AIML427 Big Data

Welcome to AIML427

Week 1: Introduction to Big Data

Dr Qi Chen and Dr Bach Hoai Nguyen

School of Engineering and Computer Science

Victoria University of Wellington

Qi.Chen@vuw.ac.nz

Bach.Nguyen@vuw.ac.nz

Outline

- Welcome to AIML427
  - Team
  - Teaching Mode
  - Assessment, Extension, and Plagiarism
  - Course Materials

- What is big data?
- Where is big come from ?
- What we can do with big data?

AIML427 Team

- Course Coordinator and Lecturer:
  - Dr Qi Chen (qi.chen@vuw.ac.nz, CO329),
    - Week 7-Week 12
    - A2 - A3 and Test
  - Dr Bach Hoai Nguyen(bach.nguyen@vuw.ac.nz, CO364),
    - Week 1- Week 6
    - A1- A2 and Test
- Lectures/Tutorials/Discussions:
  - Monday 15:10 - 16:00 – **501, Murphy**, Kelburn
  - Thursday 15:10 - 16:00 – **501, Murphy**, Kelburn
  - No tutorial or helpdesk
- You can ask questions at any time in the Lectures!
- Get to know each other!!

Teaching Mode

- The course is taught in-person (+ recording)
  - Access to lecture and tutorial recordings can be found here on VStream or through Nuku shortly after each class.

- This course has critical in-person components, and students are strongly recommended to attend lectures
  - We have In Person Test
  - We have a in person presentation and Group Work for Assignment 3

- ALL learning materials are available online: https://ecs.wgtn.ac.nz/Courses/AIML427_2024T1/

- We ***use Nuku*** for announcement, class recording, and later course/teaching evaluation.

Course Assessment

- 3 Assignments are worth and 1 test
  - Assignment_1 (20%), Due 25/03/2024, **Monday Week 5**, 11:59 pm
  - Assignment_2 (25%), Due 06/05/2024, **Monday Week 9**, 11:59 pm
  - Test (25%), 20/05/2024, **Monday Week 11**, Lecturing Time
  - Assignment_3 (30%), Due 10/06/2024, **Tuesday Week 14**, 5:00 pm

- Programming language: Java, R, Python

- To pass the course you must obtain a C- grade overall

- READ the course outline

Extensions

- We are using the "3 late days" model for this course, for assignments but not applied to Test or the group part of A3.
  - Close deadlines
  - Computer problems
  - Getting really turnt at a party
  - ...
- Automatic penalties after 3 late days
- Thus, minor extensions will not be approved.
- Don't waste your late days early on
- Medical/exceptional circumstances? Apply for extension via Submission System – preferably before the deadline

Plagiarism

- Using somebody else's work as your own, without saying so.
- This includes anything not in the lectures!
- We suggest you don't use AI tools (ChatGPT, Bing Chat, Github Copilot, Google Bard, …) to generate submitted material, or complete coursework.
- If use AI tools to check them, it is at your own risk for Plagiarism

- It's really easy to avoid.
- Just tell us if you used a resource!! (and how much)
- It is never plagiarism if you are honest.
- The penalty for getting caught can be worse than getting no marks...it's not worth it.

Course Materials

- Course web page:
  - https://ecs.wgtn.ac.nz/Courses/AIML427_2024T1/
  - **Announcements via Nuku**
- Course Information
  - *Course Outline*
  - *Lecture Schedule*
  - *Assignments*
  - *Submissions*
  - *Nuku*
  - *Reading List*
- These materials will be updated from time to time
- **Multidisciplinary 400-level course**

Data in Our Real World

- Google processes over 8.5 billion searches per day in 2022
- Facebook
  - over 1.38 billion Facebook Messenger users globally.
  - over 7 billion conversations every day

- How many emails do you think are sent per second?
  - 3,400,000.

- The Gartner forecast in 2012, predicted that Big Data would account for 6 million jobs in the US in 2015.
  - actually outgrew the estimates by a third.

- Five exabytes of information created across all of 2002 – we now create this *every two days* globally!

What is Data?

- A **data set** is a collection of data objects, where each object has a set of attributes that ***describe*** the object.

- Due to the ***interdisciplinary*** nature of big data, the terms data object and attribute have many synonyms.
  - **Data Object**: record, point, datapoint, vector, pattern, event, case, observation, example, instance, entity, ...
  - **Attribute**: feature, variable, characteristic, field, dimension, ...

  - Input: independent variable, predictor.
  - Output: dependent variable, response.

What is Big Data?

- The first (at least one of the first) documented use of the term **big data** appeared in a 1997 paper[1] written by NASA scientists.

