

CA682

Data Management & Visualisation

01 Introduction

Suzanne Little

suzanne.little@dcu.ie

Today

- Module outline
 - Assessment
 - Practicals
 - Resources and support
-
- Who are you?
 - What about data?
-
- A Data Analytics Pipeline
 - Why do we visualise data?
 - What's next?

Objectives for today:

- Get an overview of topics in CA682
- Get to know me
- Get to know a little about each other
- Ask me questions
- Experiment with technology :-)

Questions? I'll periodically stop but you can use the Q&A link at the top of the slides. Please put your name (first is fine) at the end of your question.

Housekeeping

Be polite and respectful to your fellow students and to me (have your phone on silent, stay quiet when I'm talking)

Keep the back 2 rows and the aisle seats free until after 9:15am

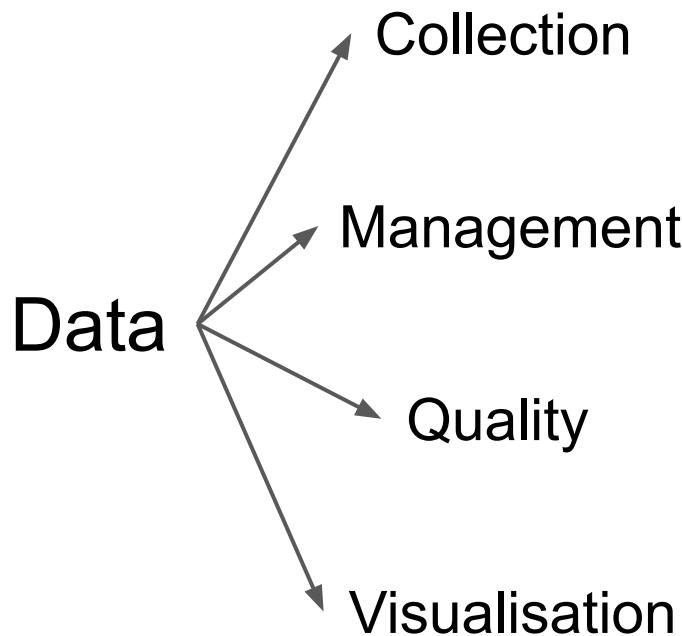
If an alarm sounds follow the exit signs outside to the assembly point

No food or drink allowed in the lecture theatre

I'll generally give you a 10-15 min break part way through

About me

Module Content Outline



Outcomes

1. Analyse the requirements of applications handling large datasets.
2. Demonstrate an ability to efficiently structure a large dataset.
3. Implement data quality measures.
4. Identify and implement appropriate data visualization techniques.

CA682 module specifics

Fully in-person delivery

7.5 credits

- Thu 09:00-10:50, lecture, discussions, case studies, activities
- Thu 11:00-11:50, labs, practical work, programming, working on assignment
- Tutors will be available in the labs to lend a hand
- plus self study and assignment work

Assessment: 25% assignment, 75% exam

Resit Category 3: if you fail then you can only retake the exam (75%), not the assignment

Resources and Support

Loop (Moodle)

- main source of communication, notes, resources
- check your student emails! (@mail.dcu.ie)
- links to shared materials on Google Drive (must be logged in @mail.dcu.ie)
- CA682? CA682A? D? E? → all use CA682 loop module

DCU Library

- Online ebooks (see Loop for a list and links to resources)

Myself

- Email is best for contacting me (suzanne.little@dcu.ie)
- Please use your @mail.dcu.ie address if possible & start the subject with **[CA682]**

Resources

There is no set textbook. Reference material will be available online.

Two books (available via the DCU library) that address Data Visualisation well:

- “Data Visualisation”, Andy Kirk (2016), Sage Publishers → 2012 version as ebook
- “Storytelling with Data”, Cole Nussbaumer Knaflc (2015), Wiley

Also: “The Data Science Handbook”, Field Cady (2017), ebook

Available on loop: slides, some notes, links to readings, practical exercises, revision quizzes, discussion topics

Assessment -- more detail later

Assignment 25%

- Create a complex data visualisation and present using a short screencast (<5mins)
- Due Friday December 1st via upload to loop
- More information later - will be in pairs

Exam 75%

- Multiple choice and short answer questions
- Past papers and examples are/will be available online
- On lab machines, during the exam session

Questions?



Today

- Module outline
- Assessment
- Practicals
- Resources and support
- Who are you?
- What about data?
- A Data Analytics Pipeline
- Why do we visualise data?
- What's next?

Getting to know you
<https://vevox.app>
103-243-683

Data Scientist



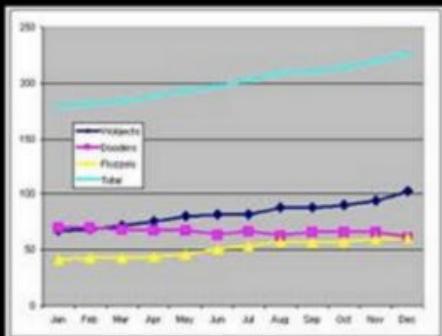
What my friends think I do



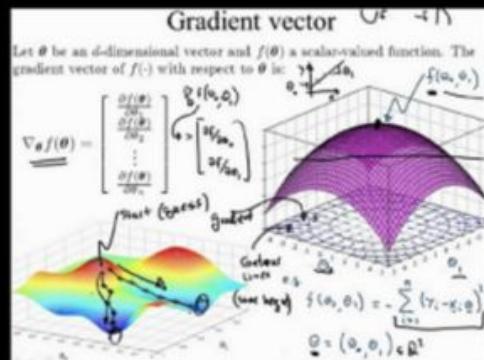
What my mom thinks I do



What society thinks I do



What my boss thinks I do



What I think I do



What I actually do

Housekeeping

Be polite and respectful to your fellow students and to me (have your phone on silent, stay quiet when I'm talking)

Keep the back 2 rows and the aisle seats free until after 9:15am

If an alarm sounds follow the exit signs outside to the assembly point

No food or drink allowed in the lecture theatre

I'll generally give you a 10-15 min break part way through

Today

- What about data?
- A Data Analytics Pipeline
- Why do we visualise data?
- What's next?

Question: Is it “data is” or “data are”?

Data is or Data are ?

<https://www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular>

Discussion: What is data?
How many potential data sources can
you think of?

<https://vevox.app>

171-231-890

Data

	A	Q	R	S	T	U	V	W
1	Total salaried empl	1995	1996	1997	1998	1999	2000	2001
32	Chile	69.40000153	70.09999847	70.40000153	69.19999695	69.19999695	69.40000153	68.59999847
33	Colombia	66.19999695	66.5	64.90000153	64.09999847	61.40000153	60.90000153	49.2999924
34	Costa Rica	71.40000153	71.19999695	69.90000153	70.90000153	71	70.80000305	68.80000305
35	Croatia		71.40000153	74.09999847	75.30000305	75.19999695	76.09999847	75.69999695
36	Cuba		84	84.30000305	83.59999847	82.69999695	81.5	80.09999847
37	Cyprus					73.69999695	73	76.30000305
38	Czech Rep.	86.09999847	86	86.09999847	85	84.5	83.90000153	84
39	Denmark	90.5	90.59999847	91.09999847	90.80000305	90.90000153	91.40000153	91.19999695
40	Djibouti					58.90000153		68.30000305
41	Dominica					56.69999695		
42	Dominican Rep.	58.29999924	59.40000153	53.90000153	53.20000076	52	56.29999924	54.90000153
43	Ecuador	53.40000153	52.5	54.20000076	53.09999847	59.29999924	59.5	59.40000153
44	Egypt	57.09999847	69.69999695	60	59.79999924	61.09999847	59.90000153	61.5
45	El Salvador	52.20000076	51.90000153	52.70000076	58.70000076	60.20000076	52.09999847	51.70000076
47	Eritrea				78.30000305			
48	Estonia	93.09999847	92.5	92	91.40000153	91.40000153	91	91.69999695
49	Ethiopia					8.199999809		
50	Fiji							
51	Finland	83.30000305	83.5	84.09999847	84.80000305	85.19999695	85.59999847	86.30000305
52	France	89.19999695	89.59999847	89.90000153	90.19999695	90.5	90.80000305	91.09999847
53	Gabon							
54	Georgia							
55	Germany	89.40000153	89.5	89.09999847	88.90000153	89.30000305	89.19999695	88.90000153
56	Greece	53.90000153	54.29999924	54.79999924	56.40000153	57.90000153	58	60.09999847
61	Honduras	49.40000153	46.09999847	46.79999924	48	46.79999924		45.5
62	Hong Kong, China	89.19999695	89.19999695	89.69999695	89.69999695	89.19999695	89.5	88.09999847
63	Hungary		85.5	85.30000305	85.80000305	87.09999847	84	84.59999847
64	Iceland	80.69999695	81.80000305	82.30000305	82.09999847	82.30000305	82	83.09999847
65	Indonesia							
66	Iran							
67	Ireland							
68	Isle of Man							

log files

social media content

photographs

microblogs

surveys

news

cctv video

movies

television

sales records

clicks

adwords

statistics

audio recordings

playlists

search terms

sensors

pedometer/activity monitor

spectrographs

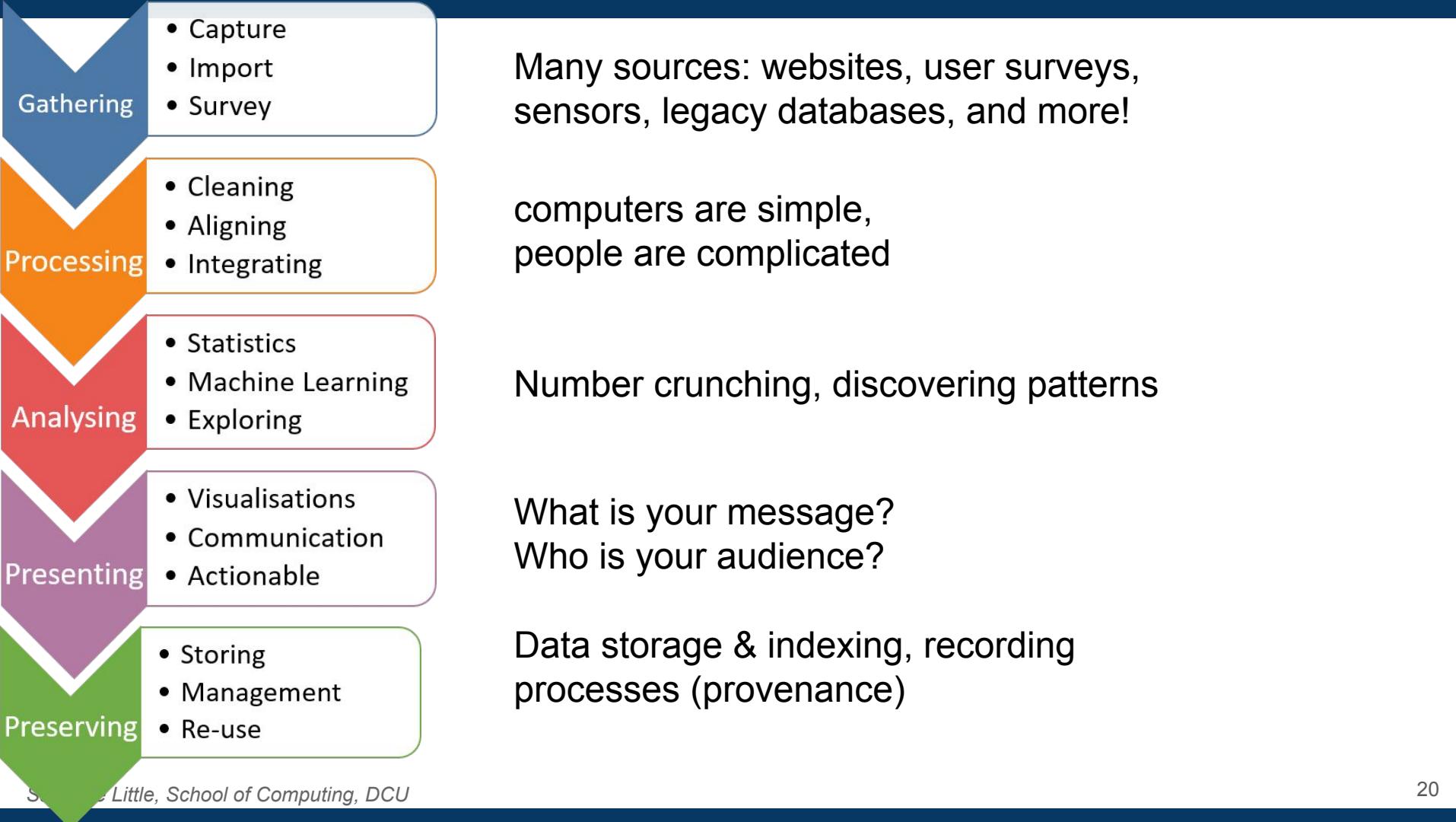
microscopy

genomes

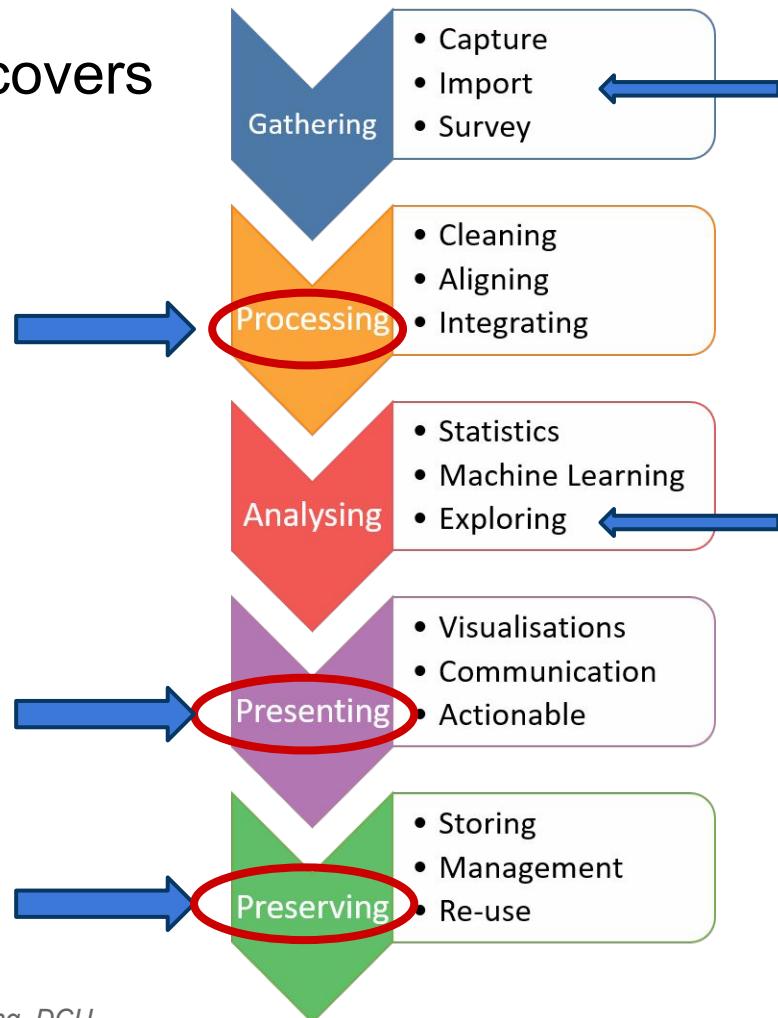
numbers



Data is collected information
(a working definition)

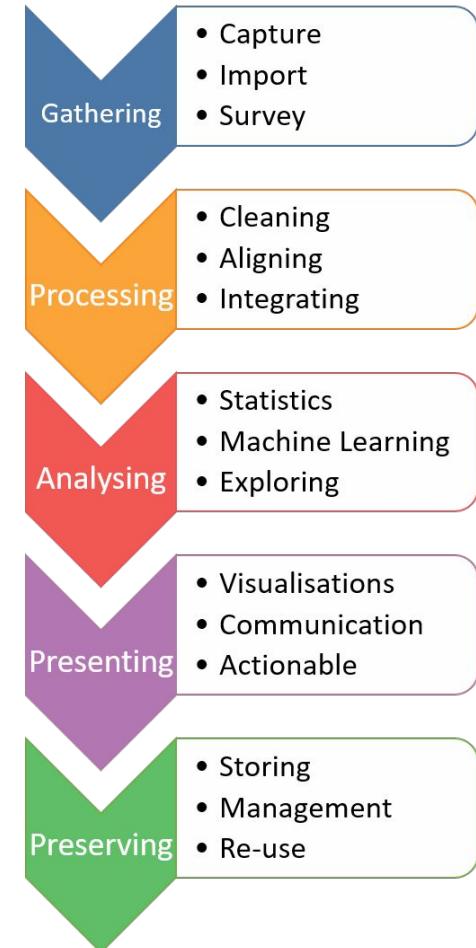


CA682 mostly covers



Topics plan (subject to change!)

1. What is data?
2. Describing data
3. Finding data
4. Cleaning data
5. Communication
6. Encoding data
7. Designing data-driven visualisations
8. Managing and Storing data + Big Data
9. Data protection & privacy
10. Final wrap up and exam information



What background do I need to have?

Or what technologies will you be teaching?

CA682 is *technology agnostic*

That is, you learn the fundamental principles and apply them using a range of tools.

Labs will include exercises using Python, Tableau, OpenRefine, Spreadsheets

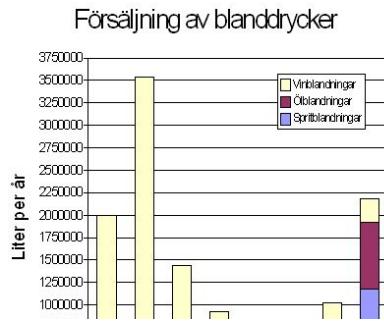
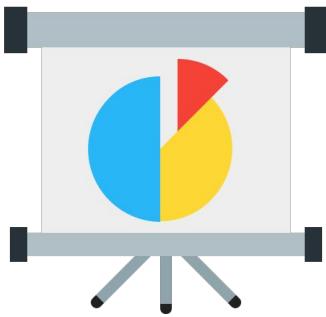
Your assignment can be completed using many different tools. The exam won't specify which tool to use.

Complete the class background skills survey and we'll discuss more next week

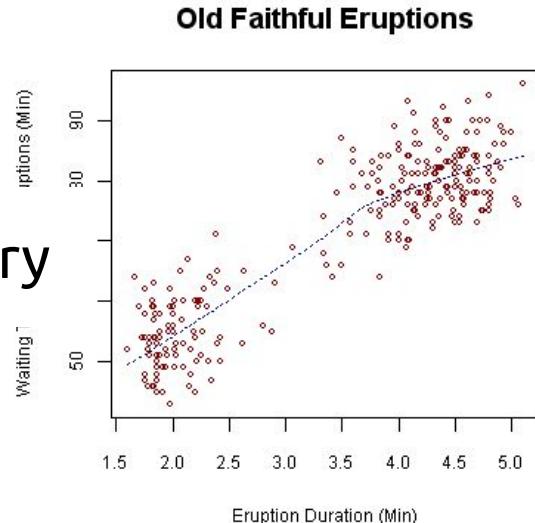
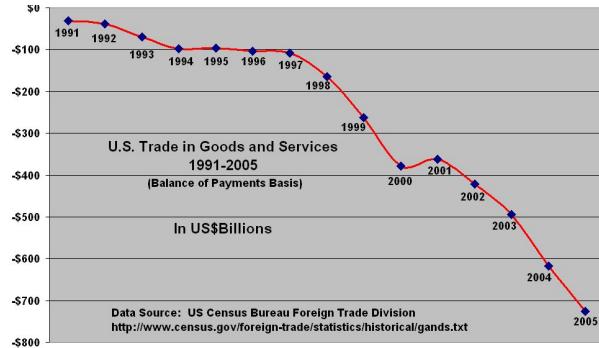
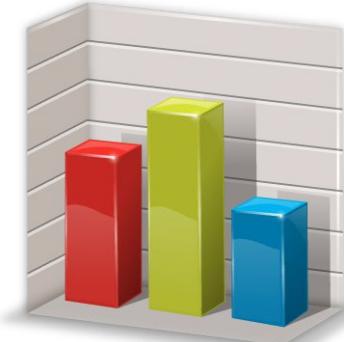
Questions?

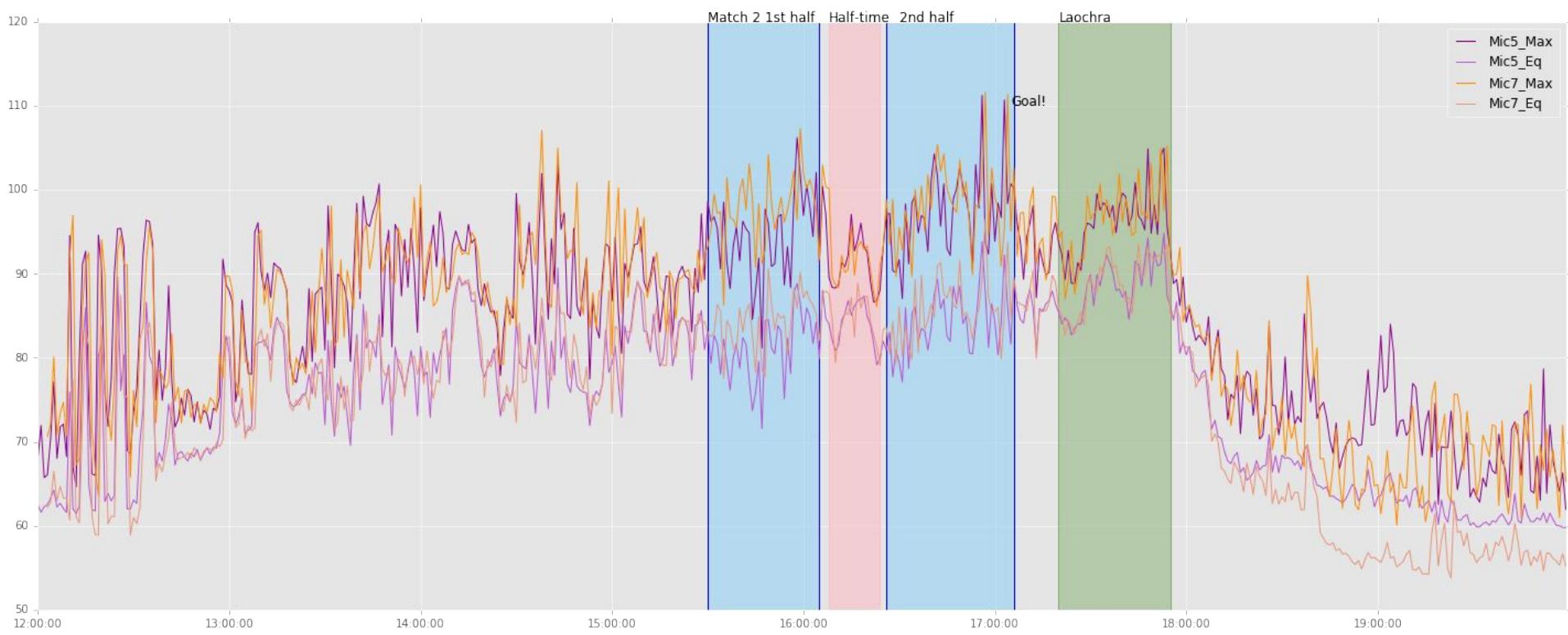


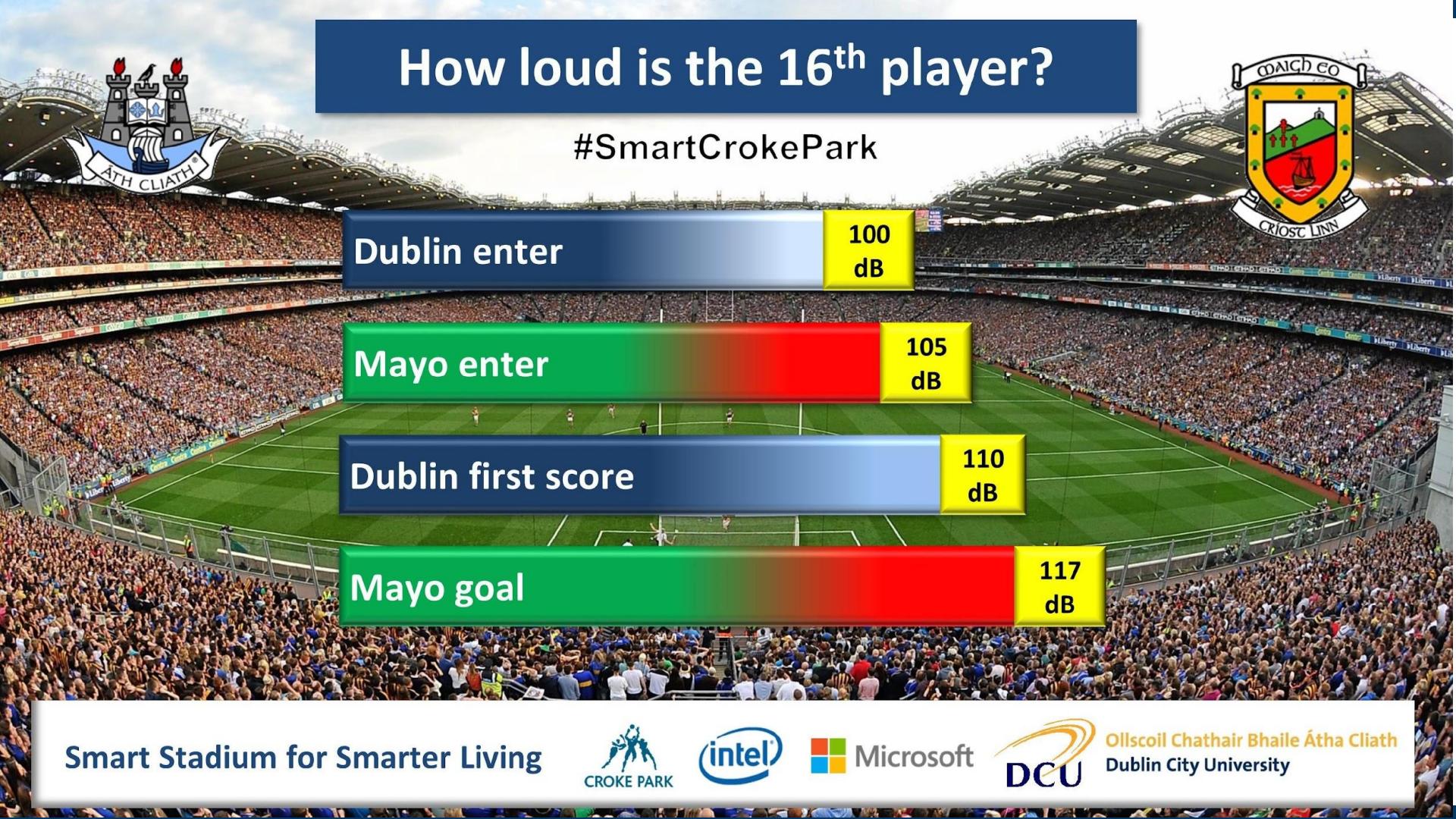
Why do we visualise data?



Exploratory vs Explanatory







How loud is the 16th player?

#SmartCrokePark

Dublin enter

100
dB

Mayo enter

105
dB

Dublin first score

110
dB

Mayo goal

117
dB

Smart Stadium for Smarter Living



Microsoft

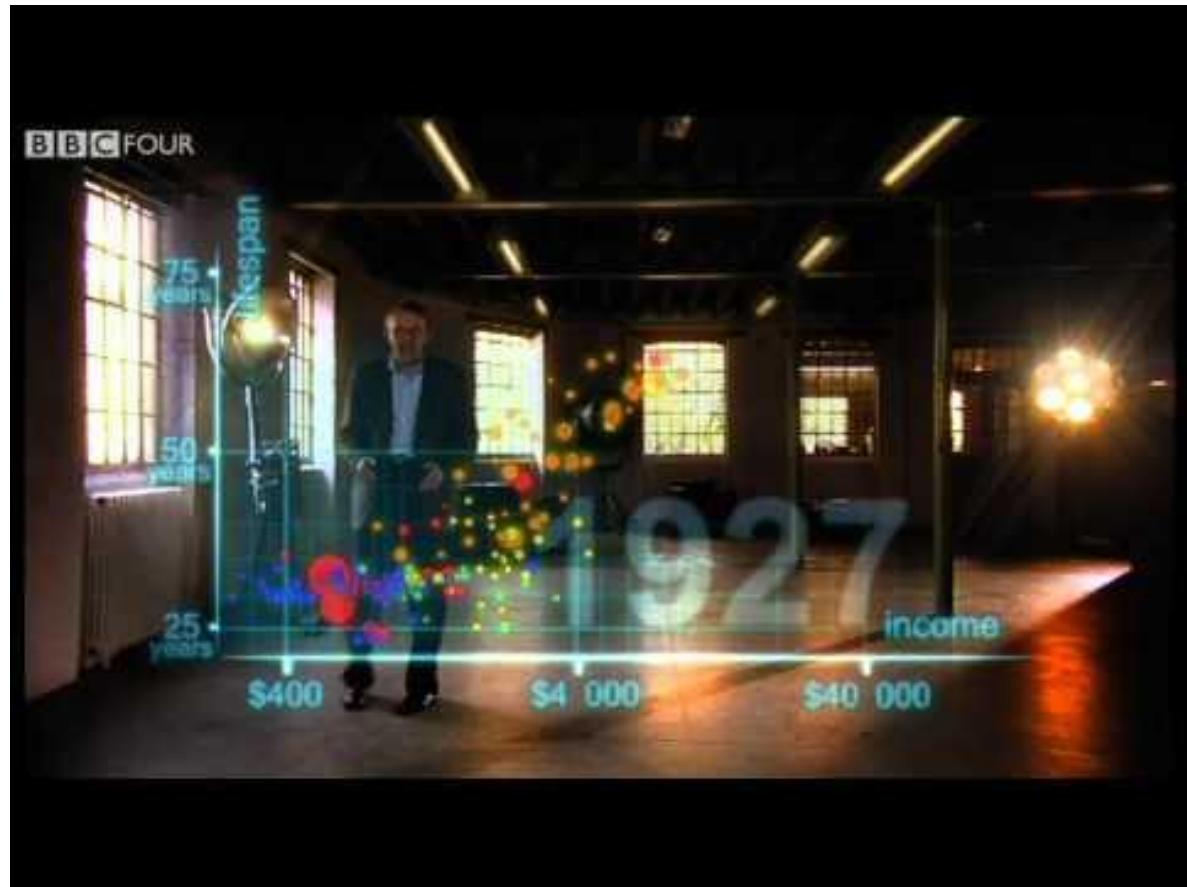


Ollscoil Chathair Bhaile Átha Cliath
Dublin City University

Questions?



Hans Rosling
See TED talks on economics
<https://youtu.be/hVimVzgtD6w>



Data Visualisation: Access to the Internet



www.gapminder.org

Today's Lab (11-12)

LG25: Family name A-I

LG26: Family name J-P

L114: Family name Q-Z

1. Complete the background skill survey -
<https://forms.gle/tNR4LUFDH8YA5vUj8>
2. Explore gapminder and answer the questions
3. Quiz on loop (no marks, just feedback)

Next week

Resources will be shared with you - linked from Loop

For this week I'll also put links to the slides and lab information on
<https://gitlab.computing.dcu.ie/slittle/ca682-content>

[Note: you must be on dcu eduroam to access gitlab from your laptop]

Before next week:

- complete the background skills survey
- upload your discoveries from gapminder
- review documents on loop (or check gitlab for links)

Questions?



Links & Resources

GapMinder - www.gapminder.org

Hans Rosling TED - <https://youtu.be/hVimVzgtD6w>

Hans Rosling BBC - <https://youtu.be/jbkSRLYSojo>

→ CA682D students could you wait please ...

CA682 Data Types

Suzanne Little

Data types

- Recall: Data is collected information (a working definition)
- Structured vs Unstructured
- Quantitative vs Qualitative
- Discrete vs Continuous
- Four levels of data
- Some special data types to watch for

Data

Example: a *person* (**object** or **entity** or **instance** or **record** or **row**) has **attributes** (or **features** or **descriptors** or **variables** or **columns**)

- Name
- Passport number
- Birth place
- Eye colour
- Shoe size



Structured

tables, organised, observations,

Row is instance, Column is attribute

Examples:

company records

scientific observation

Easier for Machine Learning to work with (kinda)

1	Total salaried empl	1995	1996	1997
32	Chile	69.40000153	70.09999847	70.40000153
33	Colombia	66.19999695	66.5	64.90000153
34	Costa Rica	71.40000153	71.19999695	69.90000153
35	Croatia		71.40000153	74.09999847
36	Cuba	84	84.30000305	83.59999847

vs

Unstructured

No hierarchy or arrangement

Raw signals that need processing

Examples:

tweets & social media posts

server logs

media (images, video, etc)

More challenging to work with. How to turn into “Structured”?



@DublinCityUni

Following

Wishing all of our new and returning students the very best of luck on their first day of lectures!

1:28 AM - 24 Sep 2018

13 Retweets 71 Likes



Special types of data to watch for

- Temporal (or Time Series)
- Geographic (or Spatial)
- Documents, Images, Video, Audio, 3D
- “Raw” data - unstructured and (sometimes) incidental

Qualitative

vs

Quantitative

Quality, Label, Trait

Categorical

Limited mathematical functions

Examples:

Country of origin

Gender

Favourite Colour

Quantity, Measurement

Numerical

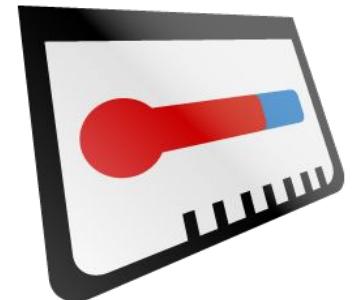
“All the maths!” (well most)

Examples:

Shoe size

Temperature

Bank balance



Quantitative

Discrete

vs

Continuous

only certain values are valid

ie: there are gaps

usually from counting

Examples:

Number of times attended

Number of crimes reported

theoretically any value is possible

depends on measuring device ability

usually from measurements

Examples:

Cholesterol level

Time required to complete task

Data types

Structured vs Unstructured

Quantitative vs Qualitative

Discrete vs Continuous

Four levels of data measurement

1. Nominal
2. Ordinal
3. Interval
4. Ratio

NOIR (Stanley Stevens)

Categorical

Nominal (name, label, category)

Gender, Department, Language

Not described by numbers

No maths except equality & set membership
mode but not mean or median

Ordinal (labels plus order)

Temperature (very hot, hot, warm, mild)

Medals (Gold, Silver, Bronze), Scale (Likert - 1 to 10), colour

Can be arranged by order but not added or subtracted, median but not mean

Measurement

Interval (numbers with proportionate spaces)

We can now talk about “difference” (+/-)

Income, Shoe size,
Temperature ($^{\circ}\text{C}$, $^{\circ}\text{F}$)

“defined interval between values but lacks a zero point”

Ratio (also numbers but with zero)

Can now multiply & divide

Age, Amount of rainfall, Book sales,
Temperature (in Kelvin), [normally counting]

Zero has meaning - no negatives

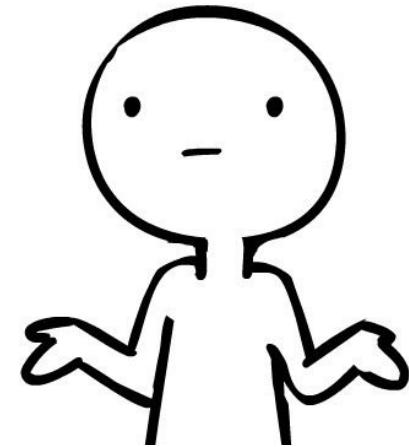
Qualitative

Quantitative

Why do we care?

Type of data determines:

- What statistics are possible/meaningful
- How data can be processed and/or stored
- Which machine learning model can be used
- Which visualisation method to use



kahoot.it - let's try identifying types ...

Data Sources

social media surveys

the internet
dcu hospitals

people

mark zuckerberg everywhere

offline forms ott platforms
travel history distributions data warehouse educational data electricity usage simulations
distributions* data logs communication sports telecast information captured observational study website cookies audio brilliant minds
wikipedia stock markets forums wearable devices covid pandemic emoji reactions google public forums
interactions newspapers sensors technology survey transactions applications smart devices internt names
television experiment universe web everything feedback football devices information video subscriptions collueagues
google map e-commerce dcu loop wearables to recordings games apps
school maps gtl bank transactions nature sports totdevices sources :)
businesses maps gt business to iot devices word of mouth game
transition cricket population sensor lot devices card swiping incident time
behaviours databases facebook dcu hospitals olympics survey data
interviews phone forms discussions network rss users cards journals user application
polls input network addresses website lot census histroy human beings
opinions cellphone towers e-commerce data digital activity multiple sources logs cctv athe internet mobiles objects
telemetry chat space actions uci laptop history mobile phone data log camera feefal emails electronic devices tesla networks
trasactional technologies books machines events reports customer records websockets trackers company everywhere!
book experience serors banking snowden education smart cities transactional data numbers online transactions censor
mark and sundar cosmic cycles instruction collecti surrounding transactionw passengerlocatorform coms from activity
customer data

Where does data come from?

Type your answer here...

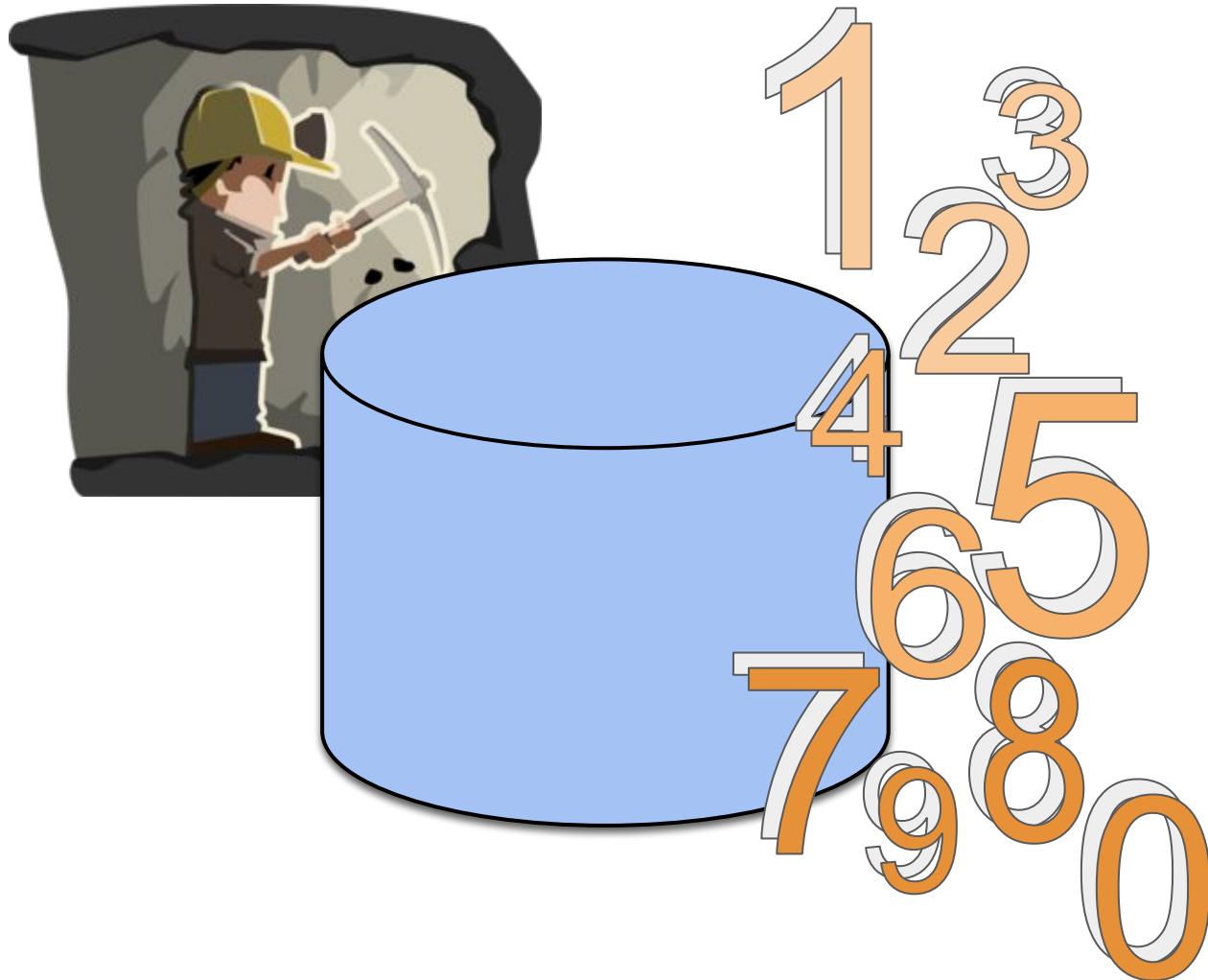
Submit

20 characters remaining

user browsing habits pollution data cars individuals employees data observation , survey sales data linkedin
 vox customer habits measurements stocks phone video, audio, database weather hand written doc any observations
 spending habits scriptures survey ecommerce sites ecommerce erp systems internet browsing literacy annual reports newspapers
 train data my fridge videos census zoom songs comms networks geographical data stock exchange
 satellites consumer surveys books everything reports internet transactions social media, iot malware data
 exam results images cookies web from humans facebook logfiles any machine diaries
 people create it!! latin bus ticket information ads library publications google nest humans iot devices data
 mobile phone feedback survey publications database heath records payroll
 audio files smart phones history human mind digital media surveys research papers encyclopedia the internet
 citizen data phone conversations databases cctv data people iot cctv us pols
 smart devices databases bio-data my oven fitbit environment camera phone records world information
 insurance data flights 2nd year students product information applications, letters banks wearable everthing
 consumer habits browsing data web, media, images hospital monitors sensors, social medi our own trail any recorded activit server data
 photograph data is everywhere multimedia files, documents, pa recorded information tracking
 humans and machines smartphones, e-commer files

Data sources

- Files
- Databases
- “The Internet”
- Open Data



Data sources: Databases

- Traditional relational db: Oracle, MySQL, Postgres, etc.
 - Tables (“relations”) of rows and columns
 - Unique key per row
 - Links between rows (“foreign key”)
 - Optimise structures (the database schema)
 - Stored procedures (queries) to speed up responses
 - Most commonly use SQL - Structured Query Language
 - `SELECT CustomerName,City FROM Customers;`
 - `SELECT CustomerName,Age FROM Customers WHERE City='Dublin';`
- In memory databases: SAP Hana (<http://hana.sap.com/abouthana.html>)
- NoSQL, document, column, graph, etc.

