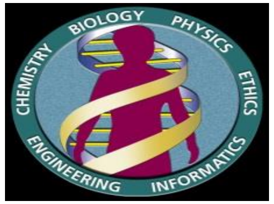
CPSC 540: Machine Learning

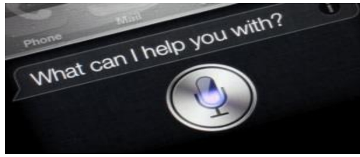
Mark Schmidt
University of British Columbia, Winter 2018
www.cs.ubc.ca/~schmidtm/Courses/540-W18

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.



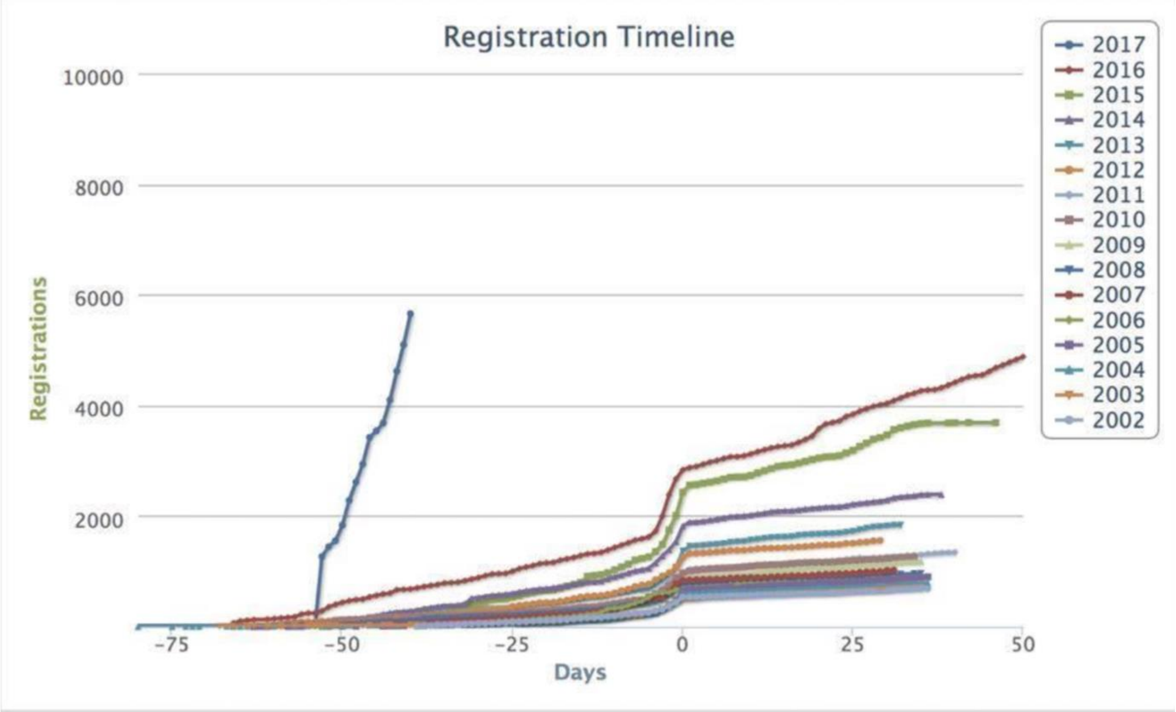


Machine Learning

- What do you do with all this data?
  - Too much data to search through it manually.
- But there is valuable information in the data.
  - Can we use it for fun, profit, and/or the greater good?
- Machine learning: use computers to automatically detect patterns in data and make predictions or decisions.
- Most useful when:
  - Don't have a human expert.
  - Humans can't explain patterns.
  - Problem is too complicated.

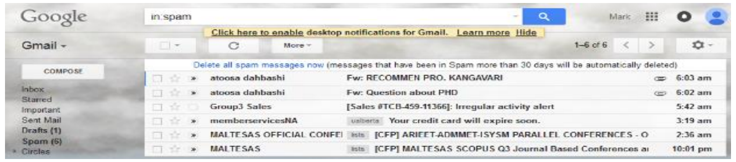
Machine Learning vs. Statistics

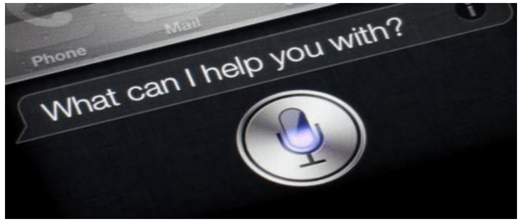
- Machine learning (ML) is very similar to statistics.
  - A lot of topics overlap.
- But ML places more emphasis on:
  1. Computation and large datasets.
  2. Predictions rather than descriptions.
  3. Non-asymptotic performance.
  4. Models that work across domains.
- The field is growing very fast:
  - Influence of $$$ too.

