Machine Learning for Data Science (CS4786)
Lecture 1

Tu-Th 8:40 AM to 9:55 AM
Klarman Hall KG70

Instructor : Karthik Sridharan

Welcome the first lecture!


WINTER IS ~~COMING~~
Here

I know its really early!


WINTER IS ~~COMING!~~
Here

But lets have fun!

The Awesome TA's

1. Cameron Benesch
2. Rajesh Bollapragada
3. Ian Delbridge
4. William Gao
5. Varsha Kishore
6. Clara Liu
7. Jeffrey Liu
8. Amanda Ong
9. Jenny Wang
10. Wilson Yoo

Course Information

- Course webpage is the official source of information: http://www.cs.cornell.edu/Courses/cs4786/2019sp
- Join Piazza: https://piazza.com/class/jr4fi8k75d571p
- TA office hours will start from week 2. Time and locations will be posted on info tab of course webpage
- Basic knowledge of python is required.

Placement Exam!

- Passing the placement exam is required to enroll
- Exam can be found at: http://www.cs.cornell.edu/courses/cs4786/2019sp/hw0.html
- Upload your solutions in PDF format via the google form indicated in the exam page
- Score on the placement exam is only for feedback, does not count towards grades for the course.

Course Grades

- Assignments: 28%
- Prelims: 20%
- Finals: 20%
- Competition: 30%
- Survey: 2%

ASSIGNMENTS

- Total of 4 assignments.
- Each worth 7% of the grade
- Will be on Vocareum (using Jupyter notebook/python)
- **Has to be done individually**

Exams

1. Prelim
    - On March 28th at STL185
    - Worth 20% of the grades.
2. Finals
    - On May 14th (see schedule for details)
    - Worth 20% of the grades.

COMPETITIONS

- One in-class Kaggle competition worth 30% of the grade
- You are allowed to work in groups **of at most 4**.
- Kaggle scores only factor in for part of the grade.
- Grades for project focus more on thought process (demonstrated through your reports)

SURVEYS

- 2 Surveys worth 1% each just for participation
- Survey will be anonymous (I will only have a list of students who participated)
- Important form of feedback I can use to steer the class
- Free forum for you to tell us what you want.

Academic Integrity

1. **0 Tolerance Policy: no exceptions**
   - We have checks in place to look for violations in Vocareum
2. If you use any source (internet, book, paper, or personal communication) cite it.
3. When in doubt cite.

Some info about class . . .

- **Would really love it to be interactive.**
- You can feel free to ask me anything
- We will have **informal** quizzes most classes
- We will have a review session for exams

Lets get started . . .

Data Deluge

- Each time you use your credit card: who purchased what, where and when

Data Deluge

- Each time you use your credit card: who purchased what, where and when
- Netflix, Hulu, smart TV: what do different groups of people like to watch

Data Deluge

- Each time you use your credit card: who purchased what, where and when
- Netflix, Hulu, smart TV: what do different groups of people like to watch
- Social networks like Facebook, Twitter, . . . : who is friends with who, what do these people post or tweet about

