Aalto University
School of Science
and Technology

Methodology for Computer Science Research
Lecture 3: Data Analysis

Boris Nechaev

Department of Computer Science and Engineering
Aalto University, School of Science
boris.nechaev@hiit.fi

October 4, 2012

What is Data Analysis?

- A set of techniques and methods for extracting useful information from raw data.
- Data analysis helps to draw conclusions that lead to actions.
- Sometimes you have a certain question you want to answer, sometimes it's "lets see if there is anything interesting" type of work.
- Exploding amount of data in the Internet increases demand for data analysis skills.
- Various forms of data analysis are used in many fields: research, engineering, business, politics, etc.

Types of Data Analysis

Exploratory data analysis (descriptive statistics)
- Describes and summarizes the data
- Helps to look for patterns
- Uses data visualization techniques

Confirmatory data analysis (inferential statistics)
- Heavily relies on formal statistical techniques
- Is used for hypothesis testing
- Usually operates with a small number of samples

Predictive data analysis (modeling)
- Trying to forecast the future
- Essentially, boils down to building a model

Challenges of Data Analysis

Huge volumes of raw data
- Storage, transfer
- Computational complexity
- Handling, analysis

Quality of data
- Data may be corrupted or partially missing
- You may even not know about this

It's very easy to make a mistake
- May lead to disastrous consequences
- There are various rules of thumb, but no general solution

What is data?

602	593.5	589.51
610.05	587	581.9
612.08	584.9	587.34
612.8	580.1	580.19
608	595.09	583.72
601.3	594.52	599.47
605.94	598.42	550.03
604.26	589.55	548.13
609	593.1	533.46
593.32	586.72	525.18

What is data?

Google share prices:

Date	Price	Date	Price	Date	Price
18.11.11	$602.00	04.11.11	$593.50	21.10.11	$589.51
17.11.11	$610.05	03.11.11	$587.00	20.10.11	$581.90
16.11.11	$612.08	02.11.11	$584.90	19.10.11	$587.34
15.11.11	$612.80	01.11.11	$580.10	18.10.11	$580.19
14.11.11	$608.00	31.10.11	$595.09	17.10.11	$583.72
11.11.11	$601.30	28.10.11	$594.52	14.10.11	$599.47
10.11.11	$605.94	27.10.11	$598.42	13.10.11	$550.03
09.11.11	$604.26	26.10.11	$589.55	12.10.11	$548.13
08.11.11	$609.00	25.10.11	$593.10	11.10.11	$533.46
07.11.11	$593.32	24.10.11	$586.72	10.10.11	$525.18

What is data?

Data is numbers, characters or other pieces of information with context

- Without knowing the context, data will look to you as a bunch of meaningless numbers
- Context guides data analysis
- Conclusions are drawn within the context of the data
- In a broader sense, data may mean any information

Is the word "data" singular or plural?

- No consensus
- IEEE Computer Society allows both as long as you are consistent (http://www.computer.org/portal/web/publications/styleguidedef)

Data Analysis workflow

Assume you already have data from somewhere (experiment, simulation, measurement, your boss, etc)

Retrieve

- SQL query, load a file, do API request

Pre-process

- Aggregate, filter, re-structure

Analyze

- Answer your questions, look for anything interesting

Present

- Write a summary, make a plot

Tools for Data Analysis

Storage

- Database, CSV, Hadoop, tcpdump, etc

Pre-processing

- Scripting langauge (e.g. Python, Perl, Excel VBA macro)

Analysis

- R environment, SAS, Excel, SPSS, Matlab

Presentation

- R environment, gnuplot, Excel, Qlikview, Tableau

Ways to present data

Text

"Last-mile latency is generally quite high, varying from about 10 ms to nearly 40 ms (ranging from 40% to 80% of the end-to-end path latency)."

