Internet Measurement and Data Analysis (2)

Kenjiro Cho

2011-10-05

review of previous class

theme of the class

- looking at the Internet from different views
  - learn how to measure what is difficult to measure
  - learn how to extract useful information from huge data sets

Class 1 Introduction

- network measurement and Internet measurement
- network management tools
- exercise: introduction of Ruby scripting language

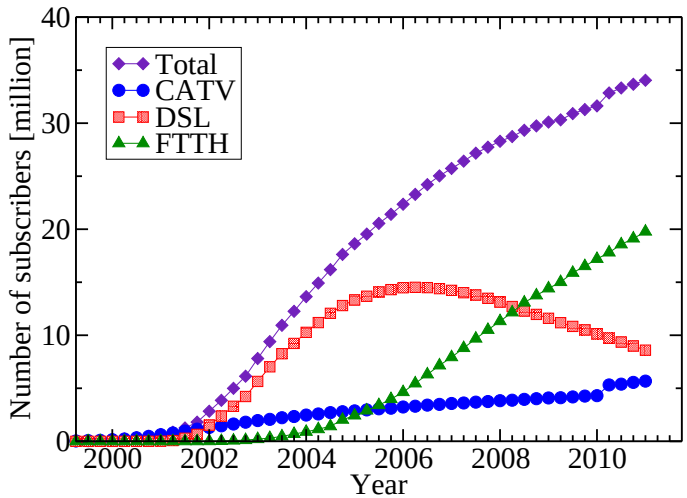
today's topics

Measuring the size of the Internet

- the number of users and hosts
- the number of web pages
- precision, errors, significant digits
- how to make good graphs
- exercise: graph plotting by gnuplot

the number of Internet users in Japan

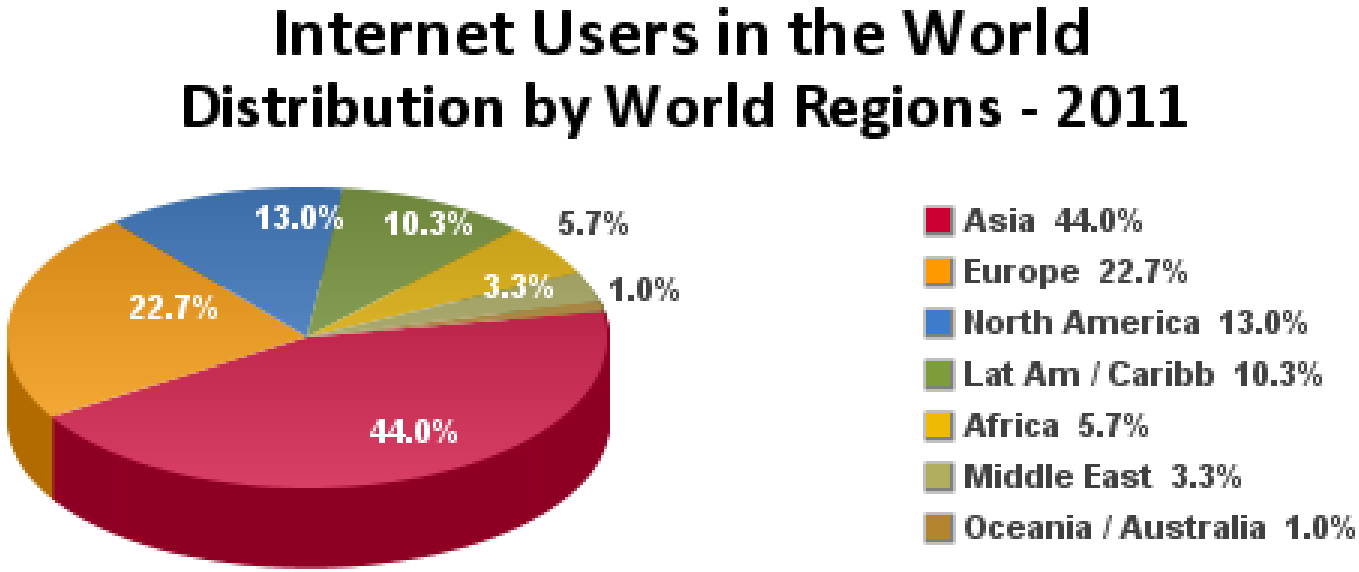
- MIC's survey on communications (総務省 通信利用動向調査)
  - 94.6 million users, population penetration 78.2% (end of 2010)
  - survey by random sampling and questionnaire
    - stratified random sampling with regions and town sizes
  - household survey: 45,120 households, 22,271 valid responses
  - total households 53.4 million (2010/03)
- MIC broadband service subscribers
  - reported by communication service providers
  - number of service subscribers: 35.5 million (2011/06)



source: MIC broadband service subscribers in Japan

the number of Internet users in the world

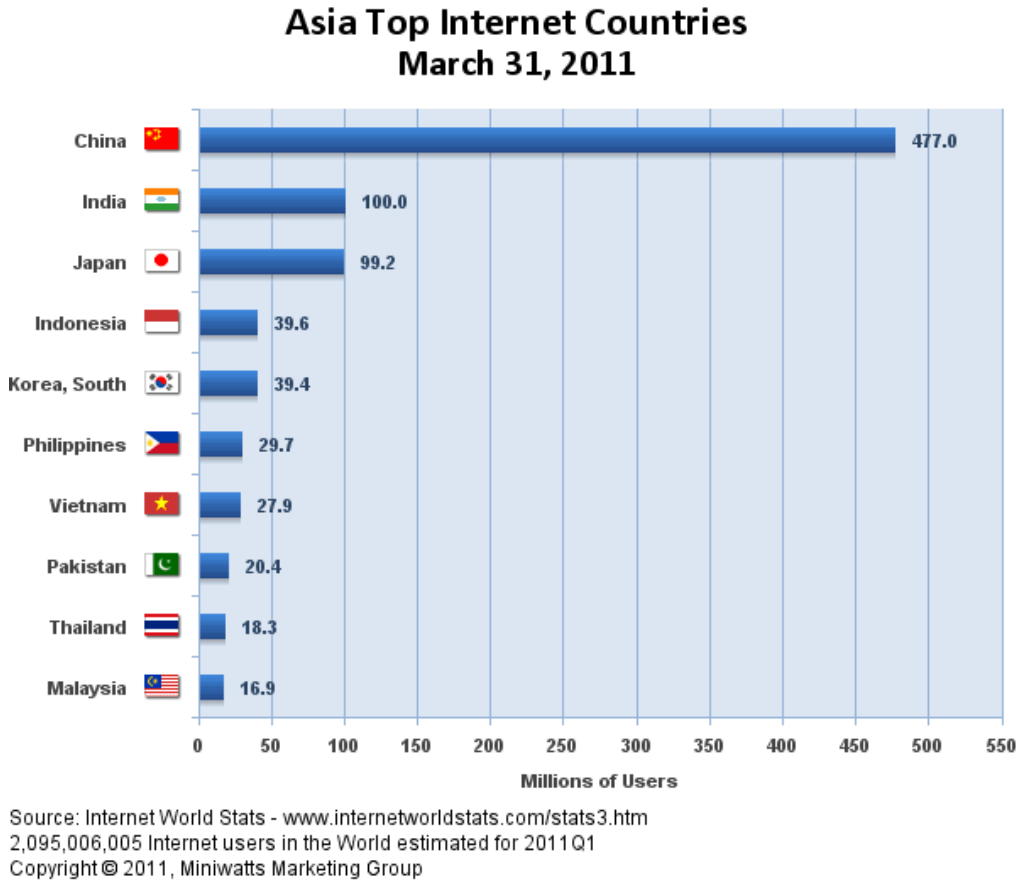
- 2,095 million, population penetration 30.2% (2011/03)



source: Internet World Stats http://www.internetworldstats.com/

the number of Internet users in Asia

- China (by far the top): 477 million, population penetration 36.3% (2011/03)



source: Internet World Stats http://www.internetworldstats.com/

the number of devices connected to the Internet

what is the definition of "connected to the Internet"?

