Data Science (CDA)

UNIT 2: Data integration and manipulation

- José Hernández Orallo, DSIC, UPV, jorallo@upv.es

Content

- Unit 2: Data integration and manipulation
  - Source types and data repositories.
  - Data gathering, integration and cleansing
  - Data property, privacy and security.
  - Data visualisation and comprehension

Source types and data repositories

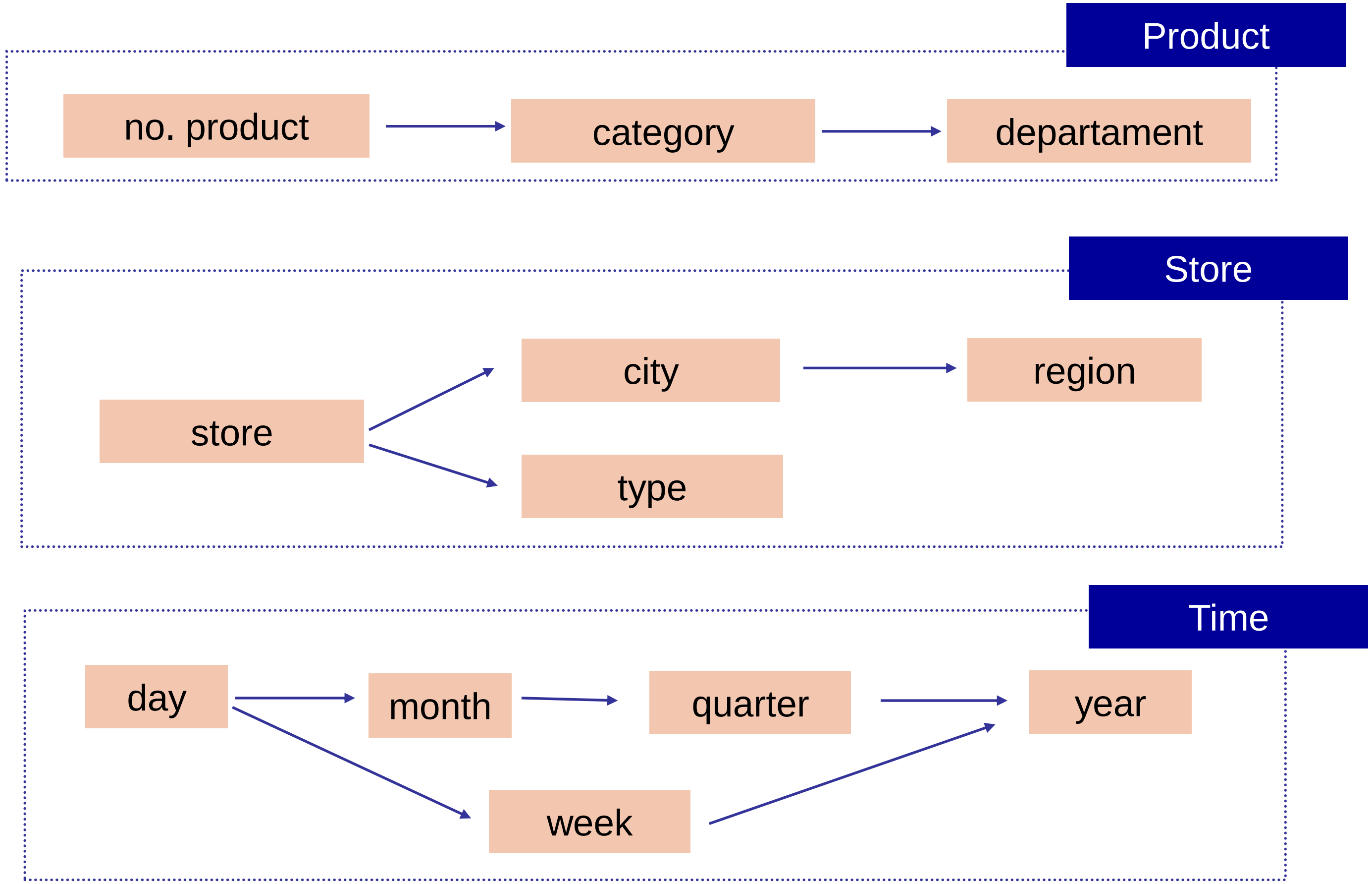
- Are the data used for transactional processing or analytical processing?

Transactional Processing	Analytical Processing
Current Data	Historical data
Detailed Data	Detailed and aggregated data
No redundancy (normalised)	Redundancy and precalculations
Medium-size databases*	Large databases
Data is updated	Data is not updated*
Repetitive processes	Unpredictable processes
Response time (msec – sec)	Response time (sec – hours*)
Systematic decisions	Strategic decisions
Many users	Few users*

  - * The distinction is not always so clear-cut.

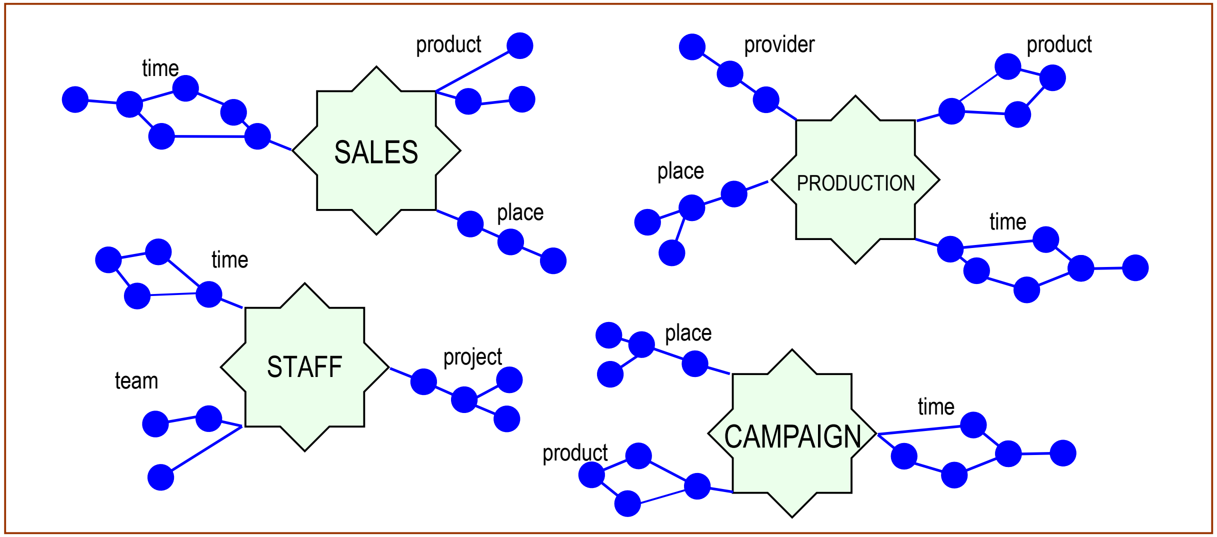
Source types and data repositories

- Traditional approach: multidimensional datawarehouse
  - Attributes are arranged into dimensions



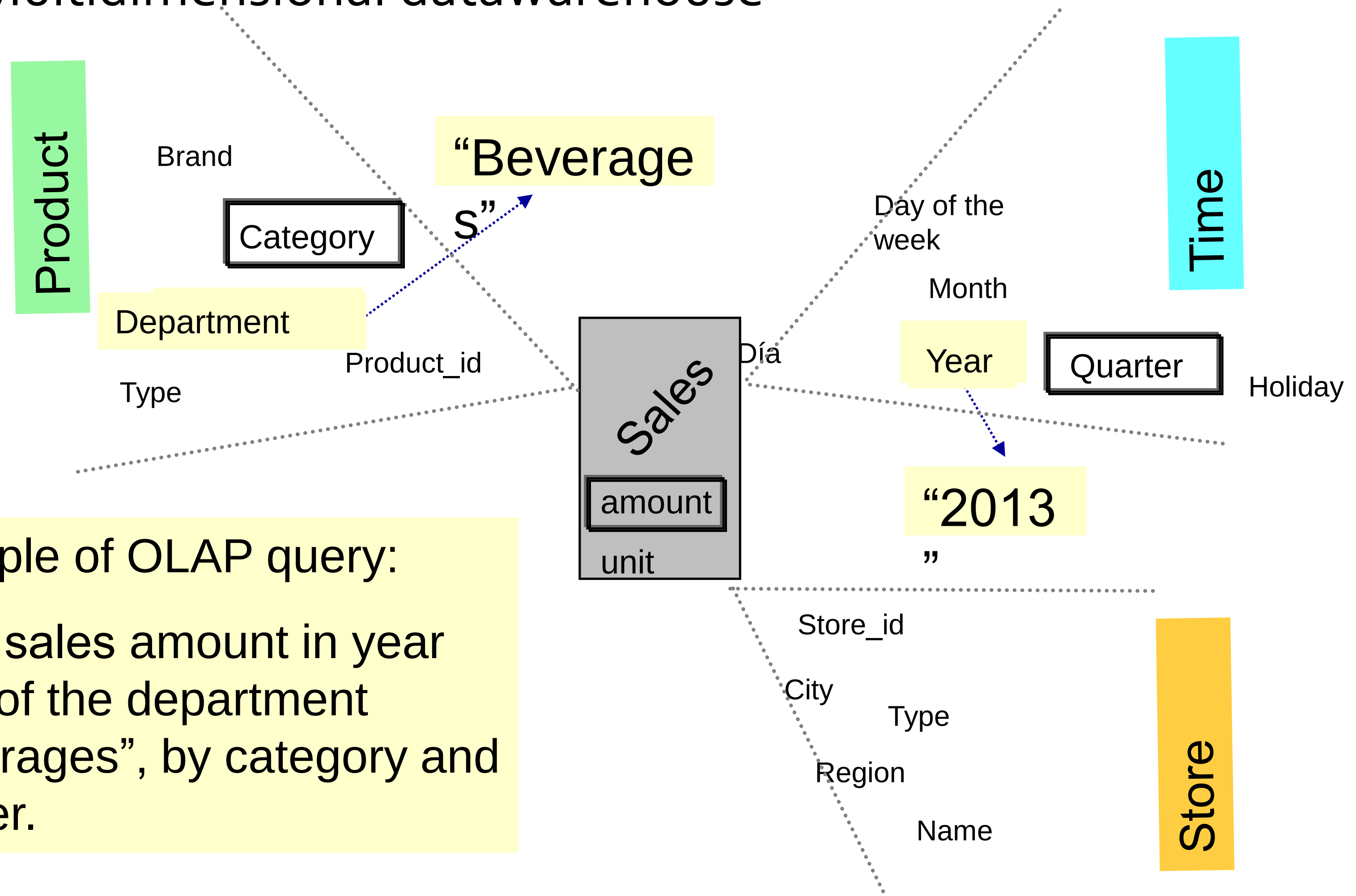
Source types and data repositories

- Multidimensional datawarehouse
  - Each schema is known as a datamart:



Source types and data repositories

- Multidimensional datawarehouse



Example of OLAP query:

"Total sales amount in year 2013 of the department "Beverages", by category and quarter.

