BigData: Data Analytics for Business and Beyond

Qiwei Yao
Department of Statistics
London School of Economics

q.yao@lse.ac.uk

- Data analytic thinking
- Principles and concepts
- Illustration with R – a versatile statistical package

Chapter 1. Introduction: Data Analytics

Data Analytics: the science of examining data in order to draw conclusions. It enables companies and organization to make Data-Driven Decisions (DDD). It can be used to verify or disprove existing hypotheses and theories.

Data Science: extract information from data such as undiscovered patterns, hidden relationships etc.

Difference: not really, though Data Analytics focuses on application and conclusions while Data Science on info extracting and knowledge discovery. But both involve making inference from data.

Ability to view various problems from a data perspective, understand the principles for extracting info from data:

intelligence quotient emotional quotient data quotient

Additional reading: Chapters 1 &2 of Provos and Fawcett (2013).

It's all about data:

- Business data: on customers, portfolio, sales, marketing, pricing, financial, risk, and fraud. For example,
  - shopping basket analysis to increase sales and cross selling
  - customer segmentation for tailored advertising and sales promotions
  - consumer data across multiple service channels (branch, web, mobile)
- Industrial process data: automate and control industrial production, manufacturing, distribution, logistics and supply chain processes.
  - sensors and actuators at the field level
  - control signals at the control level
  - operation and monitoring data at the execution level
  - schedules and indicators at the planning level
- Financial market data: historical records over long time periods and with increasing granularity (e.g. high-frequency data, limit order book, continuous market data)
- Personal data: demographics/geographics (factual) information, personality (behaviour) information, psychographics (attitudinal) information etc
- Text and unstructured data: text documents, company reports, news, messages, emails, web based data bases (the so-called deep web), in order to filter, search, extract, and structure information.

- Image data: image sensors, smartphone, satellite cameras, to find and recognize objects, analyze and classify scenes, and relate image data with other information sources

- Biomedical data: lab experiment data, DNA sequences, to understand and annotate genome functions...

- .....

New York Times story from 2004: Hurricane Frances

Hurricane Frances was on its way, threatening a direct hit on Florida's Atlantic coast.

Executives at Wal-Mart stores see the situation offered an opportunity to use DDD to predict unusual sales pattern.

Water, flashlights — common sense, no DDD required.

By mining through the trillions bytes of sales history when Hurricane Charley struck earlier, it was revealed that the sales for strawberry Pop-Tars increase 7 times, also a particular DVD film sold out (a coincident?). But the pre-hurricane top-selling item was beer.

`http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html?_r=0`

Predicting Customer Churn

Churn: a customer switches from one company to another.

It is expensive all around: one company must spend on incentives to attract a customer while another company loses revenue when the customer departs.

A marketing budget is allocated to prevent churn, as attracting new customers is more expensive than retaining existing ones.

Suppose you work for a credit card company which lose about 20% of customers per annum. Your task is to devise a detailed plan (DDD) on how and which customers be offered a retention deal.

Home equity loans of Bank of America: Failing to attract enough good customers in spite of several mail campaigns

Bank could lower interest rates to bring in more customers at the expense of lowered earnings. Existing customers may switch to the lower rates, further depressing earnings. Assuming the original rates were reasonably competitive, lower rates might bring in the disloyal customers.

Business consultants based on their marketing expertise provide the insights:

- people with college-age children want loans to pay tuition fees
- people with high but variable incomes want loans to smooth out the income fluctuations

BoA store the data of its millions of customers in a large relational database on a powerful parallel computer from `Teradata`. The data were recorded from 1914. More recent records has about 250 attributes, including income, number of children, type of home, etc.

Decision trees derived rules to classify existing customers into two categories: likely or unlikely to respond to a home equity loan offer. This adds a new attribute to each individual in the database: `likely responder` or `unlikely responder` to a home equity loan.

Then cluster analysis is performed to automatically segment the customers into groups with similar attributes. The 14 clusters were found, and many of them did not seem particularly interesting. Nevertheless one cluster has two intriguing properties:

- 39% of the people in the cluster has both business of personal accounts
- the cluster contains more than 25% of the customers with the attribute `likely responder` to a home equity loan.

People might be using home equity loans to start businesses

Change the campaign message from 'use the value of your home to send your kids to college' to 'now the house is empty, use your equity to do what you've always wanted to do'

The response rate for home equity campaigns increased from 0.7% to 7%.

Gmail is capable of classifying users into 'millions of buckets'

A new business model: free services in exchange for personal information.

Google is the world's largest advertising company: After only 15 years in business Google makes more money from ads than all the world's newspapers combined.

750 million Gmail users in October 2014, 900 million in May 2015

In late 2010 two obscure trial lawyers in Texas made what was to them a momentous discovery: ads in Gmail are correlated with keywords contained in the emails. This triggers a long lawsuit, to seek for billions of compensation from Google. The case 'ends' in July 2014 with one out-of-court payment for a single plaintiff.

- Google holds adequate user consent
- Gmail does not make much money from ads: 70+% Gmail users never click on ads.

From its earliest days Gmail was intended to be a money-making product. Instead of relying on demographic information users provided about themselves at sign up, Gmail would attempt to grasp the actual meaning of user messages and target ads accordingly.

Gmail's limitless data mining ambitions: One patent filed in June 2003 described a lengthy series of message attributes that could be used in any combination to extract the meaning of an email and select the best ads to match it.

Gmail profiles all its users: an old idea on a new (gigantic) scale – tailoring messages to specific audiences increases the advertiser's return.

Google's advertising business can be viewed as a giant but *sparse* spreadsheet with hundreds of thousands of advertisers aligned across the top and hundreds of millions of users down the left side: more than 99.99% of the cells are empty.

Nielsen PRIZM, a system developed in 1970s, use sophisticated clustering algorithms to divide Americans into such marketer-relevant buckets as Upper Crust, Blue Blood, Young Digerati, Beltway Boomers, Rustic Elders, Back Country Folks and Hard Scrabble, among dozens of others.

Google: clusters extracted by the PHIL algorithm from documents the user has viewed (web pages, inbound emails) or created (outbound emails, social media posts).