- S. Sundaresan et al, Broadband Internet Performance: A View From the Gateway, *Sigcomm'11*

- Usually used to summarize the data
- Allows to draw reader's attention to certain details
- Gives a lot of room for explanation
- May be boring and hard to comprehend

Ways to present data

Table

Table 2: Microbenchmarks for NetLord overheads

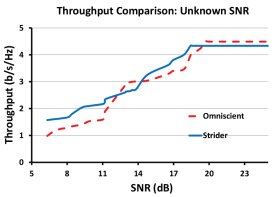
Case	Metric	PLAIN	SPAIN	NetLord
Ping (μs)	avg	97	99	98
	min/max	90/113	95/128	93/116
NetPerf 1-way (in Mbps)	avg	987.57	987.46	984.75
	min	987.45	987.38	984.67
	max	987.67	987.55	984.81
NetPerf 2-way (in Mbps)	avg	1835.26	1838.51	1813.52
	min	1821.34	1826.49	1800.23
	max	1858.86	1865.43	1835.21

- J. Mudigonda et al, NetLord: A Scalable MultiTenant Network Architecture for Virtualized Datacenters, *Sigcomm'11*

- Used to show exact values in a structured way
- Allows comparison
- There shouldn't be too many data points

Ways to present data

Graph, chart, plot

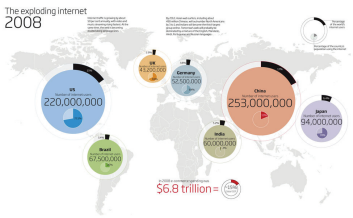


- A. Gudipati et al, Strider: Automatic Rate Adaptation and Collision Handling, *Sigcomm'11*

- "A picture is worth a thousand words"
- Can be constructed from a ton of data
- Allows to see the big picture

Ways to present data

Infographics



- NewScientist (http://www.newscientist.com/gallery/mg20227061900-exploring-the-exploding-internet/8)

- Looks super-cool!
- Usually contains much information
- Is in fact a piece of art (have to be an artist to create one)

Ways to present data

Usually different ways of presentation must be combined

- Helps to overcome limitations of each component
- E.g. a plot or a table accompanied by explanations in text
- Infographics may contain various components within themselves

The main goal is to extract meaningful information

- E.g. a good plot illustrating your point
- A piece of text answering a certain question

Data set structure

A data set consists of entities and variables.

An entity is an object described by the data (e.g. people in this classroom).

A variable is a property of an entity that can take a range of possible values (e.g. country of origin of people in this class room).

Distribution of a variable describes what values can the variable take and how often it takes these values
(e.g. Finland: 4, Sweden: 1, India: 2, China: 3, Russia: 3).

Types of variables

Categorical variable takes non-numeric values that come from a set of labels (e.g. country of origin).

Numerical variable takes numeric values that come from a set of integers, real numbers, etc (e.g. age).

Discrete variable only takes particular values (e.g. number of brothers and sisters). Categorical data is discrete.

Continuous variable can take any value (e.g. height). Continuous variable can be turned into discrete by grouping (e.g. rounding age to years instead of giving the exact age).

Types of graphs

Numerous types of graphs exist, need to know which to choose

- Analyzing a single variable
- Analyzing relationship between two variables
- Categorical vs numerical variables
- Discrete vs continuous variables
- Business vs research&engineering usage
- Number of variables vs number of dimensions
- People love graphs, they are less boring than plain text or tables

Categorical data, single variable

Example

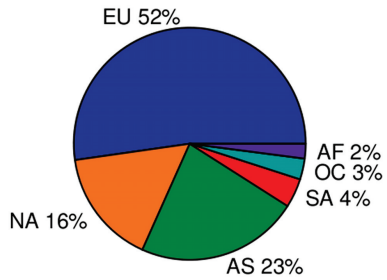
J. S. Otto et al, On Blind Mice and the Elephant – Understanding the Network Impact of a Large Distributed System, *Sigcomm'11*

The paper analyzes BitTorrent usage and its influence in the Internet. The authors collected data for 500,000 users from all over the world.

Each user's IP address was mapped to a continent. The continents are: Europe (EU), North America (NA), Asia (AS), South America (SA), Oceania (OC), Africa (AF).

How to report the shares of users from each continent?

