COS10022 – DATA SCIENCE PRINCIPLES

Dr Pei-Wei Tsai (Lecturer, Unit Convenor)
ptsai@swin.edu.au, EN508d

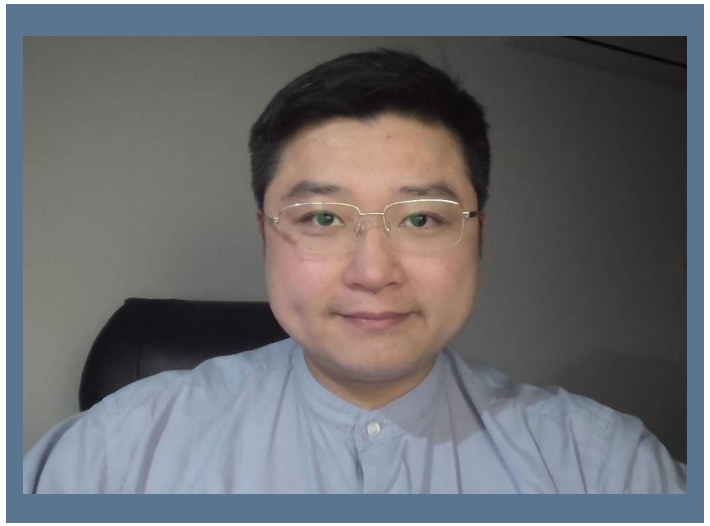
SWINBURNE UNIVERSITY OF TECHNOLOGY

Welcome to COS10022

Data Science Principles

- Asynchronized Short Videos
- Live Online Lectures
- Laboratories
- Online Tests (week 5, 9, and 12)
- Assignments (Week 4 and 10)

https://creativecommons.org/licenses/by-nc-sa/3.0/



- Lecturer, Tutor, and Convenor
  - Dr Pei-Wei Tsai
    ptsai@swin.edu.au
    Office: EN508d
  - Main Research Fields:
    - Intelligent Optimisation
    - Data Analytics
    - Machine Learning

https://creativecommons.org/licenses/by-sa/3.0/

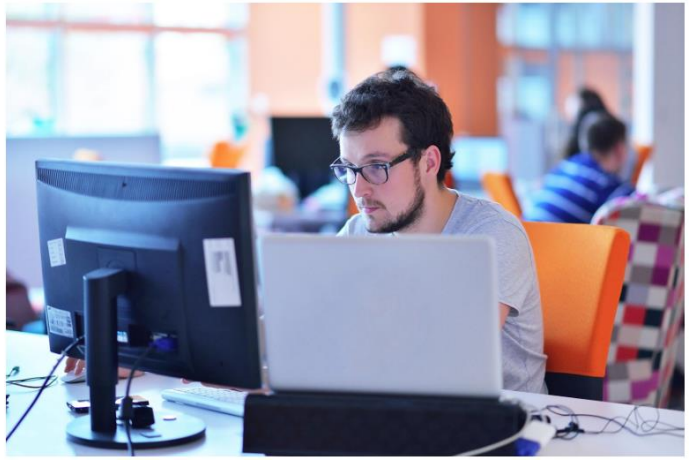
ABOUT DATA SCIENCE

indeed hiring lab | US

Occupation Spotlight

Data Scientist: A Hot Job That Pays Well

January 17, 2019 by Andrew Flowers @andrewflowers



SPOTLIGHT ON BIG DATA

Spotlight

ARTWORK Tamar Cohen, *Andrew J Buboltz*
2011, silk screen on a page from a high school yearbook, 8.5" x 12"

Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review **October 2012**

About Data Scientist

- The Sexiest Job of the 21st Century
  - by Thomas H. Davenport and D. J. Patil. (2012)
- A Hot Job that Pays Well
  - by Indeed hiring lab (on Occupation Spotlight, 2019)
- Data Science has emerged as a crucial field for organisations across various industries as they seek to make data-driven decisions.
  - by Anastasia Rashevskaya (2024).

https://www.hiringlab.org/2019/01/17/data-scientist-job-outlook/

Data Scientist

#8 in 100 Best Jobs

Data scientists use technology to glean insights from large amounts of data they collect. Read More »

Projected Jobs
59,400

Median Salary
$103,500

Education Needed
Bachelor's

2024 Data Scientist Salary Guide - Australia

Published on January 19

Data Scientist Salary Estimate: Overview

The Annual Data Scientist Salary Report for Australia

Data scientists in Australia, you're in for some exciting data insights! In 2024, the average salary for a data scientist is AU$93,760. This figure is more than just a number; it reflects the growing importance of data science in the business world.

Globally, data scientist salaries differ significantly, as per various data sources. In the US, the average is $156,717, contrasted by Australia's AU$91,703. Canada offers C$80,364, while India sees about INR 874,113. These disparities highlight the diverse valuation of data science skills

https://datablokes.com.au/blog/data-science-salaries-in-australia

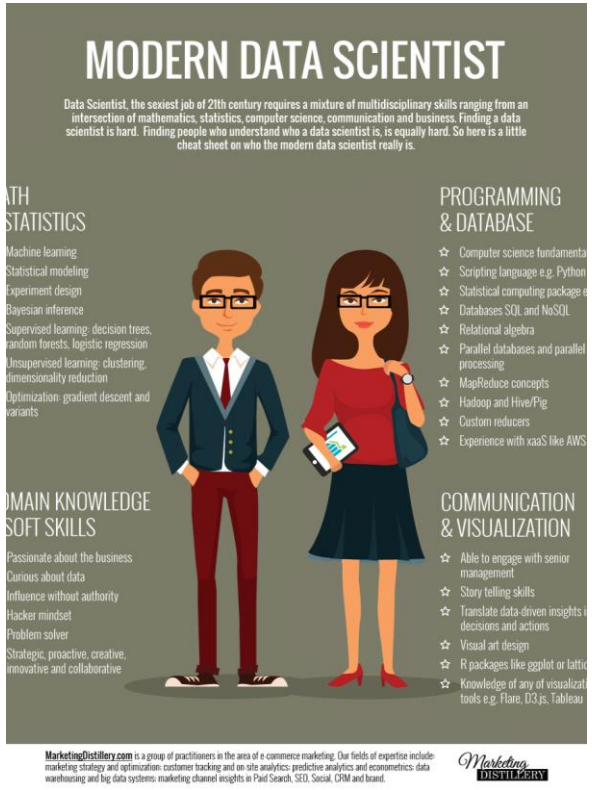
Indeed's Best Jobs of 2024

Rank	Job Title	Average Annual Salary ($75K minimum)	Jobs per 1M Total Jobs	% Change in Job Share, 2021 vs. 2024	% Containing Remote & Hybrid Phrases
1	Mental Health Technician	$77,448	1,425	1%	18%
2	Loan Officer	$192,339	1,044	3%	75%
3	Mental Health Therapist	$76,140	865	132%	41%
4	Electrical Engineer	$102,590	700	34%	19%
5	Construction Project Manager	$103,431	662	37%	10%
6	Mechanical Engineer	$96,091	552	24%	16%
7	Psychiatrist	$258,440	552	36%	15%
8	Human Resources Manager	$79,174	549	5%	13%
9	Senior Accountant	$82,811	547	18%	24%
10	Data Engineer	$130,135	532	2%	41%
11	Civil Engineer	$93,967	524	44%	22%
12	Supply Chain Specialist	$86,976	509	32%	13%

Source: Indeed's Best Jobs of 2024

About Data Scientist

- Data Scientist ranked in the top 8 in 100 best jobs.
  - U.S.News
- Data Engineer ranked in the top 10 in Indeed's best job of 2024.
  - Indeed's 2024 Workforce Insight Report
- Data Scientist's average salary in the US is AUD156,717 per year.
  - Datablokes.