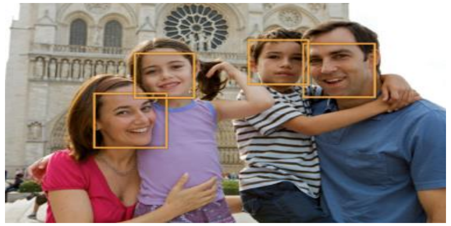


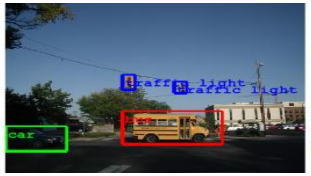
Applications

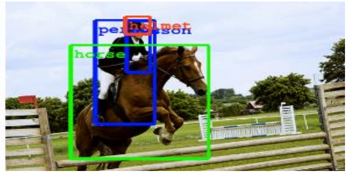
- Spam filtering.
- Credit card fraud detection.
- Product recommendation.
- Motion capture.
- Machine translation.
- Speech recognition.
- Face detection.
- Object detection.
- Sports analytics.
- Cancer subtype discovery.





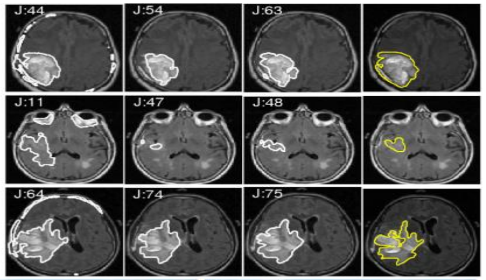


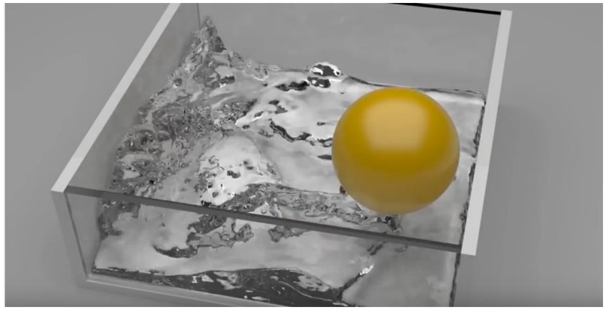


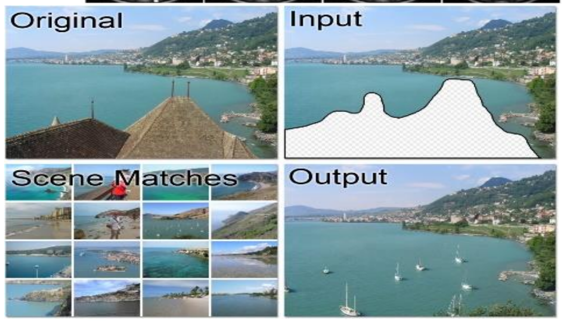


Applications

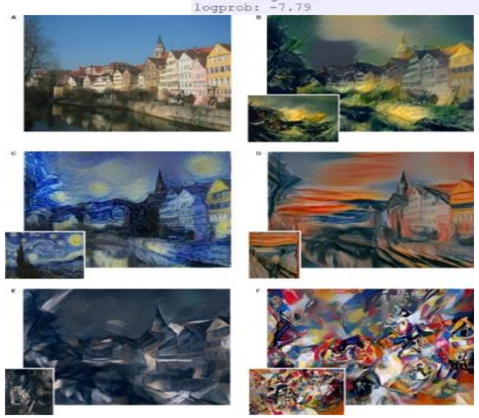
- Gene localization/functions/editing.
- Personal Assistants.
- Medical imaging.
- Self-driving cars.
- Scene completion.
- Image search and annotation.
- Artistic rendering.
- Physical simulations.
- Image colourization.
- Game-playing.







Youngsters, May 1912



(pause)

CPSC 340 and CPSC 540

- There are two ML classes: CPSC 340 and 540.
  - They are structured as one full-year course: 540 starts where 340 ends.
- CPSC 340:
  - Introductory course on data mining and ML.
  - Emphasis on applications and core ideas of ML.
  - Covers implementation of methods based on counting and gradient descent.
  - Most useful techniques that you can apply to your research/work.
- CPSC 540:
  - Research-level ML methods and theory.
  - Assumes strong background on fundamental ML concepts.
  - Assumes stronger math/CS background
  - Much more work.

CPSC 340 and CPSC 540

- Since 540 starts where CPSC 340 ends, 540 is not an introductory ML course.
- I'm not covering any of the below, and will assume you already know these concepts:
  - Calculus in matrix notation.
  - Cross-validation.
  - Probabilistic classifiers.
  - Ensemble methods.
  - Radial basis functions.
  - Kernel trick.
  - Stochastic gradient.
  - Maximum likelihood estimation.
  - MAP estimation.
  - L1-regularization.
  - Softmax loss.
  - PCA.
  - Non-negative matrix factorization.
  - Collaborative filtering
  - Deep learning.
  - Convolutional neural networks.
- If you don't know **how to implement all the above** , **you will get lost very quickly if you don't know this material**.

CPSC 340 and CPSC 540

- If you can only take one class, take CPSC 340:
  - 340 covers the most useful methods and ideas.

- If want to work in ML you should **take both courses**:
  - There is not a lot of overlap between the topics, 540 is missing a lot important topics.
  - 540 is NOT an "advanced" version of 340.
    - It just covers the methods that require more advanced math/CS background.

- It is much better to do CPSC 340 first:
  - Many people have taken CPSC 340 *after* CPSC 540 (not recommended).
  - A few people took 540 then 340 then *540 again* (REALLY not recommended).

- There will be less overlap between 340 and 540 this year:
  - 340 has required multivariable calculus as a prereq since 2016.
    - It is more advanced than it was in 2015, and much more advanced than it was before 2015.
  - I'm not covering the "diff" between 340-2015 and 340-2016 this year.

CPSC 340 and CPSC 540

- Quotes from people who probably should have taken CPSC 340:
  - "I did Coursera and have have done well in Kaggle competitions."
    - Neither of these cover calculus in matrix notation or MLE and MAP estimation.
  - "I've used SVMs, PCA, and L1-regularization in my work."
    - Sure, but do you know how to implement them from scratch?
  - "I've seen most of the 340 topics before."
    - Sure, but at what level of detail and do you know how to implement them from scratch?
  - "I want to apply machine learning in my research."
    - Great! Take 340 to learn the most useful tools and also **what can go wrong**.
  - "I took a machine learning course at my old school."
    - 340 is more broad/advanced than at most schools (talk to me if unsure).

Math Prerequisites

- Research-level ML involves a lot of math.
- You should be comfortable with:
  - Linear algebra, probability, multivariate calculus, mathematical proofs.
  - Suggested minimum requirements: Math 200, 220, 221, and 302.
- You should be able to do proofs based on:
  - Sequences of random gradient vectors.
  - Eigenvalues of second-derivative matrices.

Computer Science Prerequisites

- ML places a big emphasis on computation.
- You should be comfortable with:
  - Software engineering: reading/writing/debugging complex programs.
  - Data structures: pointers, trees, heaps, hashes, graphs.
  - Scientific computing: matrix factorization, gradient descent, condition number.
  - Algorithms and complexity:
    - Big-O, divide + conquer, randomized algorithms, dynamic programming, NP-completeness.
  - Suggested minimum requirements: CPSC 210, 221, 302, and 320:

- “I have programming experience in my work/research/courses”.
  - Great, for most people this is a poor replacement for knowing the fundamentals.
- "The early advice that you gave me to take CPSC 320 really helped me."

Prerequisite Form

- All students must submit the prerequisite form.
  - CS/ECEC/STAT grad students: submit in class/tutorial by January 10.
  - All others: submit to enroll in course.
    - I'll sign enrollment forms between lectures once I have this form.

CPSC 540: Machine Learning:
Prerequisite Form

Machine learning is a very popular topic, and it is increasingly being used in a huge variety of applications. However, the material is also very challenging because it brings together a larger number of ideas from computer science, mathematics, and statistics. Unfortunately, due to the popularity of the topic we typically have a few students register for the course who do not yet have the appropriate background. These students not only hurt themselves because they struggle with the high workload in the course, but they also hurt the experience of the other students since significant class time ends up being spent on material that should be specified as prerequisites.

While it is hard to add formal prerequisites to graduate courses because people come from such different backgrounds, we need to establish that everyone in the class has a common background. Below I give a list of courses (and important related topics) that I would ideally like a CPSC 540 student to take either before or simultaneously with CPSC 540:

- A linear algebra course like Math 221 (linear systems, eigenvalues).
- A probability course like Math 302 (conditional probability, expectations).
- A multivariate calculus course like Math 200 (gradients, optima).
- A scientific computing course like CPSC 302 (numerical solution of linear systems, condition number).
- An algorithms and complexity course like CPSC 320 (big-O notation, NP-hard problems).
- A statistical inference course like STAT 305 (linear regression, maximum likelihood estimation).

