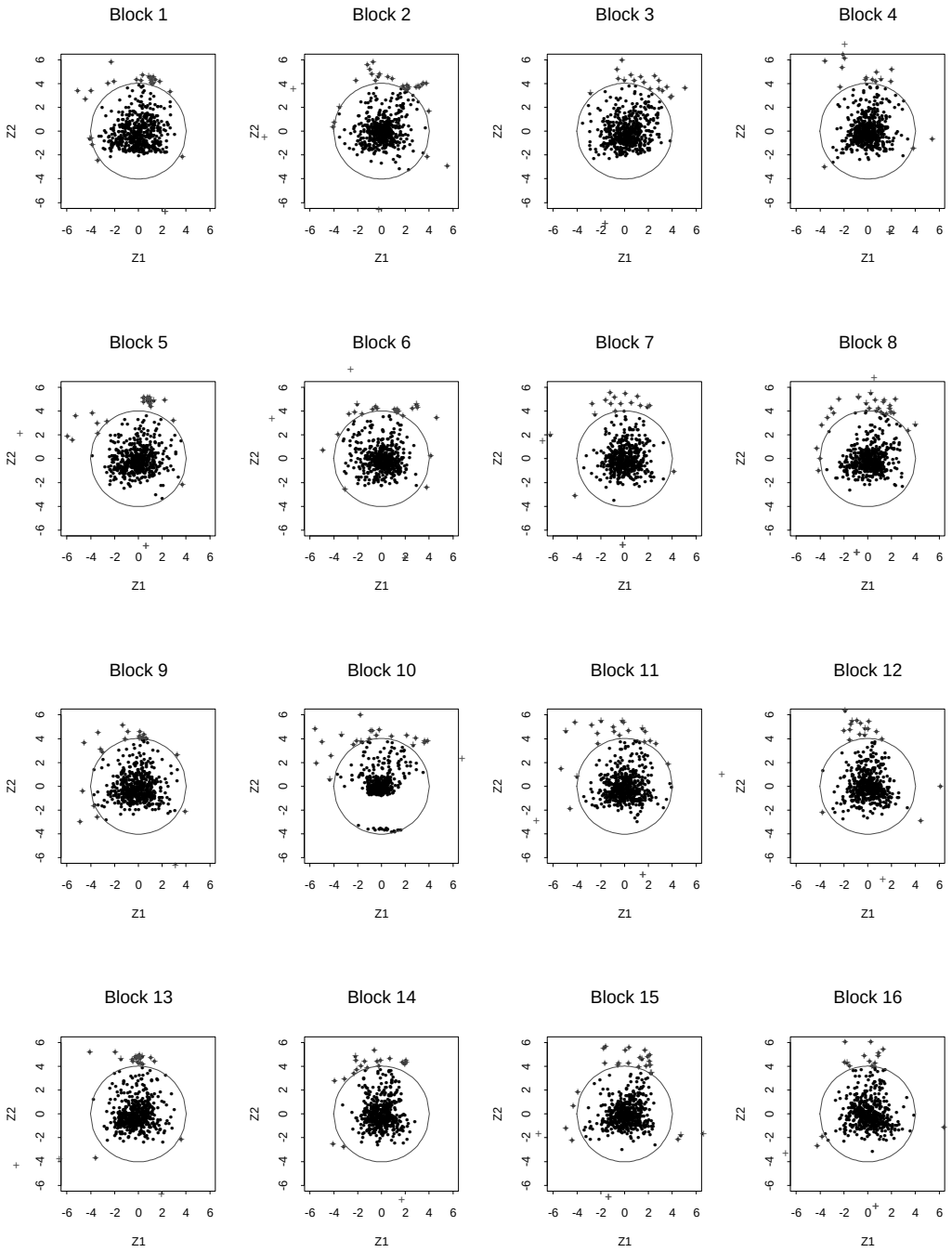
Which spots (genes) have more red than green? Multiplicity

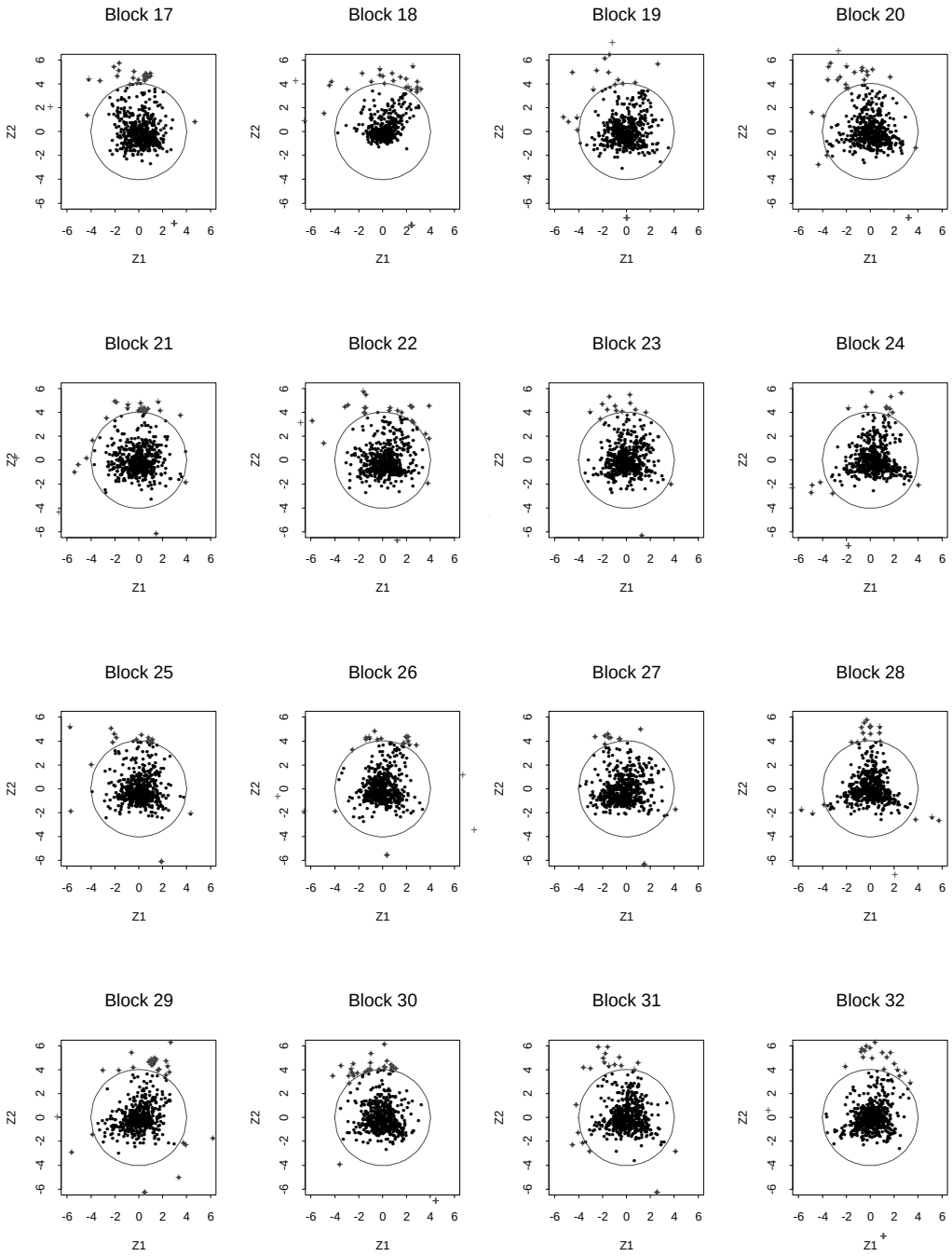
- Bonferroni not practical (α/N when $N = 50,000$ is $10^{-6} \Rightarrow z_{crit} = 4.89$)
- MC procedures control P{$\geq$ 1 False Positive} $\leq \alpha$
- Benjamini & Hochberg (JRSSB 1995): FDR Control ***Expected Proportion of false positives*** $\leq \alpha$
- We know that focusing on only the *mean* of a distribution (e.g., $W = \#$ significant by chance alone $\sim Bi(N, \alpha)$) is simpler than focusing on a quantile (100(1-α)%-point of W)

Simplified: slides "spot" genes in "blocks" of $\sim 500 - 1000$

Transform (red,green) to standard bivariate Gaussian; look for genes outside circle

Process issues; Big p, small n; correlation among p (genes)


Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Block 8
Block 9
Block 10
Block 11
Block 12
Block 13
Block 14
Block 15
Block 16
Z1
Z2
-6 -4 -2 0 2 4 6


Block 17
Block 18
Block 19
Block 20
Block 21
Block 22
Block 23
Block 24
Block 25
Block 26
Block 27
Block 28
Block 29
Block 30
Block 31
Block 32
Z1
Z2

4. Internet Traffic Data

Collected from anonymous surveillance machines outside "firewall" to monitor incoming/outgoing traffic [Marchette 2001, *Computer Intrusion Detection and Network Monitoring*]

- All internet communications are transmitted via packets
- Fundamental unit of information is **packet**
- Packet consists of data and headers that control the communications via IP, TCP
- **Flow** = exchange of packets between source-destination pair
- **Connections** = collections of flows (these data)
- After much pre-processing, data file has summary statistics on size/duration of **flows** (millions per hour)

4. Sample Data Set From George Mason Univ.

length	SIP	DIP	DPort	SPort	Npkt	Nbyte
0.23	4367	54985	443	1631	9	3211
0.27	18146	9675	3921	25	15	49
0.04	18208	28256	1255	80	6	373
1389.10	24159	17171	23	1288	845	5906
373.99	60315	37727	2073	80	1759	834778
0.13	28256	18208	80	1256	10	816
1498.11	25699	4837	9593	80	65803	35661821
0.04	18208	28256	1251	80	5	373
122.38	54985	4179	1298	443	99	85559

TCP

- Instructions for delivering/sequencing packets coming from one machine, destined for another
- Data passed through "ports" (logical, rather than physical, location; identifies connection through which data are passed between machines)
- 2^{16} = 65,536 ports per machine
  - 2^{10} Ports 0–1023: "well-known ports"
  - Registered Ports 1024–49151 (e.g., 2049 for Sun's `nfs`)
  - 2^{14} Ports 49152–65536: dynamic/private ports
- Unprotected ports are prime candidates for intrusion, so monitor amount of traffic in/out of ports

- Among the 2^{10} "well-known" Ports 0–1023:

21	`ftp`	file transfer protocol
22	`ssh/scp`	secure shell/copy
23	`telnet`	network connection
25	`smtp`	mail transmission protocol
80	`http`	conventional web port (also 8080)
110	`pop3`	pop3 mail
443	`https`	secure web encryption
554	`rtsp`	real-time streaming video/audio

