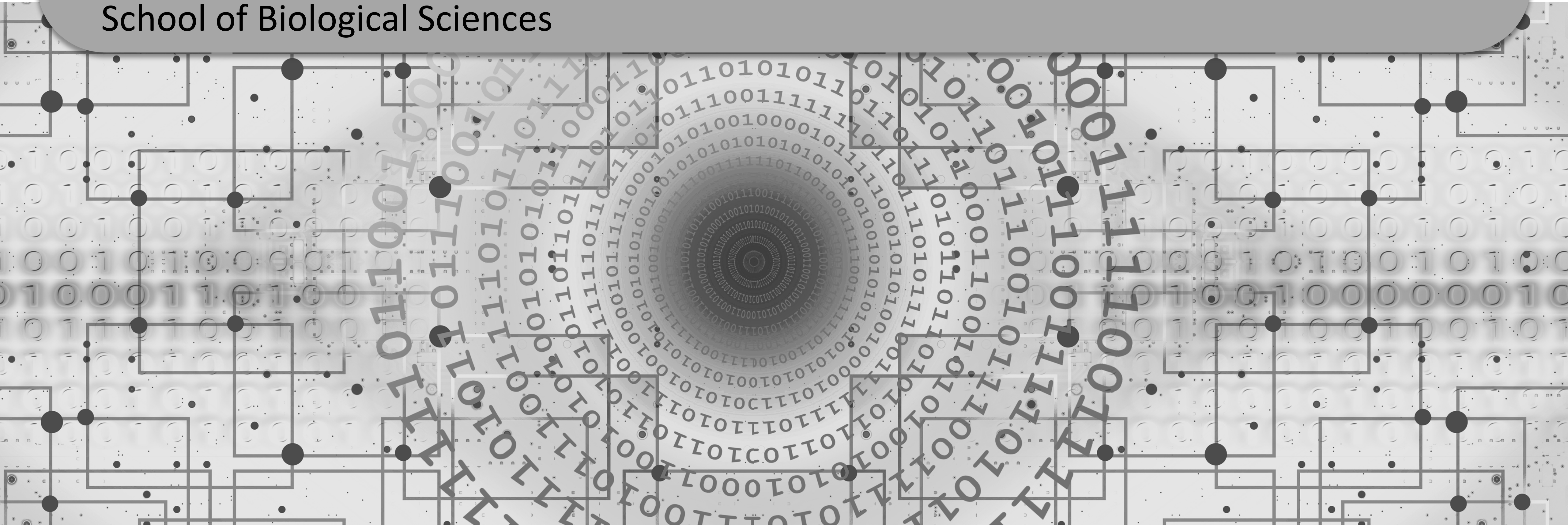
Data Science in the Singapore Landscape

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



Learning Objectives

By the end of this lecture, you should be able to:

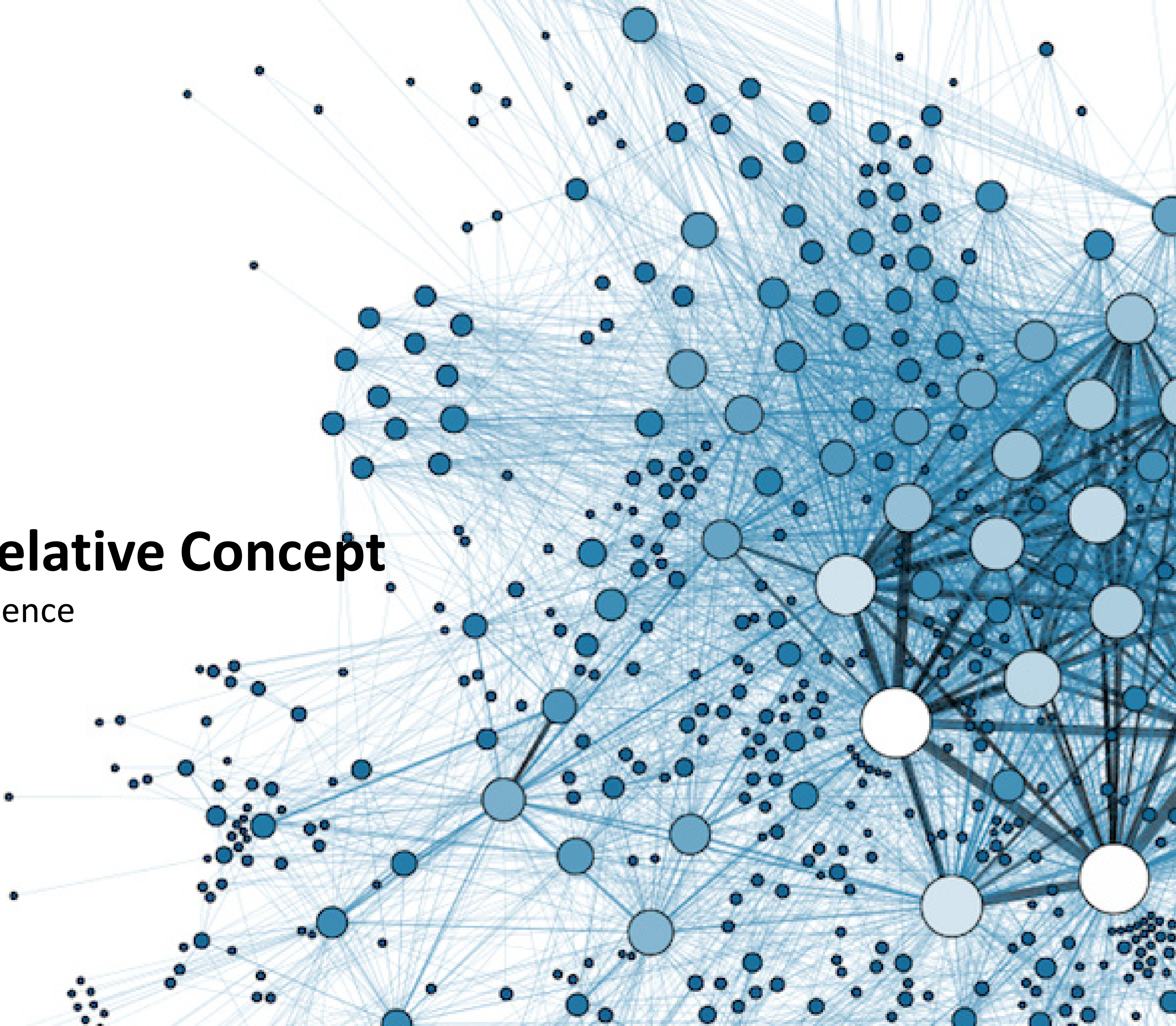
- Explain why big data is a relative concept.
- Evaluate the value of data in terms of the 4 Vs.
- Explain the importance of big data towards data science.
- Describe the data science and analytics landscape in Singapore.
- Describe how Singapore’s Smart Nation Initiatives may impact you.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Why Big Data is a Relative Concept