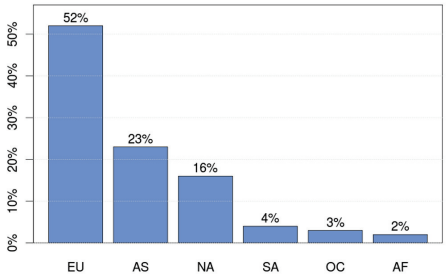
Pie chart



- J. S. Otto et al, On Blind Mice and the Elephant – Understanding the Network Impact of a Large Distributed System, *Sigcomm'11*

- Good to show relative shares
- Hard to see exact values (unless written nearby)

Bar chart (bar plot)



- Good to show exact values, relative shares
- Shows distribution of a discrete variable
- If there is no natural order, order bars by value

Categorical data, three dimensions

In the previous example we had one variable (continent) and created another dimension (share, %), i.e. the plot had two dimensions.

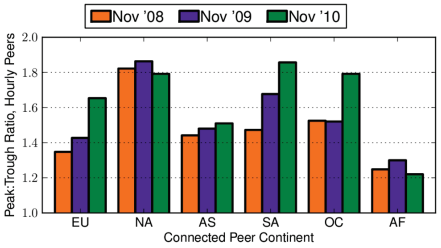
What if we have two variables and a created dimension, or three variables?

Example

J. S. Otto et al, On Blind Mice and the Elephant – Understanding the Network Impact of a Large Distributed System, *Sigcomm'11*

The authors want to check if BitTorrent usage changes with time. They want to know this for each continent. Thus, there are three variables: time, continent, usage metric. How to plot this?

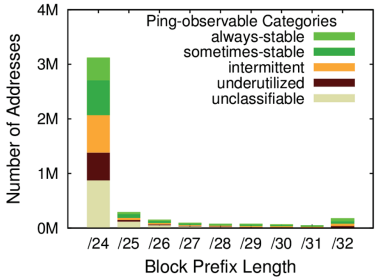
Bar chart (bar plot)



- J. S. Otto et al, On Blind Mice and the Elephant – Understanding the Network Impact of a Large Distributed System, *Sigcomm'11*

- Third dimension is encoded by color
- Easy to compare variables with each other

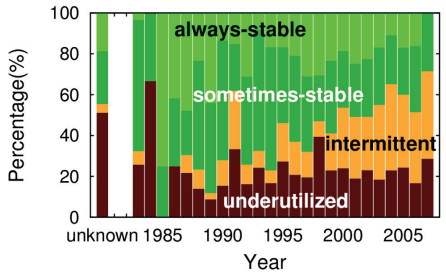
Stacked bar chart (bar plot)



- Xue Cai et al, Understanding Block-level Address Usage in the Visible Internet, *Sigcomm'10*

- Easy to see how one variable sums up within another

Stacked bar chart (bar plot)



- Xue Cai et al, Understanding Block-level Address Usage in the Visible Internet, *Sigcomm'10*

- Shows shares instead of absolute values

Numerical data, single variable

Unlike categorical variable, numerical variable can be described with various quantitative summaries.

Minimum and maximum - the smallest and the largest value the variable takes.

Mean - the average value of the variable.

Median - the value that separates the lower part of the distribution from the higher part.

Mode - the most common value the variable takes.

Mean vs median

Data: 8, 2, 3, 1, 6

Mean is $\frac{8+2+3+1+6}{5} = 4$.

To find median, sort the values and pick the middle one: 1, 2, **3**, 6, 8 - half of the values are less than the median and half are bigger.

Mean can sometimes be misleading. Consider a bar where ten programmers are hanging out. They have about the same salaries and the mean salary for all people in the bar is $3000. Bill Gates walks into the bar and mean salary skyrockets...

Mean is the "average" value, median is the "typical" value.

Measure of spread

Variance and standard deviation are the measures of spread of a variable – how far values go from the mean.

Data: 8, 2, 3, 1, 6

To find variance, first find the mean ($\mu = 4$ in this case) and then do: $\frac{(8-4)^2+(2-4)^2+(3-4)^2+(1-4)^2+(6-4)^2}{5} = \frac{16+4+1+9+4}{5} = 6.8$.

