CS4487
MACHINE LEARNING
LECTURE 1 - INTRO

2018 Semester A

Course General Info

- **Teaching Team**
  - Dr. Antoni B. CHAN
    - abchan@cityu.edu.hk, Office: AC1-G7311
    - Office hours: 1pm-2pm Mondays
  - TAs: Ms. ZHAN Xueying
    - xyzhan2-c@my.cityu.edu.hk
    - Office hours: ???, 11:50-1:00 Mondays
- **Canvas-based course site**
  - It is your own responsibility to check Canvas and University e-mail account regularly for announcements and updates.

Teaching Activities

- **Lectures** (2 hours per week)
  - present machine learning algorithms: intuition and idea. Illustrate algorithms on both toy and real-world examples.
- **Tutorial** (1 hour per week)
  - Use machine learning algorithms on small examples to gain better understanding.
- **Assignment** (2)
  - Apply machine learning algorithms to larger datasets, compare and interpret the results of different algorithms.
- **Course Project**
  - Apply machine learning to solve a real-world problem.
  - Up to 2 students per group.

Assessment

- **Coursework** (70%)
  - Tutorial exercises (10%) – due each week (before lecture)
  - Assignments (30%) – due Weeks 6 and 9.
  - Course Project (30%) – due Week 13
    - Project proposal, report, and presentation.
- **Final Exam** (30%)
- **Note:**
  - Must get at least 30% on final exam and 30% on course project to pass the course.

Programming

- **Python**
  - high-level scripting language
  - Jupyter, aka iPython Notebooks (ipynb)
    - interactive computational environment in web browser.
    - can combine text (Markdown) with Python code and output.
  - Libraries
    - numpy – arrays, linear algebra
    - scikit, scikit-learn – scientific computing, machine learning
    - matplotlib, pylab – plotting
  - Introduction later today

Resources

- **Assignment/Projects**
  - kaggle.com – a website for data science competitions.
    - Assignments will use Kaggle for evaluation
      - http://inclass.kaggle.com
    - Course projects based on current Kaggle competitions.
      - select among a list of candidates.
  - Code/Report submission
    - iPython notebooks (ipynb)

Resources

- **Textbooks**
  - Muller & Guido, "Introduction to Machine Learning with Python", O'Reilly, 2017.
  - P. Harrington, "*Machine Learning in Action*", Manning Publications Co., 2012.
- **Online Reference Books**
  - A. Rajaraman, and J. Ullman, "*Mining of Massive Datasets*", Cambridge University Press, 2011. (http://infolab.stanford.edu/~ullman/mmds.html)
  - H. Daume III, "*A course in Machine Learning*", (http://ciml.info/)

Resources

- **Other Books**
  - C.M. Bishop, "*Pattern Recognition and Machine Learning*", Springer, 2006.
  - R.O. Duda, P.E. Hart, & D.G. Stork, "*Pattern Classification (2nd Ed.)*", Wiley-Interscience, 2001.
  - T. Hastie, R. Tibshirani, and J. Friedman, "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed.)*", Springer-Verlag, 2009.
  - Deep Learning http://www.deeplearningbook.org

Course Abstract

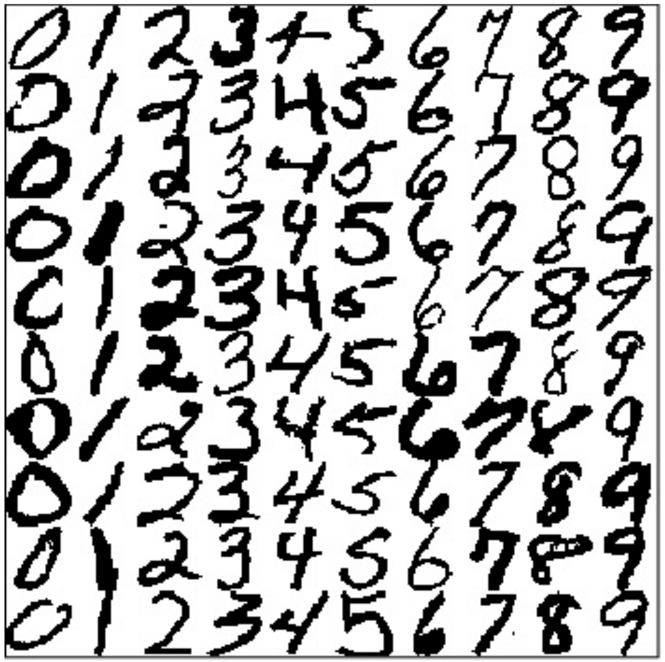
- The goal of this course is to introduce students to the field of machine learning.
  - Machine learning algorithms allow computers to automatically learn to recognize complex patterns from empirical data, such as text and web documents, images, videos, sound, sensor-data, and databases.
  - This course is intended to give a broad overview of machine learning from the practical standpoint, with a focus on applying machine learning algorithms to real-world problems.
  - At the end of the course, students will have both working knowledge of and practical experience with machine learning algorithms.

CILOs

1. Identify and explain common machine learning algorithms.
2. Apply machine learning algorithms to solve real-world problems.
3. Evaluate the effectiveness of different machine learning algorithms and discuss their advantages and disadvantages.

