CPSC 340: Machine Learning and Data Mining

Mark Schmidt

University of British Columbia, Fall 2017

www.cs.ubc.ca/~schmidtm/Courses/340-F17

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:
  - YouTube, Facebook, MOOCs, news sites.
  - Credit cards transactions and Amazon purchases.
  - Transportation data (Google Maps, Waze, Uber)
  - 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.

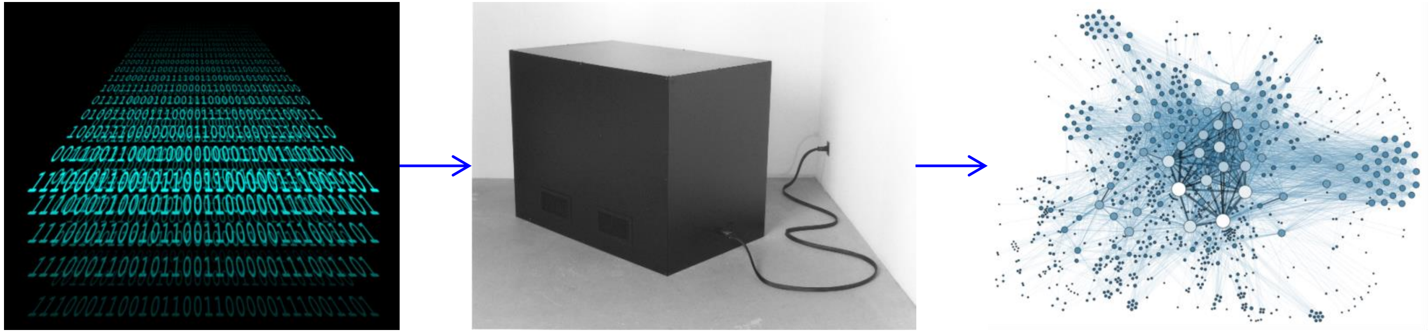
What can I help you with?

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

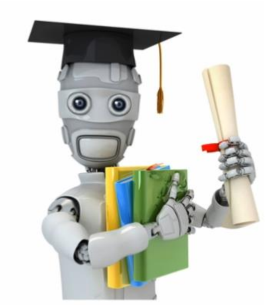
- Automatically extract useful knowledge from large datasets.



- Usually, to help with human decision making.

Machine Learning

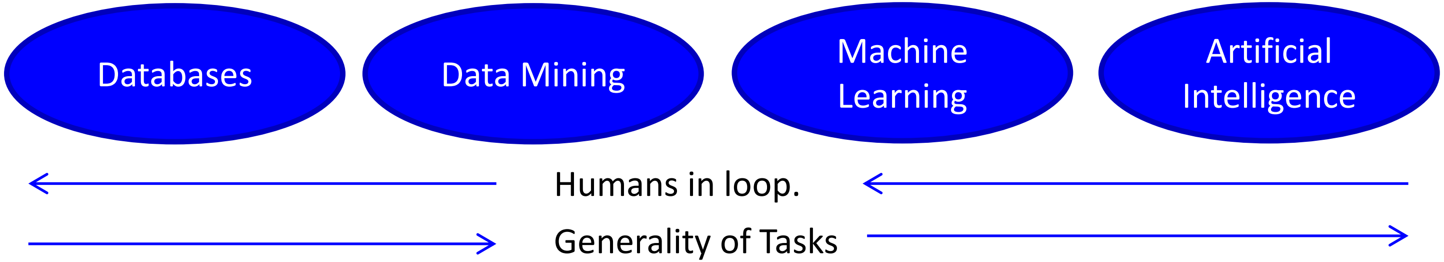
- Using computer to automatically detect patterns in data and use these to make predictions or decisions.



- Most useful when:
  - We want to automate something a human can do.
  - We want to do things a human can't do (look at 1 TB of data).

Data Mining vs. Machine Learning

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



Databases

Data Mining

Machine Learning

Artificial Intelligence

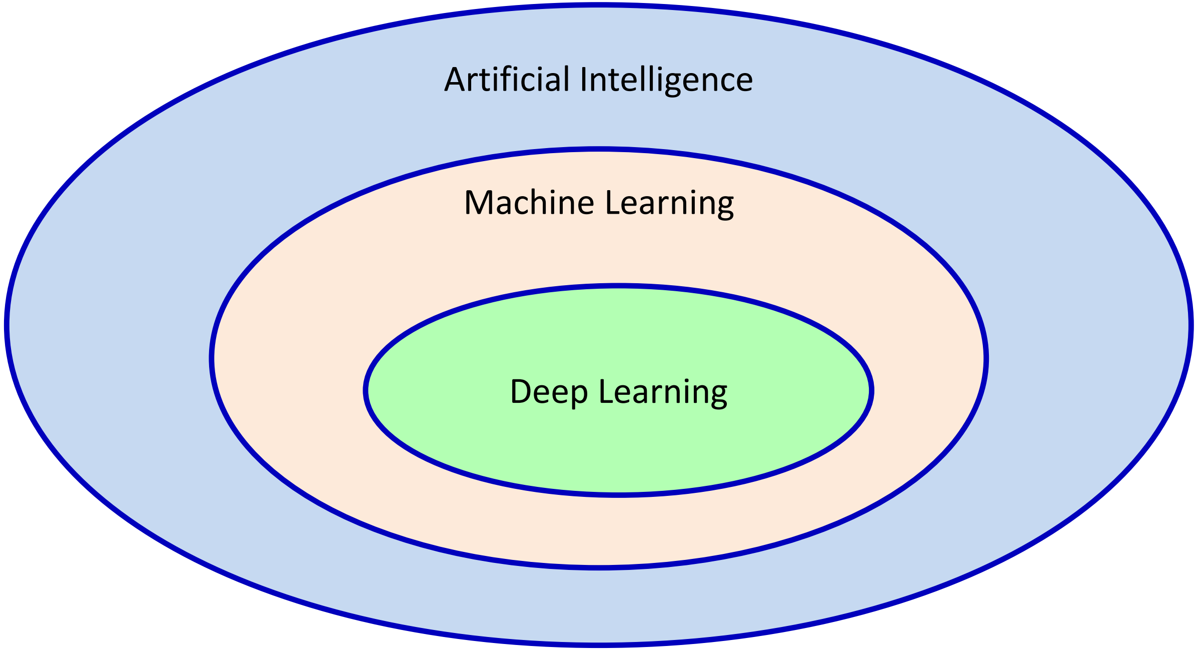
Humans in loop.

Generality of Tasks

- Both are similar to statistics, but more emphasis on:
  - Large datasets and computation.
  - Predictions (instead of descriptions).
  - Flexible models (that work on many problems).

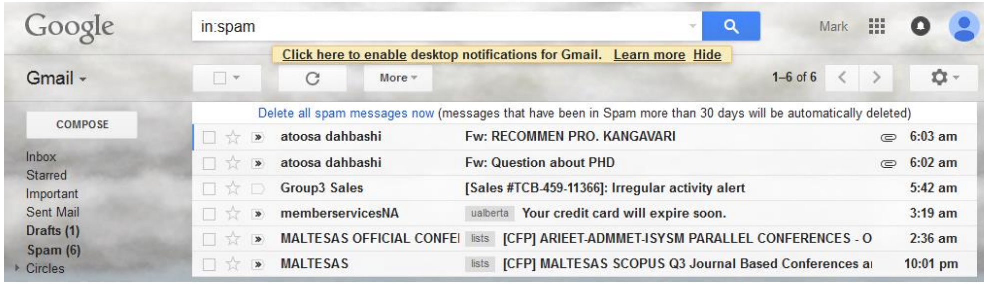
Deep Learning vs. Machine Learning vs. AI

- Traditional we've viewed ML as a subset of AI.
  - And "deep learning" as a subset of ML.



Applications

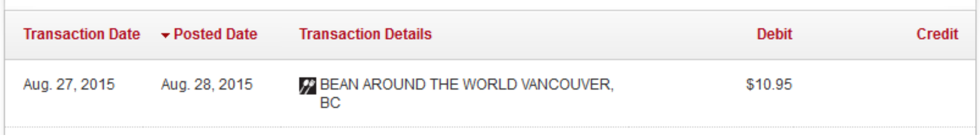
- Spam filtering:
- Credit card fraud detection:
- Product recommendation:



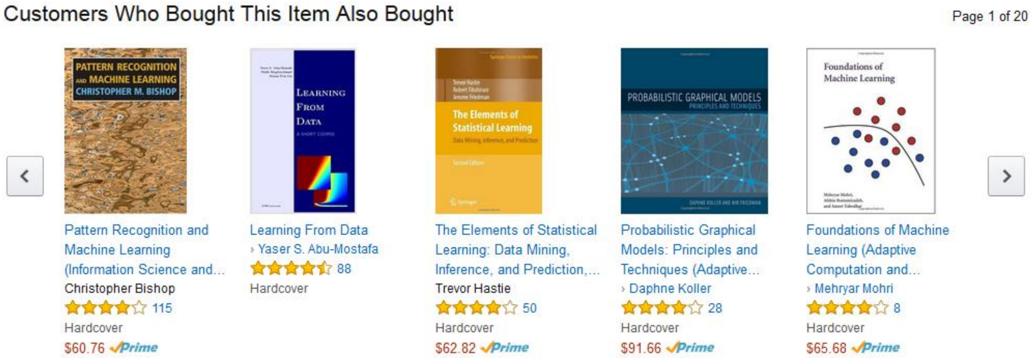
Click here to enable desktop notifications for Gmail. Learn more Hide

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

atoosa dahbashi		Fw: RECOMMEN PRO. KANGAVARI	6:03 am
atoosa dahbashi		Fw: Question about PHD	6:02 am
Group3 Sales		[Sales #TCB-459-11366]: Irregular activity alert	5:42 am
memberservicesNA	ualberta	Your credit card will expire soon.	3:19 am
MALTESAS OFFICIAL CONFEI	lists	[CFP] ARIEET-ADMMET-ISYSM PARALLEL CONFERENCES - O	2:36 am
MALTESAS	lists	[CFP] MALTESAS SCOPUS Q3 Journal Based Conferences ai	10:01 pm



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


PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
LEARNING FROM DATA
The Elements of Statistical Learning
PROBABILISTIC GRAPHICAL MODELS
Foundations of Machine Learning

Customers Who Bought This Item Also Bought

Page 1 of 20

Pattern Recognition and Machine Learning (Information Science and...
Christopher Bishop
★★★★☆ 115
Hardcover
$60.76 ✓Prime

Learning From Data
› Yaser S. Abu-Mostafa
★★★★½ 88
Hardcover

The Elements of Statistical Learning: Data Mining, Inference, and Prediction,...
Trevor Hastie
★★★★☆ 50
Hardcover
$62.82 ✓Prime

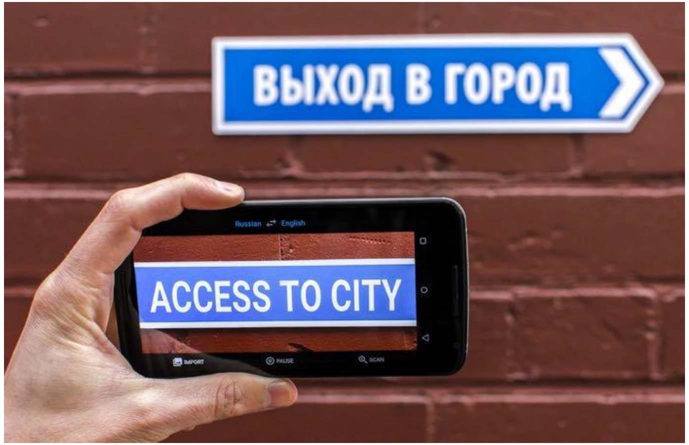
Probabilistic Graphical Models: Principles and Techniques (Adaptive...
› Daphne Koller
★★★★☆ 28
Hardcover
$91.66 ✓Prime

Foundations of Machine Learning (Adaptive Computation and...
› Mehryar Mohri
★★★★☆ 8
Hardcover
$65.68 ✓Prime

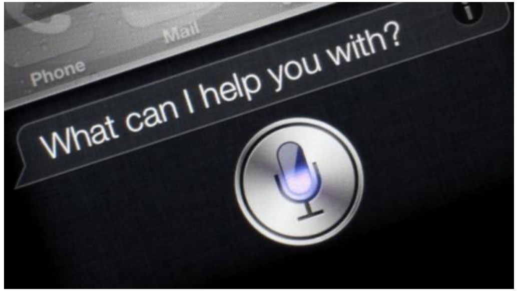
Applications

- Motion capture:

- Optical character recognition and machine translation:

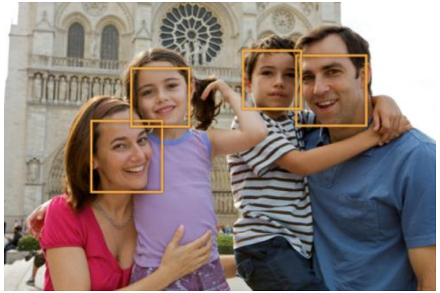


- Speech recognition:

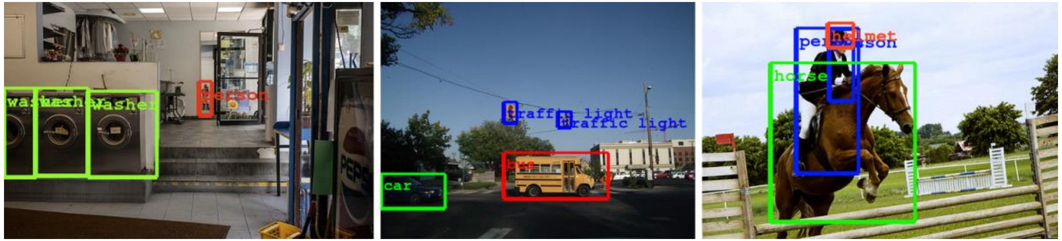


Applications

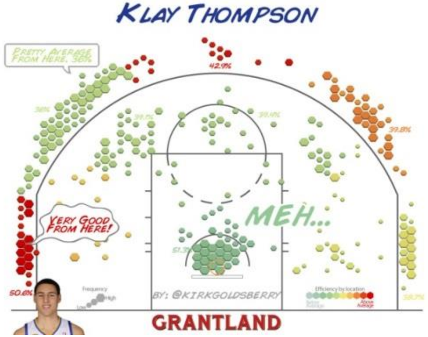
- Face detection:



- Object detection:

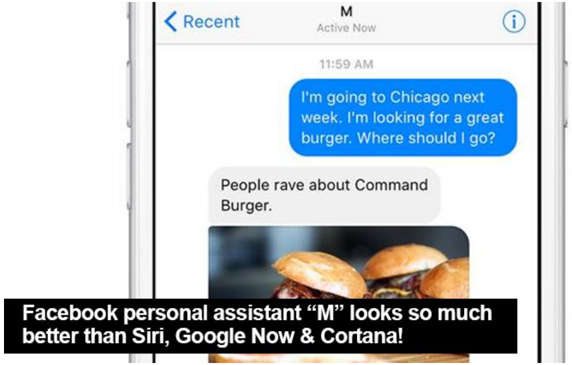


- Sports analytics:



Applications

- Personal Assistants:



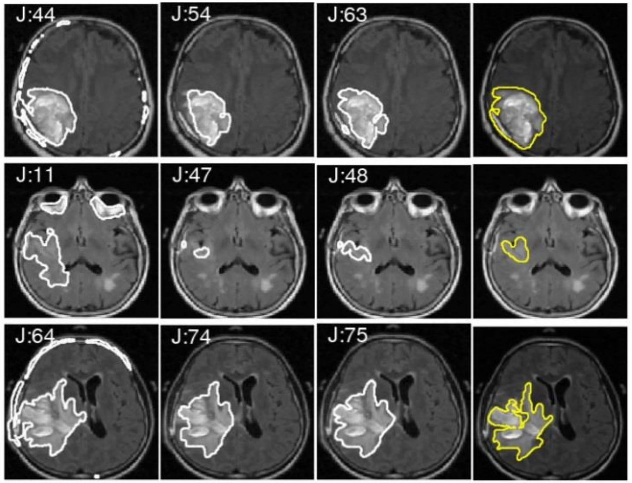
11:59 AM

I'm going to Chicago next week. I'm looking for a great burger. Where should I go?

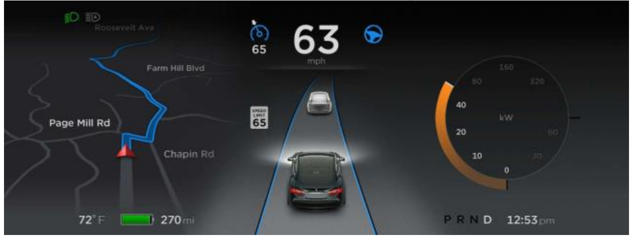
People rave about Command Burger.

Facebook personal assistant “M” looks so much better than Siri, Google Now & Cortana!

- Medical imaging:

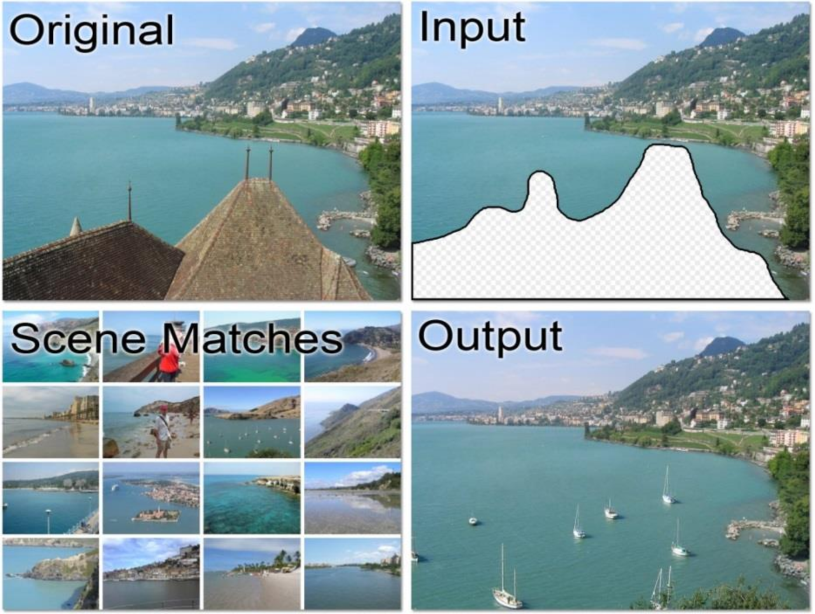


- Self-driving cars:



Applications

- Scene completion:



- Image annotation:

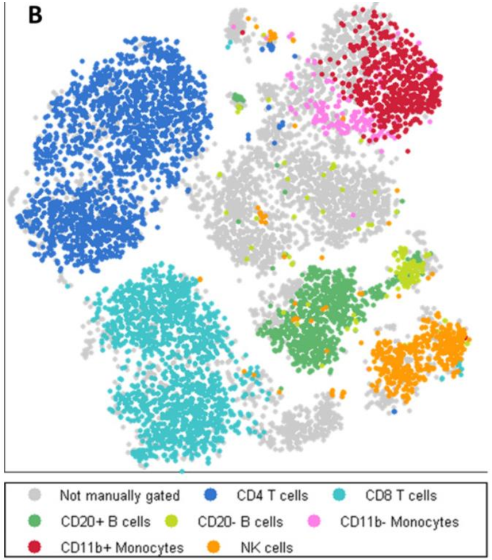
a cat is sitting on a toilet seat
logprob: -7.79

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

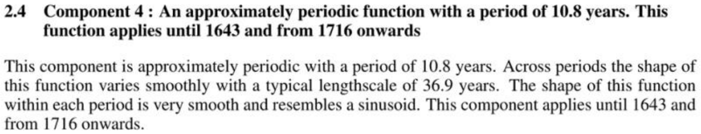
a group of people sitting at a table with wine glasses
logprob: -6.71

Applications

- Discovering new cancer subtypes:

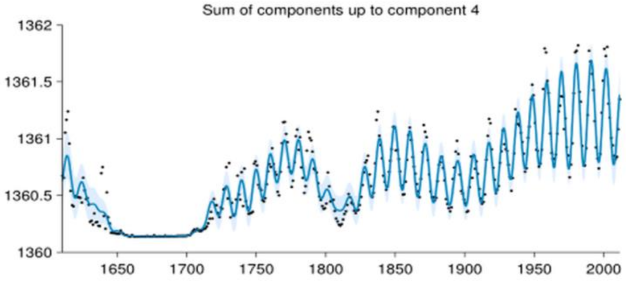


- Automated Statistician:



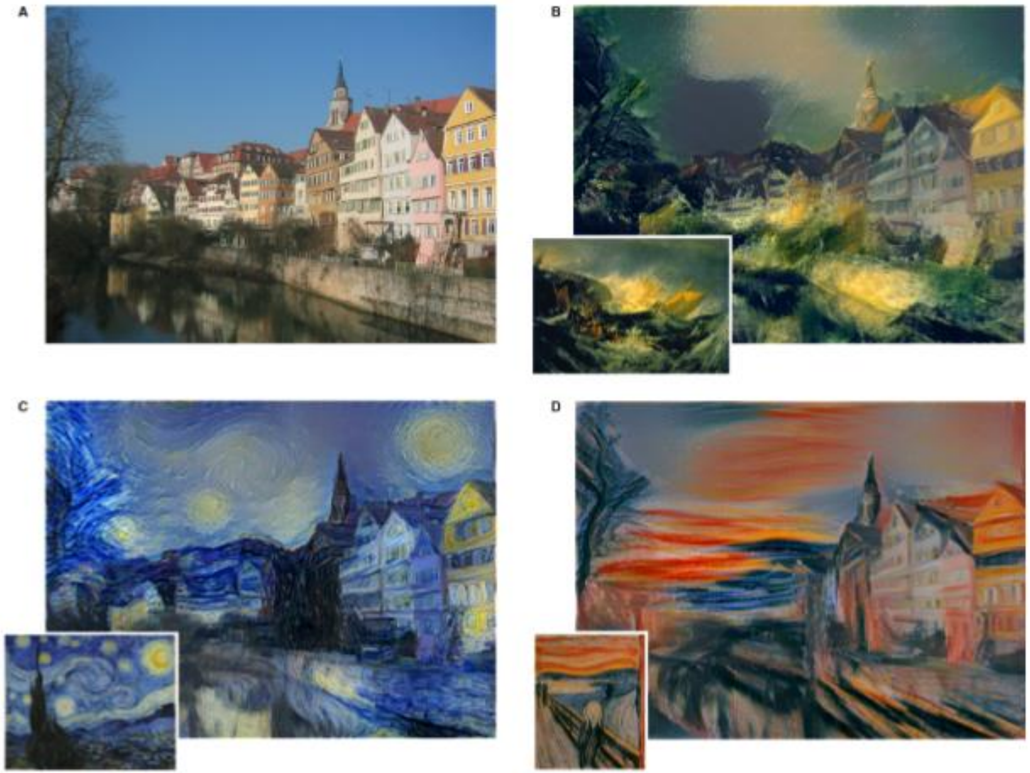
2.4 Component 4 : An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards

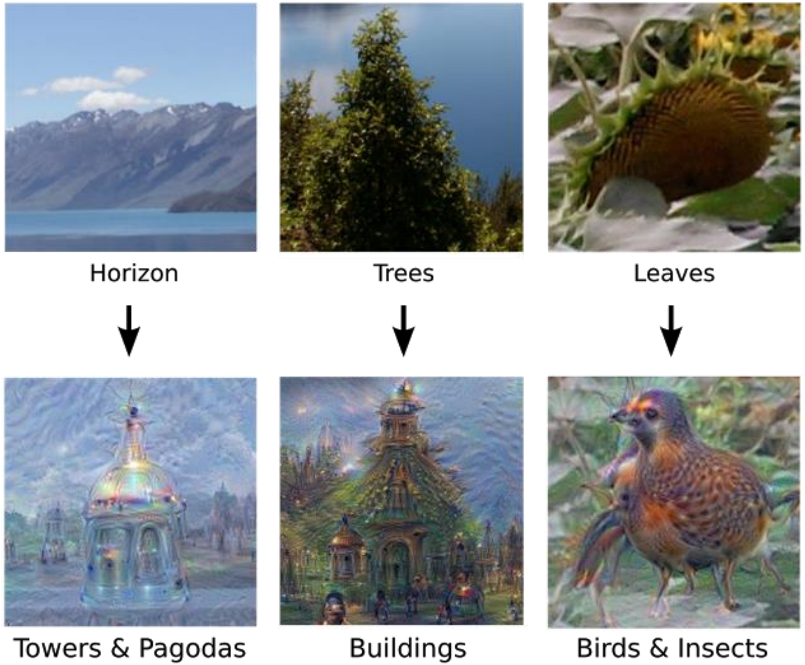
This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716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



Applications

- Mimicking artistic styles and inceptionism:



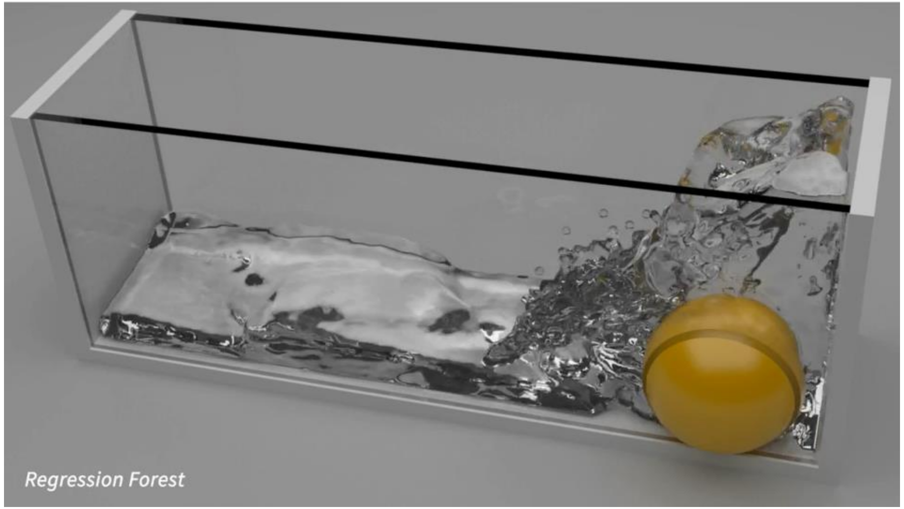


Applications

- “Deep dream”:

Applications

- Fast physics-based animation:



- Mimicking art style in video.
- Recent work on generating text/music/voice/poetry/dance.

Applications

- Beating human Go masters:

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

- But it is important to know the limitations of what you are doing.
  - "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." – John Tukey
  - A huge number of people applying ML are just "overfitting".

- We are in exciting times.
  - Major recent progress in fields like speech recognition and computer vision.
  - Things are changing a lot on the timescale of 3-5 years.
  - A bubble in ML investments.

(pause)

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 (at once).
  - "I think the prerequisites for this course should require that students have obtained at least 75% (or around there) in the required math courses. As someone who who did not excel at math, I felt severely under prepared and struggled immensely in this course, especially seeing that I have taken CPSC courses in the past with similar math requirements, but were not nearly as math heavy as CPSC340."

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 (at once).
- Compared to non-CS classes, there is a lot more programming:
  - 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 third undergrad course.
- We'll use the Julia language: Mike Gelbart uses Python.
- Take this course to learn, not to get a certain grade.

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.
    - It covers much more material than standard ML classes like Coursera.
  - CPSC 540 focuses on less widely-used methods and research topics.
    - It is intended as a continuation of CPSC 340.
    - You'll miss important topics if you skip CPSC 340.

Essential Links

- Please bookmark the course homepage:
  - www.cs.ubc.ca/~schmidtm/Courses/340-F17
  - Contains lecture slides, assignments, optional readings, additional notes.

- You should sign up for Piazza:
  - www.piazza.com/ubc.ca/winterterm12017/cpsc340/home.
  - Can be used to ask questions about lectures/assignments/exams.
  - May occasionally be used for course announcements.

- Use Piazza instead of e-mail for questions:
  - I can take a long time to respond e-mails.

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.).
  - Mining Massive Datasets (Leskovec et al.)
  - Machine Learning: A Probabilistic Perspective (Murphy).

- Most of these are on reserve in the ICICS reading room.
- List of related courses on the webpage, or you can use Google.

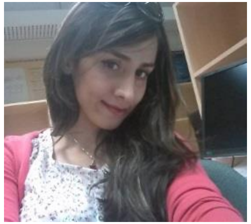
TA Cheat Sheet

- Clement Fung

- Hashemi Hooman
- Siyuan He
- Tanner Johnson
- Angad Kalra

- Xin Bei She
- Sharan Vaswani

- Nasim Zolaktaf



- Zainab Zolaktaf

Assignments and Working in Teams

- There will be 6 Assignments worth 30% of final grade:
  - Usually a combination of math and programming.
  - Submitted as a zip file using the Handin program.
    - You will need to setup a CS account to use this.

- Assignment 0 is on the webpage, and is due next Friday.

- Assignment 0 must be done individually.
- Assignments 1-5 can be done in pairs.
  - There is no commitment to keep the same pairs between assignments.

Late “Class” Policy for Assignments

- Assignments will be due at midnight “anytime on Earth” (ATE).

- If you can’t make it, you can use “late classes”:
  - For example, if assignment is due on a Friday:
    - Handing it in Friday is 0 late classes.
    - Handing it in Monday is 1 late class.
    - Handing it in Wednesday is 2 late classes.

  - You will get a mark of 0 on an assignment if you:
    - Use more than 2 late classes on the assignment.
    - Exceed 4 late classes across all assignments.
    - Submit the solutions to an assignment from a previous term.

- We’ll try to put grades on Connect within 1 week of due date.

Programming Language: Julia

- 3 most-used languages in these areas: Python, Matlab, and R.

- We will be using Julia which is similar to Matlab.
  - Except it’s free and is way faster than Python/Matlab/R.

- No, you cannot use Python/Matlab/R/etc.
  - Assignments have prepared code that we won’t translate to 3 languages.
  - TAs shouldn’t have to know 3 languages to grade.

Waiting List and Auditing

- Right now only CS students register directly.
- 181/195 seats are filled, but the room supports 250 students.

- 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!
    - They might also ask to submit a prereq form, let me know if you have issues.

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

Getting Help

- Many students find the assignments long and difficult.
- But there are many sources of help:
  - TA office hours and instructor office hours (see webpage for times).
    - Starting in the second week of class.
  - Piazza.
  - Weekly tutorials.
    - Starting in second week of class.
    - Will go through provided code, review background material, review big concepts, and/or do exercises.
    - Tutorials are optional be you must be registered in a tutorial section to stay enrolled.
  - Other students (ask your neighbor for their e-mail).
  - The web (almost all topics are covered in many places).

Midterm and Final

- In-class midterm worth 20% (tentatively scheduled for October 20) and a (cumulative) final worth 50% (some time on/before December 22)
  - Closed-book.
  - One doubled-sided ‘cheat sheet’ for midterm.
  - Two doubled-sided pages for final.
  - No need to pass the final to pass the course (but recommended).

- There will be two types of questions:
  - ‘Technical’ questions requiring things like pseudo-code or derivations.
    - Similar to assignment questions, 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.

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.

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 they’re useful if you want to take 540 or work in this area.

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

Code of Conduct

- Do not post offensive or disrespectful content on Piazza.
- If you have a problem or complaint, let me know (maybe we can fix it).
- Do not distribute any course materials without permission.
- Do not record lectures without permission.
- Acknowledge all sources, including webpages and other students.
- Think about how/when to ask for help:
  - Don’t ask for help after being stuck for 10 seconds. Make a reasonable effort to solve your problem (check instructions, Piazza, and Google).
  - But don’t wait until the 10th hour of debugging before asking for help.
- 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%.

Course Outline

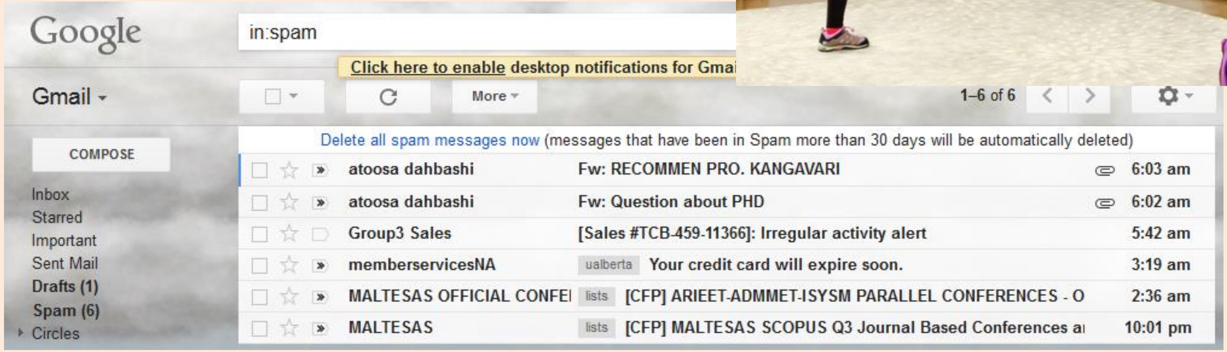
- Next class discusses data “exploratory data analysis”.

- After that, the remaining lectures focus on five topics:
  1) Supervised Learning.
  2) Unsupervised learning.
  3) Linear prediction.
  4) Latent-factor models.
  5) Deep learning.

(pause)

Supervised Learning

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



in:spam

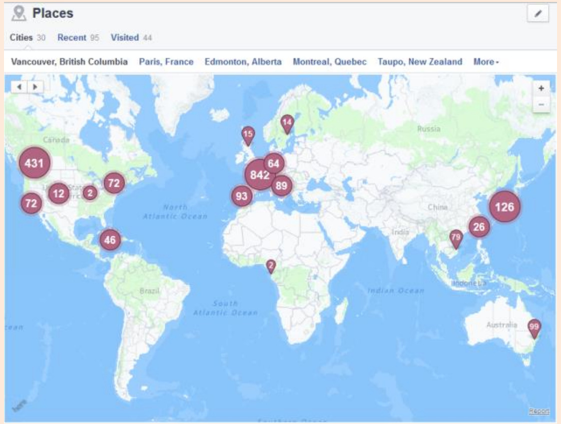
Click here to enable desktop notifications for Gmai

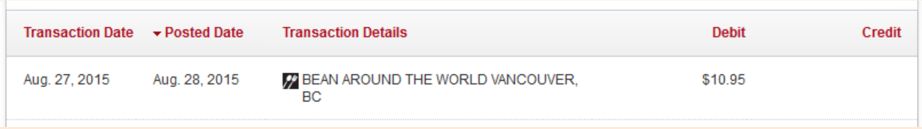
Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

atoosa dahbashi	Fw: RECOMMEN PRO. KANGAVARI	6:03 am
atoosa dahbashi	Fw: Question about PHD	6:02 am
Group3 Sales	[Sales #TCB-459-11366]: Irregular activity alert	5:42 am
memberservicesNA	ualberta Your credit card will expire soon.	3:19 am
MALTESAS OFFICIAL CONFEI	lists [CFP] ARIEET-ADMMET-ISYSM PARALLEL CONFERENCES - O	2:36 am
MALTESAS	lists [CFP] MALTESAS SCOPUS Q3 Journal Based Conferences ai	10:01 pm

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’.



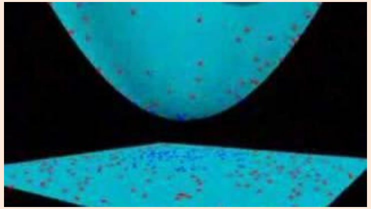


Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

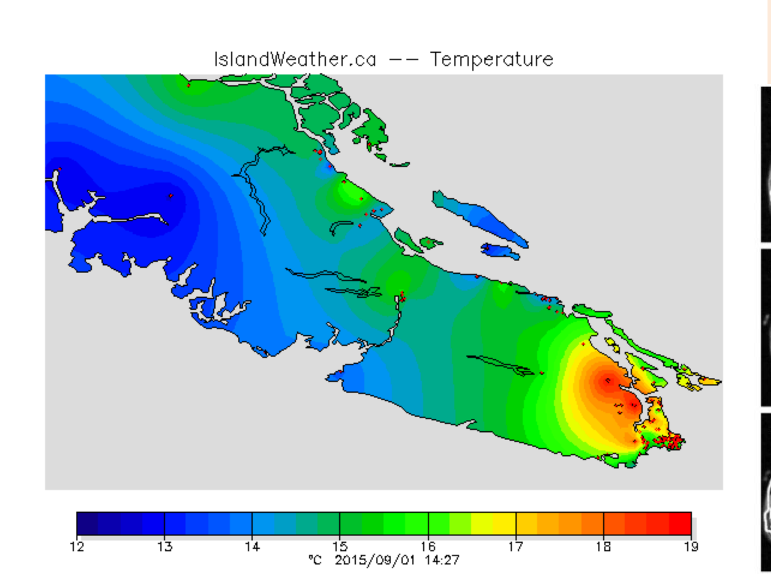
MARKET BASKET ANALYSIS

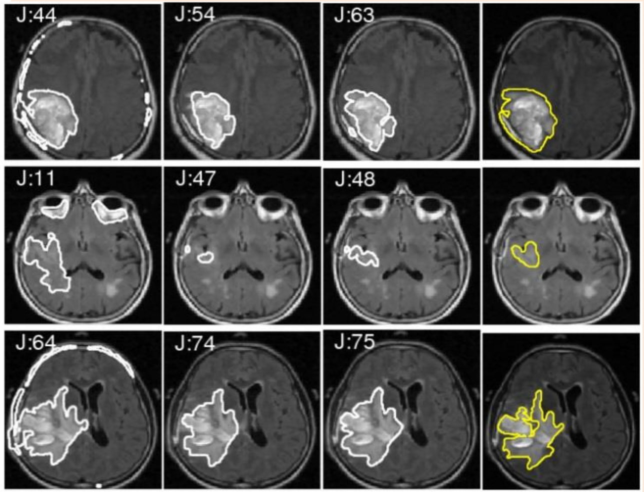
98% of people who purchased items A and B also purchased item C

Linear Prediction

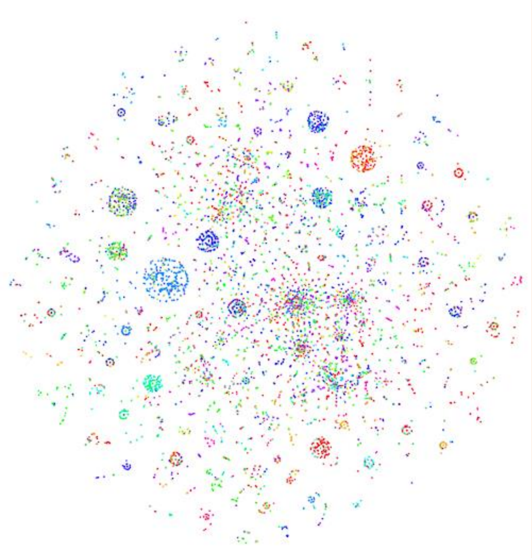


- Regression:
  - Predicting continuous-valued outputs.
- Working with very high-dimensional data.

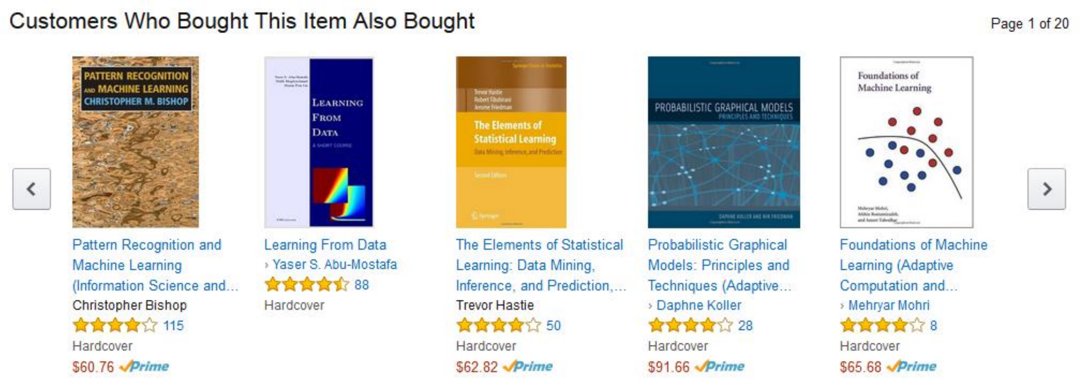




Latent-Factor Models



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



Customers Who Bought This Item Also Bought

Page 1 of 20

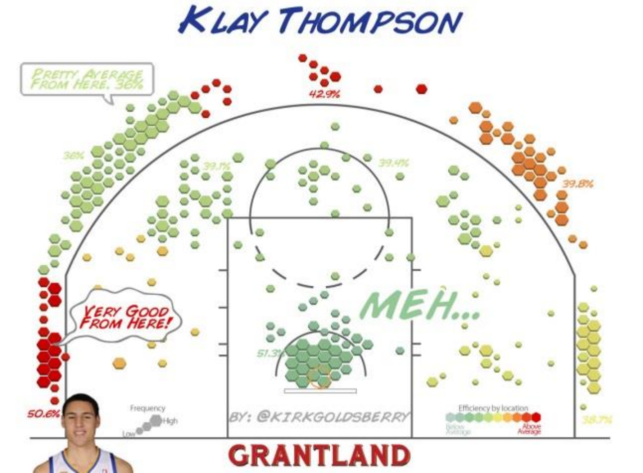
Pattern Recognition and Machine Learning (Information Science and...
Christopher Bishop
★★★★☆ 115
Hardcover
$60.76 ✓Prime

Learning From Data
› Yaser S. Abu-Mostafa
★★★★☆ 88
Hardcover

The Elements of Statistical Learning: Data Mining, Inference, and Prediction,...
Trevor Hastie
★★★★☆ 50
Hardcover
$62.82 ✓Prime

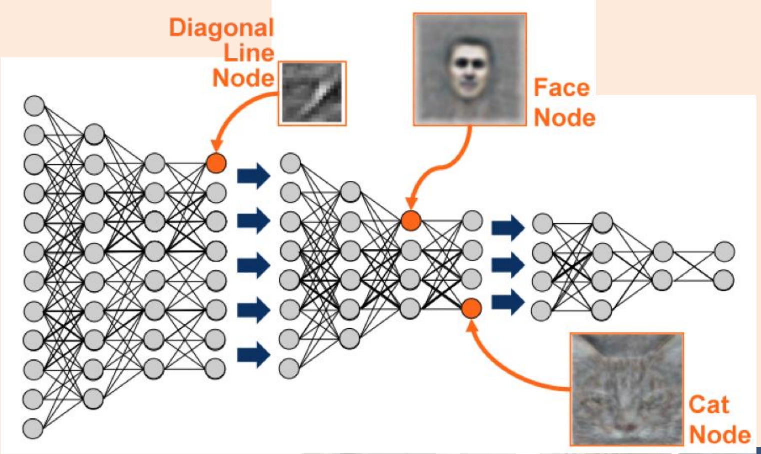
Probabilistic Graphical Models: Principles and Techniques (Adaptive...
› Daphne Koller
★★★★☆ 28
Hardcover
$91.66 ✓Prime

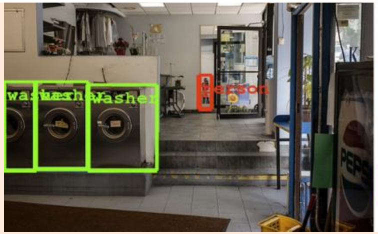
Foundations of Machine Learning (Adaptive Computation and...
› Mehryar Mohri
★★★★☆ 8
Hardcover
$65.68 ✓Prime

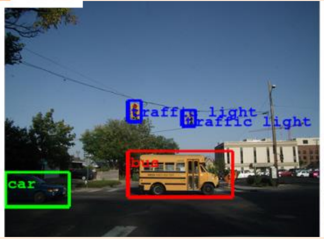


Deep Learning

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







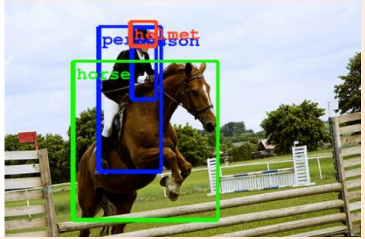
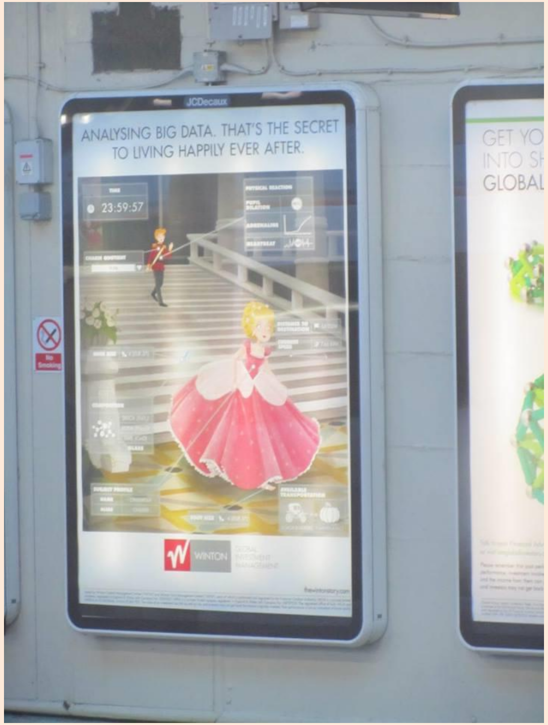


Photo I took in the UK on the way home from the “Optimization and Big Data” workshop:


JCDecaux
ANALYSING BIG DATA. THAT'S THE SECRET TO LIVING HAPPILY EVER AFTER.
23:59:57
WINTON
GET YO
INTO SH
GLOBAL