# Welcome to INFO319:
# Big Data
## Autumn 2022

Andreas L Opdahl
<Andreas.Opdahl@uib.no>

# Session 1

- Introduction to big data
- Getting started with Spark
- INFO319:
  - practical information
  - background / expectations
  - programming projects
  - essay
  - plans for the next session(s)

# Introduction to Big Data

# Why big data?

- *Convergence of several developments:*
  - the data explosion:
    - digitalisation; availability of open, social data; IoT
    - standards for data exchange and identifiers
    - cheaper mass storage and communication
    - powerful *multi-core* processing
  - last two decades:
    - new, distributed technologies for large-scale data management and analysis
    - organisational and societal impacts
- How to deal with data that are too big for one machine?

# Big data

- Popular since late 2000's
  - *buzzword, over-hyped*, mayalready be *waning*
  - but there is a (disruptive) *reality behind it*:
    - ever increasing amounts of available data
    - go beyond capabilities/capacities of established computing techniques and tools
    - calls for new understandings, techniques and tools
- Our working definition for now:
  - *the ever increasing amount of available data today that go beyond the capabilities/capacities of existing solutions and thus calls for new understandings, techniques and tools*

# Big data on four levels

- Technical-infrastructure level:
  - data collections that are too large to be straightforwardly handled with traditional mainstream data-processing techniques and tools
- Computing level:
  - data processing and storage that is highly-distributed, fault-tolerant, redundant
- Data level:
  - the “three V’s” and other letters
- Usage level:
  - new data-driven ways of managing and organising private, public, and ideal enterprises – including societies

# Material from around chapter 4 in Kitchin

- ...but also points from other chapters
- Chapter 1 is about *data*
- Chapter 2 is a call for *critical data studies*
- Chapter 3 is about *small data, infrastructures* and *brokers*
- Chapter 5 is about *open and linked data*
  - ...definitely interesting stuff :-)
- And we will go back to (some of) it
  - but we prefer to jump right into the most central themes from the start
  - chapter 4 is quite possible to read without the ones before