- can access data on the Internet
  - browse web pages
  - e-mail reachable
  - it is difficult to count the number
    - 2011: mobile phone subscribers: 5.3 billion
    - IDC report: worldwide PC shipments in 2010: 347 million
- communicate over IP protocols (including devices behind NATs)
- with global IP addresses (bi-directional access over IP)

measuring the number of computers on the Internet

goals

- to know the number of computers on the Internet
  - became difficult due to the prevailing use of NAT boxes
- to understand the usage of IP addresses
  - IP addresses are limited resources
  - inputs for IP address assignment policies
    - IPv4 addresses exhaustion

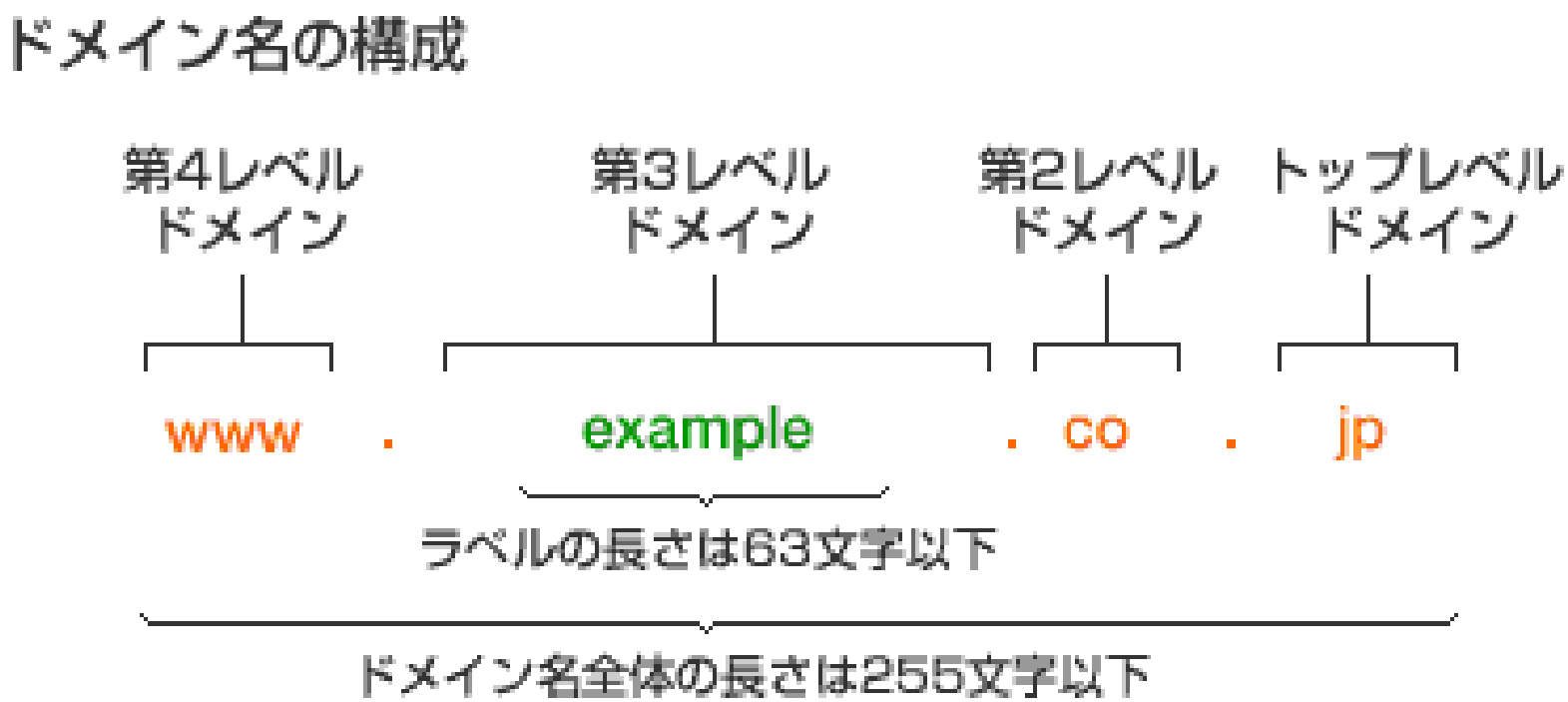
methods

- full search of the DNS name tree
- full search of the IP address space (2^{32})
- sampling for inferring the usage
  - difficult due to different usage of different address blocks

Domain Name System (DNS) basics (1/3)

from JPNIC 「ドメイン名のしくみ」

- `http://www.nic.ad.jp/ja/dom/system.html`

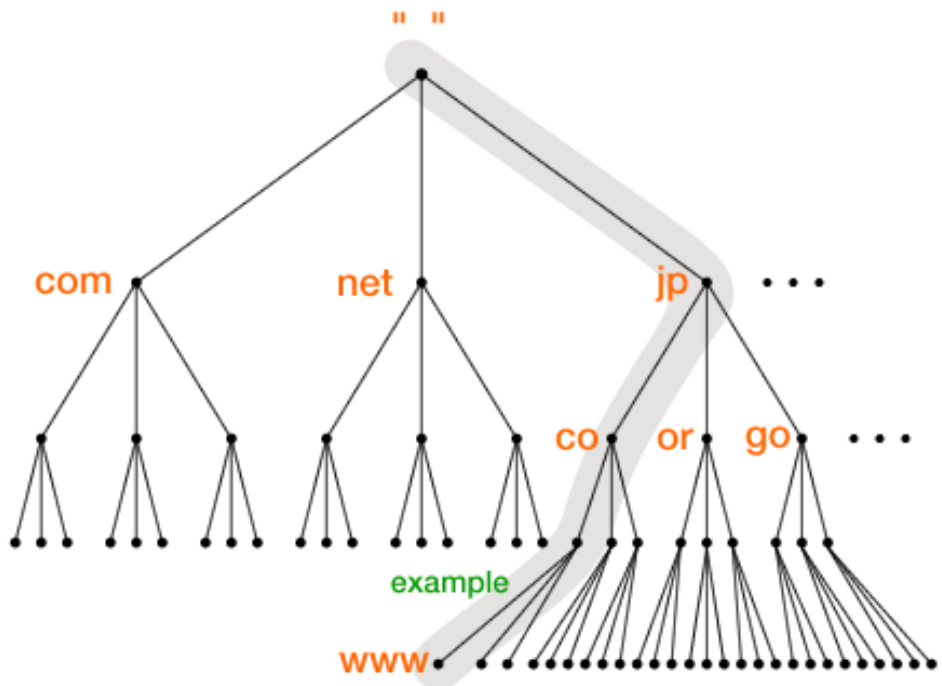


source :JPNIC

Domain Name System (DNS) basics (2/3)

structure of DNS

- a tree structure with “root” at the top
- each domain has “name servers” for managing the distributed database
  - manages mapping of domain names and IP addresses within the domain
  - also manages references to name servers of sub-domains

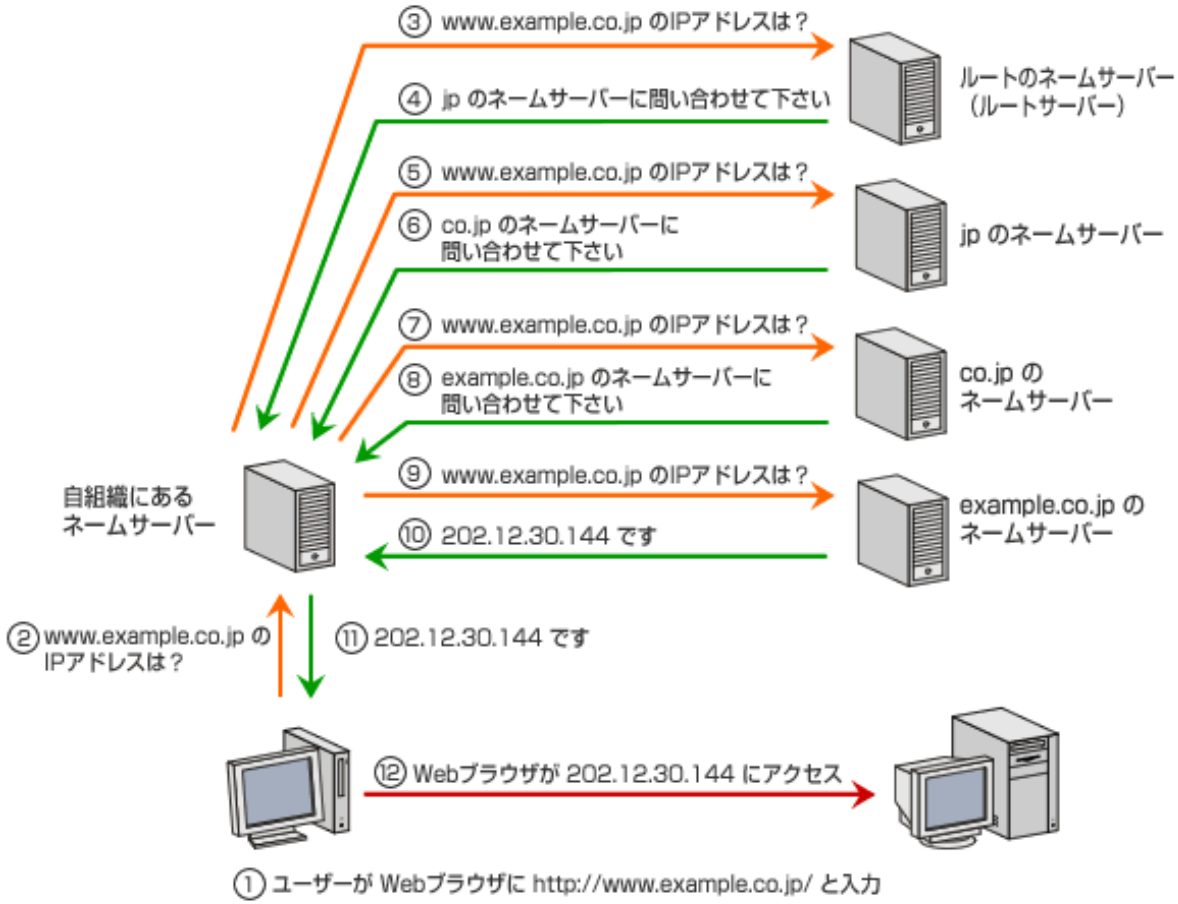


domain name space (source:JPNIC)

