CS57300 - Data Mining

Instructor: Bruno Ribeiro
TAs: Xiao Zhang and Shawn Merrill

CS57300
Purdue University

January 12, 2016

Introduction

- Course overview
- What is data mining?
- Data mining process

Course overview

Logistics

- Time and location: Tue Thu 9:00am-10:15am, Beering Hall 2280
- Instructor: Bruno Ribeiro (you can call me “Bruno” or “Professor Ri-bay-ro” but not “Professor Bruno” or “Professor”)
  ribeiro@cs.purdue.edu, LWSN 2142C, office hours: Friday 9-10:15am
- Teaching assistants: Xiao Zhang, Shawn Merril
  cs573-ta@cs.purdue.edu, office hours: TBD
- Webpage: https://www.cs.purdue.edu/~ribeirob/courses/Spring2015
- Email list: spring-2016-cs-57300-le1@lists.purdue.edu
- Prerequisites: introductory statistics course (e.g., STAT 516), basic programming skills (e.g., CS381, STAT598G)
  - Assumes you know basic probability and statistics, linear algebra, R and Python programming (see syllabus for topics list)

Workload

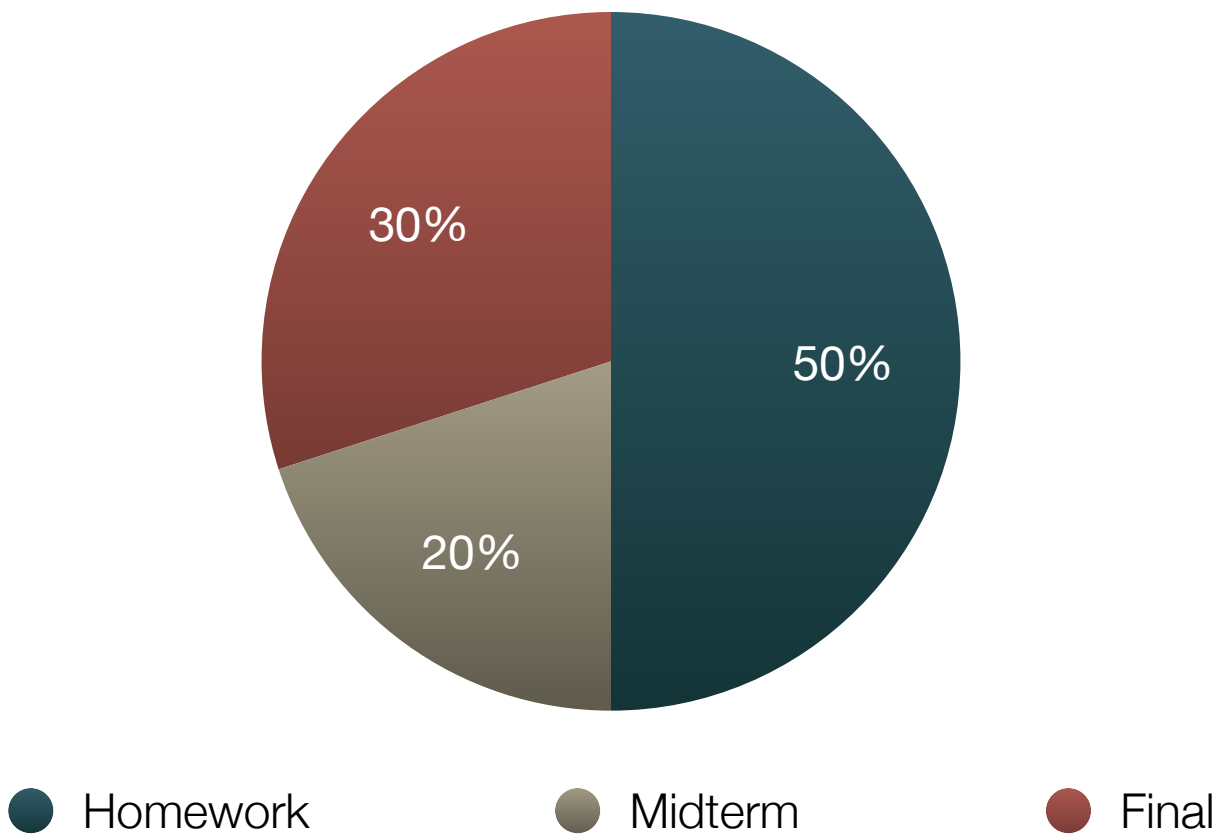
- 6 homeworks, top 5 will count.
- Assignments include written/math exercises, analysis in R, and programming assignments in python
- No homework extensions (cat late_hometwork > /dev/null)
- Homework is likely due on a Friday 11:59pm
  - But ONE updated solution OK until Sunday 11:59pm (+48 hrs)
    - Only last upload counts
  - Homework uploaded on Monday 00:00am > /dev/null
- Exams
  - One take-home midterm and one final exam
  - CS qualifier: final exam supplement

Midterm (take-home) | Tentative approach

- Students compete to predict the next song a set of users
  - Grades: Curve of best prediction ranking

Brought to you by

Grading



Why are you here?

- Because you are excited about learning the hidden secrets of data analysis
- Because it's required and you like to wake up early
- Because you want to make $$$

- In all cases you should be able to get something out of this course
- But you will have to plan ahead and work hard

Textbook

No course text book.