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

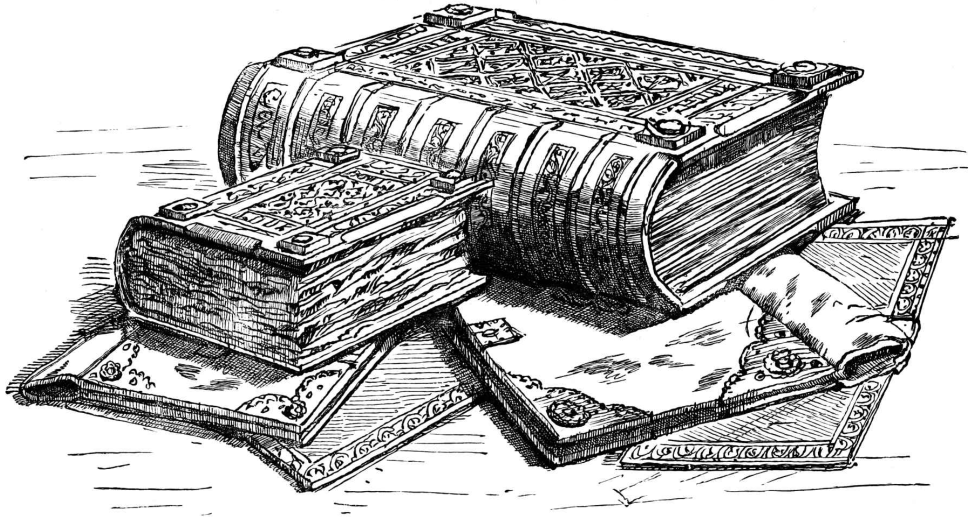


Why Big Data is a Relative Concept

You have heard a lot about the term “big data” often used in conjunction with “data science”. However,

- Big data **is not equal** to data science.
- Big data **implies** data volumes that now exist…
    - but for which we **lack** the **proper infrastructure** to store and retrieve.
    - and lack the **proper techniques** to analyse.
- In today’s world where modern day computers can comfortably handle gigabytes and terabytes of information, we naturally think of big data as within the Petabyte or Exabyte range (which goes well beyond the storage of most of our personal drives).
- But did similar data problems also exist in mankind’s pre-modern computer era?

Examples of ”Big Data” from History

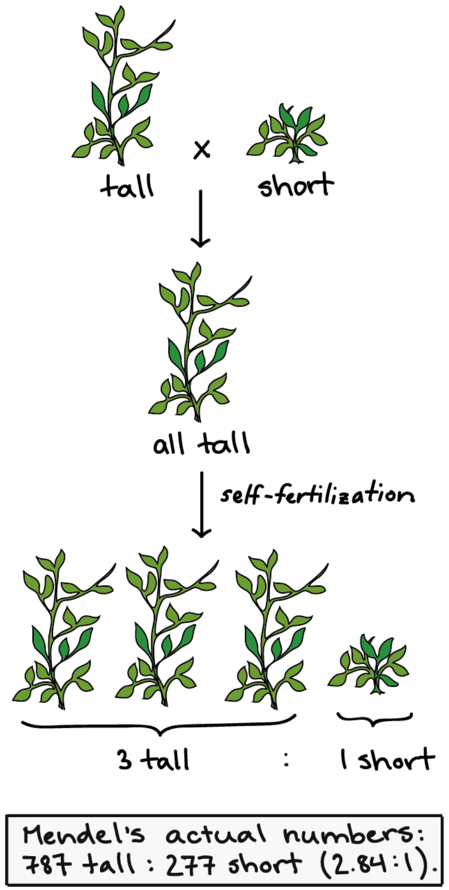


Source: https://commons.wikimedia.org/wiki/File:Domesday-book-1804x972.jpg#/media/File:Domesday-book-1804x972.jpg

Domesday Book

- **1085 A.D.: William the Conqueror** surveyed his empire
- Big data characteristics:
  - Searchable: Custodians of the book could find out data regarding to certain questions, e.g. a list of all men who have more than 100 but less than 200 sheep who did not pay their required tribute that year.
  - Included:
    1. Survey, which collected information and facts (data).
    2. The Winchester Roll: created out of the facts (frameworks: analytics and visualisations).
    3. Day of Judgment: Determination of all taxes, ownership and military obligations for hundred of years to come (extracted value, algorithms).

”Big Data” in History



Source: https://www.khanacademy.org/science/high-school-biology/hs-classical-genetics/hs-introduction-to-heredity/a/mendel-and-his-peas

- In the 1800s, Gregor Mendel studied 7 traits in peas to determine how genetic traits are passed from 1 generation to the next.
- Because genetic traits are segregated and assorted randomly, large numbers of peas would need to be collected and studied in order for the true segregation ratios to appear.
- In other words, he needed to collect, store, and systematically retrieve information on thousands of peas in order to perform his study.
- This is an example of pre-modern computer “big data” analysis.

“Big Data” in History

So will our current notions of big data remain the same ten years down the road?

Unlikely.

As we develop new technology for data storage, processing and retrieval, our ability to handle larger amounts of information will also increase.

The notion of “big data” is not simply about just its sheer size, but rather, our current limited ability to deal with it.

✓ In other words, “today’s big data is tomorrow’s small data”.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Evaluating the Value of Data in Terms of the 4 Vs

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Data is not all Born Equal

Big data is not always good data.

