

# CS102: Big Data

Tools and Techniques, Discoveries and  
Pitfalls

Spring 2017

Ethan Chan, Lisa Wang

*Lecture 8: More Data Mining & Final  
Project*

# Announcements

- Ethan's OH after class today
- Lisa's OH this week
  - Saturday 3-4.30pm, Lathrop Tech Lounge
- Updated Assignment3 Monday night, please re-download zip folder if you have old version.
- Midterm next Tuesday May 2
- Project Proposal due next Sunday May 7

# Python Tip of the Day

- Dynamic creation of nested structures (might be helpful for assignment 3)

## Dynamic creation of nested data structures

```
In [4]: friends = ["Anna", "Ben", "Mary"]
friends_favorite_food = {}
for name in friends:
    friends_favorite_food[name] = []

print friends_favorite_food

{'Ben': [], 'Mary': [], 'Anna': []}
```

```
In [6]: friends_favorite_food["Anna"].append("Guacamole")
friends_favorite_food["Anna"].append("Mango")
friends_favorite_food["Ben"].append("Strawberry")
friends_favorite_food["Ben"].append("Pesto Chicken")
friends_favorite_food["Mary"].append("Pizza")
friends_favorite_food["Ben"].append("Walnuts")

print friends_favorite_food

{'Ben': ['Strawberry', 'Strawberry', 'Pesto Chicken', 'Walnuts'], 'Mary': ['Pizza', 'Pizza'], 'Anna': ['Guacamole', 'Mango', 'Guacamole', 'Mango']}
```

# Plan for Today

- Solidify understanding of frequent itemsets and association rules
- Storytelling with data
- Final project announcement

# Recap Last Lecture

- Intro to Data Mining
- Frequent Itemsets
  - Support
- Association Rules
  - Support
  - Confidence
  - Lift

# Frequent Itemsets

Sets of items that occur together frequently in transactions.

Measure:























- *Support*: # transactions containing set / total # transactions
  - Look for sets with support > support-threshold

# Association Rules

Rule: Set1 of items  $\rightarrow$  Set2 of items

3 Measures:

- Support: *How popular is itemset Set1?*
- Confidence: *How likely will Set2 be purchased when Set1 is purchased?*
- Lift: *How likely will Set2 be purchased when Set1 is purchased, accounting for how popular Set2 is?*

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 























**Table 1. Example transactions.**

Figures & Tables from “Numsense! Data Science For The Layman” by Annalyn Ng & Kenneth Soo.



# Support























- Proportion of transactions where an itemset appears
- What is the support for {apple}?
- What is the support for {apple, beer}?
- What is the support for {milk, pear}?

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

**Table 1. Example transactions.**

# Support

- Proportion of transactions where an itemset appears
- What is the support for {apple}?  $4/8 = 0.50$
- What is the support for {apple, beer}?  $3/8 = 0.375$
- What is the support for {milk, pear}?  $1/8 = 0.125$

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

**Table 1. Example transactions.**

# Support

- Proportion of transactions where an itemset appears
- What is the support for {apple}?  $4/8 = 0.50$
- What is the support for {apple,beer}?  $3/8 = 0.375$
- What is the support for {milk, pear}?  $1/8 = 0.125$

*If support-threshold = 0.2, which itemsets to the left are frequent itemsets?*

# Support

- Proportion of transactions where an itemset appears
- What is the support for {apple}?  $4/8 = 0.50$
- What is the support for {apple, beer}?  $3/8 = 0.375$
- What is the support for {milk, pear}?  $1/8 = 0.125$

If *support-threshold* = 0.2, which itemsets to the left are frequent itemsets?

*{apple}, {apple, beer}*

*(yes, you can have itemsets of size 1)*

# Confidence

- Proportion of transactions with Set1 where Set2 also appears

$$\text{Confidence } \{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support } \{\text{🍎, 🍺}\}}{\text{Support } \{\text{🍎}\}}$$

Figure 3. Confidence measure.

