Technology and Application of Big Data

Qing LIAO(廖清)

School of Computer Science and Technology

HIT

Course Details

- Instructor:
  - Qing LIAO, liaoqing@hit.edu.cn
  - Rm. 303B, Building C
  - Office hours: by appointment
- Course web site:
  - liaoqing.me
- Reference books/materials:
  - Big data courses from University of California
  - Book: BIG DATA: A Revolution That Will Transform How We Live, Work, and Think
  - Papers
- Grading Scheme:
  - Paper Report 30%
  - Final Exam 70%
- Exam:
  - 21st July(Friday), 14:00-16:00, A502

Course Details

Exercises Lectures External Links

The Deep Learning textbook is a resource intended to help students and practitioners enter the field of machine learning in general and deep learning in particular. The online version of the book is now complete and will remain available online for free.

The deep learning textbook can now be pre-ordered on Amazon. Pre-orders should ship on December 16, 2016.

For up to date announcements, join our mailing list.

Citing the book

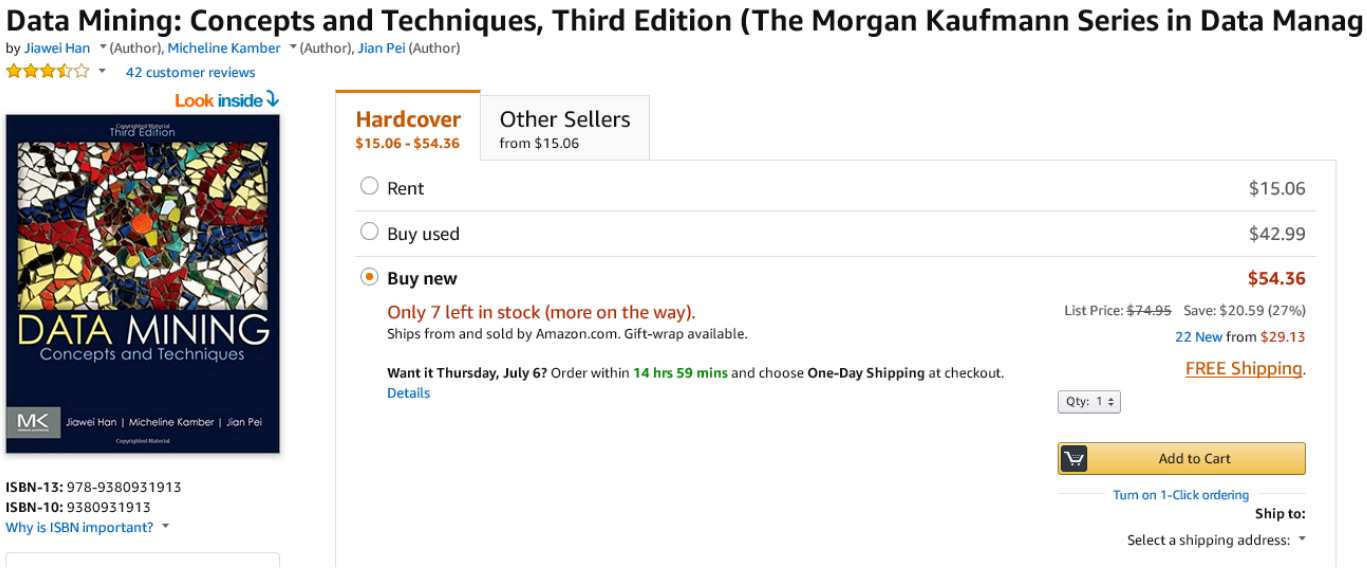
To cite this book, please use this bibtex entry:

```
@book{Goodfellow-et-al-2016,
    title={Deep Learning},
    author={Ian Goodfellow and Yoshua Bengio and Aaron Courville},
    publisher={MIT Press},
    note={\url{http://www.deeplearningbook.org}},
    year={2016}
}
```

Errata in published editions

Deep Learning

- Table of Contents
- Acknowledgements
- Notation
- 1 Introduction
- Part I: Applied Math and Machine Learning Basics
  - 2 Linear Algebra
  - 3 Probability and Information Theory
  - 4 Numerical Computation
  - 5 Machine Learning Basics
- Part II: Modern Practical Deep Networks



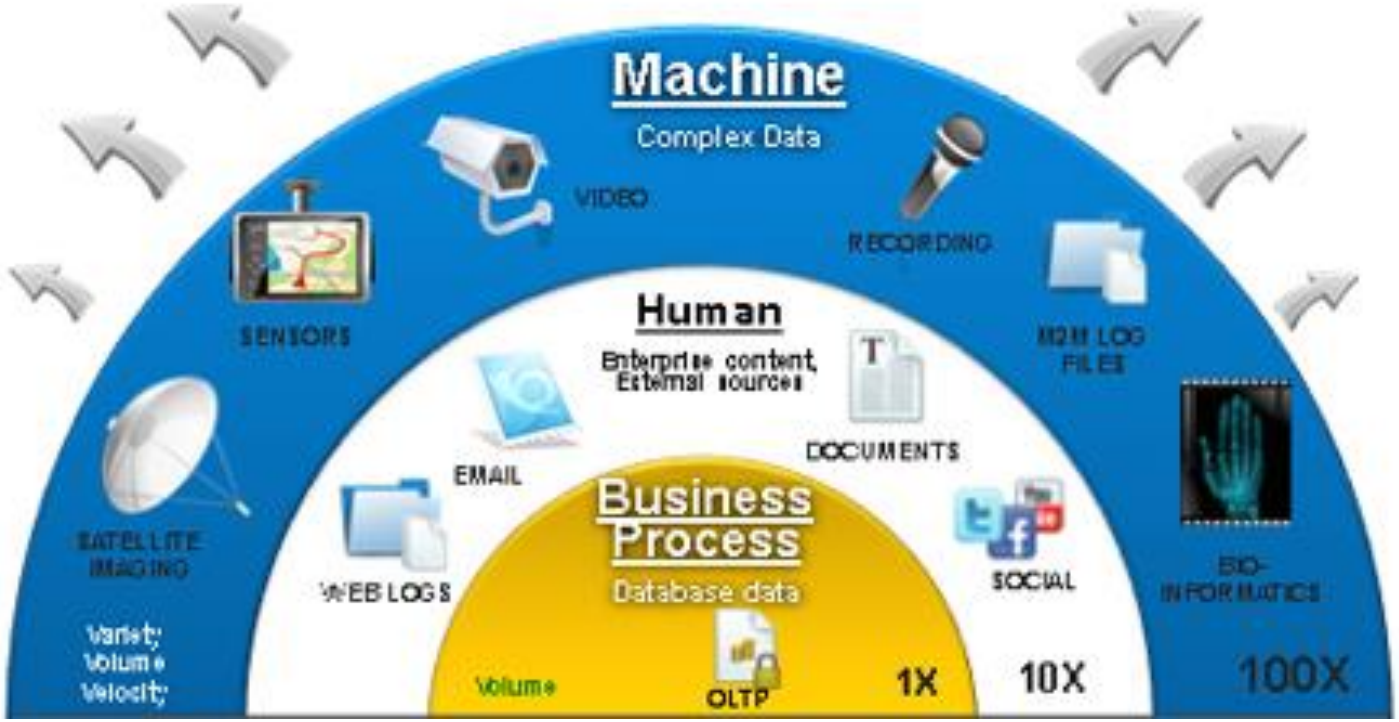
http://www.deeplearningbook.org/

What You Learnt: Overview

- Topics:
  1) Introduction of Big Data
  2) Characterizes of Big Data
  3) How to Get Value from Big Data
  4) Technologies of Big Data
  5) Applications of Big Data
- Prerequisites
  - Statistics and Probability would help
    - But not necessary
  - Machine Learning would help
    - But not necessary

