# 数据挖掘导论

# Introduction to Data Mining

## Data Preprocessing

数据智能实验室
DATA INTELLIGENCE LABORATORY
DI-LAB
浙江大学
Zhejiang University

# Agenda

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Summary

# Data Cleaning

- Missing/Inconsistent/Noisy Data in a relational table

| Student ID | Student Name | Age | GPA | Classification |
|---|---|---|---|---|
| 100122014 | Joseph | 21 | 3.5 | Junior |
| 100232015 | Patrick | 200 | 3.2 | Sophomore |
| 100122012 | Seller | 24 | 3.0 | Senior |
| 100342013 | Roger | 23 | 234 | Senior |
| 100942012 | Davis | 2.8 | 3.7 | Sophomore |
| | Travis | 23 | 3.4 | Sr |
| 100982015 | Alex | 27 | | Sophomore |
| 100982013 | Trevor | -22 | 4.0 | Senior |
| AUC2016XC | Aman | 30 | 3.5 | Jr |

Missing Data

Inconsistent Data

Noisy Data





Image source: https://medium.com/@casternxavier/get-rid-of-the-dirt-from-your-data-data-cleaning-techniques-743253b1c3e5

# Noisy Data



一个多小时行程结束以后，吴小姐傻眼了，滴滴系统发来的账单显示吴小姐要**支付540多万元的打车费用**，而打车时间显示的则是**1000多万分钟**。

# Data Inconsistency

Financial

| Employee | Salary |
|---|---|
| John | 1000 |
| | |

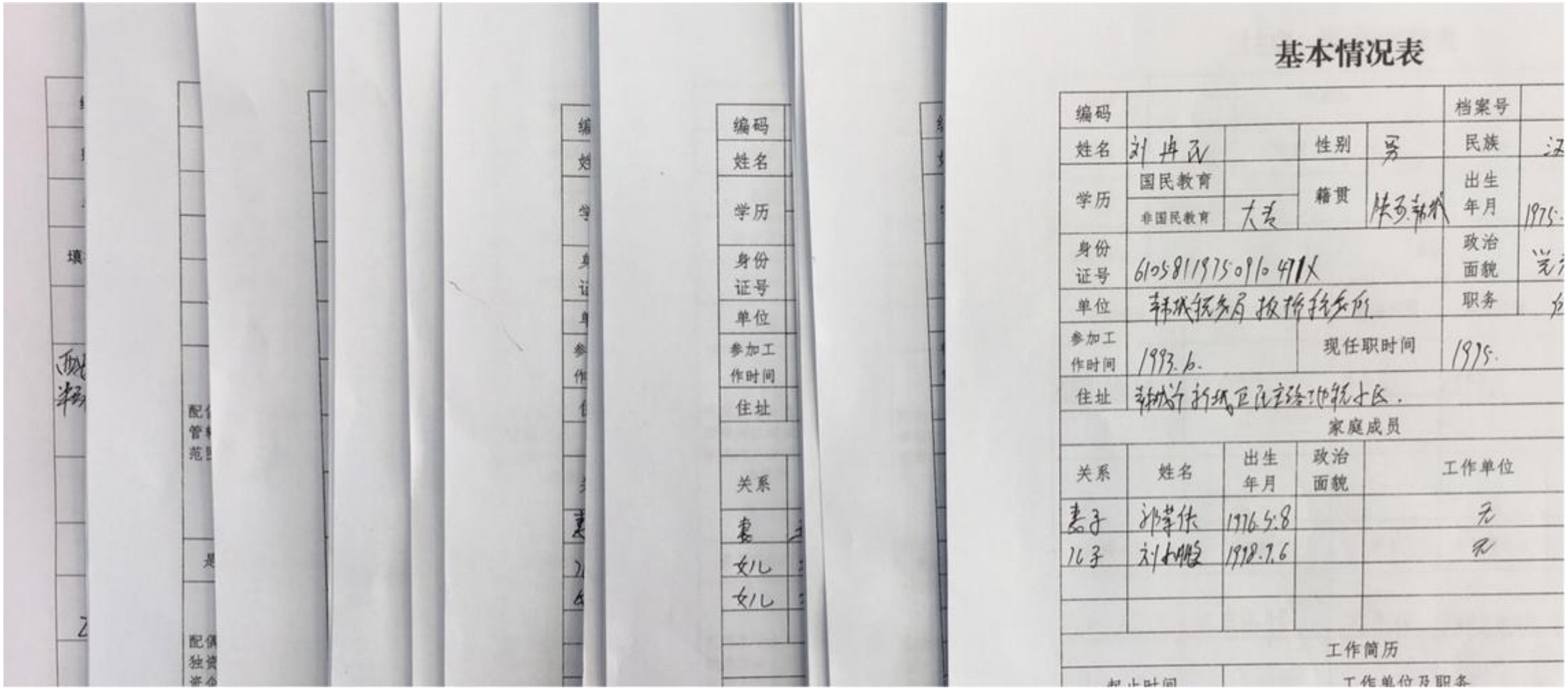
Employee → Salary

Human Resources

| Employee | Salary |
|---|---|
| John | 2000 |
| Mary | 3000 |

Employee → Salary

# Data Inconsistency



基本情况表

| 编码 | | | | 档案号 | |
|---|---|---|---|---|---|
| 姓名 | 刘再正 | | 性别 | 男 | 民族 | 汉 |
| 学历 | 国民教育 | | 籍贯 | 陕西韩城 | 出生年月 | 1975. |
| | 非国民教育 | 大专 | | | | |
| 身份证号 | 610581197509104911X | | | | 政治面貌 | 党 |
| 单位 | 韩城税务局 板桥税务所 | | | | 职务 | 办 |
| 参加工作时间 | 1993.6. | | 现任职时间 | | 1975. | |
| 住址 | 韩城市新城区正宁路地税小区. | | | | | |

家庭成员

| 关系 | 姓名 | 出生年月 | 政治面貌 | 工作单位 |
|---|---|---|---|---|
| 妻子 | 郑萍侠 | 1976.5.8 | | 无 |
| 儿子 | 刘帅骏 | 1998.7.6 | | 无 |
| | | | | |
| | | | | |

工作简历

| 起止时间 | 工作单位及职务 |
|---|---|

# Data Cleaning

## 1. Missing Data

→ Ignore

→ Fill Manually

→ Fill Computed Value

## 2. Noisy Data

→ Binning

→ Clustering

→ Machine Learning Algorithm

→ Remove Manually

## 3. Inconsistent Data

→ External References

→ Knowledge Engineering Tools





Image source: https://www.v7labs.com/blog/data-preprocessing-guide

# Binning

## 数据分箱

数据分箱（Binning）作为数据预处理的一部分，也被称为离散分箱或数据分段。分箱把数据根据一定的规则进行分组，使数据变得离散化，以增强模型的稳定性并避免过拟合

### 连续型数据分箱

| 年龄 | | 年龄组 |
|---|---|---|
| 29 | 分箱 | 18至40岁 |
| 7 | | 18岁以下 |
| 49 | | 40至60岁 |
| 12 | | 18岁以下 |
| 50 | | 40至60岁 |
| 34 | | 18至40岁 |
| 36 | | 18至40岁 |
| 75 | | 60岁以上 |
| 61 | | 60岁以上 |

### 离散型数据分箱

| 垃圾 | | 垃圾分类 |
|---|---|---|
| 报纸 | 分箱 | 可回收物 |
| 电池 | | 有害垃圾 |
| 鸡肉 | | 湿垃圾 |
| 花卉 | | 湿垃圾 |
| 胶带 | | 干垃圾 |
| 眼镜 | | 干垃圾 |
| 消毒剂 | | 有害垃圾 |
| 苹果核 | | 湿垃圾 |
| 泡沫塑料 | | 可回收物 |

