Modern (Computational) Approaches to Big Data Analytics

CSC 576

Computer Science, University of Rochester
Instructor: Ji Liu

Big Data in Academy

- SIGKDD 2014 (program page, found 14 “big data”, 50+ “large scale”) http://www.kdd.org/kdd2014/program.html
- ICML 2014 (3 of 6 tutorials are about “big data”)

  http://icml.cc/2014/index/article/17.html

Big Data in Industry

From “linkedin”, I found

- 2,107 results for data scientist positions
- 865 results for Java programmer positions
- 436 results for c++ programmer positions

What is ``Big Data''? – A Mock from a professor of psychology and behavioral economics

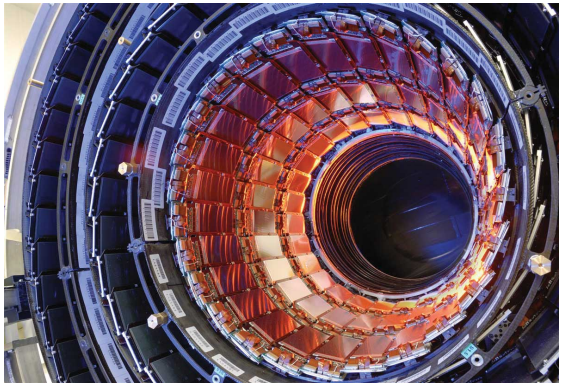
- Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it ---- Dan Ariely.

Big Data Every Where!

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Social Network

How ``Big''?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB per month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN’s Large Hydron Collider (LHC) generates 15 PB a year