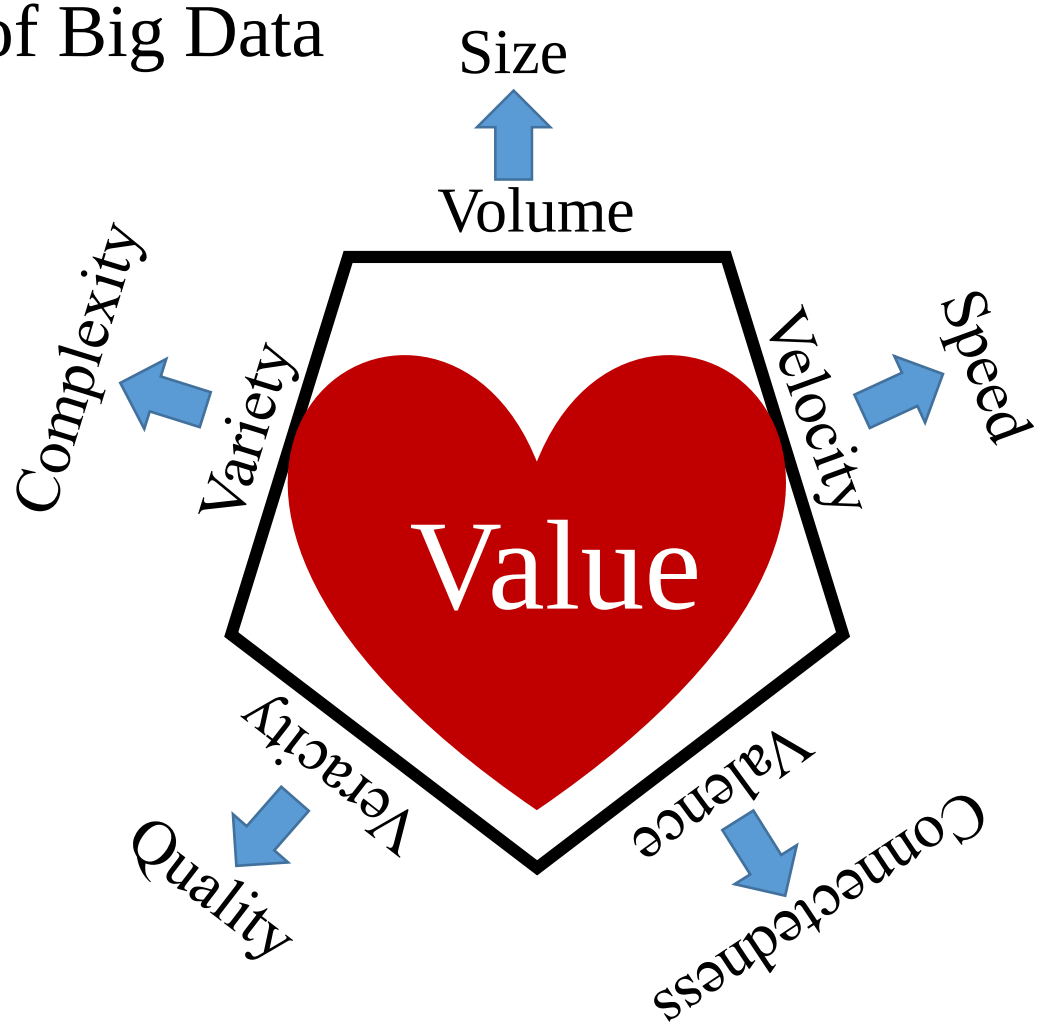
Previous Section

- Where Does Big Data Come From?



Previous Section

- Characterizes of Big Data



How to Get Value from Big Data

- Steps in the Data Science Process

Step 1: Acquire Data

Identify data sets

Retrieve data

How to Get Value from Big Data

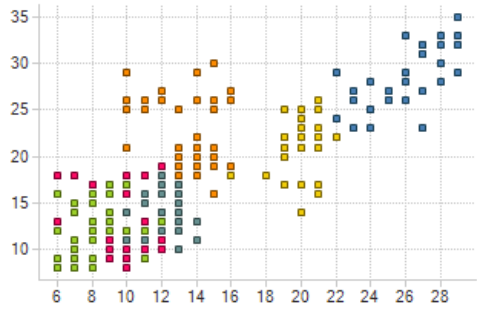
- Steps in the Data Science Process

Step 2: Prepare Data

Step 2-A: Explore

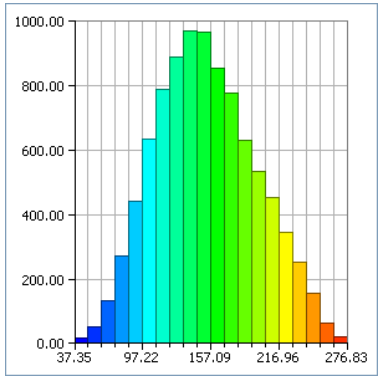
Step 2-B: Pre-process

Step 2-A: Explore Data



Understand nature of data

Preliminary analysis

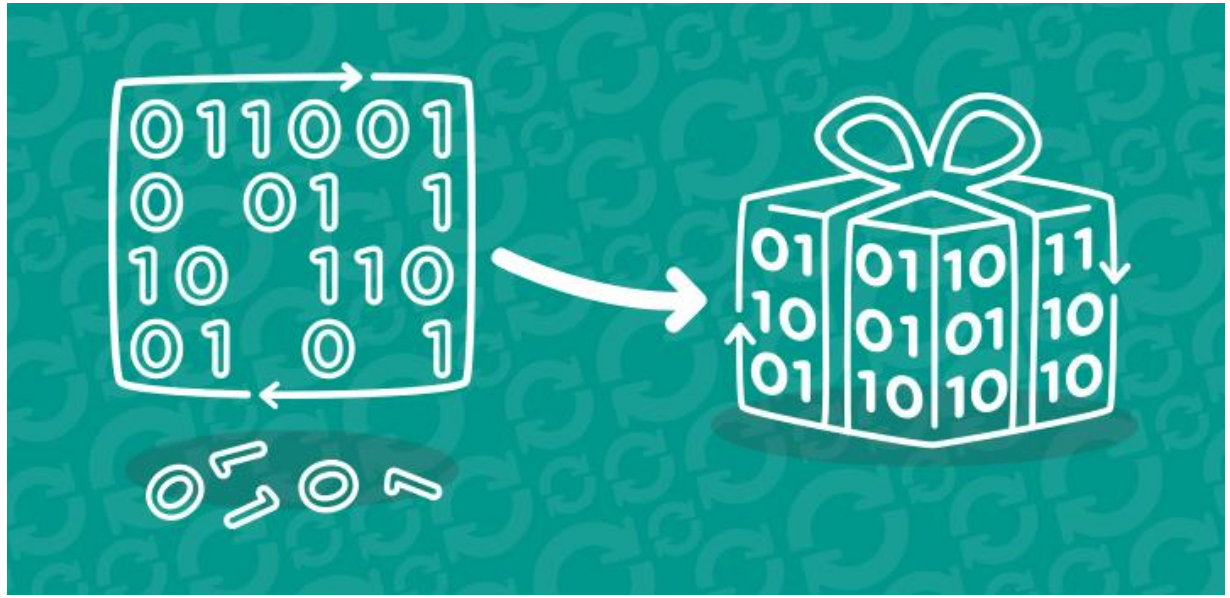


How to Get Value from Big Data

- Steps in the Data Science Process

Acquire → Prepare → Analyze → Report → ACT

Step 2-B: Pre-process Data



Clean

Integrate

Package

How to Get Value from Big Data

- Steps in the Data Science Process

Acquire → Prepare → Analyze → Report → ACT

Step 3: Analyze Data

Image courtesy of KromKrathog / FreeDigitalPhotos.net

Select analytical techniques

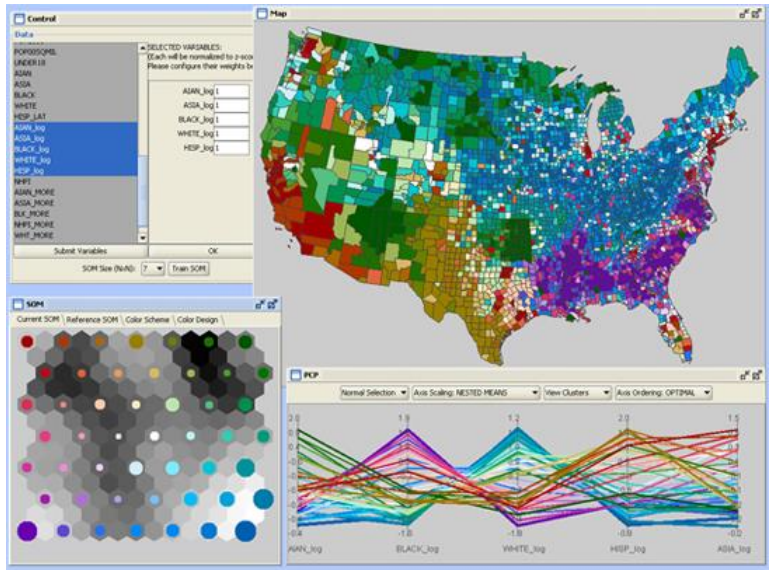
Build models

How to Get Value from Big Data

- Steps in the Data Science Process

Acquire | Prepare | Analyze | Report | ACT

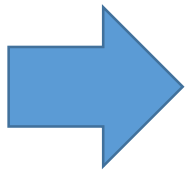
Step 4: Communicate Results



How to Get Value from Big Data

- Steps in the Data Science Process

Step 5: Apply Results



How to Get Value from Big Data

- Step1: Acquiring Data

Big Data Engineering Computational Big Data Science

Where's the data?

