香港浸會大學
HONG KONG BAPTIST UNIVERSITY

Data Science at a Glance

JOUR7280/COMM7780

Big Data Analytics for Media and Communication

Instructor: Dr. Xiaoyi Fu

Agenda

- Data Science and Data Scientists
  - Who are the data scientists?
  - The Data Science Pipeline
    - Media and Communication in the Digital Age
    - Data Processing
    - Finding a Story
    - Presenting a Story
  - What is Computational Thinking?

Data Science and Data Scientists

Data Science

- Data science is multi-disciplinary
  - Statistics
  - Computer Science
  - Visualization and communication

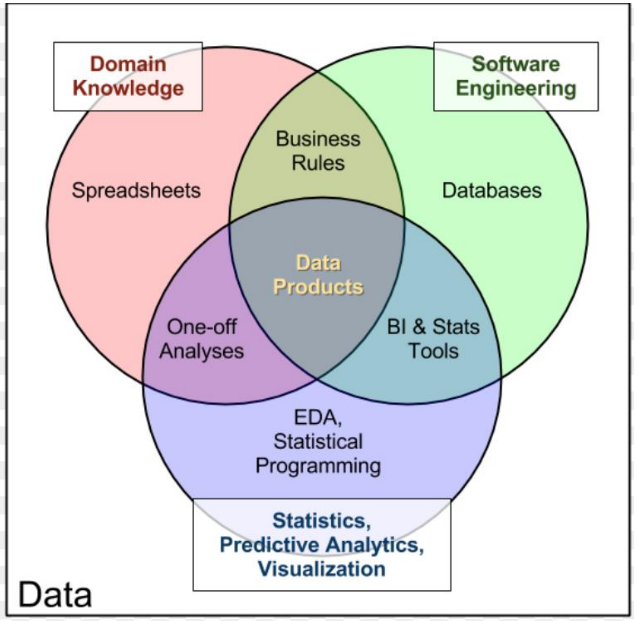
- Goal:
  - Extract knowledge and find insights from data
    - Numeric
    - Textual
    - Multimedia

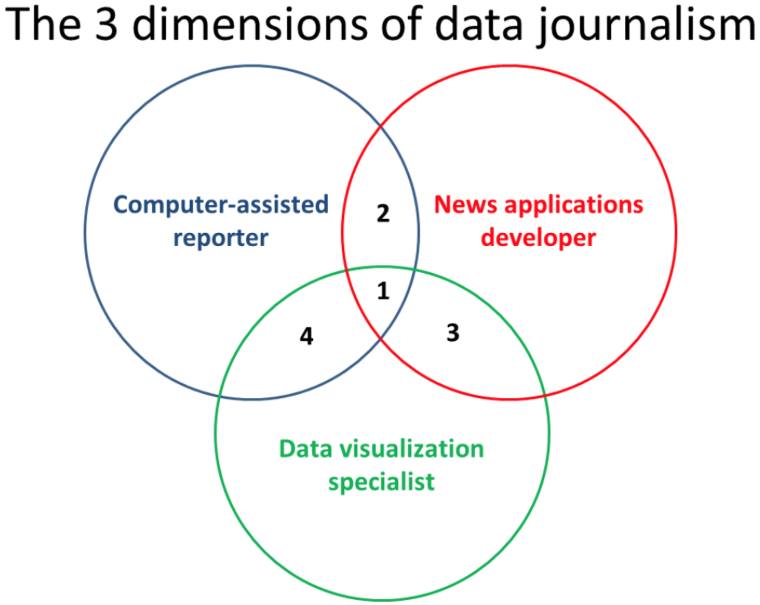
Who are they…?

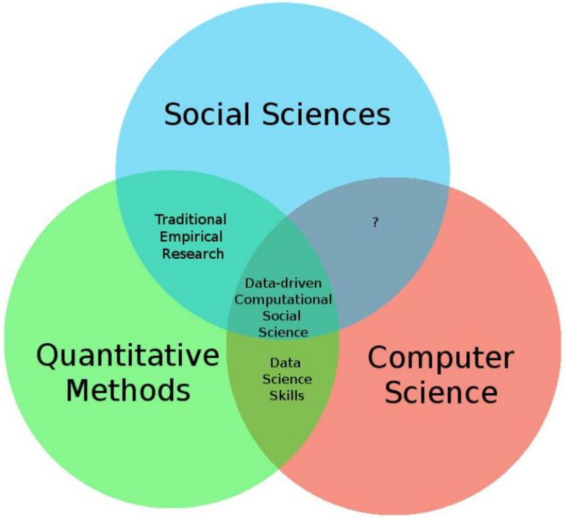
- data science and data scientists
- computational social science
- computational communication research
- digital humanities
- data-driven journalism
- computational journalism
- programmer journalism
- social informatics
- business analytics
- big data analytics
- social media analytics
- …

- 我分析42万字的歌词，为搞清楚民谣歌手们在唱些什么
- 《邪不压正》到底怎么样？我爬取了上万条网友评论进行分析
- 4天13亿！Python告诉你爆火的《我不是药神》到底神在哪里
- 3500种中西药品说明书对比
- 我们从爬取1000亿个网页中学到了什么
- 用爬虫分析互联网大数据行业薪资情况
- Tutorial: Web Scraping Hotel Prices using Selenium and Python
- How to scrape TripAdvisor.com for Hotel Data, Pricing and Reviews using Python
- …

Inter-(multi-)disciplinary Areas




The 3 dimensions of data journalism
Computer-assisted reporter
2
News applications developer
1
4
3
Data visualization specialist



Domain knowledge + motivations

- Domain knowledge
  - Theories from discipline areas
    - Journalism, advertising, marketing, management, sociology, political science, arts,…
  - Issues and topics
    - sports, fashion, popular culture, folk music, movies, cartoons, cuisine,…
- Motivations
  - "When there is a gap between the ideal situation and the reality, there is an investigation."
  - "Trend? correlation? outliers?"