01 **提高模型的稳定性与鲁棒性**

进行分箱后，特征的取值更加稳定，对模型对异常值的包容性增强

02 **防止模型过拟合**

模型提供信息的精准度降低，模型的泛化能力增强，过拟合可能性降低

03 **加快模型训练速度**

模型的复杂性降低，计算速度快，可以满足大量级数据的运算以及模型的部署、响应速度要求

04 **处理空值与缺失值**

可以将空值、缺失值进行处理，将他们单独作为一个类别进行分类

05 **增强逻辑回归的拟合力**

在将离散变量进行哑[illegible]权重与非线性，增强模型表达

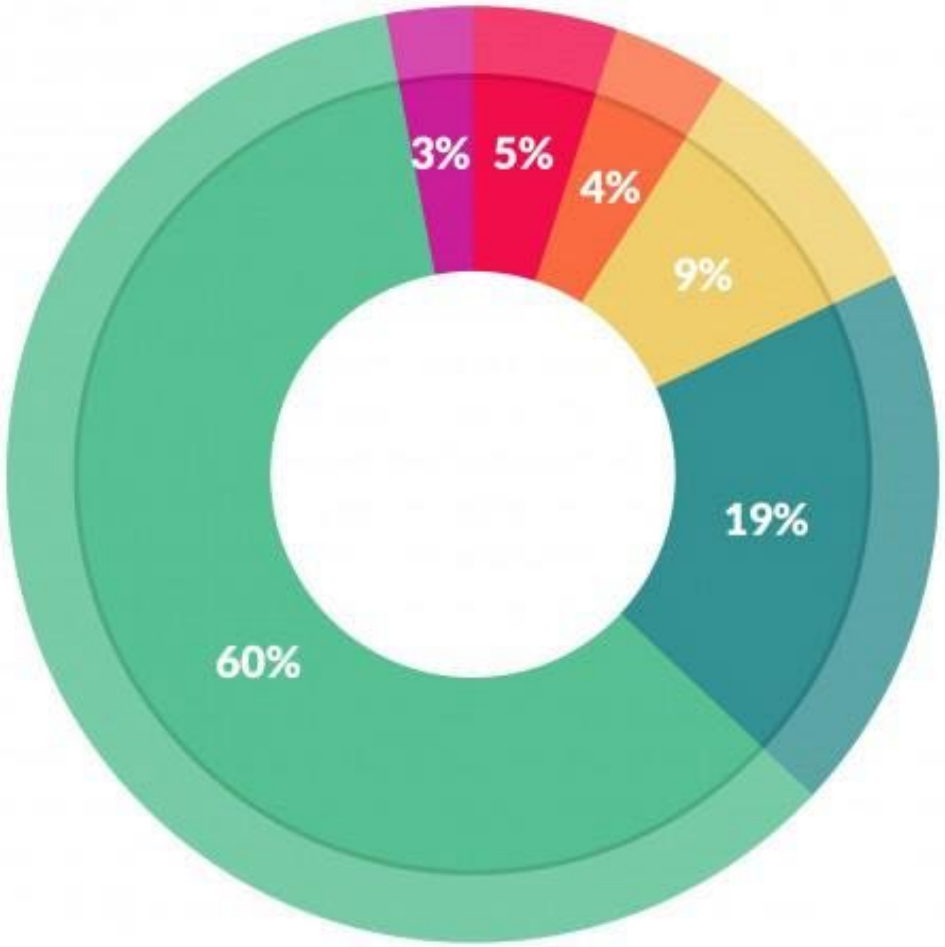






Image source: https://blog.csdn.net/Orange_Spotty_Cat/article/details/116485079

# Data Cleaning




What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%





Image source: https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/

# HoloClean

## Input

### Dataset to be cleaned

| | DBAName | Address | City | State | Zip |
|---|---|---|---|---|---|
| t1 | John Veliotis Sr. | 3465 S Morgan ST | Chicago | IL | 60608 |
| t2 | John Veliotis Sr. | 3465 S Morgan ST | Chicago | IL | 60609 |
| t3 | John Veliotis Sr. | 3465 S Morgan ST | Chicago | IL | 60609 |
| t4 | Johnnyo's | 3465 S Morgan ST | Cicago | IL | 60608 |

### Denial Constraints

c1: DBAName → Zip
c2: Zip → City, State
c3: City, State, Address → Zip

### Matching Dependencies

m1: Zip = Ext_Zip → City = Ext_City
m2: Zip = Ext_Zip → State = Ext_State
m3: City = Ext_City ∧ State = Ext_State∧
∧ Address = Ext_Address → Zip = Ext_Zip

### External Information

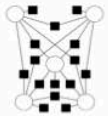
| Ext_Address | Ext_City | Ext_State | Ext_Zip |
|---|---|---|---|
| 3465 S Morgan ST | Chicago | IL | 60608 |
| 1208 N Wells ST | Chicago | IL | 60610 |
| 259 E Erie ST | Chicago | IL | 60611 |
| 2806 W Cermak Rd | Chicago | IL | 60623 |

## The HoloClean Framework

1. Error detection module

2. Automatic compilation to a probabilistic graphical model

3. Repair via statistical learning and infernece



## Output

### Proposed Cleaned Dataset

| | DBAName | Address | City | State | Zip |
|---|---|---|---|---|---|
| t1 | John Veliotis Sr. | 3465 S Morgan ST | Chicago | IL | 60608 |
| t2 | John Veliotis Sr. | 3465 S Morgan ST | Chicago | IL | **60608** |
| t3 | John Veliotis Sr. | 3465 S Morgan ST | Chicago | IL | **60608** |
| t4 | **John Veliotis Sr.** | 3465 S Morgan ST | **Chicago** | IL | 60608 |

### Marginal Distribution of Cell Assignments

| Cell | Possible Values | Probability |
|---|---|---|
| t2.Zip | 60608 | 0.84 |
| | 60609 | 0.16 |
| t4.City | Chicago | 0.95 |
| | Cicago | 0.05 |
| t4.DBAName | John Veliotis Sr. | 0.99 |
| | Johnnyo's | 0.01 |