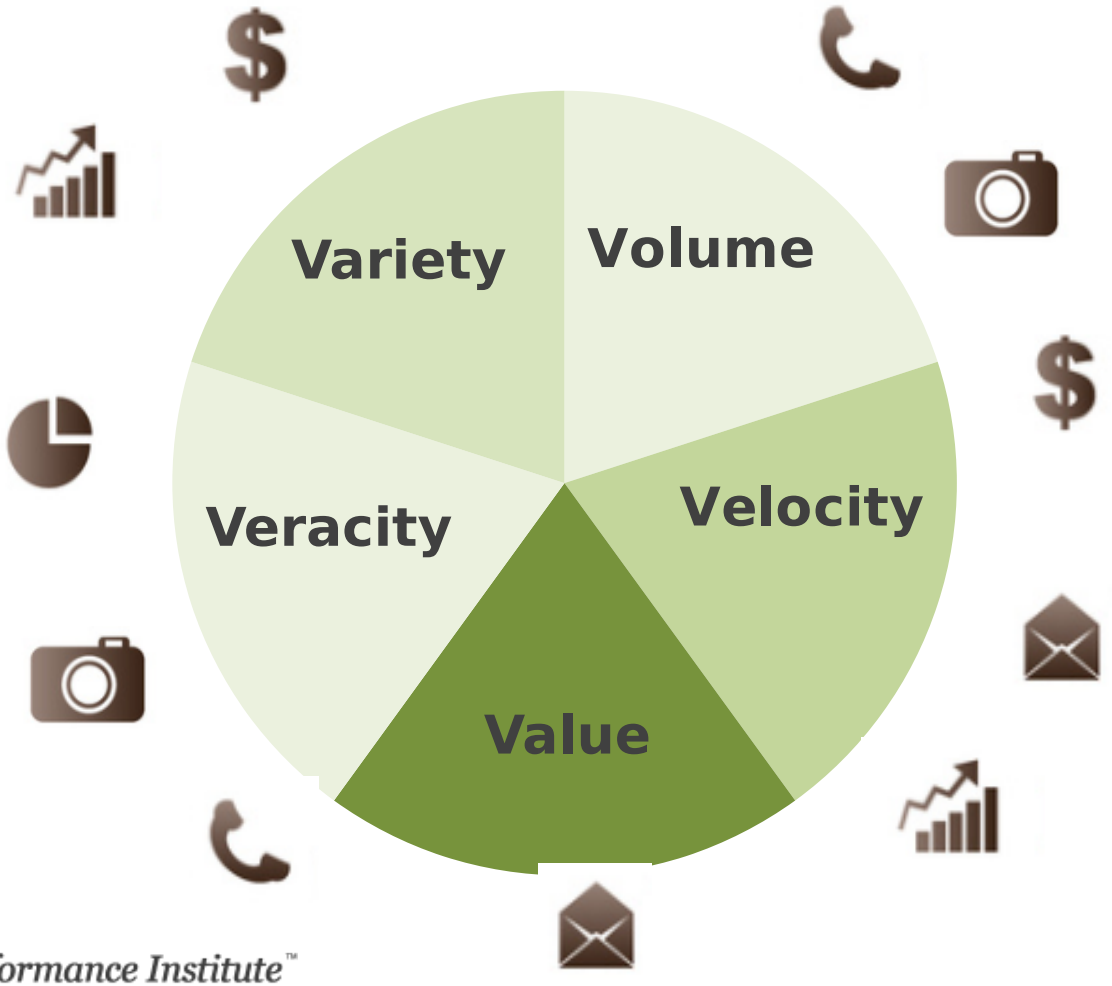
BIG DATA

VOLUME
DATA SIZE

VELOCITY
SPEED OF CHANGE

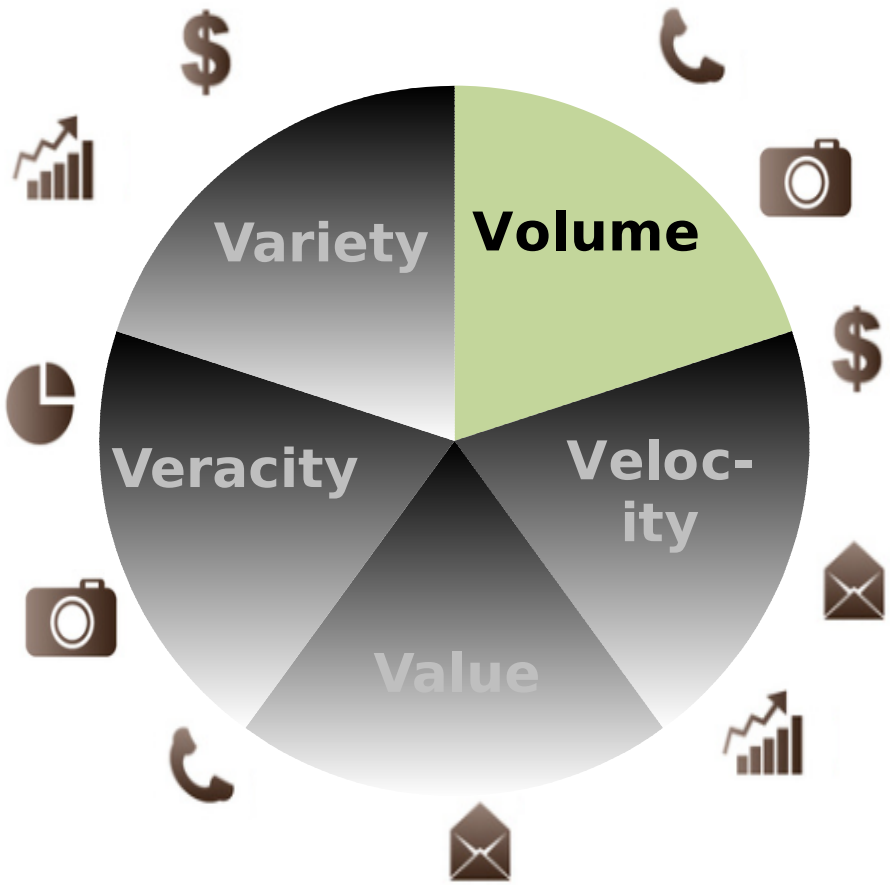
VARIETY
DIFFERENT FORMS OF DATA SOURCES

VERACITY
UNCERTAINTY OF DATA



Advanced Performance Institute™

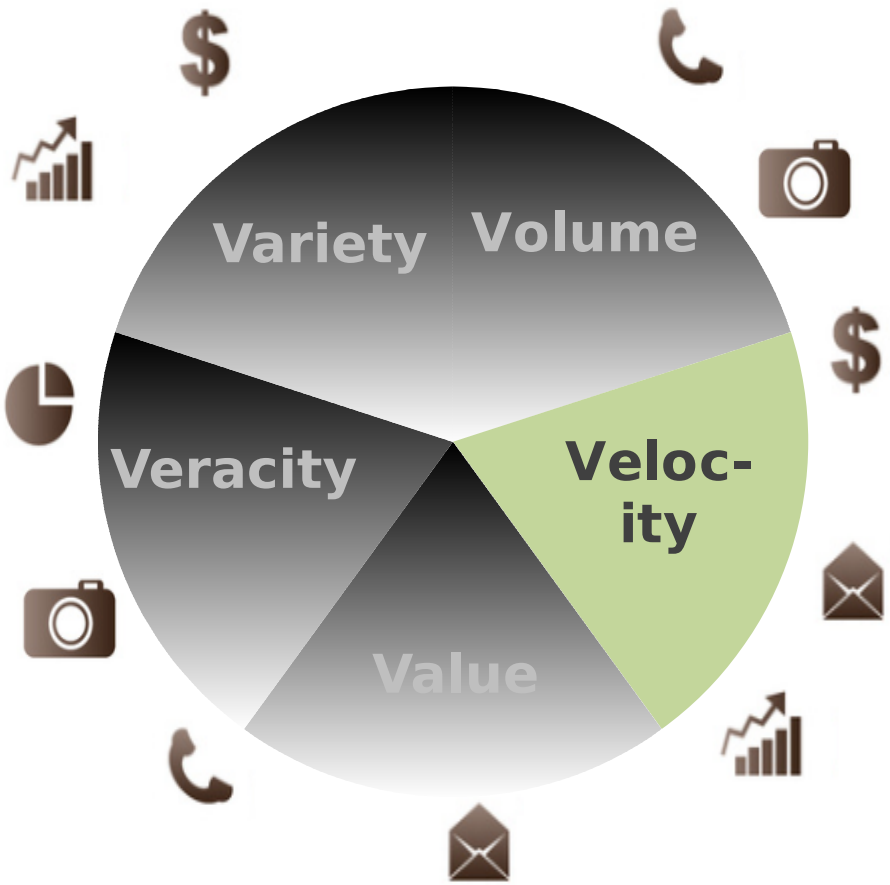
© 2014 Advanced Performance Institute, BWMC Ltd. All rights reserved.



Advanced Performance Institute™

© 2014 Advanced Performance Institute, BWMC Ltd. All rights reserved.

Volume refers to the vast amounts of data generated every second. We are not