CPSC 340:
Machine Learning and Data Mining

Mark Schmidt
University of British Columbia
Fall 2015

Outline

1) Intro to Machine Learning and Data Mining:
   - Big data phenomenon and types of data.
   - Definitions of data mining and machine learning.
   - Applications and impact.
2) Course Administrivia
3) Course Overview

Some images from this lecture are taken from Google Image Search.

Big Data Phenomenon

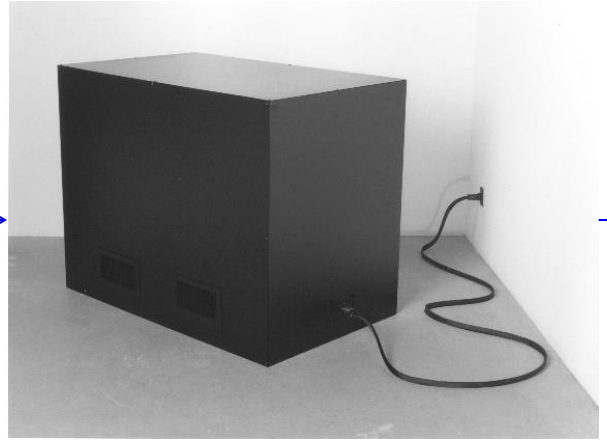
- We are collecting and storing data at an unprecedented rate.
- Examples:
  - News articles and blog posts.
  - YouTube, Facebook, and WWW.
  - Credit cards transactions and Amazon purchases.
  - Gene expression data and protein interaction assays.
  - Maps and satellite data.
  - Large hadron collider and surveying the sky.
  - Phone call records and speech recognition results.
  - Video game worlds and user actions.

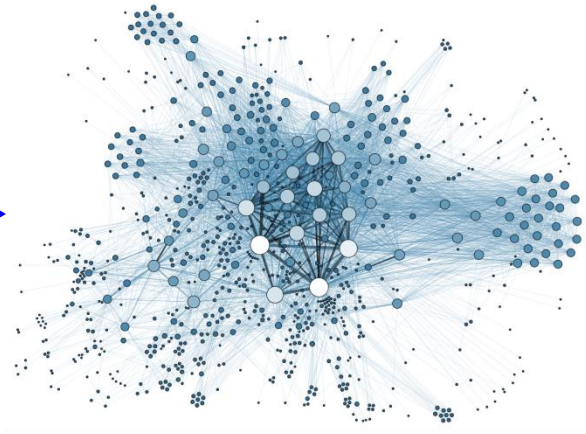
Big Data Phenomenon

- What do you do with all this data?
  - Too much data to search through it manually.
- But there is valuable information in the data.
  - How can we use it for fun, profit, and/or the greater good?
- Data mining and machine learning are key tools we use to make sense of large datasets.

Data Mining

- Automatically extract useful knowledge from large datasets.

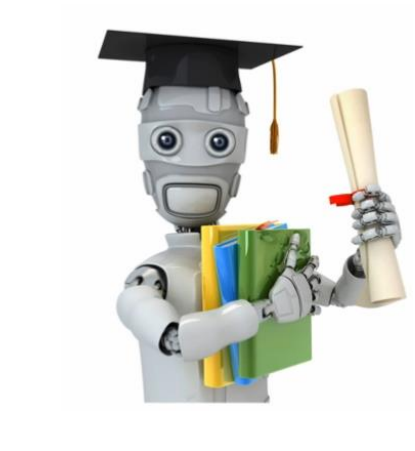




- Usually, to help with human decision making.

Machine Learning

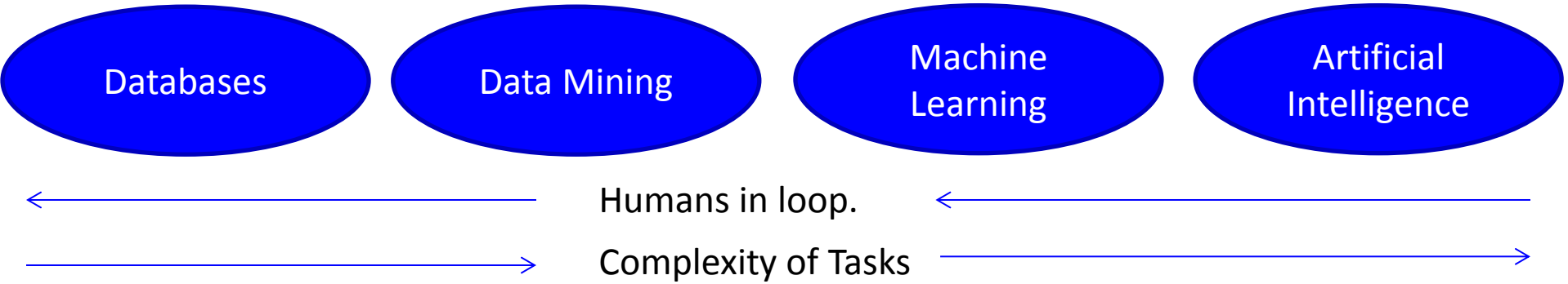
- Using computer to automatically detect patterns in data and use these to make predictions or decisions.



- Most useful when:
  - Don’t have a human expert.
  - Humans can’t explain patterns.
  - Problem is too complicated.

Data Mining vs. Machine Learning

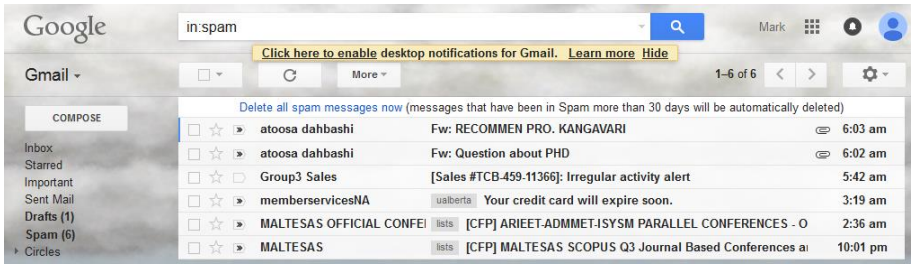
- DM and ML are very similar:
  - Data mining often viewed as closer to databases.
  - Machine learning often viewed as closer AI.



