BT4222

Mining Business Insights with Web Data

Phase 1 -- Machine Learning Fundamental

Topic 1 – **Overview about Machine Learning and Natural Language Processing**

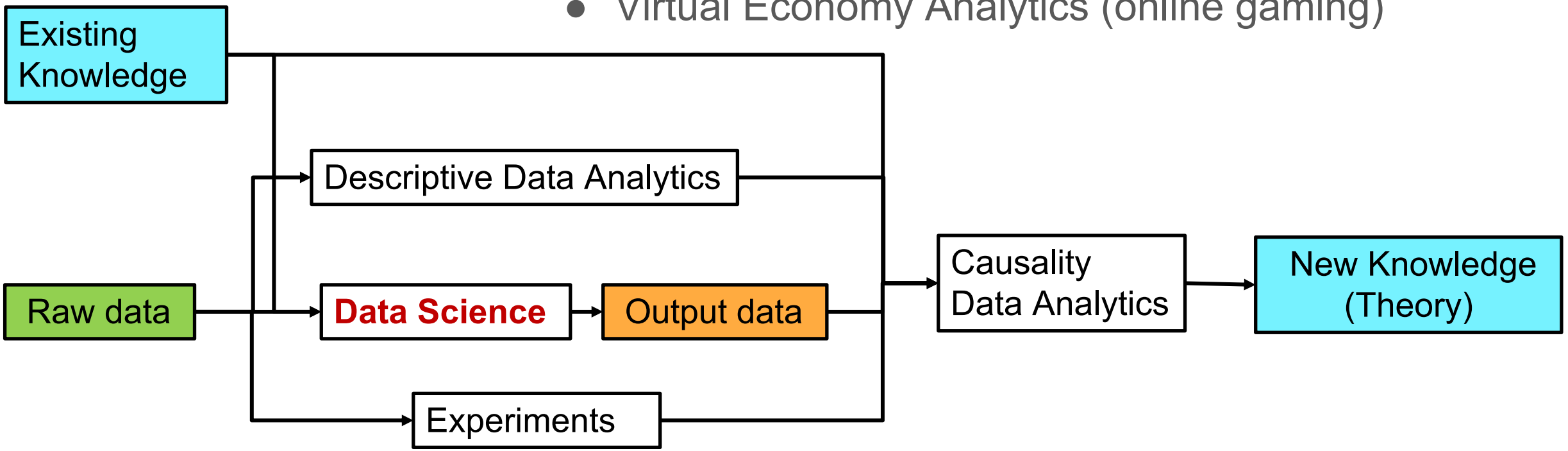
Dr. Qiuhong Wang

Term1 2022-23

Lecturer: Wang Qiuhong

Research Experience

- Cybersecurity Analytics (Attacks, Hacker forums, Internet topology, Newspaper reports, Legal documents)
- Healthcare Analytics (Hospital information systems, Online Healthcare platform)
- Virtual Economy Analytics (online gaming)



Teaching Assistant: Yu Ta

- Ph.D. student (2020 Aug intake) in Information Systems (current)
- Master in MIS, National Chengchi University, Taiwan
  - Decision and Quantitative Analysis Lab
  - Machine Learning - Recommendation system
- Office: IS Research Lab 2 [COM2-01-03]
- E0546019@u.nus.edu
- https://www.linkedin.com/in/yutanccu/

Teaching Assistant: WANG Yuchen

- Ph.D. student (2020 Aug intake) in Information Systems & Analytics
- Bachelor in IMIS, Peking University, China
- Office: IS Research Lab 2 [COM2-01-03]
- yuchen.wang@u.nus.edu
- https://www.linkedin.com/in/yuchen-wang-889bb5120/

Teaching Assistant: Zhang Xinyi

- Ph.D. student (2020 Aug intake) in Information Systems & Analytics
- Bachelor in Financial management, SCUT
- Master in Business Analytics, HKU
- Office: IS Research Lab 1 [COM2-01-02]
- xinyizhang@u.nus.edu
- https://www.linkedin.com/in/xinyi-zhang-8324b4176/

Course Introduction

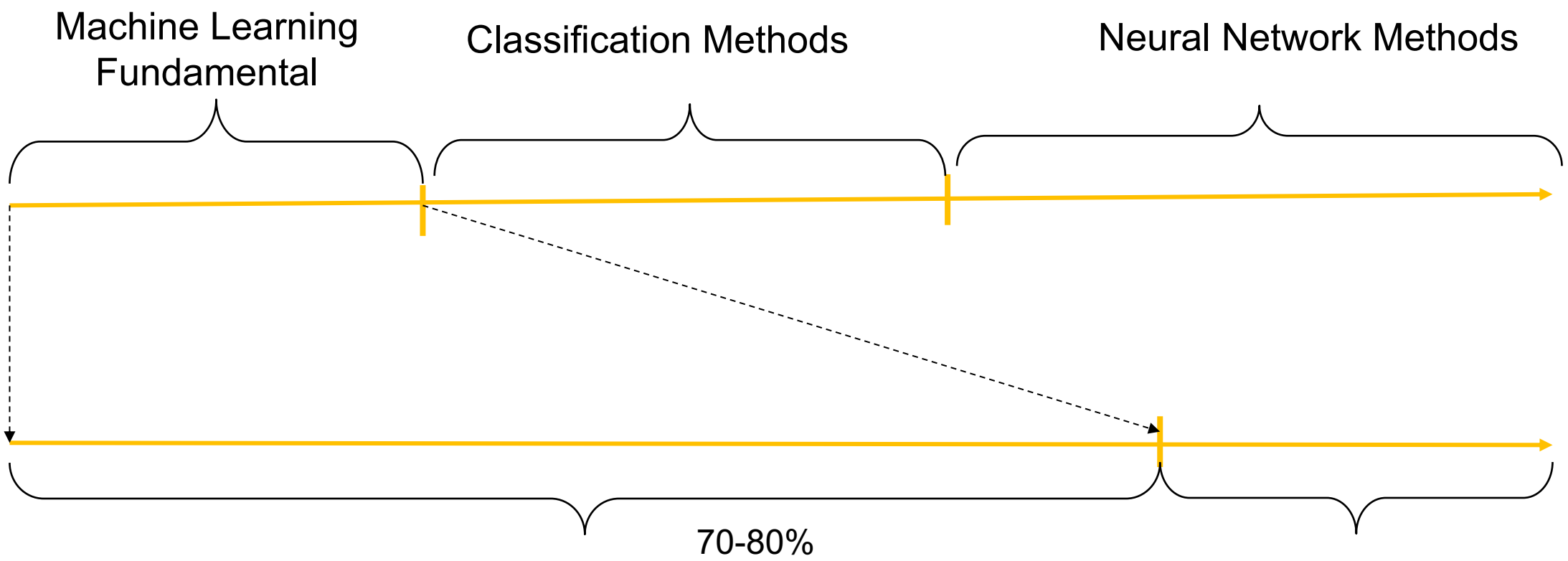
- BT4222
  - Apply ML methods on data analytics with Python
  - Understand the Intuition of ML algorithm
  - It is NOT about --
    - Python programming
    - Mathematics in ML algorithm

- How to learn
  - Class Preparation
  - In-class Participation
  - After-class Practice

Class #	Topic	Assignment & Quizzes
	Phase 1: Machine Learning Fundamental	
Week 1	• Machine Learning (ML) Introduction • Web Scraping	
Week 2	• How to Frame ML Problem? • Feature Engineering	
Week 3	• Regularization • How to Run ML Project?	• Assignment 1 (10%) • Project group formation
Week 4	• Get Start with Natural Language Processing (NLP)	
	Phase 2: Classification Methods	
Week 5	• Linear regression / Logistic regression	
Week 6	• Bayesian learning / Support vector machines	
Week 7	• Decision tree • Ensemble learning, Random forests	Assignment 2 (10%) Quiz 1 (10%)
	Phase 3: Neutral Network Methods	
Week 8	• Neutral network and deep learning	
Week 9	• Word Embedding • Project proposal presentation	Project proposal submission and presentation (5%)
Week 10	• Convolutional neutral network	Assignment 3 (10%)
Week 11	• Recurrent neutral network and other types of DNN and explainability	Quiz 2 (10%)
Week 12	• Deep learning applications in NLP	
Week 13	• Project presentation	Project deliverables submission (40%)

Attendance
Class participation
5%

Learning and Project Timeline



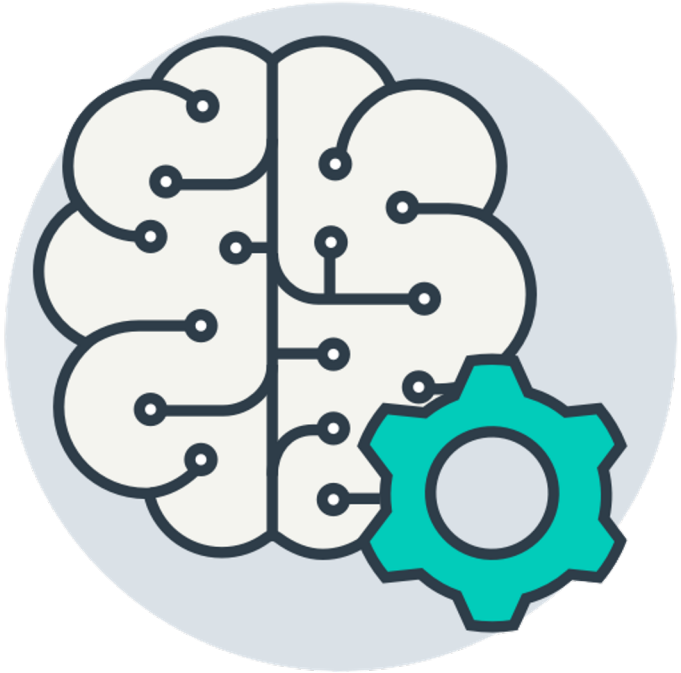
- Acquire Data
- Frame ML Problem
- Feature Engineering
- Write Report

- Model Selection and Tuning

Agenda

- Overview about Machine Learning and Natural Language Processing
- How to Frame ML Problem
- Important concepts in training machine learning models
  - The input: Feature engineering
  - The output: Regularizations
- How to Run a ML Project
- Web scraping

What is ML?



What is Human Intelligence?

Information

What to eat for lunch?

What to dress?

Image information

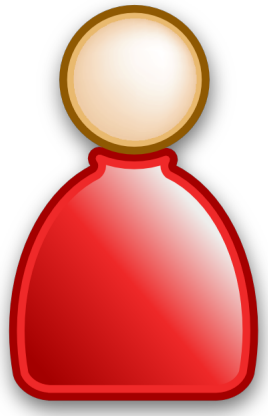
What is this picture?

Image information

What is this number?

SIA 5.52 SGD
On April 22, 2022

How much can I earn from investing in SIA?



CAT

2

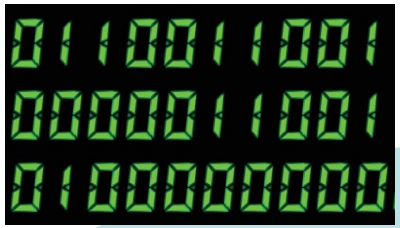
? $

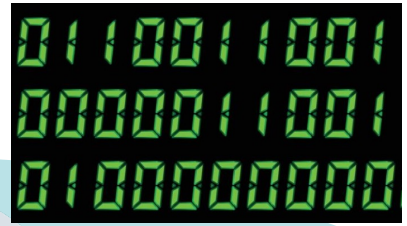
Machine Learning

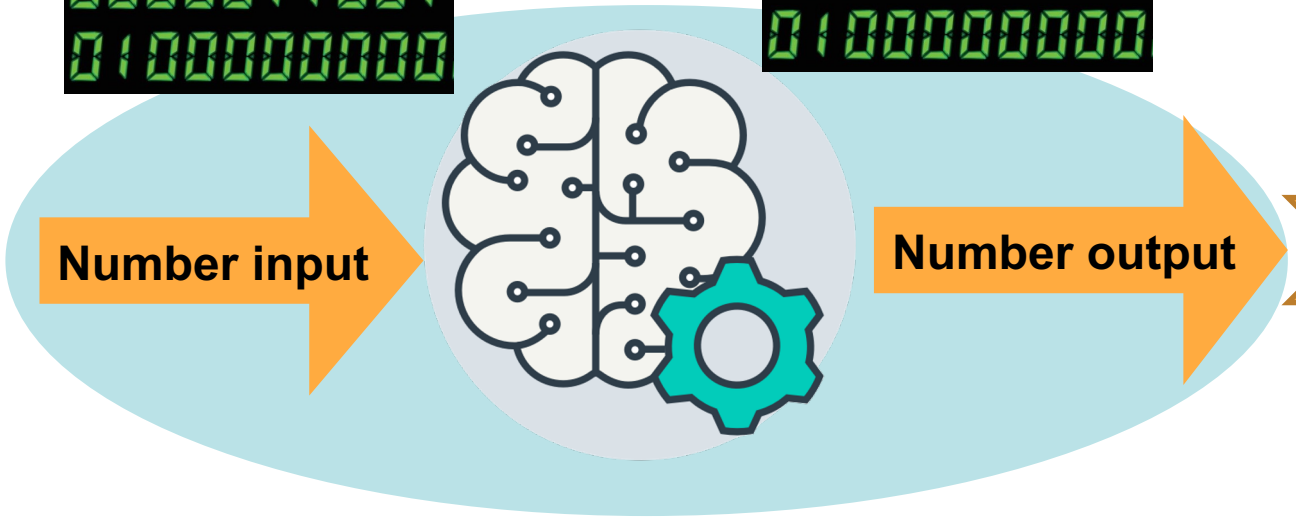
Information

Image information

Image information







Trained model

CAT

2

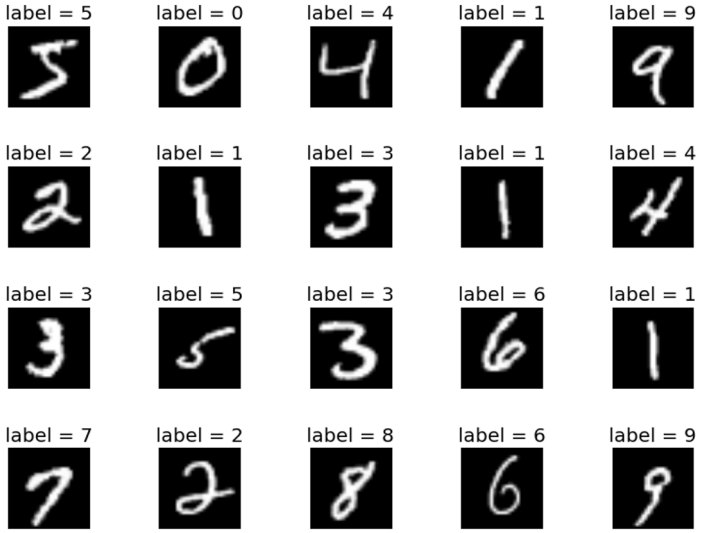
? $

SIA 5.52 SGD
On April 22, 2022

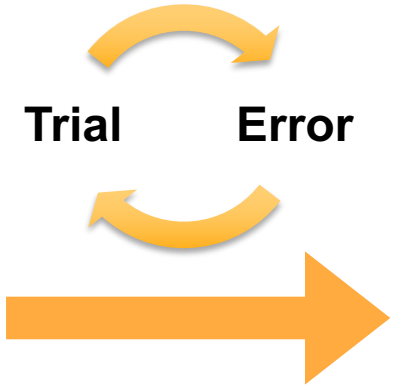
Machine Learning

Machine needs lots of training

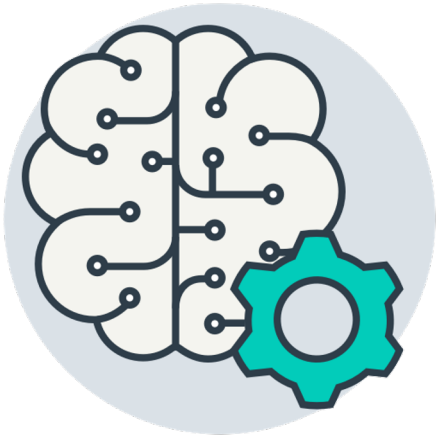
ML systems learn how to combine input to produce useful predictions on never-before-seen data.