"... *data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not in main memory (in core), or when they do not even on local disk, the most common solution is to acquire more resources.*"

[1] Cox, Michael, and David Ellsworth. "Application-controlled demand paging for out-of-core visualization." In *Proceedings. Visualization'97 (Cat. No. 97CB36155)*, pp. 235-244. IEEE, 1997.

What is Big Data?

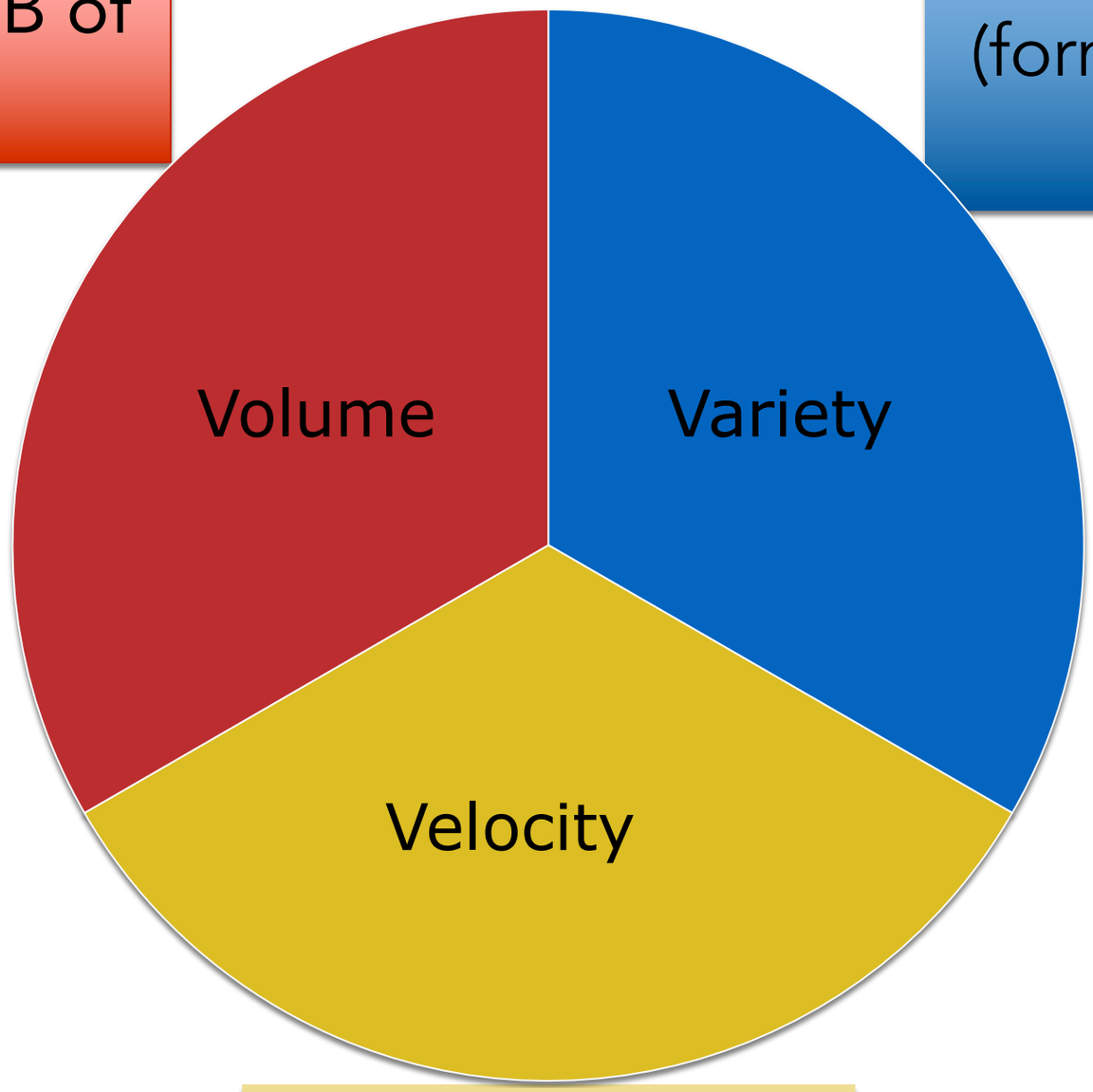
- The **Oxford** English Dictionary defines big data as:
  - "Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges."
- **Wikipedia** defines big data as:
  - "Big data is data sets that are so voluminous and complex that traditional data processing application software are *inadequate* to deal with them."
- **McKinsey** (Global research group) defined big data as:
  - "Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."

