# CPSC 340:
# Machine Learning and Data Mining

Mark Schmidt
University of British Columbia, Fall 2016
www.cs.ubc.ca/~schmidtm/Courses/340-F16

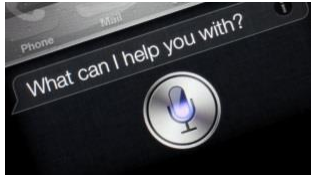
Some images from this lecture are taken from Google Image Search.

# 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.





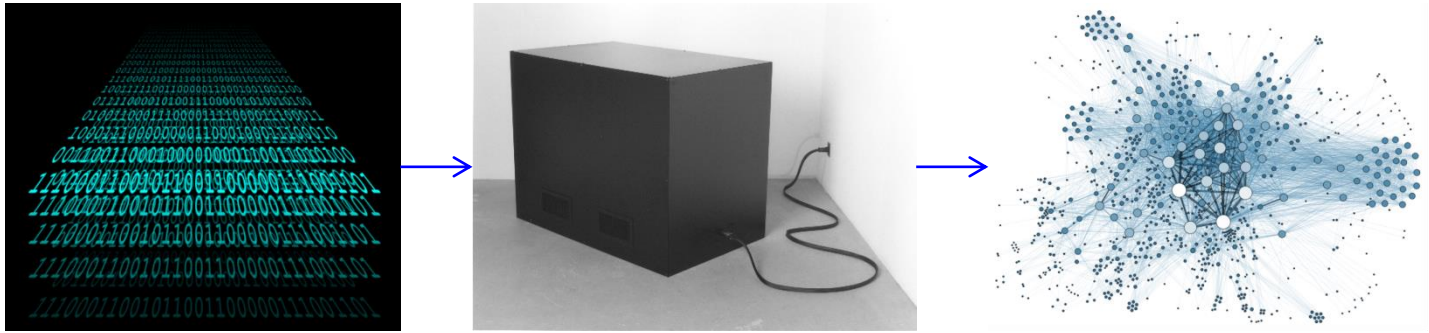


# Big Data Phenomenon

- 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

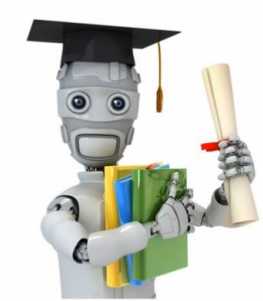
- Automatically extract useful knowledge from large datasets.

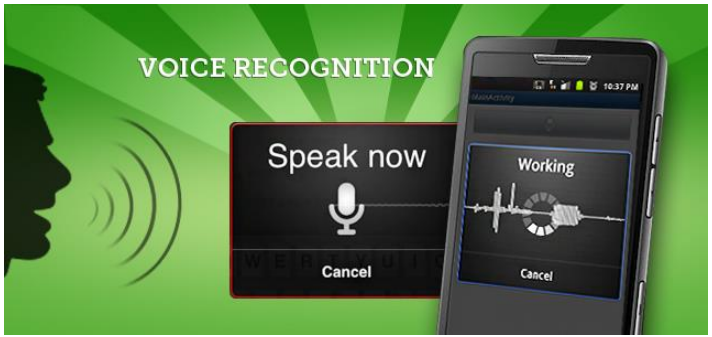


- Usually, to help with human decision making.

# Machine Learning

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

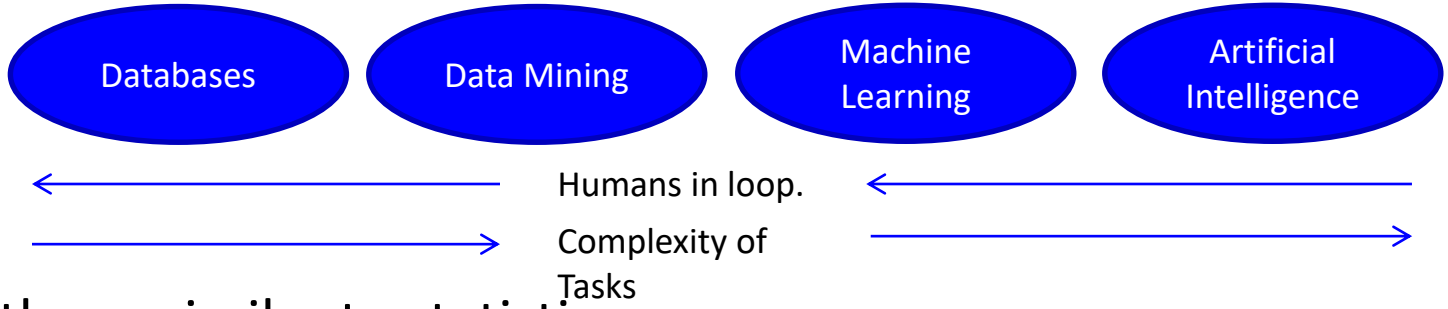





- Most useful when:
  - Don’t have a human expert.
  - Humans can’t explain patterns.
  - Problem is too complicated.