- `DPort` = **Destination Port**; `SPort` = **Source Port**

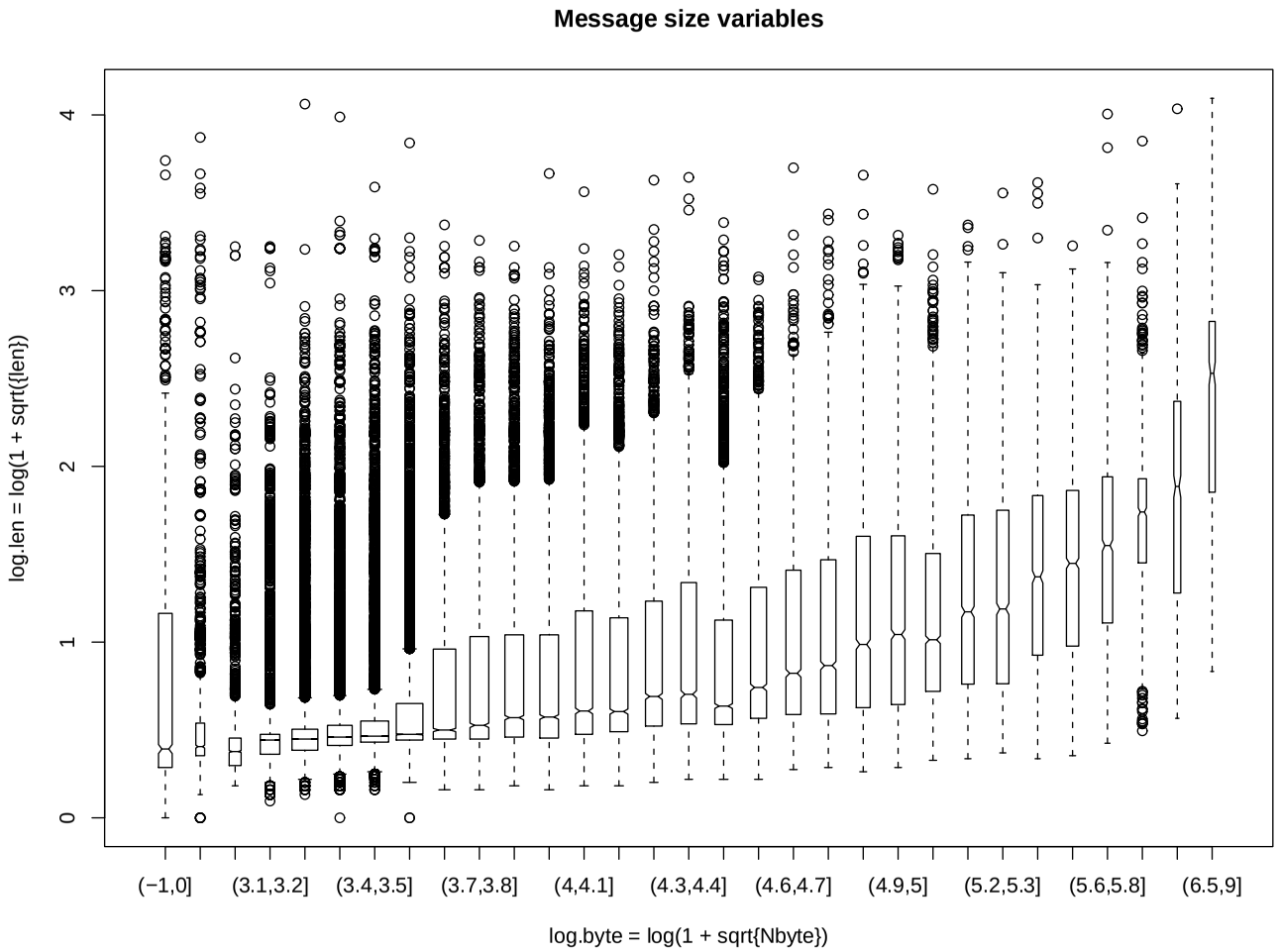
Message sizes

Three (correlated) measures of "size" of session:

- `duration` or `len` = duration (length) in seconds
- `Nbyte` = Number of bytes
- `Npkt` = Number of packets
- Highly skewed distributions: use log transformation
  $f(x) = \log(1 + \sqrt{x})$: `log.len, log.byte, log.pkt`
  ($\log(x)$ spreads many small x's too much)
- Potentially suspicious:
  Few packets, each with many bytes
  Many packets, each with very few bytes

Summary statistics over 135,605 records (1 hour in 2002)

	len	SIP	DIP	DPort	SPort	Npkt	Nbyte
min	0.0	259	259	20	20	2	0
10%	0.2	4930	4024	80	1187	9	568
25%	0.3	9765	8705	80	1369	10	860
med	0.6	20258	25164	80	1849	12	1832
75%	3.8	41282	45900	80	3681	21	7697
90%	21.5	62754	58202	80	10000	45	25161
max	3482.5	65276	65262	10000	10000	65803	35mil
#!	9101	2504	5139	380	6742	1056	29876


Message size variables
log.len = log(1 + sqrt{len})
0
1
2
3
4
(−1,0]
(3.1,3.2]
(3.4,3.5]
(3.7,3.8]
(4,4.1]
(4.3,4.4]
(4.6,4.7]
(4.9,5]
(5.2,5.3]
(5.6,5.8]
(6.5,9]
log.byte = log(1 + sqrt{Nbyte})

5. 'Visual Analytics': Exploratory Plots

Detecting "exotic" requires characterizing "typical"

- Size variables are highly skewed (already seen)
- Boxplots suggest exceesive numbers of "outliers"
- Need better display of distribution

Letter value plots

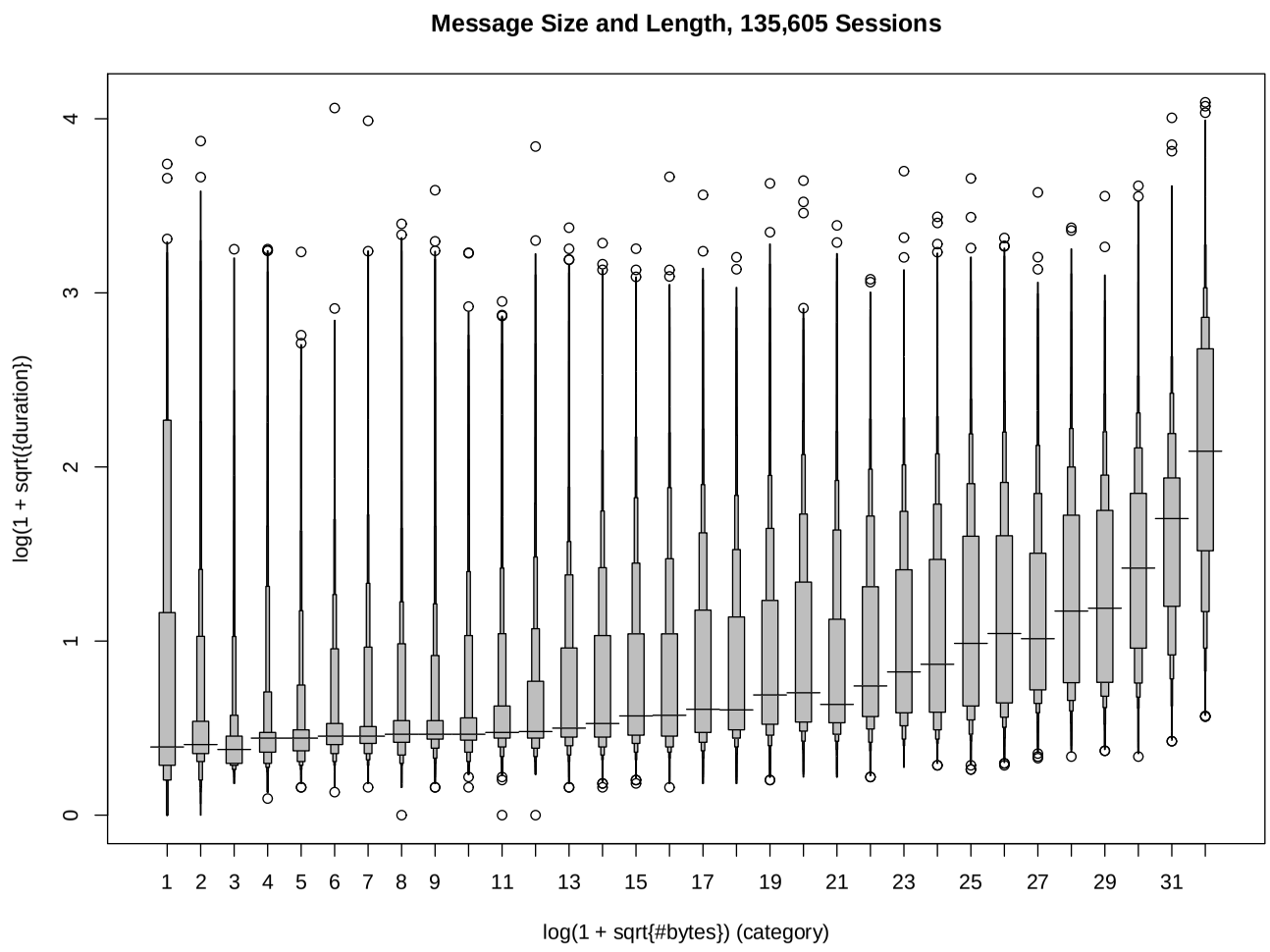
Letter value Displays

Estimate quantiles corresponding to tail areas 2^{-k}:

- Median (2^{-1}): depth $= d_M = (1 + n)/2$
- Fourths (2^{-2}): depth $= d_F = (1 + \lfloor d_M \rfloor)/2$
- Eighths (2^{-3}): depth $= d_E = (1 + \lfloor d_F \rfloor)/2$
- Etc. until extreme (depth $= 1$)
- When d_L is a half-integer, average two adjacent order statistics
- Actual tail area is closer to $(d_L - \frac{1}{3})/(n + \frac{1}{3})$ [*UREDA* §2G]
- Aymptotic correlation between adjacent LVs $\approx \sqrt{\frac{1}{2}} = 0.707$