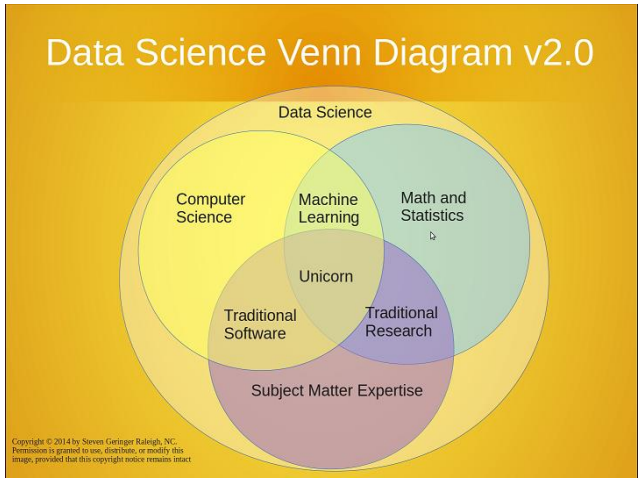
Key Questions

- What is Data Science?
- Who are Data Scientists? What do they do?
- What is Big Data?
- What drives Big Data?
- Moving from the regular data to the big data realm?

https://creativecommons.org/licenses/by-sa/3.0/

	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	◐	◐	●
Data Intuition	◐	●	◐	●
Statistics	◐	●	◐	●
Data Wrangling	○	○	●	●
Machine Learning	○	●	○	●
Software Engineering	○	◐	●	◐
Multivariable Calculus and Linear Algebra	○	●	○	◐

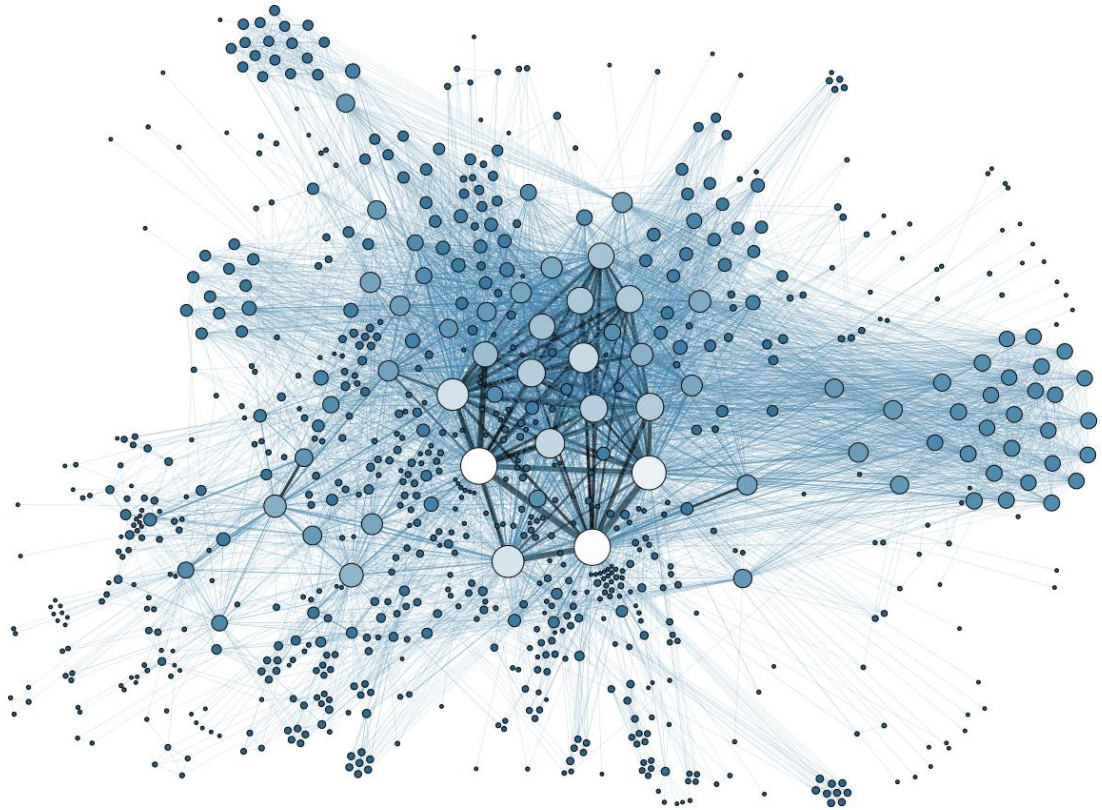
○ Not that important ◐ Somewhat important ● Very important



Data Scientist

- **Capable of analyzing and interpreting complex digital data to assist a business in its decision-making.**

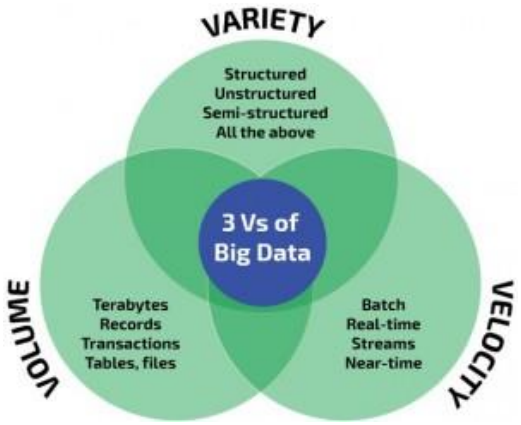
https://blog.udacity.com/2014/11/data-science-job-skills.html

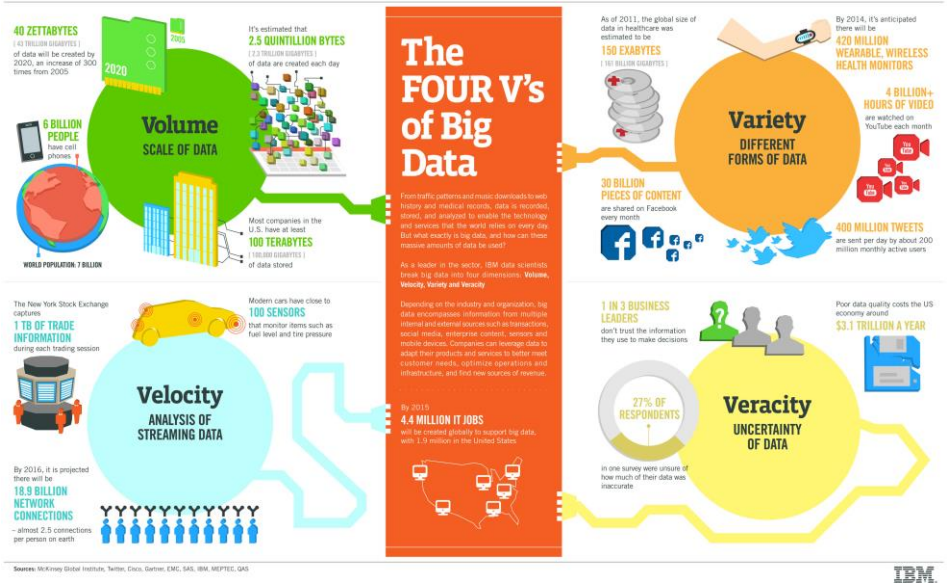


Data Scientist

- Essential sets of skills and behavioural characteristics
  1. **Quantitative skill**, e.g. mathematics, statistics.
  2. **Technical aptitude**, e.g. software engineering, machine learning, and programming skills.
  3. **Skeptial mind-set and critical thinking**, i.e. capable of examining their work critically rather than in a one-sided way.
  4. **Curious and creative**, i.e. passionate about data and finding creative ways to solve problems and portray information.
  5. **Communicate and collaborative**, e.g. able to articulate business values, a good team player.

Quantities of bytes						
Common prefix				Binary prefix		
Name	Symbol	Decimal SI	Binary JEDEC	Name	Symbol	Binary IEC
kilobyte	KB/kB	10^3	2^{10}	kibibyte	KiB	2^{10}
megabyte	MB	10^6	2^{20}	mebibyte	MiB	2^{20}
gigabyte	GB	10^9	2^{30}	gibibyte	GiB	2^{30}
terabyte	TB	10^{12}	2^{40}	tebibyte	TiB	2^{40}
petabyte	PB	10^{15}	2^{50}	pebibyte	PiB	2^{50}
exabyte	EB	10^{18}	2^{60}	exbibyte	EiB	2^{60}
zettabyte	ZB	10^{21}	2^{70}	zebibyte	ZiB	2^{70}
yottabyte	YB	10^{24}	2^{80}	yobibyte	YiB	2^{80}





Big Data

- Exceeds the processing capacity of conventional database systems
  - Too big (TB/PB level)
  - Moves too fast
  - New structure
- Four Vs of Big Data:
  - Volume
  - Velocity
  - Variety
  - *(Veracity)*

https://sites.google.com/site/iebitwiki/software/binary, https://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

Moving from the regular data to the big data realm?

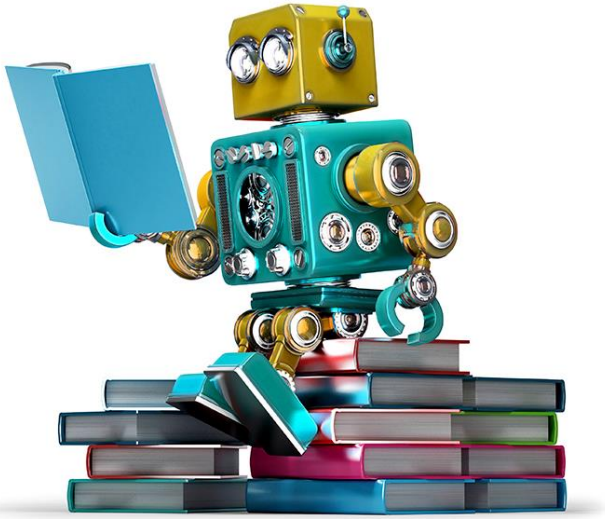
- How much data an average person generates per day?
  - Mobile phone geo-location data.
  - Social media data.
  - Communications.
  - IoT devices.
  - …
- With proper data processing techniques, new knowledge or information can be extracted form the data generated by people in their daily life.

