BIG DATA MANAGEMENT

CZ/CE4123

What are important features of big data?

Does your course schedule information belong to a kind of big data?

LEARNING OUTCOMES

- ❑ Understand what characterizes big data?
- ❑ Big data 5Vs
  - ❑ Volume
  - ❑ Velocity
  - ❑ Variety
  - ❑ Veracity
  - ❑ Value

What are important features?

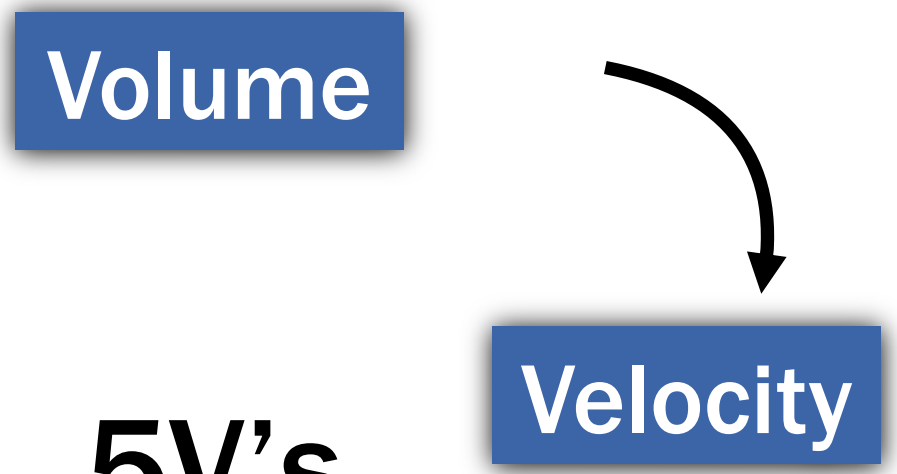
5V’s

What are important features?