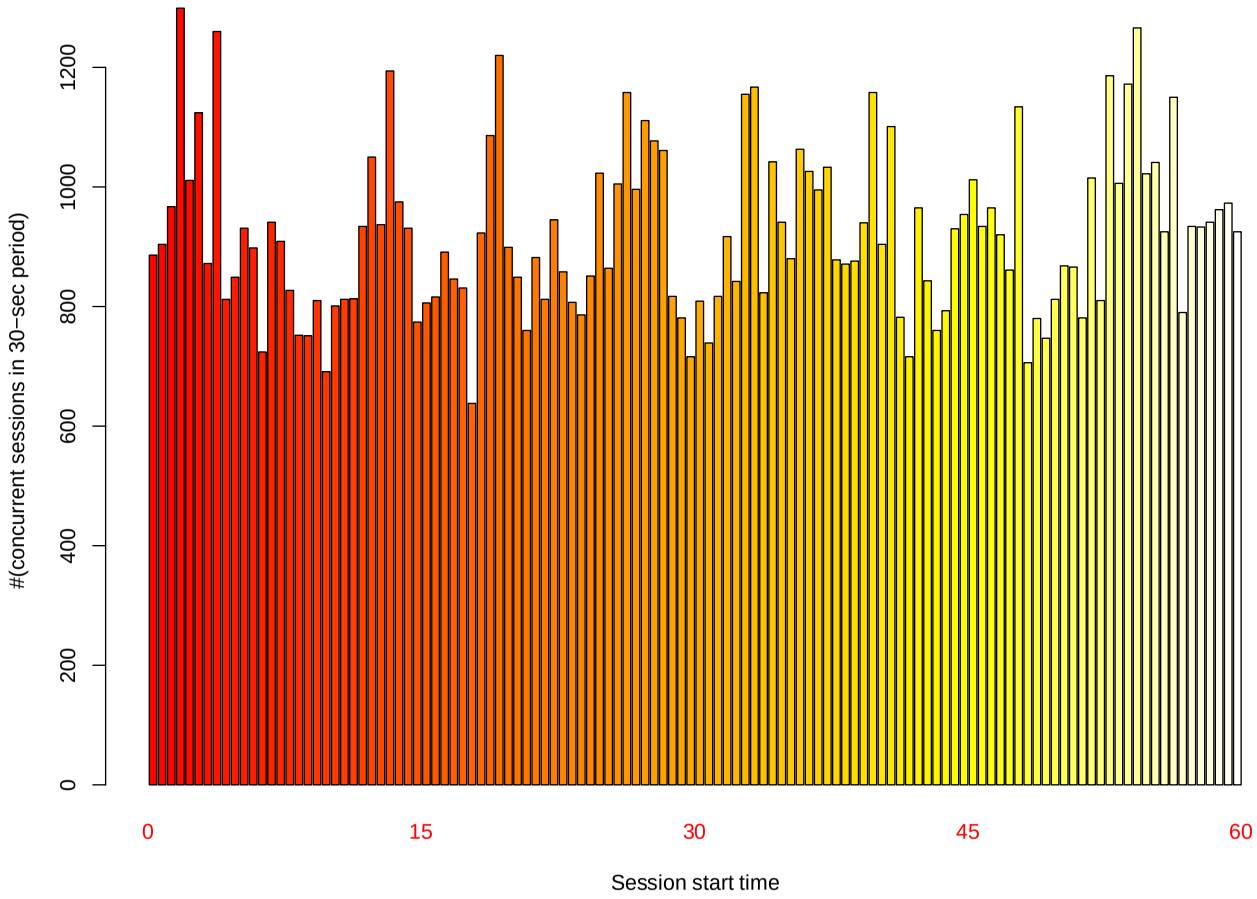
Letter value box plots

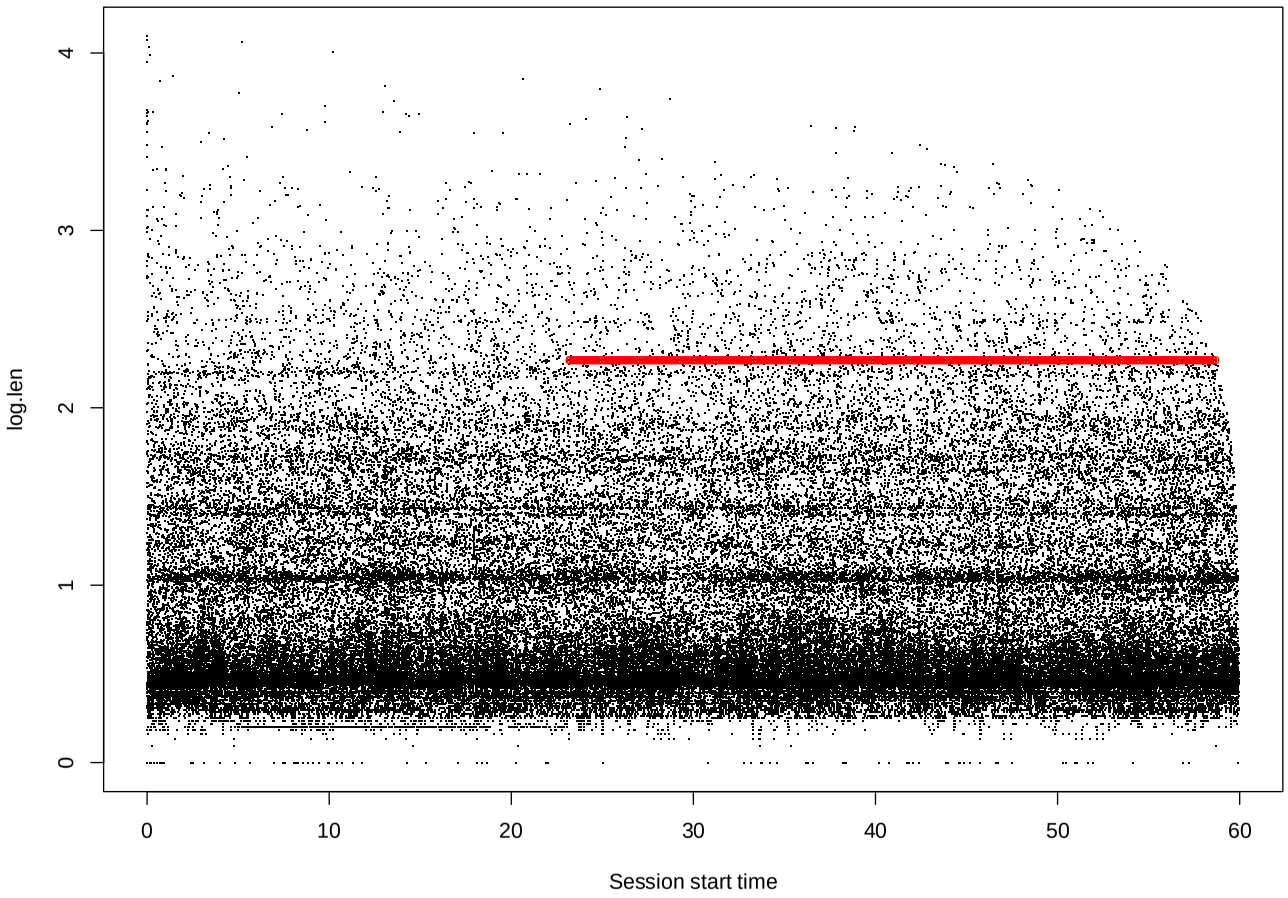
- Small data sets: Limited information about tails
- Boxplots show fourths, *extent* of data beyond fourths
- Large data set: Tail quantiles more reliable
- $\Rightarrow$ Extend boxplots to include more letter values beyond median, fourths
- "Stopping Rules": how many LVs to show?
- How to display letter values?
- Which observations are labeled as 'outliers'?
- **Plot still shows only actual data values**

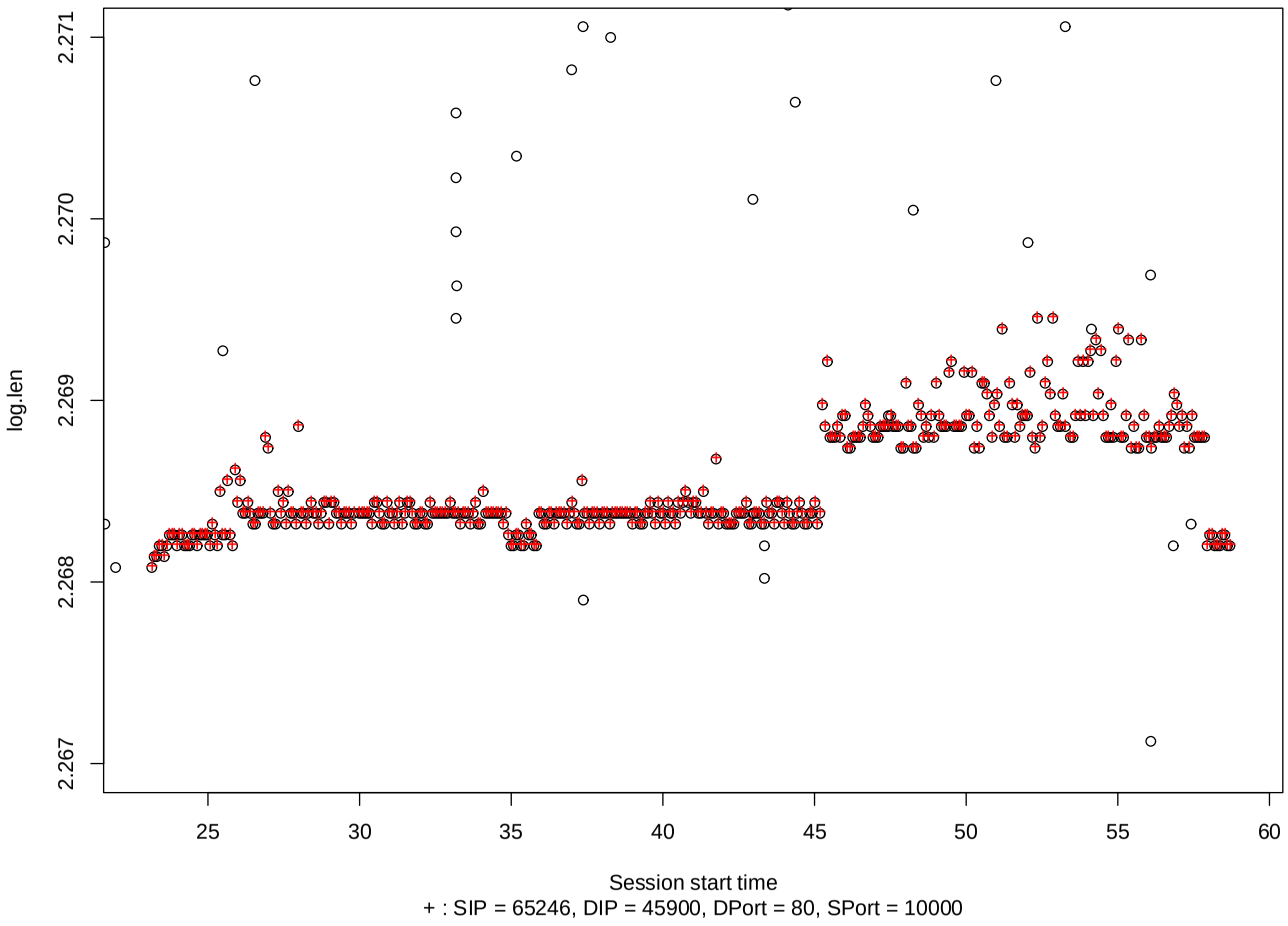

Message Size and Length, 135,605 Sessions
log(1 + sqrt{duration})
0
1
2
3
4
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31
log(1 + sqrt{#bytes}) (category)