Inbound emails to Gmail users are of particular value: the emails received obviously include messages from family and friends, social media notifications, newsletters subscribed and whatever commercial offers have made it through your spam filter settings. They also typically include a large amount of data-rich correspondence from institutions: banks, utilities, schools, tax authorities, internet providers, TV companies and, last but not least, online merchants such as Amazon, eBay or travel reservation sites where you have made purchases. Taken together, these inbound messages discriminate an individual from other users with a high degree of granularity.

What is Big Data?

From a *Google* search in 2012:

> data which are too extensive to permit iterative analysis: one-pass analysis is necessary
>
> data sets which standard database tools cannot handle
>
> data which exceeds 20% of the RAM of a given machine

From *Siri* in 2017:

> Information assets characterized by such a high volume, velocity and variety to require specific technology and analytic methods for its transformation into value

From *Wikipedia* in 2018:

> A term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy.

4V or 5V: Volume, Variety, Velocity, Value and Vanity.

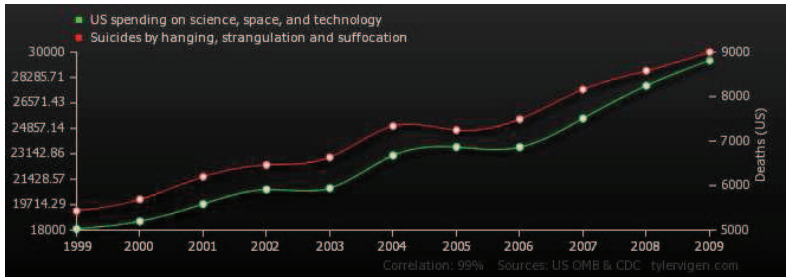
Volume: large size of data sets, or a large number of small data sets

Variety: lack of homogeneity in format, structure, quality

Velocity: high speed of data generation and data processing

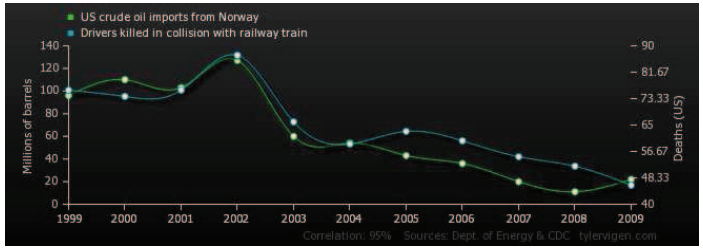
Value: valuable information buried in data sea

Vanity: 'everything' significant, spurious correlations

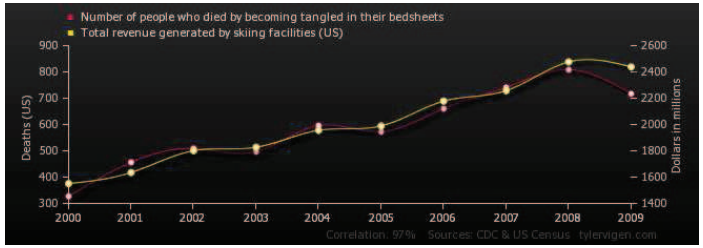


Correlation=0.992082

http://www.tylervigen.com/



Correlation=0.954509



Correlation=0.969724

Some big data stories

- Astronomy – Digital Sky Survey's archive in 2010: 140 terabytes
- Sequencing the Human Genome: 3.3 billion base pairs
- US equity markets: 7 billion shares change hands everyday
- Social network: quintillion (10^{18}) bytes per day
- Climate modelling: Coupled model intercomparison project 5th phase: more than 2 petabytes
- Google Translate: statistical machine translation; 200 billion words

Digit data expands quickly: doubling almost every 3 years

Big Data: not new!

- 1994: Wal-Mart, with over 7 billion transactions
- 1997: AT&T, with over 70 billion long distant phone call records
- 1990s: Mobil Oil, over 100 terabytes of data
- 2000: in just a few months the Sloan Digital Sky Survey collected more data than had previously been collected in the entire history of astronomy

Why now?

- automatic data capture in large scale (often secondary)
- exponential growth in computer memory and speed: making storage and computing cheap
- Data Analytics (especially Artificial Intelligence):

Data → Information → Value → Profits

Value shifting from Physical infrastructure (land, factories) to intangibles such as brands, intellectual property and now data.

Why is it exciting?

A new world, according to many!

McKinsey & Company: *'we are on the cusp of a tremendous wave of innovation, productivity, and growth, as well as new modes of competition and value capture, all driven by big data as consumers, companies, and economic sectors exploit its potential'*

Chris Anderson: *'Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves'.*

(From 'The end of theory: the data deluge makes scientific method obsolete' in Wired.)

But numbers do not speak for themselves!

= a half of  **per person & day**

Data do not speak for themselves: positive or negative framing in data presentation can change the emotional impact.

A London Underground Poster (in 2011):

> '99% of young Londoners do not commit serious youth violence'

As 1% of young Londoners (aged between 15 and 25) $\approx$ 10,000 people, the message of the above poster is equivalent to:

> 'There are 10,000 seriously violent young people in London'.

More, Messy, Good enough: 3 shifts in extracting information

- More: More data with increasing granularity not only require more advanced IT techniques, but also change our way of thinking and ambition in analysing data
- Messy: data are complex in formats and structure, varies in quality
- Good enough: often satisfied with discovering correlations, patterns instead of causality — looking for *What* instead of *Why*

Some hedge funds parse Twitter to predict stock markets

Amazon and Netflix base product recommendations on a myriad of user interactions on their sites.

Twitter, LinkedIn, Facebook map users' social graph of relationships to learn their preferences.

What's the cancer risk from bacon scanwiches?

In November 2015, the International Agency for Research in Cancer (IARC) announced: processed meat is a 'Group I carcinogen', putting bacon in the same category as cigarettes

Daily Record (a Scottish newpaper) published a headline: Bacon, Ham and Sausages Have the Same Cancer Risk as Cigarettes Warn Experts

To quell the public panic caused, IARC stated: Group I classification is about being confident that an increased risk of cancer exists, $\cdots$ 50g processed meat a day is associated with an increased relative risk of bowel cancer of 18%.

The increased absolute risk of bowel cancer is merely 1%!

The risks for non-bacon-eaters and bacon-eaters are, respectively, 6% and 7%. The relative risk is defined as relative odds-ratio: $\frac{7/93}{6/94} = 1.18$, therefore an increase of relative risk $0.18 = 18\%$.