Combination of Scribed Notes + Readings

Books that can help you:
* James, Witten, Hastie, and Tibshirani, “Introduction to Statistical Learning”. Get it online at http://www-bcf.usc.edu/~gareth/ISL

* Kevin P. Murphy, “Machine Learning: A probabilistic perspective”, MIT Press 2012. (Reserved at the Engineering Library)

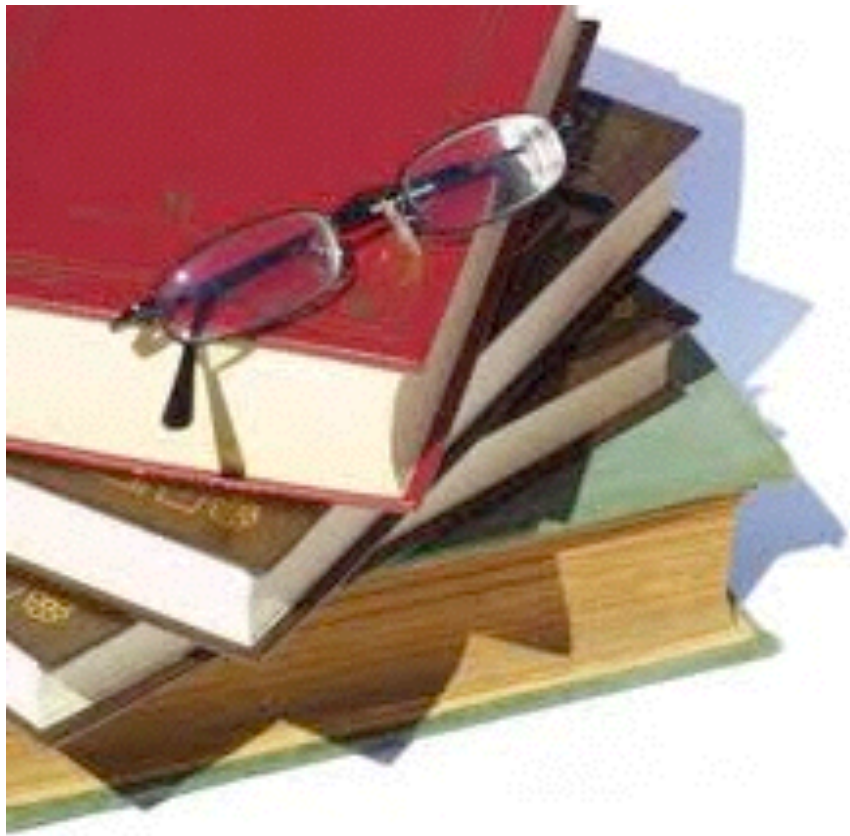
Readings / Scribe

- Readings will be announced/distributed on course webpage.
- You will be asked to scribe one of the classes with x other students
  - Counts as one homework

Registration

- The course is open to non-CS non-STAT students
- Few open spots right now
  - Go to CS department website ⇒ Courses ⇒ Registration
  - Students will be admitted on a first-come, first-serve basis

Course goals



- Identify key elements of data mining systems and the knowledge discovery process
- Understand how algorithmic elements interact
- Recognize various types of data mining tasks
- Familiarity with standard models/ algorithms
- Implement and apply basic algorithms
- Understand how to evaluate performance

What is Data Mining?

What is Data Mining?

- **Data mining** *is the science of discovering structure and making predictions in (large) data sets* (Tibshirani)
- **Discovering structure**
  - E.g., given observations $X_1, \ldots X_n$, learn some underlying group structure based on similarity
- **Making predictions**
  - E.g., given observations $(X_1, Y_1), \ldots (X_n, Y_n)$, predict Y_i from X_i

Data Mining: Not Just Applying Known Techniques

"World Map" in 1459

- Shown biased and incomplete (Columbus et al. 1492)
- Data analysis has similar problems

The Fra Mauro world map (1459)