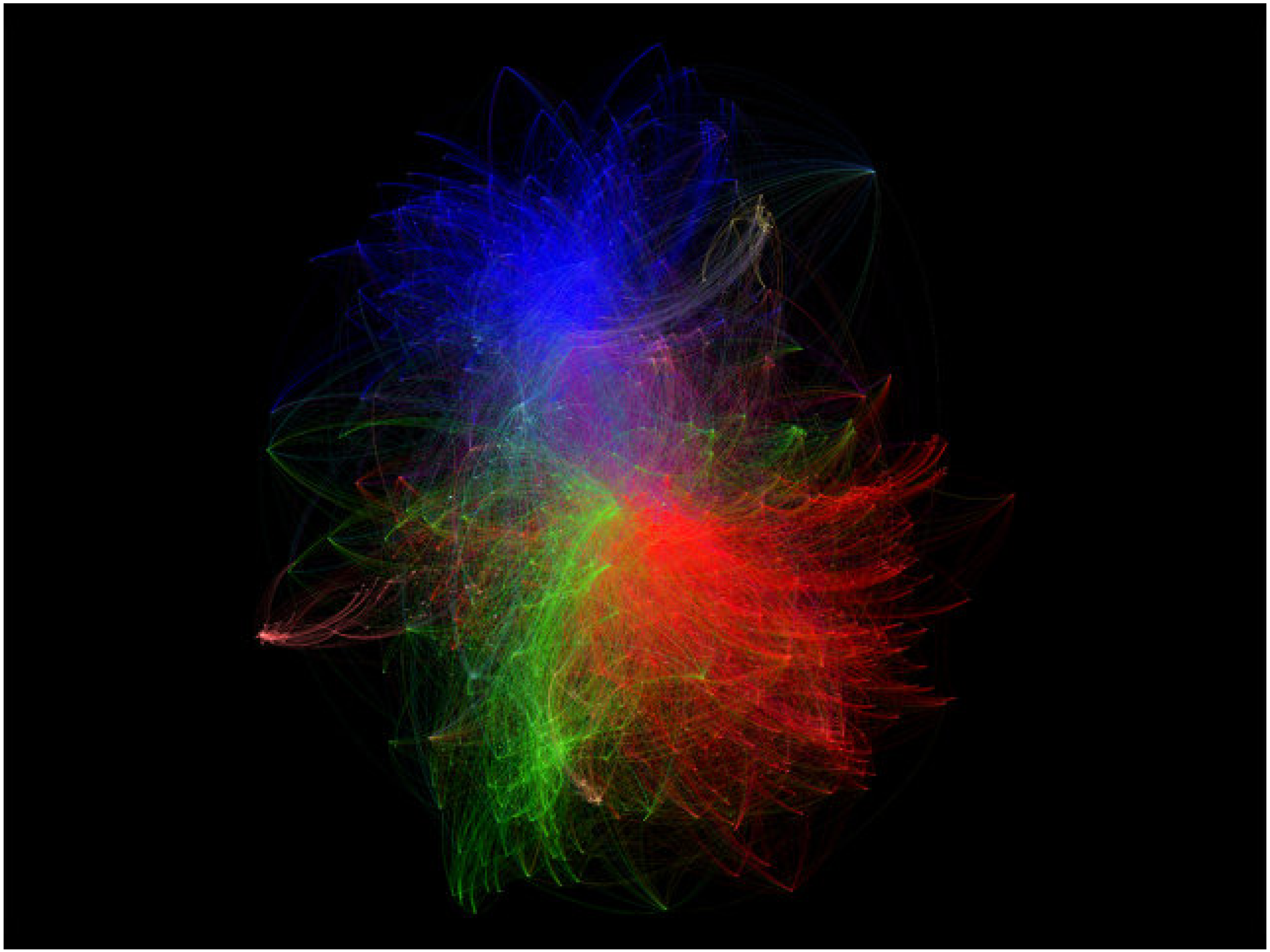
Data Deluge

- Each time you use your credit card: who purchased what, where and when
- Netflix, Hulu, smart TV: what do different groups of people like to watch
- Social networks like Facebook, Twitter, . . . : who is friends with who, what do these people post or tweet about
- Millions of photos and videos, many tagged