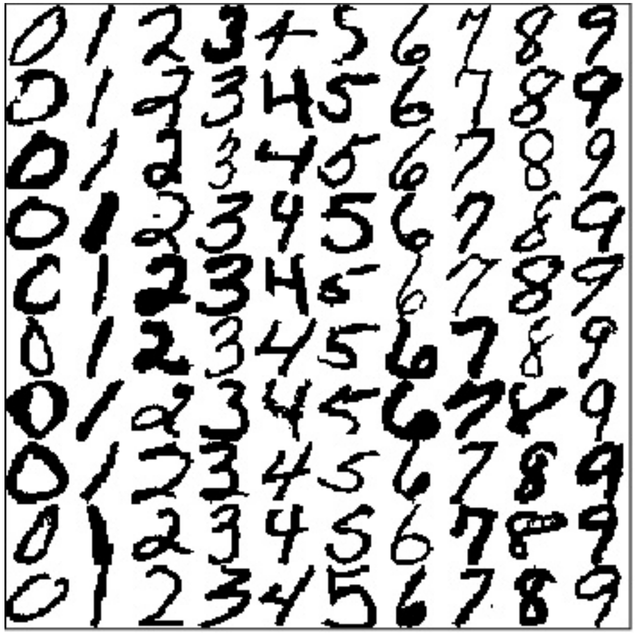
What is Machine Learning?

- Arthur Samuel, 1959
  - *Machine Learning*: field of study that gives computers the ability to learn without being explicitly programmed.
    - e.g. computer learns to play checkers by playing against itself.
- There are many applications that are difficult to program by hand.
  - Example: Recognizing handwritten digits in an image.



What is Machine Learning?

- Example: Recognizing handwritten digits in an image
  - 28x28 image → 784-dim vector
  - a lot of variations & permutations
  - difficult to identify rules & code by hand
- ML solution:
  - gather some example data.
  - train computer to discover differences automatically



What is Machine Learning?

- Tom Mitchell, 1997
  - *Well-posed Learning Problem*: "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*."
    - e.g., the computer gets better at recognizing digits as it sees more examples, as measured by the error rate.
- A closer look…

Well-posed Learning Problem

- "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*."

"class of tasks T"

learning is task-specific (recognition, clustering, etc.)

"performance measure P"

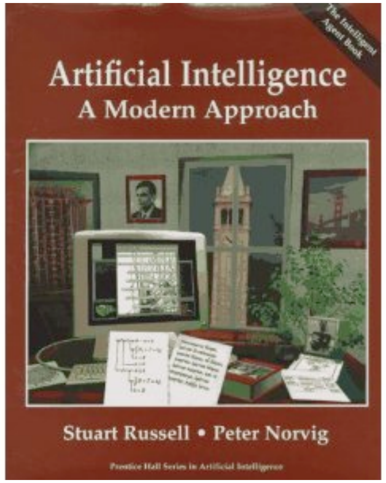
optimize a loss function (e.g., error rate), but also prevent overfitting (regularization). "generalization"

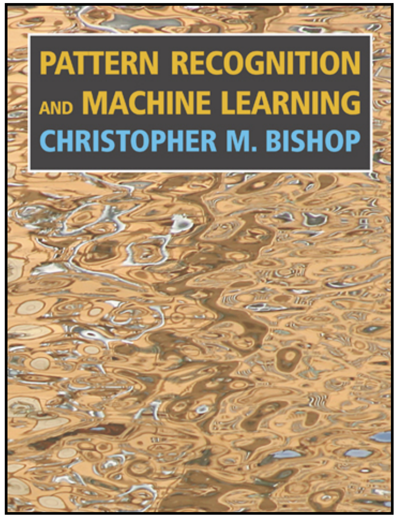
"experience E"

data-driven! More data is better!

Machine Learning vs. Artificial Intelligence

- Machine learning grew out of early work in AI
  - and other fields: statistics, physics, neuroscience, …
  - fueled by more powerful computers and more data.
- "Traditional" Artificial Intelligence (Russell-Norvig)
  - Turing test (is it a computer or a human?)
  - solving by searching (A*, α-β pruning, game playing)
  - knowledge-based (representation, reasoning, logic)
  - planning, scheduling, natural language processing
- Machine Learning (Bishop)
  - probability, statistics, Bayesian formulation
  - statistical learning theory
  - regression, classification, clustering





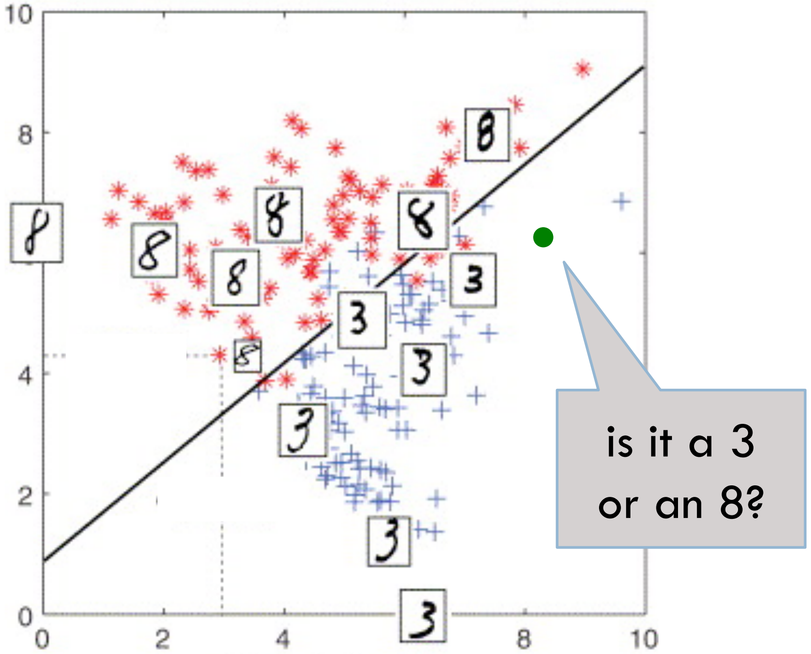
Topics in Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Learning Theory