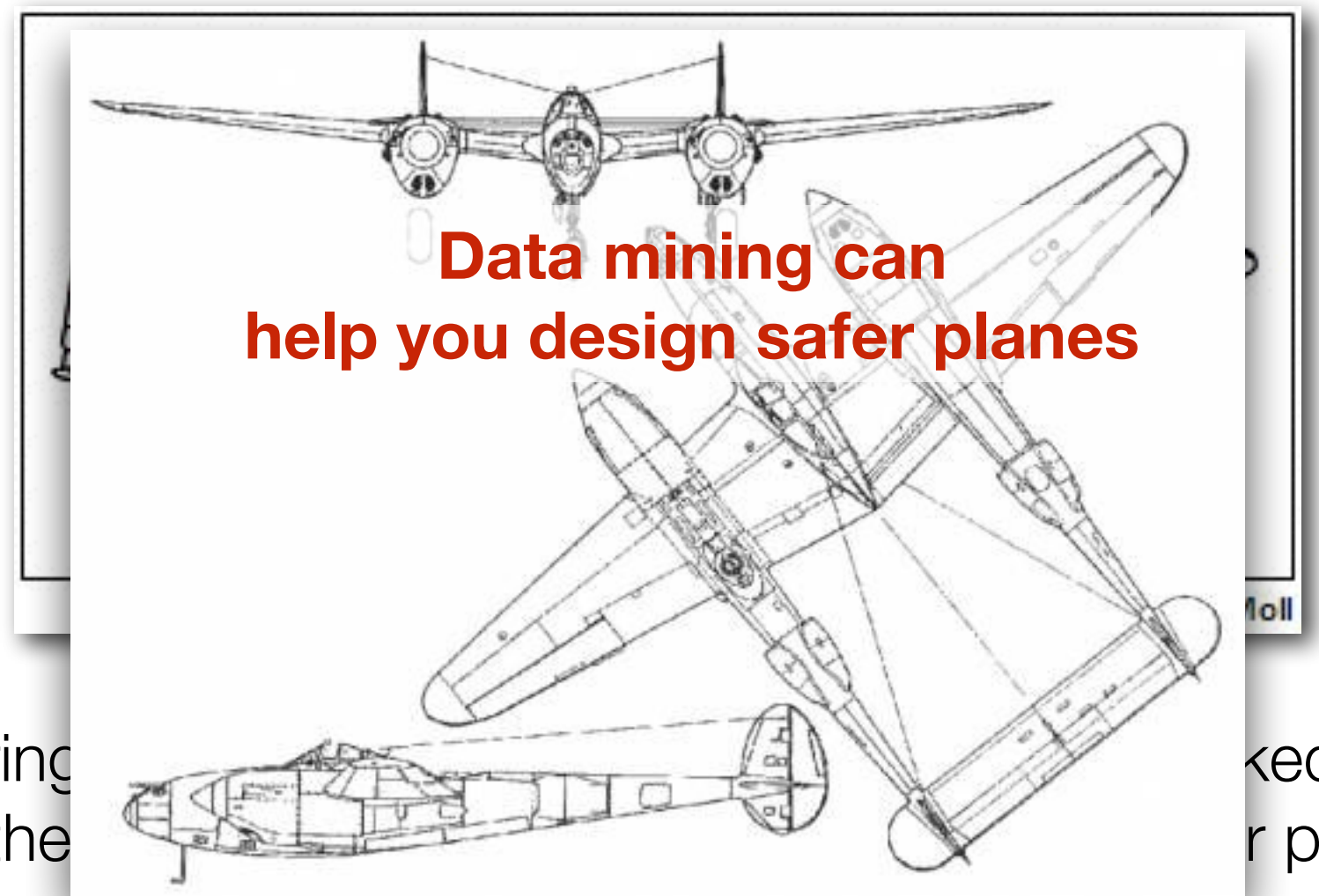
source: Wikipedia

Data Analysis without Understanding the Data

Tools to Tailored to Data

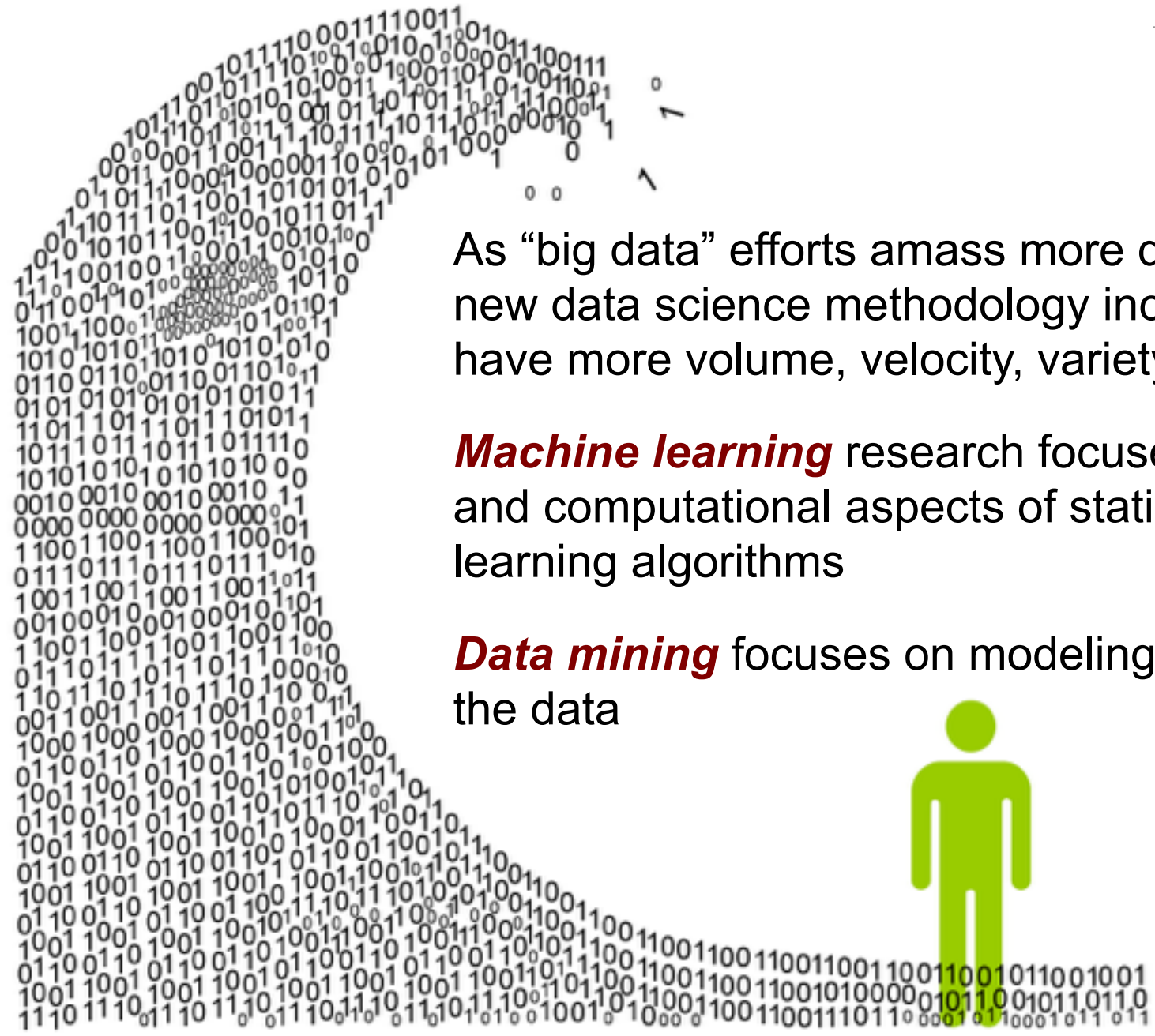
More Precise Analysis

Example



During ked to
help the r planes

The data revolution



As “big data” efforts amass more data… the need for new data science methodology increases. Data today have more volume, velocity, variety, etc.

Machine learning research focuses on the theoretical and computational aspects of statistical models and learning algorithms

Data mining focuses on modeling and understanding the data

The New York Times

Magazine

How Companies Learn Your Secrets

And among life events, none are more important than the arrival of a baby. At that moment, new parents' habits are more flexible than at almost any other time in their adult lives. If companies can identify pregnant shoppers, they can earn millions.

As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Soon after the new ad campaign began, Target's Mom and Baby sales exploded. The company doesn't break out figures for specific divisions, but between 2002 — when Pole was hired — and 2010, Target's revenues grew from $44 billion to $67 billion. In 2005, the company's president, Gregg Steinhafel, boasted to a room of investors about the company's "heightened focus on items and categories that appeal to specific guest segments such as mom and baby."