# Apple Buys Machine-Learning Startup to Improve Data Used in Siri



Image source: https://dawn.cs.stanford.edu/2017/05/12/holoclean/

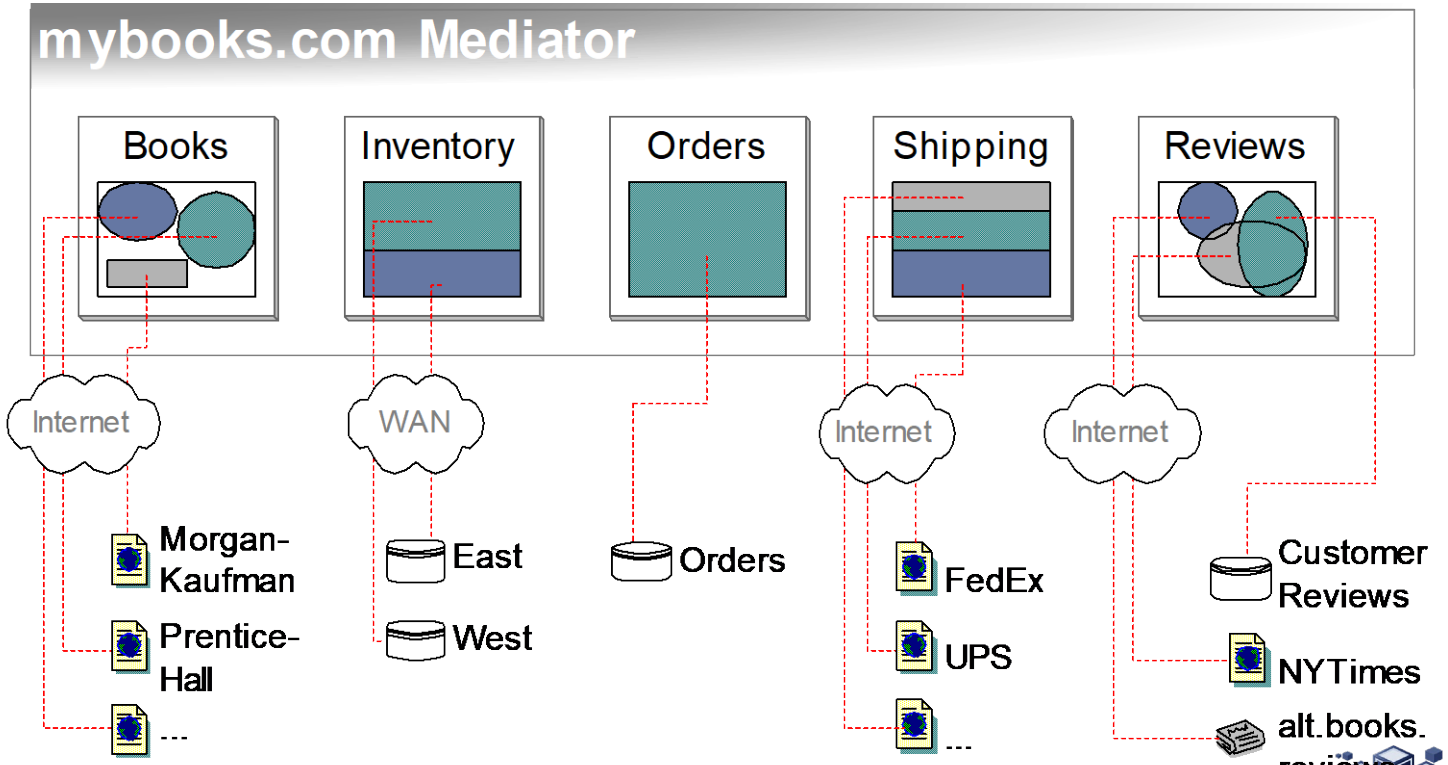
# Agenda

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Summary

# Data Integration

## Schema integration

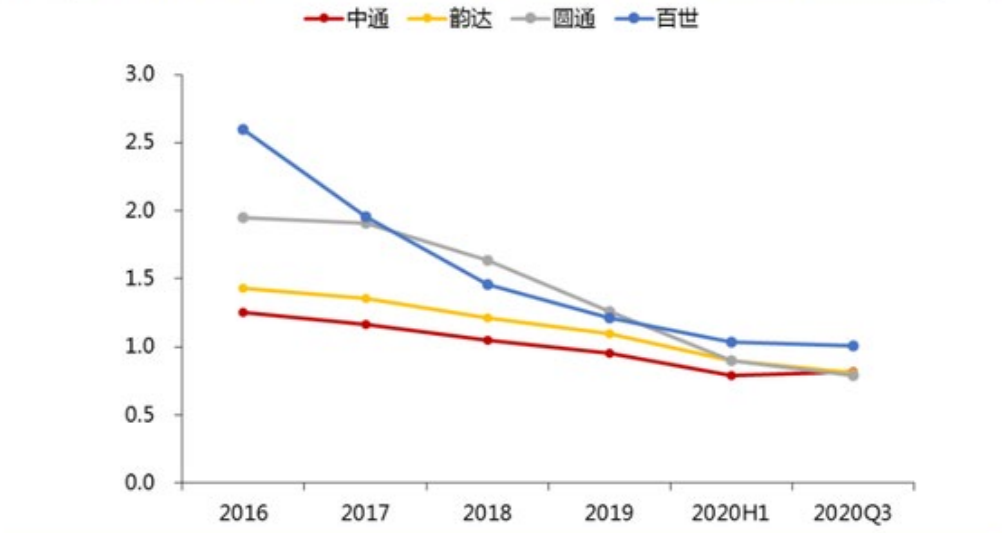
- ✓ Provide uniform access to data available in **multiple, autonomous, heterogeneous** and **distributed** data sources



# Data Integration


来源：根据公开资料查询整理　　亿欧（www.iyiou.com）

图 7：通达系快递企业单票成本（运输+中心操作，单位：元）




资料来源：公司公告，安信证券研究中心

# Data Integration

# Entity Resolution

**2018**

- [j22] Dongxiang Zhang, Mengting Ding, Dingyu Yang, Yi Liu, Ju Fan, Heng Tao Shen:
  **Trajectory Simplification: An Experimental Study and Quality Analysis.** Proc. VLDB Endow. 11(9): 934-946 (2018)
- [j21] Long Guo, Dongxiang Zhang, Yuan Wang, Huayu Wu, Bin Cui, Kian-Lee Tan:
  **CO$^2$: Inferring Personal Interests From Raw Footprints by Connecting the Offline World with the Online World.** ACM Trans. Inf. Syst. 36(3): 31:1-31:29 (2018)
- [j20] Dongxiang Zhang, Yuchen Li, Xin Cao, Jie Shao, Heng Tao Shen:
  **Augmented keyword search on spatial entity databases.** VLDB J. 27(2): 225-244 (2018)
- [j19] Yan Dai, Jie Shao, Chengbo Wei, Dongxiang Zhang, Heng Tao Shen:
  **Personalized semantic trajectory privacy preservation through trajectory reconstruction.** World Wide Web 21(4): 875-914 (2018)
- [j18] Gang Hu, Jie Shao, Zhiyang Ni, Dongxiang Zhang:
  **A graph based method for constructing popular routes with check-ins.** World Wide Web 21(6): 1689-1703 (2018)
- [c40] Lei Wang, Dongxiang Zhang, Lianli Gao, Jingkuan Song, Long Guo, Heng Tao Shen:
  **MathDQN: Solving Arithmetic Word Problems via Deep Reinforcement Learning.** AAAI 2018: 5545-5552
- [c39] Dongxiang Zhang, Mingtao Lei, Xiang Zhu:
  **SAQP++: Bridging the Gap between Sampling-Based Approximate Query Processing and Aggregate Precomputation.** DSC 2018: 258-265

# Entity Resolution

- Wei Wang 0001 — University of Waterloo, David R. Cheriton School of Computer Science,
- Wei Wang 0002 — Nanjing University, State Key Laboratory for Novel Software Technol
- Wei Wang 0003 — Chinese Academy of Sciences, Institute of Microelectronics, Laboratory China (and 4 more)
- Wei Wang 0004 — Fudan University, School of Life Science, Shanghai, China
- Wei Wang 0005 — Zhejiang University, Center for Engineering and Scientific Computation
- Wei Wang 0006 — Language Weaver, Inc.
- Wei Wang 0007 — Chinese Academy of Sciences, ThinkIT Speech Lab, Institute of Acoustic
- Wei Wang 0008 — MIT, Nonlinear Systems Laboratory, Department of Mechanical Engine
- Wei Wang 0009 — Fudan University, School of Computer Science, Shanghai Key Labora
- Wei Wang 0010 — University of California Los Angeles, CA, USA (and 1 more)
- Wei Wang 0011 — The University of New South Wales, School of Computer Science and E
- Wei Wang 0012 — Beijing Jiaotong University, Beijing Key Laboratory of Security and Pri
- Wei Wang 0013 — Peking University, Institute of Computational Linguistics, Beijing, China
- Wei Wang 0014 — Rutgers University, New Brunswick, NJ, USA
- Wei Wang 0015 (aka: Wei Chris Wang) — San Diego State University, CA, USA (and 2 more)
- Wei Wang 0016 — Beihang University, School of Automation Science and Electrical Engi
- Wei Wang 0017 — The Chinese University of Hong Kong, Mechanical and Automation Eng
- Wei Wang 0018 — University of Maryland Baltimore County, Baltimore, MD, USA
- Wei Wang 0019 — University of Naval Engineering, Wuhan, China
- Wei Wang 0020 — Alcatel Lucent Shanghai, Research and Innovation Center, Shanghai, Ch
- Wei Wang 0021 — Zhejiang University, Department of Information Science and Electron
- Wei Wang 0022 — Beijing University of Posts and Telecommunications, Key Laboratory of
- Wei Wang 0023 — Beijing Jiaotong University, School of Electrical Engineering, National
- Wei Wang 0316 — Hong Kong Polytechnic Universitym Hong Kong
- Wei Wang 0317 — Max-Planck-Institut für Physik komplexer Syste
- Wei Wang 0318 — Hebei University of Engineering, School of Infor
- Wei Wang 0319 — Shenzhen University, College of Mechatronics a
- Wei Wang 0320 — China University of Mining and Technology at B
- Wei Wang 0321 — Sichuan University, College of Electrical Enginee
- Wei Wang 0322 — Zunyi Medical University, Fifth Affiliated Hospita
- Wei Wang 0323 — Wuhan University, State Key Laboratory of Infor
- Wei Wang 0324 — University of Southampton, School of Business,
- Wei Wang 0325 — Guangzhou Regenerative Medicine and Health
- Wei Wang 0326 — Hunan Arts and Crafts Vocational College, Yiyar
- Wei Wang 0327 — Hebei GEO University, School of Mathematics a
- Wei Wang 0328 — State Information Center, Department of Inforr
- Wei Wang 0329 — University of Shanghai for Science and Technol
- Wei Wang 0330 — North China Electric Power University, State Key China
- Wei Wang 0331 — Sun Yat-sen University, School of Intelligent Sys
- Wei Wang 0332 — East China Normal University, School of Data Scie
- Wei Wang 0333 — IBM Research, Yorktown Heights, NY, USA
- Wei Wang 0334 — University of Adelaide, Australia