ABOUT THIS UNIT (COS10022)

Learning Outcomes

- After successfully completing this unit, you should be able to:

1. Appreciate the roles of data science and Big Data analytics in organisational contexts.
2. Compare and analyse he key concepts, techniques and tools for discovering, analysing, visualising and presenting data.
3. Describe the processes within the Data Analytics Lifecycle.
4. Analyse organisational problems and formulate them into data science tasks.
5. Evaluate suitable techniques and tools for specific data science tasks.
6. Develop and execute an analytics plan for a given case study.



Lab Environment

Hardware

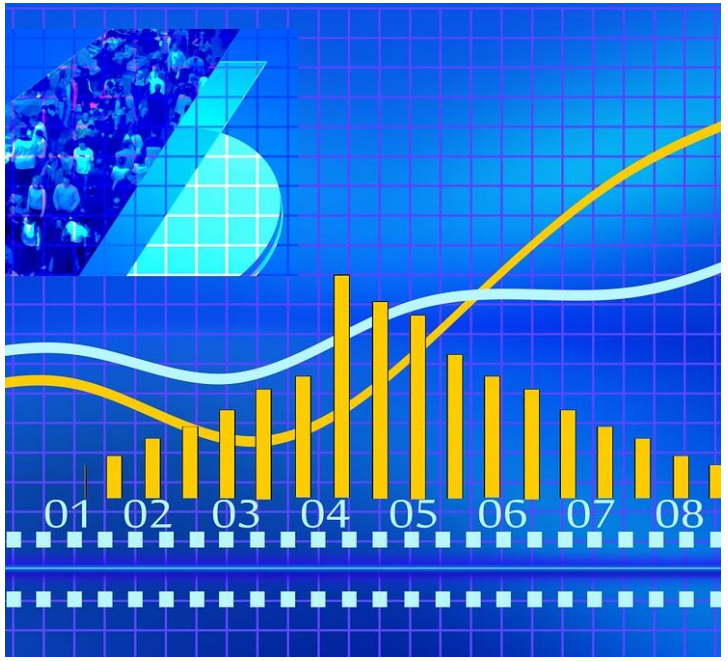
- Desktops
- You can also bring your own laptop

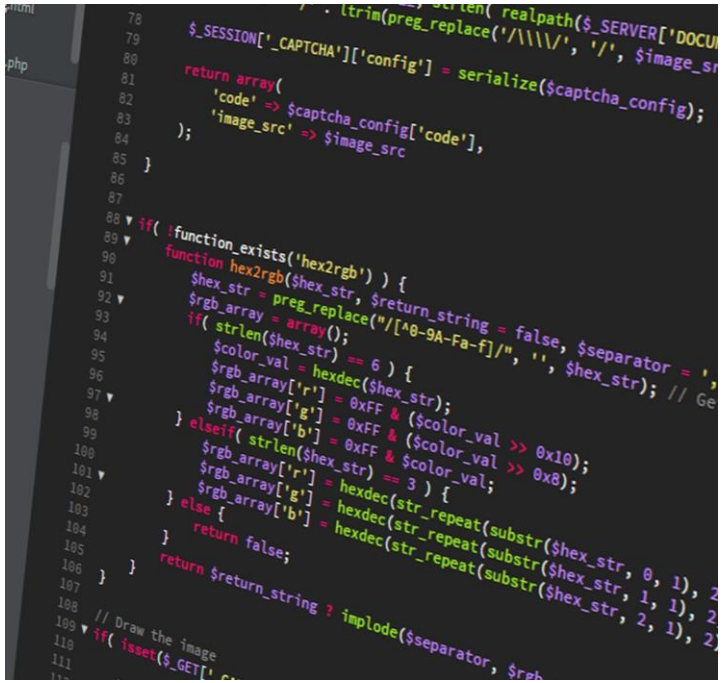
Software

- Google Colab
- KNIME (analytics libraries and tools)

Classroom Etiquette

- Usage of Personal Electronic Devices
  - Cell phones/PDAs (Set to Vibration mode or mute)
  - If your phone rings, answer it as you step out of the classroom.
- Bring your own rubbish with you when leaving the room.
- Inform the instructor and the lab partners (if any) of all absences from classroom sessions.
  - Excessive absences will be interpreted as non-attendance at the class.
- Although we encourage collaboration during the class, please treat the data files, code and lab as intellectual property of this unit and do not redistribute without the consent of the unit convenor.





Expected Background

- Strong mathematical and quantitative capability.
- Basic programming skills.

Assessment and Task Details

Assessment Task	Individual or Group Task	Weighting	Submission Due Date
Assignment 1	Individual	20%	End of Week 4
Assignment 2	Individual	30%	End of Week 10
Online Test 1	Individual	10%	Week 5 Lab
Online Test 2	Individual	10%	Week 9 Lab
Online Test 3	Individual	30%	Week 12 Lab

Requirements

Participation Requirement

- You are expected to attend all lectures and labs.
- Use the specified tool in the unit in all assignments and tests.

Minimum requirements to pass this unit

- Achieve an overall mark for the unit of 50% or more.
- Submit all assignments and participating in all tests.

Assignment Requirement

- Submit the assignment on time.
- Deductions will be applied for late submissions. 10% of the mark for that particular assignment will be deducted per day after the deadline.
- Maximum delay is 5 days.

https://creativecommons.org/licenses/by-sa/3.0/

Special Consideration?

If you encounter some difficulties such as medical issues, which affects your progress, during the study, you'll need to launch a **special consideration**.

The special consideration will only extend the deadline in a reasonable range for submitting the assignments. It will not be used to twist the evaluation criteria.

Special Consideration?

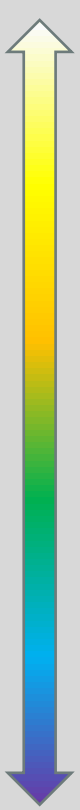
A special exam is offered to students who are granted special considerations, unless it is possible to accommodate the special considerations during the semester.

Before the end of the semester: deadline extended.
After the end of the semester: a special exam is required.

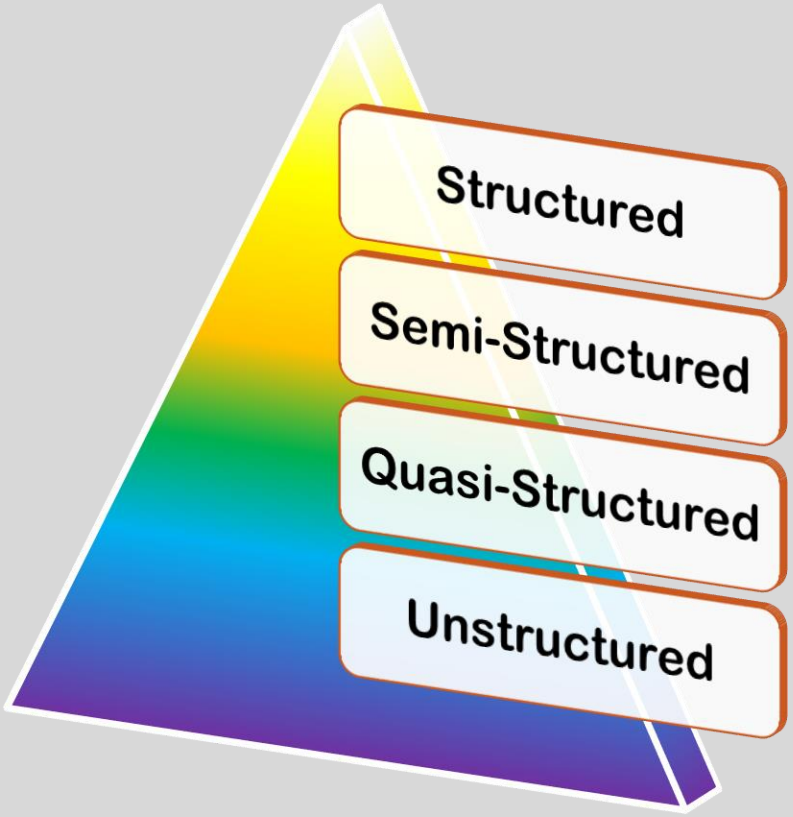
OVERVIEW OF DATA SCIENCE

Data Structures

Less Data



More Data



Relational DBMS

Extensible Markup Language (XML), JSON, CSV, etc.