Antonio Bolfo/Reportage for The New York Times

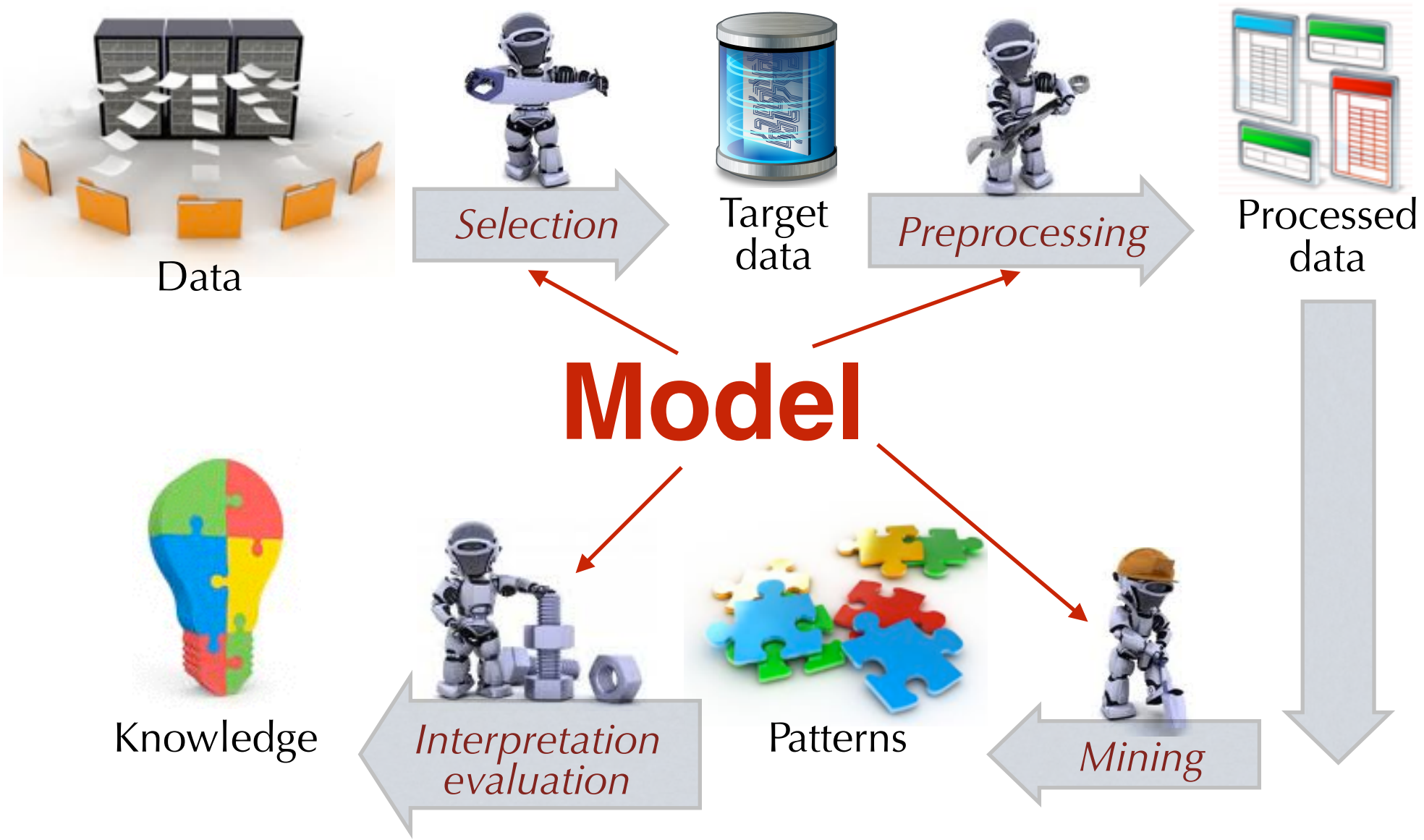
By CHARLES DUHIGG

Published: February 16, 2012 | 570 Comments

Not all about advertising:

World-changing applications: charities, healthcare, politics

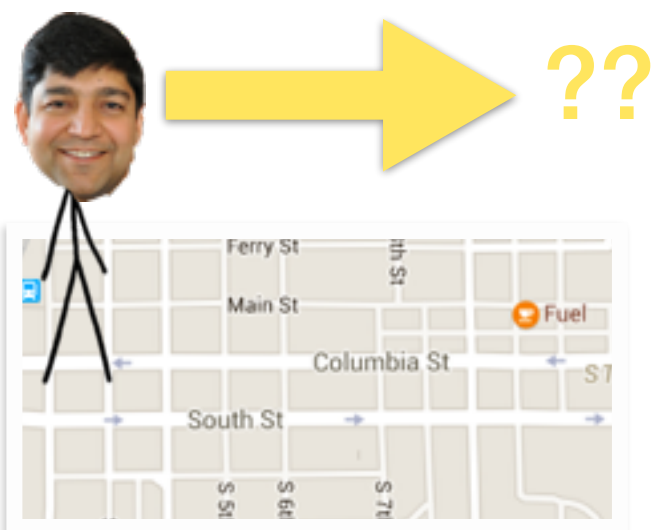
Standard view of data mining process

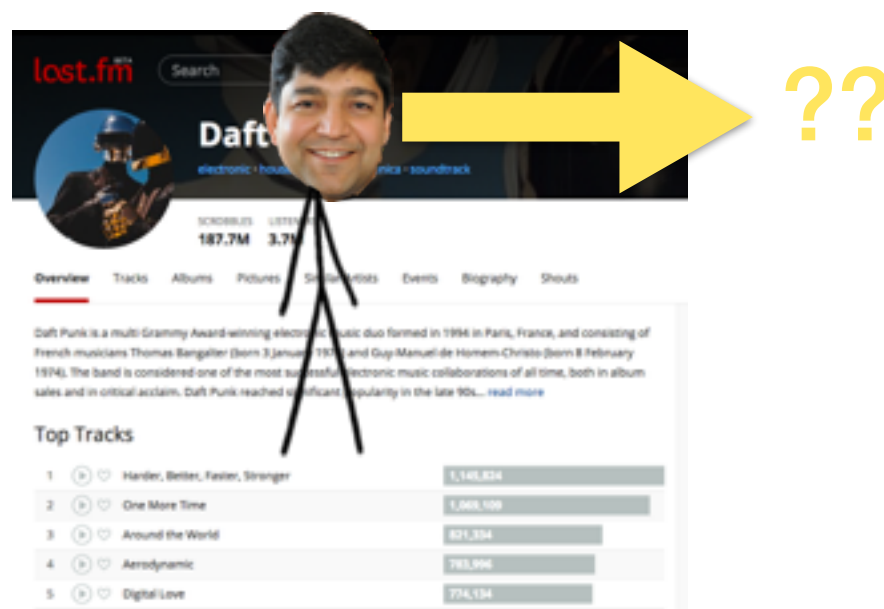


Example

Case Study: Predicting What You Will do Next

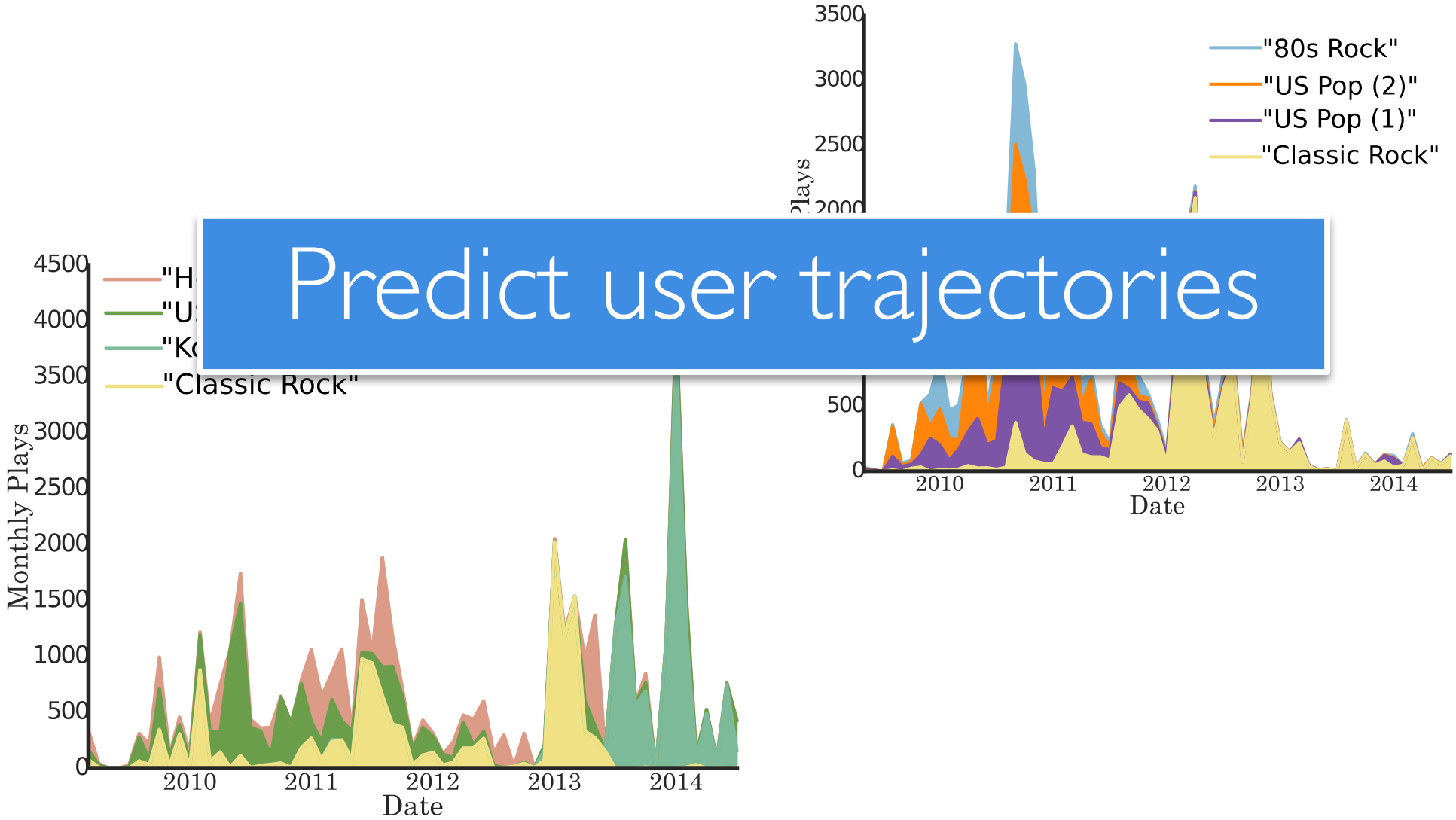
- Predict where agent will go next





Motivation (a)

last·fm
the social music revolution



Predict user trajectories

Motivation (b)

