Introduction to Data Science

~~Data Warehousing and Data Mining~~

Sunsern Cheamanunkul

sunsern.che@mahidol.edu

Mahidol University
International College

Course Structure

- Lectures
  - 2 lectures / week
  - 10-minute break after the first hour
  - Bring your laptop to lectures
    - There will be in-class exercises
- Weekly programming assignments
- Midterm Exam
- Project

Logistics

- Course website (syllabus, lecture slides, assignments, etc.)
  - https://canvas.instructure.com/courses/1112677
  - code: **JALGRW**
- Attendance is **strongly recommended**.
  - Your participation score is highly correlated with your attendance rate.

Assignments

- There will be total of ~10 assignments.
- Some of them will be continuation or improvement on the in-class exercises.
- Solutions must be submitted as Python notebooks.
- Hence, you are expected to know Python, or have ability to learn Python really fast.

Project

- Team of 1-2 students
  - Your choice!
  - Start looking for a team early!
- Can be any data science related project from any domain such as audio, images, videos, gaming, finance etc.
- You should use original datasets which mean you are the one to collect them yourself.
- However, if your project idea requires existing datasets, exceptions can be made.

Project

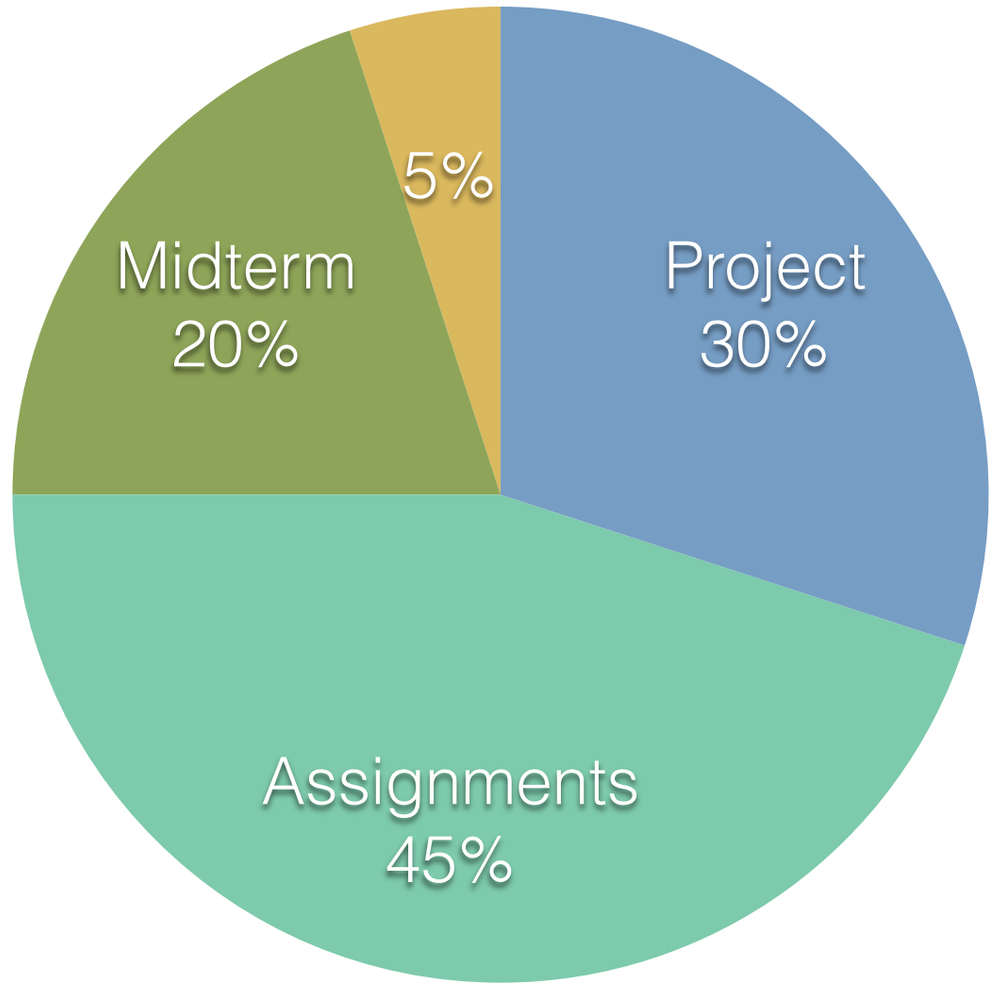
- Proposal:
  - During week 5, each team will present:
    - project overview.
    - details about data collection process.
    - what will be studied and how the analysis will be carried out.
- Checkpoint:
  - During week 8-10
  - Informal status update
- Final presentation:
  - During week 12