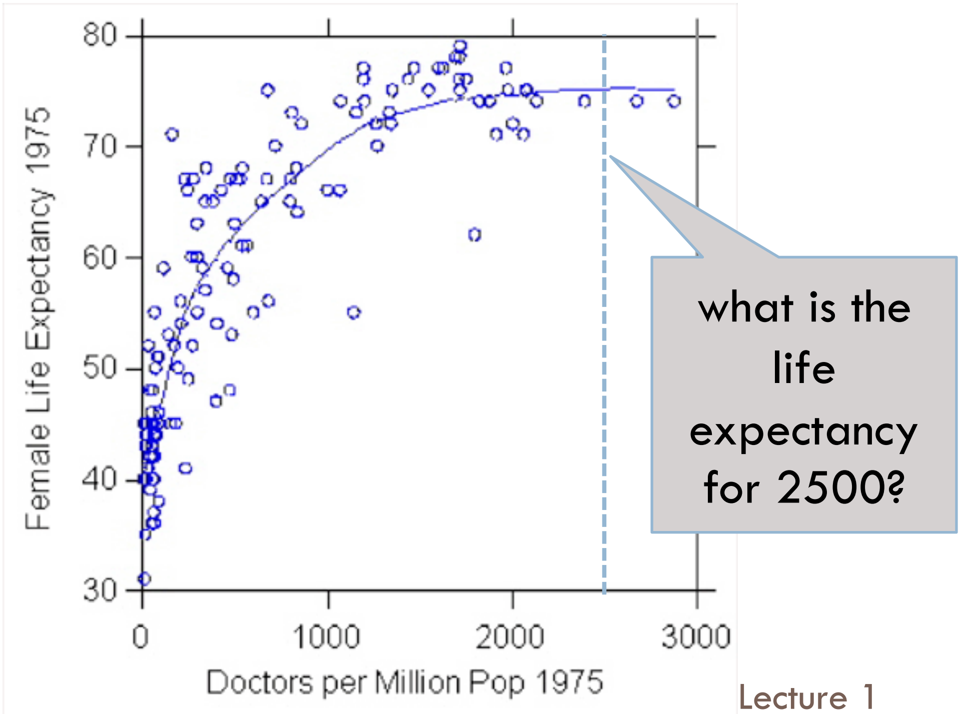
Supervised Learning

- Training data has inputs and outputs
  - e.g., digit recognition (input=image, output=digit)
  - learn a function mapping inputs to outputs