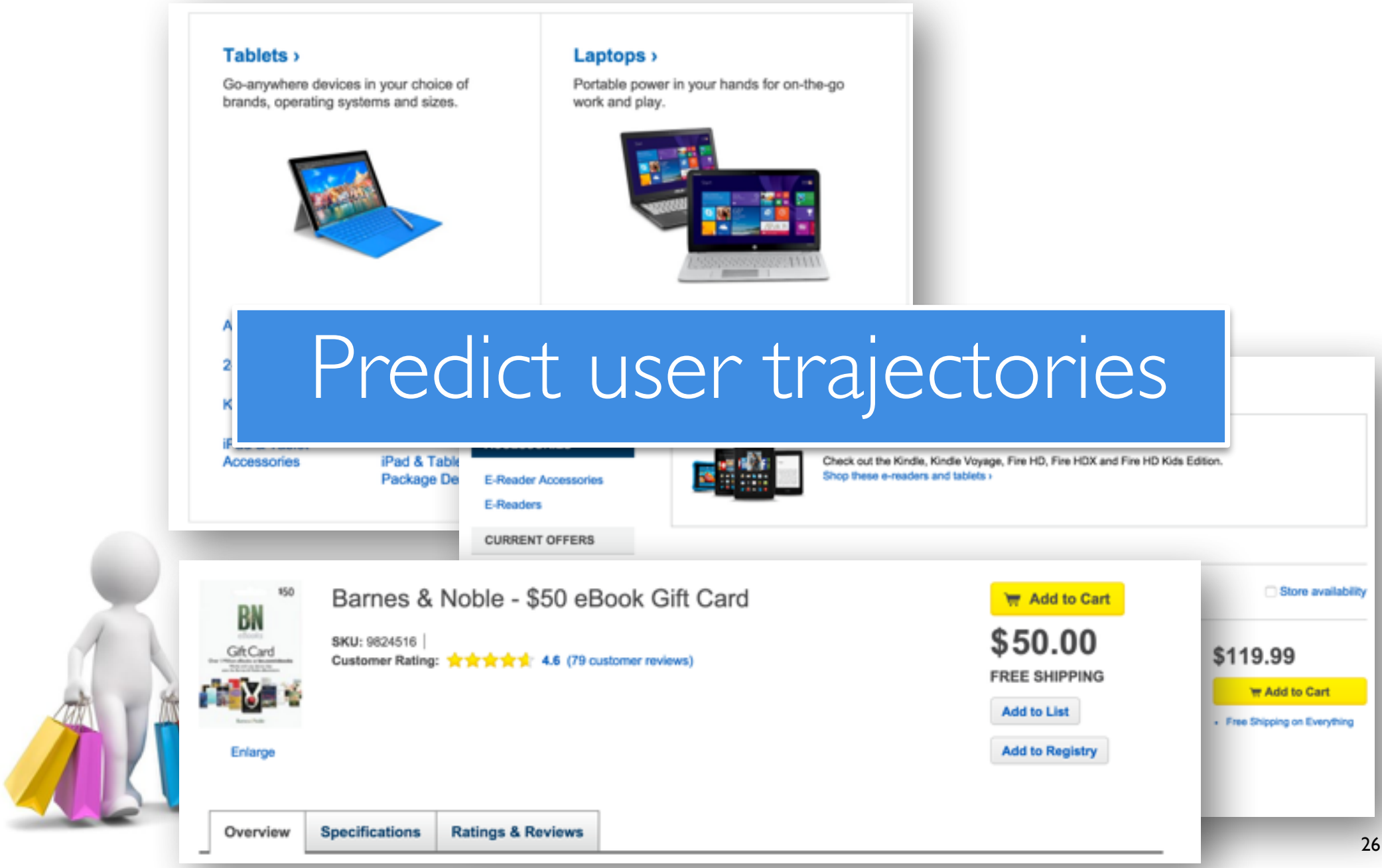
*

Predict user trajectories

* Understanding the spreading patterns of mobil González, C.A. Hidalgo and A.-L.Barabási, Science 324, 1071-1076 (2009).

brightkite
people. places. friends.

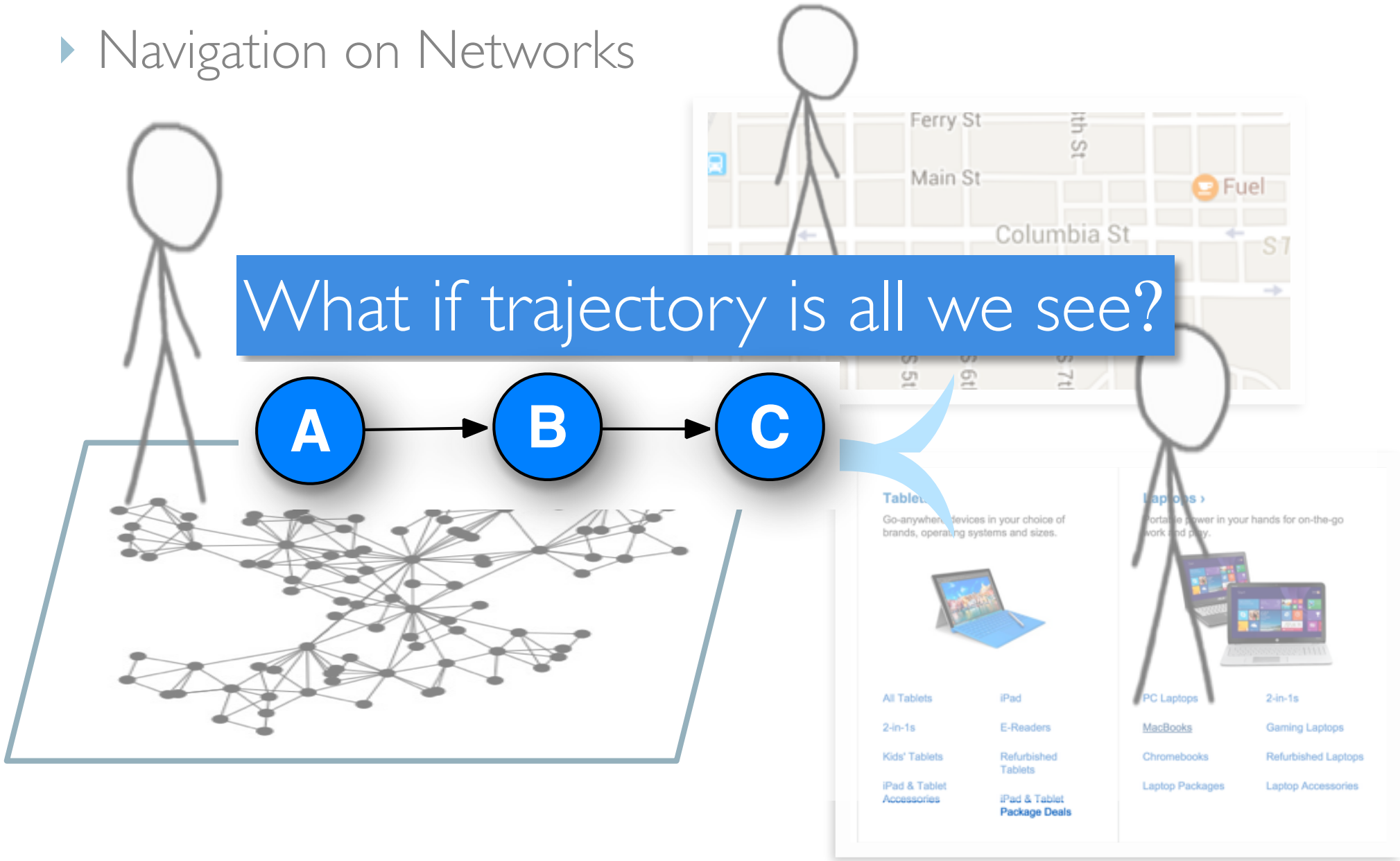
Motivation (c)