Exams

- Midterm
  - Mixed of multiple-choice and open-ended questions.
  - Exam date will be announced on Canvas. Approximately during week 7.
    - You will be tested on the concepts and materials presented in class.
- No final exam!
  - We will have final project presentation instead of the final exam.

Grading policy

- Project 30%
  - Proposal 5%
  - Checkpoint 5%
  - Final presentation 20%
- Assignments 45%
- Midterm exam 20%
- Participation 5%



Tentative Schedule

Week 1: Introduction to Data Science
Week 2: Data collection and warehousing
Week 3: Working with data
Week 4: Machine Learning
Week 5: Regression
Week 6: Nearest Neighbor, Decision Tree, SVM
— Midterm Exam —
Week 7: Clustering
Week 8: Ensemble Methods
Week 9: Natural Language Processing
Week 10: Neural Networks
Week 11: Recommender System
Week 12: Itemset Mining
— Project Presentation —

Expectations

- You will be working closely with data, a lot of data actually. Unless you can process large data sets by hand, programming skill is a must for this course.
- You should possess basic probability and statistics knowledge.
- You should be able to read API documentations.
- This is a practical course. You are expected to work on your project and assignments. They count for 75% of your grade.

Collaboration Policy

- We encourage collaboration, but do not plagiarize.
- Copying code from other people is not acceptable.
- When in doubt, cite your sources in your work.