Big Data: more data = more info? or more complexity and more noisy?

Joint force from computer science, statistics and applied mathematics is required to tackle the challenges with Big Data

Modern Data Analytics is typically teamwork!

Computer Science: manipulating data including capturing, storing, sorting, searching, selecting, aggregating, concatenating, etc

Statistics: extracting information from data, making inference

Applied Mathematics: complexity leading to many new mathematical challenges, and requiring new models and new algorithms

Big data does not mean end of 'small data'

David Hand's Power law for data set size: The probability of observing a data set of size n is inversely related to a power of n.

There are vastly more small data sets than very large ones, which are also important assets for data analytics.

'Big' is not necessarily more, useful, valuable, or interesting: it is possible to be data rich but information poor!

The future is not big data, but what we learn from it.

Your big data problem might be a small data problem in disguise, or be a large number of small data problems.

Most of the problems we want to solve are inferential.

The ultimate goal of Big Data is to forecast future.

With the advancement of big data analytics, we would be able to predict accurately the likelihood of

– a malfunction of machine (service in advance)

– a heart attack (pay more for health insurance)

– default on a mortgage (be denied a loan)

– committing a crime (perhaps be arrested in advance???)

Some Ethical issues/legal challenges with Big Data

- individual privacy versus data collection, sharing and usage: require new rules to safeguard the sanctity of the individual
- personalised service versus exploring/manipulating
- incomprehensible nature of data-driven-decisions: the data dictatorship shifts the world from causation to correlation
- self-regulation versus global law enforcement
- .....

GDPR: The General Data Protection Regulation (EU) 2016/679 (`eugdpr.org`) became effective 25 May 2018.

Cambridge Analytica – A UK political consulting firm combining data mining, data brokerage, and data analysis with strategic communication during electoral processes

Started in 2013, and closed down in May 2018 due to the Facebook-Cambridge Analytica data scandal

> Today in the United States we have somewhere close to four or five thousand data points on every individual ... So we model the personality of every adult across the United States, some 230 million people.
>
> – Alexander Nix, CEO Cambridge Analytica, October 2016.

Nix claimed that CA provided service to

- 44 US political races in 2014
- Ted Cruz' presidential campaign in 2015
- Donald Trump's presidential campaign in 2016
- `Leave.EU` in 2016

CA's role in those campaigns is controversial, and is subject to criminal investigations in USA and UK.

Political scientists and the clients (including Trump's aides) also question CA's claims about the effectiveness of its methods of targeting voters.

CA's method: 'psychographic analysis' based on data enhancement and audience segmentation techniques such as *Big Five model* of personality, to 'find your votes', and move them to action by personalized political adverts.

Big Five personality traits, also known as five-factor model or *OCEAN* or *CANOE* model:

- Openness to experience (inventive/curious vs. consistent/cautious)
- Conscientiousness (efficient/organized vs. easy-going/careless)
- Extraversion (outgoing/energetic vs. solitary/reserved)
- Agreeableness (friendly/compassionate vs. challenging/detached)
- Neuroticism (sensitive/nervous vs. secure/confident)

Channel 4 News investigation: lasted 4 months started in Nov 2017

An undercover reporter posed as a potential customer, hoping to help Sri Lankan candidates get elected. Video footage of the investigation was published on 19 March 2018. Within 7 weeks, CA collapsed as the "siege of media coverage has driven away virtually all of the Company's customers and suppliers".

From the footage, Nix said that his company uses honey traps, bribery stings, and prostitutes, for opposition research. He offered to discredit political opponents in Sri Lanka with suggestive videos using 'beautiful Ukrainian girls' and offers of bribes...

Cambridge Analytica immediately released a statement that the video footage was 'edited and scripted to grossly misrepresent' the recorded conversations and company's ethics and business practices.

On 17 March 2018, Christopher Wylie, LLB from LSE and former employee of CA, alleged in an interview with *Observer* that CA 'exploited Facebook to harvest millions of people's profiles' and used the data to target voters with personalized political adverts.

Personal data were collected via an app called 'This Is Your Digital Life' created by a Cambridge academic Dr Aleksandr Kogan. At the time, Facebook allowed app to collect data not only about app users but also their Facebook friends. As the result the personal data from 87 million people were acquired via the 0.27 million app users.

UK Information Commissioner's Office (ICO) confirmed:

> Dr Aleksandr Kogan and his company GSR, harvested the Facebook data of up to 87 million people worldwide, without their knowledge. A subset of this data was later shared with other organisations, including SCL Group, the parent company of Cambridge Analytica who were involved in political campaigning in the US.

In announcing its winding down in May 2018, Cambridge Analytica said it has 'unwavering confidence that its employees have acted ethically and lawfully', while the ICO said it would 'continue its civil and criminal investigation'.

On July 2018, several former Cambridge Analytica staff launched a new company 'Auspex International' in the field of data analytics and work in Africa and the Middle East initially.

`https://www.theguardian.com/news/series/cambridge-analytica-files`

One serious issue has raised from the story: almost irreconcilable tension between

- legal consent, i.e. ticking box for long and often incomprehensible consent and privacy statement, and
- moral and informed consent, i.e. what users actually feel comfortable with

In modern networked societies, privacy is a shared responsibility by individuals, friends, companies and governments across the global.

A critical skill in data science: decompose a data analytics problem into pieces such that each piece can be solved by a known or *newly invented* data-mining method. A data-mining task can be viewed as a process of learning from data by a computer, called machine learning, or by a statistical method, termed as statistical learning.

`Data mining, Information extraction, Knowledge Discovery: A craft`

Most frequently used data analysis methods

- Classification. *Among all customers of EDF, who are likely to switch to another energy supplier?*
- Regression (i.e. value estimation.) *How much will a given customer use the service?*
- Similarity matching. *Identify individuals who are similar to your most loyal customer group.*
- Clustering. *How should our customer care teams be structured?*
- Market-basket analysis. *Should beers be placed next to baby napkins in a supermarket?*
- Link prediction. *As you and John share 10 friends, maybe you would like to be John's friend?*
- Causal modelling. *Is the increase of sales caused by a particular advertisement?*
- Network analysis. *How do social network structures affect, disease spreading, information dissemination, human behaviour?*

Most data-mining tasks can be divided into two categories:

- *Supervised learning*: learning with a set of "training data" with known labels (such as *Classification*) or relationships (such as *regression*).
- *Unsupervised learning*: learning without training data (such as *Clustering, Co-occurrence grouping, Profiling*).