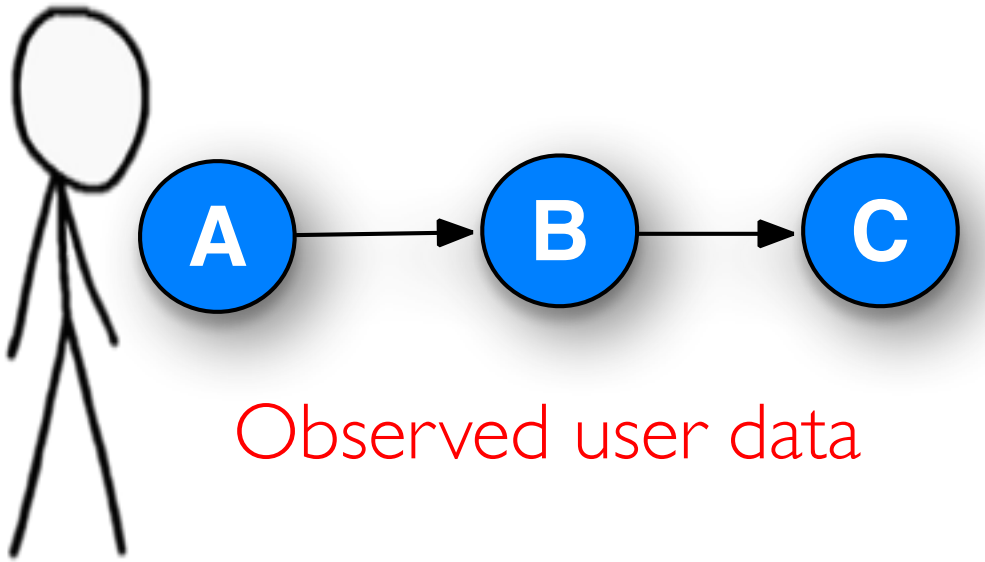
Many complex natural, physical and social systems can be represented as some form of dynamics on graphs