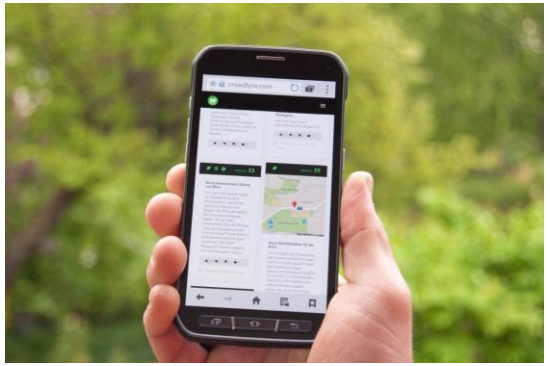
Web click stream

Text documents, PDFs, images, video, etc.

Data Devices

- **Gather** data from multiple locations.
- Continuously **generate** new data ***about*** this data.
- For each Gigabyte created for this data, an additional Petabyte of data is **created** about that data.

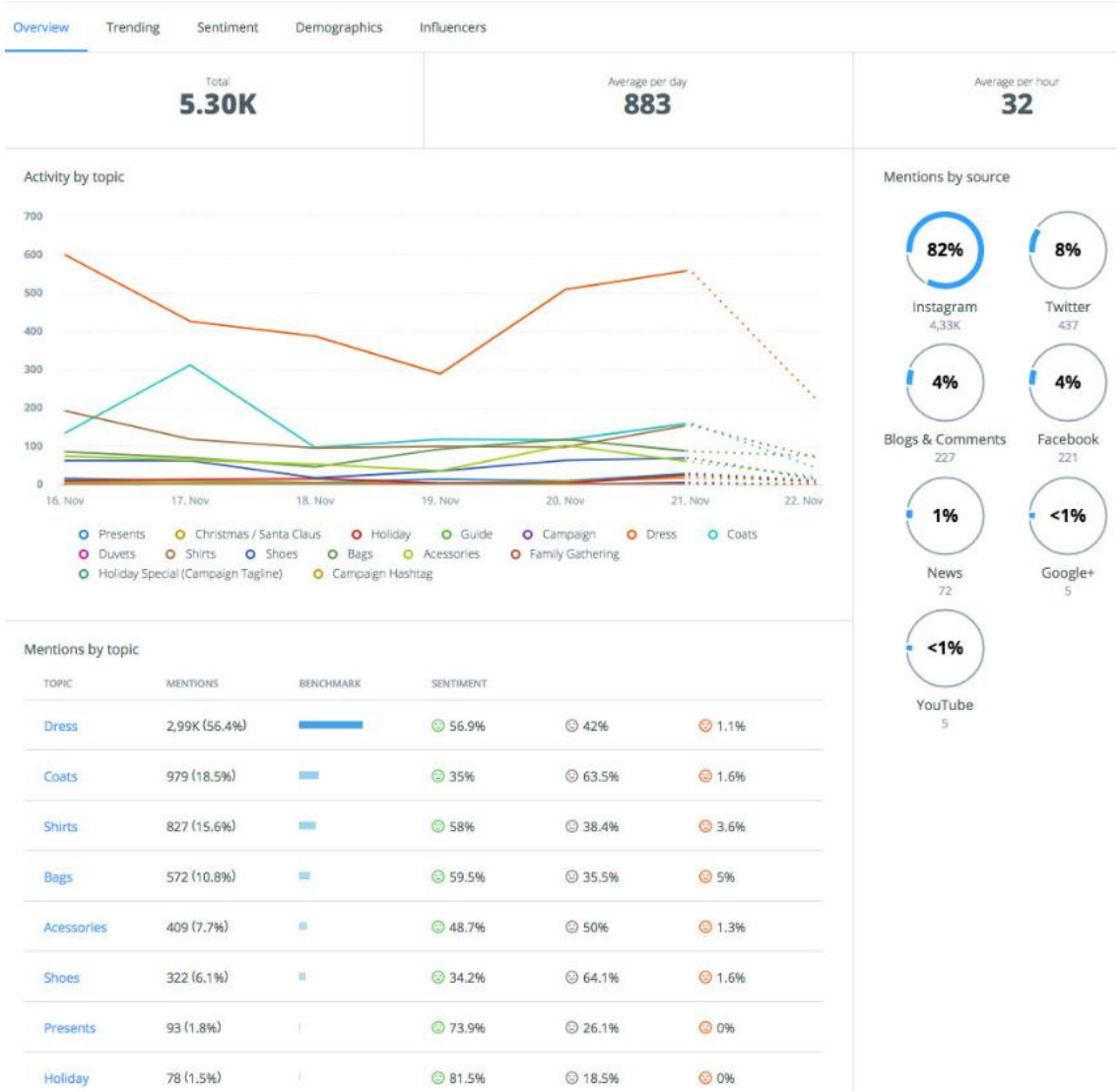
- **Consider**
  Data generated from someone playing an online video game through a PC, game console (PlayStation, Xbox, Nintendo Wii), or smartphone.



Data Collectors

- Entities that **collect** data from the device and users.

- **Consider**
- Internet Streaming provider tracks:
  - The shows a client watches
  - Which programs/channels someone will and will not pay for to watch on demand
  - The prices someone is willing to pay for premium content



Data Aggregators

- Entities that **compile** and **make sense** of the data collected by data collectors.
- **Transform** and **package** the data as products to sell.
- Example
  - Falcon

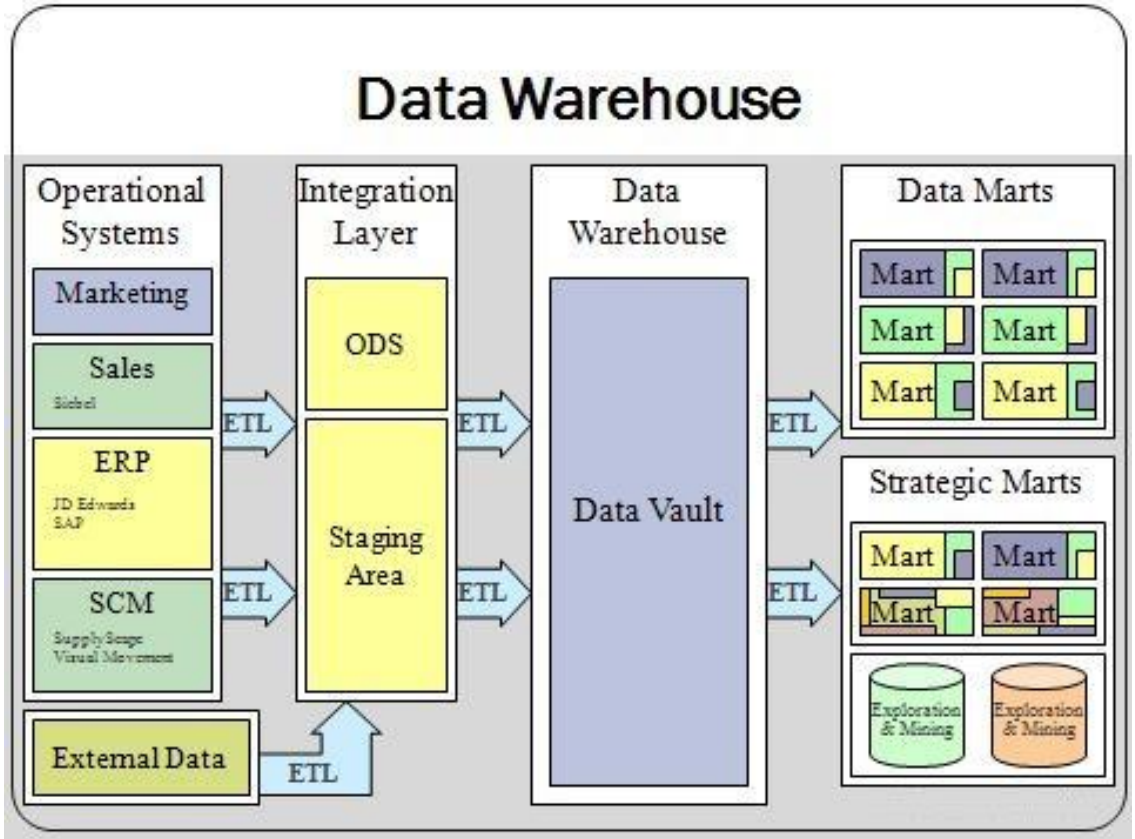
https://www.falcon.io/products/social-media-monitoring/

Data Users and Buyers

- Direct **benefactors** of the data collected and aggregated by others within the **data value chain**.