- Both similar to statistics, but less emphasis on 'correct' models and more on computation.

Applications

- Spam filtering:



- Credit card fraud detection:

Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

- Product recommendation:

Applications

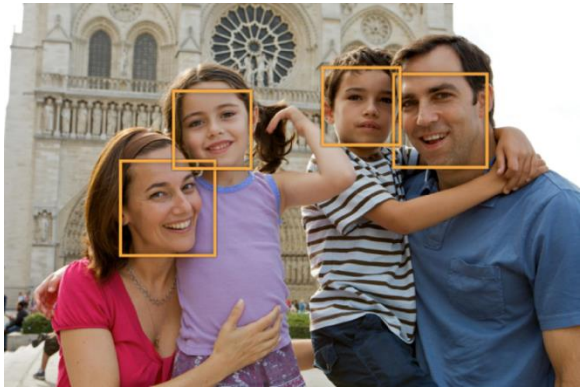
- Motion capture:

- Machine translation:

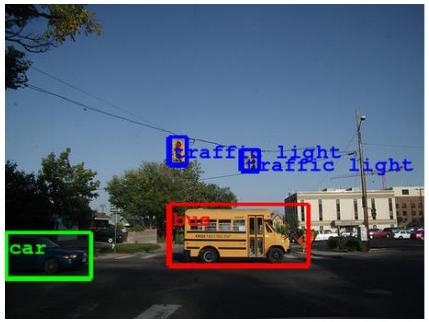
- Speech recognition:

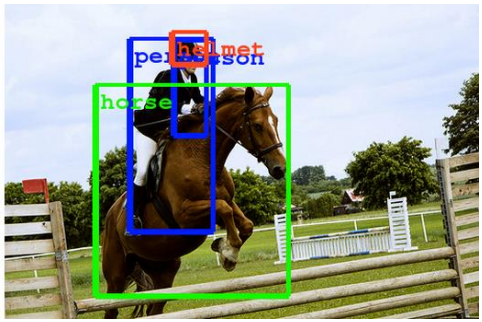
Applications

- Face detection:

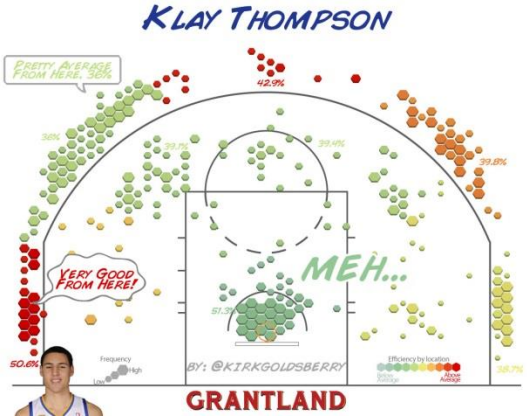


- Object detection:





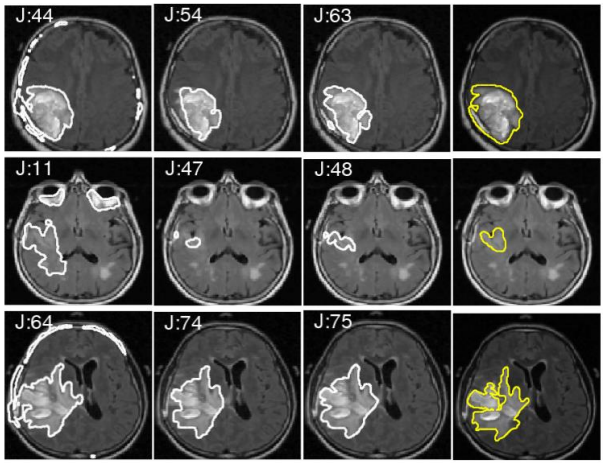
- Sports analytics:



Applications

- Personal Assistants:

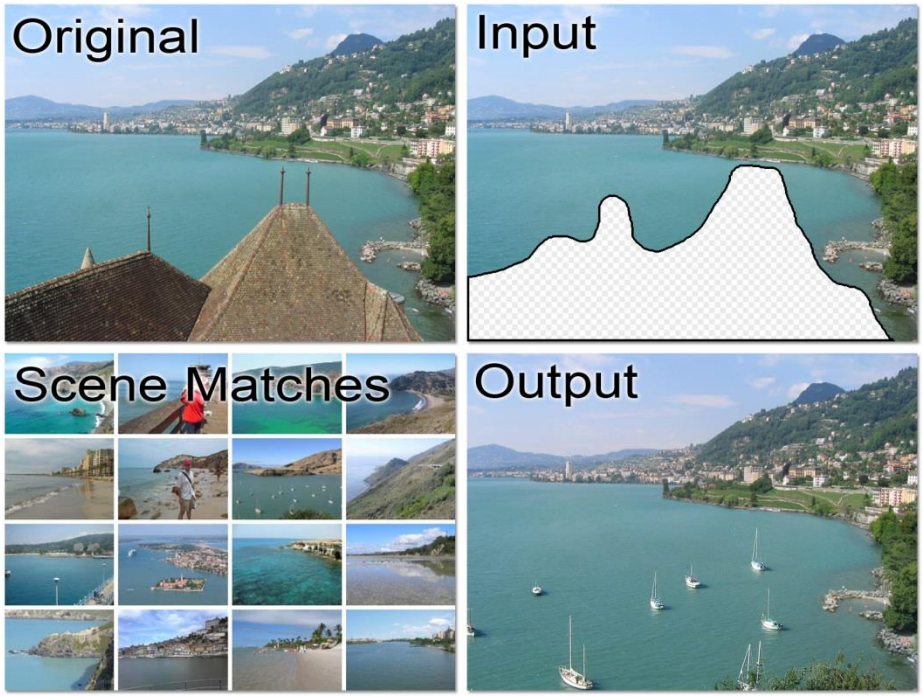
- Medical imaging:



- Self-driving cars:

Applications

- Scene completion:



- Image annotation:

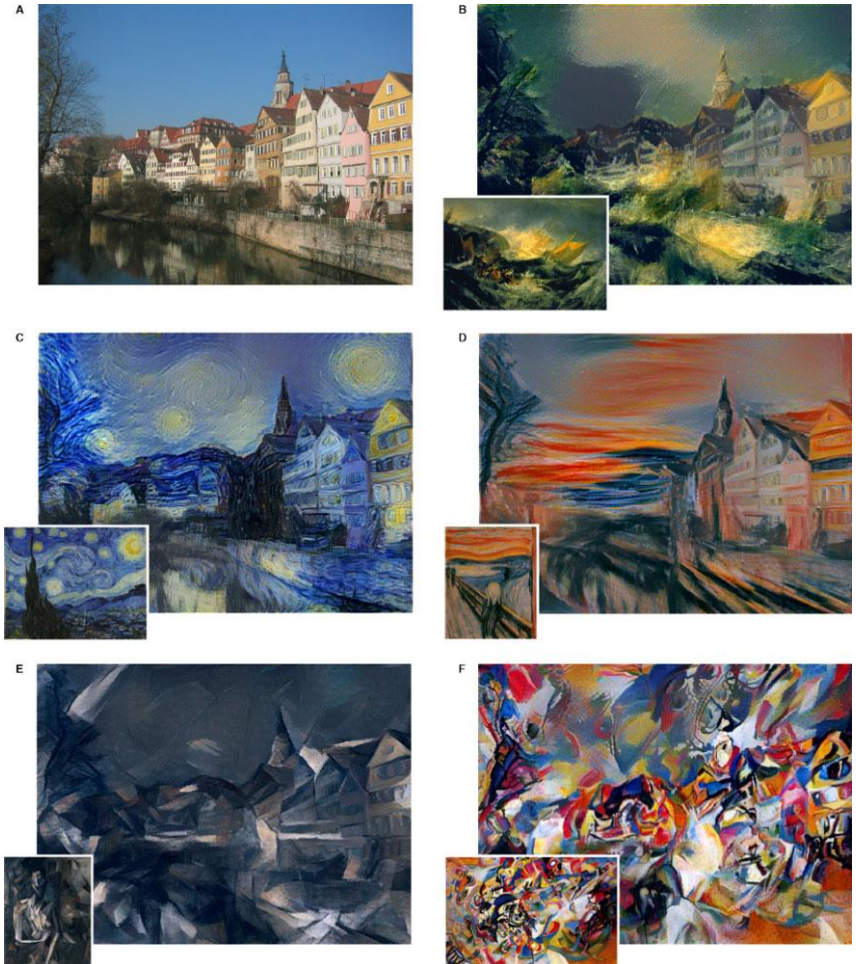
a cat is sitting on a toilet seat
logprob: -7.79

a display case filled with lots of different types of donuts
logprob: -7.78

a group of people sitting at a table with wine glasses
logprob: -6.71

Applications

- Inceptionism, mimicking art styles:



- Summary:
  - There is a lot you can do with a bit of statistics and a lot data.

- But, you should not use these methods blindly:
  - The future may not be like the past.
  - Associations do not imply causality.

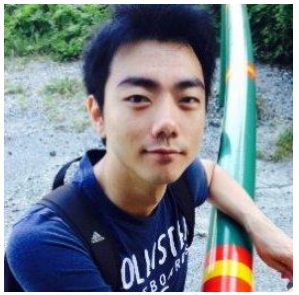
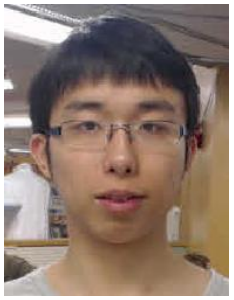
Outline

1) Intro to Machine Learning and Data Mining:
2) Course Administrivia
3) Course Overview

Location, Dates, Webpage

- Course homepage:
  - www.cs.ubc.ca/~schmidtm/Courses/340-F15
- Office hours:
  - Thursdays in ICCS 146 from 3-4.
  - Or by appointment.
- Tutorials:
  - Mondays in DMP 201 from 11-12, 2-3, and 4-5.
  - Only on weeks when assignments are due.
- Teaching Assistants:
  - Issam Laradji.
  - Sharan Vaswani.
  - Tian Qi (Ricky) Chen.
  - Yan Zhao.



CPSC 540 and Auditting 340

- There is also a graduate ML course, CPSC 540:
  - Higher workload.
  - More advanced material.
  - More implementation/theory, fewer applications.
- Auditing CPSC 340 or 540, an excellent option:
  - Pass/fail on transcript rather than grade.
  - Do 2 assignments or write a 2-page report on one technique from or attend > 90% of classes.
  - But please do this officially:
    - http://students.ubc.ca/enrolment/courses/academic-planning/audit

Textbooks

- No required textbook.
- Some books that cover related material:
  - Artificial Intelligence: A Modern Approach (Rusell & Norvig).
  - The Elements of Statistical Learning (Hastie et al.).
  - Machine Learning: A Probabilistic Perspective (Murphy).
  - Pattern Recognition and Machine Learning (Bishop).
  - All of Statistics (Wasserman).
  - Introduction to Data Mining (Tan et al.).
- There is a list of related courses on the webpage.
- You can also use Google.