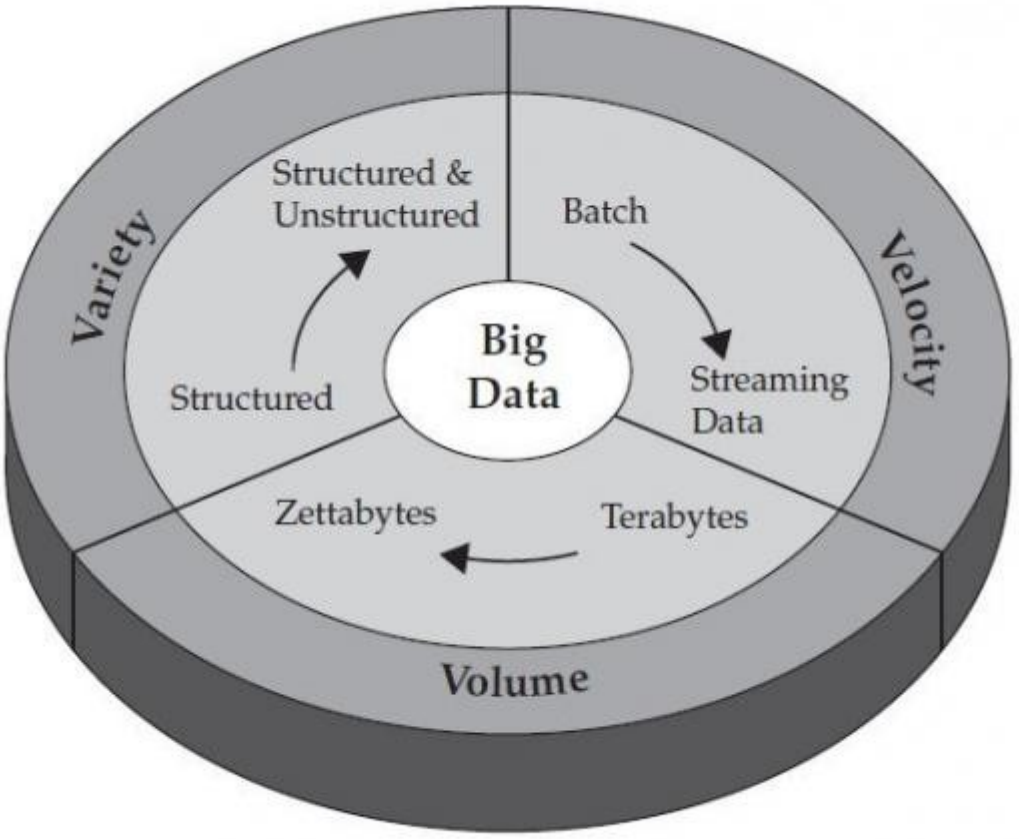



**Figure 1-1** *IBM characterizes Big Data by its volume, velocity, and variety—or simply, V³.*

# Characterstics

- The “three V's” (3V):
  - *volume*, *velocity*, *variety* – *at once*
    - old days: you could only have two of the three
  - also two more: *veracity, value*
- Other characteristics:
  - *exhaustive* in scope: *“n = all”*
  - *fine-grained* in resolution
  - *indexical*
  - *relational* in nature
  - flexible: *extensional*
  - flexible: *scalable*

# Volume

Exponentional growth in world data...

| Processor or Virtual Storage | Disk Storage |
|---|---|
| · 1 Bit = Binary Digit | · 1 Bit = Binary Digit |
| · 8 Bits = 1 Byte | · 8 Bits = 1 Byte |
| · 1024 Bytes = 1 Kilobyte | · 1000 Bytes = 1 Kilobyte |
| · 1024 Kilobytes = 1 Megabyte | · 1000 Kilobytes = 1 Megabyte |
| · 1024 Megabytes = 1 Gigabyte | · 1000 Megabytes = 1 Gigabyte |
| · 1024 Gigabytes = 1 Terabyte | · 1000 Gigabytes = 1 Terabyte |
| · 1024 Terabytes = 1 Petabyte | · 1000 Terabytes = 1 Petabyte |
| · 1024 Petabytes = 1 Exabyte | · 1000 Petabytes = 1 Exabyte |
| · 1024 Exabytes = 1 Zettabyte | · 1000 Exabytes = 1 Zettabyte |
| · 1024 Zettabytes = 1 Yottabyte | · 1000 Zettabytes = 1 Yottabyte |
| · 1024 Yottabytes = 1 Brontobyte | · 1000 Yottabytes = 1 Brontobyte |
| · 1024 Brontobytes = 1 Geopbyte | · 1000 Brontobytes = 1 Geopbyte |

See also: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

# A DAY IN DATA [2019]

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devcies - are hard to comprehend, particularly when looked at in the context of one day

## DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

| Unit | | Value | Size |
|---|---|---|---|
| b | bit | 0 or 1 | 1/8 of a byte |
| B | byte | 8 bits | 1 byte |
| KB | kilobyte | 1,000 bytes | 1,000 bytes |
| MB | megabyte | $1{,}000^2$ bytes | 1,000,000 bytes |
| GB | gigabyte | $1{,}000^3$ bytes | 1,000,000,000 bytes |
| TB | terabyte | $1{,}000^4$ bytes | 1,000,000,000,000 bytes |
| PB | petabyte | $1{,}000^5$ bytes | 1,000,000,000,000,000 bytes |
| EB | exabyte | $1{,}000^6$ bytes | 1,000,000,000,000,000,000 bytes |
| ZB | zettabyte | $1{,}000^7$ bytes | 1,000,000,000,000,000,000,000 bytes |
| YB | yottabyte | $1{,}000^8$ bytes | 1,000,000,000,000,000,000,000,000 bytes |