Domain Name System (DNS) basics (3/3)

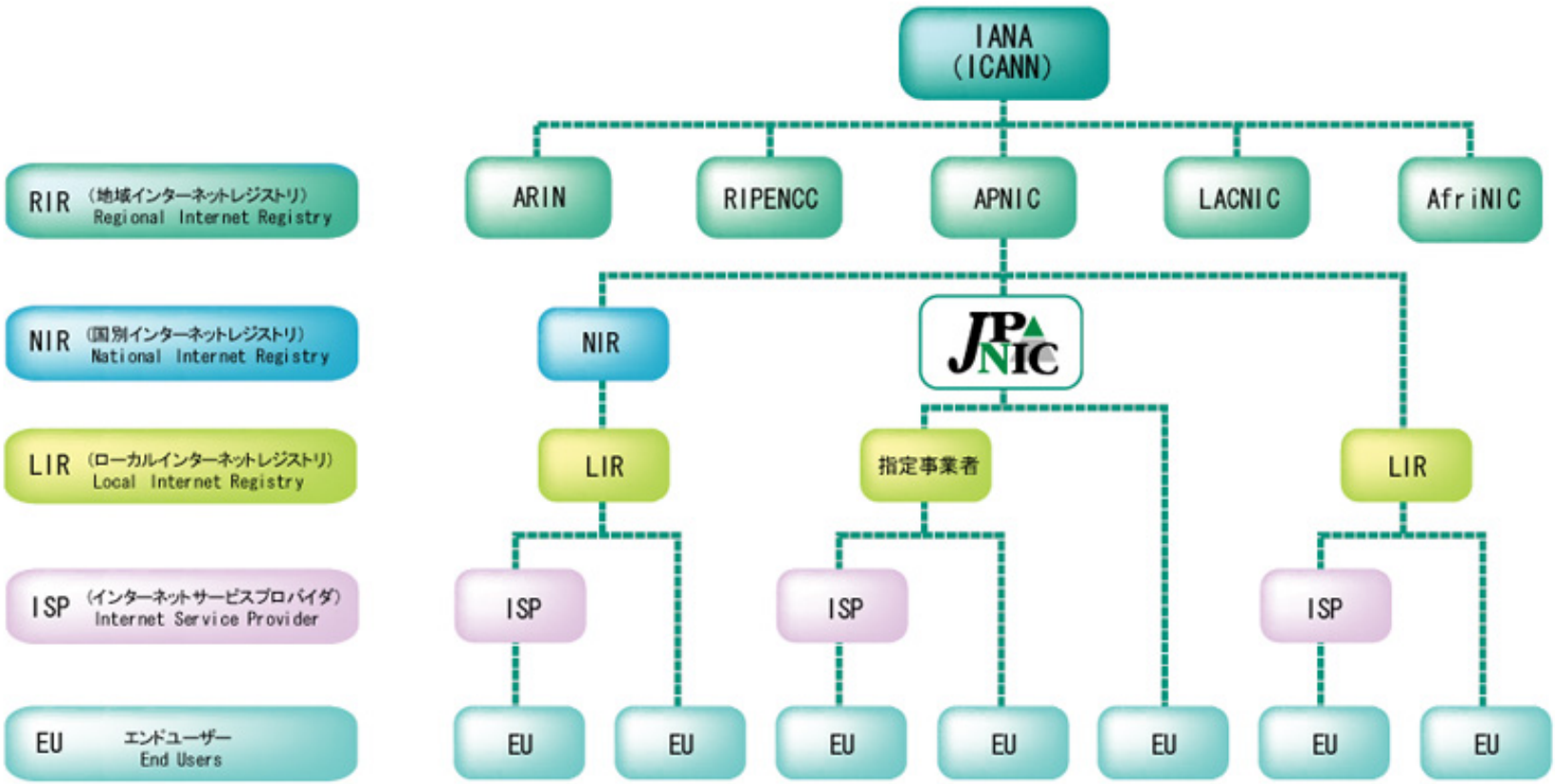
name resolution in DNS

- name resolution: from a domain name to the corresponding IP address
  - reverse lookup: from IP address to domain name (using the reverse tree)



IP address assignment management

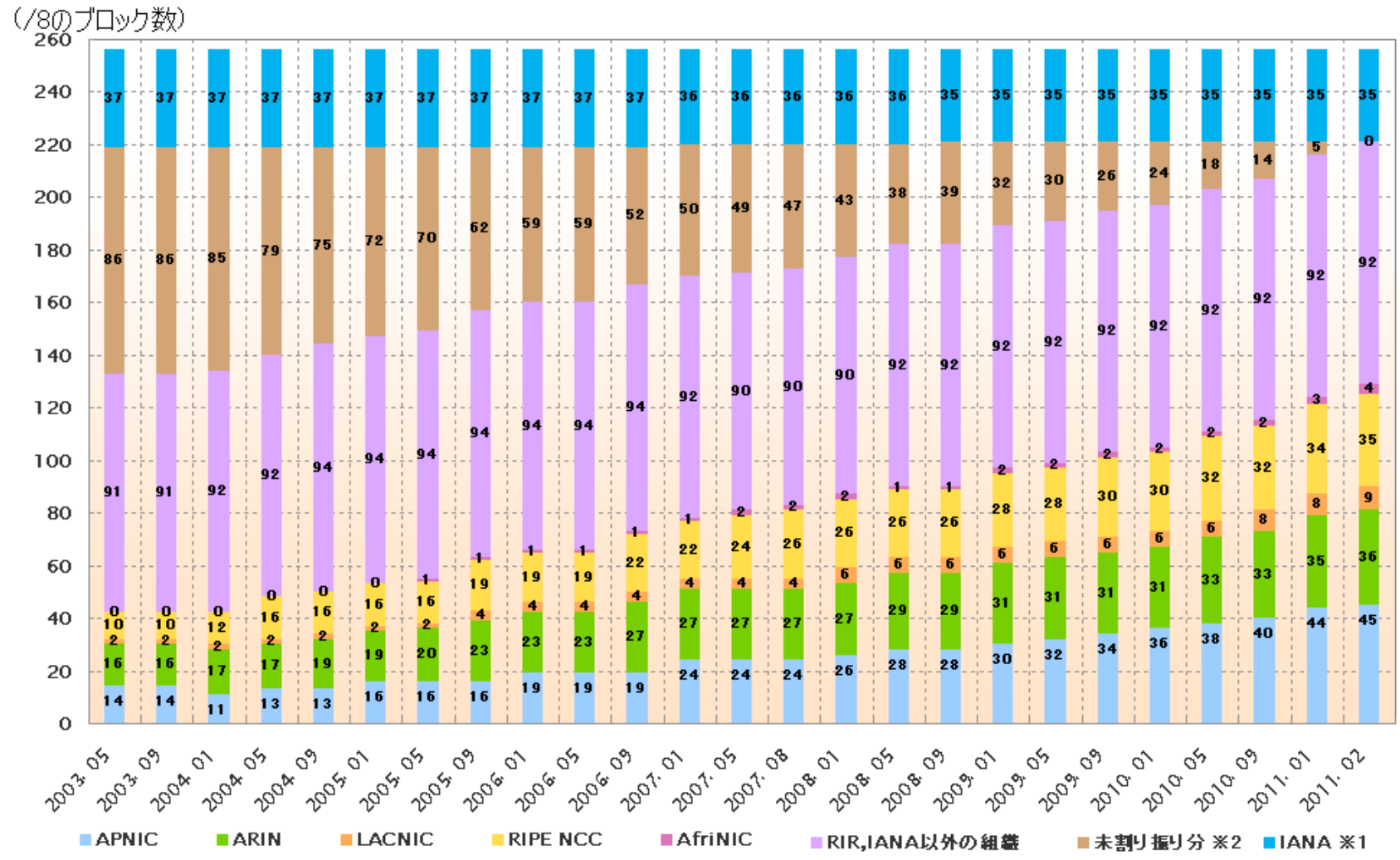
- IANA → RIR → NIR → LIR



hierarchical management of IP addresses (source:JPNIC)

exhaustion of IPv4 addresses

- 2011/2/1 exhaustion of IANA address pool
- 2011/4/15 exhaustion of APNIC and JPNIC address pool