- <u>Examples</u>
  - Corporate customers
  - Analytic services
  - Media archives
  - Advertising
  - Information brokers
  - Credit bureaus
  - Catalogue co-ops

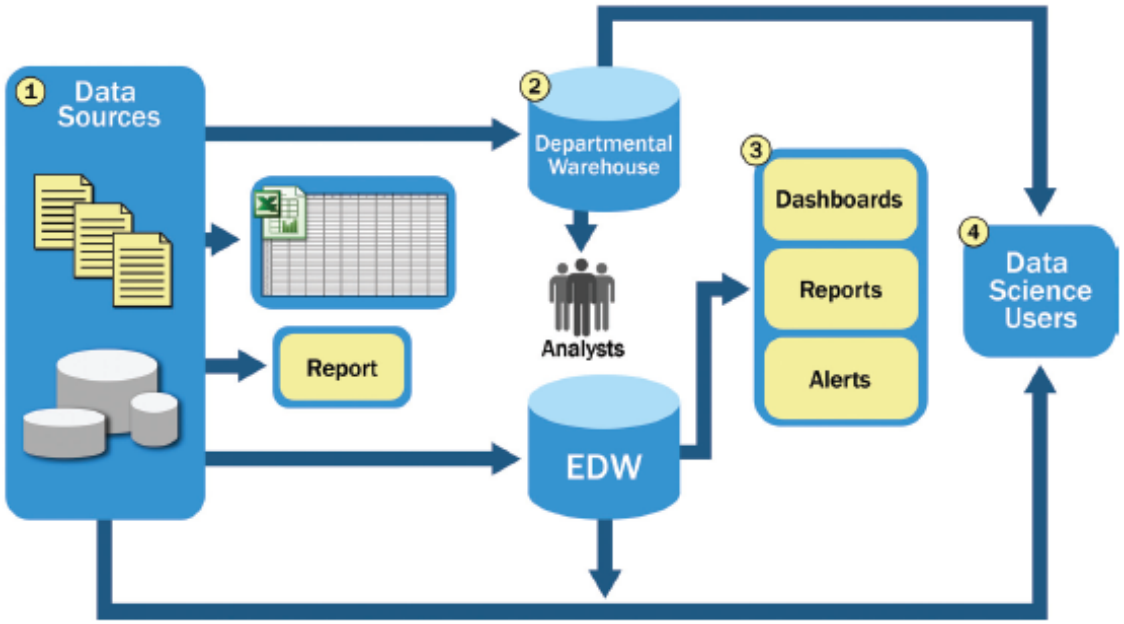
https://www.falcon.io/our-customers/



Data Science vs Enterprise Data Warehouse

- Data Warehouse (DW) is a relational database that is designed for query and analysis rather than for transaction processing. Contains selective, cleaned, and transformed historical data.
- Extraction, Transformation, and Loading (ETL)
- On-Line Analytical Processing (OLAP)
- Supports enterprise decision making

https://upload.wikimedia.org/wikipedia/commons/4/46/Data_warehouse_overview.JPG, https://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm



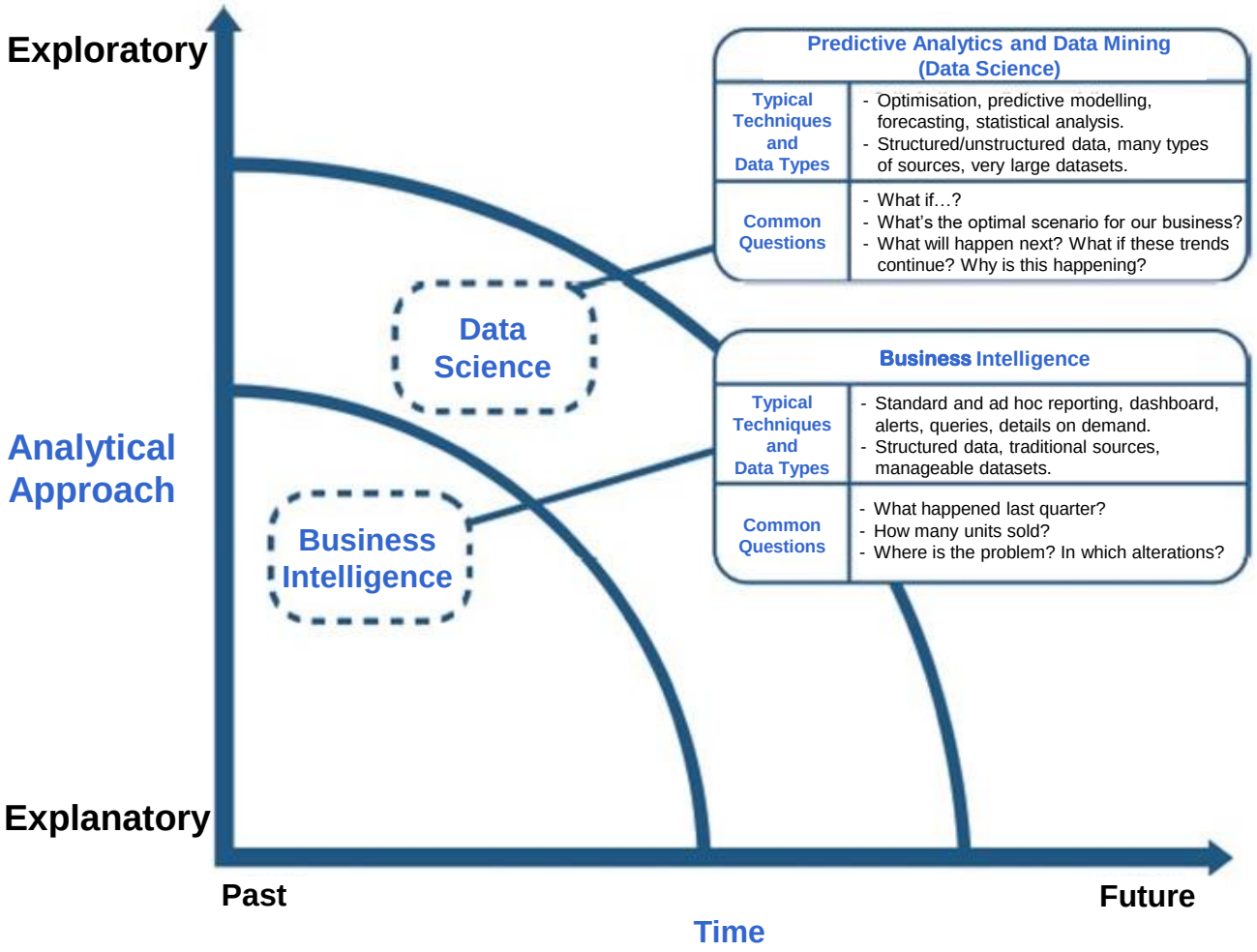
Data Science vs Enterprise Data Warehouse

- Limitations of the Enterprise DW analytics:

1. High-value data is hard to reach and leverage.
   - Low priority for data science projects.
2. Data usually moves in batches from DW to local analytics tools (e.g. R, SAS, Excel).
   - In-memory analytics; dataset size constraints.
3. Data Science projects remain isolated and ad-hoc, rather than centrally managed.
   - Data science initiatives not aligned with corporate strategic business goals.

Source: Dietrich, D. ed., 2015, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.