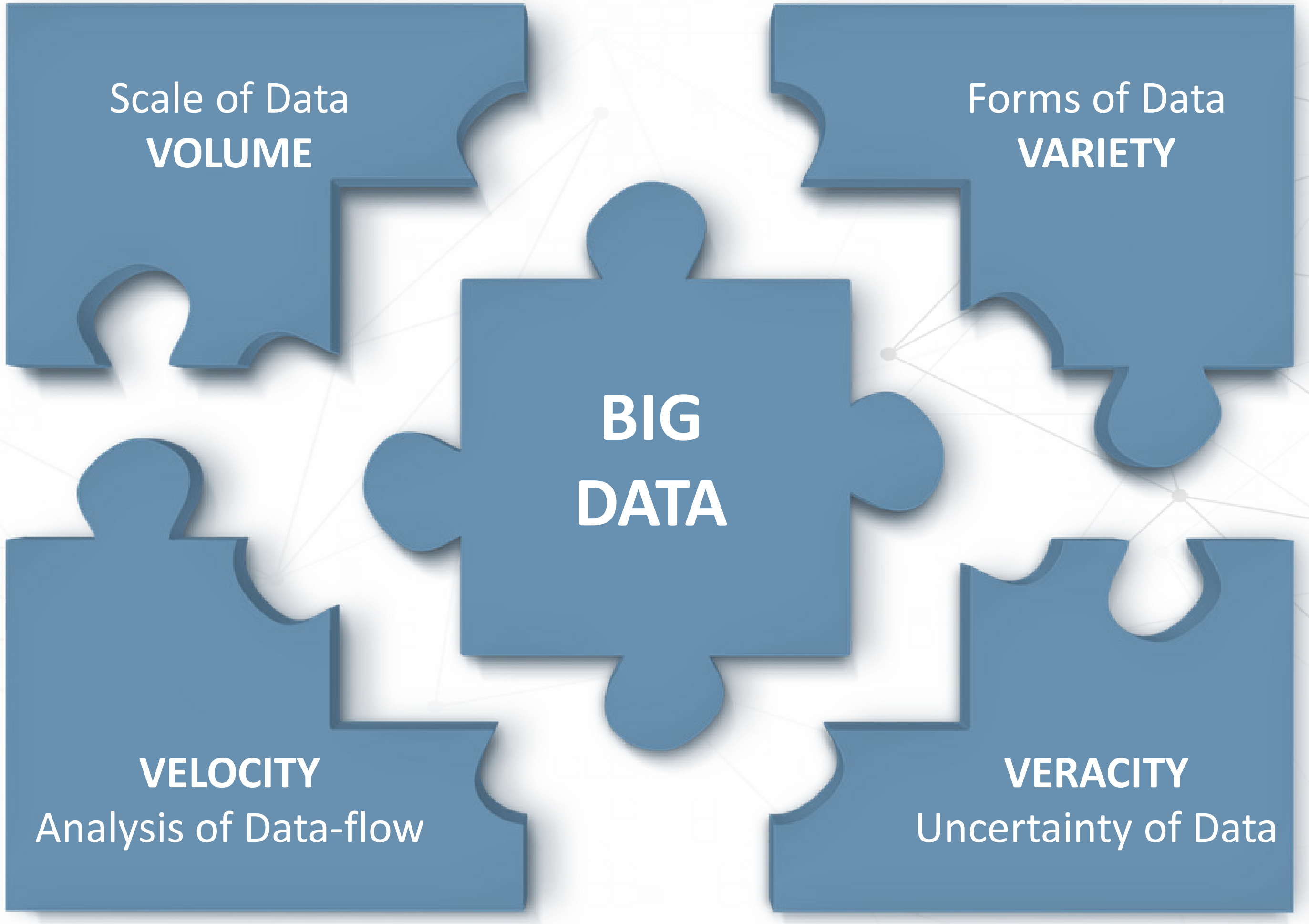
Should not mistake quantity for quality.

We may determine the worth of data using 4 metrics --- volume, velocity, variety and veracity.

Big Data and Analytics

Characteristics and properties of Big Data: 4Vs

Size/ quantity of collected and stored data, measured in tera-/ petabytes.

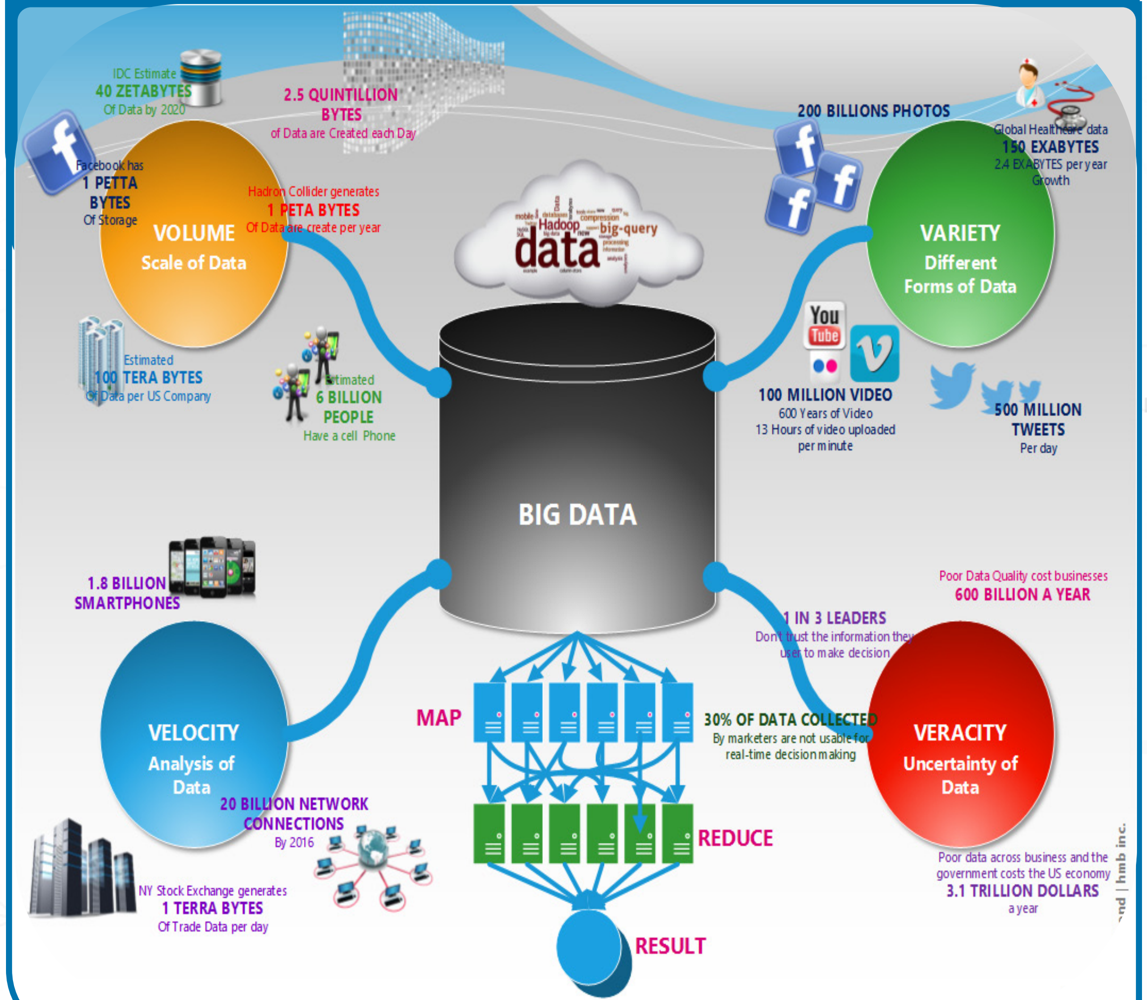


Type of data [images, text, video, audio, geo-spatial, etc.].