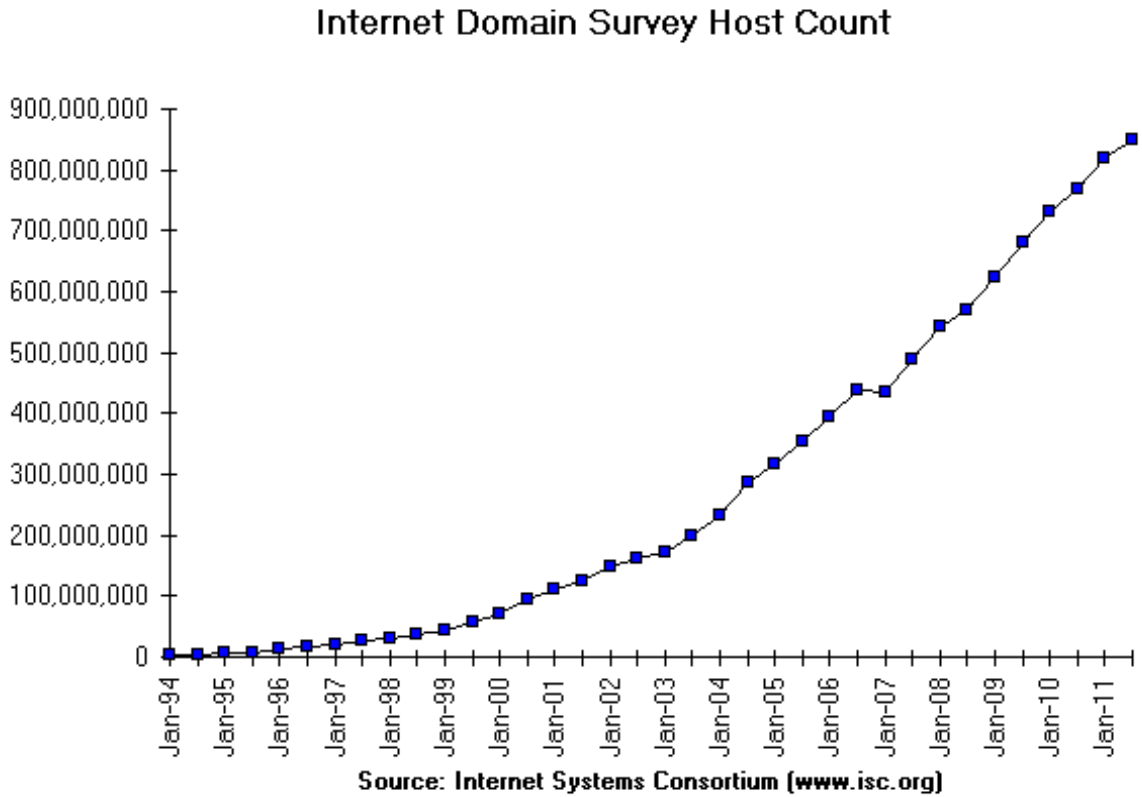


IP address assignment status by RIR (source:JPNIC)

classical method to infer host count

The ISC Domain Survey (inference from DNS)

- 850 million hosts (2011/07)



source: ISC domain survey http://www.isc.org/solutions/servey

The ISC Domain Survey

measurement method

- 1987-1997: count hosts registered in DNS (RFC1296)
  - parse the DNS delegation tree, and obtain zone data from each zone
  - count "A records" in a zone data
  - for zones prohibiting zone transfer, calibrate the count using the success ratio of zone transfer
- 1998-: count unique IP addresses registered in DNS
  - parse the reverse DNS tree, and find existing /24 blocks
  - for each /24 block found, reverse look up "PTR record" for all the IP addresses (1-254) in the block
  - existence of a PTR record doesn't mean existence of a host. randomly sample 1% of the found addresses, ping these addresses, and use the success ratio for calibration

limitations

- cannot count hosts which are not registered in DNS
- cannot know the accuracy of the calibration method
- cannot count hosts behind NATs

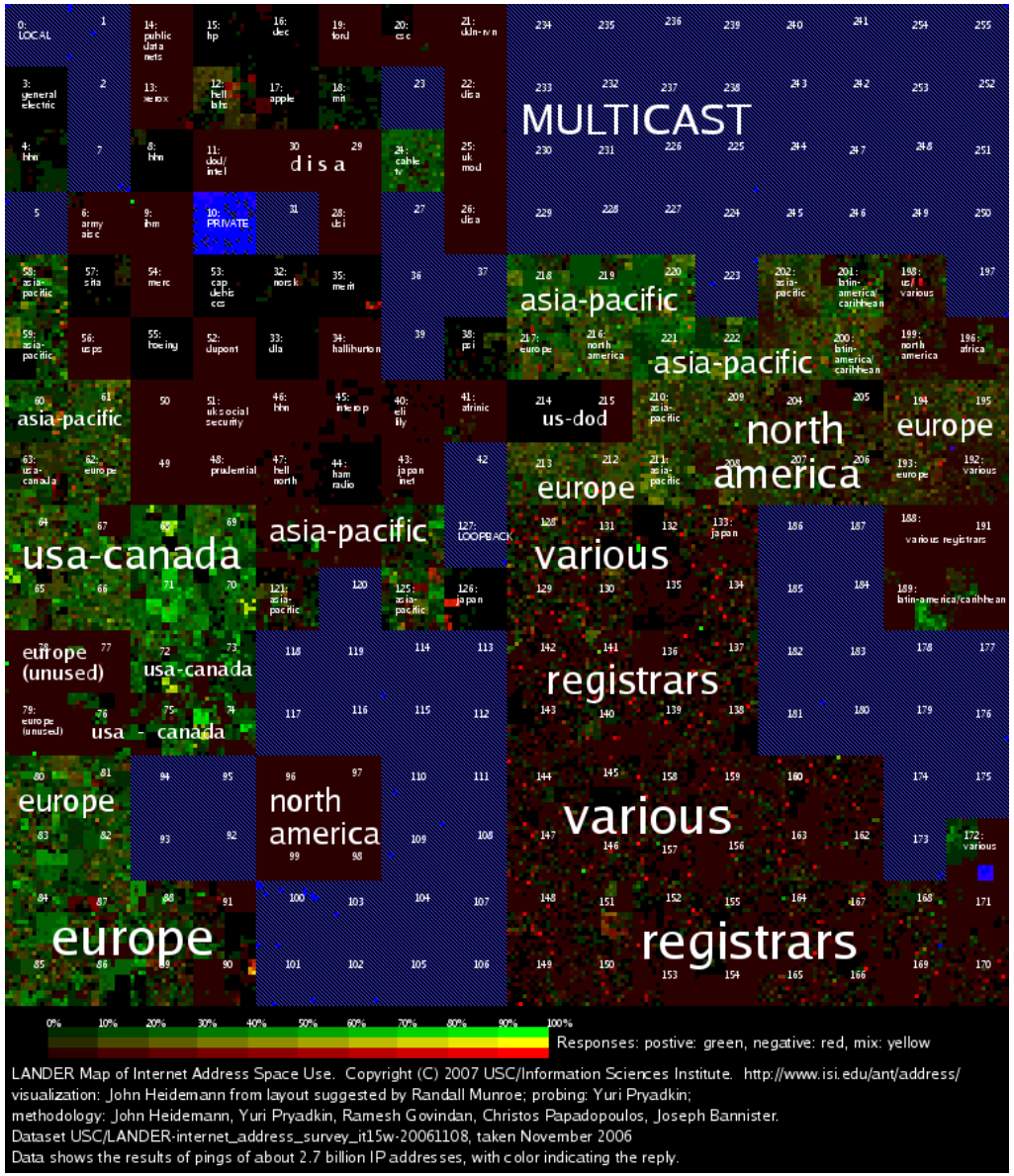
exhaustive search of IP address space

- Heidemann's measurement in 2006/11
- ping all the assigned IP addresses
- 93% of probed addresses did not respond (firewall, etc)
- a new probe is installed at SFC (2011/08)

address type	number	% of addrs	% of probed
IPv4 addresses	4,290M	100%	
reserved	1,160M	27%	
allocated	3,140M	73%	
unprobed (mcast, etc)	342M	8%	
probed	2,800M	65%	100%
replies	187M	4.4%	6.7%
positive replies	103M	3.6%	3.7%
negative replies	84M	2.0%	3.0%
non-replies	2,610M	61%	93%

J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, J. Bannister.
Census and survey of the visible Internet.
ACM IMC'08. pp169-182. Vouliagmeni, Greece. October 2008.

visualization of IP address usage

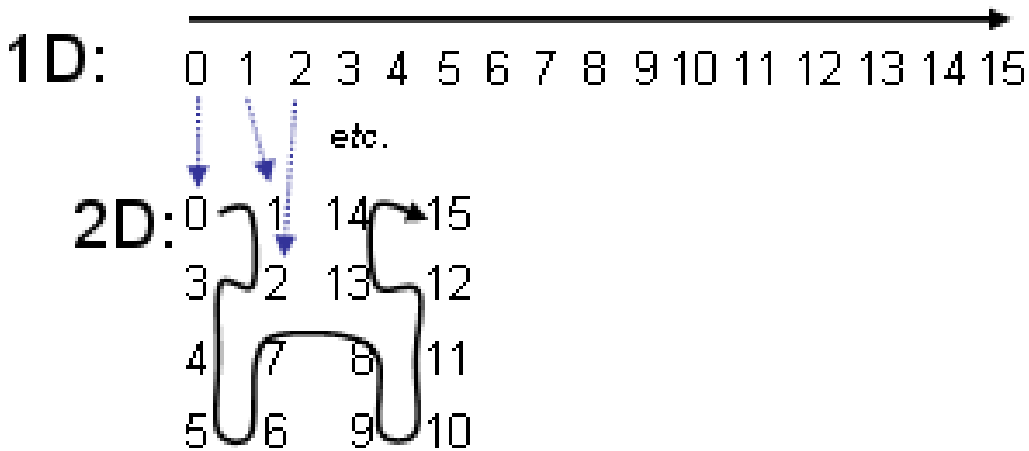


http://www.isi.edu/ant/address/

visualization of IP address usage (cont'd)

visualization technique

- IP address space visualized by Hilbert Curve (keep adjacency, recursive)
- each point shows mean of a /16 block (64k addrs)
- positive:green, negative:red, mix:yellow