# Data Mining vs. Machine Learning

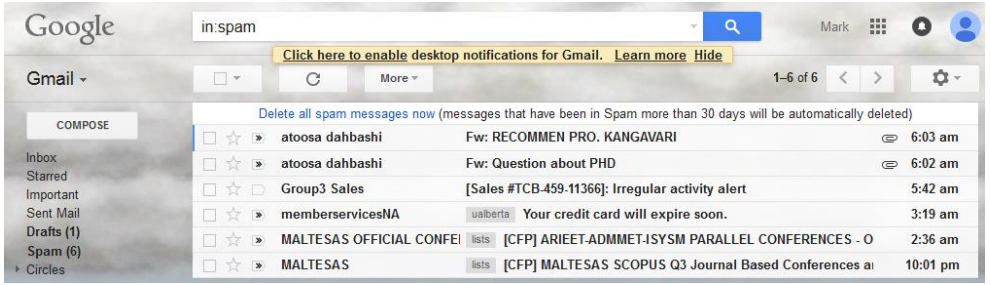
- DM and ML are very similar:
  - Data mining often viewed as closer to databases.
  - Machine learning often viewed as closer AI.



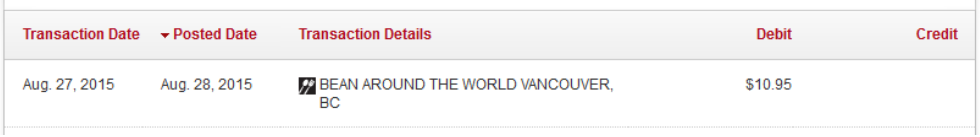

- Both are similar to statistics:
  - Less emphasis on ‘correct’ models and more focus on computation.

# Applications

- Spam filtering:




- Credit card fraud detection:



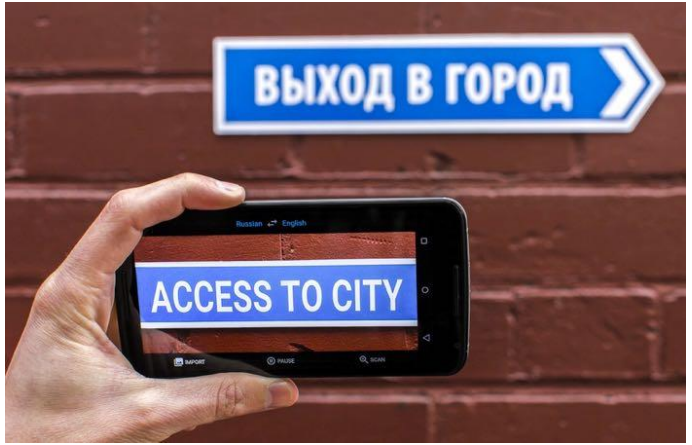
| Transaction Date | Posted Date | Transaction Details | Debit | Credit |
|---|---|---|---|---|
| Aug. 27, 2015 | Aug. 28, 2015 | BEAN AROUND THE WORLD VANCOUVER, BC | $10.95 | |

- Product recommendation:

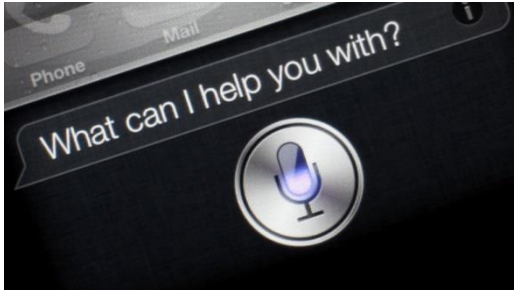
# Applications

- Motion capture:

- Optical character recognition and machine translation:

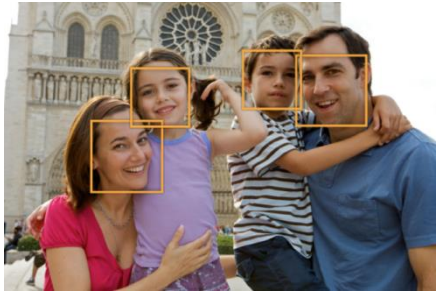



- Speech recognition:

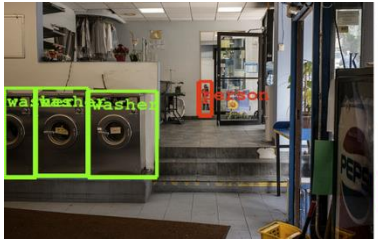


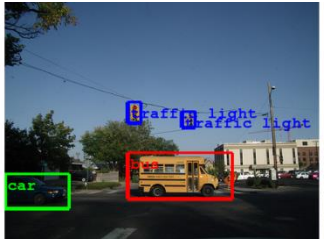
# Applications

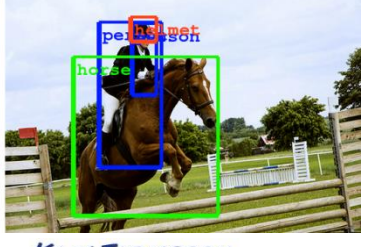
- Face detection:



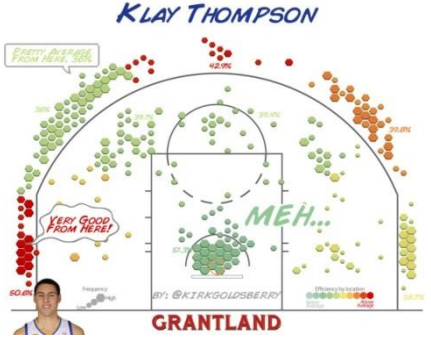
- Object detection:



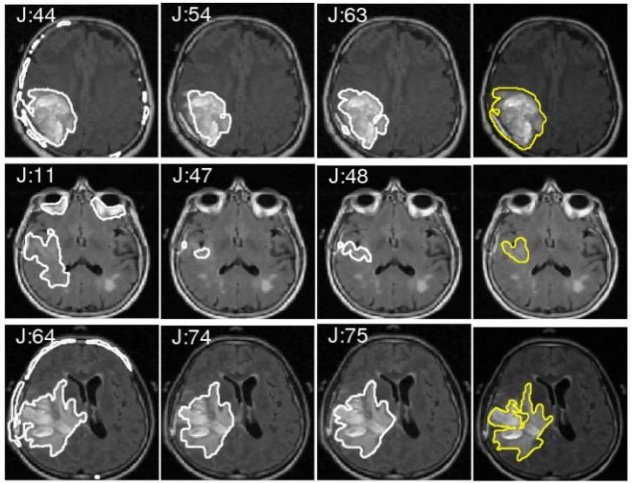

- Sports analytics:



# Applications

- Personal Assistants:

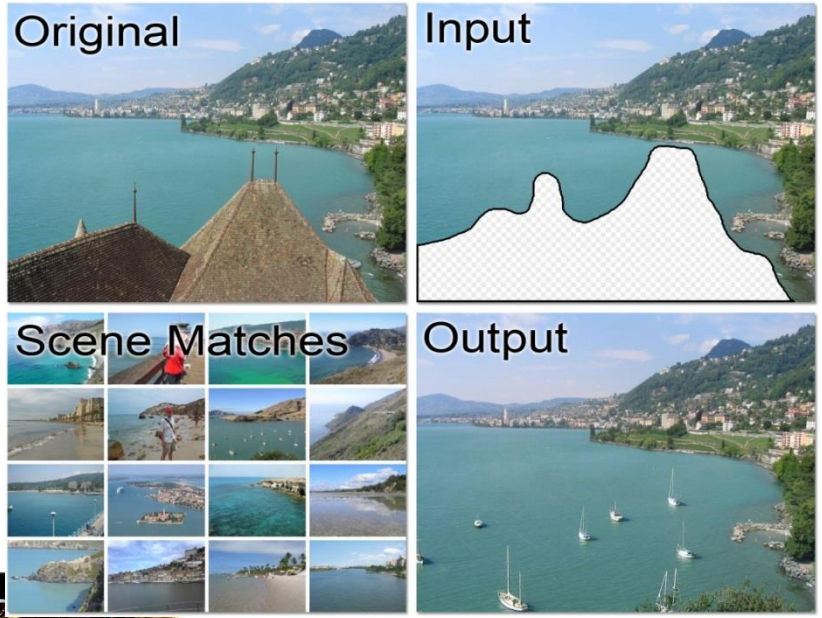

- Medical imaging:




- Self-driving cars:

# Applications

- Scene completion:




- Image annotation:

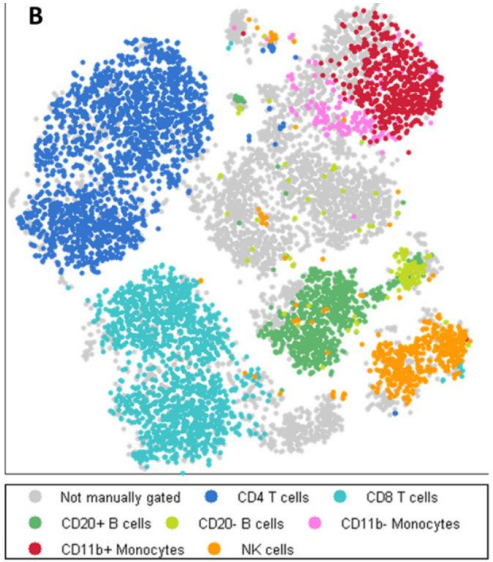

a cat is sitting on a toilet seat
logprob: -7.79


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

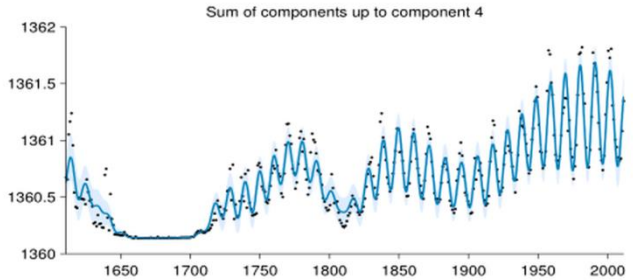
- Discovering new cancer subtypes:




- Automated Statistician:

**2.4 Component 4 : An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards**

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.



# Applications

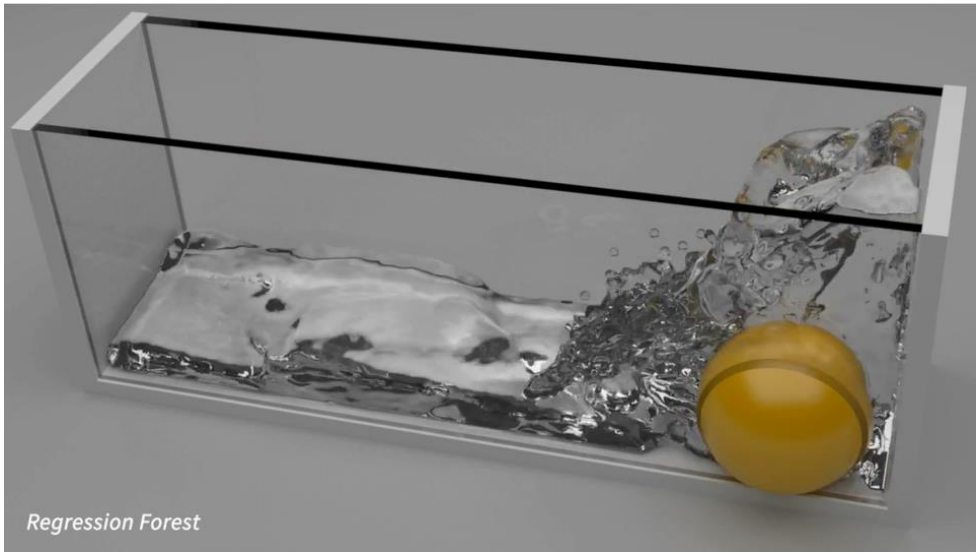
- Mimicking artistic styles and inceptionism:


# Applications

- “Deep dream”:

# Applications

- Fast physics-based animation:




- Mimicking art style in video.

# Applications

- Beating human Go masters:

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

- But, this is not magic and we can only solve certain problems:
  - “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.”
- Also, 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

# CPSC 340 vs. CPSC 540

- There is also a graduate ML course, CPSC 540:
  - More advanced material.
  - More focus on theory/implementation, less focus on applications.
  - More prerequisites and higher workload.
- For almost all students, CPSC 340 is the right class to take:
  - CPSC 340 focuses on the most widely-used methods in practice.
  - CPSC 540 is intended as a continuation of CPSC 340.
    - You'll miss important topics if you skip CPSC 340.