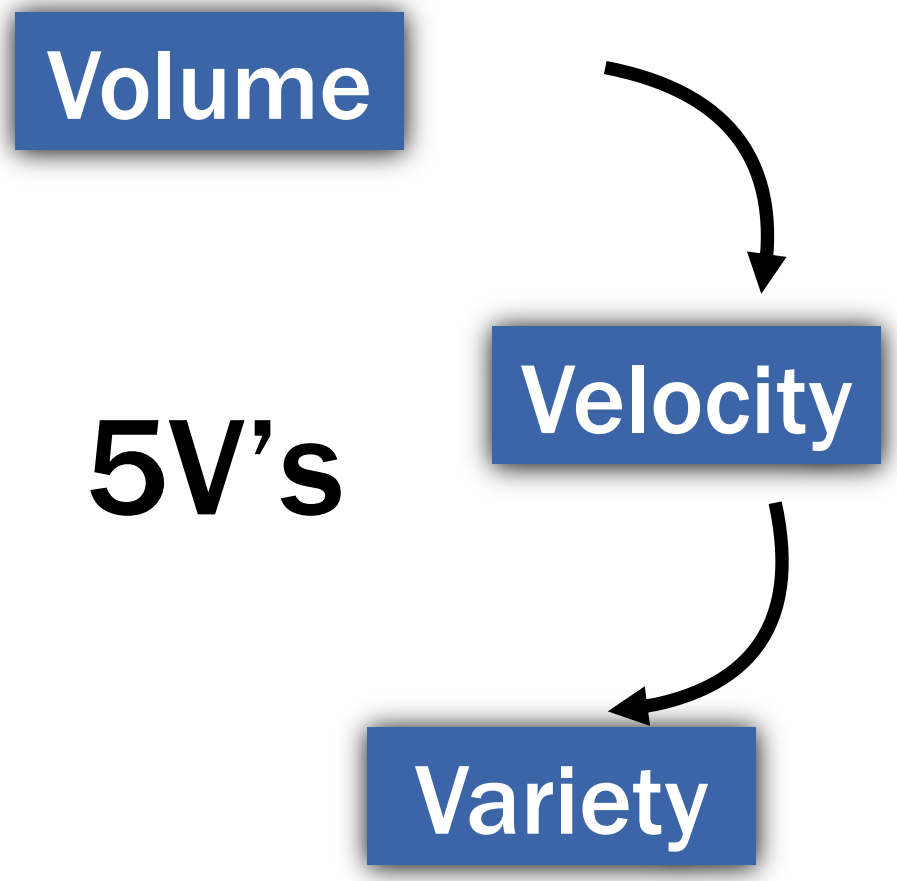
Volume

5V’s

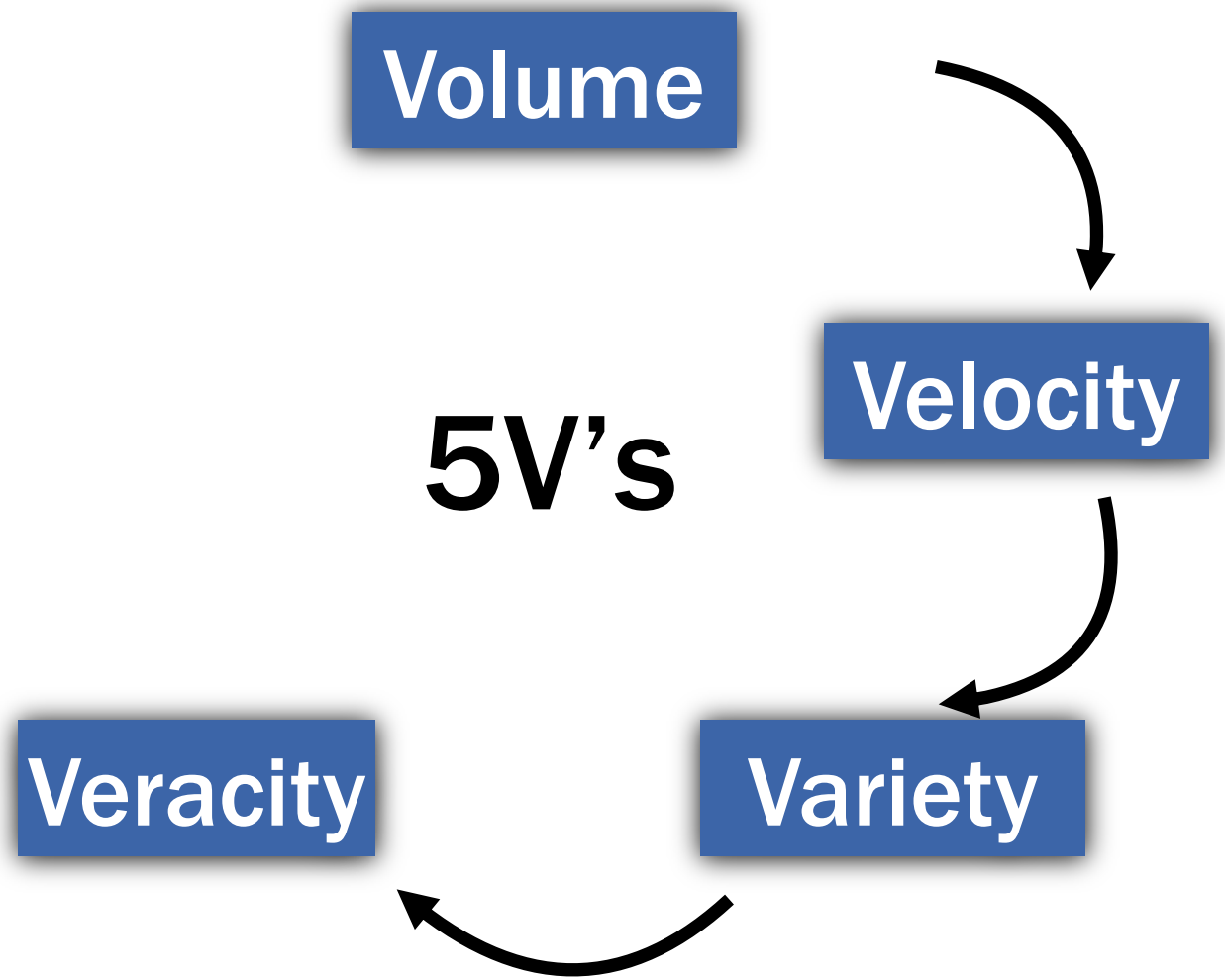
What are important features?