"This definition is intentionally subjective and incorporates a *moving definition* of how big a dataset needs to be in order to be considered big data."-- McKinsey report on big data (2011)

The 3Vs in Big Data

Size: TB, GB of data

Different types (forms, sources) of data

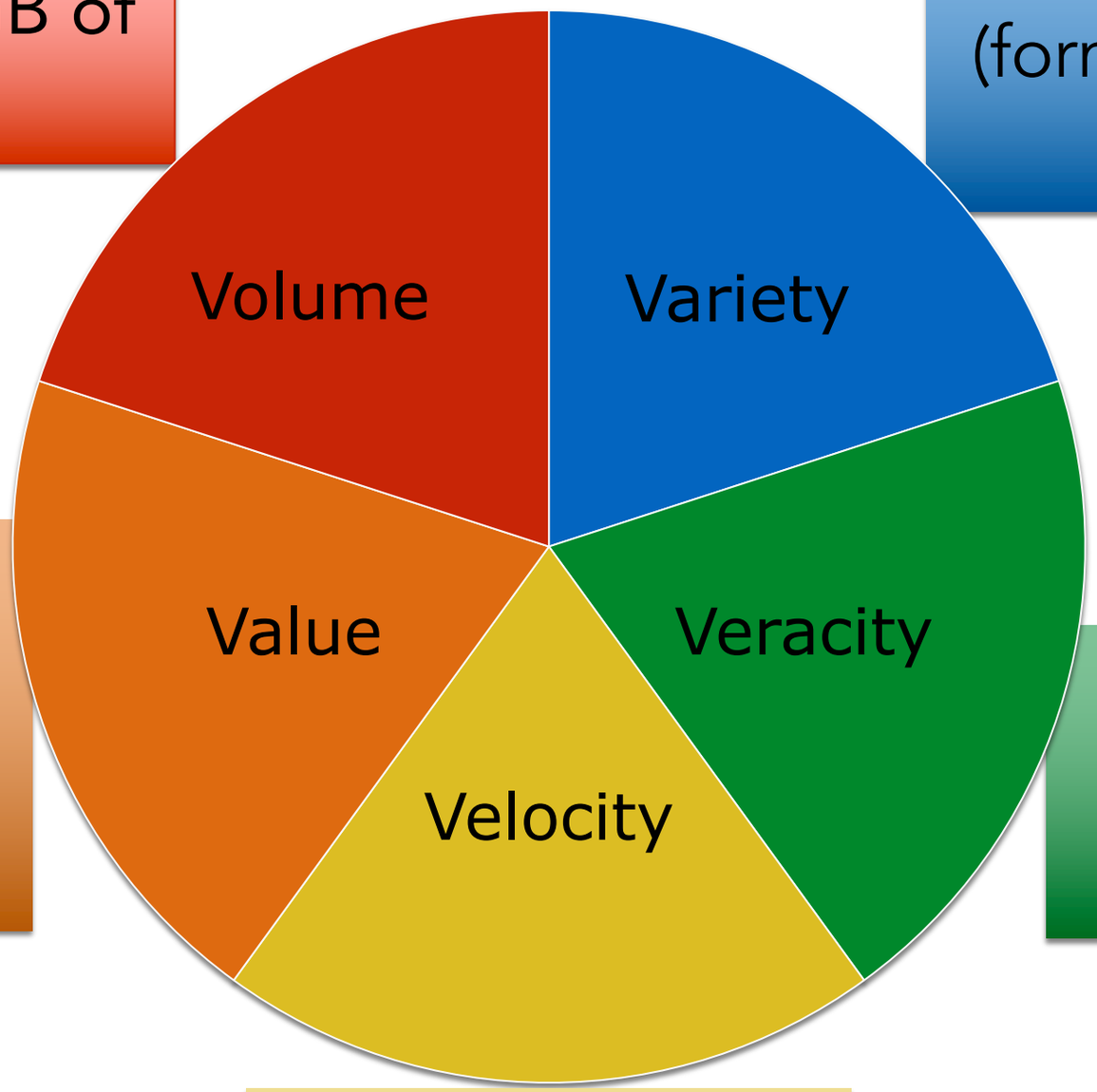


Speed: generate new data and data moves

The 5Vs in Big Data

Size: TB, GB of data

Different types (forms, sources) of data



Big data is no use unless we turn it into values

Trustworthiness, availability, authenticity

Speed: generate new data and data moves

Not all refer to "bigness".

Big Data

- Some numbers that are commonly used in big data.