Identify suitable data

Acquire all available data

Data comes from many places

…with many ways to access it

How to Get Value from Big Data

- Step1: Acquiring Data

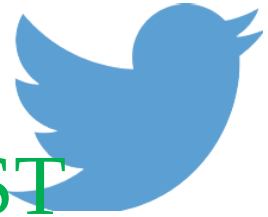
Historical weather

SQL

Current weather

WebSocket

Real-time tweets near fires



REST

How to Get Value from Big Data

- Step1: Acquiring Data

How to Get Value from Big Data

- Step 2: Prepare Data - Explore

Step 2: Prepare Data

Step 2-A: Explore

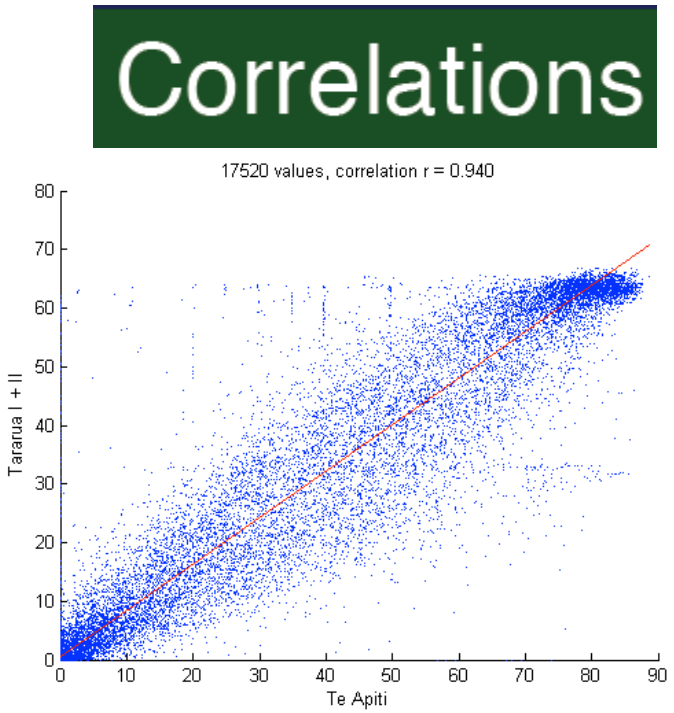
Step 2-B: Pre-process

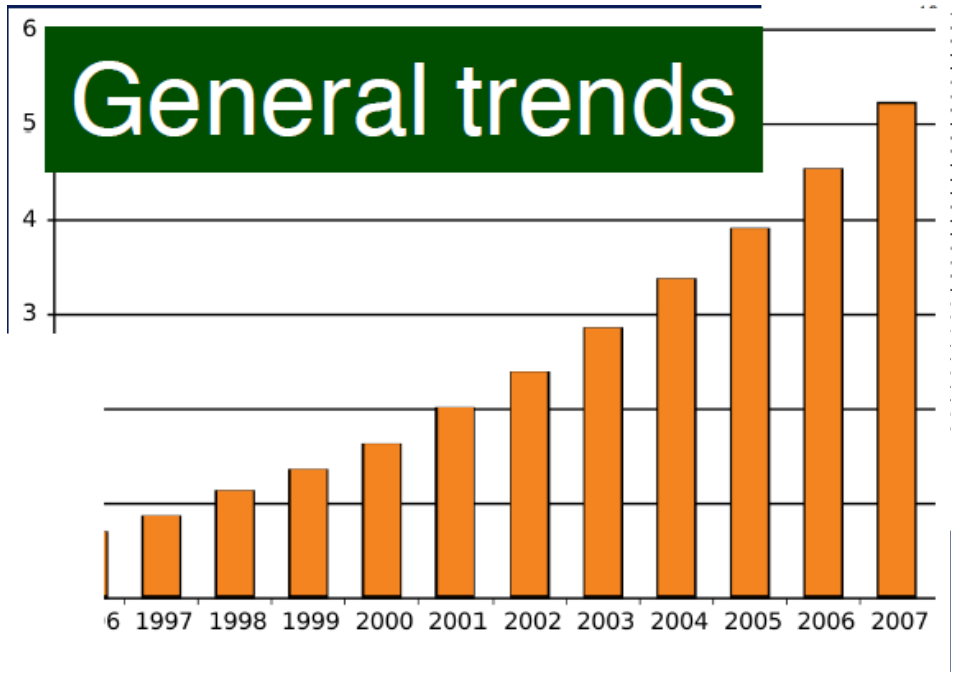
Why Explore?

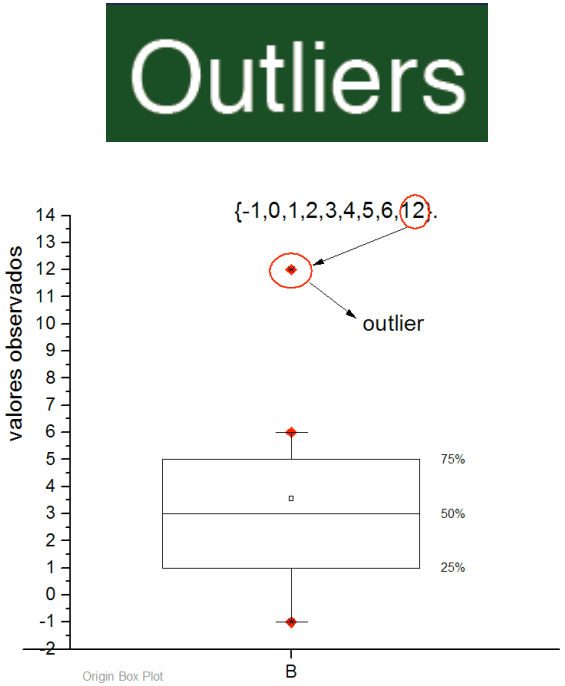
Goal: Understand your data

How to Get Value from Big Data

- Step 2: Prepare Data - Explore

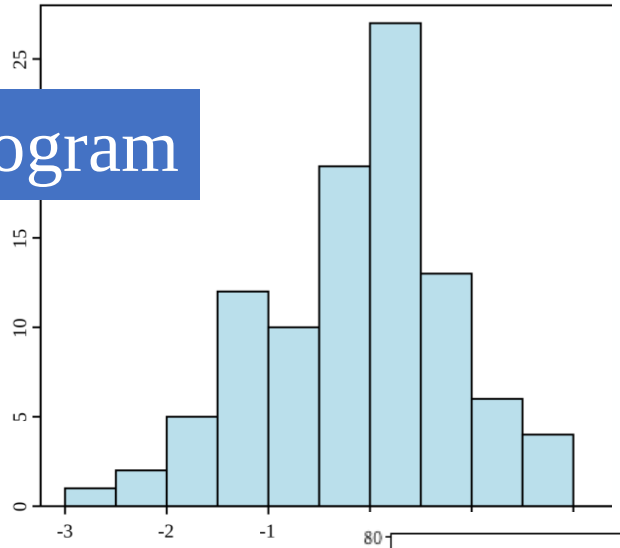




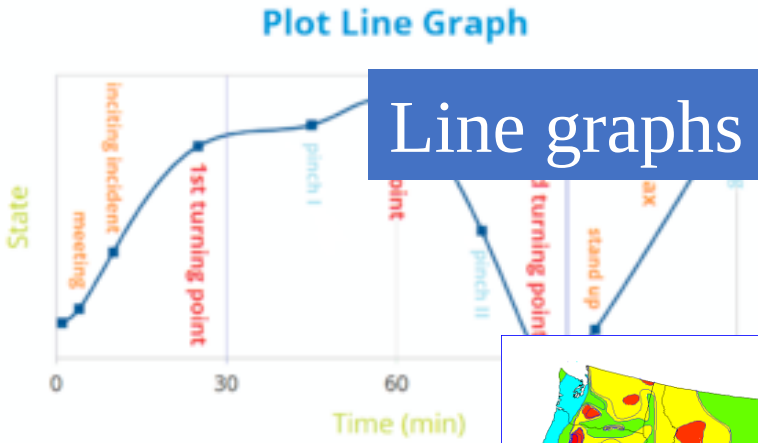


How to Get Value from Big Data

- Step 2: Prepare Data - Explore

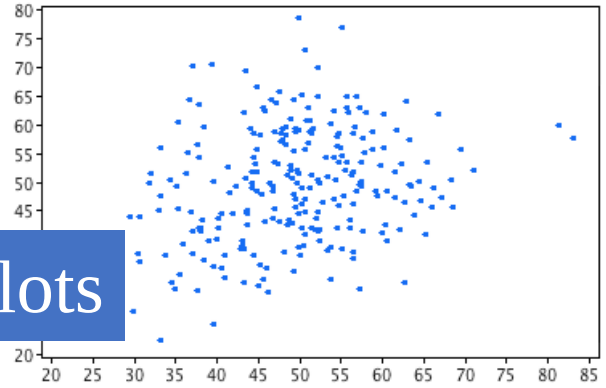


Histogram

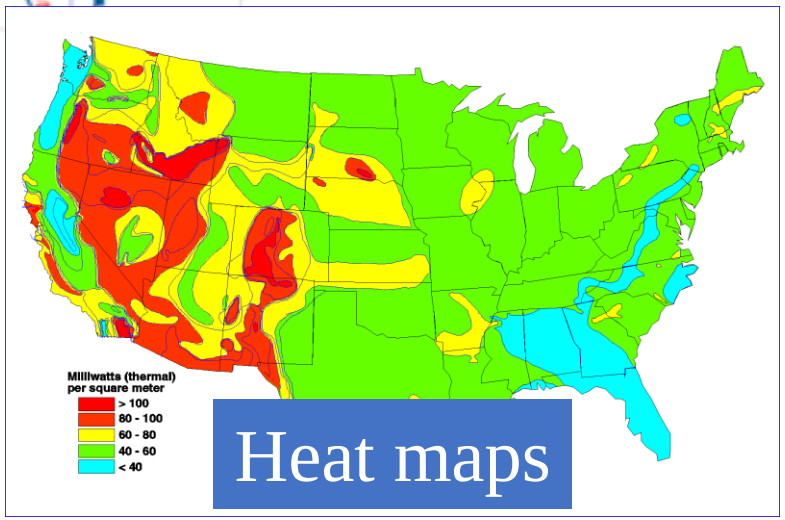


Line graphs

Visualize Your Data



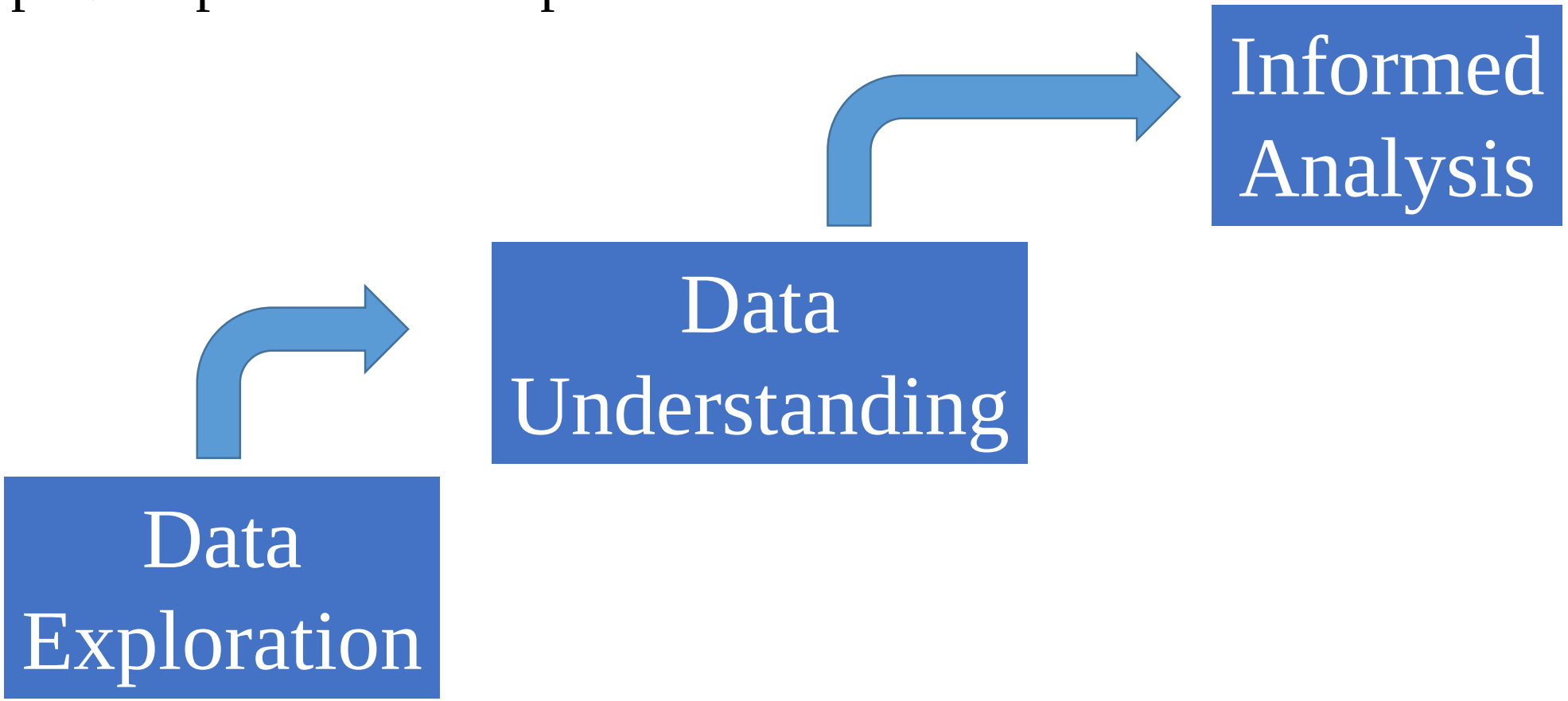
Scatter plots



Heat maps

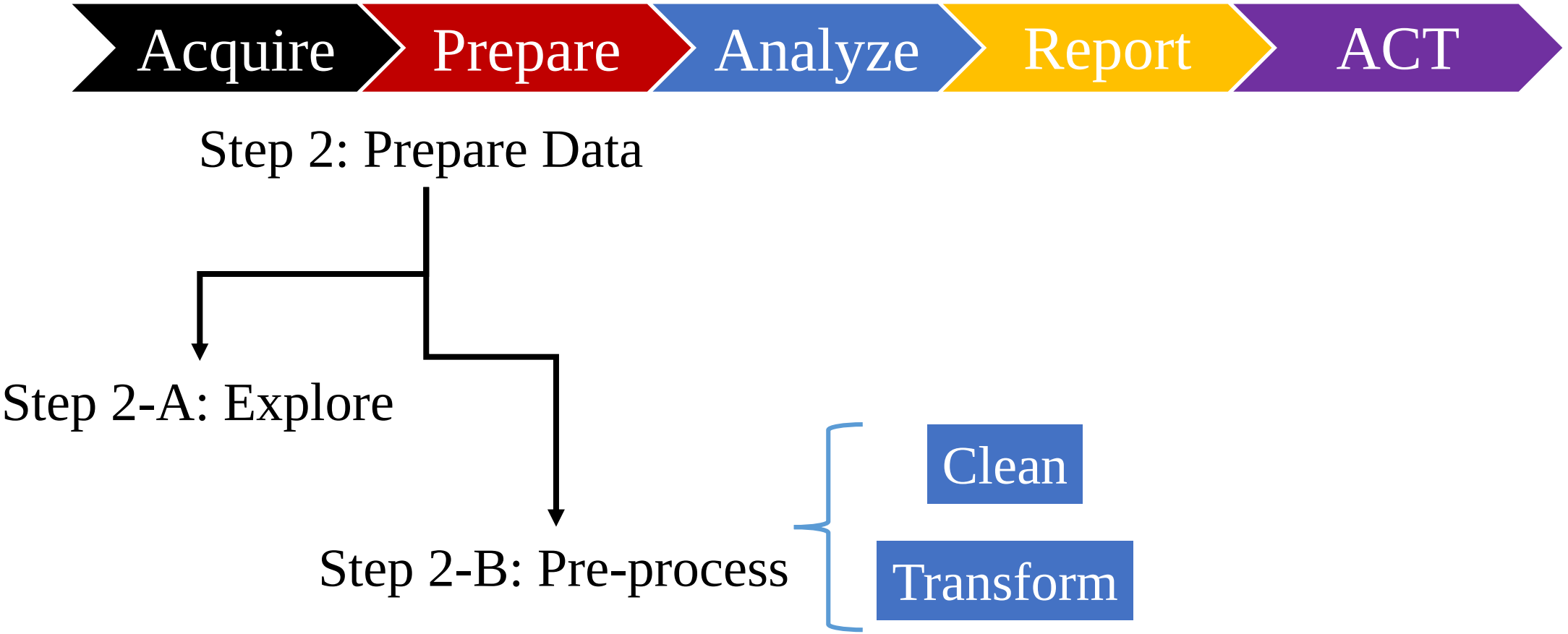
How to Get Value from Big Data

- Step 2: Prepare Data - Explore



How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process



How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process

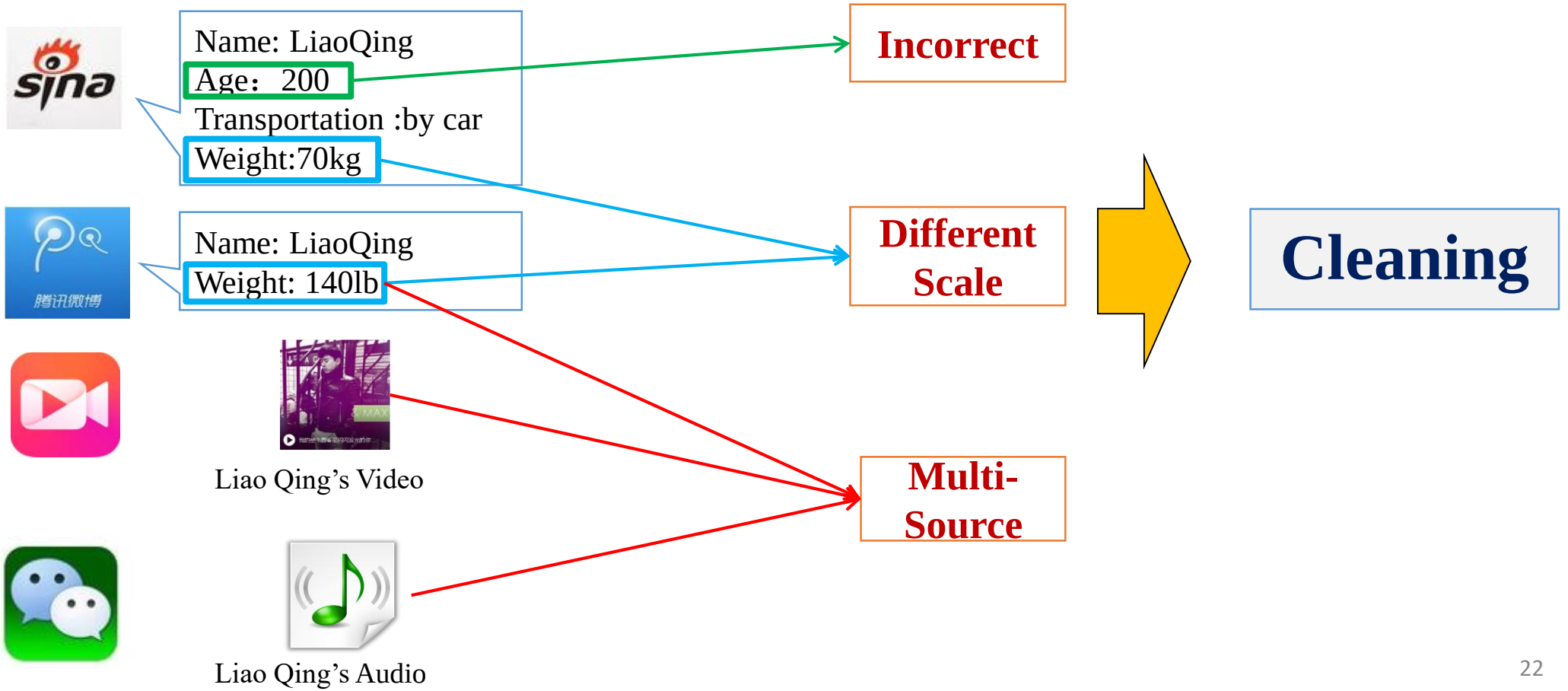
Data Quality Issues

Real-world data is messy!