Hilbert Curve

the number of web pages

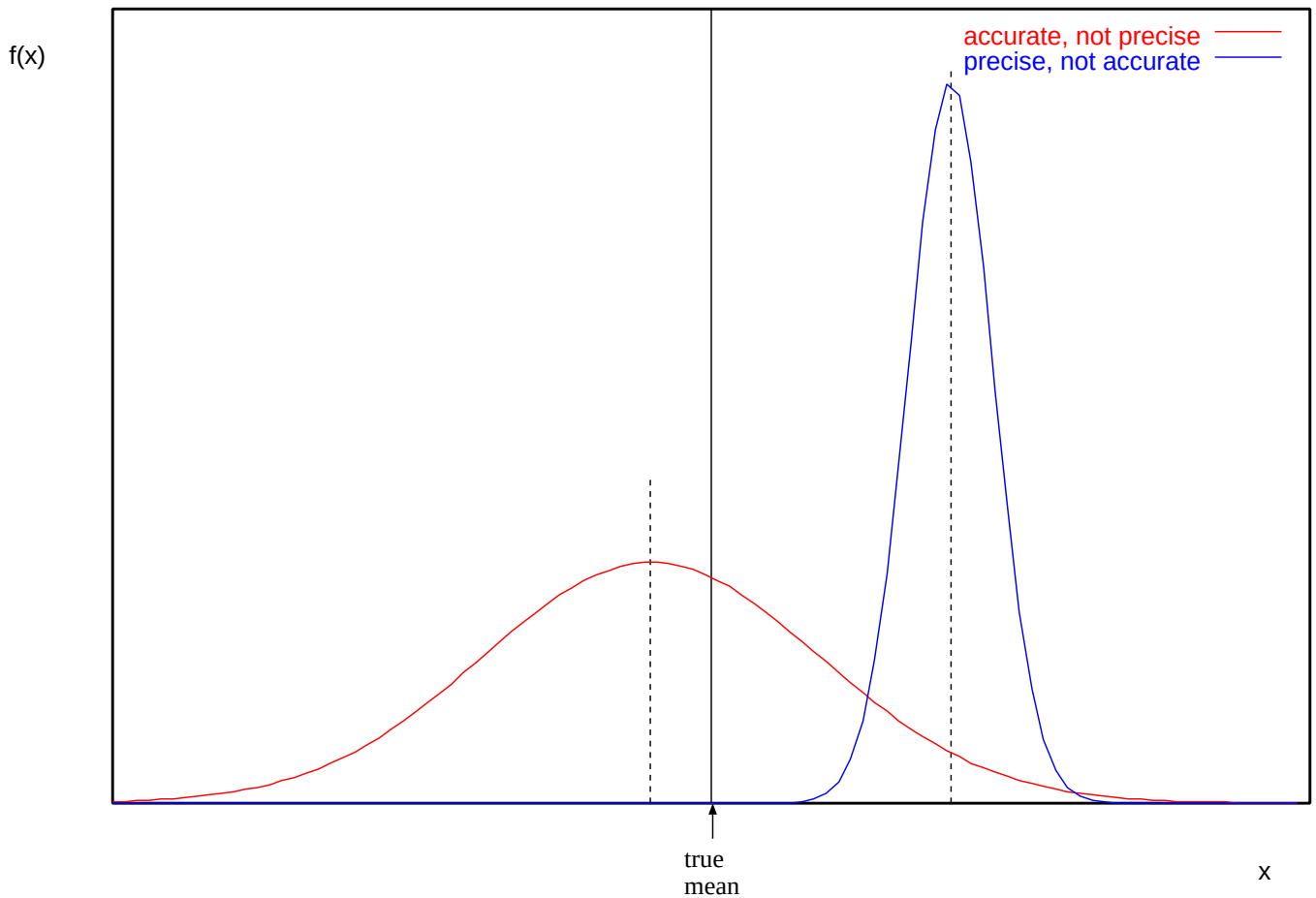
- definition of "web page"? increasing dynamic pages (calendar, etc)
- data can be collected by crawling robots
  - start from popular sites, follow links
- existing large search systems have data but they are not published
- netcraft: web server survey 227 million sites in 2010/09
- google: indexed 1 trillion (10^{12}) unique URLs in 2008
  - `http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`

accuracy, precision and errors

accuracy: how close to true value

precision: uncertainty in data

error: difference from true value, range of uncertainty



various errors

measurement errors

- systematic errors (if conditions are identified, errors could be corrected)
  - instrument error, procedural error, personal bias
- random errors (noise: accuracy can be improved by repeating measurement)

calculation errors

- round-off errors (丸め誤差)
- truncation errors (打ち切り誤差)
- loss of trailing digits (情報落ち)
- cancellation of significant digits (桁落ち)
- propagation of error (誤差の伝搬)

sampling errors

- when sampling is used, true value is usually unknown
- sampling errors: errors in estimating population characteristics

significant digits

significant digits of "1.23" is 3 ($1.225 \leq 1.23 < 1.235$)

expressions

expressions	significant digits	
12.3	3	
12.300	5	
0.0034	2	
1200	4	(vague, $1.200x10^3$)
$2.34x10^4$	3	

arithmetic

- use all the available digits during calculation
  - for manual calculation, use one more digit
- apply the significant digits to the final value

basic rules

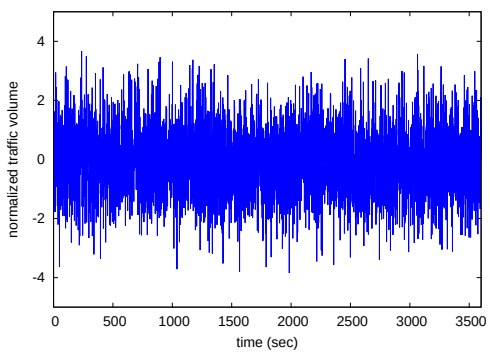
- addition/subtraction: use the smallest number of decimal places
  - $1.23 + 5.724 = 6.954 \Rightarrow 6.95$
- multiplication/division: use the smallest number of significant digits
  - $4.23 \times 0.38 = 1.6074 \Rightarrow 1.6$

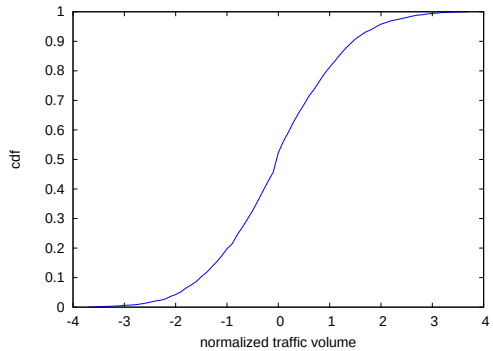
computational precision of computers

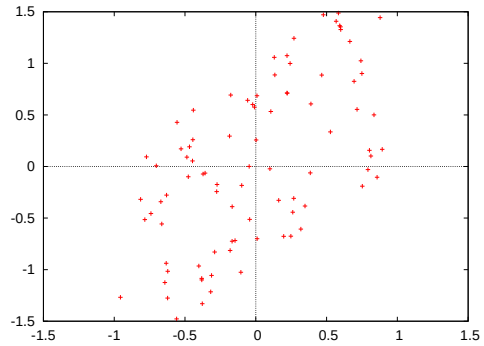
- integer (32/64bits)
  - 32bit signed integer (up to 2G)
- 32bit floating point (IEEE 754 single precision): significant digits:7
  - sign:1bit, exponent:8bits, mantissa:23bits
  - $16,000,000 + 1 = 16,000,000$!!
- 64bit floating point (IEEE 754 double precision): significant digits:15
  - sign:1bit, exponent:11bits, mantissa:52bits

graph plotting

create a set of plots using statistical techniques to intuitively understand the data







guidelines for plotting

require minimum effort from the reader

- label the axes clearly
- label the tics on the axes
- identify individual curves/bars
- select appropriate font size
- use commonly accepted practices
  - zero-origins, math symbols, acronyms
- show variation/distribution of variables
- select ranges properly
- do not present too many items in a single chart
- when comparing data sets, use appropriate normalization
- when comparing plots, use the same scale for the axes
- when using colors
  - make sure it is readable in black-and-white print
  - make sure readable on data projectors (e.g., do not use yellow)

variables in data

- univariate analysis
  - explores a single variable in a data set, separately
- multivariate analysis
  - looks at more than one variables at a time