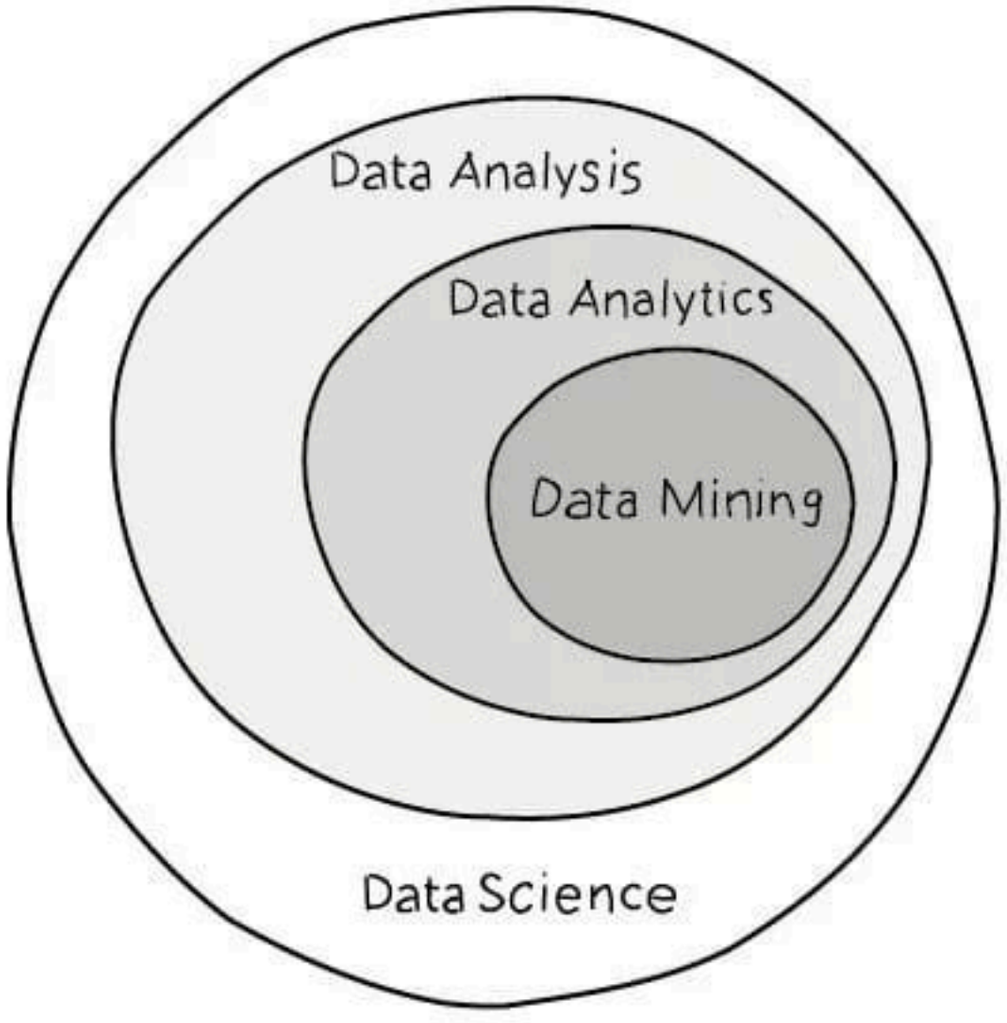
Getting help

- Office hours:
  - Monday, Wednesday 10-12pm
  - Tuesday, Thursday 4-6pm
  - You can find me either in the Science division or 1409.
- Email:
  - sunsern.che@mahidol.edu
  - sunsern@gmail.com (just in case)

Data Mining

- Data mining is a process of discovering structures, relationships, or “models” from (large) datasets.
- Why data mining?
  - We have large amount of data, but only a small fraction is knowledge.