In terms of purposes, data-mining tasks may be viewed as two types:

- Discovering/Inference: understand patterns and/or relationships within data.

  The Wal-Mart/Hurricane example: understand the sales pattern before hurricane

  Simple, transparent with easy interpretations

  Difficult with big and/or complex data (e.g. financial market).

- Predictive: learning results are used to predict unknown values.

  Which customers will respond to a particular ad?

  Prediction accuracy is of primary concern, 'black-box' methods are often used.

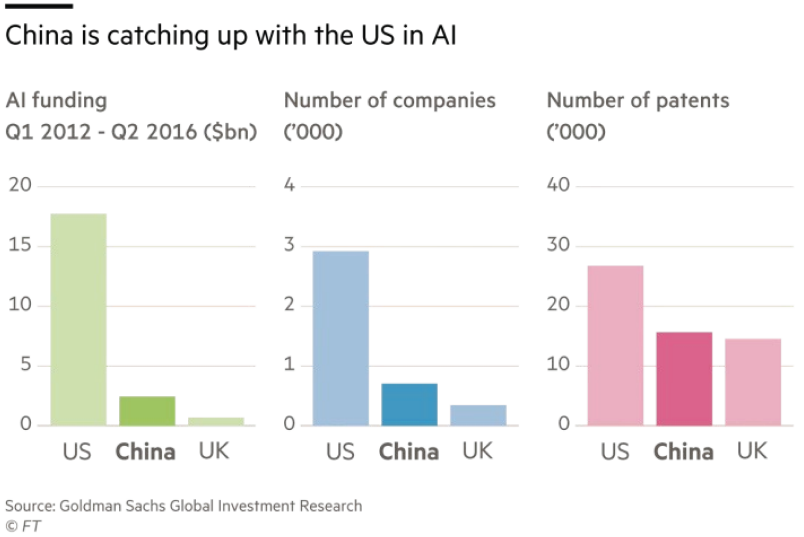
Artificial Intelligence (AI) and Deep Learning: the reincarnation since 2010 of neural networks (which was sidelined in the mid 1990s) due to

- massive improvements in computer resources
- some methodological innovations
- ideal niche learning tasks (image/video classification, speech and text processing)

AI support, or AI takeover?

Narrow AI: the tasks of which rules and sets of possibilities are unchanging

General AI: interact with and respond to multiple changing environments, requiring empathy, creativity, critical thinking, and dreaming – a distant prospect, with an unclear timeline for development



The AI arms race

Vladimir Putin: "Whoever becomes the leader in this sphere (i.e. AI) will become the ruler of the world."

In April 2018, 25 EU counties signed an agreement to collaborate on AI with the investment target of 20 billion euros by the end of 2020 (`bit.ly/2HGJ3p9`).

The UK House of Lords report "AI in the UK: ready, willing and able?" (April 2018). It was convinced that the UK can lead in AI, building on a historically strong research programme (`bit.ly/2HGHhEv`).

The 3 key factors for AI development:

Data Algorithms/Scientists High-Tech

Five principles (UK House of Lords report):

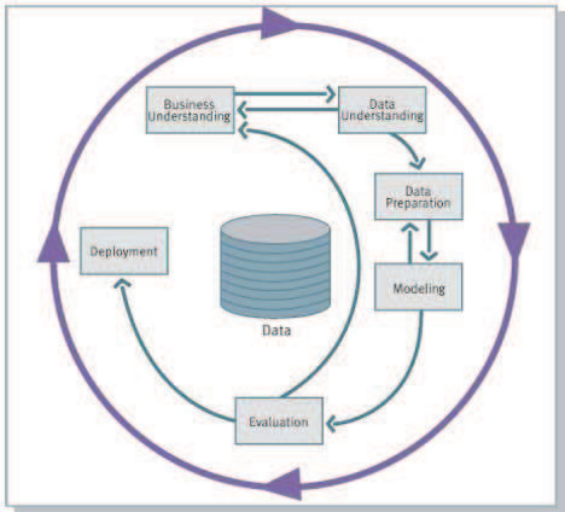
1. AI should be developed for the common good and benefit of humanity
2. AI should operate on principles of intelligibility and fairness
3. AI should not be used to diminish the data rights or privacy of individuals, families or communities
4. All citizens have the right to be educated to enable them to flourish mentally, emotionally and economically alongside AI
5. The autonomous power to hurt, destroy or deceive human beings should never be vested in AI.

Robust AI and Interpretable AI

IT aspect of Data analytics – important but not covered in this course

It is hard to imagine a working data scientist who is not proficient with some software tools!

- Computer programming: packages (R, SAS, Matlab) and programming languages (Python, JAVA, C++, C#)
- Managing data: data cleansing, data structures, databases, data querying.
- Data visualisation (such as `Tableau, ggplot2`)
- Parallel computing and distributed processing (such as Hadoop/Spark)



CRISP-DM Process diagram: iteration is the rule

Data analytical thinking: Extracting useful info from data can be treated systematically by following a process with reasonably well-defined stages. For example,

> CRISP-DM: the Cross Industry Standard Process for Data Mining.
>
> `https://the-modeling-agency.com/crisp-dm.pdf`

CRISP data mining process:

1. *Business understanding*

   Understand the problem to be solved: business projects seldom come pre-packaged as clear and unambiguous. Recasting the problem and designing a data analytic procedure is typically an iterative process (see the diagram); requiring business knowledge, data analytic creativity/experience and common sense.

   - What exactly want to do?
   - How exactly do it?
   - Which statistical techniques/methods are relevant?

   This aspect will be further elaborated late in the course.

2. *Data understanding*

   Data – the raw material from which the solution will be built.

   Strengths and limitation of data?

   Data are typically collected without explicit purposes, are often with varying degrees of reliability.

   Costs and benefits of each data source: both collecting and analysing costs.

   For example, Data mining has been used extensively for *fraud detection*

   Catching credit card fraud: supervised data mining

   Catching medicare fraud: unsupervised data mining

3. *Data Preparation* — often proceeding along with data understanding

   Data cleaning, removing outliers, inferring missing values, converting data format, data transformation etc

4. *Modelling*: apply relevant machine/statistical learning methods

5. *Evaluation*: gain confidence by assessing modelling results

   If one looks hard enough at any dataset, one will find patterns – not survive careful scrutiny.