plotting raw data

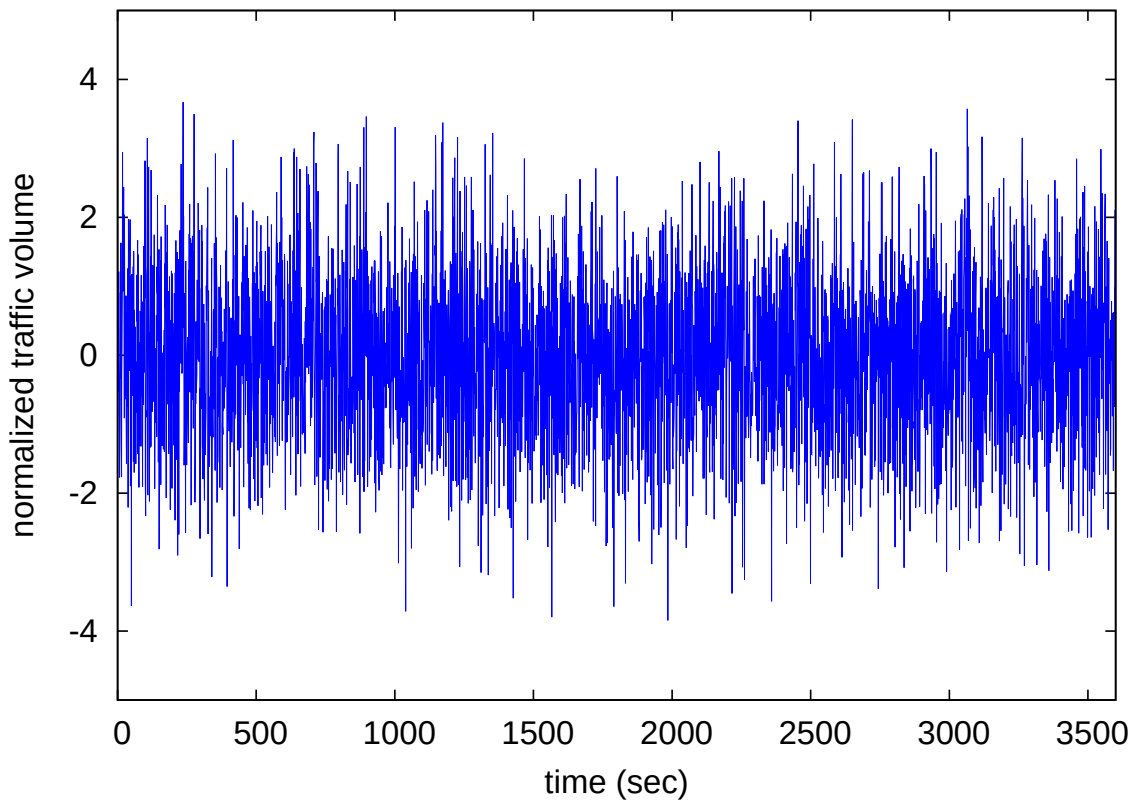
- time series plots
- histograms
- probability plots
- scatter plots

there are many other plotting techniques

time series plots

time-series plots (or other sequence plots) provides a feel for the data

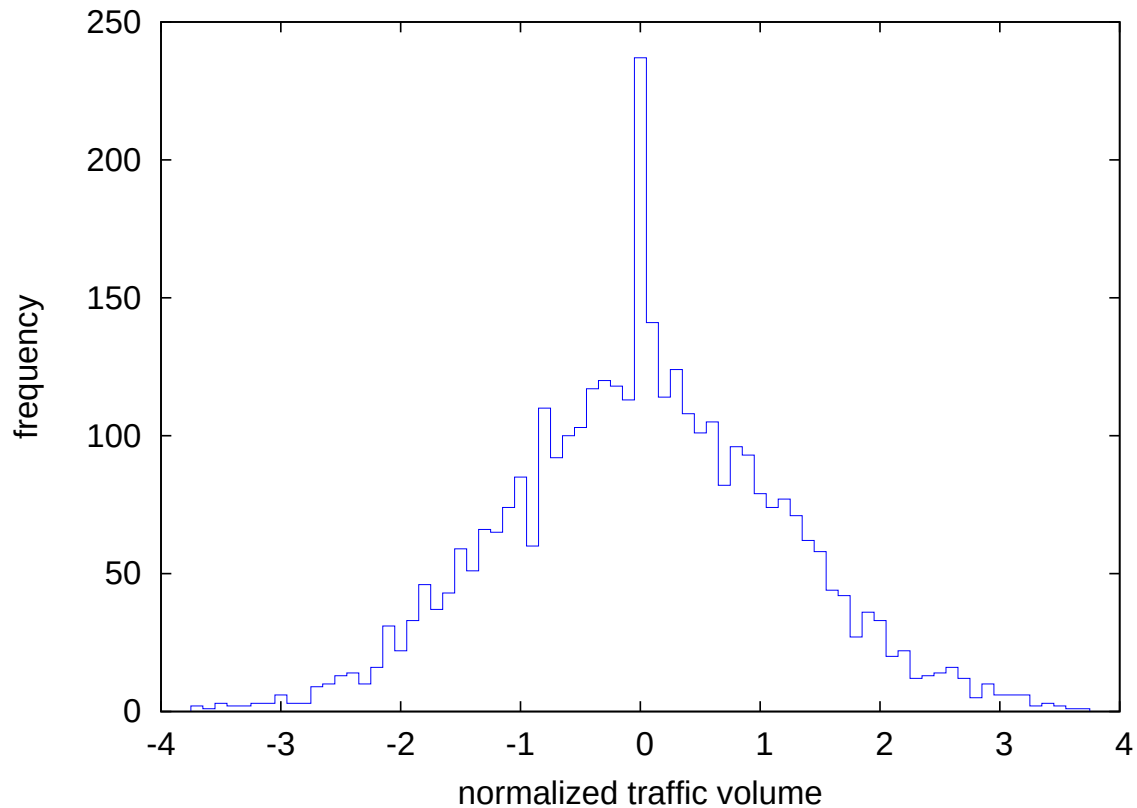
- you can identify
  - shifts in locations
  - shifts in variation
  - outliers



histograms

to see distribution of the data set

- split the data into equal-sized bins by value
- count the frequency of each bin
- plot
  - X axis: variable
  - Y axis: frequency



histograms (cont)

with histograms

- you can identify
  - center (i.e., the location) of the data
  - spread (i.e., the scale) of the data
  - skewness of the data
  - presence of outliers
  - presence of multiple modes in the data

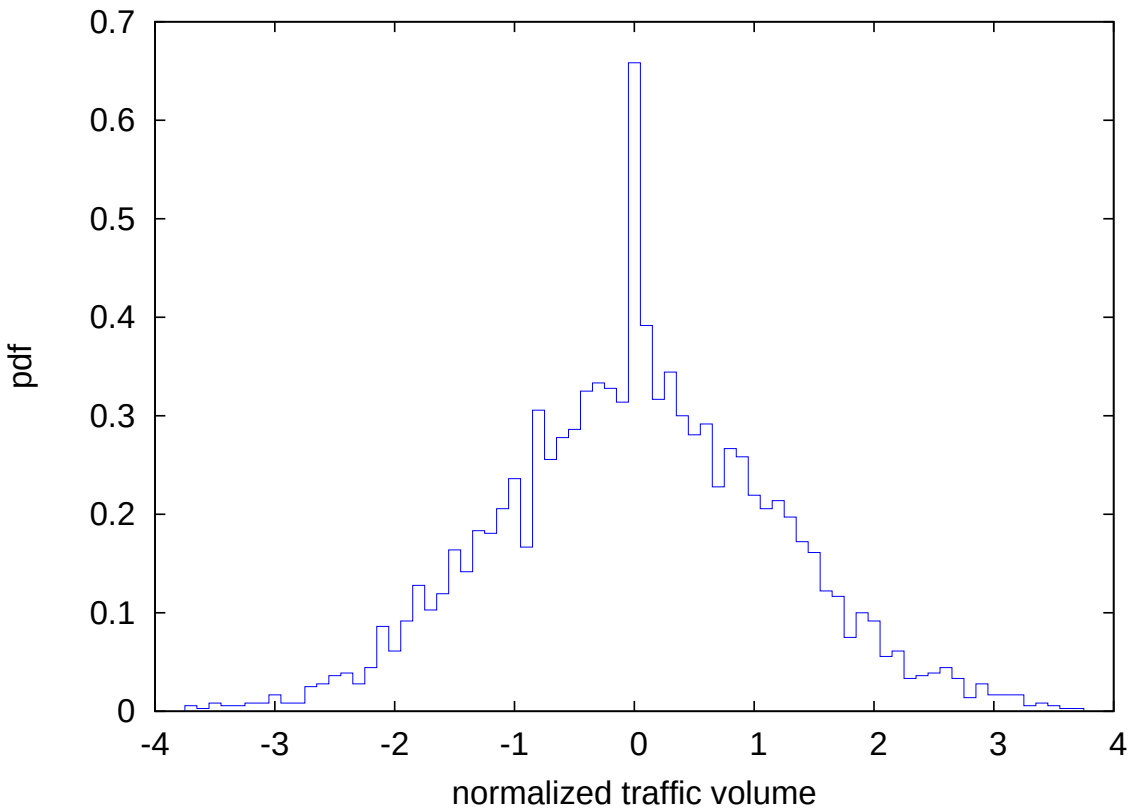
limitations of histograms

- needs appropriate bin size
  - too small: each bin doesn't have enough samples (e.g., empty bins)
  - too large: only few regions available
  - difficult for highly skewed distribution
- enough samples needed

probability density function (pdf)

