INTRO TO COS80023 BIG DATA

IRENE MOSER

Acknowledgement of Country

We respectfully acknowledge the Wurundjeri People of the Kulin Nation, who are the Traditional Owners of the land on which Swinburne's Australian campuses are located in Melbourne's east and outer east, and pay our respects to their Elders past, present and emerging.

We are honoured to recognise our connection to Wurundjeri Country, history, culture, and spirituality through these locations, and strive to ensure that we operate in a manner that respects and honours the Elders and Ancestors of these lands.

We also respectfully acknowledge Swinburne's Aboriginal and Torres Strait Islander staff, students, alumni, partners and visitors.

We also acknowledge and respect the Traditional Owners of lands across Australia, their Elders, Ancestors, cultures, and heritage, and recognise the continuing sovereignties of all Aboriginal and Torres Strait Islander Nations.

CONTENT

- Why Big Data
- Storing Big Data
- Processing, Retrieving and Analysing Big Data
- Machine Learning in Big Data
  - Including Natural Language Processing

UNIT FORMAT

- Lectures
  - Recording (mp4)
  - Online sessions (please come with questions from listening to recording)
- Face-to-face tutorials
  - Tasks to submit on Canvas
  - Marked during tutorial in discussion between student and tutor

ASSESSMENT

- ePortfolio
  - Pass tasks
  - Learning summary
  - Reflection

  - Credit tasks
  - Distinction and/or HD tasks
  - Interview (15mins)

You choose your target grade

WAIT, WHAT??

- Pass tasks
  - Pass task 1 ✓ = 1
  - Pass task 2 ✓ = 1
  - Pass task 3 ⊘ = 0
  - Pass task 4 ✓ = 1
  - Pass task 5 ✓ = 1
  - Pass task 7 ✓ = 1
  - Pass task 8 ✓ = 1
  - Pass task 9 ✓ = 1
  - Learning summary ✓ = 1
  - Reflection ✓ = 1

Fail!

Will not be marked!

- Credit tasks
- Distinction and/or HD task project
- Interview

ASSESSMENT

Grade	Requirements		
Pass	- Pass tasks marked as complete (= 1 mark) - Learning summary submitted with acceptable content		
Credit	- Pass level requirements achieved - Credit tasks marked as complete (= 1 mark)		
Distinction	- Credit level requirements achieved - Project report submitted - Project contains practical component - Project report/interview at D level	High Distinction	- Credit level requirements achieved - Project report submitted - Project contains practical component - Project report is of advanced standard with good results and logical discussion - Interviewee provides competent, comprehensive answers

D/HD PROJECT

1. Fill in Project Plan under Assignments
   - What do you want to find out (question you want to answer)
   - How will you do it?
   - What tools will you use?
2. Approved by tutor
3. Do the investigative work
4. Write and submit a report

ASSESSMENT

Pass Tasks	Lots of support and guidance (not a lot of googling)
Credit Tasks	More independent study, figuring things out (more googling!)
Distinction /High Distinction Project	Largely independent study (lots of googling!)

What does it mean to have a HD?

+ Portfolio document

+ Interviews

FINAL GRADING

Fail Does not meet Pass standard.		Portfolio not submitted. or One or more Pass tasks not signed off as Complete, or Fails to demonstrate coverage of all unit learning outcomes.				
Pass Pass tasks are marked as Complete. Learning summary report submitted.	**50 P** All Learning outcomes covered, but Pass tasks very late or learning summary poor.	**53 P** Meets Pass, tasks on time, but poor reflections or learning summary.	**55 P** Meets Pass with acceptable reflections and learning summary, tasks on time.	**57 P** Meets Pass with good reflections and learning summary, tasks on time.		Learning Summary Report + Pass tasks
Credit Passed, and all Credit tasks are Complete.	**60 C** Meets Pass, but tasks poor or late, or learning summary poor.	**63 C** Meets Credit, but some issues with tasks, reflections, or learning summary.	**65 C** Meets Credit, with generally good tasks, reflections and learning summary.	**67 C** Meets Credit with good tasks, reflections and learning summary.		+ Credit tasks
Distinction Passed, all Credit tasks are Complete, D/HD plan approved, report has been submitted, interview attended, project at D level.	**70 D** Meets credit, but major flaws with D/HD project report, tasks late or learning summary poor.	**73 D** Meets Distinction, but some issues with the design or implementation, reflections or learning summary.	**75 D** Meets Distinction, with good design and implementation, reflections and learning summary.	**77 D** Meets Distinction, with well thought-through design and implementation, reflections and learning summary.		+ D/HD report + 15 min Interview
High Distinction As under Distinction, but report of higher quality, i.e. substantial investigation documented. Report informative and understandable.		**83 HD** Excellent outcomes, good discussion of investigation and details, but some weaknesses with report.	**85 HD** Excellent outcomes, discussion of investigation, good insights or comparison with alternative methods.	**87 HD** All outcomes are excellent with very high quality finish, including the discussion of alternatives.		+ Excellent quality
High Distinction, plus a research report that demonstrates ability to conduct a small research project, analyse findings and make conclusions of a substantial or demanding project.		**93 HD** All outcomes are excellent, research substantial and report with good insights or comparisons.	**95 HD** All outcomes are excellent, research substantial and report with excellent insights or comparisons	**97 HD** All outcomes are excellent, including investigation method, report and analysis.	**100 HD** Something special.	+Outstanding quality