Assignments

- 6 Assignments worth 25% of final grade:
  - Written portion and Matlab programming.
  - Due at the start of Friday class:
    - September 18 (Friday of next week), October 2, October 16, November 6, November 20, December 4.
  - You can have up to 3 total 'late classes':
    - Handing in an assignment on Monday counts as one.
    - Handing in on Wednesday counts as two.
    - After that, you will get a mark of 0 for late assignments.
    - Examples:
      - you can hand in A1, A4, and A5 one day late.
      - you can hand in A2 two days late and A4 one day late.
      - you can hand in A1 three days late, and all others on time.

Getting Help

- Tutorials on Mondays before assignments due.
- Piazza for assignment/course questions:
  - piazza.com/ubc.ca/winterterm12015/cpsc340
- If you do not have access to Matlab:
  - Ask for a CS guest account.
  - Purchase Matlab through the bookstore or online.
  - Use the free alternative Octave.
- You can work in groups and use any source, but:
  - Hand in your own homework.
  - Acknowledge all sources, including other students.

Midterm and Final

- Midterm details:
  - 30% of final grade
  - In class October 30.
  - Closed book, two-page double-sided ‘cheat sheet’.
- No ‘tricks’ or ‘surprises’:
  - Given a list of things you need to know how to do.
  - Mostly minor variants on assignment questions.
- If you miss the exam, see me with a doctor’s note or other relevant documentation.
- Final will follow same format:
  - 45% of final grade.
  - Cumulative.

Lecture Style and Instructor Evaluation

- I feel that I learn/teach better when using the board.
  - Slows down the lecture.
  - Makes the lecture adaptive.
- This term: hybrid “writing on slides” approach.
- About recording:
  - Do not record without permission.
  - All class material will be available online.
- Topics/readings will be posted before each class.
- October 12, we’ll do an unnofficial instructor evaluation:
  - Will let me adapt lecture/assignment style.

Outline

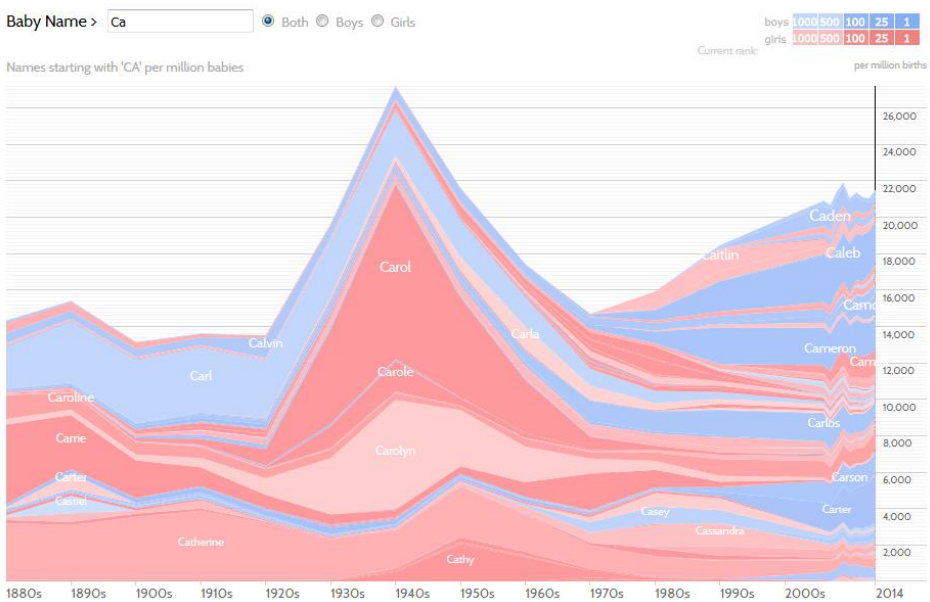
1) Intro to Machine Learning and Data Mining:
2) Course Administrivia
3) Course Overview

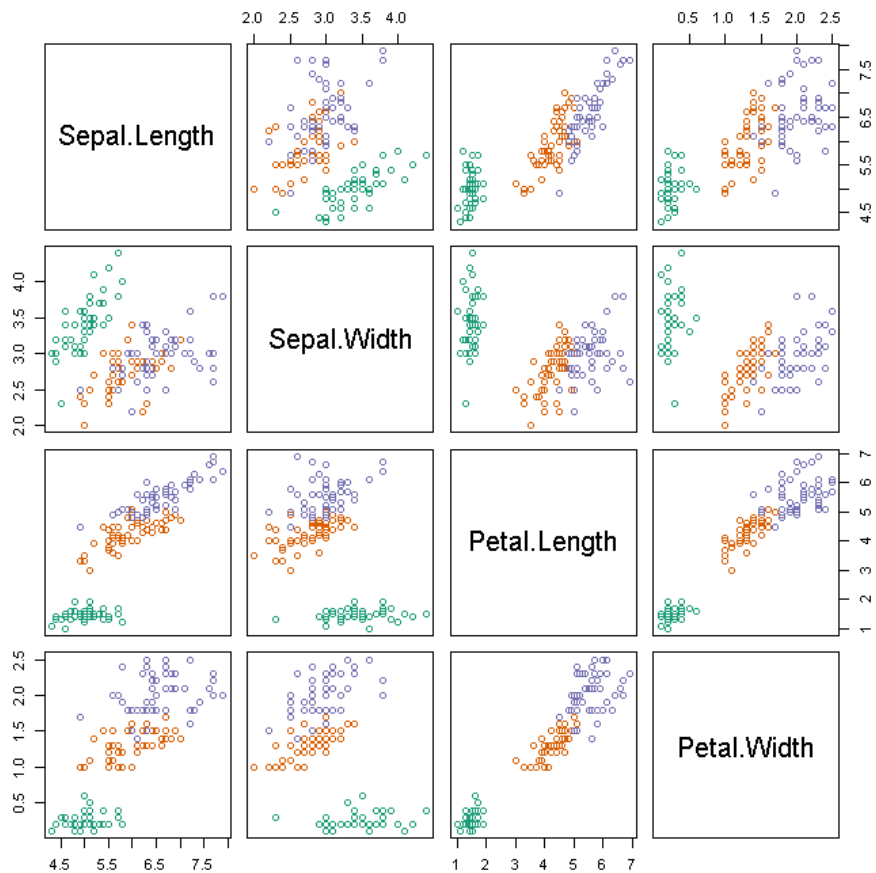
Course Outline

1) Data preprocessing and exploration.
2) Frequency-based supervised learning.
3) Data clustering and association rules.
4) Linear prediction, regularization, and kernels.
5) Outlier detection, dimensionality reduction, and visualization.
6) Neural networks and deep learning.
7) Link analysis and collaborative filtering.
8) Sequences and time-series.