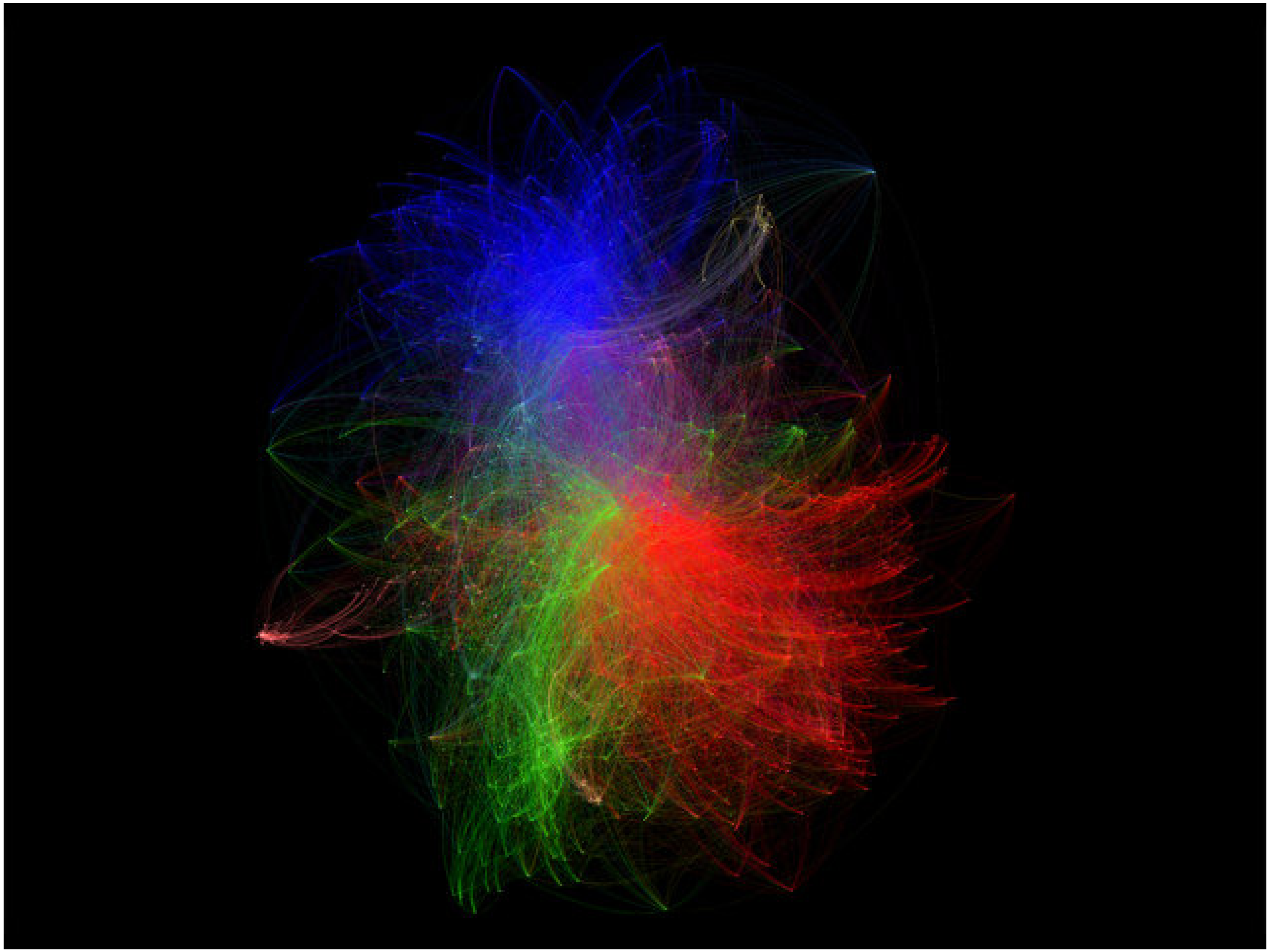
DATA DELUGE

- Each time you use your credit card: who purchased what, where and when
- Netflix, Hulu, smart TV: what do different groups of people like to watch
- Social networks like Facebook, Twitter, . . . : who is friends with who, what do these people post or tweet about
- Millions of photos and videos, many tagged
- Wikipedia, all the news websites: pretty much most of human knowledge

Guess?



Social Network of Marvel Comic Characters!



by Cesc Rosselló, Ricardo Alberich, and Joe Miro from the University of the Balearic Islands

What can we learn from all this data?

