資料分析與學習基石
Fundamental of Data Analytics

Hung-Yu Kao (高宏宇)
***I**ntelligent **K**nowledge **M**anagement Lab*

Institute of Medical Informatics,
Dept. of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan

Hung-Yu Kao (高宏宇)

- 成功大學資訊工程系教授
- 成功大學電機資訊學院副院長
- 中華民國 人工智慧學會 常務監事
- 中華民國 計算語言學會 副理事長
- 人工智慧學校講師

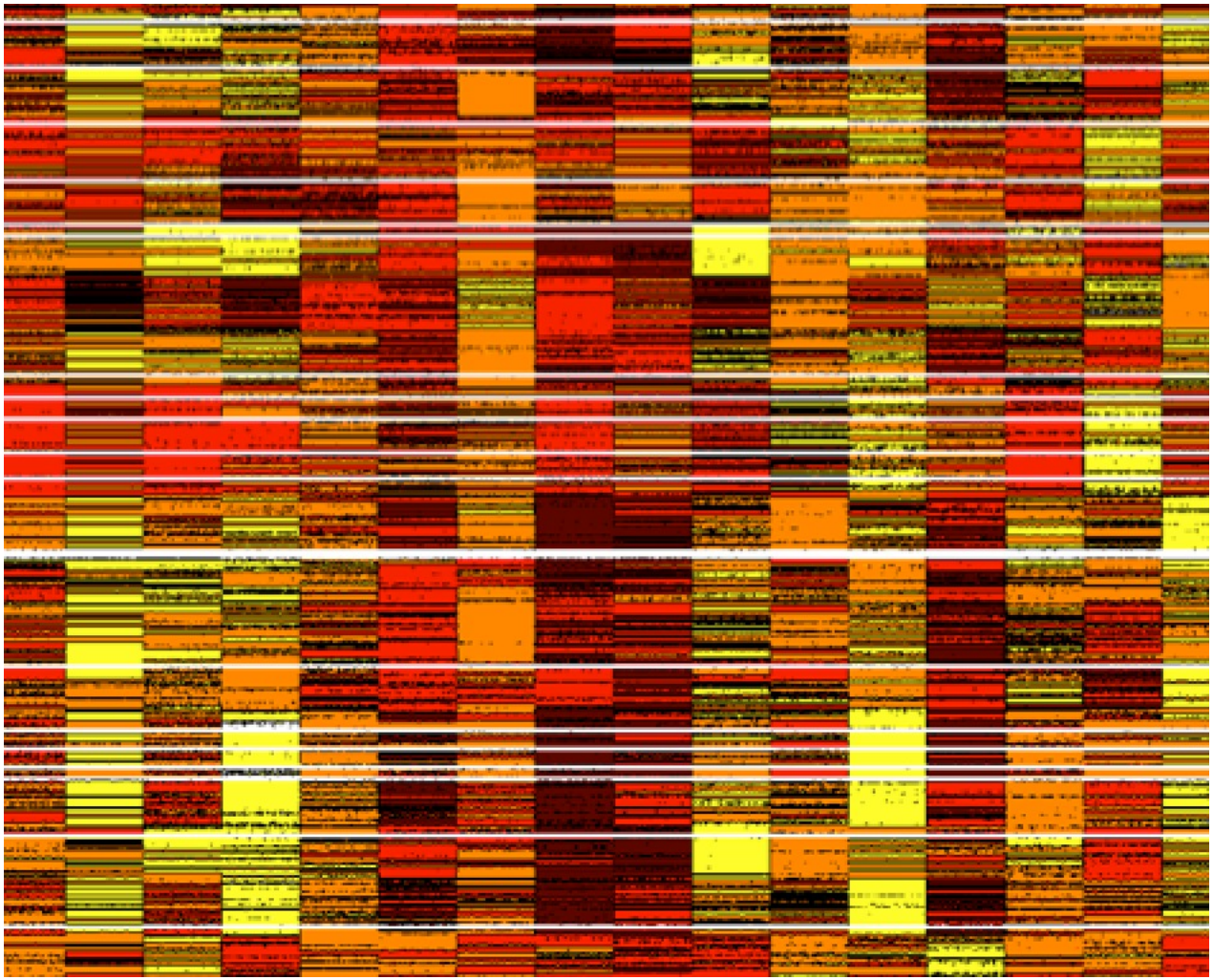
Data Competition Award

- BioCreative, Rank 1, 2010, 2011, 2015
- TREC Blog, Rank 3, 2010
- ACL SemEval **Rumor Detection Competition**, Score Rank1, 2017
- ACL SemEval Argument Reasoning Comprehension Test, Rank 2, 2018
- CIKM AnalyticCup Short Text Matching, Rank 2, 2018
- WSDM Fake News Classification, Rank 3, 2019
- ACL/Google AI Gender Bias for Natural Language Processing Rank 4, 2019
- WSDM **Visual Question Answering** Challenge, Rank 4, 2022

