COM5507 Social Media Data Acquisition and Processing
Week 4. Data Science Pipeline & Project Implementation

Lectured by: Dr. Xinzhi ZHANG
Research Assistant Professor, Department of Journalism
Hong Kong Baptist University
26 Sep 2018 @ CityU M4003

Agenda

- Data science pipeline: the work flow
- Finding a story
  - from issues and cases
  - from an “investigation”
- Presenting a story
  - from a theoretical perspective
  - from data exploration
- Project implementation (1)
  - Exercise 1
  - A proposal

THE PIPELINE

Data science in action: a revisit

- Defining 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

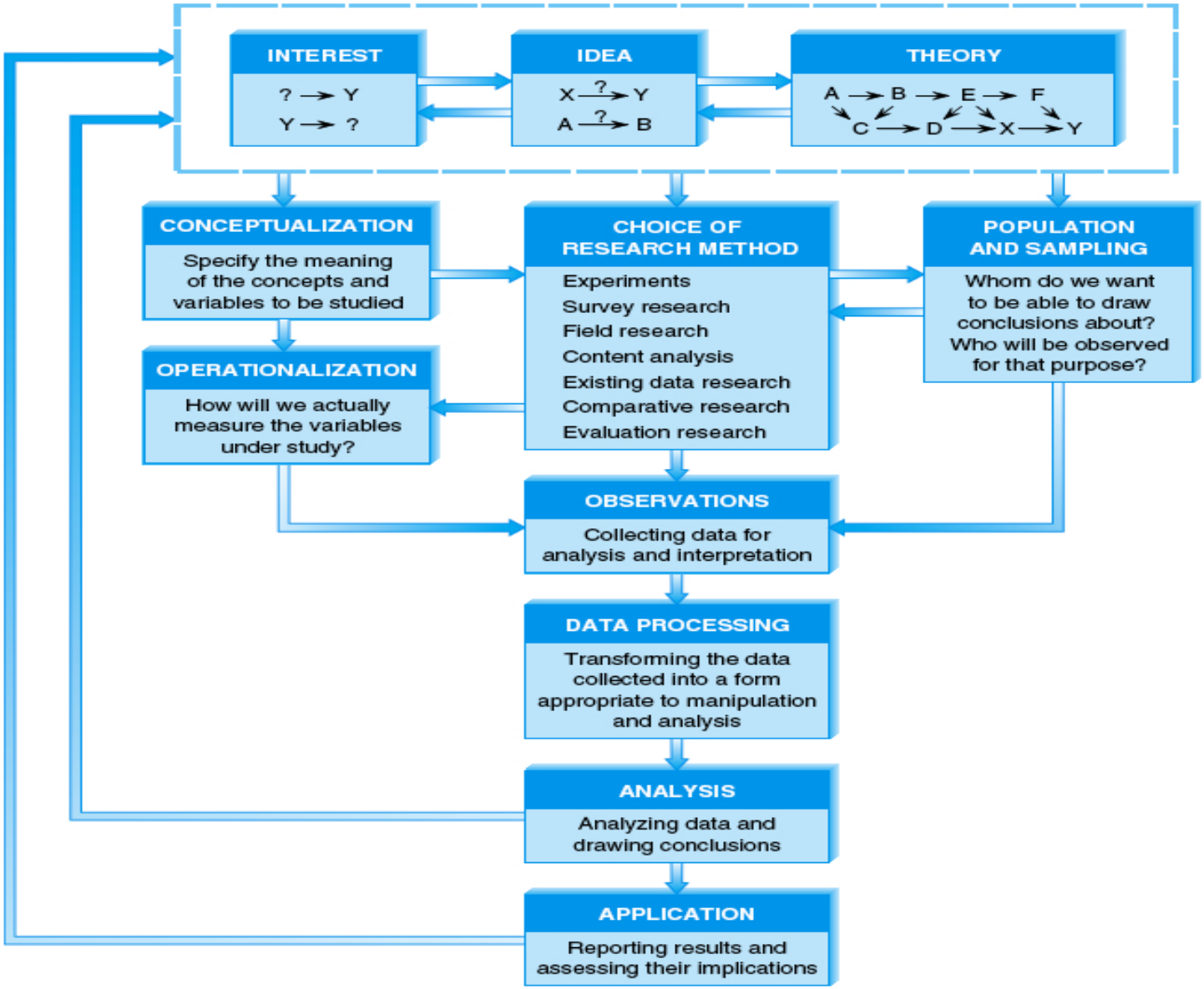
Data science pipeline - a verbal explanation

- **The "OSEMN Pipeline"**
- **O —** Obtaining our data
- **S** — Scrubbing / Cleaning our data
- **E —** Exploring / Visualizing our data will allow us to find patterns and trends
- **M —** Modeling our data will give us our predictive power as a wizard
- **N —** Interpreting our data
- - Reference

Data science pipeline - a graphical illustration

- Zacharias Voulgaris (2017): The Data Science Pipeline - Data Science: Mindset, Methodologies, and Misconceptions