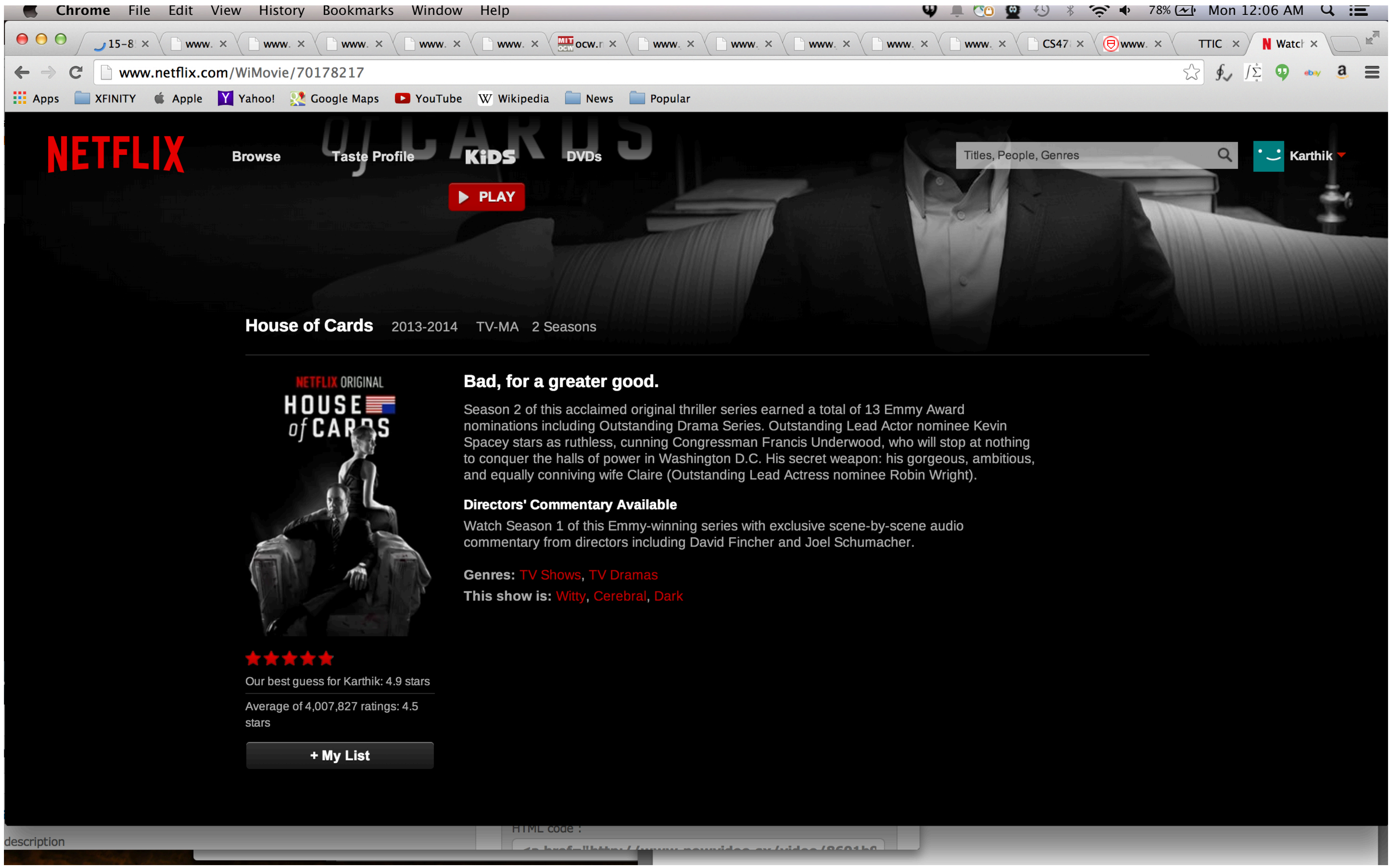
What is Machine Learning?

Use data to automatically learn to perform tasks better.

Close in spirit to T. Mitchell's description

WHERE IS IT USED ?

Movie Rating Prediction



Where is it used ?