# Reasons NOT to take this class

- For many people, this course is a LOT of work.
  - Some people spend tens of hours per assignment.
- Compared to typical CS classes, there is a lot more math:
  - Requires linear algebra, probability, and multivariate calculus.
  - Course is harder this year because of new calculus requirement.
- Compared to non-CS classes, there is a lot more CS:
  - This is not a class about running other people's software packages.
  - You are going to make/modify implementations of methods.
- Instructor: this is only my second undergrad course.
- Matlab: next semester might use Python instead.

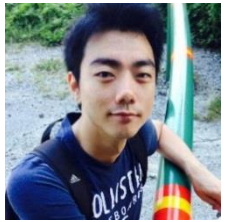
# Webpage, Piazza, Office Hours, Tutorials

- Course homepage:
  - www.cs.ubc.ca/~schmidtm/Courses/340-F16
- Piazza for assignment/course questions:
  - www.piazza.com/ubc.ca/winterterm12016/cpsc340/home
  - Office hours:
    - Tuesday at 2-3 (ICICS 146) and 3:30-4:30 (DLC Table 4), Wednesdays 4-5 (ICICS X337)
    - Or by appointment.
- Optional weekly tutorials:
  - Start in second week of class (September 12).
  - Mondays 4-5 and 5-6, Tuesdays 4:30-5:30, and Wednesdays 9-10.
  - Cover mix of tutorial material and exercises to help with assignments.
  - You must be registered in a tutorial section to stay enrolled.

# The Teaching Assistants (are outstanding)

- Reza Babanezhad

- Tian Qi (Ricky) Chen



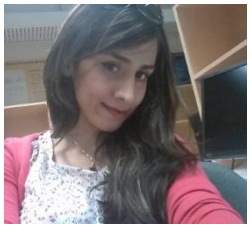
- Issam Laradji

- Robbie Rolin

- Alireza Shafaei

- Moumita Roy Tora

- Nasim Zolaktaf



- Zainab Zolaktaf

# Waiting List and Auditing

- The SSC currently lists this class as full at 160 students.
- But the room supports 188 students (possibly more)
- We're going to start registering people from the waiting list.
  - Being on the waiting list is the only way to get registered:
    - https://www.cs.ubc.ca/students/undergrad/courses/waitlists
  - You might be registered without being notified, be sure to check!

- Because the room is full, we may not have seats for auditors.
  - If there is space, I'll describe (light) auditing requirements then.

# Textbooks

- No required textbook.
- I'll post relevant sections out of these books as optional readings:
  - Artificial Intelligence: A Modern Approach (Rusell & Norvig).
  - Introduction to Data Mining (Tan et al.).
  - The Elements of Statistical Learning (Hastie et al.).
  - Machine Learning: A Probabilistic Perspective (Murphy).
- List of related courses on the webpage, or you can use Google.

# Assignments

- 6 Assignments worth 25% of final grade:
  - Written portion and Matlab programming.
  - Submitted as a PDF file using the Handin program.
  - You can have up to 4 total “late classes”:
    - For example, if assignment is due on Wednesday:
      - Handing it in before Wednesday class is 0 late classes.
      - Handing it in before Friday class is 1 late classes.
      - Handing it in before Monday class is 2 late classes.
      - Handing it in before Wednesday class is 3 late classes.
  - You will get a mark of 0 on an assignment if you:
    - Use more than 3 late classes on the assignment.
    - Exceed 4 late classes across all assignments.

# Getting Help

- There are many sources of help on the assignments:
  - Weekly tutorials, office hours, Piazza, other students.
- 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.
    - Let me know about any Octave incompatibilities in the assignments.
    - Julia might work, too.
- You can work in groups and use any source, but:
  - Hand in your own homework.
  - Acknowledge all sources, including webpages and other students.

# Midterm and Final

- Midterm details:
  - 30% of final grade
  - In class, October 28.
  - 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 doctor’s note or relevant documentation.
- Final will follow same format:
  - 45% of final grade.
  - Cumulative.

# Lecture Style and Lecture Slides

- The course we will cover a lot of topics:
  - Some topics will not be covered in much depth.
    - But we'll go into depth on a few key recurring issues.
    - To keep things sane, I'll give you a list of topics to know for the midterm/final.
  - It can be better to know many methods than learning only a few in detail:
    - I'll explain why when we discuss the "best" machine learning algorithm.
    - Some class time will be devoted to important ideas that you won't be tested on.
- All class material will be available online or on Piazza.
  - I'll try to post topics/readings before each class.
  - After class, I'll post annotated/updated slides.
  - Do not record without permission.
- In early October, we'll do an unnofficial instructor evaluation:
  - Will let me adapt lecture/assignment/tutorial style.