No score points?

See "Portfolio Format and Assessment Criteria" document on Canvas

TIMELINE

- Pass/Credit tasks
  - Start work in the week of the tutorial
    - e.g. Week 1
  - Make corrections if feedback received
  - Last chance to have it marked in the following tutorial
    - e.g. Week 2
- Timeliness considered in final marking
- After due date, no feedback
  - You must get it right on your own.

FIRST HALF OF SEMESTER

Week	Teaching and Learning Activity	Student Task or Assessment
1	Unit structure and expectations Big Data – Opportunities and Challenges	Start Pass Tasks 1
2	Big Data Storage Introduction to Azure	Start Pass Tasks 2 Pass tasks 1 marked as correct
3	Big Data Processing and ETL	Start Pass Tasks 3 Pass tasks 2 marked as correct
4	Big Data Parallelisation and Distribution	Start Pass Tasks 4 Pass tasks 3 marked as correct
5	Big Data Information Retrieval	Start Pass Tasks 5 Pass tasks 4 marked as correct
6	Industry speaker	Start Credit Tasks 6* Pass tasks 5 marked as correct

SECOND HALF OF SEMESTER

Mid-Semester Break		
7	Big Data Machine Learning	Start Pass Tasks 7 Credit Tasks 6 marked as correct*
8	Big Data Classifiers	Start Pass Tasks 8 Pass Tasks 7 marked as correct
9	Big Data – Natural Language Processing	Start Pass Tasks 9 Pass Tasks 8 marked as correct
10	Big Data – Advanced topics and D/HD topics	Start Credit Tasks 10* Pass Tasks 9 marked as correct
11	D/HD project consultation	Start D/HD Project* Credit Tasks 10 marked as correct*
12	D/HD project consultation	Work on D/HD Task

*optional

THE GOOD BIT

- You can choose how high you want to aim
  - If your pass tasks are late, just make sure you achieve a pass
  - If you are up for a project, study the dataset and decide on an investigation
  - Submit a plan on Canvas and have a tutor approve it
- You know where you are going at all times
  - If you have your pass tasks signed off (= set to 1 on Canvas) and your portfolio document submitted, no surprise fails!
  - If you stick to your project plan, you know if you can achieve your D or HD

QUESTIONS?

Log in to the first
Live Online session!

BIG DATA

OPPORTUNITIES AND CHALLENGES

LEARNING OBJECTIVES

- At the end of this presentation, you should be able to
  - explain what we mean by Big Data;
  - explain what opportunities Big Data has to offer;
  - explain how we tap into those opportunities;
  - understand how IoT and social media have contributed to Big Data;
  - explain what difficulties we can face making use of Big Data.

WHAT IS BIG DATA?

- ”A data set that does not fit on a single computer”

Volume

Variety

Velocity

Veracity

Valence

SWINBURNE UNIVERSITY OF TECHNOLOGY

BIG DATA IN NUMBERS

Single page .docx
≈ 50kB

Terabyte = 10^{12}

≈ 20 million single page .docx files

Can still fit on a very large server

Petabyte = 10^{15}

≈ 20000 million single page .docx files

Definitely does not fit on a single computer

WHY ARE PEOPLE INTERESTED IN IT?

Example

- Information
  - We can collect information about our customers to better understand what they want.
    - Lucy has just bought a motorbike. Maybe she wants insurance, or boots, or a jacket.
- ”360° view”
  - We can combine data about a customer to know better what they want.
    - According to her Facebook page, Lucy plans a road trip. She needs pannier bags.
- Automation
  - We can use programs to do useful things with this information.
    - An algorithm can work out what Lucy is up to and mail her ads about pannier bags.

