Data Literacy & Open Data

Pili Hu

Symposium on Digital Scholarship @HKBU

Case: "the lord of bar chart" -- Trump

## Case: deadly flu or not?

<table>
<thead>
<tr>
<th></th>
<th>2017 Flu</th>
<th>2003 SARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>7.4M</td>
<td>6.7M</td>
</tr>
<tr>
<td>Death</td>
<td>315</td>
<td>299</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2017 Flu</th>
<th>2003 SARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosed</td>
<td>14713</td>
<td>1755</td>
</tr>
<tr>
<td>Death</td>
<td>315 (2.14%)</td>
<td>299 (17.3%)</td>
</tr>
</tbody>
</table>
Case: Most supported member in HK Legco (12'-16')

Total Yes: 2537
Total motions: 1066
Potential total Yes: 74620
Yes Rate: 3.4%

Most motions are proposed around June

“Mainly filibustering on budgets”

http://legco.initiumlab.com/
Case: Gender discrimination?

<table>
<thead>
<tr>
<th>Department</th>
<th>Male</th>
<th>Female</th>
<th>Male Ratio</th>
<th>Female Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admitted</td>
<td>Denied</td>
<td>Admitted</td>
<td>Denied</td>
</tr>
<tr>
<td>A</td>
<td>512</td>
<td>313</td>
<td>89</td>
<td>19</td>
</tr>
<tr>
<td>B</td>
<td>313</td>
<td>207</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>C</td>
<td>101</td>
<td>205</td>
<td>202</td>
<td>391</td>
</tr>
<tr>
<td>D</td>
<td>138</td>
<td>279</td>
<td>131</td>
<td>244</td>
</tr>
<tr>
<td>E</td>
<td>40</td>
<td>138</td>
<td>94</td>
<td>299</td>
</tr>
<tr>
<td>F</td>
<td>22</td>
<td>351</td>
<td>24</td>
<td>317</td>
</tr>
<tr>
<td>Total</td>
<td>1126</td>
<td>1493</td>
<td>557</td>
<td>1278</td>
</tr>
</tbody>
</table>
Case

One researcher finds that those who start a fight are more likely to die. The correlation is very high!

What do we learn from this?

Never start a fight if you could help.

The reality:

Those who are alive always say the other one started fighting!

(This case is fricticious)
Data literacy

Data literacy is the ability to **read, understand**, create and communicate data as information:

- Survival rate after fighting: Sample bias
- Flu: Conditional probability v.s. Joint probability
- Trump campaign: Misleading charts
- Gender discrimination: drill down
- Legco vote: domain knowledge
Some antiques to get started...

Darrell Huff, 1954

Willard C. Brinton, 1939
We have good intellectual assets on data literacy,
But lack essential and deliberate training
Data Journalism as a training ground for data literacy

- Don't have to be data journalist to learn data journalism
- Idea is more important than tools
- Critical thinking on data
- Many convenient tools to help you get started quickly

<table>
<thead>
<tr>
<th>Data Collection:</th>
<th>Data Analysis</th>
<th>Data Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Scraper (Chrome extension)</td>
<td>● Google spreadsheet</td>
<td>● Google Sheets</td>
</tr>
<tr>
<td>● Copy and paste</td>
<td>○ Basic manipulation</td>
<td>● Google Fusion table</td>
</tr>
<tr>
<td>● Google Cache</td>
<td>○ Function</td>
<td>● Google Drawing</td>
</tr>
<tr>
<td>● Wayback Machine</td>
<td>○ Pivoting</td>
<td>● Tableau</td>
</tr>
<tr>
<td>● Import.io</td>
<td>○ Charting</td>
<td>● Baidu Echart</td>
</tr>
<tr>
<td></td>
<td>○ Scripting (.gs)</td>
<td>● Infogr.am</td>
</tr>
<tr>
<td></td>
<td>● Google fusion table</td>
<td>● Kumu</td>
</tr>
<tr>
<td>Data Cleaning</td>
<td></td>
<td>● Netlytic</td>
</tr>
<tr>
<td>● Google Refine</td>
<td></td>
<td>● AI / PS</td>
</tr>
<tr>
<td>● Google Sheets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>● PDF tables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>● Small PDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>● Tabula</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Source: Pili Hu, HKBU DNN Inargrual Talk, Dec 2015)
Quiz: What is this distribution?
What crossed the top search term in the world?

Oct 2016
Data Journalism: leverage open data, data analytics and infographics to make impactful stories
Data Journalism as the motivation for open data

Open data for data journalism:

● Data is the cooking material
● Features:
  ○ Free access/ no cost;
  ○ machine readable;
  ○ permissive license for derivatives

Data journalism for open data:

● Applications and justifications
● Data validators
Legco voting analysis, @Initium Media, 2015

- 200,000 voting records
  - Open Data from Legco official website
- 70 members
- ~3000 motions
- Heatmap to visualise

http://legco.initiumlab.com/matrix
Data sounds project @Initium Lab, 2016

News: hottest year in history

Data from HK Observatory

http://datasounds.initiumlab.com/
Census 2011

Map the Census Data
To begin, select a feature to overlay on the map

% of population over 65

http://hkncensus11.hupili.net/#/choropleth  Community project: ODHK, CODE4HK, etc
Open data status quo in Hong Kong