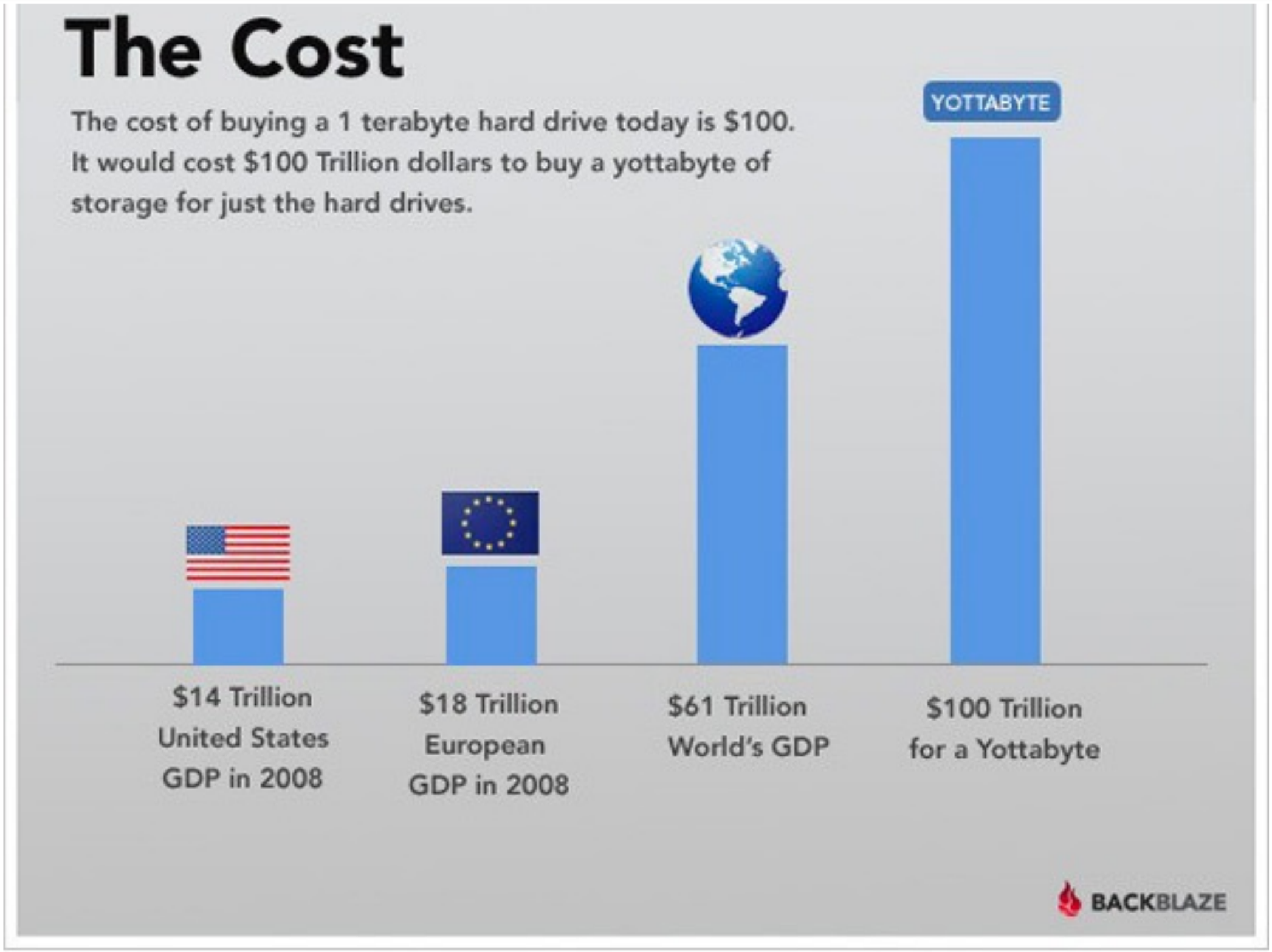
Value	Name	Number
1000	kilobyte	1,000
1000^2	megabyte	1,000,000
1000^3	gigabyte	1,000,000,000
1000^4	terabyte	1,000,000,000,000
1000^5	petabyte	1,000,000,000,000,000
1000^6	exabyte	1,000,000,000,000,000,000
1000^7	zettabyte	1,000,000,000,000,000,000,000
1000^8	yottabyte	1,000,000,000,000,000,000,000, 000

What is a terabyte and petabyte?

- What is a terabyte?
  - The size of an inexpensive (< $100) external hard drive.
  - 220 DVDs.
  - 4.5 million books.
  - two years of MP3 music.
  - 16 million Facebook photos.