- Inconsistent values
- Duplicate records
- Missing values
- Invalid data
- Outliers

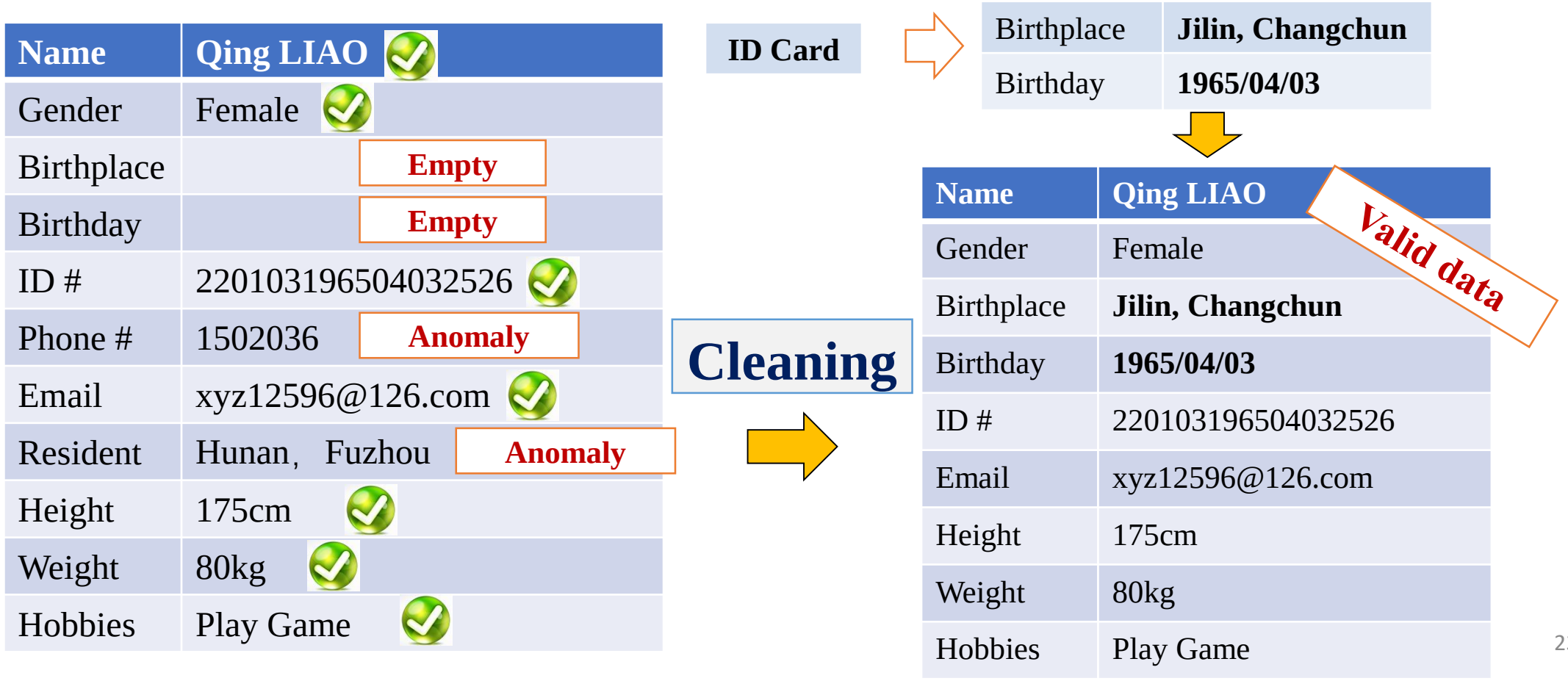
How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process



How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process



How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process

Data Source 1	**(Name, family phone #, home address, office phone # , office address)**
Data Source 2	**(Family name, given name, nick name, phone #, address, QQ #)**
Data Source 3	**a: (id, name); b: (id, Private phone #, office phone #) no address**
Data Source 4	**(Family name, given name, Private phone #, , home address)**
Data Source 5	**(Nick name, name, zip code、 phone #, City ， street)**

Different Format

Target Format **(Nick name, name, phone #, address, QQ #)**

Data Source 1	**(~~Nick name~~, name, phone #, address, ~~QQ #~~)**
Data Source 2	**(Nick name, name, phone #, address, QQ #)**
Data Source 3	**(~~Nick name~~, name, phone #, address, ~~QQ #~~)**
Data Source 4	**(~~Nick name~~, name, phone #, address, ~~QQ #~~)**
Data Source 5	**(Nick name, name, phone #, address, ~~QQ #~~)**

Pattern Alignment

How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process

Record Association

Web Page

Table Format

Category	Camera
Brand	Sony
Model	DSC-H300
Type	Telephoto Lens
Pixels	21 Million photo pixel
Color	Black
Feature	35X Telephoto Lens