  - Automated data collection tools, larger and cheaper storage devices, computerized society
- Data sources:
  - Business: web, e-commerce, transactions, stock market, etc.
  - Science: sensors technology, bioinformatics, simulations, etc.
  - User-generated contents: news, blogs, photos, youtube, etc.

Data Science



Harvard
Business
Review

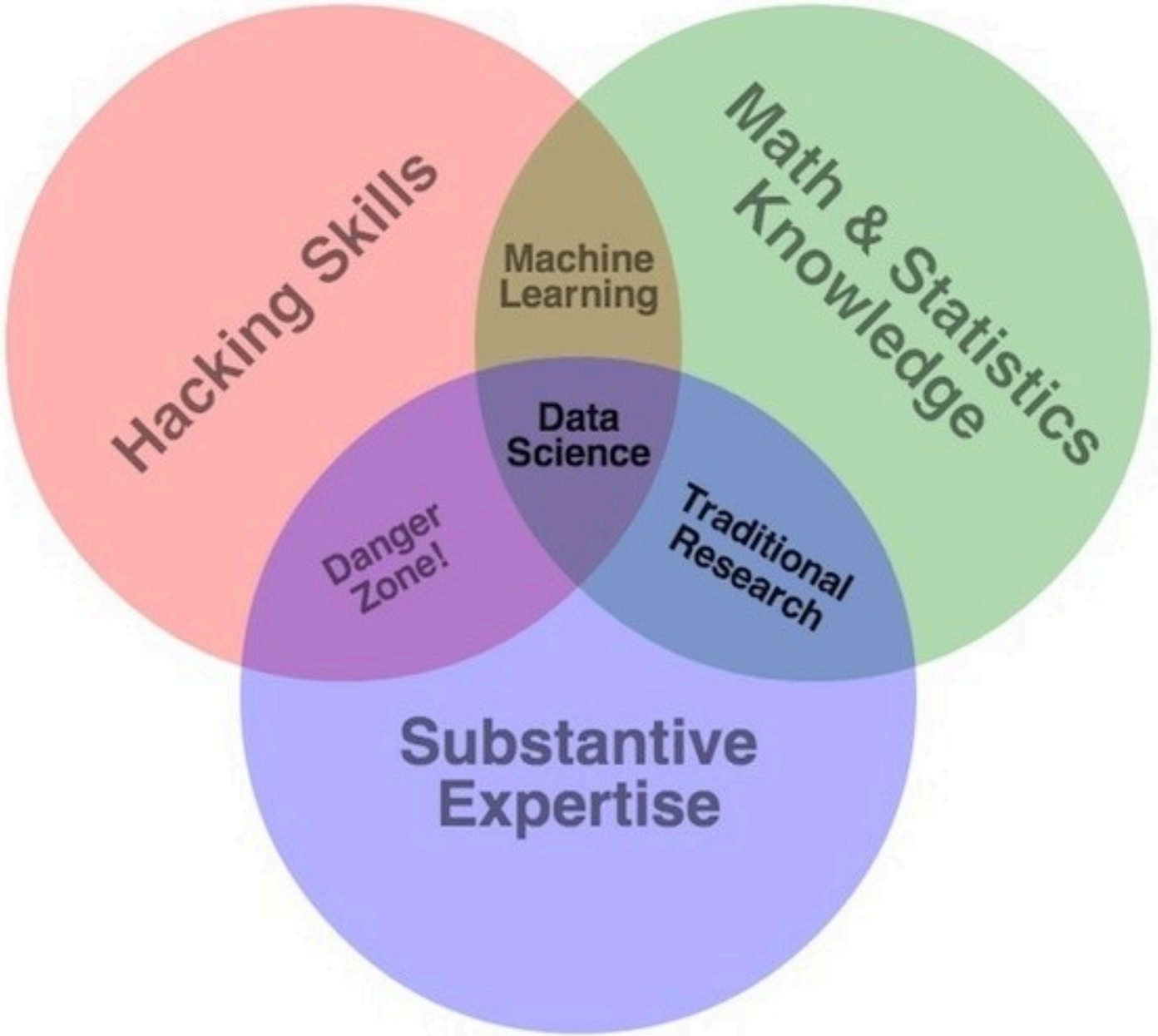
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Introduction to Data Science