Foremost: A question

> “*The most important thing in data science is the question;*
>
> *The second most important is the data;*
>
> *Often the data will limit or enable the questions;*
>
> *But having data can’t save you if you don’t have a question.* ”
>
> *Jeffrey Leek, JHU*

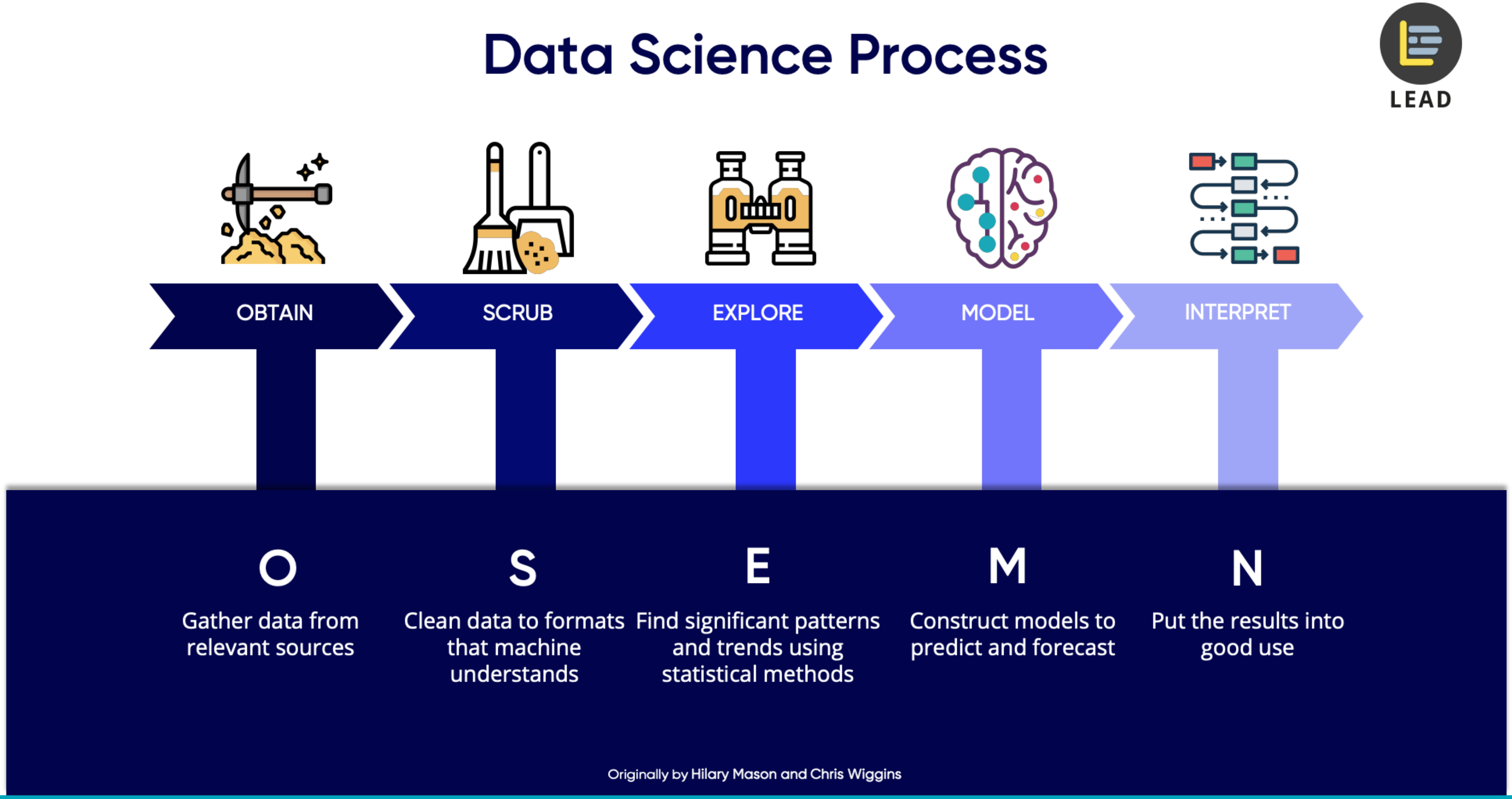
The Pipeline

Data science in action

- **Define the problem**
- Scouting the data sources
- Accessing to and collecting the data
- (Pre-)processing and cleaning the data
- Exploring the data
- Analyzing the data
- Interpreting the results
- Offering insights and solutions
- …

Data science pipeline - a verbal explanation

- The “OSEMN Pipeline”



Data Science Process

O Gather data from relevant sources

S Clean data to formats that machine understands

E Find significant patterns and trends using statistical methods

M Construct models to predict and forecast

N Put the results into good use

Originally by Hilary Mason and Chris Wiggins

Media and Communication in the Digital Age

Digital Media

- Digital Media: when the information is encoded in digital format
  - Numerical
  - Textual
  - Digital Image
  - Digital Video
  - Social information
  - Networks
  - etc.