The formula for variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$.

Standard deviation σ is simply a square root of variance.

You can calculate what happened to variance when Bill Gates walked into the bar.

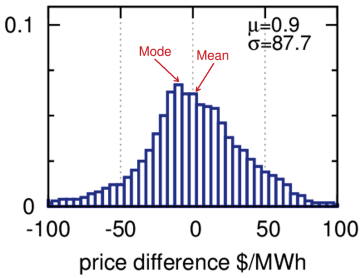
Continuous variable

Unlike a discrete variable, a continuous variable doesn't have predefined disjoint values (categories). How to plot the distribution of a continuous variable? Create categories!

The idea is to break the whole interval (from min to max) of the continuous variable into contiguous regions of the same length. Then count how many data points fall into each region and plot it along Y-axis.

Alternatively, the area of the bin drawn above the region may be proportional to the number of data points falling into the region. In this case the total area of all bins must be 1.

Histogram



- A. Qureshi et al, Cutting the Electric Bill for Internet-Scale Systems, *Sigcomm'09*

- Very useful to examine distribution of a continuous variable
- It is important to choose the right number of bins

Examining the distribution

Study the pattern

- Distribution shape
- Symmetry and skewness (in a symmetric distribution mean and median are close to each other)
- Number of modes
- Spread
- May help to choose the right model for the data

Look for outliers

- Outlier is a data point that doesn't fit into the overall pattern
- Outliers may indicate errors or unexpected findings
- Bill Gates' salary is an outlier

Percentiles (quantiles, quartiles)

Percentile is a generalization of median (50th percentile).

Nth percentile is the value which is bigger than N% of the smallest values in the data set and is less than 100%-N% of the biggest values in the data set.

0th percentile and 100th percentile are minimum and maximum of the data set.

25th percentile (denoted Q_1) and 75th percentile (denoted Q_3) are called lower and upper quartiles. Median is denoted as Q_2 (second quartile).

Another measure of spread: interquartile range $IQR = Q_3 - Q_1$.

Percentiles (quantiles, quartiles)

Example

"Figure 5 presents ... the number of flows a MLA or worker node participates in concurrently (defined as the number of flows active during a 50ms window). When all flows are considered, the median number of concurrent flows is 36, ... The 99.99th percentile is over 1,600, and there is one server with a median of 1,200 connections. When only large flows (> 1MB) are considered, the degree of statistical multiplexing is very low - the median number of concurrent large flows is 1, and the 75th percentile is 2."

- M. Alizadeh et al, Data Center TCP (DCTCP), *Sigcomm'10*

Five-number summary

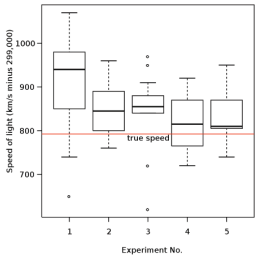
A useful way to summarize a distribution.

Consists of five numbers:

- The minimum
- The lower quartile
- The median
- The upper quartile
- The maximum

Sometimes also includes the mean.

Boxplot



- Michelson-Morley experiment (http://en.wikipedia.org/wiki/File:Michelsonmorley-boxplot.svg)

- Shows five-number summary
- Easy to compare several variables with the same range

Whiskers

Whiskers in a plot may mean different things:

- The minimum and the maximum
- Some low and big percentile
- The smallest value within 1.5 IQR of the lower quartile, and the biggest value within 1.5 IQR of the upper quartile (values outside the range are plotted as considered outliers and plotted with dots)
- One standard deviation above and below the mean of the data
- Confidence interval

The description of the plot should tell what whiskers mean.

Whiskers

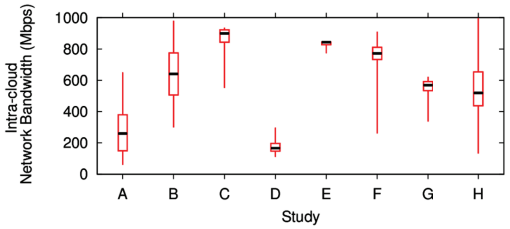


Figure 1: Percentiles (1-25-50-75-99th) for intra-cloud network bandwidth observed by past studies.

