

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

Deep Learning

An MIT Press book

Ian Goodfellow and Yoshua Bengio and Aaron Courville

[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](#)
 - [6 Deep Feedforward Networks](#)

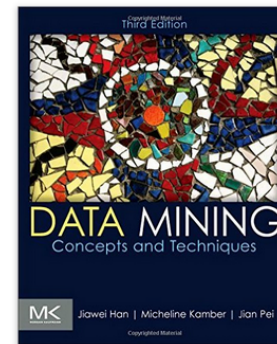
<http://www.deeplearningbook.org/>

Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Manag

by [Jiawei Han](#) (Author), [Micheline Kamber](#) (Author), [Jian Pei](#) (Author)

★★★★☆ 42 customer reviews

[Look inside](#)



ISBN-13: 978-9380931913

ISBN-10: 9380931913

[Why is ISBN important?](#)

Hardcover
\$15.06 - \$54.36

Other Sellers
from \$15.06

☐ Rent \$15.06

☐ Buy used \$42.99

☒ **Buy new** **\$54.36**

Only 7 left in stock (more on the way).

Ships from and sold by Amazon.com. Gift-wrap available.

List Price: \$74.95 Save: \$20.59 (27%)

22 New from \$29.13

Want it Thursday, July 6? Order within 14 hrs 59 mins and choose One-Day Shipping at checkout.

[Details](#)

FREE Shipping.

Qty: 1

[Add to Cart](#)

[Turn on 1-Click ordering](#)

Ship to:

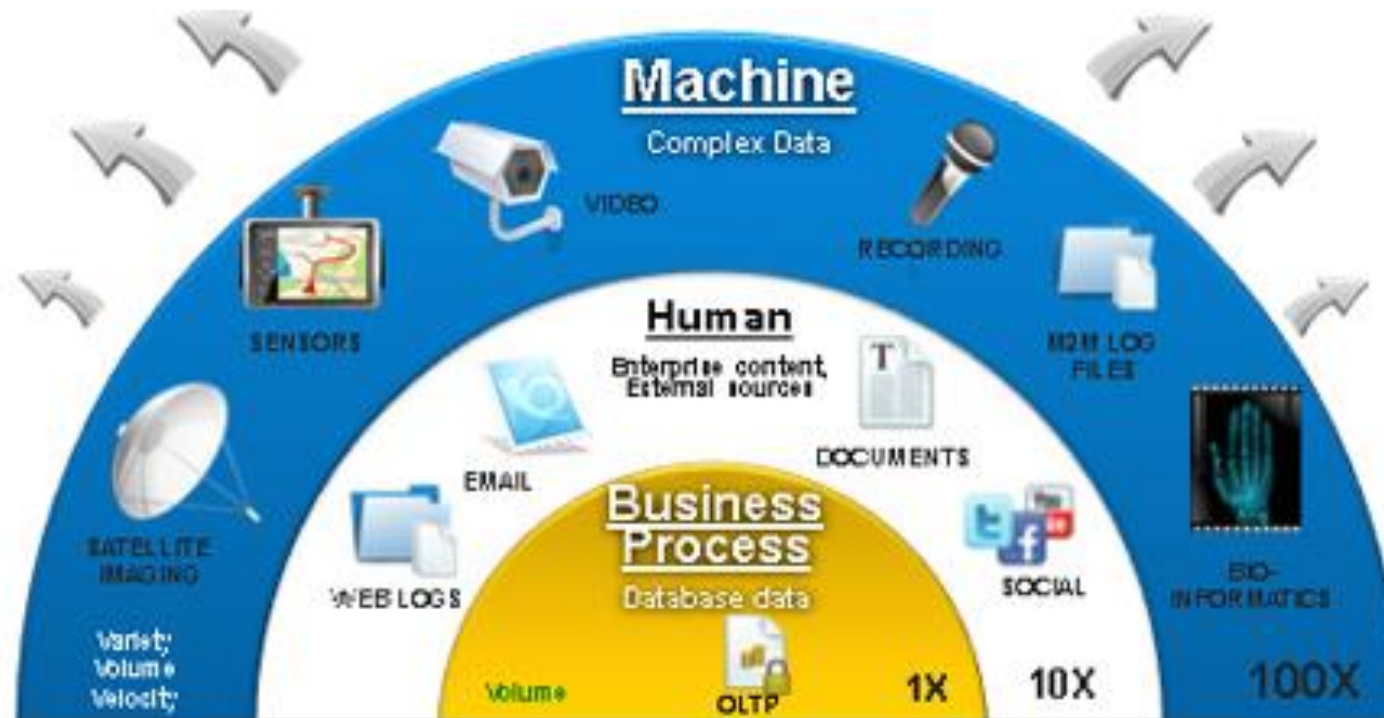
Select a shipping address:

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

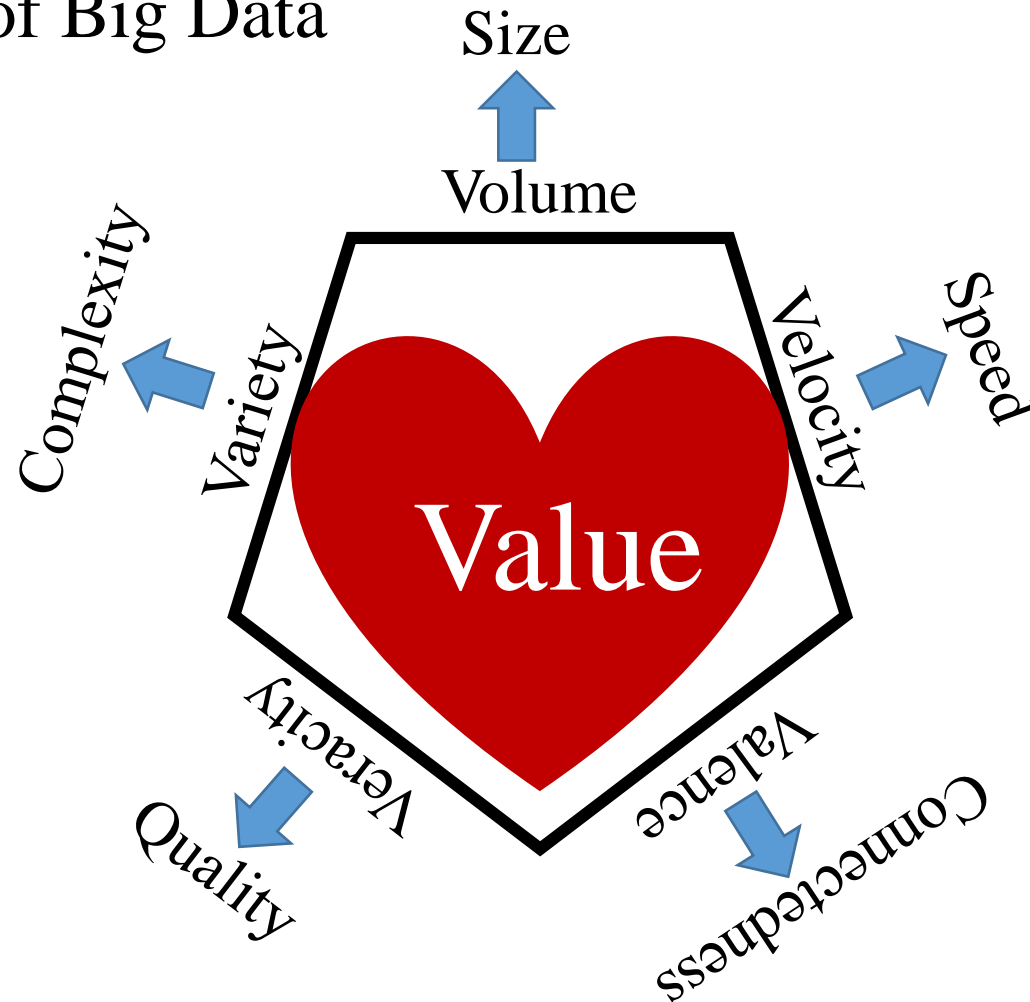
Previous Section

- 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

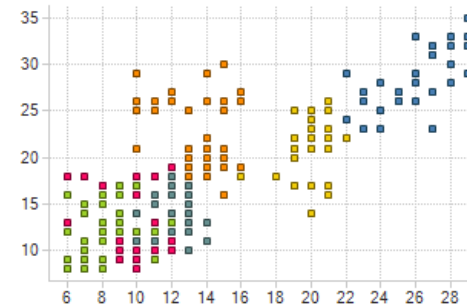


Step 2: Prepare Data

Step 2-A: Explore

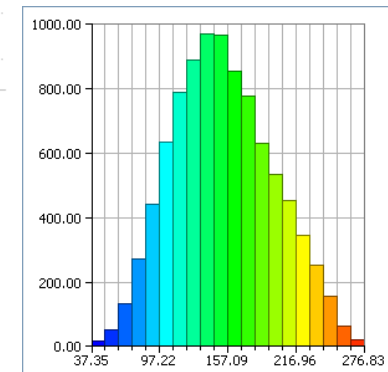
Step 2-B: Pre-process

Step 2-A: Explore Data



Preliminary analysis

Understand nature of data

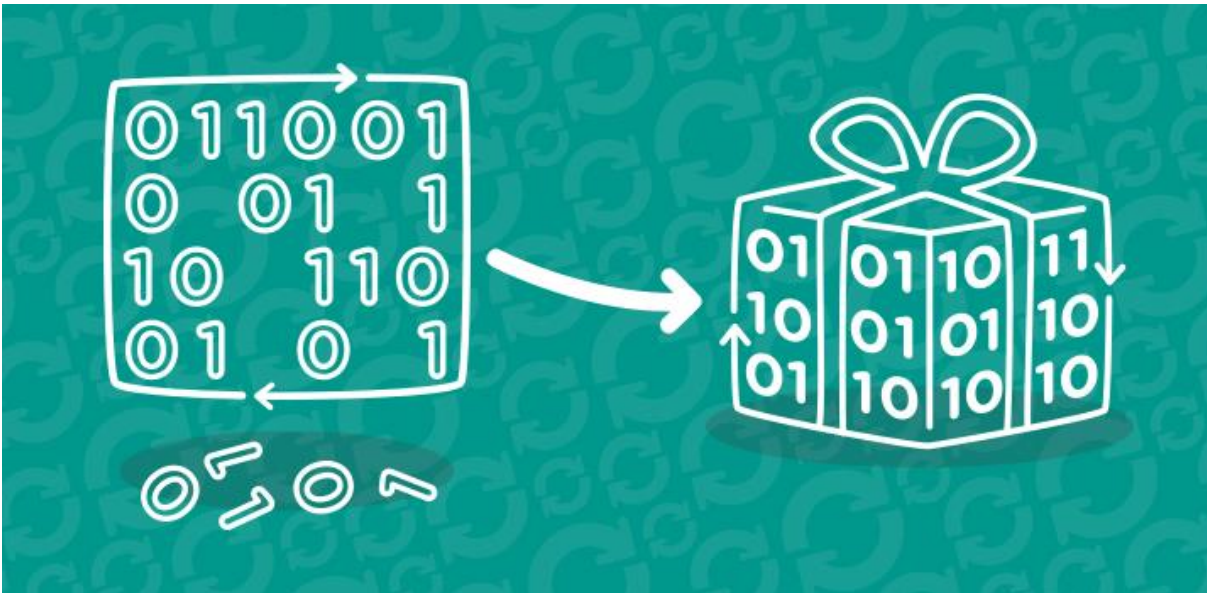


How to Get Value from Big Data

- Steps in the Data Science Process



Step 2-B: Pre-process Data



Clean
Integrate
Package

How to Get Value from Big Data

- Steps in the Data Science Process



Step 3: Analyze Data



Select analytical techniques

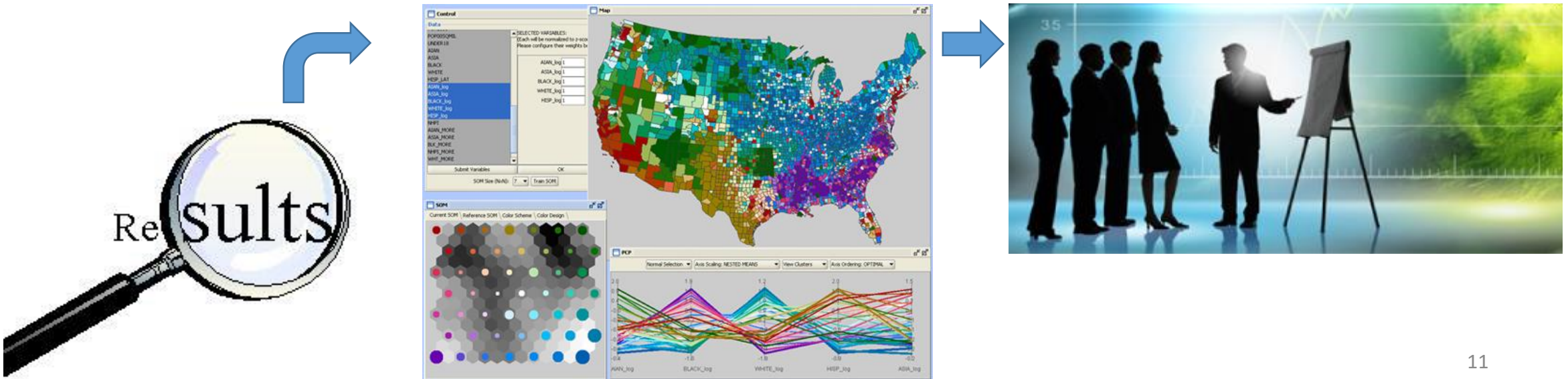
Build models

How to Get Value from Big Data

- Steps in the Data Science Process



Step 4: Communicate Results

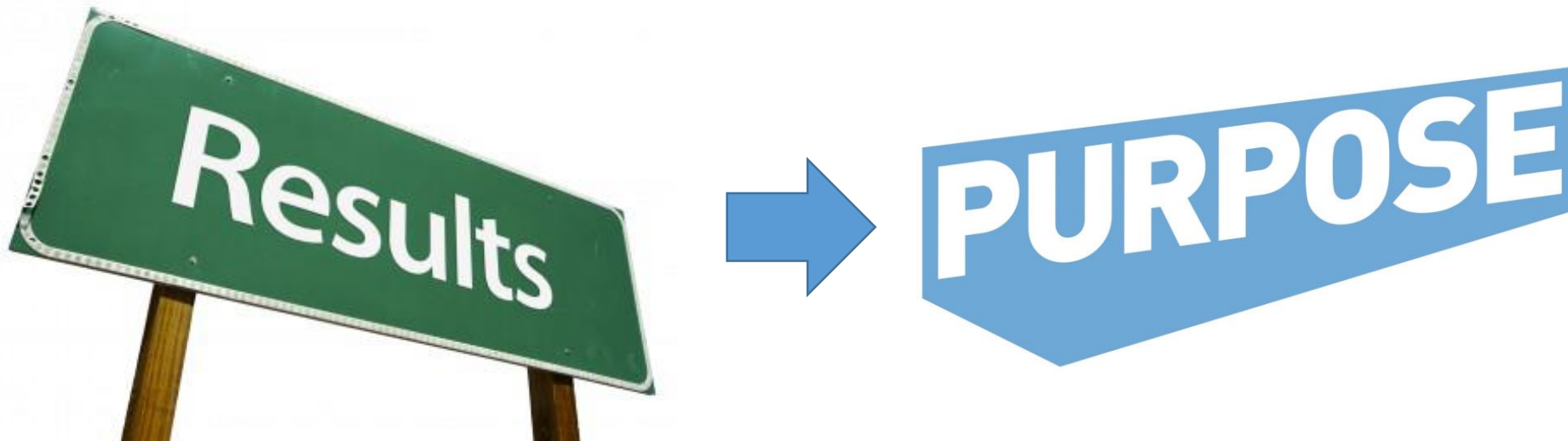