- What is the confidence for {apple} → {beer}?
- For {milk, beer} → {rice}?

Transaction 1	🍎 🍺 🍲 🍗
Transaction 2	🍎 🍺 🍲
Transaction 3	🍎 🍺
Transaction 4	🍎 🍏
Transaction 5	🍼 🍺 🍲 🍗
Transaction 6	🍼 🍺 🍲
Transaction 7	🍼 🍺
Transaction 8	🍼 🍏

Table 1. Example transactions.

# Confidence

- Proportion of transactions with Set1 where Set2 also appears

$$\text{Confidence } \{\text{🍏} \rightarrow \text{🍺}\} = \frac{\text{Support } \{\text{🍏, 🍺}\}}{\text{Support } \{\text{🍏}\}}$$

Figure 3. Confidence measure.

- What is the confidence for {apple} → {beer}?  $3/4 = 0.75$
- For {milk, beer} → {rice}?  $2/3 = 0.67$

Transaction 1	🍏 🍺 🍲 🍗
Transaction 2	🍏 🍺 🍲
Transaction 3	🍏 🍺
Transaction 4	🍏 🍏
Transaction 5	🍼 🍺 🍲 🍗
Transaction 6	🍼 🍺 🍲
Transaction 7	🍼 🍺
Transaction 8	🍼 🍏

Table 1. Example transactions.

# Confidence

- Proportion of transactions with Set1 where Set2 also appears

$$\text{Confidence } \{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support } \{\text{🍎, 🍺}\}}{\text{Support } \{\text{🍎}\}}$$

Figure 3. Confidence measure.

*If confidence-threshold = 0.7, which relations on the left form association rules?*

- What is the confidence for {apple} → {beer}?  $3/4 = 0.75$
- For {milk, beer} → {rice}?  $2/3 = 0.67$

# Confidence

- Proportion of transactions with Set1 where Set2 also appears

$$\text{Confidence } \{\text{🍏} \rightarrow \text{🍺}\} = \frac{\text{Support } \{\text{🍏, 🍺}\}}{\text{Support } \{\text{🍏}\}}$$

Figure 3. Confidence measure.

- What is the confidence for  $\{\text{apple}\} \rightarrow \{\text{beer}\}$ ?  $3/4 = 0.75$
- For  $\{\text{milk, beer}\} \rightarrow \{\text{rice}\}$ ?  $2/3 = 0.67$

*If confidence-threshold = 0.7, which relations on the left form association rules?*























$\{\text{apple}\} \rightarrow \{\text{beer}\}$



What is a drawback of  
the confidence  
measure?

Confidence can misrepresent the importance  
of an association.

Is there really a strong association {apple} →  
{beer}?

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

**Table 1. Example transactions.**

# Lift to the rescue!

# Lift

- Similar to confidence, but accounting popularity of Set2.
- Divide confidence by support of Set2.

$$\text{Lift} \{ \text{🍏} \rightarrow \text{🍺} \} = \frac{\text{Support} \{ \text{🍏}, \text{🍺} \}}{\text{Support} \{ \text{🍏} \} \times \text{Support} \{ \text{🍺} \}}$$

Figure 4. Lift measure.

Transaction 1	🍏 🍺 🍲 🍗
Transaction 2	🍏 🍺 🍲
Transaction 3	🍏 🍺
Transaction 4	🍏 🍏🍏
Transaction 5	🍼 🍺 🍲 🍗
Transaction 6	🍼 🍺 🍲
Transaction 7	🍼 🍺
Transaction 8	🍼

Table 1. Example transactions.

- Lift > 1: positive association, Set2 likely bought, if Set1 bought
  - Lift = 1: no association
  - Lift < 1: negative association, Set2 unlikely bought, if Set1 bought
- What is the lift for {apple} → {beer}?
  - {apple} → {pear}?