Classification

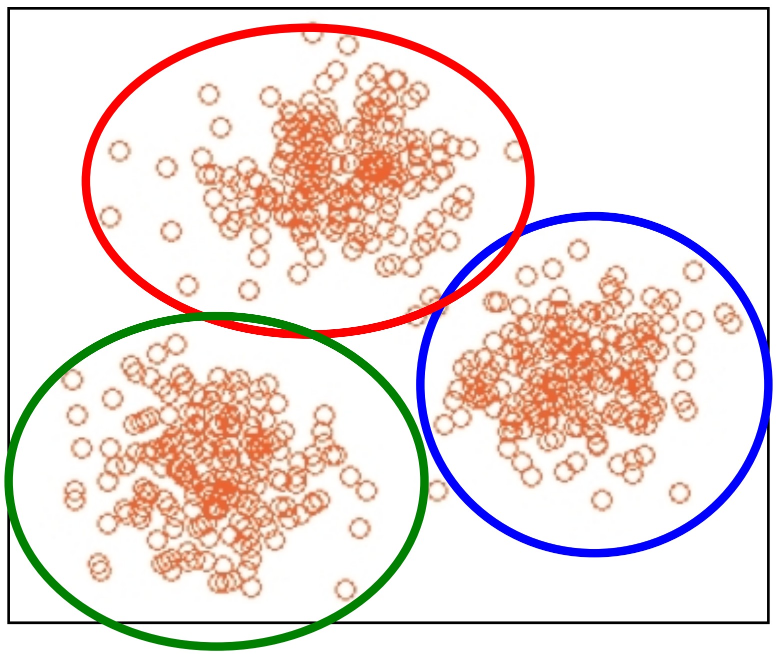


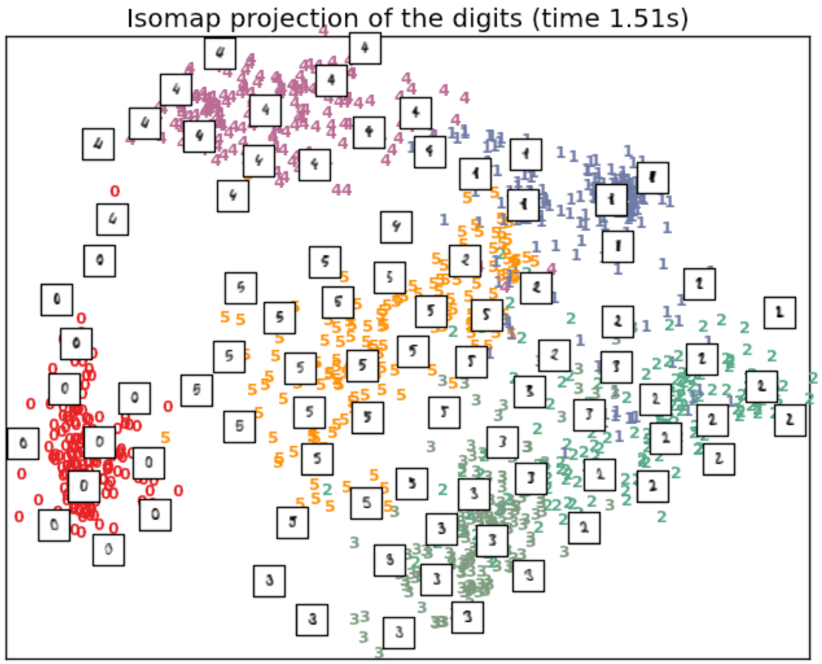
Regression



Unsupervised Learning

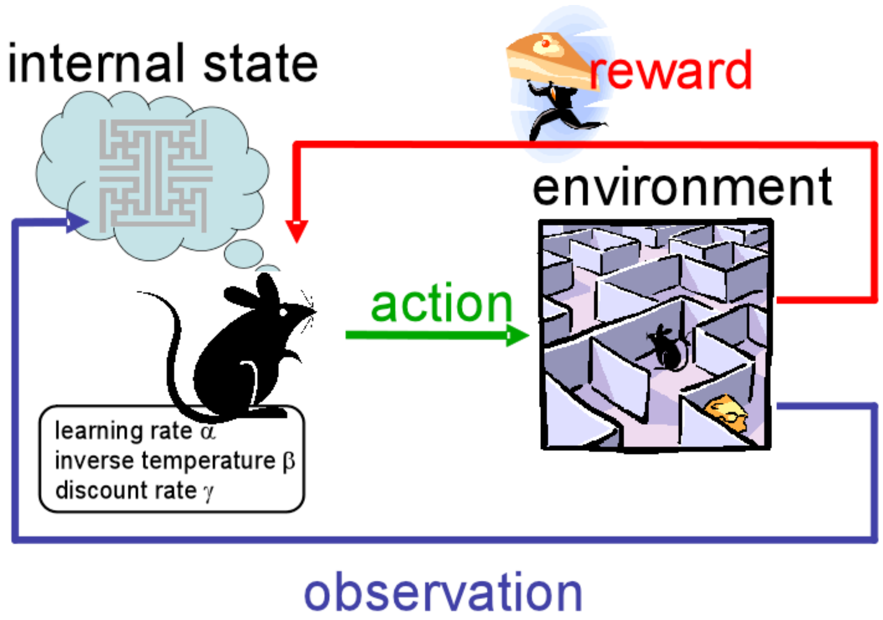
- Training data only has inputs (no outputs)
  - e.g., collection of web documents
  - clustering - discover groups of similar examples.
  - visualization - project high-dim data to 2 or 3-dimensions.





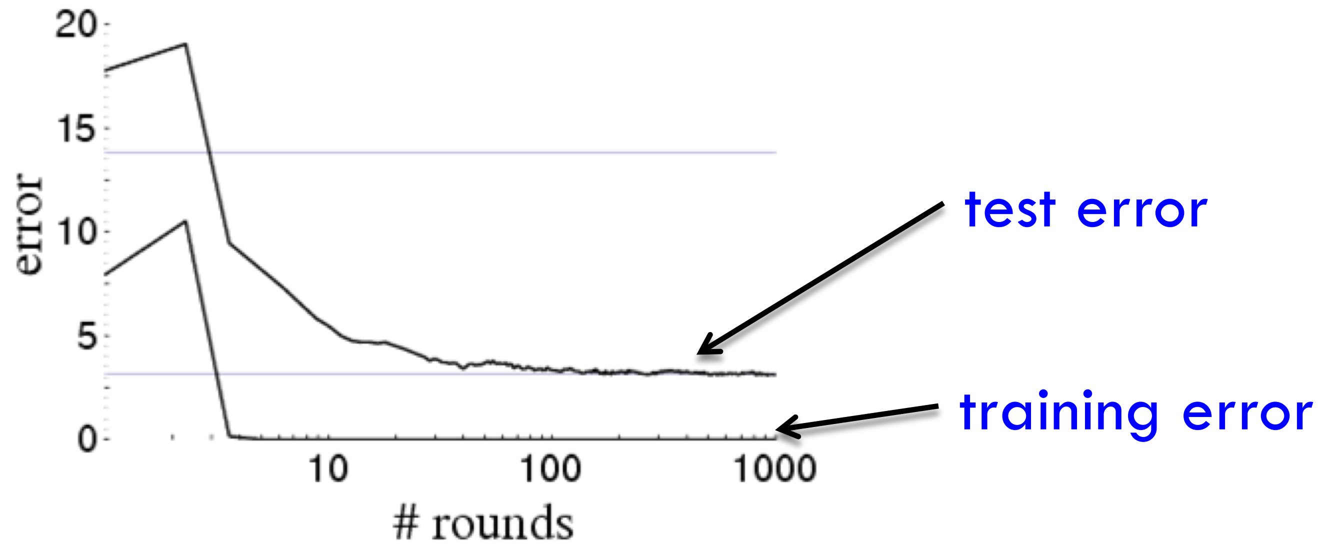
Reinforcement Learning

- Make a sequence of actions, given current states
  - e.g. a robot interacting with its environment
  - Maximize the reward
    - at some point, receive a reward or a punishment.
    - actions may also affect future reward.



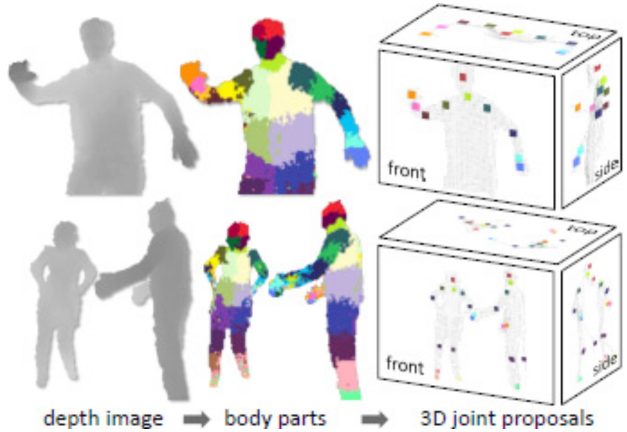
Learning Theory

- Why does machine learning work?
  - performance guarantees – bounds on the expected test error.
  - What types of functions can be represent by an algorithm, and how much data do we need?