How to Get Value from Big Data

- 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

TWITTER POPULARITY OF U.S. PRESIDENT CANDIDATES

More than 50,000 Tweets analyzed



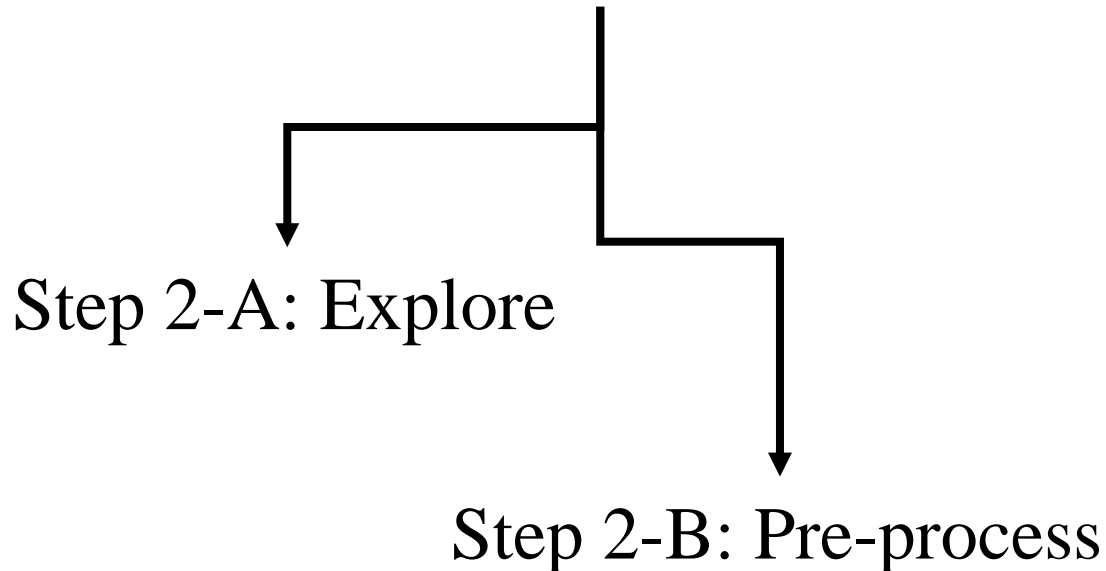
L. Feicho

How to Get Value from Big Data

- Step 2: Prepare Data - Explore



Step 2: Prepare Data



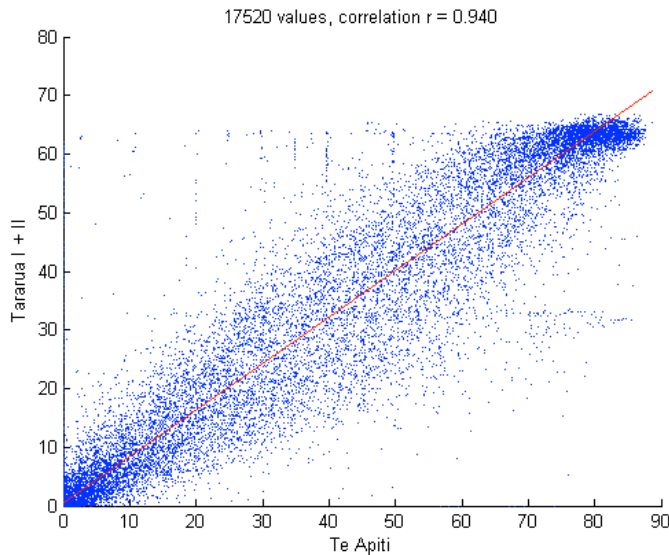
Why Explore?

Goal: Understand your data

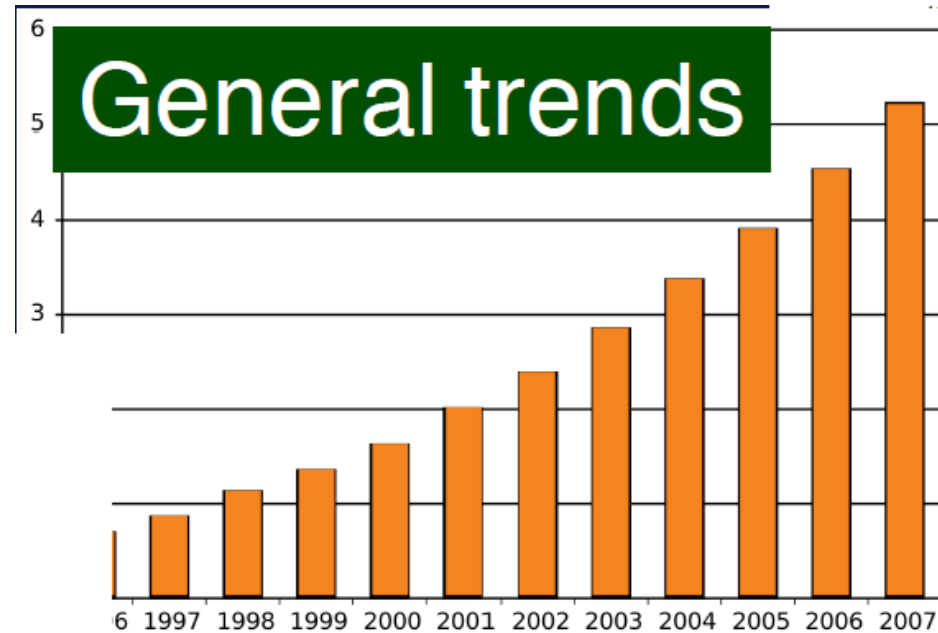
How to Get Value from Big Data

- Step 2: Prepare Data - Explore

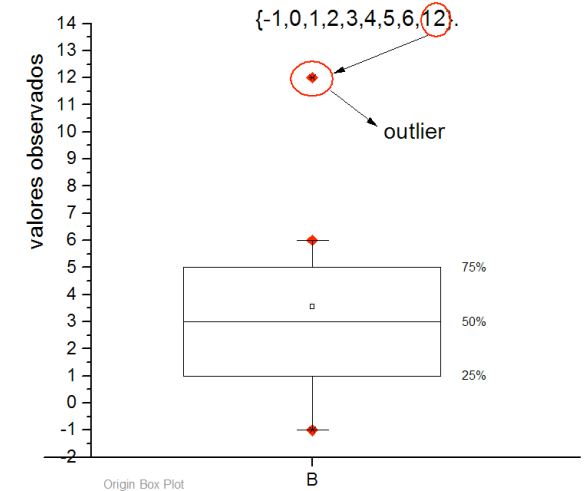
Correlations



General trends



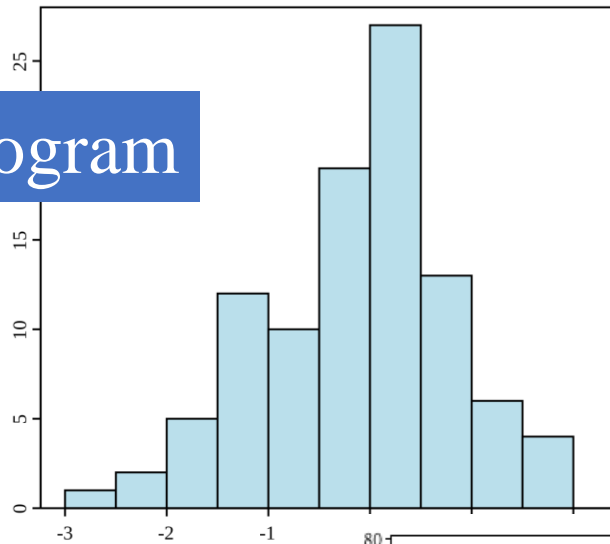
Outliers



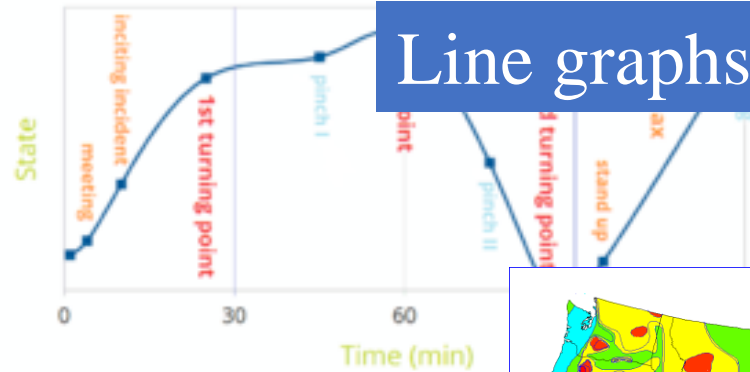
How to Get Value from Big Data

- Step 2: Prepare Data - Explore

Histogram



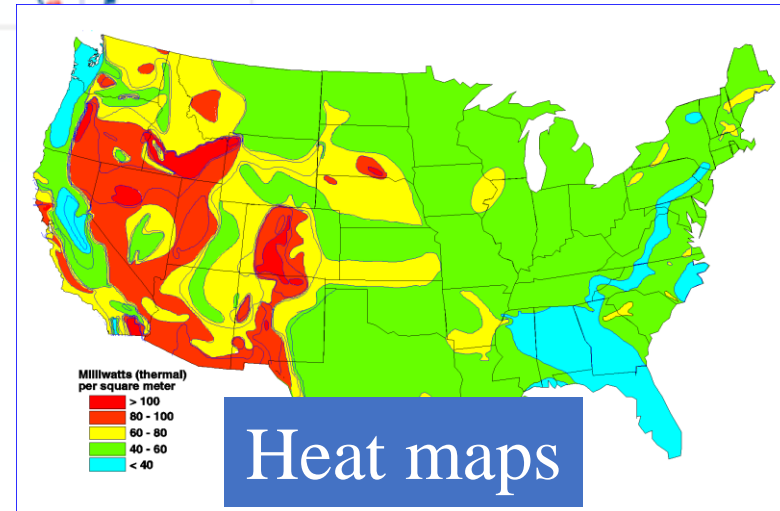
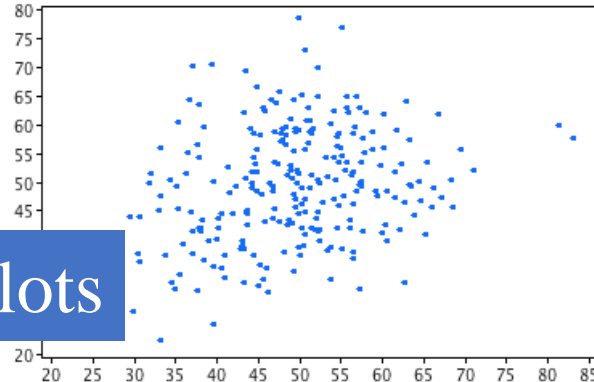
Plot Line Graph