Speed/ rate of data transfer between source and destination/ data availability for analysis.

Data quality/ accuracy, uncertainties impact confidence data.

Big Data and Analytics



Source: https://tomyrhymond.files.wordpress.com/2014/08/big-data.png

Characteristics and properties of Big Data or the 4 Vs are as follows:

1. **Volume**: Size/ quantity of collected and stored data, measured in tera-/ petabytes.
2. **Velocity**: Speed/ rate of data transfer between source and destination, available for analysis.
3. **Variety**: Type of data (images, text, video, audio, geo-spatial, etc.).
4. **Veracity**: Data quality/ accuracy: uncertainties impact confidence data.

Big Data and Analytics

How the data is rated along the 4 Vs will determine what you can do with it. For example, if you are performing a genomic analysis of 1000 cancer patients against 1000 normal patients:

Metric	How does it rate?	Why?
Volume	Relatively small	Although this is considered a big and expensive experiment in biological terms, it is extremely small, relative to modern day notions of big data. If each patient's data is only 1 GB. Then all the data can fit into 1000 x 1 x 2 = 2TB. A 12 TB solid state drive is cheap these days.
Velocity	Fast	Most of the data are stored locally. And so, velocity is essentially limited by the memory speeds of the computer, which is quite fast.
Variety	There is only 1 kind of data	A genomics experiment only looks at the expression of 20,000 genes per patient. This is only 1 dimension of information, and other information such as the epigenetics and protein expression are also useful but presents challenges for data integration and combined analysis.
Veracity	Poor	Technical and biological bias will inevitably exist. Technical bias includes batch effects and instrumentation artifacts. Biological bias may arise due to inherent differences amongst individuals that can lead to false positive findings.

Big Data and Analytics

5) Value: Importance and usefulness of data (e.g., for business decisions).

6) Validity: Data authenticity: correctness or accuracy of data used to extract result in the form of information.

7) Variability: Data differentiation efficiency: differentiation between noisy and important data.

8) Venue: Different platforms: various types of data derived from different sources via different platforms (e.g., personnel systems, private and public clouds).

9) Vocabulary: Data terminology: data models; semantics that describes data structure.

10) Vagueness: Confusion over meaning of data and used tools.

11) Volatility: Duration of usefulness of stored data.

12) Visualisation: Data act/ Data process: representing abstract data.

13) Virality: Spreading speed: rate at which data is broadcast by a user and received by different users.

14) Viscosity: Lag of event: time difference between event occurrence and description of the event.

15) Complexity: Correlation of Data derived from different sources and at different times to synergise changes quickly.

No need to memorise these. Only good to know. But must know the first 4 Vs.

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

The Data Science and Analytics Landscape in Singapore

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Data Analytics and Data Science

A viable distinction

Data science relates to the various theories on how to derive insight from data.

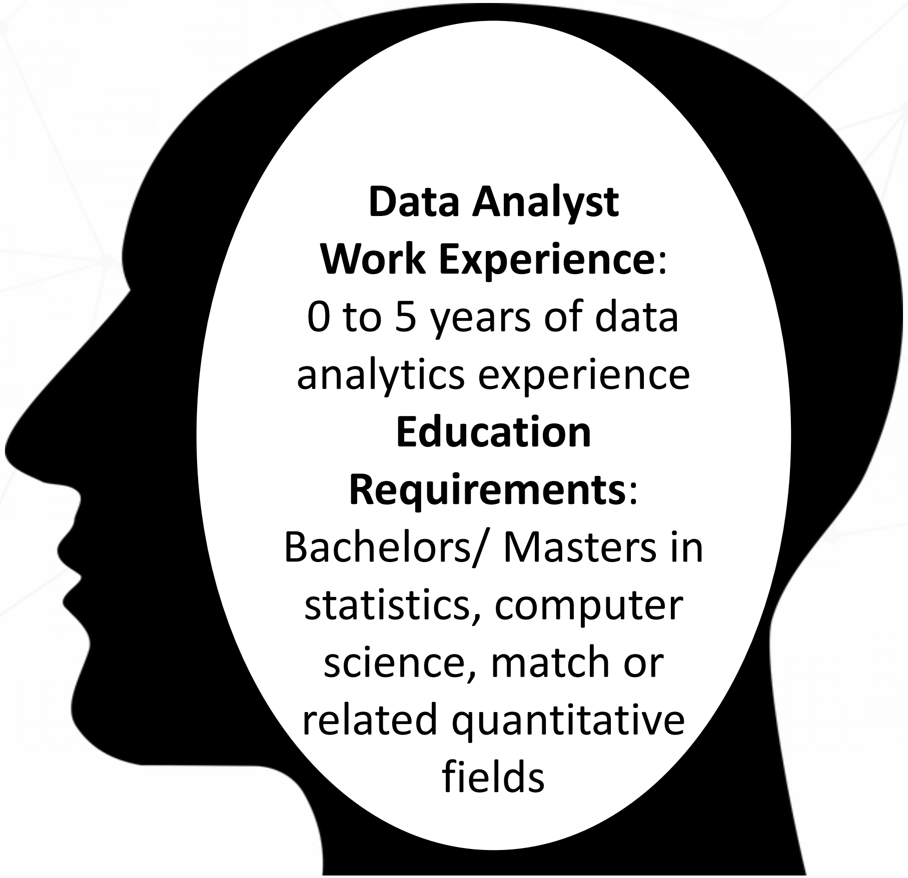
- It often has a “scientific/ academic” connotation.
- Normally considered higher level.

Data analytics is application of theory into action.