- H. Ballani et al, Towards Predictable Datacenter Networks, *Sigcomm'11*

- A boxplot with a different type of whiskers

Whiskers

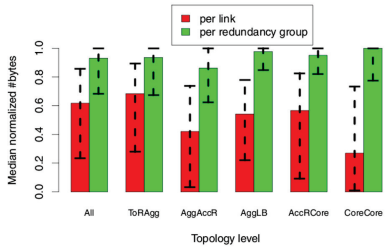
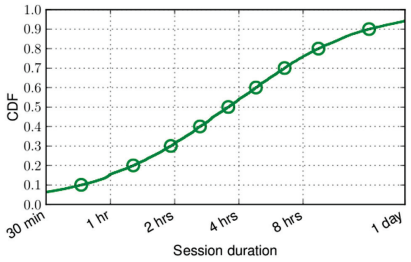


Figure 17: Normalized bytes (quartiles) during failure events per link and across redundancy group compared across different layers in the data center topology.

- P. Gill et al, Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, *Sigcomm'11*

- Not only boxplots can have whiskers

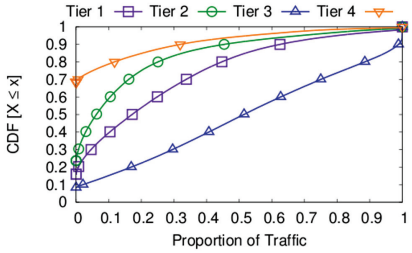
Cumulative distribution function



- J. S. Otto et al, On Blind Mice and the Elephant – Understanding the Network Impact of a Large Distributed System, *Sigcomm'11*

- Describes distribution by plotting quantiles (go from 0 to 1)
- An alternative to histogram
- Extremely popular in research literature

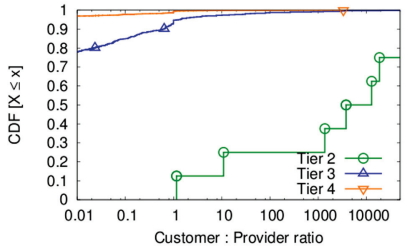
Cumulative distribution function



- J. S. Otto et al, On Blind Mice and the Elephant – Understanding the Network Impact of a Large Distributed System, *Sigcomm'11*

- Unlike histogram, allows to compare multiple variables
- Line shifted to the right in general means higher values

Logarithmic scale



- J. S. Otto et al, On Blind Mice and the Elephant – Understanding the Network Impact of a Large Distributed System, *Sigcomm'11*

- If variable spans multiple orders of magnitude, use log scale (usually powers of ten: 0.01, 0.1, 1, 10, 100, ...)
- Log scale can be used not only in CDF plots

Relationship between two numerical variables

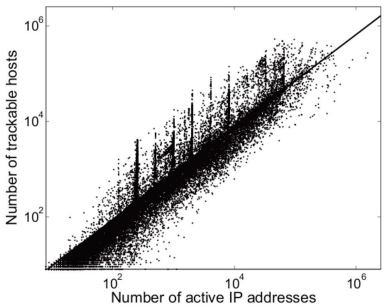
Example

Yinglian Xie et al, De-anonymizing the Internet Using Unreliable IDs, *Sigcomm'09*

The authors try to track hosts using application-level data. Due to proxies and NATs multiple hosts can share a single IP address. Authors monitored numerous IP ranges and in each range counted active IP addresses and active hosts.

What is the relationship between these two variables?

Scatter plot



- Yinglian Xie et al, De-anonymizing the Internet Using Unreliable IDs, *Sigcomm'09*

- Shows whether growth of one variable is accompanied by growth of the other

Correlation

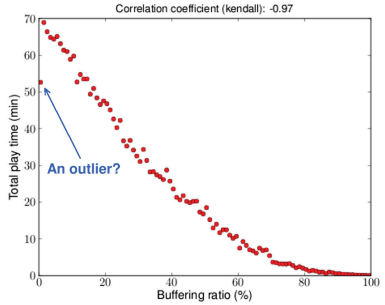
The measure of relationship between two variables is called correlation coefficient.