# Lift

- Similar to confidence, but accounting popularity of Set2.
- Divide confidence by support of Set2.

$$\text{Lift} \{ \text{🍏} \rightarrow \text{🍺} \} = \frac{\text{Support} \{ \text{🍏}, \text{🍺} \}}{\text{Support} \{ \text{🍏} \} \times \text{Support} \{ \text{🍺} \}}$$

Figure 4. Lift measure.

Transaction 1	🍏 🍺 🍲 🍗
Transaction 2	🍏 🍺 🍲
Transaction 3	🍏 🍺
Transaction 4	🍏 🍏
Transaction 5	🍼 🍺 🍲 🍗
Transaction 6	🍼 🍺 🍲
Transaction 7	🍼 🍺
Transaction 8	🍼

Table 1. Example transactions.

- Lift > 1: positive association, Set2 likely bought, if Set1 bought
- Lift = 1: no association
- Lift < 1: negative (anti) association, Set2 unlikely bought, if Set1 bought
- What is the lift for {apple} → {beer}? 1.0
- {apple} → {pear}? 2.0

*Data in the Real World*

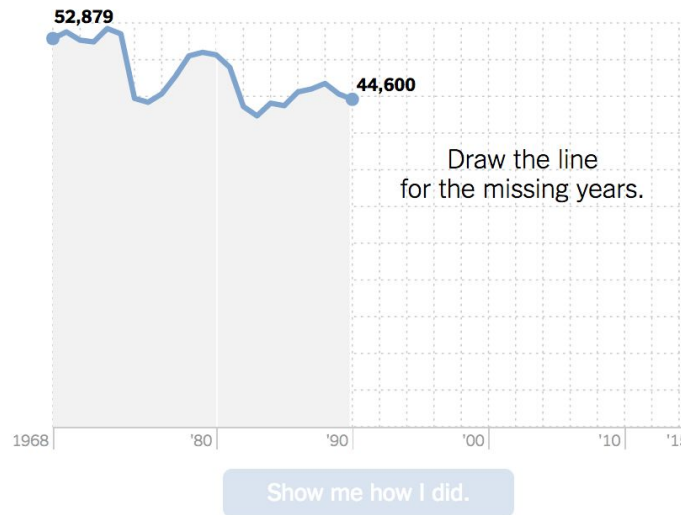
# Storytelling with Data

# You Draw It: Just How Bad Is the Drug Overdose Epidemic?

By JOSH KATZ APRIL 14, 2017

How does the surge in drug overdoses compare with other causes of death in the U.S.? Draw your guesses on the charts below.

Since 1990, the number of Americans who have died every year from **car accidents**...



I don't want to play; just tell me the answers.

Source: New York Times.