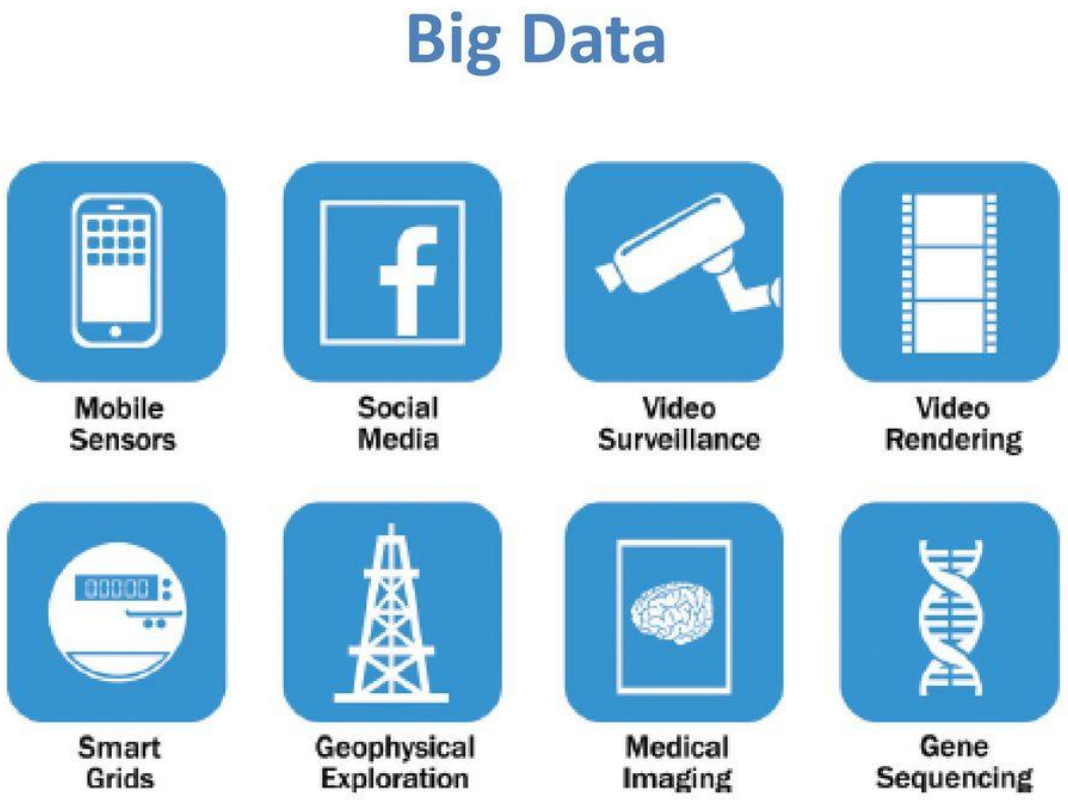
Data Science and Business Intelligence



DATA SCIENCE VS BUSINESS INTELLIGENCE

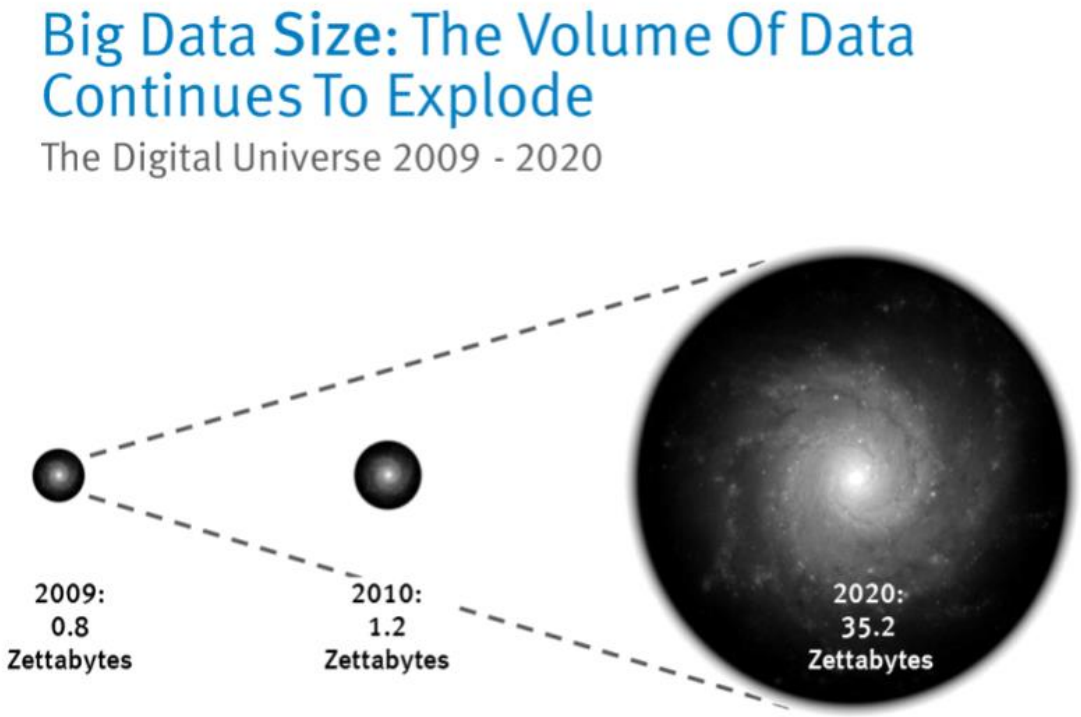
Source: Dietrich, D. ed., 2015, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.

BIG DATA ECOSYSTEM



DRIVERS OF BIG DATA

Source: Dietrich, D. ed., 2015, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.