Correlation coefficient varies from -1 to 1. Positive correlation means that as one variable grows, the other one does so too. Negative correlation means that as one variable grows, the other one decreases.

There are several different correlation coefficients, for linear and non-linear dependence.

Strong correlation doesn't imply causation! Children's shoe size is strongly correlated with their spelling skills. This doesn't mean that bigger feet make them smarter.

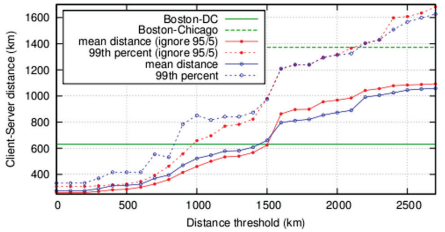
Correlation



- F. Dobrian et al, Understanding the Impact of Video Quality on User Engagement, *Sigcomm'11*

- Outliers can influence correlation coefficient

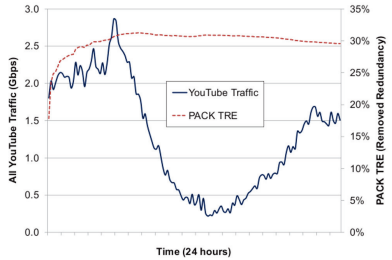
Line plot



- A. Qureshi et al, Cutting the Electric Bill for Internet-Scale Systems, *Sigcomm'09*

- A scatter plot with points connected with lines
- Good to show several variables with the same range

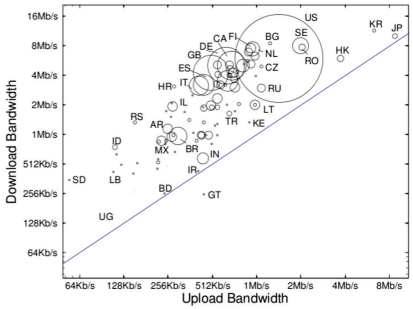
Time series



- Eyal Zohar et al, The Power of Prediction: Cloud Bandwidth and Cost Reduction, *Sigcomm'11*

- A line plot with time as one dimension
- Numerous techniques for analysis of time series exist

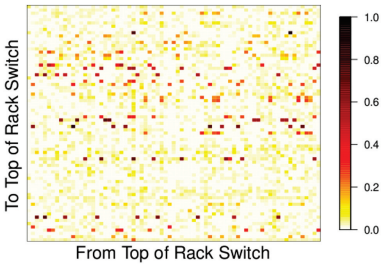
Bubble plot



- C. Kreibich et al, Netalyzr: Illuminating The Edge Network, *IMC'10*

- Circle radii show number of data points at this coordinate
- More suitable for discrete data

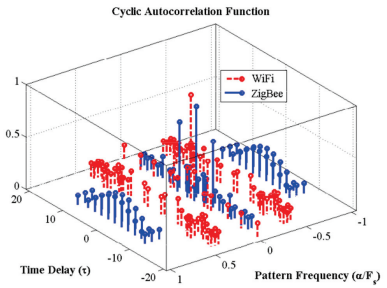
Heat map



- Daniel Halperin et al, Augmenting Data Center Networks with Multi-Gigabit Wireless Links, *Sigcomm'11*

- Color shows number of data points at this coordinate
- More suitable for continuous data

Now in 3D!



- Steven Hong et al, DOF: A Local Wireless Information Plane, *Sigcomm'11*

- Don't forget to put on your 3D glasses!
- 3D to 2D projection may make 3D plots hard to understand

General tips

Unfortunately, you can never be sure your analysis is totally correct (same as in programming, you can never be sure you didn't make any bugs).

Several things that help you do as little mistakes as possible:

- Be careful
- Learn the data before you start the analysis
- Break the whole task into subtasks
- Manually check results of each step (subtask)
- Have everything automated (manual work leads to errors)
- Have a master script – you run it, it reproduces all results
- If possible, keep raw and intermediate data