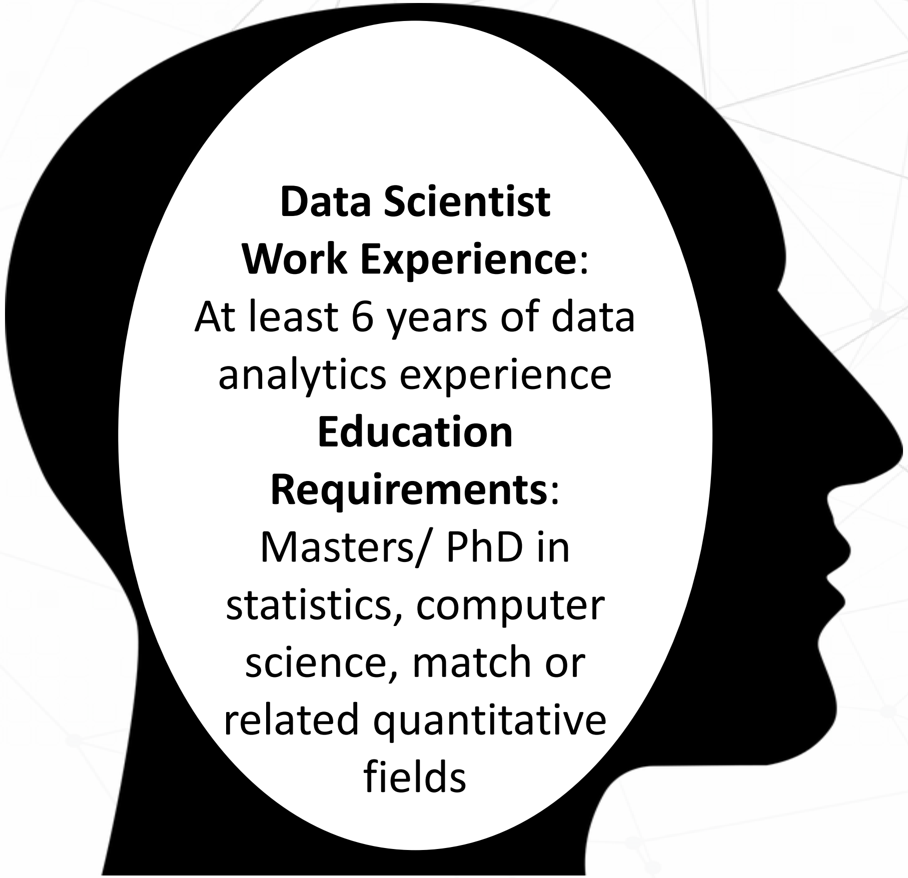
- It often has a “business” connotation.
- Normally considered lower level.

Data Analytics and Data Science

You can also discern higher and lower level according to the job scoping as defined by our own Economic Development Board.



Data Analyst
Work Experience:
0 to 5 years of data analytics experience
Education Requirements:
Bachelors/ Masters in statistics, computer science, match or related quantitative fields



Data Scientist
Work Experience:
At least 6 years of data analytics experience
Education Requirements:
Masters/ PhD in statistics, computer science, match or related quantitative fields

Data Analytics and Data Science

Business intelligence versus data analytics and data science.

In industry, you will also hear the term business intelligence being strewn around quite liberally.

And usually used in the same breadth as data analytics and data science.

Data Analytics and Data Science

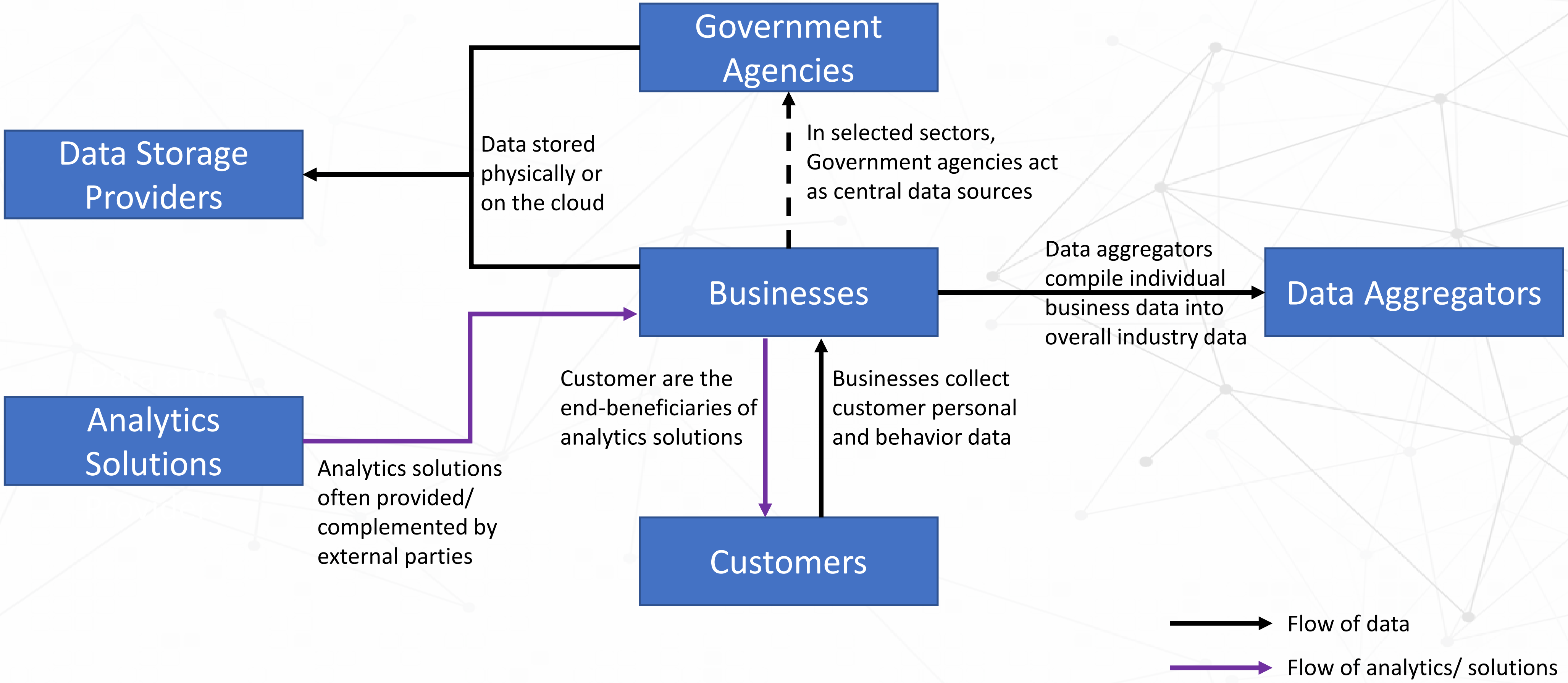
Business intelligence versus data analytics and data science.

- Business Intelligence (“BI”) focuses on historical trend analysis and patterns (e.g. what products are most commonly sold, what are customer preferences).
    - Superficial description and exploration of data.
- Data analytics and data science focuses on how to make predictions and change.
    - How to derive deeper insight into the data?
- “BI is needed to run the business while analytics is needed to change the business.” --- *Pat Roche, VP of Engineering, Magnitude Software.*

Key Players in the Data Landscape

Key Players	How do they get involved?
Government Agencies	• Facilitating data collection and dissemination, or implementation of analytics solutions. • Government involvement in data collection and analytics in a sector can significantly affect how advanced the sector is in using data analytics.
Businesses	• Backbone of the data and analytics landscape in each sector. • The users or adopters of data analytics.
Customers	• Generate data for businesses. • Providing data on personal information, transaction habits and preferences to businesses.
Data Analytics Solutions Providers	• Provide either the solutions to analytics or the tools to perform analytics. • These range from companies that perform actual analytics, to software providers, mobile applications developers, and hardware developers.
Data Storage Providers	• Physical (within premises) or on the cloud.
Data Aggregators	• Compile industry-level data and provide these to businesses that do not have an understanding of the overall industry (modern day consultants/ data wizards).