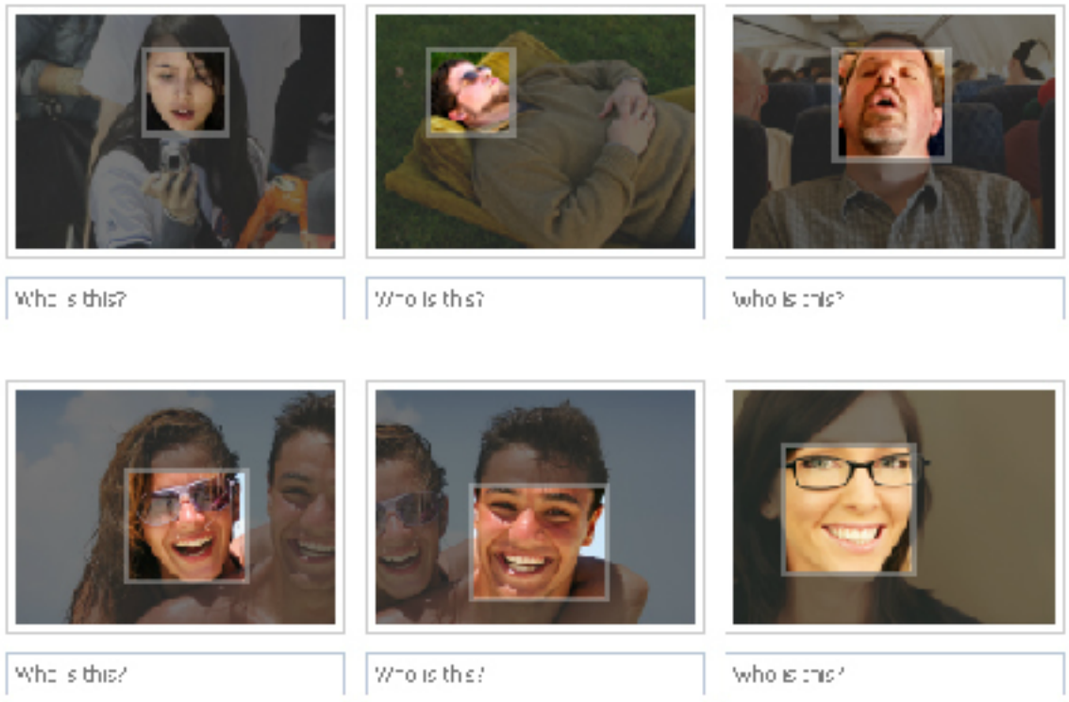
Types of modeling

- Pattern mining

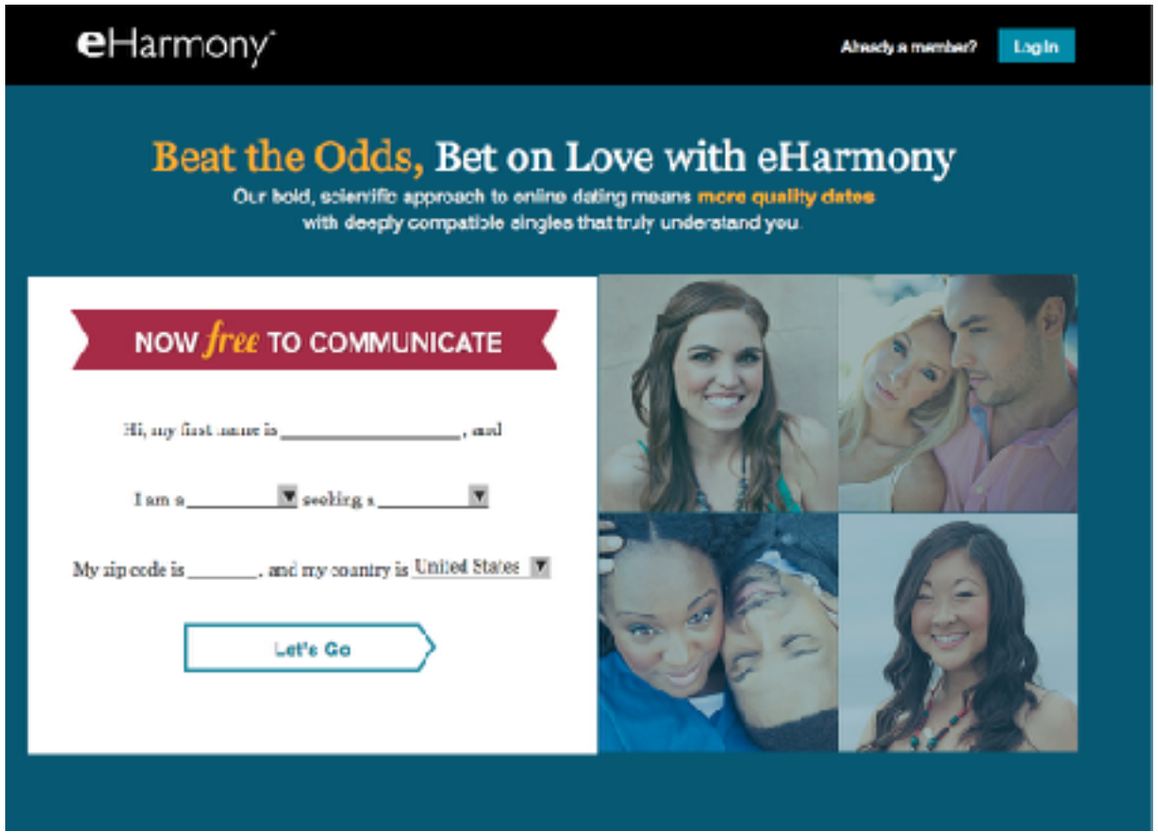
amazon.com