Source types and data repositories

- OLAP operators

quarter | category | amount

year

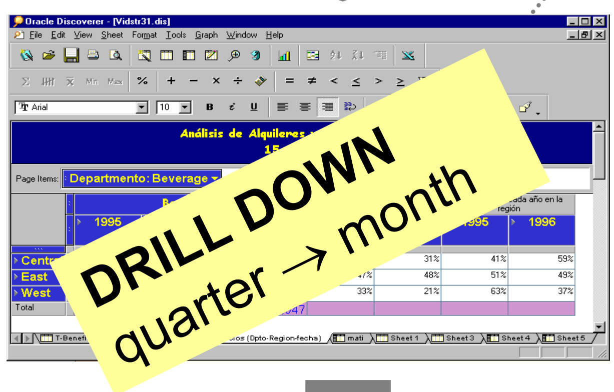
quarter

month

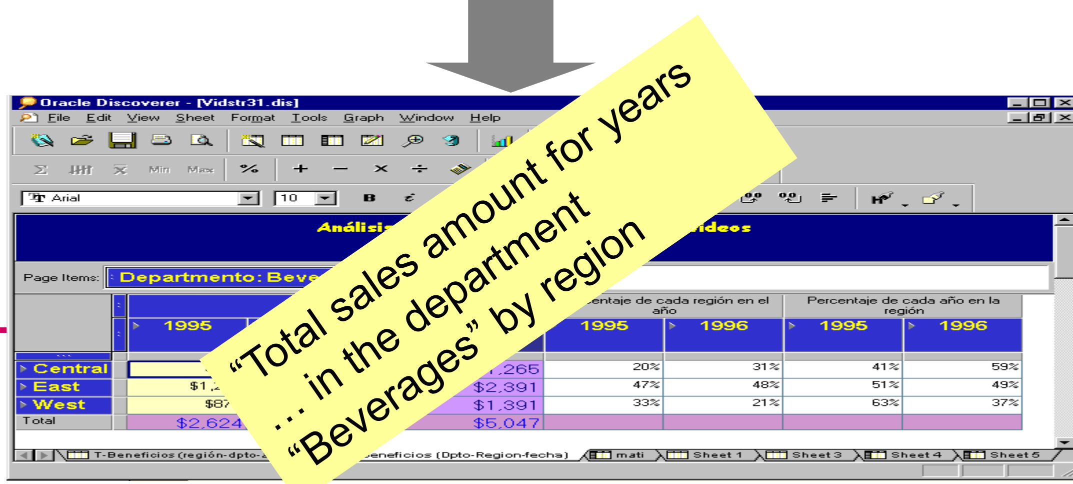
day

ROLL-UP

DRILL-DOWN



The DRILL operation is performed over the original report!



Source types and data repositories

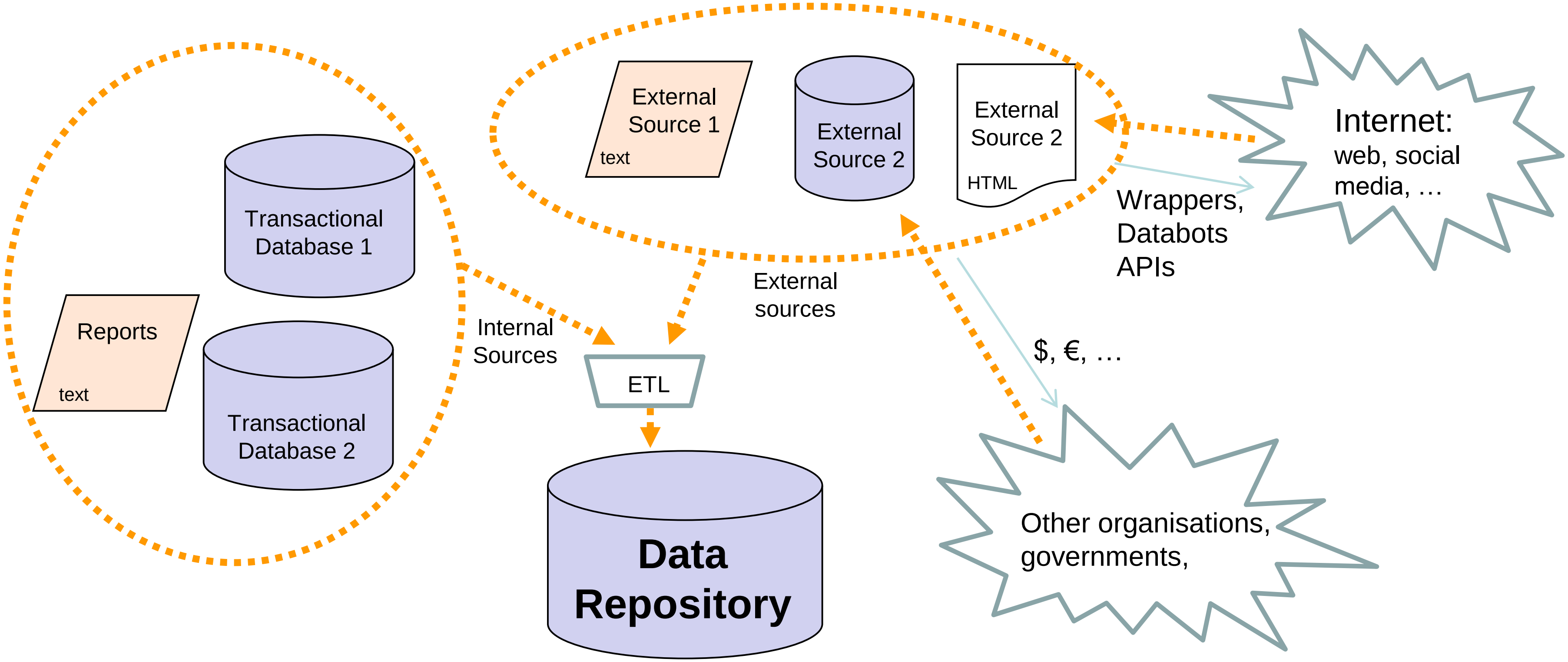
- OLAP tools do not infer patterns

OLAP queries	Data mining
What is the accident rate among smokers and non-smokers?	Which are the best predictors for accidents?
What is the average telephone bill of my current customers vs. the ex-customers who quit the company?	Will X leave the company? Which factors affect churn?
What is the average daily purchase amount between stolen credit cards operations and legitimate ones?	What purchase patterns are associated with credit card frauds?

- Do I need a DW for Analysis?
  - In DM (modelling):
    - Granularity is usually higher than in DW
    - Information to be recorded into the database must be carefully planned beforehand.
      - If we now realise we need the age of the customer and we haven't recorded it, the problem has a difficult solution now.
    - External sources are very important.
    - The effort may not compensate for the benefits of a single data science project.

Source types and data repositories

- But I need a repository (a "silo")
  - Need to integrate internal and external data

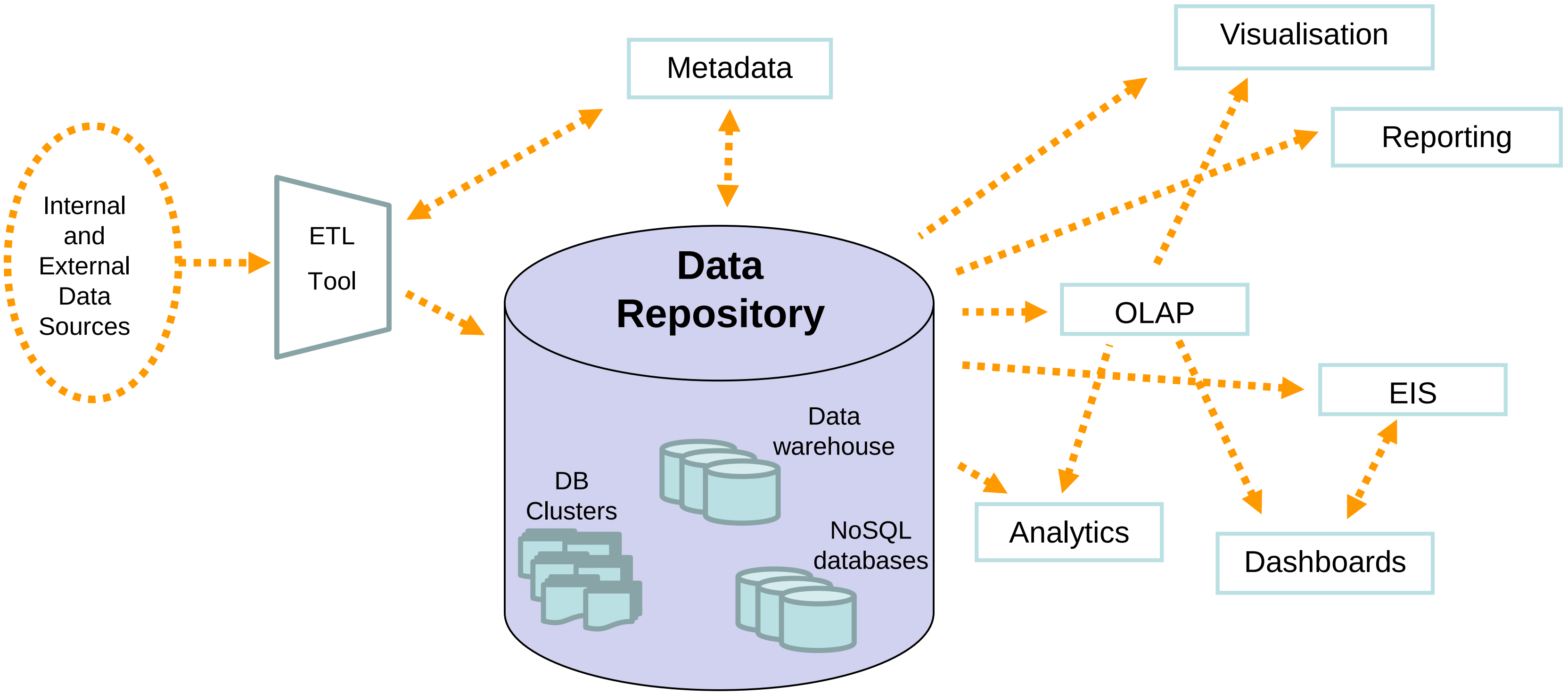


Source types and data repositories

- Why do we need external sources?
  - Demographic data, sociological studies, general economical data, ...
  - Business data, competitors, ..
  - Calendars, weather, traffic, TV/sport schedule, catastrophes, ...
  - External information is frequently sold and bought.
- Many data science projects only (or mostly) work with external data.
  - Remember the in/out cases.

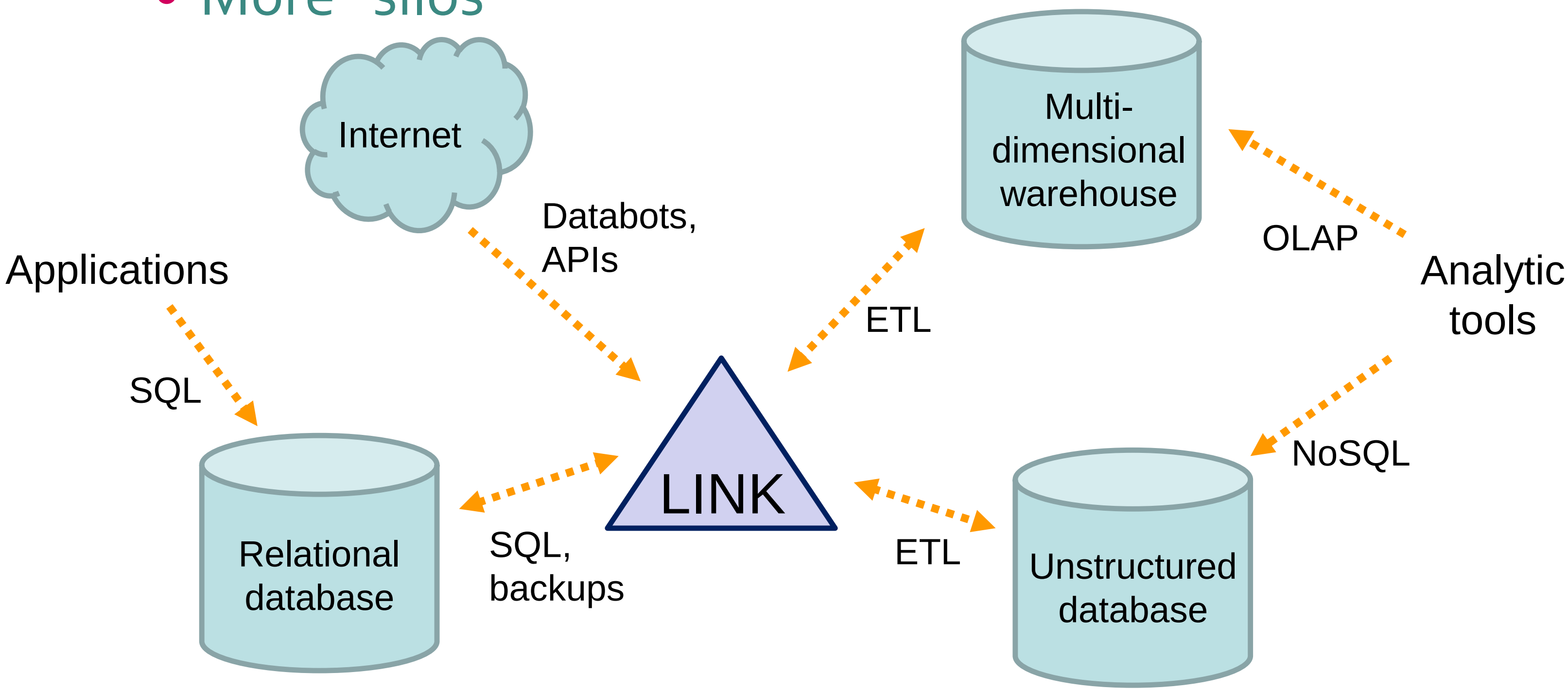
Source types and data repositories

- Several repositories:
  - When many tools are integrated.
    - Things become more complicated.