# Outline

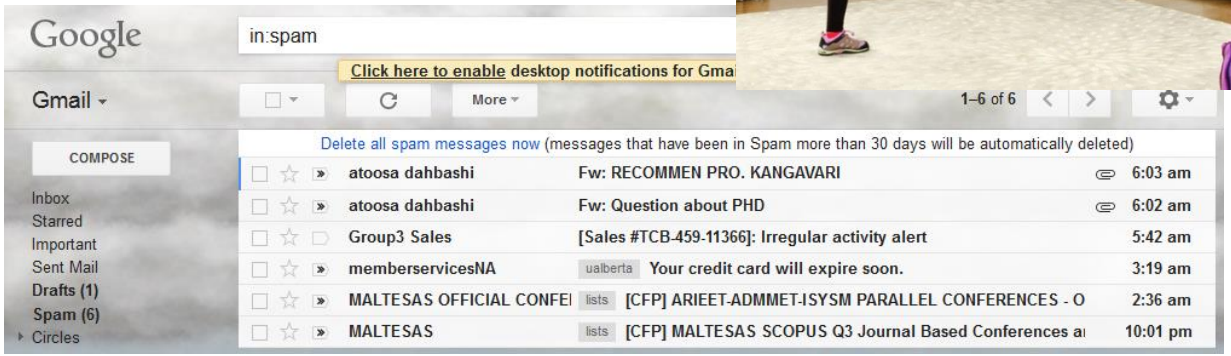
1) Intro to Machine Learning and Data Mining:
2) Course Administrivia
3) Course Overview

# Course Outline

- Next class discusses data exploration, cleaning, and preprocessing.
- After that, the remaining lectures focus on the six topics:
  1) Supervised Learning.
  2) Unsupervised learning.
  3) Linear prediction.
  4) Latent-factor models.
  5) Deep learning.
  6) Density estimation.

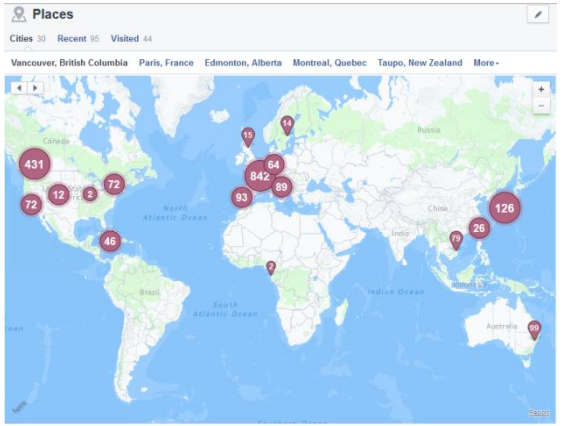
# Supervised Learning

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



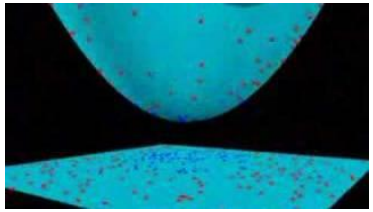
# Unsupervised Learning

- Clustering:
  - Find groups of `similar' items in data.
- Examples:
  - Are there subtypes of tumors?
  - Are there high-crime hotspots?
- Outlier detection:
  - Finding data that doesn't belong.
- Association rules:
  - Finding items frequently 'bought together'.