# Handling Redundancy in Data Integration

News Recommendation App

Search Engine

# Agenda

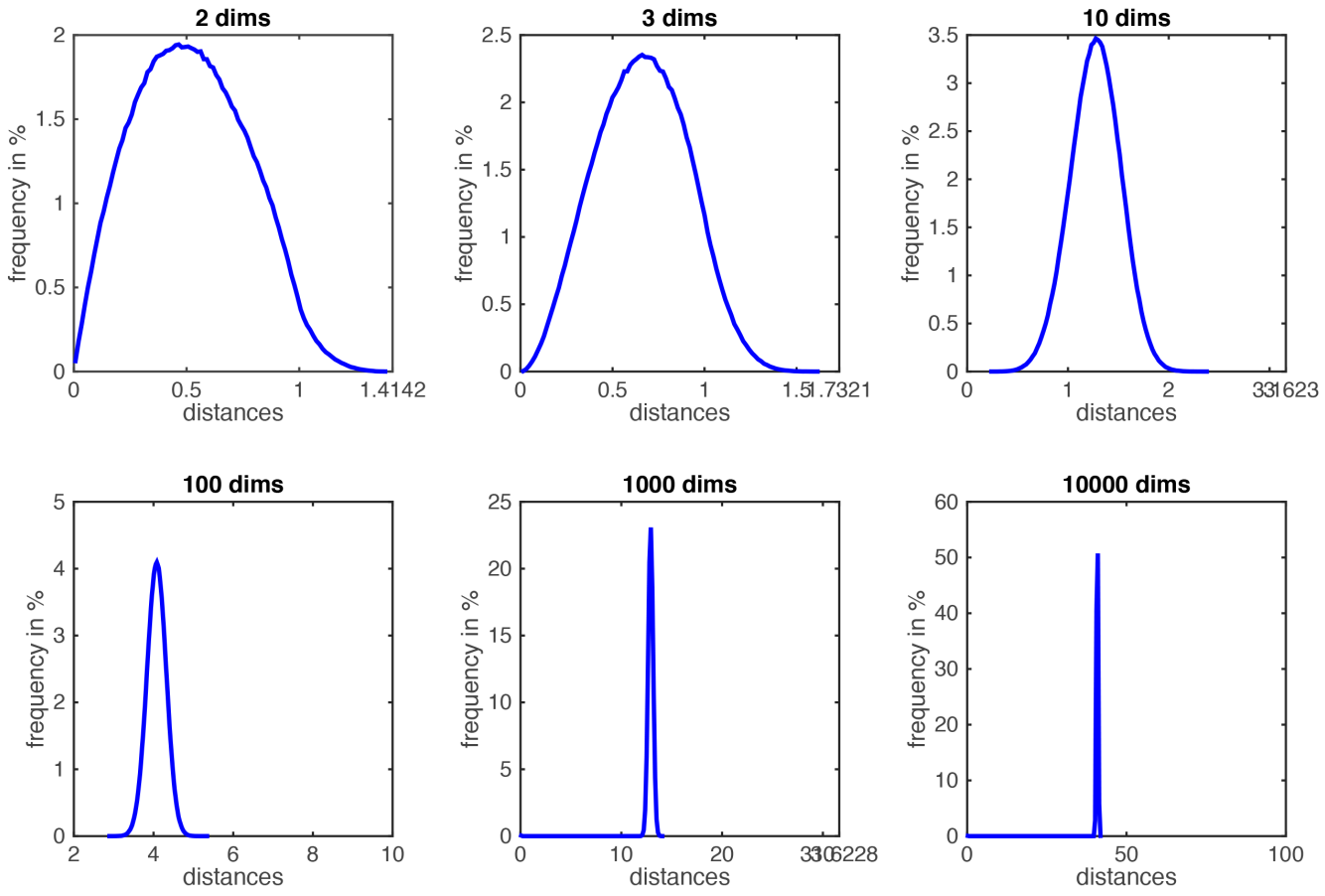
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Summary

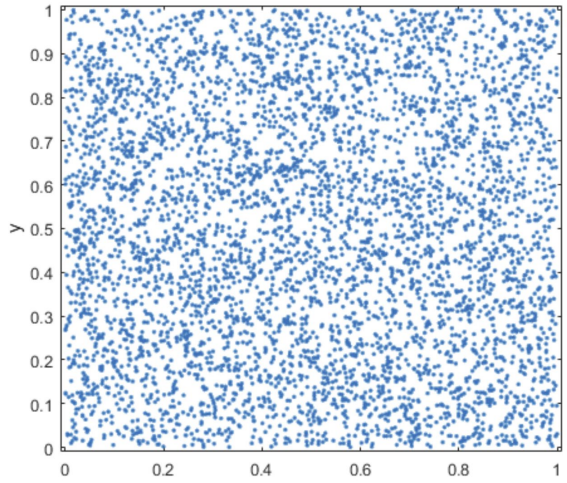
# Data Reduction

- **What is data reduction?**
  - ✓ Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical result.
- **Why data reduction?**
  - ✓ A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

# Data Reduction 1: Dimensionality Reduction

## Curse of dimensionality



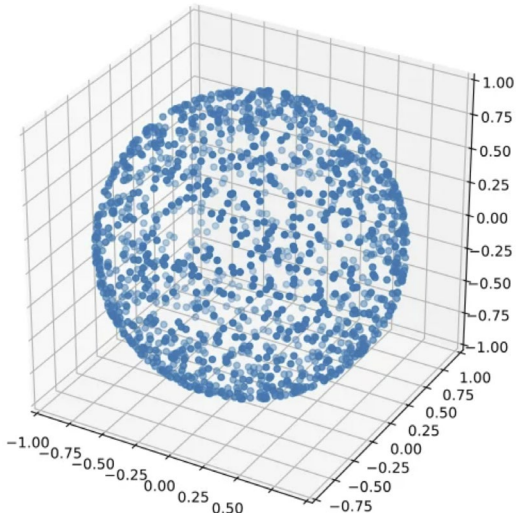



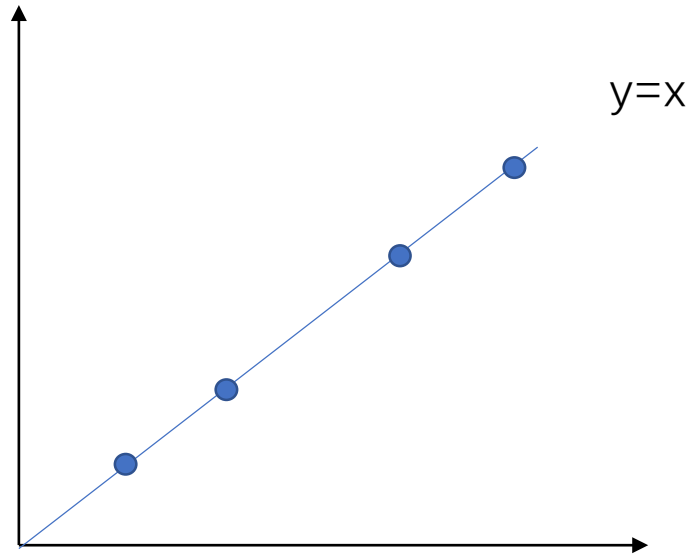


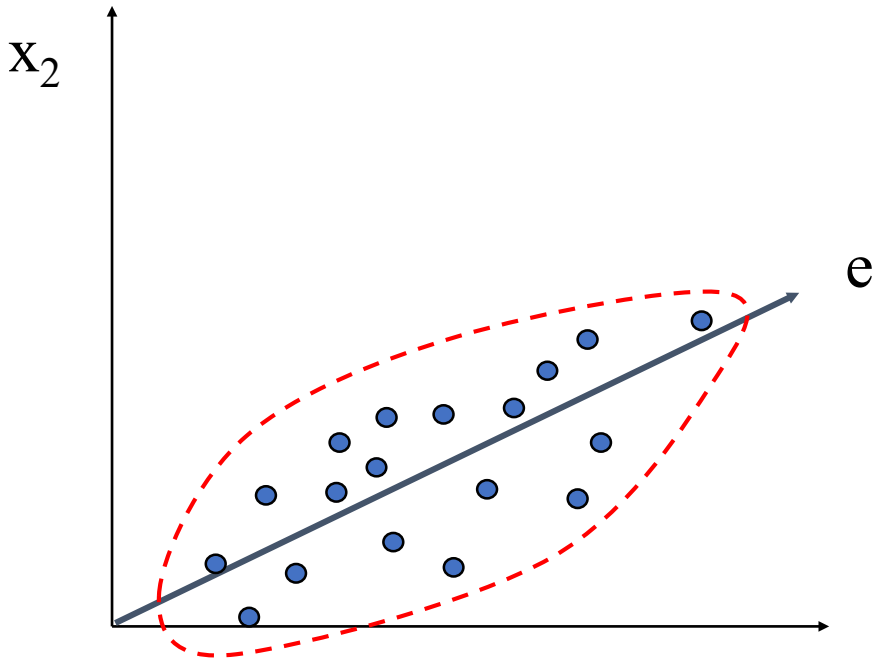


Image source: https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote02_kNN.html