Data science pipeline - a social science' perspective



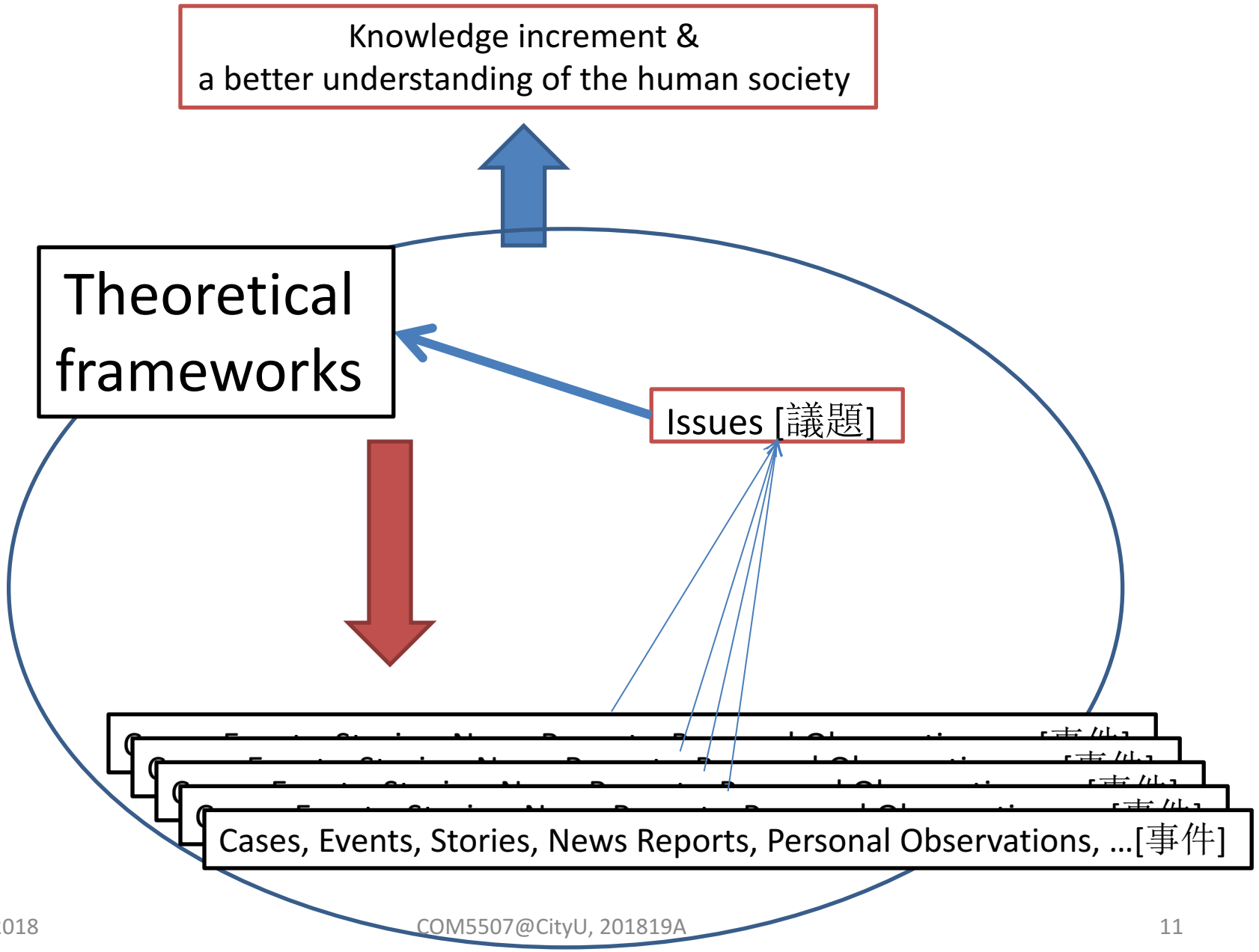
Foremost: A question

- “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*

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


Knowledge increment &
a better understanding of the human society
Theoretical frameworks
Issues [議題]
Cases, Events, Stories, News Reports, Personal Observations, …[事件]

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).

Defining a problem/finding a story

- Alexis Ulrich: Using Data Journalism to Generate Content Ideas [URL]
- Case: the speeding cops [URL]
- Case: NYT: Deaths at Rail Crossings

- 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?