Labeled dataset



training

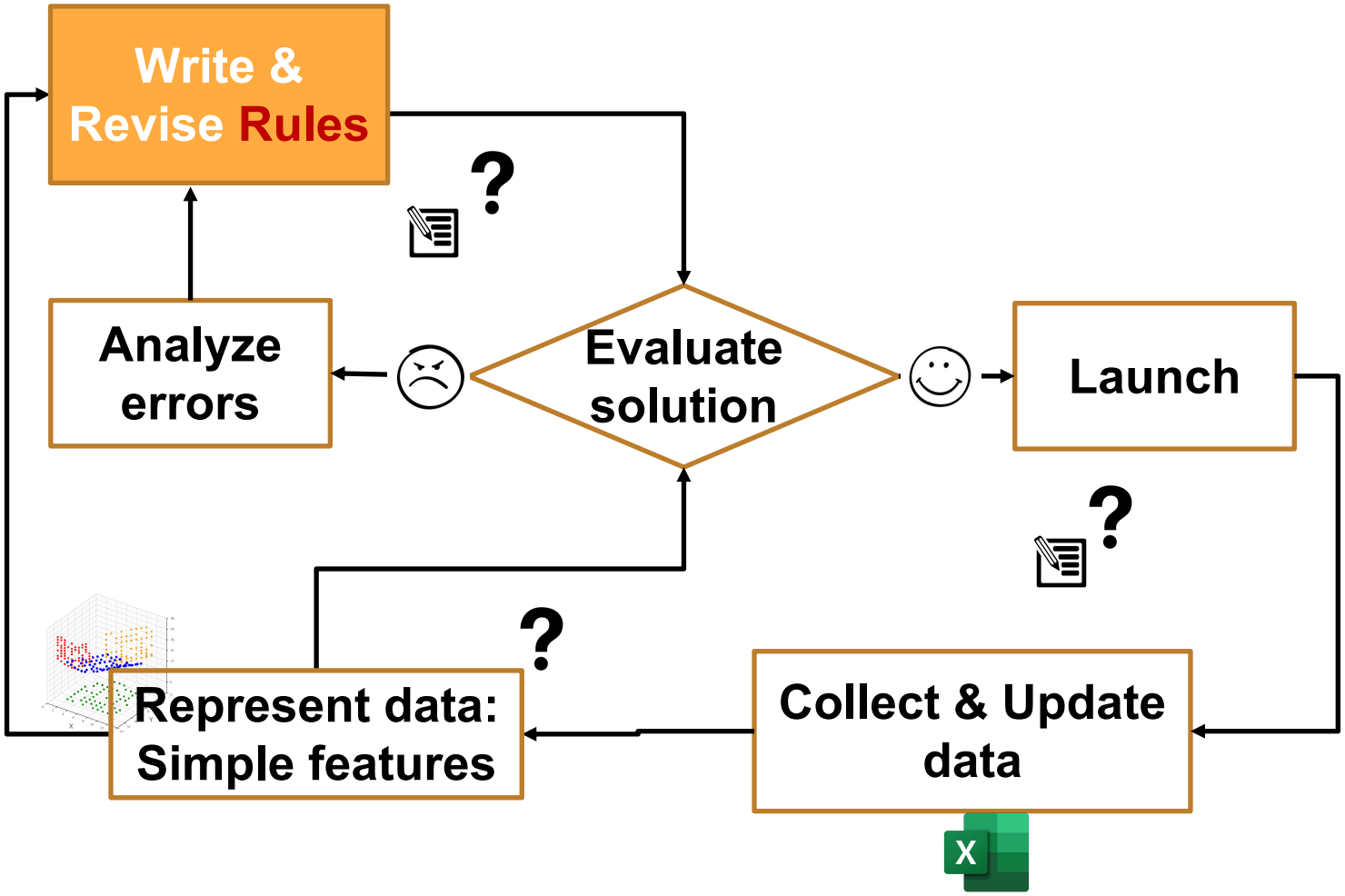


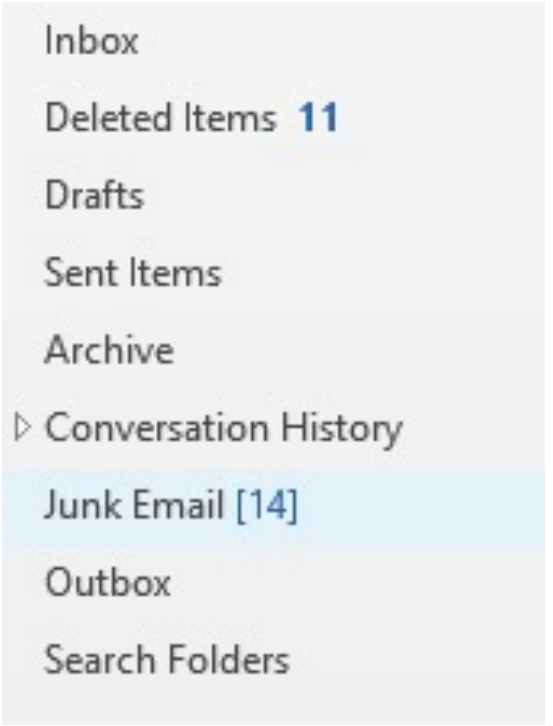
Model

Inference means applying the trained model to unlabeled examples and make useful predictions (y')

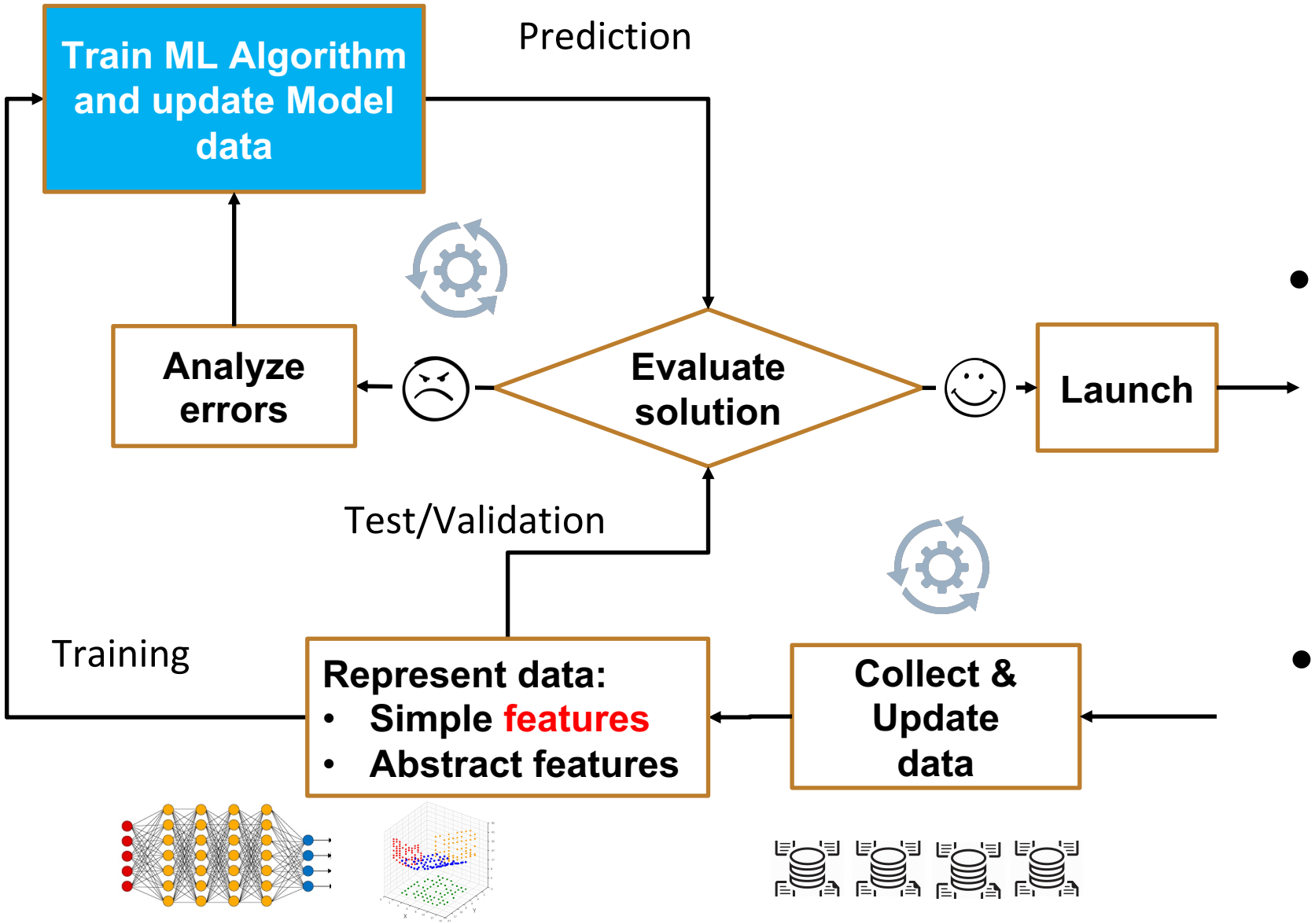
During **training**, the model is fed with labeled examples and gradually learns the mathematical relationships between features and label.

Why Use Machine Learning?





Why Use Machin Learning?



- ML can automatically keep learning from the large scale of new data and update their algorithms
- Machine learning can work for problems that either are too complex for traditional approaches or have no known algorithm.
- ML algorithms can be inspected to see what they have learned, sometimes.

Machine Learning

Model to detect spam

Subject	Content	Sender	Date	Spam
xxx	yyyy	aa@hotmail.com	9/11/22	0
xxx	yyyy	gg@maldas.net	8/11/22	1
xxxx	yyyy	xx@mail.com	9/11/22	1
xxxxx	yyyyyyy	yy@hk.edu.hk	10/11/22	0
xxxxxx	yyyyyyy	wangdt@nus.edu.sg	11/11/22	0
xxx	yyyyyyyy	tanll@mas.gov.sg	12/11/22	0

- A **classification** model predicts discrete values.
- A **regression** model predicts continuous values

A **feature** is an input variable, $x_1, x_2, \ldots x_N$

A **label** is the thing we're predicting, y

Model to predict housing price

District	Floor Size (sqft)	Tenure	Top	Price (SGD)
Woodlands	1052	99-year Leasehold	2000	900,500
Woodlands	998	99-year Leasehold	2000	820,000
Toa Payoh	1119	99-year Leasehold	2010	1,560,800
Toa Payoh	1256	99-year Leasehold	2010	1,890,000
China Town	870	Free Hold	2017	1,400,000
China Town	1100	Free Hold	2017	1,980,000

labeled examples
{features, label}: $(x_1, x_2, \ldots x_N, y)$

unlabeled examples
{features, ?}: $(x_1, x_2, \ldots x_N, ?)$

Machine Learning: Algorithm vs. Model

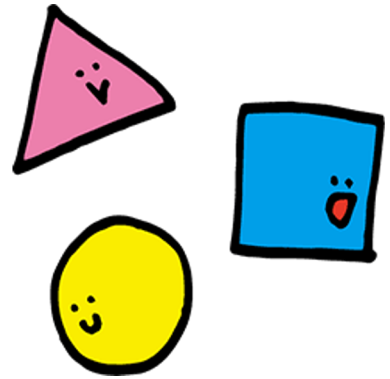
- Algorithm: Procedures that are implemented in code and are run on data to create a machine learning model
- Model: Outputs by algorithms and are comprised of model data and a prediction algorithm
  - Model data: the rules, numbers, and any other algorithm-specific data structures required to make predictions.

Algorithm	Model
Linear regression	a model comprised of a vector of coefficients with specific values
Decision tree	a model comprised of a tree of if-then statements with specific values
Neural network	a model comprised of a graph structure with vectors or matrices of weights with specific values.

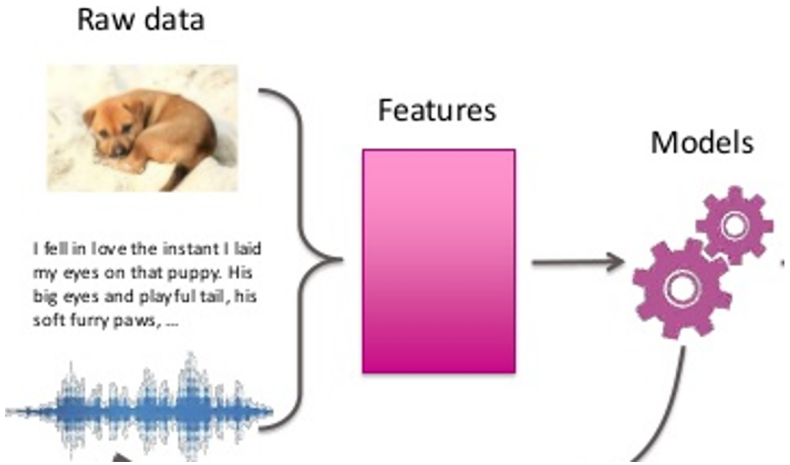
Machine Learning

Representation → Evaluation → Optimization

Given a task: how to classify these following shapes:



- Our system should work as:
  - Input: Image
  - Representation: Number of corners.
  - Model: Fed with representation and based on mathematical models or rules to make prediction
- Designing features is a complex process, which require a deep domain expertise.
- Deep learning is the method which tries to learn features by the model itself.



ML: Representation → **Evaluation** → Optimization

Keras API reference / Losses

- **Empirical risk minimization (ERM)**

Choose the prediction model that minimizes loss on the training set.

✓ **Empirical**

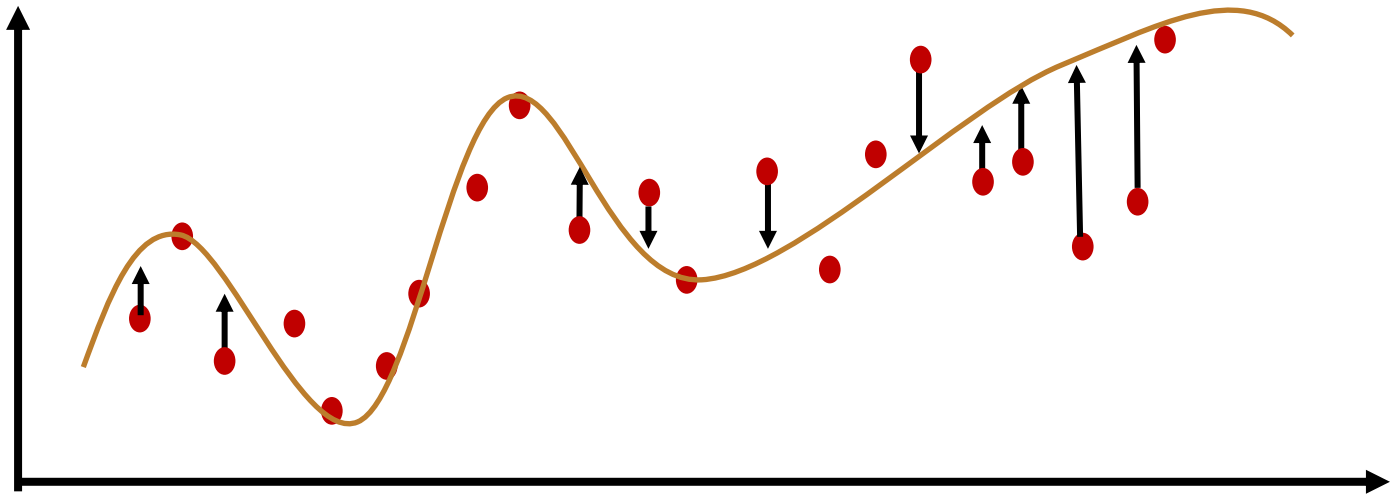
We do not know the true distribution ***P*** over (X, Y); We replace it by the training set that comes from ***P***

✓ **Risk**

It is measured by a Loss Function ***L*** that informs us how much it "hurts" to make the prediction $\hat{y} = f(x)$ when the true output is y.