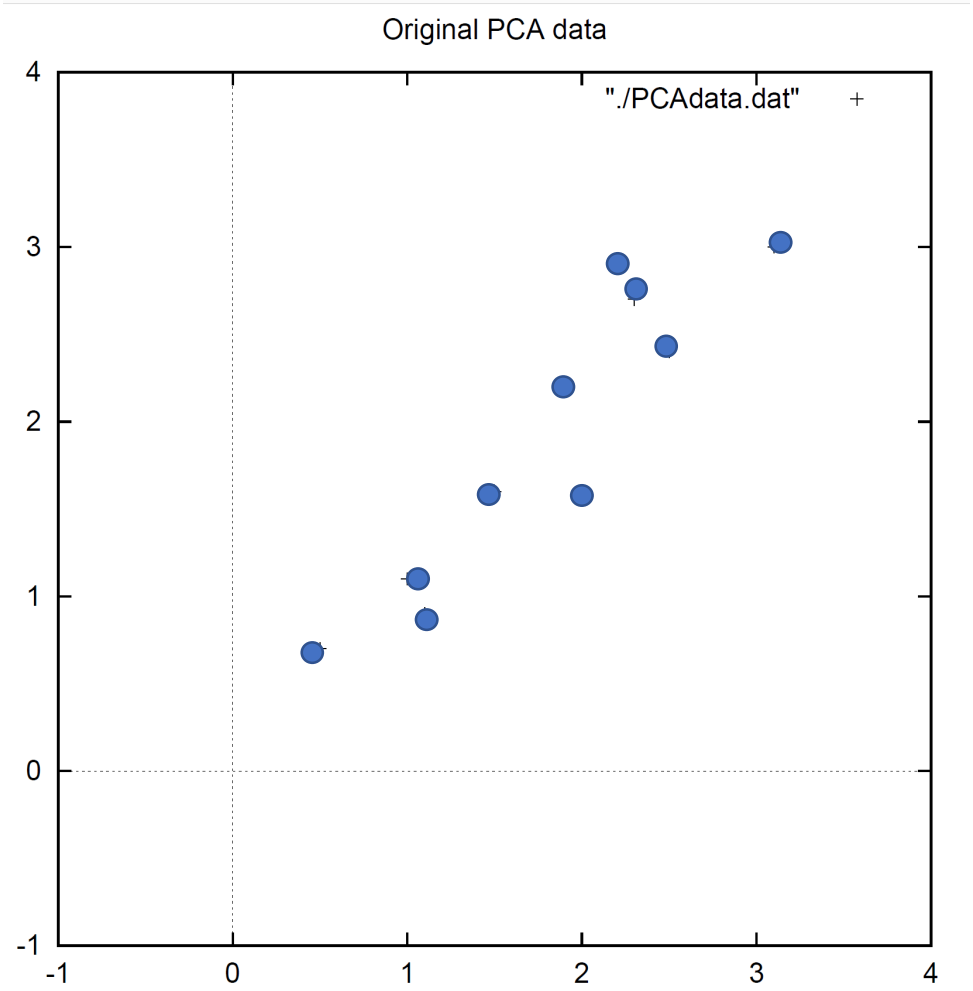
# Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.



# Principal Component Analysis (PCA)




Data =

| $x$ | $y$ |
|---|---|
| 2.5 | 2.4 |
| 0.5 | 0.7 |
| 2.2 | 2.9 |
| 1.9 | 2.2 |
| 3.1 | 3.0 |
| 2.3 | 2.7 |
| 2 | 1.6 |
| 1 | 1.1 |
| 1.5 | 1.6 |
| 1.1 | 0.9 |

DataAdjust =

| $x$ | $y$ |
|---|---|
| .69 | .49 |
| -1.31 | -1.21 |
| .39 | .99 |
| .09 | .29 |
| 1.29 | 1.09 |
| .49 | .79 |
| .19 | -.31 |
| -.81 | -.81 |
| -.31 | -.31 |
| -.71 | -1.01 |

Example from : http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

# Principal Component Analysis (PCA)

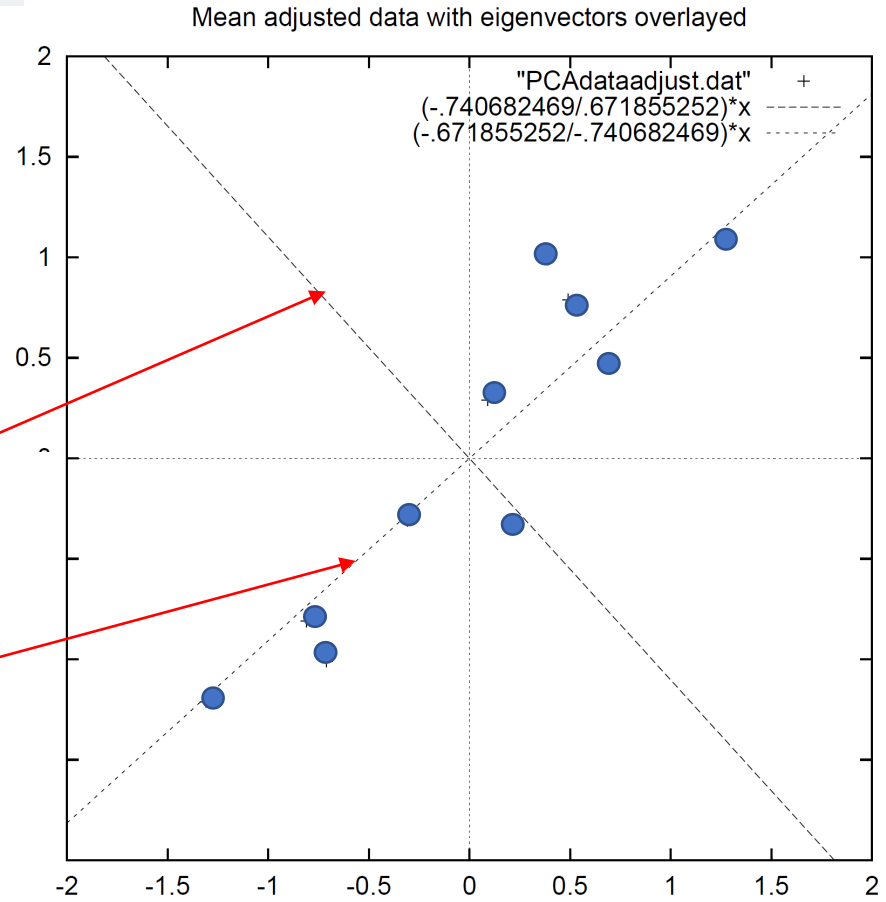
- Calculate the covariance matrix $cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

- Calculate eigenvectors and eigenvalues of the matrix *cov*

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$



# Principal Component Analysis (PCA)

- Remove less significant component

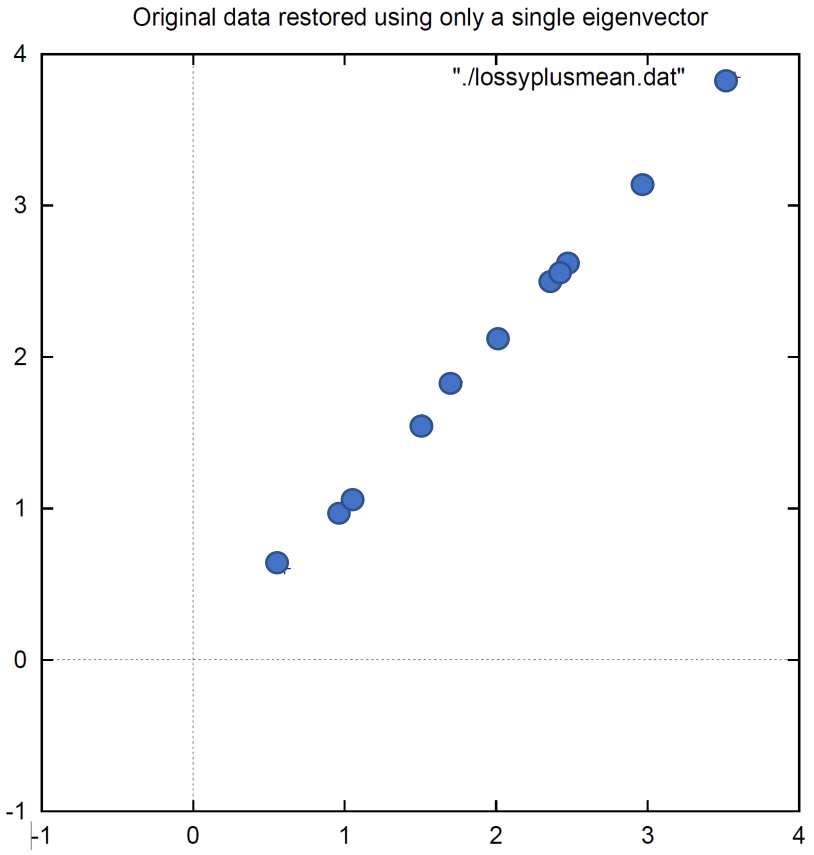
$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