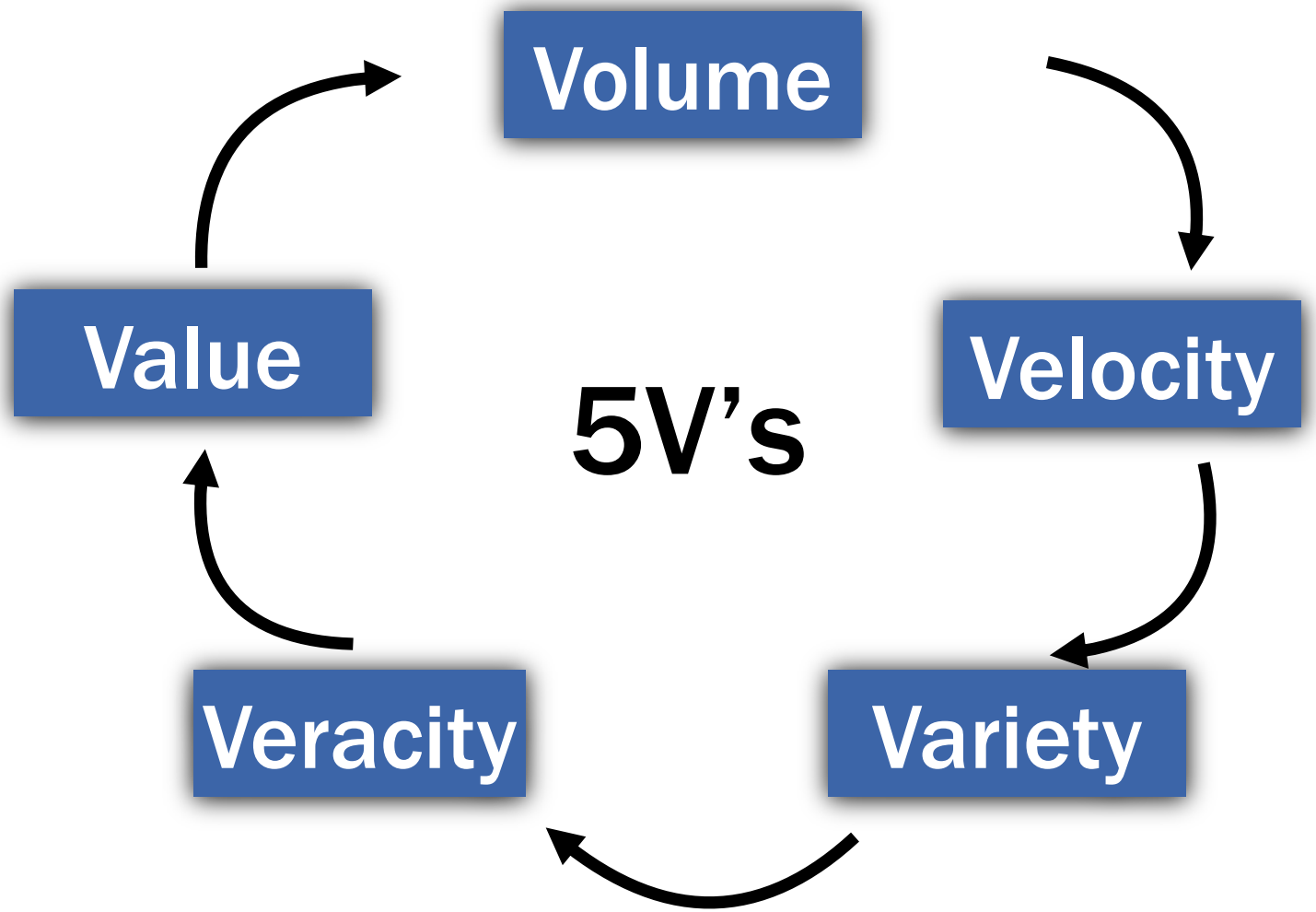
What are important features?



What are important features?



What are important features?