Digital Communication

- Delivering digital media information through digital channels
  - Documents
  - Websites
  - Streaming Media
  - Social Networks Platforms
  - etc.
- Who is delivering the information
  - Journalists
  - Business Communicators
  - Private Enterprises and Public Sector Organisations

Examples

- Data Journalism
  - America is more diverse than ever — but still segregated [Link]
  - Fivethirtyeight: Lionel Messi Is Impossible [Link]
  - SCMP: Brexit, how Britain voted [Link]
  - NYT: The Scale of the President's Budget [Link]
  - HUFFPOST: Gun Ownership [Link]
  - "Unfounded" - The Globe and Mail, 2017 [Link]
    - [Data Journalism Award, 2017, Investigation of the Year]
  - Data Journalism Award 2020 Winners [Link]
  - WHO: Malaria Report 2018 [Link]
  - PwC Fintech Report 2017 [Link]

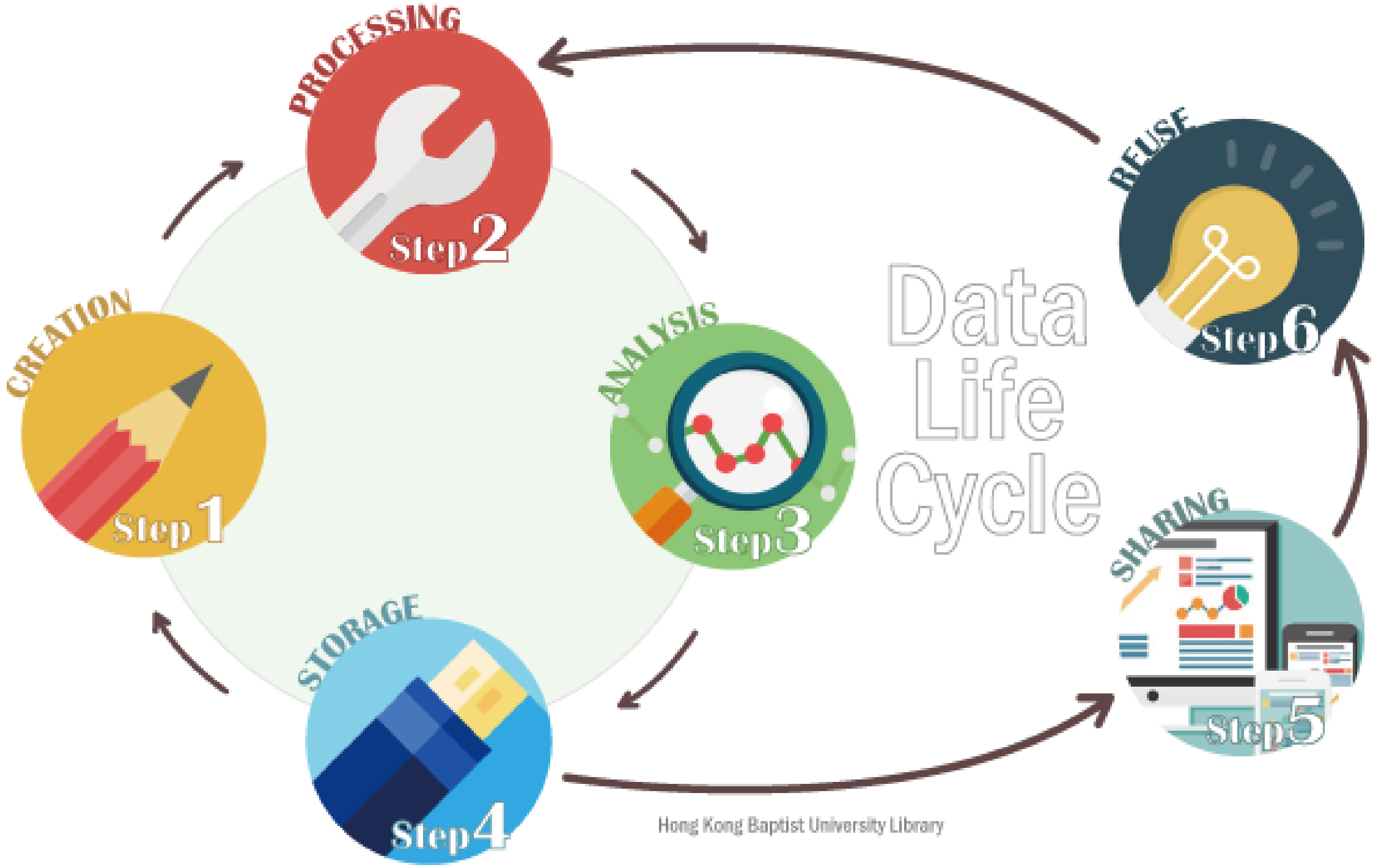
Data Processing

Data processing

- What is data processing?
    - Preprocessing, processing, data cleaning, data scrubbing
    - Data processing
        - Transfer the collected data for further statistical analysis and interpretation
    - Characteristics
        - Reproducible
        - Transparent
        - Automated

Python and Data Life Cycle

- Guide at the HKBU library: https://hkbu.libguides.com/data-analytics



Hong Kong Baptist University Library

Python and Data Life Cycle

Data life cycle	Major tasks	"Non-coding-based"	The role of Python
1. Data creation	Data collection (via fieldworks or simulation)	Netlogo	Data simulation Web scraping
2. Data processing	Data cleaning	OpenRefine	Python programming
3. Data analysis	Various types of data analysis (data exploration, data visualization, statistical analysis…)	SPSS, STATA, EXCEL, Gephi	Python, Numpy, Scipy, Pandas, Matplotlib
4. Data storage	Data archiving and version control	Git	
5. Data sharing	Open data via data repositories	GitHub	/
6. Data re-use	Re-using the data	/	/