Source types and data repositories

- The repositories ("silos") need to be connected:
  - New technologies typically complement old ones
    - But do not substitute them.
    - More "silos"



  - Need for links between the "silos".

Source types and data repositories

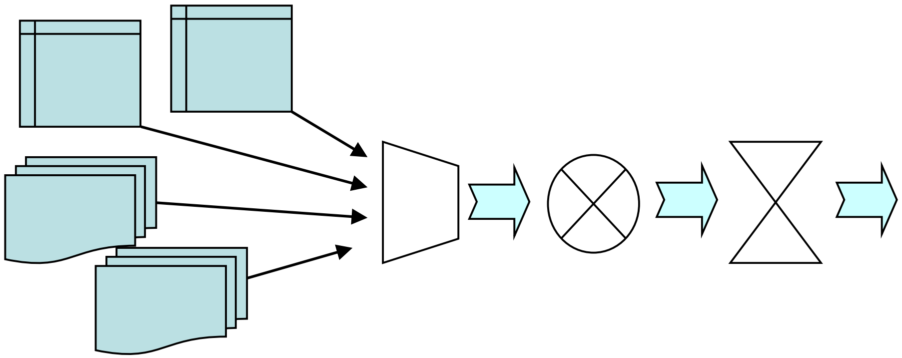
- Do I need Hadoop / Spark / MongoDB / TensorFlow?
  - For a small or medium-sized organisation…
    - This is still the most common configuration:
      - A transactional relational database with SQL.
      - A multidimensional data warehouse with OLAP tools.
      - ETL tools.
      - Dashboards.
      - A data mining tool.

> On many occasions, classical business intelligence tools are enough.

- Putting all together in a repository is not enough
  - Data integration is a complex process
- Integration is not sufficient
  - Data preparation…
- All this includes:
  - Data comprehension
  - Data cleansing
  - Data transformation
  - Data selection

This stage usually takes half of the time/effort from the overall D2K process.

- The result of the preparation process:



MINABLE VIEW

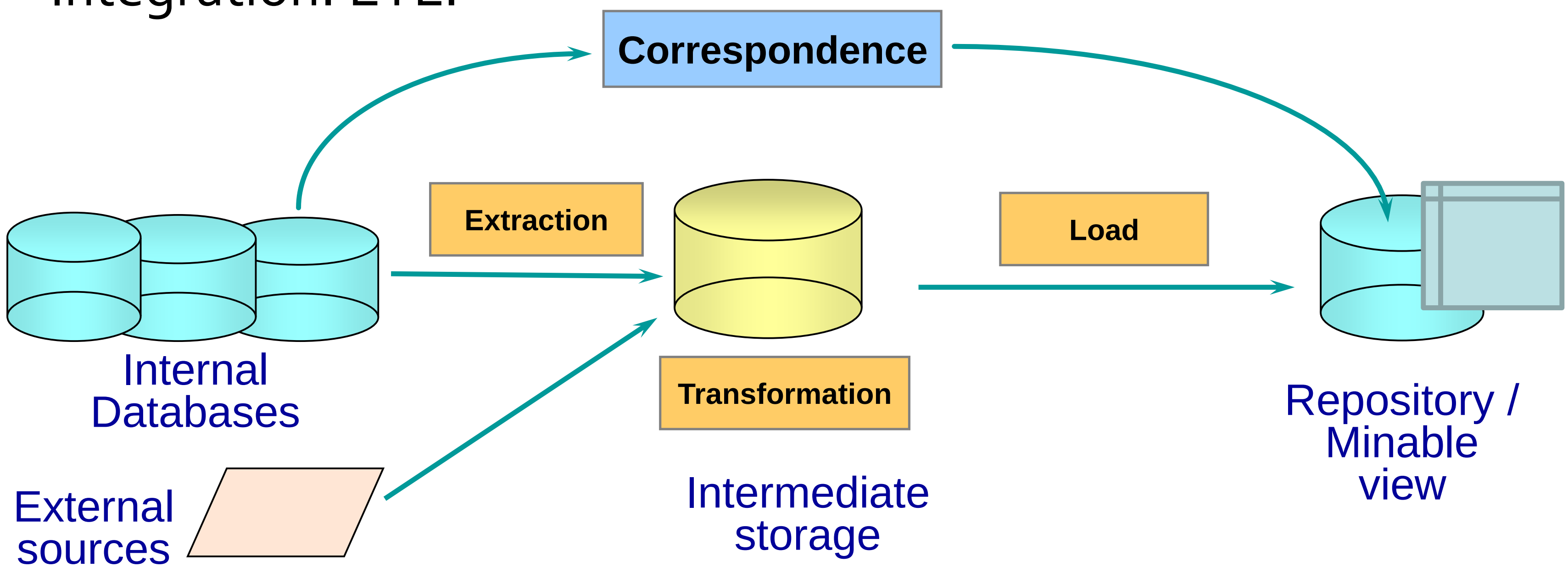
Idc	D-credit (years)	C-credit (euros)	Salary (euros)	Own house	Default account	…	Good customer
101	15	60000	2200	Yes	2	…	no
102	2	30000	3500	Yes	0	…	yes
103	9	9000	1700	Yes	1	…	no
104	15	18000	1900	No	0	…	yes
105	10	24000	2100	No	0	…	no
...	…	…	…	…	…	…	…

Minable view: set of data which includes all and only the interest variables for the given problem in the adequate format

- Sometimes just known as "dataset" + "task description"
- Task description (unit 3)
  - Predictive: classification or regression
  - Descriptive: clustering, association, correlations.

- Integration: ETL:
  - We can do the integration with the use of languages and scripts (e.g., R, Python, etc.) or specialised tools.
  - The system is known as ETL (Extraction - Transformation - Load)
    - Functions of the ETL:
      - Initial load
      - Maintenance through timely refreshment.
    - It usually makes the integration through intermediate steps or repositories.

- Integration: ETL:

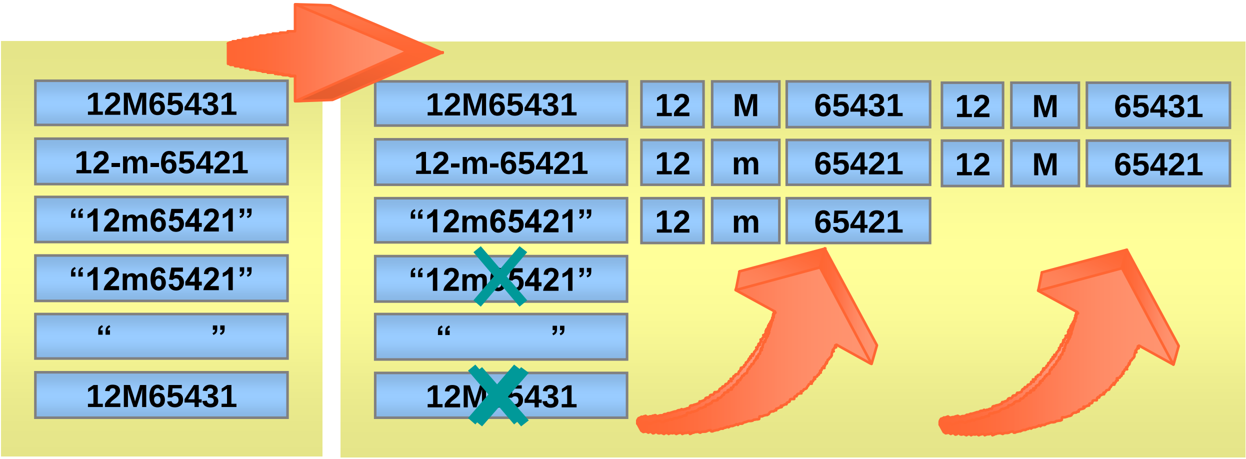


- The intermediate storage allows for:
  - Performing transformations without stopping the original databases or the databots/scripts downloading the data
  - Storing metadata
  - Easing the integration of internal and external data.

- Extraction
  - Use of scripts and programs designed to extract the data from the sources.
    - SQL and DB scripts to load from databases
    - APIs to download from web services and applications
    - Wrappers when no API is available.
    - Databots when we need to perform search on the Internet
    - Special scripts for non-structured data.

- Extraction
  - What if data can change.
    - We need to identify these changes.
    - Methods:
      - Reload everything
      - Compare old and new data.
      - Use of time stamping.
      - Use of triggers or other event-driven code.
      - Use of a log.
      - Several of the above.

- Transformation
  - Example: product code



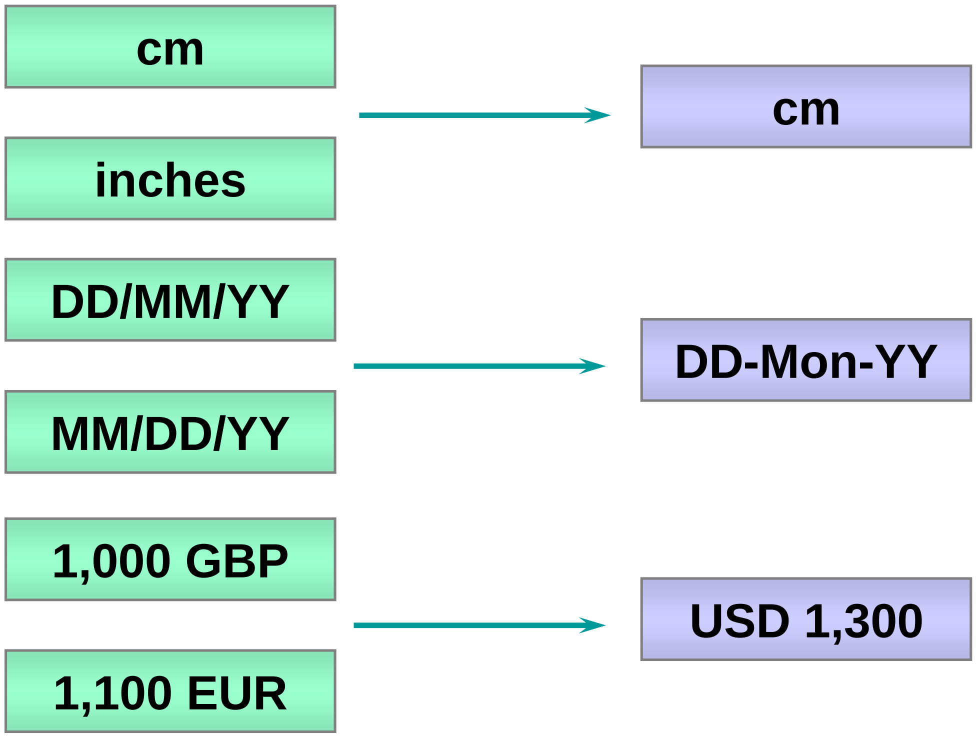
- Transformation
  - Example: inconsistent codings