✓ **Loss Function**

Loss function is to measure the difference between prediction and ground truth incorporating our preference.



ML: Representation → **Evaluation** → Optimization

- **Regression Losses**

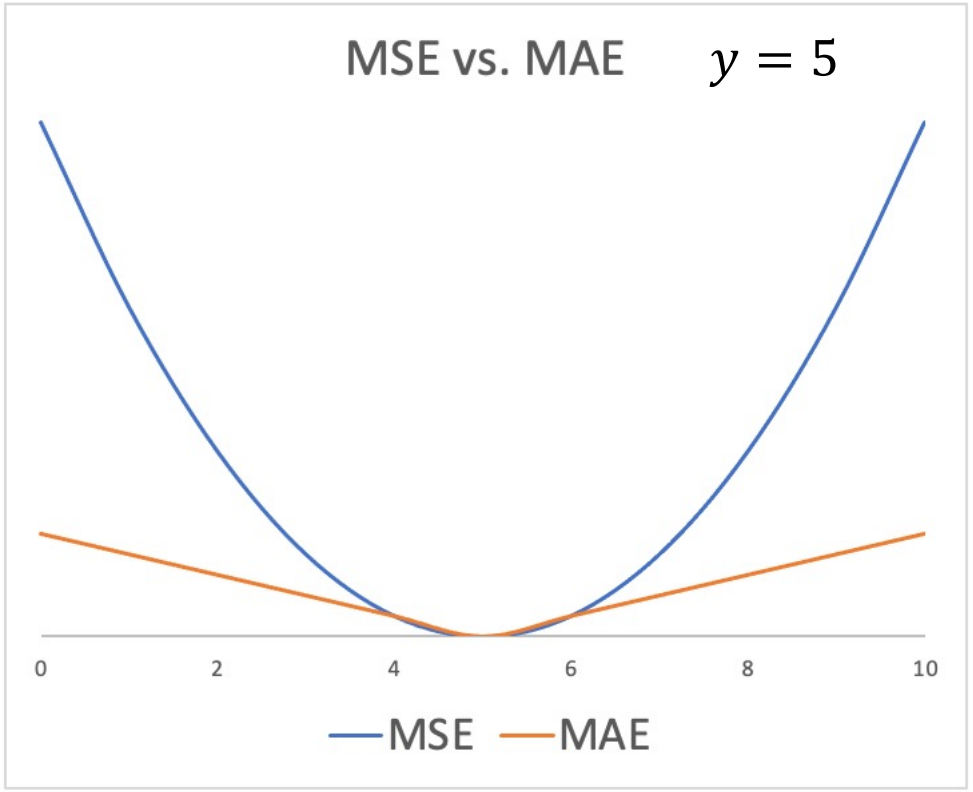
Mean Square Error/Quadratic Loss/L2 Loss

$$MSE = \frac{1}{N} \sum_{(x_i, y_i) \in D}^{n} (y_i - \hat{y}_i)^2$$

Mean Absolute Error/L1 Loss

$$MAE = \frac{1}{N} \sum_{(x_i, y_i) \in D}^{n} |y_i - \hat{y}_i|$$

What is the difference?



$MSE = (5 - \hat{y})^2$ $MAE = |5 - \hat{y}|$

Note: this is for a single instance, $N = 1$

ML: Representation → **Evaluation** → Optimization

- **Classification Losses**

Cross Entropy Loss

Multi-Class Classification, $K > 2$

$H = -\sum_{k=1}^{K} y_{ik} \log(\hat{y}_{ik})$, where

For a sample i,

K is the number of possible classes;

y_{ik} is the ground truth probability of class k, y_{ik}= 0 or 1;

$\hat{y}_{ik}$ is the predicted probability of class k

Binary Classification, $K = 2$

$H = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$

$Entropy\ of\ event\ E$

$h(E) = -\log_2 p(E), unit\ is\ bits$

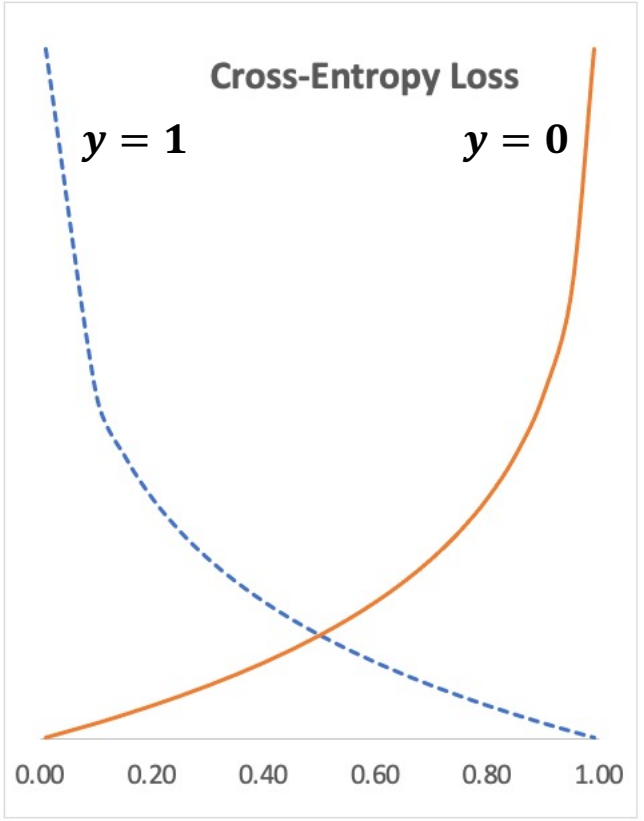
Training dataset *N*:

$$H = -\sum_{n=1}^{N} \sum_{k=1}^{K} y_{ik} \log(\hat{y}_{ik})$$

Cross Loss for Binary Classification: $H = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$

y_i=1, $H = -\log(\hat{y}_i)$

Actual probability	Predicted probability of being positive	Cross Entropy
1	0.01	6.64
1	0.10	3.32
1	0.15	2.74
1	0.20	2.32
1	0.25	2.00
1	0.30	1.74
1	0.35	1.51
1	0.40	1.32
1	0.45	1.15
1	0.50	1.00
1	0.55	0.86
1	0.60	0.74
1	0.65	0.62
1	0.70	0.51
1	0.75	0.42
1	0.80	0.32
1	0.85	0.23
1	0.90	0.15
1	0.95	0.07
1	0.99	0.01



y_i=0, $H = -(1 - y_i) \log(1 - \hat{y}_i)$

Actual probability	Predicted probability of being positive	Cross Entropy
0	0.01	0.01
0	0.10	0.15
0	0.15	0.23
0	0.20	0.32
0	0.25	0.42
0	0.30	0.51
0	0.35	0.62
0	0.40	0.74
0	0.45	0.86
0	0.50	1.00
0	0.55	1.15
0	0.60	1.32
0	0.65	1.51
0	0.70	1.74
0	0.75	2.00
0	0.80	2.32
0	0.85	2.74
0	0.90	3.32
0	0.95	4.32
0	0.99	6.64

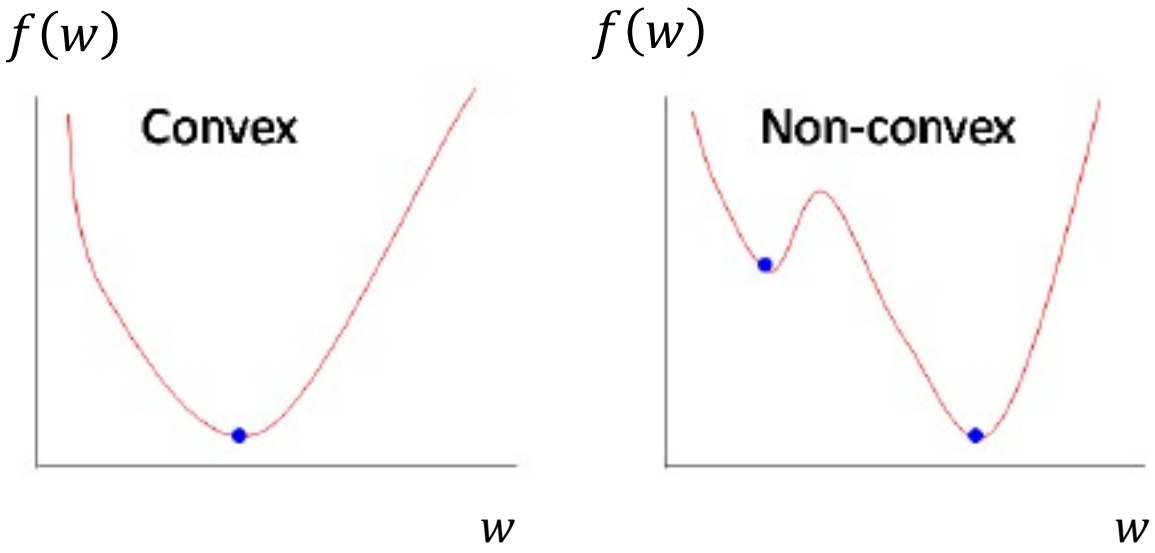
ML: Representation → Evaluation → **Optimization**

- **Optimization -- Gradient Descent**
  - To converge at the minimum loss, the loss function must be convex and differentiable
  - To minimize loss, the value of model parameter should move towards reducing loss

Given features X and Labels Y, and a linear model,

$$\hat{Y} = wX + b$$

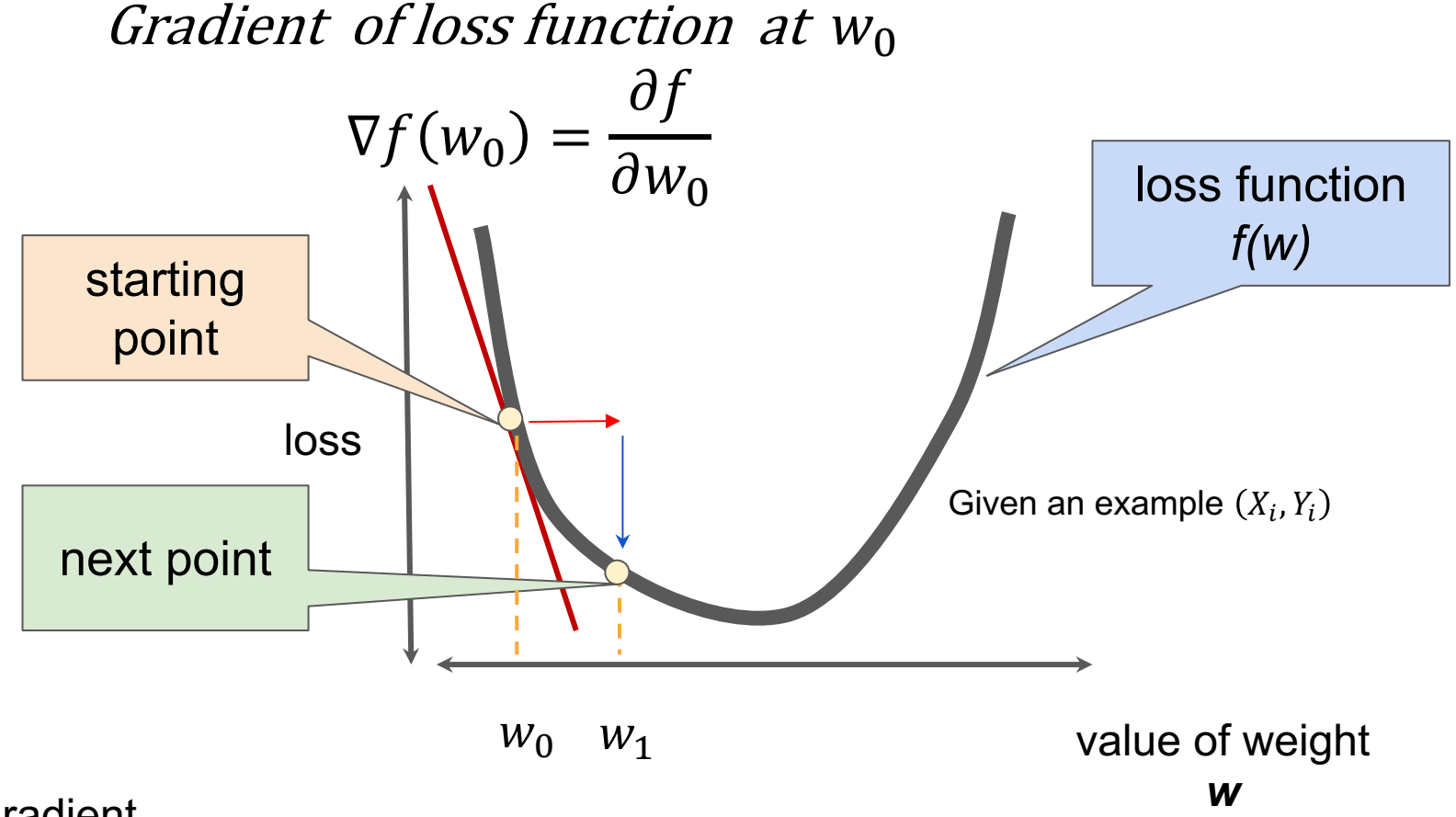
the loss function $f(w, b, X, Y)$ is determined by the coefficient w and bias b.



Assume *b is fixed*

ML: Representation → Evaluation → **Optimization**

- **Optimization -- Gradient Descent**
  - To converge at the minimum loss, the loss function must be convex and differentiable
  - To minimize loss, the value of model parameter should move towards reducing loss