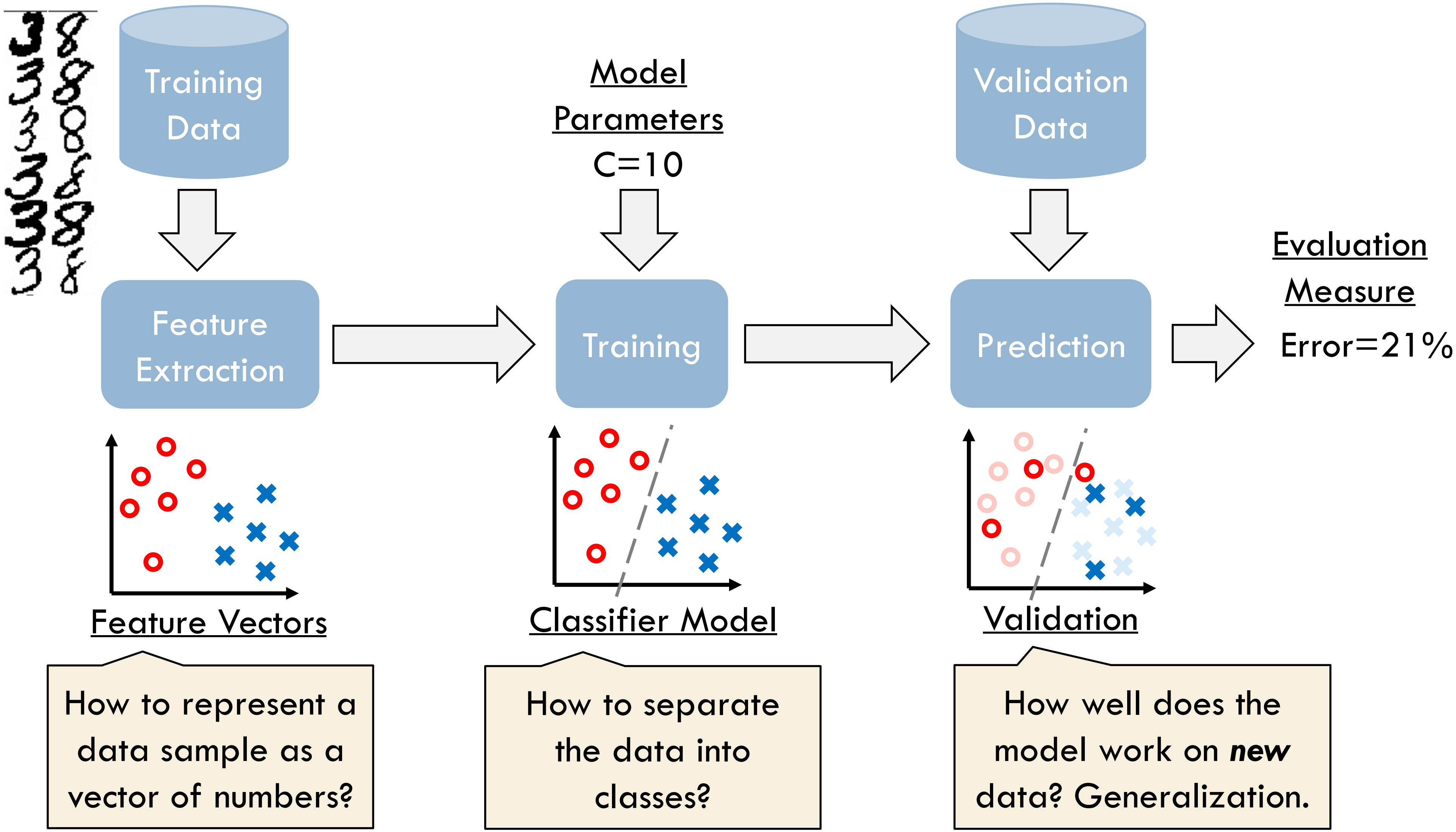


ML in the Real World

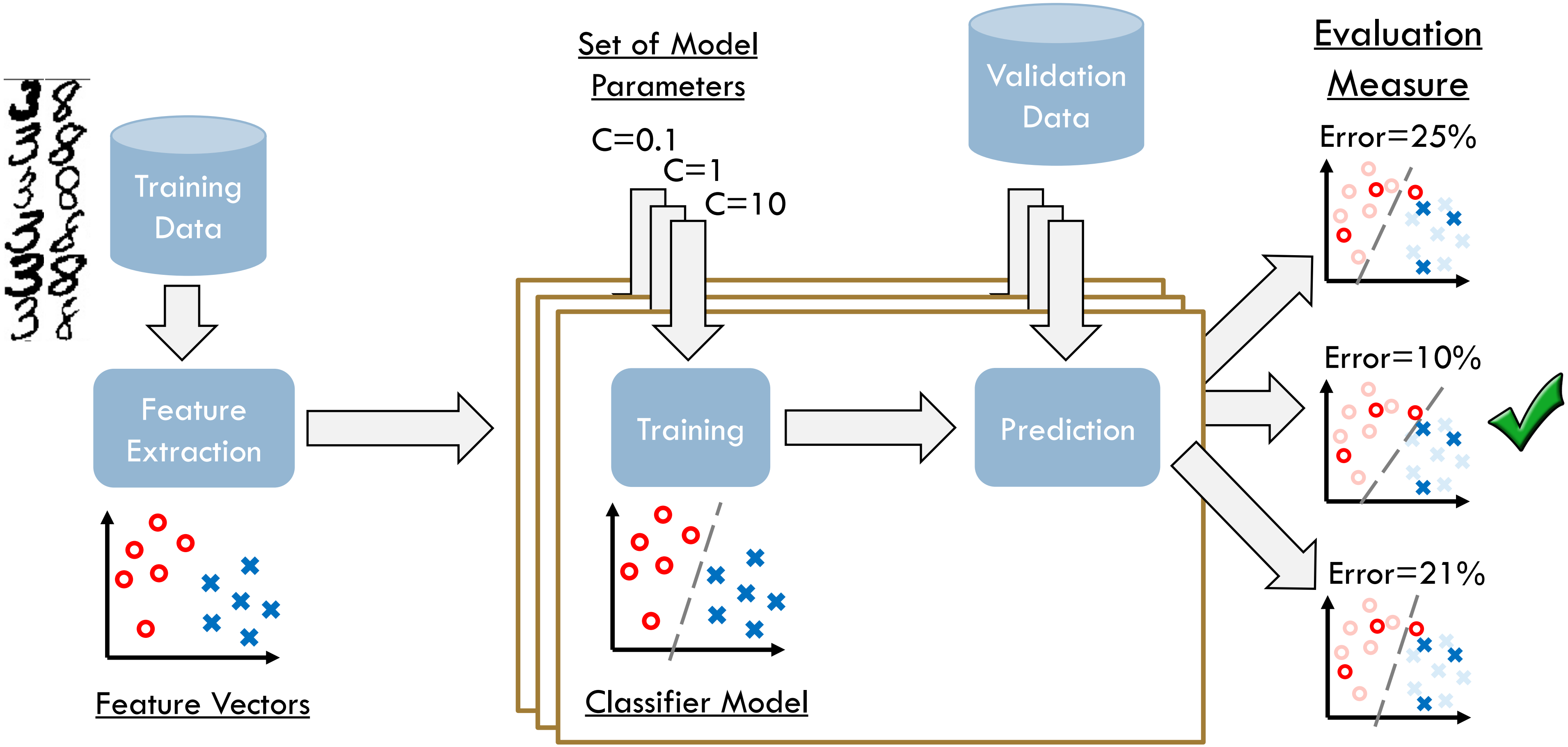
- Google
  - spam email classifier, speech recognition, machine translation, image annotation, AlphaGo
- Face detection & recognition
  - digital cameras, Google street view, Facebook
- Business
  - credit card fraud detection
  - stock trading (portfolio optimization)
- Recommendation systems
  - Netflix, Amazon
- Human pose recognition (Kinect)
- Controllers (reinforcement learning)
- Self-driving Cars



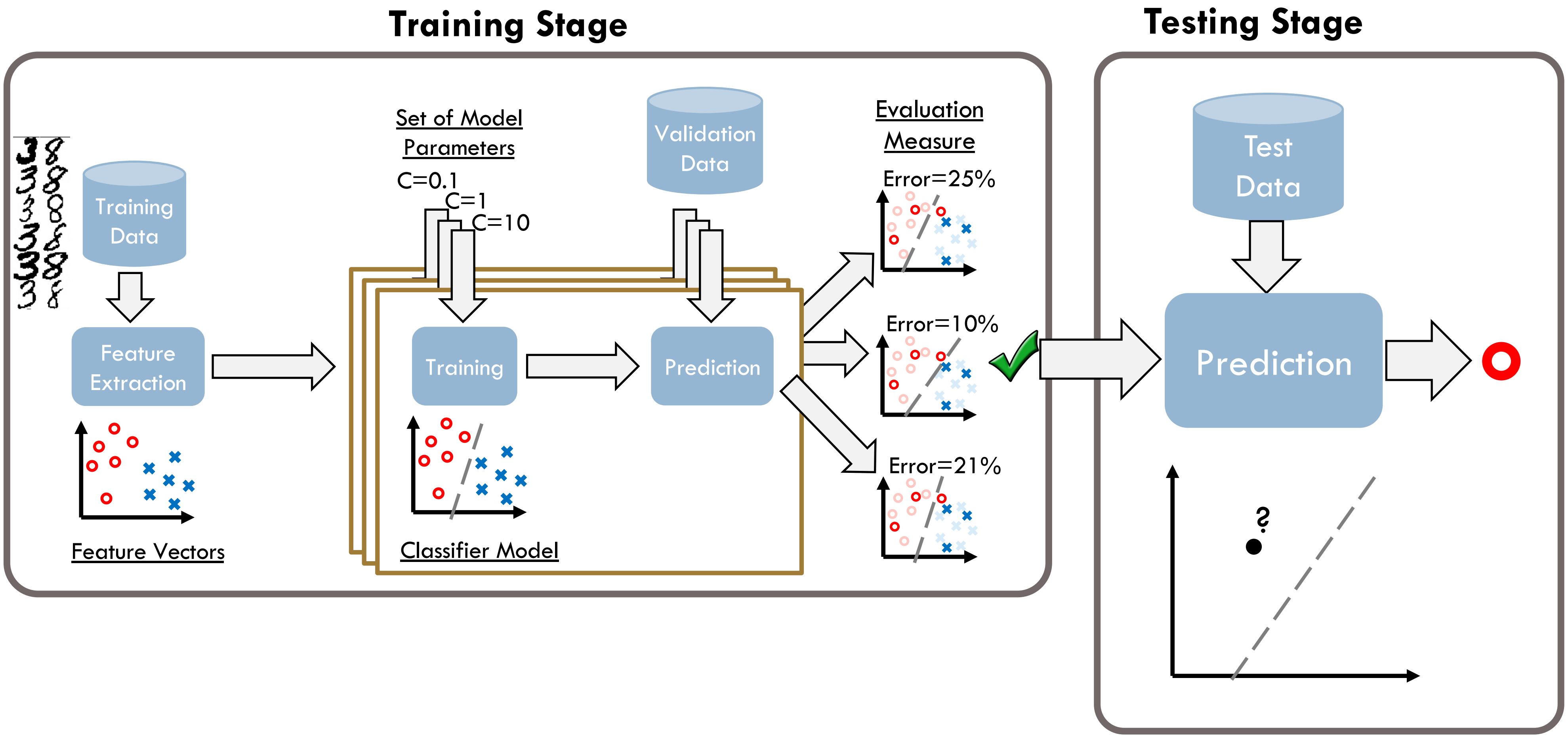
ML Training Pipeline



Model Selection Pipeline

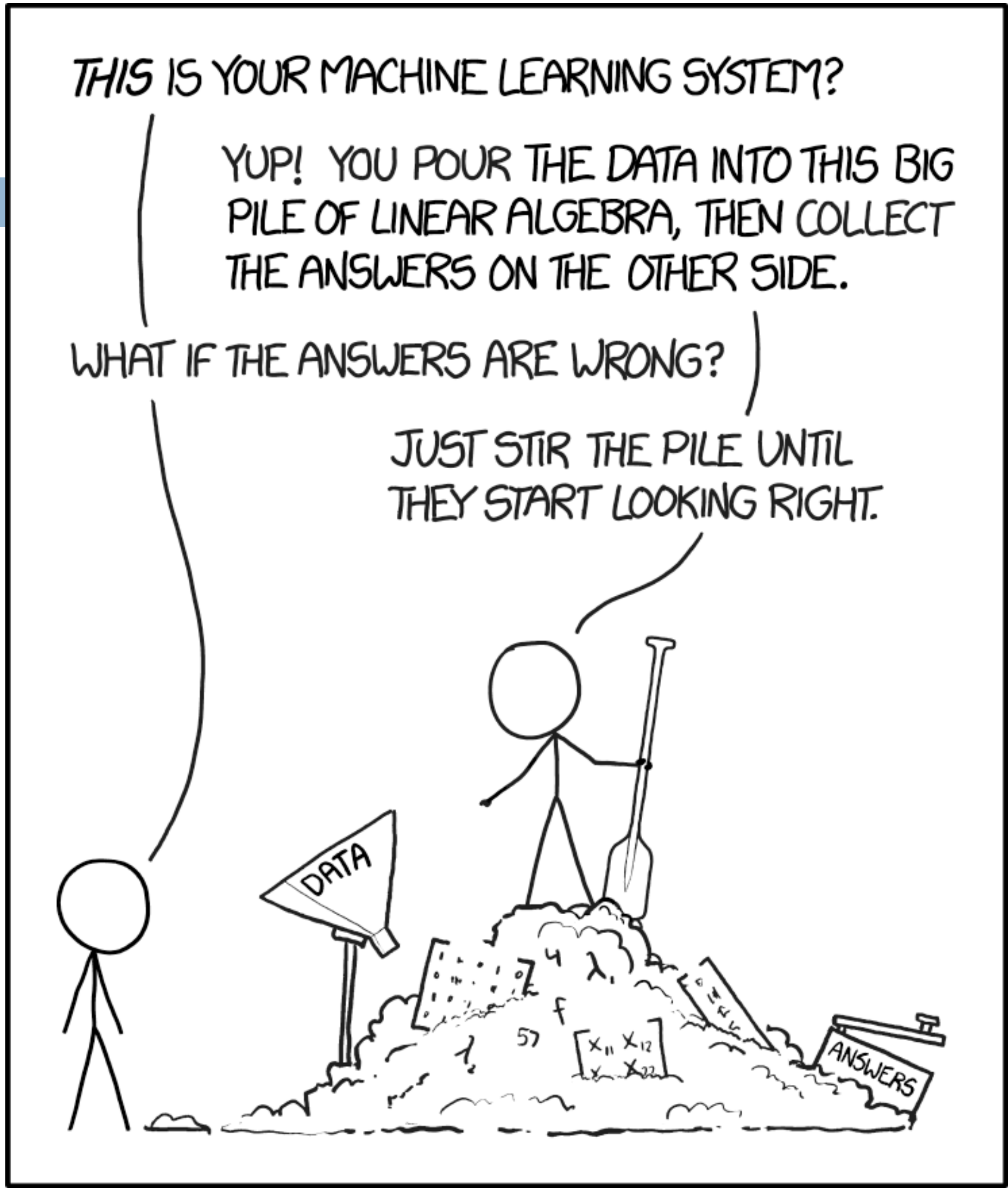


Training-Testing Stages



Building Blocks for ML

- What are the tools needed for Machine Learning?
  - *linear algebra*
    - matrices, inverse, eigenvector, SVD, …
  - *probability & statistics*
    - random variables, expectation, Bayes' theorem, …
  - *optimization algorithms*
- Don't worry, we will review these as necessary.
  - we need these tools to understand *how* ML works.
  - the algorithms we will use are already implemented in the ML libraries.

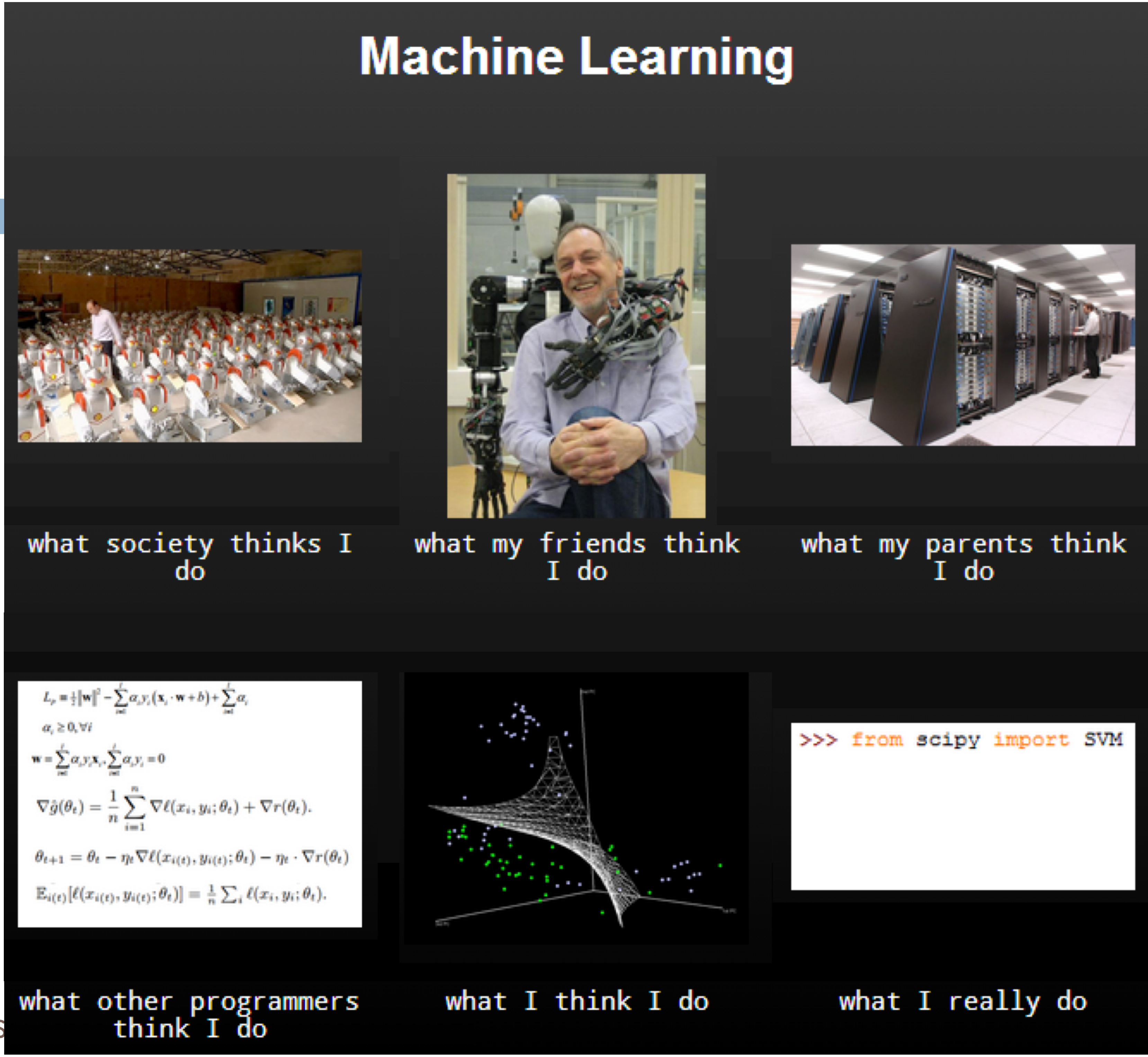


https://xkcd.com/1838/

Wk	Lecture Topic (2018)	Tutorial	Assessment
1	Lec 1: Introduction / Python	Tut 1	
	Supervised Learning (Weeks 2-6)		
2	Lec 2: Probabilistic Models & Bayes Classifiers	Tut 2	
3	Lec 3: Discriminative Classifiers (LR & SVM)	Tut 3	A1 out
4	Lec 4: Nonlinear Classifiers (KSVM, AdaBoost & RF)	Tut 4	
5	*Holiday (National Day)*		
6	Lec 5: Regression	Tut 5	A1 due, A2 out
	Unsupervised Learning (Weeks 7-9)		
7	Lec 6: Density-based Clustering	Tut 6	
8	Lec 7: Linear Dimensionality Reduction	Tut 7	
9	Lec 8: Non-Linear Dim. Reduction & Manifold Embedding	Tut 8	A2 due, Pr. out
	Other Topics (Weeks 10-11)		
10	Lec 9: Neural Networks & Deep Learning	Tut 9	
11	Project meetings	Pr. meet.	
12	Lec 10: More Deep Learning, Distributed Computing	Tut 10	
13	Project presentations	Pr. Pres.	Pr. Due

Academic Honesty

- CityU has *Rules of Academic Honesty* and has required all students to complete an online tutorial on subject and declare your understanding
- Plagiarism…
  - It is serious fraud to plagiarize others' work.
  - Punishment ranges from warning to course failure.
- How to prevent plagiarism...
  - Finish the assignments by yourself! You have to write the program/solution yourself.
    - okay to talk about *how* to do the problem with your classmates.
    - **Protect your code**; don't give it away as a "reference" copy.
  - In plagiarism cases, we treat the giver and the copier as both guilty.
  - You hurt your own grades by not reporting cheating.
- As instructors...
  - We have responsibility to report academic dishonesty cases so as not to compromise the quality of education.
  - We take suspected plagiarism cases very seriously.


Machine Learning
what society thinks I do
what my friends think I do
what my parents think I do
>>> from scipy import SVM
what other programmers think I do
what I think I do
what I really do