talking Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. This makes most data sets too large to store and analyse using traditional database technology. New big data tools use distributed systems so that we can store and analyse data across databases that are dotted around anywhere in the world.

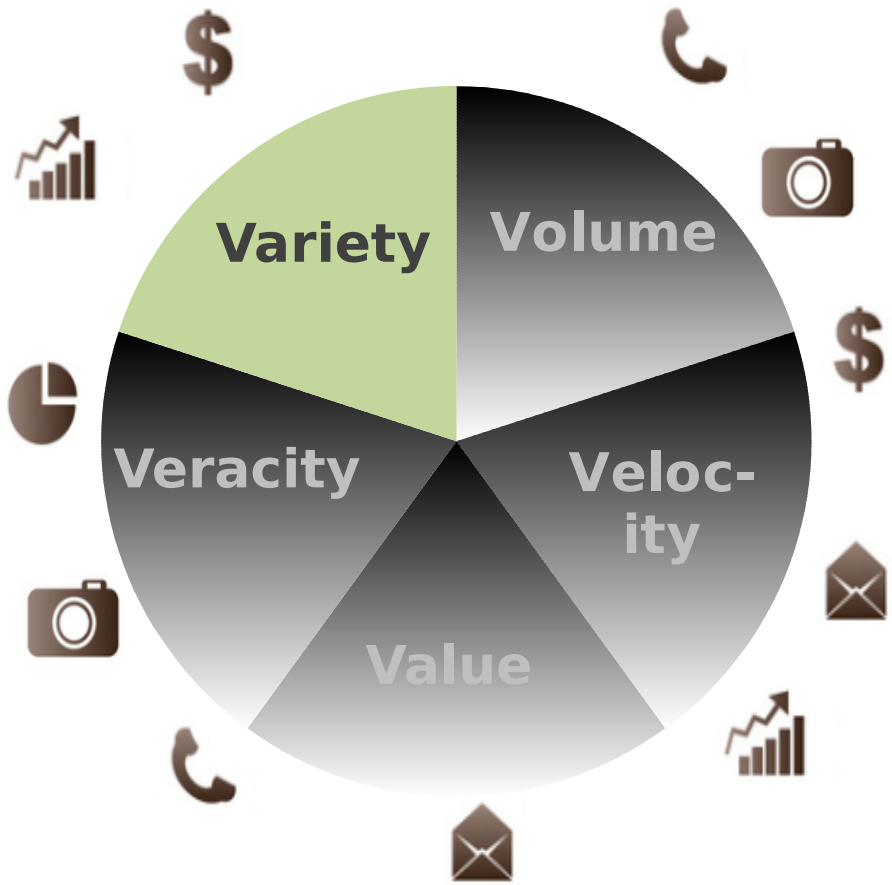


Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology allows us now to analyse the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

Advanced Performance Institute™

© 2014 Advanced Performance Institute, BWMC Ltd. All rights reserved.