Data Preprocessing and Exploration,

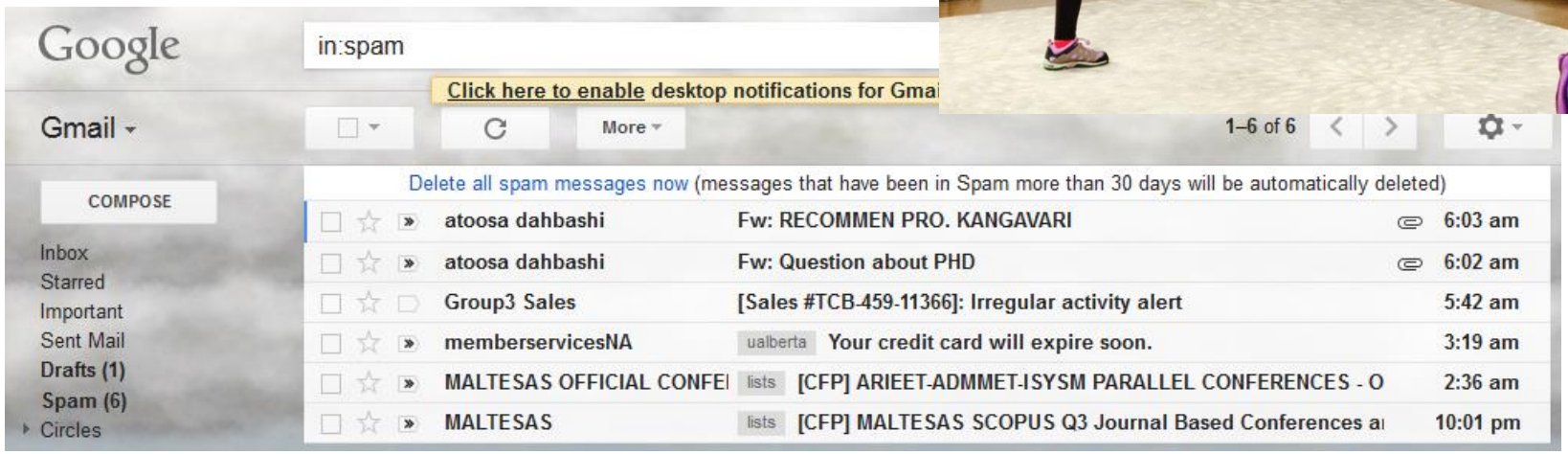
- Types of data and data structures.
- Issues with data quality.
- Summary statistics.
- Data visualization.





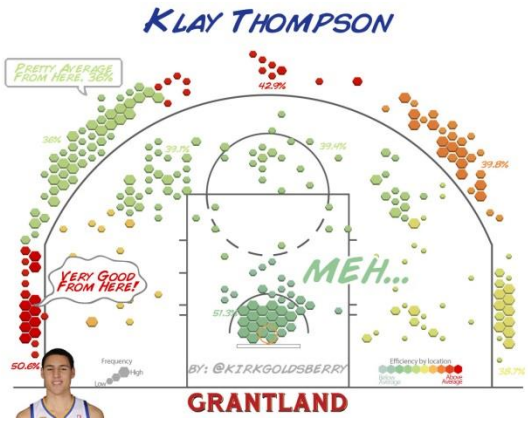
Frequency-Based Supervised Learning

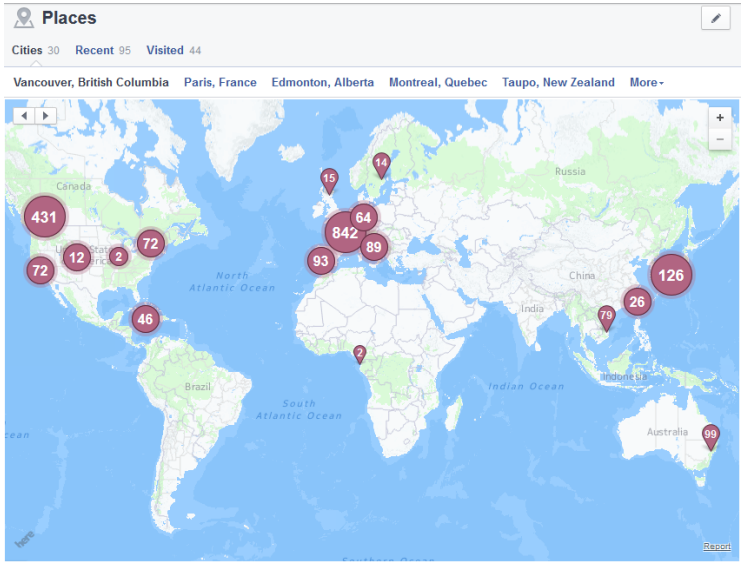
- Classification:
  - Given an object, assign it to predefined ‘classes’.
- Examples:
  - Spam filtering.
  - Body part recognition.