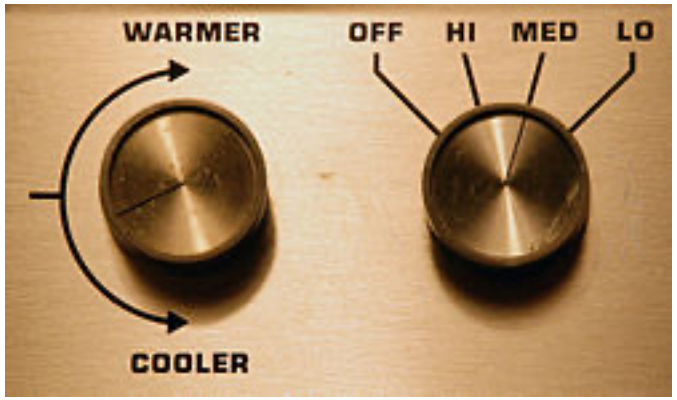


Adjust *w* in the reverse direction of gradient

$$w_1 = w_0 - \eta \nabla f(w_0)$$

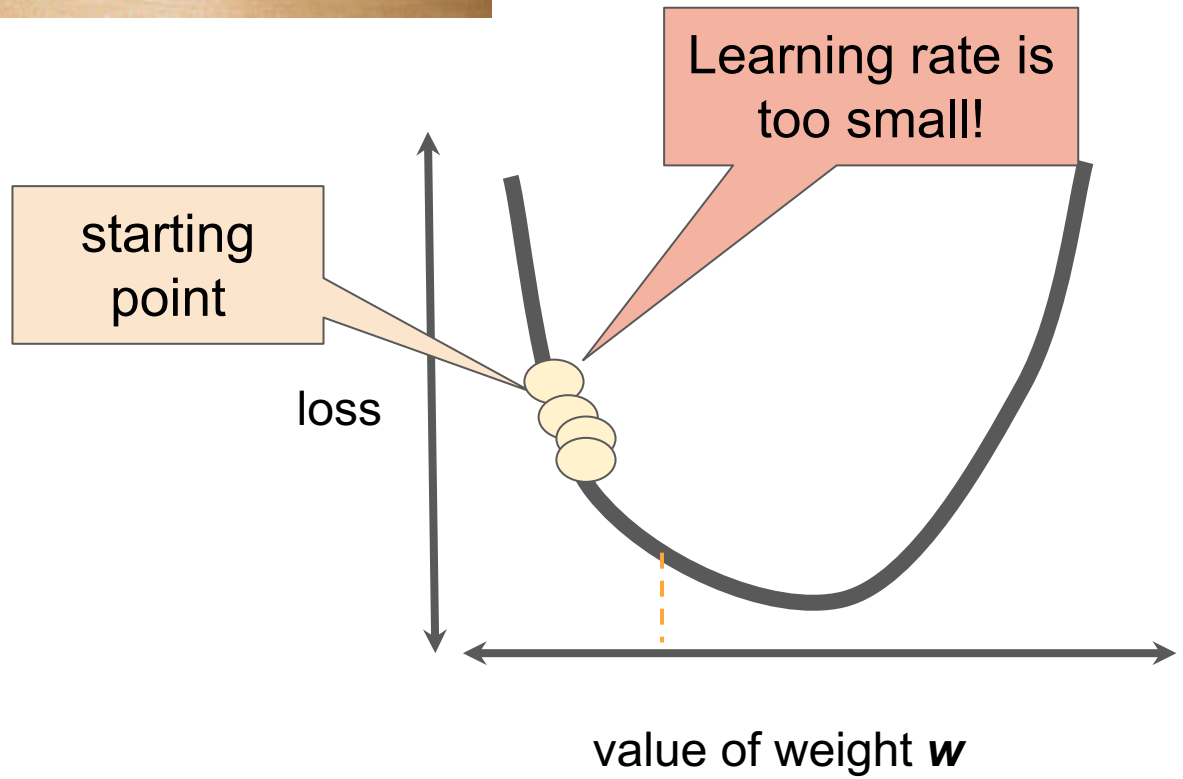
Learning rate: η

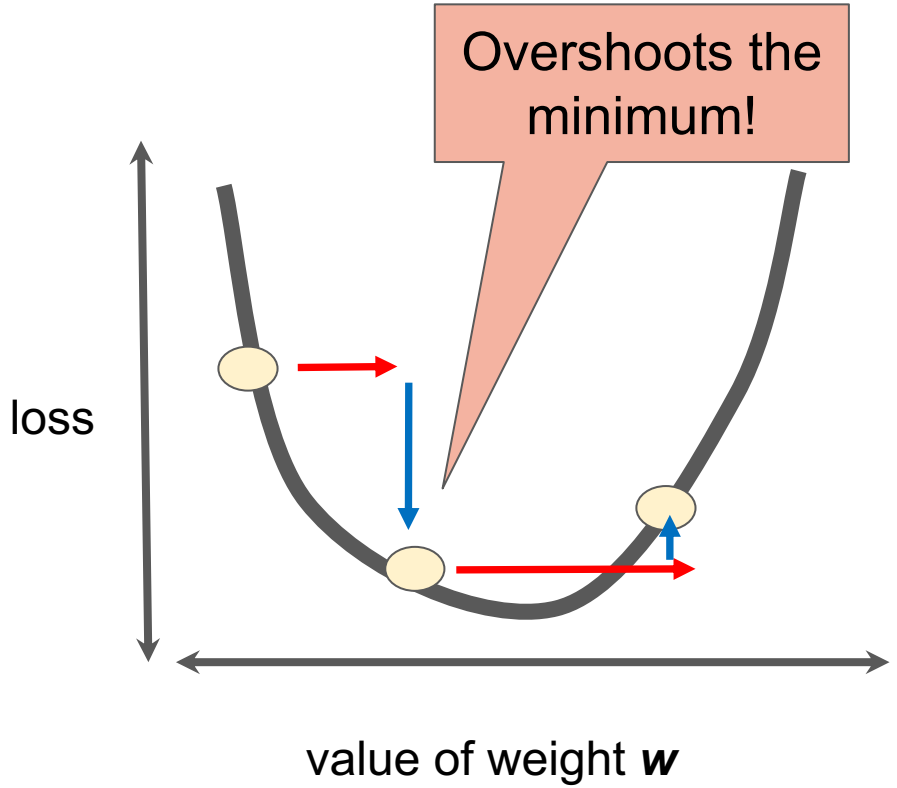
ML: Representation → Evaluation → **Optimization**

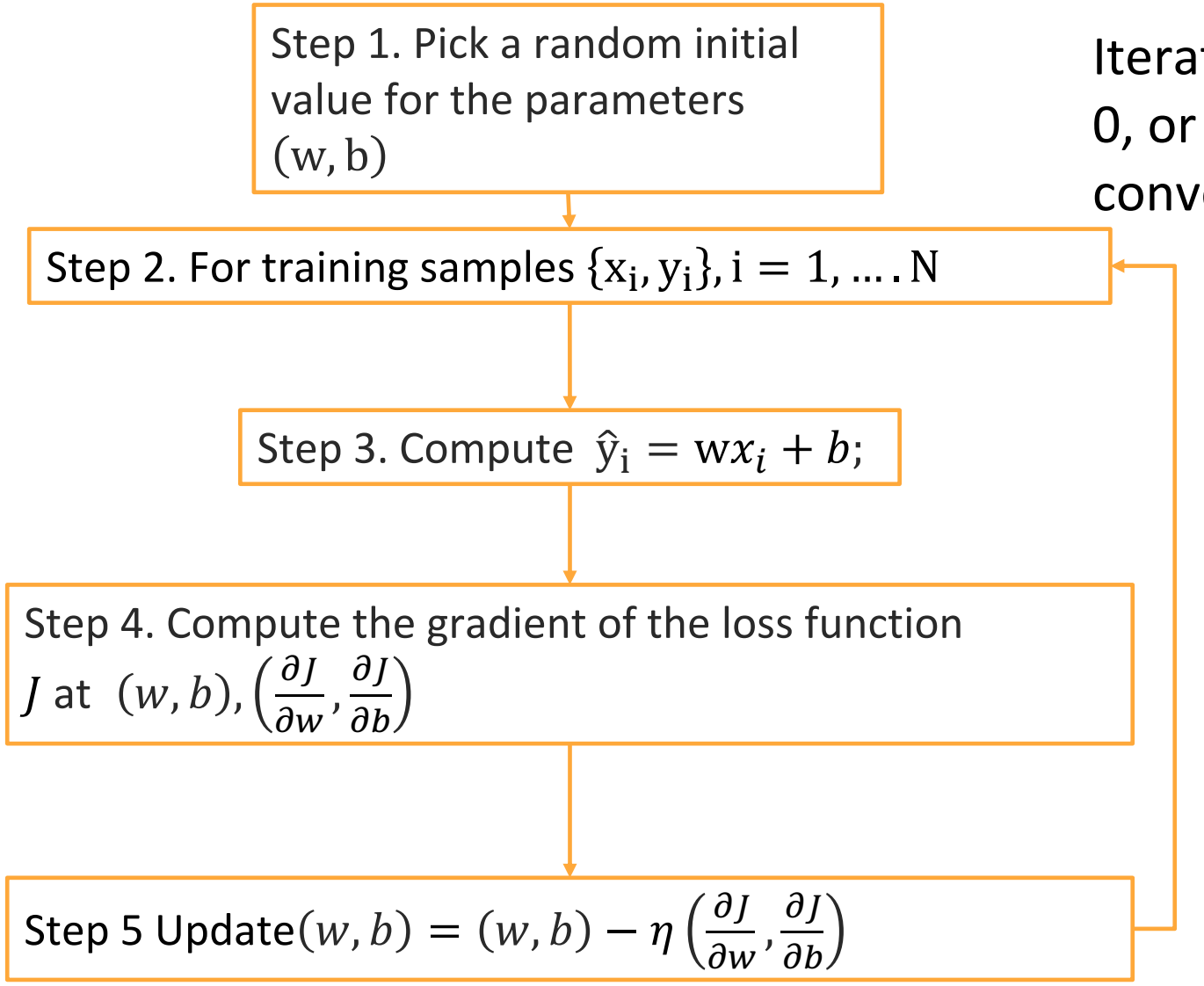


- **Hyperparameter**
  - Learning Rate

Playground







Iterate till gradient is almost 0, or is less than a convergence threshold

$$\hat{y}_i = wx_i + b$$

$$J = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\frac{\partial J}{\partial \hat{y}_i} = -\frac{2}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J}{\partial w} = \frac{\partial J}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial w} = -\frac{2}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i$$

$$\frac{\partial J}{\partial b} == \frac{\partial J}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial b} = -\frac{2}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

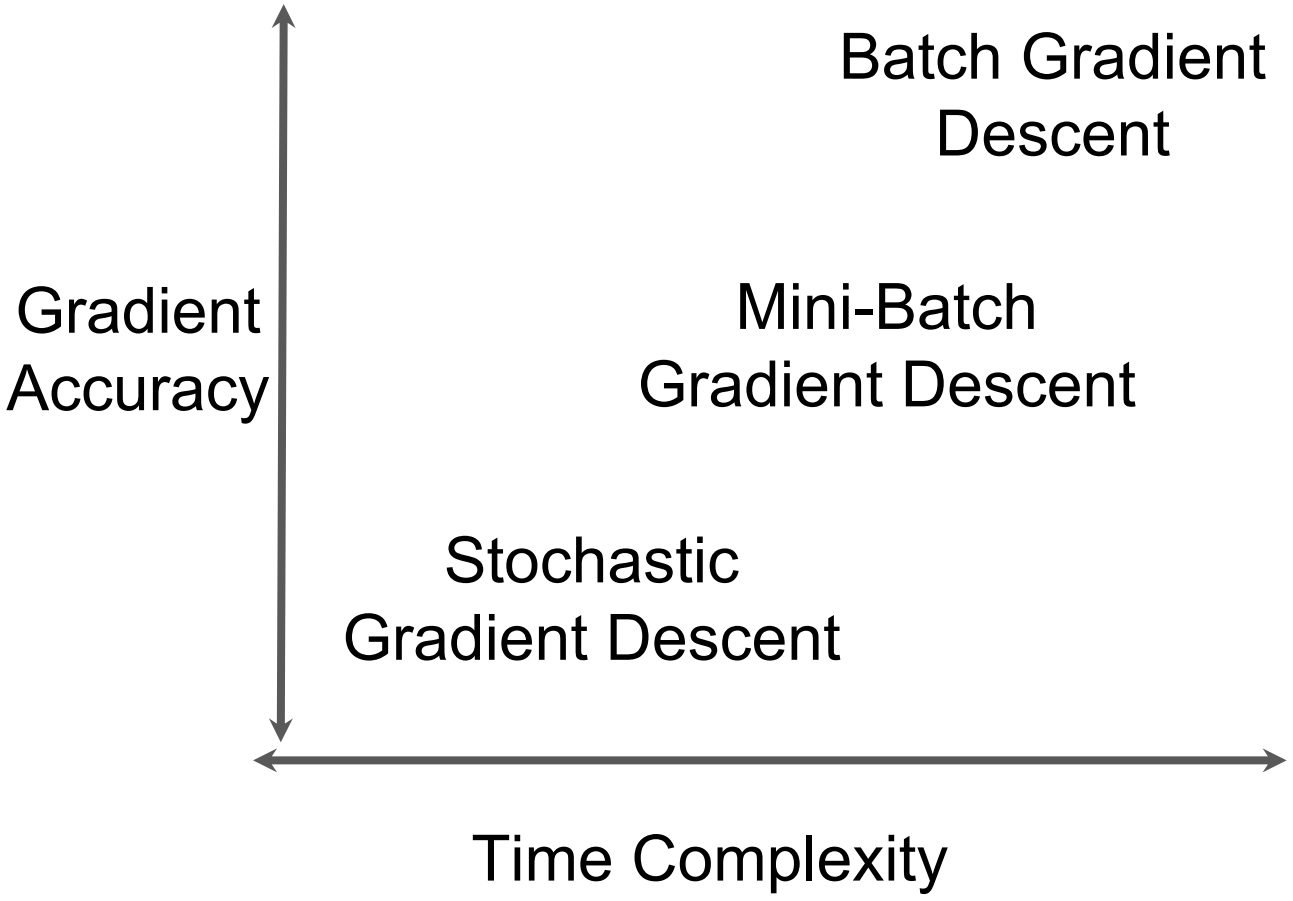
How much is the amount of computation?

- We have 10,000 data points and 20 features.
- Suppose the sum of squared residuals consist of just 10,000 terms

- We need to do 10000 predictions for any updated value of each of the 20 feature value, and repeat calculations in Step 4 and 5.
- Having 1000 iterations, the total computation is more than 10000*20*1000

ML: Representation → Evaluation → **Optimization**

- How to calculate gradient over sample data?



Sum up over all examples on each iteration when performing the updates to the parameters.

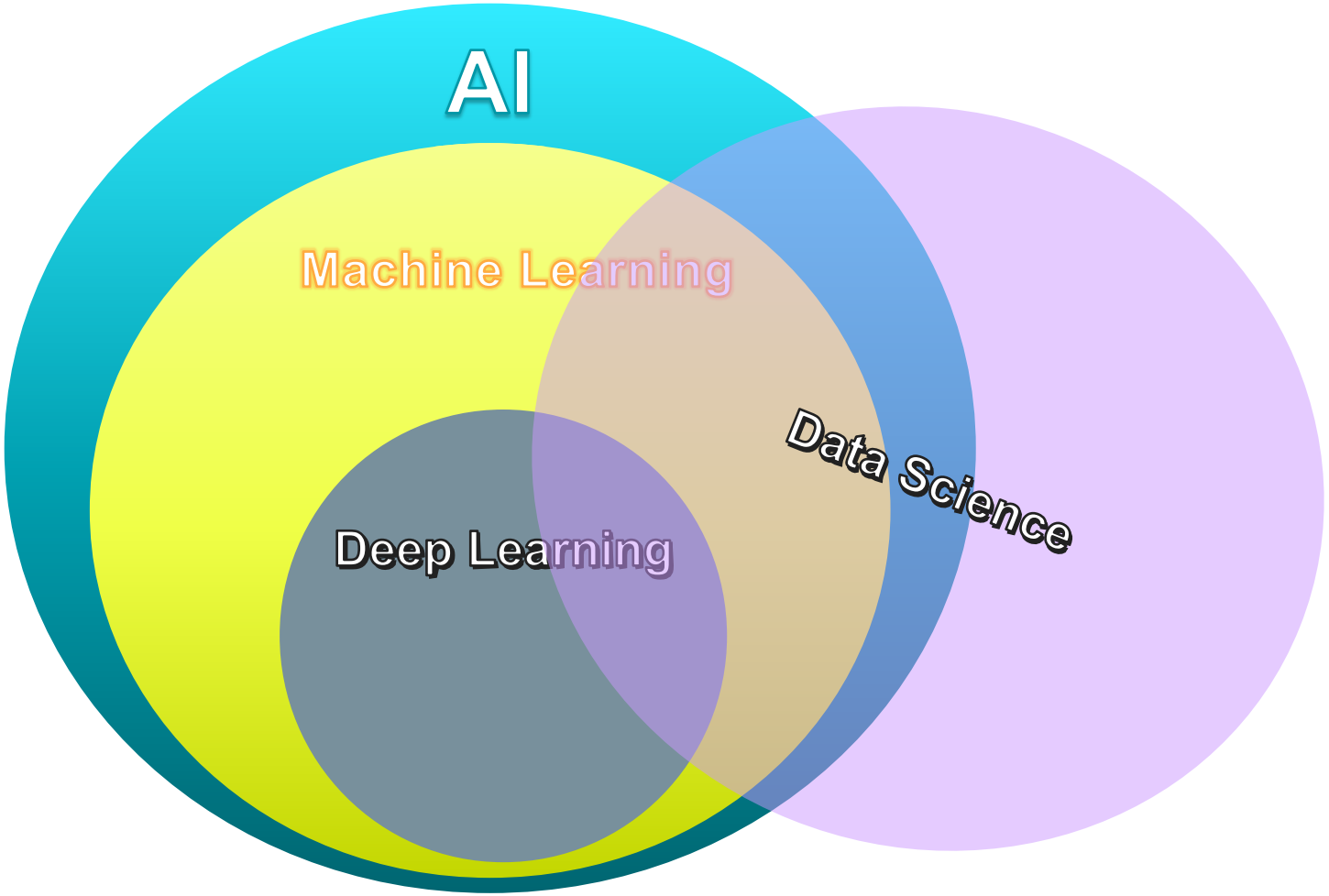
Partition the training data set into b mini-batches based on the batch size; pick up a batch on each iteration and sum up over examples in a batch

Learning happens on single example in each iteration

Machine Learning: Summary of basic concepts

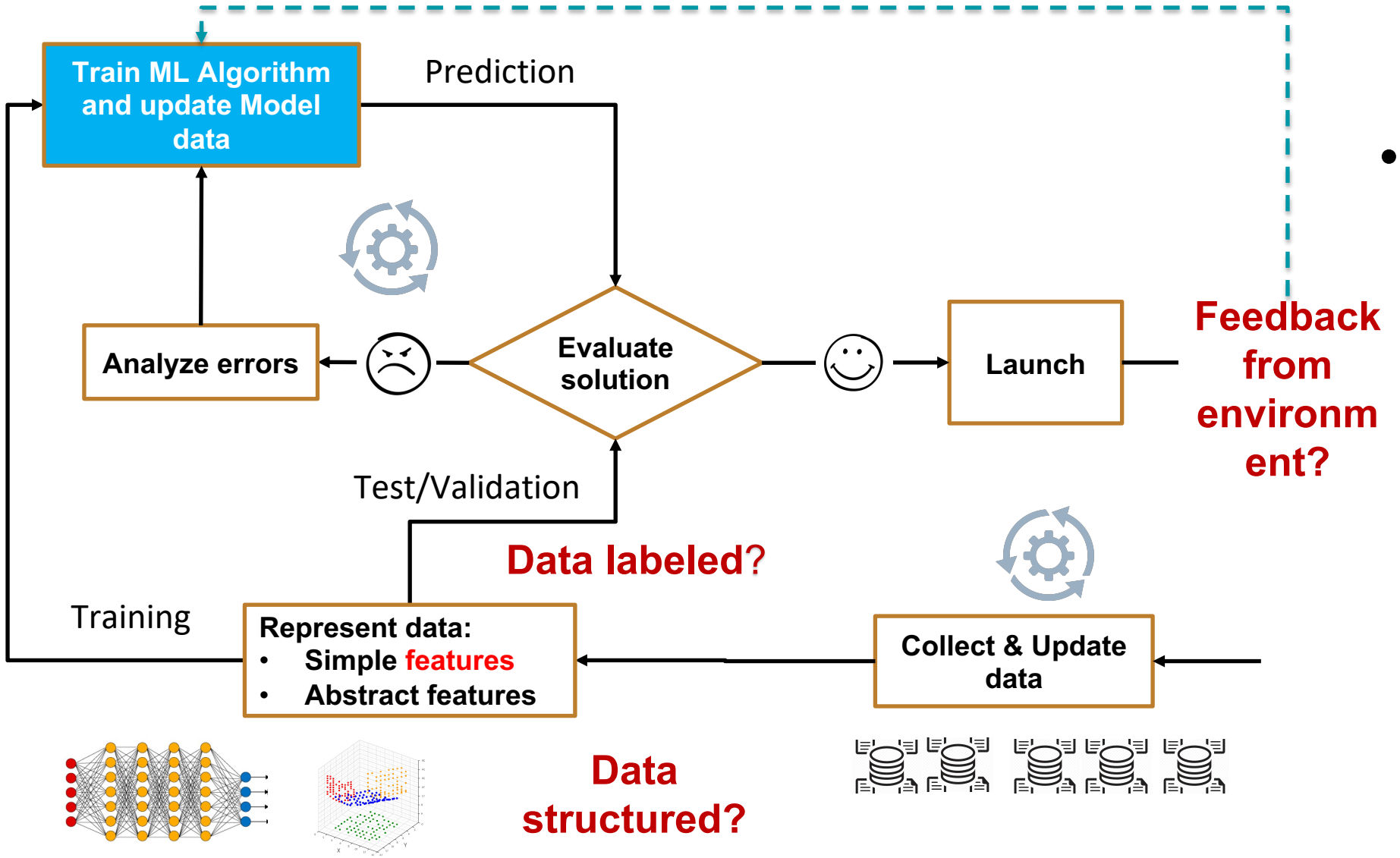
- Label and Feature
- Regression and Classification
- Model and Algorithm
- Representation
- Empirical Risk Minimization
- Gradient Descent
- Hyperparameter and Learning Rate
- Batch, mini-Batch, Stochastic Gradient Descent

What is what



- **Artificial Intelligence**: Intelligence exhibited by machines to mimic a human behavior
- **Machine Learning**: Computers being able to learn without hand-coding each step
- **Deep Learning**: a class of machine learning algorithms that uses neural net and multiple layers to progressively extract higher-level features from the raw input.
- **Data Science**: Methods, processes, and systems to extract insights from data

Types of Machine Learning Systems



- Learn from experience – the environment's feedback resulted from the previous action – Reinforcement learning

- Supervised vs. Unsupervised Labeled dataset specifies that some input and output parameters are already mapped

- Learn from (semi) structured data vs. unstructured data – deep learning

Structured vs. Unstructured Data

- Structured Data

  Highly organized and stored in a predefined format, e.g. relational databases and spreadsheets.
- Unstructured Data

  Data that isn't organized in a pre-defined manner, e.g. web pages, emails, videos, Facebook posts, etc.
- Semi-structured Data

  Data falls in between structured and unstructured data. There is no formal structure like in relational databases or data tables, but there are tags or other kinds of markers that separate semantic elements and define the hierarchy of records and fields within the data, e.g. JSON and XML files.

Structured or Unstructured Data?

	A	B	C
1	**students**		
2	**student_id**	**student_name**	**gpa**
3	2538	John Smith	3.5
4	2541	Mary Sue	4
5	2542	Tony Stark	3.8
6			

```
students: [
 {
 "student_id" : 2538,
 "student_name" : "John Smith",
 "gpa" : 3.5
 },
 {
 "student_id" : 2541,
 "student_name" : "Mary Sue",
 "gpa" : 4
 },
 {
 "student_id" : 2542,
 "student_name" : "Tony Stark",
 "gpa" : 3.8
 }
]
```

mary.sue@gmail.com

Hey

Hey Mary,
I just saw the results: a 4.0 GPA, you nerd! Let me know if the plan for Saturday still stands.

Tony S.

CVE-2022-28749 Detail

Current Description

Zooms On-Premise Meeting Connector MMR before version 4.8.113.20220526 fails to properly check the permissions of a Zoom meeting attendee. As a result, a threat actor in the can join the meeting without the consent of the host.

Hide Analysis Description

Analysis Description

Zooms On-Premise Meeting Connector MMR before version 4.8.113.20220526 fails to properly check the permissions of a Zoom meeting attendee. As a result, a threat actor in the can join the meeting without the consent of the host.

Severity

CVSS Version 3.x | CVSS Version 2.0

CVSS 3.x Severity and Metrics:

NVD

NIST: NVD

Base Score: 4.3 MEDIUM

Vector: CVSS:3.1/AV:N/AC:L/PR:L/UI:N/S:U/C:N/I:L/A:N

https://www.cvedetails.com/cve/CVE-2018-12359/

There are **131** matching records.

Displaying matches **1** through **20**.

1 2 3 4 5 6 7 > >>

Vuln ID	Summary	CVSS Severity
CVE-2021-40150	The web server of the E1 Zoom camera through 3.0.0.716 discloses its configuration via the /conf/ directory that is mapped to a publicly accessible path. In this way an attacker can download the entire NGINX/FastCGI configurations by querying the /conf/nginx.conf or /conf/fastcgi.conf URI. **Published:** July 17, 2022; 7:15:08 PM -0400	*V3.1:* 7.5 HIGH *V2.0:*(not available)
CVE-2021-40149	The web server of the E1 Zoom camera through 3.0.0.716 discloses its SSL private key via the root web server directory. In this way an attacker can download the entire key via the /self.key URI. **Published:** July 17, 2022; 6:15:08 PM -0400	*V3.1:* 5.9 MEDIUM *V2.0:*(not available)
CVE-2022-28749	Zooms On-Premise Meeting Connector MMR before version 4.8.113.20220526 fails to properly check the	*V3.1:* 4.3 MEDIUM *V2.0:* 4.0 MEDIUM

```
{
   "glossary": {
      "title": "example glossary",
  "GlossDiv": {
         "title": "S",
   "GlossList": {
            "GlossEntry": {
               "ID": "SGML",
    "SortAs": "SGML",
    "GlossTerm": "Standard Generalized Markup Language",
    "Acronym": "SGML",
    "Abbrev": "ISO 8879:1986",
    "GlossDef": {
                  "para": "A meta-markup language, used to create markup
languages such as DocBook.",
     "GlossSeeAlso": ["GML", "XML"]
               },
    "GlossSee": "markup"
            }
         }
      }
   }
}
```

The same text expressed as XML:

```
<!DOCTYPE glossary PUBLIC "-//OASIS//DTD DocBook V3.1//EN">
 <glossary><title>example glossary</title>
  <GlossDiv><title>S</title>
   <GlossList>
    <GlossEntry ID="SGML" SortAs="SGML">
     <GlossTerm>Standard Generalized Markup Language</GlossTerm>
     <Acronym>SGML</Acronym>
     <Abbrev>ISO 8879:1986</Abbrev>
     <GlossDef>
      <para>A meta-markup language, used to create markup
languages such as DocBook.</para>
      <GlossSeeAlso OtherTerm="GML">
      <GlossSeeAlso OtherTerm="XML">
     </GlossDef>
     <GlossSee OtherTerm="markup">
    </GlossEntry>
   </GlossList>
  </GlossDiv>
 </glossary>
```

Output -- What can ML do?

Classification

Which of N labels?
`cat, dog, horse, or bear`

Regression

Predict numerical values
`(e.g. click-through-rate)`

Clustering

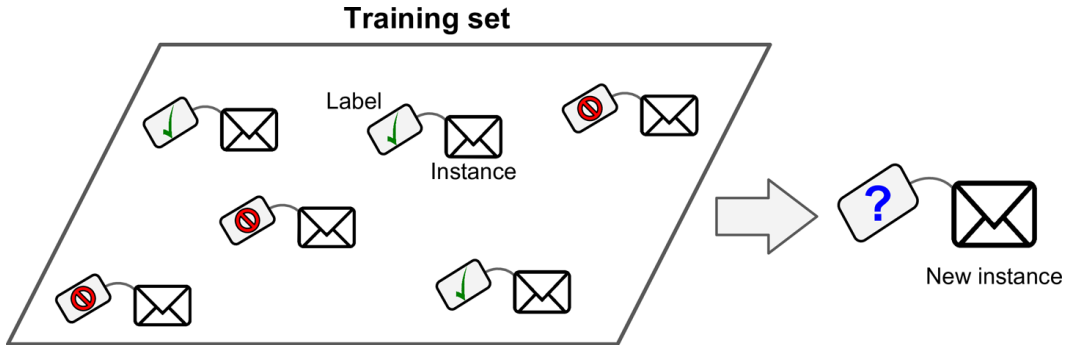
Most similar other examples
Most relevant documents
(unsupervised)

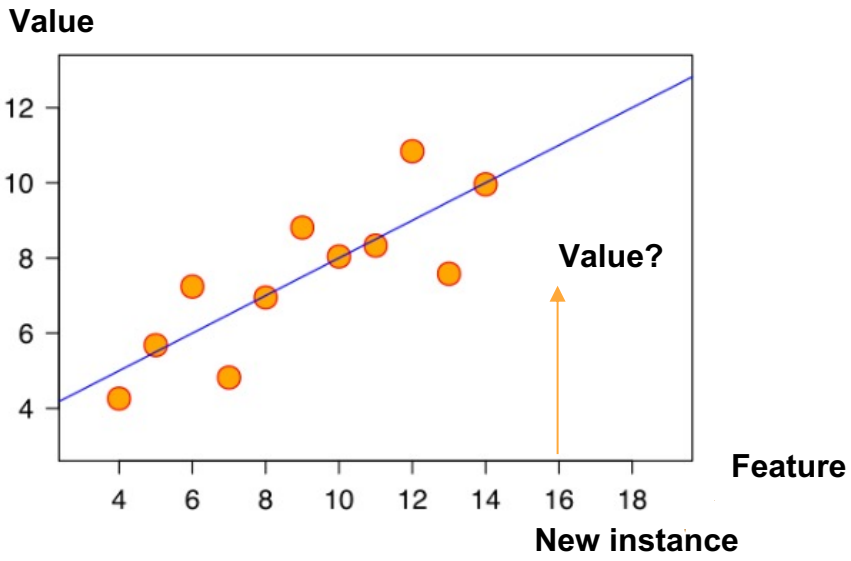
Generation

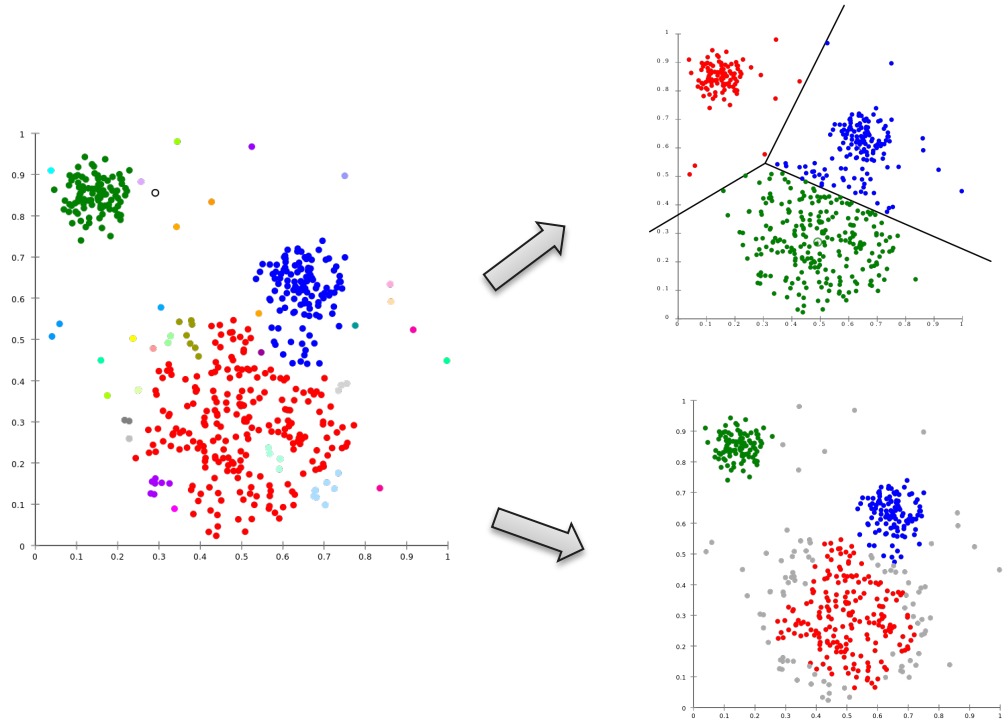
Complex output
`(e.g. image captions, translations)`

Types of Machine Learning Systems cont.

- Supervised vs. Unsupervised



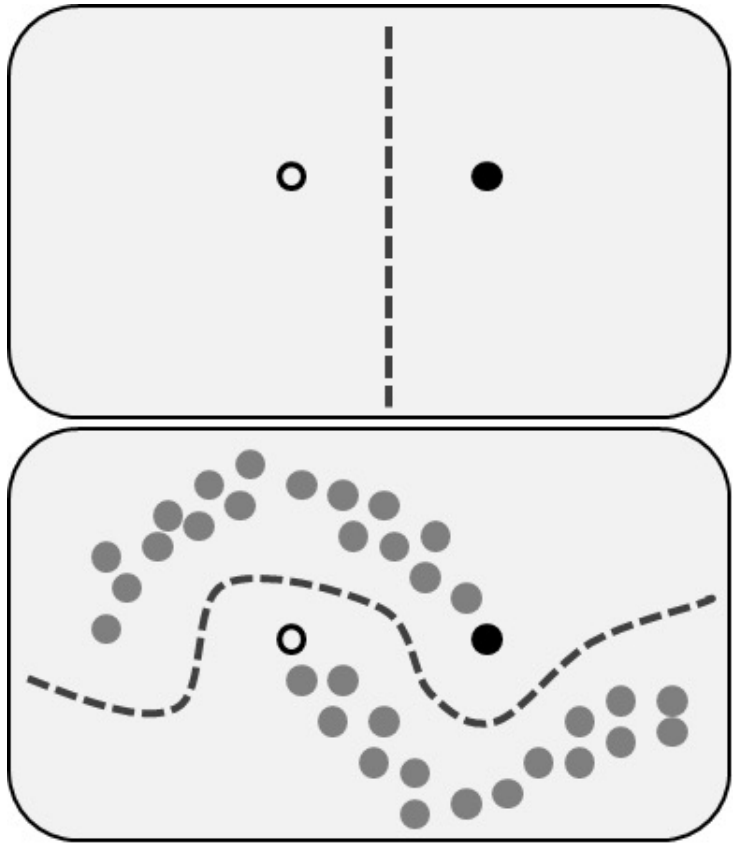




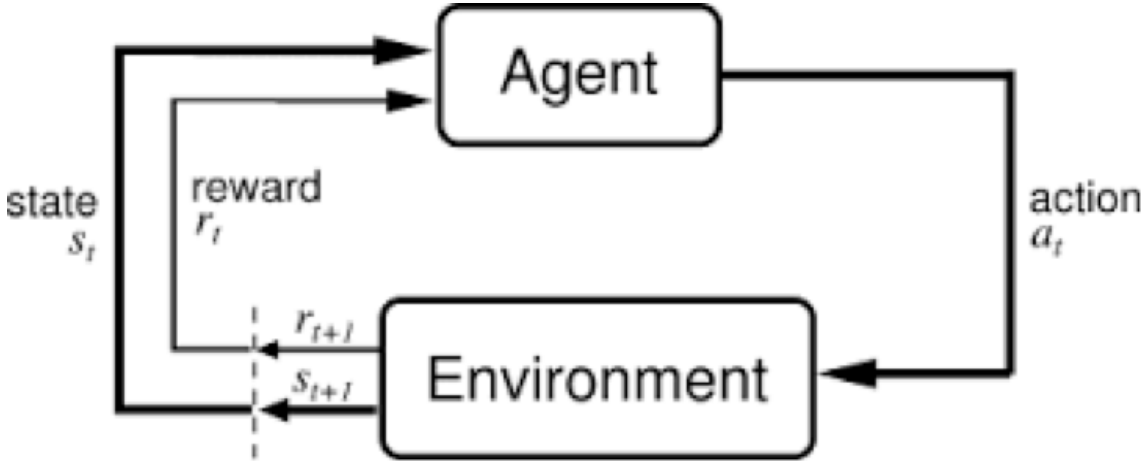
Source of picture: https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/ch01.html

Types of Machine Learning Systems cont.

- Semi-Supervised Learning



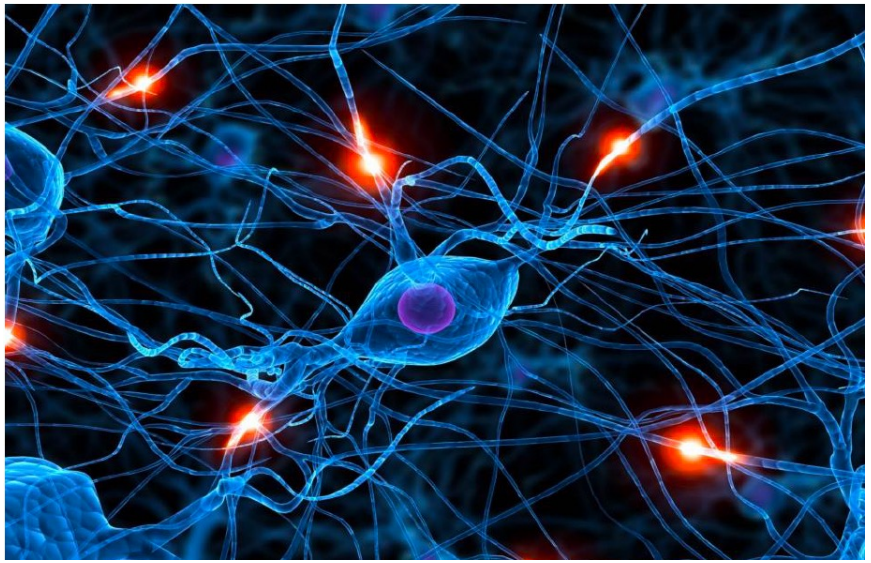
- Reinforcement Learning

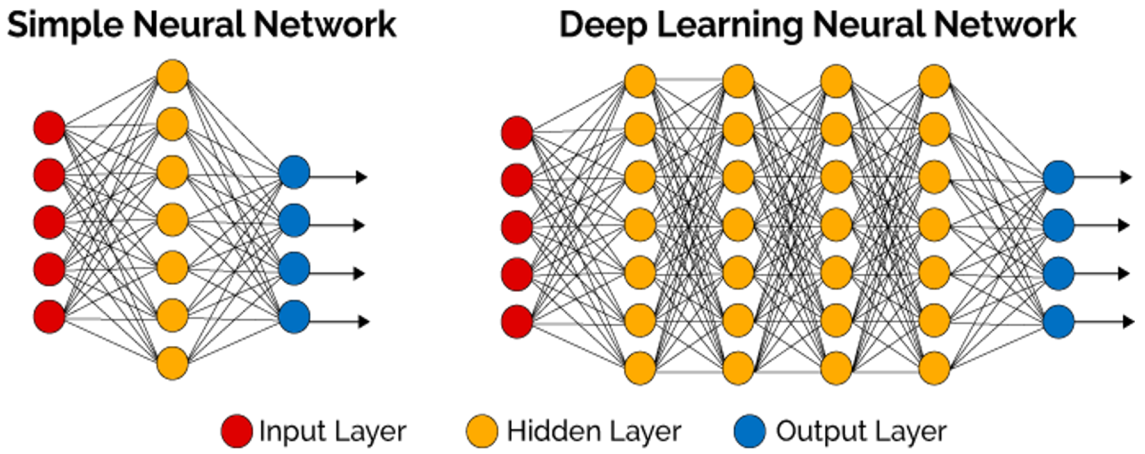


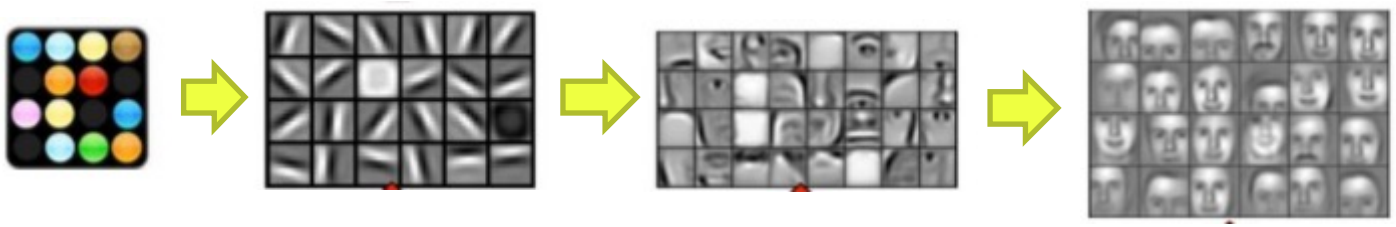
*Credit: Sutton & Barto

Types of Machine Learning Systems -- Deep Learning

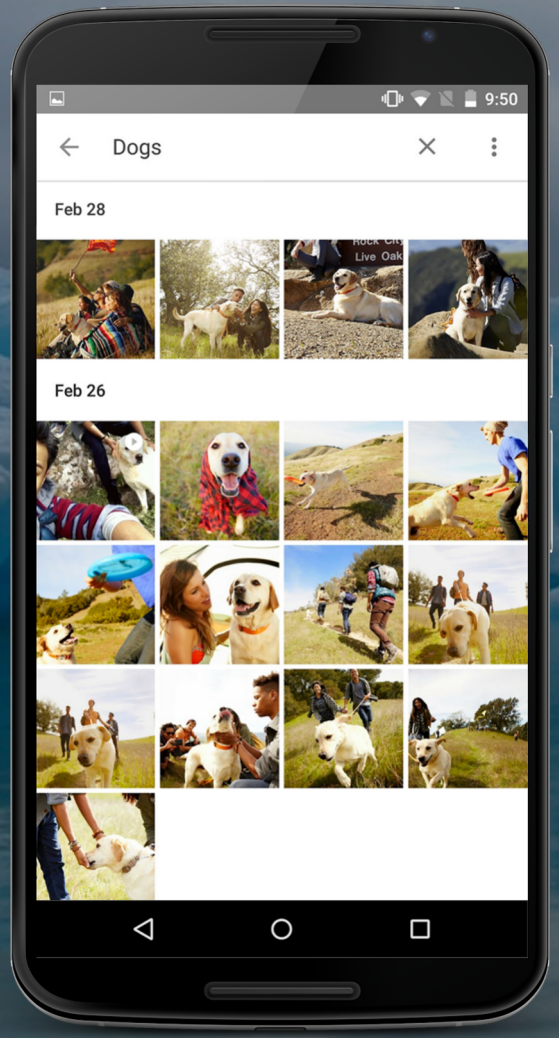
- Representation learning attempts to automatically learn good features or representation
- It will learn multiple levels of representation
- From “raw” inputs x





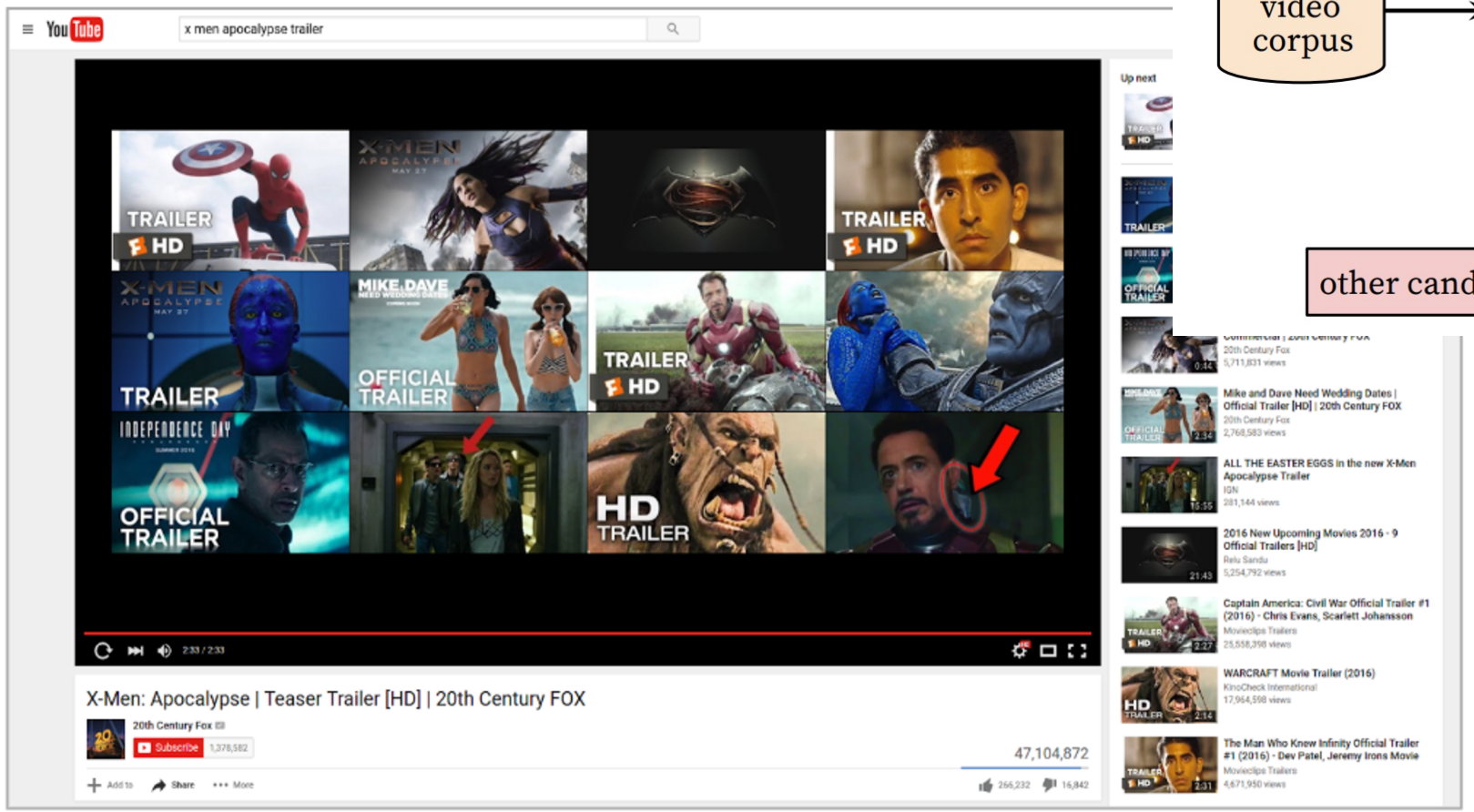


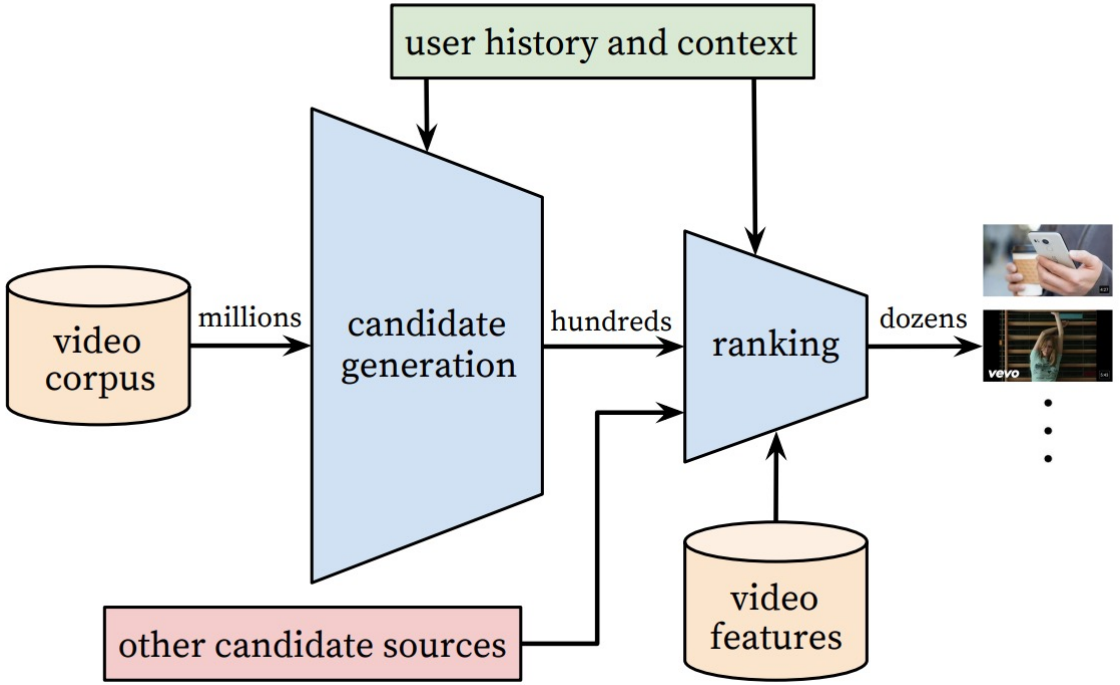
Google Photos



[Improving Photo Search: A Step Across the Semantic Gap]

YouTube Recommendations

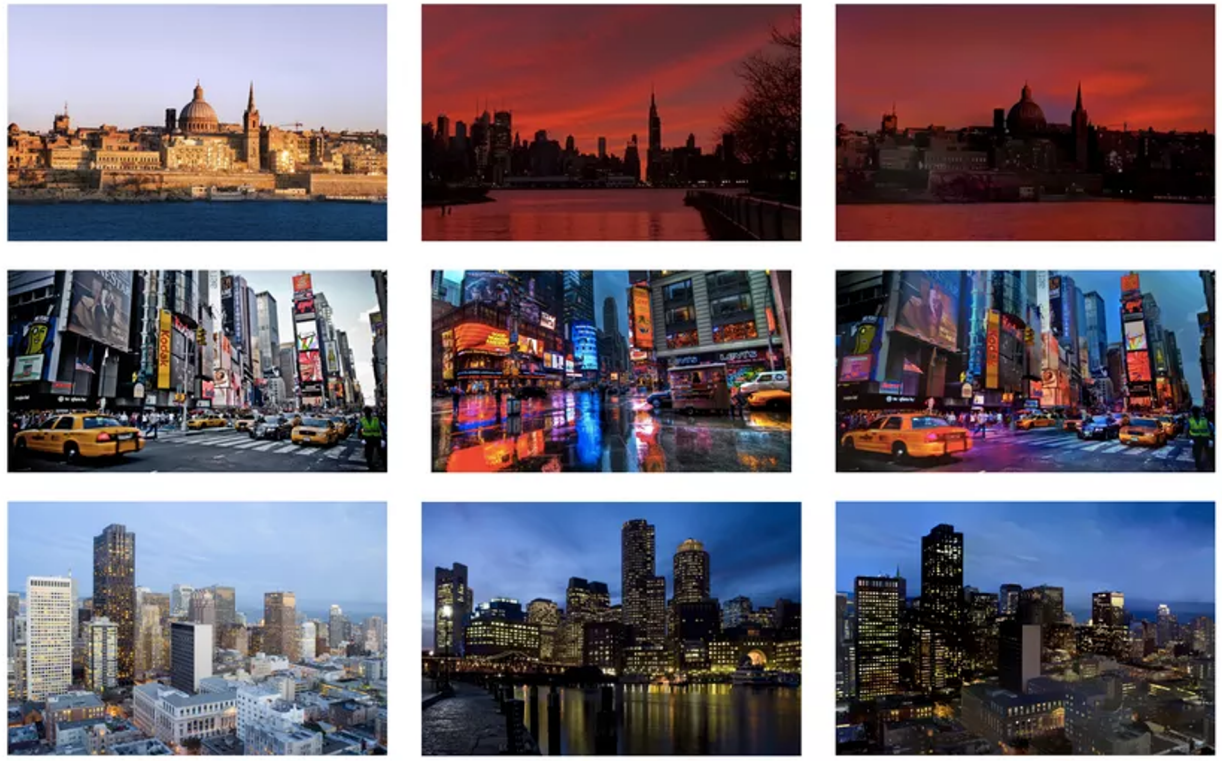




[Deep Neural Networks for YouTube Recommendations]

Deep Learning For Arts

Style transfer based on Deep Learning: use one image to stylize another.



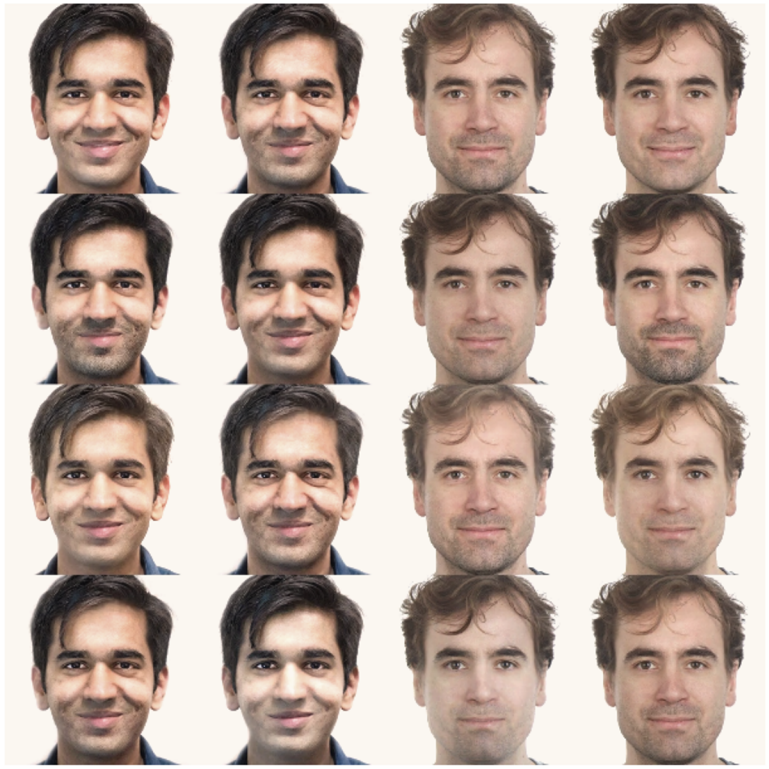
Original photo Reference photo Result

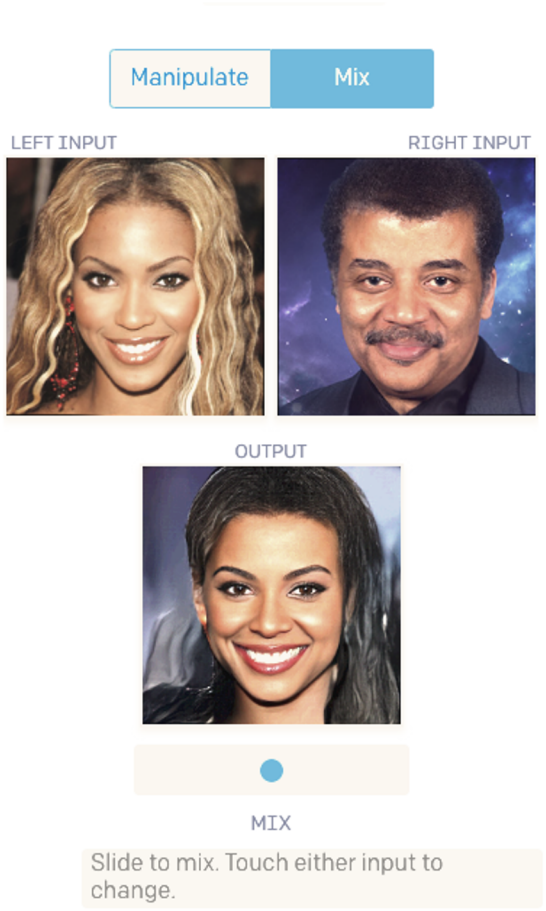
The now iconic examples from Figure 2 of Gatys et al (2015).

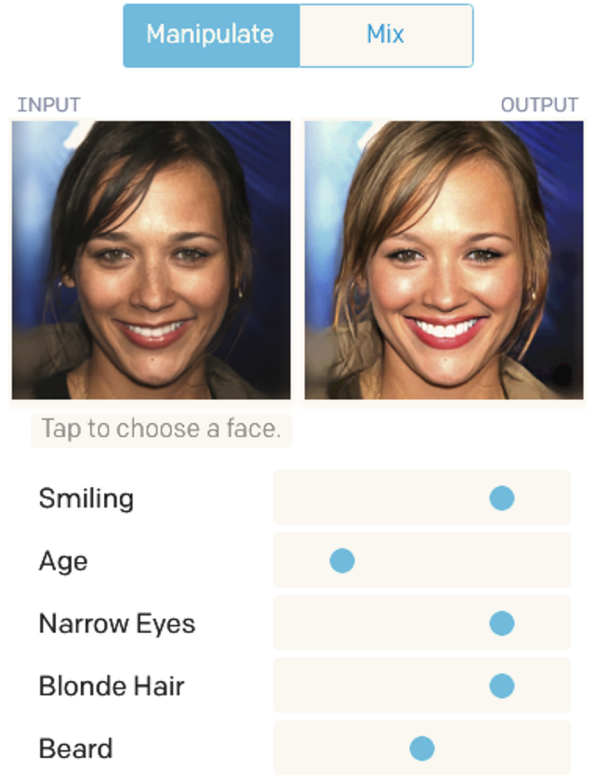
Deep Learning For Data Generation

https://openai.com/blog/glow/

Glow, a reversible generative model using invertible 1*1 convolutions, learns a latent space where certain directions capture attributes like age, hair color, and so on. (Kingma & Dhariwal 2018)

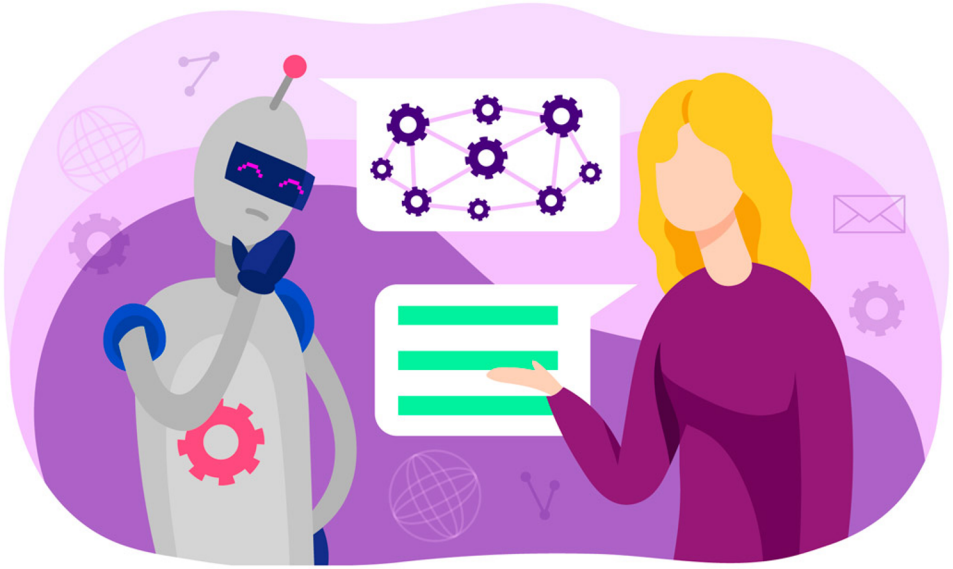






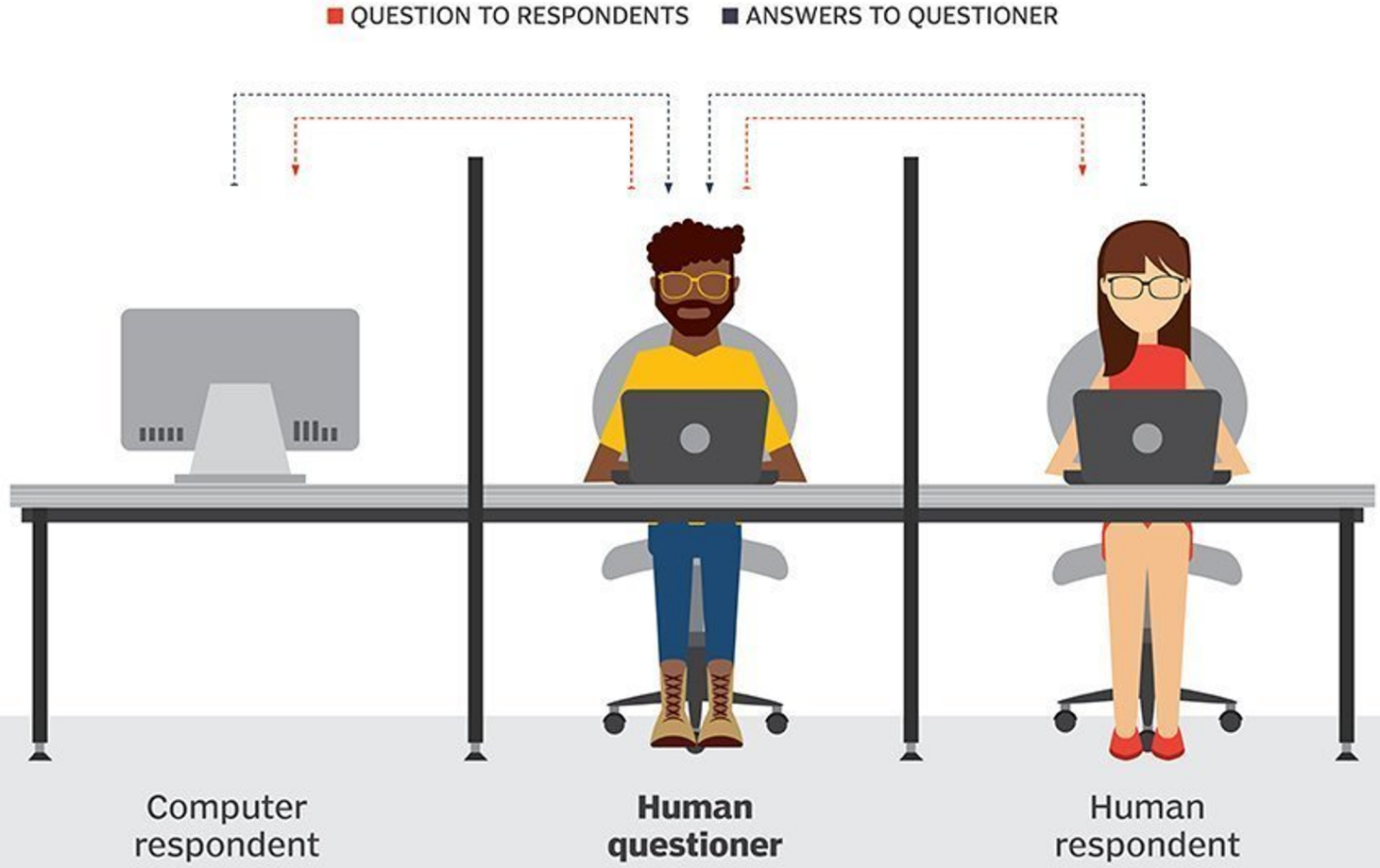
What is Natural Language Processing (NLP)?

- Natural Language Processing: a subfield of the broader Artificial Intelligence (AI) discipline, with three sub-topics:
  - Computer Science
  - Artificial intelligence
  - Linguistics
- NLP enables computers to understand and process human languages.
- One definition of AI-complete is perfect language understanding.



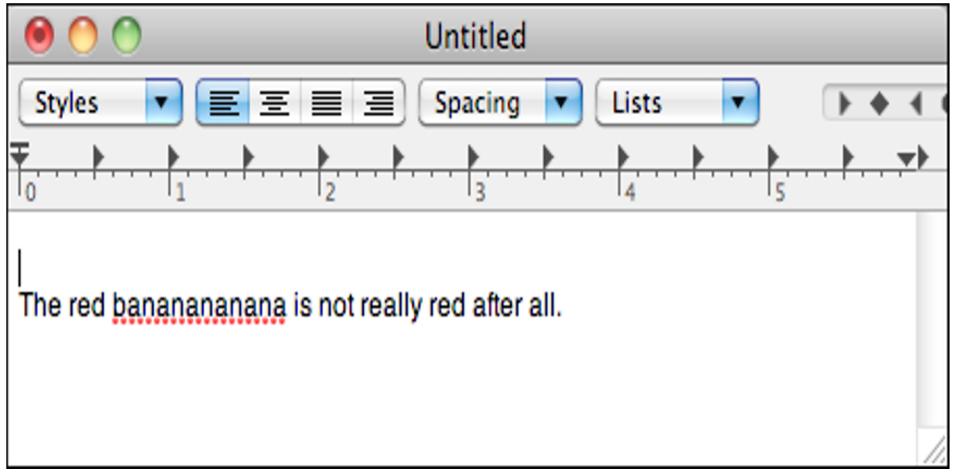
Turing test

During the Turing test, the human questioner asks a series of questions to both respondents. After the specified time, the questioner tries to decide which terminal is operated by the human respondent and which terminal is operated by the computer.

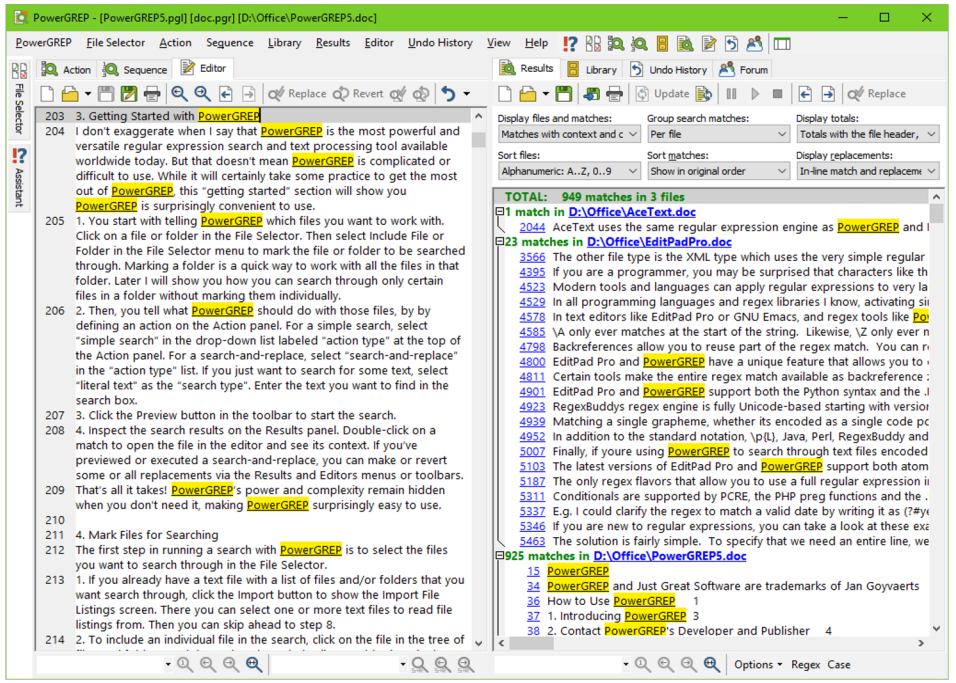


Easy NLP Tasks

- Spell Checking



- Keyword Search



Medium-Level NLP Tasks

- Name Entity Recognition

Figure 1: An example of NER application on an example text

https://nlp.stanford.edu/software/CRF-NER.html

- Convert unstructured text into a well structured document

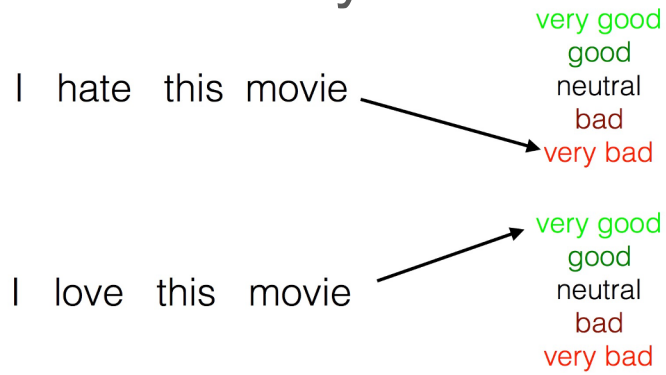
Medium-Level NLP Tasks

- Topic Classification

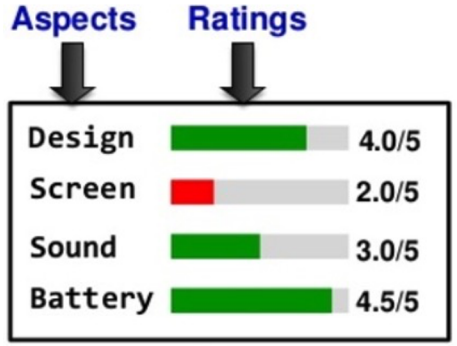
- Assign topic into each document/piece of text

Hard NLP Tasks

- Sentiment Analysis



- Aspect-based sentiment Analysis



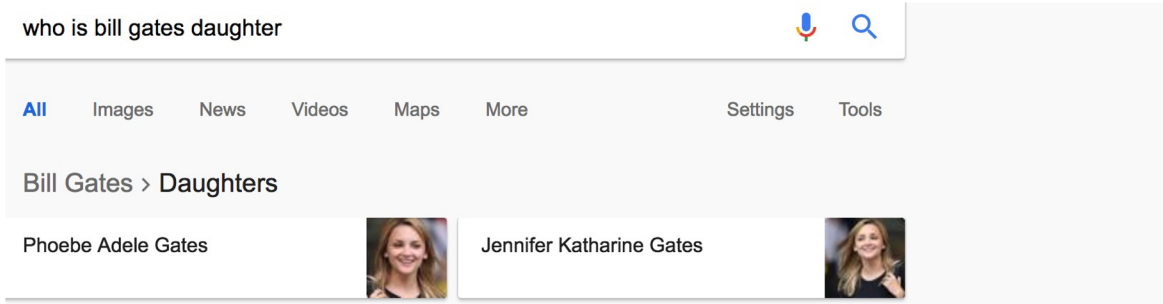
- Analyze opinions/sentiment behind text

https://komprehend.io/sentiment-analysis

- A stock investor scanning news about a company to assess overall market
- sentiment
- An individual scanning tweets about the launch of a new phone to decide the
- prevailing sentiment
- A political party analyzing social media feeds to assess the sentiment regarding
- their candidate

Hard NLP Tasks

- Machine Translation
- Question Answering

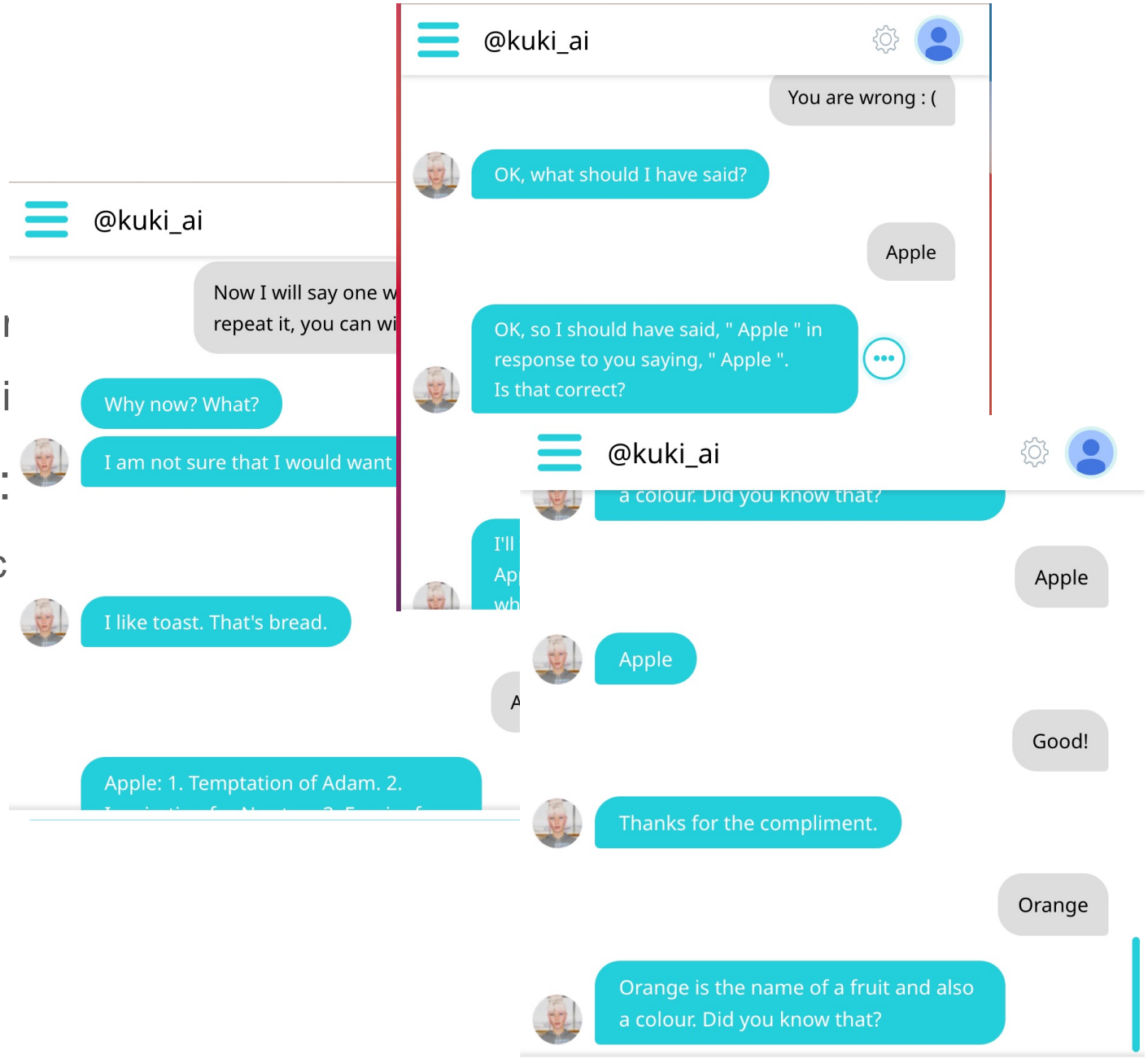


- Visual Question Answering

https://visualqa.org/

Chatbot

- Traditional approaches:
  - Hand-craft knowledge base ar
  - Can not address out-of-domai
- Deep learning approaches:
  - Neural language models whic
- Try a popular chatbot
  - https://chat.kuki.ai/chat



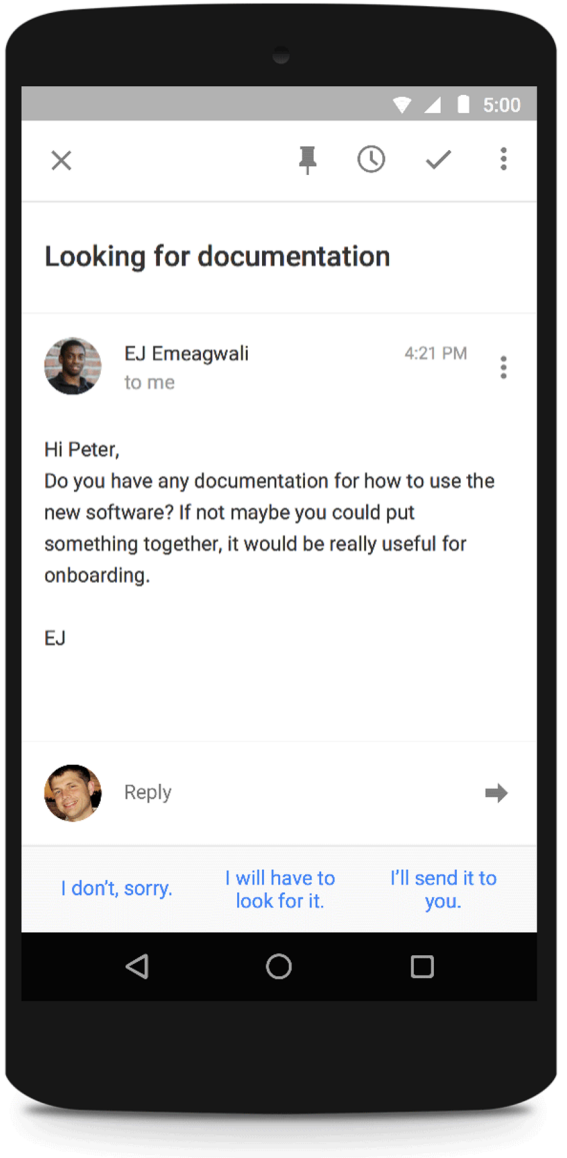
Smart Reply

April 1, 2009: April Fool's Day joke

Nov 5, 2015: Launched Real Product

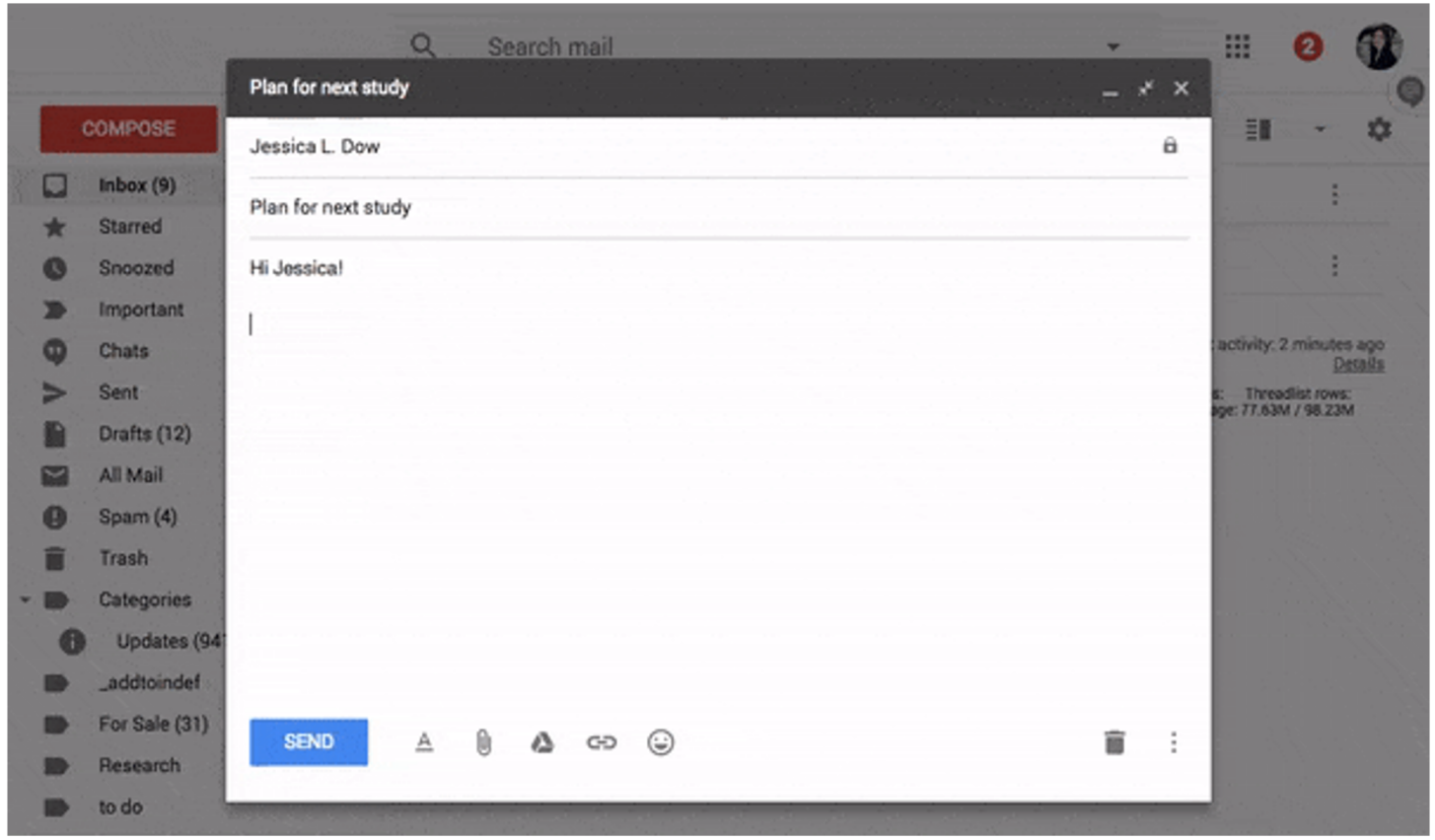
Feb 1, 2016: ~20% of mobile replies

[Smart Reply: Automated Response Suggestion for Email]



Smart Compose

Deep learning now helps reducing the typing effort by predicting words/phrases based on the previous context



[Smart Compose: Using Neural Networks to Help Write Emails]

Google Translate