- normalize the frequency (count)
  - sum of the area under the histogram to be 1
  - divide the count by the total number of observations times the bin width
- probability density function: probability of observing x

$$f(x) = P[X = x]$$



cumulative distribution function (cdf)

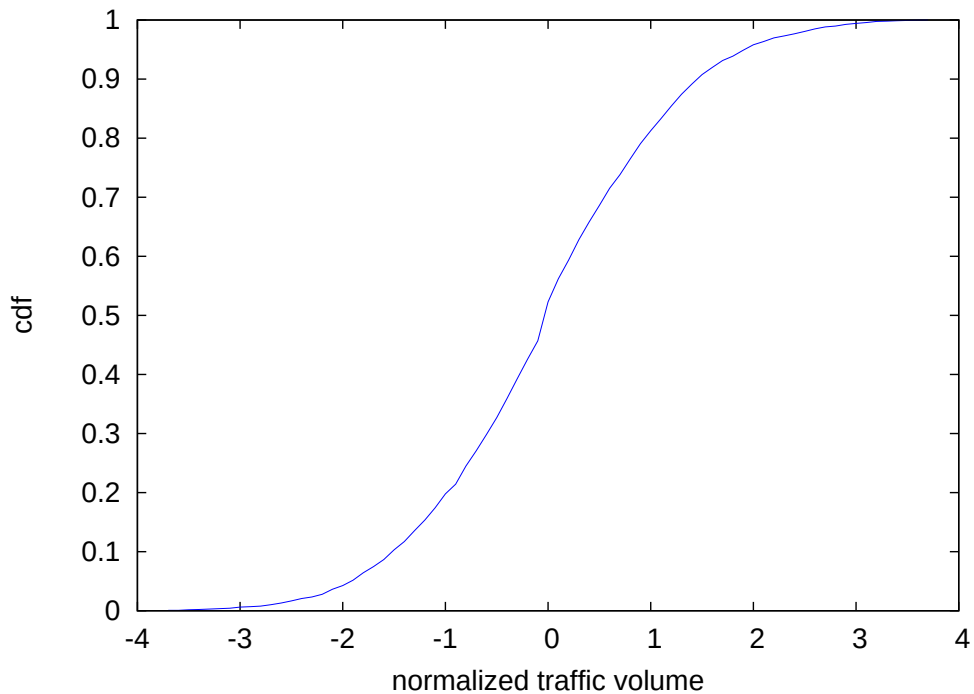
- density function: probability of observing x

$$f(x) = P[X = x]$$

- cumulative distribution function: probability of observing x or less

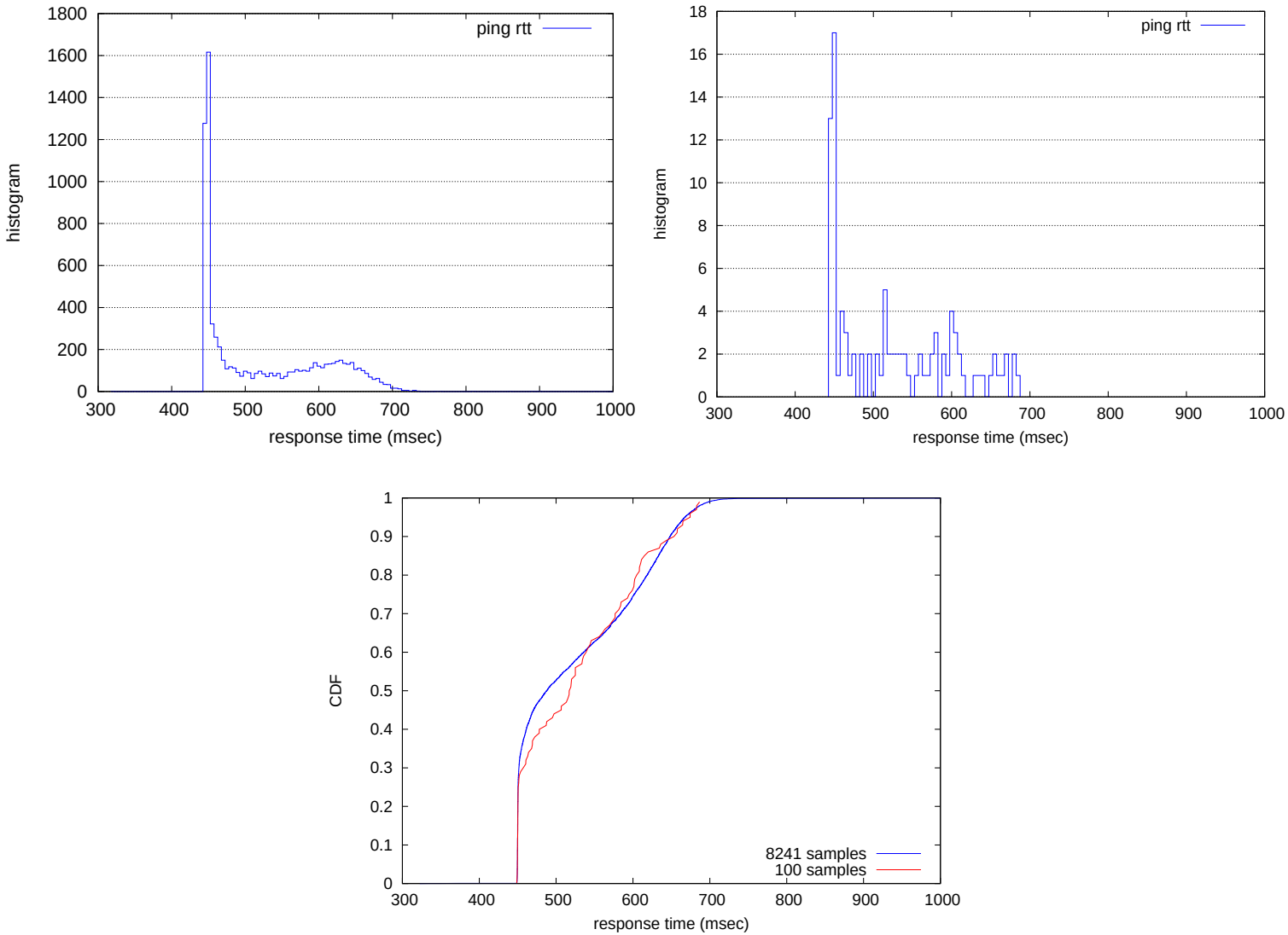
$$F(x) = P[X <= x]$$

- better than histogram when distribution is highly skewed, sample count is not enough, or outliers are not negligible



histogram vs cdf

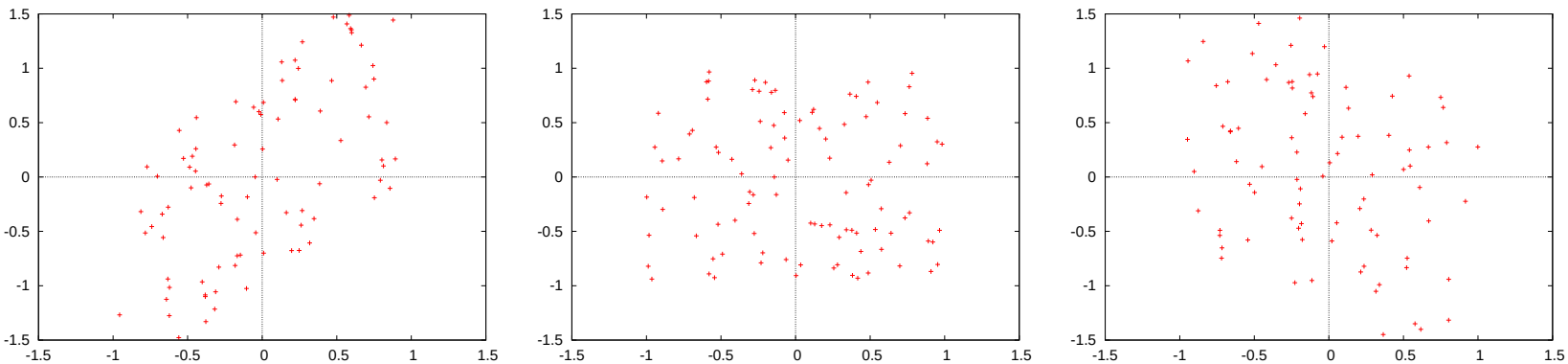
- no need to worry about bin size or sample count for cdf



original data (left), 100 samples (right), cdfs (bottom)

scatter plots

- explores relationships between 2 variables
  - X-axis: variable X
  - Y-axis: corresponding value of variable Y
- you can identify
  - whether variables X and Y related
    - no relation, positive correlation, negative correlation
  - whether the variation in Y changes depending on X
  - outliers
- examples: positive correlation 0.7 (left), no correlation 0.0 (middle), negative correlation -0.5 (right)



examples: positive correlation 0.7 (left), no correlation 0.0 (middle), negative correlation -0.5 (right)

plotting tools

- gnuplot
  - command-line tool suitable for automated plotting
  - http://gnuplot.info/
- grace
  - comes with graphical user interface
  - powerful for fine-tuning the output
  - http://plasma-gate.weizmann.ac.il/Grace/

exercise: gnuplot

- plotting a simple graph using gnuplot
- sample data from a book: P. K. Janert "Gnuplot in Action"
  - `http://web.sfc.keio.ac.jp/~kjc/classes/sfc2011f-measurement/marathon.txt`
  - `http://web.sfc.keio.ac.jp/~kjc/classes/sfc2011f-measurement/prices.txt`