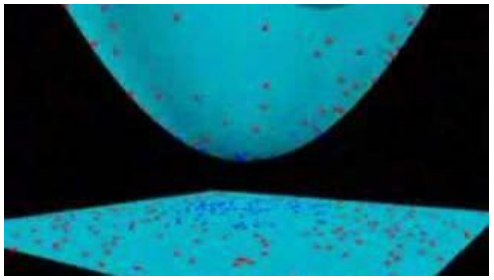
Data Clustering and Association Rules

- Clustering:
  - Find groups of `similar' items in data.
- Examples:
  - Are there subtypes of tumors?
  - Are there high-crime hotspots?
- Association rules:
  - Finding items frequently 'bought together'.

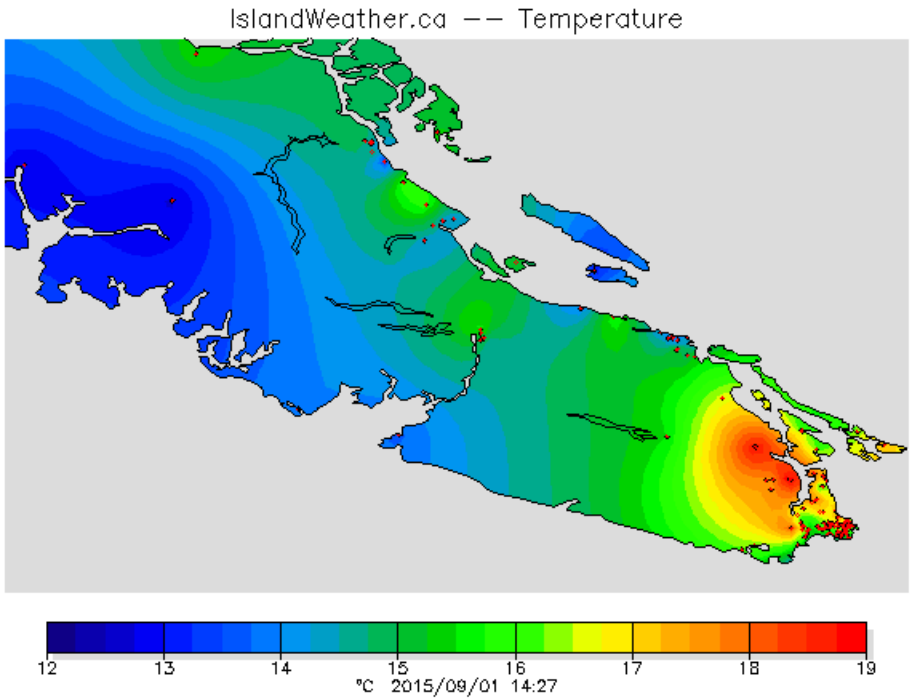


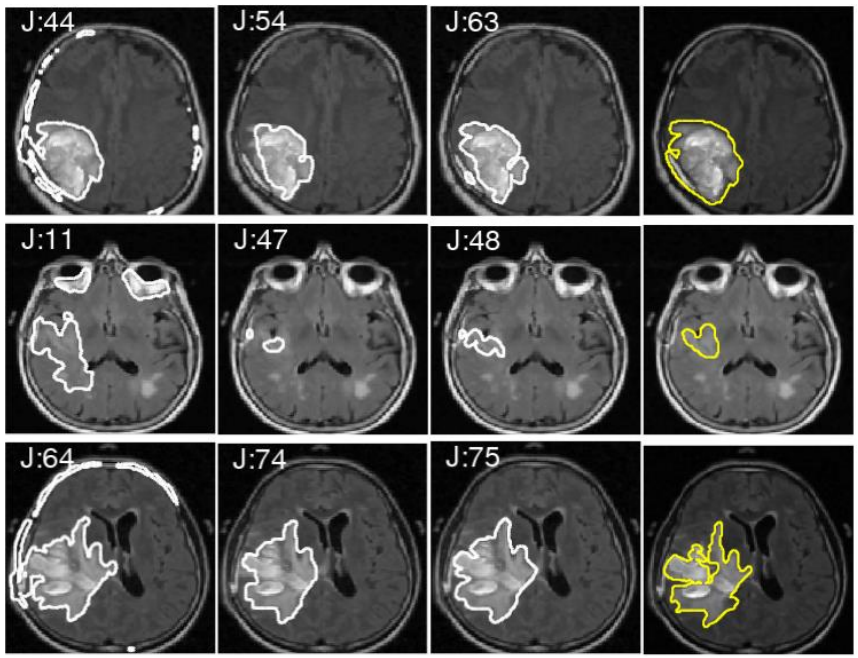


Linear Prediction, Regularization, and Kernels



- Regression:
  - Predicting continuous-valued outputs.
- Working with very high-dimensional data.