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

**463EB**

of data will be created every day by 2025

IDC

**500m**

tweets are sent every day

Twitter

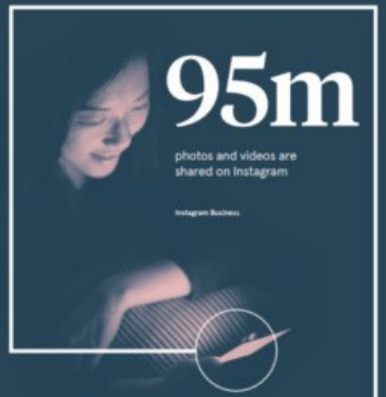
**4PB**

of data created by Facebook, including

**350m** photos

**100m** hours of video watch time

Facebook Research



**95m**

photos and videos are shared on Instagram

Instagram Business

**65bn**

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

**320bn**

emails to be sent each day by 2021

**306bn**

emails to be sent each day by 2020

**294bn**

billion emails are sent

Radicati Group

**3.9bn**

people use emails

**4TB**

of data produced by a connected car

Intel

Searches made a day → **5bn**

Searches made a day from Google → **3.5bn**

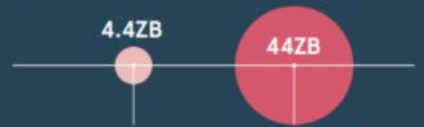
Smart Insights


**28PB**

to be generated from wearable devices by 2020

Statista

## ACCUMULATED DIGITAL UNIVERSE OF DATA




PwC

https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

RACONTEUR

# More examples (2022)

- Each day, Google processes 8.5 billion searches.
- WhatsApp users exchange up to 65 billion messages daily.
- The world will produce slightly over 180 zettabytes of data by 2025.
- 80-90% of the data we generate today is unstructured.
- It would take a person approximately 181 million years to download all the data in the Internet
- Social media accounts for 33% of the total time spent online.
- Facebook has almost two billion daily active users.
- Tweeps send over 870 million tweets per day.

https://techjury.net/blog/big-data-statistics/

# Velocity and variety

- Velocity:
  - created rapidly, in or near real time
  - analysis on the fly, not always storing it all
- Variety:
  - structured, semi-structured and unstructured
  - new sources such as
    - natural language; microblog and other messages; social media conversations; sensor data; photos; video and sound recordings; PDFs/scans...
  - some temporal, some spatial, some both, some neither
  - some socially networked, some thematically grouped

# Veracity and value

- Veracity:
  - the thrustwortiness of data: quality
    - accuracy, correctness, provenance
  - big data quality is uneven and can be low
    - e.g., microblog streams
  - how and when can volume make up for quality?
- Value
  - how to make value out of the data?
    - both commercial and societal
  - e.g.: understand/serve customers/citizens; optimise business processes; “nowcasting”; assess teaching effectiveness; societal safety; detect cyber crime...

# Exhaustiveness, resolution, indexicality

- Exhaustiveness
  - capturing and analysing data about everyone/-thing
    - instead of *sampling*
- Fine-grainedness in resolution
  - aiming to be as fine-grained as possible
  - collecting, storing and analysing smallest data points
    - instead of storing aggregate values
- Indexicality
  - unique identifiers for everyone and everything
    - trying to match different identifiers for the same person or thing (e.g., user names/handles)
  - *using IRIs to identify resources on the Web of Data*

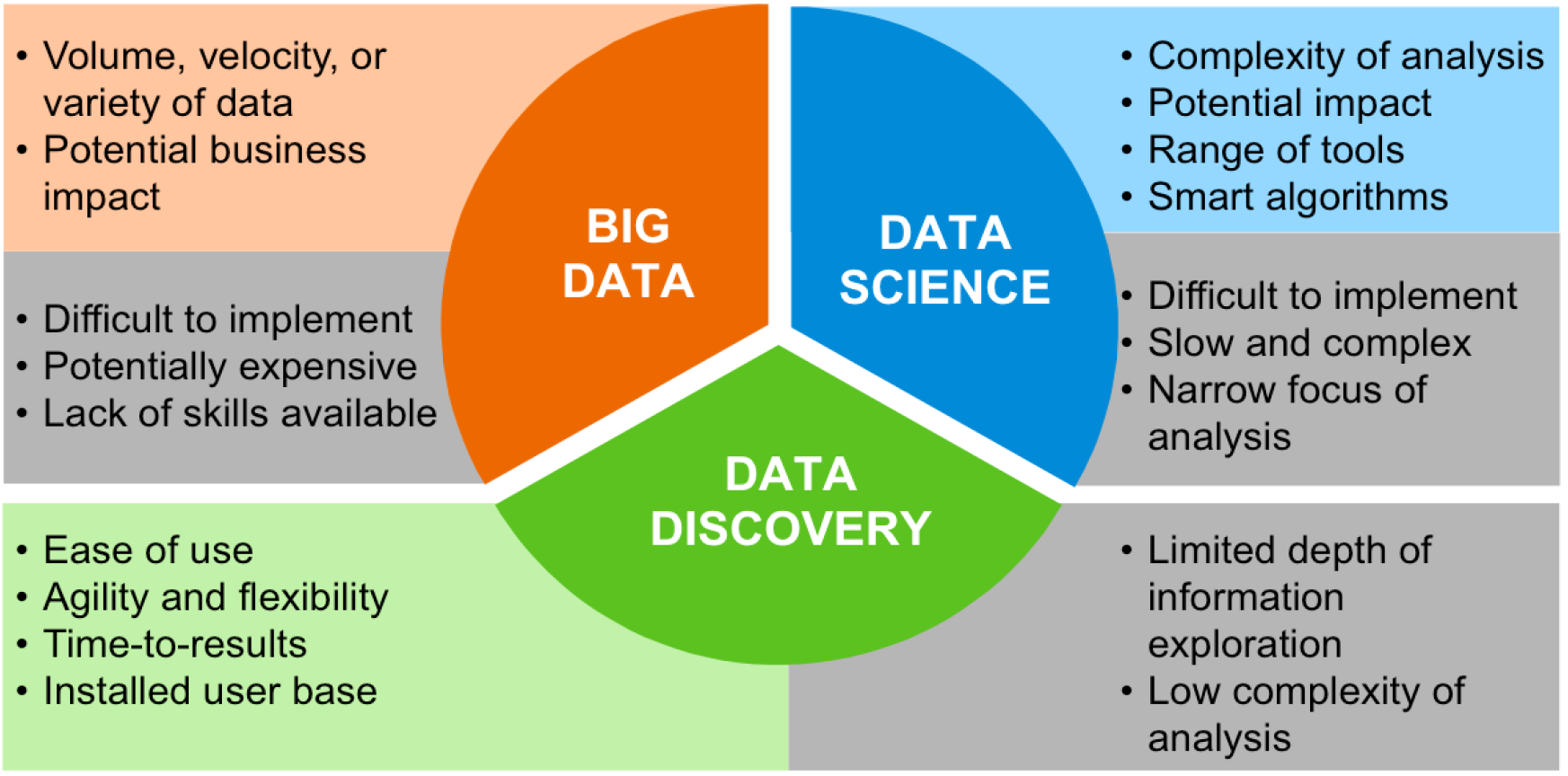
# Relationality

- People and things are described in ways that make them combinable with
  - other related persons and things
  - other descriptions of the same persons and things
- *Using IRIs to relate resources on the Web of Data*
  - *owl:sameAs, skos:narrower, skos:related etc.*

# Flexibility

- Extensionality
  - easy to add new data to the data set
  - *adding new triples to a graph and new graphs to a data set is easy in the Web of Data*
- Scalability:
  - big datasets should be able to scale rapidly
  - use of grid computing, cloud servers, NOSQL databases (Not-Only SQL)

# Putting big data to use



# Social shaping of data

- Big data are socially shaped:
  - they are a *representation* and a *sample*
  - there are no “raw”, only “cooked” big data
- Field of view: placement and settings of capture devices
- Sampling frame: who/what is available for collection
- Technology and platform: surveys, sensors, lenses, prompts, layout
- Context: generation circumstance
- Data ontology: how the data are calibrated and classified
- Regulatory evironment: privacy, data protection, security
- Also: big data tend to be opportunistic

# Big data as a disruption

- Disruptive technology:
  - a technology that displaces established ones, and shakes up existing or creates new industries
  - e.g., PCs, the internet, digital media, social media
- Big data is disruptive
  - it creates new data-driven organisation forms
  - new ways of doing research and science
  - new ways of creating and maintaining products and services
  - new threats to privacy and social order
- *...too easy to shrug off (just) as a hype/buzzword*

# Data-driven organisations

- "The next phase of the knowledge economy, reshaping the mode of production" (RK, p. 16)
  - *inward:* monitor, evaluate performance in real time; reduce waste and fraud; improve strategy, planning and decision making
  - *outward:* design new commodities, identify and target new markets, implement dynamic pricing, realise untapped potential, gain competitive advantage
- Goals: run more intelligently; flexibility and innovation; reduced risk, cost, losses; improved customer exper., return on investment, profit
- Changing *organisational practice* in *all these areas*
  - and in *a coordinated / integrated way*

# New ways of doing business

- Marts (Walmart, Kohl's): analyse sales, pricing, eoconomic, demographic and weather data to tailor local product selection and price markdowns
- Amazon: an authoritative KG-driven one-stop site for all the products in the world in context
- Online dating: sift through personal characterstics, reactions and communications to improve matches
- NY Police: analyse data on past arrests, paydays, sporting events, weather and holidays to deploy officers optimally
- Professional sports: massaging sports statistics to spot undervalued players
- Education: analyse data from learning management systems to improve teaching / studying