Line graphs

Visualize Your Data

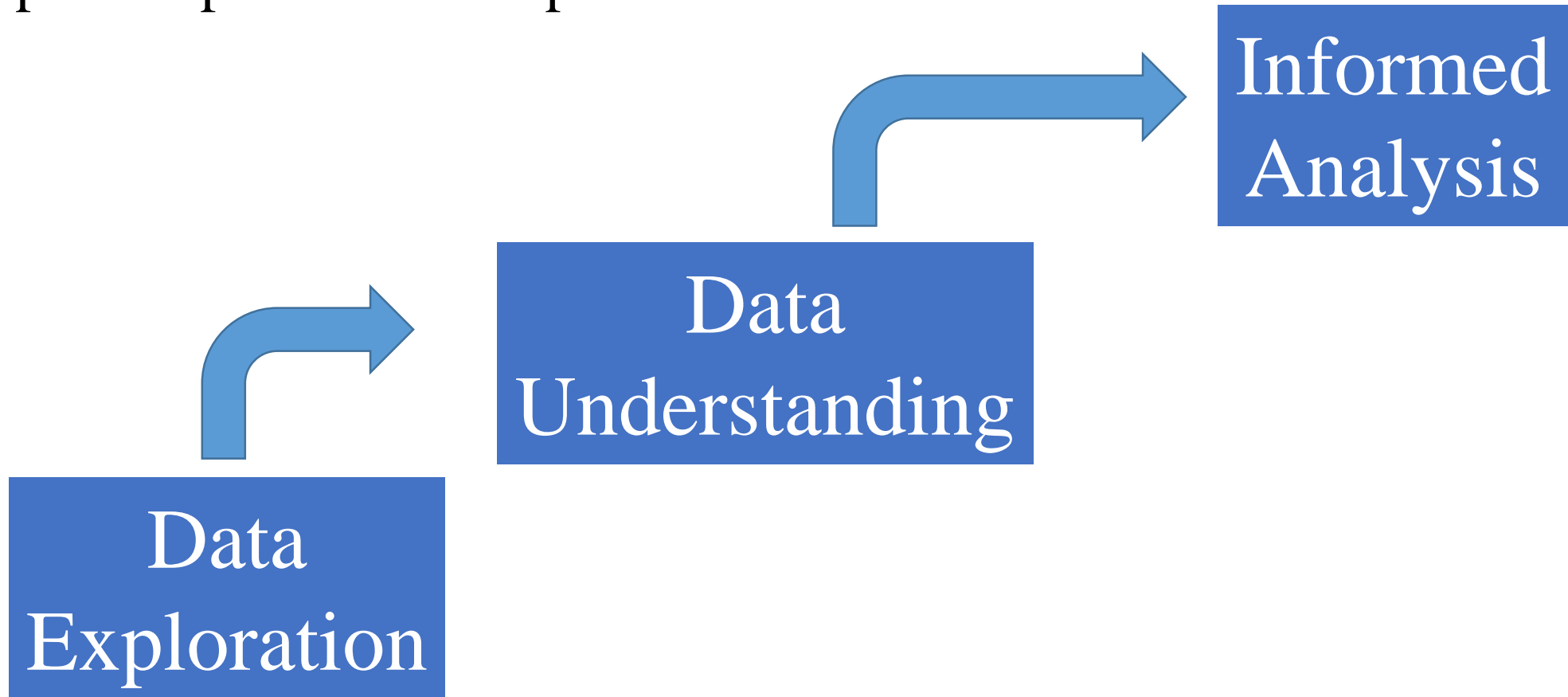
Scatter plots



Heat maps

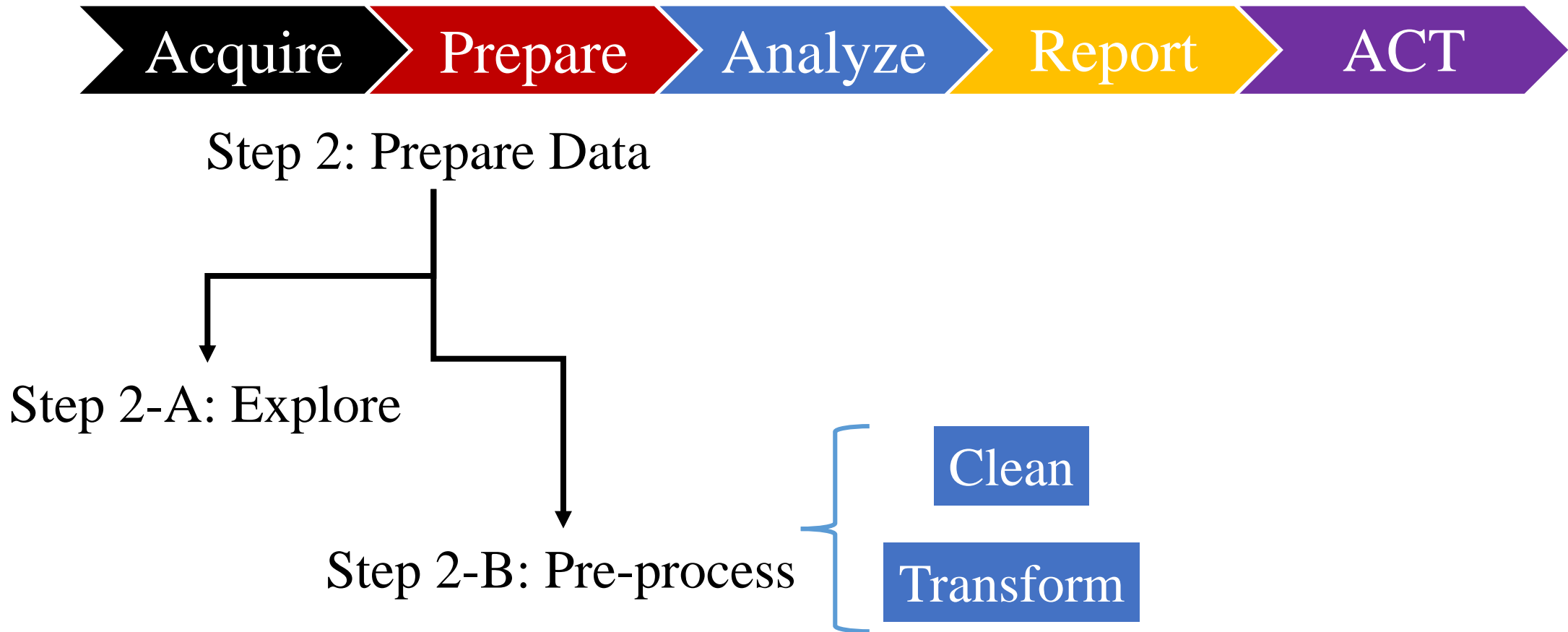
How to Get Value from Big Data

- 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

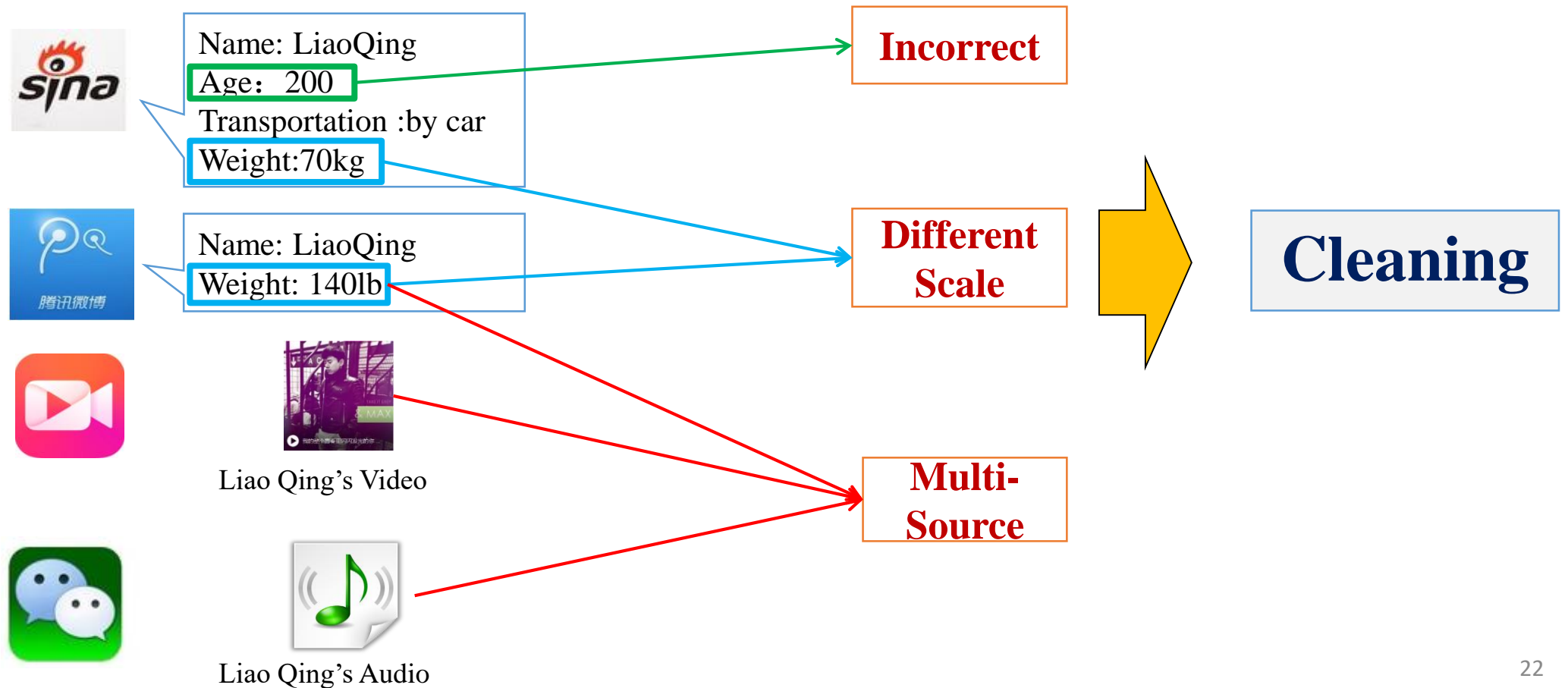
Missing values

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

Name	Qing LIAO	✓
Gender	Female	✓
Birthplace		Empty
Birthday		Empty
ID #	220103196504032526	✓
Phone #	1502036	Anomaly
Email	xyz12596@126.com	✓
Resident	Hunan, Fuzhou	Anomaly
Height	175cm	✓
Weight	80kg	✓
Hobbies	Play Game	✓

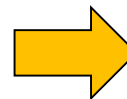
ID Card



Birthplace	Jilin, Changchun
Birthday	1965/04/03



Cleaning



Name	Qing LIAO
Gender	Female
Birthplace	Jilin, Changchun
Birthday	1965/04/03
ID #	220103196504032526
Email	xyz12596@126.com
Height	175cm
Weight	80kg
Hobbies	Play Game

Valid data

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

Web Page



Color

Brand

Sony/索尼

Model

DSC-H300

Type

数码长焦照相机

35倍变焦 大广角

Resolution

2010万像素

13人付款

20条评论



限量抢购 包邮 送MP3 同一购三

正品Sony/索尼DSC-H300数码小单反相机

Feature

35倍长焦大陆行货全国联保

48人付款

129条评论

Record Association

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

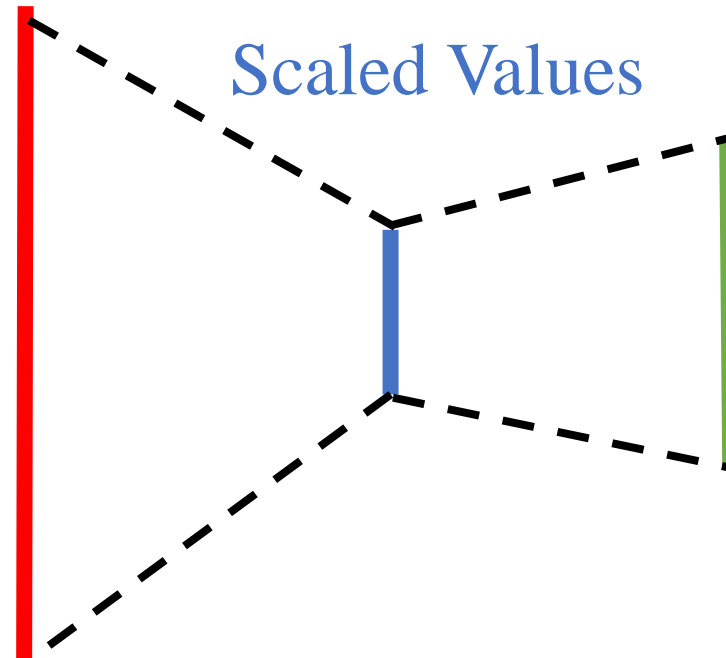
How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process