- What is a petabyte?
  - 20 million four-drawer filing cabinets filled with text.
  - 13.3 years of HD video.
  - 2000 years of MP3 music.
  - The number of photos (3MB, standard size/shape) that, when placed side-by-side, would go around the equator almost twice.
  - The number of cells in the human body is ≈ 100 trillion (10^{14}). If one bit (0 or 1) represents a cell, then you would have enough cells in a petabyte for ≈ 80 people.

The Cost



https://www.theatlantic.com/technology/archive/2011/05/infographic-how-big-is-a-yottabyte/239034/

Volume

- Volume: the magnitude/size of data.
- Volume is relative, varies with time and depends on data type making it impractical to define a specific threshold value (terabyte and above?)

  - According to IBM, 2.5 exabytes of data was generated every day in 2012.

  - On a single flight, the engines on a Boeing 787 could generate half a terabyte of data (used for predictive maintenance).

  - Square Kilometre Array Radio Telescope will generate 14 exabytes of data per day.

  - How many photos or posts Facebook store and process?

Variety

Variety: refers to the structural heterogeneity in a data set.

- Structured: can be stored, accessed and processed in the form of fixed format
  - tabular data found in spreadsheets and relational databases
  - Estimated to be ≈ 5 − 10% of all existing data

- Unstructured: data that is not organized in a predefined manner
  - heterogeneous data source containing a combination of simple text files, images, videos etc

- Semi-Structured: can contain both the forms of data
  - user defined tags making them machine-readable.

Velocity

Velocity: the rate/speed at which data are being generated and the speed at which it should be analyzed.

- Nuclear physics experiments at the Large Hadron Collider at CERN generate 40 terabytes of data every second.

- Wal-Mart can process more than one million transactions per hour.

- Facebook processes up to one million photos per second.

Data can be analysed in real-time using location, demographics, past buying patterns ...

Velocity

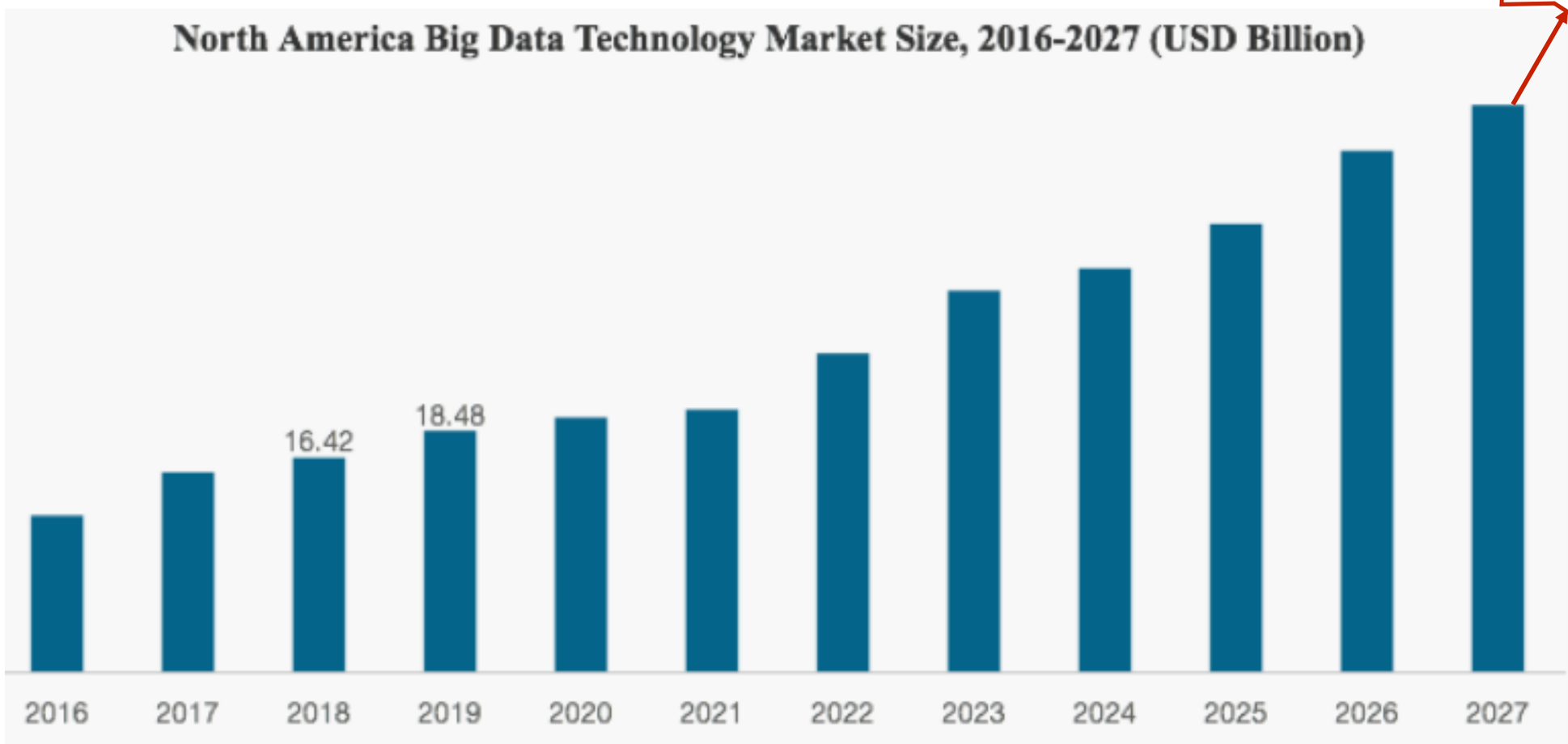

44,000,000
MESSAGES PROCESSED
486,000
PHOTOS
26
NEW REVIEWS
POSTED ON YELP
120
NEW ACCOUNTS
OPENED ON
LINKEDIN
MORE THAN
140
SUBMISSIONS
ON REDDIT
MORE THAN
2,315,000
SEARCHES
3,125,000
243,055
MORE THAN
21,000,000
MESSAGES SENT
70,000
VIDEO MESSAGES
SHARED
Share
MORE THAN
3,000,000
ITEMS ARE
SHARED
MORE THAN
18,000
MATCHES MADE
MORE THAN
195,000
MINUTES OF AUDIO CHATTING
ON WECHAT
972,000
DAILY SWIPES
ON TINDER
MORE THAN
69,500
HOURS OF
VIDEO WATCHED
ON NETFLIX
NETFLIX
IN
60
SECONDS
MORE THAN
150,000,000
E-MAILS ARE SENT
GO-Globe
CUSTOM WEB DEVELOPMENT
MORE THAN
48,000
APPS DOWNLOADED
ON IPHONE
MORE THAN
95,000
APPS DOWNLOADED
ON ANDROID
MORE THAN
430,000
TWEETS SENT
MORE THAN
2,700,000
VIDEO VIEWS AND
139,000 HOURS
OF VIDEO WATCHED
MORE THAN
39,300
HOURS OF MUSIC
LISTENED
AROUND
56,000
PHOTOS
UPLOADED
9,800
ARTICLES PINNED
ON PINTEREST
MORE THAN
280,000
SNAPS SENT
ON SNAPCHAT
MORE THAN
100
NEW DOMAINS
REGISTERED
14 NEW
SONGS ADDED
ON SPOTIFY
MORE THAN
300 HOURS
OF VIDEO ARE UPLOADED

Value

Value: comes from the analysis of big data.

- Not related to bigness
- Big data is no use unless we turn it into values
- McKinsey states that the annual value of big data to U.S. Health Care is $300 billion.
- Potential annual consumer surplus from using personal location data globally was estimated at $600 billion.

Value



Big Data companies are forecast to see dramatic revenue increases in the years ahead.

Veracity

Veracity: accuracy and trustworthiness of the data.

- Important even though it doesn't refer to bigness!
- Note: this is a property that all data sets should have — big or small!

Veracity is particularly important for big data for several reasons:

- Data Fusion: Lots of structured, semi-structured and unstructured data are fused together to create big data.
  - These data can be in different formats, different units, different quality, have missing values,...
  - Potentially creating all sorts of conflicts in the data.

Veracity

- **Sampling Bias**: some sampling units are unintentionally more or less likely to be included in the sample than others. (Note: sampling bias can improve or worsen your results!)

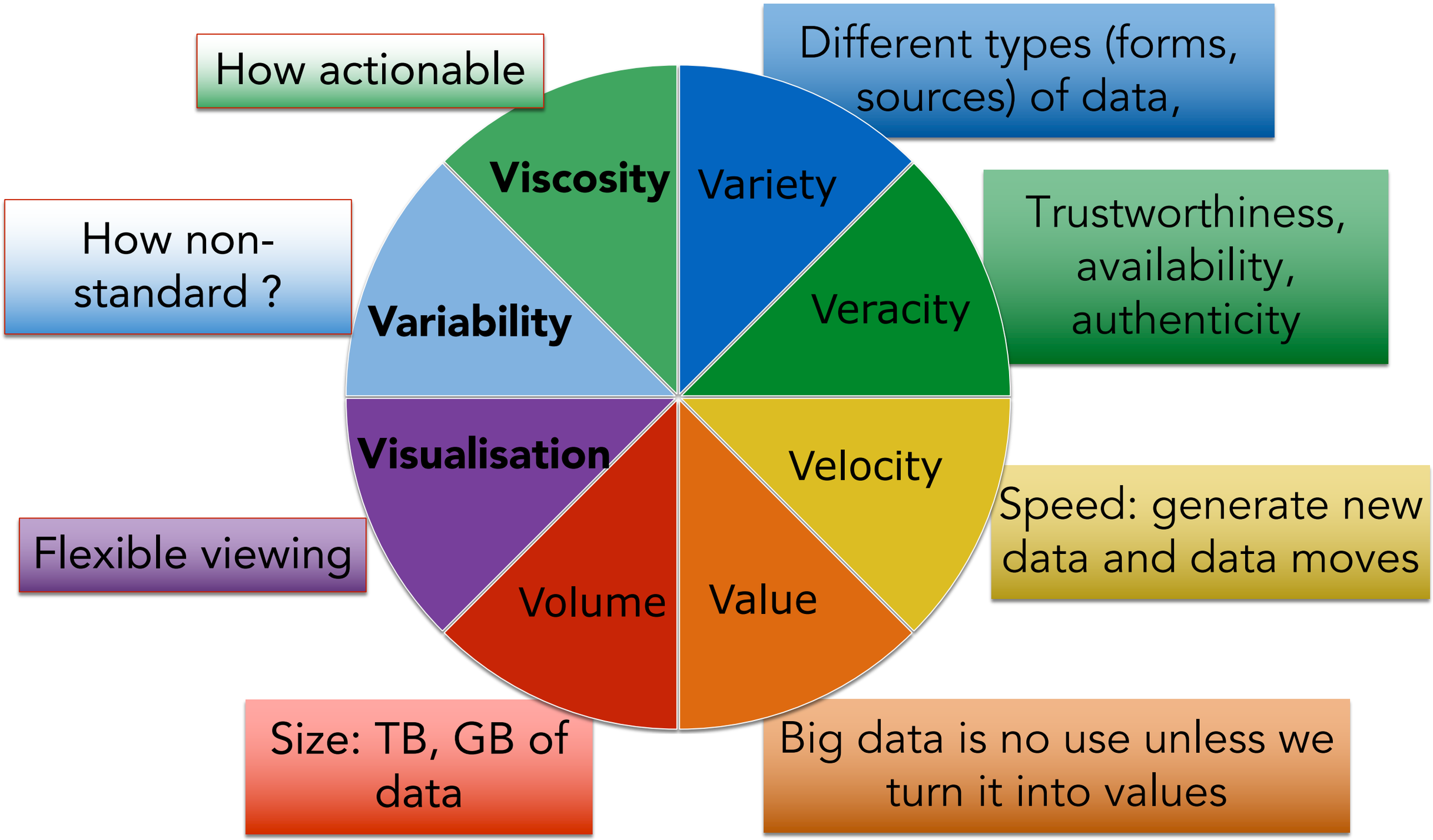
  - Should not be confused with *sampling error*, which is the error due to observing a sample rather than the entire population.

  - Under-coverage Bias: is when there are too few observations from a subset of the population.

  - E.g. Sentiment analysis: analyse every tweet on Twitter for a given day to draw conclusions about the public mood.

    -Twitter users are not representative of the population as a whole!

The 8Vs in Big Data



Not all refer to "bigness".

Where big data comes from

- In 2016, Facebook allowed developers to integrate Chatbots to its messenger services.
  - From 30,000 bots deployed in first 6 months to over 100,000 today,
  - the platform is processing 2 billion messages every month.

Where big data comes from?

- Some experts have estimated that 90% of all the data in the world today was produced within the last two years.
- Data has always been generated, but it is now cheap to store (≈ $100 for a terabyte hard drive).

- IBM estimates that 2.5 exabytes of data is generated every day.
- A single Jet engine can generate *10+terabytes* of data in *30 minutes* of flight time. With many thousand flights per day, generation of data reaches up to many *Petabytes.*

Where big data comes from?

- Science:
  - Astronomy, Particle Physics, Biology, satellite imagery, genomics, environmental data, transportation data, …

- Humanities and Social Sciences
  - Scanned books, historical documents, social interactions data,

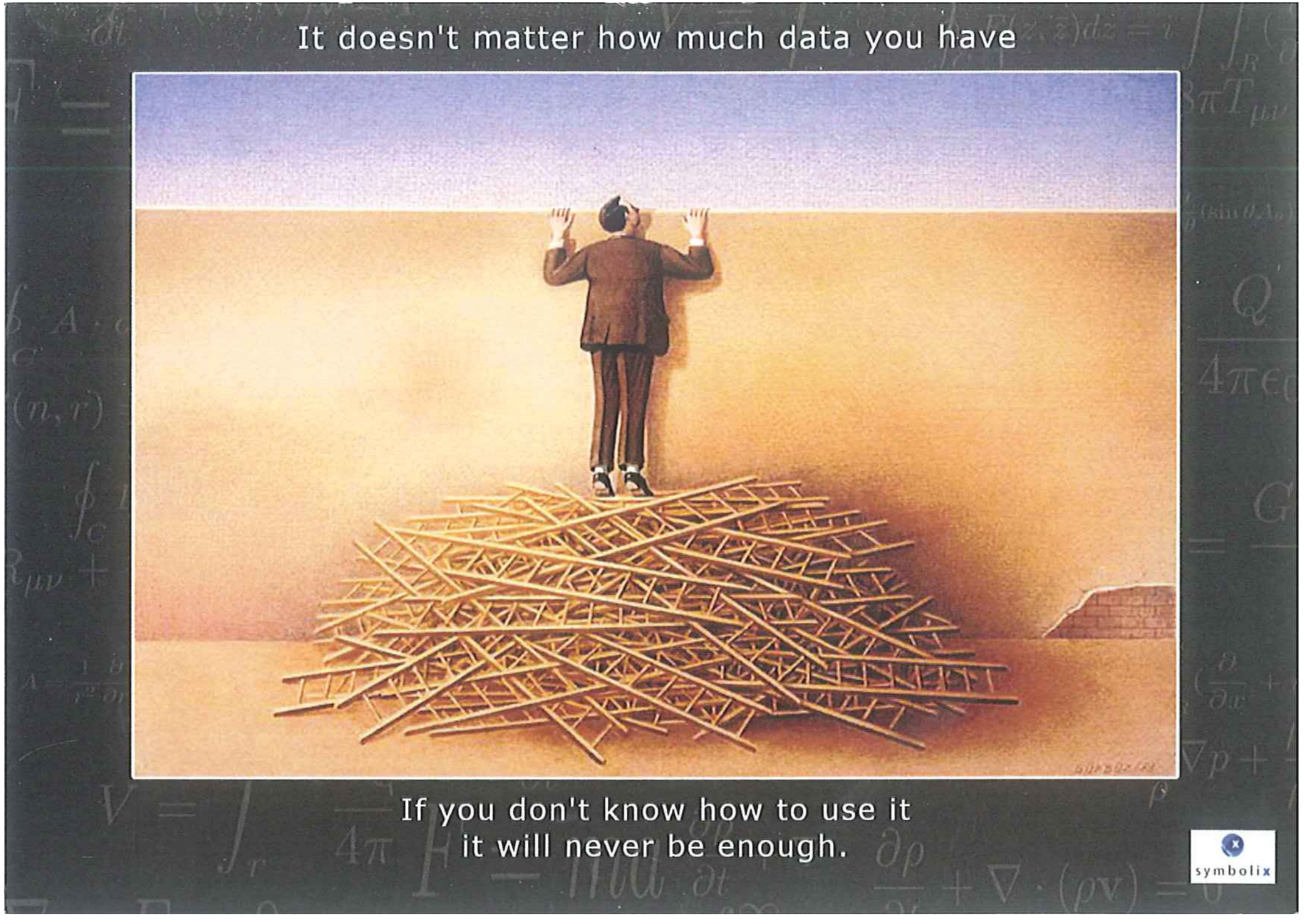
- Social media:
  - *500+terabytes* of new data get ingested into the databases of social media site Facebook, every day.
  - This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

- SMART PHONES:
  - Our personal digital exhaust.

Where big data comes from?

- Internet of Things (IoT): (physical objects that are connected to the internet)
  - For example, there are now countless digital sensors worldwide in industrial equipment, cars, electrical meters and shipping crates. They can measure and communicate location, movement, vibration, temperature, humidity, even chemical changes in the air.
  - There were *3.7 billion connected "things" in use in 2014*.
- Online Transactions:
  - Amazon, Target, Wal-Mart, Dominoes, Corporate sales, stock market transactions, census, airline traffic,
  - The New York Stock Exchange generates about *one terabyte of new trade data per day*.
- Medicine
  - MRI & CT scans, patient records, …
  - Sensors

What can we do with Big Data?



What can we do with Big Data?

- Data Mining, Machine Learning, Statistical Learning, ... techniques can be used to discover interesting patterns, associations and knowledge from big data.

- Learning from big data is an **interdisciplinary** task:
  - Statistics; Computer Science; and Mathematics;