Reasons Not to Take This Course

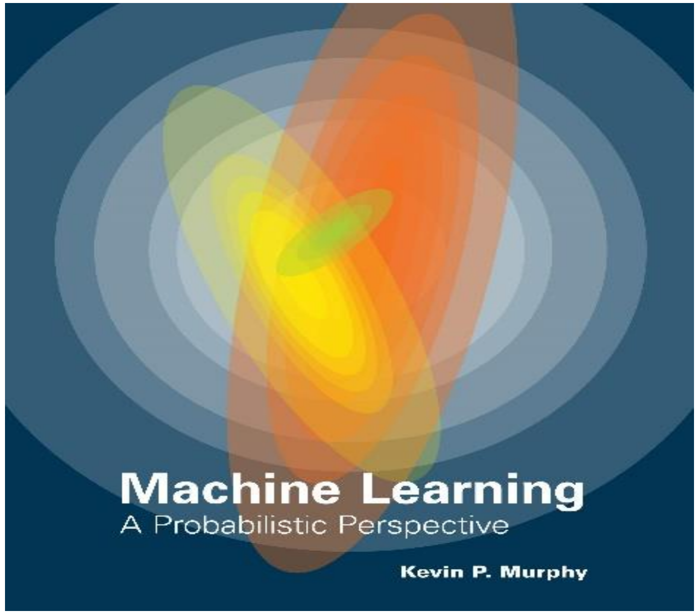
- High workload:
  - “This course's workload was a bit more than I would have liked. It seems like this course takes twice the amount of time as another course.”
- Haven’t taken CPSC 340:
  - You’ll be missing half of the story, you won’t know many of the most important methods, and a lot of stuff will seem random.
- Missing prerequisites (or low grades in prereq courses):
  - It’s better to improve your MATH/CSPC backgroud, and take the course later.
    - Many topics will make a lot more sense as you won’t be filling in background.
  - “I know too much math” said nobody ever.
  - “I’m too good at computer science”, see above (and think $$$ if necessary).

Auditing and Recording

- Auditing 540, an excellent option:
  - Pass/fail on transcript rather than grade.
  - Do 1 assignment or write a 2-page report on one technique from class or attend > 90% of classes.
  - But please do this officially:
    - http://students.ubc.ca/enrolment/courses/academic-planning/audit
- About recording lectures:
  - Do not record without permission.
  - All class material will be available online.
  - Videos of material from first month of a previous section are here:
    - https://www.youtube.com/watch?v=p4EnVHSml4U

Textbook and Other Optional Reading

- No textbook covers all course topics.
- The closest is Kevin Murphy's "Machine Learning".
  - But we're using a very different order.



- For each lecture:
  - I'll give relevant sections from this book.
  - I'll give other related online material.
- There is a list of related courses on the webpage.

Textbook and Other Optional Reading

- Other good machine learning textbooks:
  - All of Statistics (Wasserman).
  - Elements of Statistical Learning (Hastie et al.).
  - Pattern Recognition and Machine Learning (Bishop).

- Good textbooks on specialized topics from this course:
  - Convex Optimization (Boyd and Vandenberghe).
  - Probabilistic Graphical Models (Koller and Friedman).
  - Deep Learning (Goodfellow et al.).
  - Bayesian Data Analysis (Gelman).

- Some of these are on reserve at the ICICS reading room.

Grading

- 40%: 5 assignments (written, math, and Julia programming).
- 30%: Final (date will be placed here when known).
- 30%: Course project (due date will be placed here when known).
  - There will be no post-course grade changes based on grade thresholds:
    - 49% will not be rounded to 50%, and 71% will not be rounded to 72%.

- No, you can’t do the assignments in Python, R, Matlab, and so on.
  - Julia is free and way faster than Python/R/Matlab.
  - Assignments have prepared code that we won’t translate to 3 languages.
  - TAs shouldn’t have to know 3 languages to grade
- For the course project, you can use any language.

Assignments

- Due at midnight on days where we have lectures:
  - First assignment due next Wednesday (January 10).
    - Subsequent assignments due every 3 weeks.
- Start early, the assignments are a lot of work:
  - Previous students estimated that each assignments takes 6-25 hours:
    - The was heavily correlated with satisfying prerequistes.
    - Please look through the assignment in previous offerings to see length/difficulty.
- Assignment 1 should be done on your own.
- Assignments 2-5 can be done in groups of 1 to 3.
  - Hand in one assignment for the group.
  - But each member should still know the material.