Difficult Problem Due to Hidden Network Constraints

- Navigation on Networks

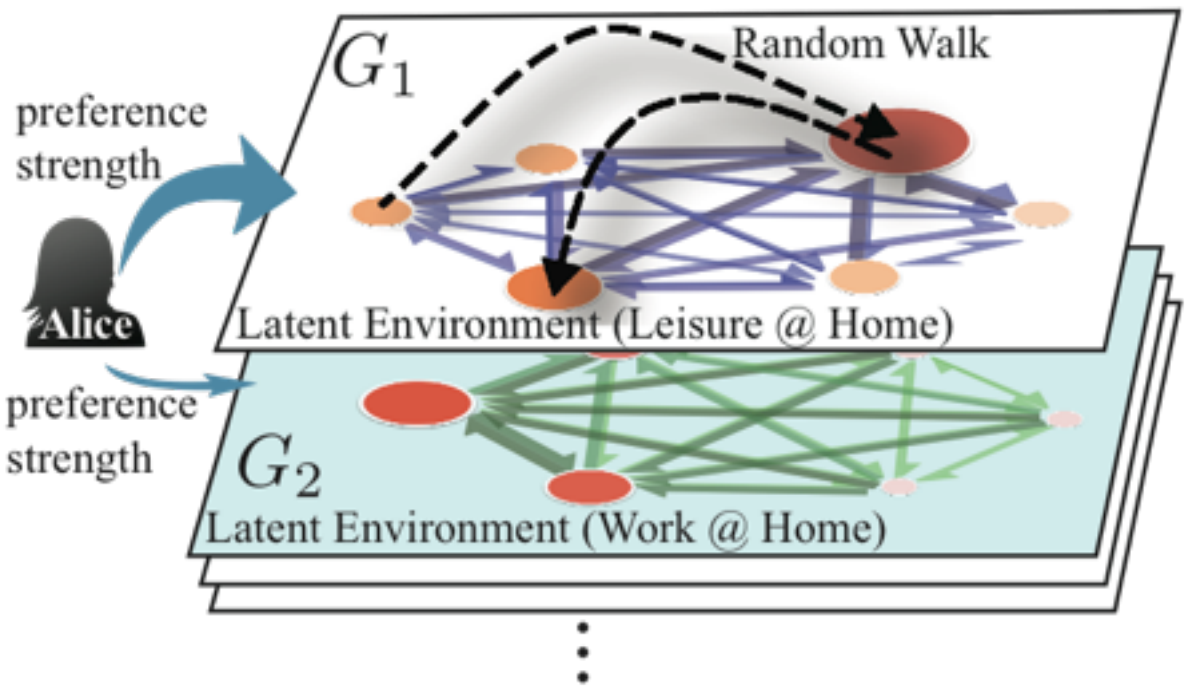


Learns Model from Data

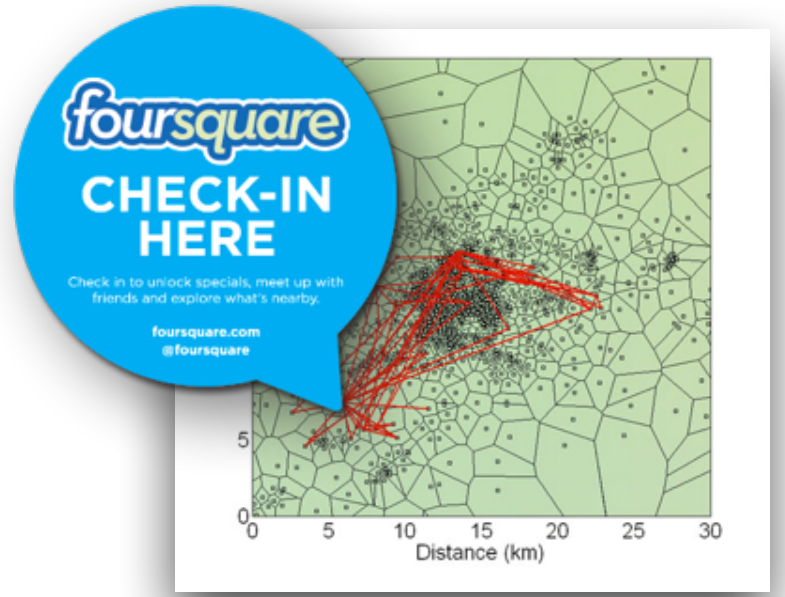


Observed user data

Model



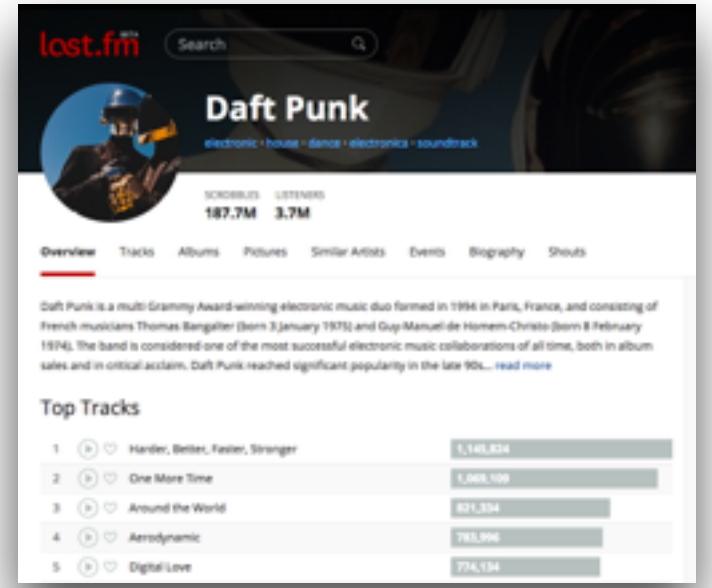
Foursquare prediction example

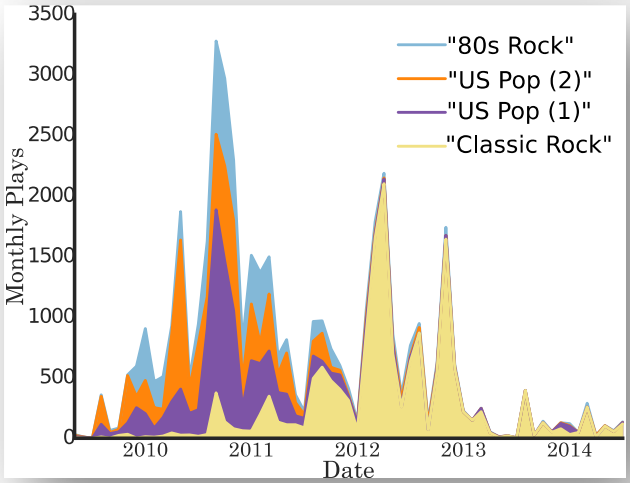


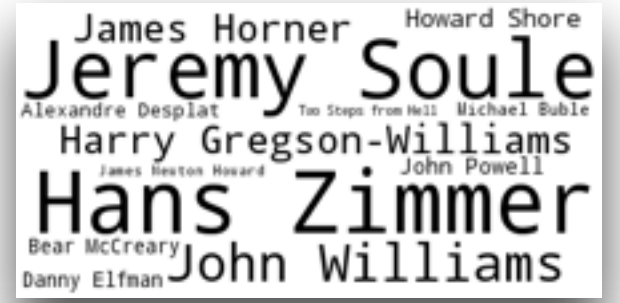
Past Approaches

Example of Uncovered Relations

Last.fm: Uncovered Music Genres

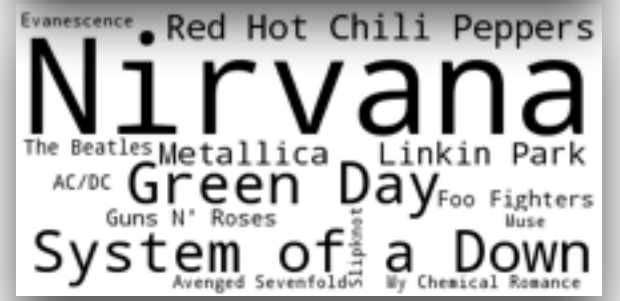




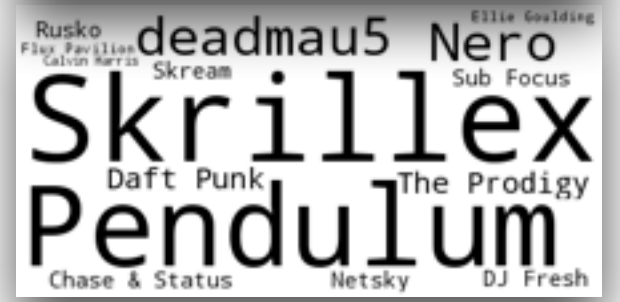


Movie Soundtracks

Trash Metal



90's Rock



Electro House

Topics

- Elements of data mining algorithms
- Statistics and background
- A/B Testing
- Link analysis & Prediction
- Classification tasks
- Collaborative filtering
- Latent variable models
- Model selection
- Hierarchical Models + Deep Learning

Questions?