   Qualitative evaluation: model may be accurate in lab, but much less so in actual business context.

   Fraud detection, Spam detection etc typically produce too many false alarms. (How much would it cost to deal with the false alarms? What would be the cost in customer dissatisfaction?)

   Leading to a better business understanding – iterations

   Qualitative evaluation: make models comprehensible to stakeholders who need to 'sign off' before any deployment

6. *Deployment*

   Deployment the results and, increasingly, the data mining techniques themselves in real use, in order to adapt to the changing world.

   - A model used for real-life predicting the likelihood of churn to help churn management
   - A fraud detection model is used to alarm possible fraud cases
   - Online advertising and recommendation etc

Identifying likely buyers for a sport-utility vehicle: a case study
by Wei-Xiong Ho & Joseph Harder, Southern Illinois University

In 1992, one of the big three U.S. auto makers entrusted SIU to develop an *expert system* to identify likely buyers of a particular sport-utility vehicle.

The initial challenge was to improve response to a direct mail campaign for a particular model. The campaign involved sending an invitation to a prospect to come test-drive the new model. Anyone accepting the invitation would be offered a free pair of sunglasses.

Very few people were returning the response card or calling the toll-free number for more information, and few of those that did ended up buying the vehicle.

A lot of money could be saved by not sending the offer to people unlikely to respond!

Merging data sets

As is often the case when the data to be mined is from several different sources:

mail file : a mailing list containing names and addresses of about a million people who had been sent the promotional mailing

Appendix of *mail file* contains zip codes with demographic and *psychographic* characterizations of the neighborhoods associated with the zip codes.

response file : a list of people who sent back the response card

call file : a list of people who called the toll-free number from more information

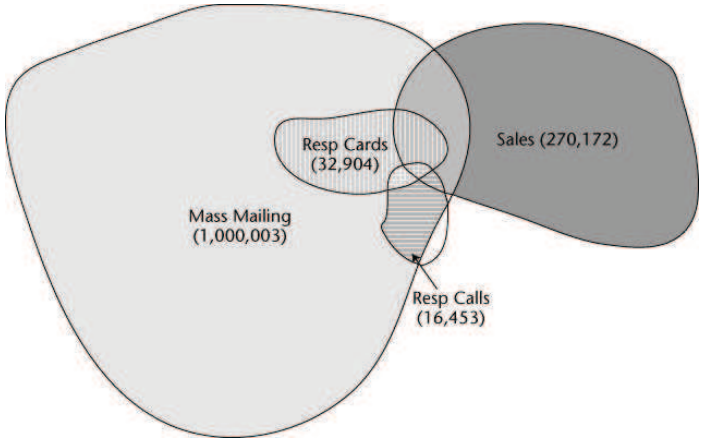
sales file : a list of people who bought cars within the 3-months following the mailing, containing the info on names, addresses, model purchased

Linking the response cards back to the original mailing file was simple because the mail file contained a nine-character key printed on the response cards.

Telephone responders presented more of a problem since their reported name and address might not exactly match their address in the database, and there is no guarantee that the call even came from someone on the mailing list since the recipient may have passed the offer on to someone else.

The initial response rate 5%: Of 1,000,003 people who were sent the mailing, 32,904 responded by sending back a card and 16,453 responded by calling the toll-free number

The total sales in the period is 270,172. An automated name-matching program with loosely set matching standards discovered around 22,000 apparent matches between people who bought the car of the advertised model and people who had received the mailing. Hand editing reduced the intersection to 4,764 people.



Simple classification

success was defined as *received a mailing and bought the car* and failure was defined as *received the mailing, but did not buy the car*.

A series of trials was run using decision trees and neural networks. The tools were tested on various kinds of training sets. Some of the training sets reflected the true proportion of successes in the database, while others were enriched to have up to 10 percent successes and higher concentrations might have produced better results.

The neural network did better on the sparse training sets, while the decision tree tool appeared to do better on the enriched sets.

An improved two-stage approach: First, a neural network determined who was likely to buy a car, any car, from the company. Then, the decision tree was used to predict which of the likely car buyers would choose the advertised model.

This two step process proved quite successful. The hybrid data mining model combining decision trees and neural networks missed very few buyers of the targeted model while at the same time screening out many more nonbuyers than either the neural net or the decision tree was able to do.

Too simplistic! For example, people who test-drive one model, but end up buying another should be in a different class than nonresponders, or people who respond, but buy nothing. People who did not received ad but bought the car are in an even more interesting group.

Resulting actions

Armed with a model that could effectively reach responders the company decided to take the money saved by mailing fewer pieces and put it into improving the lure offered to get likely buyers into the showroom. Instead of sunglasses for the masses, they offered a nice pair of leather boots to the far smaller group of likely buyers. The new approach proved much more effective than the first.

Iterating the Cycle

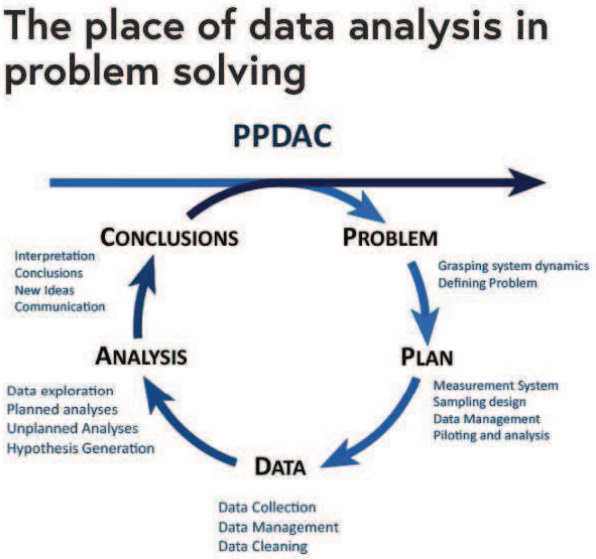
The approach described above used only a limited number of broad-brush variables and was crude and too simplistic by today's standard, in spite of its success in improving the effectiveness of a direct marketing campaign for a big-ticket item like an automobile.

The next step is to gather more data, build better models, and try again!