Python and Data Life Cycle

Data life cycle	Major tasks	“Non-coding-based”	The role of Python
1. Data creation	Data collection (via fieldworks or simulation)	Netlogo	Data simulation Web scraping
2. Data processing	Data cleaning	OpenRefine	Python programming
3. Data analysis	Various types of data analysis (data exploration, data visualization, statistical analysis…)	SPSS, STATA, EXCEL, Gephi	Python, Numpy, Scipy, Pandas, Matplotlib
4. Data storage	Data archiving and version control	Git	
5. Data sharing	Open data via data repositories	GitHub	/
6. Data re-use	Re-using the data	/	/

Finding a Story

Defining a problem/finding a story

- Cases and issues
- The “beat”
    - Check the course offered by any journalism school: “Beat reporting”
- Domain knowledge

Defining a problem/finding a story

- What have been found?
- What have been “covered?”

Defining a problem/finding a story

- An investigation often arises when a reporter perceives a difference between what is (the observed reality) and what should be (as articulated in law or policy) (Broussard, 2015);
- A high-impact investigative story looks at a situation where what is differs from what should be, and explains why (Broussard, 2015).
- *"When there is a difference between the ideal case and the reality, there is an investigation."*

Defining a problem/finding a story

- Alexis Ulrich: Using Data Journalism to Generate Content Ideas [URL]
- Other quick thoughts:
  - Entertainment: the producer-celebrity relationships? the contents of the lyrics?
  - Education: tuition fee? educational outcomes? articulation rates?
  - Society and technology: the "Python mania" and knowledge gaps?
  - Medical and public health
  - Sports: most likely outlier stories?

Presenting a Story

Presenting a theory/story

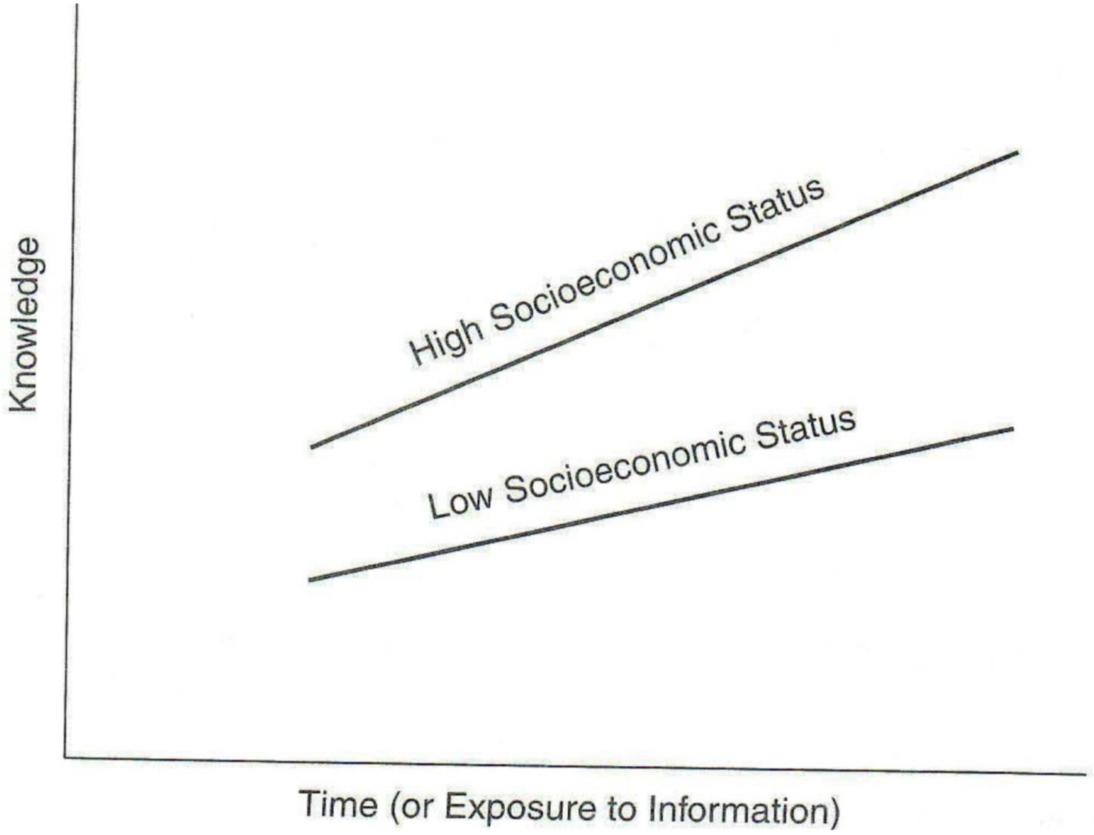
- Three ways to present a theory (perhaps a news story as well)
- The knowledge gap hypothesis (Tichenor, Donohue, & Olien, 1970)
  - The **knowledge gap hypothesis** explains that knowledge, like other forms of wealth, is often differentially distributed throughout a social system. (wiki)
  - Before Social Networking apps!

Presenting a theory/story

1. In written texts:
   - As information diffuses into a society, members of privileged sectors will learn knowledge at a faster rate than members of less-privileged sectors.

Presenting a theory/story

2. In a graphical illustration (data visualization, infographic, information visualization)



Presenting a theory/story

3. In a mathematical formula
   - Knowledge = K
   - Time = T
   - Social Eco' Status = S

$$K = b_0 + b_1 * T + b_2 * S + b_3 * T * S$$

- All the three ways are presenting the same theory.
- For a news story, it may also be able to present in three different ways.