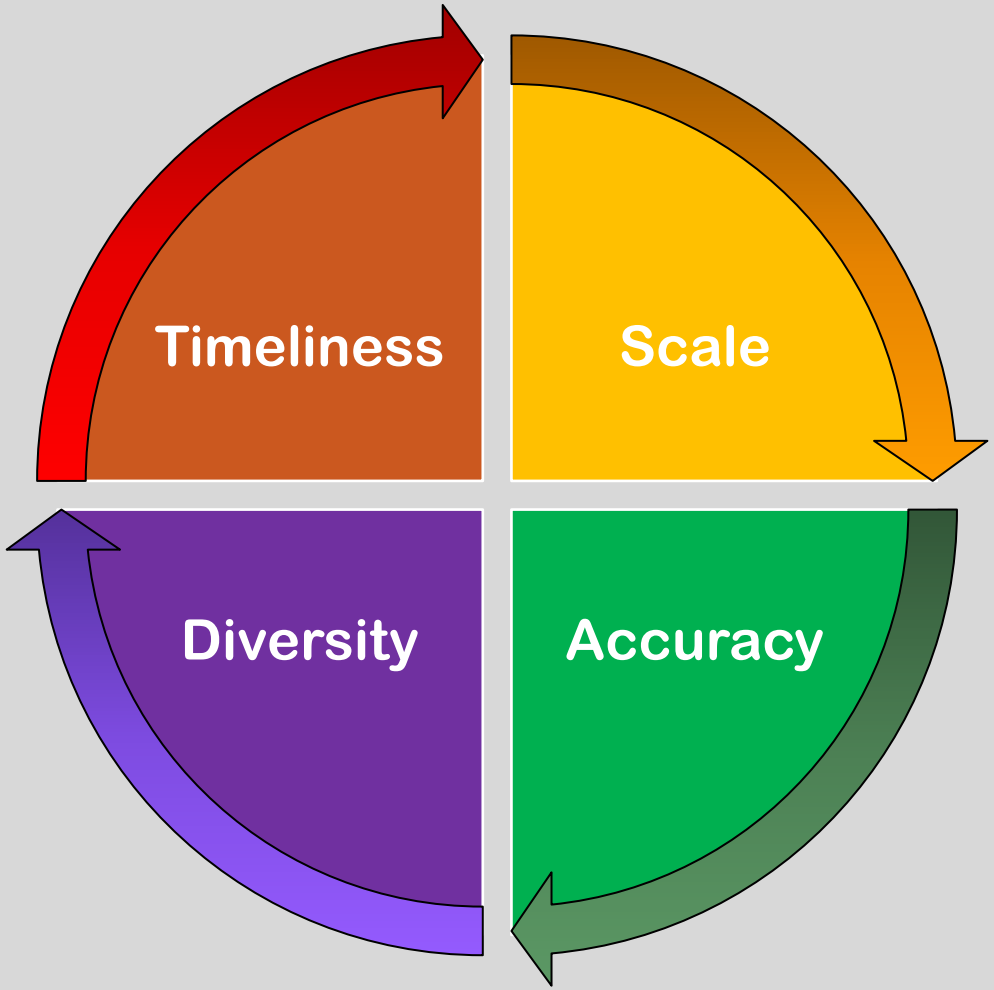


Key Characteristics of Big Data

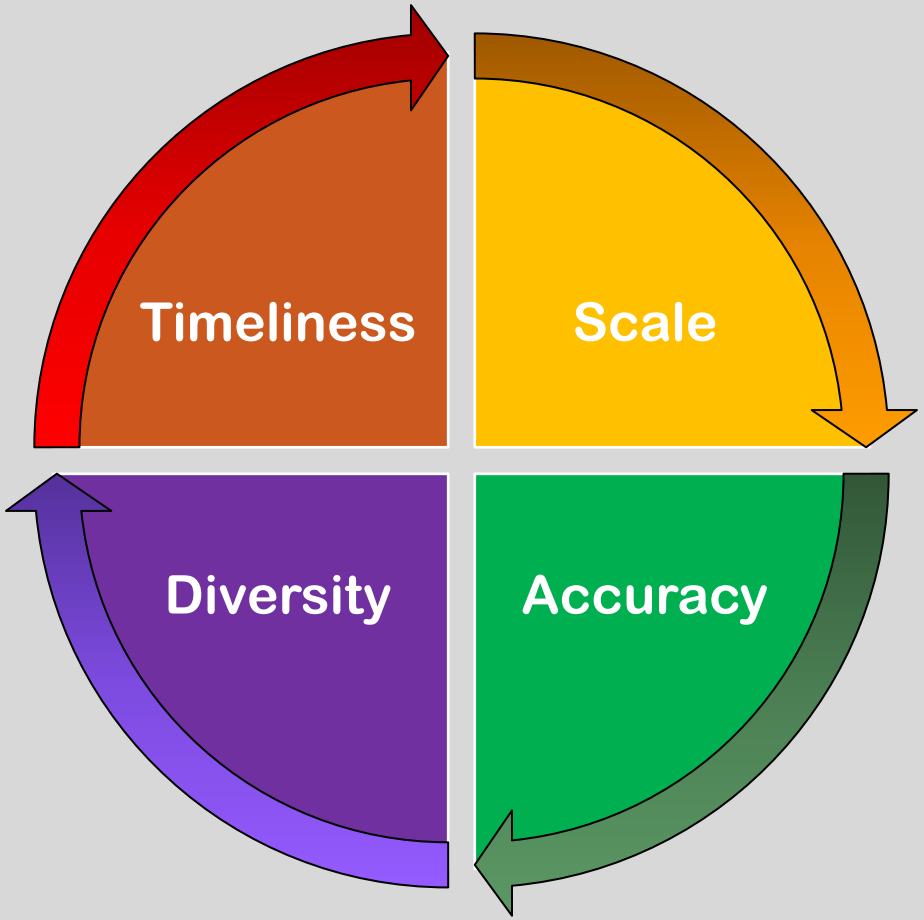
- **Data Volume**
  - **44x increase from 2010 to 2020. (1.2 Zetta Bytes to 35.2 ZB)**
- **Processing Complexity**
  - **Changing data structures.**
  - **Use cases warranting additional transformations and analytical techniques.**
- **Data Structure**
  - **Greater variety of data structures to mine and analyse.**

Big Data: What is

- Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value. – by M. James, C. Michael, B. Brad, B. Jacques, D. Richard, R. Charles, and H. Angela, 2011. Big data: the next frontier for innovation, competition, and productivity. *The McKinsey Global Institute*.



Big Data 4 Vs Revisit



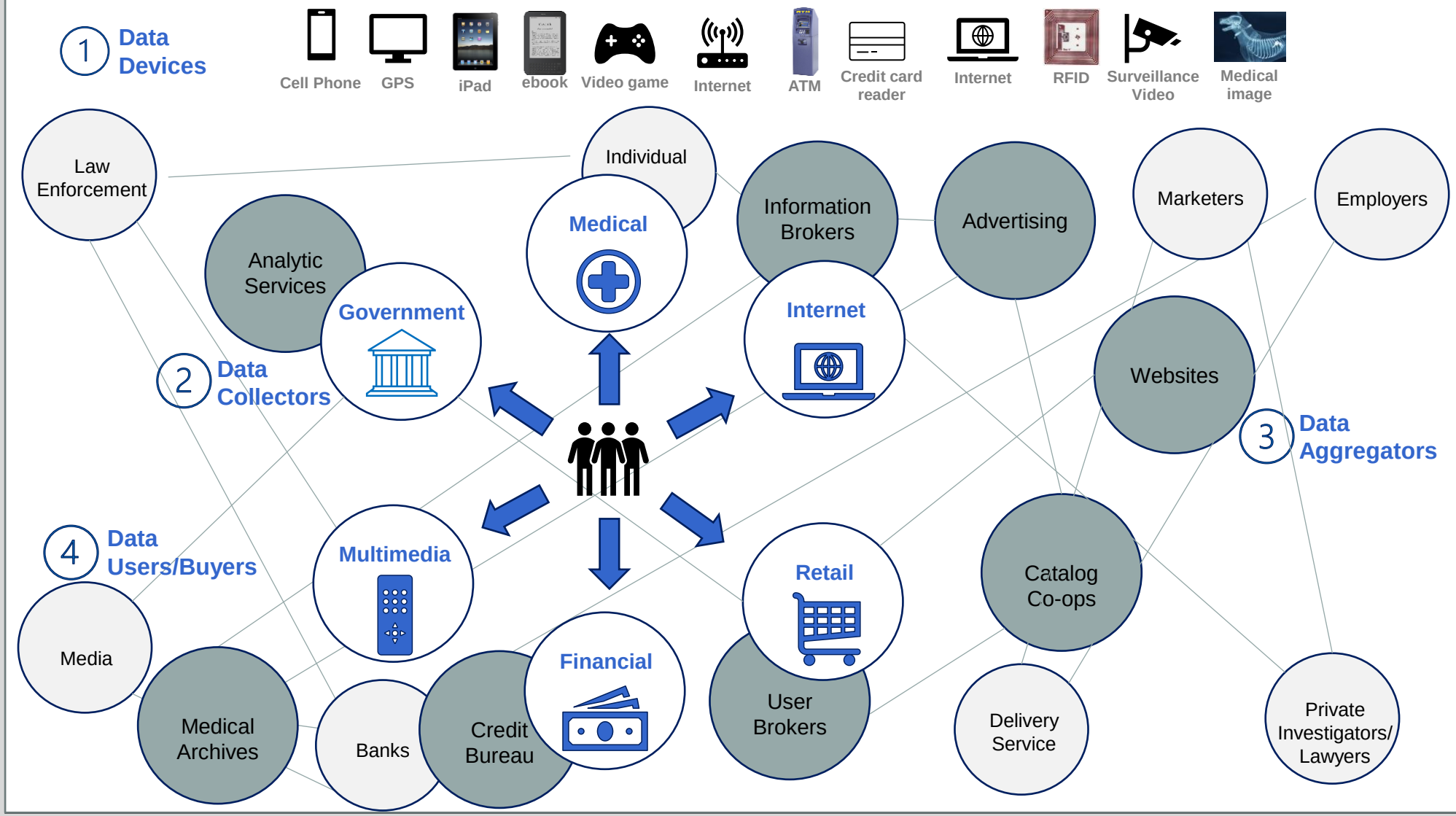
Volume
- Scale

Variety
- Diversity
- Distribution

Velocity
- Timeliness

Veracity
- i.e. pertaining to the accuracy of data.

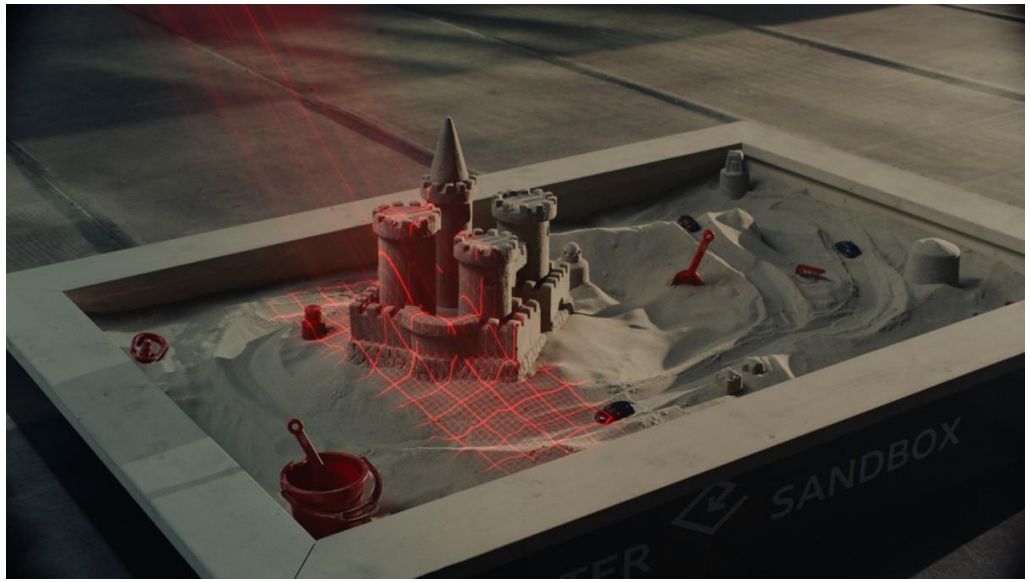
Emerging Big Data Ecosystem



Source: Dietrich, D. ed., 2015, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.