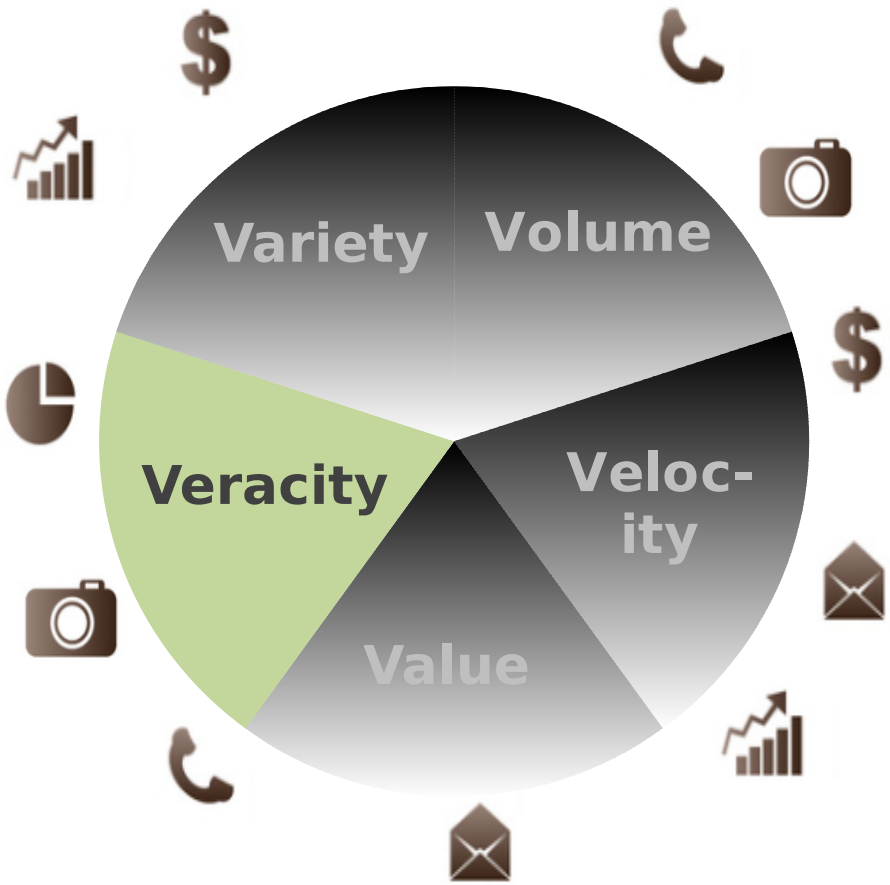
We see increasing variety of data types:



Advanced Performance Institute™

© 2014 Advanced Performance Institute, BWMC Ltd. All rights reserved.

Variety refers to the different types of data we can now use. In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world's data is unstructured (text, images, video, voice, etc.) With big data technology we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

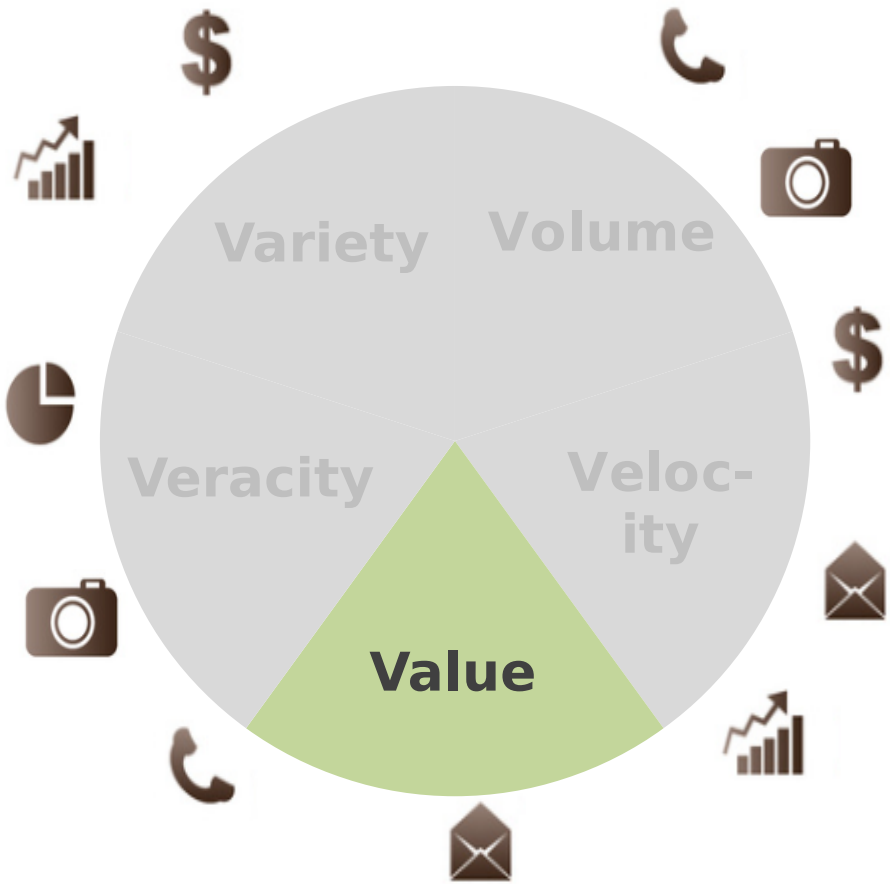


Veracity refers to the messiness or trustworthiness of the data. With many forms of big data quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but technology now allows us to work with this type of data.

Advanced Performance Institute™

© 2014 Advanced Performance Institute, BWMC Ltd. All rights reserved.

Value – The most important V of all!



Then there is another V to take into account when looking at Big Data: Value!

Having access to big data is no good unless we can turn it into value.

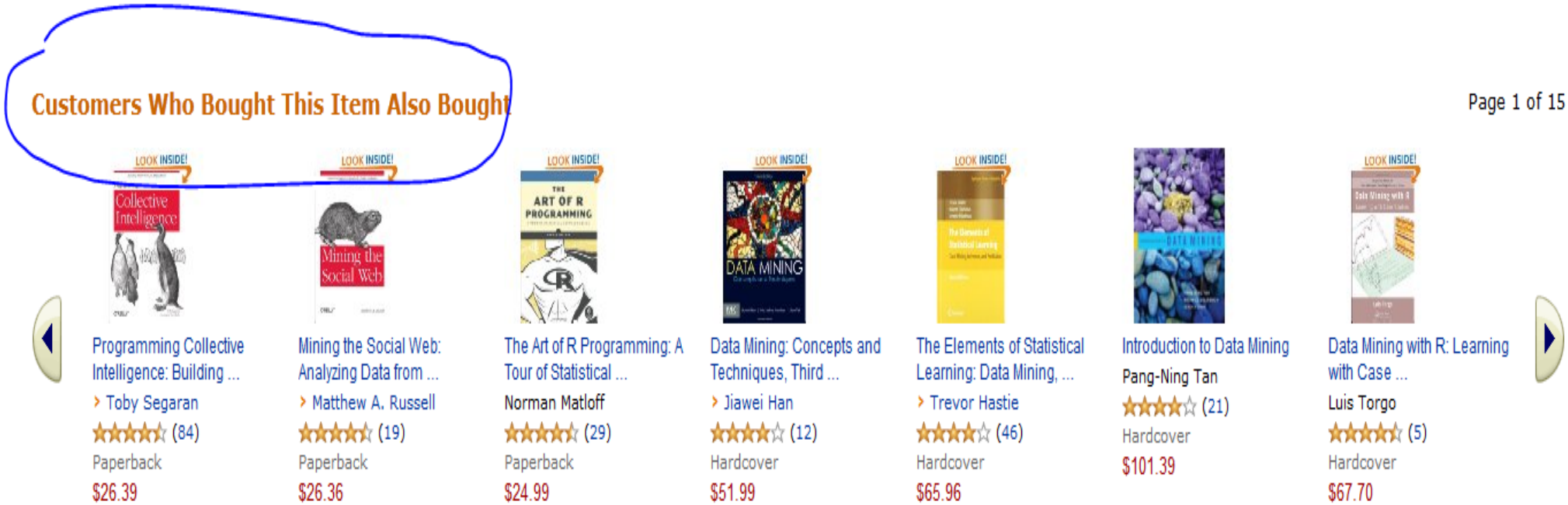
Companies are starting to generate amazing value from their big data.