How the Data “Eco-system” in Singapore Looks Like



Government Agencies

Data Storage Providers

Data stored physically or on the cloud

In selected sectors, Government agencies act as central data sources

Businesses

Data aggregators compile individual business data into overall industry data

Data Aggregators

Analytics Solutions

Analytics solutions often provided/ complemented by external parties

Customer are the end-beneficiaries of analytics solutions

Businesses collect customer personal and behavior data

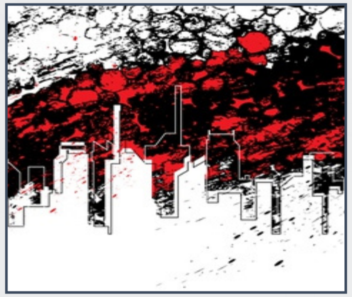
Customers

Flow of data

Flow of analytics/ solutions

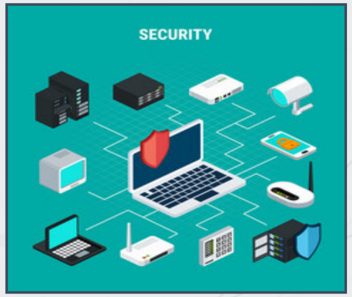
Data Analytics Sectors

Sectors for which data analytics is particularly important:



Urban Services

Consumer/ Retail



Security

Manufacturing

Finance

Supply Chain/ Logistics

Healthcare

Healthcare Sector and Data Science

- The healthcare sector in Singapore is burdened by an ageing population and an increased chronic disease burden.
- Lowering healthcare costs and managing limited capacity are two main areas of focus in the sector.
- We may improve on these efforts via careful data collection and performing appropriate analyses.
- We shall touch briefly on the mode of data collection (database) and what is being done with the data (without going into technical details).

Healthcare Sector and Data Science (Database)

- The Integrated Health Information Systems (IHiS), a wholly owned subsidiary of MOH Holdings, champions the implementation of analytics and other technologies among public healthcare providers.
- A prime example is the National Electronic Health Record (NEHR), a central portal for patient data. The portal reduces the chances of wrong diagnoses from incomplete patient data and allows patients to avoid unnecessary duplicate tests.
- All public healthcare providers participate in NEHR. But this is optional for private (a data blind spot).
- Reasons for limited participation include issues with incompatible IT systems, patients' preference to keep medical conditions private and the individual practice and ownership structure of private healthcare providers.

Healthcare Sector and Data Science (Database)

- Patient data are regarded with strict confidentiality, and only authorised healthcare professionals who are directly involved in patient care have access to the data. Patient participation in the NEHR is on an opt-out basis.

- Public healthcare providers utilise cloud storage, through IHiS's Health Cloud, a private cloud platform. This allows analytics solutions to be applied more quickly and cheaply.

Healthcare Sector and Data Science (Database)

The NEHR records patients' medical histories, including:

- Admission and visit history
- Laboratory results
- Radiology results
- Hospital inpatient discharge summaries
- Medication history
- History of past observations
- Immunisations
- Allergies and adverse drug reactions

Benefits of NEHR:

- Reduces the chances of inaccurate diagnoses.
- Reduce healthcare costs for patients.

Healthcare Sector and Data Science (Analytics)

- Most current attempts are observational (non-predictive).

- Some predictive analytics are also being explored:
  - **Managing limited supply of inpatient capacity**, through the use of remote patient monitoring and providing healthcare services outside the hospital.
  - **Optimising hospital facilities and resources** through dashboards that monitor in real-time utilisation of operating theatres and other facilities to minimise down town and understanding trends in patient demand to manage bed, doctor and nurse capacity.
  - **Drawing relationships** between aggregate patient health indicators with the risk of disease, and engaging patients who are deemed as high risk before they fall sick.
  - **Use of predictive analytics in personalised medicine**, predicting a person’s risk of contracting a critical disease in the future, based on his genetic makeup.

Impediments to Greater Use of Data Analytics and Sharing

- Across all sectors, stakeholders have raised the lack of skilled labour as the main impediment to greater use of data analytics. In particular, it is the lack of experienced labour with industry experience that is impeding the growth of analytics, as graduates are not provided with sufficient guidance.

- Other impediments relate to the costs of setting up infrastructure and IT systems to support the use of data analytics. Individual companies do not immediately see the benefits of such investments, and often, a lot of the initiatives are Government-led.

- There are also requests for there to be greater transparency and proper frameworks of the types of data that can be shared, especially those relating to personal customer details. This would facilitate greater data sharing.

You Need to Know about PDPA

- The Personal Data Protection Act (PDPA) in Singapore protects data that can be used to identify individuals. Before collecting, using or disclosing such data, organisations have to ensure that customer consent is obtained for that specific and reasonable purpose.

- The three key focus of the PDPA are:
  - **Consent** – organisations are only allowed to collect, use or disclose personal data with the individual's knowledge and consent.
  - **Purpose** – organisations are only allowed to collect, use or disclose personal data if the individual has been informed of the purpose of data collection, use or disclosure.
  - **Reasonableness** – organisations are only allowed to collect, use or disclose personal data for reasonable purposes.

The PDPA does not apply to Government agencies.

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Singapore’s Smart Nation Initiative and You

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Data Analytics and Smart Nation

In 2017,

The data analytics industry is expected to add **S$1 billion** in value to the Singapore economy.

Singapore aims to become the **Big Data Hub** in Asia.

Singapore aims to train **2,500 multidisciplinary data analytics professionals** to join the workforce.

But What is ‘Smart Nation’?

- Development of digital infrastructure in main population centres and the meshing together of digital and analogue infrastructure.
- Primary focus on infrastructure.
- Future plans: Digital technologies can benefit the lives of individual citizens, support effective day-to-day elements of our lives in terms of work, home and community, towards “smarter living”.
- In other words, we want to use data-centric approach to improve the way we function as a society and nation.

The 3 Ds of Smart Nation

The process of converting data or information into digital format.

Digitisation

The use of digital technologies to provide new opportunities and value.

Digitalisation

Integration of digital technologies into everyday life; digitisation of everything that can be digitised based on data, analysis and extraction of “meaning” of data, and conversion of this “meaning” into opportunities and value.

Digital Society

What Makes a “Smart Nation”?

It can be loosely defined as the automation of all aspects of urban life to achieve greater efficiency.