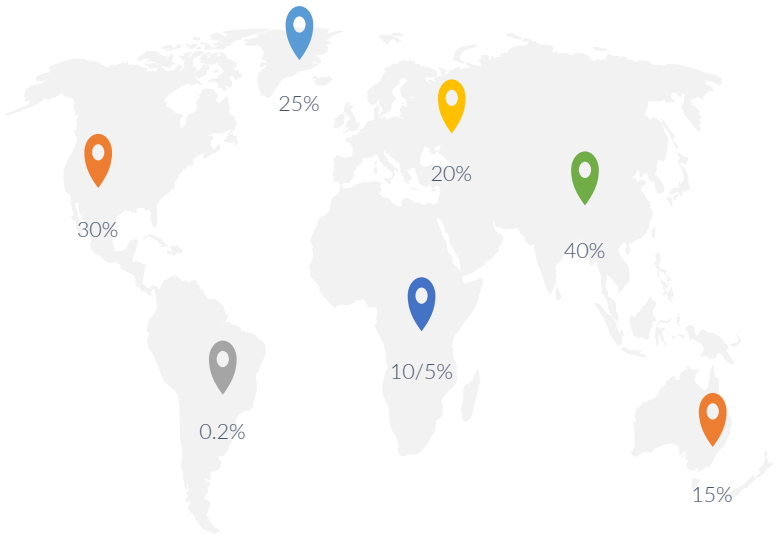
NLP related publication (2019-2021)

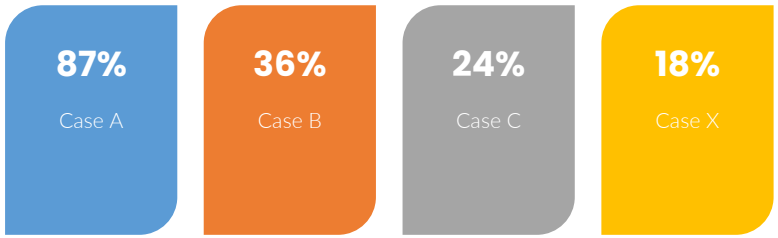
- ACL x 3, ACL workshop x 2, Coling x 2, AAAI x 1, AACL x 2, WSDM x 1, ACML x 1,

What is this? What do you get from this figure?