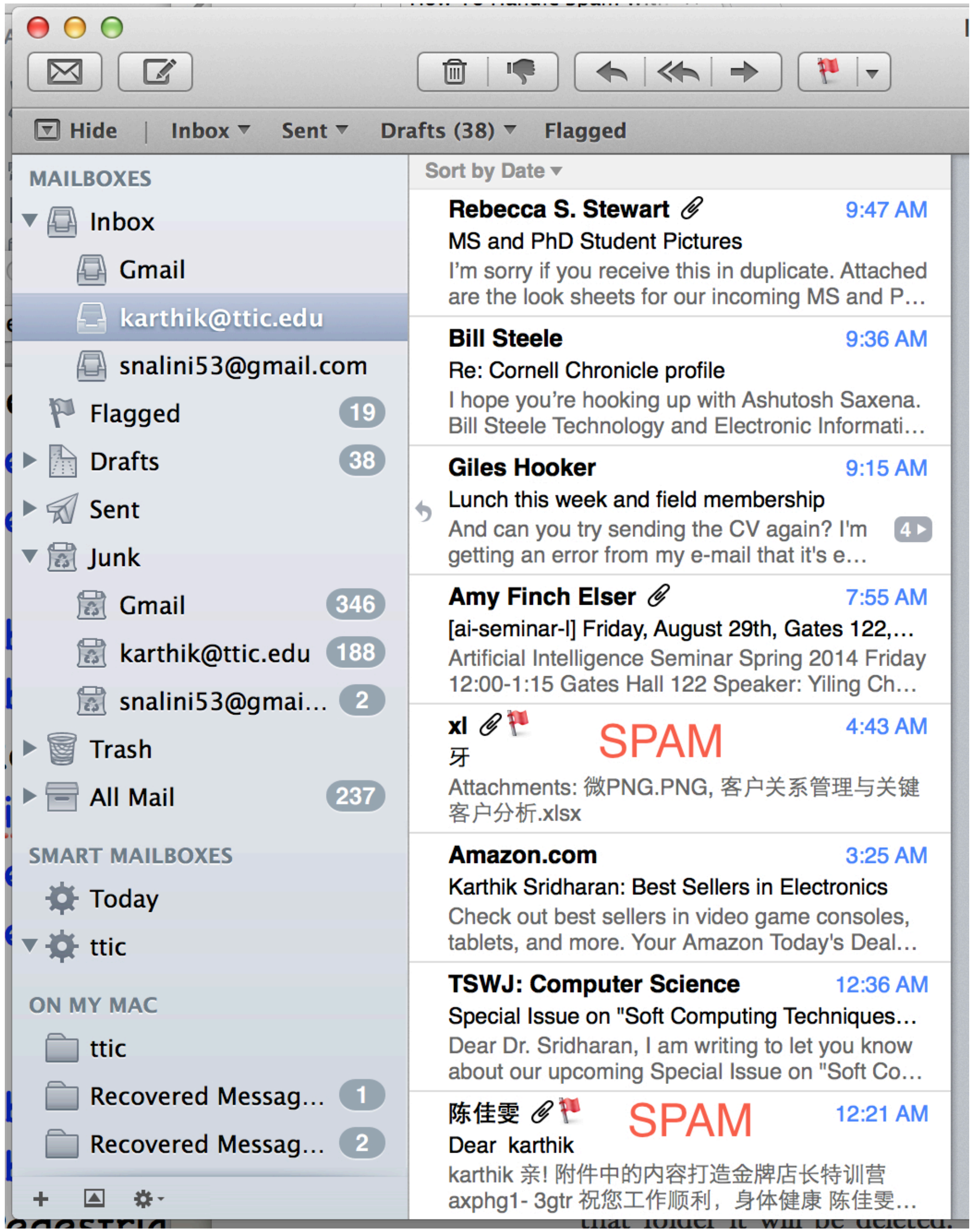
Pedestrian Detection

WHERE IS IT USED ?

Market Predictions

WHERE IS IT USED ?

Spam Classification



More Applications

More Applications

- Each time you use your search engine

More Applications

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings

More Applications

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings
- Biometrics: reason you shouldn't smile

More Applications

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings
- Biometrics: reason you shouldn't smile
- Recommendation systems: what you may like to buy based on what your friends and their friends buy

More Applications

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings
- Biometrics: reason you shouldn't smile
- Recommendation systems: what you may like to buy based on what your friends and their friends buy
- Computer vision: self driving cars, automatically tagging photos

More Applications

- Each time you use your search engine
- Autocomplete: Blame machine learning for bad spellings
- Biometrics: reason you shouldn't smile
- Recommendation systems: what you may like to buy based on what your friends and their friends buy
- Computer vision: self driving cars, automatically tagging photos
- Topic modeling: Automatically categorizing documents/emails by topics or music by genre
- …

Course Synopsis

- Primary focus: Unsupervised learning

Course Synopsis

- Primary focus: Unsupervised learning
- Roughly speaking 4 parts:

Course Synopsis

- Primary focus: Unsupervised learning
- Roughly speaking 4 parts:
  1. Dimensionality reduction:
     Principle Components Analysis, Random Projections, Canonical Components Analysis, Kernel PCA, tSNE, Spectral Embedding

Course Synopsis

- Primary focus: Unsupervised learning

- Roughly speaking 4 parts:

  1. Dimensionality reduction:
     Principle Components Analysis, Random Projections, Canonical Components Analysis, Kernel PCA, tSNE, Spectral Embedding

  2. Clustering:
     single-link, Hierarchical clustering, k-means, Gaussian Mixture model

Course Synopsis

- Primary focus: Unsupervised learning
- Roughly speaking 4 parts:
  1. Dimensionality reduction:
     Principle Components Analysis, Random Projections, Canonical Components Analysis, Kernel PCA, tSNE, Spectral Embedding
  2. Clustering:
     single-link, Hierarchical clustering, k-means, Gaussian Mixture model
  3. Probabilistic models and Graphical models
     Mixture models, EM Algorithm, Hidden Markov Model, Graphical models Inference and Learning, Approximate inference

Course Synopsis

- Primary focus: Unsupervised learning
- Roughly speaking 4 parts:

  1. Dimensionality reduction:
     Principle Components Analysis, Random Projections, Canonical Components Analysis, Kernel PCA, tSNE, Spectral Embedding
  2. Clustering:
     single-link, Hierarchical clustering, k-means, Gaussian Mixture model
  3. Probabilistic models and Graphical models
     Mixture models, EM Algorithm, Hidden Markov Model, Graphical models Inference and Learning, Approximate inference
  4. Socially responsible ML
     Privacy in ML, Differential Privacy, Fairness, Robustness against polarization

Unsupervised Learning

Given (unlabeled) data, find useful information, pattern or structure

Unsupervised Learning

Given (unlabeled) data, find useful information, pattern or structure

- Dimensionality reduction/compression : compress data set by removing redundancy and retaining only useful information

Unsupervised Learning

Given (unlabeled) data, find useful information, pattern or structure

- Dimensionality reduction/compression : compress data set by removing redundancy and retaining only useful information
- Clustering: Find meaningful groupings in data

UNSUPERVISED LEARNING

Given (unlabeled) data, find useful information, pattern or structure

- Dimensionality reduction/compression : compress data set by removing redundancy and retaining only useful information
- Clustering: Find meaningful groupings in data
- Topic modeling: discover topics/groups with which we can tag data points

Dimensionality Reduction

Given n data points in high-dimensional space, compress them into corresponding n points in lower dimensional space.

Why dimensionality reduction?

Why dimensionality reduction?

- For computational ease
    - As input to supervised learning algorithm
    - Before clustering to remove redundant information and noise

Why dimensionality reduction?

- For computational ease
  - As input to supervised learning algorithm
  - Before clustering to remove redundant information and noise
- Data visualization

WHY DIMENSIONALITY REDUCTION?

- For computational ease
  - As input to supervised learning algorithm
  - Before clustering to remove redundant information and noise
- Data visualization
- Data compression

Why dimensionality reduction?

- For computational ease
  - As input to supervised learning algorithm
  - Before clustering to remove redundant information and noise
- Data visualization
- Data compression
- Noise reduction

Dimensionality Reduction

Desired properties:

Dimensionality Reduction

Desired properties:

1. Original data can be (approximately) reconstructed

Dimensionality Reduction

Desired properties:

1. Original data can be (approximately) reconstructed

2. Preserve distances between data points

Dimensionality Reduction

Desired properties:

1. Original data can be (approximately) reconstructed
2. Preserve distances between data points
3. “Relevant” information is preserved

Dimensionality Reduction

Desired properties:

1. Original data can be (approximately) reconstructed
2. Preserve distances between data points
3. “Relevant” information is preserved
4. Redundant information is removed

Dimensionality Reduction

Desired properties:

1. Original data can be (approximately) reconstructed
2. Preserve distances between data points
3. "Relevant" information is preserved
4. Redundant information is removed
5. Models our prior knowledge about real world

Based on the choice of desired property and formalism we get different methods

Sneak Peek

- Linear projections
- Principle component analysis