f , m

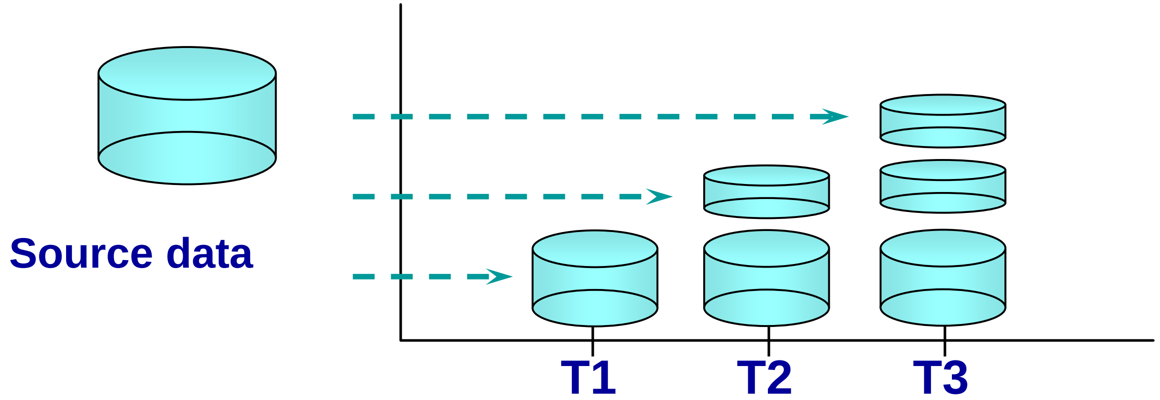
1 , 0 → f, m

male, female

- Transformation
  - Example: different units



- Load
  - Periodically.
    - If the source is the transactional database we need to find the “load windows” (e.g., at night) so that the database doesn’t suffer overload.
    - We load that incrementally.
    - We can delete very old data to have a reasonably recent window of historical data.



- ETL software:
  - Use of languages: R, python, ...
  - Use of tools

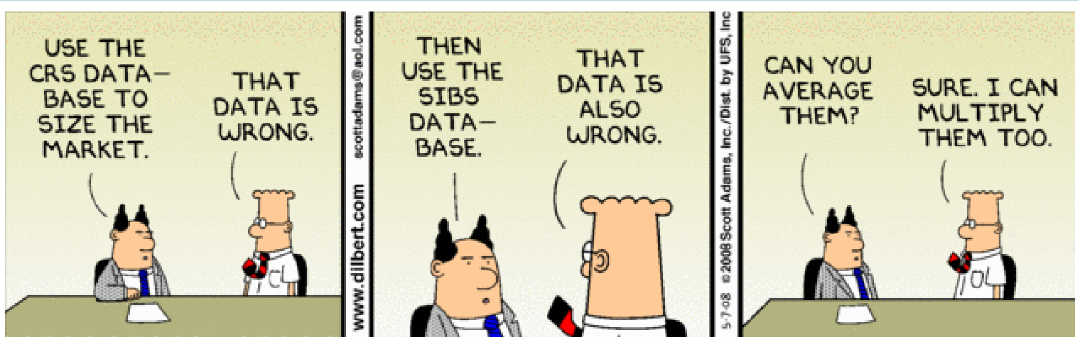
	IBM DataStage	Informatica PowerCenter	Microsoft SSIS	Oracle Data Integrator (ODI) 12c	SAS ETL Studio	Talend Open Studio (Enterprise Edition)	Talend Open Studio (Free Edition)
Version	9	9.1.0	SqlServer Integration Services 2012	12c	9.1.3	5.3.1	5.3.1
TYPE OF LICENSE OFFERINGS							
Commercial	✓	✓	✓	✓	✓	✗	✗
Open Source	✗	✗	✗	✗	✗	✗	✓
Open-Source Commercial	✗	✗	✗	✗	✗	✓	✗
OTHER CAPABILITIES							
Support for Data Integration in the Cloud	✓	✓	✓	✓	✓	✓	✓
Support for Hadoop	✓	✓	✓	✓	✓	✓	✓
Direct Access to Data Quality Module	YES, Separately Purchasable Module	YES, Separately Purchasable Module	YES, Separately Purchasable Module	YES, Separately Purchasable Module	YES, Separately Purchasable Module	YES, Separately Purchasable Module	NO
Centralized Metadata Repository	✓	✓	✓	✓	✓	✓	✗
TYPE OF ACCESS TO VERSION CONTROL							
Concurrent Versions System (CVS)	Direct Access	Configuration Required for Access	Configuration required for Access	Configuration Required for Access	Direct Access	NO	NO

CloverETL Pentaho Kettle

* From http://www.quickdatainsights.com/data-integration-tools-comparison-2/etl-tools-comparison-matrix/

- Once the data is loaded…
  - Are we ready? No
    - Is data quality good?
      - Cleansing
    - Is the data as we want for the minable view?
      - More transformation
    - Is the data that we need?
      - Selection

Integration from different sources may conceal data quality issues:



- Data Cleansing
  - Possible actions against outliers or missing values:
    - ignore.
    - filter (eliminate or replace) the column.
    - filter the row.
    - replace the value by an average or predicted value.
    - segment the rows between correct data and the rest, and work separately.
    - discretise numerical attributes.
    - give up and modify the data quality policy for the next time.

- Transformation/selection/reduction:
  - Global transformation: e.g. exchange rows and columns.
  - Attribute creation or modification:
    - Discretisation and numerisation.
    - Normalisation.
    - Derived attributes.
  - Attribute reduction.
  - Selections:
    - Vertical (over features / attributes):
      - Feature selection.
    - Horizontal (over instances):
      - Sampling.

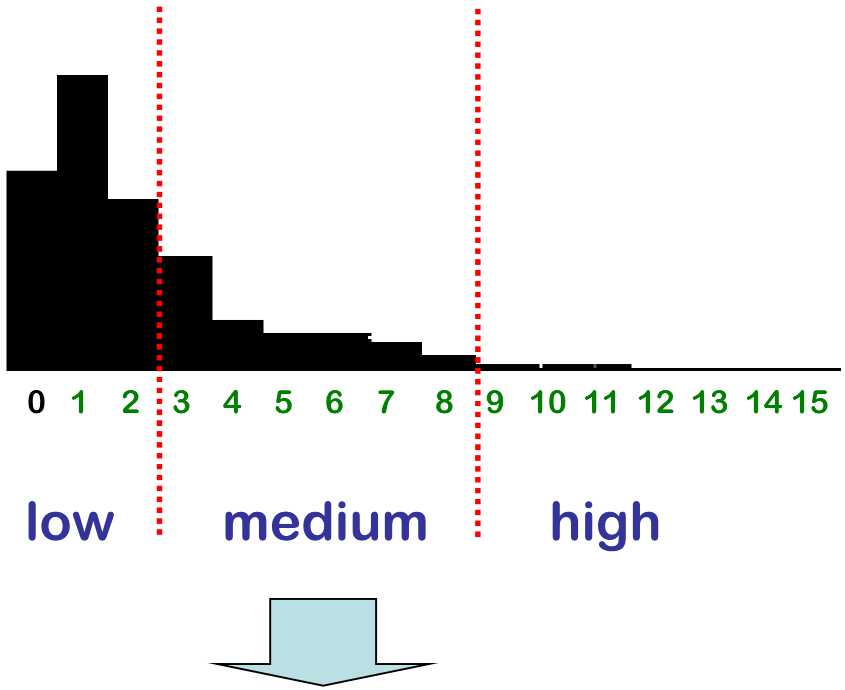
- Transformation/selection/reduction:
  - Attribute creation
    - A good knowledge of the domain is the most important issue to create good derived attributes:

Examples:

- **height^2/weight (obesity index)**
- **debt/earnings**
- **passengers * miles**
- **credit limit - balance**
- **population / area**
- **minutes of use / number of telephone calls**
- **activation_date - application_date**
- **number of web pages visited / total amount purchased**

- Transformation/selection/reduction:
  - Discretisation

Example: attribute "weektickets" (numerical, 1 … 15).



New attribute "weekticketsNOM" (nominal: low, medium, high).

- Transformation/selection/reduction:
  - Numerisation

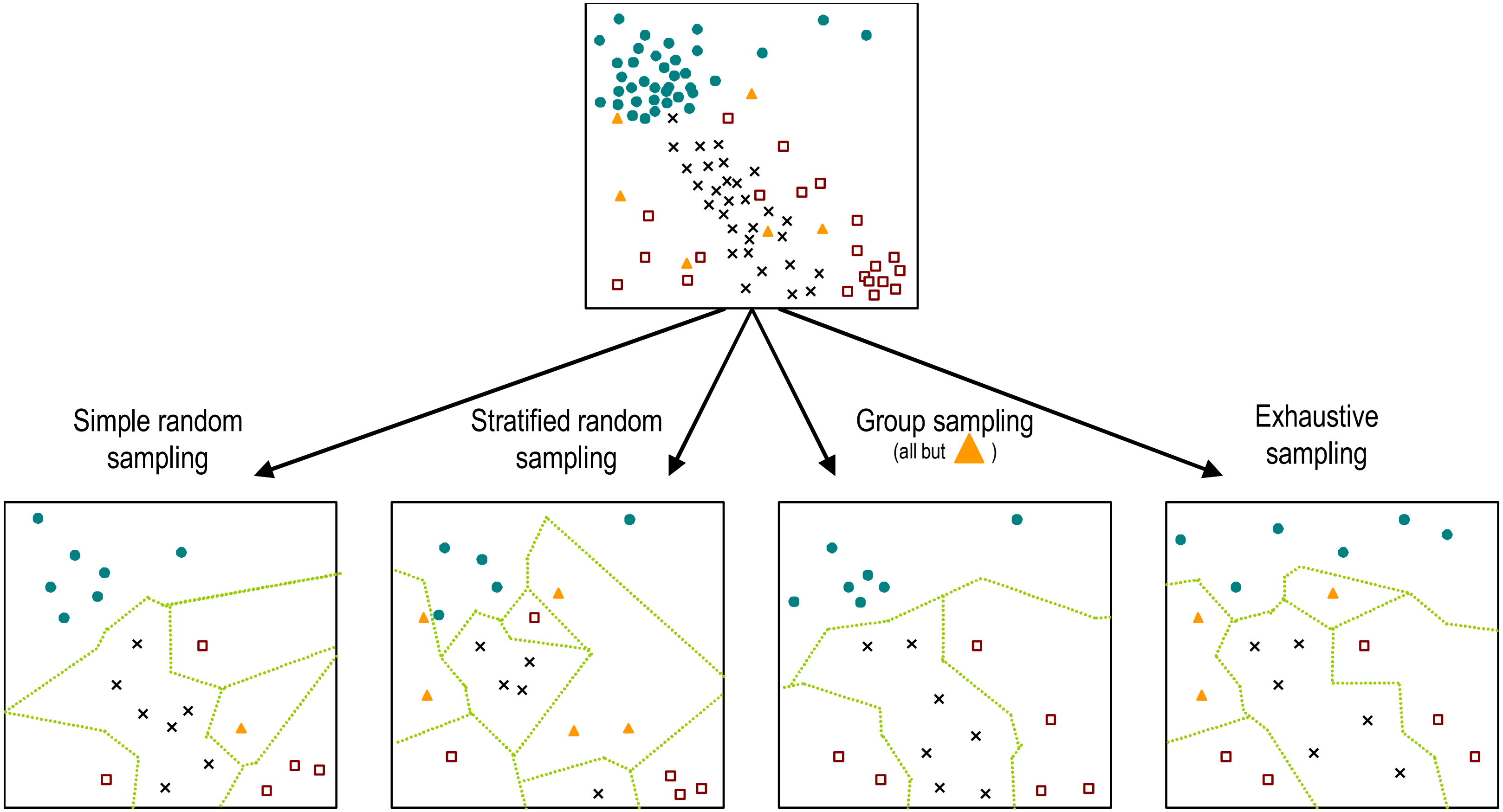
- **Numerisation “1 to n” (or n-1) (a.k.a. “one-hot encoding”):**
  - **EXAMPLE: Convert the field “card” with values: { “VISA”, “4B”, “Amer”, “Maestro” } into four binary fields.**

- **Numerisation “1 to 1”:**
  - **EXAMPLE: if we have four categories such as {child, young, adult, senior} we can create one attribute with values from 1 to 4.**

- Transformation/selection/reduction:
  - Attribute reduction by transformation
    - Well-known techniques such as:
      - principal component analysis (PCA).
        - » PCA transforms the *m* original attributes into a new set of attributes *p* where $p \leq m$.
        - » It is a geometrical projection.
        - » New attributes are independent from each other, and they are ordered by information relevance.