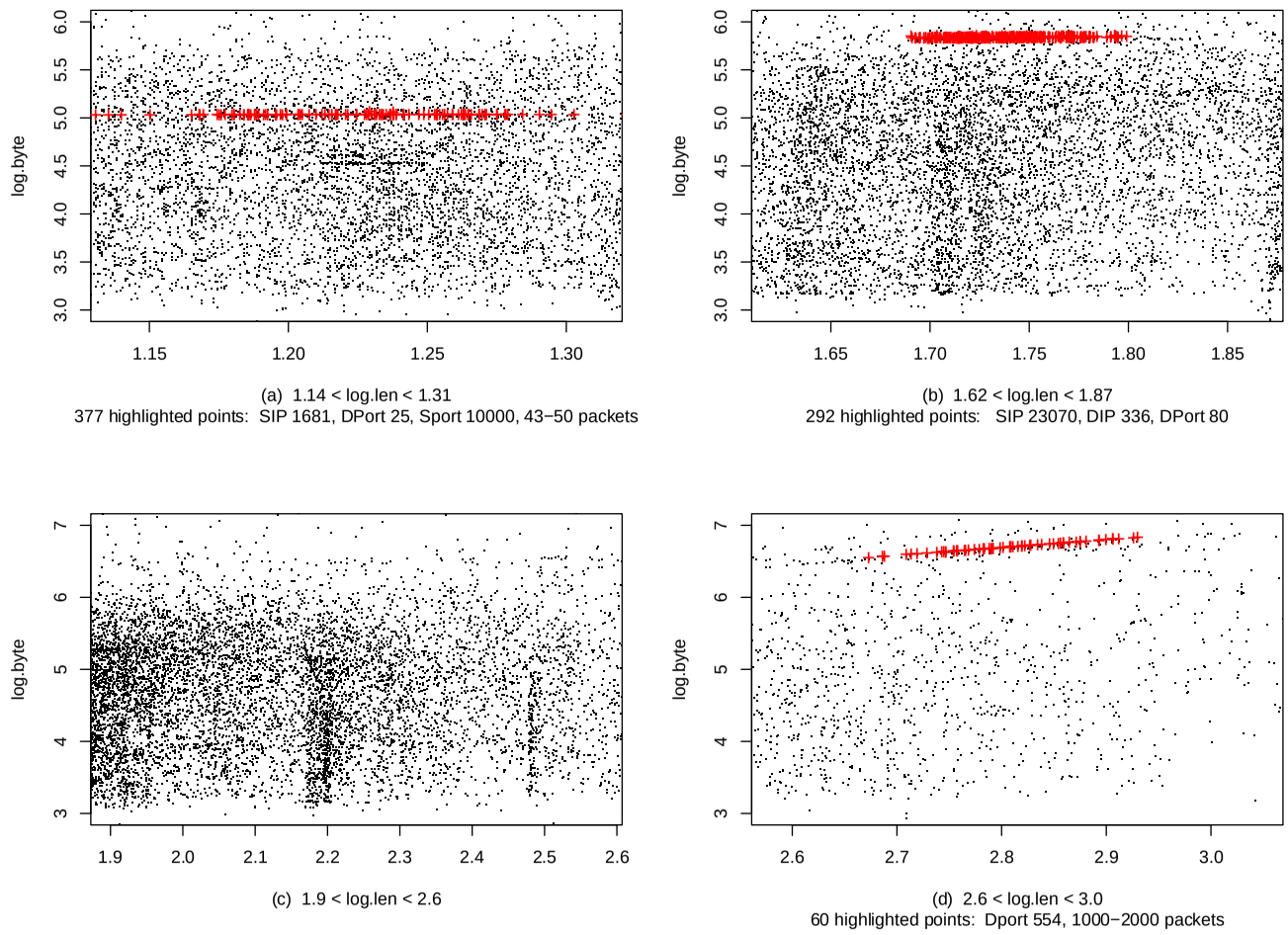
More exploratory plots

- Preponderance of relatively short sessions (next)
- Number of active sessions in 120 successive 30-second non-overlapping intervals (mean 923, SD 140, $\approx 3\sigma$ limit 1343, max 1299 = expected max of 120 N(923,140^2) variates)
- Plot of `log.len` vs *Session Start Time* should be relatively uninteresting


#(concurrent sessions in 30−sec period)
0
200
400
600
800
1000
1200
0
15
30
45
60
Session start time


log.len
0
1
2
3
4
0
10
20
30
40
50
60
Session start time


log.len
2.267
2.268
2.269
2.270
2.271
25
30
35
40
45
50
55
60
Session start time
+ : SIP = 65246, DIP = 45900, DPort = 80, SPort = 10000

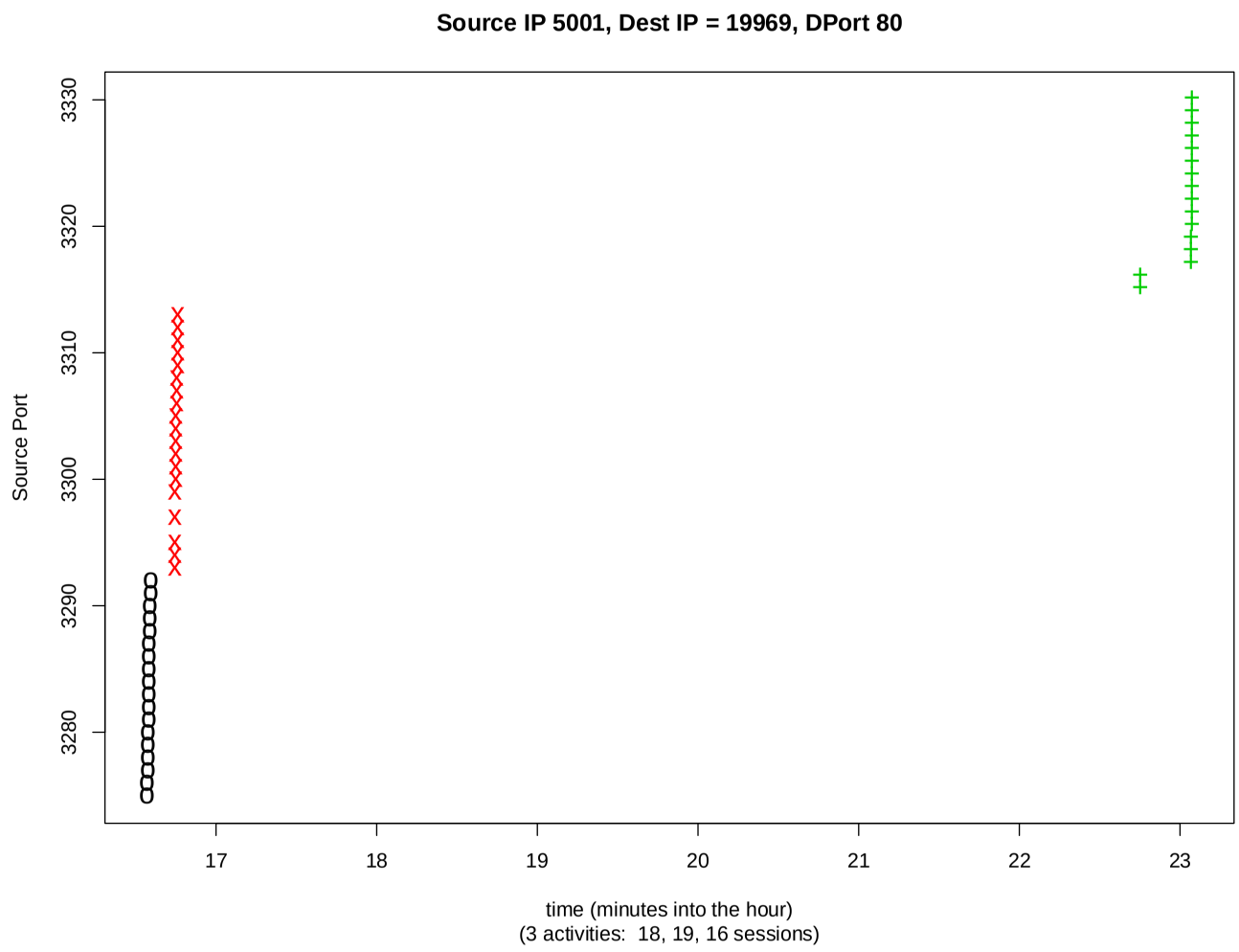

log.byte
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1.15
1.20
1.25
1.30
(a) 1.14 < log.len < 1.31
377 highlighted points: SIP 1681, DPort 25, Sport 10000, 43−50 packets
log.byte
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1.65
1.70
1.75
1.80
1.85
(b) 1.62 < log.len < 1.87
292 highlighted points: SIP 23070, DIP 336, DPort 80
log.byte
3
4
5
6
7
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
(c) 1.9 < log.len < 2.6
log.byte
3
4
5
6
7
2.6
2.7
2.8
2.9
3.0
(d) 2.6 < log.len < 3.0
60 highlighted points: Dport 554, 1000−2000 packets

6. Single sessions → 'activity' sessions

- Single sessions are too detailed and not sufficiently summarized
- Example: `SIP` 5001 invokes `DPort` 80 (Web)
  `http://...` pulls in images from ports on other servers
- Example: `SIP` invokes `DPort` 21 (ftp)
  `ftp` invokes a control session (like `telnet`), extended duration but few bytes/packets until file transfer request (e.g., `get` or `send`), which initiates many further sessions
- Goal: summarize these individual sessions, all related to single activity, into one "activity session"

Example of Activity session

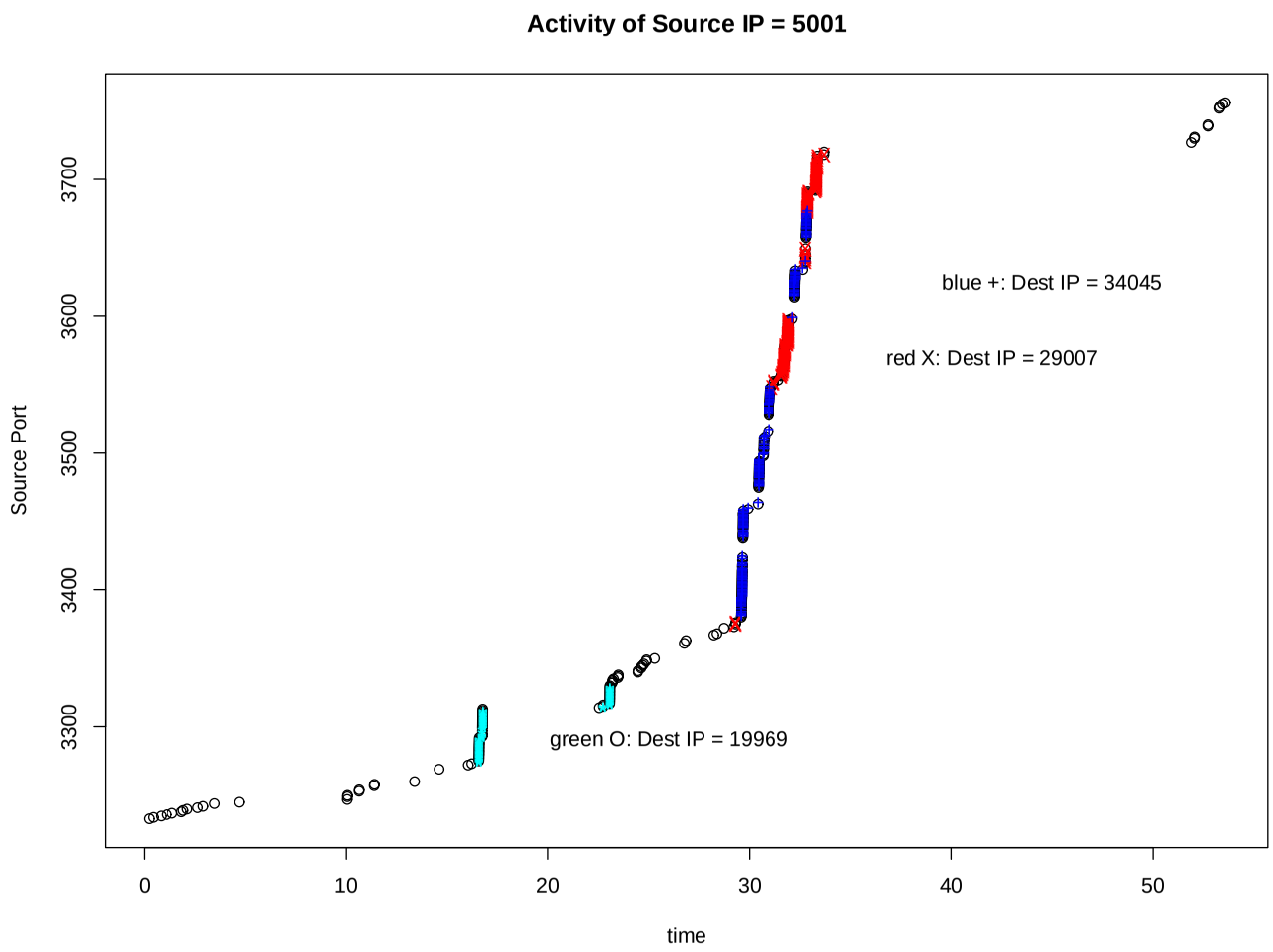
- `SIP` 5001 (probably 1 user, could be 2 wireless users)
- 348 sessions (mostly in minutes 30–32):
  262 web (80), 79 `DPort`=2523, 7 other `DPorts`
  146 `DIP`=34045, 91 `DIP`=29007, 53 `DIP`=19969, 58 other
- `DIP` 19969: 3 "activities" consisting of 18, 19, 16 sessions; time between sessions within activities < 0.18 seconds; time between "activities" is 9 seconds and 360 seconds (6 minutes)
- Define "activity": Same `SIP, DIP, DPort`, similar (increasing) `SPort`, time between sessions < 1 second