A cautionary remark. If you look too hard at data, you may find something which might not generalize beyond this particular data set

Overfitting, spurious correlation, incidental causality

Formulating data analytical solutions and evaluating the results involves thinking carefully about the context in which they will be used. Intuition, creativity, common sense, and domain knowledge can be brought to bear.



Descriptive Data Analysis in R (I)

What is *R*: an environment for data analysis and graphics based on *S* language

- a full-featured programming language
- freely available to everyone (with complete source code)
- easier access to the means of handling BigData such as parallel computation, Hadoop, distributed computation.
- official homepage: http://www.R-project.org

1.1 Installation

Installing *R*: R consists of two major parts: the base system and a collection of (over 10K) user contributed add-on packages, all available from the above website.

To install the base system, click on `download R` at http://www.R-project.org, then choose a mirror site close to you. Follow the link that describes your operating system: *Windows, Mac, or Linux*.

Note. The base distribution comes with many high-priority add-on packages such as graphic systems, linear models etc.

After the installation, one may start R by doubt-clicking the logo 'R' on your desktop in Windows or Mac. An R-console will pops up with a prompt character like '>'. You can input commands in the R language at the prompt.

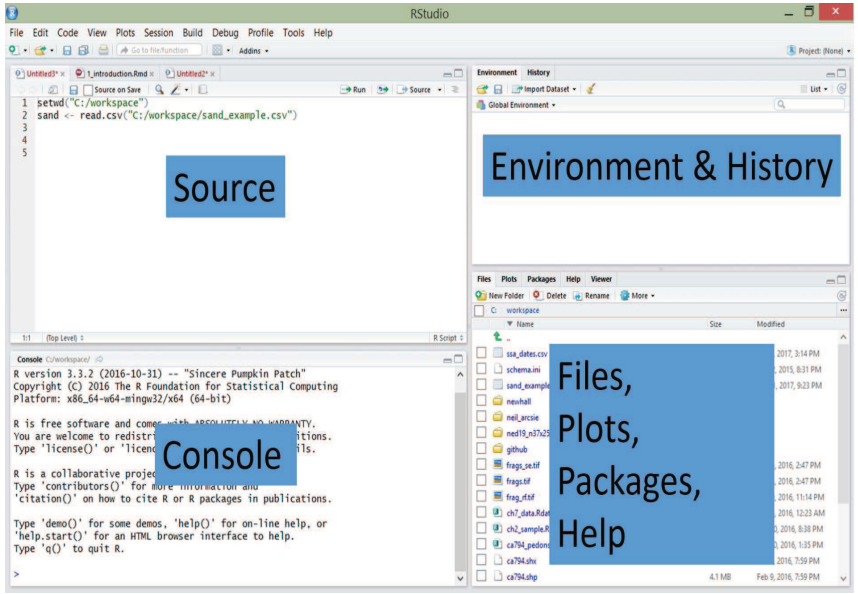
Nowadays most people use `RStudio` to use R, which is an application providing more user-friendly interface with R.

To install `RStudio`, click on `download` at `https://www.rstudio.com`. Once installed, you can open `RStudio` like any other program on your computer, usually by clicking an icon on your desktop.

When you open RStudio, a window appears with three panes in it. The largest pane is a console window. This is where you run R codes and see results.

The console window is almost exactly what you see if you ran R directly. The real work happens here.

Hidden in the other panes are a text editor, a graphics window, a debugger, a file manager, and much more.



R may be used as a calculators. Of course it can do much more. Try out in the console window:

```
> sqrt(9)/3 -1
```

To quit R or RStudio, type at the prompt '`q()`'.

To define a vector x consisting of integers $1, 2, \cdots, 100$

```
> x <- 1:100
> x
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
 [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
 [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
 [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 [91]  91  92  93  94  95  96  97  98  99 100
> sum(x)
> [1] 5050
```

Or we may also try

```
> y <- (1:100)^2
> y
  [1]     1     4     9    16    25    36    49    64    81   100   121   144
 [13]   169   196   225   256   289   324   361   400   441   484   529   576
 [25]   625   676   729   784   841   900   961  1024  1089  1156  1225  1296
 [37]  1369  1444  1521  1600  1681  1764  1849  1936  2025  2116  2209  2304
 [49]  2401  2500  2601  2704  2809  2916  3025  3136  3249  3364  3481  3600
 [61]  3721  3844  3969  4096  4225  4356  4489  4624  4761  4900  5041  5184
 [73]  5329  5476  5625  5776  5929  6084  6241  6400  6561  6724  6889  7056
 [85]  7225  7396  7569  7744  7921  8100  8281  8464  8649  8836  9025  9216
 [97]  9409  9604  9801 10000
> y[14]     # print out the 14-th element of vector y
```

```
[1] 196
```

One may also try `x+y, (x+y)/(x+y), help(log), log(x)` etc.

Additional packages can be installed directly from the R prompt. Information on the available packages is available at

http://cran.r-project.org/web/views/
http://cran.r-project.org/web/packages/

For example, one may install ggplot2 – a package for elegant graphics for data analysis.

```
> install.packages("ggplot2")
> library("ggplot2")  # To load all the objects in the package
                      #  into the current session
```

Note. In the bottom right window of RStudio, you may click on `Packages` → `Install` ··· to install the package to the same effect.

1.2 Help and documentation

To start help manual, click onn `help` alsp in the bottom right window. Then click on `Packages` to access the manuals for installed packages.

Alternatively online manual: `http://cran.r-project.org/manuals.html`

To access the info on an added-on package: `help(package="ggplot2")`

Quick access to the manual for 'mean': `help(mean)`, or `?mean`

Also try `?plot`, `?qplot`, `?sd`, `?summary`

R Newsletter: http://cran.r-project.org/doc/Rnews/

R FAQ: http://cran.r-project.org/faqs.html

Last but not least, `google` whatever questions often leads to most helpful answers

1.3 Data Import/Export

Working directory: a directory/folder from/to which R imports and outports files. You may change your working directory by clicking on

```
Session -> Setting Working Directory
```

with RStudio. For a direct R session, click on `File -> Change dir...`. For example, I create on my laptop `D:\bigData` as my working directory for this course.