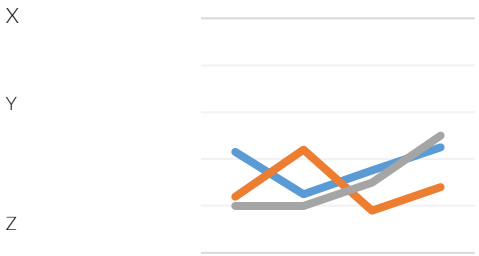


Distribution A

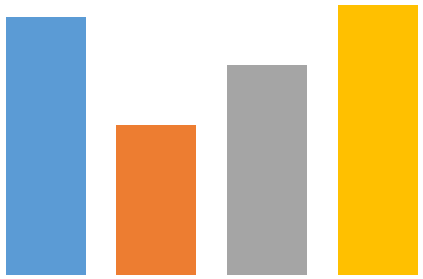




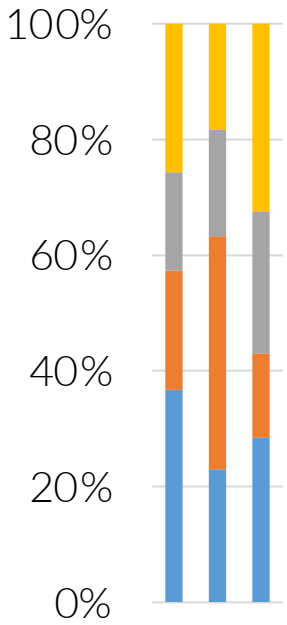
Distribution B



Distribution C

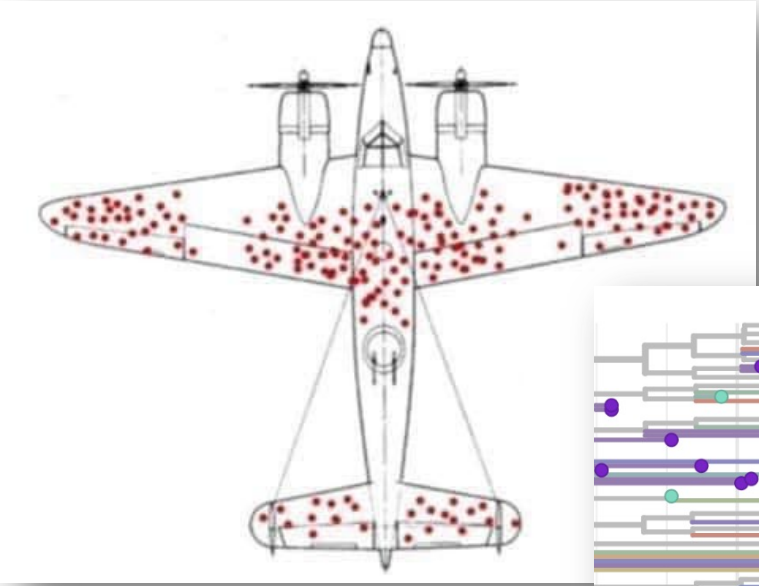


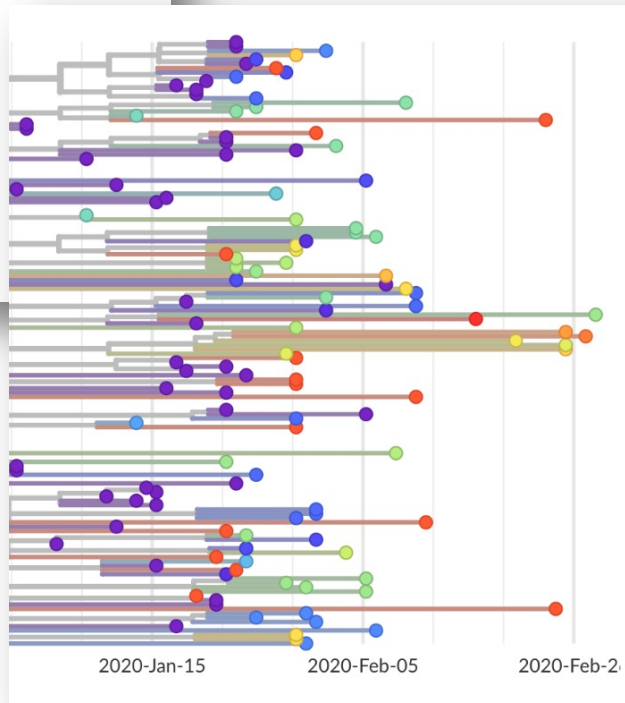
Task A

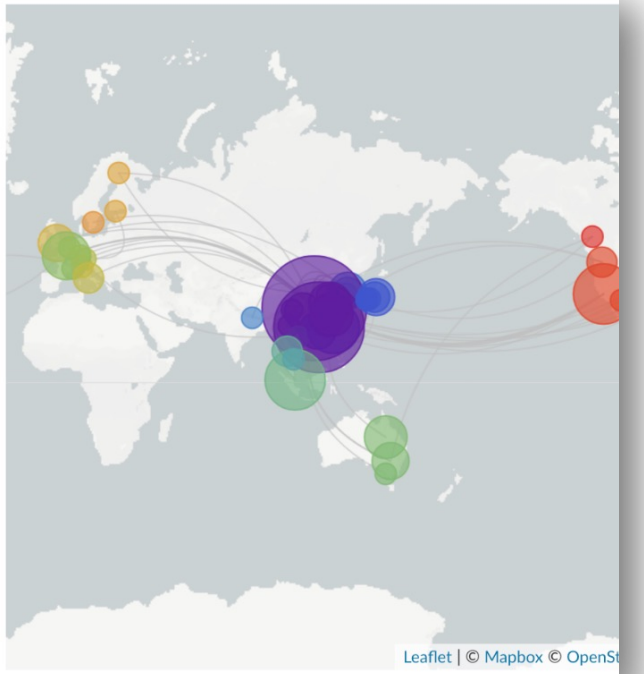


Task B