Types of modeling

- Predictive modeling



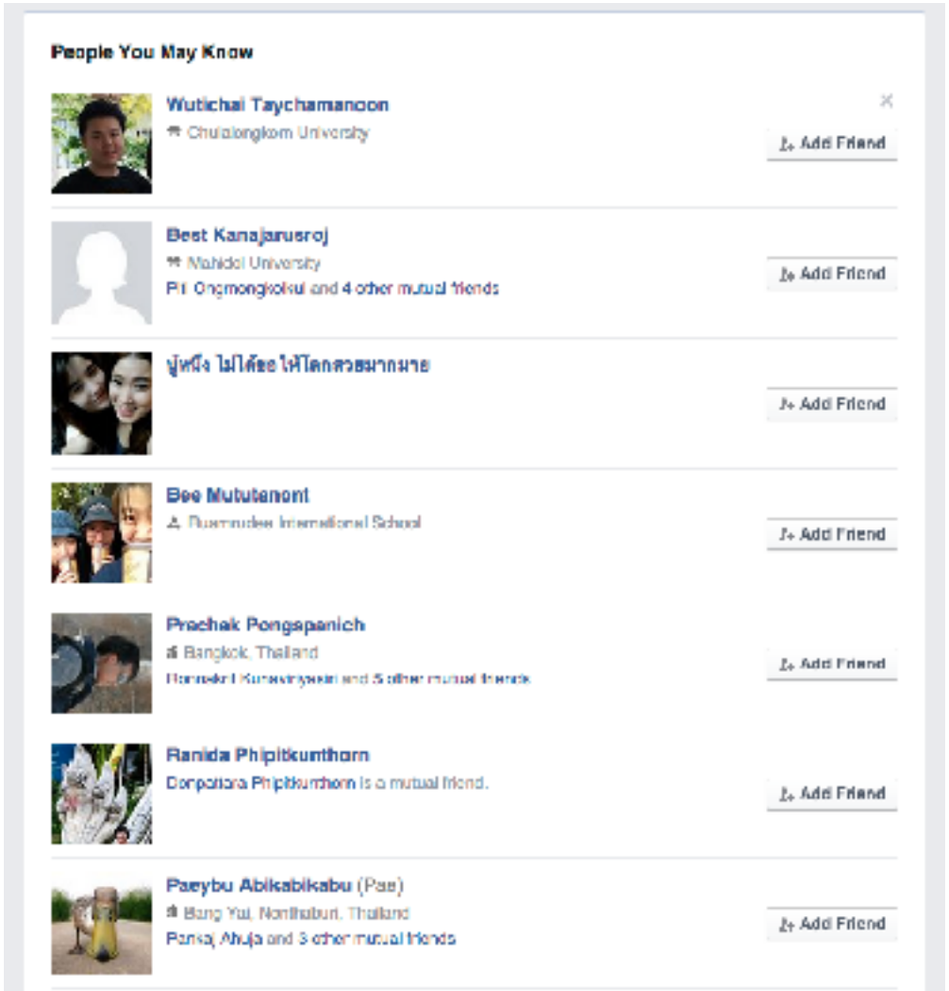
Facebook's DeepFace — facial recognition algorithm