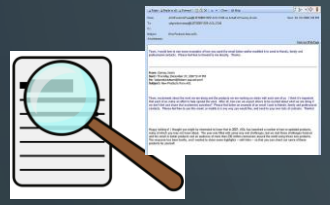
HOW TO BENEFIT FROM BIG DATA - BUSINESS INTELLIGENCE (ANALYTICS)

- Organisations can analyse the data to find out what is happening
- Descriptive analysis
  - Find out what is happening
- Predictive analysis
  - Find out what is likely to happen in the future
- Prescriptive analysis
  - A mix of descriptive and predictive

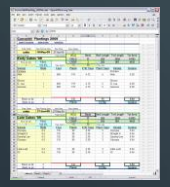
HOW TO BENEFIT FROM BIG DATA

- Organisations can just sell it!
  - Other organisations pay big money if the data is useful to them.
- Google, X, Facebook, Whatsapp, Telstra, TomTom
  - all collect data on a large scale.
  - Facebook has sold user data to Cambridge Analytica
  - TomTom and Telstra collect travel data using mobile phones
- Scientific research.

WHERE DOES BIG DATA COME FROM?



- People (Online activity, Social Media)
  - Emails, tweets, posts, google searches, documents, pictures, videos



- Organisations
  - Data about business transactions, products.

- Devices (”IoT”)
  - Smart meters, sensors

BIG DATA OPPORTUNITY: LINKING ONLINE DATA

Example

360° view

- We already know that Lucy has bought a motorbike.
  - Her birthday is also coming up.
  - She has a mother who is likely to want to buy her a present.
  - She has a sister who is pregnant.
  - Her sister's Facebook page says it's a boy.
- If we can mine and extract all this information
  - We can advertise motorcycle gear to Lucy's mother as a birthday present.
  - We can advertise baby boys' jumpsuits to Lucy for her sister's baby shower.

Targeted advertising

WHAT IS THE INTERNET OF THINGS?

The Internet of Things (IoT) is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.

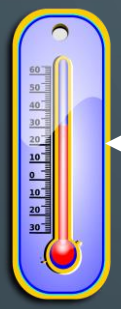
OPPORTUNITIES OF BIG DATA: IOT APPLICATION

Example

- Many people do not like living in nursing homes.
  - Elderly residents are at risk of falls.
  - If a person is injured, lying on the floor, they cannot reach a phone to alert help.
- IoT can employ sensors and machine learning to keep frail people safe.
  - Wearable technology and sensors record where a person spends time.
  - Machine learning algorithms learn what 'normal' behaviour is for a person.
  - When a person's behaviour becomes 'abnormal', help is called.

OPPORTUNITIES OF BIG DATA: IOT APPLICATION

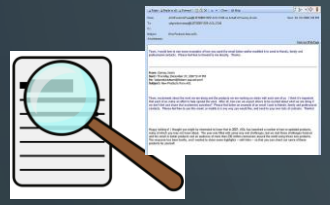
Example



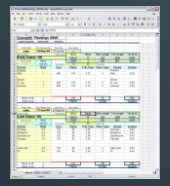
Air conditioner can work out best temperature by accessing body temperature, ambient temperature and historical data

Internet

WHAT FORMAT DOES BIG DATA COME IN?