Volume

large amount of data

Volume large amount of data

The monthly message data volume

300 billion messages ~ 30TB

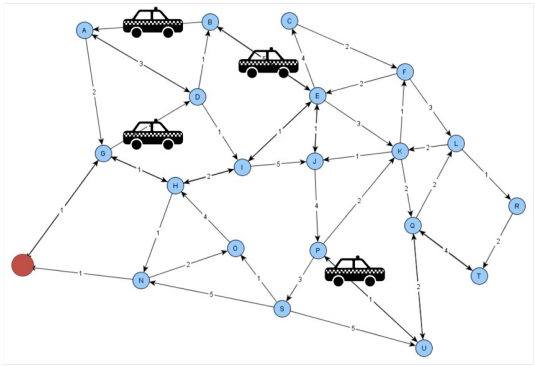
large!

Velocity

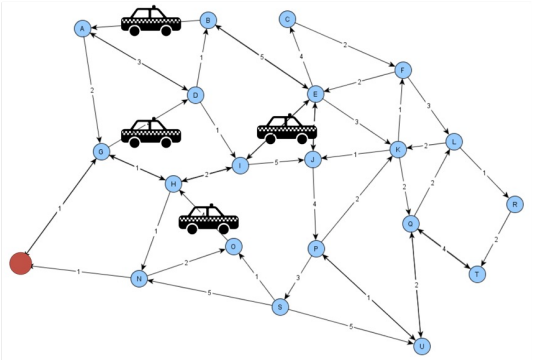
fast data generation

Velocity fast data generation

T1



T2



Example:

Taxi locations updated every 5 seconds

Frequent!

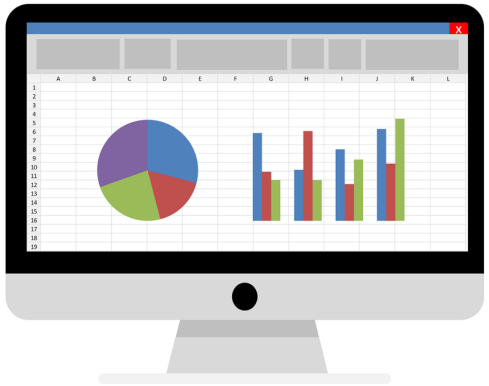
Variety

various data types/sources

Variety various data types

Structured data

Structured data concerns all data which can be stored in relational database in a table with rows and columns.



Excel

Unstructured data

Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database.

Video

Variety various data sources

1. Internet

2. Crowd sourcing
   Wikipedia, forums

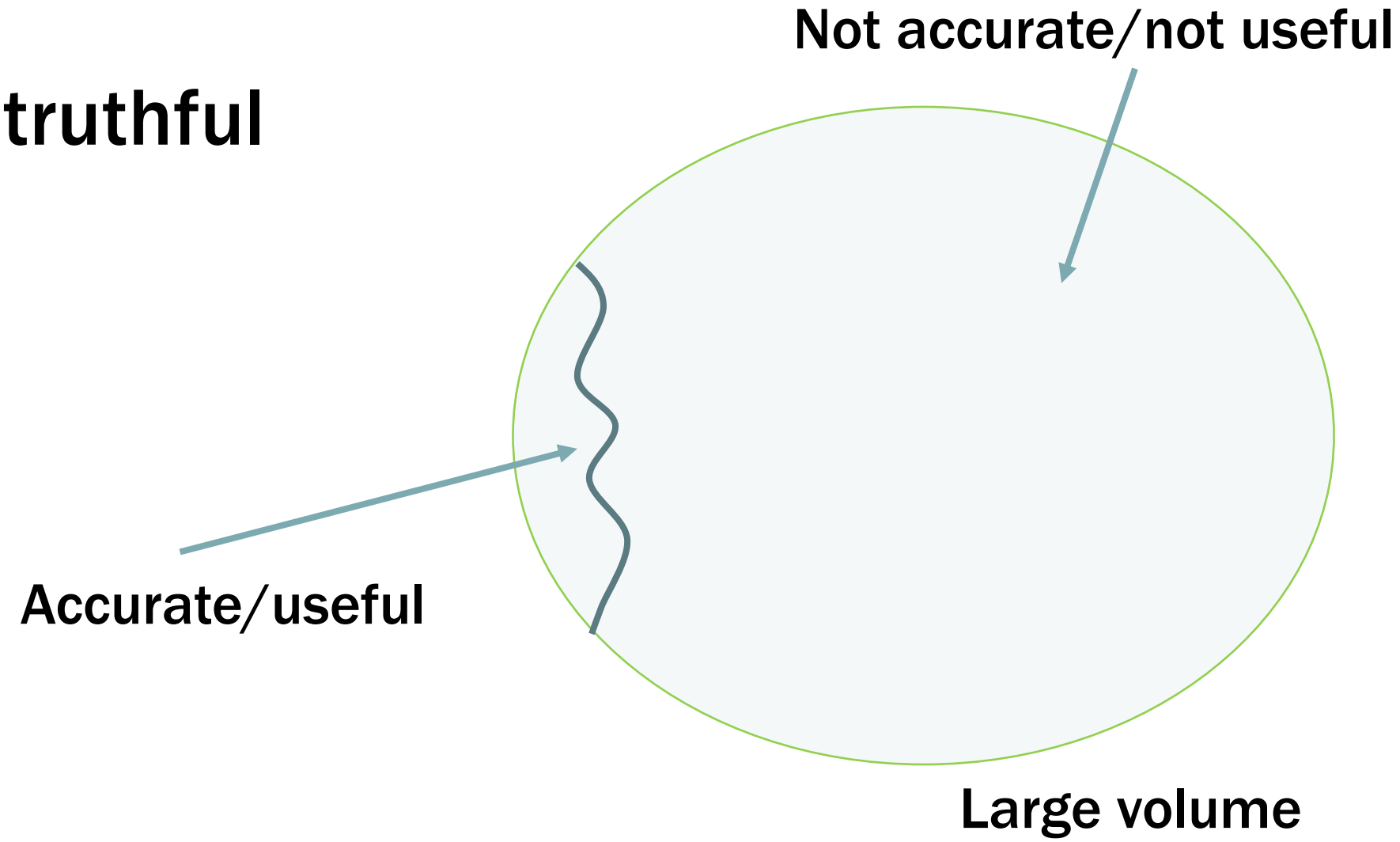
3. Social networks
   Facebook, instagram