The easiest form of data to import into R is a simple text file. The primary function to import from a text file is `scan`. You may check out what 'scan' can do: `> ?scan`

Create a plain text file 'simpleData', in your working directory, as follow:

```
This is a simple data file, created for illustration
of importing data in text files into R
1 2 3 4
5 6 7 8
9 10 11 12
```

It has two lines of explanation and 3 lines numbers. The R session below imports it into R as a vector `x` and 3×4 matrix `y`, perform some simple operations. Note the flag `skip=2` instructs R to ignore the first two lines in the file.

Note. R ignores anything after '#' in a command line.

```
> x <- scan("simpleData.txt", skip=2)
> x # print out vector x
 [1]  1  2  3  4  5  6  7  8  9 10 11 12
> length(x)
[1] 12
```

```
> mean(x); range(x) # write 2 commands in one line to save space
[1] 6.5
[1]  1 12
> summary(x)    # a very useful command!
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.75    6.50    6.50    9.25   12.00

> y <- matrix(scan("simpleData.txt", skip=2), byrow=T,
    ncol=4)
> y   # print out matrix y
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
> dim(y)    # size of matrix y
[1] 3 4
> y[1,]    # 1st row of y
[1] 1 2 3 4
```

```
> y[,2]    # 2nd column of y
[1]  2  6 10
> y[2,4]   # the (2,4)-th element of matrix y
[1] 8
```

A business school sent a questionnaire to its graduates in past 5 years and received 253 returns. The data are stored in a plain text file 'Jobs' which has 6 columns:

C1: ID number
C2: Job type, 1 - accounting, 2 - finance, 3 - management, 4 - marketing and sales, 5 -others
C3: Sex, 1 - male, 2 - female
C4: Job satisfaction, 1 - very satisfied, 2 - satisfied, 3 - not satisfied
C5: Salary (in thousand pounds)
C6: No. of jobs after graduation

```
IDNo.  JobType  Sex    Satisfaction  Salary     Search
1        1       1       3             51      1
2        4       1       3             38      2
3        5       1       3             51      4
4        1       2       2             52      5
... ...
```

We import data into R using command `read.table`

```
> jobs <- read.table("Jobs.txt"); jobs
          V1      V2  V3           V4     V5     V6
1        IDNo. JobType Sex Satisfaction Salary Search
2          1       1   1            3     51      1
3          2       4   1            3     38      2
4          3       5   1            3     51      4
   ... ...
> View(jobs)    # display data properly in the top left window
> dim(jobs)
[1] 254   6
> jobs[1,]
     V1      V2  V3           V4     V5     V6
1 IDNo. JobType Sex Satisfaction Salary Search
```

We repeat the above again by taking the 1st row as the names of variables (`header=T`) and the entries in 1st column as the names of the rows (`row.names =1`).

```
> jobs <- read.table("Jobs.txt", header=T, row.names=1)
> dim(jobs)
[1] 253   5
> names(jobs)
[1] "JobType"  "Sex"   "Satisfaction"   "Salary"  "Search"
> class(jobs)
[1] "data.frame"
> class(jobs[,1]); class(jobs[,2]); class(jobs[,3]);
   class(jobs[,4]); class(jobs[,5])
[1] "integer"
[1] "integer"
[1] "integer"
[1] "integer"
[1] "integer"
```

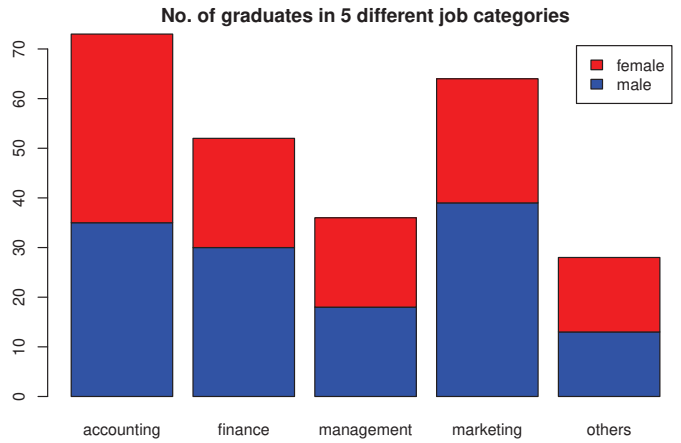
Since the first three variables are nominal, we may specify them as 'factor', while "Salary" can be specified as 'numeric':

```
>  jobs <- read.table("Jobs.txt", header=T, row.names=1,
    colClasses = c("integer", "factor", "factor", "factor",
    "numeric", "integer"))
> class(jobs[,1]); class(jobs[,2]); class(jobs[,3]);
      class(jobs[,4]); class(jobs[,5])
[1] "factor"
[1] "factor"
[1] "factor"
[1] "numeric"
[1] "integer"
```

Note. we need to specify the class for the row name variable (i.e. 1st column) as well.

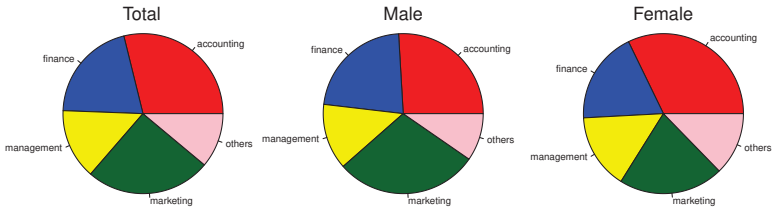
Now we do some simple descriptive statistical analysis for this data.

```
> table(jobs[,1])
 1  2  3  4  5
73 52 36 64 28   # No. of graduates with 5 different JobTypes
> t <-table(jobs[,2], jobs[,1], deparse.level=2)  # store table in t
> t
         jobs[, 1]
jobs[, 2]  1  2  3  4  5
        1 35 30 18 39 13    # No. of males with 5 different JobTypes
        2 38 22 18 25 15    # No. of females with 5 different JobTypes
> 100*t[1,]/sum(t[1,])
       1        2        3        4        5
25.92593   22.22222   13.33333   28.88889   9.62963
        # Percentages of males with 5 different JobTypes
> 100*t[2,]/sum(t[2,])
       1        2        3        4        5
32.20339   18.64407   15.25424   21.18644   12.71186
        # Percentages of females with 5 different JobTypes
> barplot(t, main="No. of graduates in 5 different job categories",
         legend.text=c("male", "female"), names.arg=c("accounting",
    "finance", "management", "marketing", "others"))
    # draw a bar-plot
```



The barplot shows the difference in job distribution due to gender. We may also draw pie-plots, which are regarded as less effective.