Data Analytics, everywhere







Data Analytics, everywhere

Data

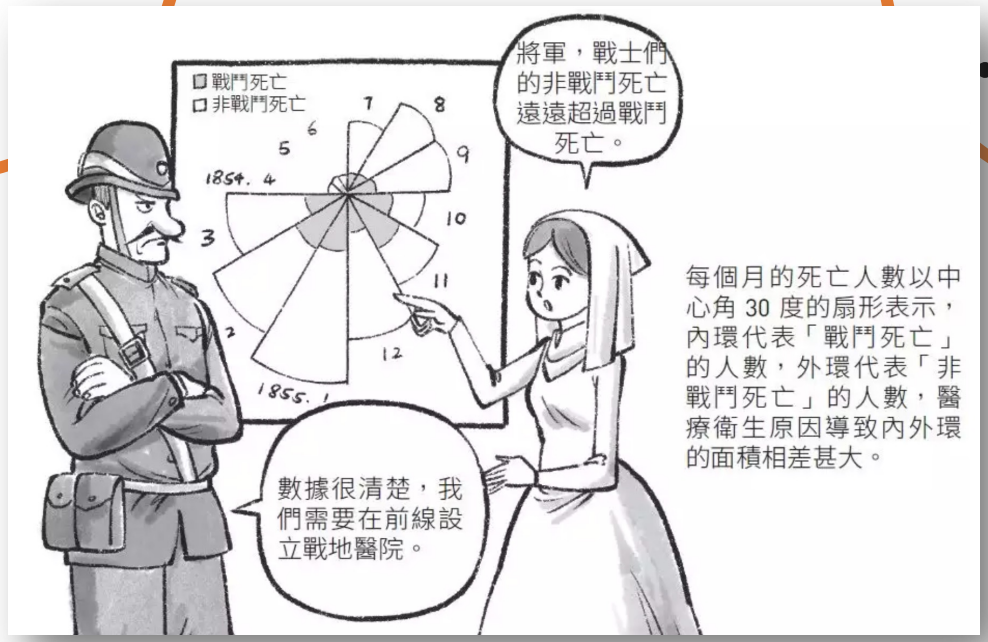
- Database
- (Big) Data processing
- (A)IOT

Data Analytics

Value !

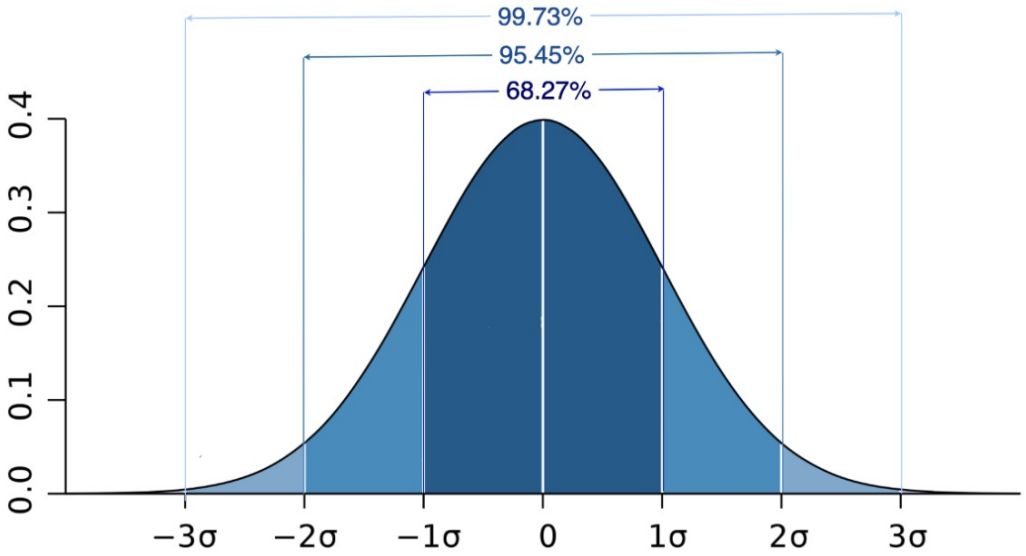
Analytics

- SQL
- Statistics
- Machine Learning
- AI



歷史上第1張「極座標圓餅圖」（polar area diagram）

Is this course related to Statistics?