Data sources: the Internet

- Crawlers or spiders
 - Scraping data from semi-structured sources
 - Parse HTML
 - Match Patterns to extract data
 - Identify links (repeat)
- URL
 - Files and databases on the web
 - Many libraries and apps will accept either a local path or url
- How many file formats?

Open data

Public data, shared and freely available

Why?

Why not?

Examples of open data

Some examples of projects that use open data are:

- [Plantwise - Lose less, feed more](https://www.plantwise.org/) (<https://www.plantwise.org/>)
- [Humanitarian OpenStreetMap](https://www.hotosm.org/) (<https://www.hotosm.org/>)
- [OpenGLAM](https://openglam.org/) (<https://openglam.org/>)

Or on a less elevated topic ... the [Great British Public Toilet Map: open geospatial data!](#)

Where to find open data?

- <https://data.gov.ie/>
- <http://www.dublindashboard.ie/>
- <https://www.google.com/publicdata/directory>
- <https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/>

Also lots of datasets for learning data science:

- <https://www.kaggle.com/datasets>
- <https://github.com/datasets>

Exercise

[Data.gov](#) is the portal for the US Government's Open Data. Browse the portal and find a dataset to answer the following questions.

1. What format is the dataset available in?
2. How many features (attributes or columns) does the data have?
3. Are the features mostly categorical (qualitative) or numerical (quantitative)?
4. What is a question that this dataset could help you answer? (you don't need to provide the answer!)

What's on Loop?

Material from last week plus Formal Data Management Lifecycles document

Slides & Notes on Data Types

Notes on Files

→ Exercise: Data Formats - Files

Notes on Open Data

→ Exercise: using open data

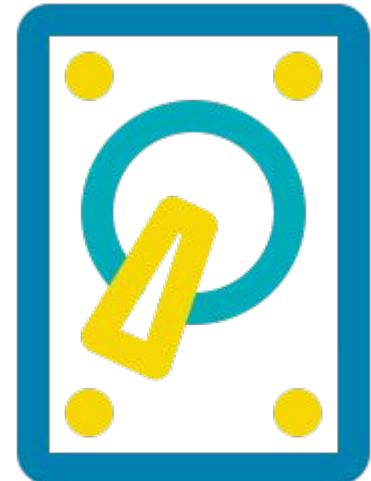
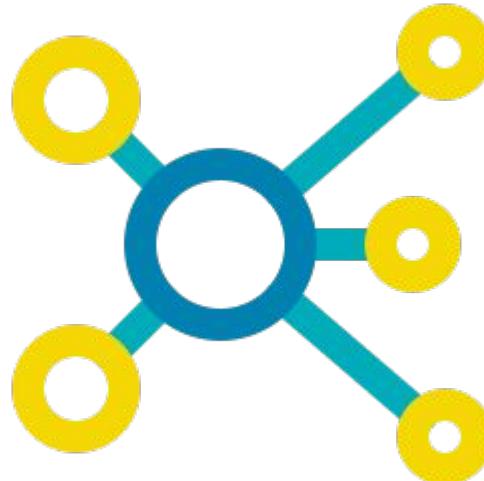
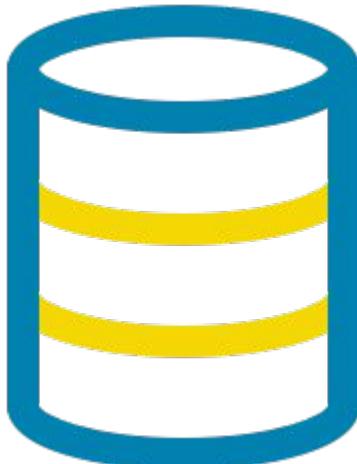
Linked Data (RDF & SPARQL) → Includes Exercise: using SPARQL on DBpedia

Overview of Big Data (3Vs)

(or is it 4Vs ...)

Suzanne Little

You may have heard the term but just
how **big** is “big data”?



Big data

Characterised by 3 'V's

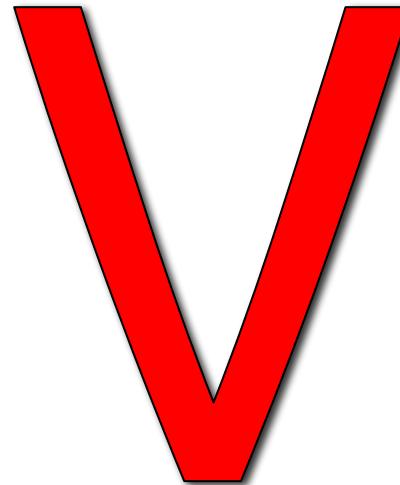
Volume

Variety

Velocity

a 4th V is sometimes added ...

April 1st post [42 Vs of Big data!](#)



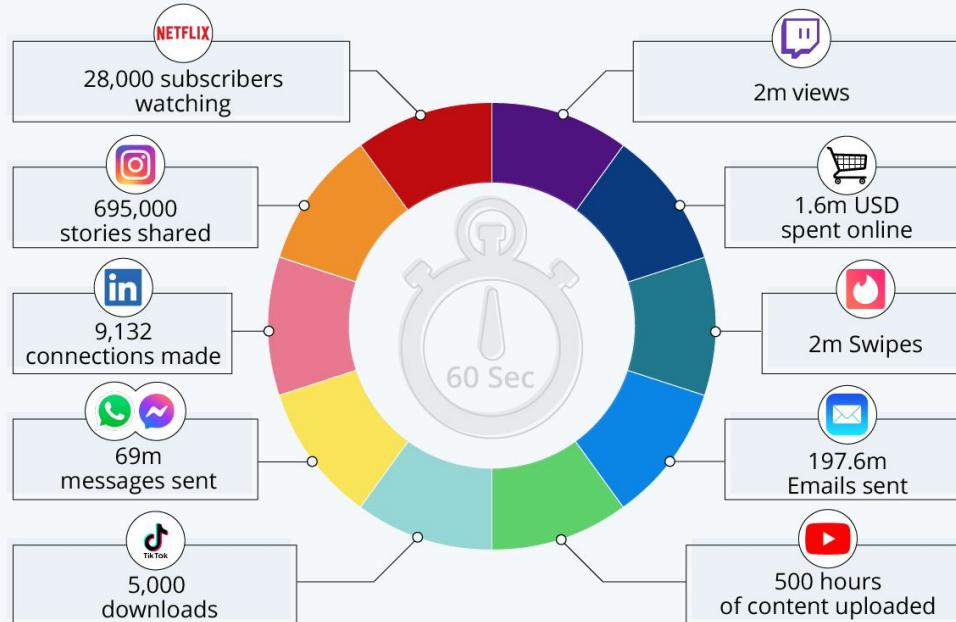
Big data: Volume

- Refers to the **amount** of data
- For big data, varies from terabytes to petabytes to zetabytes
- Can you open the whole dataset in your PC? Probably not big.
- In 2008, Google was already processing 20,000 terabytes of data (20 petabytes) a day
- In 2018, Google processes 40,000 searches per second!
- Social media produces vast quantities of data per minute!

<https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#672fd38760ba>

A Minute on the Internet in 2021

Estimated amount of data created
on the internet in one minute



Source: Lori Lewis via AllAccess



statista

[https://www.statista.com/chart/25443/
estimated-amount-of-data-created-on-
the-internet-in-one-minute/](https://www.statista.com/chart/25443/estimated-amount-of-data-created-on-the-internet-in-one-minute/)

2020 This Is What Happens In An Internet Minute



2019 This Is What Happens In An Internet Minute



Big data: Variety

- Refers to **differing** types and data sources
- Structured, semi-structured and unstructured data
- Organisations may need to combine data from many different sources
- With the proliferation of analytics and sensor data, the variety of data is expanding rapidly
- Q: How many digital cameras do you own?

Big data: Velocity

- Data in motion -- **dynamic**, temporal
- We may need to process data as it arrives
 - ...because we cannot store such volumes
 - ...because we need timely processing
- Related notion of latency (lag-time)
 - The time between data being generated and processed
 - Some applications, such as fraud detection are highly time-sensitive.

Big data: The 4th V - Veracity

- Introducing the notion of data **uncertainty**
 - Veracity refers to how reliable the data is
- Is the data correct?
- Is it out-of-date?
- Is it complete?
- Data cleansing helps greatly
 - Using techniques such as data fusion, stochastic models
 - But with the huge volumes of data being generated – errors will slip in

40 ZETTABYTES

[43 TRILLION GIGABYTES]
of data will be created by
2020, an increase of 300
times from 2005

6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION

Volume SCALE OF DATA

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



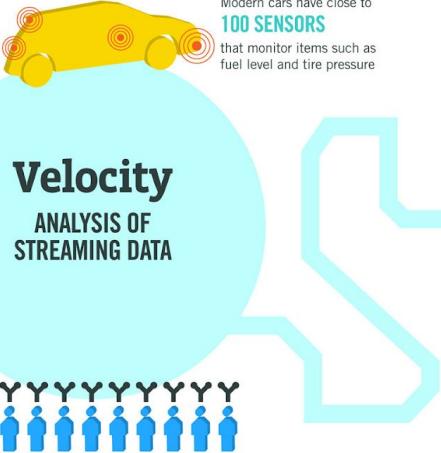
Most companies in the
U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION

during each trading session



Modern cars have close to
100 SENSORS
that monitor items such as
fuel level and tire pressure



Velocity ANALYSIS OF STREAMING DATA

By 2016, it is projected
there will be

**18.9 BILLION
NETWORK CONNECTIONS**

- almost 2.5 connections
per person on earth



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume**, **Velocity**, **Variety** and **Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of
data in healthcare was
estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



Variety DIFFERENT FORMS OF DATA

**30 BILLION
PIECES OF CONTENT**

are shared on Facebook
every month



Variety DIFFERENT FORMS OF DATA

By 2014, it's anticipated
there will be
**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**

4 BILLION+
HOURS OF VIDEO
are watched on
YouTube each month



400 MILLION TWEETS
are sent per day by about 200
million monthly active users

Poor data quality costs the US
economy around

\$3.1 TRILLION A YEAR



**1 IN 3 BUSINESS
LEADERS**

don't trust the information
they use to make decisions



**27% OF
RESPONDENTS**

in one survey were unsure of
how much of their data was
inaccurate

Veracity UNCERTAINTY OF DATA

Resources

Introduction to Big Data (O'Reilly) -

<http://orm-atlas2-prod.s3.amazonaws.com/pdf/e11376d1c19a651736042656f2aae705.pdf>

03 Describing Data

CA682

suzanne.little@dcu.ie

Today

Data Sources

What is metadata?

Metadata exercise

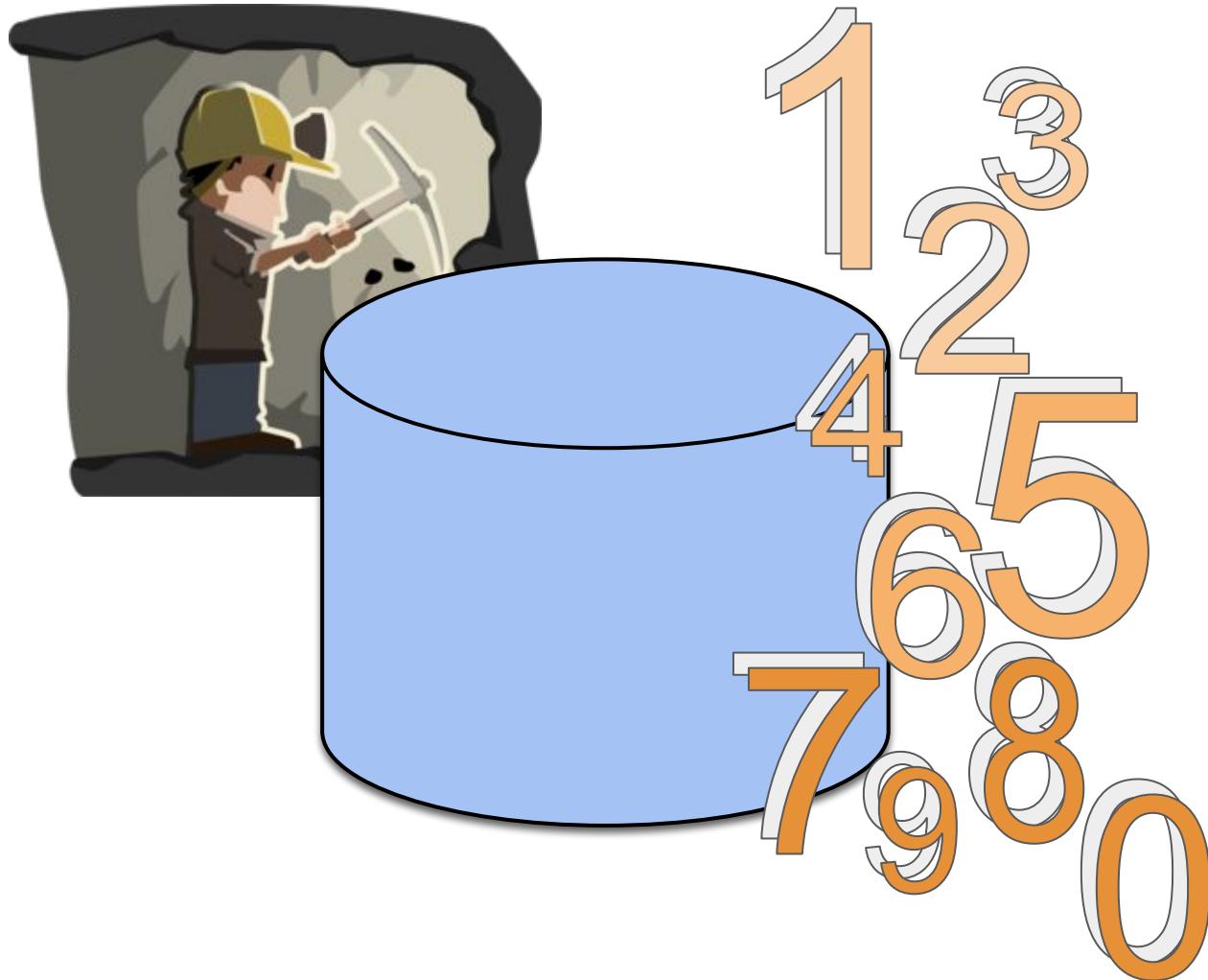
Assignment

Next week: Big data & an example



Data sources

- Files
- Databases
- “The Internet”
- Open Data



10 common data science file formats

1. CSV (& TSV & TXT)
2. JSON
3. XLS or XLSX
4. SQL
5. PDF
6. HTML
7. DOC or DOCX
8. HDF5
9. ZIP (or GZ or TGZ)
10. XML

How many do you know how to process?

Data sources: Databases

More to come!

- Traditional relational db: Oracle, MySQL, Postgres, etc.
 - Tables (“relations”) of rows and columns
 - Unique key per row
 - Links between rows (“foreign key”)
 - Optimise structures (the database schema)
 - Stored procedures (queries) to speed up responses
 - Most commonly use SQL - Structured Query Language
 - SELECT CustomerName,City FROM Customers;
 - SELECT CustomerName,Age FROM Customers WHERE City='Dublin';
- In memory databases: SAP Hana (<http://hana.sap.com/abouthana.html>)
- NoSQL, document, column, graph, etc.

Data sources: the Internet

- Crawlers or spiders
 - Scraping data from semi-structured sources
 - Parse HTML
 - Match Patterns to extract data
 - Identify links (repeat)
- URL
 - Files and databases on the web
 - Many libraries and apps will accept either a local path or url
- How many file formats?

*Exercise &
Information
on loop*

Open data

Public data, shared and freely available

Why?

Why not?

Examples of open data

Some examples of projects that use open data are:

- [Plantwise - Lose less, feed more](https://www.plantwise.org/) (<https://www.plantwise.org/>)
- [Humanitarian OpenStreetMap](https://www.hotosm.org/) (<https://www.hotosm.org/>)
- [OpenGLAM](https://openglam.org/) (<https://openglam.org/>)

Or on a less elevated topic ... the [Great British Public Toilet Map: open geospatial data!](#)

Where to find open data?

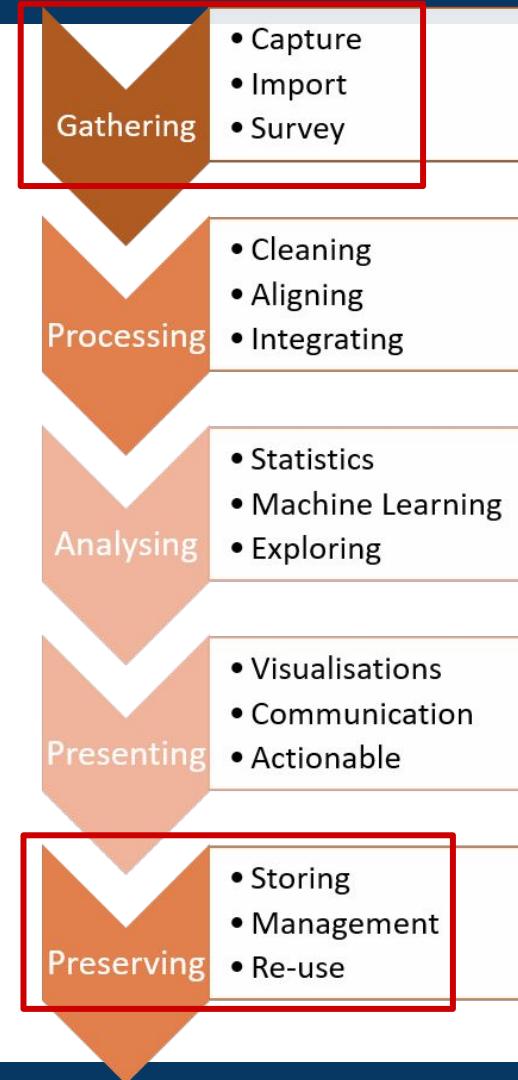
- <https://data.gov.ie/>
- <http://www.dublindashboard.ie/>
- <https://www.google.com/publicdata/directory>
- <https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/>

Also lots of datasets for learning data science:

- <https://www.kaggle.com/datasets>
- <https://github.com/datasets>

Data

- Data is collected information
- Where does data come from?
 - Files
 - The Internet
 - Databases



Metadata

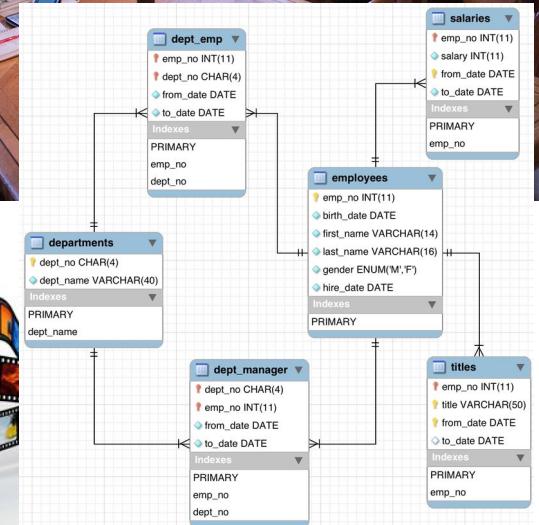
What is metadata?

“data about data”

“inferior type of cataloging”

Why is it useful?

*“Information on the organization of the data, data domains, and the relationship between them”
(Baeza-Yates)*



Metadata



Campo dei Miracoli
La Torre di Pisa
Field of Miracles, Pisa, Italy
Leaning Tower
Pisa, Italy
Sunny day in Pisa
My holidays
Old building

ExifVersion 0220
CompressedBitsPerPixel (5, 1)
DateTimeOriginal 2006:10:01 17:50:11
DateTimeDigitized 2006:10:01 17:50:11
MaxApertureValue (107, 32)
MeteringMode 5
Flash 80
FocalLength (7272, 1000)
ApertureValue (128, 32)
FocalPlaneXResolution (3264000, 286)
Make Canon
Model Canon PowerShot S80
Orientation 1
YCbCrPositioning 1
SensingMethod 2
XResolution (180, 1)
YResolution (180, 1)
ExposureTime (1, 640)
ColorSpace 1
FNumber (40, 10)
DateTime 2006:10:01 17:50:11
ExifImageWidth 3264
FocalPlaneYResolution (2448000, 214)
ExifImageHeight 2448

<https://readexifdata.com/>

Metadata

Why is it useful?

Find, Locate, Identify, Select,
Obtain, Navigate

Use data

Rights and data management

Doctorow on Meta-utopia

- People lie
- People are lazy
- People are stupid
- People delude themselves
- Schemas aren't neutral
- Metrics distort or limit
- There's more than one way!

<http://www.well.com/~doctorow/metacrap.htm>

Types of metadata

DESCRIPTIVE metadata

what the information object is about; inherently intrinsic properties

ADMINISTRATIVE metadata

who, what, why, where of the object's creation and management; inherently extrinsic properties

STRUCTURAL metadata

information about the structure, format, and composition of the thing being described; can be intrinsic or extrinsic

Metadata (Part 2)

5th Oct 2023

Exercise

Describe your favourite book or movie.

What qualities did you use? Title, Author, Year, Genre, Characters ?

How would you use the “metadata” to Find, Locate, Identify, Select, Obtain, Navigate

Exercise: Metadata about your favourite book

Title: Pride and Prejudice

Author: Jane Austen

Summary: "Pride and Prejudice is an 1813 novel of manners by English author Jane Austen. The novel follows the character development of Elizabeth Bennet, the protagonist of the book, who learns about the repercussions of hasty judgments and comes to appreciate the difference between superficial goodness and actual goodness."

Description: 1916 red hardback printing

Location: storage box 12

Metadata



Campo dei Miracoli
La Torre di Pisa
Field of Miracles, Pisa, Italy
Leaning Tower
Pisa, Italy
Sunny day in Pisa
My holidays
Old building

ExifVersion 0220
CompressedBitsPerPixel (5, 1)
DateTimeOriginal 2006:10:01 17:50:11
DateTimeDigitized 2006:10:01 17:50:11
MaxApertureValue (107, 32)
MeteringMode 5
Flash 80
FocalLength (7272, 1000)
ApertureValue (128, 32)
FocalPlaneXResolution (3264000, 286)
Make Canon
Model Canon PowerShot S80
Orientation 1
YCbCrPositioning 1
SensingMethod 2
XResolution (180, 1)
YResolution (180, 1)
ExposureTime (1, 640)
ColorSpace 1
FNumber (40, 10)
DateTime 2006:10:01 17:50:11
ExifImageWidth 3264
FocalPlaneYResolution (2448000, 214)
ExifImageHeight 2448

What is being described?

Two separate dimensions the metadata can be associated with:

- Abstraction hierarchy
- Granularity

An example abstraction hierarchy

WORK - an abstract entity; the distinct intellectual or artistic creation; it has no single material manifestation

EXPRESSION - the multiple realizations of a work in some particular medium or notation, where it can actually be perceived

MANIFESTATION - each of the formats of an expression that have the same appearance; but not necessarily the same implementation

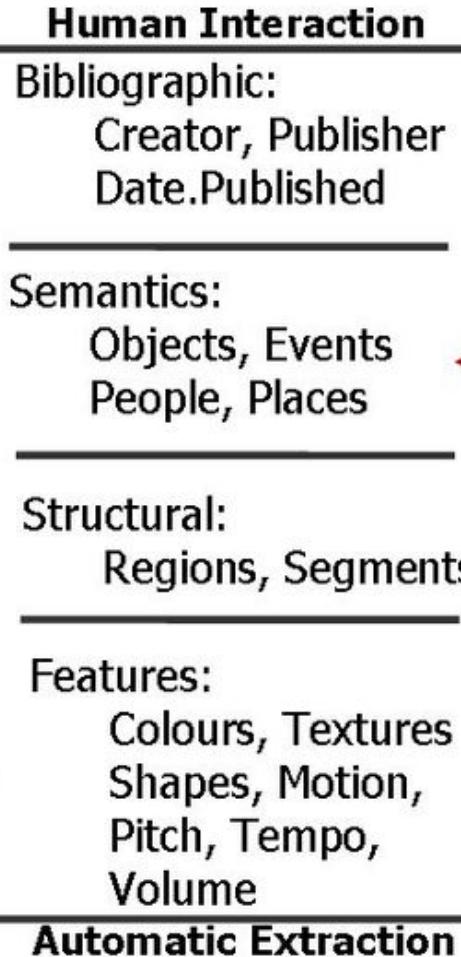
ITEM - a single exemplar of a manifestation; if we distinguish this level it is because otherwise identical manifestations have some differentiation

Metadata granularity

Heterogeneous
Complex Multimedia
Data

- video
- images
- graphs
- SMIL files
- web pages

semantic gap



**Semantic search and retrieval e.g.
"Give me high porosity fuel cell images"**

So how much metadata do we need?

Consider the **tradeoffs** between organization (adding, duplicate detection, storage) and retrieval (query, search)

Not all documents / resources need the same amount of metadata

Could someone else understand and use your dataset?

Where does metadata come from?

Simple

EXIF, document headers, time stamps, “ad hoc” labels

Structured

Adhering to a standard

Professional

Created by a librarian or curator

Crowd Sourced

hashtags, comments

Metadata standards

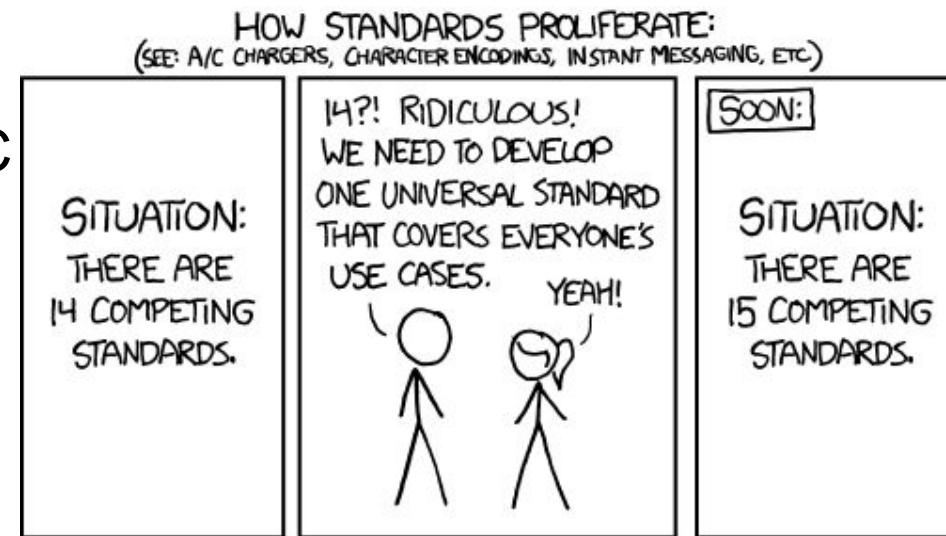
Specify the structure - Syntax or Content?

Many exist, not always compatible

MARC, Dublin Core, MPEG-7

Publishers: ISO, RFC, IEEE, W3C

<https://xkcd.com/927/>



Example: Dublin Core

Proposed in 1995 as standard set of metadata elements, simple enough to be supplied by document's author rather than professional curator

DC is the set of elements, described abstractly and all optional

Semantics of DC established by international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields

Specifications of how to use it in numerous syntaxes (especially XML and RDF) and languages

Dublin Core

TITLE

IDENTIFIER

SUBJECT

CREATOR - makes the content

CONTRIBUTOR

PUBLISHER

DATE

DESCRIPTION

LANGUAGE

TYPE - nature or genre

RIGHTS

SOURCE - if derived from something

RELATION

COVERAGE

AUDIENCE

Go back to your movie or book description. Can you see what a Dublin Core description would look like?

<http://www.dublincore.org/documents/2000/07/16/usageguide/generic/>

Problems

“Some information may appear to belong in more than one metadata element”

“There is potential semantic overlap between some elements”

“There will occasionally be some judgment required from the person assigning the metadata”

<https://catalog.data.gov/dataset/consumer-complaint-database>

Today

Open data exercise

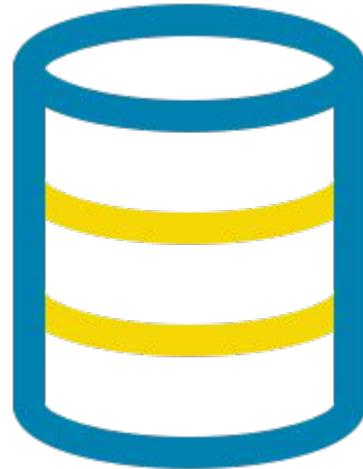
What is metadata?

Metadata exercise

<break>

Assignment

Big data & an example



Visualisation Assignment - loop has specification

Assignment due Friday Dec 1st 23:59

In pairs, create a visualisation that **illustrates a point, answers a question or tells a story (explanatory)**.

Short report (following the template) plus screencast video showing and describing your visualisation

A simple chart on limited data is not sufficient. Remember it is worth 25%.
Marking criteria are included in the description.

Use any tool or tools but remember you want to demonstrate your skill.

Assignment Marking Criteria

1. Dataset [30%]:
 - a. “big” data
 - b. showing either data cleaning **or** transformation **or** integration.
2. Visualisation [50%]:
 - a. suitable graph choice;
 - b. difficulty level;
 - c. good design/style;
 - d. use of interactivity **or** animation.
3. Report [20%]:
 - a. follows instructions and template;
 - b. good abstract;
 - c. critique and reflection.

Assignment - good things to know

Present 1 explanatory graph. No dashboards, please.

Keep your report concise, following the template, and your video screencast brief.

Not allowed to use the MovieLens, Chicago Crime, New York Traffic or derivative datasets.

Critique & reflect is very important.

Objective: Assessing your data processing, graph *selection* and *design* skills.

Planning your visualisation assignment **[OPTIONAL]**

Link in Loop

Submit a brief description (500 words) of your idea.

What data? What question or idea? What tools?

I'll try to give individual feedback but may end up summarising for the class.

Submit before Nov 8th (so I can comment in class)

Today's labs

Exercises available on loop for Open Data, Linked Data, Jupyter Notebooks, reading data using pandas library and Scraping data from the web. Review the “Documents”.

Jupyter/Pandas: There is a video introduction on loop or I will be in LG25 at the start to give an introduction & again next week in LG26

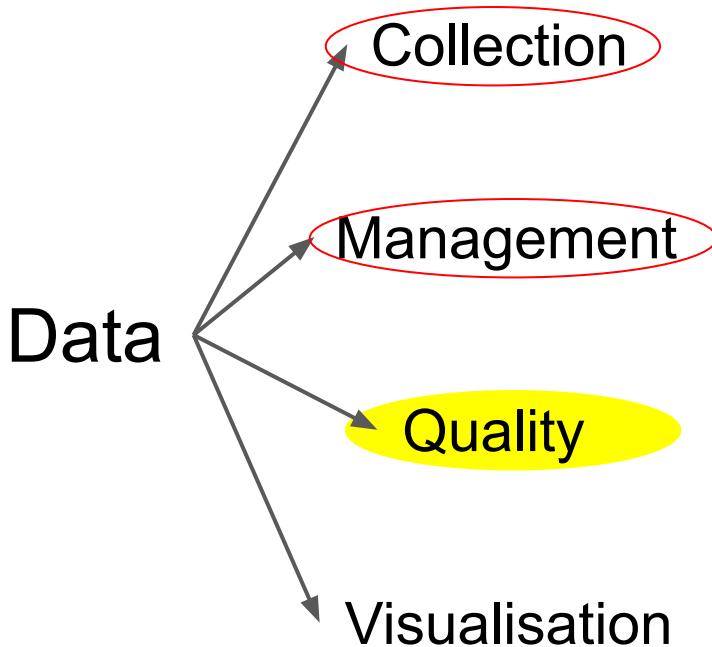
These exercises should cover this week and next week.

If you don't have python background then I suggest datacamp

05 Data Quality & Cleaning

suzanne.little@dcu.ie

Recap



Outcomes

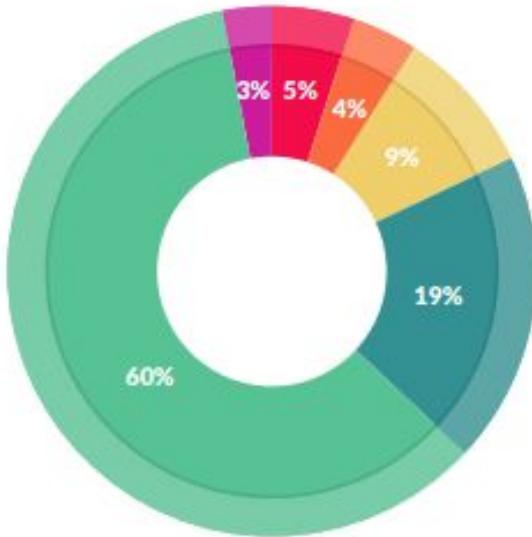
1. Analyse the requirements of applications handling large datasets.
2. Demonstrate an ability to efficiently structure a large dataset.
3. Implement data quality measures.
4. Identify and implement appropriate data visualization techniques.

CA682: where are we?

- Introduction: A Data Analytics Pipeline
 - Formal Data Management Lifecycles
- Data Collection: what is data? where does it come from?
 - data from files (text or binary, open or proprietary)
 - data types (SvU, QvQ, DvC, NOIR)
- Big data: 4 Vs
- Open Data
- Metadata - what is it? what is it used for?

Today: Data Quality & Cleaning → Exercises: Spreadsheets, OpenRefine

Next: Data Visualisation



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Overview

- Defining “data quality”
- Examples and causes of “Bad Data”
- Measuring data quality
- Tools for data cleaning

Objectives:

1. Managing data projects
2. Cleaning data



Data Quality and Cleaning

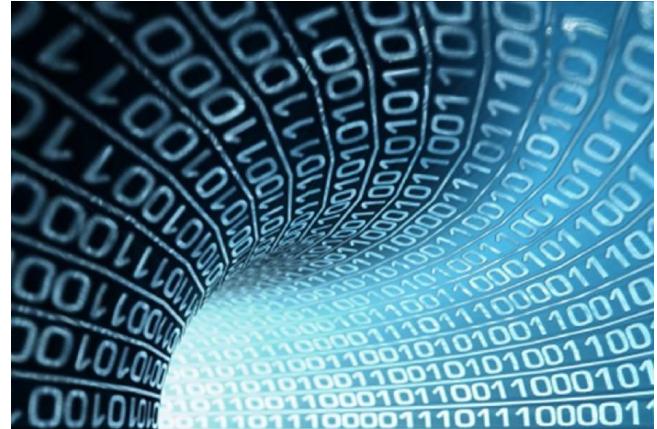
Computers are absolute (1 or 0)

People are complicated

Data generated by people is complicated and often messy

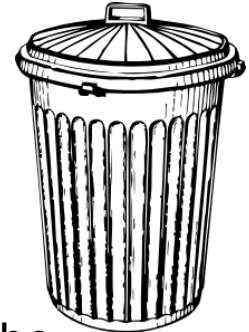
How do we **assess data quality?**

How can we **clean data** for storage and processing?



Data Quality

High quality data is free from both errors and artefacts.



Error: data that is missing or lost due to the capture process and cannot be recovered.

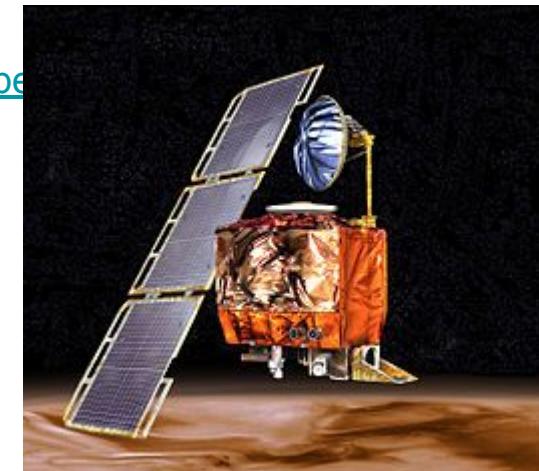
Artefact: something that has been introduced into the dataset during the gathering, processing, integration or cleaning activities.

Poor quality data may be due to **individual** (one off) or **collective** (systemic) issues.

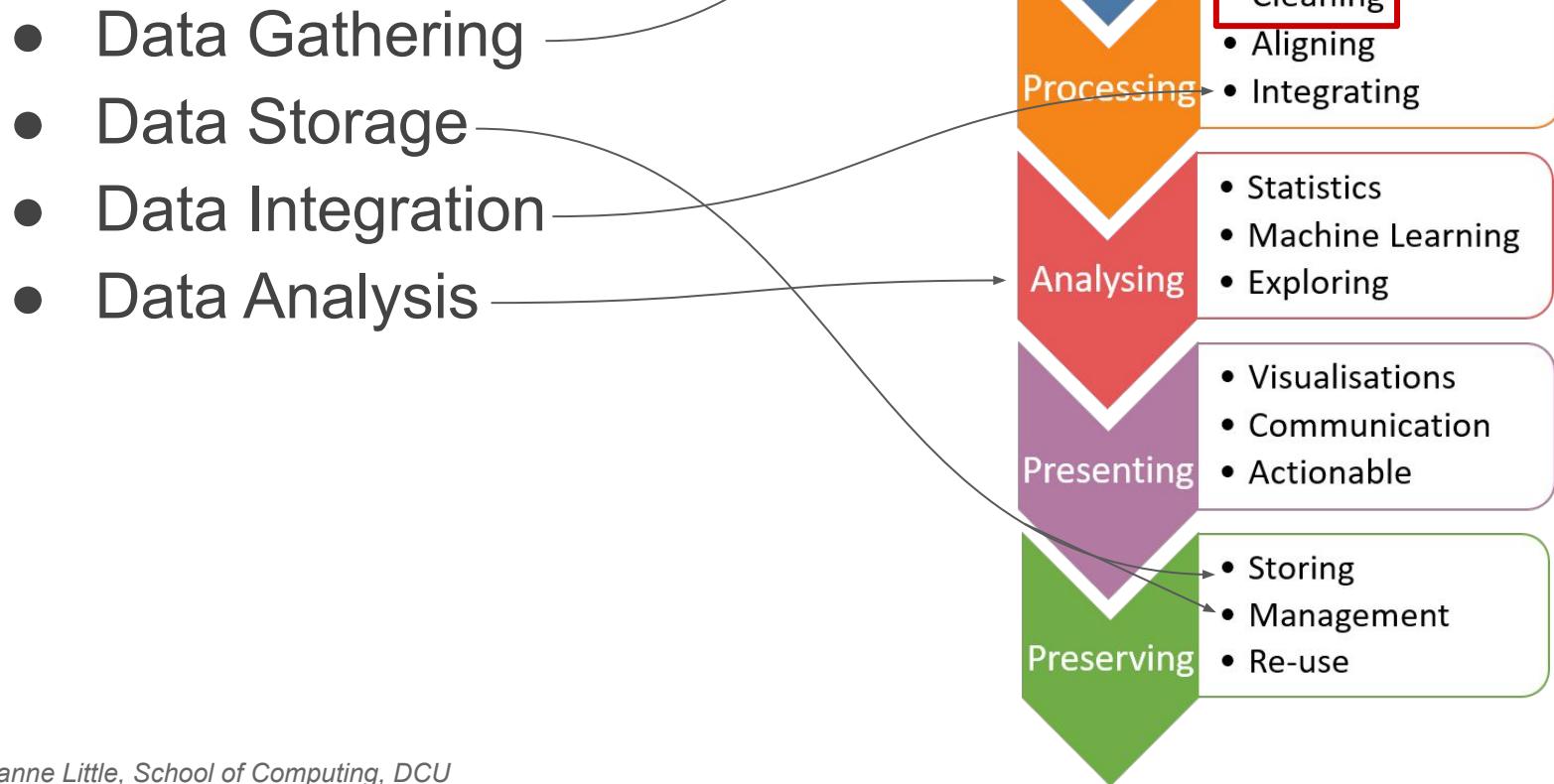
What happens when the data quality is poor?

Mars orbiter: In 1999, the \$125 million Mars climate orbiter was destroyed when incorrect units used by a contractor and NASA engineers (metric vs imperial) caused it to have the incorrect trajectory.

https://www.vice.com/en_us/article/qkvzb5/the-time-nasa-lost-a-mars-orbiter-because-of-a-unit-mixup



Where do problems occur?





- Capture
- Import
- Survey

Data Gathering

- Manual entry errors or artefacts caused by typos, lazy people etc.
 - Eg. DUC or DCU
 - Eg. duplicate entry, duplicate entry, duplicate entry
- Poor survey or interface design
 - Eg. What colour is your toothbrush? The only options are red, green or blue! What if it's multicoloured?
 - Eg. A drop down for entering your age only goes back to 1923. No one alive older than 100?
- No standards for format or controlled vocabulary for fields
 - Eg. DCU or Dublin City University?
 - Eg. centimeters or inches?



- Capture
- Import
- Survey

Data Gathering: solutions

Preemptive:

- build in integrity checks and entry constraints (process architecture)
- Process management - reward accurate human data entry, data sharing, data stewards, redundancy

Retrospective:

- Cleaning focus (duplicate removal, merge/purge, name & address matching, field value standardization)
- Diagnostic focus (automated detection of glitches)



Data Delivery

Destroying or mutilating information by inappropriate pre-processing

- Inappropriate aggregation
- Nulls converted to default values

Loss of data:

- Buffer overflows
- Transmission problems (network overload)
- No checks

→ [“Understanding packet loss for sound monitoring in a smart stadium IoT testbed”](#)



Data Delivery: solutions

- Build reliable transmission protocols
 - ◆ Use a relay server
- Verification
 - ◆ Checksums, verification parser
 - ◆ Do the uploaded files fit an expected pattern?
- Relationships
 - ◆ Are there dependencies between data streams and processing steps
- Interface agreements
 - ◆ Data quality commitment from the data stream supplier.

Data Storage

- Format conversion errors
 - string or float?
 - 1918 or 2018?
 - rounding or approximation (e.g., database field limited to integers)
- No metadata recorded
 - What does the field mean?

Also possible to have technical issues

- Transmission errors (network dropout)
- Disk failure or corruption



- Document and publish
- Data exploration and retrospective checking
- Assume that the worst might happen!





Data Integration

Combine data sets (acquisitions, across departments, organisations)

Common source of problems

- Heterogenous data : no common key, different field formats, approximate matching
- Different definitions : What is a customer, an account, a family, ...
- Time synchronization : Does the data relate to the same time periods? Are the time windows compatible?
- Legacy data : IMS, spreadsheets, ad-hoc structures, binary data
- Sociological factors : Reluctance to share – loss of power



Data Integration: solutions

- Commercial Tools
 - ◆ Significant body of research in data integration
 - ◆ Many tools for address matching, schema mapping are available.
- Data browsing and exploration
 - ◆ Many hidden problems and meanings : must extract metadata.
 - ◆ View before and after results : did the integration go the way you thought?
- Google Fusion Tables, Spreadsheets, Software Libraries

Data Retrieval



- Capture
- Import
- Survey

Exported data sets are often a *view* of the actual data. Problems occur because:

- Source data not properly understood
- Need for derived data not understood
- Just plain mistakes
 - Inner join vs. outer join
 - Understanding NULL values
- Computational constraints
 - E.g., too expensive to give a full history, we'll supply a snapshot.



Data Mining and Analysis

What are you doing with all this data anyway?

Problems in the analysis

- Scale and performance
- Confidence bounds? 0.95, 0.99?
- Attachment to models
- Insufficient domain expertise
- Casual empiricism (use of an arbitrary number to support a pre-conception)
 - 85% of all statistics are made up on the spot!

Conan Doyle: “I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”
(Sherlock Holmes)



Data Analysis: solutions

→ Data exploration

- ◆ Determine which model does what

Engage your brain!
(smell test or commonsense)

→ the analysis part of the feedback loop

Questions & Discussion



Untitled spre...

**Share**

File Edit View Insert Format D



100%



.00



Default (Ari...



A12



1 0.595773514

2 0.190379357

3 0.137266402

4 0.810291551

5 0.72615007

6 0.749070919

7 0.996801633

8 0.229031203

9 0.942622668

10 0.660208972

11 #DIV/0!

12

13

14

Error

Evaluation of function
MOYENNE caused a divide by
zero error.



testsml



Overview

- Defining “data quality”
- Examples and causes of “Bad Data”
- Measuring data quality
- Tools for data cleaning



Conventional measures of data quality

Accuracy : The data was recorded correctly.

Completeness : All relevant data was recorded.

Uniqueness : Entities are recorded once.

Timeliness : The data is recent or kept up to date.
Date published vs Data captured ...

Consistency : The data agrees with itself (internal).

Credibility : The data comes from a recognised (or official) source.



Problems with conventional measures

Unmeasurable: Accuracy and completeness are extremely difficult, perhaps impossible to measure.

Context independent: No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.

Incomplete: What about interpretability, accessibility, metadata, analysis, etc.

Vague: The conventional definitions provide no guidance towards practical improvements of the data.

So what do you do to measure data quality?

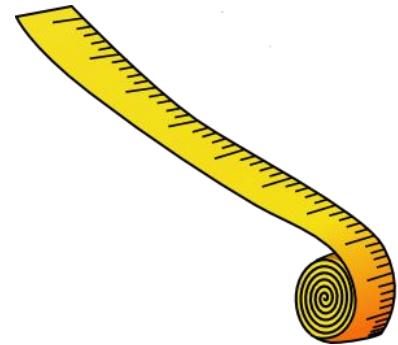
Some options ...

- Inventory (expensive)
- Using a proxy measure such as tracking customer complaints
- Applying formal measures of accessibility
- Using test cases with known results and checking for glitches or errors in analysis
- Successfully completing an end-to-end process (e.g., data ingestion, processing, indexing, querying and summarisation)

Data quality constraints

- Many data quality problems can be captured by **static constraints** based on the schema.
 - Nulls not allowed, field domains, foreign key constraints, etc.
- Many others are due to **problems in workflow**, and can be captured by **dynamic constraints**
 - E.g., orders above \$200 are processed by Biller 2
- The constraints follow an 80-20 rule
 - A few constraints capture most cases, thousands of constraints to capture the last few cases.
- Adherence to constraints are **measurable**.

Data quality metrics



- Measure to improve → incentivise
 - Indicates what is wrong and how to improve
 - Realize that DQ is a messy problem, no set of numbers will be perfect
- Types of metrics
 - Static vs. dynamic constraints
 - Operational vs. diagnostic
- A very large number metrics are possible → choose the most important ones
- Warning: Metrics can give incentives for bad behavior → throw away data that doesn't join.

Methods for data cleaning

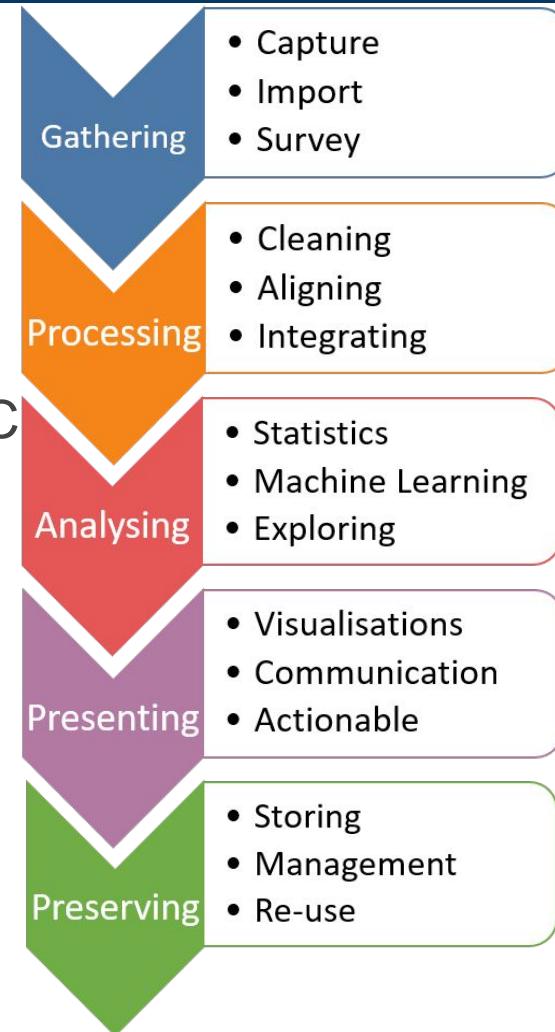
1. Implement process mandates (fix the human problem)
 - a. Including schema or rule restrictions to enforce format
2. Custom tools written in a General Purpose language (hack a script)
 - a. Good for one-off quick fixes (cleaning that only happens once)
3. Off the shelf tools such as Spreadsheets, OpenRefine (formerly Google Refine)
 - a. Form of exploratory data analysis and cleansing



Using something like Jupyter or R-studio is a mix of 2 & 3 as you can interactively explore and refine but also document your quick fix for future use.

Application

- A: House price census missing Collins Ave
 - B: Forex data for EUR-GBP missing May 2016
 - C: Temperature values entered in °F rather than °C
 - D: Entries overwritten when merging company records
1. Which phase or task is the likely cause?
 2. Is it an error or an artefact?
 3. What solutions could you use?
 4. What might be possible consequences?



More examples

- Sound monitoring packets lost when network overloaded
- See also <http://okfnlabs.org/bad-data/> for a few examples
- “Excel: Why using Microsoft's tool caused Covid-19 results to be lost”
- Mis-calculation examples (not always “bad” data),
<https://www.bbc.com/news/magazine-27509559>

Practical suggestions for initial data quality checks

- Missing values, records or variables - are empty cells no value (0) or no measurement (null)? How should they be handled?
- Erroneous values - typos or values that are clearly out of place (gender value in age column)
- Inconsistencies - capitalisation, units of measurement
- Duplicate records
- Out of date - e.g., age will have changed
- Leading or trailing spaces! Windows or Linux end of line characters
- Format of dates - DD/MM/YYYY, MM/DD/YYYY, ?? Excel based or Unix based
- “Sanity checks” - look for extreme values or outliers, count how many records

Tools for Data Cleaning

Many options

Spreadsheets (Google, Excel, etc)

Purpose built tools (RapidMiner, TableauPrep, OpenRefine)

General Purpose Language (Python, R, etc)

Some things to consider:

- How big is your data?
- How frequently are you cleaning? Once off or regularly?
- How will you document your work?

OpenRefine

- Until late 2012, Google Refine, now no longer maintained by Google
- Offline (desktop) tool that runs in the browser
- OpenRefine can:
 - Import/export a range of data types
 - Explore data (use graphs to check distributions and outliers)
 - Clean and transform data
 - Fix errors in fields
 - Use heuristics to group data and spot errors
 - Call services to enhance your data (e.g. Geocoding)
 - What OpenRefine calls *reconciling*
 - Handle large (-ish) datasets
- <http://openrefine.org/>

Summary

Poor quality data has big consequences

If you can, control for quality in gathering and integration before you do analysis!

Build a “tool kit” of programs, methods and questions to satisfy yourself that your data is clean before you analyse or visualise.

Develop your sense of when, where and how problems might happen
(How can you do this? Read the references, think about the causes)

Next steps

Document linked from loop has detailed summary and other links

Suggest doing the European Data Portal Course on Data Cleaning, linked from loop

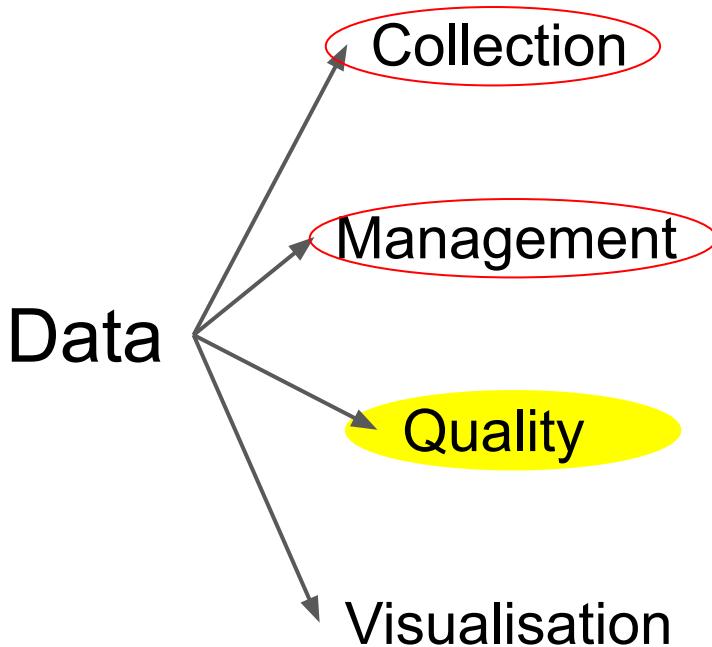
Three sets of exercises to perform data cleaning:

1. Google Sheets
2. OpenRefine
3. Python/Pandas in notebooks

05 Data Quality & Cleaning

suzanne.little@dcu.ie

Recap



Outcomes

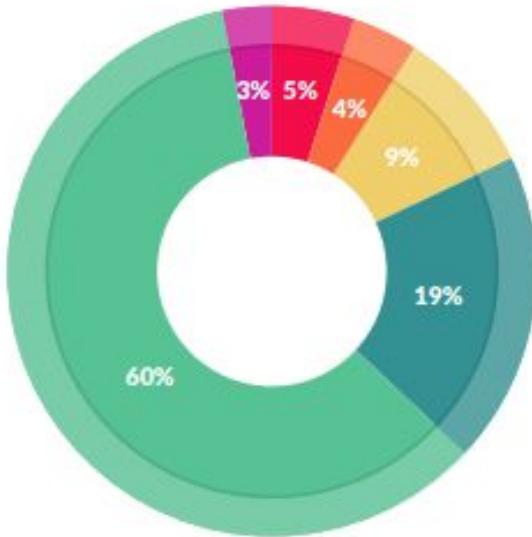
1. Analyse the requirements of applications handling large datasets.
2. Demonstrate an ability to efficiently structure a large dataset.
3. Implement data quality measures.
4. Identify and implement appropriate data visualization techniques.

CA682: where are we?

- Introduction: A Data Analytics Pipeline
 - Formal Data Management Lifecycles
- Data Collection: what is data? where does it come from?
 - data from files (text or binary, open or proprietary)
 - data types (SvU, QvQ, DvC, NOIR)
- Big data: 4 Vs
- Open Data
- Metadata - what is it? what is it used for?

Today: Data Quality & Cleaning → Exercises: Spreadsheets, OpenRefine

Next: Data Visualisation



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Overview

- Defining “data quality”
- Examples and causes of “Bad Data”
- Measuring data quality
- Tools for data cleaning

Objectives:

1. Managing data projects
2. Cleaning data



Data Quality and Cleaning

Computers are absolute (1 or 0)

People are complicated

Data generated by people is complicated and often messy

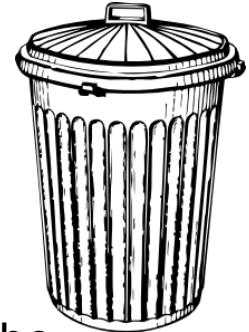
How do we **assess data quality?**

How can we **clean data** for storage and processing?



Data Quality

High quality data is free from both errors and artefacts.



Error: data that is missing or lost due to the capture process and cannot be recovered.

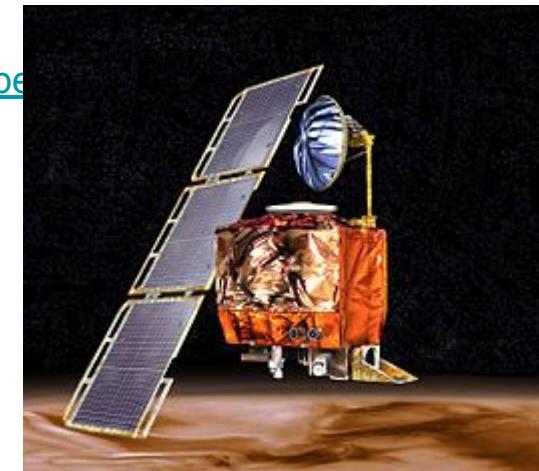
Artefact: something that has been introduced into the dataset during the gathering, processing, integration or cleaning activities.

Poor quality data may be due to **individual** (one off) or **collective** (systemic) issues.

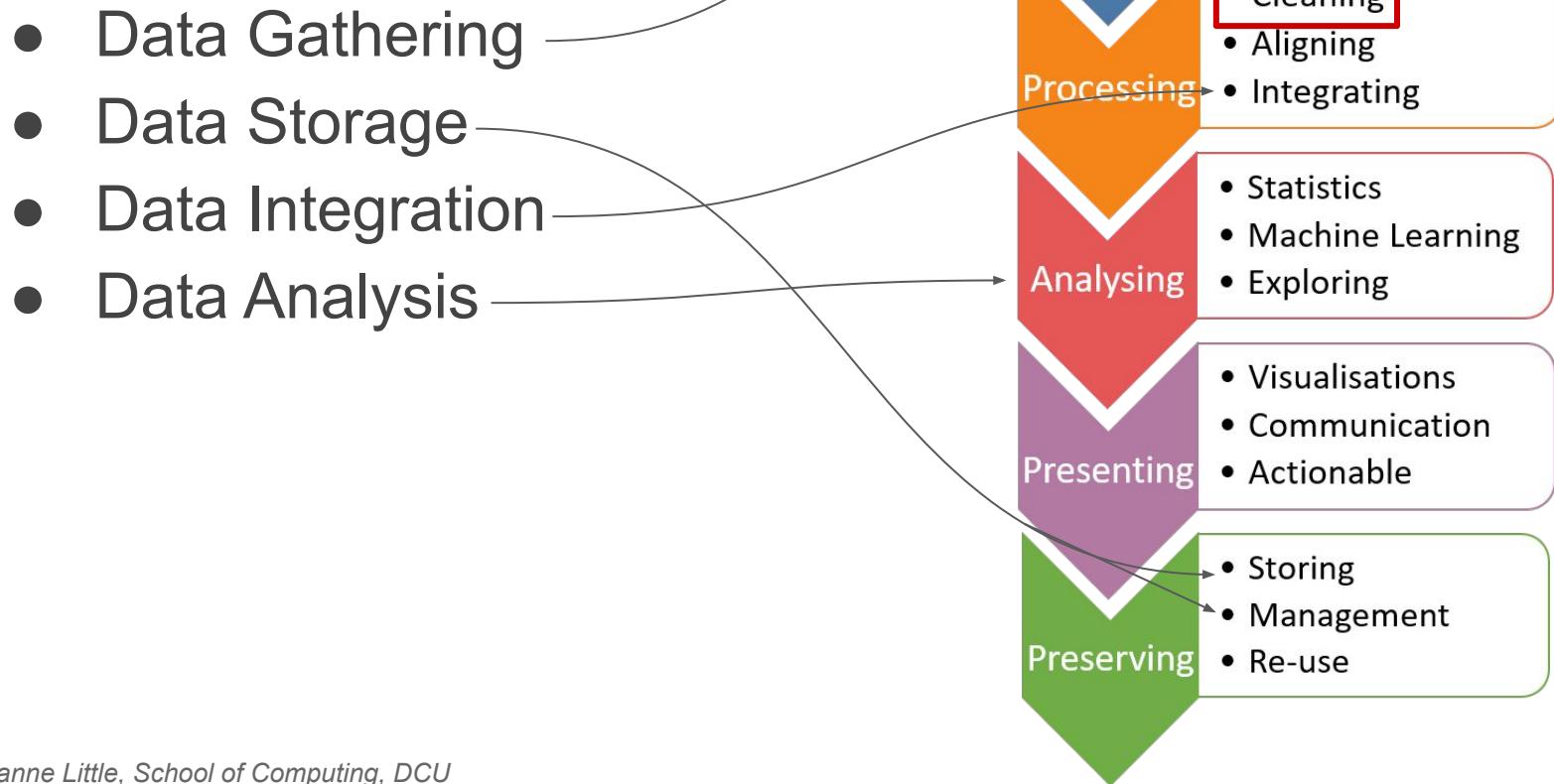
What happens when the data quality is poor?

Mars orbiter: In 1999, the \$125 million Mars climate orbiter was destroyed when incorrect units used by a contractor and NASA engineers (metric vs imperial) caused it to have the incorrect trajectory.

https://www.vice.com/en_us/article/qkvzb5/the-time-nasa-lost-a-mars-orbiter-because-of-a-unit-mixup



Where do problems occur?





- Capture
- Import
- Survey

Data Gathering

- Manual entry errors or artefacts caused by typos, lazy people etc.
 - Eg. DUC or DCU
 - Eg. duplicate entry, duplicate entry, duplicate entry
- Poor survey or interface design
 - Eg. What colour is your toothbrush? The only options are red, green or blue! What if it's multicoloured?
 - Eg. A drop down for entering your age only goes back to 1923. No one alive older than 100?
- No standards for format or controlled vocabulary for fields
 - Eg. DCU or Dublin City University?
 - Eg. centimeters or inches?



- Capture
- Import
- Survey

Data Gathering: solutions

Preemptive:

- build in integrity checks and entry constraints (process architecture)
- Process management - reward accurate human data entry, data sharing, data stewards, redundancy

Retrospective:

- Cleaning focus (duplicate removal, merge/purge, name & address matching, field value standardization)
- Diagnostic focus (automated detection of glitches)



Data Delivery

Destroying or mutilating information by inappropriate pre-processing

- Inappropriate aggregation
- Nulls converted to default values

Loss of data:

- Buffer overflows
- Transmission problems (network overload)
- No checks

→ [“Understanding packet loss for sound monitoring in a smart stadium IoT testbed”](#)



Data Delivery: solutions

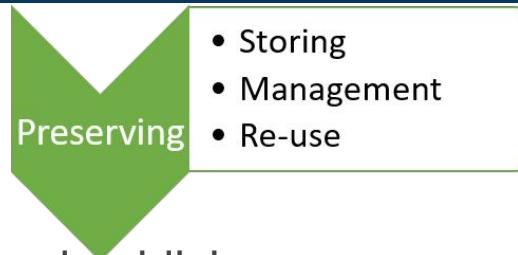
- Build reliable transmission protocols
 - ◆ Use a relay server
- Verification
 - ◆ Checksums, verification parser
 - ◆ Do the uploaded files fit an expected pattern?
- Relationships
 - ◆ Are there dependencies between data streams and processing steps
- Interface agreements
 - ◆ Data quality commitment from the data stream supplier.

Data Storage

- Format conversion errors
 - string or float?
 - 1918 or 2018?
 - rounding or approximation (e.g., database field limited to integers)
- No metadata recorded
 - What does the field mean?

Also possible to have technical issues

- Transmission errors (network dropout)
- Disk failure or corruption



- Document and publish
- Data exploration and retrospective checking
- Assume that the worst might happen!





Data Integration

Combine data sets (acquisitions, across departments, organisations)

Common source of problems

- Heterogenous data : no common key, different field formats, approximate matching
- Different definitions : What is a customer, an account, a family, ...
- Time synchronization : Does the data relate to the same time periods? Are the time windows compatible?
- Legacy data : IMS, spreadsheets, ad-hoc structures, binary data
- Sociological factors : Reluctance to share – loss of power



Data Integration: solutions

- Commercial Tools
 - ◆ Significant body of research in data integration
 - ◆ Many tools for address matching, schema mapping are available.
- Data browsing and exploration
 - ◆ Many hidden problems and meanings : must extract metadata.
 - ◆ View before and after results : did the integration go the way you thought?
- Google Fusion Tables, Spreadsheets, Software Libraries

Data Retrieval



- Capture
- Import
- Survey

Exported data sets are often a *view* of the actual data. Problems occur because:

- Source data not properly understood
- Need for derived data not understood
- Just plain mistakes
 - Inner join vs. outer join
 - Understanding NULL values
- Computational constraints
 - E.g., too expensive to give a full history, we'll supply a snapshot.



Data Mining and Analysis

What are you doing with all this data anyway?

Problems in the analysis

- Scale and performance
- Confidence bounds? 0.95, 0.99?
- Attachment to models
- Insufficient domain expertise
- Casual empiricism (use of an arbitrary number to support a pre-conception)
 - 85% of all statistics are made up on the spot!

Conan Doyle: “I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”
(Sherlock Holmes)



Data Analysis: solutions

→ Data exploration

- ◆ Determine which model is best

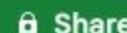
Engage your brain!
(smell test or commonsense)

→ the analysis part of the feedback loop

Questions & Discussion



Untitled spre...



File Edit View Insert Format D



100%



.00



123 Default (Ari...



10



A12



	A	B	C	D	E	F
1	0.595773514					
2	0.190379357					
3	0.137266402					
4	0.810291551					
5	0.72615007					
6	0.749070919					
7	0.996801633					
8	0.229031203					
9	0.942622668					
10	0.660208972					
11	#DIV/0!					
12						
13						
14						

Error

Evaluation of function
MOYENNE caused a divide by
zero error.



testsml



Overview

- Defining “data quality”
- Examples and causes of “Bad Data”
- Measuring data quality
- Tools for data cleaning



Conventional measures of data quality

Accuracy : The data was recorded correctly.

Completeness : All relevant data was recorded.

Uniqueness : Entities are recorded once.

Timeliness : The data is recent or kept up to date.
Date published vs Data captured ...

Consistency : The data agrees with itself (internal).

Credibility : The data comes from a recognised (or official) source.



Problems with conventional measures

Unmeasurable: Accuracy and completeness are extremely difficult, perhaps impossible to measure.

Context independent: No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.

Incomplete: What about interpretability, accessibility, metadata, analysis, etc.

Vague: The conventional definitions provide no guidance towards practical improvements of the data.

So what do you do to measure data quality?

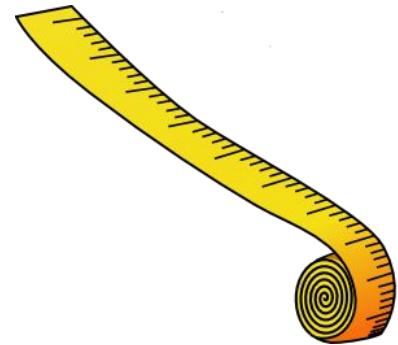
Some options ...

- Inventory (expensive)
- Using a proxy measure such as tracking customer complaints
- Applying formal measures of accessibility
- Using test cases with known results and checking for glitches or errors in analysis
- Successfully completing an end-to-end process (e.g., data ingestion, processing, indexing, querying and summarisation)

Data quality constraints

- Many data quality problems can be captured by **static constraints** based on the schema.
 - Nulls not allowed, field domains, foreign key constraints, etc.
- Many others are due to **problems in workflow**, and can be captured by **dynamic constraints**
 - E.g., orders above \$200 are processed by Biller 2
- The constraints follow an 80-20 rule
 - A few constraints capture most cases, thousands of constraints to capture the last few cases.
- Adherence to constraints are **measurable**.

Data quality metrics



- Measure to improve → incentivise
 - Indicates what is wrong and how to improve
 - Realize that DQ is a messy problem, no set of numbers will be perfect
- Types of metrics
 - Static vs. dynamic constraints
 - Operational vs. diagnostic
- A very large number metrics are possible → choose the most important ones
- Warning: Metrics can give incentives for bad behavior → throw away data that doesn't join.

Methods for data cleaning

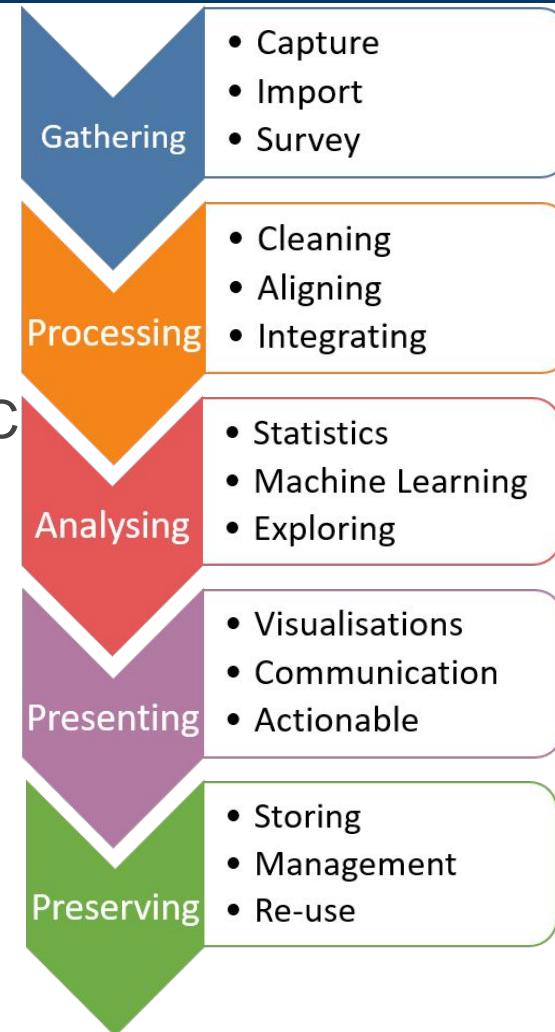
1. Implement process mandates (fix the human problem)
 - a. Including schema or rule restrictions to enforce format
2. Custom tools written in a General Purpose language (hack a script)
 - a. Good for one-off quick fixes (cleaning that only happens once)
3. Off the shelf tools such as Spreadsheets, OpenRefine (formerly Google Refine)
 - a. Form of exploratory data analysis and cleansing



Using something like Jupyter or R-studio is a mix of 2 & 3 as you can interactively explore and refine but also document your quick fix for future use.

Application

- A: House price census missing Collins Ave
 - B: Forex data for EUR-GBP missing May 2016
 - C: Temperature values entered in °F rather than °C
 - D: Entries overwritten when merging company records
1. Which phase or task is the likely cause?
 2. Is it an error or an artefact?
 3. What solutions could you use?
 4. What might be possible consequences?



More examples

- Sound monitoring packets lost when network overloaded
- See also <http://okfnlabs.org/bad-data/> for a few examples
- “Excel: Why using Microsoft's tool caused Covid-19 results to be lost”
- Mis-calculation examples (not always “bad” data),
<https://www.bbc.com/news/magazine-27509559>

Practical suggestions for initial data quality checks

- Missing values, records or variables - are empty cells no value (0) or no measurement (null)? How should they be handled?
- Erroneous values - typos or values that are clearly out of place (gender value in age column)
- Inconsistencies - capitalisation, units of measurement
- Duplicate records
- Out of date - e.g., age will have changed
- Leading or trailing spaces! Windows or Linux end of line characters
- Format of dates - DD/MM/YYYY, MM/DD/YYYY, ?? Excel based or Unix based
- “Sanity checks” - look for extreme values or outliers, count how many records

Tools for Data Cleaning

Many options

Spreadsheets (Google, Excel, etc)

Purpose built tools (RapidMiner, TableauPrep, OpenRefine)

General Purpose Language (Python, R, etc)

Some things to consider:

- How big is your data?
- How frequently are you cleaning? Once off or regularly?
- How will you document your work?

OpenRefine

- Until late 2012, Google Refine, now no longer maintained by Google
- Offline (desktop) tool that runs in the browser
- OpenRefine can:
 - Import/export a range of data types
 - Explore data (use graphs to check distributions and outliers)
 - Clean and transform data
 - Fix errors in fields
 - Use heuristics to group data and spot errors
 - Call services to enhance your data (e.g. Geocoding)
 - What OpenRefine calls *reconciling*
 - Handle large (-ish) datasets
- <http://openrefine.org/>

Summary

Poor quality data has big consequences

If you can, control for quality in gathering and integration before you do analysis!

Build a “tool kit” of programs, methods and questions to satisfy yourself that your data is clean before you analyse or visualise.

Develop your sense of when, where and how problems might happen
(How can you do this? Read the references, think about the causes)

Next steps

Document linked from loop has detailed summary and other links

Suggest doing the European Data Portal Course on Data Cleaning, linked from loop

Three sets of exercises to perform data cleaning:

1. Google Sheets
2. OpenRefine
3. Python/Pandas in notebooks

06 Data Visualisation *Communication*

CA682
suzanne.little@dcu.ie

CA682: Status and coming up next

So far: generic data process, sources & formats, describing data, open data, big data, metadata, data quality & cleaning

Today: Data Visualisation → Communication

Next:

- Data Visualisation: Choosing Charts
- Data Visualisation: Good design

Today: Data Visualisation

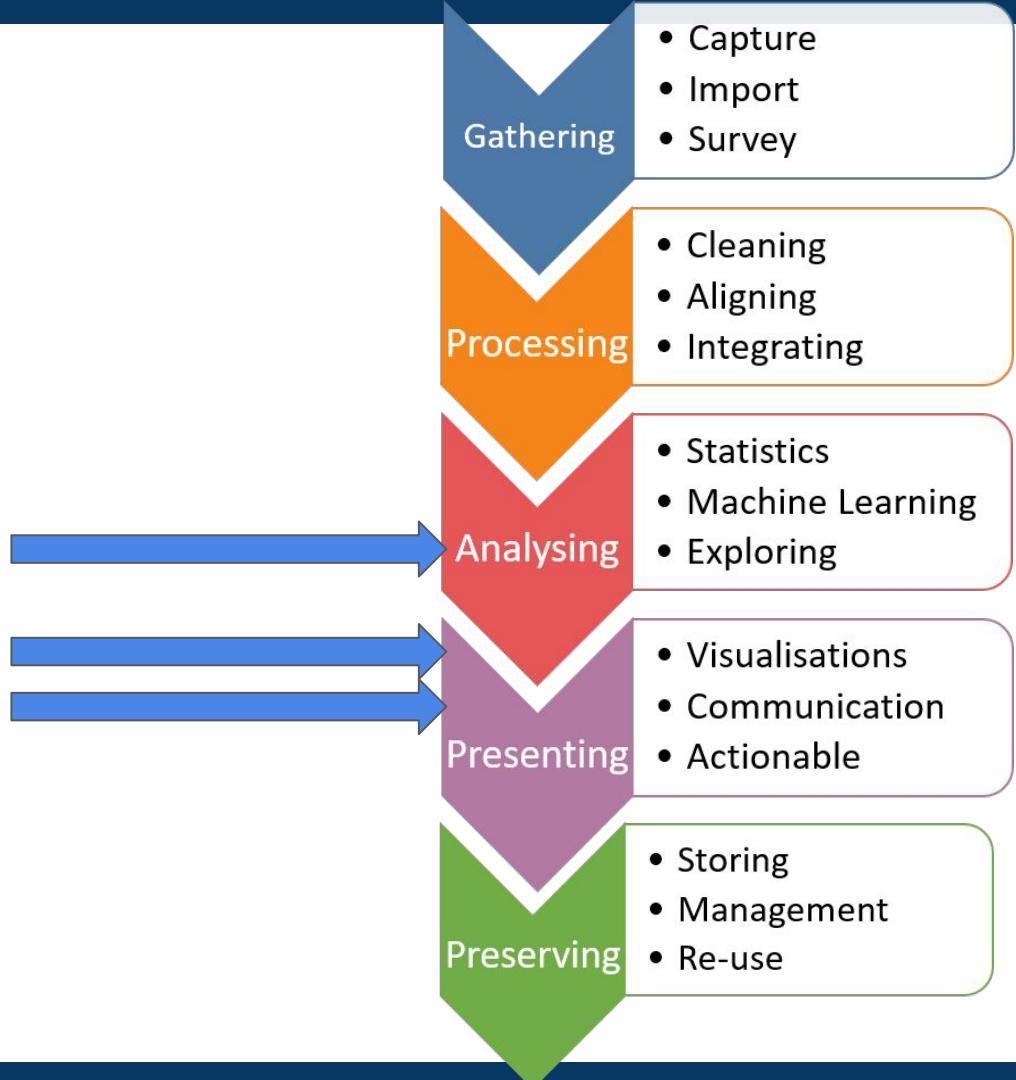
- Communication
- Motivations for visualisations
- Analysing communication
- “Data Visualisation: Good things to know”

Data Visualisation

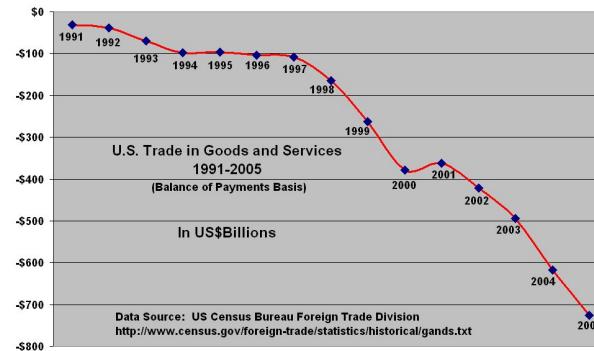
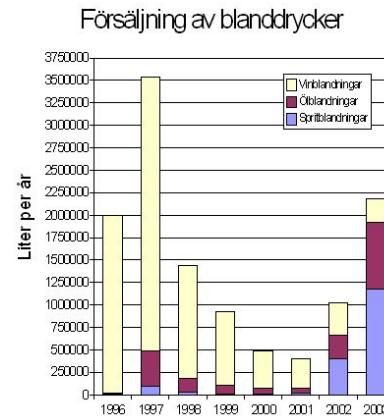
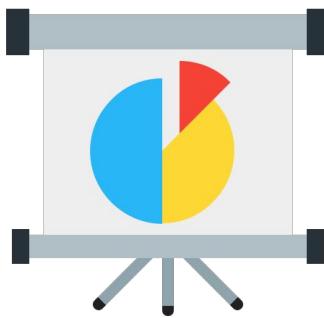
Why visualise?

Analysis tool: Exploration

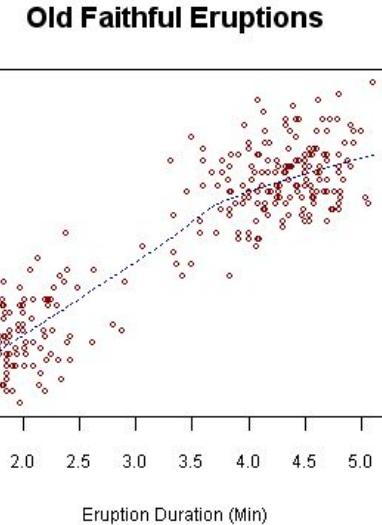
Communication
(Explanation)



What do I mean by “data visualisation”?



Waiting Time Between Eruptions (Min)



But also diagrams, maps, infographics, dashboards, tables etc. that are based on data.

Consider ...

“... data visualization is not an exact science. There is rarely, if ever, a single right answer or single best solution. It is much more about using heuristic methods to determine the most satisfactory solutions.”

p20, Andy Kirk, Data Visualisation - a successful design process (2012)

<http://site.ebrary.com/lib/dublincu/detail.action?docID=10642563>

→ **This doesn't apply to marking your assignment and exam!**

What is “good” visualisation?

<https://www.menti.com/> → 4420 7756



Provide 1 or 2 word descriptions for what makes a good visualisation
(you can submit up to 8 descriptions).

I'll show the results later in the session.

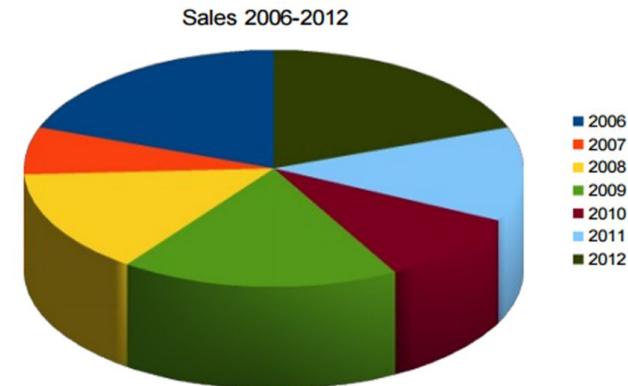
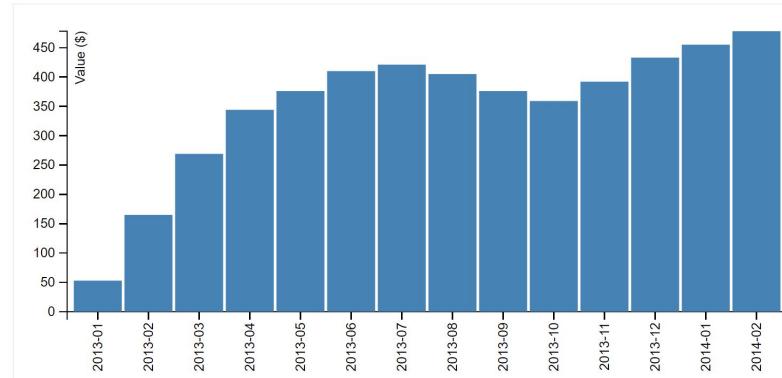


Figure 1 – a 3d pie-chart

menti.com
4420 7756



Graphic Communication

AKA Visual Communication

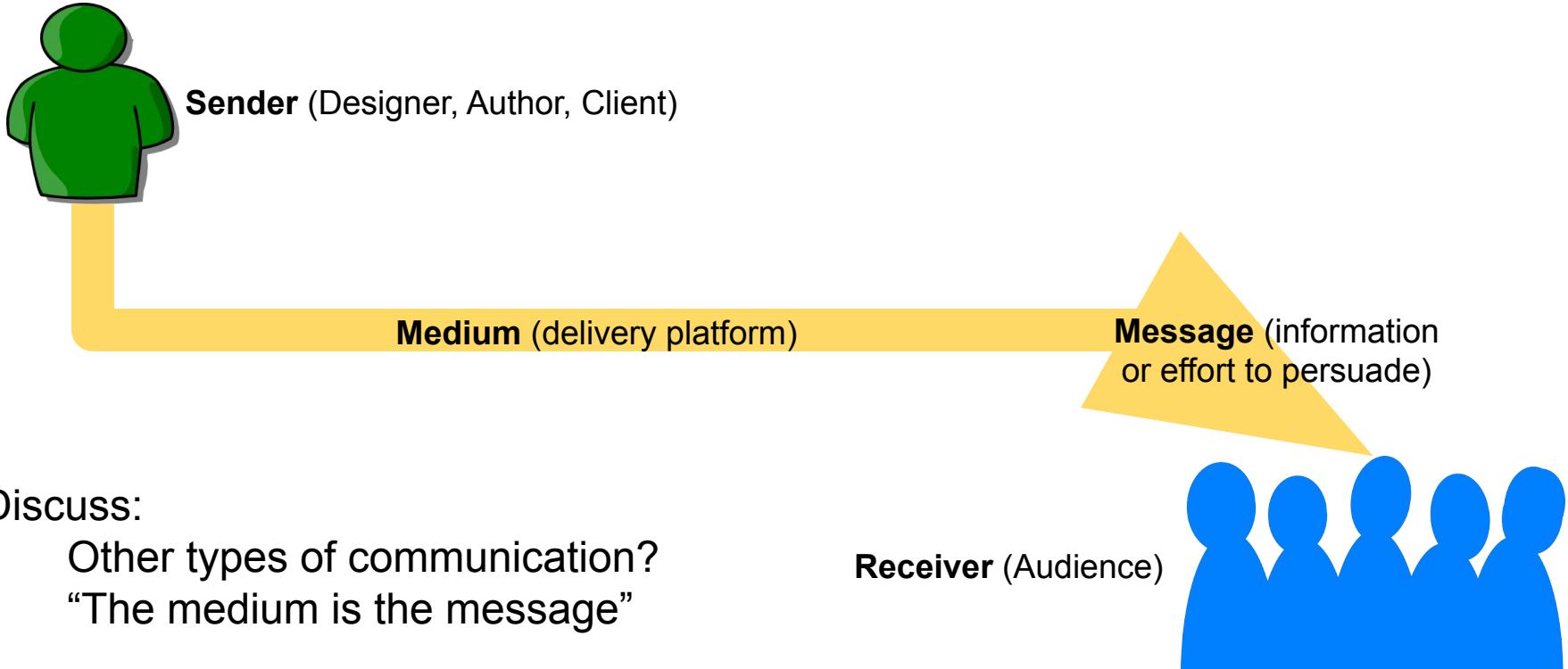


Cave Paintings → Written Word → Illuminated Manuscripts → Printed Word →
Digital Age → Modern Media

Media: “*all forms of printed paper or material (books, magazines, newspapers, brochures, flyers, signage, and billboards), the Internet, mobile phones and handheld devices, television, radio, CDs and DVDs, videos, video games, films, ...*”[Ref: 1]

A key part of “postindustrial” information economies

Graphic Communication



Graphic Communication: Stages of Understanding

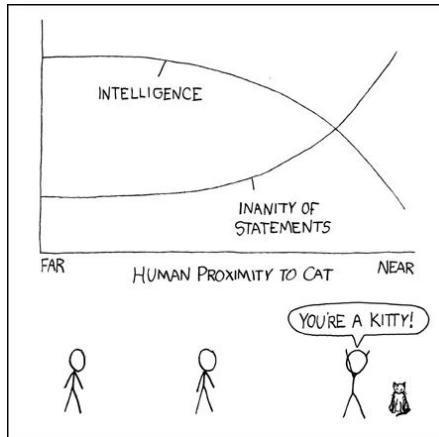
- Sensing → your brain seeing colours and shapes
- Perceiving → what does it show? big, small, bright, red,
- Interpreting → what does it mean? increasing, smaller, good, bad
- Comprehending → what does it mean **to me?** relevance, consequences



Graphic Communication Goals

- Information
- Persuasion
- Education
- Entertainment

<http://www.informationisbeautiful.net/visualizations/snake-oil-supplements/>



WORLD OCTOPUS DAY

THE GIANT PACIFIC OCTOPUS CAN WEIGH MORE THAN 600 POUNDS

ALL SPECIES ARE VENOMOUS, BUT THE BLUE-RINGED OCTOPUS IS THE ONLY ONE DANGEROUS TO HUMANS, RESPONSIBLE FOR AT LEAST TWO DEATHS.

OCTOPUSES VS. OCTOPI
THE PLURAL IN ENGLISH IS "OCTOPUSES," BUT THE GREEK PLURAL FORM "OCTOPODES" IS SOMETIMES USED. "OCTOPI," WHILE COMMONLY USED, IS CONSIDERED INCORRECT.

OCTOPUSES ARE ABOUT 90% MUSCLE

THE GIANT PACIFIC OCTOPUS CAN INHABIT DEPTHS OF UP TO 5,000 FEET

OCTOPUSES CAN QUICKLY CHANGE THE COLOR AND TEXTURE OF THEIR SKIN

300 RECOGNIZED SPECIES OF OCTOPUS

NATIONAL AQUARIUM | [Facebook](#) [Twitter](#) [YouTube](#) [Instagram](#) [Flickr](#) [Pinterest](#) [RSS](#) | aqua.org

Goals: Information

Structured data

Design principles:

Structure is key

Level of detail - macro v micro

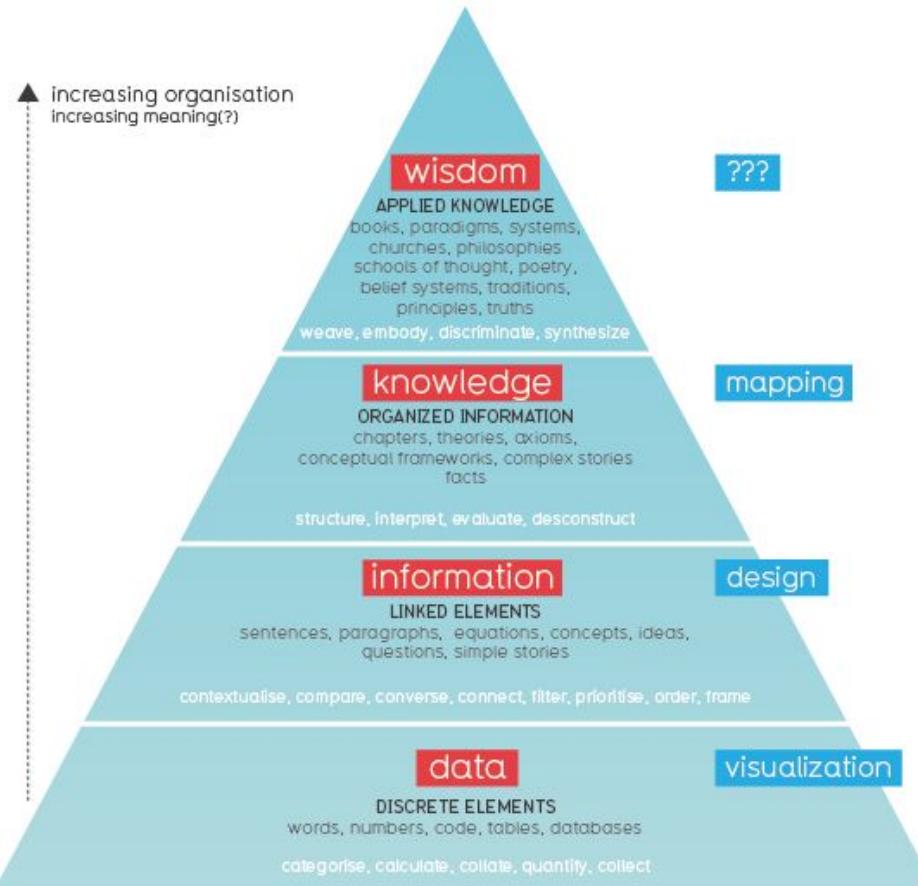
Layout, Colour - **credibility**



Important to have good source data - high quality

Hierarchy Of Visual Understanding?

Just playing. Something in this?



informationisbeautiful.net

Goals: Persuasion

Communication to elicit a particular response



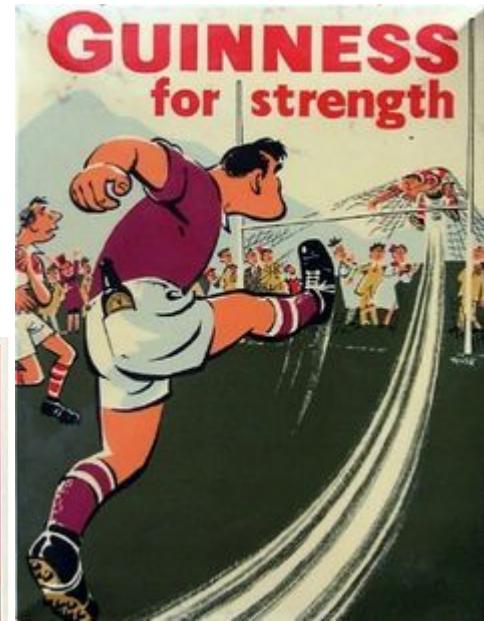
E.g., Advertising → using information to present a message

Appeal: Factual (rational) vs Emotional (values, opinions, attitudes)

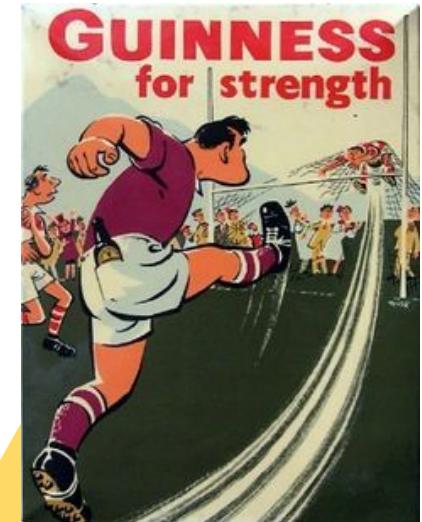
Design principles:

Research audience

Illustrations, themes, colours, grouping → attract viewer's eye

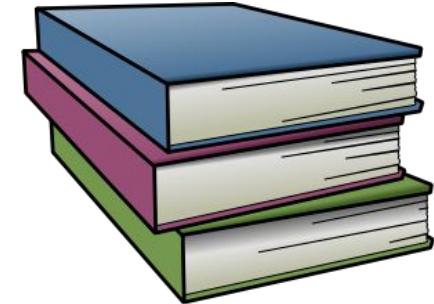


Graphic (Visual) Communication



Goals: Education

transferring knowledge and skills



Textbooks, online learning resources, brochures, movies

Design principles:

Divide information into chunks (hierarchy - trees, chapters, etc.)

Legibility is key

Progressive disclosure

Goals: Entertainment

pleasure, diversion, amusement

art, video games, film, television, ebooks

Design principles:

focus on narrative

how constructed (lighting, layout, multimodality) → the medium

Style ...

What Makes a Good Visualization?

explicit (implicit)



information
the art of journalism
structuring & harmonising information

visualization
the art of design
structuring & harmonising visuals

What makes a good visualisation?

Edward Tufte

“Graphical excellence ... gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space”

Stephen Few - “Show Me the Numbers” (2012)

Well told stories - simple, seamless, informative, true, contextual, familiar, concrete, personal, emotional, actionable, sequential

Gregor Aitsch - drivenbydata.net (NY Times, Graphics Editor)

“Know the rules, before you break them ...”

Cole Nussbaumer Knaflic - “Storytelling With Data” (2015)

“Data visualization is the process of turning information into pictures for a specific purpose.”

Kirk's principles of Good Data Visualisation

Good data visualisations are:

1. Trustworthy
 - a. Don't use inappropriate colour **palettes** or **fonts**
 - b. Don't include unnecessary chart junk
2. Accessible
 - a. Useful and understandable
 - b. Reward vs Effort (complexity is sometimes okay!)
3. Elegant
 - a. Thorough (get the little details right!)
 - b. Stylish

Andy Kirk, "Data Visualisation" (2016)

Questions?

Questions on exercises?

Cleaning with spreadsheets or OpenRefine or Python

Q: Which do you prefer? You can give the answer via menti.com

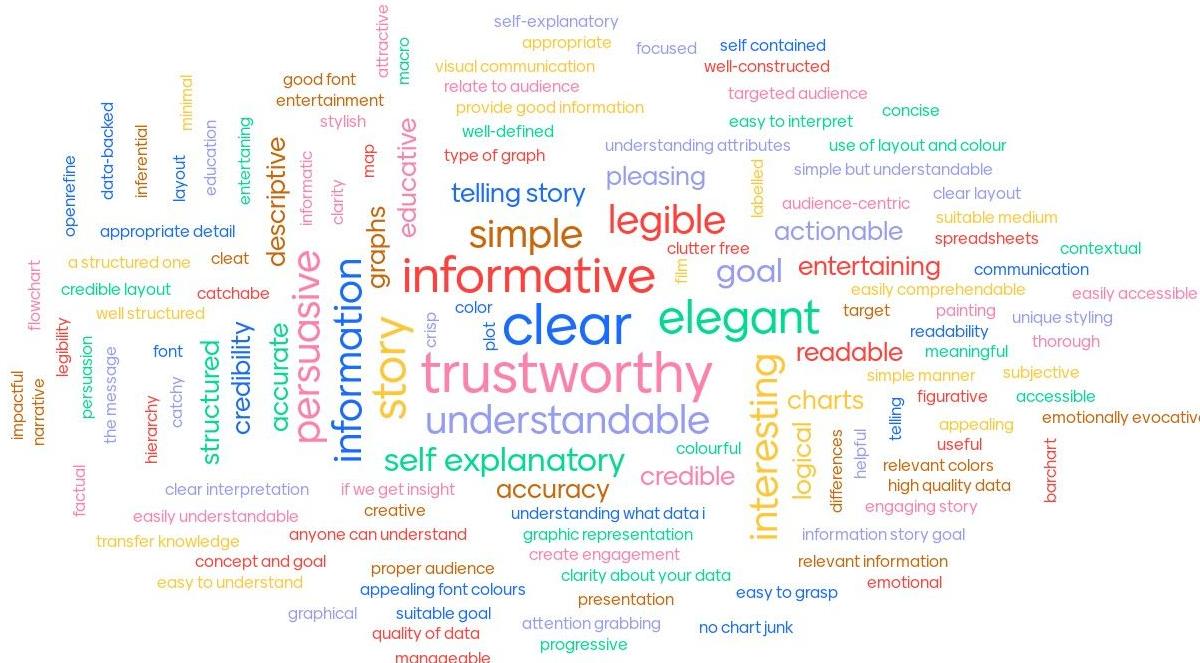
menti.com
4420 7756





What is "good" visualisation?

2022/2023 class



CHOCOLATE

The first box of Valentine's Day chocolates was created by famed British chocolatier **Richard Cadbury**, in **1868**

Chocolate sales for Valentine's Day total over
ONE BILLION DOLLARS (US)
every year

71%
of North American
eaters prefer
MILK
CHOCOLATE

Americans eat
12 POUNDS
of chocolate
ANNUALLY

A survey con-
ducted by the
**Chocolate
Manufacturers
Association**
revealed that
**50 PERCENT
OF WOMEN**
will likely give a
gift of chocolate
for **Valentine's
Day**.

As an **elixir for love**, chocolate
has been believed throughout
history to bring smiles to the
broken-hearted and to prompt
amorous feelings in both
men and women.



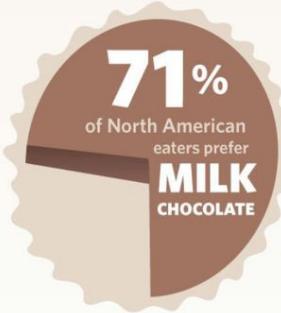
Motivation? Information, Education, Persuasion and/or Entertainment

- Data?
- Illustrations?
- Structure?
- Design?

CHOCOLATE

The first box of Valentine's Day chocolates was created by famed British chocolatier **Richard Cadbury**, in **1868**

Chocolate sales for Valentine's Day total over **ONE BILLION DOLLARS (US)** every year



A survey conducted by the **Chocolate Manufacturers Association** revealed that **50 PERCENT OF WOMEN** will likely give a gift of chocolate for **Valentine's Day**.



As an **elixir for love**, chocolate has been believed throughout history to bring smiles to the broken-hearted and to prompt amorous feelings in both men and women.

With all of the chocolate eaten on Valentine's Day there is bound to be a few stains.

Find out how to remove chocolate and more at **Clorox.com**

Motivation? Information, Education, Persuasion and/or Entertainment

- Data?
- Illustrations?
- Structure?
- Design?

Tools for Visualisation?



Excel/Google sheets

Photoshop/GIMP

Powerpoint

Tableau

PowerBI

Qlikview

R - ggplot

Pandas .plot()

Python - matplotlib, seaborn,
bokeh

Plot.ly

D3.js & other javascript libraries

[Overview of Python Visualisation Libraries \(with example notebooks\)](#)

Resources

[Ref 1] John Dimarco, Digital Design for Print and Web An Introduction to Theory, Principles, and Techniques, (Part 1 only), <https://www-dawsonera-com.dcu.idm.oclc.org/abstract/9780470639184>

- David McCandless, The Beauty of Data Visualization (TED talk)
http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization?language=en
- ProPublica Guides, Data Style Guide,
<https://github.com/propublica/guides/blob/master/news-apps.md> - very good source of general design rules, sections on Accuracy, Axes, Charts, Colors, Legends, Maps, Money, Numbers, Sources and Time are relevant to CA682.
- History of Visual Communication, http://www.citrinitas.com/history_of_viscom/
- Does Comic Sans Benefit People with Dyslexia?
<https://www.boia.org/blog/does-comic-sans-benefit-people-with-dyslexia>

Books on Visualisation

Andy Kirk, “Data Visualisation” (2016)

Cole Nussbaumer Knaflic, “Storytelling with Data” (2015)

Stephanie D. H. Evergreen, “Effective Data Visualisation” (2017) ← Business view

Alberto Cairo, “The Truthful Art” (2016) ← Journalistic view

Stephen Few, “Show Me the Numbers” (2012)

Edward Tufte, many ...

07 Data Visualisation: Encoding Data

CA682

suzanne.little@dcu.ie

Today: Data Visualisation (part II)

- Recap
- Data representation
- Chart Types
- Choosing your chart

Recap

- What is a “good” visualisation?
- 4 components of communication
 - Sender, Message, Medium, Receiver
- 4 types of Graphic Communication
 - Information, Persuasion, Education, Entertainment
- Opinions on good visualisation

concept explanation graphical communication visual form structural data

view colours line chart magnifying illuminating communicates idea

easy to explain outstanding easy to remember easy to recall sort data by values

readable methodical feels good provides summary

natural simple to understand interesting metrics

calm grab attention advertising

good legend attributes

graphics

presentable vibrant

labeled graph hierarchy

graphical facts connection seems good sheet

concise

accessible general accuracy graphical tables connection

captures details relativity

growth enter immediate minimalist

outcome contrasting colors

communicative illustration

inclusive user-friendly

keyword everything legible eyecatcher

pictures something mats

explainable goal-driven comprehend

factual exploratory

good design better

good design

appealing insightful

narrative summary

insightful sensing

descriptive apples

explanatory charts

interesting useful layout

charts statistics

contextual

catchy

eye catching

graphics

good legend attributes

attractive

beautiful user friendly

easy to read true

legibility

colourful oranges graphs

small trend shapes

interactive

credible detailed

understandable

story

structured tool

rich

trustworthy

simple

clear

elegant

informative

goal easy

colorful

accrue

representational

structure

distinguishable histogram

usefulness encapsulating

fast

character

clean data

color

informatic

medium

animation

meaningful

big chungus

eye catchy

solution

technical

attractive

graphic

precise

explainable but concise

data representation

including legend skew

good data pie chart

detailed explanation

high quality focus on narrative

derive insights

with specific goals

knowledge

interpret

cool

less text

persuasive

elegance

behaviour

pattern

concept

succinct

analysis responsive

visual

consistency

simple and clear

should be legible

understand

iphone advertisement

entertaing

story telling

helpful

color categories

effective colour scheme

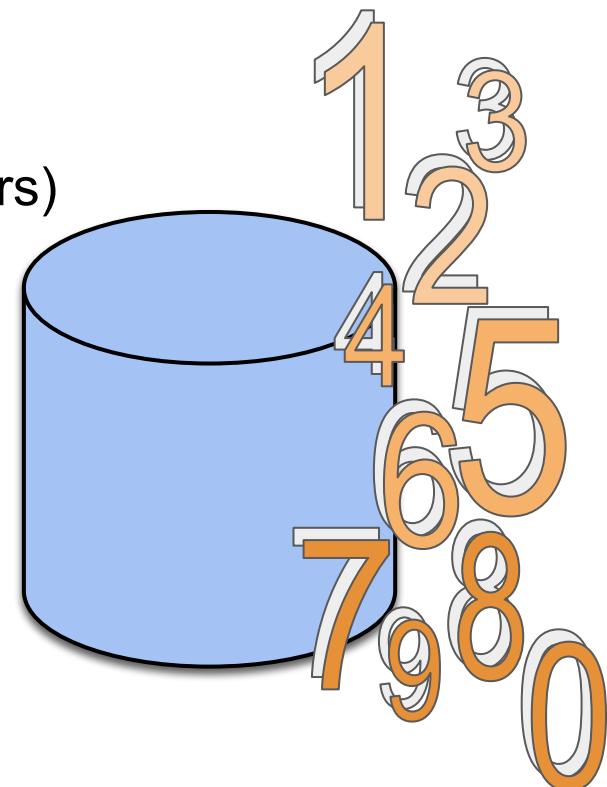
able to tell story

easy understanding

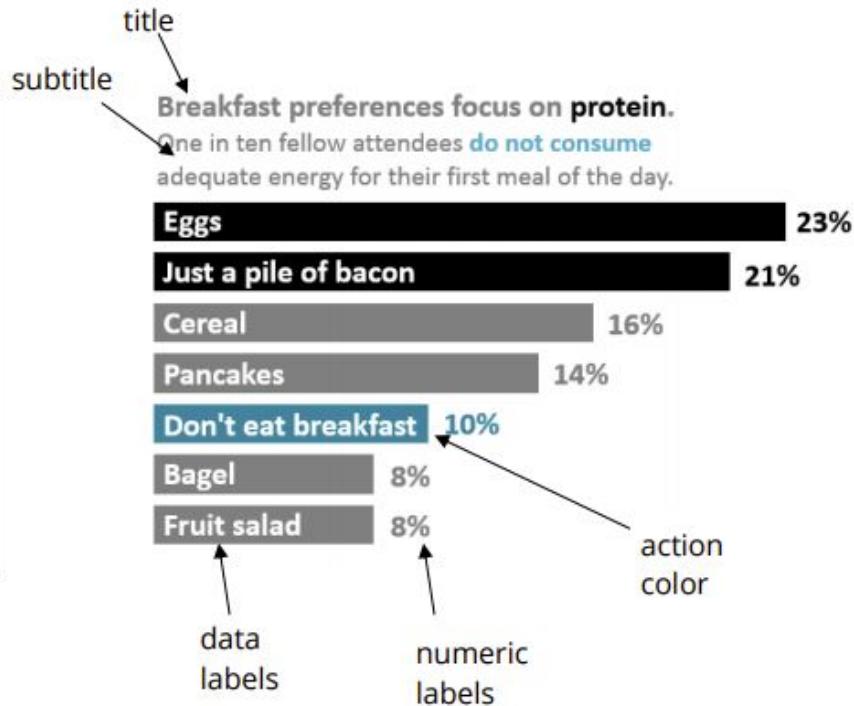
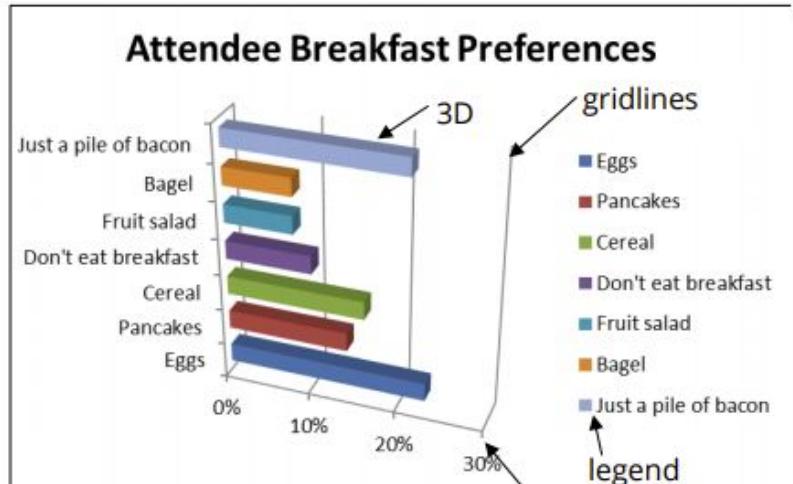
self exploratory

Categorising data

- Structured or Unstructured
 - Tabulated or Raw
- Qualitative (description) or Quantitative (numbers)
- Text (categorical), Numbers (numeric)
→ Documents, Images, Video, Audio, 3D
- Discrete or Continuous
- Nominal, Ordinal, Interval, Ratio
- Temporal (or Time Series)
- Geographic (or Spatial)



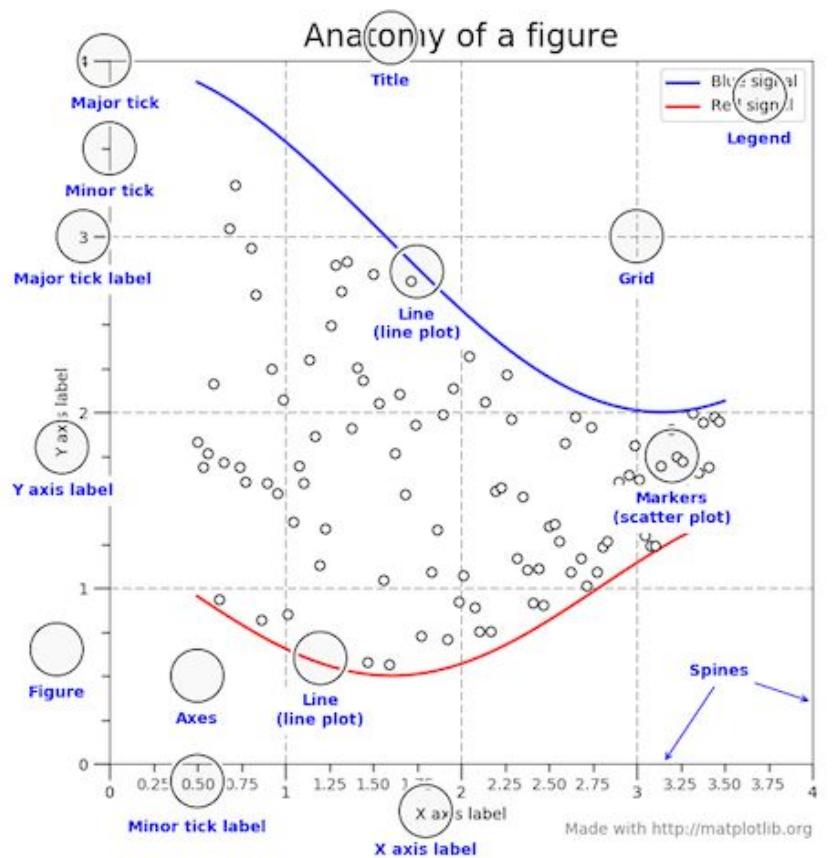
What is a graph? (or chart)



<http://stephanieevergreen.com/updated-data-visualization-checklist/>

What is a graph? (or chart)

<https://matplotlib.org/examples/showcase/anatomy.html>



encode? represent in an alternative way

A graph *encodes* data

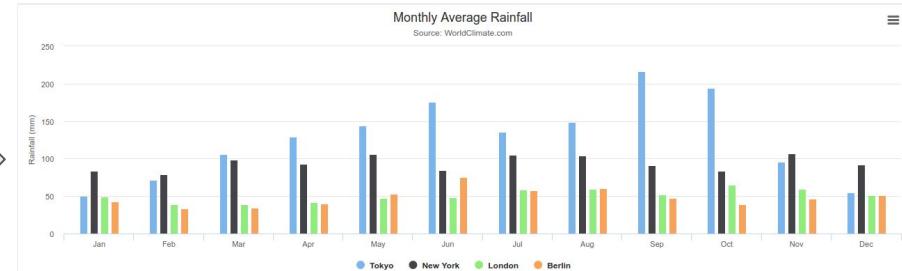
2 100 39 91 93 98 94 89 30 82

name, age, id, colour, language



}

encoded



→ Marks & Attributes

Data representation: Marks

Point



No variation

Eg. Quantity through position
(scatter plot)

Data representation: Marks

Point



No variation

Line



1 dimension

Eg. Quantity through position
(scatter plot)

Eg. Quantity through variation in
size (bar chart)

Data representation: Marks

Point



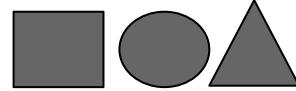
No variation

Line



1 dimension

Area



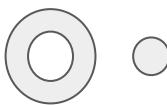
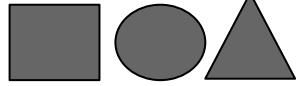
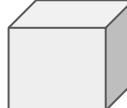
2 dimensions

Eg. Quantity through position
(scatter plot)

Eg. Quantity through variation in
size (bar chart)

Eg. Quantity through size and
position (bubble chart)

Data representation: Marks

Point		No spatial variation	Eg. Quantity through position (scatter plot)
Line		1 spatial dimension	Eg. Quantity through variation in size (bar chart)
Area		2 spatial dimensions	Eg. Quantity through size and position (bubble chart)
Form		3 spatial dimensions	Eg. Quantity through variation in size/volume (proportional shape)

Data representation: Attributes

Quantitative

Position

Size (length, area, volume)

Angle/Slope

Quantity

Colour: Saturation

Colour: Lightness

Pattern

Motion

Categorical

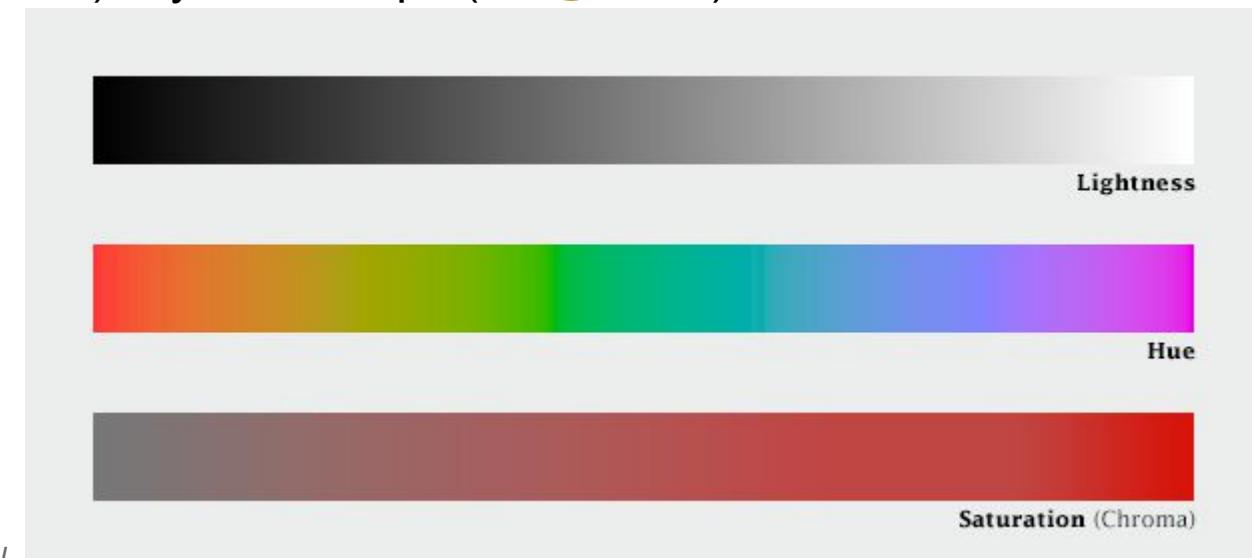
Colour: Hue

Symbol/Shape (⊕ ☹ ↗ Ω)

Relational

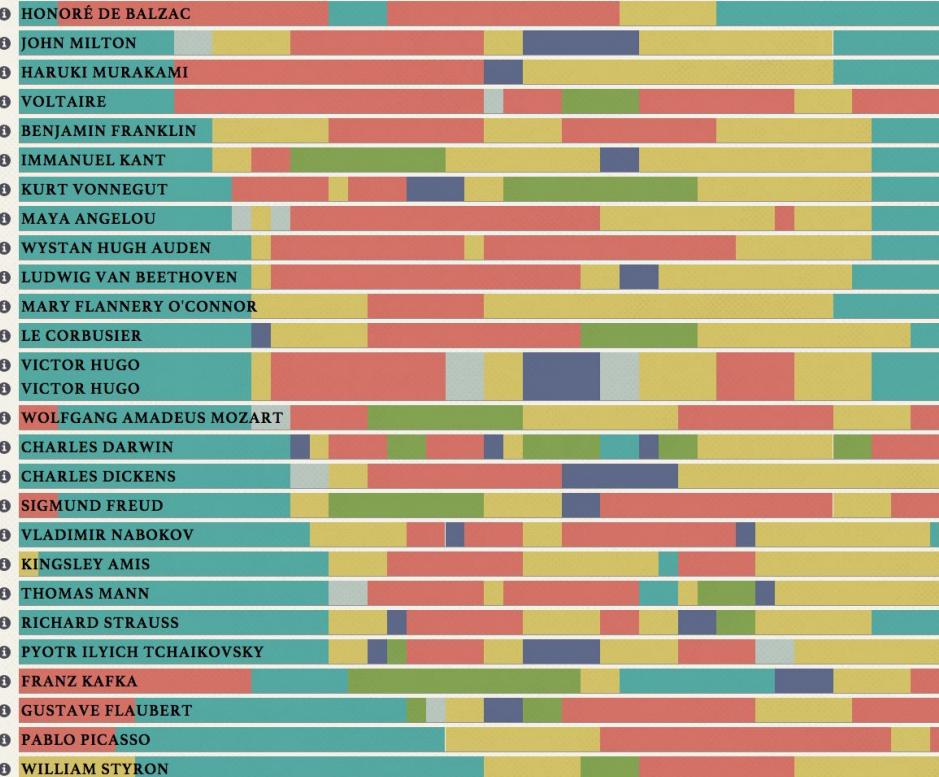
Connection/Edge

Containment



THE DAILY ROUTINES OF FAMOUS CREATIVE PEOPLE

Turns out great minds don't think alike. Discover how some of the world's most original artists, writers and musicians structured their day, based on 'Daily Rituals' by Mason Currey. Filter the different categories by toggling on or off, and hover over the colored bars to learn more about the daily routines.



<https://podio.com/site/creative-routines>

(charts, graphs, tables, figures, maps, plot, diagram, ...)

Chart types

Categorical: comparing categories and distributions of quantitative values

Hierarchical: Charting part-to-whole relationships and hierarchies

Relational: Graphing relationships to explore correlations and connections

Temporal: Showing trends and activities over time

Spatial: Mapping spatial patterns through overlays and distortions

Charts (a curated selection) - Categorical

- Bar graph: comparisons of quantitative values from different categories
- Dot plot: Like bar but use a point or symbol to indicate the value so can include colour, area, shape to capture extra dimensions.
- Circle packing: comparisons of values using area, shape, colour, layout
- Polar chart: (also radar or spider) radially plotted bar chart showing 3+ quantitative measures

comparison

Charts (a curated selection) - Categorical

- Bar graph: comparisons of quantitative values from different categories
- Dot plot: Like bar but use a point or symbol to indicate the value so can include colour, area, shape to capture extra dimensions.
- Circle packing: comparisons of values using area, shape, colour, layout
- Polar chart: (also radar or spider) radially plotted bar chart showing 3+ quantitative measures

comparison

- Box-and-Whisker plot: common in statistical analysis
- Histogram (not a bar chart): frequency and distribution
- Word cloud: frequency of concepts

distribution

Quick word on histograms and box/whiskers

Statistical graphs - very useful and powerful!

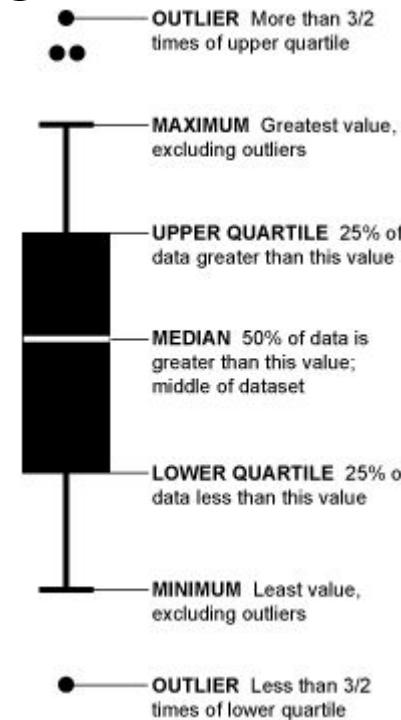
Do not confuse histogram with bar chart

Learn how to read a box/whisker plot

But remember non-experts often struggle with them

Histogram → rich visualisation of distributions

Boxplot → comparing distributions between several groups



Charts (a curated selection) - Hierarchical

- Pie charts: how quantities make up a whole
- [Waffle charts](#): aka square pie, coloured grid squares to show quantities
- [Stacked bar chart](#): breakdown values within bar
- [Treemap](#): enclosed hierarchical display
- Venn diagram: relationships between sets and collections

part-to-whole

Charts (a curated selection) - Hierarchical

- Pie charts: how quantities make up a whole
- [Waffle charts](#): aka square pie, coloured grid squares to show quantities
- [Stacked bar chart](#): breakdown values within bar
- [Treemap](#): enclosed hierarchical display
- Venn diagram: relationships between sets and collections

- [Dendrogram](#): aka tree hierarchy, layout tree, clusters. Node-link diagram showing hierarchical relationships across multiple layers

} part-to-whole
} hierarchies

Charts (a curated selection) - Relational

- [Scatter plot](#): relationship between quantitative values for two categories
- [Bubble plot](#): relationship between 3 qualitative values (area, x position, y position)
- Heat map: quantitative values between 2 categorical dimensions (colour coded)
- Matrix chart: quantitative values between 2 categorical dimensions
- [Sankey diagram](#): categorical composition and qualitative flows

connections

Charts (a curated selection) - Temporal

- Line chart: change in quantitative values over time
- Area chart: coloured in line chart :-)
- [Stream graph](#): continuous changes in qualitative values in different categories over time

trends

Charts (a curated selection) - Temporal

- Line chart: change in quantitative values over time
- Area chart: coloured in line chart :-)
- Stream graph: continuous changes in qualitative values in different categories over time
- Gantt chart: start, finish & duration of difference categorical activities

trends

activities

Charts (a curated selection) - Spatial

Map projections - think about flattening an orange peel

<http://geoawesomeness.com/best-map-projection/>

https://youtu.be/KUF_Ckv8HbE

Charts (a curated selection) - Spatial

- Choropleth map: (aka heat map) quantitative values for distinct spatial regions
- Isarithmic map: (aka contour map) quantitative values linking spatial regions
- Proportional symbol map: represent values by proportionally sized areas overlayed on map

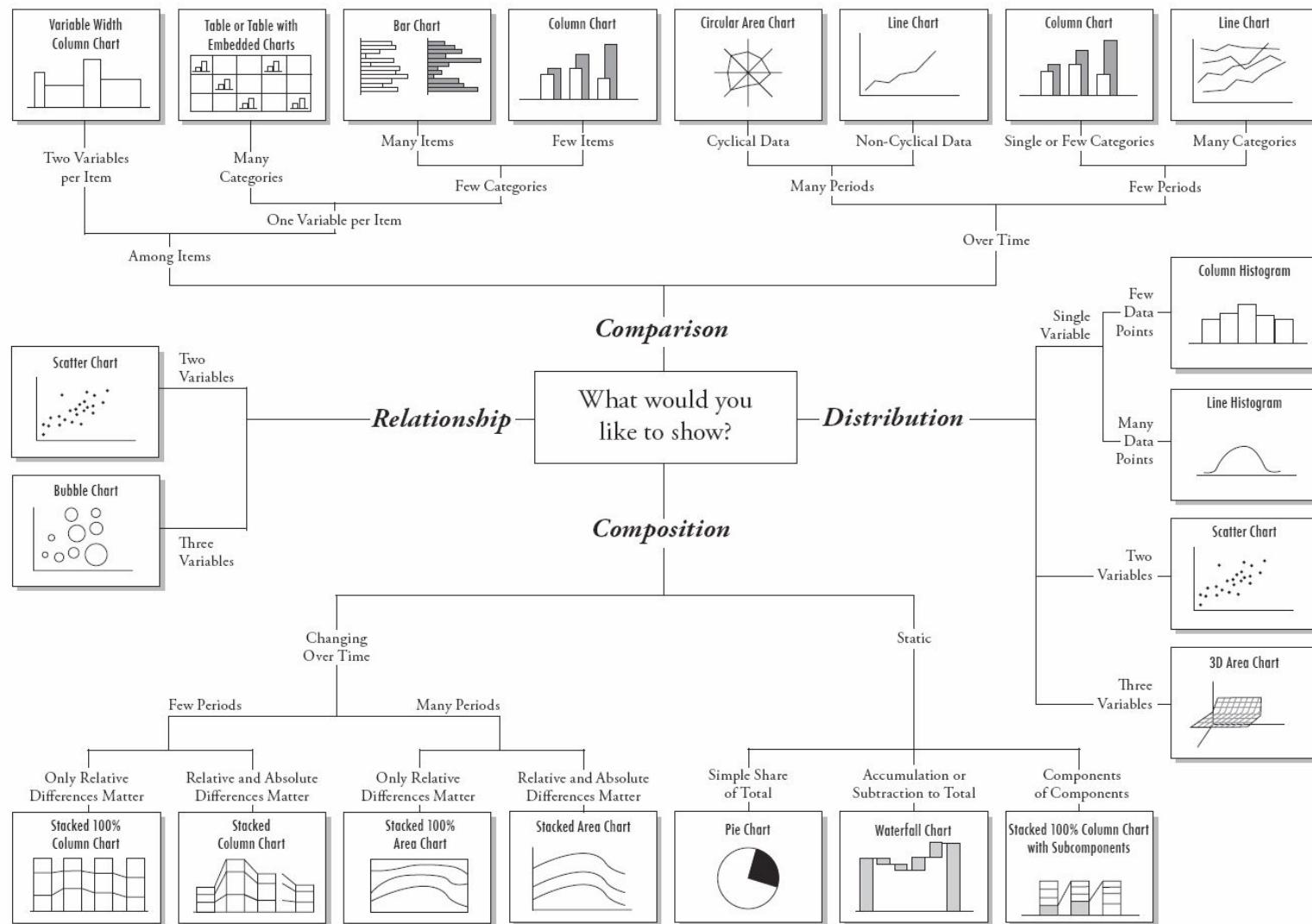
overlays

Charts (a curated selection) - Spatial

- Choropleth map: (aka heat map) quantitative values for distinct spatial regions
 - Isarithmic map: (aka contour map) quantitative values linking spatial regions
 - Proportional symbol map: represent values by proportionally sized areas overlayed on map
-
- Area cartogram: distort map spatial regions to show value
 - Dorling Cartogram/Grid map: arrange regular shapes into map using colour to indicate category

overlays

distortions



Exercise

In pairs/threes, discuss the best graph type for your question.

Categorical: Comparing categories and distributions of quantitative values

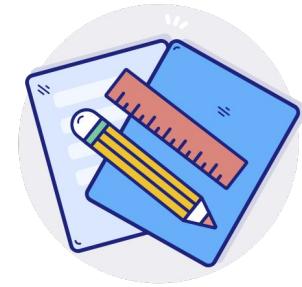
Hierarchical: Charting part-to-whole relationships and hierarchies

Relational: Graphing relationships to explore correlations and connections

Temporal: Showing trends and activities over time

Spatial: Mapping spatial patterns through overlays and distortions

Tools for Visualisation?



Excel/Google sheets

Photoshop/GIMP

Powerpoint

Tableau

PowerBI

Qlikview

R - ggplot

Pandas .plot()

Python - matplotlib, seaborn,
bokeh

Plot.ly

D3.js & other javascript libraries

[Overview of Python Visualisation Libraries \(with example notebooks\)](#)

Tools to create visualisations -

<https://loop.dcu.ie/mod/page/view.php?id=1651443>

- Programming languages
 - Document discussing Python Libraries on Loop
<https://loop.dcu.ie/mod/url/view.php?id=1651494>
- Dedicated tools (many!)
 - Tableau - <https://www.tableau.com/academic/students>
- Web-based, interactive options like D3.js and other Javascript libraries (exercises will be on loop for next week)

D3.js

- <https://d3js.org/>
- Data Driven Documents
- JavaScript library
- Transform data to standard web formats (HTML, SVG, CSS)
- Good for interactive and dynamic browser visualisations
- “D3 does not replace the browser’s toolbox, but exposes it in a way that is easier to use”

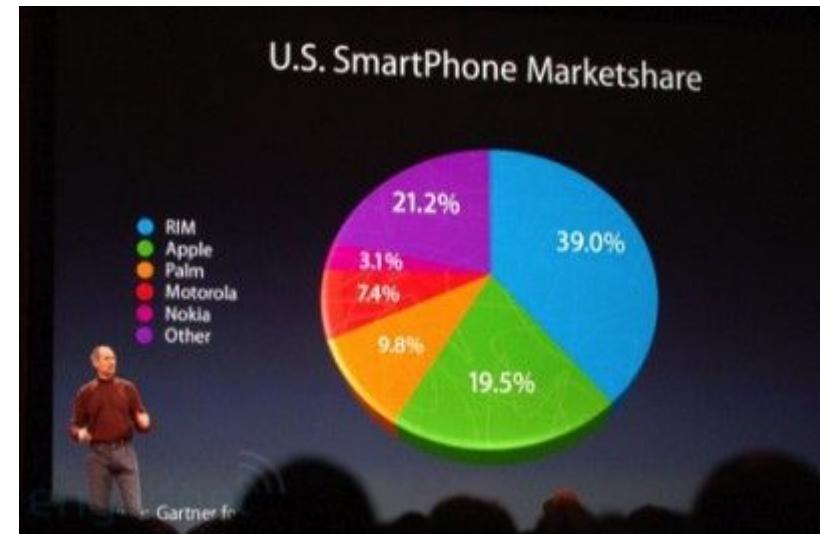
D3.js doesn't ...

- support older browsers
- generate prepared visualisations for you (unlike Excel, Tableau etc.)
 - so you generally don't do **Processing** or **Analytics** in D3.js
- ~~handle bitmaps (non-vector graphics) like the tiles on Google Maps (although there are ways around this)~~ Use [leaflet.js](#)
- hide your original data - it's all sent to the browser (client) to do the graph generation. So be sure you want it exposed!

Critiquing designs - reading for week 7

<https://www.washingtonpost.com/graphics/politics/2016-election/trump-charts/>

<https://simonrogers.net/2013/03/15/a-conversation-with-stephen-few-about-data-visualisation-kind-of/>



Resources

Chapter 6 of Data Visualisation (Andy Kirk) or Chapter 5 of ebook
(<http://site.ebrary.com/lib/dublincu/Doc?id=10642563>) covers most of the types.

“A tour through the Visualization Zoo” (ACM publication)

<http://queue.acm.org/detail.cfm?id=1805128> - See also interactive presentation on loop (<https://loop.dcu.ie/mod/resource/view.php?id=79908>)

“The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”,
Ben Schneiderman,

http://www.interactiondesign.us/courses/2011_AD690/PDFs/Shneiderman_1996.pdf

More resources

<https://dsaber.com/2016/10/02/a-dramatic-tour-through-pythons-data-visualization-landscape-including-ggplot-and-altair/>

Some examples using python,

<https://towardsdatascience.com/5-quick-and-easy-data-visualizations-in-python-with-code-a2284bae952f>

<https://datavizcatalogue.com/> Interactive website with a catalogue of different chart types

<https://chaione.com/blog/building-blocks-graphs/> Good summary of main graph components

08 Data Visualisation - human vision

suzanne.little@dcu.ie

Today

Choosing Charts Recap

Human visual processing

Perception

Attention

Sometimes a “chart” is not even needed!

In our recent technology workshop
42% identified as female
and 57%* as male.
*1% declined to answer

Gender balance of workshop participants
12th July 2019

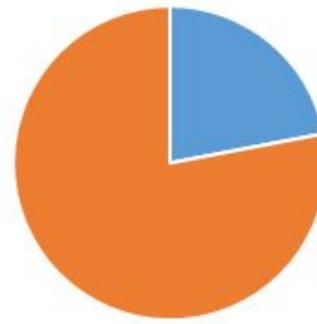


Chart types

Categorical: comparing categories and distributions of quantitative values

Hierarchical: Charting part-to-whole relationships and hierarchies

Relational: Graphing relationships to explore correlations and connections

Temporal: Showing trends and activities over time

Spatial: Mapping spatial patterns through overlays and distortions

CHRTS - the classic examples ([full list](#))

Categorical - Bar chart

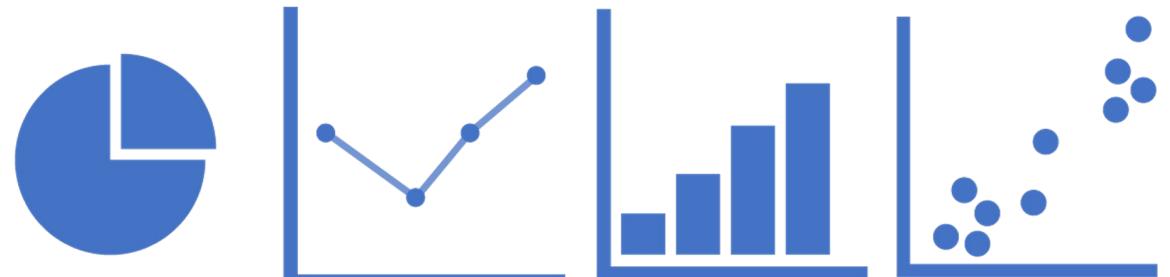
"There's a strand of the data viz world that argues everything could be a bar chart. That's possibly true but also possibly a world without joy." Amanda Cox, Editor, The Upshot

Hierarchical - Pie chart

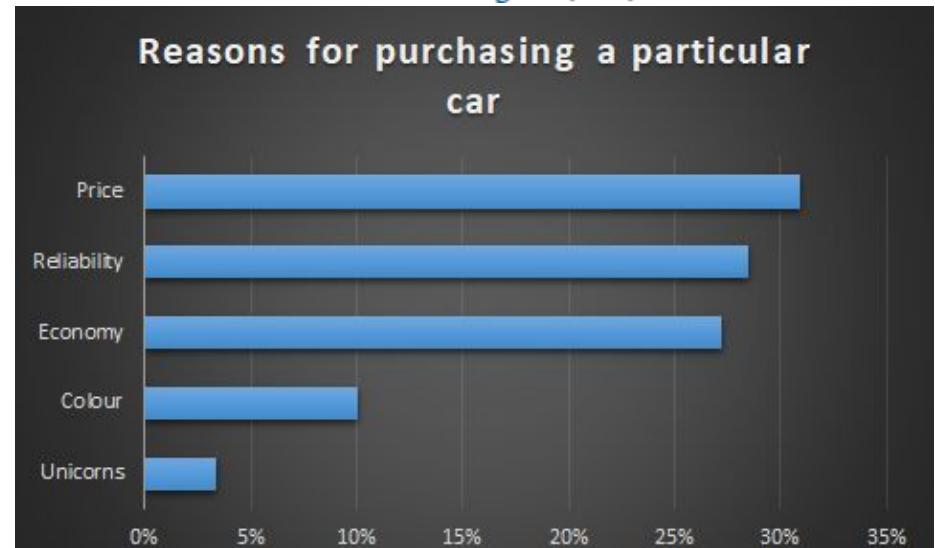
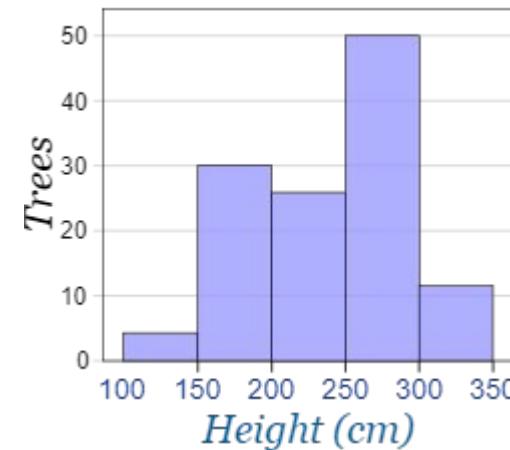
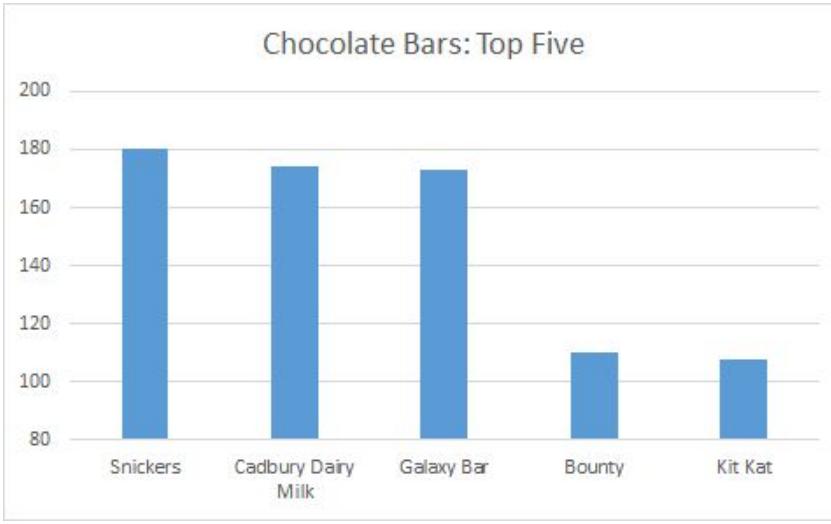
Relational - Scatterplot

Temporal - Line chart

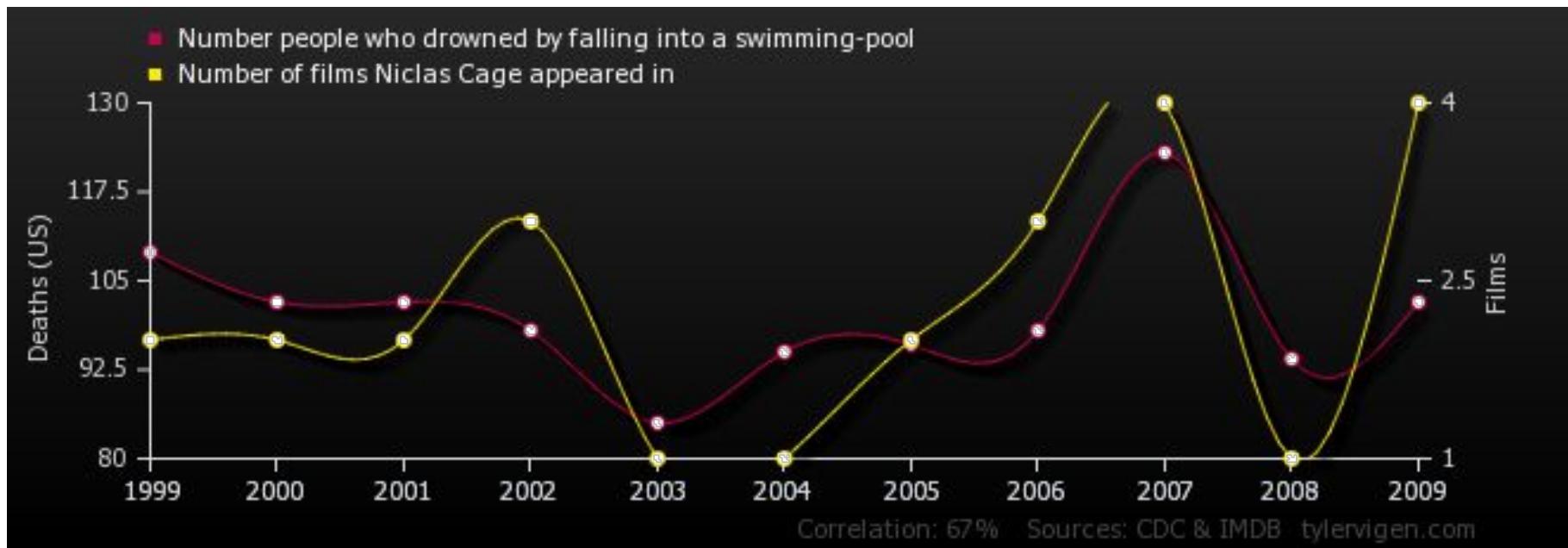
Spatial - Maps



Bar chart or Histogram?



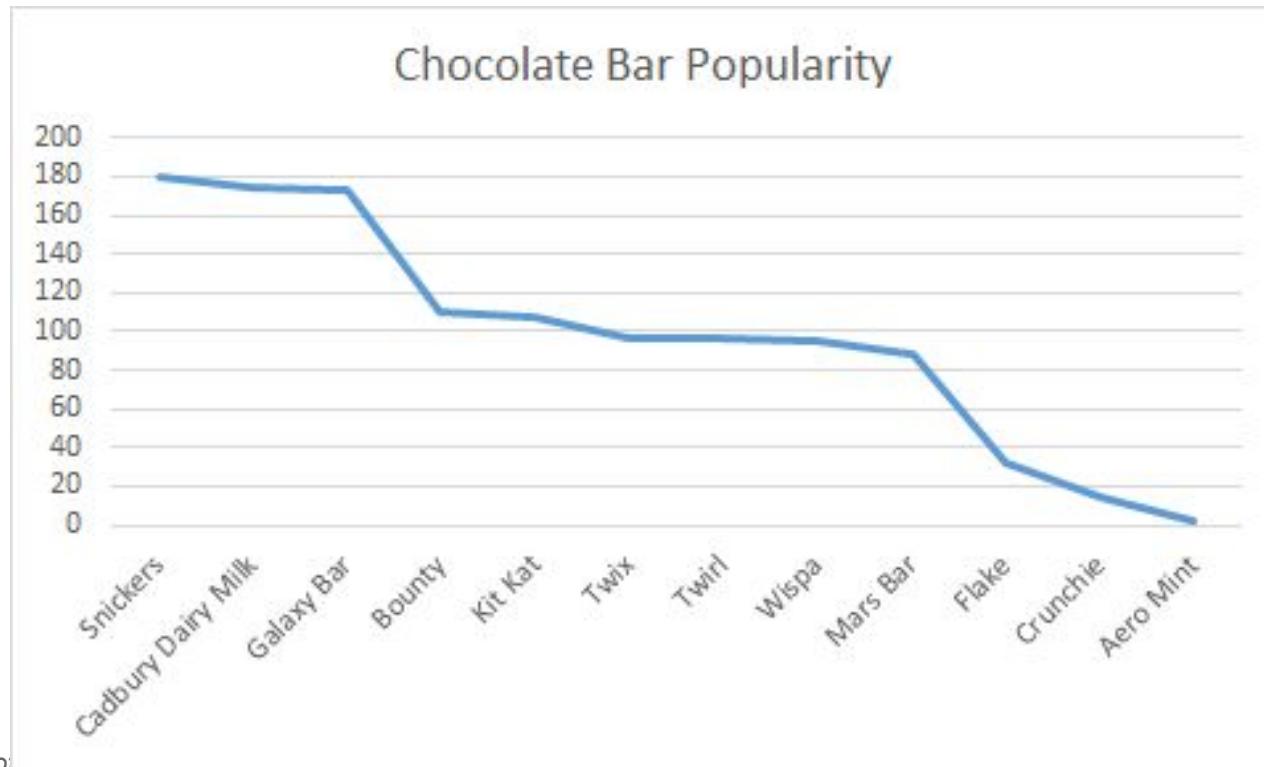
Beware of spurious correlations!



<https://tylervigen.com/old-version.html>

Temporal - Line charts

Be careful that the attribute on the x-axis is connected!



Another risk of spatial maps ...



<https://www.fastcompany.com/90572489/u-s-election-maps-are-wildly-misleading-so-this-designer-fixed-them>

Exercise

In pairs/threes, discuss the best graph type for your question.

Categorical: Comparing categories and distributions of quantitative values

Hierarchical: Charting part-to-whole relationships and hierarchies

Relational: Graphing relationships to explore correlations and connections

Temporal: Showing trends and activities over time

Spatial: Mapping spatial patterns through overlays and distortions

Tools for Visualisation?



Excel/Google sheets

R - ggplot

Photoshop/GIMP

Pandas .plot()

Powerpoint

Python - matplotlib, seaborn,
bokeh

Tableau

Plot.ly

PowerBI

Qlikview

D3.js & other javascript libraries

[Overview of Python Visualisation Libraries \(with example notebooks\)](#)

Tools to create visualisations -

<https://loop.dcu.ie/mod/page/view.php?id=2169731>

- Programming languages
 - Document discussing Python Libraries on Loop
- Dedicated tools (many!)
 - Tableau - <https://www.tableau.com/academic/students>
- Web-based, interactive options like D3.js and other Javascript libraries

D3.js

- <https://d3js.org/>
- Data Driven Documents
- JavaScript library
- Transform data to standard web formats (HTML, SVG, CSS)
- Good for interactive and dynamic browser visualisations
- “D3 does not replace the browser’s toolbox, but exposes it in a way that is easier to use”

D3.js doesn't ...

- support older browsers
- generate prepared visualisations for you (unlike Excel, Tableau etc.)
 - so you generally don't do **Processing** or **Analytics** in D3.js
- ~~handle bitmaps (non-vector graphics) like the tiles on Google Maps (although there are ways around this)~~ Use [leaflet.js](#)
- hide your original data - it's all sent to the browser (client) to do the graph generation. So be sure you want it exposed!

Today

~~Choosing Charts~~

Human visual processing

Perception

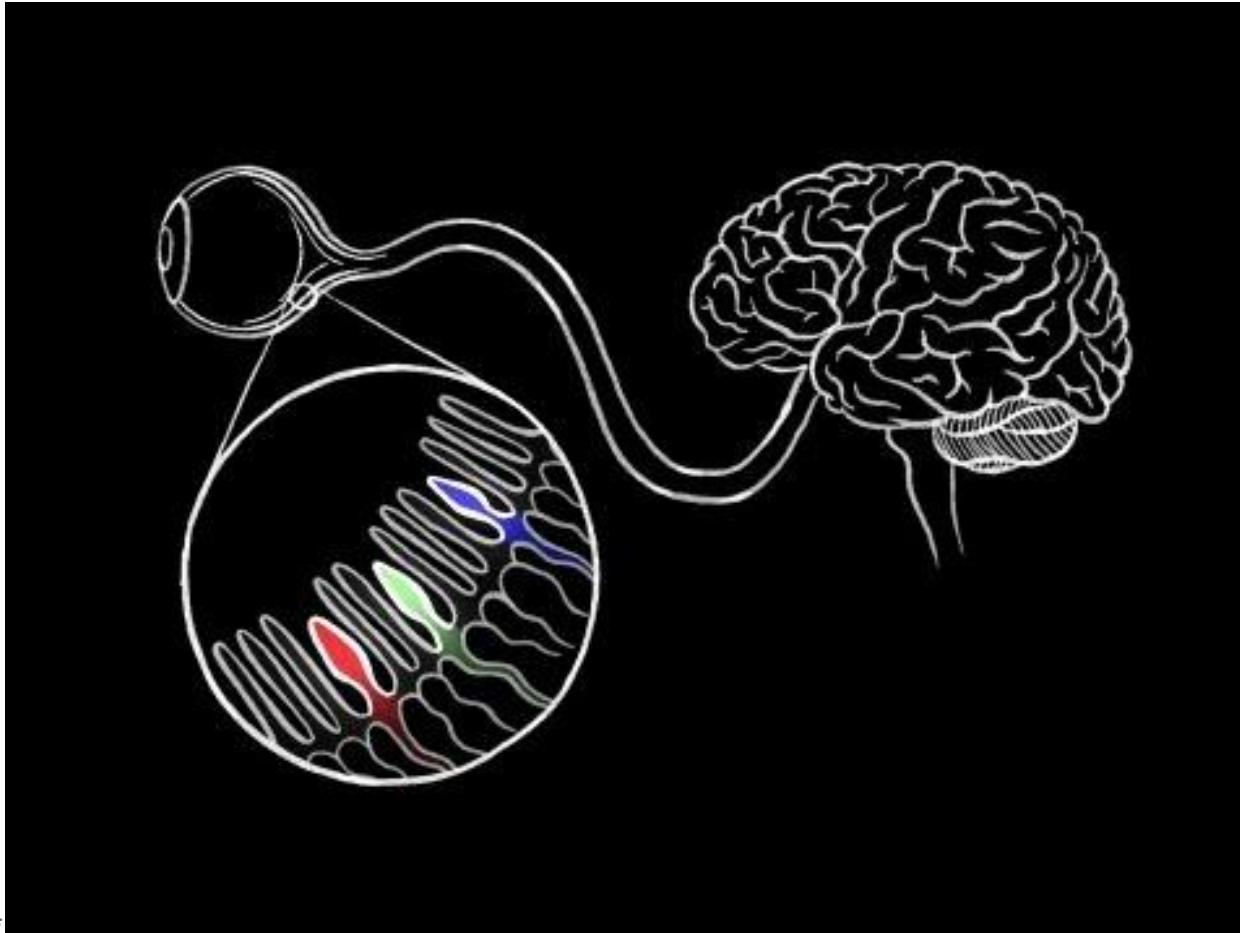
Attention

Graphic Communication: Stages of Understanding

- Sensing → your brain seeing colours and shapes
- Perceiving → what does it show? big, small, bright, red,
- Interpreting → what does it mean? increasing, smaller, good, bad
- Comprehending → what does it mean **to me?** relevance, consequences



How we see colour - https://www.youtube.com/watch?v=l8_fZPHasdo



How we see

Light enters through Lens to focus on Retina

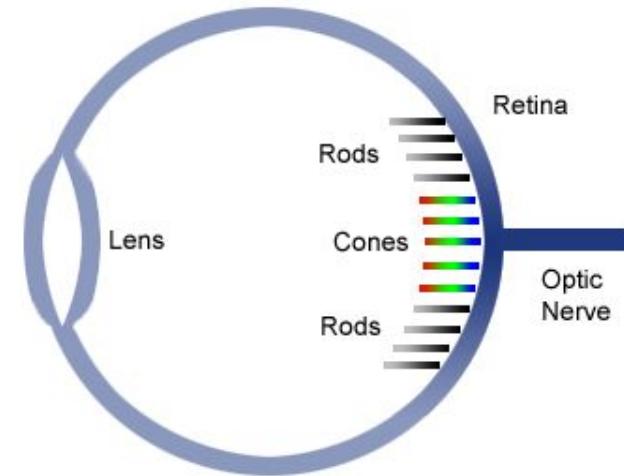
(Lens actually inverts the image!)

Triggers nerve impulses

Rods = low light; B&W; peripheral, movement

Cones = colour; 3 types: RGB

Processed in visual cortex (back of brain)



How we see

Binocular vision

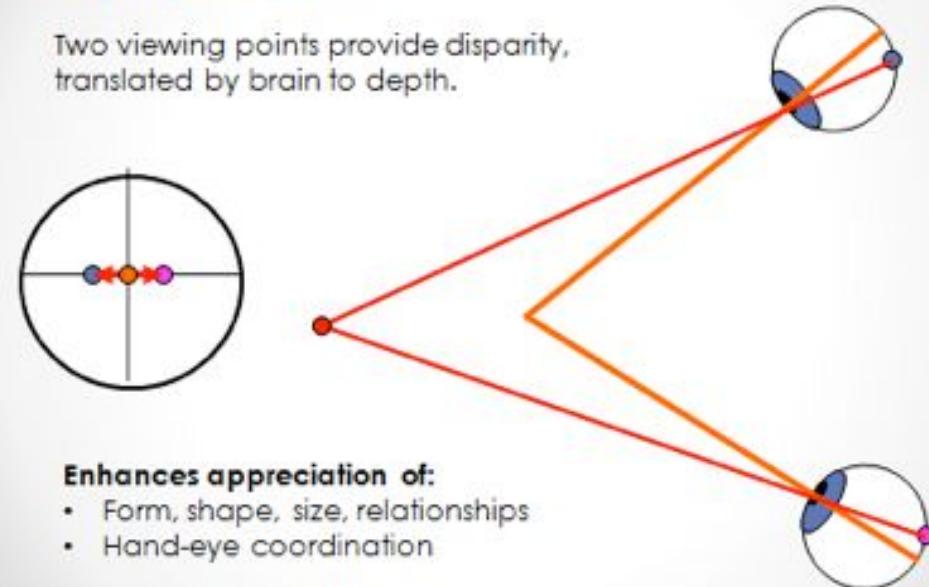
Depth perception due to two different images (L & R)

Useful for hand-eye coordination

Other important cues for depth ...

Stereopsis (3D Vision) – Frontal position of the eyes allows both eyes to work together to create depth perception.

Two viewing points provide disparity, translated by brain to depth.



Enhances appreciation of:

- Form, shape, size, relationships
- Hand-eye coordination



Depth Cues - Binocular Vision

Most important cue is **Binocular Disparity** (or the binocular parallax)

Images sensed by our two eyes are slightly different and this difference is used to determine depth

Exploiting this gives 3D movies - 1 eye sees the red while the other sees the blue due to lens filters. Or use polarised lens for the same effect.

Also **Convergence** - the difference in direction of our eyes when looking at closer objects (slightly pointing inwards)

<http://sciencelearn.org.nz/Contexts/Light-and-Sight/Sci-Media/Video/How-we-see-3D>

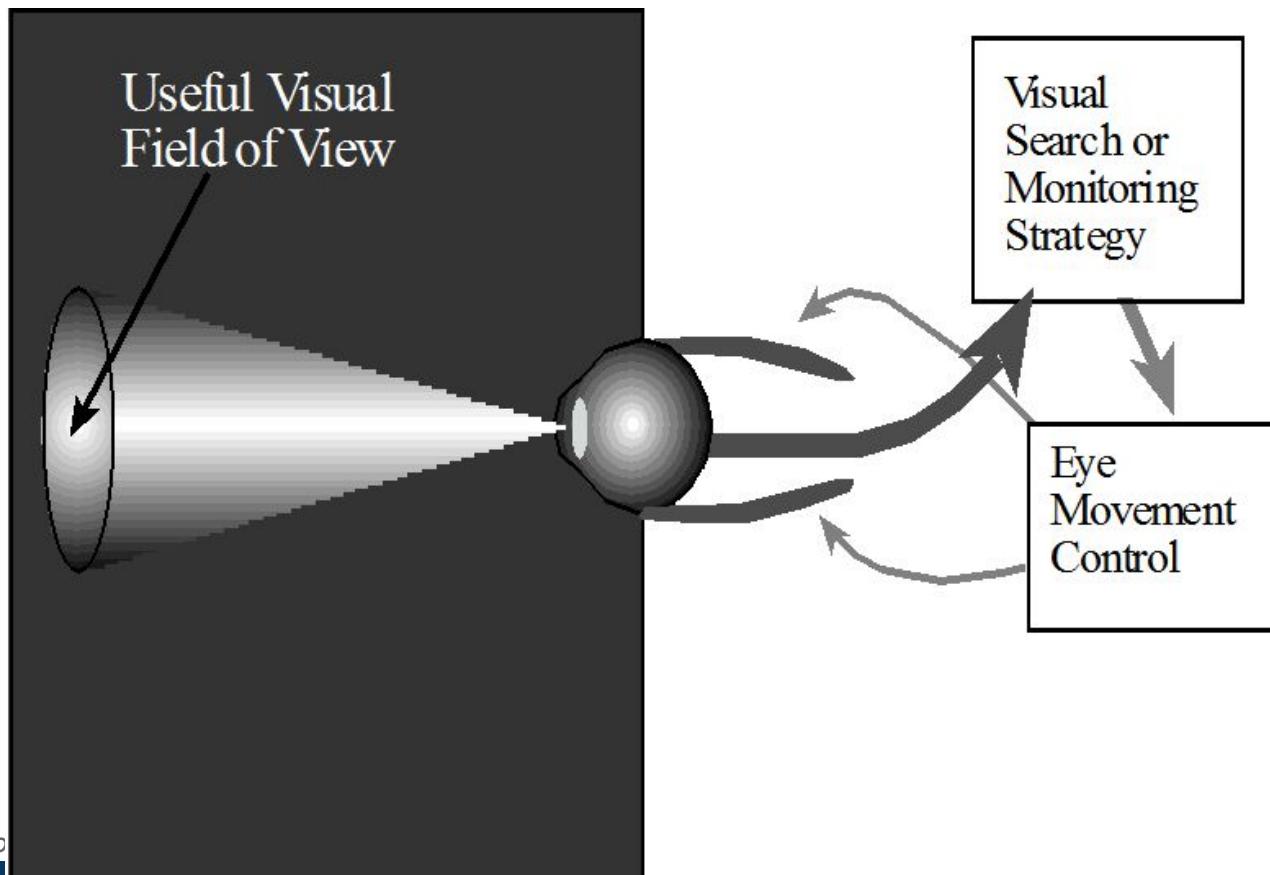
Depth Cues - Monocular

Can still perceive depth with one eye

- **Occlusion** – the blocking of more distant objects by closer objects (overlapping)
- **Relative size** – when viewing an object of known size (e.g., people) the brain compares the sensed size to the known size to estimate the distance of the object
- **Aerial haze** (texture) – object on the far horizon (e.g., mountains) look hazy due to particles in the air
- **Accommodation** – tension of the muscle that changes the focal length of the lens of the eye (weak cue)
- **Motion parallax** – similar to binocular parallax, by moving the head slightly the differences in the sensed images (even from 1 eye) can be used to judge depth.

<http://www.eruptingmind.com/depth-perception-cues-other-forms-of-perception/>

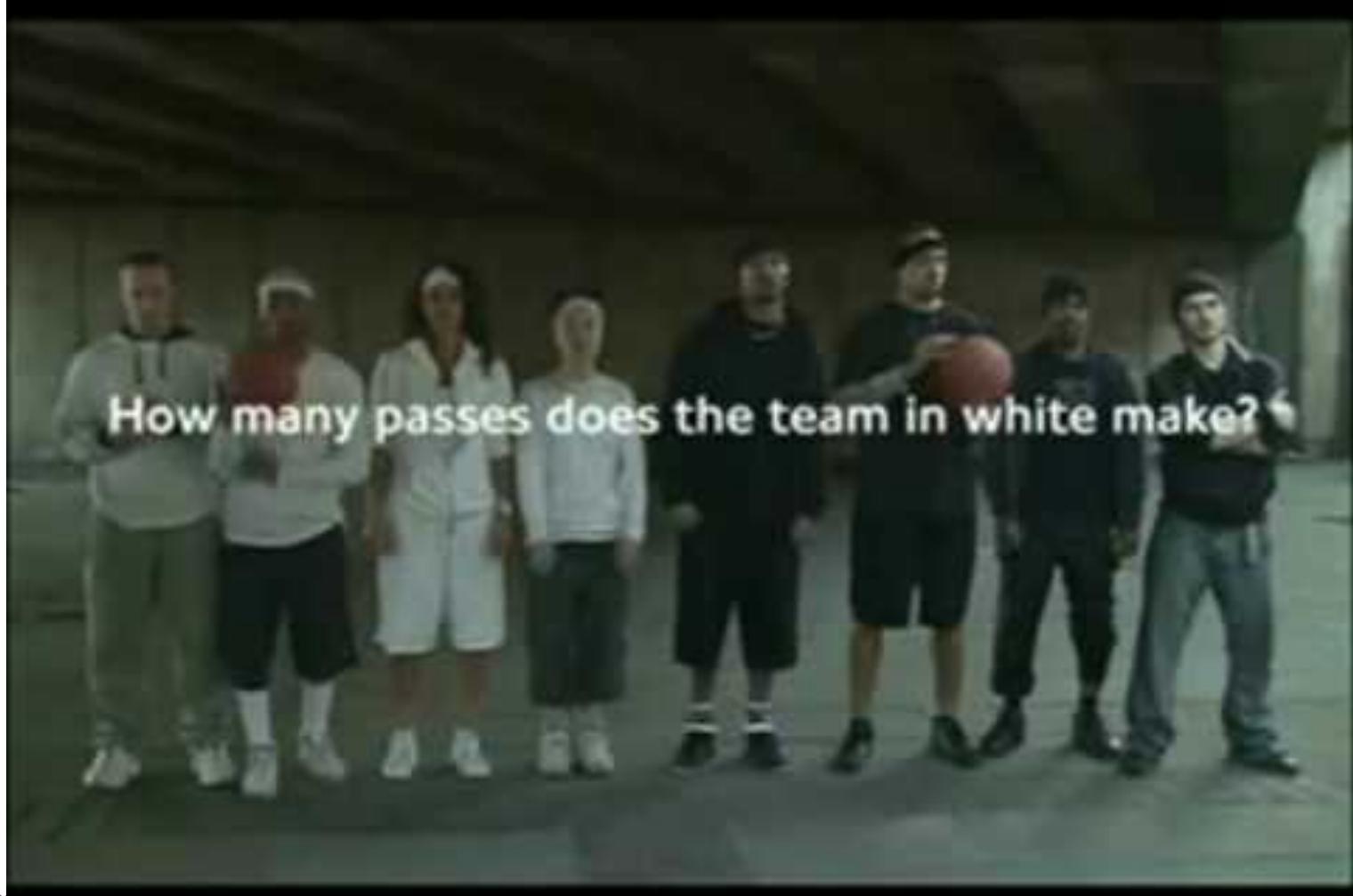
Attention - Searchlight model



Attention - Searchlight properties

- Searchlight Size varies with
 - Data density
 - Stress level (distraction, motivation)
- Attention operators work within searchlight beam
- Attention = Tunable Filter
- Eye movements 3/sec – series of saccades*
- **Popout** effects (general attention)
- Segmentation effects (dividing up the visual field)
→ Guide Attention

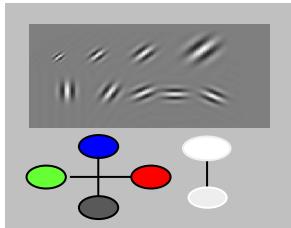
*a rapid movement of the eye between fixation points



Parallel Processes

Feature extraction:

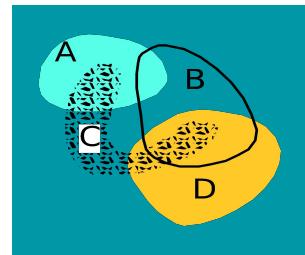
- Orientation
- Texture
- Colour
- Motion



Detection: Edges, Regions, 2D shape

Transitory state

Bottom-up, data driven



Serial Processes

Object recognition: visual attention & memory important.

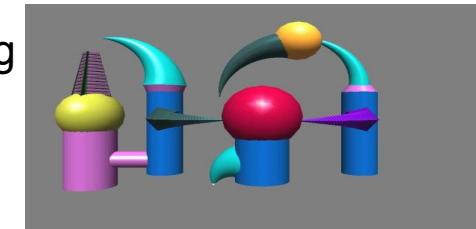
Uses both short-term memory and long-term memory

Short-term memory (chunking):

$$7 \pm 2 = 5 \text{ to } 9 \text{ Objects}$$

More emphasis on symbols

Top-down processing



Pre-attentive processing

Some visual properties detected very rapidly

< 200-250ms

When designing visualisations these features/properties are:

immediately perceived

can mislead the viewer

How many ‘2’s?

035219248730515
708029135238051
337920714842415
119665329034538

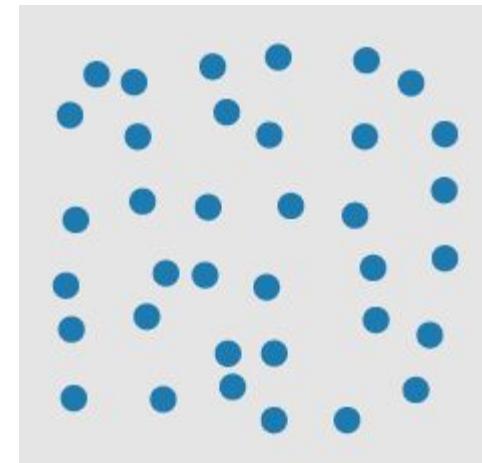
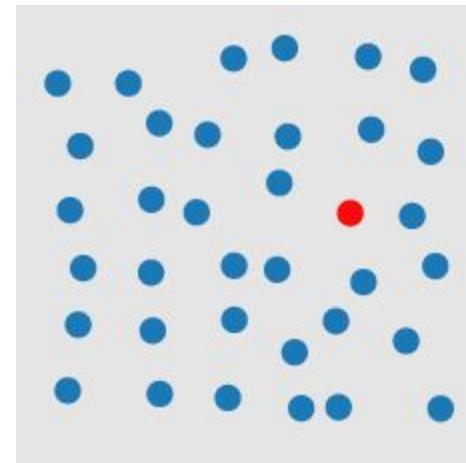
How many '8's?

035219248730515
708029135238051
337920714842415
119665329034538

Pre-attentive processing - Colour

Target (red circle)

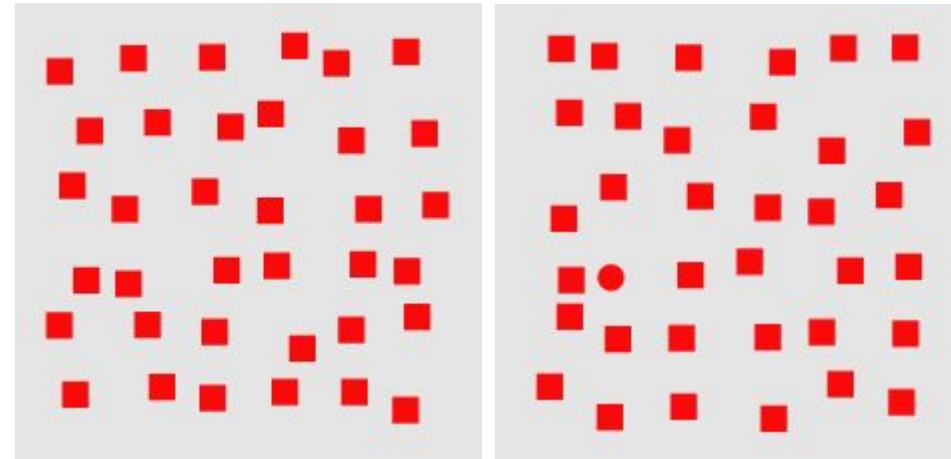
Distractors (blue circles)



Pre-attentive processing - Shape

Target (red circle)

Distractors (red squares)

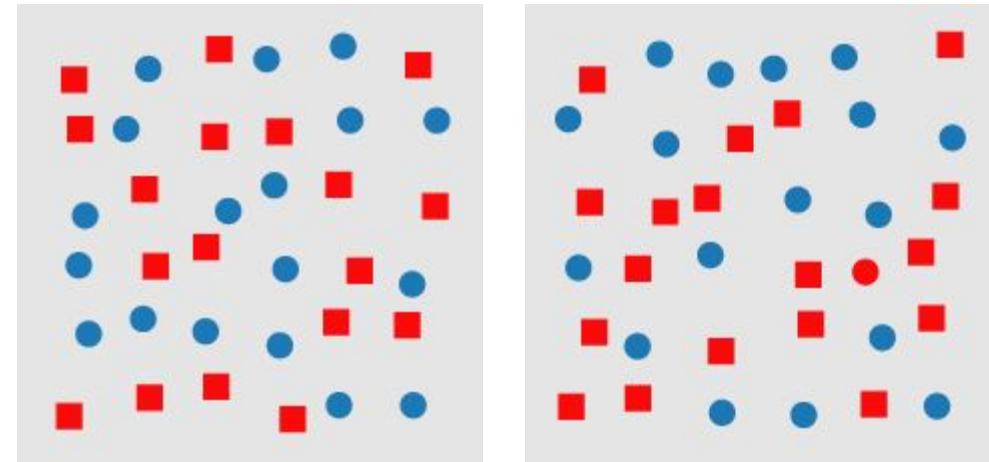


Pre-attentive processing - Conjunction Target

Cannot be detected
pre-attentively!

Target (red circle)

Distractors (blue circles & red squares)



Use Pre-attentive Processing

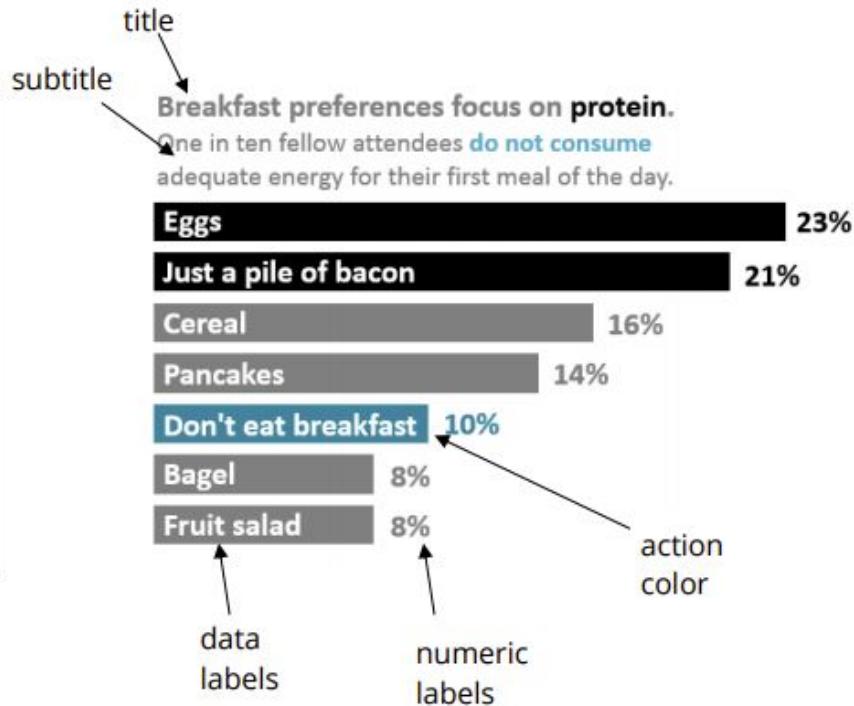
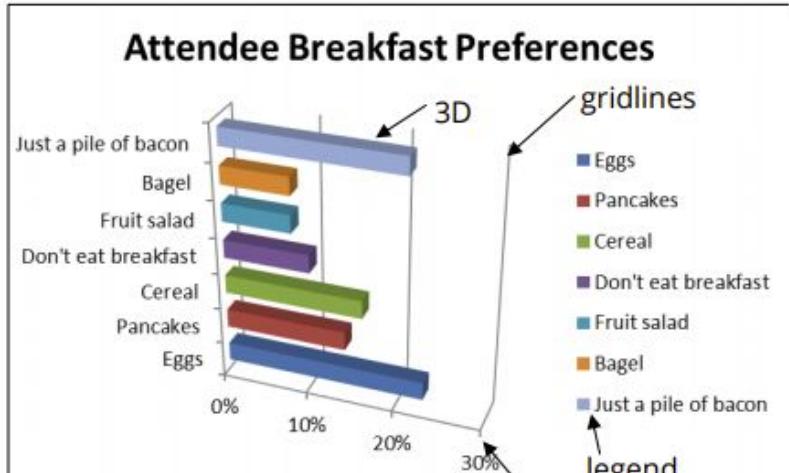
Target must stand out in simple dimension

- Color
- Simple Shape = orientation, size
- Motion
- Depth



<http://www.csc.ncsu.edu/faculty/healey/PP/index.html>

Using pre-attentive features



<http://stephanieevergreen.com/updated-data-visualization-checklist/>

2016 Berlin Marathon

<https://interaktiv.morgenpost.de/berlin-marathon-2016/>

References

How we see:

<http://sciencelearn.org.nz/Contexts/Light-and-Sight/Sci-Media/Video/How-the-eye-works>

In-depth Perception: <http://www.csc.ncsu.edu/faculty/healey/PP/index.html>

Pre-attentive processing: https://infovis-wiki.net/wiki/Preattentive_processing

Depth cues:

http://www.hitl.washington.edu/projects/knowledge_base/virtual-worlds/EVE/III.A.1.c.DepthCues.html

Today's Labs

Finish the Python Visualisation Libraries activity and the “[Notebook: Create a graph using Python](#)”

Review the other visualisation libraries: D3.js & Tableau (documents on loop)

Complete the replicate a graph exercise and upload your result to loop

Work on your visualisation assignment

09 Data Visualisation III - design

CA682

suzanne.little@dcu.ie

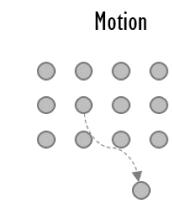
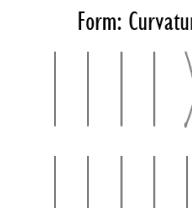
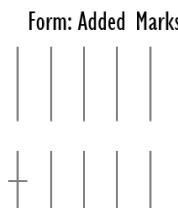
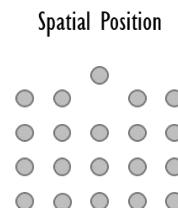
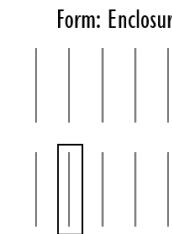
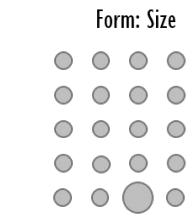
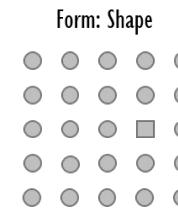
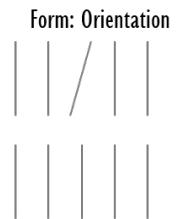
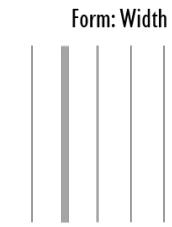
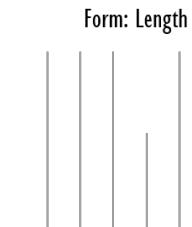
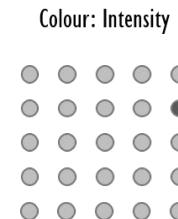
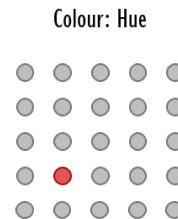
Today

Colour

Gestalt

Critique

Reminder: pre-attentive visual features



https://docs.google.com/document/d/1gcosfoduHt_VlGz3Q80taKxyMbBV0oDjTbWlfK3KcS8/edit?usp=sharing

Colour

Color Name	Hex Code RGB	Decimal Code RGB
Purple	800080	128,0,128
Fuchsia	FF00FF	255,0,255
Lime	00FF00	0,255,0
Teal	008080	0,128,128
Aqua	00FFFF	0,255,255

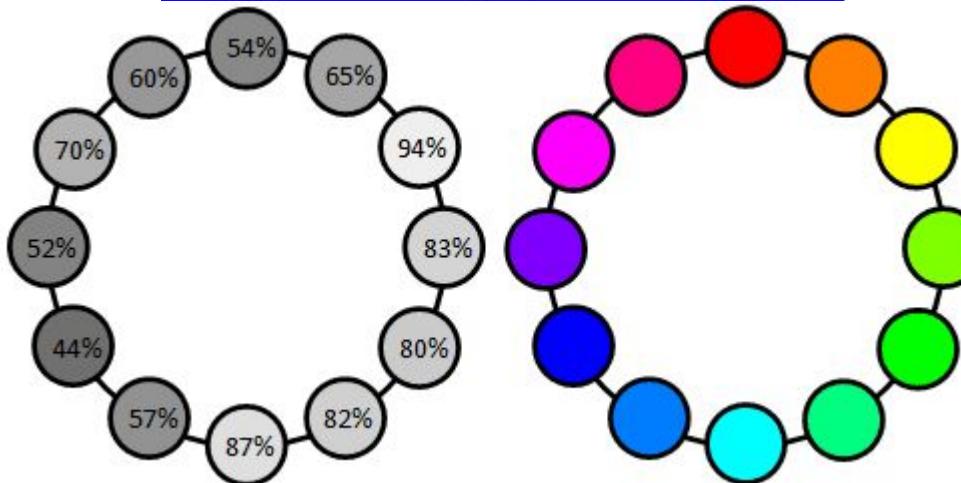
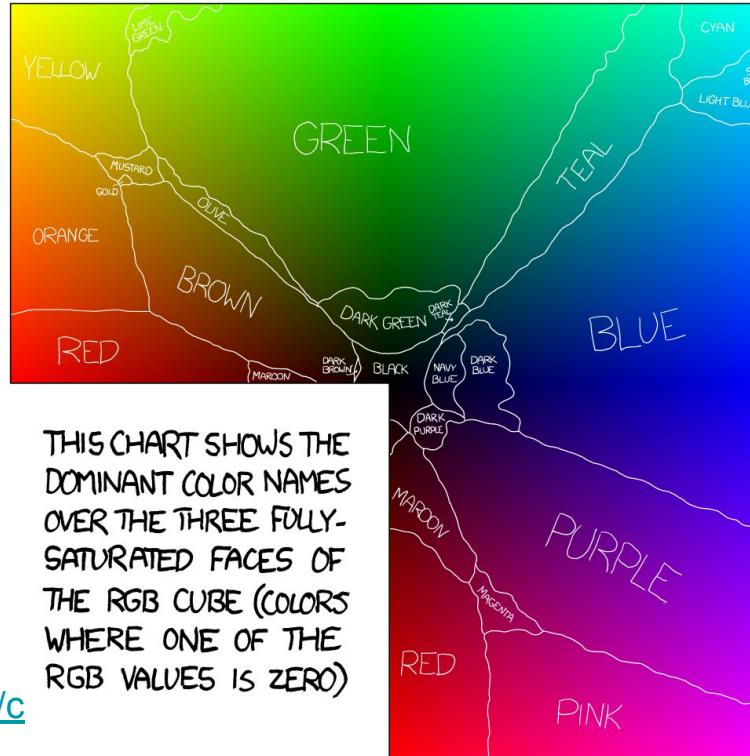


Fig: Luminance values of core colours (different hues) (from: <http://www.workwithcolor.com/color-luminance-2233.htm>)

Do you see the same blue as I do?

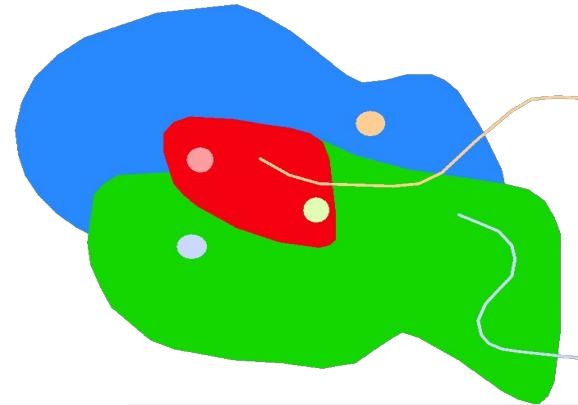
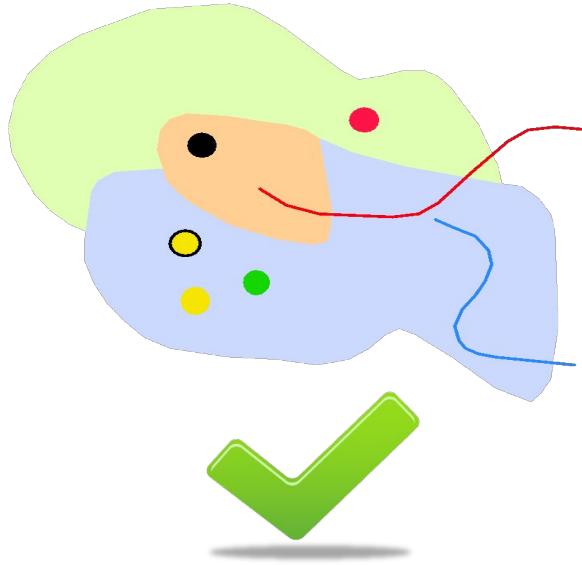


THIS CHART SHOWS THE DOMINANT COLOR NAMES OVER THE THREE FULLY-SATURATED FACES OF THE RGB CUBE (COLORS WHERE ONE OF THE RGB VALUES IS ZERO)

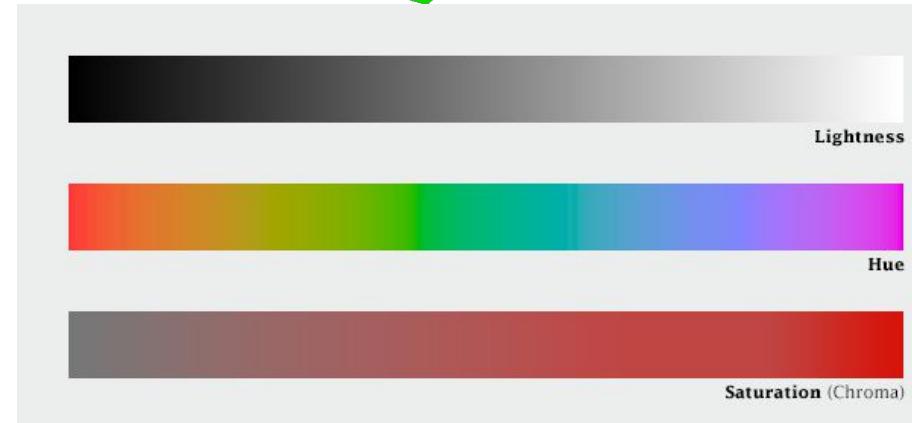
<https://blog.xkcd.com/2010/05/03/color-survey-results/>

Colour Coding

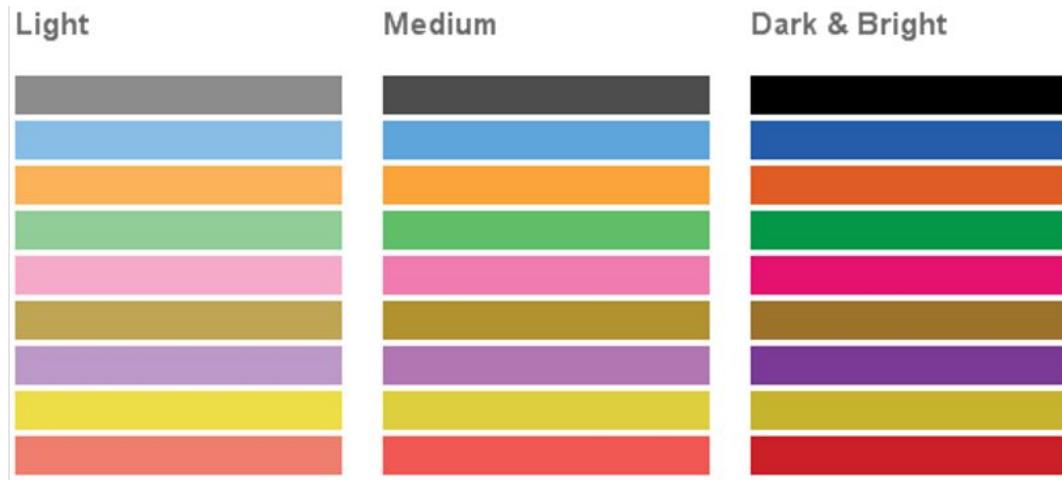
<https://color.method.ac/>



Large areas = low saturation
Small areas = high saturation



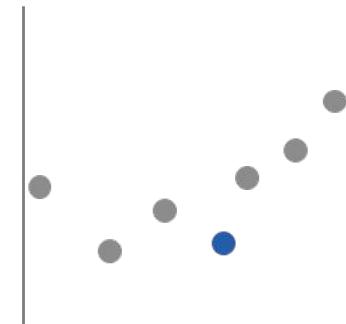
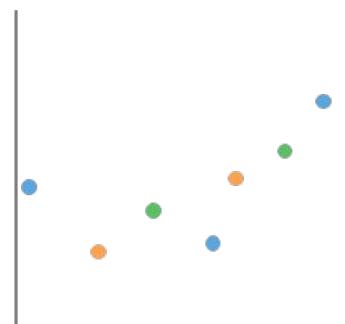
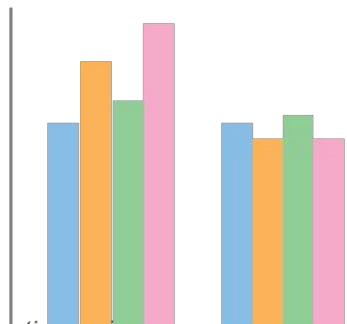
Colour Palettes suggested by Stephen Few



Luminance: higher
(whiter)
Large data encoding
objects

Luminance: mid-
range
Small data encoding
objects

Luminance: lower
(darker)
Highlight and draw
attention



Country Level Sales Rank Top 5 Drugs

Rainbow distribution in color indicates sales rank in given country from #1 (red) to #10 or higher (dark purple)

Country	A	B	C	D	E
AUS	1	2	3	6	7
BRA	1	3	4	5	6
CAN	2	3	6	12	8
CHI	1	2	8	4	7
FRA	3	2	4	8	10
GER	3	1	6	5	4
IND	4	1	8	10	5
ITA	2	4	10	9	8
MEX	1	5	4	6	3
RUS	4	3	7	9	12
SPA	2	3	4	5	11
TUR	7	2	3	4	8
UK	1	2	3	6	7
US	1	2	4	3	5

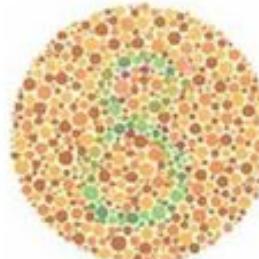
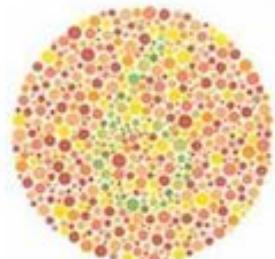
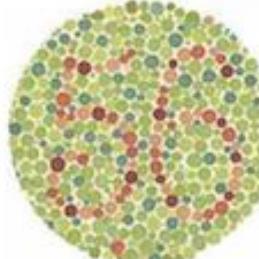
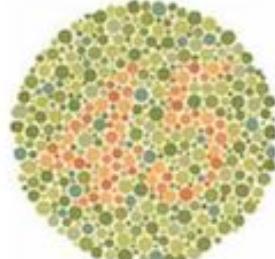
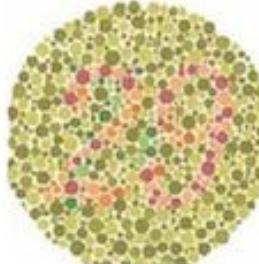
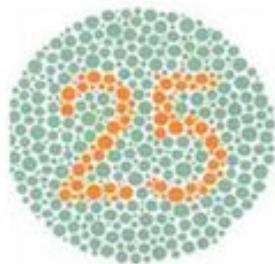
Top 5 drugs: country-level sales rank

RANK	1	2	3	4	5+
COUNTRY DRUG	A	B	C	D	E
Australia	1	2	3	6	7
Brazil	1	3	4	5	6
Canada	2	3	6	12	8
China	1	2	8	4	7
France	3	2	4	8	10
Germany	3	1	6	5	4
India	4	1	8	10	5
Italy	2	4	10	9	8
Mexico	1	5	4	6	11
Russia	4	3	7	9	12
Spain	2	3	4	5	11
Turkey	7	2	3	4	8
United Kingdom	1	2	3	6	7
United States	1	2	4	3	5

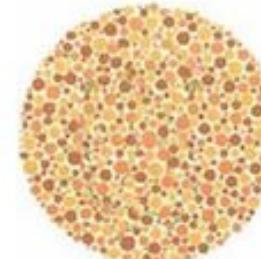
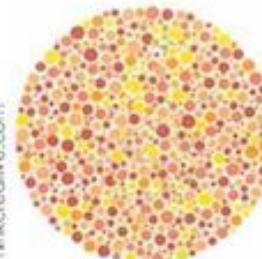
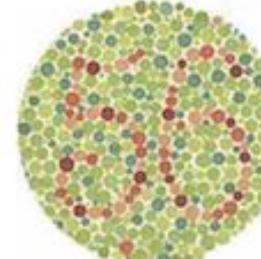
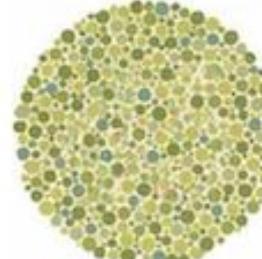
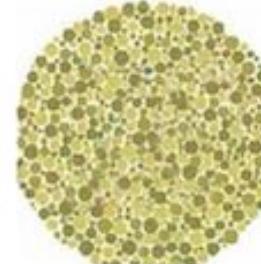
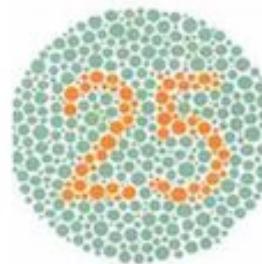
Fig: Example of how restraint in use of colour improves clarity (from Knaflic, 2015)

Ishihara Test For Color Blindness

What People With Regular Vision See



What Red-Green Color Blind People See



Colour - Take home messages

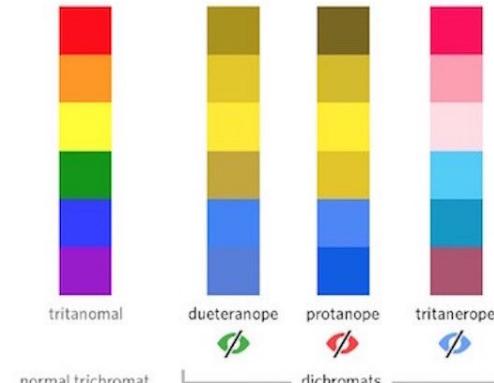
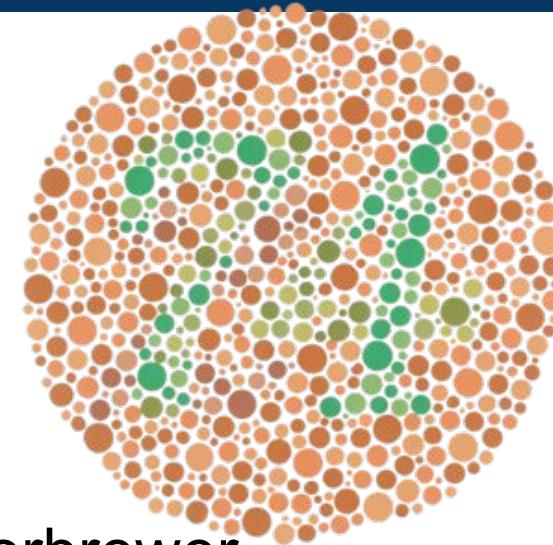
- Thoughtfully & Sparingly
- Vary the **Luminance** for Detail, Shape and Form
- Use colour **Hue** attribute for **Categorization** - few colours
- **Strong** (high saturation) colours for Small Areas
- **Contrast** in luminance with background
- **Subtle** colours (low saturation) for Large Areas

Checkout: <http://contrastrebellion.com>



Colour - Take home messages

- Use design rules to choose good colours
 - should it be viewable in B&W?
 - check for colour-blindness
- Tools like paletton (paletton.com) and Colorbrewer (colorbrewer2.org) can help

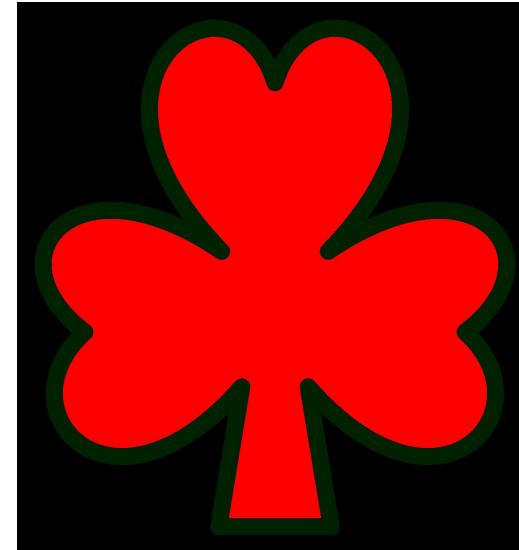


Document on Using Colour:

<https://docs.google.com/document/d/119nETfGiWly8VGcJhOg>

[n=sharing](#)
Suzanne Little, School of Computing, DCU

What else is important about colour?



Emotional and cultural meanings -

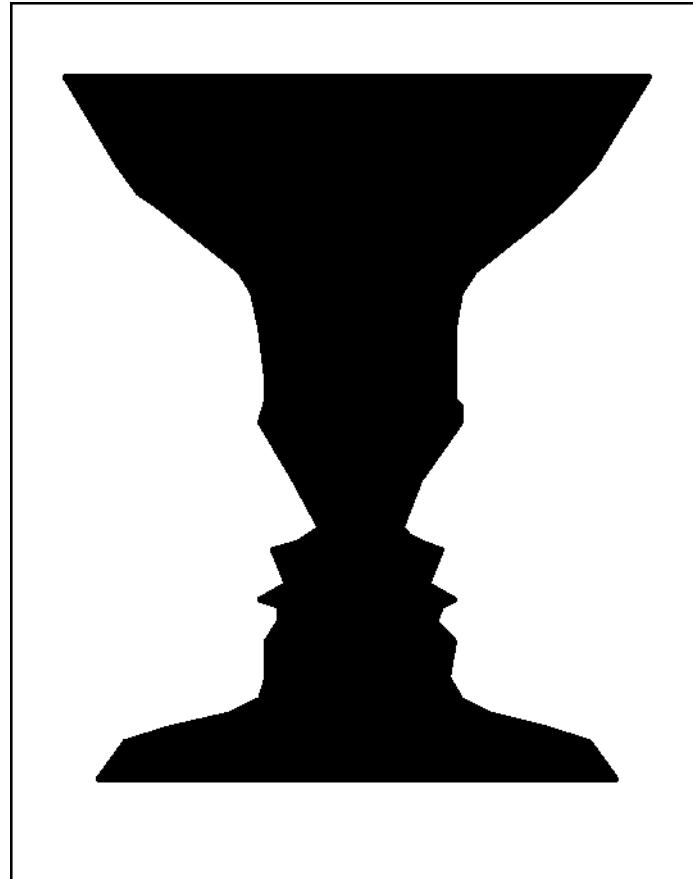
<https://www.informationisbeautiful.net/visualizations/colours-in-cultures/>

Gestalt Laws

“Unified whole” - theory of visual perception

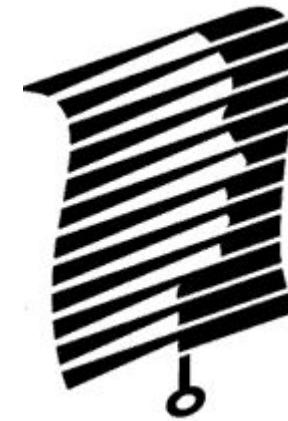
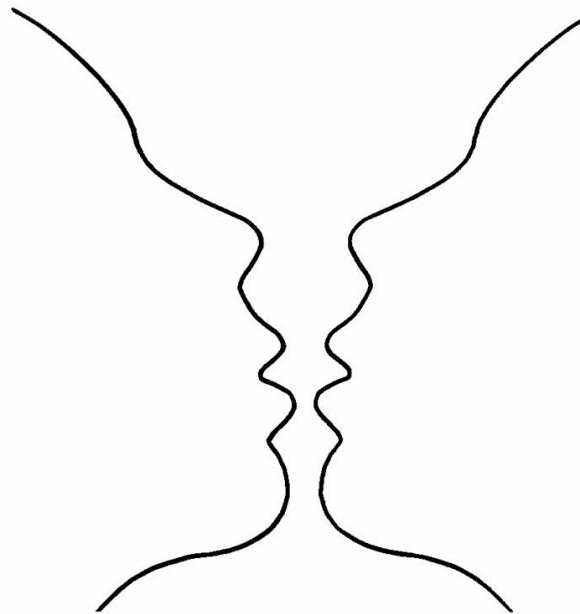
Max Westheimer, Kurt Koffka and Wolfgang Kohler (1912)

- Proximity
- Similarity
- Enclosure
- Closure
- Continuity
- Connection



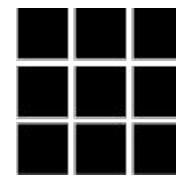
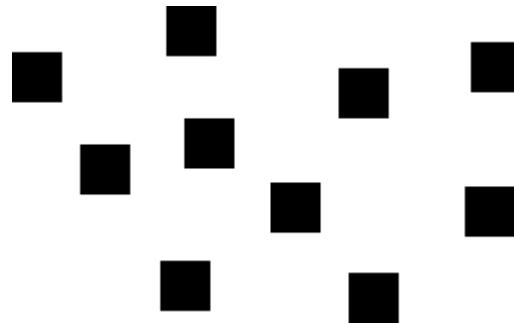
Gestalt - Figure & Ground

Differentiating an object (figure) from its surrounding area (ground)



Gestalt - Proximity

Elements placed together are seen as a whole

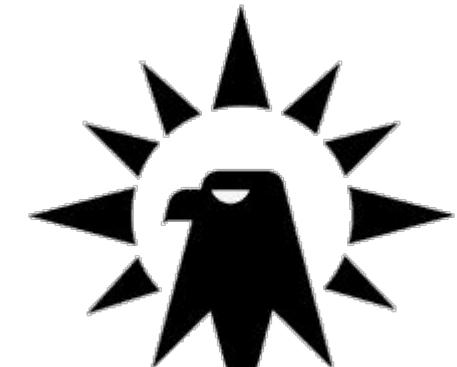
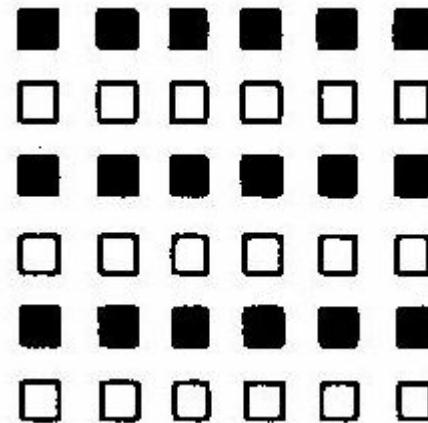


Gestalt - Similarity

Elements that are similar to each other are seen as a whole

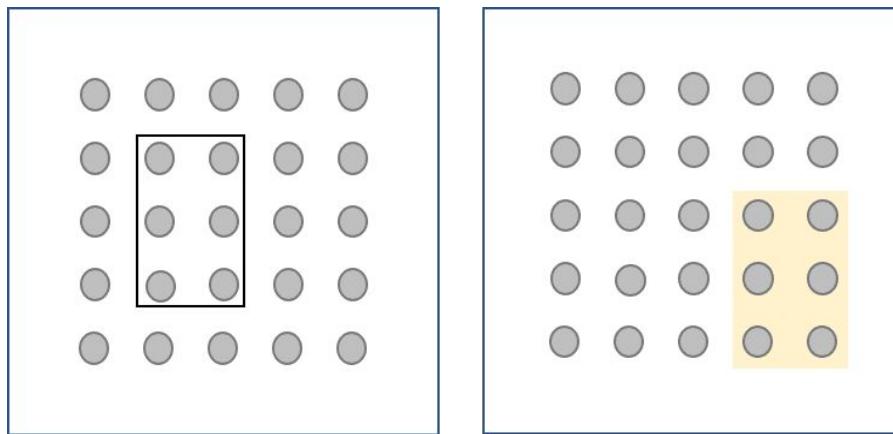


Anomaly



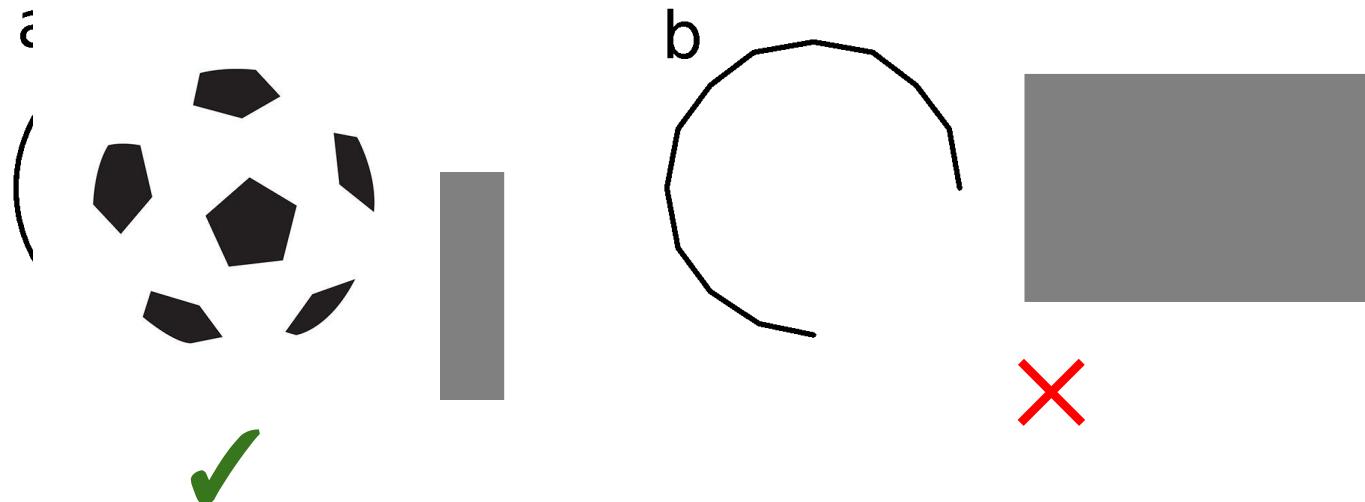
Gestalt - Enclosure

Objects that appear to have a boundary around them are perceived as a group



Gestalt - Closure

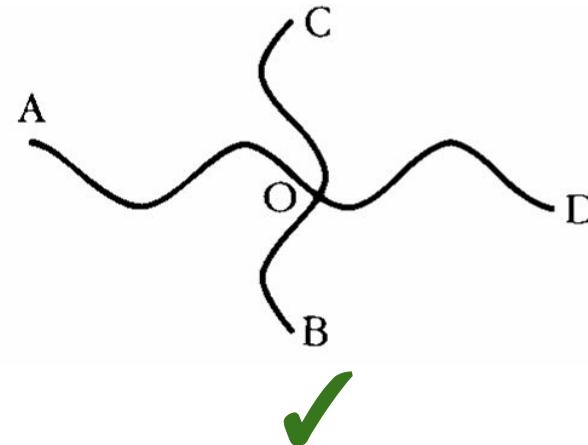
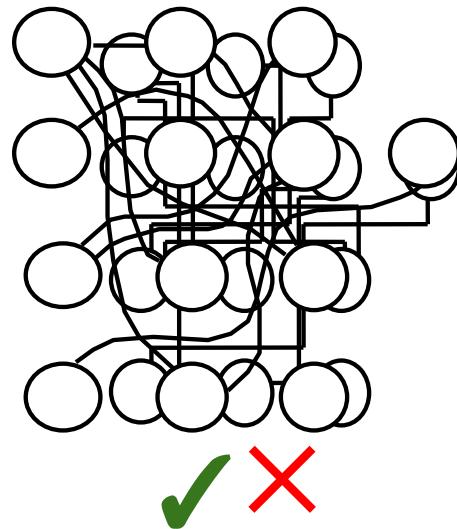
A boundary with no interruptions re-enforces the figure and enhances grouping



Note: People are very good at completing the boundary with enough indicators!

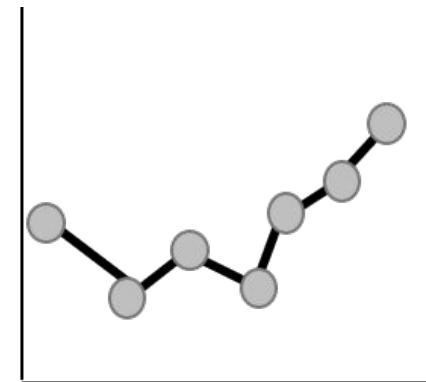
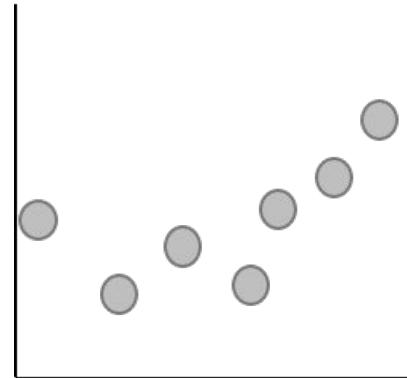
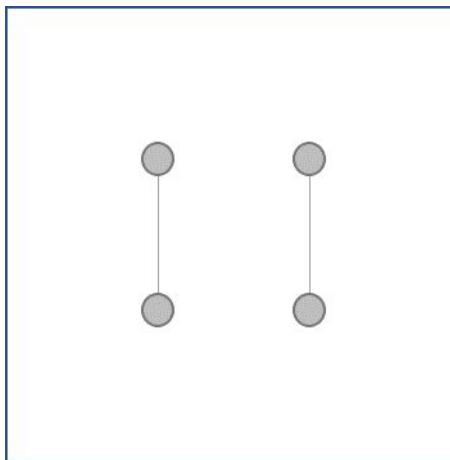
Gestalt - Continuity

Sequential and/or continuous lines enhance grouping of elements



Gestalt - Connection

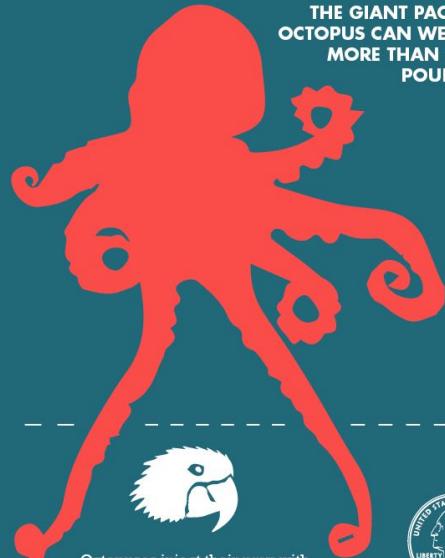
Objects that are connected (e.g., by a line) are perceived as a group. Similar to enclosure we can draw an explicit link between objects using lines to connect them.



Summary: Why does how we see matter?

- Making most of visual cues [preattentive attributes]
- Not overloading the visual information channel [cognitive processing load]
- Not excluding or miscommunicating with your audience [colour, visibility and accessibility]
- The problems with simulating effective depth, motion and using 3D effects in a 2D visualisation [depth perception]
- Engaging attention [use preattentive attributes and gestalt principles]
- Not diluting your message [distraction, gestalt principles]

WORLD OCTOPUS DAY



THE GIANT PACIFIC OCTOPUS CAN WEIGH MORE THAN 600 POUNDS

Octopuses inject their prey with venom using a beak similar to a bird's made from the same tough material as a lobster shell.



ALL SPECIES ARE VENOMOUS, BUT THE BLUE-RINGED OCTOPUS IS THE ONLY ONE DANGEROUS TO HUMANS, RESPONSIBLE FOR AT LEAST TWO DEATHS.

OCTOPUSES VS. OCTOPI

THE PLURAL IN ENGLISH IS "OCTOPUSES," BUT THE GREEK PLURAL FORM "OCTOPODES" IS SOMETIMES USED. "OCTOPI," WHILE COMMONLY USED, IS CONSIDERED INCORRECT.



OCTOPUSES ARE ABOUT 90% MUSCLE

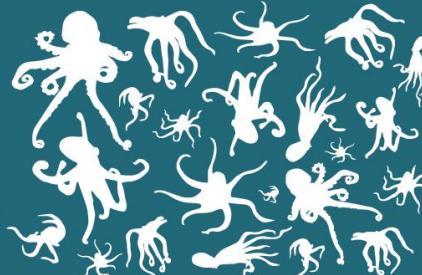
THE GIANT PACIFIC OCTOPUS CAN INHABIT DEPTHS OF UP TO 5,000 FEET



A mature female octopus can have up to 280 suckers on each arm! Each sucker contains thousands of chemical receptors, with sensitivities to both touch and taste.

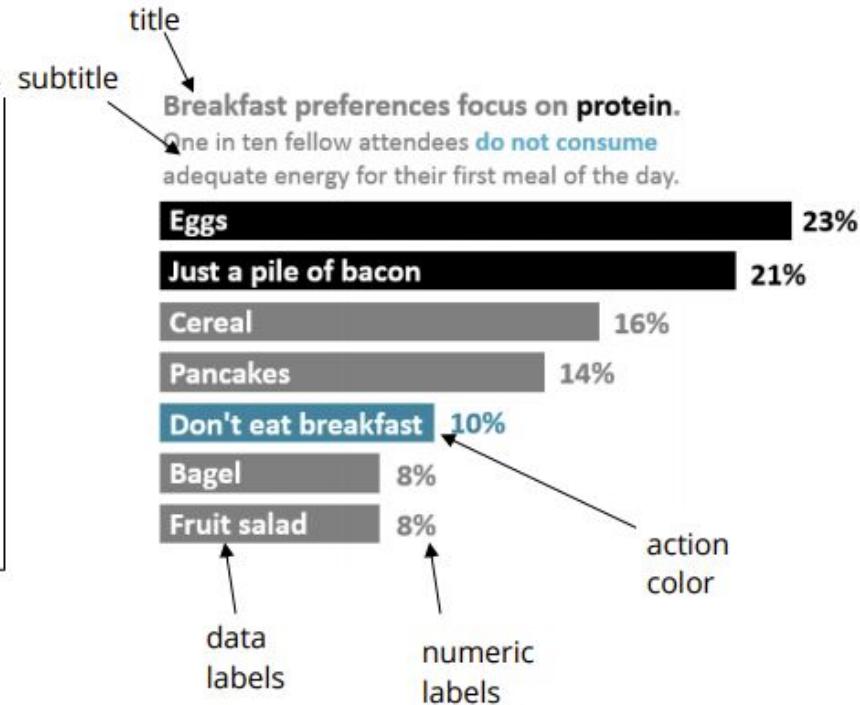
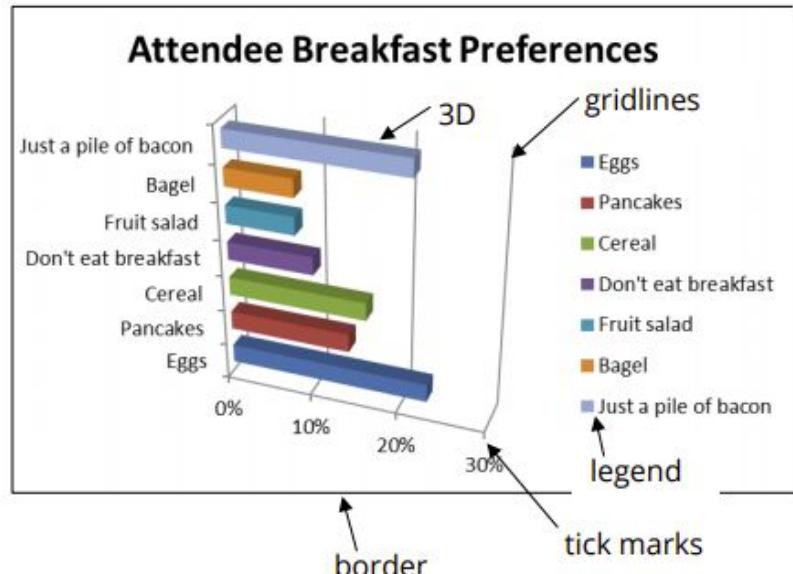
OCTOPUSES CAN QUICKLY CHANGE THE COLOR AND TEXTURE OF THEIR SKIN

300
RECOGNIZED
SPECIES
OF OCTOPUS



NATIONAL AQUARIUM | [Facebook](#) [Twitter](#) [Instagram](#) [YouTube](#) [Flickr](#) [Tumblr](#) [Pinterest](#) [RSS](#) | aqua.org

Evergreen: Data Visualisation Checklist



<http://stephanieevergreen.com/updated-data-visualization-checklist/>

Suzanne Little, School of Computing, DCU

Resources

<https://youtu.be/LIzuJqZ797U> -- gestalt principles

<https://www.stat.auckland.ac.nz/~ihaka/120/Lectures/lecture03.pdf> -- good bad graphs

<http://sciencelearn.org.nz/Contexts/Light-and-Sight/Sci-Media/Video/How-the-eye-works>

S. Few, “Practical rules for using colors in charts” (2008)

http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf

<http://www.informationisbeautiful.net/visualizations/what-makes-a-good-data-visualization/>

<http://www.csc.ncsu.edu/faculty/healey/PP/index.html>

http://www.infovis-wiki.net/index.php/Preattentive_processing

<http://www.users.totalise.co.uk/~kbroom/Lectures/gestalt.htm>

<http://graphicdesign.spokanefalls.edu/tutorials/process/gestaltprinciples/gestaltpri n c.htm>

http://www.hitl.washington.edu/projects/knowledge_base/virtual-worlds/EVE/III.A.1_c.DepthCues.html

10 Data Management

CA682

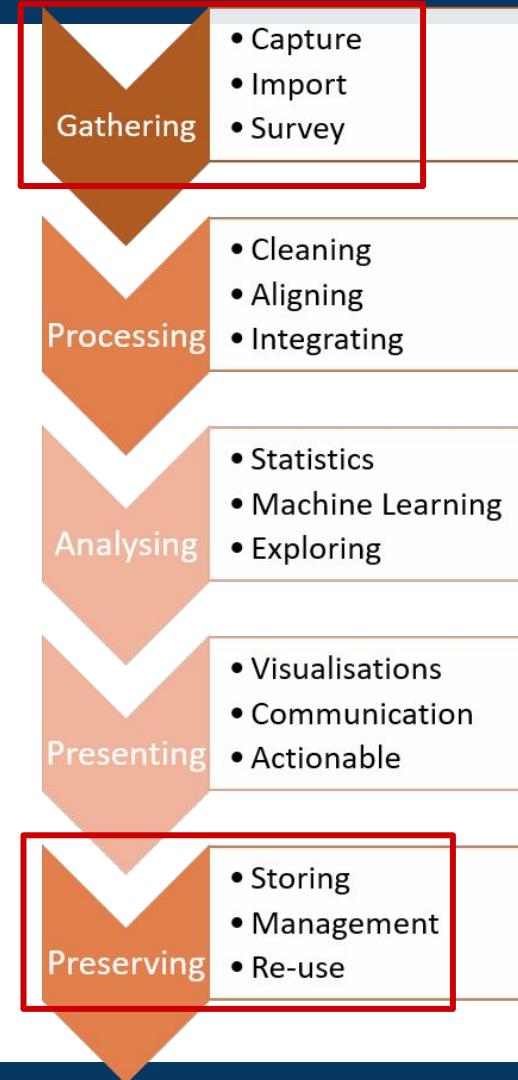
suzanne.little@dcu.ie

Today

- Comments on assignment (after break)
- Data Storage overview
- What storage approach should you use for ...?

Data

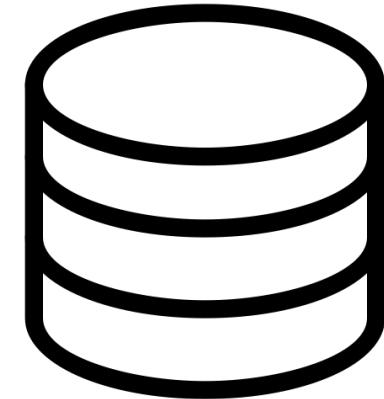
- Data is collected information
- Where does data come from?
 - Files
 - The Internet
 - Databases



Data storage some options

Database management tools

Why? Data persistence



Find relevant information quickly

Select required data subset for further analysis

Apply analysis function to (large, distributed?) datasets

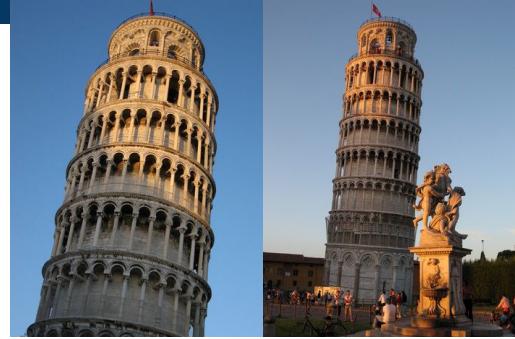
Calculate regular reports as data is updated

Ensure **consistency** and **protection** of data

Scenario: my photo collection

Data consists of:

- CSV file with list of photo id, path to EXIF file & path to jpg file
 - Folder with JPG & EXIF files organised into subfolders of year & event name (e.g, 2018/Italy_trip_2018, 2006/Pisa2006, 2007/AustChristmas)
1. How do I find the photo of the leaning tower taken Oct 2006?
 2. How do I delete the photo of the leaning tower taken Oct 2006?
 3. How do I find all photos of the leaning tower of Pisa?



What makes a “database”?

1. Structure - tables, documents, “chunks”
2. Minimise redundancy - efficient storage, normalisation
3. Maintain consistency - updates, transactions, deletions
4. Multiple user and concurrent access
5. Query options - eg. language like SQL, SPARQL, Cypher
6. Security

Data storage approaches

Database: structured set of data that can be accessed, managed and updated (easily)

1. Relational (traditional & modern)
2. Column
3. MPP, Data Warehouse
4. NoSQL
5. Big Data (MapReduce, Hadoop - HDFS)

1 Relational DBs (the all-purpose solution for not-that-big data)

- Structured according to the relationship between data
- Tables, Records and Columns
- Relationship facilitates searching, organisation & reporting

Consider:

- “adequate for all tasks but not excellent at any of them” - ?
- easy to use
- low resource requirements
- well-supported by all software
- familiar
- not suitable for really big data

Eg: Oracle, PostgreSQL,
MySQL/MariaDB, sqlite, IBM
DB2, Microsoft SQL Server

2 Columnar stores?

- inversion of a row store: indexes become data & data becomes indexes
- for aggregations and transformations of highly structured data
- good for BI, analytics, some archiving but not data mining
- moderately big data (0.5-100TB) → compression
- slow to add new data / purge data
- Eg: Cassandra, Bigtable, HBase, PostgreSQL (option)

<https://database.guide/what-is-a-column-store-database/>

3 DW & MPP

Data Warehouse (**DW**)

- a centralized repository
- stores data from multiple information sources
- transformed into a common, multidimensional data model
- efficient querying and analysis

https://www.datawarehouse4u.info/index_en.html

Massively Parallel Processing Database (**MPP**)

- optimized to be processed in parallel
- many operations performed by many processing units at a time.

OLAP, OLTP, DW ??

OLTP: On-Line Transactional Processing

Operations: INSERT, UPDATE, DELETE

OLAP: On-Line Analytical Processing

Information: complex analytics, aggregations, batch

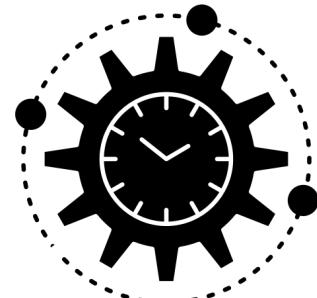
DW: Data Warehouse



OLTP



OLAP



OLTP

vs Data Warehouse/OLAP

- many single-row writes
- current data
- queries generated by user activity
- < 1s response times
- 1000's of users

- few large batch imports
- years of data
- queries generated by large reports
- queries can run for minutes/hours
- 10's of users

OLTP

vs

Data Warehouse/OLAP

big data
for many
concurrent
requests to
small amounts
of data each

big data
for low
concurrency
requests to very
large amounts
of data each

4 NoSQL

- Non-relational or sometimes “not only SQL” -
<https://en.wikipedia.org/wiki/NoSQL>
- Eg: key-value, document, object or graph-based data stores
- Eg: MongoDB, Solr, HBase, Splunk, Neo4j
- Why?
 - Large volumes of structured, semi-structured, and unstructured data
 - Quick iteration
 - Efficient, scale-out architecture instead of expensive, monolithic architecture

<https://www.devbridge.com/articles/benefits-of-nosql/>

5 Map/Reduce

Data is big

Readily available computing power is small to moderate (because of cost)

Disk space is relatively cheap (although working memory can be a constraint)

How to construct algorithms to make the most of resources?

Before Map/Reduce

Large-scale data processing was difficult:

- Managing hundreds or thousands of processors
- Managing parallelization and distribution
- I/O Scheduling
- Status and monitoring
- Fault/crash tolerance

So what is Map/Reduce

Programming paradigm, used by Google, for processing and generating large data sets with a parallel, distributed algorithm on a compute cluster.

- Data processing broken up in Map and Reduce stages
- Popular implementation is Apache Hadoop
- Key contributions:
 - many small machines tackle jobs not possible even with large ones
 - scalability
 - fault-tolerance
 - execution optimisation

Twinkle, twinkle,
little star

MAP

twinkle 2
little 1
star 1

How I wonder
what you are

MAP

how 1
I 1
wonder 1
what 1
you 1
are 1

Up above the
world so high

MAP

up 1
above 1
the 1
world 1
so 1
high 1

Like a diamond
in the sky

MAP

like 1
a 1
diamond 1
in 1
the 1
sky 1



Summary

- MapReduce enabled distributed/parallel processing of big data on commodity hardware (cheap, scalable, reliable)
- Apache Hadoop is a collection of supporting tools for Big Data Analytics
- Approach Big Data differently
 - Don't always gather → clean → index → store (or the traditional ETL operation)
 - Rather, ingest (continuously) and process at point-of-need, move computation to many copies of data (redundancy)

What about elasticsearch and ELK?

- Elasticsearch: distributed search (& analytics) engine
- Ingest JSON files, full text search options → document
- AWS and Apache Lucerne
- Used for log analytics, full-text search, security intelligence, business analytics, and operational intelligence

What about elasticsearch and ELK?

ELK stack or Elastic stack - <https://www.elastic.co/elk-stack>

- Elasticsearch, Logstash, and Kibana. Plus Beats.
- Logstash: server-side data processing pipeline that ingests data from multiple sources
- Kibana: visualize data with charts and graphs
- Beats: lightweight data processes ('tail log file')
- Modern form of Data Warehousing for documents?

Actual databases ...

<http://db-engines.com/en/ranking>

416 systems in ranking, November 2023

Rank			DBMS	Database Model	Score		
Nov 2023	Oct 2023	Nov 2022			Nov 2023	Oct 2023	Nov 2022
1.	1.	1.	Oracle	Relational, Multi-model	1277.03	+15.61	+35.34
2.	2.	2.	MySQL	Relational, Multi-model	1115.24	-18.07	-90.30
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	911.42	+14.54	-1.09
4.	4.	4.	PostgreSQL	Relational, Multi-model	636.86	-1.96	+13.70
5.	5.	5.	MongoDB	Document, Multi-model	428.55	-2.87	-49.35
6.	6.	6.	Redis	Key-value, Multi-model	160.02	-2.95	-22.03
7.	7.	7.	Elasticsearch	Search engine, Multi-model	139.62	+2.48	-10.70
8.	8.	8.	IBM Db2	Relational, Multi-model	136.00	+1.13	-13.56
9.	9.	↑ 10.	SQLite	Relational	124.58	-0.56	-10.05
10.	10.	↓ 9.	Microsoft Access	Relational	124.49	+0.18	-10.53
11.	11.	↑ 12.	Snowflake	Relational	121.00	-2.24	+10.84
12.	12.	↓ 11.	Cassandra	Wide column, Multi-model	109.17	+0.34	-8.96

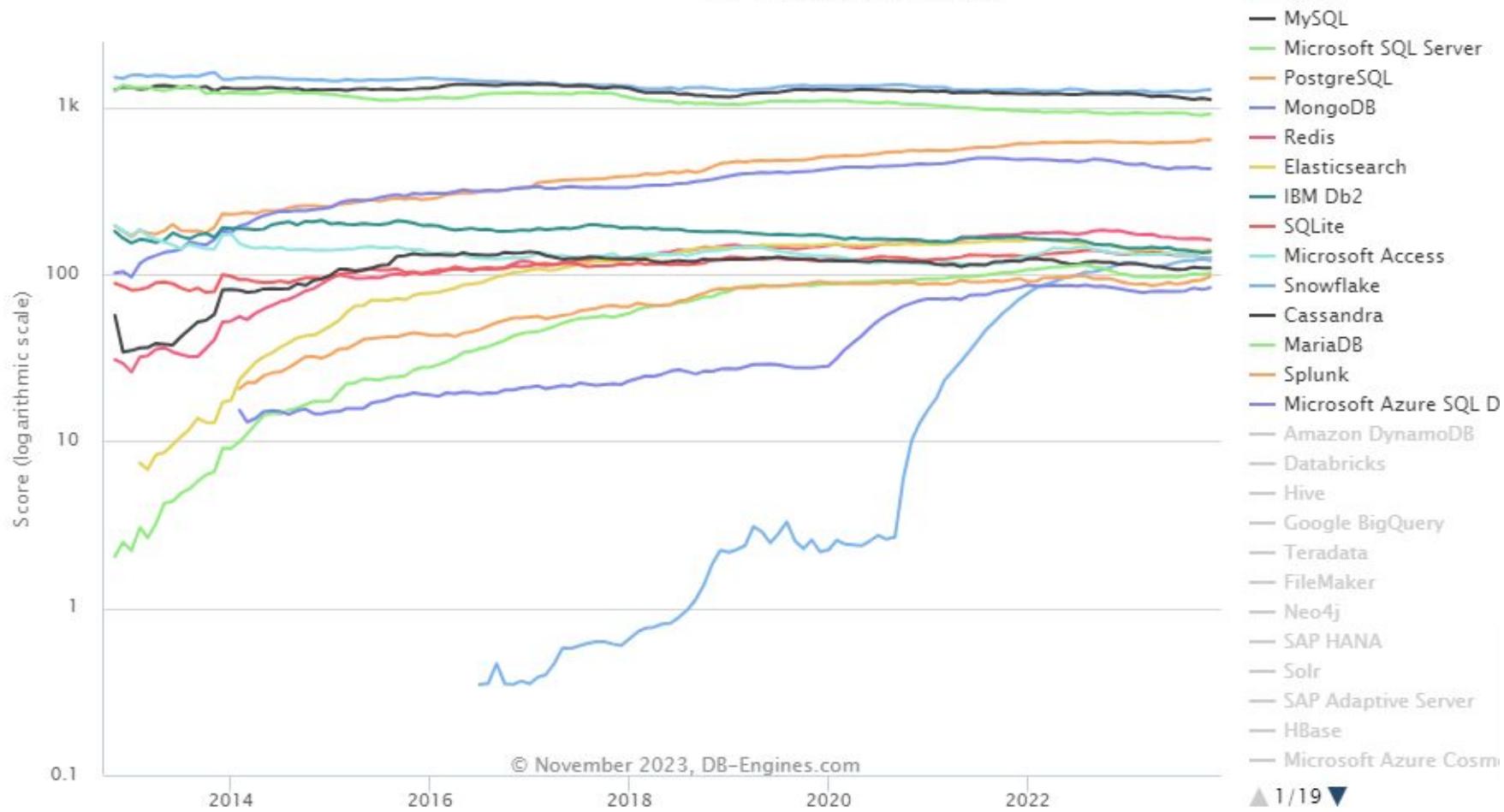
Actual databases ...

<http://db-engines.com/en/ranking>

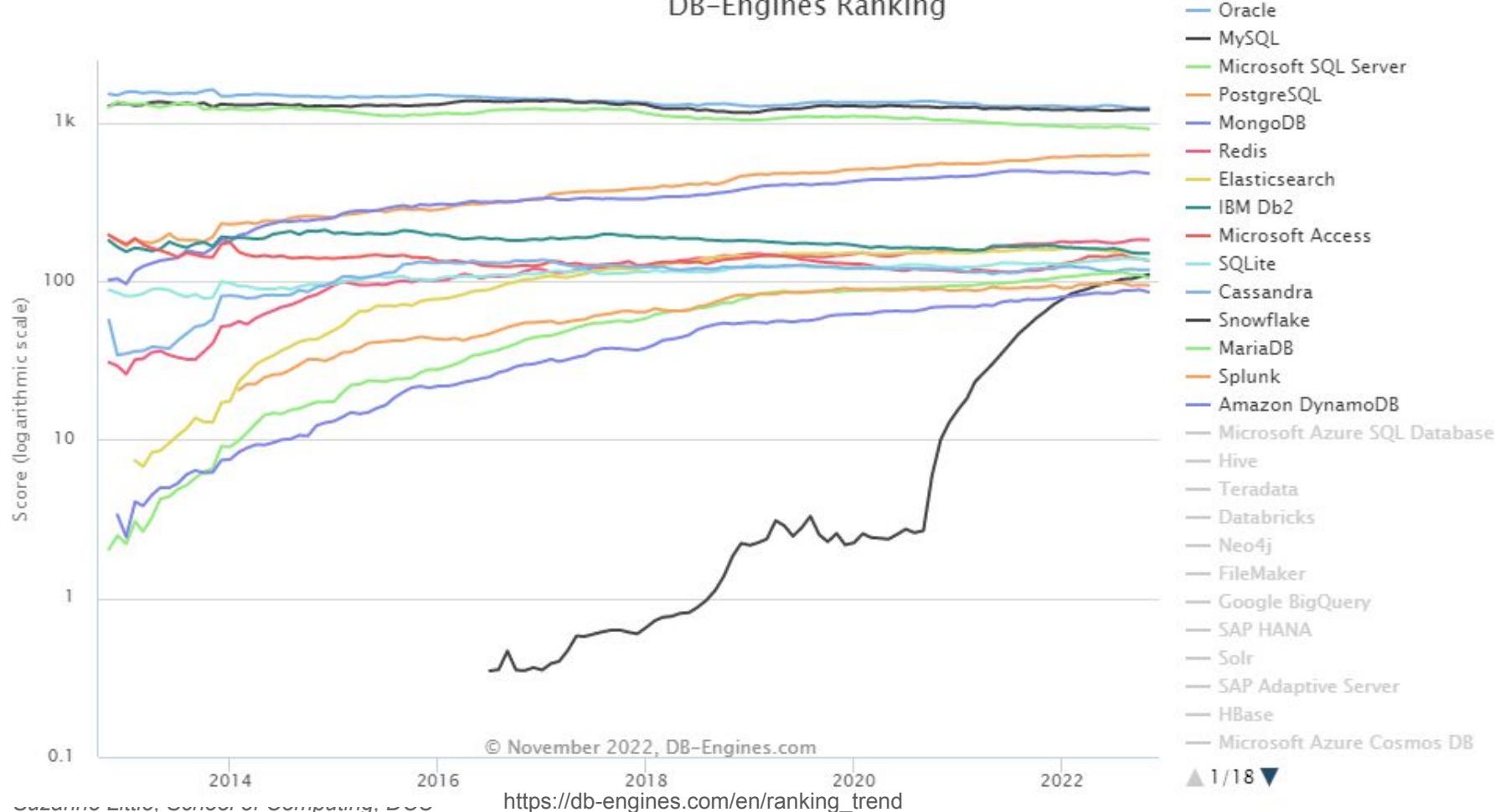
397 systems in ranking, November 2022

Rank			DBMS	Database Model	Score		
Nov 2022	Oct 2022	Nov 2021			Nov 2022	Oct 2022	Nov 2021
1.	1.	1.	Oracle	Relational, Multi-model	1241.69	+5.32	-31.04
2.	2.	2.	MySQL	Relational, Multi-model	1205.54	+0.17	-5.98
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	912.51	-12.17	-41.78
4.	4.	4.	PostgreSQL	Relational, Multi-model	623.16	+0.44	+25.88
5.	5.	5.	MongoDB	Document, Multi-model	477.90	-8.33	-9.45
6.	6.	6.	Redis	Key-value, Multi-model	182.05	-1.33	+10.55
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model	150.32	-0.74	-8.76
8.	8.	↓ 7.	IBM Db2	Relational, Multi-model	149.56	-0.10	-17.96
9.	9.	↑ 11.	Microsoft Access	Relational	135.03	-3.14	+15.79
10.	10.	↓ 9.	SQLite	Relational	134.63	-3.17	+4.83
11.	11.	↓ 10.	Cassandra	Wide column	118.12	+0.18	-2.76
12.	↑ 13.	↑ 18.	Snowflake	Relational	110.15	+3.43	+45.97

DB-Engines Ranking



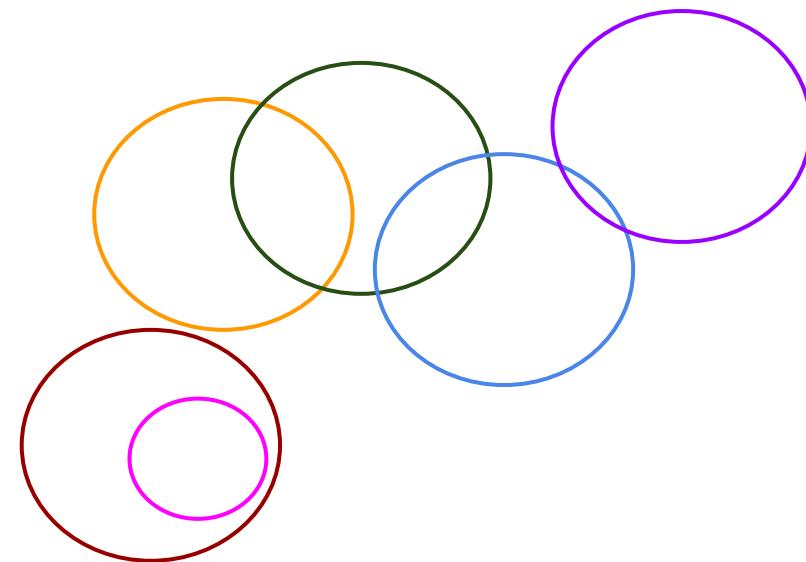
DB-Engines Ranking



Data storage approaches

Database: structured set of data that can be accessed, managed and updated (easily)

1. Relational (traditional & modern)
2. Column
3. MPP, Data Warehouse
4. NoSQL
5. Big Data (MapReduce, Hadoop)



Database management - Some questions to ask

1. How much data do I have now? What rate will I get new data?
2. Is the data structured? What format is the data?
3. Does the data need to be processed before loading?
4. How many queries will be run? Will they be concurrent? How many users?
5. What questions will the users be trying to answer? Do I know these questions?
6. Do I need to perform complex calculations on the data?
7. Do I have metadata or catalogue information? Is there a domain standard I can use?

Exercise - Which data storage method?

1. Sales and Customer data for a Small-to-Medium-Enterprise (SME)
2. Website logs, audience profiles, content generation for a media organisation
3. Building Management System - CCTV, power, floor plans, energy usage, work rosters, emergency plans, alarms, other sensors, etc.

Simple relational

DW/MPP

NoSQL → Column, Graph, Document?

Map/Reduce

ELK / Elasticstack

Hang on! What do I need to know?

The characteristics (pros/cons) of different types of data management methods and some examples

Specific terms and acronyms related to data management

How to approach a data collection and storage task

Eg: What questions to ask; What attributes to look for

Solutions are often multipart (“polyglot persistence”)

Labs today (LG25 & LG26)

Three options

1. Assignment
2. Datacamp
3. Or play with Map/Reduce:

[https://nbviewer.org/github/phelps-sq/python-bigdata/blob/master/src/
main/ipynb/spark-mapreduce.ipynb](https://nbviewer.org/github/phelps-sq/python-bigdata/blob/master/src/main/ipynb/spark-mapreduce.ipynb) (example of word counting using
python)

References

DCU library ebook: R. Stevens, Beginning Database Design Solutions (Part 1 only) -

[https://ebookcentral-proquest-com.dcu.idm.oclc.org/lib/dcu/reader.action?ppg=39
&docID=427853&tm=1543835305326](https://ebookcentral-proquest-com.dcu.idm.oclc.org/lib/dcu/reader.action?ppg=39&docID=427853&tm=1543835305326)

References

Apache Hadoop, <https://hadoop.apache.org/>

Chapter 10: Advanced Analytics—Technology and Tools: MapReduce and Hadoop, “Data Science and Big Data Analytics : Discovering, Analyzing, Visualizing and Presenting Data” DCU Library:

<https://capitadiscovery.co.uk/dcū/items/dda-17/EBC1908952>

Apache Hive Essentials (brief Overview of ecosystem)

https://subscription.packtpub.com/book/application_development/9781788995092/1/ch01lvl1sec14/overview-of-the-hadoop-ecosystem

Udacity course on Big Data + Map/Reduce (Hadoop) -

https://youtu.be/DEQNknALf_8?list=PLAwxTw4SYaPkXJ6LAV96gH8yxIfGaN3H-

A World
Leading SFI
Research
Centre



Visual analytics from big data generated by instrumented vehicles

Dr Suzanne Little

Insight SFI Research Centre for Data Analytics
School of Computing, Dublin City University

Insight

SFI RESEARCH CENTRE FOR DATA ANALYTICS

HOST INSTITUTIONS



FUNDED BY:



Instrumented Vehicles EU-funded Projects

Cloud-LSVA



Large-Scale Video Analytics
Annotation & search of video
data for ADAS & cartography
H2020 ICT “Big Data”



VI-DAS



Vision Inspired Driver Assistance System
Object tracking & path
prediction for safer driving
H2020 Mobility





~15-20 TB/day
~300 hrs /minute

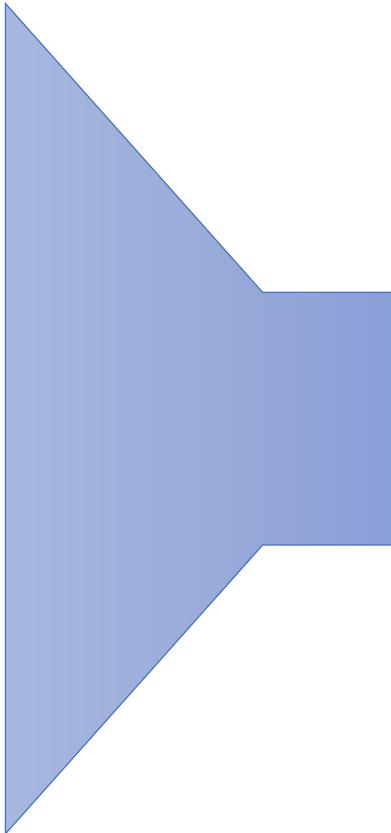
Lossy content
Large amount of files
Worldwide upload points



ADAS Context
Open-road Acquisition

~10-50 TB/day/vehicle
~8 hr collection window

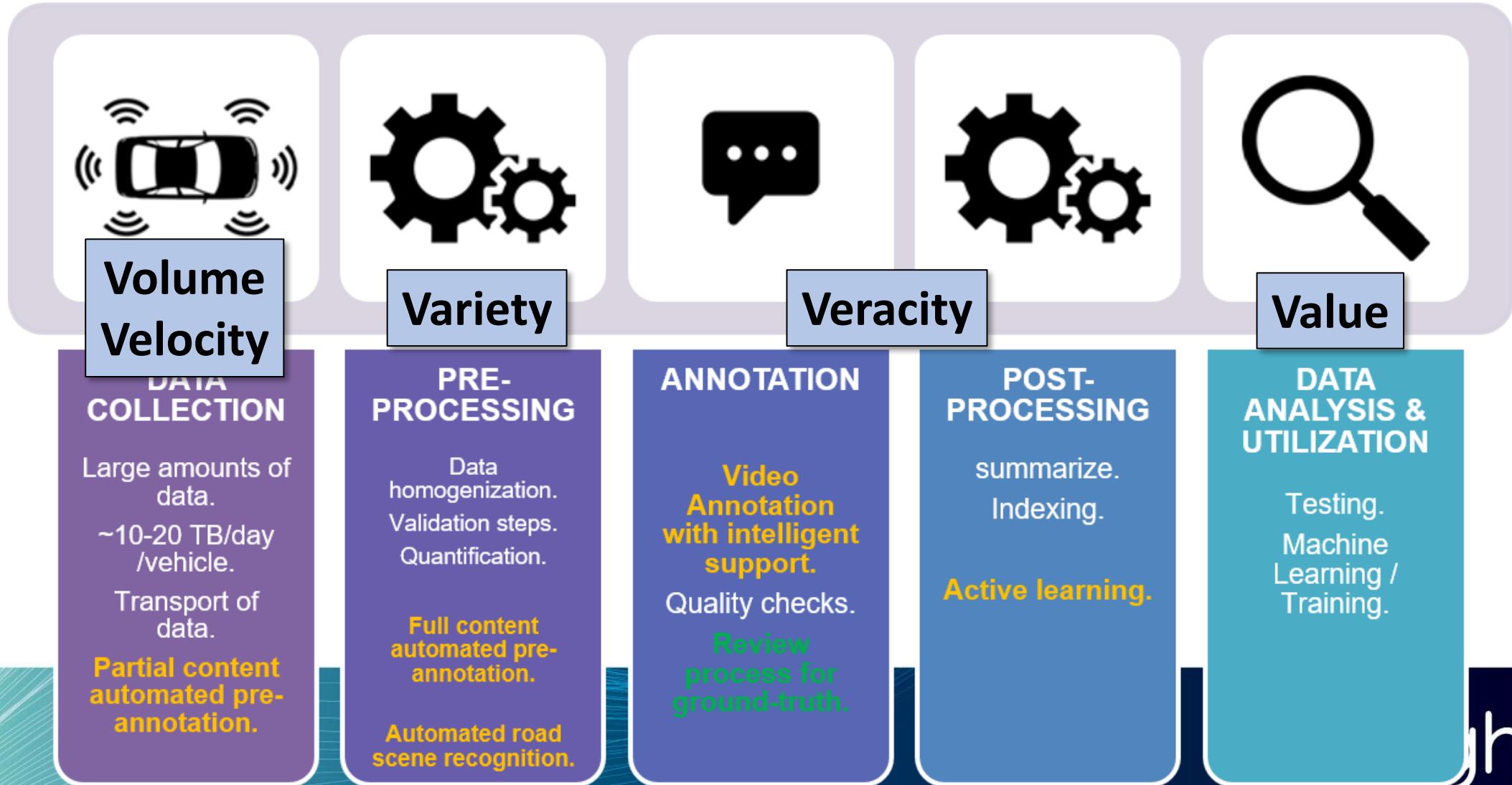
Lossless content
Reduced amount of files
Limited upload points



Big Data

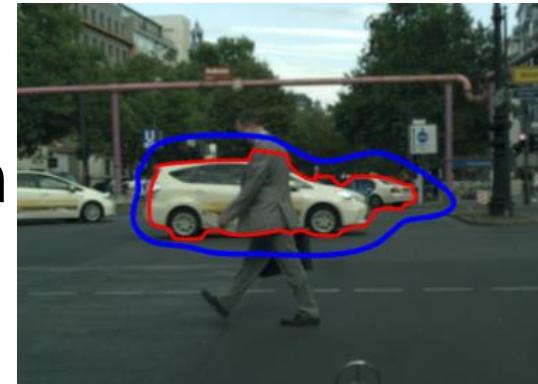
Volume
Velocity
Variety

Visual Analytics for Instrumented Vehicles

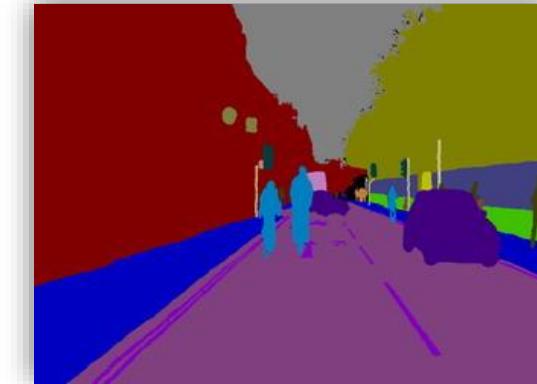


Visual Analytics for Instrumented Vehicles

Augmented Segmentation

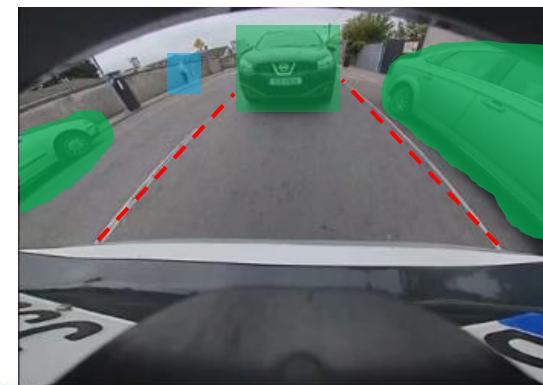


Semi-automatic Annotation



Semantic Segmentation

Object Detection, Tracking & Classification



Path Prediction



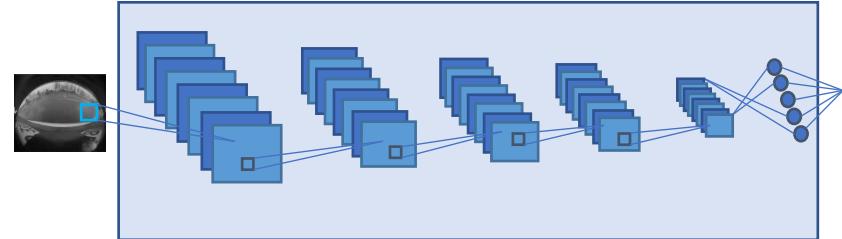
Situation Classification

Semantic Search

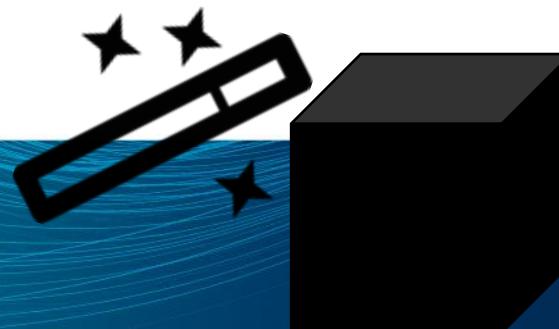
Understanding high-volume, high-speed multimedia data – a computer scientist's view

- Big data in ADAS means fully manual annotation is infeasible
- Data that's not annotated is not usable
- Annotation via Machine Learning?
- Deep learning: a step change in computer vision

Deep Learning



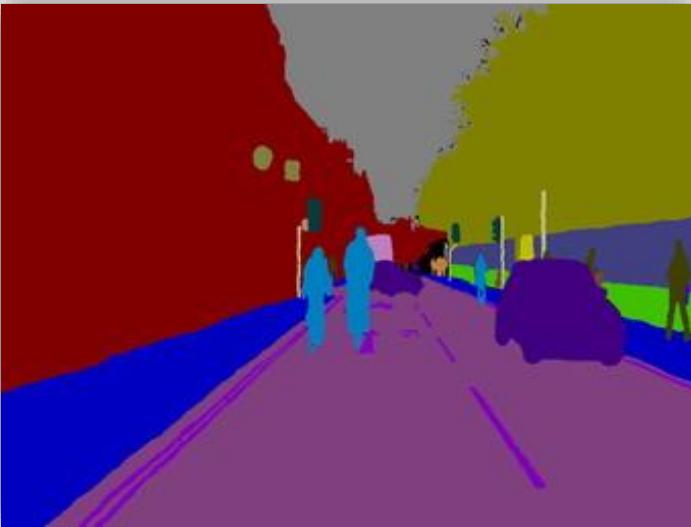
- Rebranding of an old idea?
- Possible because of powerful hardware (GPUs)
- Requires large quantities of labelled input
- Can require a lot of working memory
- Configuring, optimising complexity and performance is (still) a bit of an art



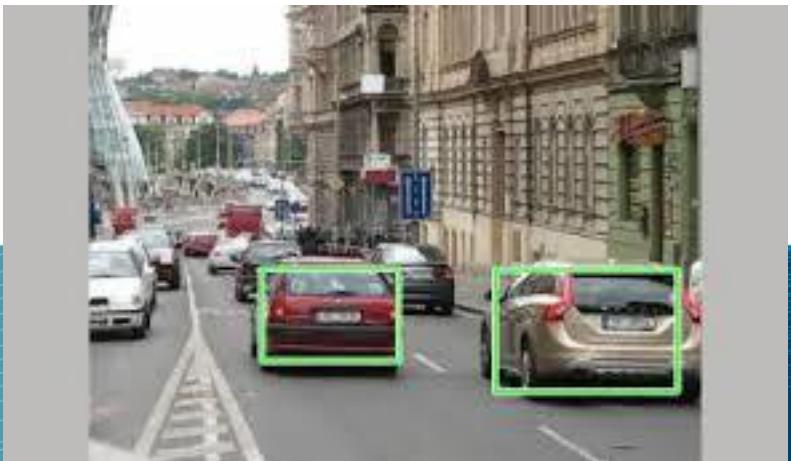
Annotating Video



Road scene



Ground-truth

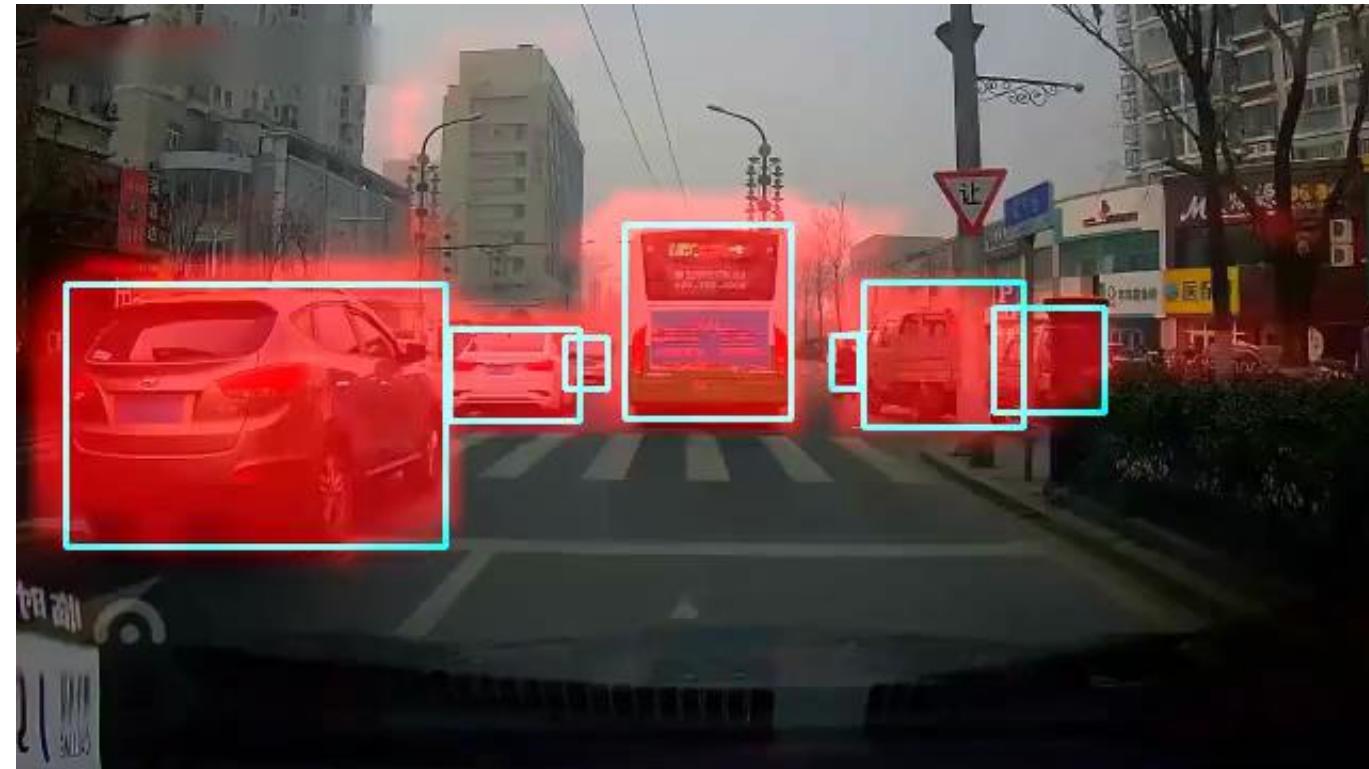


Scene level
Static Objects
Dynamic Objects
Background
Actors
Vehicles

Void	Building	Wall	Tree	VegetationMisc
Fence	Sidewalk	ParkingBlock	Column_Pole	TrafficCone
Bridge	SignSymbol	Misc_Text	TrafficLight	Sky
Tunnel	Archway	Road	RoadShoulder	LaneMkgsDriv
LaneMkgsNonDriv	Animal	Pedestrian	Child	CartLuggagePram
Bicyclist	MotorcycleScooter	Car	SUVPickupTruck	Truck_Bus
Train	OtherMoving			



Improved object tracking efficiency using saliency



- Object detection, classification & tracking at high accuracy takes time
- Use region proposals based on generic saliency
- Result: pipeline that achieves high accuracy with lower processing time

Data?

- **Volume**
 - Research datasets like [KITTI](#): ~5GB, ~7400 training images
- **Annotations**
 - Not always available
 - Not always sufficiently detailed (bounding box vs pixel-level segmentation)
- **Velocity? Data already captured so simulation only**
- **Variety**
 - Synchronised Sensor Data (LIDAR, telemetry, GPS, etc.)
 - Context ...





Saleh, K., Hossny, M., & Nahavandi, S. (2018). Effective vehicle-based kangaroo detection for collision warning systems using region-based convolutional networks. *Sensors*, 18(6).

Volvo's driverless cars 'confused' by kangaroos

⌚ 27 June 2017

f 📲 🐦 📧 Share

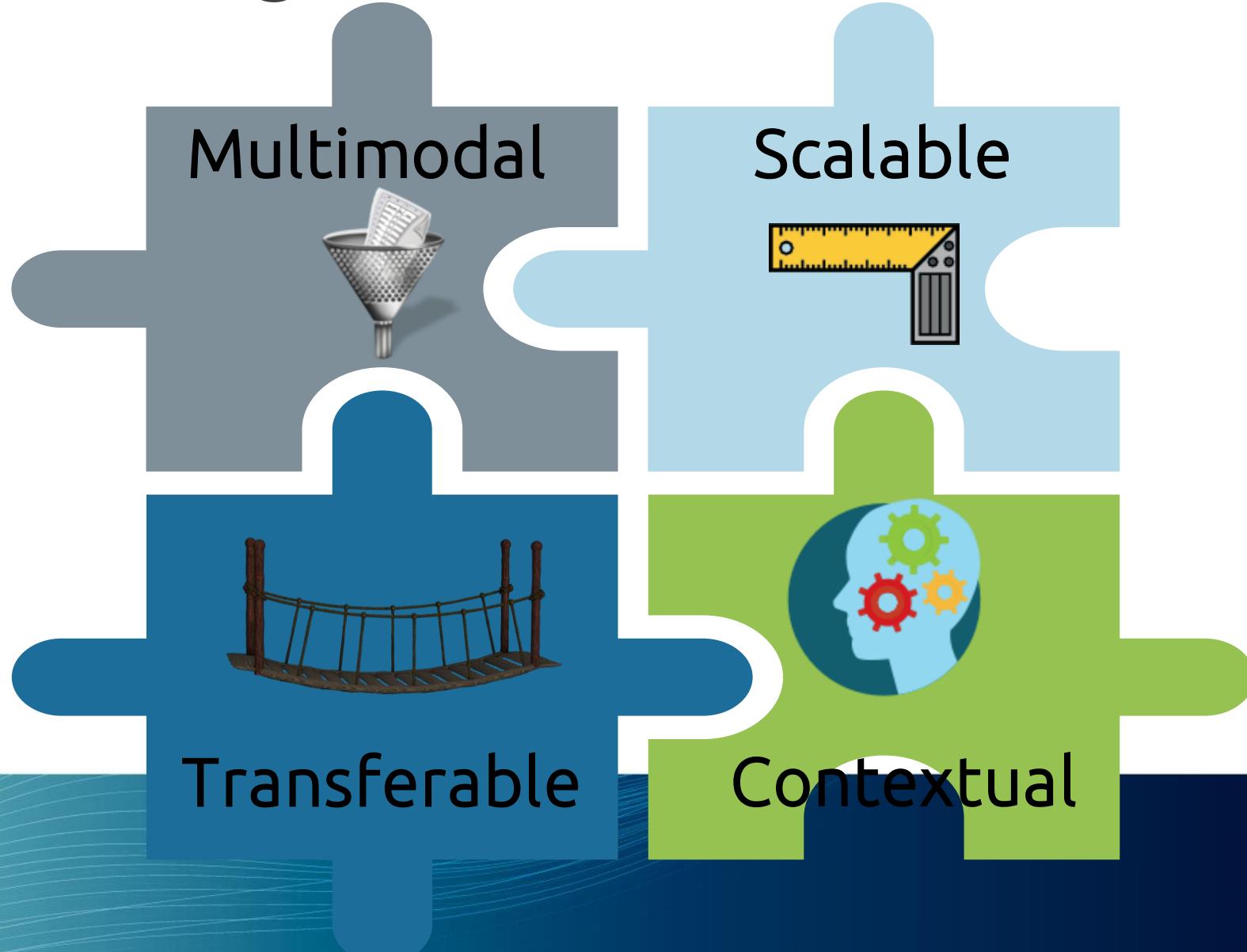


There are more than 20,000 kangaroo strikes each year in Australia

Volvo's self-driving technology is struggling to identify kangaroos in the road.

<https://www.bbc.com/news/technology-40416606>

Key challenges



Thanks to ...



Multimedia Analytics Team at Insight DCU

Prof. Noel O'Connor

Prof. Alan Smeaton

Dr Kevin McGuinness

Enric Moreu (Deep Clip)

Dian Zhang

Jaime Fernandez Roblero

Venkatesh Gurram Munirathnam

Dr Feiyan Hu

Contact:

Suzanne Little

suzanne.little@dcu.ie

@suz_research

computing.dcu.ie/~slittle

This work has been partly funded by Science Foundation Ireland under grant number SFI/12/RC/2289
and the EU H2020 Projects VI-DAS (grant number 690772) and Cloud-LSVA (grant number 688099).

A World
Leading SFI
Research
Centre



Data Visualisation: Good Things to Know

Suzanne Little
School of Computing
Dublin City University

Insight

SFI RESEARCH CENTRE FOR DATA ANALYTICS

HOST INSTITUTIONS



FUNDED BY:

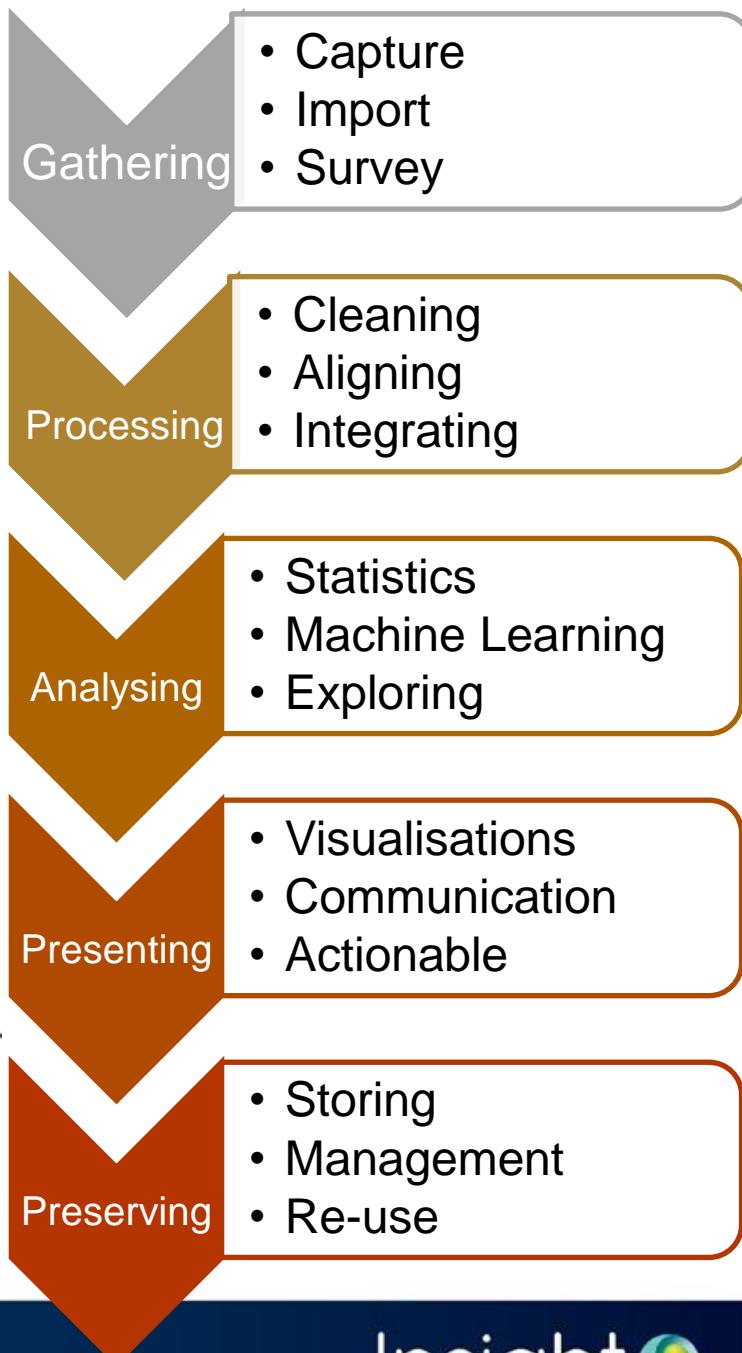
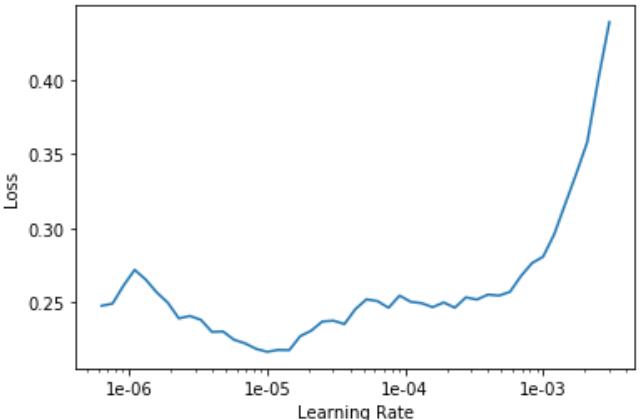
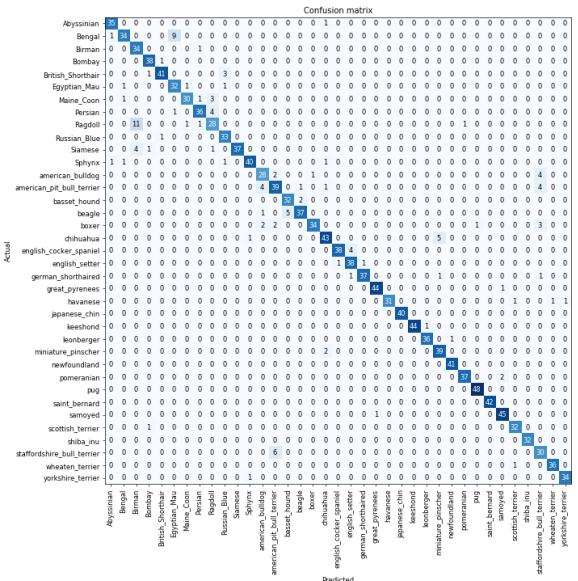


Outline

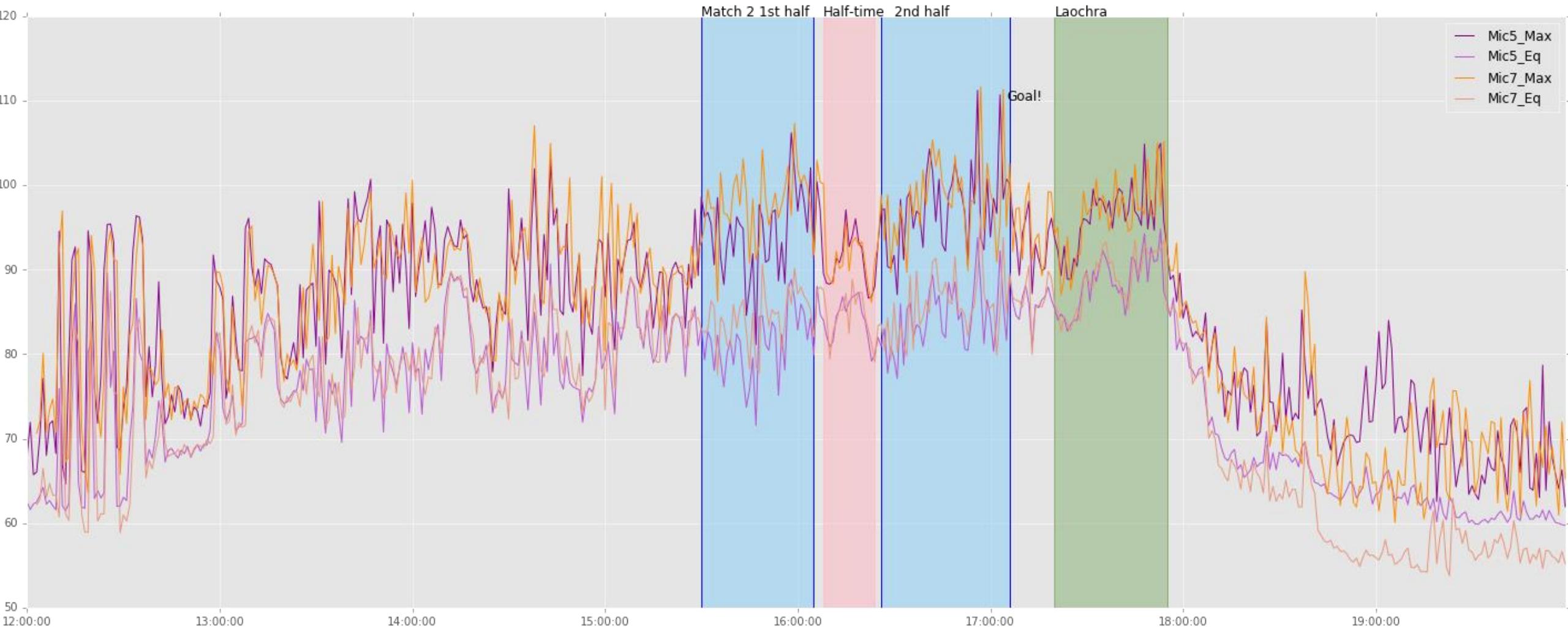
- Why visualise data?
- Good things to know
 - Pie charts
 - 3D
 - Area
 - Axes
 - Clutter
- Good Visualisation?

Why visualise data?

- To explore and analyse
- To communicate



Decibel Levels, Croke Park, 2016





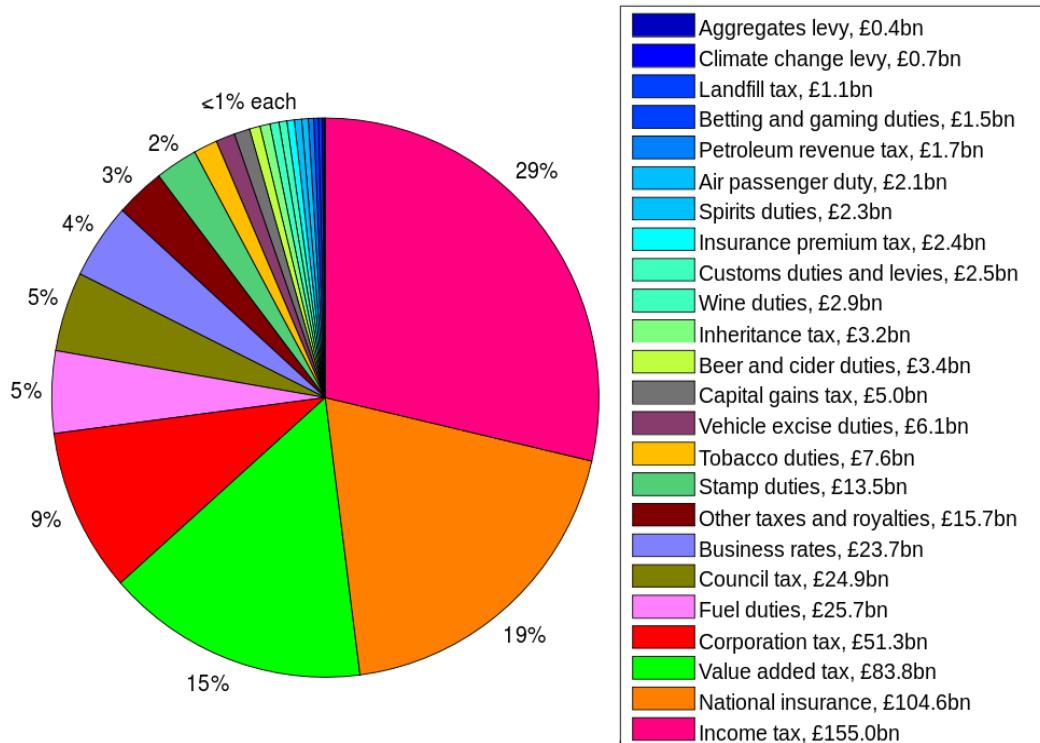
Smart Stadium for Smarter Living

Outline

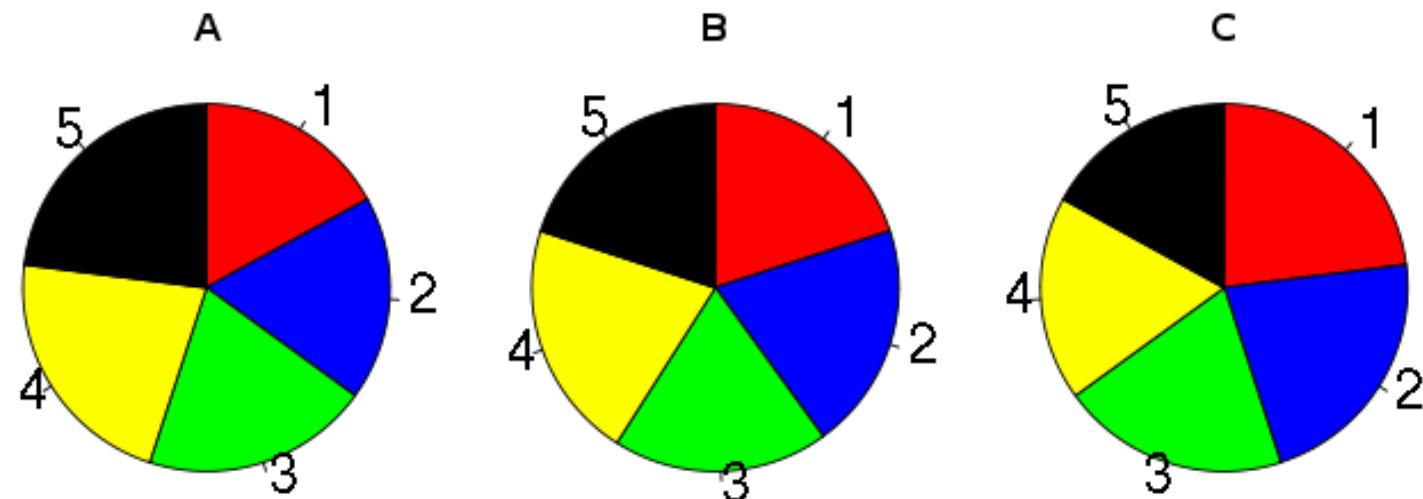
- Why visualise data?
- Good things to know
 - Pie charts
 - 3D
 - Area
 - Axes
 - Clutter
- Good Visualisation?

Good things to know: Pie Charts

Almost never a good idea ...



https://en.wikipedia.org/wiki/Taxation_in_the_United_Kingdom

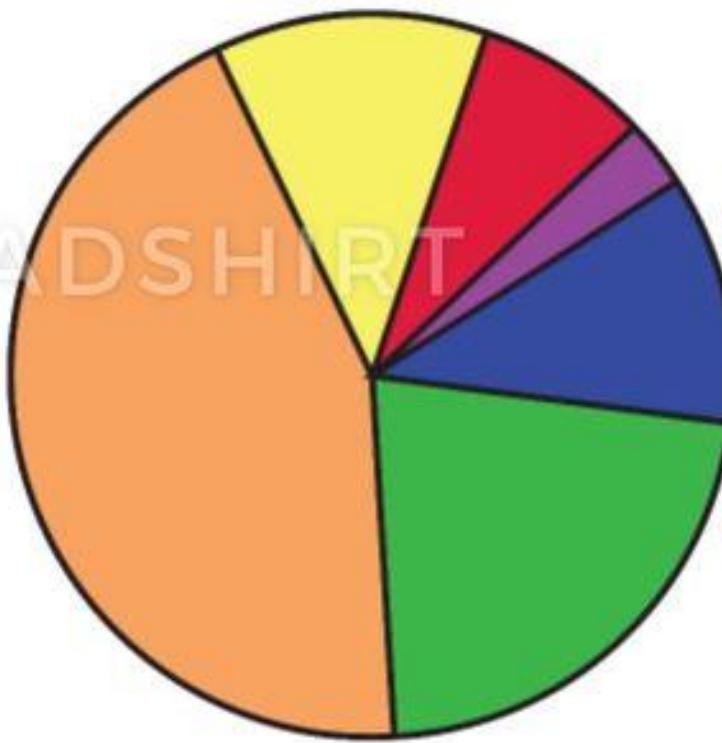


If you *must* use a pie chart

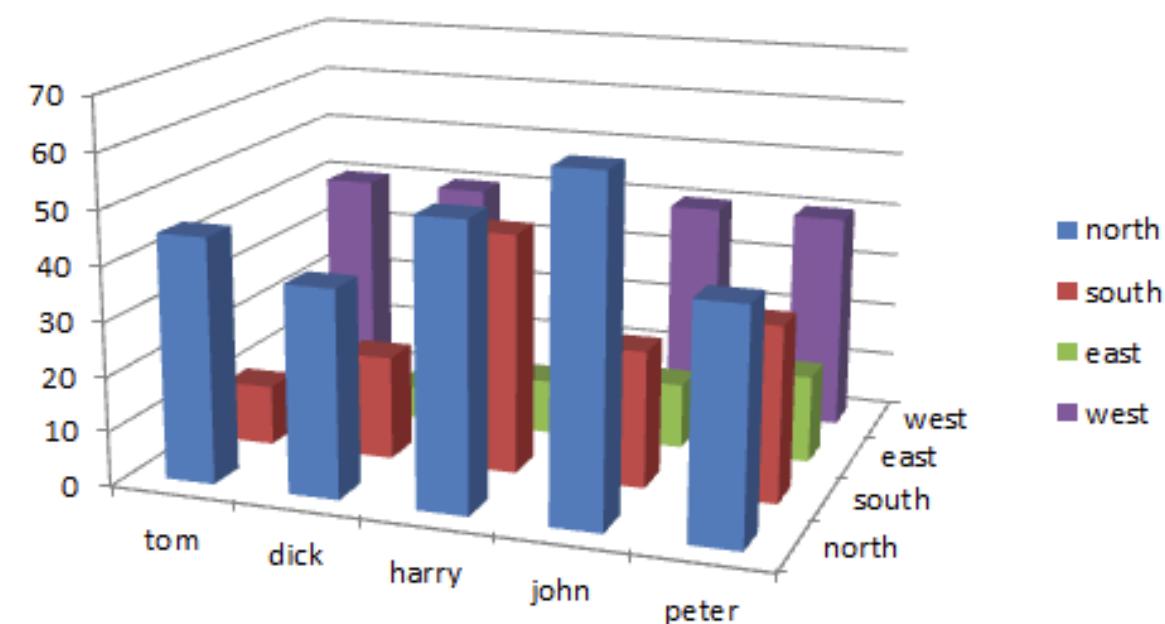
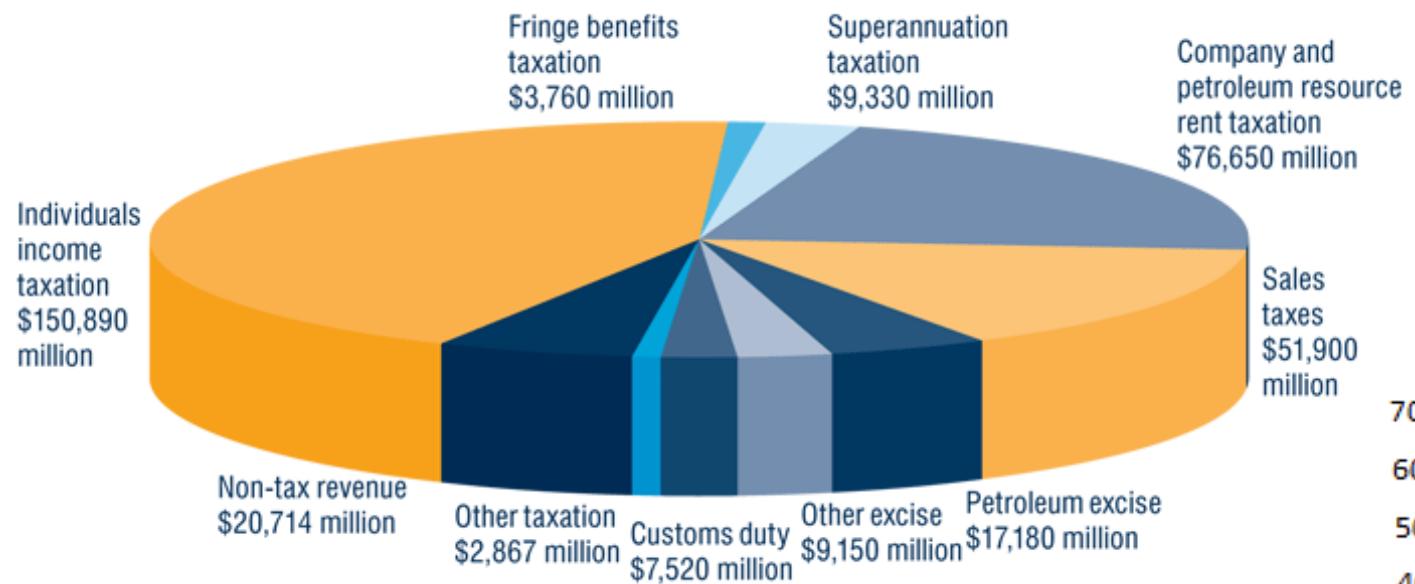
- Only for parts of a whole (ie, 100% divided into categories)
- No more than 5 slices
- Label carefully and clockwise, decreasing in size
 - Minimise user effort
- Never 3D and exploded isn't good either!
 - But it looks cool ... ☹

THINGS I WILL NEVER DO

- Give you up
- Let you down
- Run around & hurt you
- Make you cry
- Say goodbye
- Tell a lie and hurt you

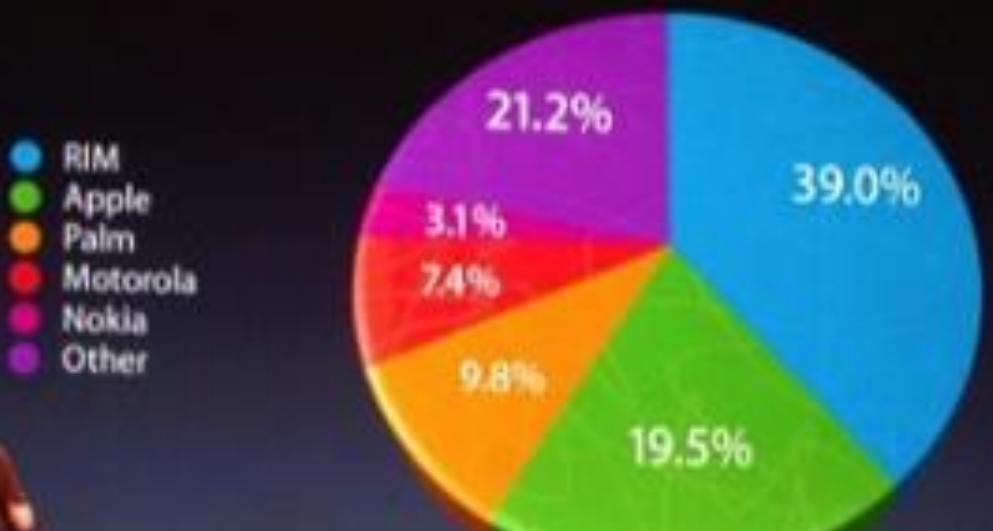


Good things to know: 3D



http://www.budget.gov.au/2011-12/content/overview/html/overview_46.htm

U.S. SmartPhone Marketshare



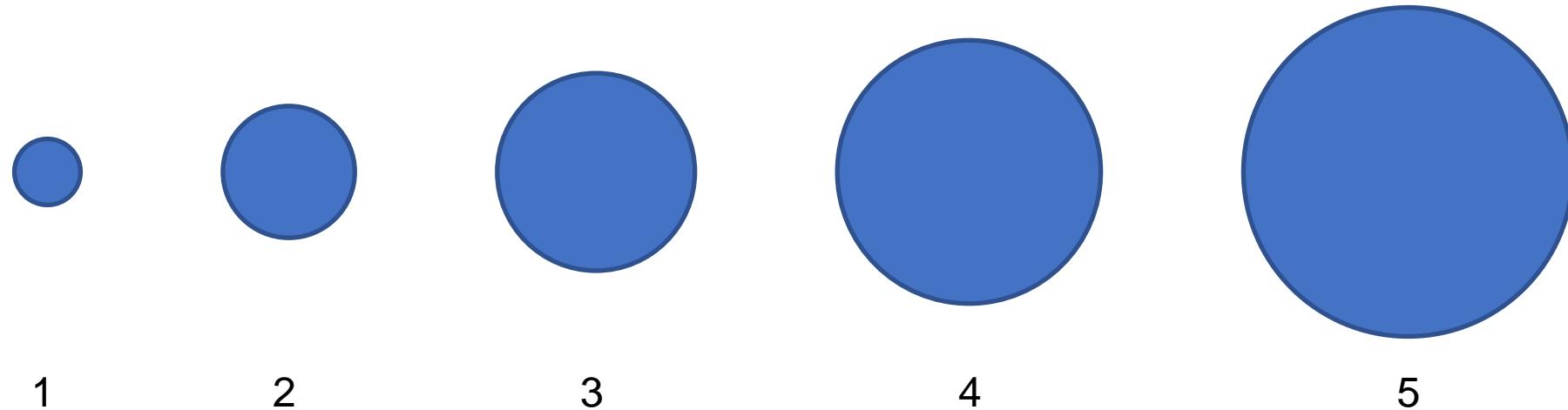
Another (infamous) example
Jobs at Engadget 2008
Compare 21.2% and
19.5%
Which is bigger?

A good example?

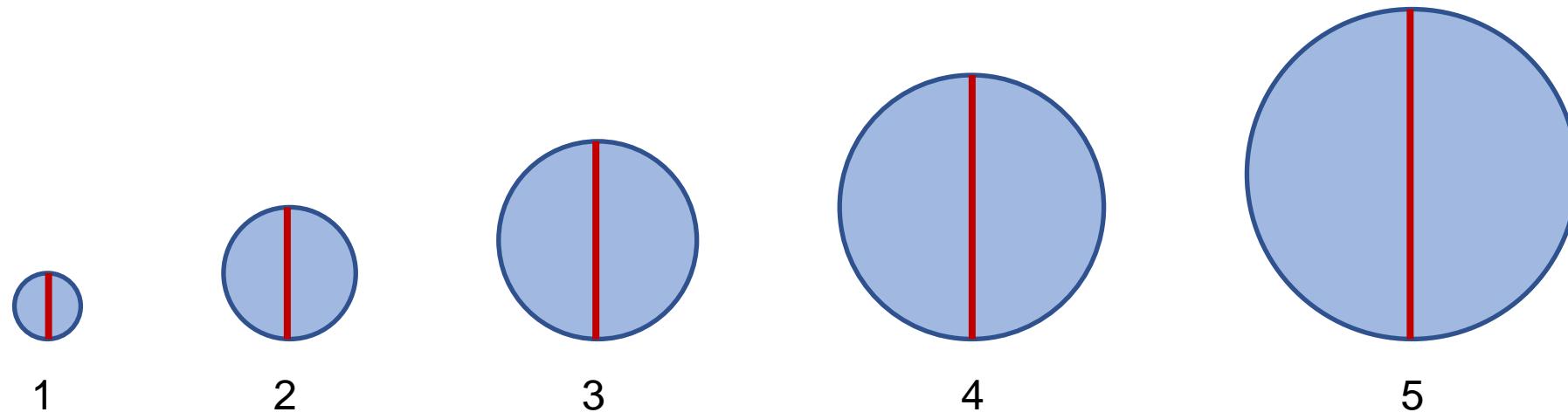


Good things to know: understanding area

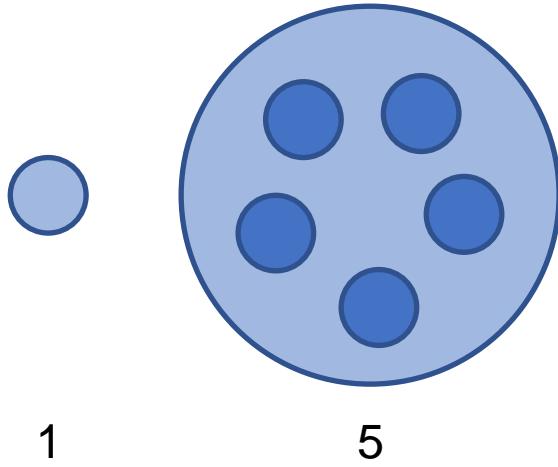
Be careful when turning numbers into shapes ...



Good things to know: understanding area

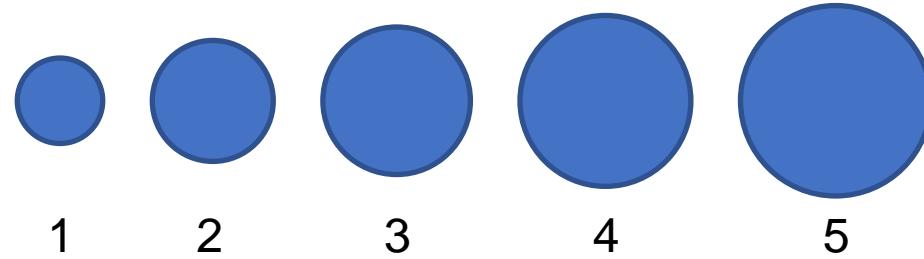


Good things to know: understanding area

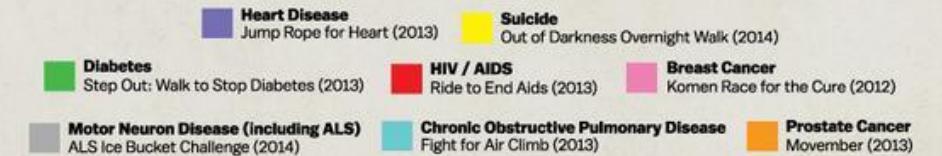


Good things to know: understanding area

$$A = \pi r^2$$



WHERE WE DONATE VS. DISEASES THAT KILL US



- What does colour indicate?
- How much do your eyes move?
 - Is it easy to compare?
- Highly specific numbers
- Area or diameter?
- Data from 2011 but dates are 2013/2014?

WHERE WE DONATE VS. DISEASES THAT KILL US

Heart Disease Jump Rope for Heart	Suicide Out of Darkness Overnight Walk	Breast Cancer Komen Race for the Cure
Diabetes Step Out: Walk to Stop Diabetes	HIV / AIDS Ride to End Aids	Prostate Cancer Movember
Motor Neuron Disease (including ALS) ALS Ice Bucket Challenge	Chronic Obstructive Pulmonary Disease Fight for Air Climb	

MONEY RAISED



DEATHS (US)



<http://www.vox.com/2014/8/20/6040435/als-ice-bucket-challenge-and-why-we-give-to-charity-donate>

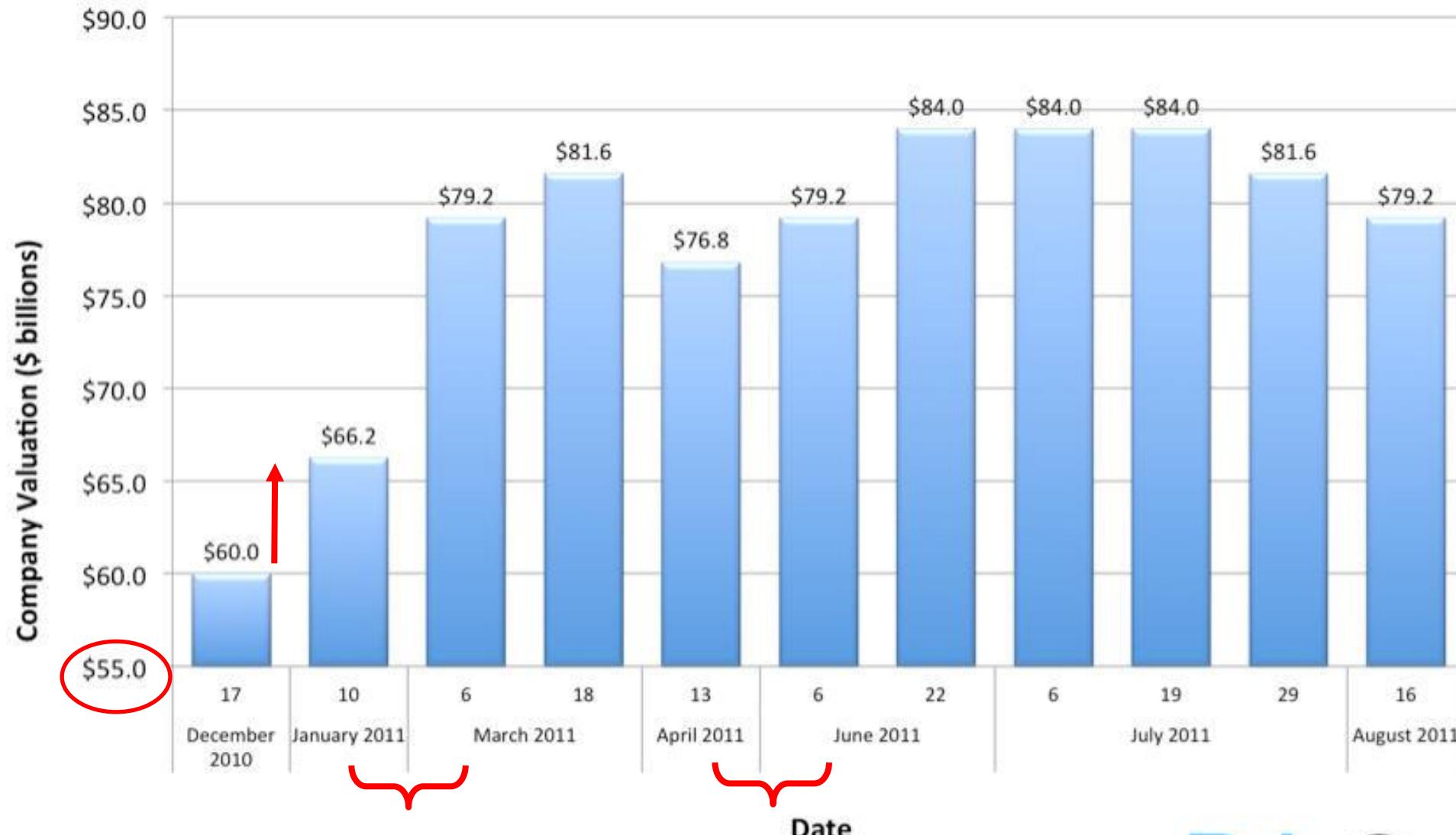
Good things to know: axes

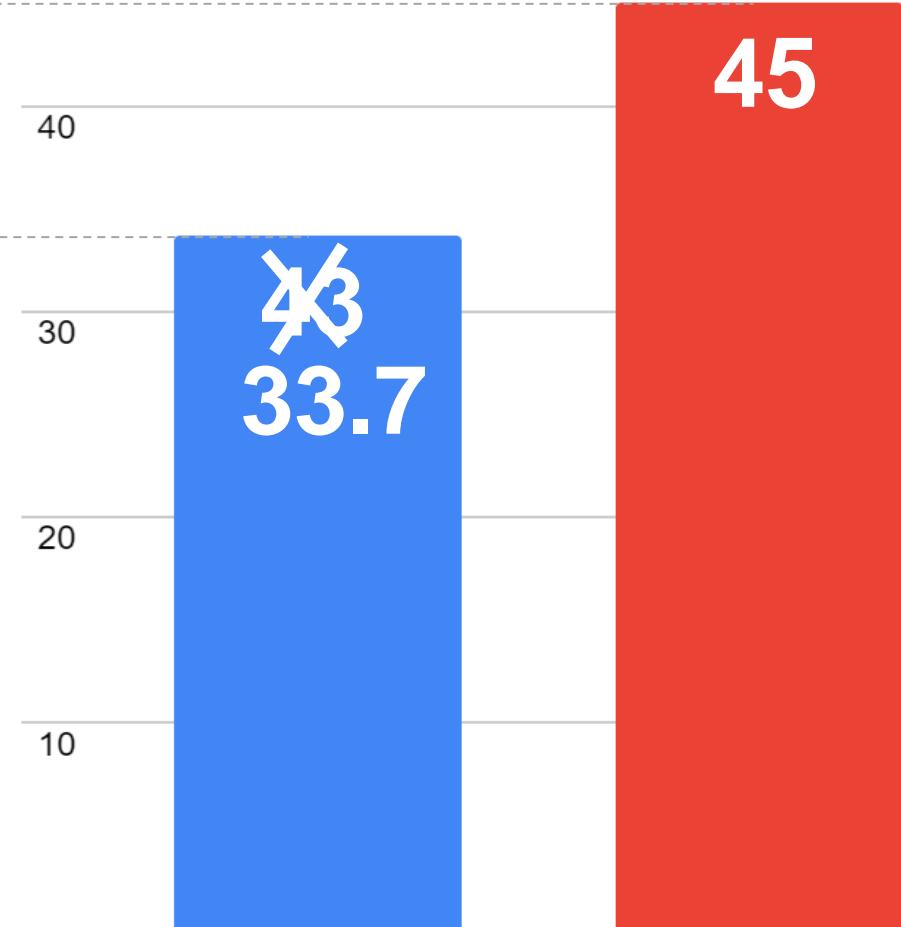
- Label properly
- Include units
- Be careful with exponential
- Where's your baseline? Start from zero.



<https://thenode.biologists.com/non-zero-baselines-the-good-the-bad-and-the-ugly/resources/>

Facebook, Inc.: Company Valuation





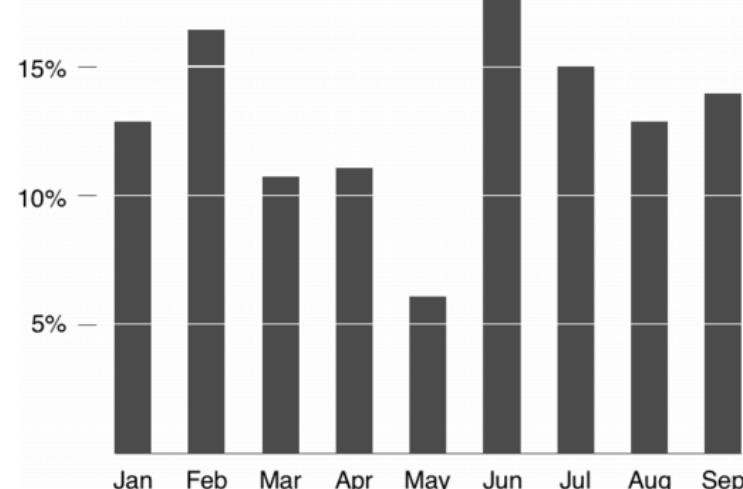
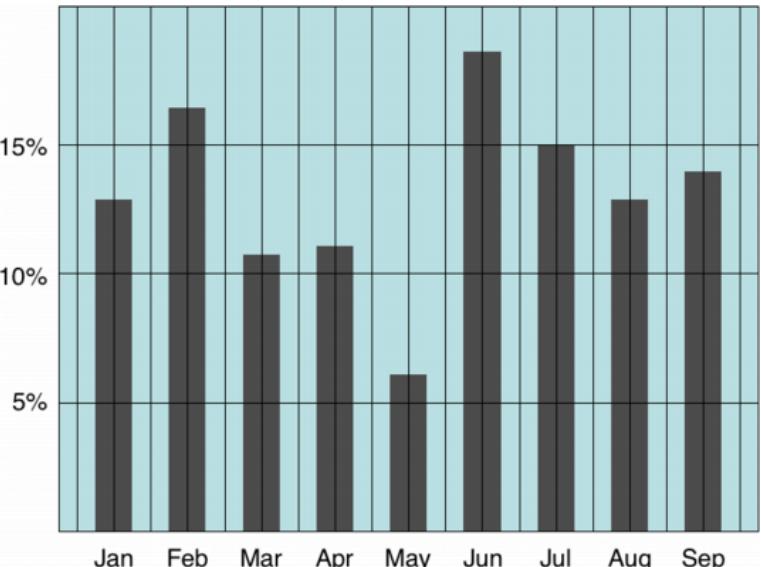
<https://www.washingtonpost.com/graphics/politics/2016-election/trump-charts/>

Good things to know: Remove Clutter

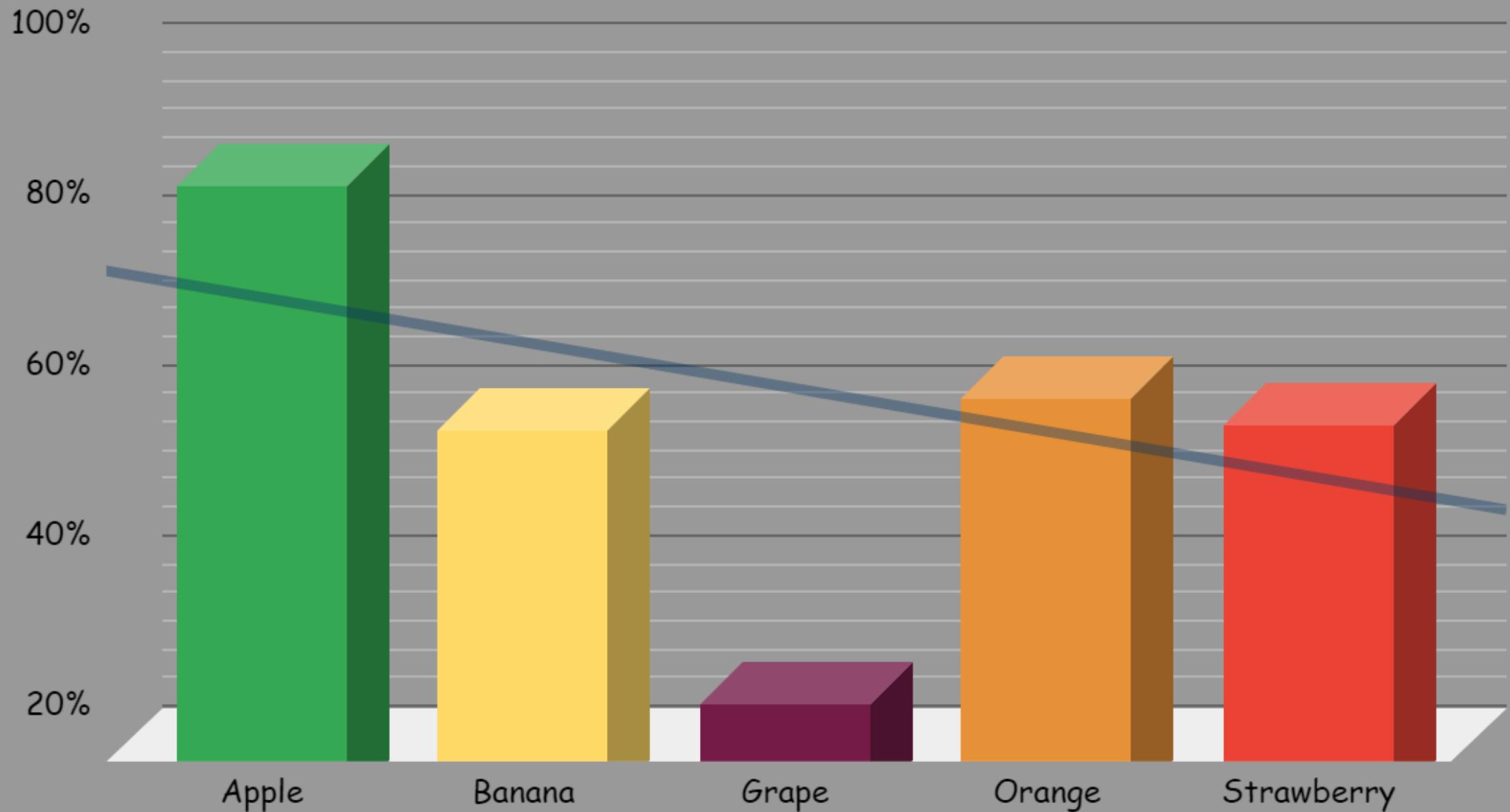
Edward Tufte: data to ink ratio

Remove all pixels not directly related to data

- Highly minimalist approach



Popularity of fruit flavours



Take away points

- What are you trying to say?
- Pie (and donuts!) in moderation only!
- Apply data-to-pixel method to remove unnecessary elements (Marie Kondo your graph?)
- Title should state the message/conclusion and right align
- Be kind to your viewer! Attention not distraction.