How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process

Data Munging

Dimensionality Reduction

Data Manipulation

Transformation

Feature Selection

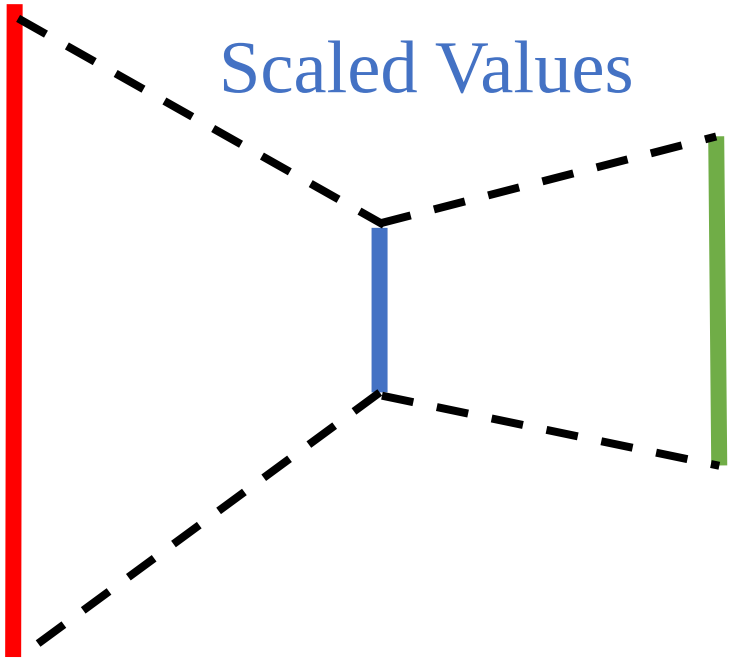
Scaling

How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process

Scaling

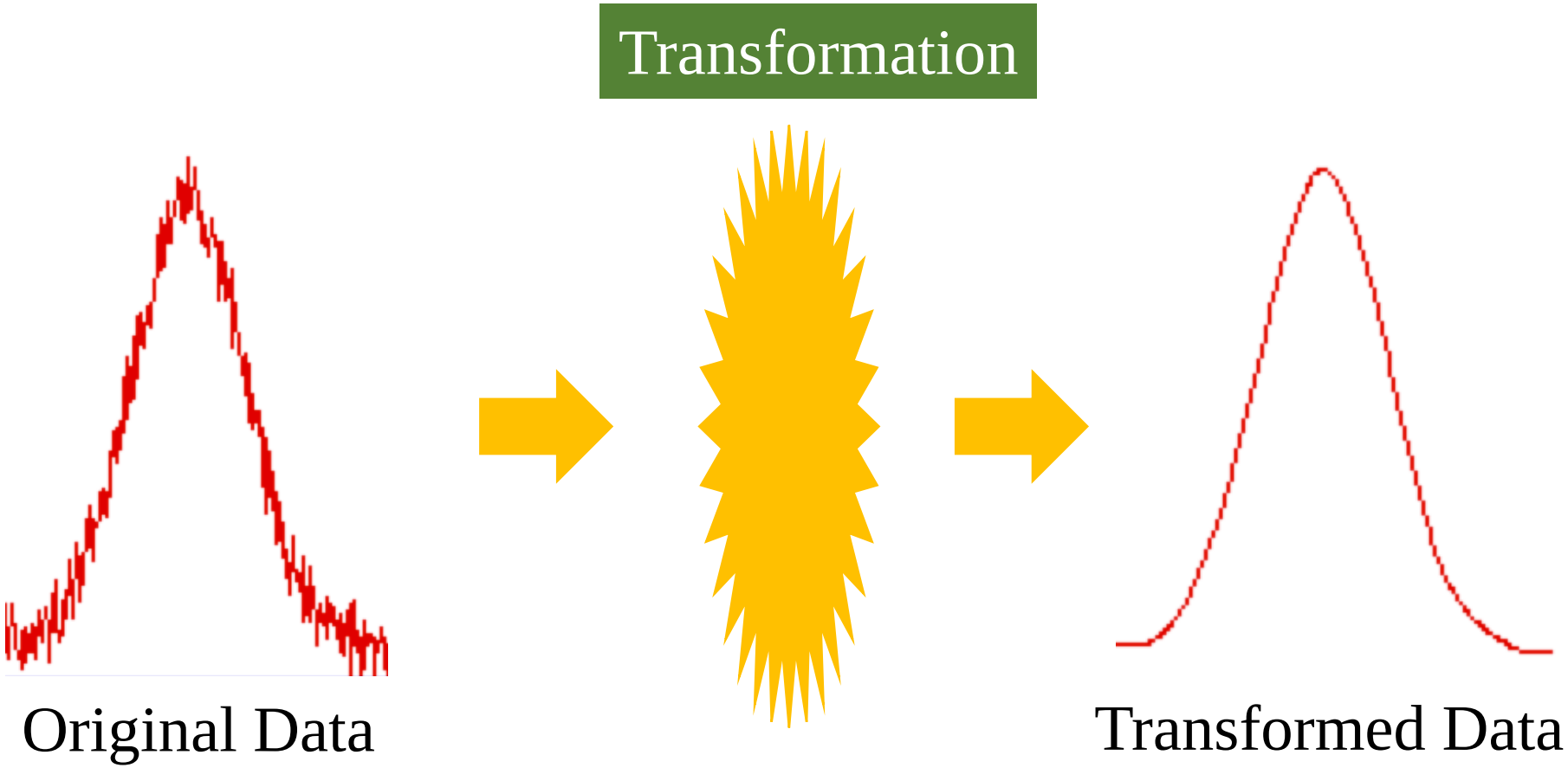
KG



Pounds

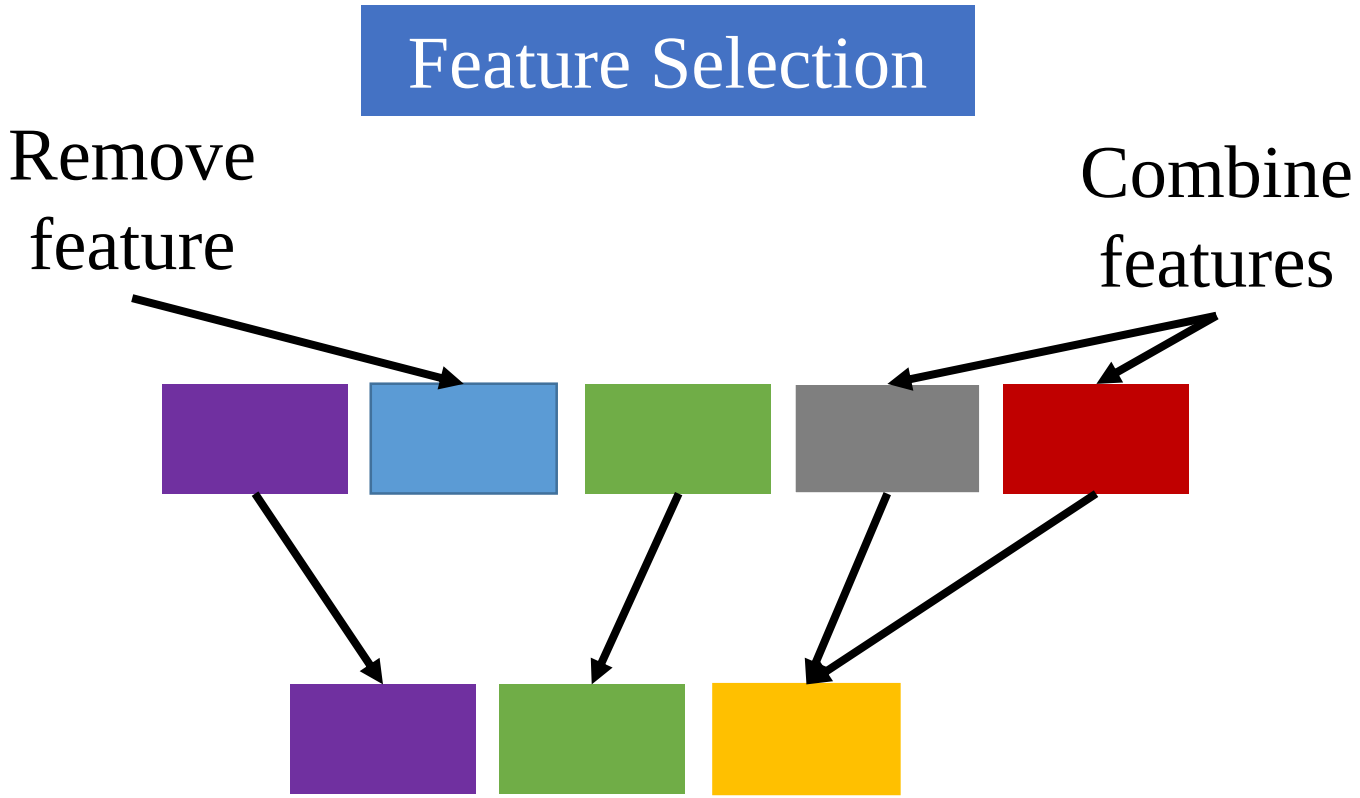
How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process