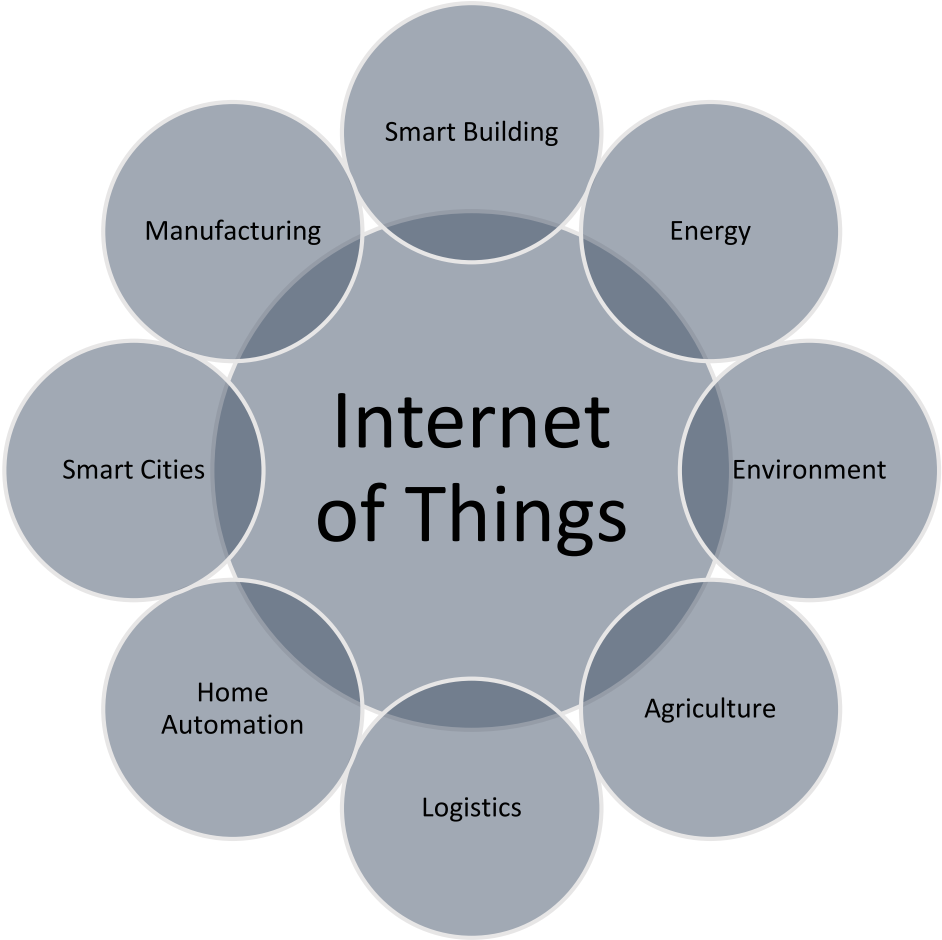
Aspects of urban life may include transportation systems, education, healthcare, etc. create a better life experience for citizens (improved quality of life).

However it goes further…

Singapore’s key resource is human capital. And a smart nation must run on smart people who can manage these new technologies.

Education and training also needs to change to meet these new requirements. Digital Literacy in University for all undergraduate degrees. Lifelong learning schemes such as Skills Future.

Internet of Things (IoT)



- Smart Nation is not just about the data.
- It is also about the smart devices that interface with us.
- How these smart devices deal with data and generate data, and how they communicate with each other is known as the “Internet of Things”.
- <u>A smart device should be able to</u>:
  - Sense or monitor aspects of the real world, such as temperature, lighting, the presence or absence of people or objects, or human behaviour (movement, sound, heart beat…).
  - Able to transmit and receive information from other devices or from the internet.
  - Examples: household appliances, fitness smart devices, smart alarm clock, monitors, e.g. to alert caregivers.

“Smart Nation” Initiatives

Some Key Projects Linked to the “Smart Nations” Initiatives

	2017	2018	2019	2020	2021	2022
National Digital Identity (NDI)		• SingPass mobile token • MyInfo extended to more government and private esector services		• NDI platform operational • Digital signatures for paperless transactions		• Widespread adoption of NDI
E-payments	• PayNow • Account-based ticketing for public transport	• Common payment QR standard	• E-payments in hawker centres • Roll-out of 25,000 unified point-of-sales terminals			
Smart Nation Sensor Platform (SNSP)		• Wireless sensor network deployment in Orchard Road and selected areas		• Smart connected lamp post trial		• City-level sensor data available for industry and public
Smart Urban Mobility	• More timely bus arrivals with common public bus fleet management system	• Trial of on-demand bus services				• Deploy autonomous vehicles to enhance public transport (Earliest)
Moments of Life (MOL)		• MOL app for families with young children	• Expanded suite of services for families with young children			

Smart Nation Sensor Platform (SNSP)

Integrated nation-wide sensor platform (wireless sensor network) to improve:

- Municipal services
- Urban planning
- City-level operations
- Public security
- Responsive and reliable public transport

Examples:

- **Wireless Sensor Network (trial):** over 500 sensors in Yuhua to transmit water usage data from smart water meters → enabling users to access near real-time water usage data and detect water leaks through a mobile app → empowers users to save water.
- **Drowning detection at swimming pools (pilot):** system using computer vision for drowning detection and continuous surveillance of activities in the pool; alert lifeguards to react faster to swimmers in distress → prevent drownings.
- **Personal Alert Button (trial):** develop cost-effective, lightweight, easy-deployable elderly help button solution to replace current pull cord system.
- **Smart Lamp Posts (Lamppost-as-a-Platform, Laap; trial):** hosting sensors on lamp posts; use crowd analytics and environmental sensors to measure air quality, rainfall and water levels. Analysis of sensor data using various techniques, including AI. Goal: improve policy making and service delivery for citizens and businesses.