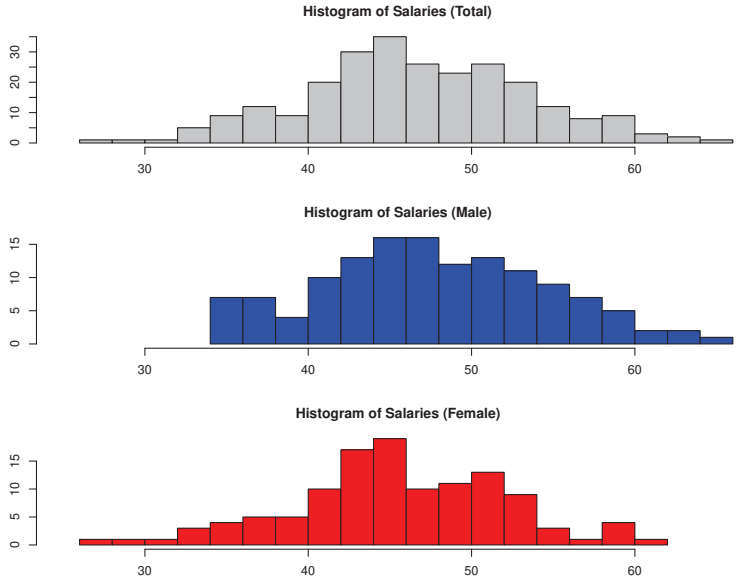
```
> pie(t[1,]+t[2,],label=c("accounting","finance","management",
      "marketing","others")); text(0,1, "Total", cex=2)
> pie(t[1,],label=c("accounting","finance","management",
       "marketing","others")); text(0,1, "Male", cex=2)
> pie(t[2,],label=c("accounting","finance","management",
       "marketing","others")); text(0,1, "Female", cex=2)
```



Now let look at the salary (`jobs[,4]`) distribution, and the impact due to gender.

```
> mSalary <- jobs[,4][jobs[,2]==1]
                # extract the salary data from male
> fSalary <- jobs[,4][jobs[,2]==2]
                # extract the salary data from female
> summary(jobs[,4]); summary(mSalary); summary(fSalary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.00   43.00   47.00   47.13   52.00   65.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.00   44.00   48.00   48.11   53.00   65.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.00   42.25   46.00   46.00   51.00   61.00
> hist(jobs[,4], col="gray", nclass=15, xlim=c(25,66),
                main="Histogram of Salaries (Total)")
      # plot the histogram of salary data
> hist(mSalary, col="blue", nclass=15, xlim=c(25,66),
                main="Histogram of Salaries (Male)")
> hist(fSalary, col="red", nclass=15, xlim=c(25,66),
                main="Histogram of Salaries (Female)")
```

You may also try stem-and-leaf plot: `stem(jobs[,4])`



To export data from R, use `write.table` or `write`.

To write `jobs` into a plain text file 'Jobs1.txt':

```
> write.table(jobs, "Jobs1.txt")
```

which retains both the row and column names. Note the different entries in the file are separated by spaces.

We may also use

```
> write.table(jobs, "Jobs2.txt", row.names=F, col.names=F),
> write.table(jobs, "Jobs3.txt", sep=",")
```

Compare the three output files.

Note that the values of factor variables are recorded with " ". To record all the levels of factor variables as numerical values, we need to define a pure numerical data.frame first:

```
> t <- data.frame(as.numeric(jobs[,1]), as.numeric(jobs[,2]),
        as.numeric(jobs[,3]), jobs[,4], jobs[,5])
> write.table(t, "Jobs4.txt")
```

The file "Jobs4.txt" contains purely numerical values.

Note. (i) (i) Working directory — a directory/folder from/to which R imports and outports files. You may change your working directory by clicking on

```
Session -> Setting Working Directory
```

(ii) Saving a session — when you quit an R session `q()`, you will be offered an option to 'save workspace image'. By clicking on "yes", you will save all the objects (including data sets, loaded functions from added-on packages etc) in your R session. You may continue to work on this session by directly double-clicking on the image file (with the last name `RData`) in your working directory.

Alternatively you may save work done in an R session including all objects, use "save.image" in console window:

```
 > save.image("filename.RData")
```

the file must have "RData" as its last name.

To save all the commands used in an R session only, type in console

```
> savehistory("filename.Rhistory")
```

the file must have "Rhistory" as its last name.

A useful tip: Create a separate working directory for each of your R projects.

1.4 Organising an Analysis

An R analysis typically consists of executing several commands. Instead of typing each of those commands on the R prompt, we may collect them into a plain text file – an R-script. It can be resulted from editting an `Rhistory`-file saved from `savehistory`. For example, the file "jobsAnalysis.r" in my working directory reads like:

```
jobs <- read.table("Jobs.txt", header=T, row.names=1)
        # File "Jobs.txt" is in the working directory now
mSalary <- jobs[,4][jobs[,2]==1]
fSalary <- jobs[,4][jobs[,2]==2]
summary(jobs[,4])
summary(mSalary)
summary(fSalary)
par(mfrow=c(3,1))   # display 3 figures in one column
hist(jobs[,4], col="gray", nclass=15, xlim=c(25,66),
```

```
         main="Histogram of Salaries (Total)")
hist(mSalary, col="blue", nclass=15, xlim=c(25,66),
        main="Histogram of Salaries (Male)")
hist(fSalary, col="red", nclass=15, xlim=c(25,66),
        main="Histogram of Salaries (Female)")
```

You may carry out the project by sourcing the file into an R session:

```
> source("jobAnalysis.r", echo=T)
```

Also try `source("jobAnalysis.r")`.

Editing, including cleaning up, and saving the commands from the current session can be easily done by clicking on `History` in the top right window (i.e. Environment window of RStudio). Then highlight those you would like to save, and click on the buttom of `To Source`. All the highlighted contents will be displayed in the top left window. You can then further edit and save them into a file (i.e. an R-script) which can used again later.

To execute the commands in "jobAnalysis.r", open the file by clicking on `File -> Open File`. The contents of the file will be displayed in the top left window. Highlight the commands you like to execute, click on `Run`-buttom.