- Transformation/selection/reduction:
  - Feature selection
    - Choice of *p* variables from *m* variables:
      - Filter methods: selection is made independently from data mining models.
      - Wrapper methods: selection is made using a data mining model.

- Transformation/selection/reduction:
  - Sampling: Reduce the number of rows/instances



- The end of the preparation process
  - Remember the aim was a minable view
    - Only ready when the data has been prepared, cleansed and selected.
    - And we have a clear task (unit 3).

(when data is complex (text, multimedia, ...), we no longer talk about a view but the "minable dataset").

  - Let's see some examples of minable views...

- Example: Bank agent

Must I grant a mortgage to this customer?

Historical Data:

cId	Credit-p (years)	Credit-a (euros)	Salary (euros)	Own House	Defaulter accounts	…	Returns-credit
101	15	60.000	2.200	yes	2	…	no
102	2	30.000	3.500	yes	0	…	yes
103	9	9.000	1.700	yes	1	…	no
104	15	18.000	1.900	no	0	…	yes
105	10	24.000	2.100	no	0	…	no
...	…	…	…	…	…	…	…

Data Mining

Pattern / Model:

If Defaulter-accounts > 0 **then** Returns-credit = no
If Defaulter-accounts = 0 **and** [(Salary > 2.500) **or** (credit-p > 10)] **then** Returns-credit = yes

- Example: Supermarket manager

When customers buy eggs, do they also buy oil?

Historical Data:

BasketId	Eggs	Oil	Nappies	Wine	Milk	Butter	Salmon	Endive	...
1	yes	yes	no	yes	no	yes	yes	yes	...
2	no	yes	no	no	yes	no	no	yes	...
3	no	no	yes	no	yes	no	no	no	...
4	no	yes	yes	no	yes	no	no	no	...
5	yes	yes	no	no	no	yes	no	yes	...
6	yes	no	no	yes	yes	yes	yes	no	...
7	no	no	no	no	no	no	no	no	...
8	yes	yes	yes	yes	yes	yes	yes	no	...
...	...	...	...	...	...	...	...	...	...

Data Mining

Pattern / Model:

Eggs → Oil : Confidence = 75%, Support = 37%

- Example: Personal Manager

What kind of employees do I have?

Historical Data:

Id	Salary	Married	Car	Children	Rent/ Owner	Union	Off sick/year	Work years	Gender
1	10000	yes	no	0	Rent	no	7	15	M
2	20000	no	yes	1	Rent	yes	3	3	F
3	15000	yes	yes	2	Owner	yes	5	10	M
4	30000	yes	yes	1	Rent	no	15	7	F
5	10000	yes	yes	0	Owner	yes	1	6	M
6	40000	no	yes	0	Rent	yes	3	16	F
7	25000	no	no	0	Rent	yes	0	8	M
8	20000	no	yes	0	Owner	yes	2	6	F
15	8000	no	yes	0	Rent	no	3	2	M
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Data Mining

Pattern / Model:

- **Group 1:** Without children and in a rented house. Low participation in unions. Many days off sick.
- **Group 2:** Without children and with car. High participation in unions. Few days off sick. More women and in rented houses.
- **Group 3:** With children, married and with car. More men and usually house owners. Low participation in unions.

- Example: Trader in a retail company

How many TVs do we expect to sell next month?

Historical Data:

PRODUCT	Month–12	...	Month–4	Month–3	Month–2	Month–1	Month
Flat TV 30’	20	...	52	14	139	74	?
BlueRay	11	...	43	32	26	59	?
PlayStation	50	…	61	14	5	28	?
Five star fridge	3	…	21	27	1	49	?
Three star fridge	14	...	27	2	25	12	?
…	…	…	…	…	…	…	...

Data Mining

Pattern / Model:

Linear Model: TV Sales for Next Month:

$$V(Month)_{flatTV} = 0.62 \cdot V(Month\text{-}1)_{flatTV} + 0.33 \cdot V(Month\text{-}2)_{flatTV} + 0.12 \cdot V(Month\text{-}1)_{BlueRay} - 0.05$$

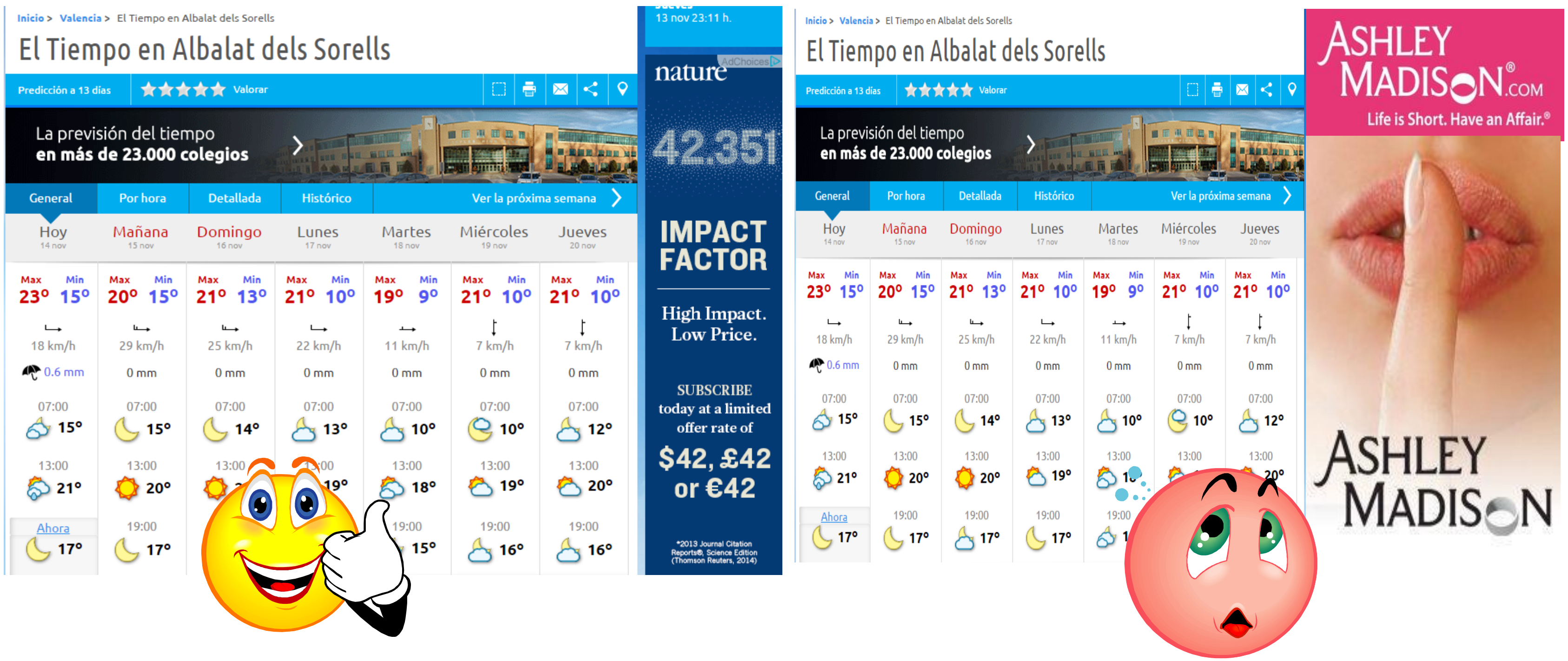
- "Tension between privacy and improving business decisions" (Provost & Fawcett 2013)
  - We need ethical principles
    - Google's motto: "Don't be evil".
  - We need laws for those who lack ethical principles.
    - Not only privacy laws are applicable here.

Ethos and laws change from country to country.

- Ethics:
  - Some data science is about **discrimination**.
    - E.g., is acceptable that *only one group* of customers receive an offer?
      - Only those living in a particular city?
        - » To sell a football T-shirt
      - Only rich customers?
        - » To sell diamonds
      - Only women?
        - » To sell a women's magazine
      - Only men?
        - » To sell anti hair loss lotion
      - Only atheist people? (inferred by their tweets or purchase history)
        - » To sell a book by Dawkins
      - Only white people? (inferred by names and surnames)
        - » To sell a sunscreen lotion

- Data collection and ownership:
  - Have you authorised a company to use your information?
    - For the original company's purpose or for any other purpose?
      - E.g. Telefónica is selling data about people's location.
  - Are you going to sacrifice privacy for a better service?
    - Good recommendations only work if you allow the service to track your purchases.
  - Get a free service by losing some privacy?
    - Cookies, filling a long form, etc.
  - Is data downloaded with scrapers or databots from the Internet legal?
    - You need to check the webpage and the law in the country.
  - Open Data
    - It may have restrictions about what to do with the data.

- Data privacy and personalisation
  - Companies know where you are and what you do.
    - Some recommender systems (google ads) disclose user's preferences (sometimes private) every time (e.g., browsing)
      - Me browsing in front of my students…



- Data privacy and personalisation
  - For instance, retailer *Target* found a teenage girl was pregnant and sent her coupons for baby clothes and cribs.*

Was Target wrong in using analytics to identify pregnant women from changes in their buying behavior ? [364 votes total]	
Yes, Target was wrong - companies should not try to infer sensitive personal data such as pregnancy (61)	17%
No, Target had a right to use consumer shopping information, and was doing it effectively (270)	74%
Not sure (33)	9%

http://www.kdnuggets.com/polls/2012/was-target-wrong-using-analytics-identify-pregnancy.html

  - Would you answer the same if you're told that her father, who didn't know she was pregnant, actually became aware about her pregnancy when the coupons for baby clothes and cribs (cots) arrived??!!
    - Is the problem in the pattern or in the use of the pattern?

* http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0

- Data Privacy and Anonymisation
  - Many data science projects do not work with personalised data.
  - Data is anonymised or aggregated.
- Is data anonymisation safe?
  - Possible leaks:
    - "The second Netflix prize was cancelled because researchers have found how to identify several participants"*,**
    - "AOL release of its anonymized search logs led to a similar fiasco" *, ***

* http://www.datasciencecentral.com/profiles/blogs/big-data-vs-privacy-where-is-the-line
** Arvind Narayanan, Vitaly Shmatikov "How To Break Anonymity of the Netflix Prize Dataset", http://arxiv.org/abs/cs/0610105
*** http://en.wikipedia.org/wiki/AOL_search_data_leak

- Security about data and knowledge.
  - We may relax security when analysing the data
    - We focus on efficient processing
      - We're less careful about users, firewalls, etc.
    - Even if anonymised, unauthorised access is a critical issue.
  - Even if a company is strict about security, they collect too much information.
    - What if it is hacked?
      - The hacker's motto is not going to be: "don't be evil".
    - It is not sufficient to trust an organisation but also to be sure that the organisation is secure!

- What is Information Visualisation?