- People (online data)
  - Mostly text – unstructured or semistructured

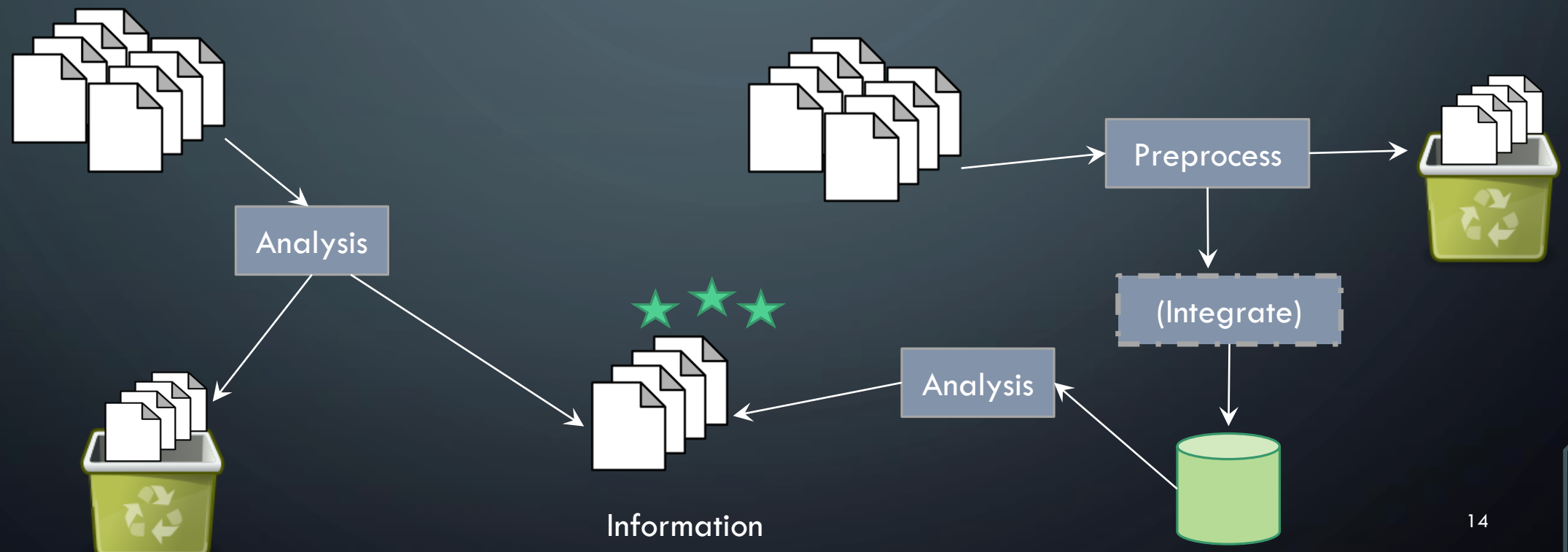


- Organisations
  - Mostly transactions - structured

- Devices (IoT)
  - Readings, pictures – unstructured, structured or semistructured

HOW DO WE DEAL WITH BIG DATA?