Smart Lamp Posts

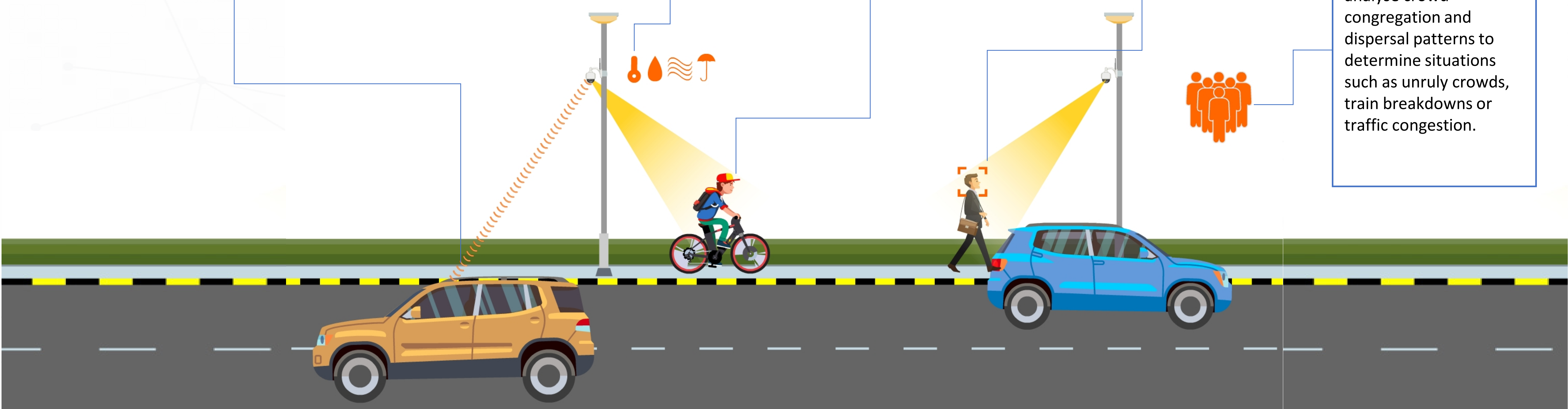
Autonomous vehicle
Real-time kinematic technologies mounted on lamp posts will provide line-of-sight connection to self-driving vehicles, to determine their precise location for navigation and to avoid collisions.

Environmental Sensors
Sensors mounted on lamp posts will be able to collect environmental data, including temperature, humidity, air quality and rainfall. The data is sent to self-driving cars to improve their situational awareness of road conditions.

Personal Mobility
Camera and artificial intelligence-based video analytics systems mounted on lamp posts will be able to determine if a mobility device or bicycle is travelling at more than 15kmh on footpaths, which is illegal. The data will be captured and sent to the relevant agency.

Facial detection
Camera and artificial intelligence-based video analytics systems mounted on lamp posts will have the ability to index faces to determine gender, race and age, as well as perform facial matching against databases.

Crowd analytics
The lamp post-mounted systems will be able to analyse crowd congregation and dispersal patterns to determine situations such as unruly crowds, train breakdowns or traffic congestion.



Visit this link to learn more about what the smart lamp posts can do.

Recreated from Source: GOVTECH

Smart Urban Mobility

Using digital technologies to enhance comfort, convenience and reliability of public transport systems and support a car-lite Singapore:

- **Improved commuting experience**: Analysing anonymised data obtained from commuters' fare cards and identifying commuter hotspots to manage bus fleets more efficiently.
- **Contactless fare payment for public transport:** Hands-free ticketing technology at fare gates (MRT/ buses) allow elderly, families with young children, and commuters with mobility challenges to enter and exit train stations without the need to tap fare cards at the gates → **more inclusive public transport system.**
- **Autonomous (on-demand) shuttles**:
  - Strengthen intra-town connectivity.
  - Enhance mobility for commuters (particularly the elderly and persons with disabilities).
- **Autonomous Vehicles**: Optimising the use of limited road space for more efficient, safe, reliable and enhanced transportation.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Summary

BS0004 Introduction to Data Science

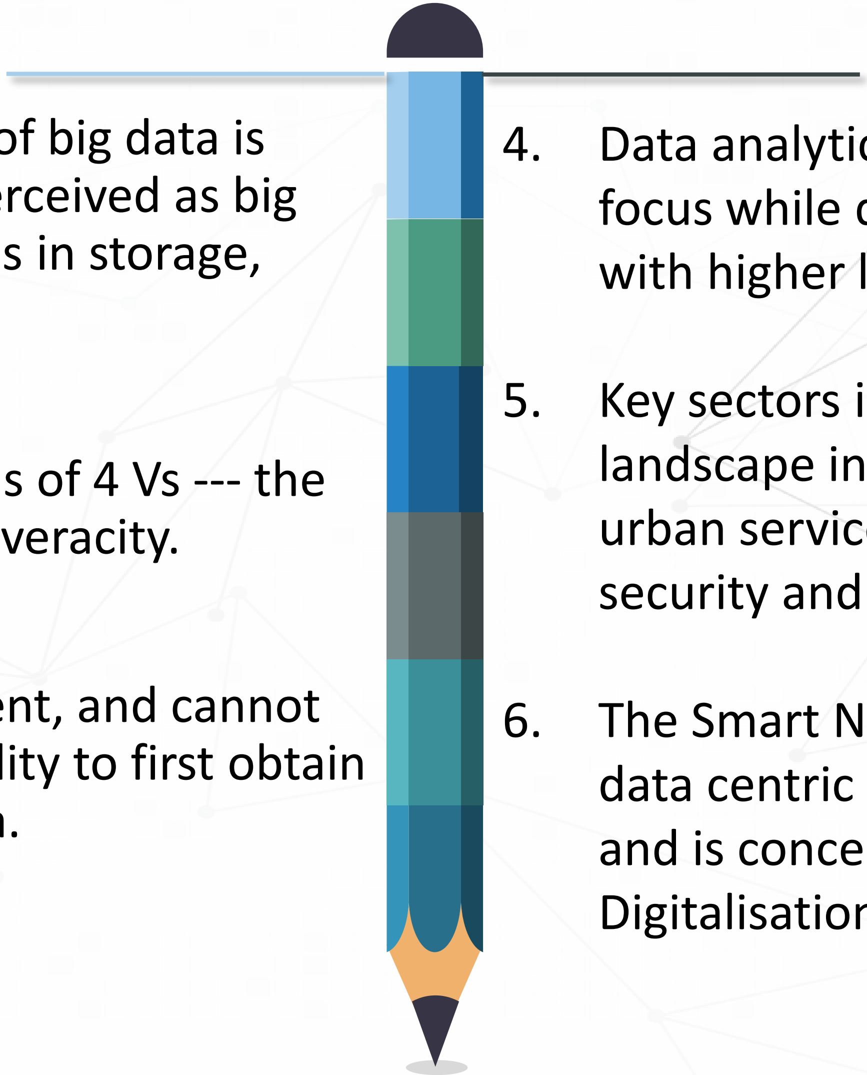
Dr Wilson Goh
School of Biological Sciences

Key Takeaways from this Topic

1. The concept and perception of big data is relative --- the data is only perceived as big because of existing limitations in storage, handling and analysis.

2. We may think of data in terms of 4 Vs --- the volume, velocity, variety and veracity.

3. Data science is data-dependent, and cannot be done well without the ability to first obtain a good amount of useful data.



4. Data analytics has an applications-based focus while data science is more concerned with higher level theories.

5. Key sectors in the data science/ analytics landscape in Singapore include healthcare, urban services, logistics, retail, finance, security and manufacturing.

6. The Smart Nation Initiative is about taking a data centric approach towards urban life and is concerned with 3Ds: Digitisation, Digitalisation and Digital Society.