- Get the reduced data

| $x$ | $y$ |
|---|---|
| .69 | .49 |
| -1.31 | -1.21 |
| .39 | .99 |
| .09 | .29 |
| 1.29 | 1.09 |
| .49 | .79 |
| .19 | -.31 |
| -.81 | -.81 |
| -.31 | -.31 |
| -.71 | -1.01 |

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

- Get the data back




$$RowOriginalData = (RowFeatureVector^T \times FinalData) + OriginalMean$$

# PCA for Image Compression

d=1

d=2

d=4

d=8

d=16

d=32

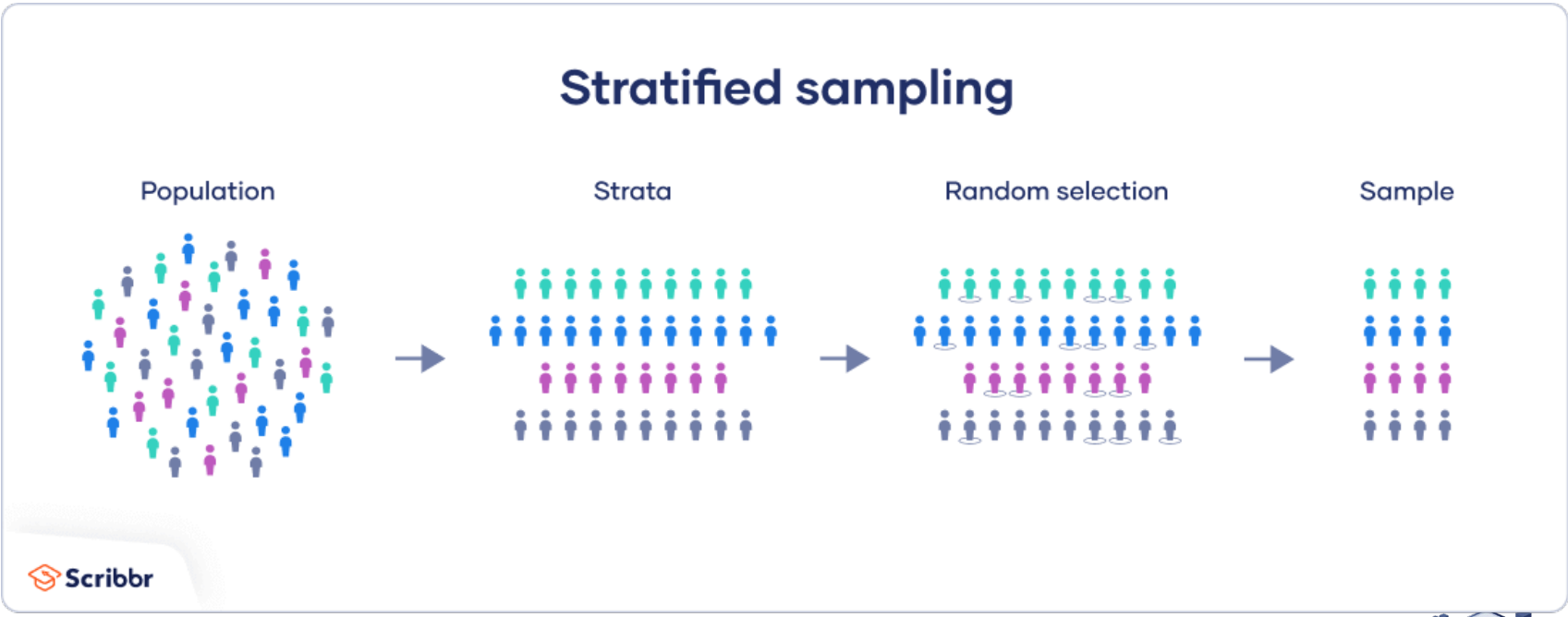
d=64

d=100

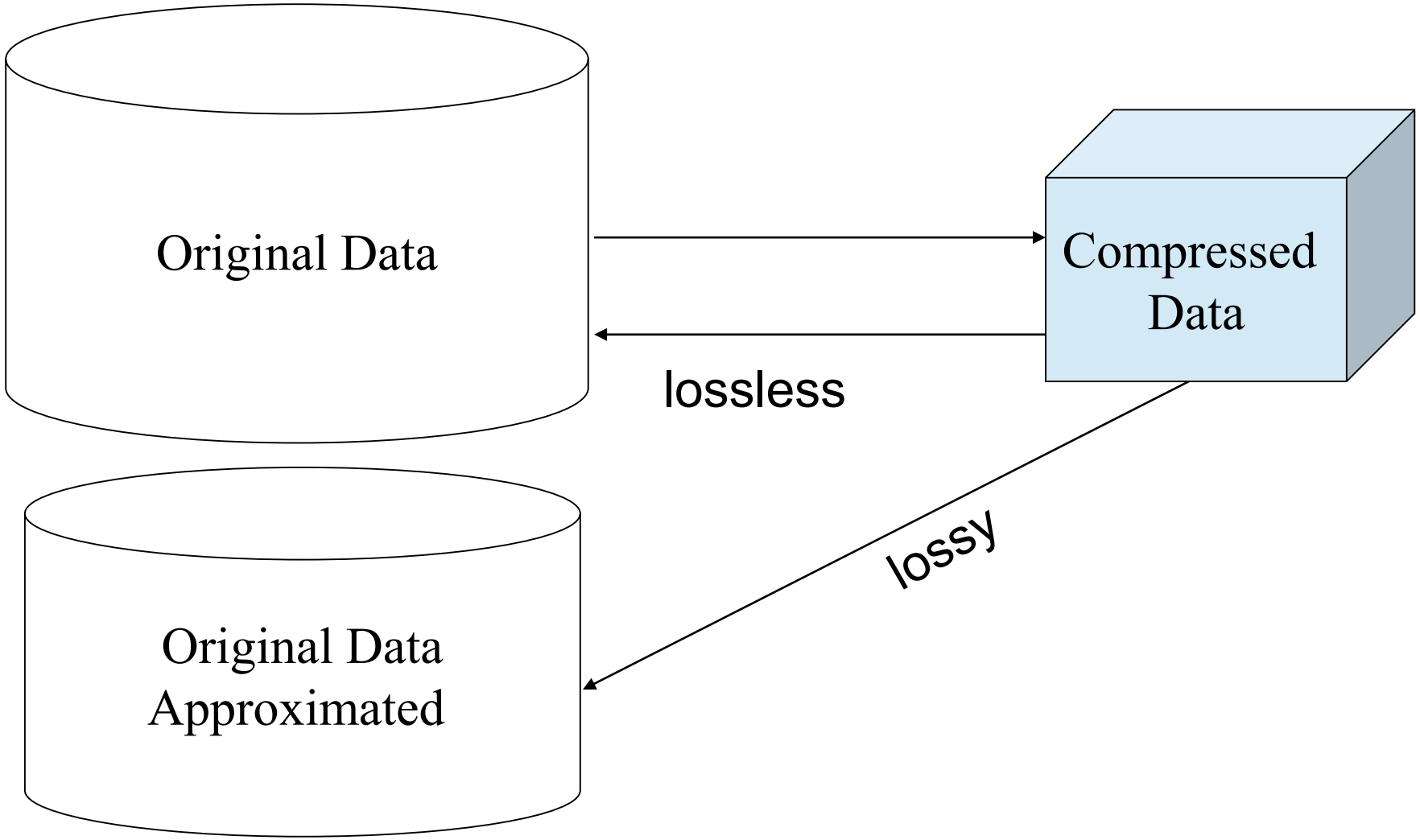
Original Image

# Data Reduction 2: Numerosity Reduction

- **Sampling:**
  - ✓ Obtaining a set of samples to represent the whole data set N
  - ✓ Allow an algorithm to run in complexity that is potentially sub-linear to the size of the data
  - ✓ Simple random sampling may have very poor performance in the presence of skew