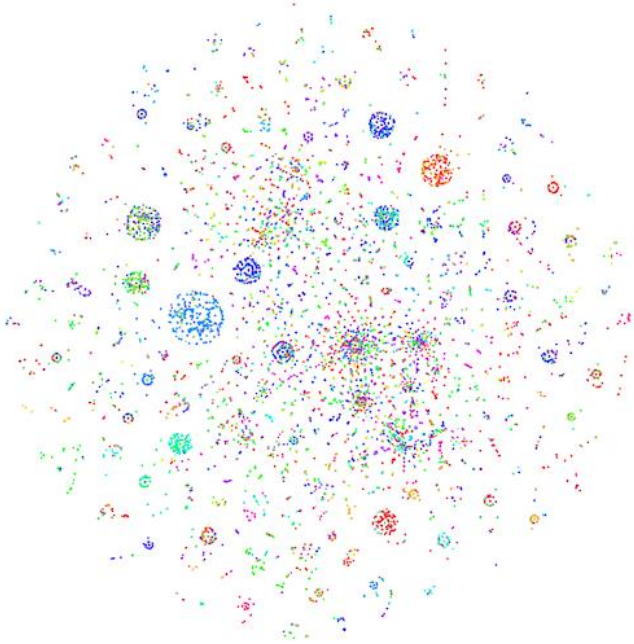


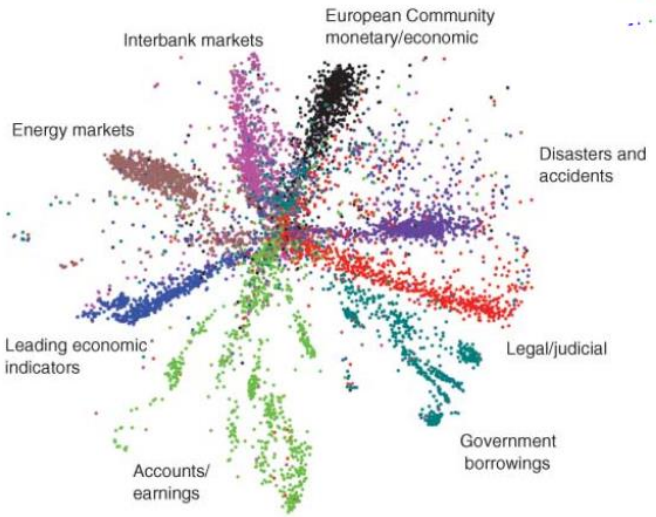


Outlier Detection, Dimensionality Reduction, and Visualization

Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

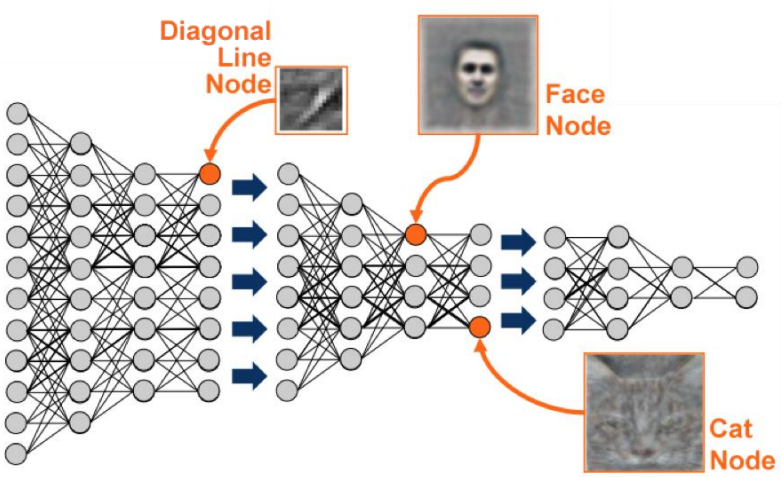
- Outlier detection:
  - Finding data that doesn't belong.
- Dimensionality reduction:
  - Low-dimensional representations .
- Visualization:
  - Displaying high-dimensional relationships.

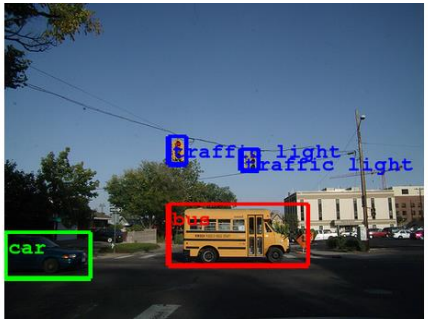


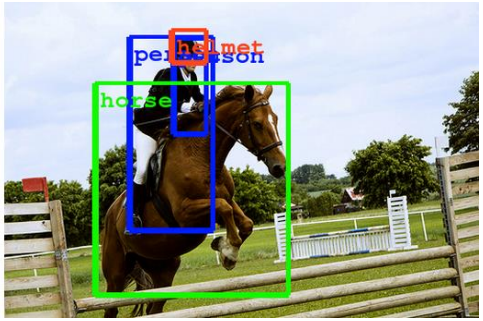


Neural Networks and Deep Learning

- ML when you have a lot of data/computation but don’t know what is relevant.