Late Assignment Policy

- You have up to 4 total "late classes".
- Example:
  - Assignment 1 is due Wednesday January 10.
  - You can use 1 late class to hand it in Friday January 12.
  - You can use 2 late classes to hand it in Monday January 15.
  - If you need multiple late days for Assignment 1, consider dropping the course.

- FAQ:
  - You cannot use more than 2 "late classes" on any one assignment (0 after that).
  - You cannot use more than 4 total "late classes" throughout the term (0 after that).
    - Otherwise, there is no penalty for using "late classes".
  - You can use late classes on the assignments/project, but not the exam.
  - Number of late classes for a group:
    - If group member 'i' has c_i late classes, group can use at most ceil(mean(c_i)).

Assignment Issues

- No extensions will be considered beyond the late days.
  - Also, since you can submit more than once, so you have no excuse not to submit something preliminary by the deadline.

- Further, due to limited TA hours, these issues are a 50% penalty:
  - Missing names or student IDs on assignments.
  - Corrupted .zip submission files or not using a .zip file.
  - Submitting the wrong assignment (year or number).
  - Incorrect assignment names in submission files.
  - Not including answers in the correct location in the .pdf file.

Cheating and Plagiarism

- Read about UBC’s policy on “academic misconduct” (cheating):
  - http://www.calendar.ubc.ca/Vancouver/index.cfm?tree=3,54,111,959

- When submitting assignments, acknowledge all sources:
  - Put “I had help from Sally on this question” on your submission.
  - Put “I got this from another course’s answer key” on your submission.
  - Put “I copied this from the Coursera website” on your submission.
  - Otherwise, this is plagiarism (course material/textbooks are ok with me).

- At Canadian schools, this is taken very seriously.
  - Could receive 0 in course, be expelled from UBC, or have degree revoked.

Getting Help

- Piazza for assignment/course questions:
  - https://piazza.com/ubc.ca/winterterm22017/cpsc540
- Instructor office-hours:
  - Tuesdays 3:00-4:00 (ICICS 146) or by appointment (starting next week).
- TA office hours: TBA.
- Weekly tutorials:
  - Run by TAs covering related material.
  - Mondays 5:00-6:00 (DMP 110, starting next week).
- Teaching Assistants:
  - Reza Babanezhad.
  - Raunak Kumar.
  - Alireza Shafaei.

Final Exam

- Final exam details:
  - Date will be written here (eventually).
  - Closed book, three-page double-sided "cheat sheet".
  - No requirement to pass the final (but recommended).

- Do not miss the final.
  - I don't control when the final is, don't make travel plans before April 25th.

- There will be two types of questions:
  - 'Technical' questions requiring things like pseudo-code or derivations.
    - Similar to assignment questions, and will only be related topics covered in assignments.
  - 'Conceptual' questions testing understanding of key concepts.
    - All lecture slide material except "bonus slides" is fair game here.

Course Project

- Course projects can be done in groups of 2-3 and have 3 parts:
  1. Project proposal (due with Assignment 4).
  2. Literature review (due with Assignment 5).
  3. Coding, experiments, application, or theory (due late April).
     - More details coming later in term, and I don’t care if you switch groups during the term.

Lectures

- All slides will be posted online (before lecture, and final version after).

- Please ask questions: you probably have similar questions to others.
  - I may deflect to the next lecture or Piazza for certain questions.

- Be warned that the course we will move fast and cover a lot of topics:
  - Big ideas will be covered slowly and carefully.
  - But a bunch of other topics won't be covered in a lot of detail.

- Isn't it wrong to have only have shallow knowledge?
  - In this field, it's better to know many methods than to know 5 in detail.
    - This is called the "no free lunch" theorem: different problems need different solutions.
    - If you why something is important, and the core ideas, you can fill in details later.

Course Outline

- We'll cover the following core machine learning research topics:
  - Large-scale machine learning (my research area).
  - Structured prediction (machine learning with multiple outputs).
    - Density estimation.
    - Graphical models.
    - Recurrent neural networks.
  - Bayesian methods.
- Topics needed to understand machine learning research papers.

- Some of these are not the "usual" machine learning topics.
  - Most of the "usual" topics are covered in CPSC 340 (overview of topics).

Bonus Slides

- I will include a lot of "bonus slides".
  - May mention advanced variations of methods from lecture.
  - May overview big topics that we don't have time for.
  - May go over technical details that would derail class.

- You are not expected to learn the material on these slides.
  - But you may find them interesting or useful in the future.

- I'll use this colour of background on bonus slides.