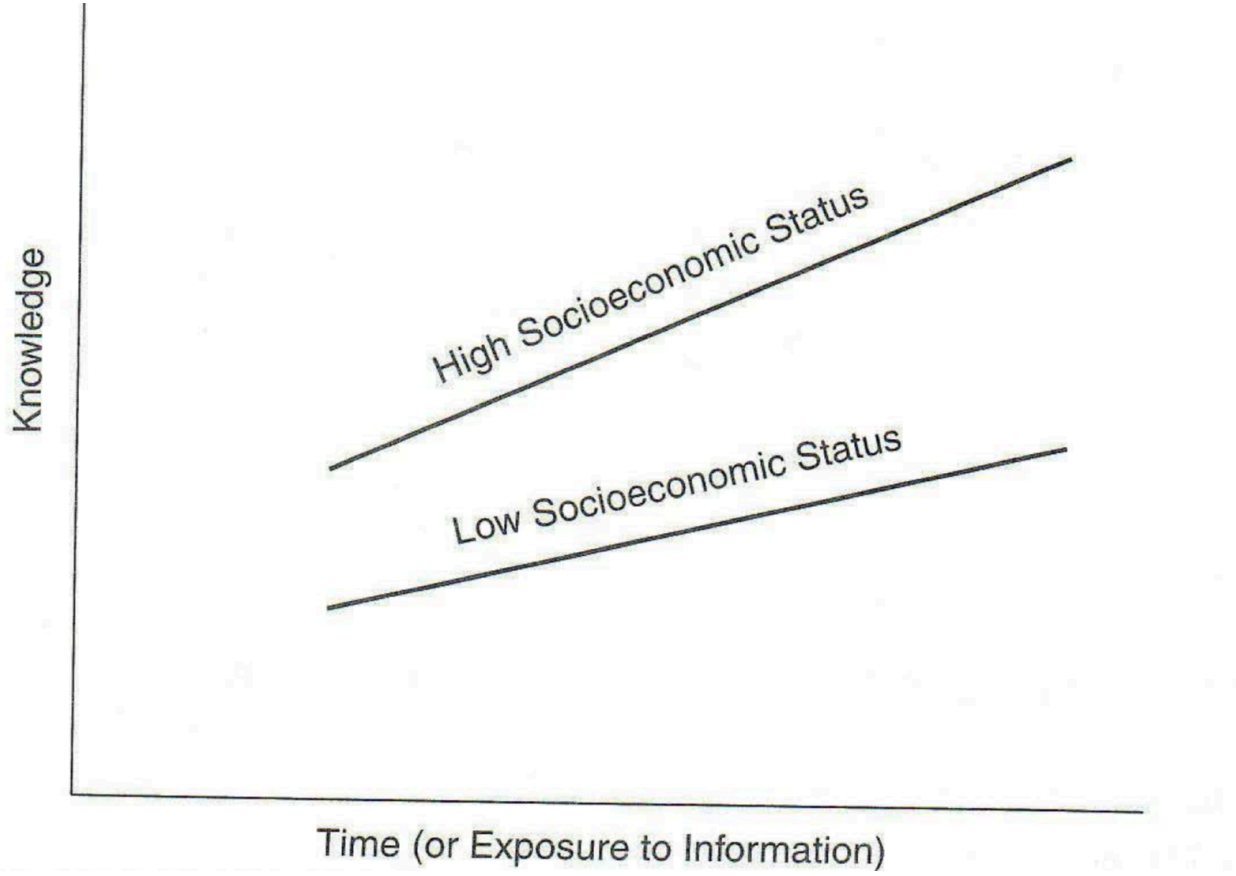
PRESENTING A STORY

Presenting a theory/story

- Three ways to present a theory (perhaps a news story as well)
- The knowledge gap hypothesis (Tichenor, Donohue, & Olien, 1970)
- (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
    - $K = b_0 + b_1 * T + b_2 * S + b_3 * T * S$

Presenting a theory/story

- 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.

  - *Side note: What are the three assumptions of a “causal relationship?”*

Presenting a theory/story: from data exploration

- Also on finding the "stories" (by C. Anderson)
  - "Between the unique and the pattern"

- A classic example [Video]
- Another example by Fivethirtyeight: *Lionel Messi Is Impossible* [Link]

The potential outcomes

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

Kicking off and getting onboard

PROJECT IMPLEMENTATION (1)

Exercises 1a

- Seeking the storytelling possibilities
  - identify a “text” story, i.e., an investigative report in SCMP, WSJ, Guardian, FT (Chinese), Nanfang’s, Jiemian, and seek possible ways to turn it into a data-driven story (try to add some data source and charts);
- try to answer these questions:
  - What is the problem/gap defined by the news story?
  - What is the data source? Can we have more or less data sources?
  - Can we add or remove the data analysis methods?
  - (for this course) Why or why not use web scraping?
  - Can the presentation layout be changed? If so, how? If not, why?

Exercises 1b

- Seeking the storytelling possibilities
  - identify a “data-driven journalism” piece, i.e., a data story from HK01, Initium, Bloomberg Interactive, Guardian Interactive, FT data, Caixin data, and seek possible ways to turn it into a text story (“de-datafication” – try to “remove” as many charts and tables as you can).
- try to answer these questions:
  - What is the problem/gap defined by the news story?
  - What is the data source? Can we have more or less data sources?
  - Can we add or remove the data analysis methods?
  - (for this course) Why or why not use web scraping?
  - Can the presentation layout be changed? If so, how? If not, why?

A research/reporting proposal

1. Questions / objectives
   - Which issue/case/beat do you want to cover?
   - What is the “problem” you can identify?
2. What has been said / literature review
   - Reviewing existing works (week 5 – 6’s tasks)
3. What are you going to do?
   - Data sources
   - Data collection plan
   - Measurement
4. How do you find answers?
   - Data analysis (planned)
5. Labor distribution among the team members
6. Storytelling & outputs
   - Medium of outputs (text? graphics? webpage? video?)