- Access only once
- Access several times



CHALLENGES: HOW TO STORE

Speed

Security

Safety

Scalability

CHALLENGES: HOW TO INTERPRET

Synonyms — throw ≈ toss

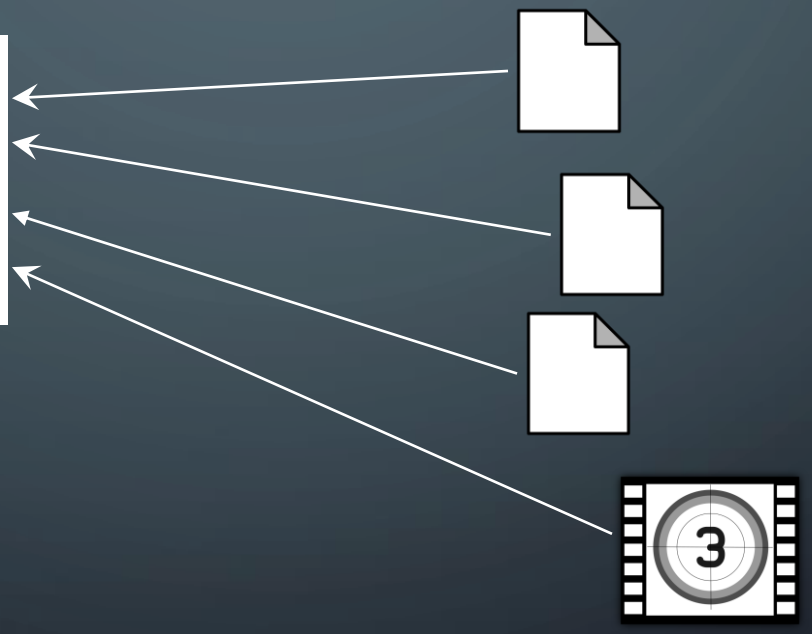
Spelling — requset ≈ request

Homographs — bow

CHALLENGES: HOW TO INTEGRATE

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

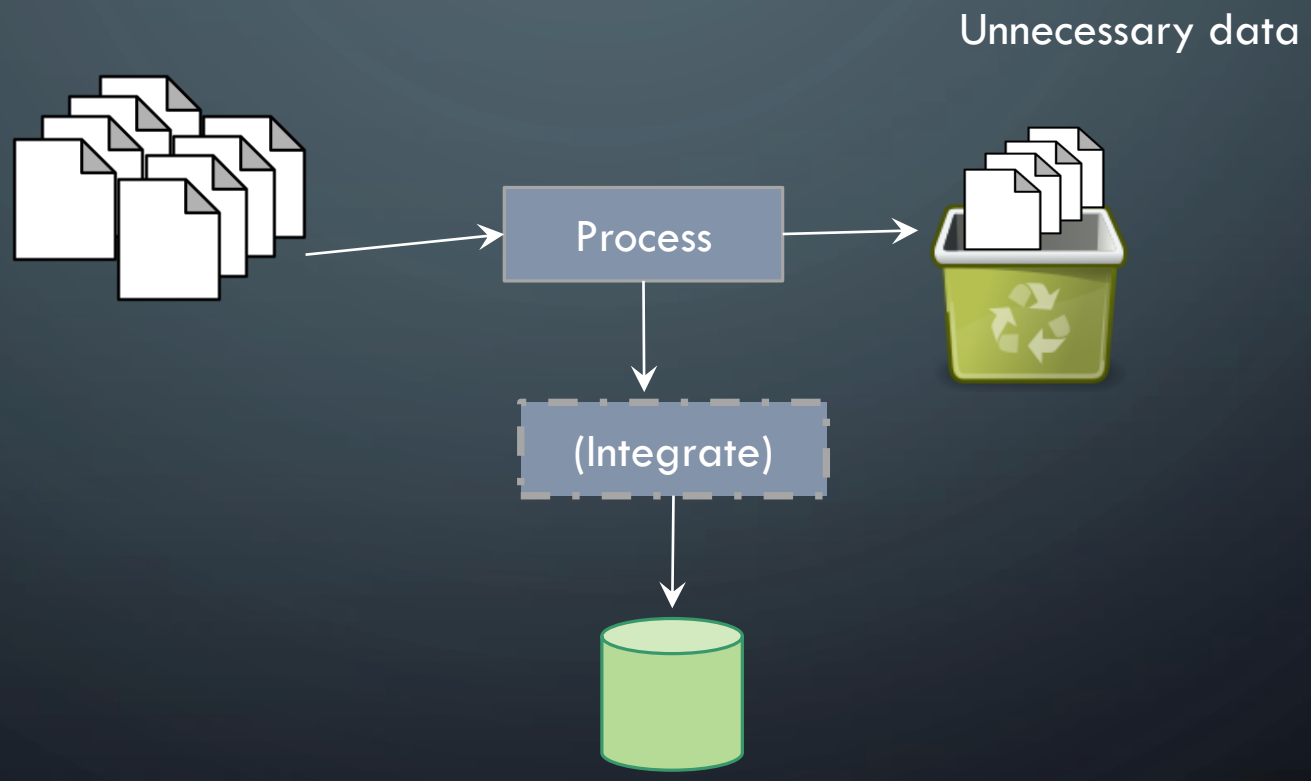
structured



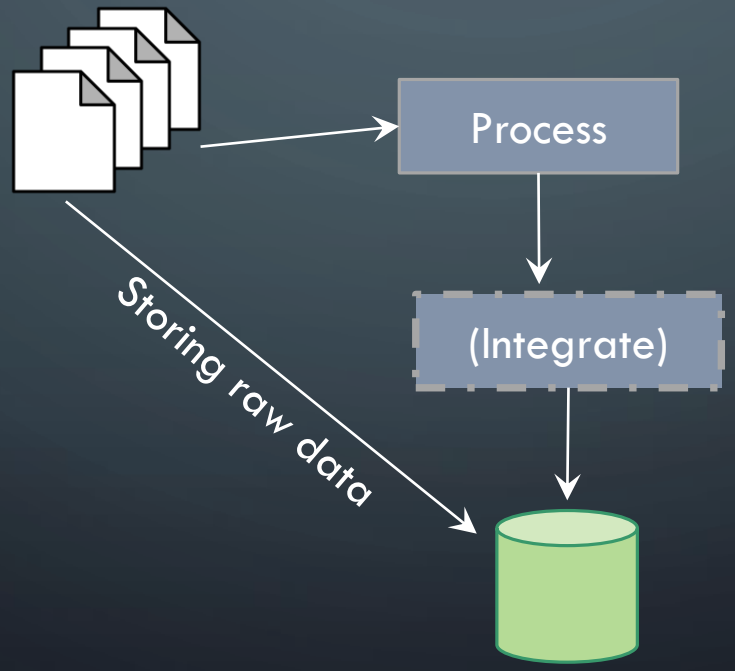
Whose email
or tweet
or FB post is it?

unstructured

CHALLENGES: WHAT TO KEEP



CAN WE JUST KEEP EVERYTHING?