Presenting a theory/story: from data exploration

- Finding the “stories” (by Jonathan Stray)
  - The “outlier stories”: An outlier is a value that is different from all the others.
  - The "trend stories": A trend is a pattern through time.
  - The “correlation stories”: A correlation is when two variables change together.

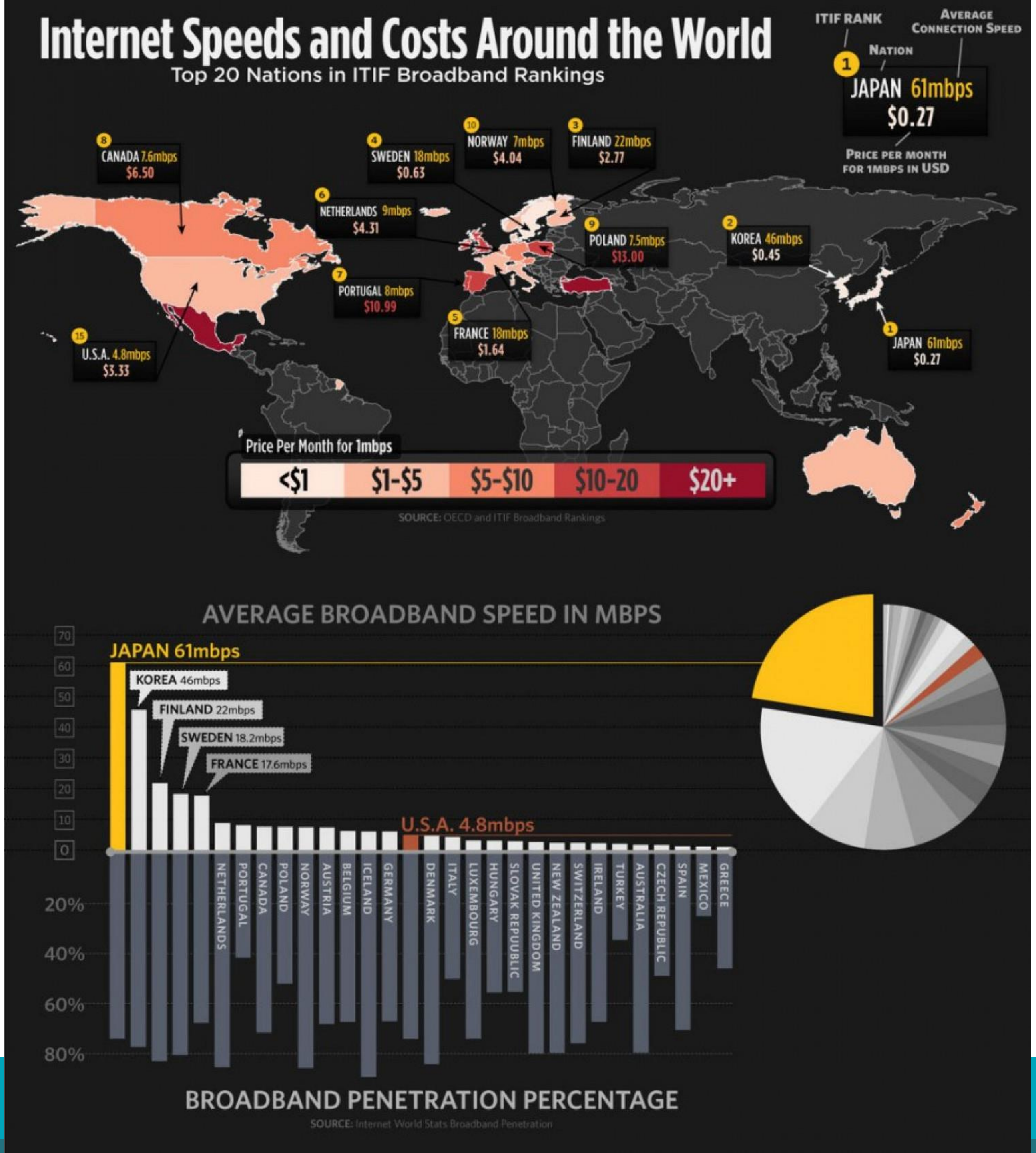
- *Exercise: Find an example for each category*