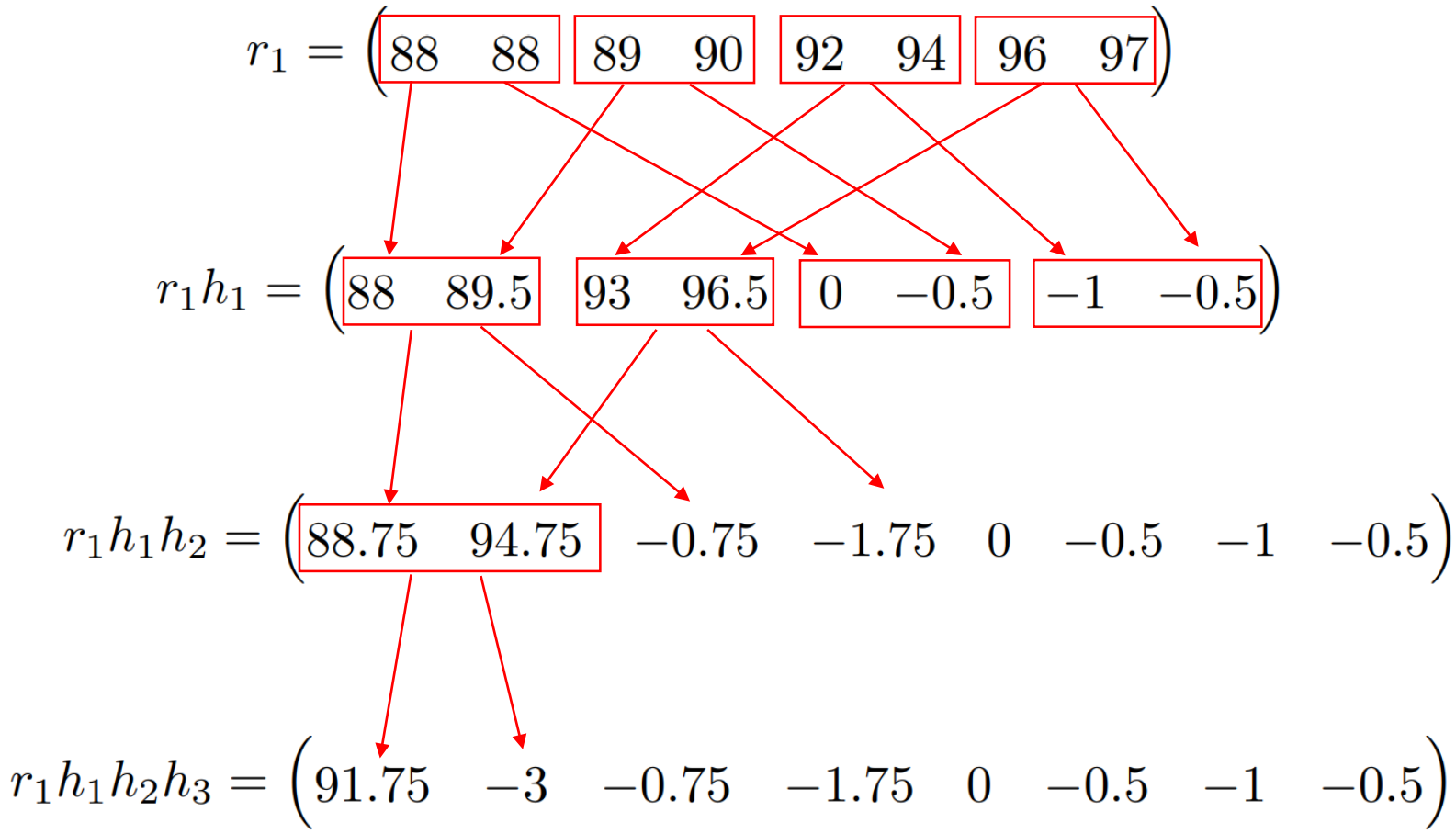

# Data Reduction 3: Data Compression



# Wavelet Transform

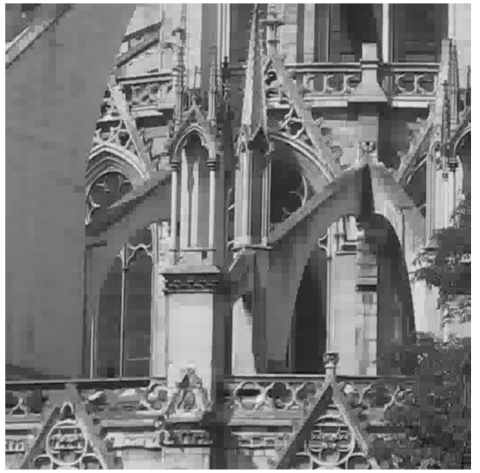
$$A = \begin{pmatrix} 88 & 88 & 89 & 90 & 92 & 94 & 96 & 97 \\ 90 & 90 & 91 & 92 & 93 & 95 & 97 & 97 \\ 92 & 92 & 93 & 94 & 95 & 96 & 97 & 97 \\ 93 & 93 & 94 & 95 & 96 & 96 & 96 & 96 \\ 92 & 93 & 95 & 96 & 96 & 96 & 96 & 95 \\ 92 & 94 & 96 & 98 & 99 & 99 & 98 & 97 \\ 94 & 96 & 99 & 101 & 103 & 103 & 102 & 101 \\ 95 & 97 & 101 & 104 & 106 & 106 & 105 & 105 \end{pmatrix}$$

# Wavelet Transform

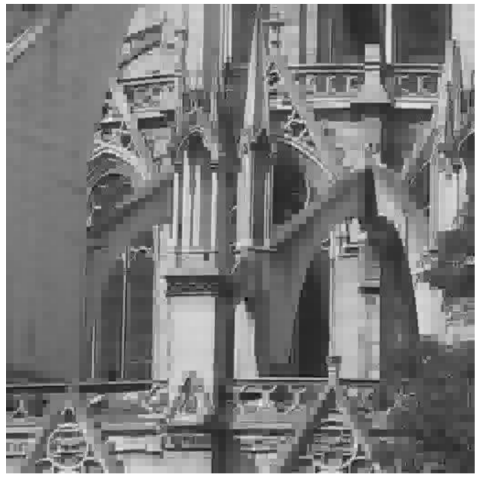


# Wavelet Transform

$$\begin{pmatrix} 96 & -2.0312 & -1.5312 & -0.2188 & -0.4375 & -0.75 & -0.3125 & 0.125 \\ -2.4375 & -0.0312 & 0.7812 & -0.7812 & 0.4375 & 0.25 & -0.3125 & -0.25 \\ -1.125 & -0.625 & 0 & -0.625 & 0 & 0 & -0.375 & -0.125 \\ -2.6875 & 0.75 & 0.5625 & -0.0625 & 0.125 & 0.25 & 0 & 0.125 \\ -0.6875 & -0.3125 & 0 & -0.125 & 0 & 0 & 0 & -0.25 \\ -0.1875 & -0.3125 & 0 & -0.375 & 0 & 0 & -0.25 & 0 \\ -0.875 & 0.375 & 0.25 & -0.25 & 0.25 & 0.25 & 0 & 0 \\ -1.25 & 0.375 & 0.375 & 0.125 & 0 & 0.25 & 0 & 0.25 \end{pmatrix}$$