Big Data vs Enterprise DW/Business Intelligence

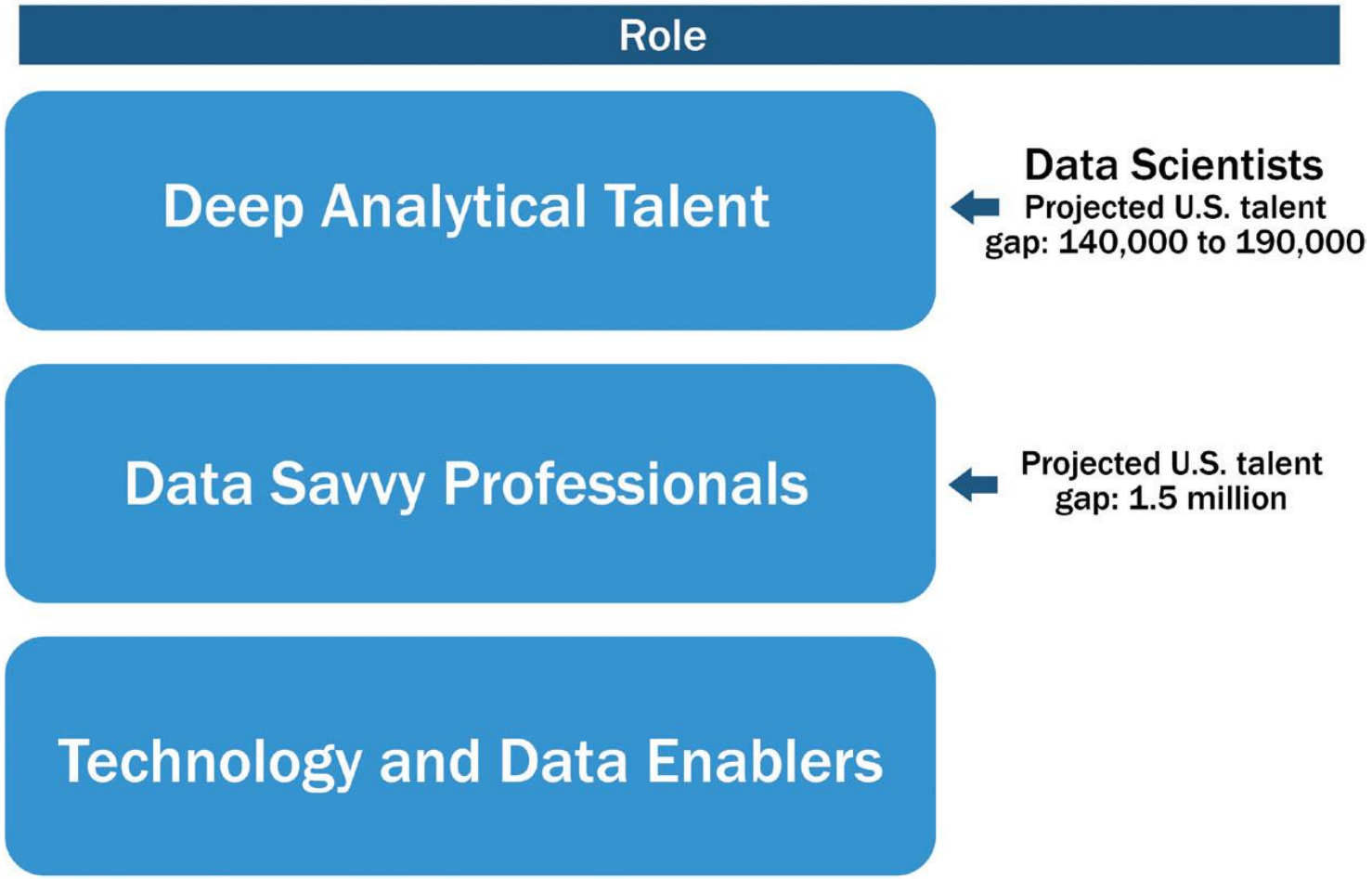
- The four Vs of Big Data will not work well with the traditional Enterprise Data Warehouse.
  - Centralised, purpose-built space. (lack of agility)
  - Supports Business Intelligence and reporting. (restrict robust analyses)
  - Analysts must depend on IT group and DBAs for data access. (lack of control)
  - Analyst must spend significant time to aggregate and dis-aggregate data from multiple sources. (reduces timeliness)
- To succeed, Big Data analytics require different approaches.



Analytic Sandbox (*Workspaces*)

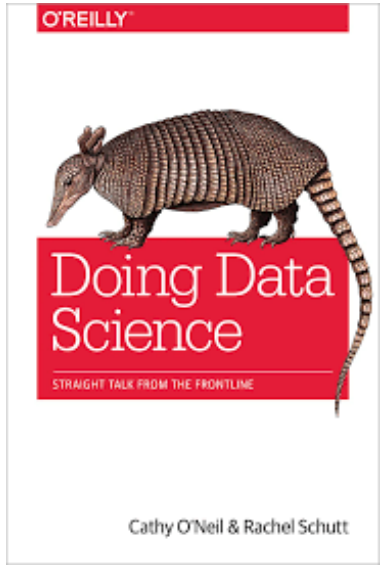
- Resolve the conflict between the needs of analysts and the traditional EDW or other formally managed corporate data.
- Data assets gathered from multiple sources and technologies for analysis.
- Enables flexible, high performance analysis in nonproduction environment.
- Reduces costs and risks associated with data replication into “shadow” file systems.
- “Analyst owned” rather than “DBA owned.”

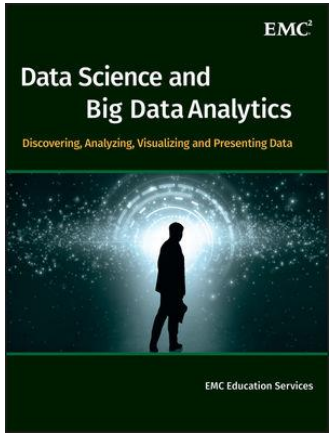
Three Key Roles of The New Data Ecosystem

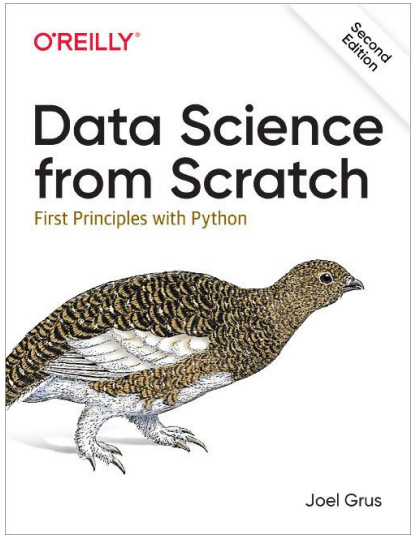


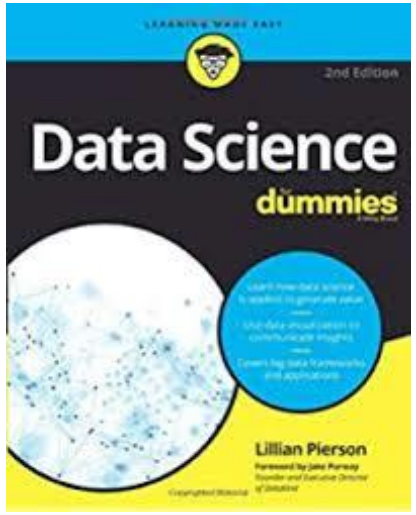
Note: Figures above reflect a projected talent gap in US in 2018, as shown in McKinsey May 2011 article "Big Data: The Next Frontier for Innovation, Competition, and Productivity"

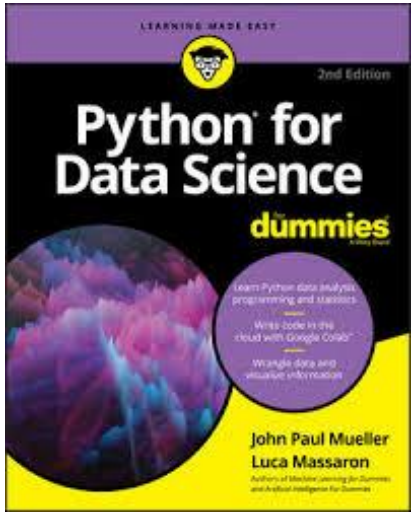
Bid Data: Key Roles











Texts and Resources

- Unless stated otherwise, the materials presented in this lecture are taken from:
  - Dietrich, D. ed., 2015. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. EMC Education Services.
  - Schutt, R. and O'Neil, C., 2013. *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc.
  - Pierson, L., *Data Science for Dummies, 2nd Edition*, John Wiley & Sons © 2017
  - Mueller, J. P. and Massaron, L., *Python for Data Science for Dummies, 2nd Edition*, John Wiley & Sons © 2019 (432 pages), ISBN:9781119547624
  - Joel Grus, 2019. *Data Science from Scratch – First Principles with Python, 2nd Edition*, O'Reilly Media, Inc.

Unit Summary

- Unit structure
- Requirements, Expectations, and Rules
- About Data Science
- Big Data Ecosystem