Advanced Performance Institute™

© 2014 Advanced Performance Institute, BWMC Ltd. All rights reserved.

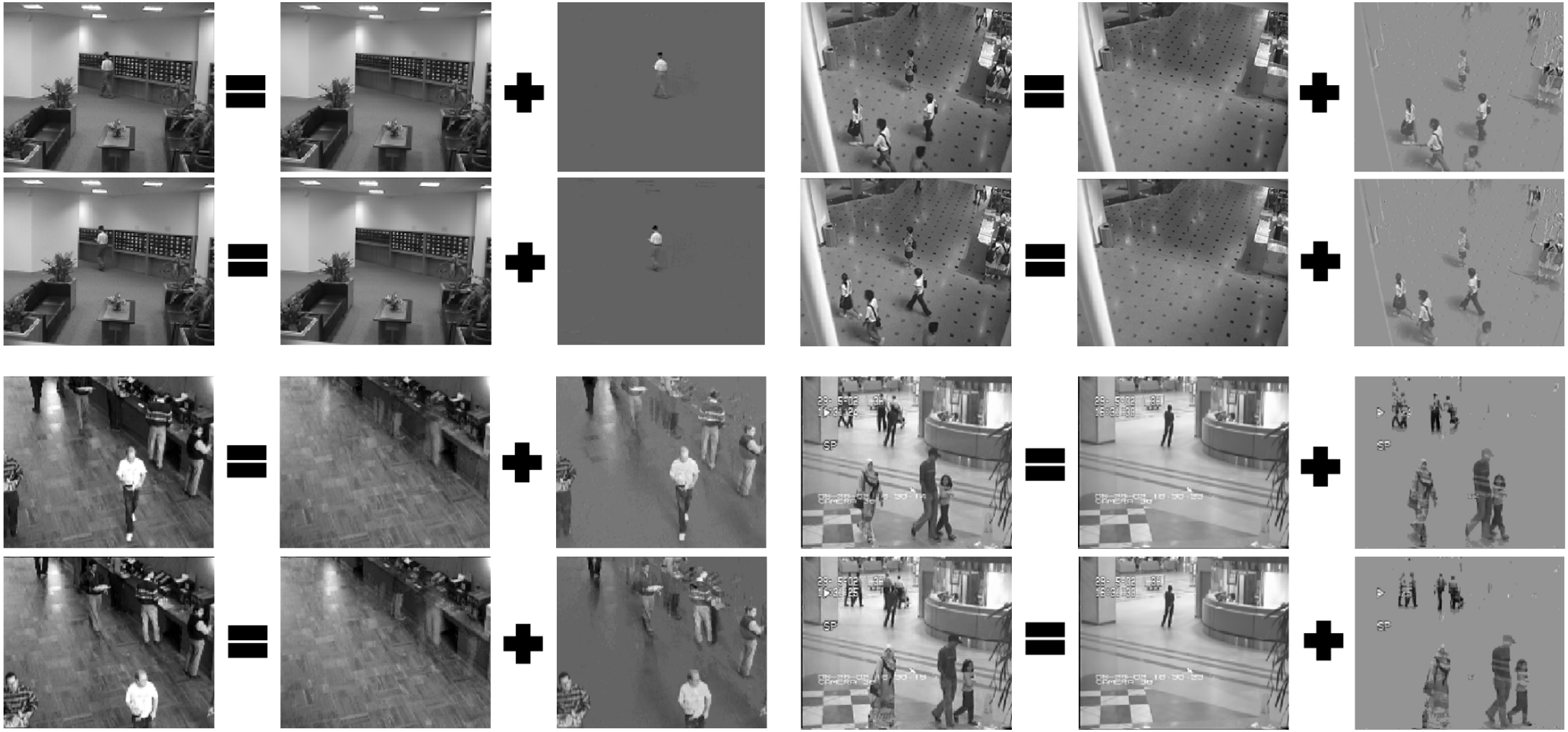
Recommendation System – Example 1

Recommendation System – Example 2



Video Analysis

Video Surveillance



Steps of Data Analysis

- Pose a problem
- Collect data – raw and dirty data
- Pre-process data (like extract feature) – clean data
- Design mathematical model (formulation)
- Find a solution
- Evaluation