10:1



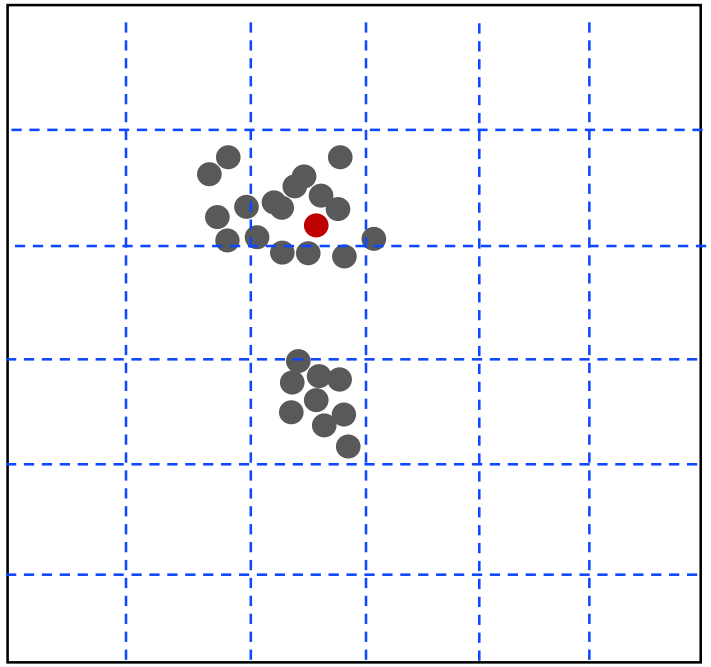
30:1

50:1

# Agenda

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Summary

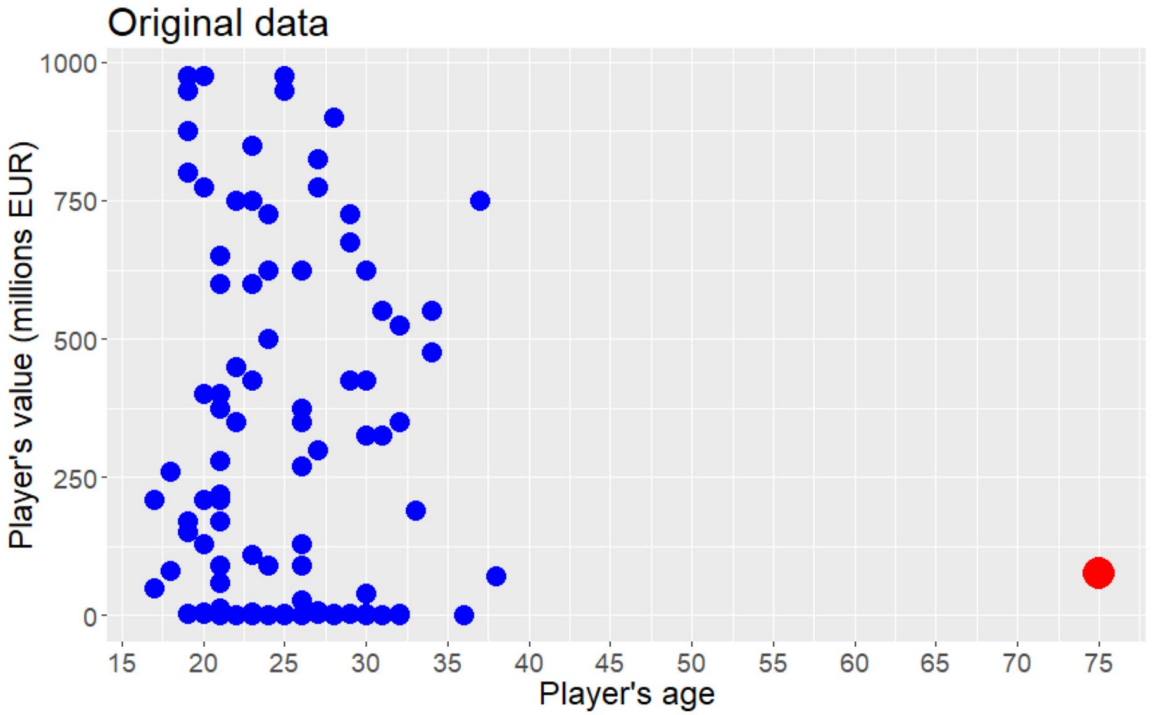
# Grid-Index

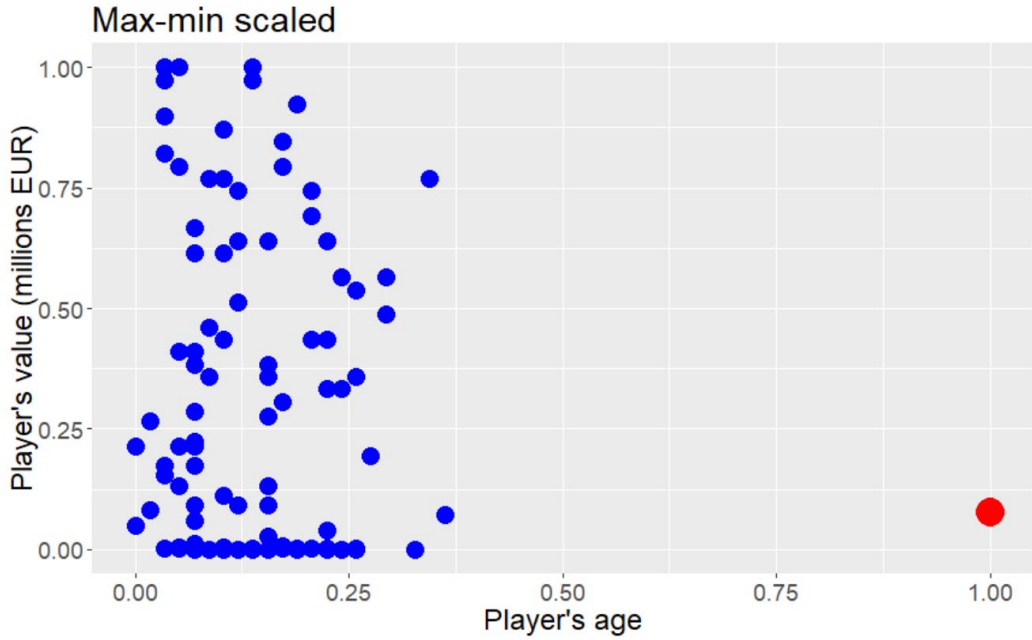



# Normalization: Min-Max Scaling

- **Maps a numerical value x to the [0,1] interval**
  - ✓ Ensures that all features will share the exact same scale
  - ✓ Does not cope well with outliers

$$x' = \frac{x - min}{max - min}$$



# Z-score Normalization

- **Maps a numerical value x to a new distribution with mean=0 and standard deviation=1**
  - ✓ More robust to outliers
  - ✓ Normalized data may be on different scales

$$x' = \frac{x - \mu}{\sigma}$$

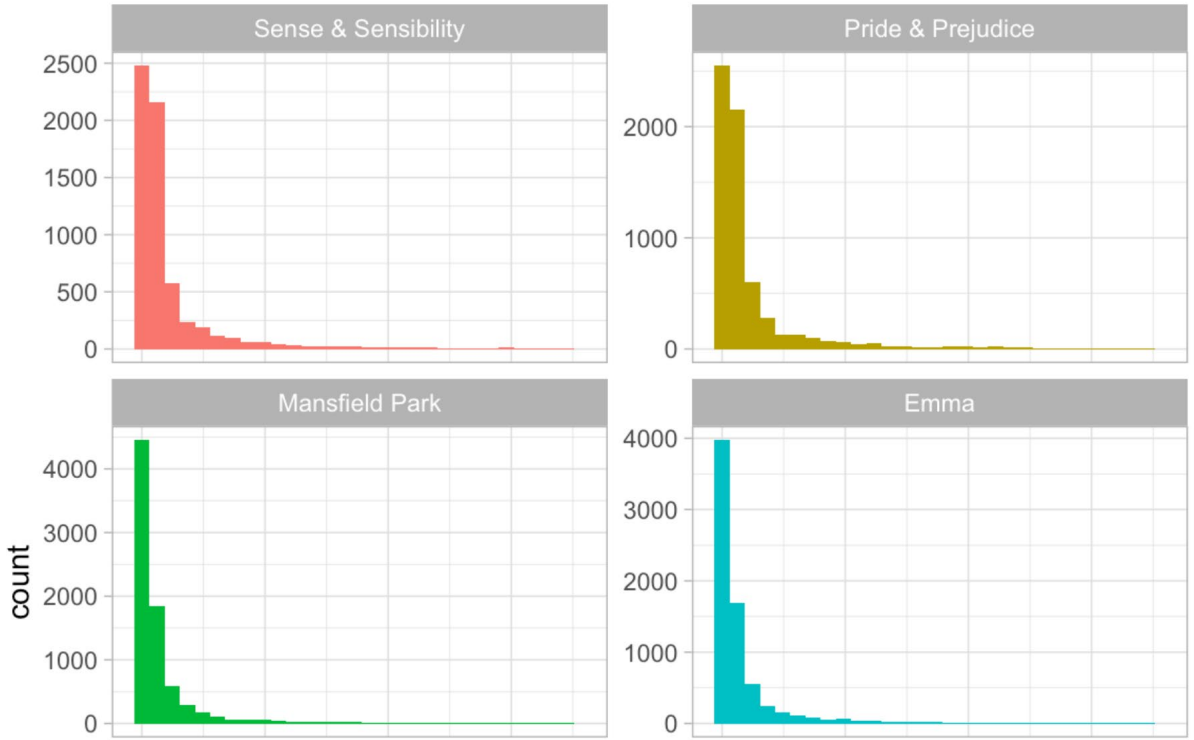
# Term Frequency （Log-based Normalization）

## ☐ Usually follow Zipf distribution

✓ In natural language, there are a small number of very high-frequency words and a large number of low-frequency words

- The log frequency weight of term $t$ in $d$ is defined as follows:

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$




Term frequency distribution in Jane Austen's novels





Image source: https://www.tidytextmining.com/tfidf.html

# Agenda

- ☐ Data Cleaning
- ☐ Data Integration
- ☐ Data Reduction
- ☐ Data Transformation
- ☐ Summary

# Summary

- **Data cleaning**
  - ✓ Missing data, noisy data, inconsistent data
- **Data integration**
  - ✓ Schema integration, entity resolution, duplicate removal
- **Data reduction**
  - ✓ Dimensionality reduction, numerosity reduction, data compression
- **Data transformation**
  - ✓ Normalization, discretization