How to Get Value from Big Data

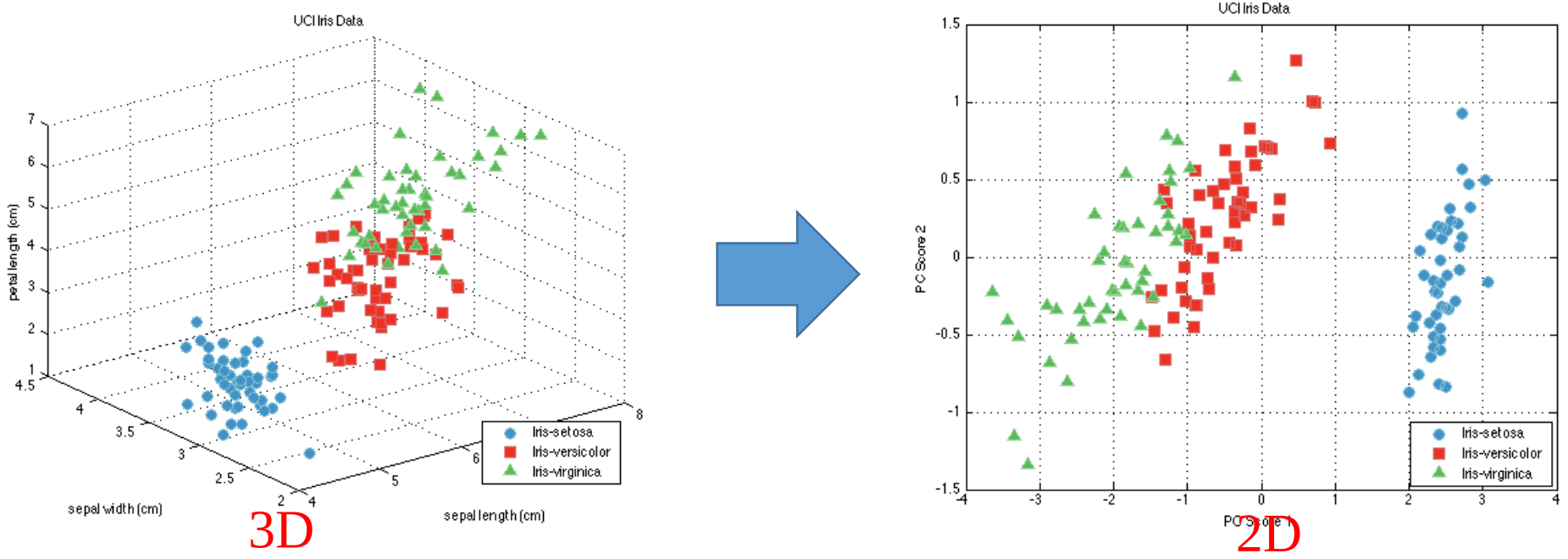
- Step 2: Prepare Data - Pre-process



How to Get Value from Big Data

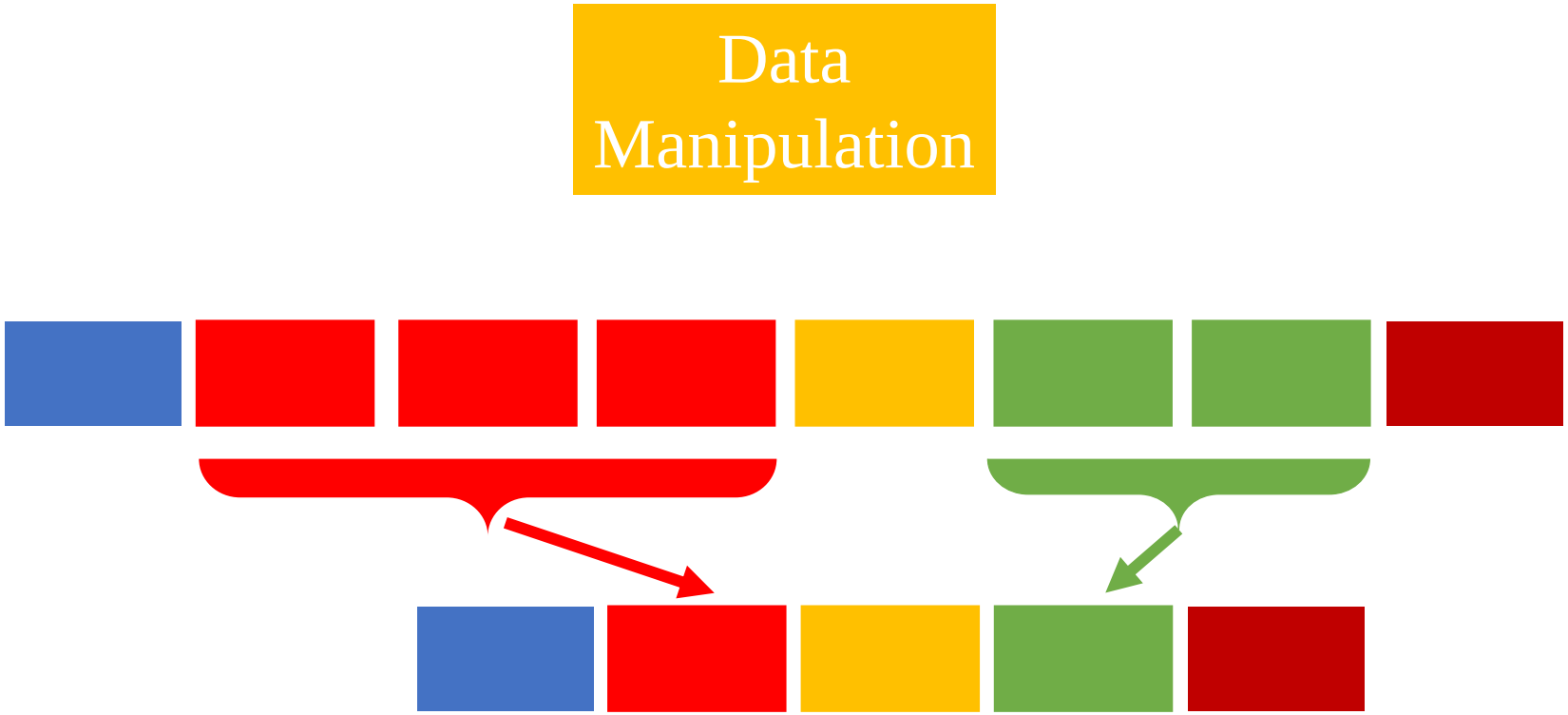
- Step 2: Prepare Data - Pre-process

Dimensionality Reduction



How to Get Value from Big Data

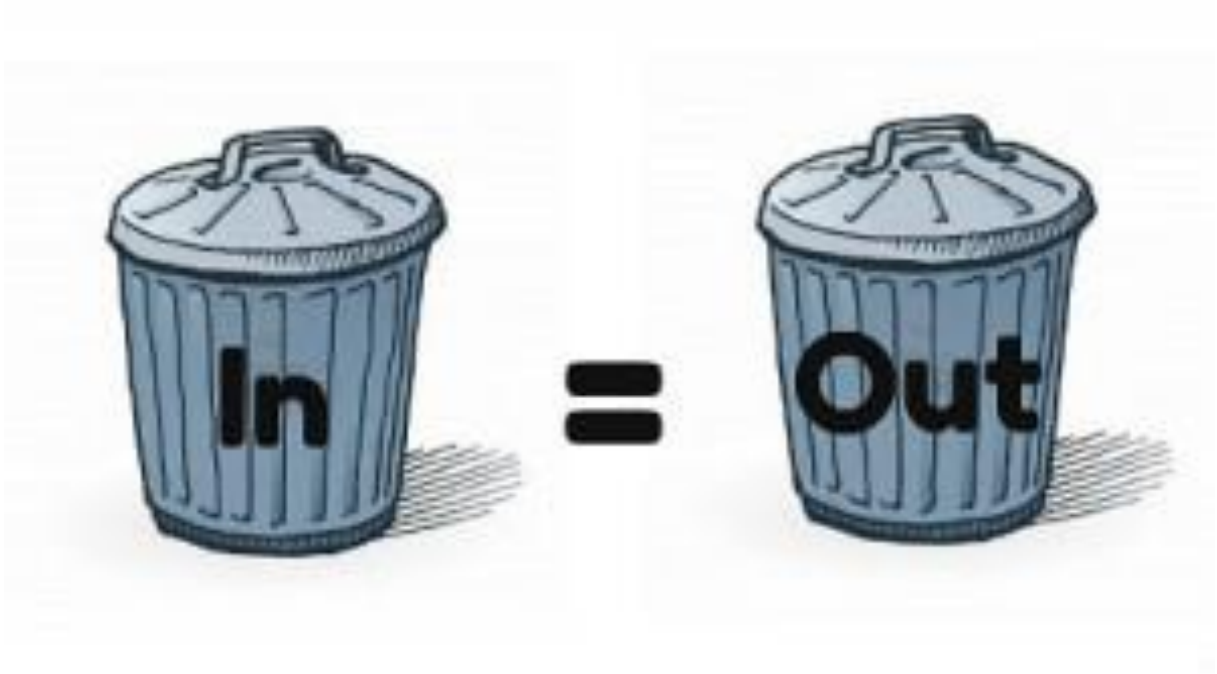
- Step 2: Prepare Data - Pre-process



How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process

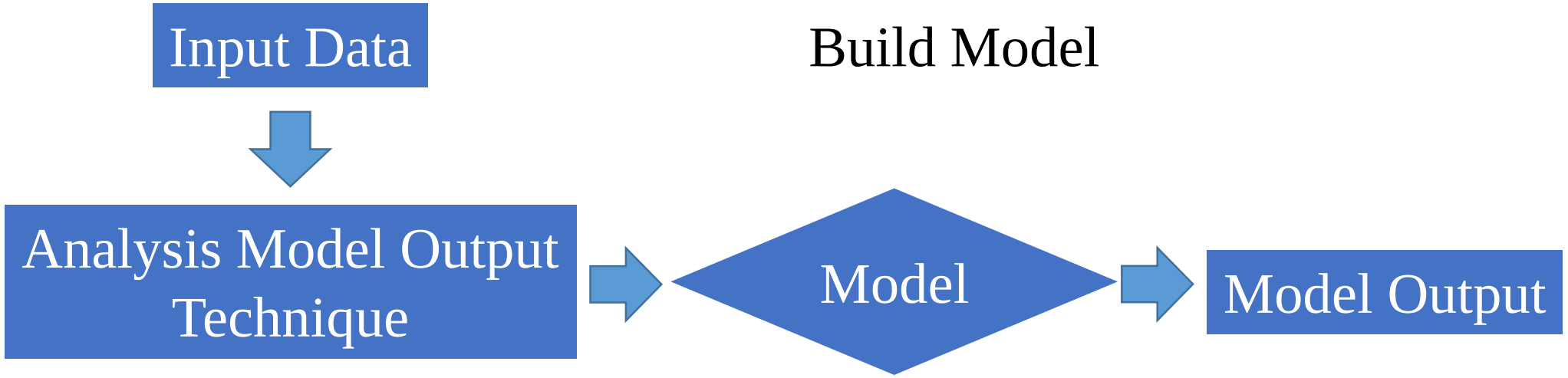
Garbage in = Garbage out



Data preparation is very important for meaningful analysis!