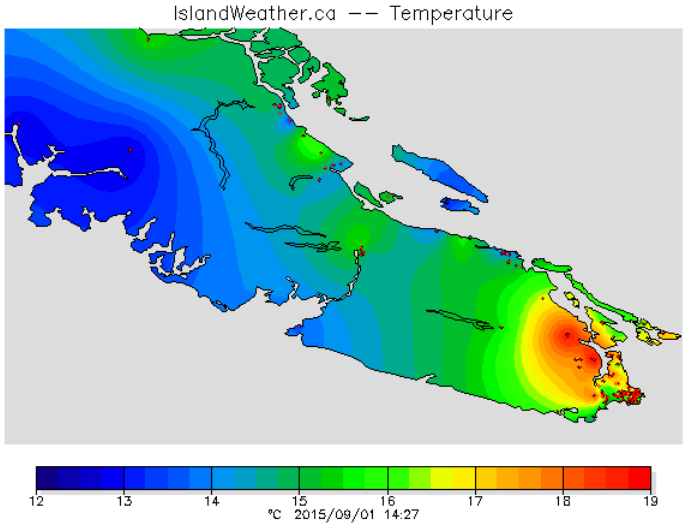


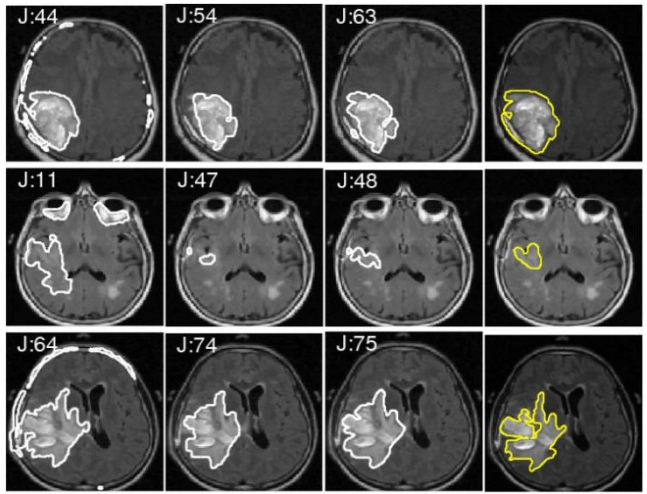

| Transaction Date | Posted Date | Transaction Details | Debit | Credit |
|---|---|---|---|---|
| Aug. 27, 2015 | Aug. 28, 2015 | BEAN AROUND THE WORLD VANCOUVER, BC | $10.95 | |

# Linear Prediction

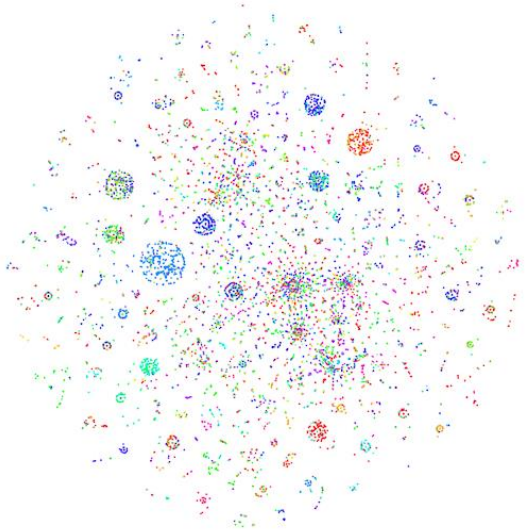


- Regression:
  - Predicting continuous-valued outputs.
- Working with very high-dimensional data.

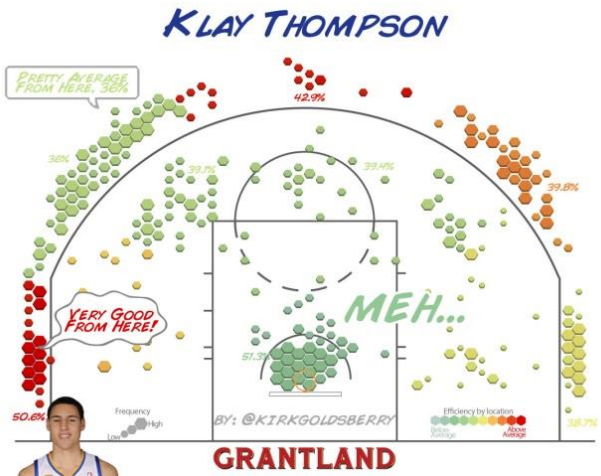


# Latent-Factor Models

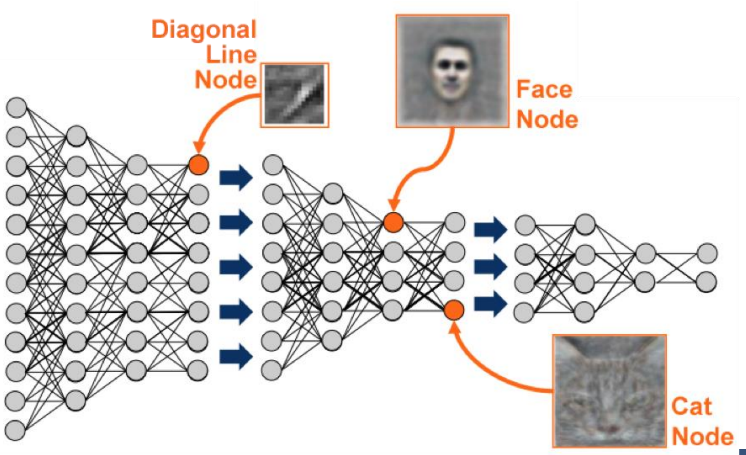


- Principal component analysis and friends:
  - Low-dimensional representations.
  - Decomposing objects into “parts”.
  - Visualizing high-dimensional data.
- Collaborative filtering:
  - Predicting user ratings of items.