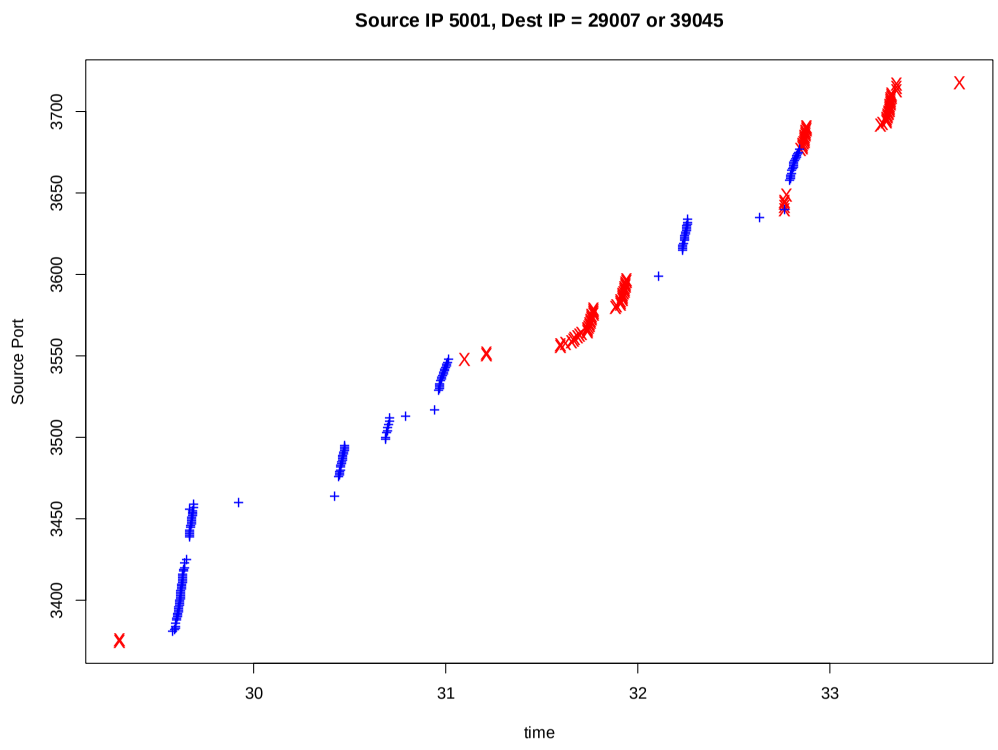

Source IP 5001, Dest IP = 19969, DPort 80
Source Port
3280
3290
3300
3310
3320
3330
17
18
19
20
21
22
23
time (minutes into the hour)
(3 activities: 18, 19, 16 sessions)

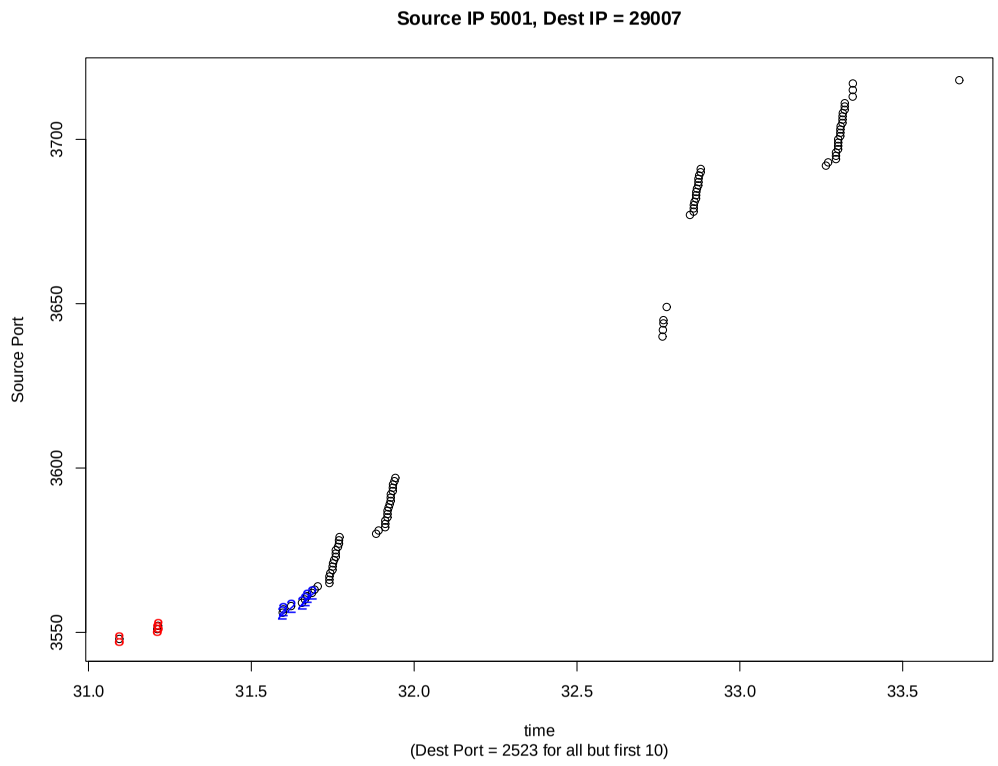
Summarize "activities" for `SIP` 5001, `DIP` 19969, `DPort` 80:

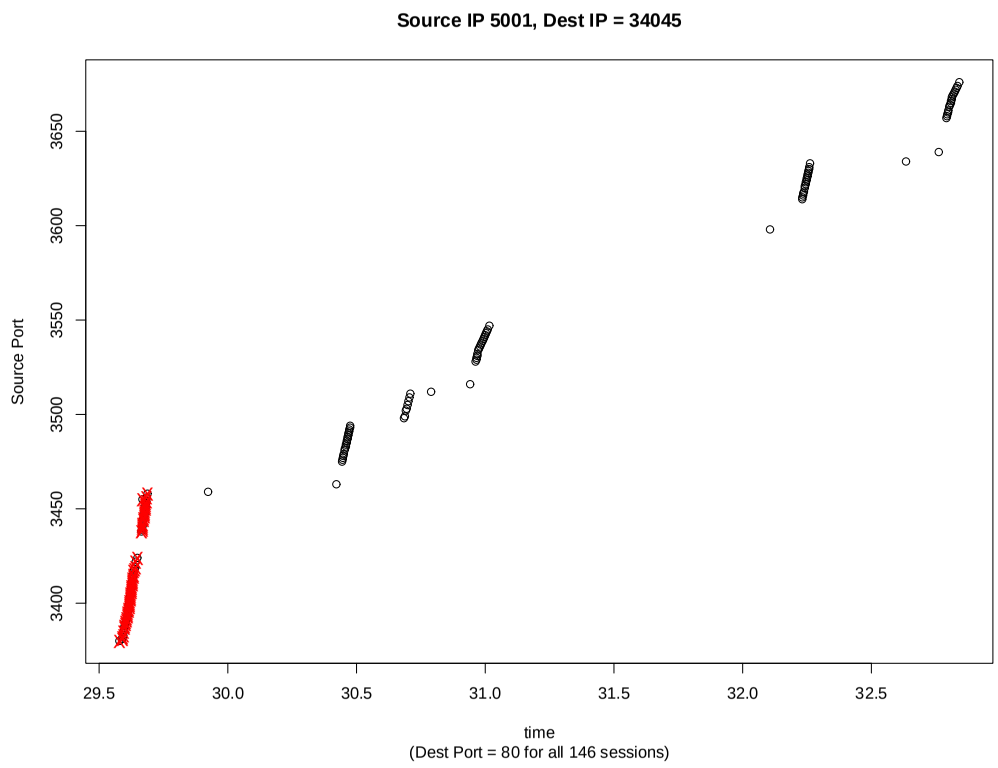
1. Duration 3.14 seconds, #packets = 612, #bytes=373,915
2. Duration 2.65 seconds, #packets = 473, #bytes=249,811
3. Duration 1.49 seconds, #packets = 264, #bytes= 76,073

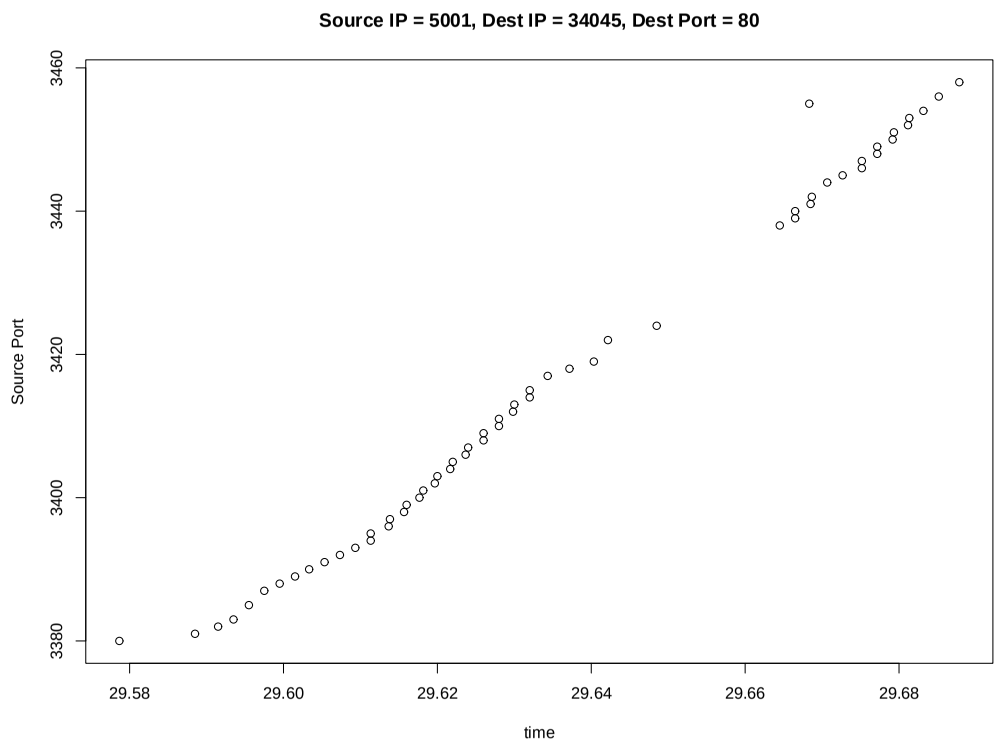
Likewise, 146 sessions where `DIP`=34045, and 91 sessions where `DIP`=29007, can be collated into "activities"


Activity of Source IP = 5001
Source Port
3300
3400
3500
3600
3700
time
0
10
20
30
40
50
blue +: Dest IP = 34045
red X: Dest IP = 29007
green O: Dest IP = 19969


Source IP 5001, Dest IP = 29007 or 39045
Source Port
3400
3450
3500
3550
3600
3650
3700
30
31
32
33
time


Source IP 5001, Dest IP = 29007
Source Port
3550
3600
3650
3700
31.0
31.5
32.0
32.5
33.0
33.5
time
(Dest Port = 2523 for all but first 10)


Source IP 5001, Dest IP = 34045
Source Port
3400
3450
3500
3550
3600
3650
29.5
30.0
30.5
31.0
31.5
32.0
32.5
time
(Dest Port = 80 for all 146 sessions)


Source IP = 5001, Dest IP = 34045, Dest Port = 80
Source Port
3380
3400
3420
3440
3460
29.58
29.60
29.62
29.64
29.66
29.68
time

Conclusions:

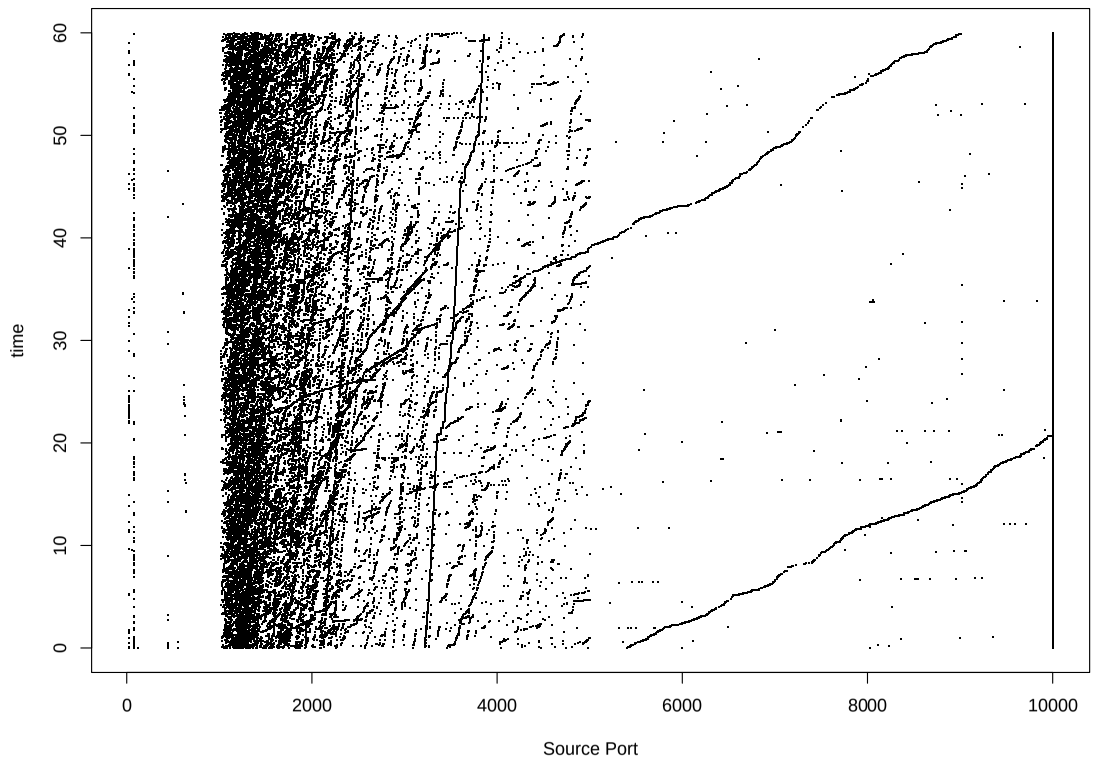
- "User" or "Activity" sessions are the natural units to study
- Modeling these "activity" session will aid in our understanding Internet traffic data
- These models will yield better tools for identifying "exotic" or subversive behavior
- Current research: Davis, KK, Marchette

Evolutionary Displays for Internet data

- Exploratory plots are useful for modeling activity sessions
- Adapt evolutionary plots to summarized data
- Can we summarize as quickly as the data arrive?
- Fast algorithms can do little more than compare and add (linear operations)
- Robust methods usually rely on medians and sorts
- Fast + robust $\Rightarrow$ compare and (keep or discard), follow by linear operations

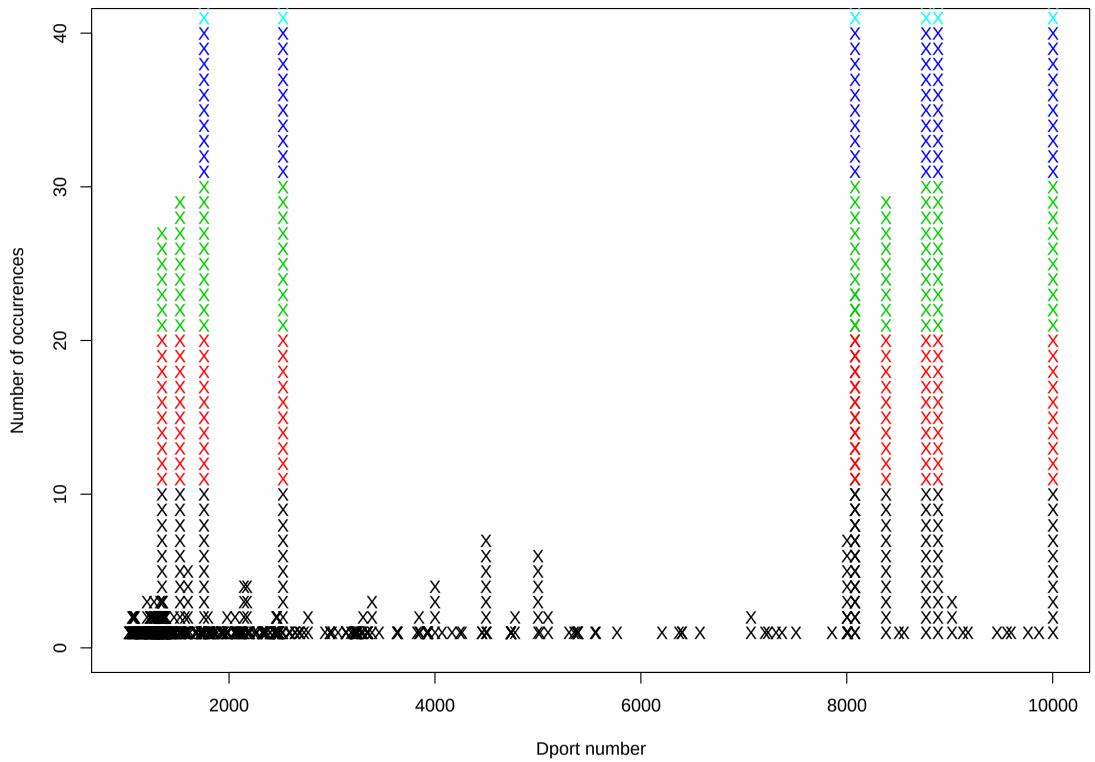
Waterfall Diagrams (Wegman and Marchette 2003)

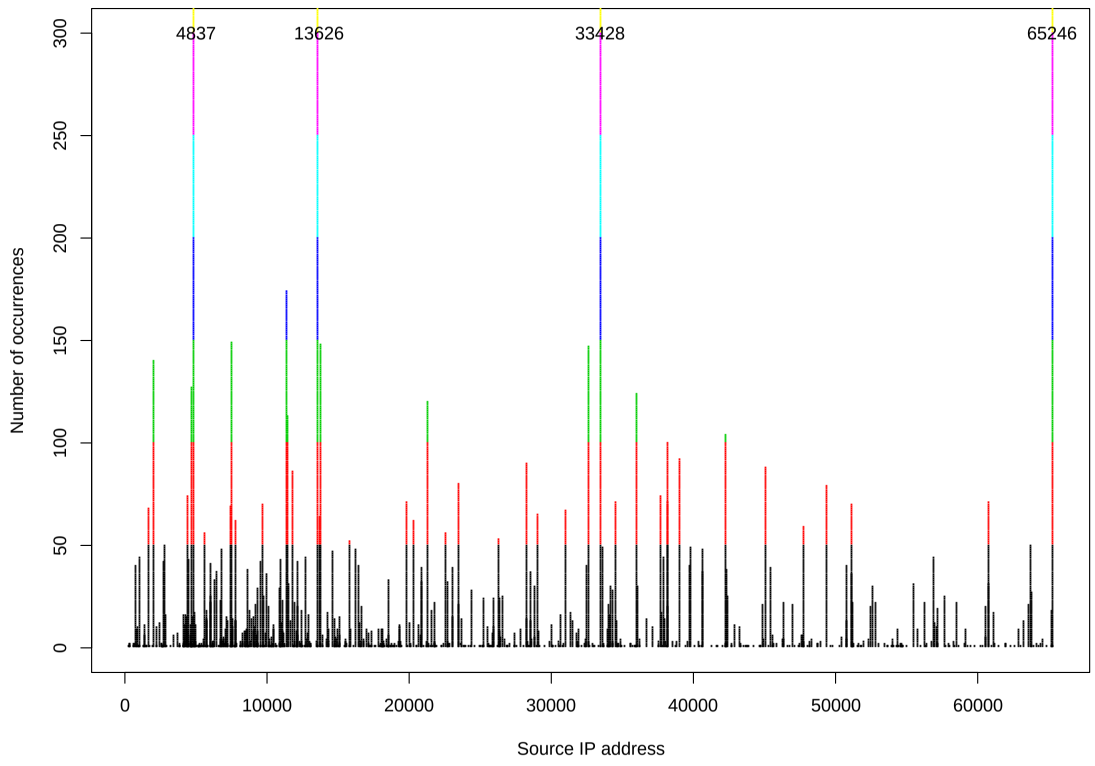
- "Streaming plot": Plot a point at location (s, t), where t (time) = session start time (starts at 0, continues upward), s is a source IP (`SIP`) or source port (`SPort`)
- `SIP`: 4837 (occurs 4,754 times), 13525 (occurs 4,448 times), 65246 (occurs 12,150 times)
- `SPort`: Trends *across* plot may indicate scanning `SPort`s
- Useful for monitoring attempted access: For a given session (exchange of packets), initial port may be assigned arbitrarily; subsequent ones assigned by incrementing pattern characteristic of operating system. Attacker can tell from pattern of `SPort` increments about operating system


time
0
10
20
30
40
50
60
0
2000
4000
6000
8000
10000
Source Port

Skyline plots

- "Streaming plot": Plot access of `DPorts` or `SPorts`
- Recall: 353 of the 380 `DPorts` (92.9%) occur < 5 times; the 3 most frequent `DPorts` are expected (`http, https, smtp`)
- Plot `X` when `DPort` is accessed (apart from "well-known" ports 0–1023); red `X` if count > 10
- Likewise for `SPort` but higher "control limit" 12% (4%) of the 6742 unique `SPorts` occur > 50 (100) times
- Building a "skyline"


Number of occurrences
0
10
20
30
40
2000
4000
6000
8000
10000
Dport number


4837
13626
33428
65246
Number of occurrences
0
50
100
150
200
250
300
0
10000
20000
30000
40000
50000
60000
Source IP address

5. High-Energy Physics (HEP)

- HEP experiments: Colliding beams of particles (MeV,GeV)
- Each collision (event) $\Rightarrow$ sub-particles (products)
- Huge detector ("bandshell" of wires): detects sub-particle
- Thousands of computer programs:
  - reconstruct particle tracks; connect particles w/tracks
  - estimate particle lifetime, momentum, mass
  - identify (?) particles (ex: +, mass 0.14GeV/c^2 $\Rightarrow$ 'π^+')

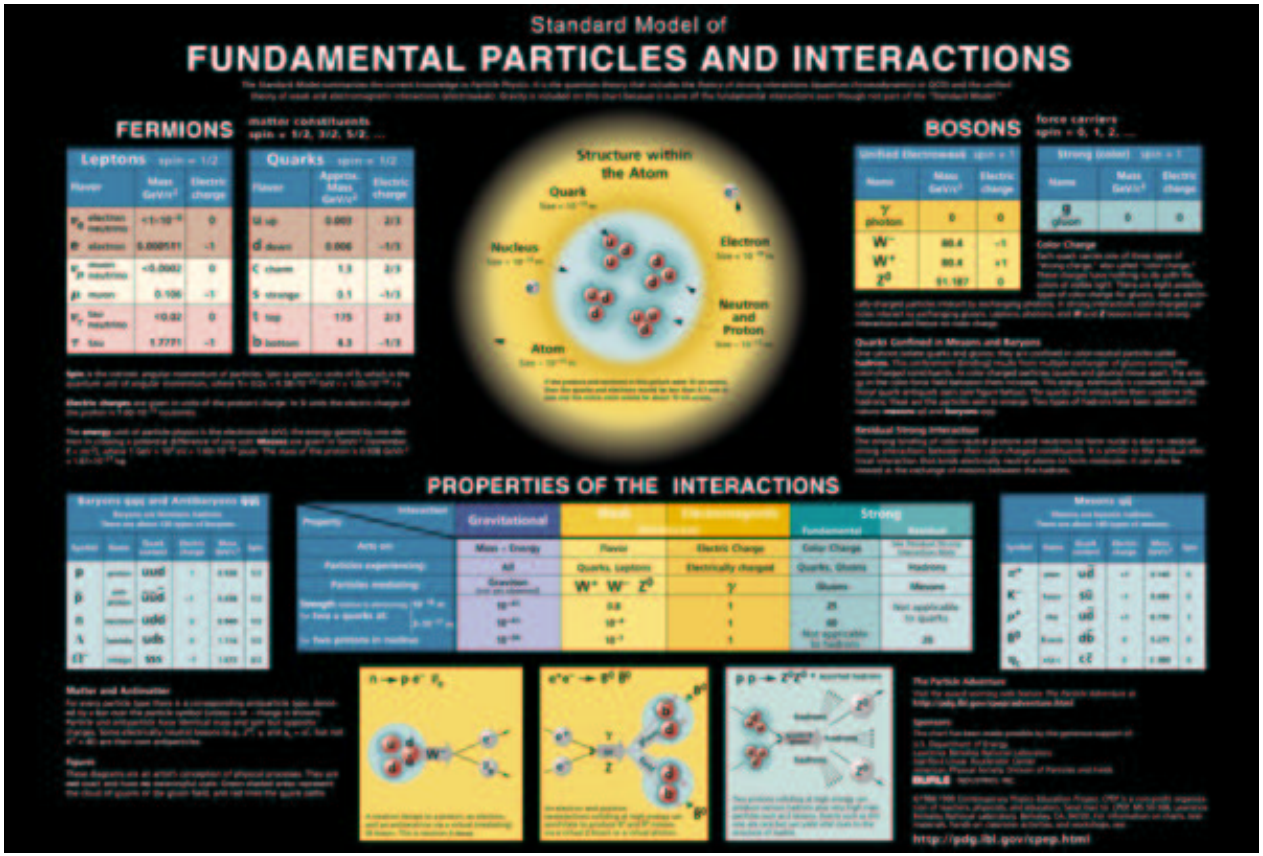
Goals:

- What happened? (decay type)
- How? (Parameter of model for specific decay)

Examples from chart of Standard Model (`http://CPEPweb.org`)

1. neutron (udd) β-decay to a proton (uud), an electron (e^-), and an antineutrino ($\bar{\nu}_e$)
2. electron-positron (e^+e^-) collision $\Rightarrow$ meson pair $B^0\bar{B}^0$ via a virtual γ photon or virtual Z boson
3. proton pair (pp) collision $\Rightarrow$ many hadrons (baryons qqq or $\bar{q}\bar{q}\bar{q}$ or mesons $q\bar{q}$) and bosons (force carriers)

Masses, charges, spins of q's (u,d,t,b,c,s), plus theory of behavior (Standard Model), determine masses, charges, spins of ~ 120 currently known baryons (~ 140 mesons)


Standard Model of
FUNDAMENTAL PARTICLES AND INTERACTIONS
FERMIONS
matter constituents
spin = 1/2, 3/2, 5/2, ...
Leptons spin = 1/2
Quarks spin = 1/2
Structure within the Atom
Quark
Nucleus
Electron
Neutron and Proton
Atom
BOSONS
force carriers
spin = 0, 1, 2, ...
Unified Electroweak spin = 1
Strong (color) spin = 1
Color Charge
Quarks Confined in Mesons and Baryons
Residual Strong Interaction
PROPERTIES OF THE INTERACTIONS
Gravitational
Strong
Mesons qq̄
Baryons qqq and Antibaryons q̄q̄q̄
Matter and Antimatter
The Particle Adventure
http://pdg.lbl.gov/cpep.html

- Hundreds of thousands of possible decays (events)
- ~ 100 events per nanosecond
- Data collection for 1 day would fill 6 CDs
- Massive filtering steps ("triggers") to discard data from 99+% of events
- SLAC saves data from ~ 100 events per hour
- Sensor resolution: may miss some particles from event
- Mis-ID: may connect particle to wrong event

SLAC collides beams of electrons, at energies designed to generate events with many b's

Example: Features of "target" decay of interest:

1. b's are produced ($\sim$20% of saved events)
2. b's decay into **neutral** B-mesons ($B^0\bar{B}^0$, not B^+B^-)
3. B^0 or $\bar{B}^0 \to \rho^+\rho^- \to \pi^0\pi^+ \cup \pi^0\pi^-$

Questions

- How often do such events occur?
- How long do sub-particles live? (e.g., 10^{-13} sec)
- Did anything else occur?

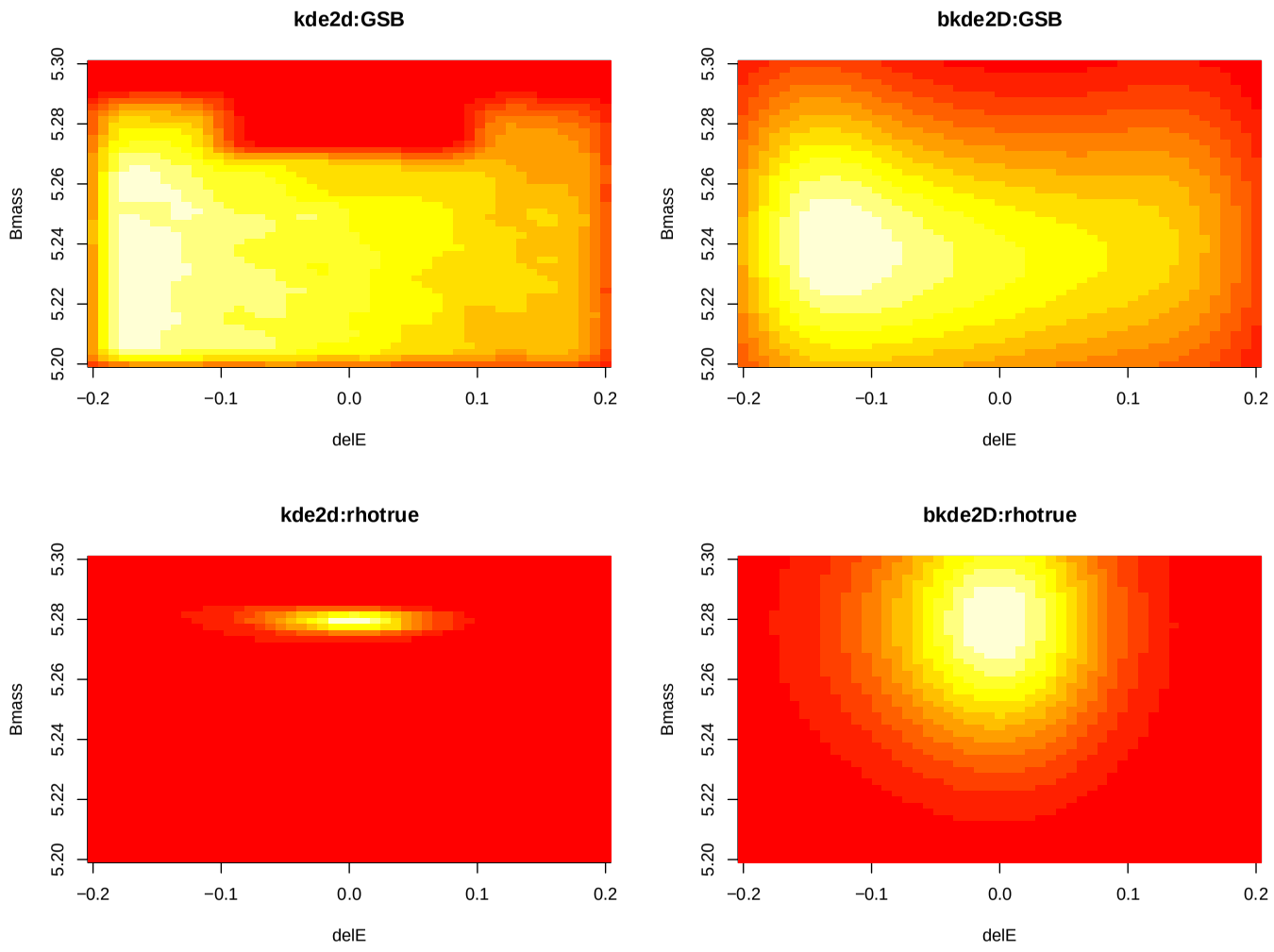
- Guiding principle:
  Better estimates of decay rates if we have less "noise"
- Use data to rule out likely "imposters"
  (# sub-particles, final $E_f <$ or $\approx E_0$, etc.)
- Data: ΔE, m_B, $Tthr$, m_{ρ^+}, m_{ρ^-}, H_1, H_2
  — *assuming our target decay occurred!*
- Physicists use *simulations* to assess what the data from our target event should look like (univariate pdfs); discard events where data lie outside "cuts" of "likely" regions on each variable
- Some *real* "imposter" data is generated from events at different energies; assume frequency of "imposter" events is the same at energy where target decay occurs
- Likelihood ratio test on all possible "imposters" not practical