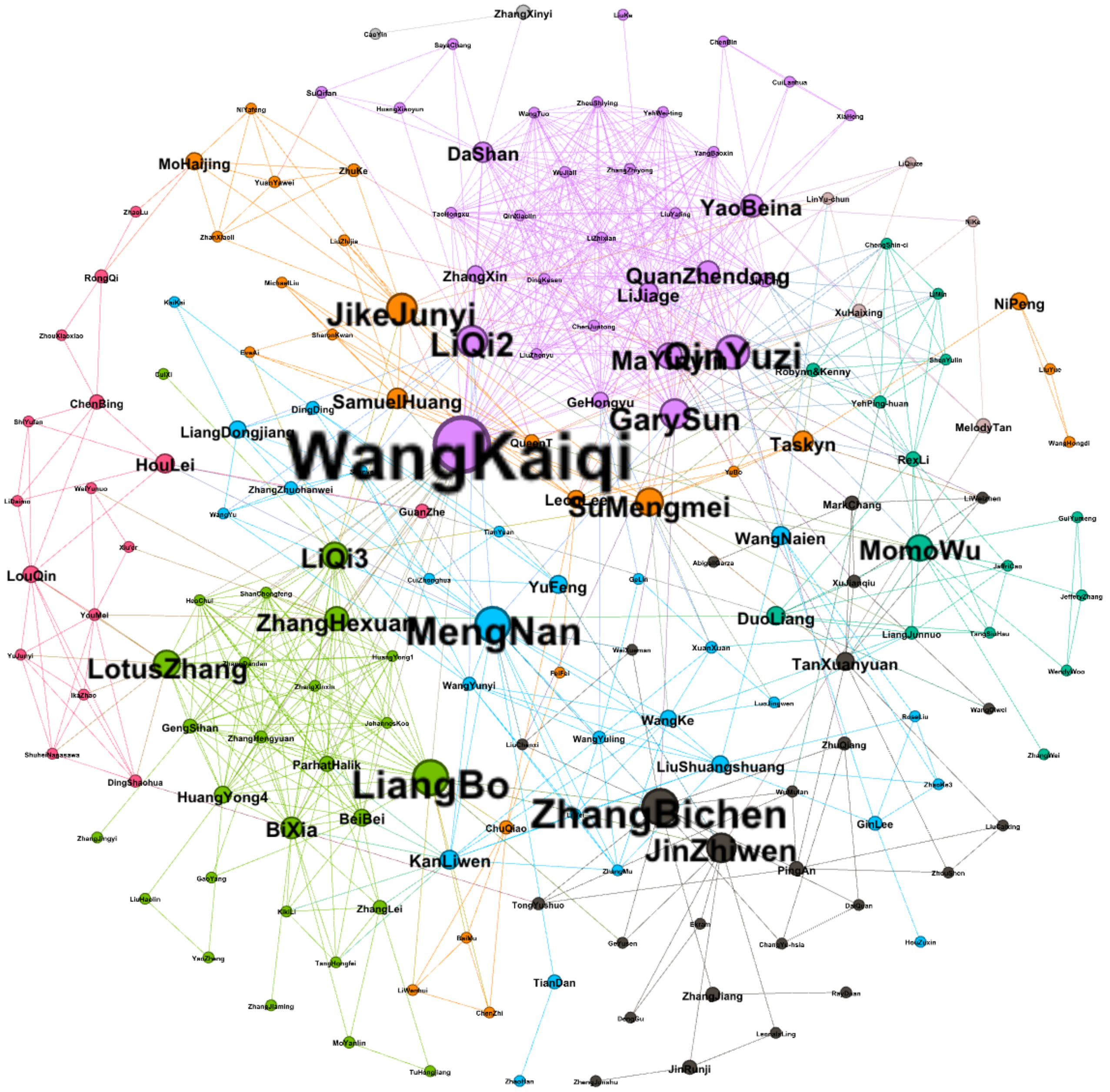
The potential outcomes

- A social science' research report
- An investigative report
- A business report
- Infographics
- Visualization
  - Interactive visualization (allowing user exploration)
  - Presentation visualization (does not support user input)
  - A combined type: interactive storytelling (web-based)