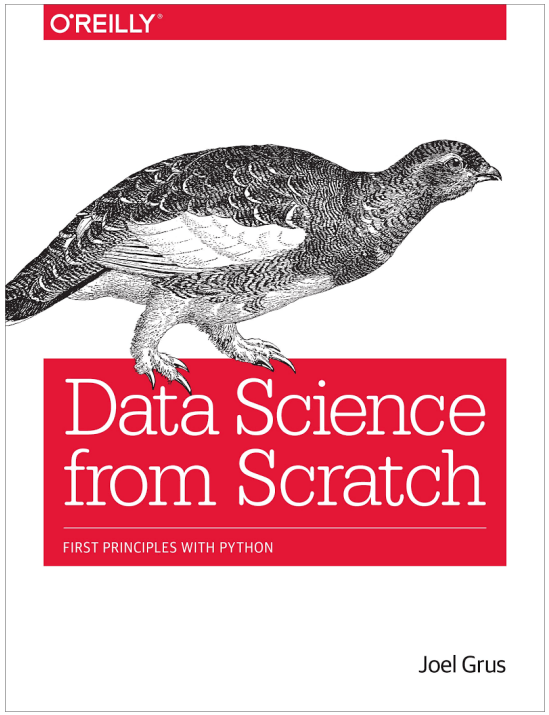
$$\text{t-score} = \frac{\bar{d} - \mu_d}{\text{SE}/\sqrt{n}} = \frac{74 - 0}{13.2/\sqrt{n}}.$$

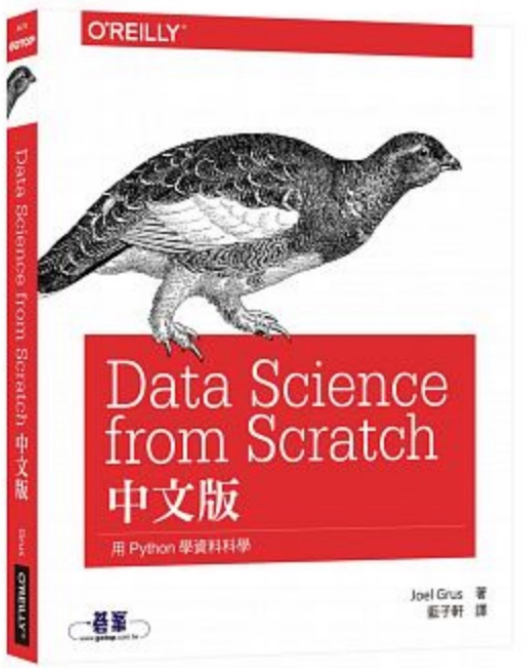
In this course

- Basics
  - Python usage
  - Data Visualization
  - Related Linear Algebra / Statistics / Probability (few)
- Inference & Learning
  - Gradient Descent
  - Data Processing
  - Machine Learning
    - KNN, NB, Regression, Decision Trees, NN
  - Clustering
  - NLP
  - Network Analysis

Text book & References

- **Data Science from Scratch: First Principles with Python, Joel Grus**





Schedule

Lecture

Tutorial

Reporting

Exam

Week	*Topics*
1 (2/16)	Syllabus
2	Data Analytics Fundamental, Homework 1
3 (3/2)	Python Basics for data analytics (tutorial)
4	Classification / Homework 2
5	Classification / Evaluation
6	Classification / Evaluation / Homework 3
7 (3/30)	AI CUP 2023 (tutorial)
8 (4/6)	Vacation
9	Natural Language Processing
10	Natural Language Processing
11	AI CUP Competition Report (1/2)
12	Final Project (tutorial)
13	Unsupervised Learning
14 (5/18)	Unsupervised Learning / Homework 4
15 (5/25)	Exam
16 (6/1)	AI CUP Competition Report (2/2)
17 (6/8)	Final Project Report (1/2)
18 (6/15)	Final Project Report (2/2)

Grading

- Homework
  - single x **4** 45 %
  - group x **1** 20% (from kaggle / Industry)
    - Some selected datasets from Kaggle
    - Kaggle Kernel construction
    - From your own
    - #group members = 3 ~ 4
- **1** midterm: 15 %
- **1** competition (AI CUP): 20 % (group)

Instructors and TAs

- *Instructors:*
  - *Hung-Yu Kao 資訊系館 12F, Room 65C11 (hykao@mail.ncku.edu.tw)*
  - *https://ikmlab.csie.ncku.edu.tw/advisor.html*
- *TAs:*
  - *資訊系館 9F, Room 65903 IKM lab.*
  - *nckudm@gmail.com*

- *Course website NCKU Moodle*