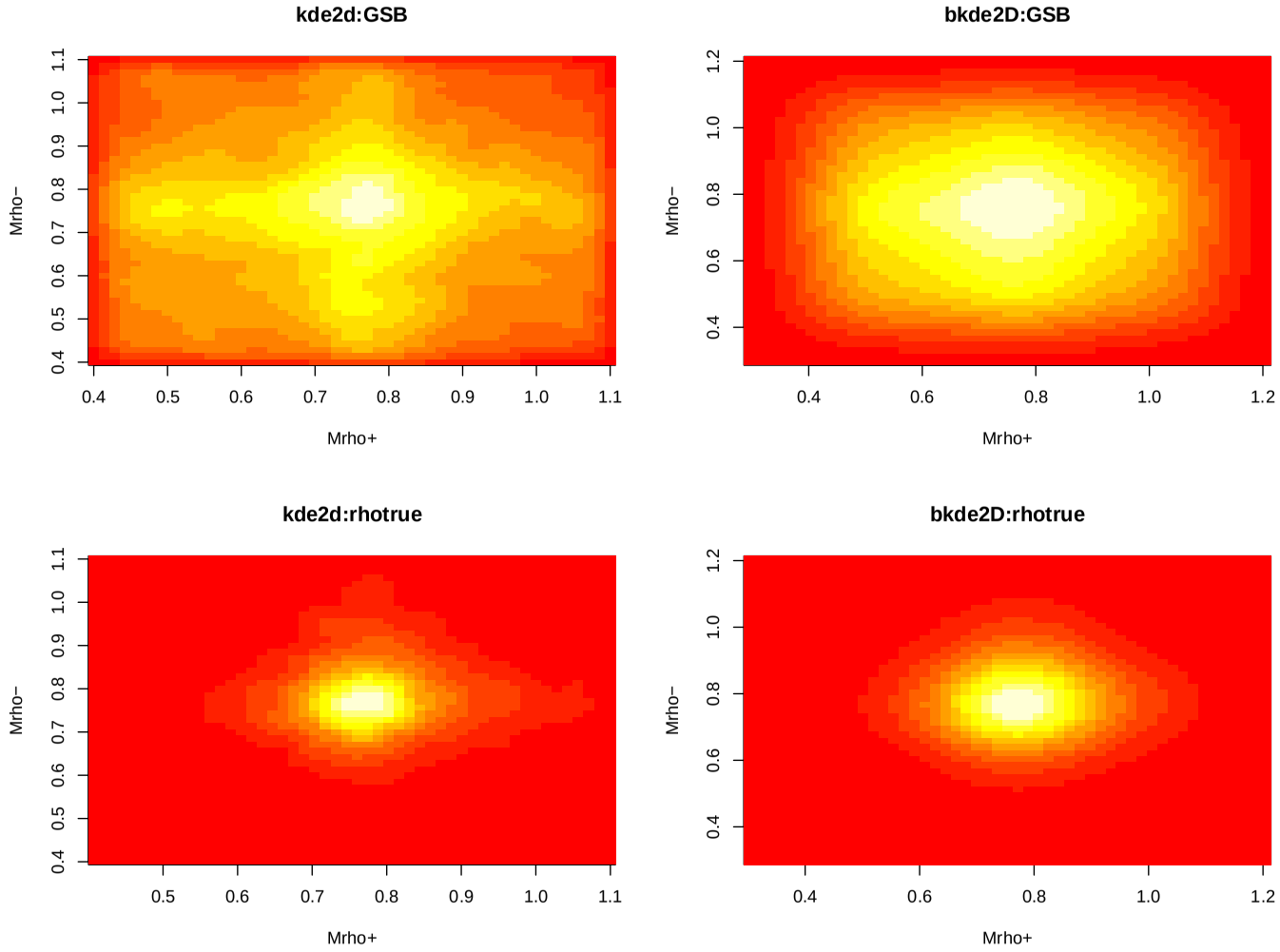
- But we might be able to reduce "background" by considering "likelihoods" from "top 20" most likely imposter decays

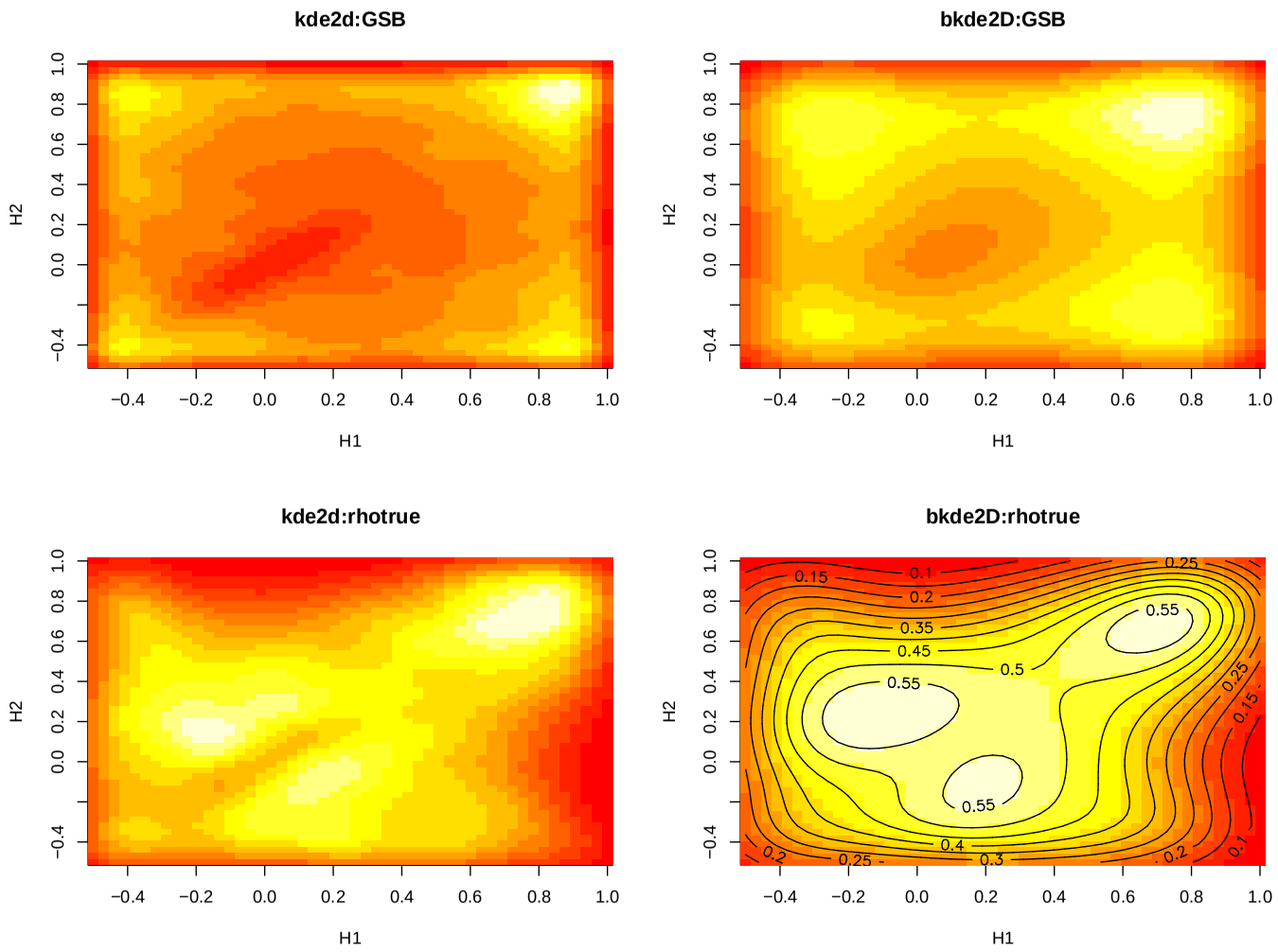
How realistic are the simulated data sets?

Can we approximate the "likelihoods" of the "top 20" events?

How well do mixture models fit with non-Gaussian components?


kde2d:GSB
bkde2D:GSB
kde2d:rhotrue
bkde2D:rhotrue
Bmass
delE
5.20
5.22
5.24
5.26
5.28
5.30
−0.2
−0.1
0.0
0.1
0.2


kde2d:GSB
bkde2D:GSB
kde2d:rhotrue
bkde2D:rhotrue
Mrho−
Mrho+
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2


kde2d:GSB
bkde2D:GSB
kde2d:rhotrue
bkde2D:rhotrue
H1
H2
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55

From Cybersecurity to Counterterrorism: Challenges

Data

- Massive data sets *and*
- Massive data sources
- Szewczyk (ACAS'04): Land purchases in Florida + Disney stock prices ⇒ more Mickey

Monitoring

- Inspection of *each unit* not feasible ⇒ Sampling plan
- Need "sampling plan" for entire data sets
- Inspection of *each process step* not feasible ⇒ Key indicators
- Need "key indicators" for analysis steps

Methods: Control Charts

- Traditional focus: low data rates, single process
- Today: Many processes, high rates

Example: Accumulated weight of checked baggage

- Cusum, EWMA for detecting small changes fast
- Excellent baseline information for intelligent choice of cusum, EWMA chart parameters

Example: Rare events in combination

- # cash paying customers within 3 hours
- Univariate/Multivariate(?) Poisson chart
- Shewhart charts for sudden changes quickly

Summary and future work

- New data types/structures lead to advances in science
- Information age $\Rightarrow$ Excellent opportunities for collaborations among statisticians, computer scientists, engineers
- Streaming data require:
  - much pre-processing to be interpretable
  - much summarization so they can be displayed
  - fast, **scalable** processing algorithms
- Streaming data offer new challenges to statisticians:
  - data acquisition, storage, distribution
  - fast algorithms and meaningful displays
  - new methods or combinations of classical+robust ones
- We still need exploratory plots:

detecting "exotic" requires characterizing "typical"

- EDA helps to identify natural "units" for study
- We need new tools & displays for streaming data, but ...
- Displays will be monitored by non-statisticians, so interpretation must be clear:

  "Churchill Eisenhart ... defined practical power as the product of the mathematical power by the probability that the procedure will be used. A compact procedure may well be used so much more often as to more than compensate for its loss of mathematical power."

  — J.W. Tukey, "A Quick, Compact, Two-Sample Test to Duckworth's Specifications," *Technometrics* 1(1), p.32

References

Benjamini, Y.; Hochberg, Y. (1995): "Controlling the false discovery rate...", *JRSSB* 57: 289–300.

Marchette, David J. (2001), *Computer Intrusion Detection and Network Monitoring*, Springer, New York.

KK + EJW (2006), "Visualizing 'typical' and 'exotic' Internet traffic data," *CSDA* 50: 3721-43.

Tukey, J.W. (1962), "The future of data analysis," *AMS* 1962; 1–67.

Wegman, Edward J.; Marchette, David J. (2004), "Statistical analysis of network data for cybersecurity," *Chance*, 9–19.