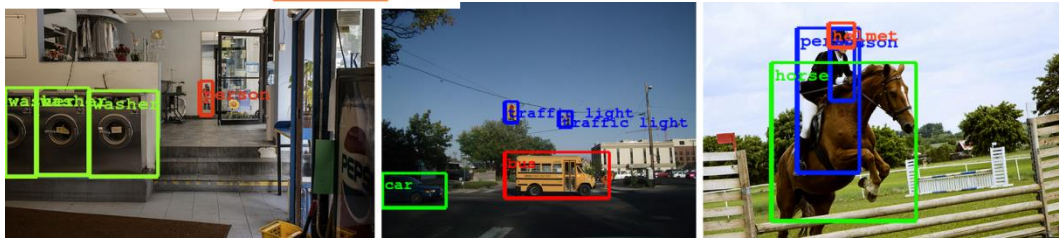



# Deep Learning

- Neural networks: Brain-inspired ML when you have a lot of data/computation but don't know what is relevant.

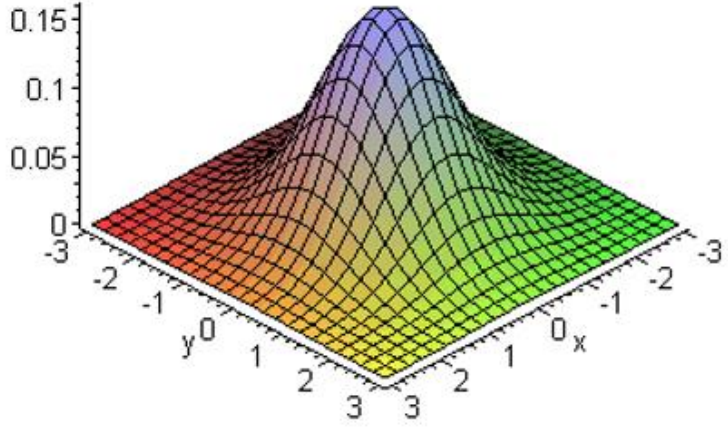






# Density Estimation

- Density estimation:
  - Modeling the probability of a complex event happening.
  - Modeling dependencies over time.



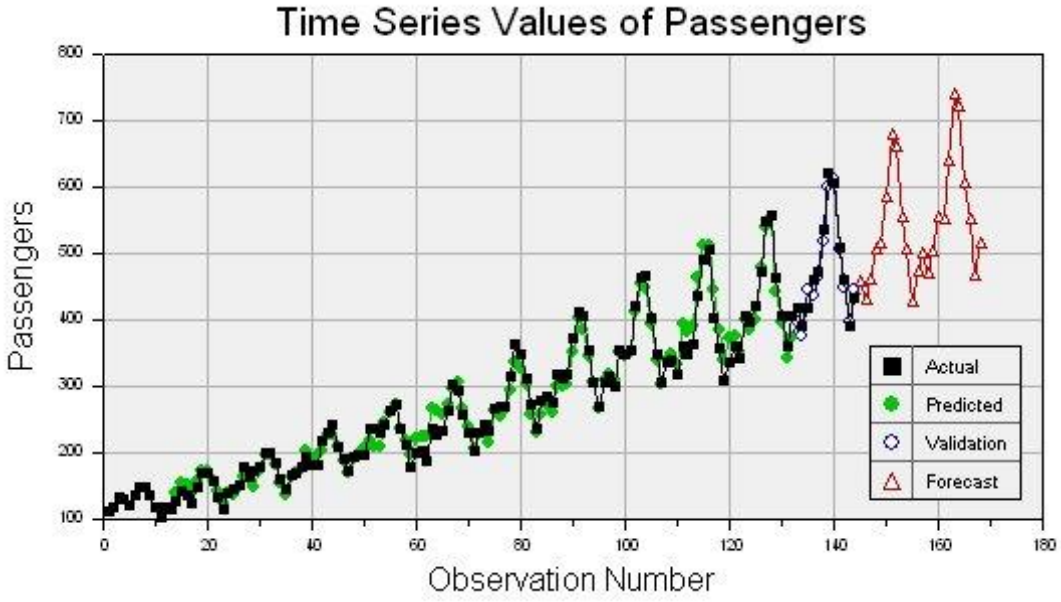



Photo I took in the UK on the way home from the “Optimization and Big Data” workshop: