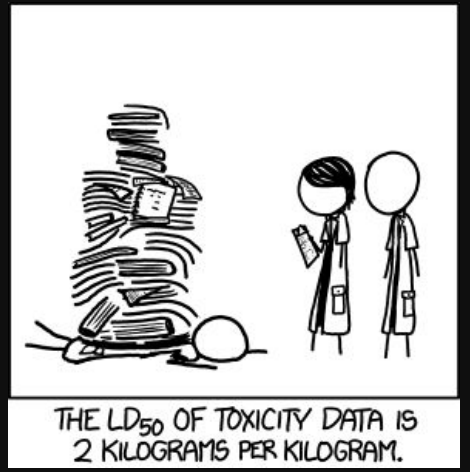
Session 1

Data at Scale: An Introduction

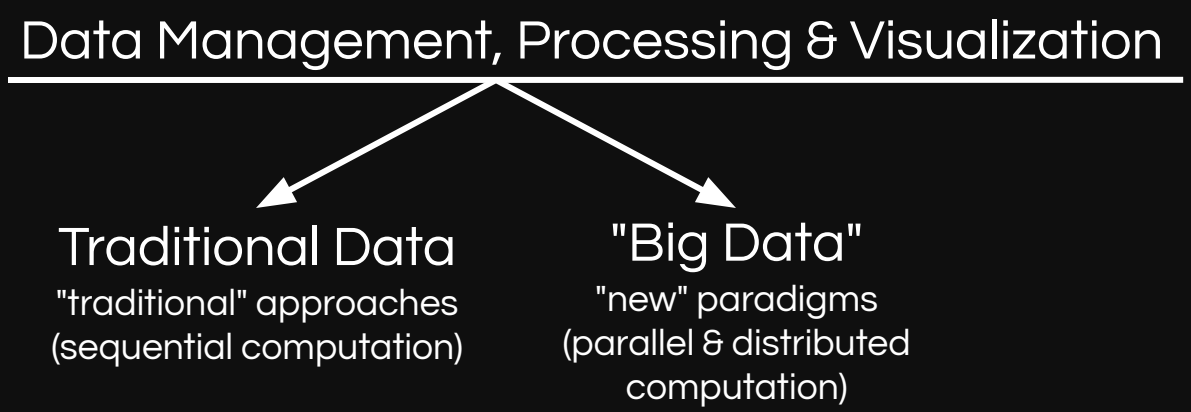


What do I mean by ***Data at Scale*** anyway?

- Data *(that uses computers to be **managed** and **processed**)*

So, in this module

The different options to:



But what is Big Data anyway...

Data has been around for a long time....

1975 Conference on Very Large Databases was established

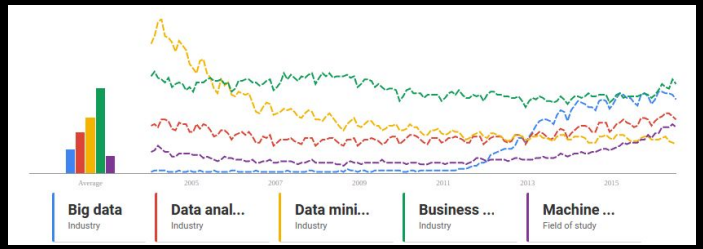
Turning data into useful information has also been around for a long time....

- Data analytics
- Data mining

- Knowledge discovery
- Machine learning
- Statistical modelling...

Why is there a new term...

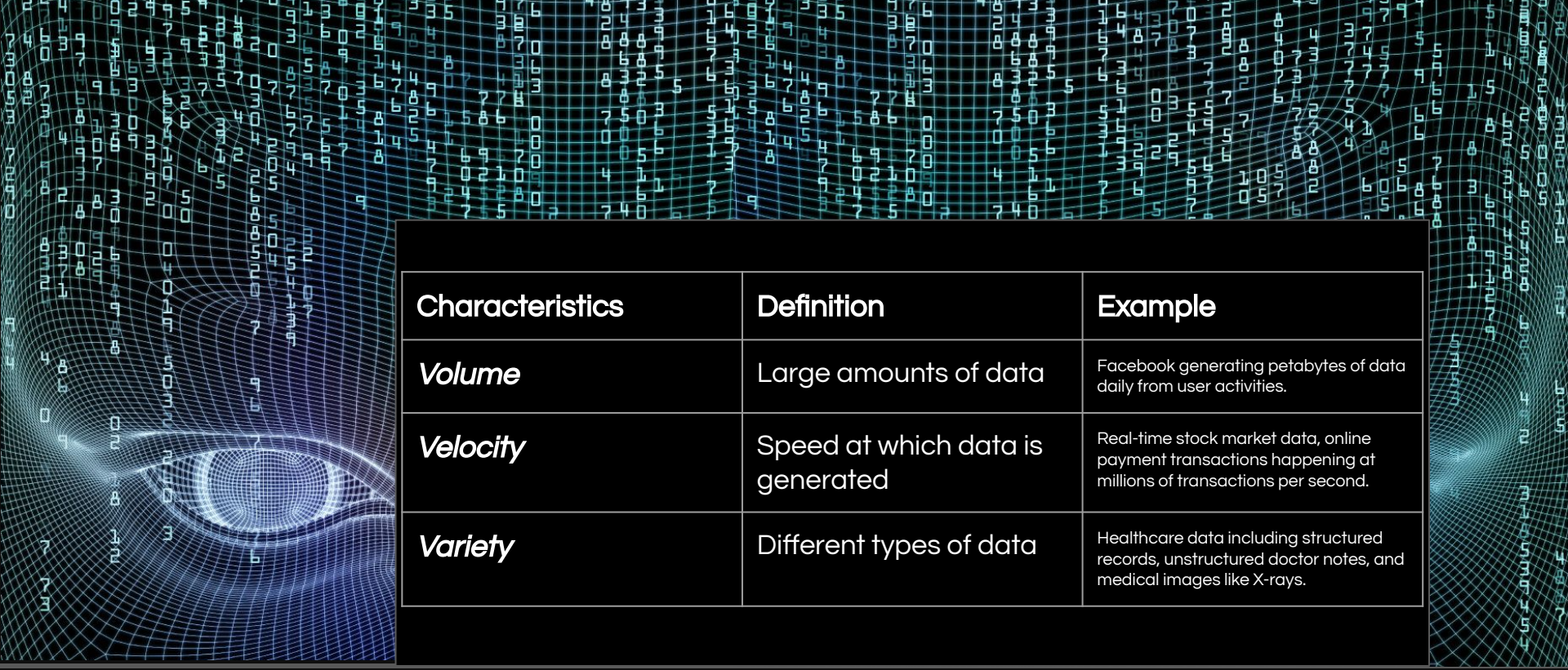
Popularity of terms over time (Google Trends)



…. simply a **buzzword** for "problems that **currently** won't simply scale up by throwing more (of the same) computational power at it"

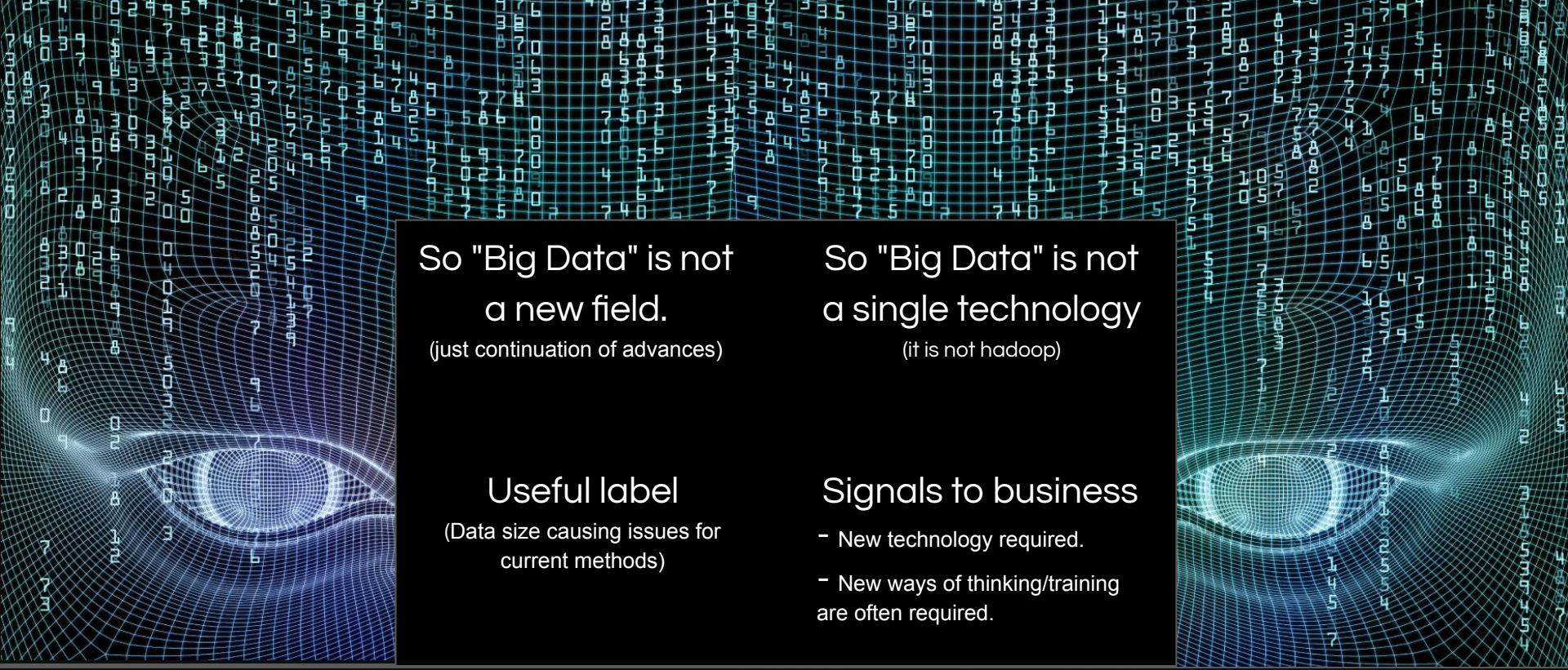
Previous buzzwords for the same thing: **large data, very large data**

Future buzzwords for the same thing: **Massive data?? Gigantic data??**



Characteristics	Definition	Example
Volume	Large amounts of data	Facebook generating petabytes of data daily from user activities.
Velocity	Speed at which data is generated	Real-time stock market data, online payment transactions happening at millions of transactions per second.
Variety	Different types of data	Healthcare data including structured records, unstructured doctor notes, and medical images like X-rays.

3 Vs model of Big Data, Doug Laney, 2001



So "Big Data" is not a new field.

(just continuation of advances)

So "Big Data" is not a single technology

(it is not hadoop)

Useful label

(Data size causing issues for current methods)

Signals to business

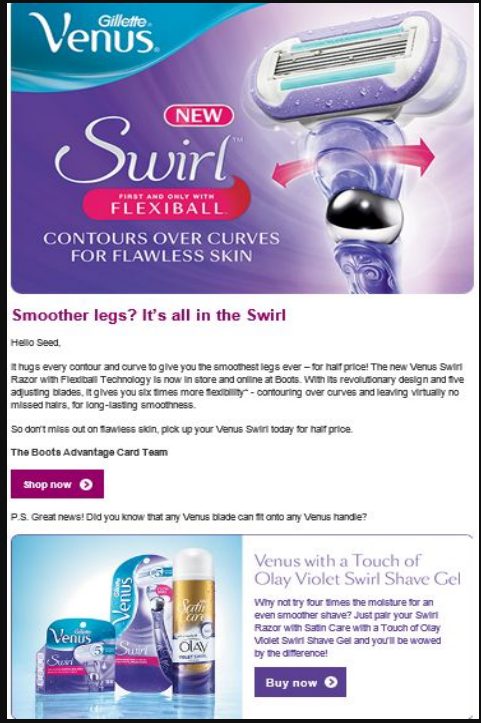
- New technology required.
- New ways of thinking/training are often required.

Why?

to underpin analytics…

To sell intelligently….

New Product Prediction (targeted marketing)

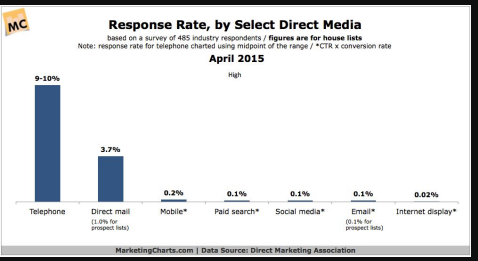


Why?

to underpin analytics…

To sell intelligently….

Some notes on the study below
Mix of B2B (52%), B2C (32%) & unknown
Study run in 2012
House = current/former customers



Boots

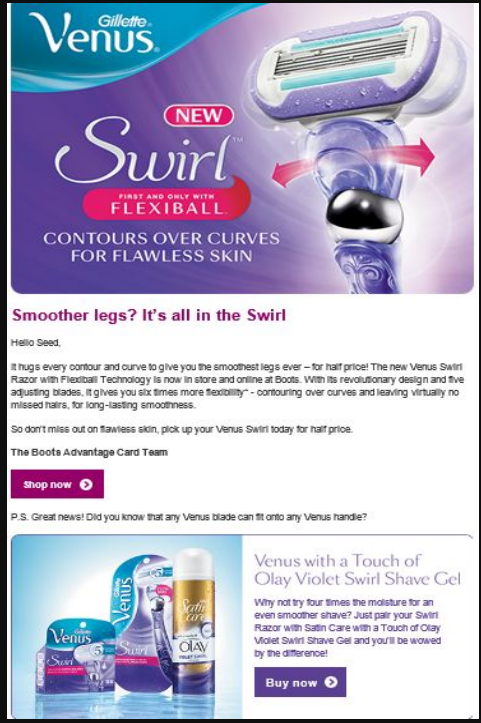
New Product Prediction (targeted marketing)

Kroger

Runs significant direct mailing campaigns

In 2013 it was reported[1] they had a redemption rate of over 70% (uplift though?…)

Worked with dunnhumby



[1] https://www.forbes.com/sites/tomgroenfeldt/2013/10/28/kroger-knows-your-shopping-patterns-better-than-you-do/#2784b20d746a

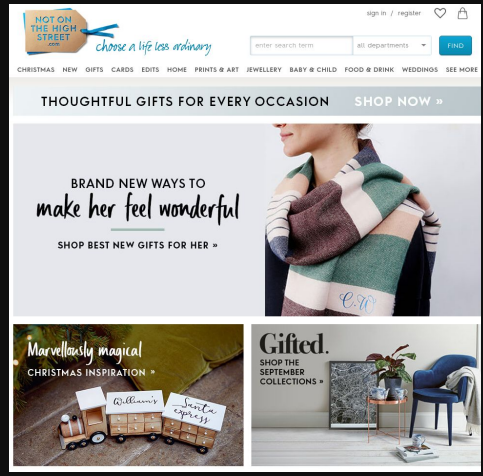
Sell intelligently

Why?

to underpin analytics...

To adjust staffing & supply chain strategies...

Predicting company's near future prospects



Sell intelligently

Why?

to underpin analytics…

To adjust staffing & supply chain strategies…

Predicting company's near future prospects

The Weather Company:

100,000 dedicated weather sensors

~10 billion forecast points / day (sensors + smartphones etc)

Weather can have non-trivial impact on sales.

Also impacts energy companies supply/demand forecasting.

Why?

to underpin analytics…

Sell intelligently

Adjust staffing & supply chain strategies

To minimize equipment & asset failures…

Pratt & Whitney
A United Technologies Company

Aim: Reduce unplanned aircraft engine maintenance.

Each engine:
~ 100 parameters per snapshot
soon ~5000 parameters

Preventative maintenance
i.e. predict maintenance requirements

Petabytes of data
1PB= 1000 TB
~4000 of the hard drives in your laptop

Why?

to underpin analytics…

To leverage customer lifetime value…

Sell intelligently

Adjust staffing & supply chain strategies

Minimize Equipment & Asset Failures

Pratt & Whitney
A United Technologies Company

Depending on how you define customer….

Segmentation
(predicted lifetime value)

Tiered incentives

Multichannel marketing campaign

Additionally : forecasting regional demand for fleet placements & pricing

More information:
https://www.informationweek.com/it-leadership/big-data-6-real-life-business-cases

Why?

to underpin analytics...

But is data management, processing & visualization **really** worth caring about...

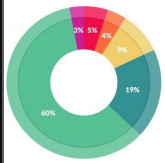
Sell intelligently

Adjust staffing & supply chain strategies

Minimize Equipment & Asset Failures

Leverage Customer Lifetime Value

DATA MANAGEMENT



Data scientists time:

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*

CrowdFlower via Forbes 2017

DATA PROCESSING

*72% of business and analytics leaders **aren't satisfied** with how long it takes to retrieve the insights they need from data* Alteryx

- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

DATA VISUALIZATION

"Data **visualization** is the key to **actionable insights** "
Head of Bus. Intel. & Data Analytics at AccuWeather

"**Actionable Insights** : The Missing Link Between Data And **Business Value** "
Director Data Strategy (Domo), in Forbes 2016

"74% of firms say they want to be "data-driven," **only 29%** say they are good at **connecting analytics to action** " Forrester 2016

DATA ANALYTICS IN BUSINESS

*Through 2017, **60% of data projects will fail** to go beyond piloting and experimentation and will be abandoned*
Gartner

***Only 27%** the executives surveyed described their data initiatives as successful*
Capgemini

*65% of CEOS think their organisation is able to **interpret only a small proportion** of the information to which they have access*
The Economist

N/LAB: *Data at Scale*

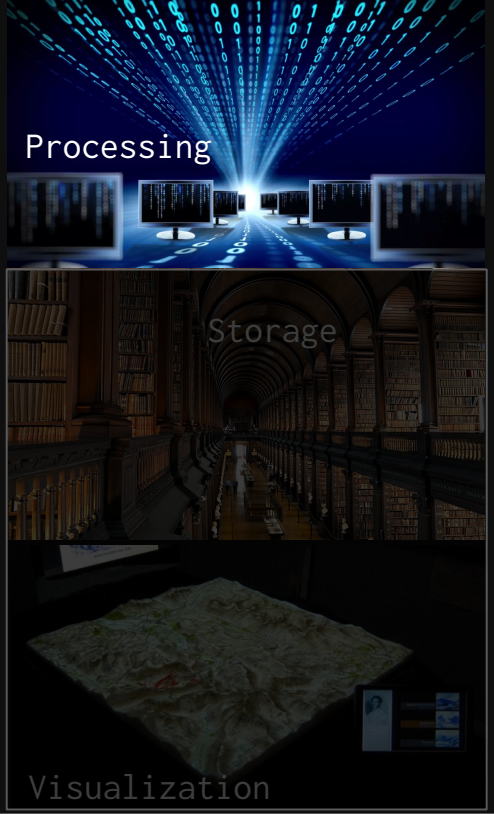
Dr Evgeniya Lukinova

How?