> *"The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system." (Hearst 2003)*

- Some criteria that visualisation has to fulfill:
  - ***Based on (non-visual) data***. That means that the data must come from something that is abstract or at least not immediately visible.
  - ***Produce an image***. That means that the visual must be the primary means of communication.
  - ***The result must be readable and recognisable***. That means that the visualisation must provide a way to learn something about the data.

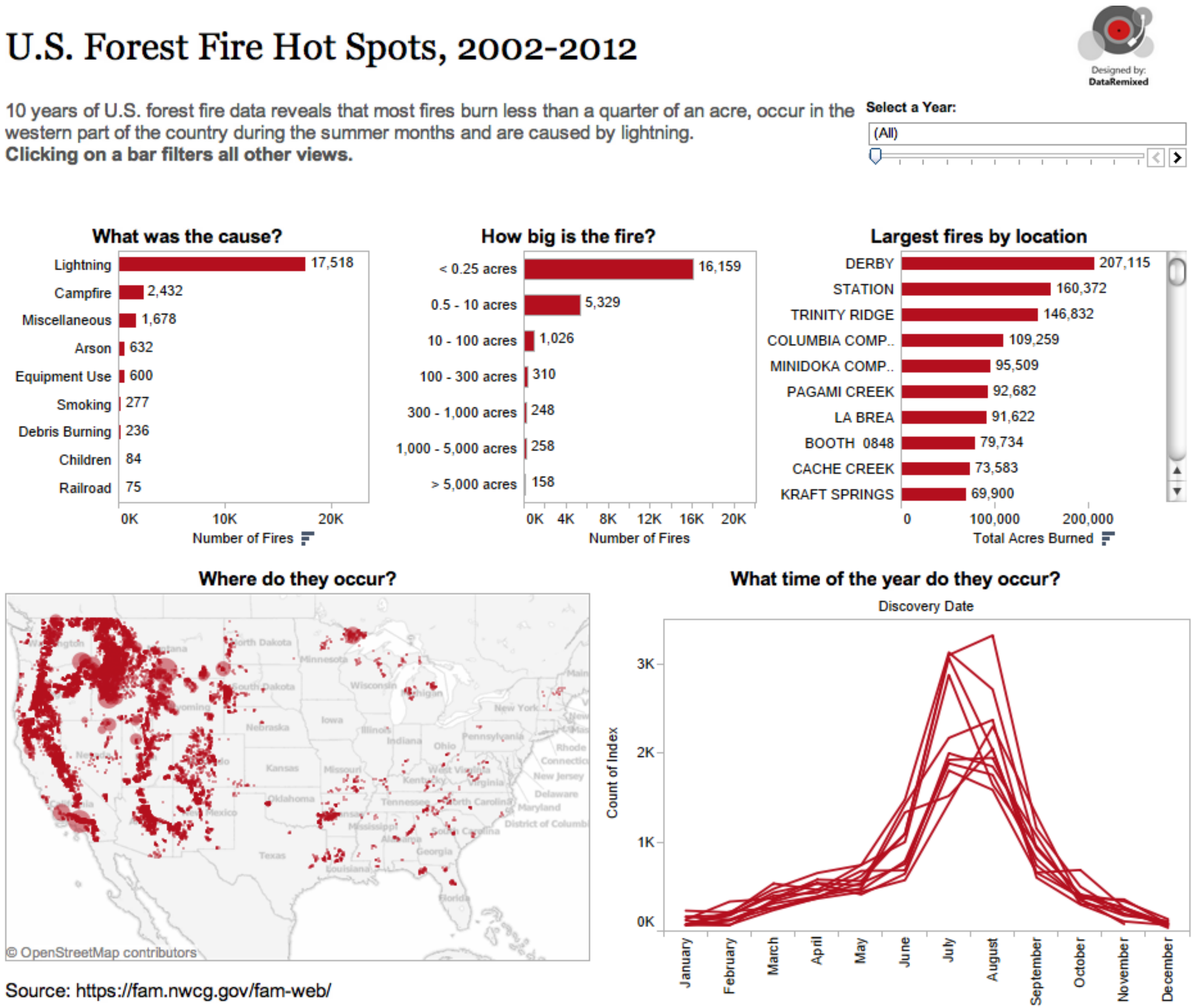
Visualisation and data understanding

- With big datasets, how to understand them?

state	high	bs	adv
Wyoming	91.8	23.8	7.9
Minnesota	91.5	31.5	10.3
Alaska	91.4	26.6	9
New Hampshire	91.3	32	11.2
Vermont	91	33.1	13.3
Montana	90.8	27.4	8.3
Iowa	90.5	25.1	7.4
Hawaii	90.4	29.6	9.9
Utah	90.4	28.5	9.1
Maine	90.2	26.9	9.6
North Dakota	90.1	25.8	6.7
South Dakota	89.9	25.1	7.3
Nebraska	89.8	27.4	8.8
Wisconsin	89.8	25.7	8.4
Kansas	89.7	29.5	10.2
Washington	89.7	31	11.1
Colorado	89.3	35.9	12.7
Oregon	89.1	29.2	10.4
Massachusetts	89	38.2	16.4
Connecticut	88.6	35.6	15.5
Idaho	88.4	23.9	7.5
Maryland	88.2	35.7	16
Michigan	87.9	24.6	9.4
Pennsylvania	87.9	26.4	10.2
Ohio	87.6	24.1	8.8
Delaware	87.4	28.7	11.4
New Jersey	87.4	34.5	12.9
District of Columbia	87.1	48.5	28

  - Take better advantage of human perceptual system.
  - Convert information into a graphical representation.

- Goals of Information Visualisation:
  - Make large datasets coherent
    (Present huge amounts of information compactly)
  - Present information from various viewpoints
  - Present information at several levels of detail
    (from overviews to fine structure)
  - Support visual comparisons
  - Tell stories about the data


U.S. Forest Fire Hot Spots, 2002-2012
Designed by: DataRemixed
10 years of U.S. forest fire data reveals that most fires burn less than a quarter of an acre, occur in the western part of the country during the summer months and are caused by lightning.
Clicking on a bar filters all other views.
Select a Year:
(All)
What was the cause?
Lightning 17,518
Campfire 2,432
Miscellaneous 1,678
Arson 632
Equipment Use 600
Smoking 277
Debris Burning 236
Children 84
Railroad 75
0K 10K 20K
Number of Fires
How big is the fire?
< 0.25 acres 16,159
0.5 - 10 acres 5,329
10 - 100 acres 1,026
100 - 300 acres 310
300 - 1,000 acres 248
1,000 - 5,000 acres 258
> 5,000 acres 158
0K 4K 8K 12K 16K 20K
Number of Fires
Largest fires by location
DERBY 207,115
STATION 160,372
TRINITY RIDGE 146,832
COLUMBIA COMP.. 109,259
MINIDOKA COMP.. 95,509
PAGAMI CREEK 92,682
LA BREA 91,622
BOOTH 0848 79,734
CACHE CREEK 73,583
KRAFT SPRINGS 69,900
0 100,000 200,000
Total Acres Burned
Where do they occur?
© OpenStreetMap contributors
What time of the year do they occur?
Discovery Date
Count of Index
3K
2K
1K
0K
January
February
March
April
May
June
July
August
September
October
November
December
Source: https://fam.nwcg.gov/fam-web/

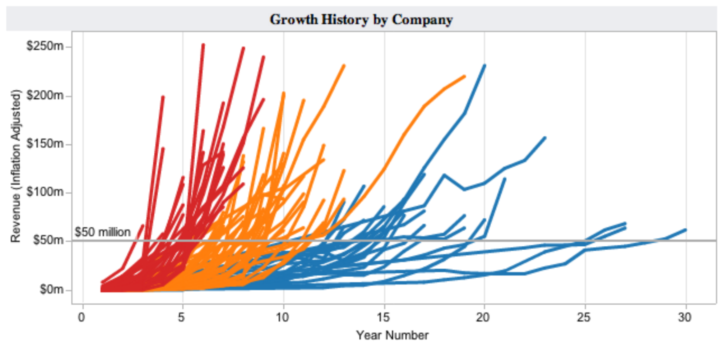
Visualisation and data understanding

How fast do successful tech companies grow? The Wall Street Journal posted this visualization that compares the performance of 100 fast growing software companies.

Tale of 100 Entrepreneurs

Click to interact

Rocket Ship · Hot Company · Slow Burner



Company Details

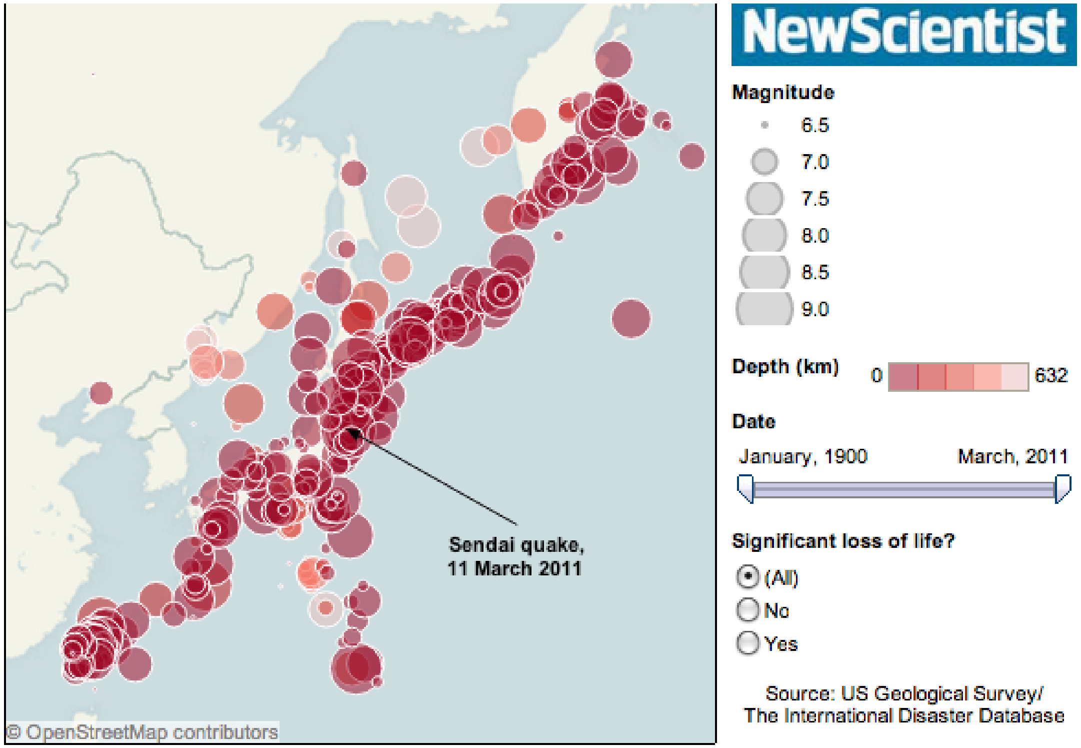
Growth Group	Company Name	Avg. Year Founded	Years to $50m	Revenue (2008)
Rocket Ship	Activision	1979	4	$3,026m
	Adobe	1982	6	$3,580m
	Autodesk	1982	5	$2,172m
	Blackboard	1997	5	$312m
	Cadence	1983	6	$1,039m
	Checkpoint	1993	5	$808m
	China Digital	2004	4	$55m
	Cognizant	1994	5	$2,816m
	Electronic Arts	1982	6	$3,665m
	Interwoven	1995	6	$260m
	McAfee	1989	6	$1,600m
	MedAssets	1999	5	$189m
	Mentor Graphi..	1981	4	$880m
	Novell	1983	3	$957m
	OpenTV	1996	6	$116m
	Rackspace	1998	6	$532m

Select Segment:

- ☑ (All)
- ☑ Business Intelligence
- ☑ CAD / EDA
- ☑ Communications
- ☑ Content Management
- ☑ CRM
- ☑ Database
- ☑ Entertainment
- ☑ ERP
- ☑ Network / Infrastructure/ EAI
- ☑ Operating Systems
- ☑ Security
- ☑ Services
- ☑ Supply Chain
- ☑ Systems Management
- ☑ Vertical Application

http://www.tableausoftware.com/public/gallery/taleof100

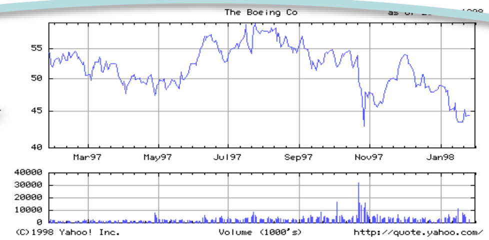
EARTHQUAKES IN JAPAN SINCE 1900



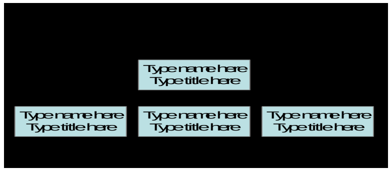
Quake	Total deaths	Quake deaths	Tsunami deaths	Magnitude	Depth (km)
Great Kantō, 1 September 1923	145,144	143,000	2,144	7.9	25
Great Hanshin, 16 January 1995	5,297	5,297		6.9	21
Fukui, 28 June 1948	5,131	5,131		7.3	20

http://www.tableausoftware.com/public/gallery/all

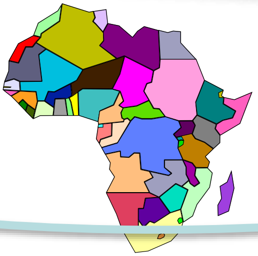
- Graphs



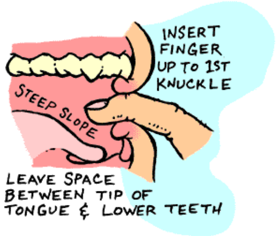
- Charts



- Maps

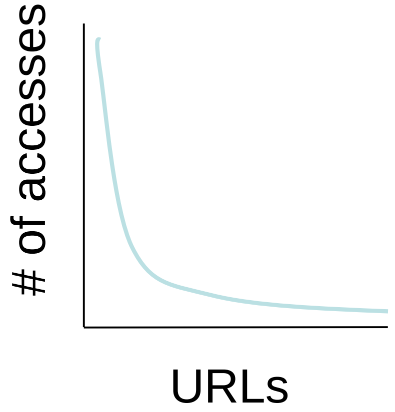


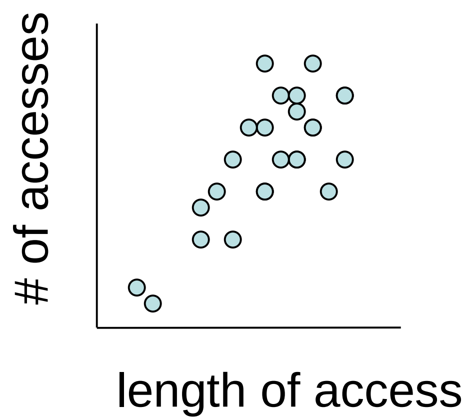
- Diagrams

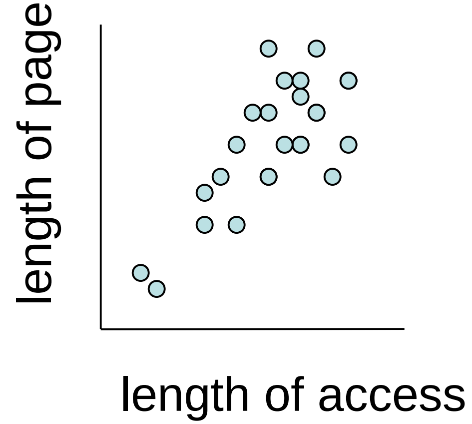


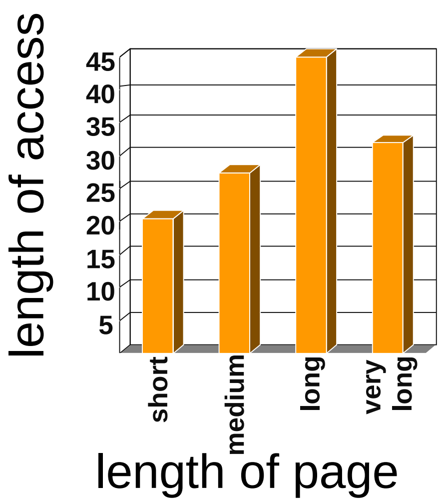
[Kosslyn]

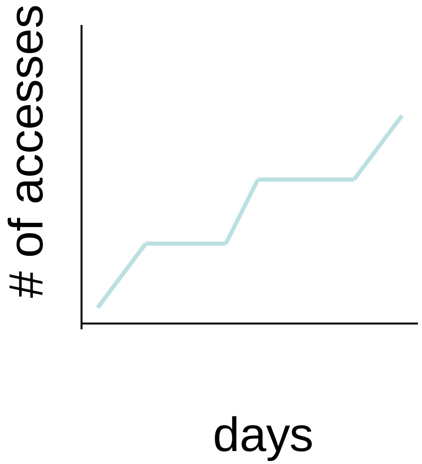
- Common graphs

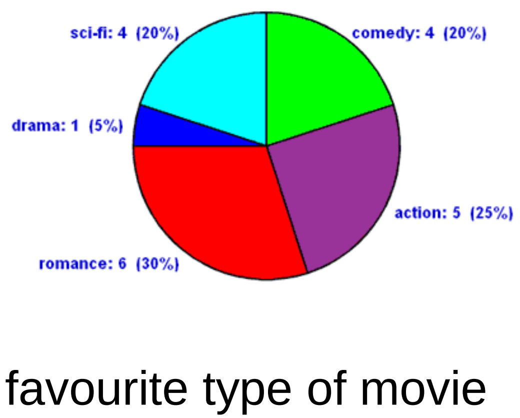












When to use which type?

- Line graph
  - x-axis requires quantitative variable
  - Variables have contiguous values
- Bar graph
  - comparison of relative point values
- Scatter plot
  - convey overall impression of relationship between two variables
- Pie Chart
  - Emphasising differences in proportion among a few numbers

- In order to visualise data:
  - Map data sets to visual attributes (also known as data encoding)

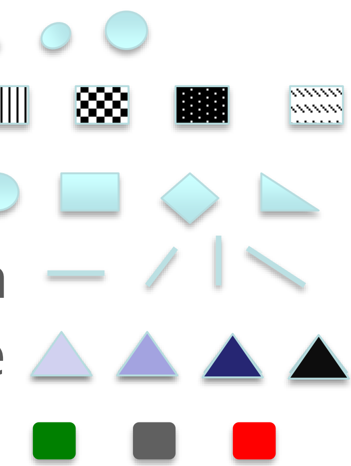
- Process:
  1. Classify data types
  2. Determine which visual attributes represent data types most effectively

- There are three **basic types of data**: something you can just differentiate, something you can order and something you can count.
  - **Nominal or Categorical**
    a limited (and usually fixed) collection values without an inherent order.
      yellow, red, green
      DAS, MFC, FLA, ISE
  - **Ordered or Ordinal**
    values than can be sorted by a rank order but not at measurable intervals.
      low, medium, high
      C, B, A, A+
  - **Quantitative or Numeric**
    numbers or real
      1,2,3,4,8.
      speed, distance, duration,…

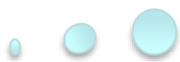
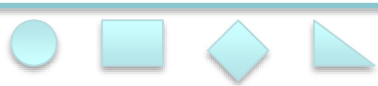
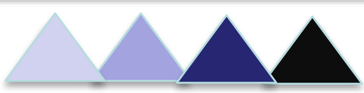
Mapping Data Types to Visual Attributes (Michael Dubakov)

- Planar variables: X and Y represented in a bi-dimensional graph (with two axis). ➔ (Position in Bertin's Visual Attributes)

- **Retinal** variables:
  - Size
  - Texture
  - Shape
  - Orientation
  - Color Value
  - Color Hue

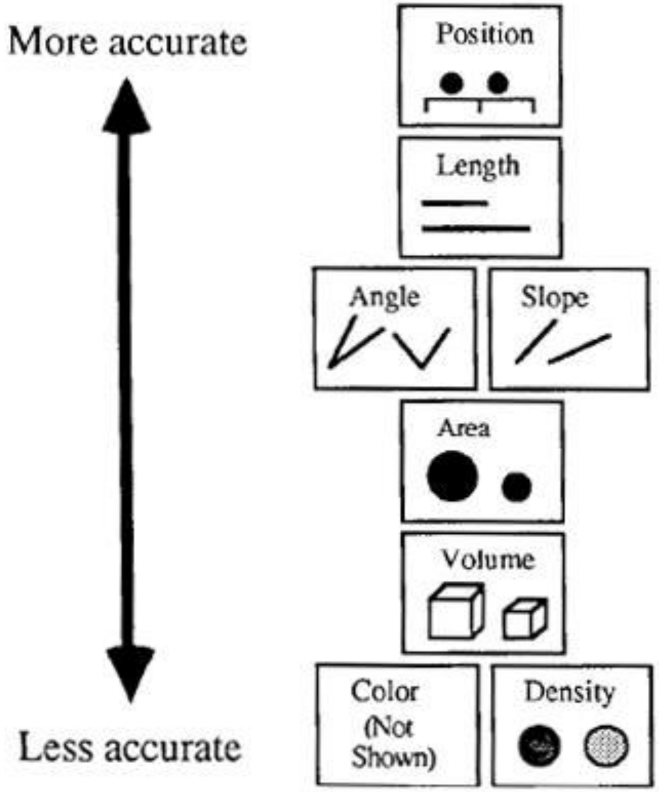


How to Apply the Retinal Variables to Data? (Bertins' Levels of Organisation)

	N	O	Q
X and Y	X	X	X
	X	X	X
	X		
	X		
	X	X	X
	X	X	X
	X		

N: nominal
O: ordinal
Q: quantitative

How to Apply the Retinal Variables to Data? [Mackinlay 88 from Cleveland & McGill]



- Some properties have intrinsic meaning:
  - Value (Greyscale) is perceived as ordered

    Darker → More
  - Size / Length / Area

    Larger → More
  - Position

    Leftmost → first, Topmost → first
  - Hue is perceived as unordered
    - Encode nominal values

- How many variables can be depicted in an image?
  - Univariate data (vectors, factors)
    - bar plots, scatter plots, box plots, line plots
  - Bivariate data
    - scatter plot, line plots
  - Trivariate data
    - 3D scatter plot is possible
      - two variables can map to points (scatter plots, maps, ...
      - third variable must use color, size, shape
  - Multidimensional data
    - use a different visual variable for each dimension

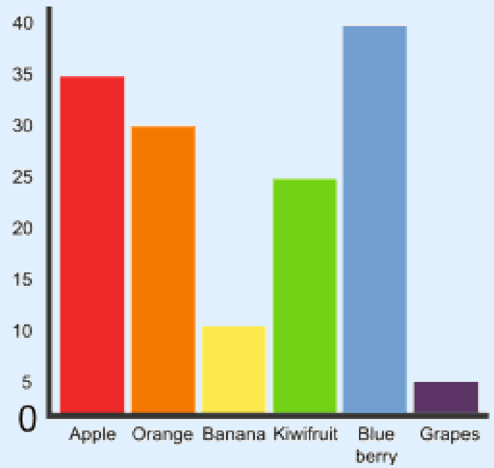
Example

Example: Most Popular Fruit

A survey of 145 people revealed their favorite fruit:

Fruit:	Apple	Orange	Banana	Kiwifruit	Blueberry	Grapes
People:	35	30	10	25	40	5

And here is the bar graph:



For that group of people Blueberries are most popular and Grapes are the least popular.

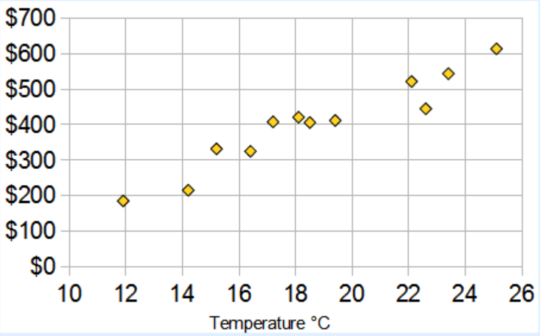
* from https://www.mathsisfun.com/data/bar-graphs.html

- How many variables can be depicted in an image?
  - Univariate data (vectors, factors)
    - bar plots, scatter plots, box plots, line plots
  - Bivariate data
    - scatter plot, line plots
  - Trivariate data
    - 3D scatter plot is possible
      - two variables can map to points (scatter plots, maps, ...
      - third variable must use color, size, shape
  - Multidimensional data
    - use a different visual variable for each dimension

Example

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

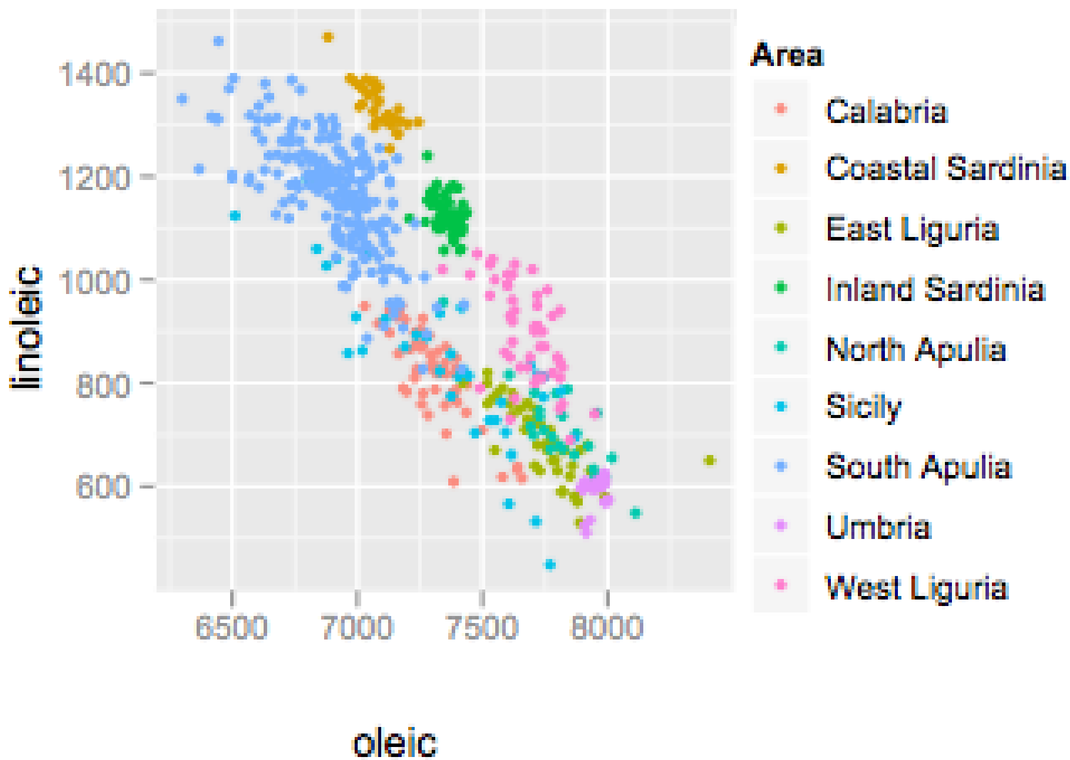
And here is the same data as a Scatter Plot :



* from https://www.mathsisfun.com/data/bar-graphs.html

- How many variables can be depicted in an image?
  - Univariate data (vectors, factors)
    - bar plots, scatter plots, box plots, line plots
  - Bivariate data
    - scatter plot, line plots
  - Trivariate data
    - 3D scatter plot is possible
      - two variables can map to points (scatter plots, maps, ...
      - third variable must use color, size, shape
  - Multidimensional data
    - use a different visual variable for each dimension

Example

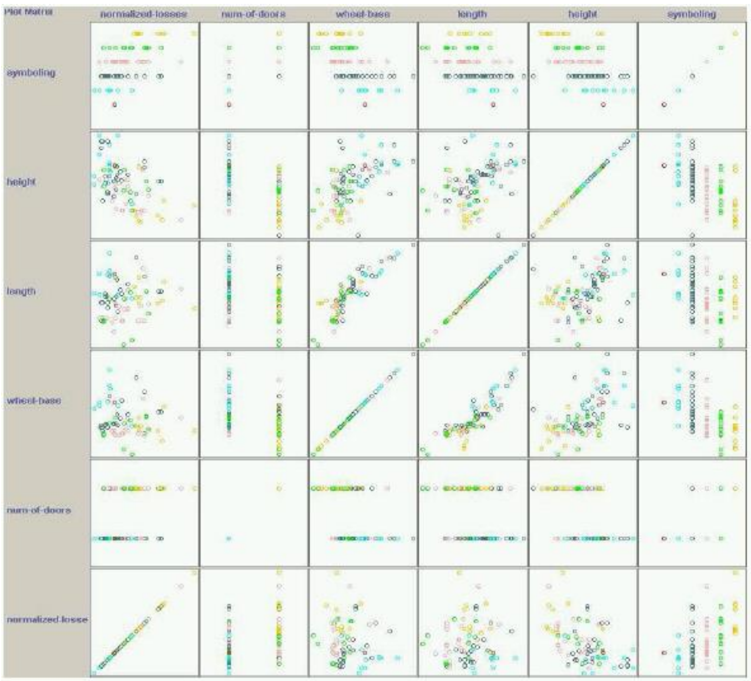


* from http://streaming.stat.iastate.edu/~dicook/multivariatelectures/students/graphics-class.pdf

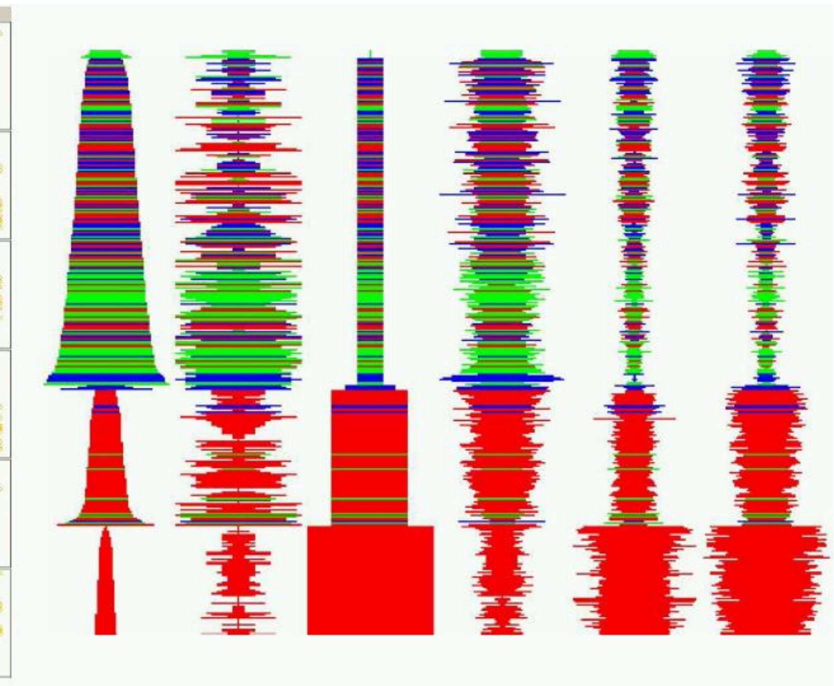
- How many variables can be depicted in an image?
  - Univariate data (vectors, factors)
    - bar plots, scatter plots, box plots, line plots
  - Bivariate data
    - scatter plot, line plots
  - Trivariate data
    - 3D scatter plot is possible
      - two variables can map to points (scatter plots, maps, ...
      - third variable must use color, size, shape
  - Multidimensional data
    - use a different visual variable for each dimension

Example

scatterplot

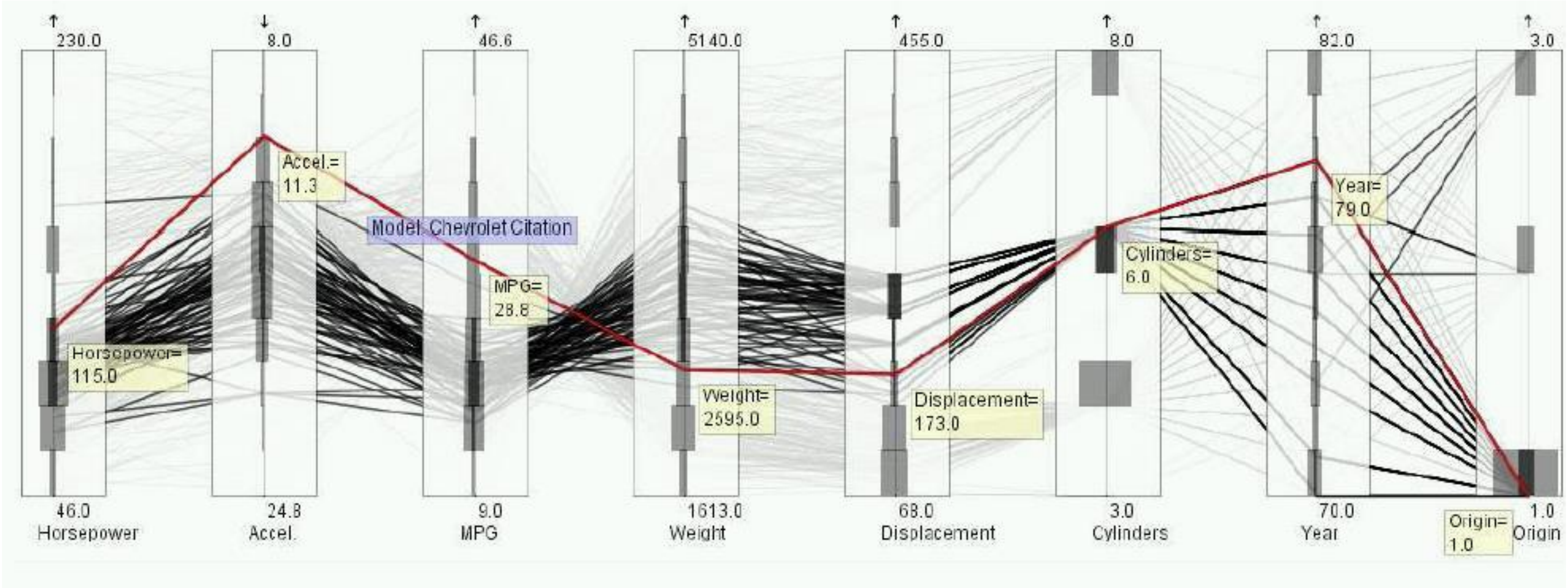


surveyplot



Example

Parallel coordinates



Some examples

- Olympics Games
    - Visual encoding variables:
        - Colour: continent
        - Size: medals count
        - X and Y: the world map
- Basketball Teams Performance
    - Visual encoding variables:
        - X and Y: basketball court map
        - Colour: points per region
        - Size: number of attempts
- Usain Bolt vs. The World
    - Visual encoding variables:
        - Colour: natural colors used to encode bronze, silver and gold medals
        - X: metres behind Bolt (quite an unusual but very impressive metric)
        - Y: year

https://www.targetprocess.com/articles/visual-encoding/

- Exercises and recommended readings.

  - Graphics in R: the ggplot2 package
    http://www.statmethods.net/advgraphs/ggplot2.html
    http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

  - Examples
    - https://towardsdatascience.com/10-viz-every-ds-should-know-4e4118f26fc3
    - https://towardsdatascience.com/10-viz-every-ds-should-know-4e4118f26fc3
    - https://python-graph-gallery.com/