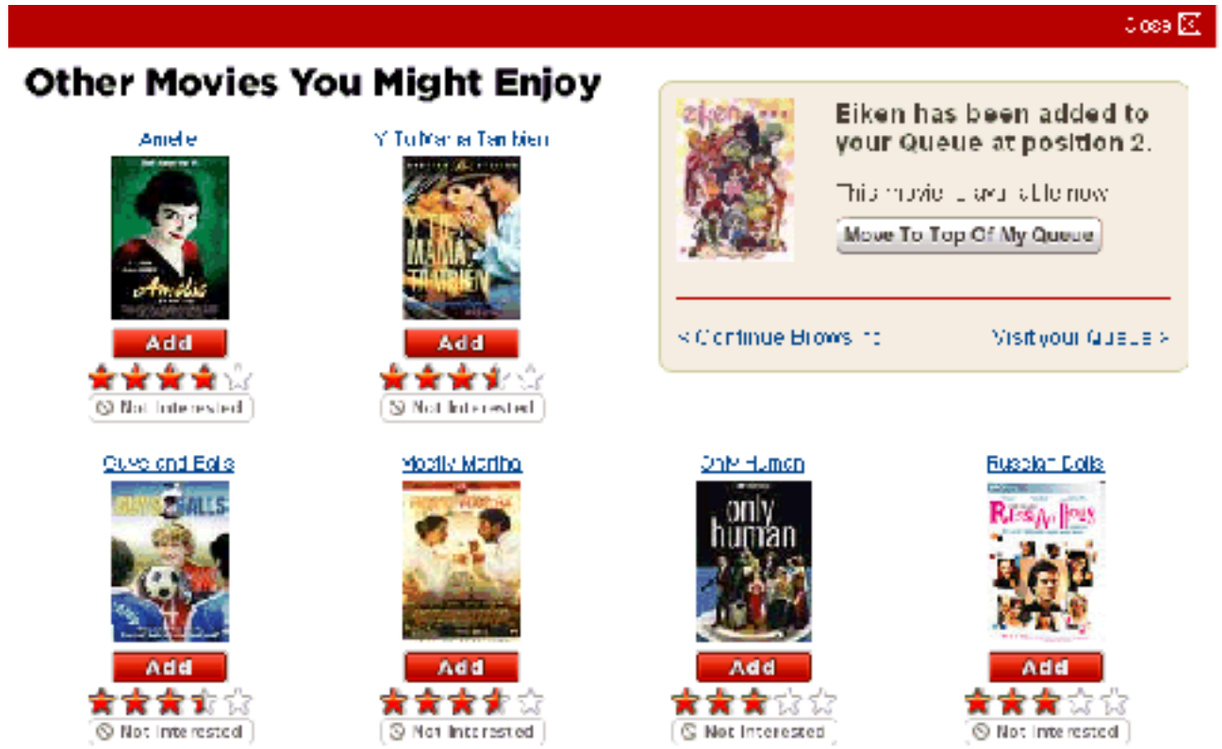
eHarmony — a dating site

Types of modeling

- Clustering



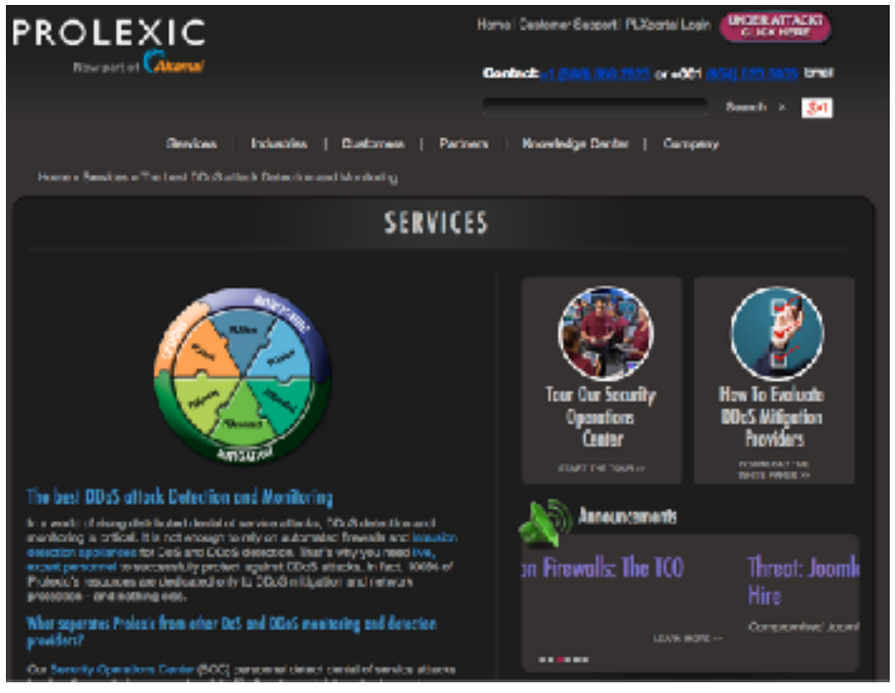
Facebook friend suggestions



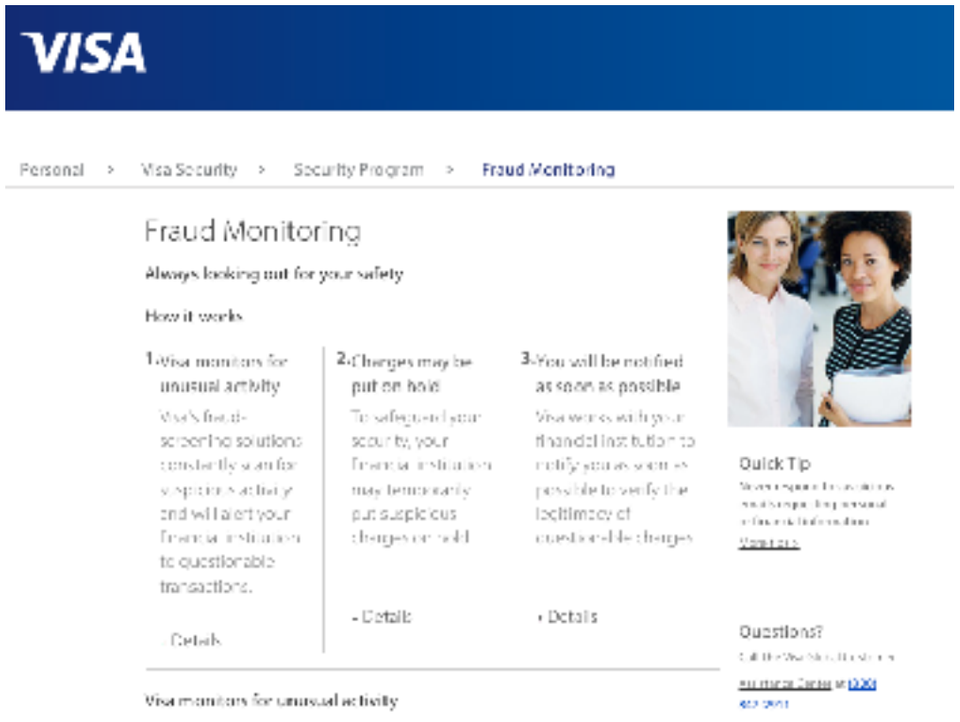
Netflix movie recommendations

Types of modeling

- Anomaly detection



DDOS attack monitoring



Credit card fraud detection

Typical workflow

Data collection

- Acquiring data from (multiple) sources
- Storing / Warehousing
- Cleaning
- Manipulating
- Feature extraction / engineering

Model training

- Pattern mining
- Clustering
- Classification
- Regression

Visualizing results

- Summary
- Tables
- Plotting
- Infographic

Statistical Limits on Data Mining

- When you have a large amount of data, and you look for specific things in the data. You will find them, even if they data is completely random.
- These findings are considered “bogus”.
- This is known as Bonferroni’s principle

Example of Bonferroni's Principle

- Suppose there are some "evil-doers" out there and we want to detect them before they do bad things.
- We have reasons to believe that these evil-doers often gather at a hotel to come up evil plans.
- Assume the following:
  - There are **1 billion people** who might be evil-doers.
  - Everyone goes to a hotel **1 day in 100 days**.