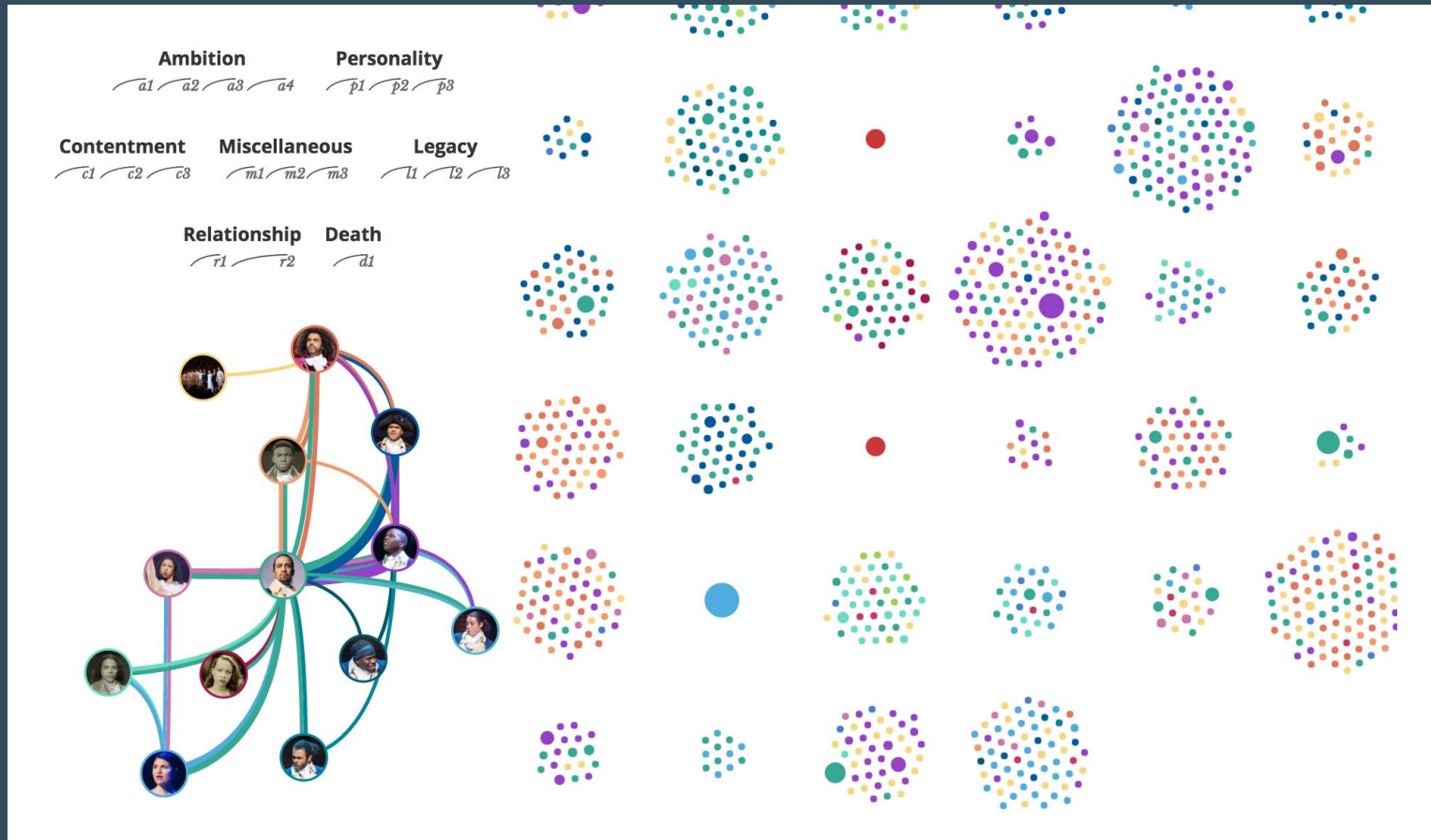
<https://www.nytimes.com/interactive/2017/04/14/upshot/10000005019985.app.html>

# In Class Discussion Points

1. Interactive
2. Engage much deeper, you think more about the problems, question on your assumptions
3. Some people used regression, trying to continue the pattern
4. Based on your own general knowledge, it questions ur assumptions on trends



# Visualization of Hamilton



Source: Pudding.cool. <https://pudding.cool/2017/03/hamilton/>

# Final Project!

- Goals:
  - Use the tools & techniques you've learned in this class to analyze a dataset that interests you
  - Tell a compelling story with visualizations
  - Do something cool / meaningful with data!
- 2-3 students per team
  - Higher expectations for teams of 3
- All details about final project on the website:  
<https://web.stanford.edu/class/cs102/projects/finalproject>

# Final Project: Datasets

Some datasets you could analyze:

- Poverty Statistics (World Bank)
- World Health (World Bank)
- Indicators on Women and Men (United Nations Statistics Division)
- Startups: Funding and Acquisitions (Crunchbase)
- Crime and Socioeconomic indicators (City of Chicago)
- Walmart historic sales (Kaggle)
- Kickstarter Full Dataset (171k+ projects)
- Yelp Full Dataset (140k+ projects)

# Deliverables

**Proposal:** May 7, 11:59 PM

**Short Presentations in Class on Jun 6, Slides  
due Jun 5, 11:59 PM**

**Final Report:** Jun 6, 11:59 PM

All instructions on final project page. Let us know if you want to discuss your project ideas!