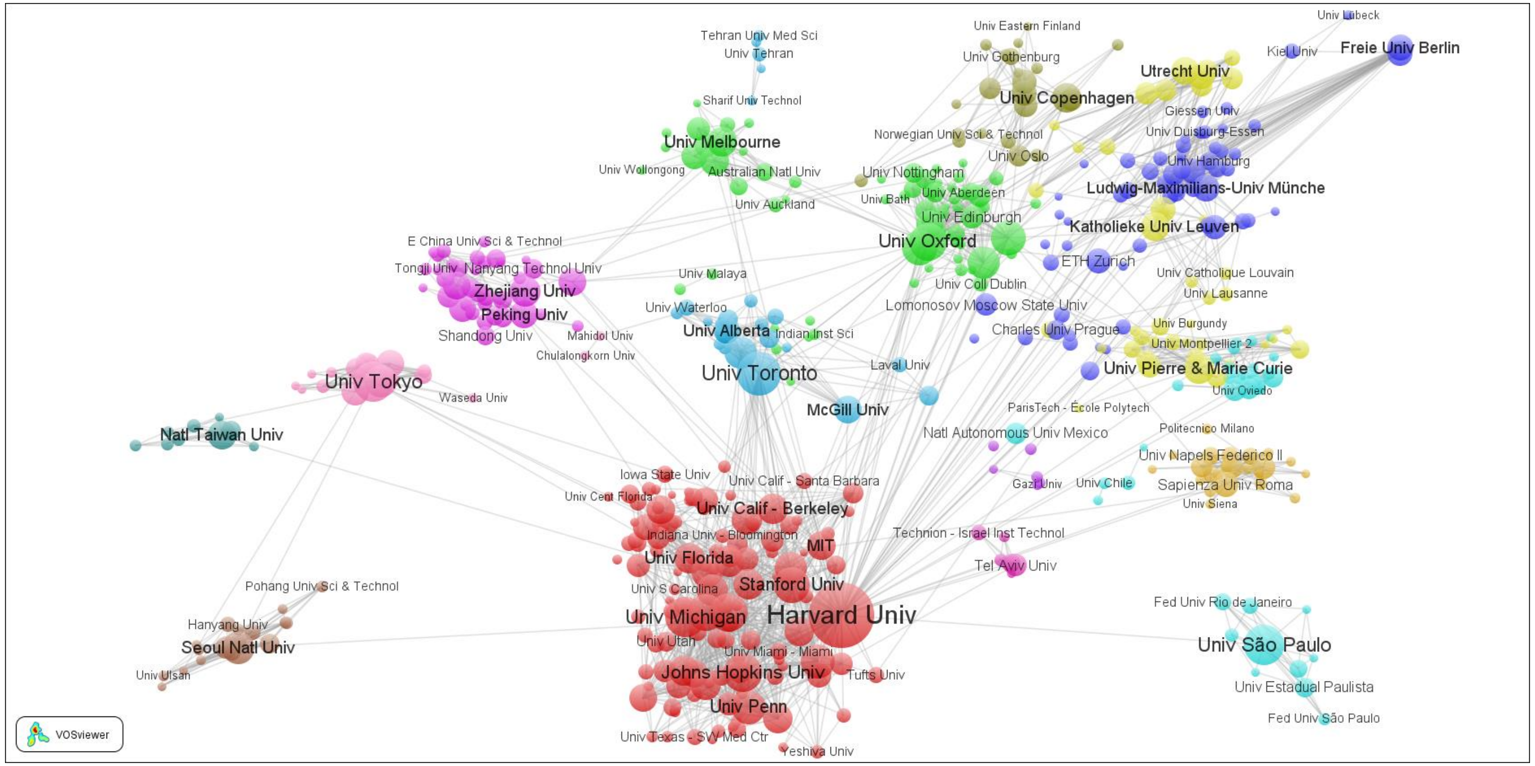
Infographics

- source





Zhang (2017).
Figure 2. A snapshot of the contestant network of the Voice of China, illustrating the co-cover process (note: only the connected section of the graph is illustrated. The figure was plotted by Gephi).



Co-author + university collaboration network

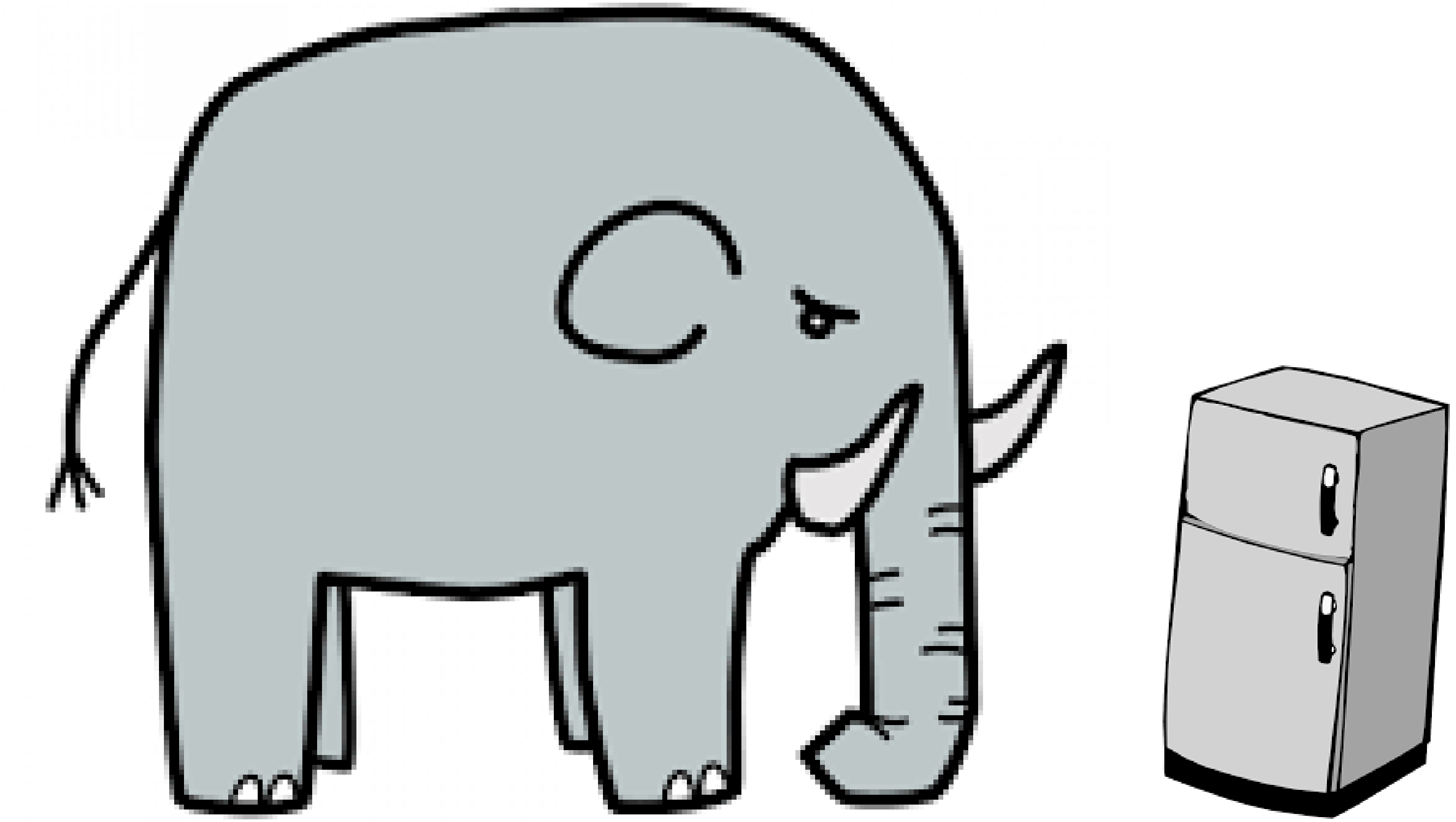
Computational Thinking

Computational Thinking

- Approach to problem solving
- Express problems and solutions to be executed by a computer
- Involves
  - Problem analysis
  - Abstraction
  - Designing automated computations