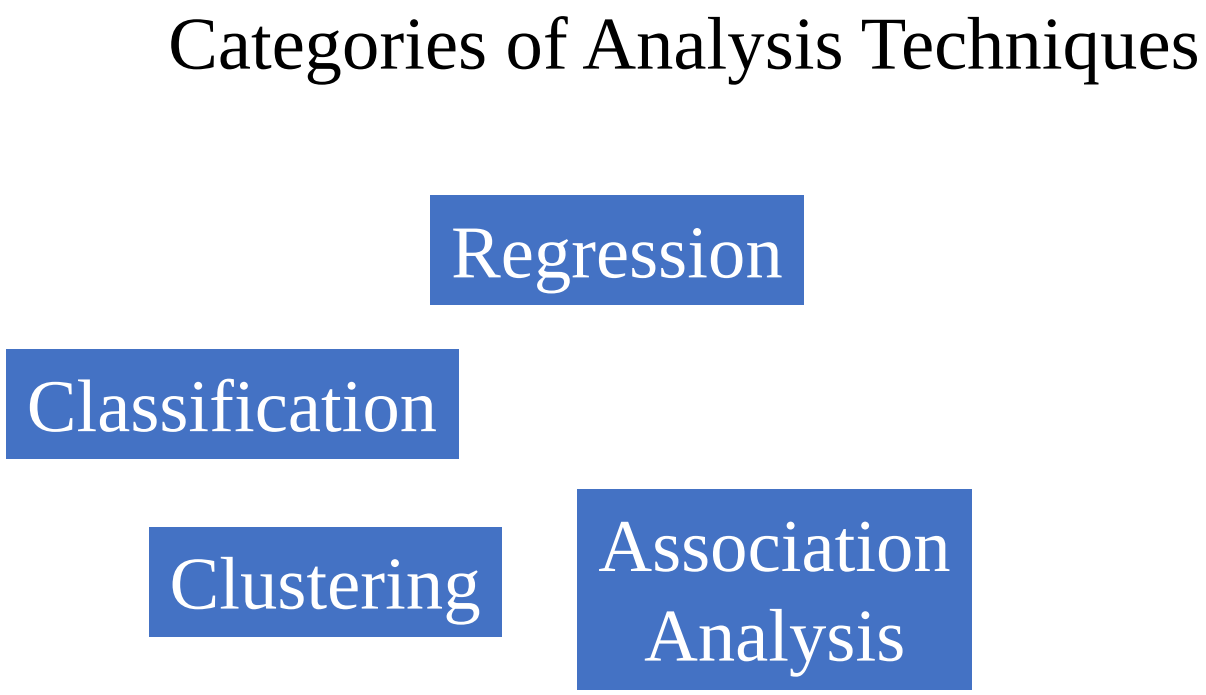
How to Get Value from Big Data

- Step 3: Analyze Data



How to Get Value from Big Data

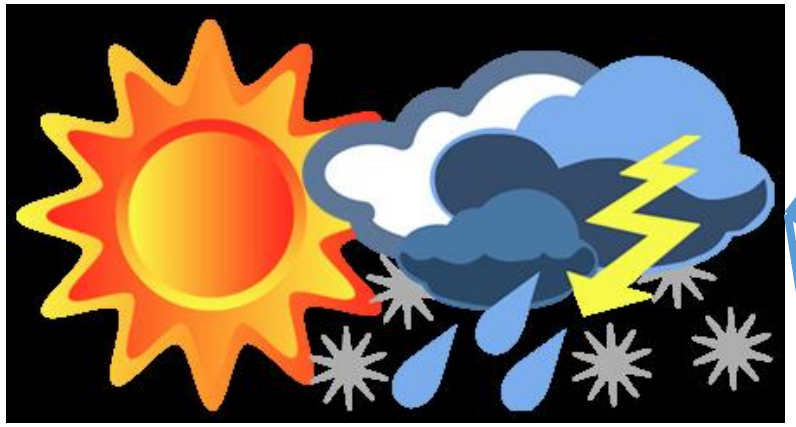
- Step 3: Analyze Data



How to Get Value from Big Data

- Step 3: Analyze Data

Classification



Rainy

Sunny

Cloudy

Windy

Goal: Predict category

How to Get Value from Big Data

- Step 3: Analyze Data

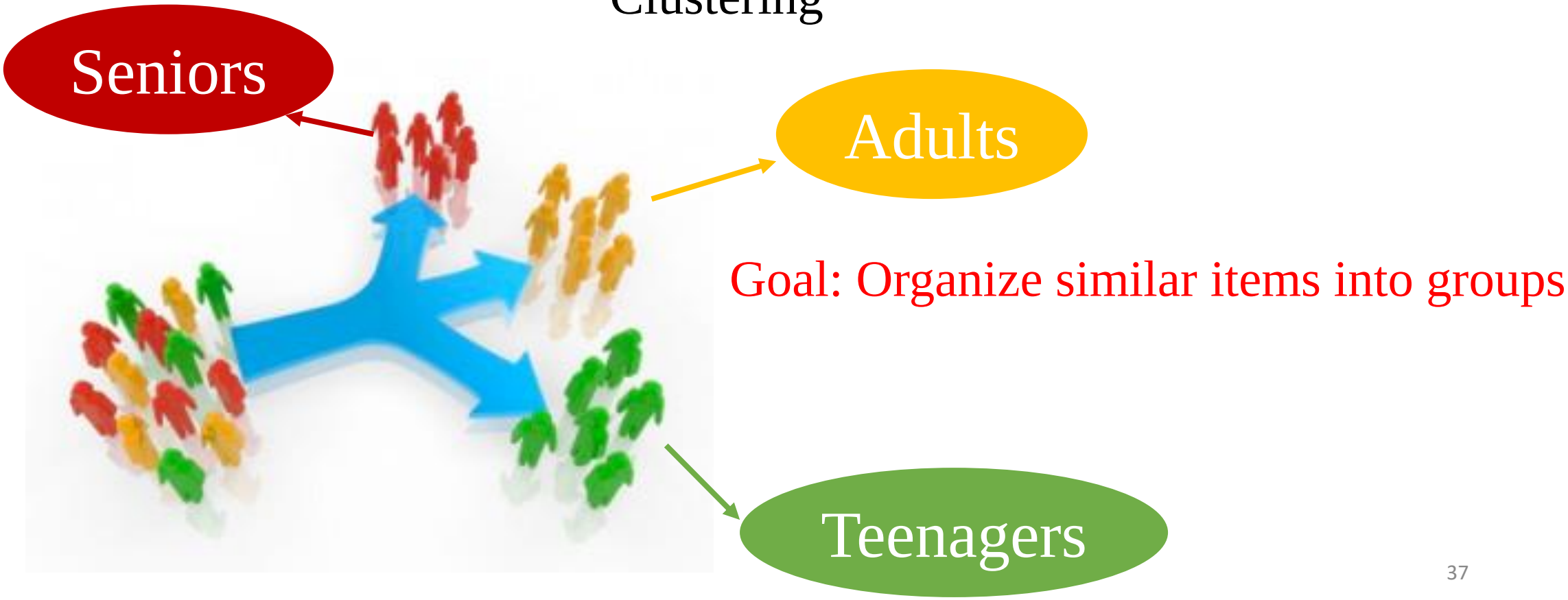
Regression

Goal: Predict numeric value

How to Get Value from Big Data

- Step 3: Analyze Data

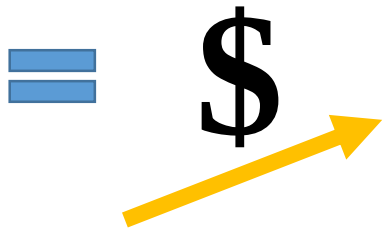
Clustering



Goal: Organize similar items into groups

How to Get Value from Big Data

- Step 3: Analyze Data



Association Analysis

Goal: Find rules to capture associations between items

How to Get Value from Big Data

- Step 3: Analyze Data

Select technique

Build model

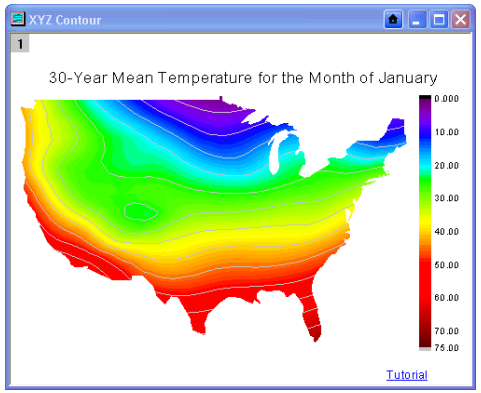
Validate model

Classification
Regression
Clustering
Association
Analysis
Graph Analytics

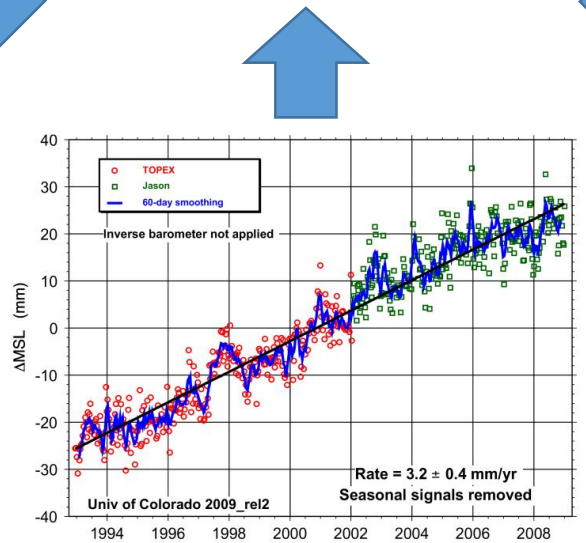
How to Get Value from Big Data

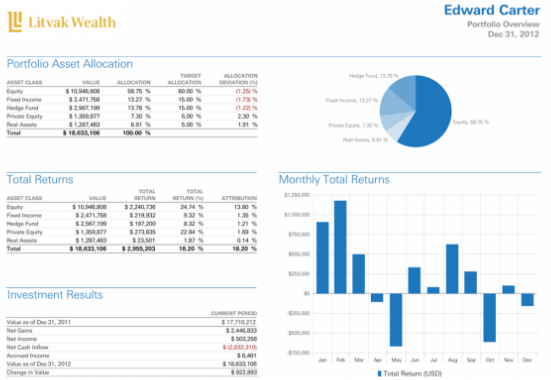
- Step 4: Reporting Insights

Acquire → Prepare → Analyze → Report → ACT



How to Present





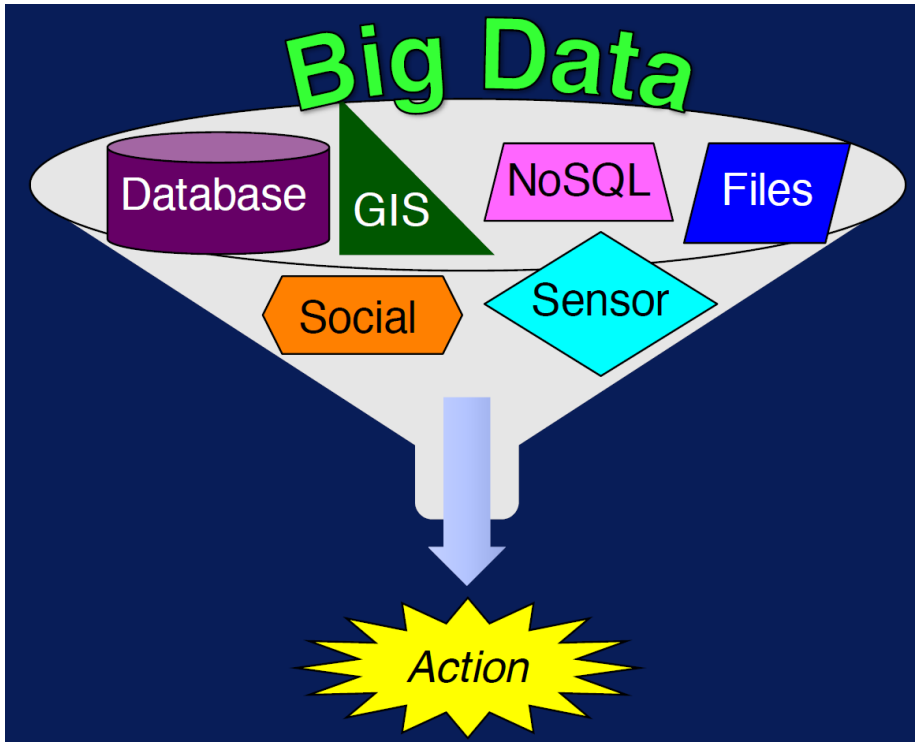
How to Get Value from Big Data

- Step 4: Reporting Insights

Visualization Tools

How to Get Value from Big Data

- Step 5: Insights into Action



How to Get Value from Big Data

- Step 5: Insights into Action