4. Sensors
   Mobile phones, GPS

Image source: http://avtecmedia.com/images/blog/social-networking-logos.jpg

Veracity

accurate and truthful



Veracity

accurate and truthful

High veracity ≈ High data quality

Veracity

accurate and truthful

Statistical bias

Software bugs

Untrustworthy source

Veracity

accurate and truthful

Statistical bias

Software bugs

Untrustworthy source

Some data points are given more weightage than others.

Veracity

accurate and truthful

Statistical bias

Software bugs

Untrustworthy source

Data as software output are distorted.

Veracity

accurate and truthful

Statistical bias

Software bugs

Untrustworthy source

Can you name some trustworthy data sources?

Veracity accurate and truthful

Statistical bias

Some data points are given more weightage than others.

Software bugs

Data as software output are distorted.

Untrustworthy source

Can you name some trustworthy data sources?

Text books, Research papers, Professional magzines

Veracity accurate and truthful

Name ambiguity

Who is **Michael Jordan**?

Veracity accurate and truthful

In many people' mind:

Former NBA star Michael Jordan

Veracity accurate and truthful

Machine learning
people may think about

Prof. Michael Jordan

Great contributions in ML/AI

Veracity

accurate and truthful

Sentence ambiguity

Summary of veracity

Big data should have high veracity;
that usually means the data are highly accurate and trustworthy,
particularly

1. do not have statistics bias,
2. are not generated by software with bugs,
3. come from reliable data source,
4. have no/little data ambiguation

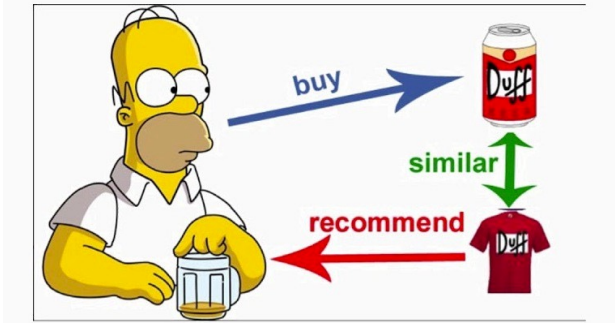
Value

Benefits from analyzing the data

Value

Benefits from analyzing the data

Recommendation



Retail Marketing

Health Care

Computer Vision

Discussions & Questions

We finish big data 5V's!

Next lecture:

Data models