Many competing tools

How?

Many competing tools

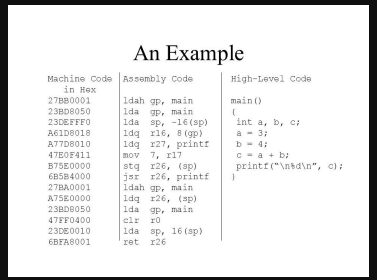


How to describe a problem?

- To humans?
- Computers?

Best way?

- Some history

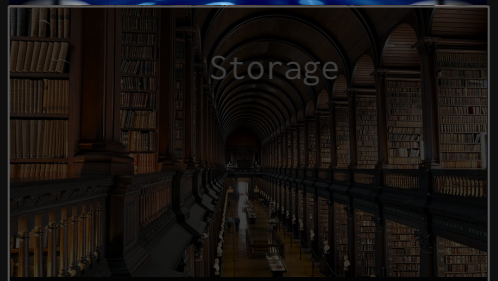


An Example

Machine Code in Hex	Assembly Code	High-Level Code
27BB0001	ldah gp, main	main()
23BD8050	lda gp, main	{
23DEFFF0	lda sp, -16(sp)	int a, b, c;
A61D8018	ldq r16, 8(gp)	a = 3;
A77D8010	ldq r27, printf	b = 4;
47E0F411	mov 7, r17	c = a + b;
B75E0000	stq r26, (sp)	printf("\n%d\n", c);
6B5B4000	jsr r26, printf	}
27BA0001	ldah gp, main	
A75E0000	ldq r26, (sp)	
23BD8050	lda gp, main	
47FF0400	clr r0	
23DE0010	lda sp, 16(sp)	
6BFA8001	ret r26	

How?

Many competing tools



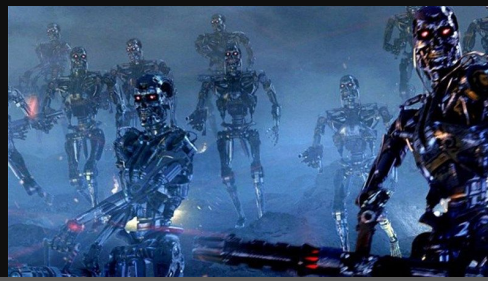
How to describe a problem?

- A "paradigm"
- Expression of problem within the paradigm (hidden detail, syntax)

Simplification of expression
=
automatic concept translation by rules

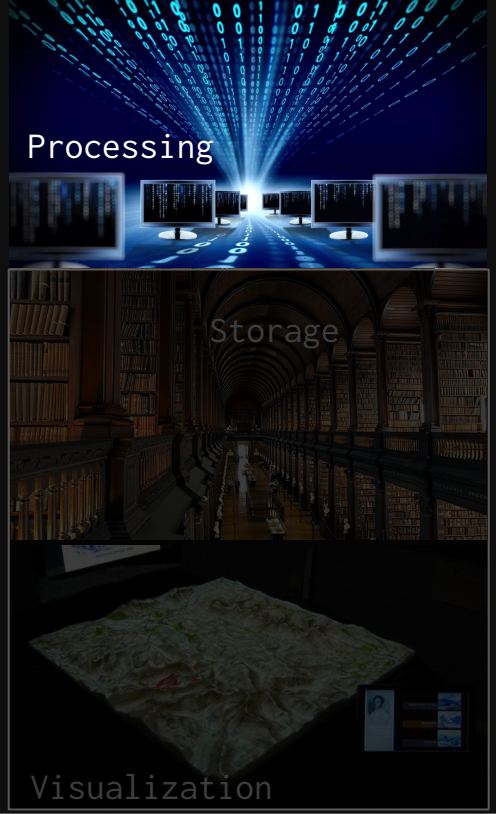
Better for humans
vs.
better for machines

Some concepts can be **automated easily** resulting in efficient machine code. **Some can't.**



How?

Many competing tools



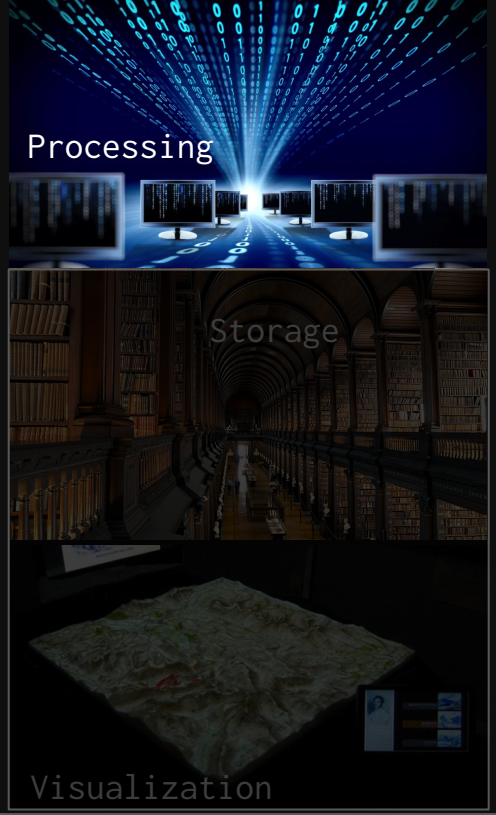
How to describe a problem?

For sequential processors some standard **paradigms** have "won".

Still significant variance in syntax and in the level of "simplified expression".

How?

Many competing tools



How to describe a problem?

Describing **non-sequential** processors - **hard for humans!!!**

No simple & "complete" paradigm.

Paradigms make some tasks:

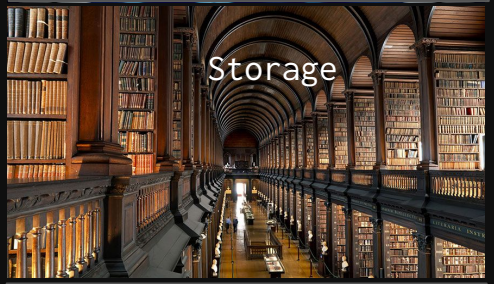
easy to describe , while other tasks **hard/impossible** .

easy to translate to **efficient code** , while other tasks **hard/impossible** .

+ syntax variation
+ level of abstraction

How?

Many competing tools

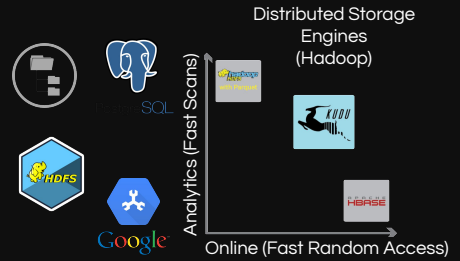


How to ~~describe a problem~~?

...organise data in a physical space?

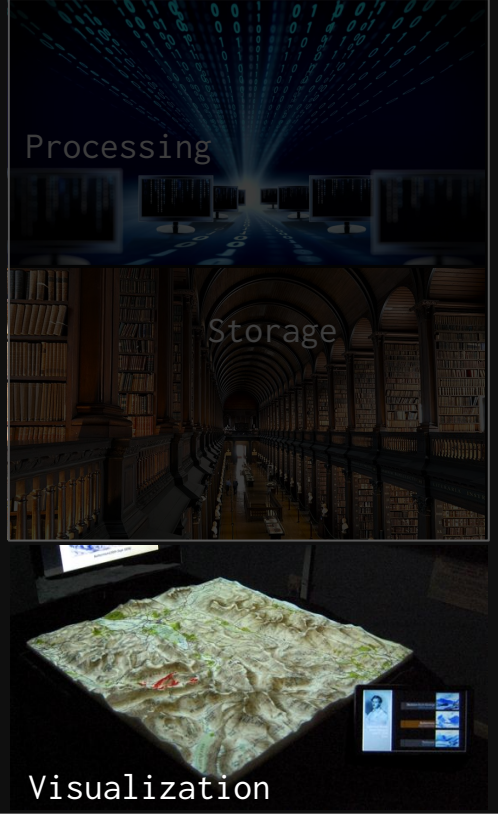
Trades speed of access with ease to access

→ too much of a trade-off, not all processing viable



How?

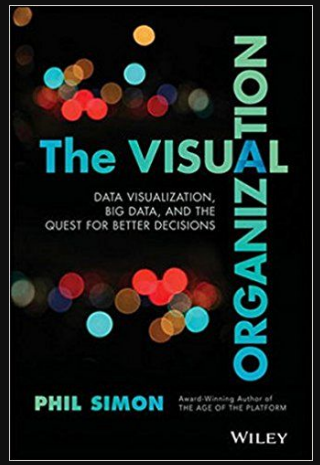
Many competing tools



"Data **visualization** is the key to **actionable insights** "
Head of Business Intelligence and Data Analytics (AccuWeather)

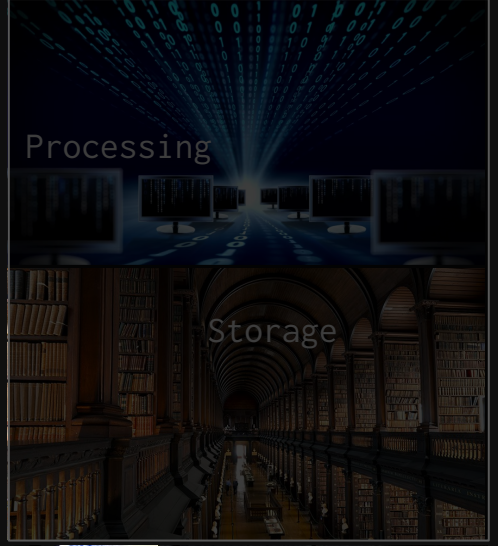
Wide range of visualizations.

Wide range of tools.



How?

Many competing tools

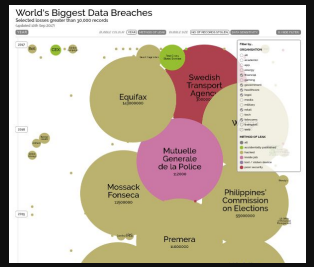


"Data **visualization** is the key to **actionable insights** "

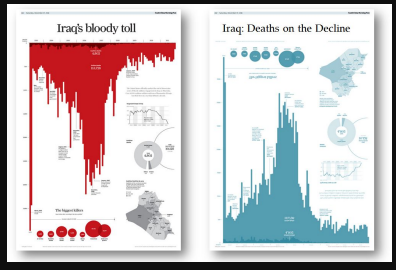
Head of Business Intelligence and Data Analytics (AccuWeather)

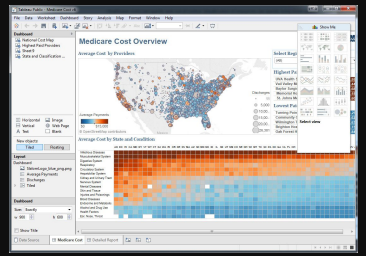
Exploratory vs. Explanatory

Lots of types of visualizations

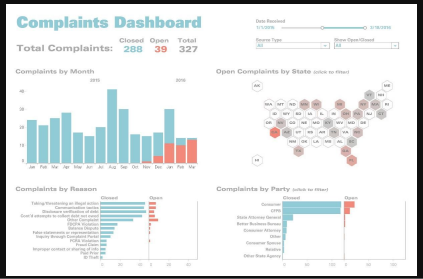


interpretability, & expectations





Simplicity, familiarity



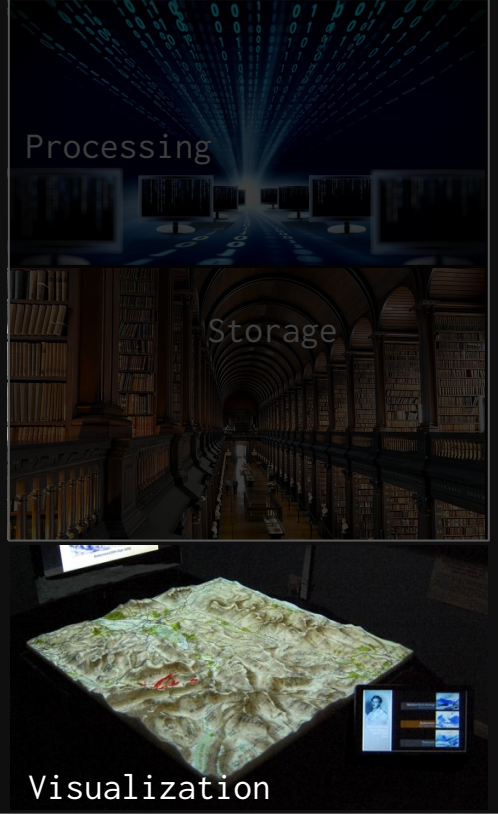
World's Biggest Data Breaches Interactive Visualization:
http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

N/LAB: *Data at Scale*

Dr Evgeniya Lukinova

How?

Many competing tools



"Data **visualization** is the key to **actionable insights** "
Head of Business Intelligence and Data Analytics (AccuWeather)

Exploratory vs. Explanatory

Leads to lots of tools/software
(desktop? web? mobile? print?)

tableau SOFTWARE

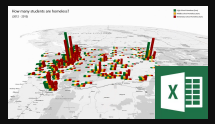
Google Analytics

ORACLE Visual Analyzer

FOLIUM
https://github.com/wrobstory/folium
"Python Data. Leaflet.js Maps."

Simplicity, familiarity & interpretability, & expectations

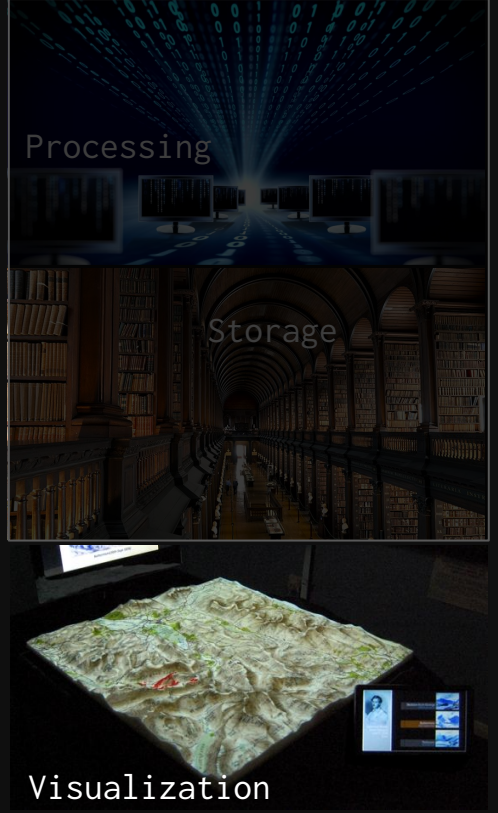
CARTODB
Geospatial on the cloud



python

How?

Many competing tools

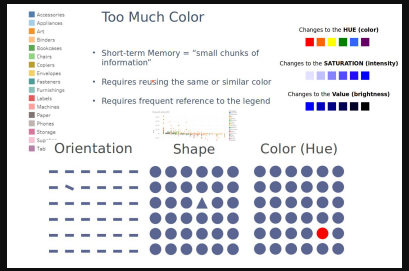


"Data **visualization** is the key to **actionable insights** "

Head of Business Intelligence and Data Analytics (AccuWeather)

Exploratory vs. Explanatory

Design choices / concepts



The Duell's Rules for Actionable Visualizations

The question to answer must be identifiable

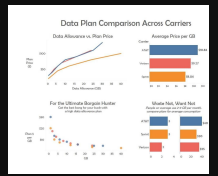
The data needed must be available

The visualization should be tailored to the person who will use the information

The story uncovered in the visualization should be evident

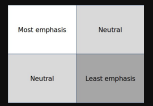
The action required should be clear

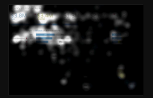
Simplicity, familiarity



interpretability, & expectations

[1]





[1] https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/

Data Management, Processing & Visualization*

	Normal Data	Big Data
Management (data storage)	Digital files | Hierarchical databases | Network databases Relational Databases: PostgreSQL, SQLite, IBM DB2, MySQL, Microsoft SQL Server, ORACLE DATABASE	Parquet, mongoDB, APACHE HBASE, HDFS, Google, Cockroach DB (Scalable, Survivable, Consistent And open source, of course.)
Processing	python, Java, C++, SQL, R, orange, WEKA The University of Waikato, scikit-learn, SPSS Modeler, rapidminer	HIVE, hadoop, STORM, NVIDIA CUDA, cloudera IMPALA, H2O.ai, beam, APACHE Spark, APEX
Visualization	FusionCharts, D3, Many language specific packages/frameworks, R + ggplot2, + python, FOLIUM (https://github.com/wrobstory/folium "Python Data. Leaflet.js Maps."), Other frameworks, + a b l e a u SOFTWARE, Google Analytics, ORACLE Visual Analyzer, Q, CARTODB Geospatial on the cloud	

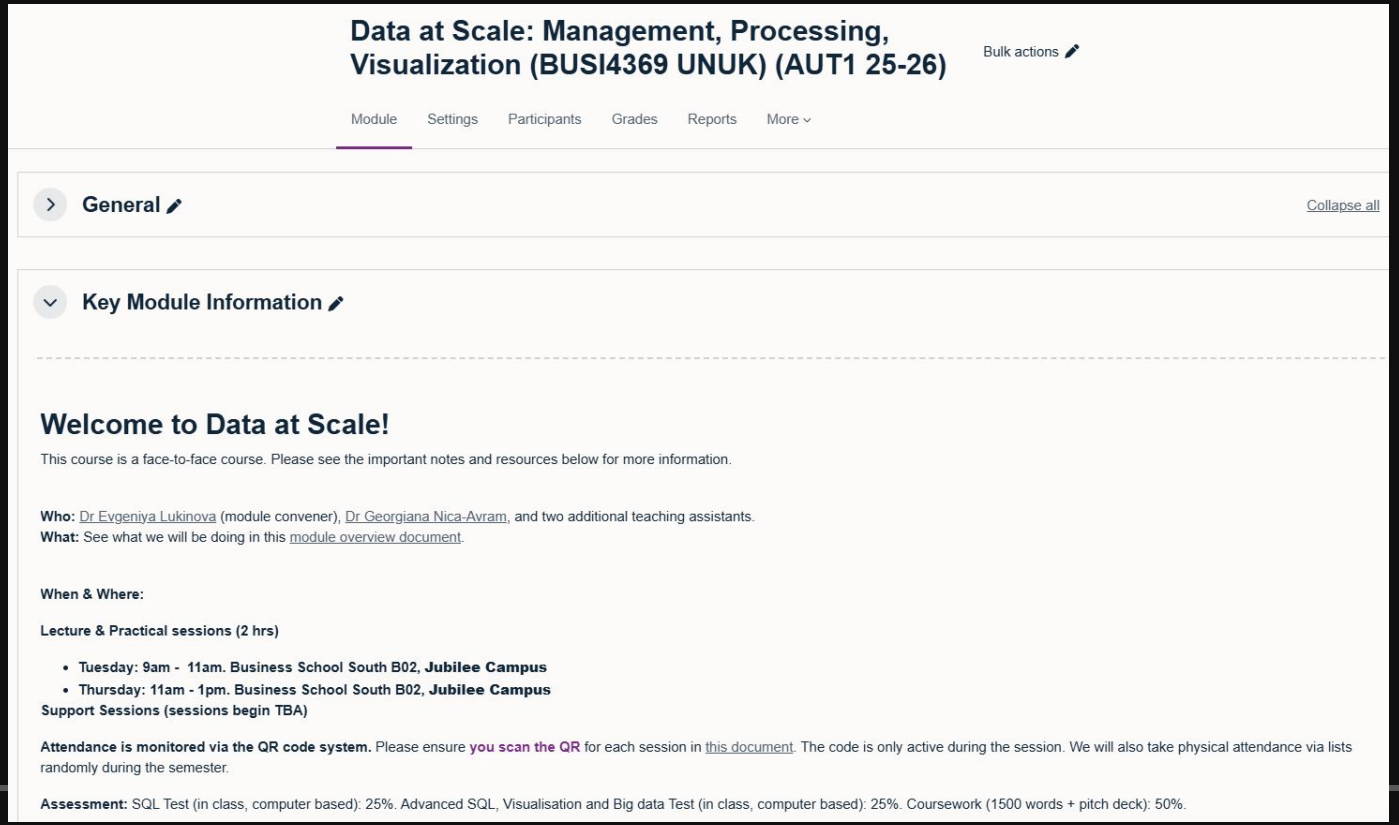
*Representative list

Module overview

NLAB: *Data at Scale*

Dr Evgeniya Lukinova

What is expected of you: Use Moodle!



What is expected of you: Use Moodle!

Data at Scale: At a glance.

Note: Schedule is subject to change.

Lecturers: Dr. Evgeniya Lukinova (module convenor) and Dr Georgiana Nica-Avram

Week 1	**Session 1 (Lecture & Practical):** The what and why of "data at scale". Introduction to the first concept of data storage and processing *(aka welcome to the course and the basics)*. **Session 2 (Lecture):** Introduction to relational and object based paradigms for storing and processing data at scale *(aka why relational databases will be part of your future job as an analyst)*. **Session 2 Vis (Lecture):** The theory behind making good visualisations *(aka how to visually manipulate people, for good…. or at least to help you get your message across)*.
Week 2	**Session 3 (Lecture):** Graph types and dashboards, the when, where and how you should use them building on the theory behind making good visualisations previously discussed. Part 1 *(aka, let's make nice pictures that drive "actionable insights")*. **Session 4 (Lecture & Practical):** An in-depth look and how-to guide to using Tableau for visualisation *(aka, creating visualisations to drive your point home)*.
Week 3	**Session 5 (Lecture & Practical):** SQL I: Foundations of relational databases in practice *(aka the basis of 20%+ of most data analytics jobs)*. **Session 6 (Lecture):** The theory behind relational database design and how it affects you as an analyst even though database design is not your job. How to read the (often poor) documentation left by database designers for the database you will use *(aka, how to work with what you've been given in your job)*.

What is expected of you: Use Moodle

Attendance will monitored via QR codes…
Although the google doc with QR codes is shared, we will transition to live codes with the new system SEAtS!

Assessments and Feedback

25% SQL Test (in-class, week 6). [open book, open web, no ChatGPT or similar, no collaboration with others]

25% Advanced SQL, Visualisation and Big Data Test (in-class, week 11). [closed book, 1 single-sided A4 page of handwritten notes allowed]

50% Coursework. Data analysis task based on a real-world based scenario and a data set (per group). 1500 word **individual** report and a **group** presentation.

Why 3 assessments? We opt for smaller, more regular assessment as continued study and feedback is the optimal way to learn the technical content (in this case SQL).

Deadline Date for Submission of Coursework
Due: 3PM, Thursday, 8th January 2026

What is expected of you: Work hard.

This is a technical module in a ***qualification*** .

You will learn and be certified to know how to undertake data analysis and visualization.

This takes time and practice, **even after you understand the higher level concepts** .

What is expected of you: Work hard.

This is a technical module in a ***qualification*** .

You will learn and be certified to know how to undertake data analysis and visualization.

This takes time and practice, **even after you understand the higher level concepts** .

Lectures
Guide you in what you need to learn, providing material and pointing to further resources.

Practicals
→ Consolidate the technical elements.

→ Provide real-world business use case examples.

What is expected of you: Work hard.

This is a technical module in a ***qualification*** .

You will learn and be certified to know how to undertake data analysis and visualization.

This takes time and practice, **even after you understand the higher level concepts** .

Lectures
Guide you in what you need to learn, providing material and pointing to further resources.

Practicals
→ Consolidate the technical elements.

→ Provide real-world business use case examples.

➔ Make sure you finish your practicals in your own time.

➔ Use the extra material if you are not confident with components of the course.

➔ Use the support session.

➔ Ask for help if you need it.

What is expected of you: Work hard.

Total learning time:
200 hours

Contact time:
44 - 55 hours [with support session]

Background study / coursework / revision:
145 - 156 hours

- ➔ Make sure you finish your practicals in your own time.
- ➔ Use the extra material if you are not confident with components of the course.
- ➔ Use the support session.
- ➔ Ask for help if you need it.

What is expected of you: Don't cheat.

Plagiarism means to pass off someone else's work, intentionally or unintentionally, as your own.

This might be by copying or paraphrasing someone's published or unpublished work without proper acknowledgment, or representing someone's artistic or technical work or creation as your own.

University of Nottingham
UK | CHINA | MALAYSIA

UK
China
Malaysia

Study | Research | Business | Global | About | A–Z | keyword(s)

University of Nottingham > Studying effectively > Writing > Avoiding plagiarism

Studying Effectively

- Home
- Studying at university
- Types of teaching
- Being organised
- Reading and interpreting sources and data
- **Writing**
  - Writing tasks at university
  - Strategies for writing
  - Referencing and citing
  - Avoiding plagiarism
    - Do you understand plagiarism
- Preparing for assessment

Avoiding plagiarism

Plagiarism means to pass off someone else's work, intentionally or unintentionally, as your own.

This might be by copying or paraphrasing someone's published or unpublished work without proper acknowledgment, or representing someone's artistic or technical work or creation as your own.

The University's policy on plagiarism

An act of Academic Misconduct is, generally speaking, any action in which may give a student an unpermitted academic advantage; as such, it is not acceptable in a scholarly community The most common examples of acts of Academic Misconduct are plagiarism, cheating in exams, collusion, and fabricating results or data. It can be, however, anything that gives you an unfair advantage in an assessment.

Incidences of placgiarism will first be addressed within the School, and they may apply penalties such as giving you a mark of zero for the piece of work concerned. The University's Academic Misconduct Committee has the power to apply a range of penalties for serious or repeated cases, including terminating your course.

Tips for avoiding plagiarism

- Don't just copy

Academic integrity

Quicklinks

Test your skills
- Plagiarism quiz

Further reading

Studying at university
- Academic integrity and plagiarism

Writing
- Referencing and citing

https://www.nottingham.ac.uk/studyingeffectively/writing/plagiarism/index.aspx

What is expected of you: Don't cheat.

An act of Academic Misconduct is, generally speaking, **any action in which may give a student an unpermitted academic advantage;** as such, it is not acceptable in a scholarly community.

The most common examples of acts of Academic Misconduct are **plagiarism, cheating in exams, collusion, and fabricating results or data**. It can be, however, **anything that gives you an unfair advantage in an assessment** .

University of Nottingham
UK | CHINA | MALAYSIA

UK China Malaysia

Study Research Business Global About A–Z keyword(s)

University of Nottingham > Studying effectively > Writing > Avoiding plagiarism

Studying Effectively

- Home
- Studying at university
- Types of teaching
- Being organised
- Reading and interpreting sources and data
- **Writing**
  - Writing tasks at university
  - Strategies for writing
  - Referencing and citing
  - Avoiding plagiarism
    - Do you understand plagiarism
- Preparing for assessment

Avoiding plagiarism

Plagiarism means to pass off someone else's work, intentionally or unintentionally, as your own.

This might be by copying or paraphrasing someone's published or unpublished work without proper acknowledgment, or representing someone's artistic or technical work or creation as your own.

The University's policy on plagiarism

An act of Academic Misconduct is, generally speaking, any action in which may give a student an unpermitted academic advantage; as such, it is not acceptable in a scholarly community The most common examples of acts of Academic Misconduct are plagiarism, cheating in exams, collusion, and fabricating results or data. It can be, however, anything that gives you an unfair advantage in an assessment.

Incidences of placgiarism will first be addressed within the School, and they may apply penalties such as giving you a mark of zero for the piece of work concerned. The University's Academic Misconduct Committee has the power to apply a range of penalties for serious or repeated cases, including terminating your course.

Tips for avoiding plagiarism

- Don't just copy

Academic integrity

Quicklinks

Test your skills
- Plagiarism quiz

Further reading

Studying at university
- Academic integrity and plagiarism

Writing
- Referencing and citing

https://www.nottingham.ac.uk/studyingeffectively/writing/plagiarism/index.aspx

What is expected of you: Don't cheat.

DON'T DO IT.

The University takes it very seriously. It protects the qualification you are all trying to earn.

→ Zeros. More pressure than before.
→ End of course.
→ Black marks on Record.
→ Ruined references from us when you apply for your job after.


University of Nottingham
UK | CHINA | MALAYSIA

UK
China
Malaysia

Study Research Business Global About A–Z keyword(s)

University of Nottingham > Studying effectively > Writing > Avoiding plagiarism

Studying Effectively

Home
Studying at university
Types of teaching
Being organised
Reading and interpreting sources and data
Writing
Writing tasks at university
Strategies for writing
Referencing and citing
Avoiding plagiarism
Do you understand plagiarism
Preparing for assessment

Avoiding plagiarism

Plagiarism means to pass off someone else's work, intentionally or unintentionally, as your own.

This might be by copying or paraphrasing someone's published or unpublished work without proper acknowledgment, or representing someone's artistic or technical work or creation as your own.

The University's policy on plagiarism

An act of Academic Misconduct is, generally speaking, any action in which may give a student an unpermitted academic advantage; as such, it is not acceptable in a scholarly community The most common examples of acts of Academic Misconduct are plagiarism, cheating in exams, collusion, and fabricating results or data. It can be, however, anything that gives you an unfair advantage in an assessment.

Incidences of placgiarism will first be addressed within the School, and they may apply penalties such as giving you a mark of zero for the piece of work concerned. The University's Academic Misconduct Committee has the power to apply a range of penalties for serious or repeated cases, including terminating your course.

Tips for avoiding plagiarism

- Don't just copy

Academic integrity

Quicklinks

Test your skills
- Plagiarism quiz

Further reading

Studying at university
- Academic integrity and plagiarism

Writing
- Referencing and citing

https://www.nottingham.ac.uk/studyingeffectively/writing/plagiarism/index.aspx

What is expected of you: Don't cheat.

DON'T DO IT.

The University takes it very seriously. It protects the qualification you are all trying to earn.

→ Zeros. More pressure than before.
→ End of course.
→ Black marks on Record.
→ Ruined references from us when you apply for your job after.


University of Nottingham
UK | CHINA | MALAYSIA

UK
China
Malaysia

Study Research Business Global About A–Z keyword(s)

University of Nottingham > Studying effectively > Writing > Avoiding plagiarism

Studying Effectively

Home
Studying at university
Types of teaching
Being organised
Reading and interpreting sources and data
Writing
Writing tasks at university
Strategies for writing
Referencing and citing
Avoiding plagiarism
Do you understand plagiarism
Preparing for assessment

Avoiding plagiarism

Plagiarism means to pass off someone else's work, intentionally or unintentionally, as your own.

This might be by copying or paraphrasing someone's published or unpublished work without proper acknowledgment, or representing someone's artistic or technical work or creation as your own.

The University's policy on plagiarism

An act of Academic Misconduct is, generally speaking, any action in which may give a student an unpermitted academic advantage; as such, it is not acceptable in a scholarly community The most common examples of acts of Academic Misconduct are plagiarism, cheating in exams, collusion, and fabricating results or data. It can be, however, anything that gives you an unfair advantage in an assessment.

Incidences of placgiarism will first be addressed within the School, and they may apply penalties such as giving you a mark of zero for the piece of work concerned. The University's Academic Misconduct Committee has the power to apply a range of penalties for serious or repeated cases, including terminating your course.

Tips for avoiding plagiarism

- Don't just copy

Academic integrity

Quicklinks

Test your skills
- Plagiarism quiz

Further reading

Studying at university
- Academic integrity and plagiarism

Writing
- Referencing and citing

https://www.nottingham.ac.uk/studyingeffectively/writing/plagiarism/index.aspx

What is expected of you: Don't cheat.

Quick note about AI (e.g., ChatGPT)

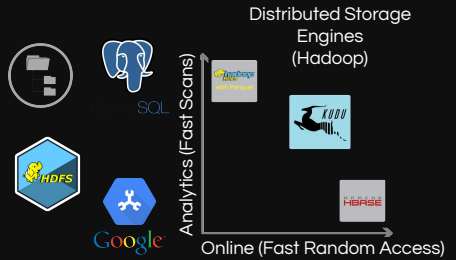
- The university considers the unauthorised use of AI tools false authorship.
- It will be clearly communicated to you whether you may use AI tools in assessment and how you are permitted to do so.
- **PLEASE ADHERE!**

https://www.nottingham.ac.uk/currentstudents/news/using-ai-tools-in-your-studies

A gentle start...

How to organise data in physical space?

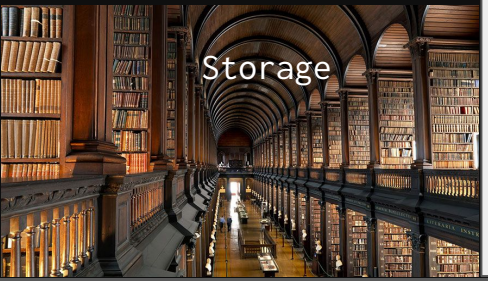
Data analytics starts with loading data from somewhere!



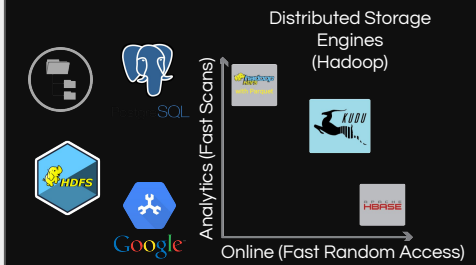
A gentle start...

How to organise data in physical space?

Data analytics starts with loading data from somewhere!



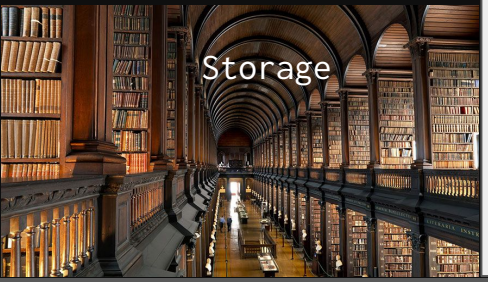
Option 1 (today):
→ Store data as "objects" (**file**)
→ Store objects in a hierarchy (**tree structure**)



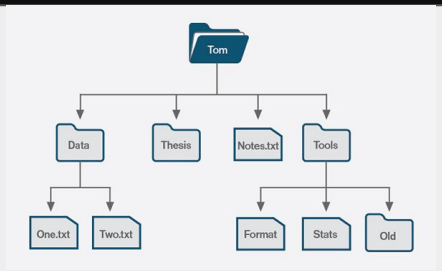
A gentle start...

How to organise data in physical space?

Data analytics starts with loading data from somewhere!



Option 1 (today):
→ Filesystem of traditional computers



Windows

mac OS

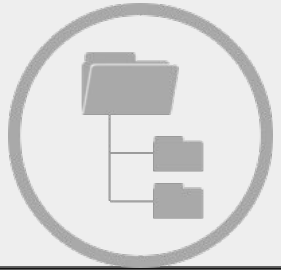
ubuntu

A gentle start...

How to organise data in physical space?

Data analytics starts with loading data from somewhere!

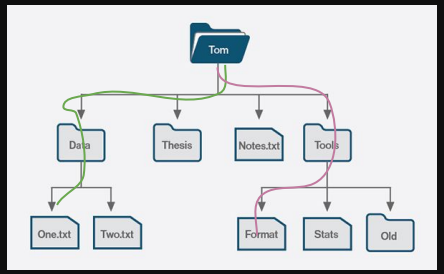
Option 1 (today):
→ Store data as "objects" (**file**)
→ Store objects in a hierarchy (**tree structure**)



Benefits:
+ Store arbitrary things
+ Conceptually simple
+ Tree structure means single path locates all files

Negatives:
- Coming next, when we look at alternatives

Objects stored as trees.

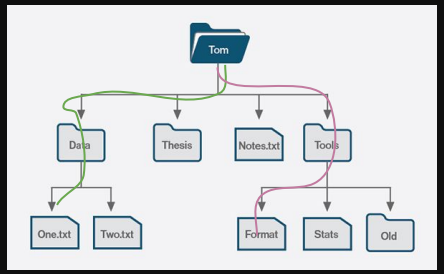


Each object's location can be defined by a path.

Tom/Data/One.txt

Tom/Tools/Format

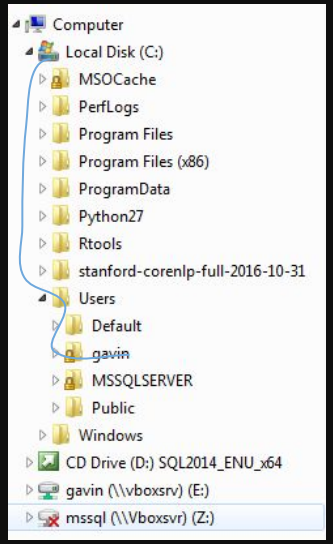
Objects stored as trees.



Each object's location can be defined by a path.

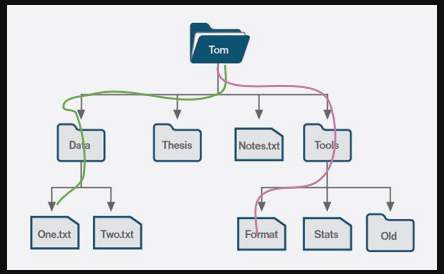
Tom/Data/One.txt

Tom/Tools/Format



C:\Users\gavin

Objects stored as trees.



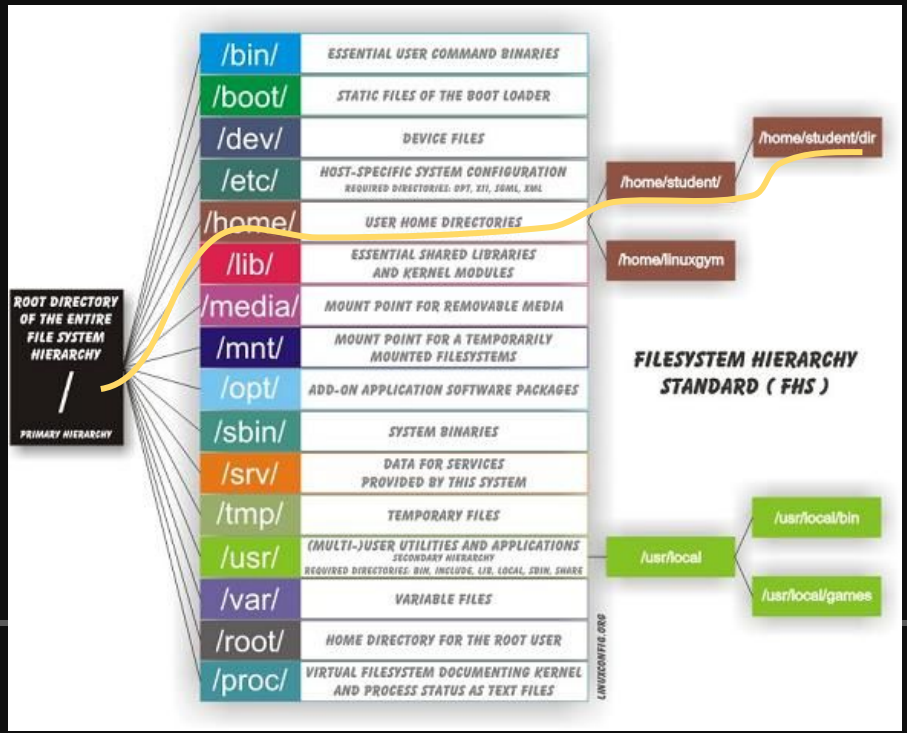
Each object's location can be defined by a path.

Tom/Data/One.txt

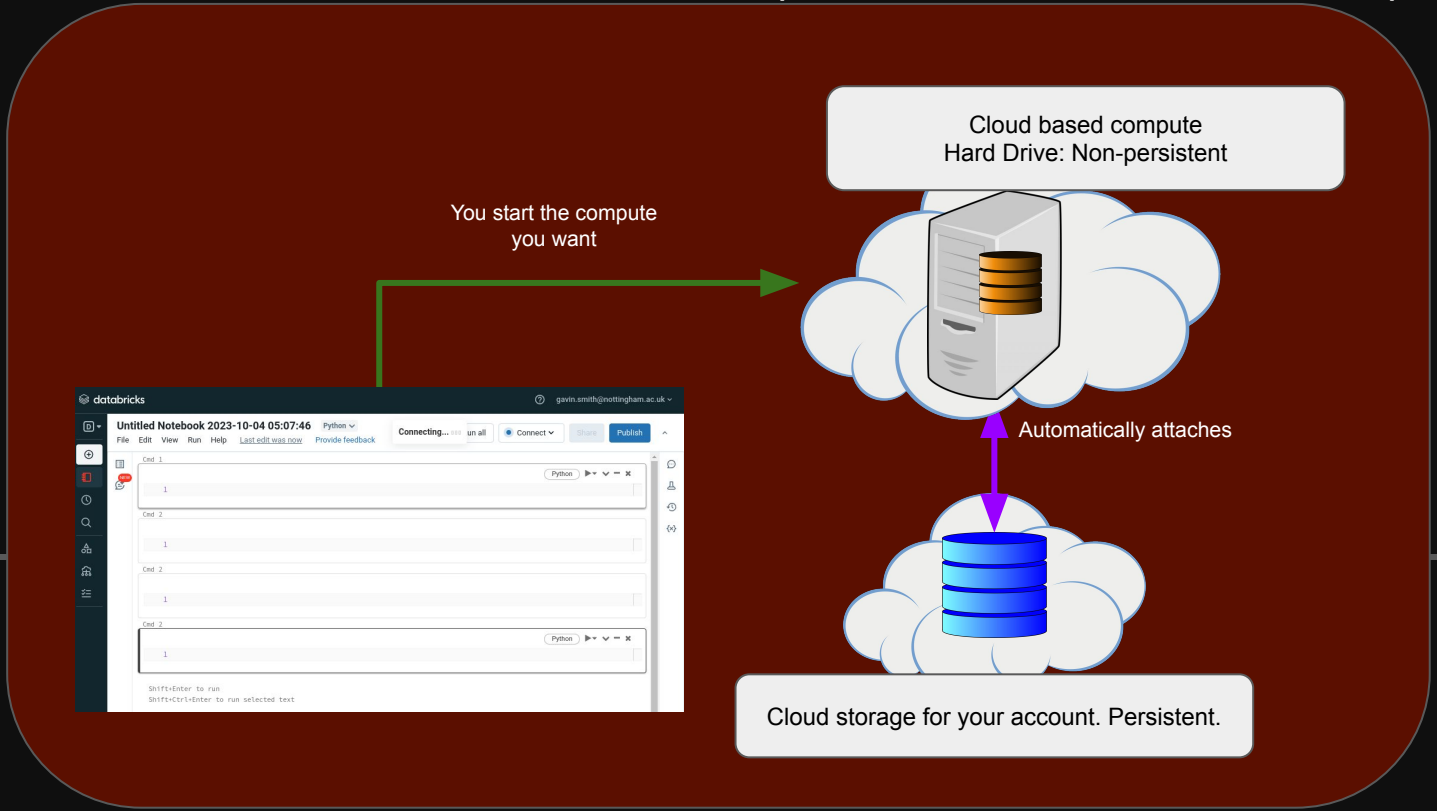
Tom/Tools/Format

/home/student/dir

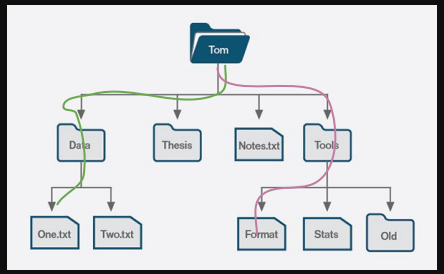
ubuntu



In cloud based systems we often have more than one "hierarchical object store" (filesystem).
… But Databricks Free Edition is restricted to one workspace, one metastore, one object store.

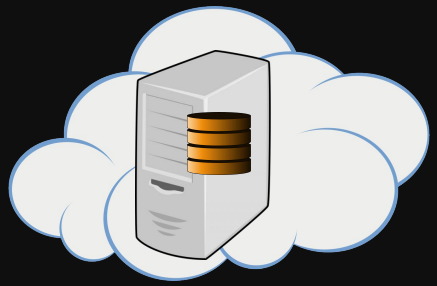


Objects stored as trees.



Tom/Data/One.txt

Tom/Tools/Format



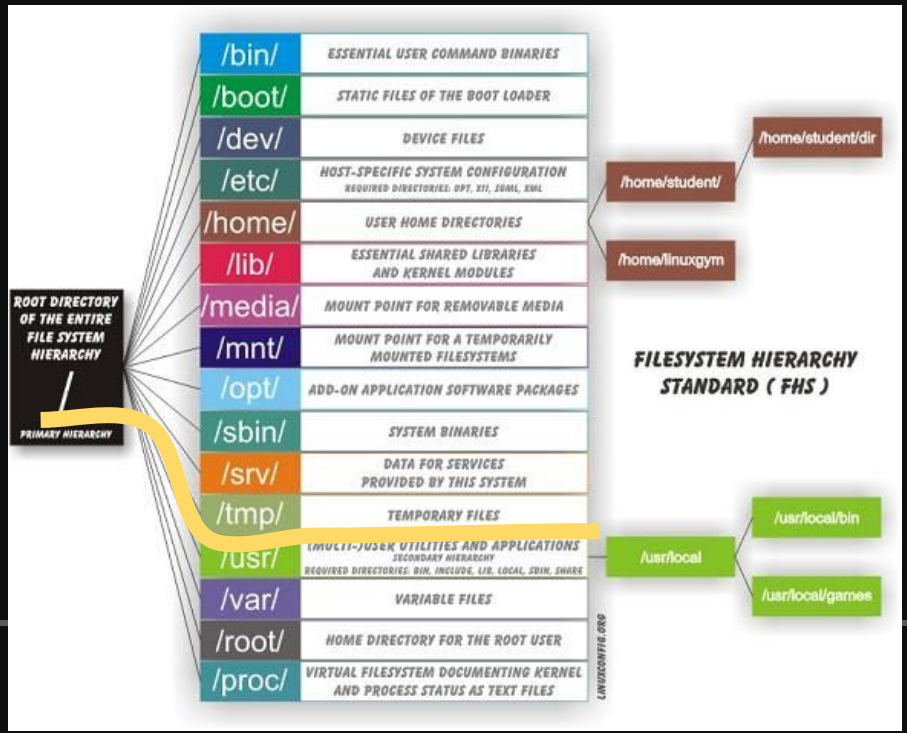
Temporary compute has a hard drive and runs a linux version.

Each object's location can be defined by a path.

Can use the temporary folder if needed (**typically won't**).

/tmp/

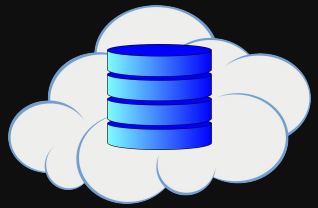
databricks



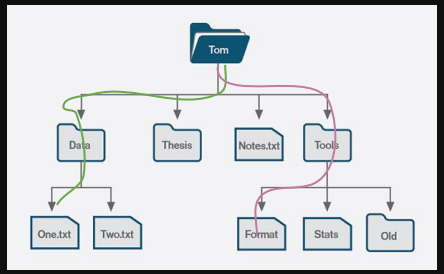
NLAB: *Data at Scale*

Dr Evgeniya Lukinova

Objects stored as trees.

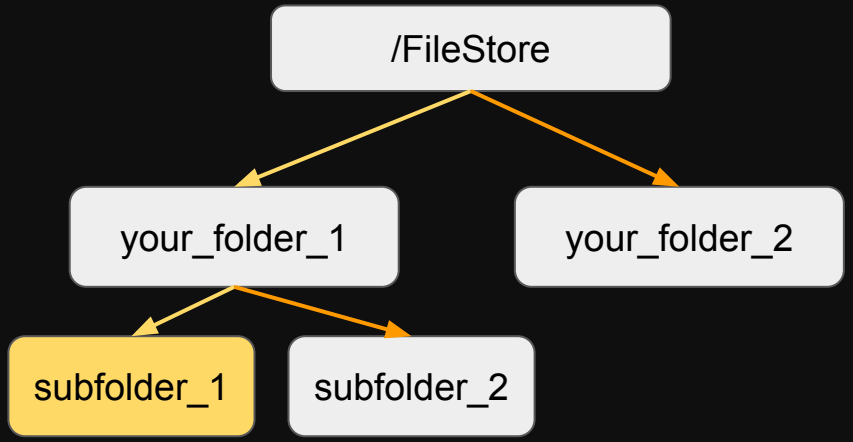


databricks



Persistent cloud filestore also stores objects as a tree.

Start with just a root node "/Filestore"

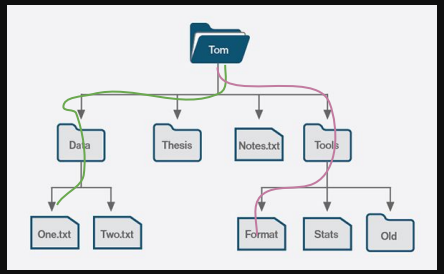


Tom/Data/One.txt

Tom/Tools/Format

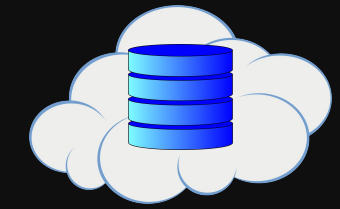
/Filestore/your_folder1/subfolder1

Objects stored as trees.



Tom/Data/One.txt

Tom/Tools/Format

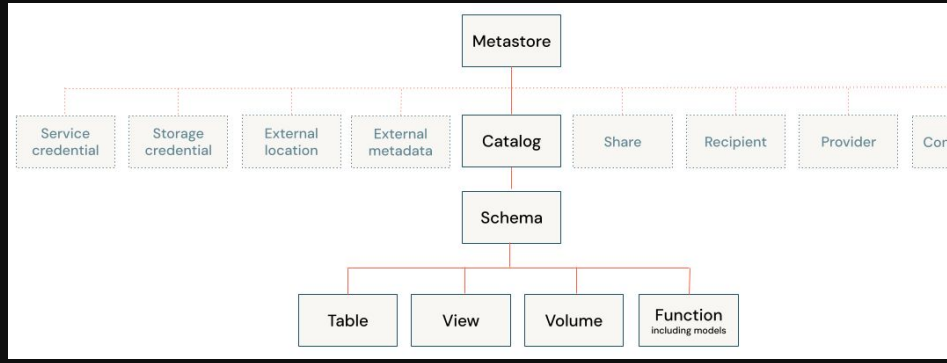


databricks

Persistent cloud storage also stores objects as a tree.

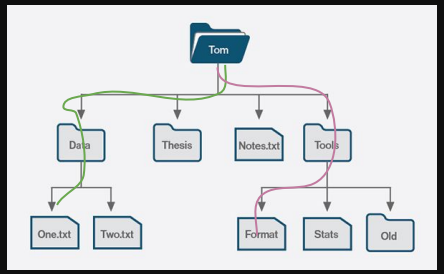
The root is the **Unity Catalog** metastore, under which you define catalogs, schemas, and tables.

catalog.schema.table



https://docs.databricks.com/aws/en/data-governance/unity-catalog/

Objects stored as trees.



Each object's location can be defined by a path.

Tom/Data/One.txt

Tom/Tools/Format

Why do I care?

To store / load data for analytics you need to know where it is.

Selecting files via a graphical interface doesn't scale.

Graphical interfaces often cannot handle large data…. as we'll see in the practicals!

Try the Practical!