Introduction to Machine Learning & others

Dr. Jinjun Xiong
(jx2308@columbia.edu)

Outline

- Course logistics
- Introduction to machine learning
- Conclusion

About this class

- Lectures: every Friday 4 – 6 pm
- Office hours
  - By appointments only (In-person office hours are 30 minutes before each lecture)
  - Otherwise, E-mail
- Class TAs
  - Han
  - Terry
  - Kevin
- Class website
  - ?

Programming requirements

- You don’t need any programming background in order to understand the lecture content
- But you are highly recommended to learn at least one of many programming languages (Python, R, Matlab, C/C++, etc) in order to perform well on most assignments
- I will not be biased toward any programming language, i.e., I won’t use any particular programming examples to teach you the ML concepts
  - This also means that I won’t teach you any programming skills as well
- But I’ll let our TAs to hold different tutorial sessions (optional) to help you get things started if needed
  - Python (2^{nd} week), R (3^{rd} week), Matlab (4^{th} week)

Grading

- Three graded in-class quizzes, 5 points each = total 15 points
  - Randomly chosen 3 in-class quizzes for grading
- Six homework assignments, 5 points each = total 30 points
- Midterm project: 20 points
- Final project: 20 points
- Participation & attendance: 15 points

- Bonus points
  - 2 points for midterm project
  - 3 points for final project
  - 3 points for in-class presentation

Grading scale

- No pre-determined distribution

A+	98 % and above
A	93 – 97.9%
A-	90 – 92.9%
B+	87 – 89.9%
B	83 – 86.9%
B-	80 – 82.9%
C+	77 – 79.9%
C	73 – 76.9%
C-	70 – 72.9%
D	60 – 69.9%
F	59.9% and below

Homework Assignments

Homework	Assignment Date	Due Date	Content
#1	Friday, 1/27/2017	Friday, 2/10/2017	• Review and critique existing visualization from Data Visualization Gallery • Create visualization for some existing dataset, using any visualization tools
#2	Friday, 2/10/2017	Friday, 2/24/2017	• Background knowledge about linear algebra, complexity, gradient, Hessian, and Newton's method for optimization
#3	Friday, 2/24/2017	Friday, 3/3/2017	• Linear regression • Clustering
Midterm project report	Friday, 2/17/2017	Sunday, 3/12/2017	
#4	Friday, 3/24/2017	Friday, 4/7/2017	• Logistic regression • Support vector machine • Decision tree
#5	Friday, 4/7/2017	Friday, 4/21/2017	• Ensemble learning and random forest
#6	Friday, 4/21/2017	Friday, 4/28/2017	• NLP • Deep learning
Final project report	Friday, 3/31/2017	Sunday, 4/30/2017	

Midterm & final projects

- Two options
  - Two separate topics: one for midterm and one for final
  - One big topic with two-phase progress reports, but graded separately
- Topics can be chosen from a list of suggested topics (to be given per the schedule), or you can propose your own
  - In the latter case, please start to think about it now (and discuss with me ASAP so you may start early)
- The projects will most likely involve some kind of programming + reports
  - Theoretic hypothesis with proof is welcome too
  - Some ideas can turn into a quality conference paper, so be bold ☺
- Team size of 1 – 3 students
  - Talk to me if you want to form a bigger team with justification
  - Start to form your team now (& get to know your fellow students)
  - Team must be registered before you start the project

Midterm & final project grading

- Project reports should indicate clearly the contribution of each team member
  - For the part that is not easy to separate individual's contributions, a percentage should be
- Project reports should also include a section about how well the team work together
  - Link to the way the team is deciding on each team member's contribution
- Grading
  - 50% of each individual's contribution to the project
  - 50% of team's contribution (all team members share the same grades for this part)
    - Team work (including ways to resolve conflicts if any) is highly valued
- Bonus points will be given to either the team or individuals, depending on the quality of work

In-class presentation for the 3 bonus points

- Starting the 5th week (2/13-2/19), I will have three 10-minutes in-class slots reserved for student presentation
  - Total 30 slots (= 3 slots/week * 10 weeks)
  - Sign-up deadline: Friday, 2/10
- Format
  - The student presenter
    - Choose a dataset from the public domain
    - Explain, explore & visualize the dataset
    - Suggest a few ML techniques (learnt so far) that can be used to learn useful information from the dataset
    - Justify your suggestion with reasons
  - The student audience
    - Optionally to submit ONE write-up to review and critique ONE of the in-class presentations
      - What is good about the presentation, and what could be done better

Syllabus

- 1st week (1/17-1/22): Introduction
- 2nd week (1/23-1/29): Data visualization and exploration
  - The last 30 mins reserved in-class tutorial on Python led by TA
- 3rd week (1/30-2/5): Data preprocessing
  - The last 30 mins reserved in-class tutorial on R led by TA
- 4th week (2/6-2/12): Background
  - The last 30 mins reserved in-class tutorial on Matlabled by TA
  - Friday, 2/10, deadline to register your volunteer in-class presentations
- 5th week (2/13-2/19): Linear regression
  - The last 30 mins reserved for student in-class presentations
  - Friday, 2/17, deadline to register your mid-term and final project team members
  - Friday, 2/17, selection of mid-term projects finalized

Syllabus

- 6th week (2/20-2/26): Clustering
  - The last 30 mins reserved for student in-class presentations
- 7th week (2/27-3/5): Perceptron & logistic regression
  - The last 30 mins reserved for student in-class presentations
- 8th week (3/6-3/12): Support vector machine
  - The last 30 mins reserved for student in-class presentations
- Spring break (3/13 – 3/19): no classes
- 9th week (3/20-3/26): Decision tree
  - The last 30 mins reserved for student in-class presentations
- 10th week (3/27-4/2): Ensemble learning and random forest
  - The last 30 mins reserved for student in-class presentations
  - Friday, 3/31, selection of final projects finalized

Syllabus

- 11th week (4/3-4/9): Hyper parameter tuning
  - The last 30 mins reserved for student in-class presentations
- 12th week (4/10-4/16): Deep learning
  - The last 30 mins reserved for student in-class presentations
- 13th week (4/17-4/23): Natural language processing
  - The last 30 mins reserved for student in-class presentations
- 14th week (4/24-4/30): IBM Watson and Bluemix
  - The last 30 mins reserved for student in-class presentations

Outline

- Course logistics
- Introduction to machine learning
- Conclusion

The world of computing has changed dramatically since …

JEOPARDY!

The IBM Challenge

What is Jeopardy!

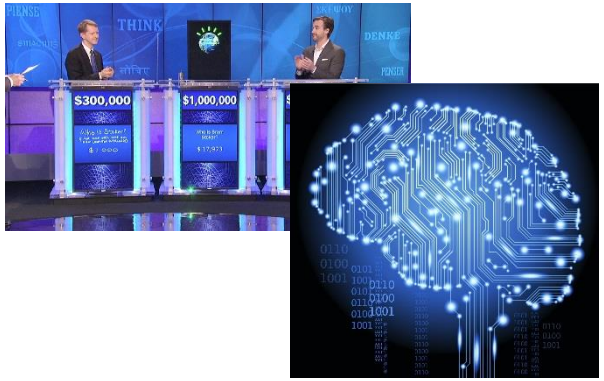
- An American quiz show since 1964
- Answer-and-question format
  - Contestants are presented with clues in the form of answers
  - Contestants must phrase their responses in question form

***Category**: General Science*

***Clue**: When hit by electrons, a phosphor gives off electromagnetic energy in this form*

***Answer**: What is light?*

Transformation of Computing



Cognitive computing

Save time to comprehend information
Easy to make decision

1950: Alan Turing proposes the Turing Test as a measure of machine intelligence.
1956: The first Dartmouth College summer AI conference. The name artificial intelligence is coined by John McCarthy.
1965: ELIZA, an interactive program that carries on a dialogue in English language on any topic.
1969: Perceptrons, a book considered by some to mark the beginning of the AI winter of the 1970s
1997: The Deep Blue chess machine (IBM) defeats the (then) world chess champion, Garry Kasparov.
2004: DARPA introduces the DARPA Grand Challenge requiring competitors to produce autonomous vehicles.
2011: IBM's Watson computer defeated Jeopardy! champions Rutter and Jennings.
2016: Google's AlphaGo won 4-1 against Lee Sedol, one of the top Go players in the world

Internet

Save time to communicate information
Easier access to more information

1945: Vannevar Bush's article "As We May Think" predicts the evolution of hypertext.
1958: Advanced Research Projects Agency (ARPA) is created by US Department of Defense
1969: The first ARPANET node is installed at UCLA Network Measurement Center.
1971: The first ARPANET network email message is transmitted.
1972: Dialog offers the first publicly available online research service.
1979: USENET emerges as a collection of user-submitted messages
1982: ARPANET shifts to TCP/IP.
1986: NSFNET replaces ARPANET as the main government network linking universities and research facilities.
1993: The HTML 1.0 standard published. World Wide Web released. Mosaic, first graphical browser introduced.
1999: The Google search engine is launched
2003:The amount of information transmitted globally over the Internet is projected to double each year.
2004: 55% of adult internet users have broadband at home or work.
2006: Twitter is founded
2007: Apple's iPhone released
2013: Tenth anniversary of the International Internet Preservation Consortium.

Digitization

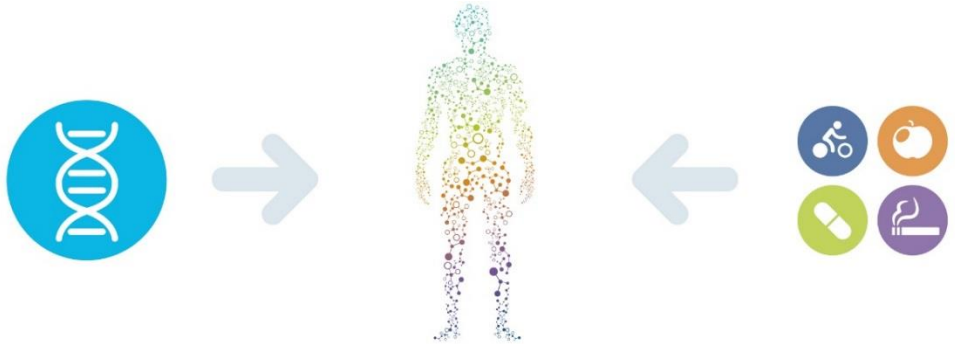
Save time to search information
Easy access to information

1881:J.S. Billings suggests to Herman Hollerith that a mechanical system based on cards be used to tabulate the Census.
1890: Hollerith develops a punch card system used with the 1890 Census.
1923: Enigma machine
1924: Hollerith's "Computer Tabulating Recording Company" renamed "International Business Machines Corporation" (IBM).
1938: First use of the term digital applied to a computer
1964: IBM System/360 is announced
1965: Moore's Law established
1970: IBM System/370 is introduced with the notion of a virtual machine, allowing users to share mainframe resources.
1979: WordStar software becomes the first commercially successful word processor
1983: Apple's Lisa is introduced, the first commercial microcomputer with a graphical user interface.
2009: The World Digital Library is launched. All Television broadcasting in the U.S. went digital by June 12, 2009
2013: Digital Public Library of America launched.

[http://www.dpworkshop.org/dpm-eng/timeline/viewall.html, https://en.wikipedia.org/wiki/Timeline_of_artificial_intelligence]

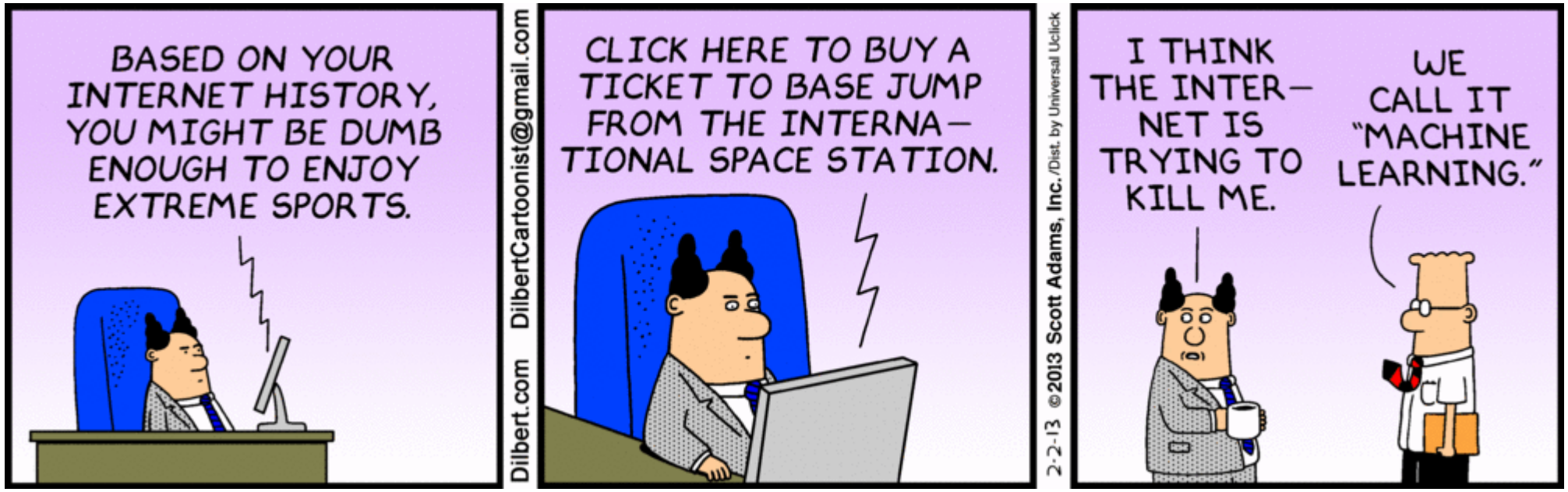
Everyone in town is talking about things related to machine learning

- Social media
- Self-driving cars
- Trading
- Precision medicine
- …



https://stacksocial.com/sales/quant-trading-using-machine-learning
https://www.geico.com/more/wp-content/uploads/self-driving-car-post.jpg
http://www.northeastern.edu/careers/jobs-internships/social-media/
http://www.metabolon.com/precision-medicine.aspx

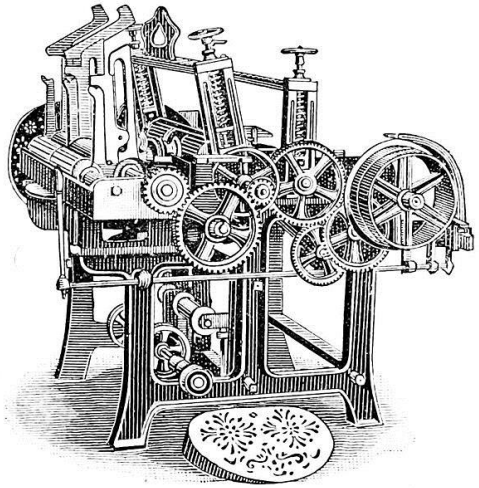
What is machine learning?

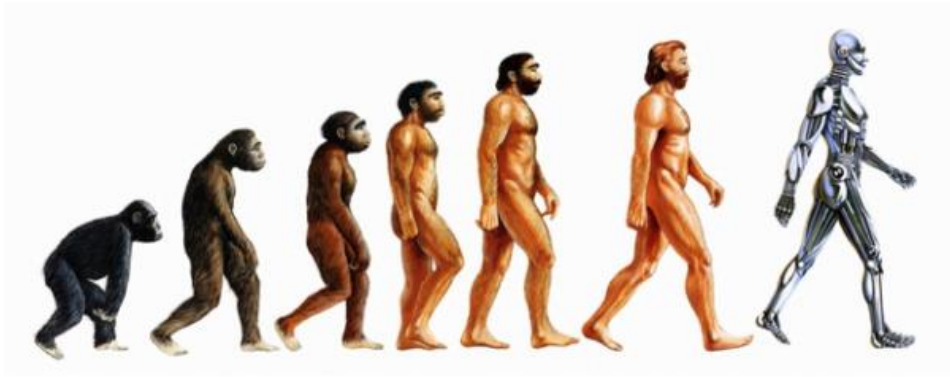


[http://dilbert.com/strip/2013-02-02]

What is machine learning anyway, seriously?

- It turns out that this can be a very tricky question
- What is learning?
- What is machine?
- What does it mean when we ask if a machine CAN learn?





[https://www.linkedin.com/pulse/20140409101445-23292849-screen-size-is-the-disruptive-change-in-personal-computing]
[https://commons.wikimedia.org/wiki/File:PSM_V39_D312_A_gilling_machine.jpg]
[https://news.webhosting.info/ibm-opens-new-data-center-in-south-korea/]
[http://www.3plearning.com/human-evolution-human-beings/]

A formal definition of machine learning

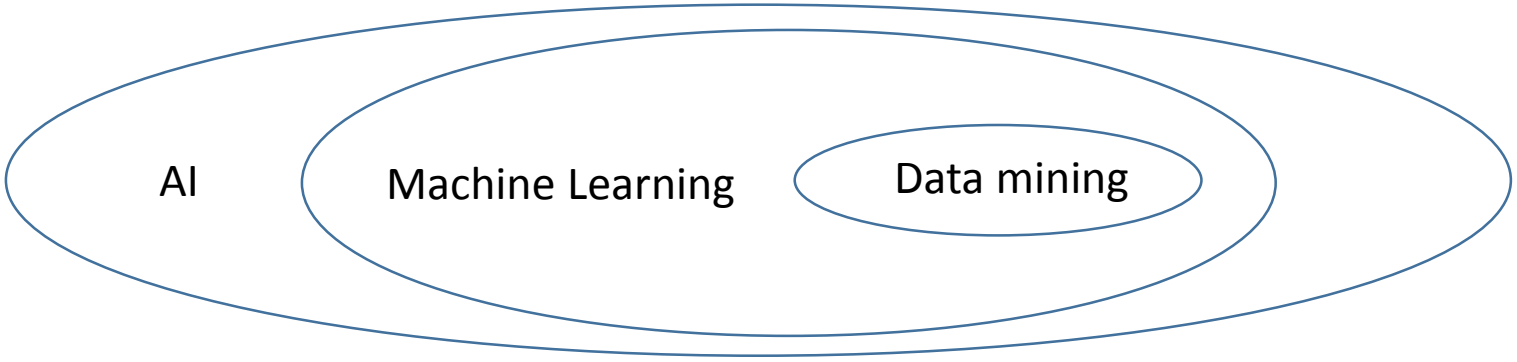
- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

- For example, learn to play checkers
  - T: play checkers
  - P: % of games won in world tournament
  - E: opportunity to play against self

[“Machine Learning,” T. Mitchell, 1997]

Artificial intelligence, machine learning and data mining

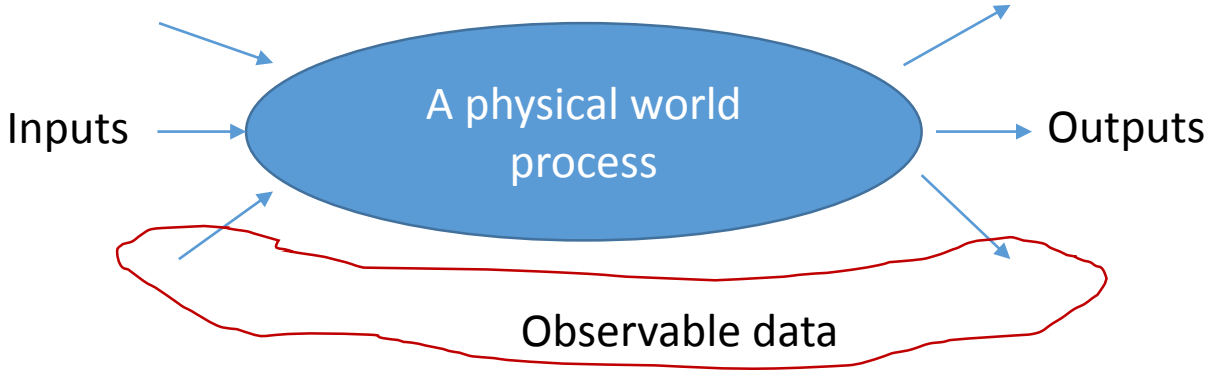
- Artificial intelligence: intelligence exhibited by machines
  - Artificial general intelligence (AGI): intelligence of a machine that could successfully perform any intellectual task that a human being can
    - Strong AI
  - Applied AI: the use of software to study or accomplish specific problem solving or reasoning tasks
    - Weak AI
- Data mining: the process of discovering patterns: automatically or semi-automatically, in large quantities of data



Very crude conceptual distinction of AI, machine learning and data mining

Machine learning: a general view

- A general view of data we collected
  - The physical world process can be anything of interests
    - Social behavior
    - Human's state of mind while speaking
    - Natural phenomena (tree growing, mushroom types, etc.)
  - The physical world process can be time varying as well
    - Most of the problems we deal with in this course do NOT consider the time varying effects
- Not all inputs and outputs are observable
  - We collect a subset of observable data with repeated "experience"

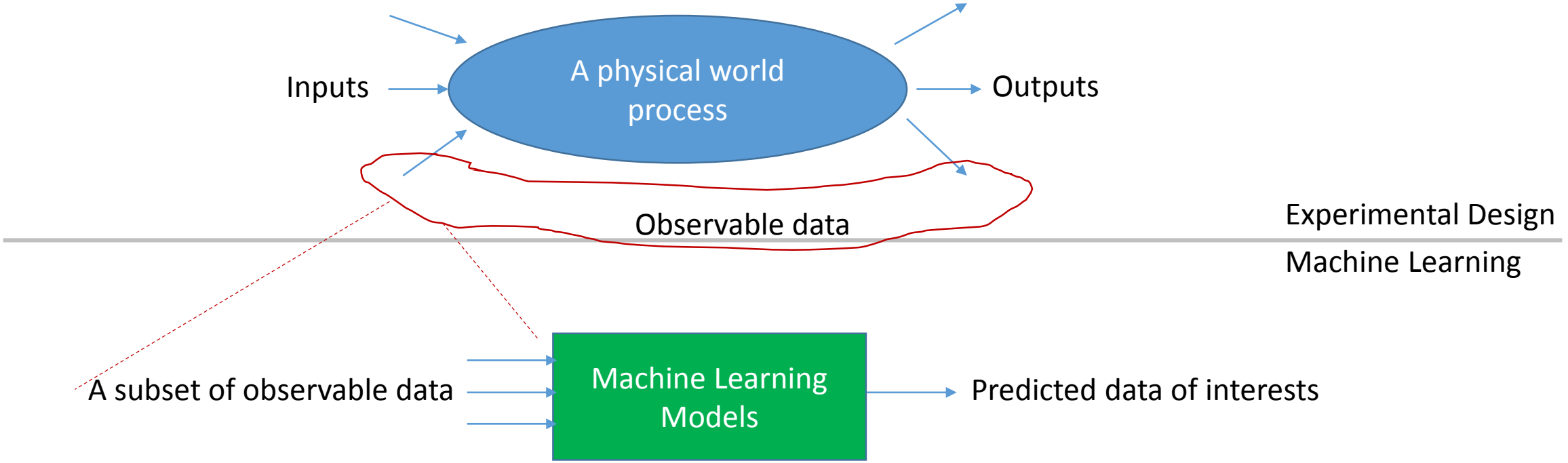


Machine learning: things to watch out for

- The physical world process can be
  - Unknown (e.g., human social behaviors, stock buying and selling)
  - Known, but too complicated to simulate in time (e.g., semiconductor physics)
- When we put the collected observable data together, we don't necessarily know what data are inputs and what data are outputs
- But we typically know what interested data to predict, i.e., what class of tasks to perform, based on what observable data
  - The data to be predicated and the observable data do NOT necessarily have any causal effects
    - "The fire emergence" vs "The alarm sounds of fire engines"
- Machine learning is to provide that bridge as a surrogate model (in part) of the physical world process

Machine learning models

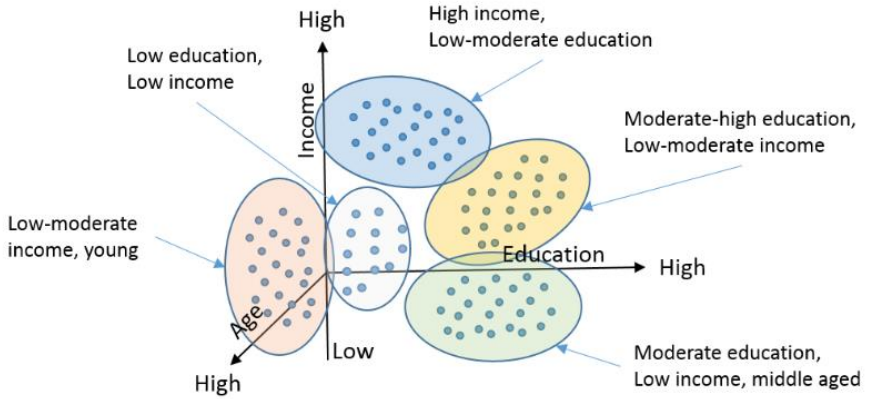
- Though ML models do not necessarily depend on the knowledge about the physical world process, imparting such knowledge to the design of ML model will definitely help



What does a machine learning model look like?

- Model can be considered as a way to represent “knowledge” in some structural form
  - A functional form
  - A cluster form that partitions data
  - A set of decision rules

$$y = f(x)$$



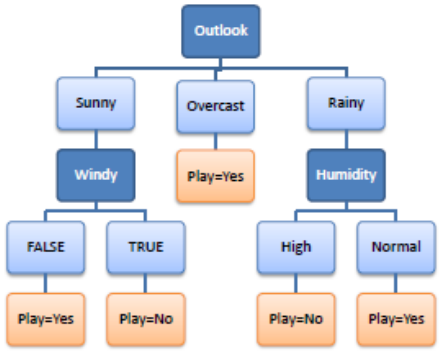
R_1: IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

R_2: IF (Outlook=Sunny) AND (Windy=TRUE) THEN Play=No

R_3: IF (Outlook=Overcast) THEN Play=Yes

R_4: IF (Outlook=Rainy) AND (Humidity=High) THEN Play=No

R_5: IF (Outlook=Rain) AND (Humidity=Normal) THEN Play=Yes



Outline

- Course logistics
- Introduction to machine learning
- Conclusion

What will be covered in this course

- A set of most popular machine learning models that have been proven useful to solve a wide range of learning tasks
  - Some models are derived based on the knowledge of the original physical world process, but they are later extended to solve other problems that do not necessarily resemble the original physical world process (or “do we care?”)
  - Many more ML models can be found from the literature that is beyond the scope of this course
- Some fundamental concepts that underpin many of those ML models
  - Mastering those fundamental concepts will help you to better understand the various algorithms people have designed to solve the ML problems
- A strong mental readiness to tackle any machine learning problems in your future career endeavor
  - The fundamentals are not that hard to understand
  - Do not be fooled by many fancy terminologies and mathematics

Your roles after learning this course

- ML end users: use the ML application developed by someone else with your application data
  - Know how to interpret ML results
- ML application developers: use existing ML packages and develop your own applications
  - Know how to select models (pros and cons) and how to train models with existing ML packages
- ML model developers: modify existing models or suggest new models based on application knowledge (more mathematic background)
  - Use existing ML packages to incorporate new models
- ML algorithm developers: who faster and better ways of training ML models (more computer science background)
  - Implement the algorithm in existing software packages or develop new one

Homework - optional

- Based on the conceptual framework below, please come up with a few examples from your experience where machine learning can be of use
  - Define what is the physical world process & what data is observable
  - Define what data of interests to predict
  - Define the relationship between the physical world process and your proposed machine learning model (i.e., how the ML model should roughly look like?)