Bulk

Security

Privacy

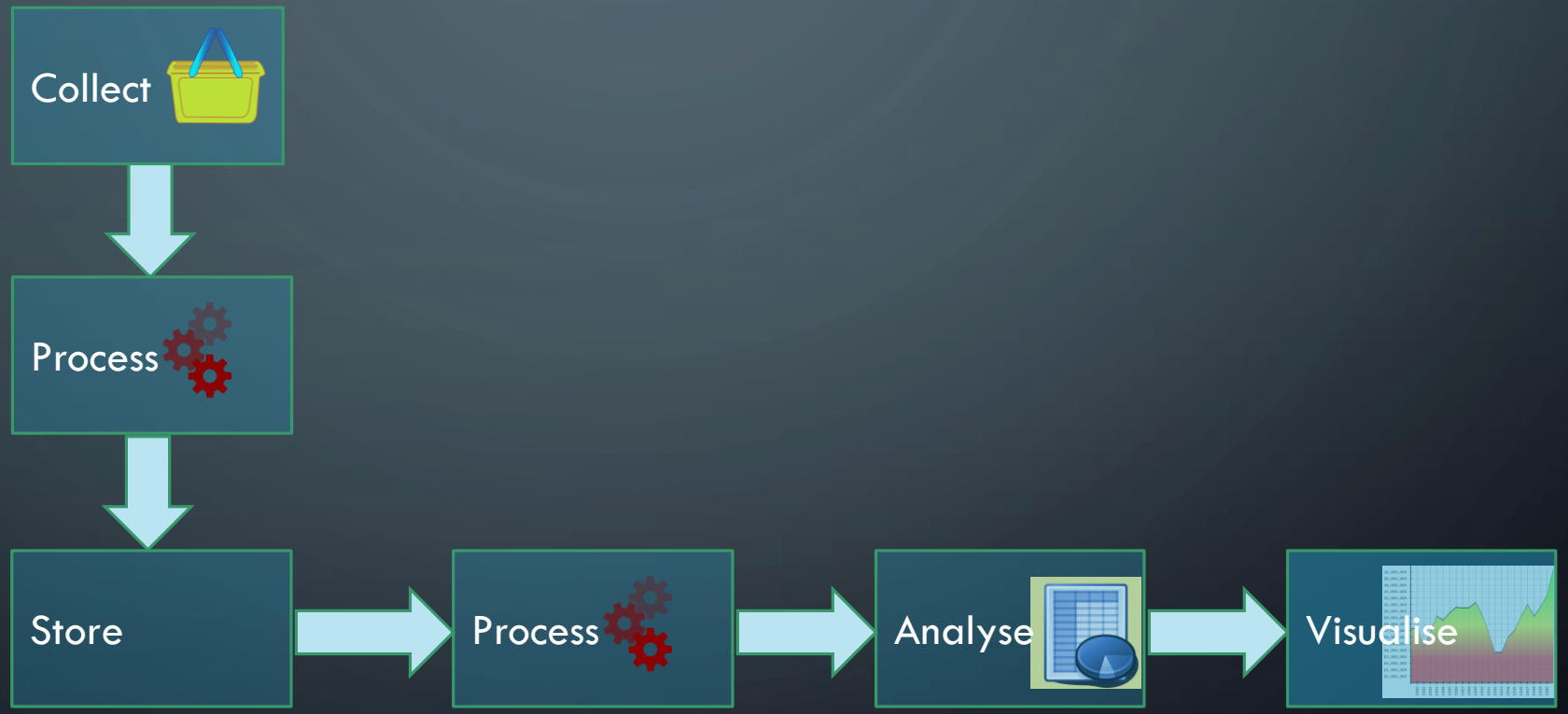
PRIVACY AND GDPR

- European Union's General Data Protection Regulation
  - Collect minimal data
  - Use only for original purpose
  - De-identify if possible
  - Keep within EU
  - Disclose data breaches immediately
  - Right of access

Fines are substantial

Effective May 2018

THE BIG DATA PROCESS



SUMMARY

- Big Data is valuable to many people.
- It comes form diverse sources and is often collected for a purpose.
- Integrating diverse data sources often leads to maximum benefit.
- Big Data often comes with challenges:
  - How to store
  - How to integrate
  - How to interpret
  - What to keep
  - How to secure