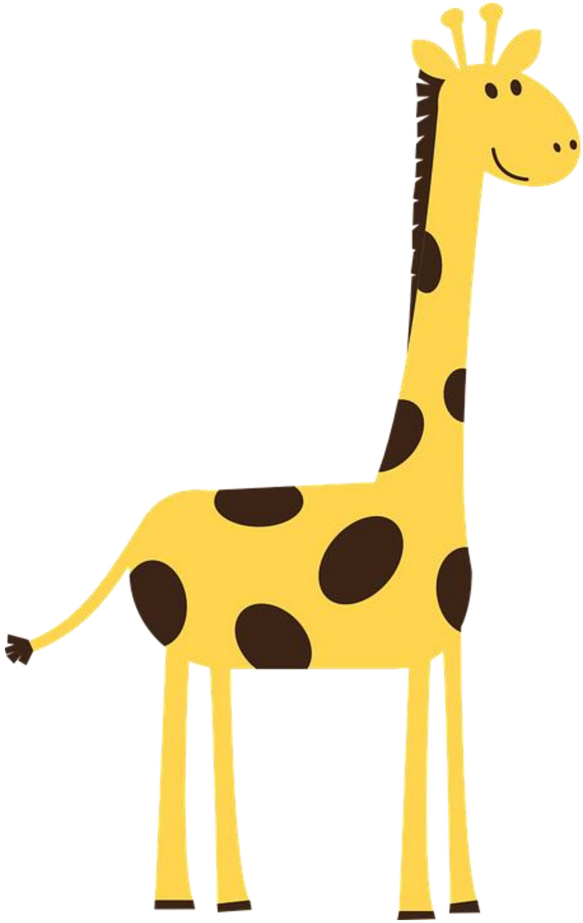
Computational Thinking

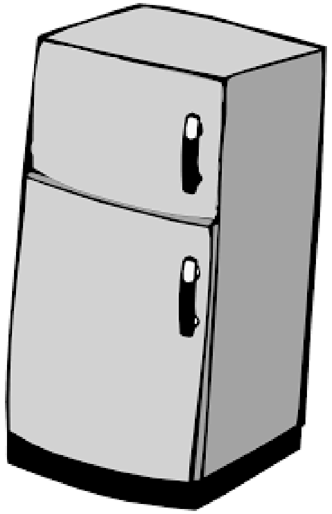
- Problem: How to put an elephant into a fridge?



Computational Thinking

- Problem: How to put a giraffe into a fridge?





Computational Thinking

- Problem: The Lion King calls for a general meeting of all the animals. Someone is missing. Who?

Computational Thinking

- Problem: An explorer wants to cross a river full of crocodiles. What solution you suggest?

Computational Thinking

- Some tips for computational thinking approach to problem solving
    - Split the problem in more simple problems and then combine the solutions
        - Divide and conquer
    - Try to find a general solution
        - Is there a general problem that can be solved?
    - Memory is important to extract knowledge and logical conclusions

Why programming?

- Increasing market demand
- The digital transformation of the newsrooms
- The increasingly presence of AI, machine learning, and data science in the media communication context
- Internet is a rich database - digital footprints (digital trace)
- The price and difficulty of collecting and storing massive online data has dramatically reduced.
- The field itself is becoming more interdisciplinary.

Am I getting lost?

- Ask for help as soon as possible
- Search on the Internet
- Compare with your peers
- Ask the instructor/TA
- 莫做伸手党

Tips: How to ask for help?

- Asking reproducible questions (other can understand your question and rework it on their own machines)
- What is the question you are going to answer?
- What steps did you use to find out the answer?
- What is the expected output?
- What do you see instead?
- What version and operating system are you using?
- What are the data analytical tools/functions you are using?
- What other solutions have your thought about?
- Reference: Eric Steven Raymond: How To Ask Questions The Smart Way [Must read, URL here]

Tips: How to ask for help?

- Be polite: others do not have the obligations to help you
- Be explicit: Try to be as specific and detailed as you can! Don't ask too general questions
- Following up and post solutions - helping others, knowledge increments

Tips: Where to look for help?

- Ready-made:
  - Software's manuals and helping documents
  - Official tutorials
- Online sources:
  - Stack overflow
  - GitHub pages
  - Google and Google scholar
  - Course forums
  - WeChat or Twitter public accounts
  - Online courses
- Offline sources
  - A skilled friend?
  - Workshops, seminars, hackathons, meetups

Tips: Practice, practice, practice!

- Get your hands dirty!
- Make errors!
  - Errors are normal, learn from them!
  - A funny reading about programmers in Chinese: [Link]
- Try again!

References: Python programming

- Python 3.8.x documentation (latest stable version). https://docs.python.org/3/
- Elkner, J., Downey, A. B., & Meyers, C. (2016). How to Think Like a Computer Scientist: Learning with Python. [Link]
- Severance, C. (2017). Programming for everybody (getting started with Python). *University of Michigan,* [Link].
- VanderPlas, J. (2016), A Whirlwind Tour of Python, OReilly Media Inc. [Link]

References: Data Visualization

- VanderPlas, J. (2016). Python data science handbook: essential tools for working with data. O'Reilly Media, Inc.
- Munzner, T. (2014). Visualization analysis & design. CRC Press - Taylor & Francis Group.
- Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011.
- Knaflic, Cole N. Storytelling with Data: A Data Visualization Guide for Business Professionals. Hoboken, New Jersey: Wiley, 2015.
- Stephanie Evergreen. Effective Data Visualization: The Right Chart for the Right Data, SAGE Publications, Inc, 1st edition, 2016.

References: A batch of GitHub “Repos”

- # “Repos” on general data science
  - Data-X@Berkerly
  - Computational Sociology @ Duke

- # data science based on python
  - WhirlwindTourofPython
  - PythonDataScienceHandbook

Thank You