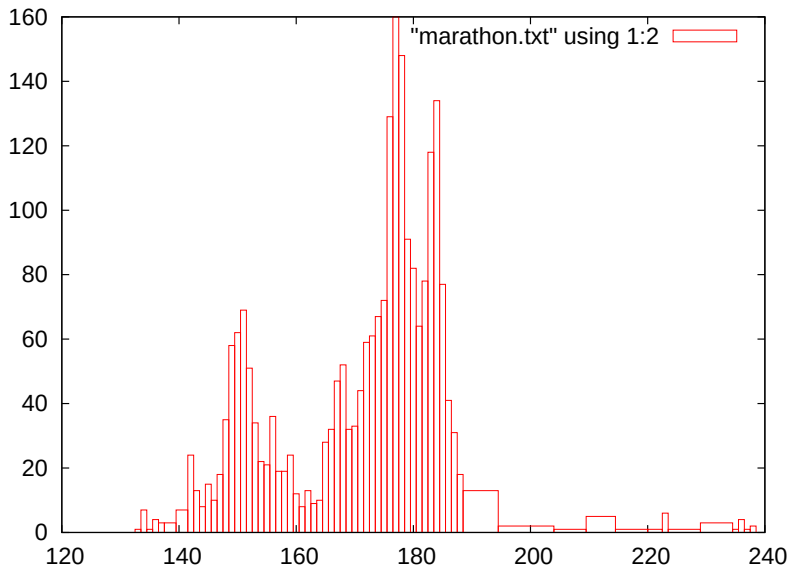
histogram

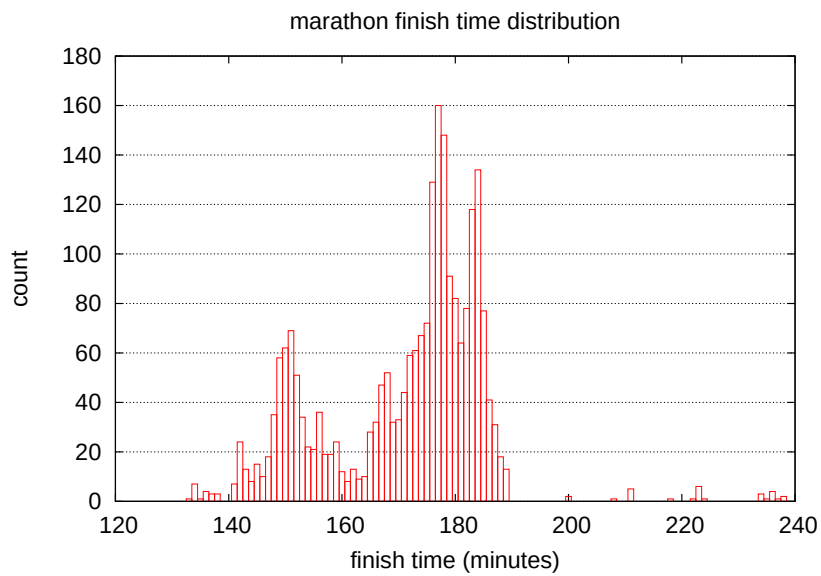
- distribution of finish time of a city marathon

```
plot "marathon.txt" using 1:2 with boxes
```

make the plot look better (right)

```
set title "marathon finish time distribution"
set boxwidth 1
set xlabel "finish time (minutes)"
set ylabel "count"
set yrange [0:180]
set grid y
plot "marathon.txt" using 1:2 with boxes notitle
```

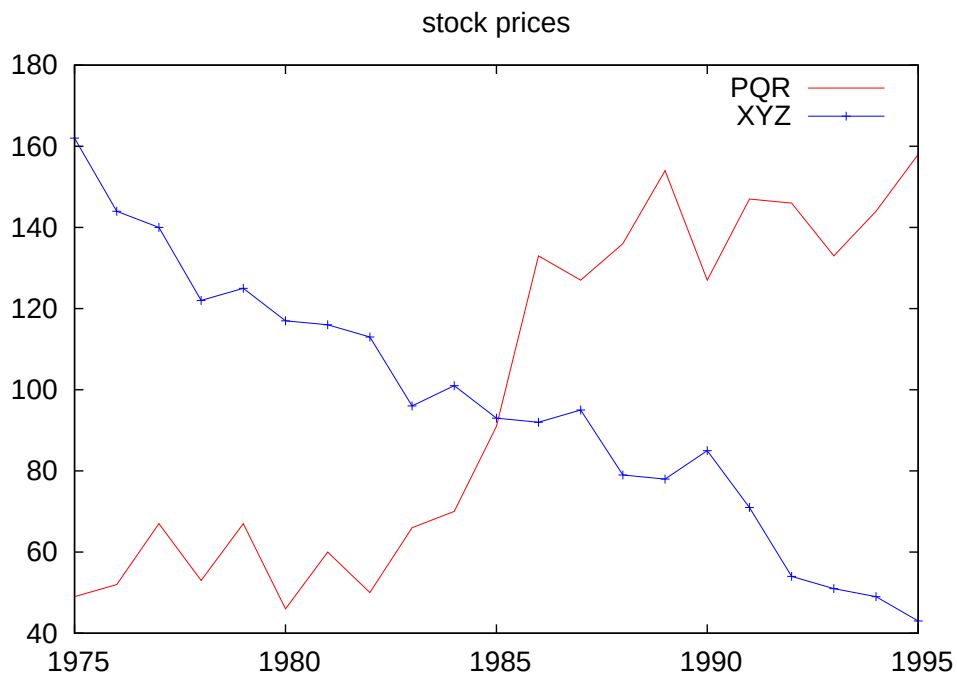




time-series plot

- stock prices over time: PQR and XYZ

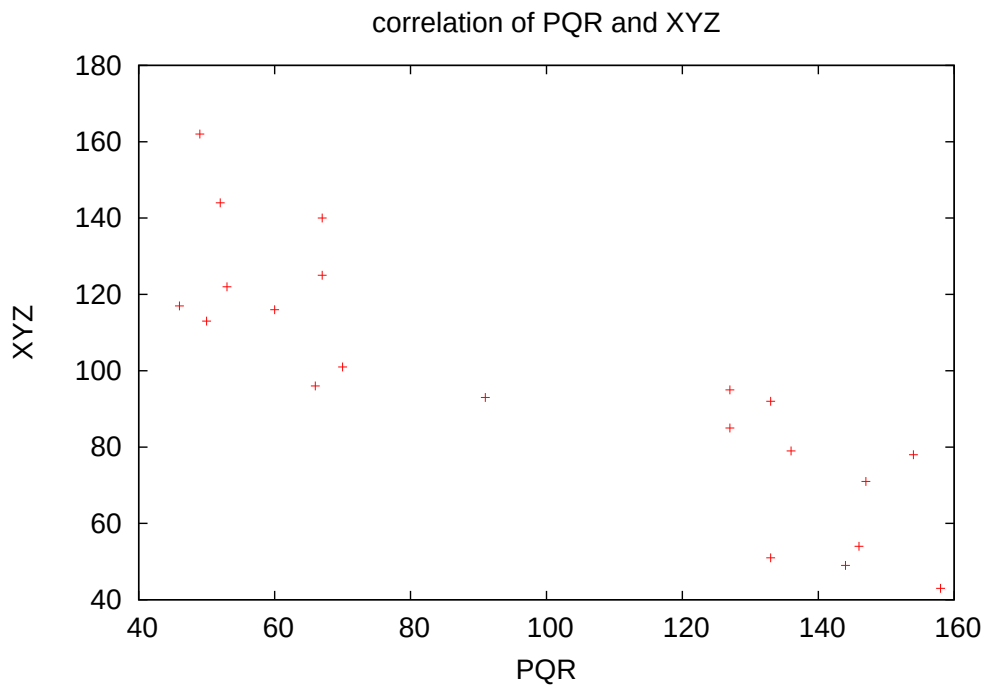
```
set title "stock prices"
plot "prices.txt" using 1:2 title "PQR" with lines, \
"prices.txt" using 1:3 title "XYZ" with linespoints linetype 3
```



scatter plot

- correlation of stock prices: PQR and XYZ

```
set title "correlation of PQR and XYZ"
set xlabel "PQR"
set ylabel "XYZ"
plot "prices.txt" using 2:3 notitle with points
```



summary

Measuring the size of the Internet

- the number of users and hosts
- the number of web pages
- precision, errors, significant digits
- how to make good graphs
- exercise: graph plotting by gnuplot

next class

Class 3 Data recording and log analysis (10/12)

- data format
- log analysis methods
- exercise: log data and regular expression