  - A hotel holds 100 people. Hence, there are 100,000 hotels — enough to hold 1% of people who visit a hotel on a given day.
  - We look at the hotels records for **1000 days**

Example of Bonferroni's Principle

- We want to find out who, on two different days, were both at the same hotel.
- First, suppose there are no evil-doers. Everyone picks hotels at random.
- The probability of any two people both deciding to visit a hotel is 0.01 * 0.01 = 0.0001.
- The chance that they will visit the same hotel is $0.0001 / 100000 = 10^{-9}$
- The chance of that happening on two different days is 10^{-18}

Example of Bonferroni's Principle

- Now we consider data from 1 billion people.
- The number of pairs of people is $\binom{10^9}{2} \approx 5 \times 10^{17}$
- The number of pairs of days is $\binom{1000}{2} \approx 5 \times 10^5$
- The number of events that look like evil-doing is $5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} = 250,000$
- Suppose, there really are only 10 pairs of evil-doers. That means there must be something wrong with our method of finding evil-doers.

Next Step

- Setting up software stack
  - IPython (notebook)
  - Numpy
  - Scipy
  - Pandas
  - etc.
- Easier to just install Anaconda:
  - https://www.continuum.io/downloads
- Now go through the Numpy and Pandas tutorial

These might interest you…

- Kaggle competitions
- Hackathons