Some continuous force:

- Data studio @ HKSTP: http://datastudio.hkstp.org/
- Open Data Hong Kong: https://www.opendatahk.com/
- Open Source Hong Kong: https://opensource.hk/
- Civic Hacking: https://g0vhk.io/
Taiwan Ranks 1st in OKFN Open Data Index

<table>
<thead>
<tr>
<th>Rank</th>
<th>Place</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Taiwan</td>
<td>78%</td>
</tr>
<tr>
<td>2</td>
<td>United Kingdom</td>
<td>76%</td>
</tr>
<tr>
<td>3</td>
<td>Denmark</td>
<td>70%</td>
</tr>
<tr>
<td>4</td>
<td>Colombia</td>
<td>68%</td>
</tr>
<tr>
<td>5</td>
<td>Finland</td>
<td>67%</td>
</tr>
<tr>
<td>5</td>
<td>Australia</td>
<td>67%</td>
</tr>
<tr>
<td>7</td>
<td>Uruguay</td>
<td>66%</td>
</tr>
<tr>
<td>8</td>
<td>United States</td>
<td>64%</td>
</tr>
<tr>
<td>8</td>
<td>Netherlands</td>
<td>64%</td>
</tr>
<tr>
<td>10</td>
<td>Norway</td>
<td>63%</td>
</tr>
<tr>
<td>10</td>
<td>France</td>
<td>63%</td>
</tr>
</tbody>
</table>

Source: Pili Hu, 2016, Open Data & Smart City @HKPC

http://index.okfn.org/place/
### Main Problem of Hong Kong’s PSI

<table>
<thead>
<tr>
<th>Rank</th>
<th>Place</th>
<th>National Statistics</th>
<th>Government Budget</th>
<th>Legislation</th>
<th>Procurement Tenders</th>
<th>Election Results</th>
<th>National Map</th>
<th>Weather forecast</th>
<th>Pollutant Emissions</th>
<th>Company Register</th>
<th>Location Datasets</th>
<th>Water Quality</th>
<th>Land Ownership</th>
<th>Government Spending</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>Jamaica</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>42%</td>
</tr>
<tr>
<td>37</td>
<td>Albania</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>42%</td>
</tr>
<tr>
<td>37</td>
<td>Hong Kong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>42%</td>
</tr>
<tr>
<td>40</td>
<td>Luxembourg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41%</td>
</tr>
<tr>
<td>41</td>
<td>Indonesia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40%</td>
</tr>
<tr>
<td>42</td>
<td>Thailand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39%</td>
</tr>
<tr>
<td>42</td>
<td>Greece</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39%</td>
</tr>
<tr>
<td>44</td>
<td>Israel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38%</td>
</tr>
<tr>
<td>44</td>
<td>Rwanda</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38%</td>
</tr>
<tr>
<td>44</td>
<td>Puerto Rico</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38%</td>
</tr>
<tr>
<td>47</td>
<td>Turkey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37%</td>
</tr>
<tr>
<td>47</td>
<td>Georgia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37%</td>
</tr>
</tbody>
</table>

---

**Legislation in Hong Kong**

- It’s not openly licensed
- It’s not machine readable
- It’s free
- It’s not available in bulk
- It’s up to date
- It’s online [here](#)
- It’s digital
- It’s public
- Data does not exist

In this case, the questions about ‘available in bulk’ and ‘open licensing’ are very pertinent. …

---

**Machine Readability and Open License → This is improving today!**

Actionable item #1: Keep up the conversation

**Government**
- What is the output if we open our data?
- Too much manpower to open data

**Academics**
- What is happening out there?

**Industry**
- Data is corporate asset, even if the company is funded by gov

**Citizen**
- Just want the information...

**Community**
- How can the people appreciate our work?
- How can gov open more data?

Source: Pili Hu, 2016, Open Data & Smart City @HKPC
Actionable item #2: A community/academics data repo with continuous maintenance

CKAN, the world’s leading Open Source data portal platform

CKAN is a powerful data management system that makes data accessible – by providing tools to streamline publishing, sharing, finding and using data.

WHAT IS CKAN?
CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available. Learn more

WHY CKAN?
CKAN is open source, free software. This means that you can use it without any license fees, and you retain all rights to the data and metadata you enter.

HOW TO CONTRIBUTE?
CKAN version releases are coordinated, tested and deployed by the tech team. Being an open source project, CKAN and its extensions are developed by a large community of people.
Actionable item #3: Contribute topics, ideas, datasets, data tools, for data journalism projects

A collection of interactive web stories from our students:

http://interactive.jour.hkbu.edu.hk/

- > 10 data related courses
- New concentration: Data and Media Communication (DMC)
Actionable item #4:
DJ Bootcamp + Contest in Jan 2019

See Jan 2018 version:

New format: (TBD)

- Hackathon
- Make a data & media project in 3 days
- Mini workshops to cover:
  - Data collection
  - Data analytics
  - Data visualization
  - Journalism investigation
  - In-depth writing
  - Design thinking
  - Business plan

- Joint university effort
Contact

- Pili Hu
- hupili.net
- hupili@hkbu.edu.hk ; e@hupili.net
- https://github.com/hupili
- Search real name on all social networks

My blog on Wechat:

pwords