Scaling



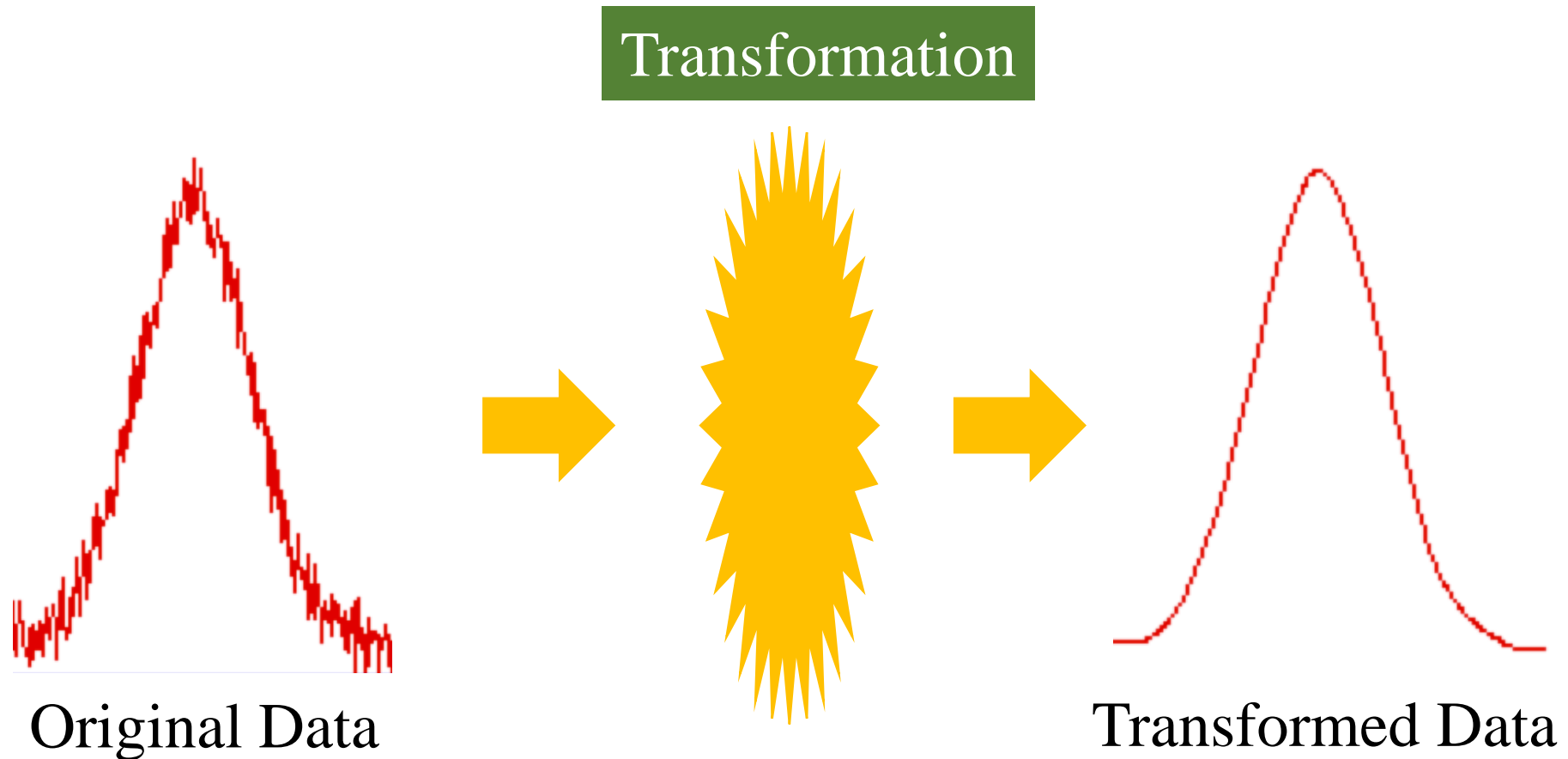
KG



Pounds

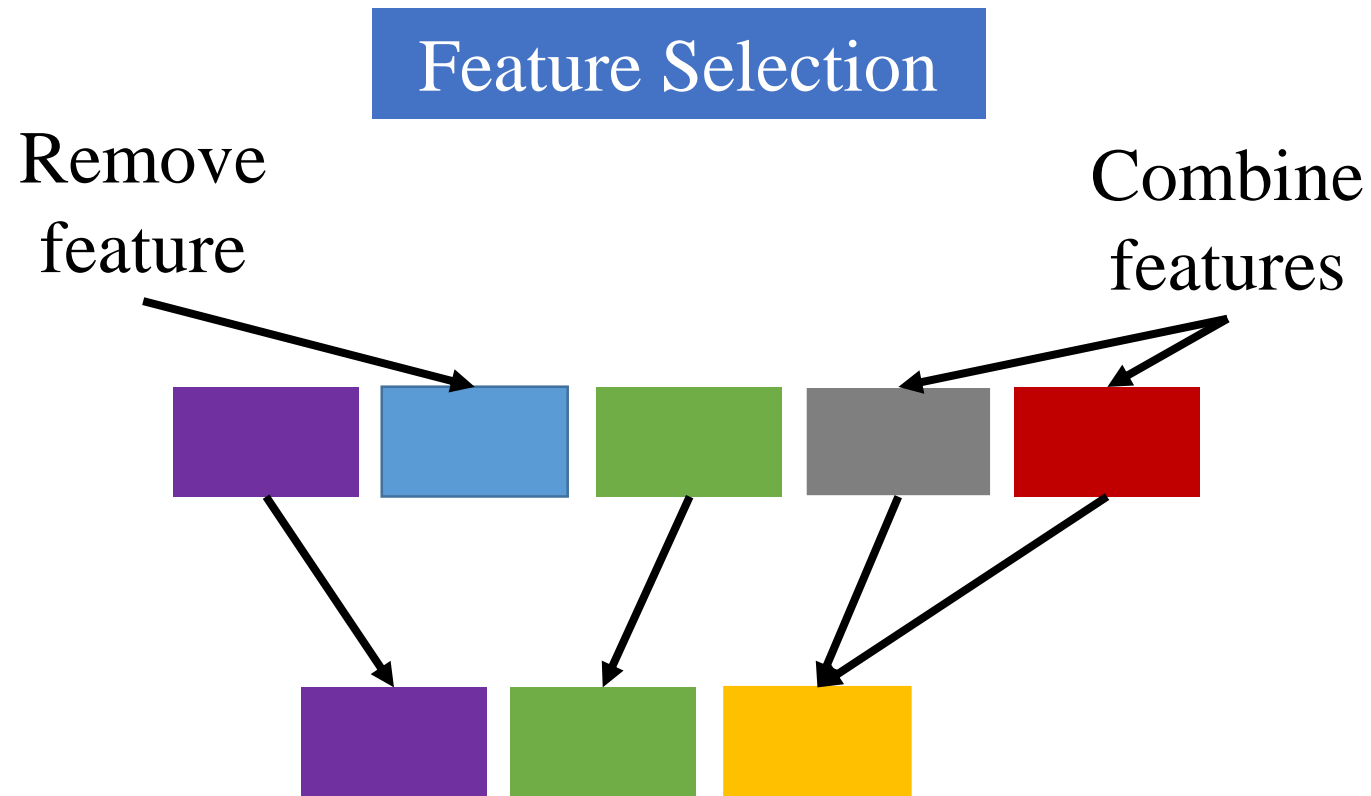
How to Get Value from Big Data

- Step 2: Prepare Data - Pre-process



How to Get Value from Big Data

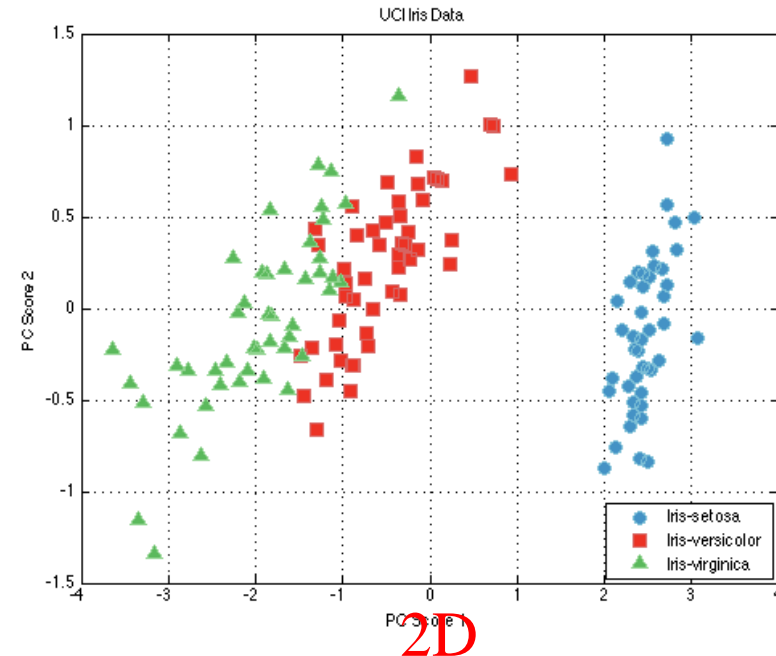
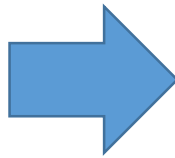
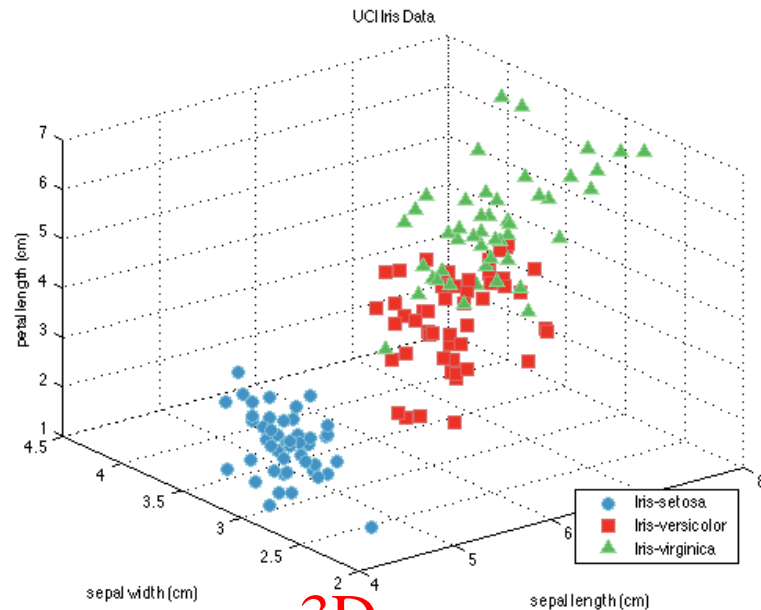
- Step 2: Prepare Data - Pre-process



How to Get Value from Big Data

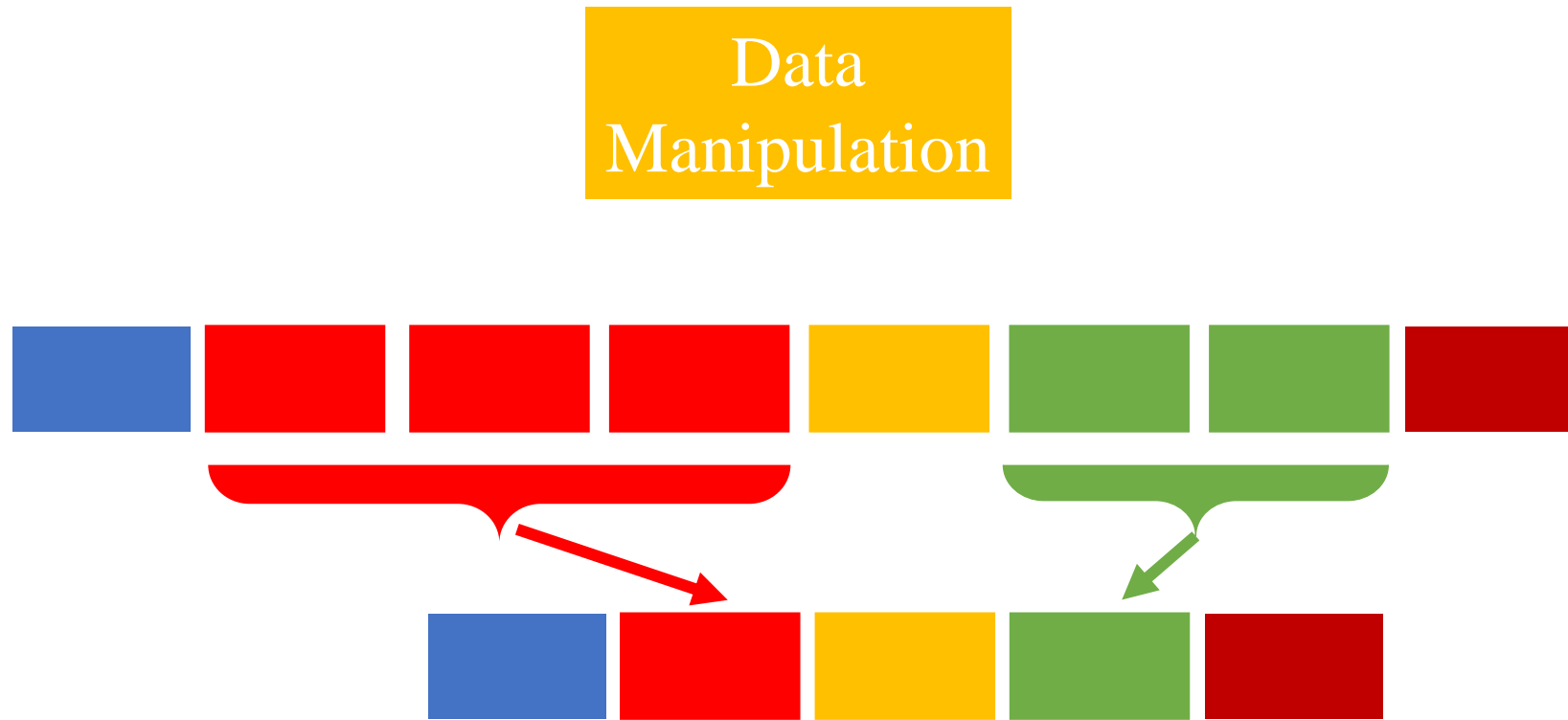
- Step 2: Prepare Data - Pre-process

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

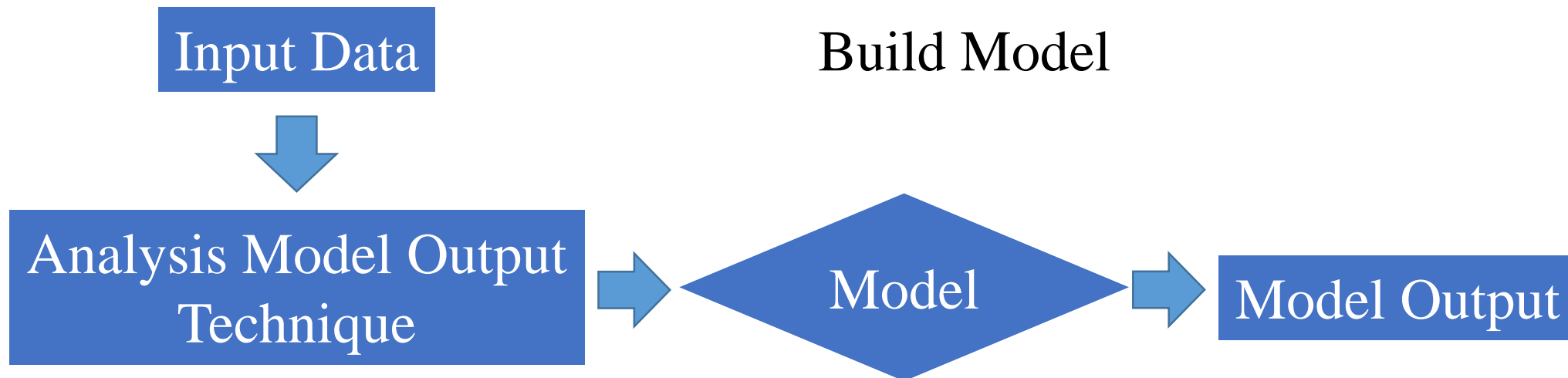
Garbage in = Garbage out



Data preparation is
very important for
meaningful analysis!

How to Get Value from Big Data

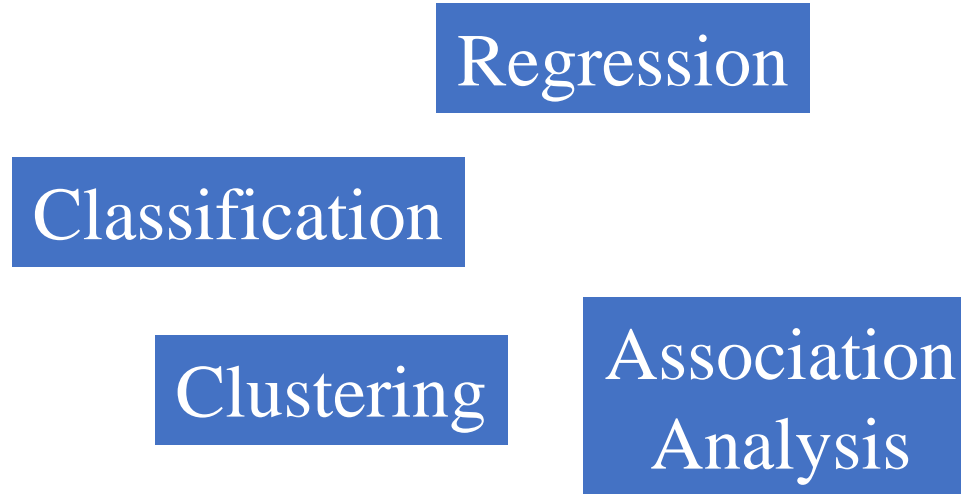
- Step 3: Analyze Data



How to Get Value from Big Data

- Step 3: Analyze Data

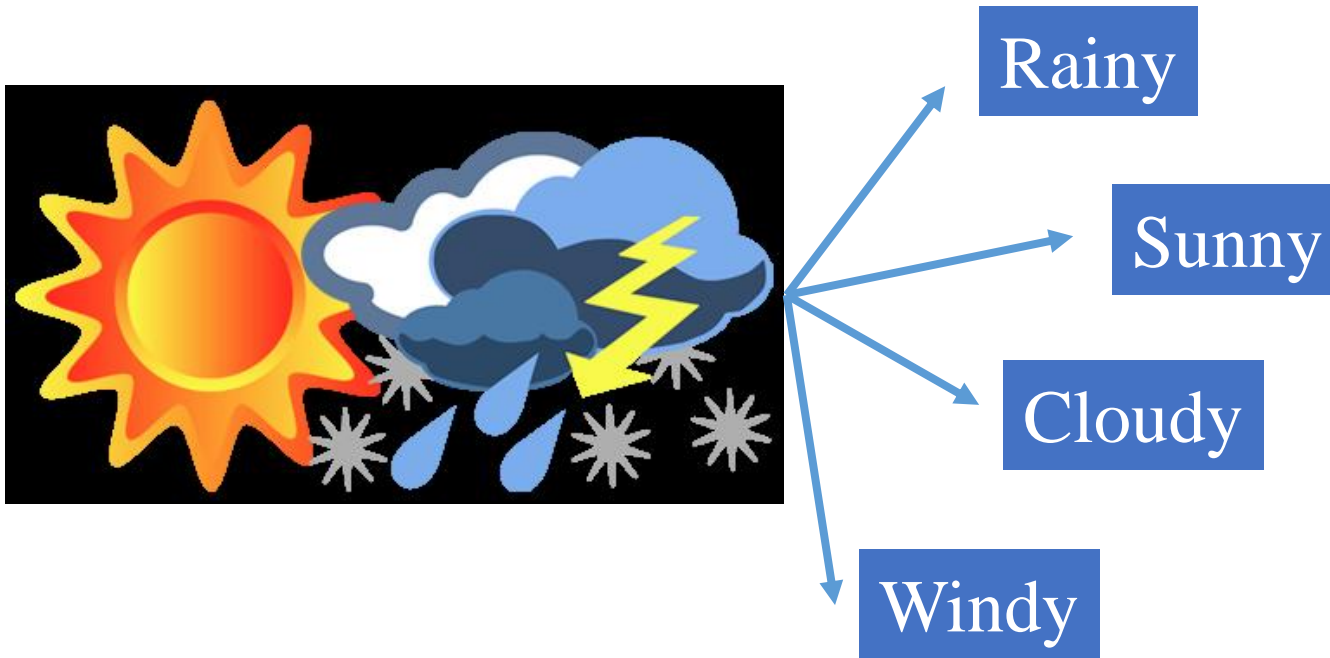
Categories of Analysis Techniques



How to Get Value from Big Data

- Step 3: Analyze Data

Classification



Goal: Predict category

How to Get Value from Big Data

- Step 3: Analyze Data

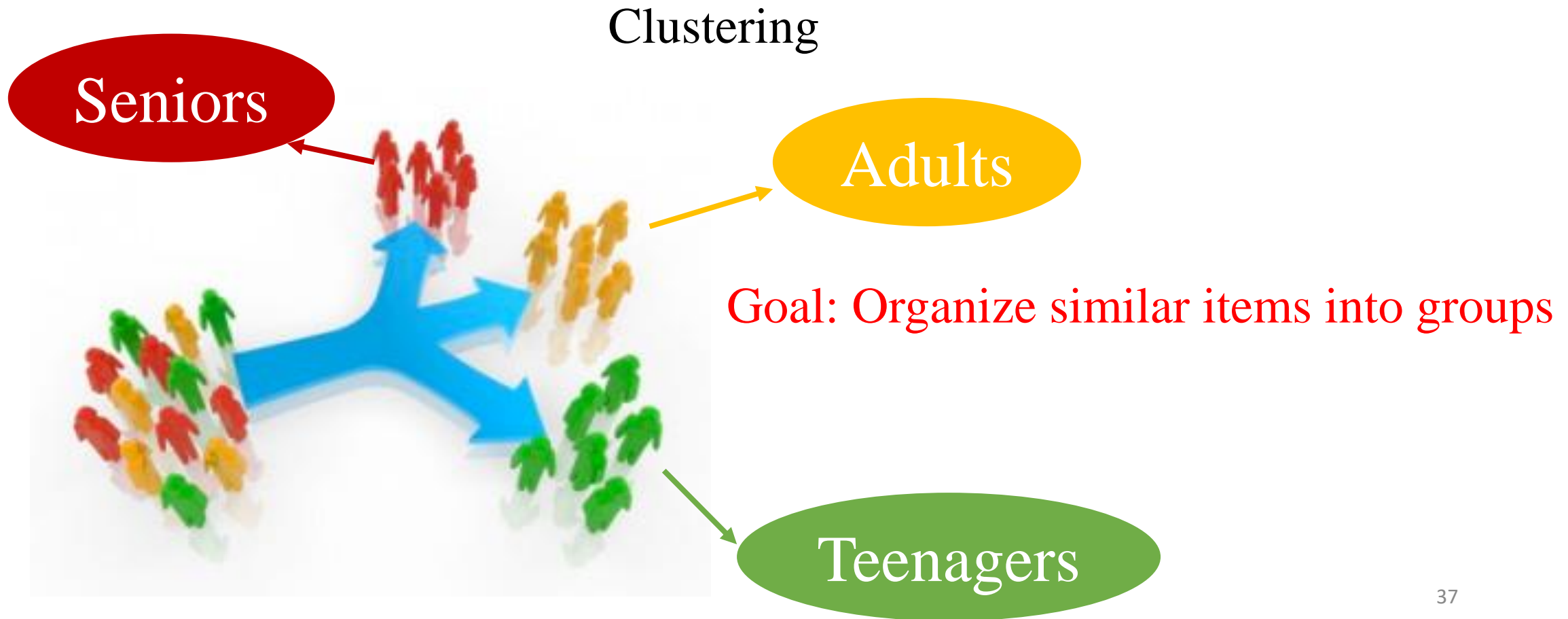
Regression



Goal: Predict numeric value

How to Get Value from Big Data

- Step 3: Analyze Data



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

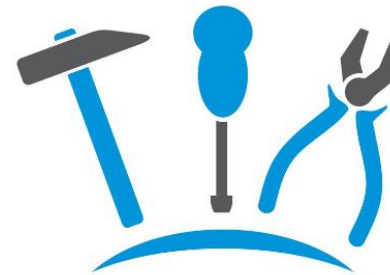


Build model



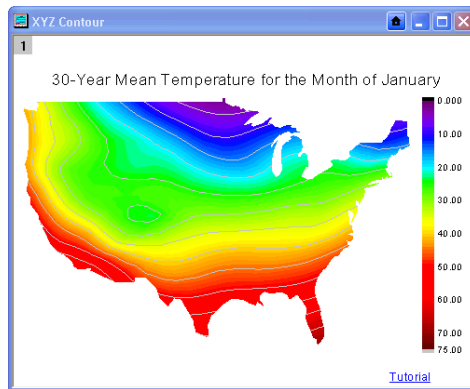
Validate model

Classification
Regression
Clustering
Association
Analysis
Graph Analytics

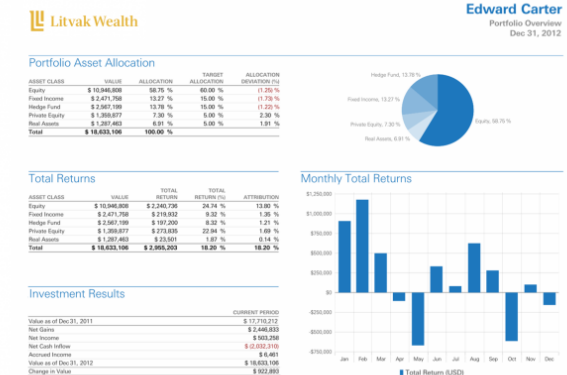
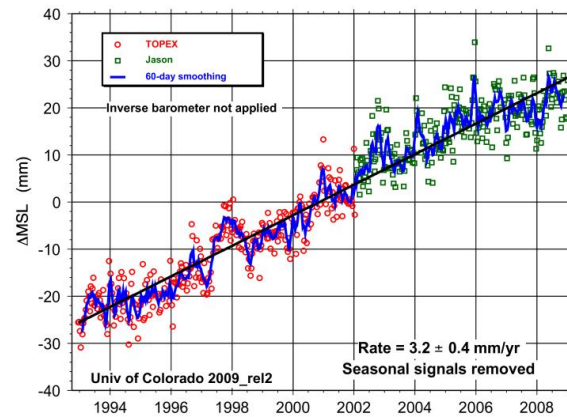
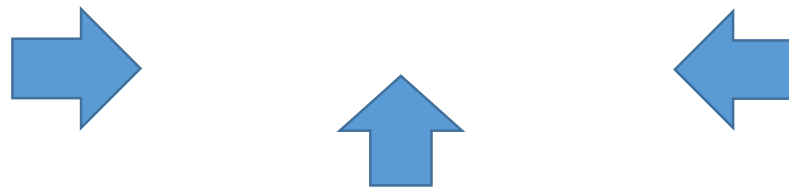


How to Get Value from Big Data

- Step 4: Reporting Insights



How to Present



How to Get Value from Big Data

- Step 4: Reporting Insights



Visualization Tools



python



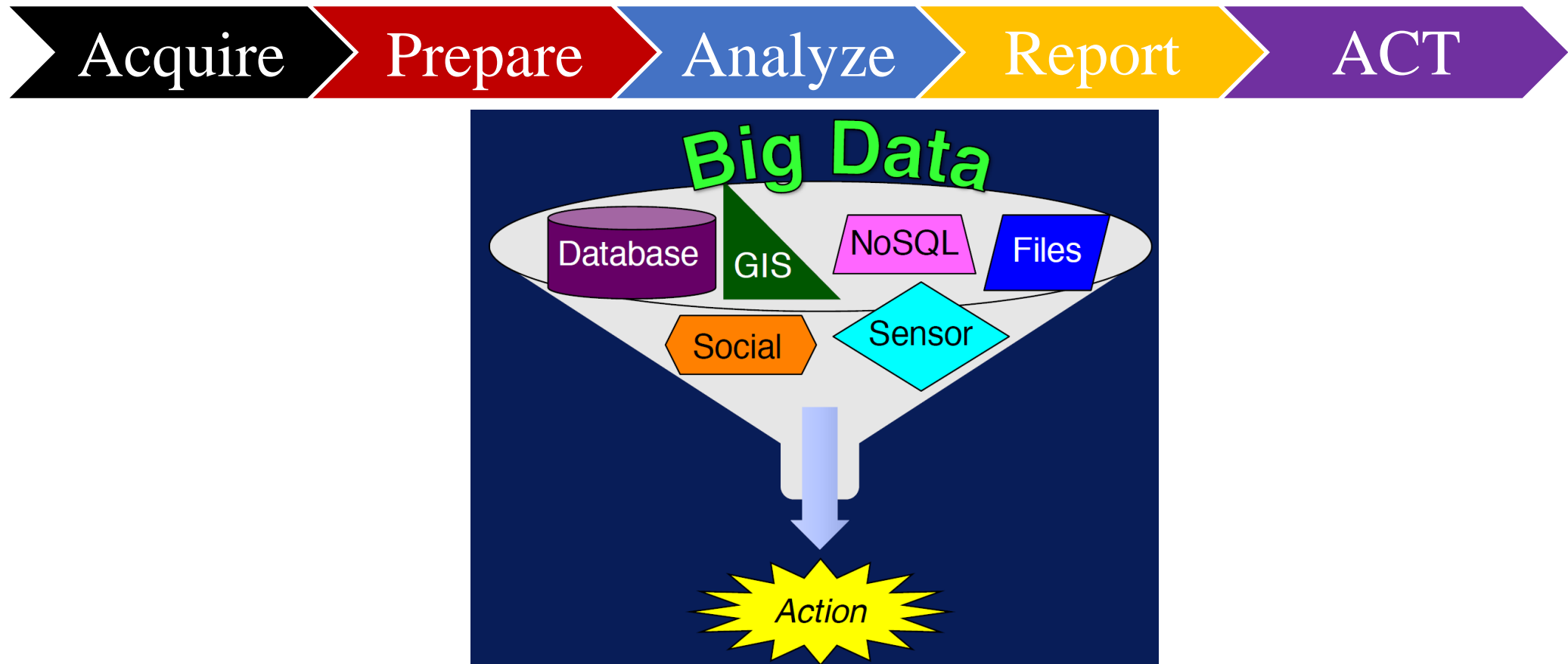
Google
Developers
CHARTS

Timeline ^{JS}

Beautifully crafted timelines that are easy
and intuitive to use.

How to Get Value from Big Data

- Step 5: Insights into Action



How to Get Value from Big Data

- Step 5: Insights into Action