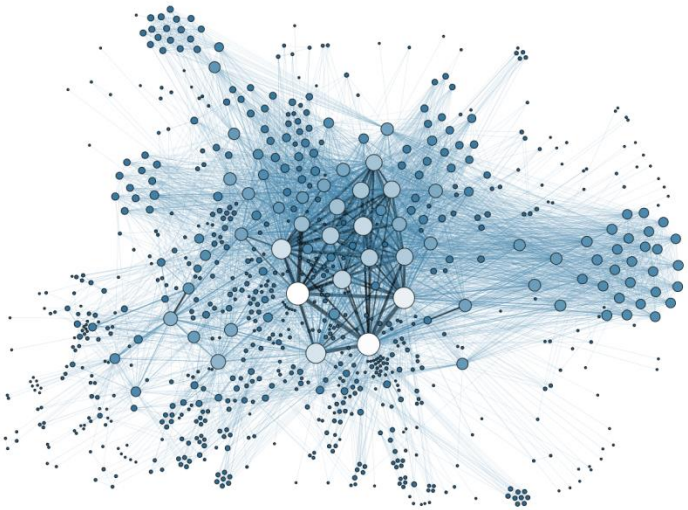


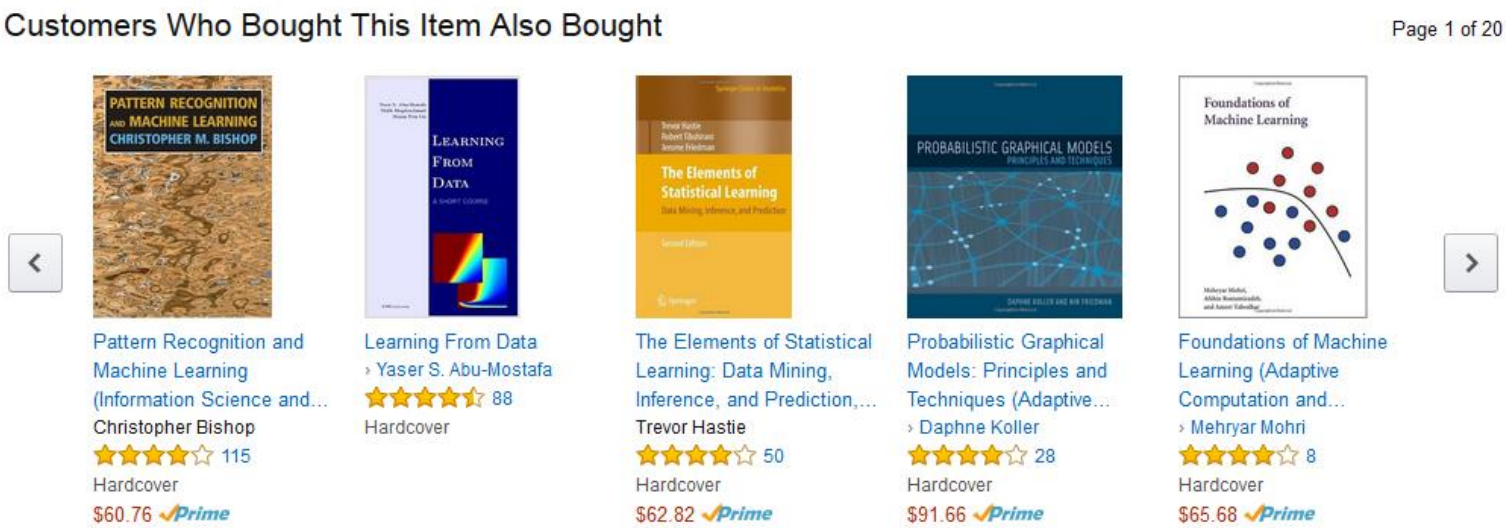




Link Analysis and Collaborative Filtering

- Link analysis:
  - Evaluate relationships between nodes in graph.
- Collaborative filtering:
  - Predict user rating of items.





Sequences and Time Series

- How can we model trends over time?

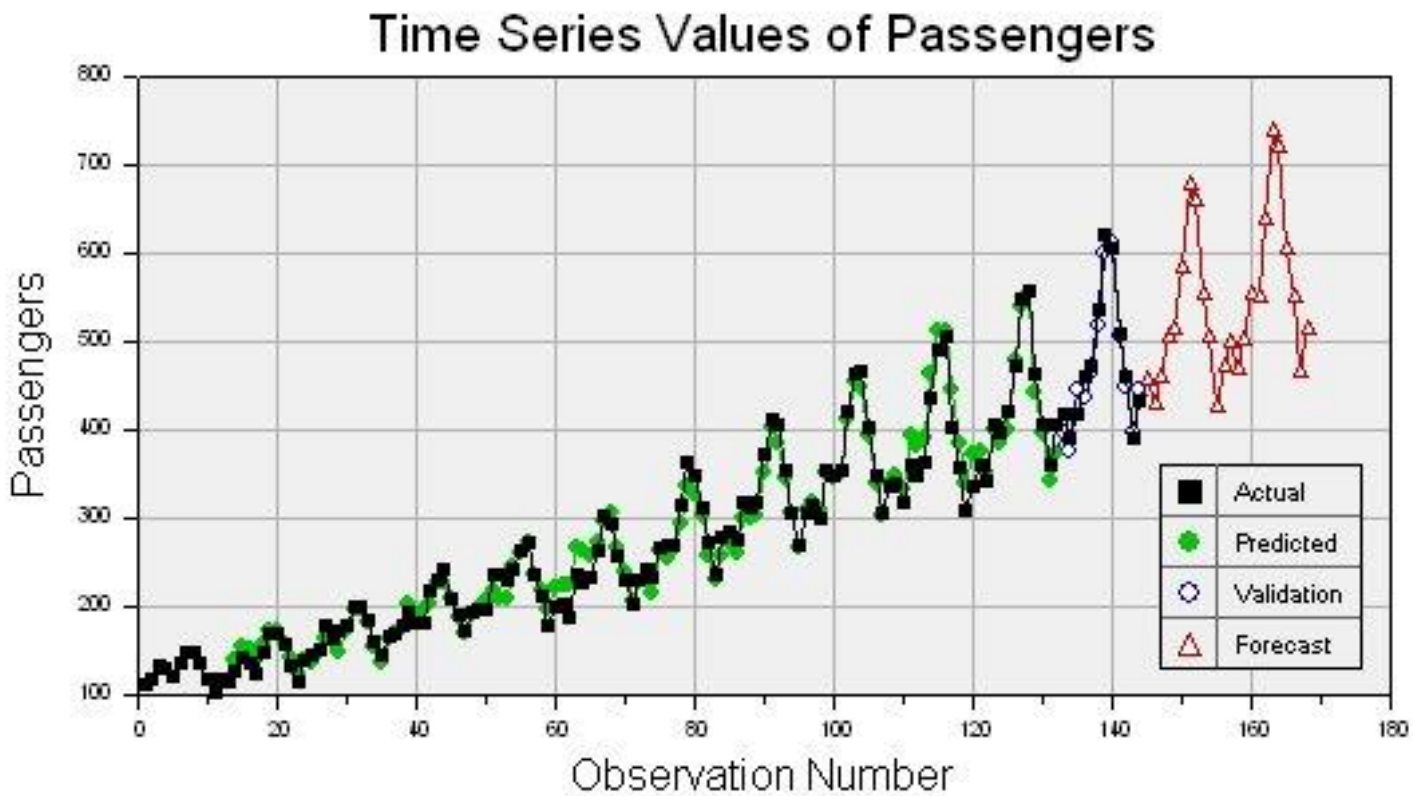


Photo I took in the UK on the way home from the “Optimization and Big Data” workshop: