CSC4007 Advanced Machine Learning

Lesson 01: Introduction

by Vien Ngo
EEECS / ECIT / DSSC

Outline

- Organization
- Outline of the module
- Introduction to machine learning
- What is a learning problem?
- Introduction to supervised learning
- Introduction to unsupervised learning

Outline

- **Organization**
- Outline of the module
- Introduction to machine learning
- What is a learning problem?
- Introduction to supervised learning
- Introduction to unsupervised learning

CSC4007: Organization

- Lectures:
  - Tuesday : 14:00 - 16:00
  - Thursday: 13:00 - 14:00
- Practical:
  - Thursday: 15:00 – 17:00
- Assignment (two):
  - Assignment 1: Release at week 15 (30% module marks), deadline at week 19
  - Assignment 2: Release at week 19 (30% module marks), deadline at week 23
- Final exam: 40% module marks in April

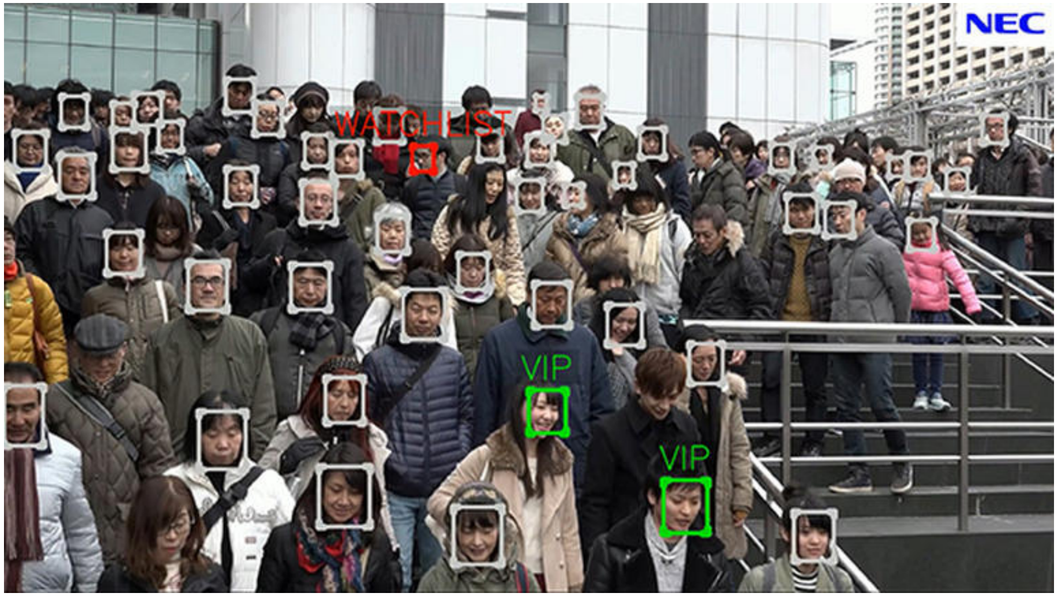
Outline of the Course

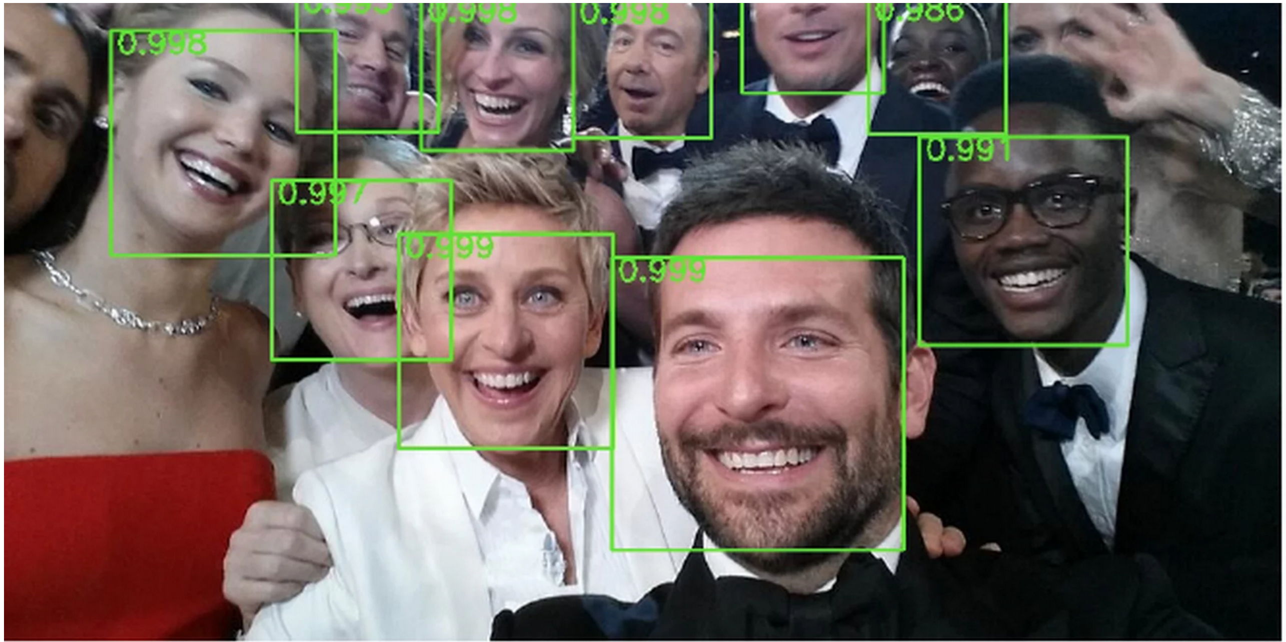
- Introduction to learning problem
- Basics on linear algebra
- Regression (week 3)
- Classification (week 4)
- Kernel methods, support vector machine (week 5)
- Unsupervised learning (week 6+7)
- Enssemble methods (week 8)
  - Boosting, random forest
- Neural networks (week 9+10)
  - Multilayer perception, back-propagation, CNN, LSTM

Outline

- Outline of the module
- **Introduction to machine learning**
- Components of learning
- Introduction to supervised learning
- Introduction to unsupervised learning
- Feature representation

Face detection on commercial products





Autonomous Driving

Machine learning definition

Arthur Lee Samuel (1901-1990)

- The term "machine learning" was first coined in 1959

  Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

checker game

Machine learning definition

Arthur Lee Samuel (1901-1990)

- The term "machine learning" was first coined in 1959

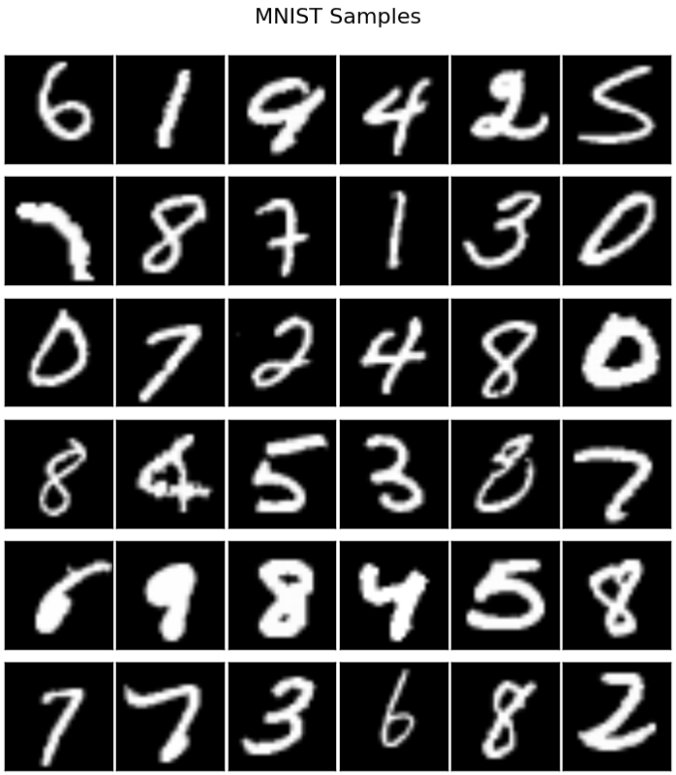
  Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

- Tom Mitchell (Computer scientist, CMU) proposed a more precise definition in 1998:

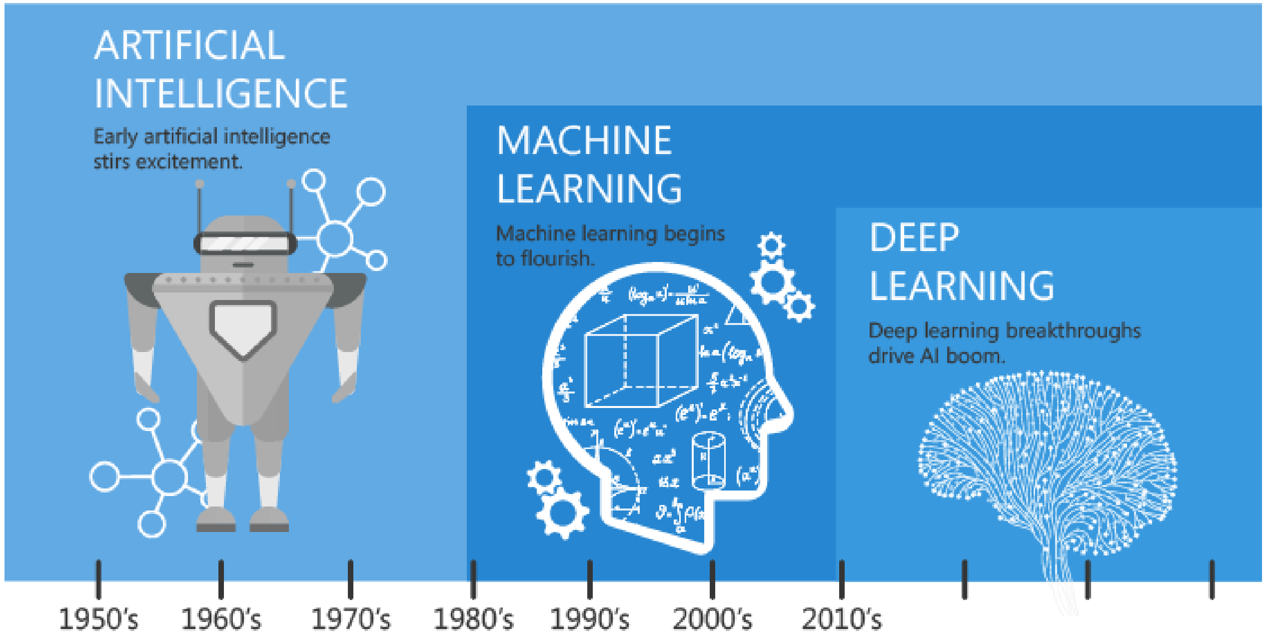
  *Well posed Learning Problem: A computer program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.*

Example 1

- Recognizing hand-written digits taken from US zip codes
  - **task T** is to label/classify an input image to digit numbers
  - the **performance measure P** could be the percentage correct classifications
  - **experience E** is the set of already labeled image



Machine learning



from Kapil's blog

- **Artificial Intelligence**: Techniques enable computer mimic human or other animal intelligence
- **Machine Learning**: a subfield in AI that uses statistical models to enable machine to improve at tasks with experience
- **Deep learning**: a subset of ML that uses a cascade of multiple layers of nonlinear processing units, e.g. neural networks, for feature extraction and transformation.

When to apply ML?

- Machine Learning consists of synthetic approaches that looks for **patterns** in data to solve different **predictive tasks**
- When to apply?
  - Human expertise is absent (i.e. when you can not code the rules)

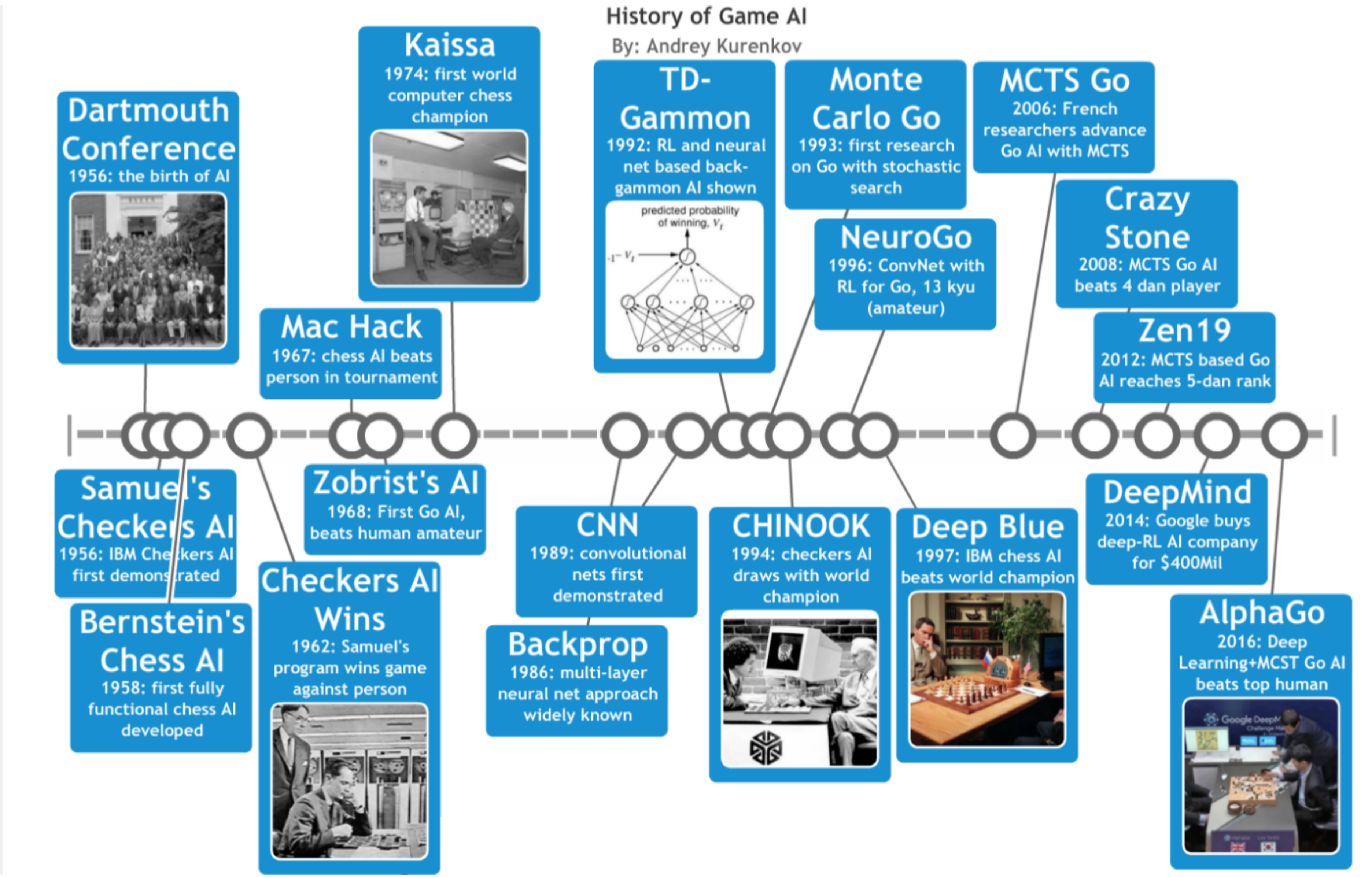
When to apply ML?

- Machine Learning consists of synthetic approaches that looks for **patterns** in data to solve different **predictive tasks**
- When to apply?
  - Human expertise is absent (i.e. when you can not code the rules)
  - Humans are unable to explain their expertise (speech recognition, vision, language)

When to apply ML?

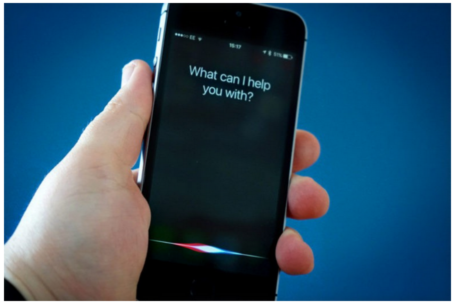
- Machine Learning consists of synthetic approaches that looks for **patterns** in data to solve different **predictive tasks**
- When to apply?
  - Human expertise is absent (i.e. when you can not code the rules)
  - Humans are unable to explain their expertise (speech recognition, vision, language)
  - The problem size is very large (i.e. when you can not scale)

Machine Learning is everywhere



Skynet Today

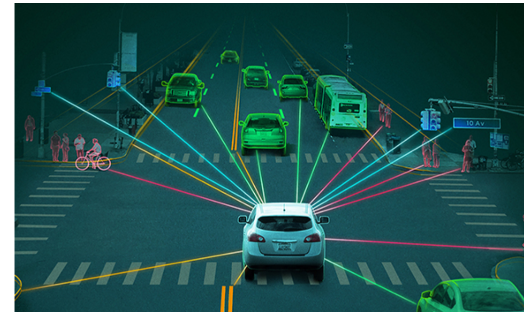
Machine Learning is everywhere



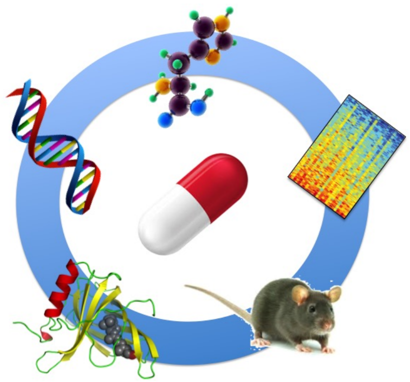
Voice assistance

NETFLIX

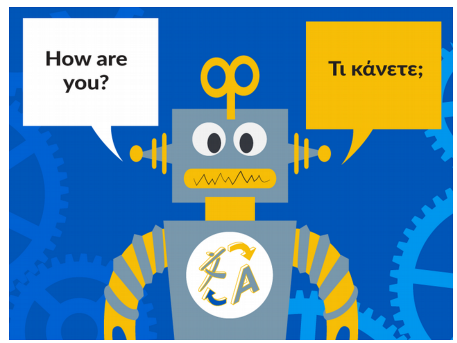
Recommendation system



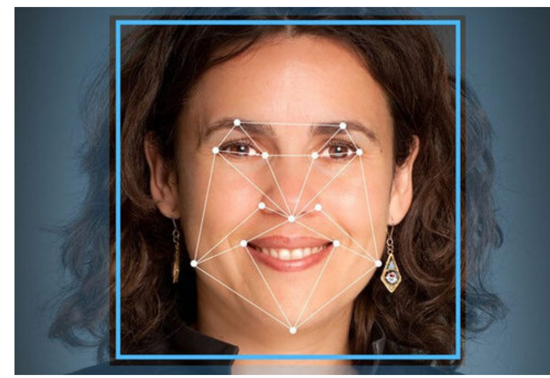
Autonomous driving



Drug discovery

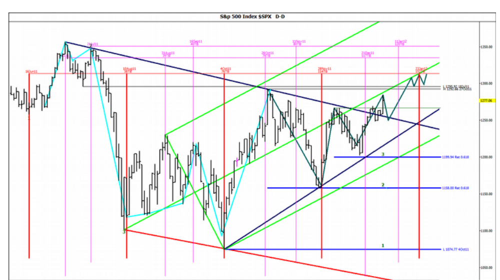


Machine translation



Face recognition

Disease dianogsis



hedge fund stock prediction

Assistive robotics

Machine Learning is everywhere

- **Forecasting** (e.g. sales, energy demand prediction, sales)
- **Inputting** missing data (e.g. Netflix/Amazon recommendations)
- **Detecting** anomalies (e.g. intruders, virus mutations)
- **Classifying** (e.g. credit risk assessment, cancer diagnosis)
- **Ranking** (e.g. Google search, personalization)
- **Summarizing** (e.g. social media sentiment, product review/opinion)
- **Decision making** (e.g, AI, robotics, trading)

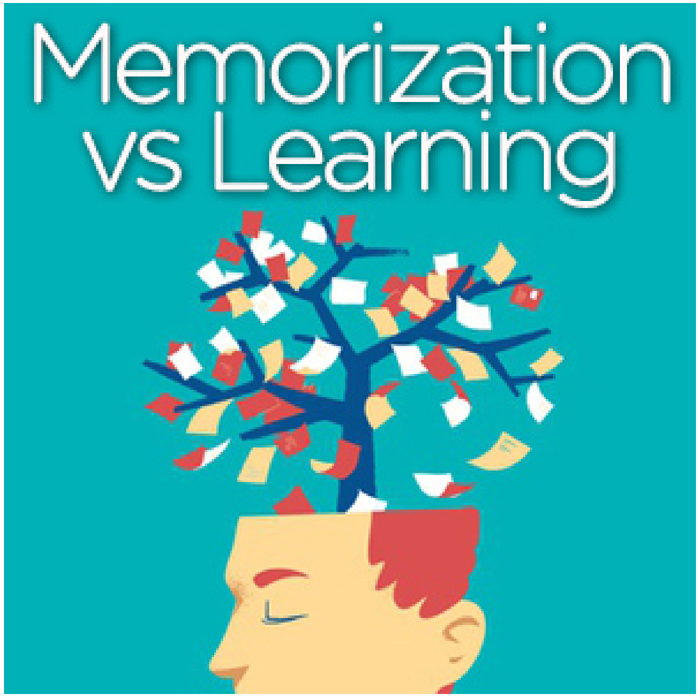
(Nando de Freitas's ML lecture)

Outline

- Outline of the module
- Introduction to machine learning
- **Components of learning?**
- Introduction to supervised learning
- Introduction to unsupervised learning
- Feature representation

Components of learning

- How are things learned?
  - Memorization (declarative knowledge)
    - Accumulation of individual facts
    - Cons: memory to store facts, time to observe facts
  - Generalization (imperative knowledge)
    - Deduce new facts from old facts
    - Cons: accuracy of deduction process



Components of learning

- Basic learning paradigm
  - Observe set of examples: **training data**
  - Infer something about process that generated that data
  - Use inference to make predictions about previoulsy unseen data: **test data**

Components of learning

- Example: credit approval
  - Applicant information:

age	23 years
gender	male
annual salary	$30,000
years in residence	1 year
years in job	1 year
current debt	$15,000
···	···

Approve credit?

Yaser Abu-Mostafa - LFD Leture 1

Given data of many previous customers and their credit assessment (5 years later)

Components of learning

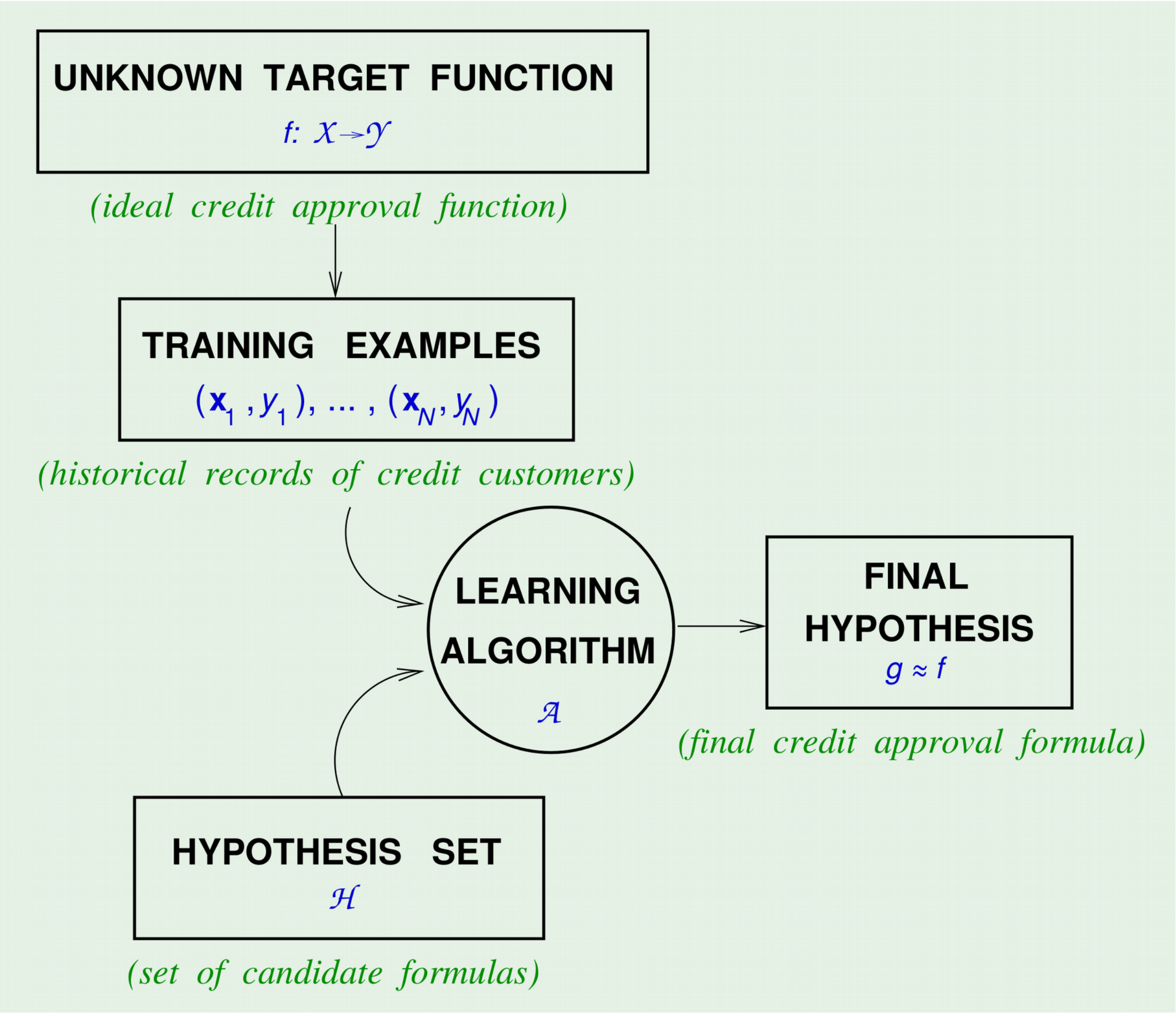
- Example: credit approval
  - Formalization:

- Input: $\mathbf{x}$ (*customer application*)
- Output: y (*good/bad customer?*)
- Target function: $f : \mathcal{X} \to \mathcal{Y}$ (*ideal credit approval formula*)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ (*historical records*)

$\downarrow \quad \downarrow \quad \downarrow$

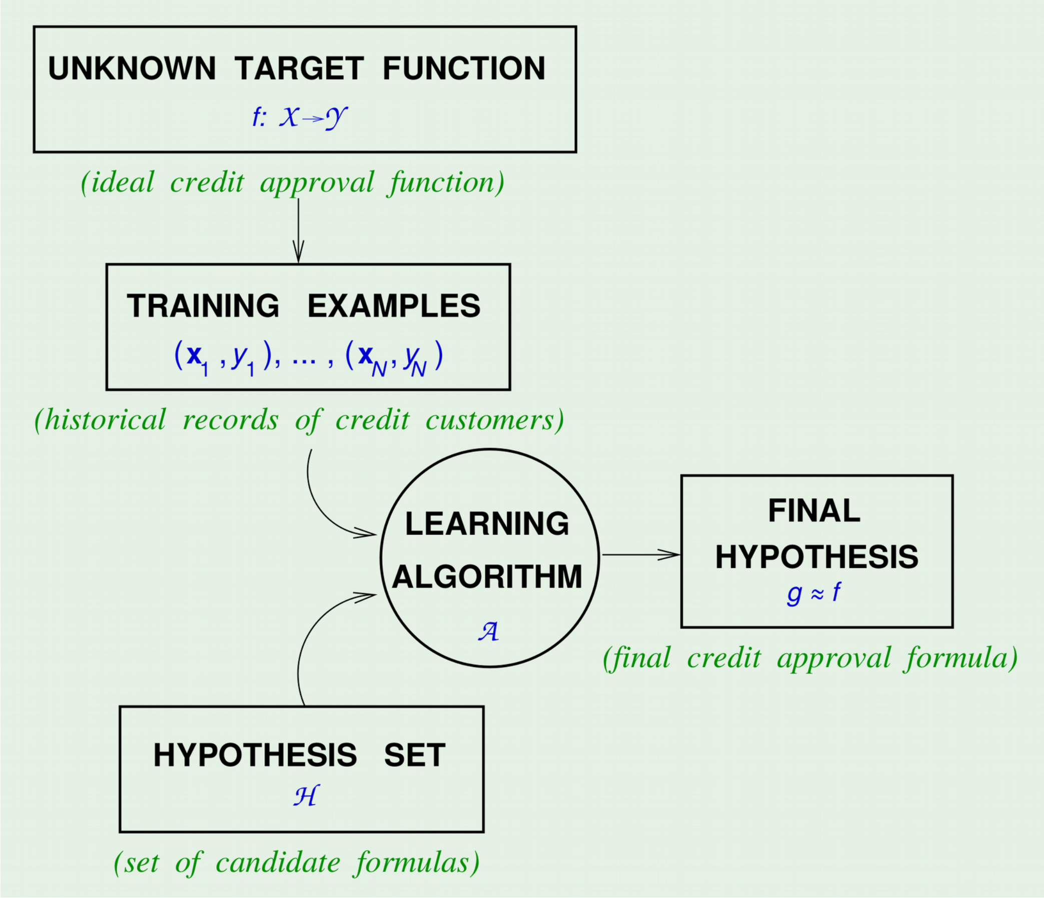
- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ (*formula to be used*)

Components of learning

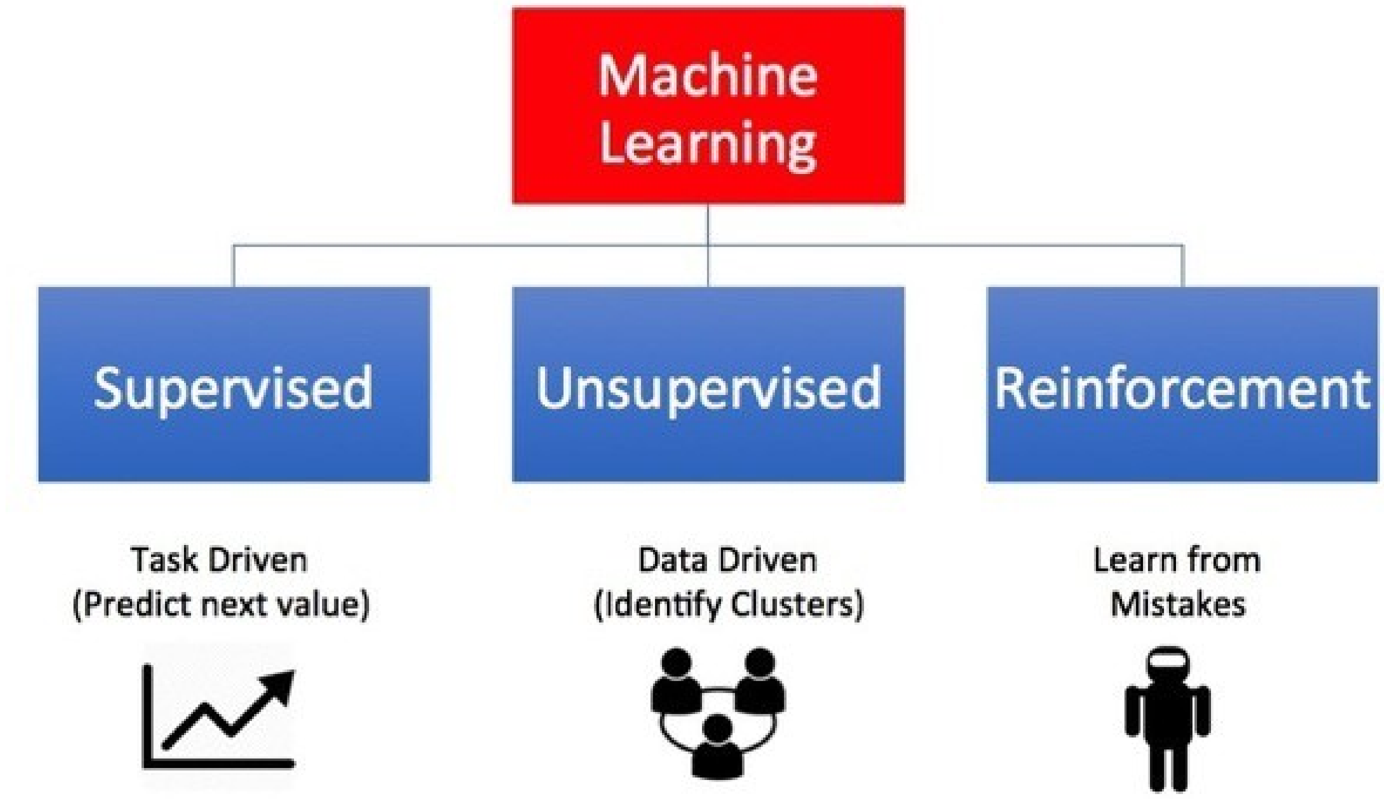


Components of learning

- Two main components of the learning problem:
  - **The Hypothesis Set**: linear functions, feedforward networks, etc.
  - **Learning Algorithms:** supervised learning, unsupervised learning, reinforcement learning



Machine learning algorithms



Supervised learning : when given **labeled data**
Unsupervised learning : when given **unlabaled data**
Reinforcement learning: learn by **trial and error** when given **evaluative feedback**

RL will not be covered in CSC4007

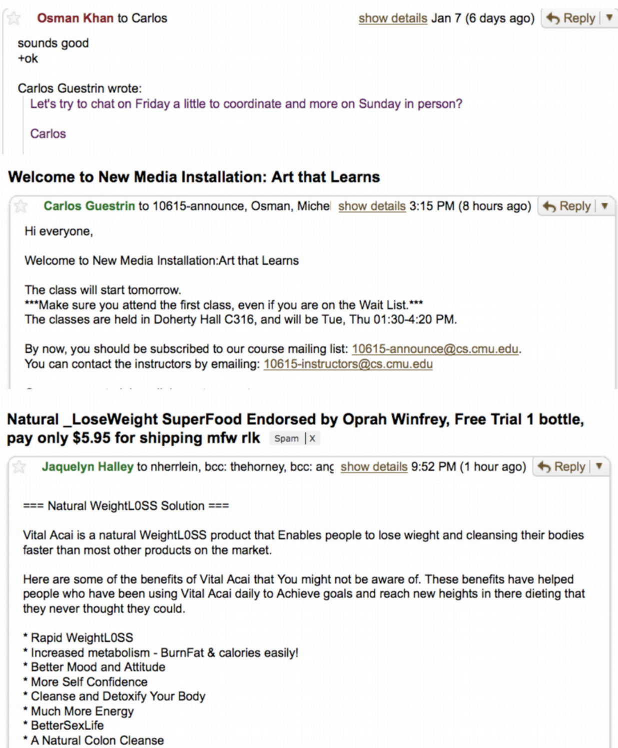
Outline

- Outline of the module
- Introduction to machine learning
- Components of learning?
- **Introduction to supervised learning**
- Introduction to unsupervised learning
- Feature representation

Introduction to supervised learning

- Spam filtering: given an email → **spam** or **not spam**

data **prediction**



→

Spam
vs.
Not Spam

Bag of words

$e.g.\ \ x = [\text{winner} = 2,\ \ \text{prize} = 1,\ \$dd = 2,\ \ldots]$

Spam

- Bag of words example:

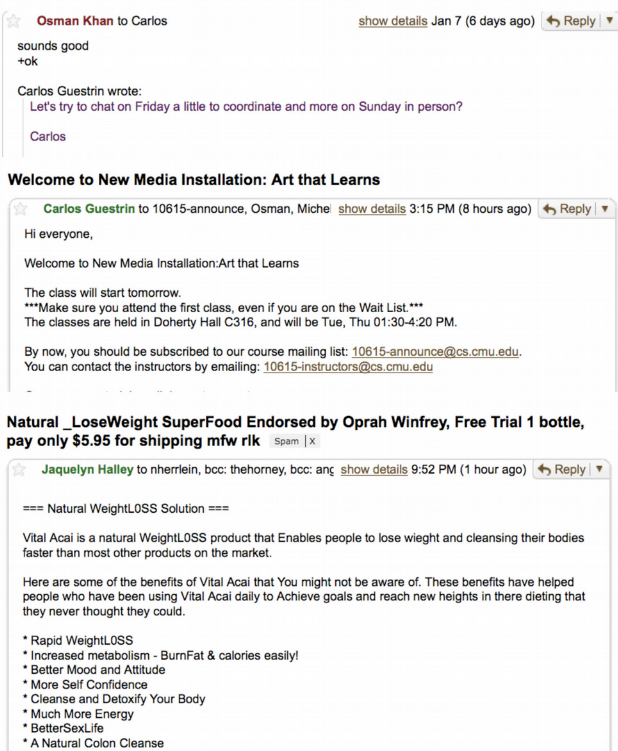
```
Sentence 1 = John likes to watch movies. Mary likes movies too.
BoW1 = {"John":1,"likes":2,"watch":1,"movies":2,"Mary":1,"too":1,other:1};

Sentence 2 = John also likes to watch football games.
BoW2 = {"John":1,"also":1,"likes":1,"watch":1,"football":1,"games":1,other:1};
```

Introduction to supervised learning

- Spam filtering: given an email → **spam** or **not spam**

data


Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos

Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Michel show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | x

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: ang show details 9:52 PM (1 hour ago) Reply

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

→

prediction

Spam
vs.
Not Spam

→

Classification
Predict discrete values

Outputs={0,1}

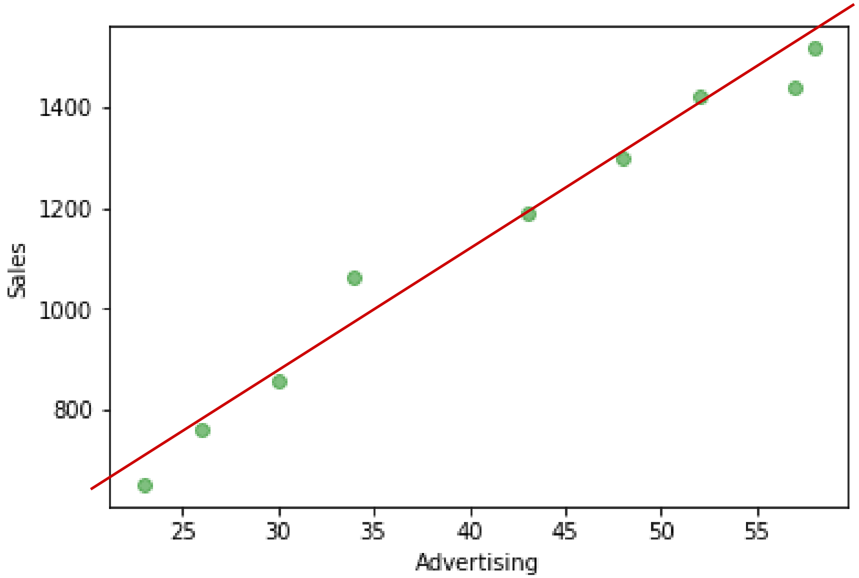
Introduction to supervised learning

- Sale prediction: advertising → sales (million EUR)

Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

Regression

Predict continuous values (sales)



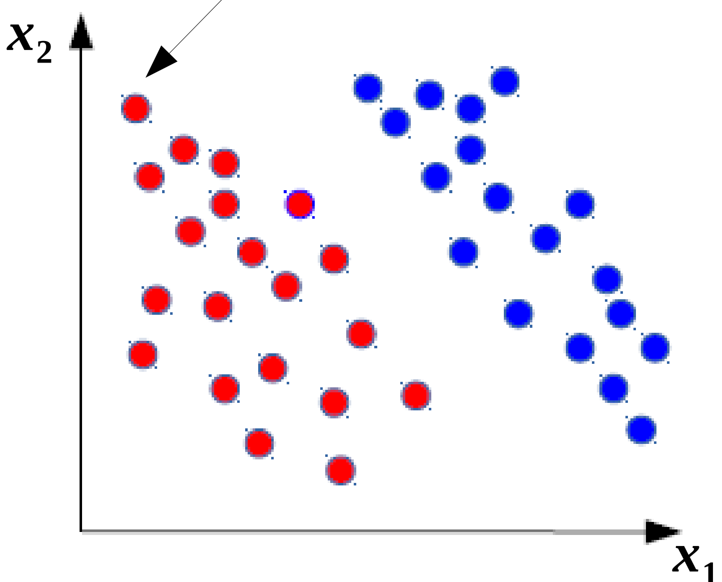
Outline

- Outline of the module
- Introduction to machine learning
- Components of learning?
- Introduction to supervised learning
- **Introduction to unsupervised learning**
- Feature representation

Introduction to unsupervised Learning

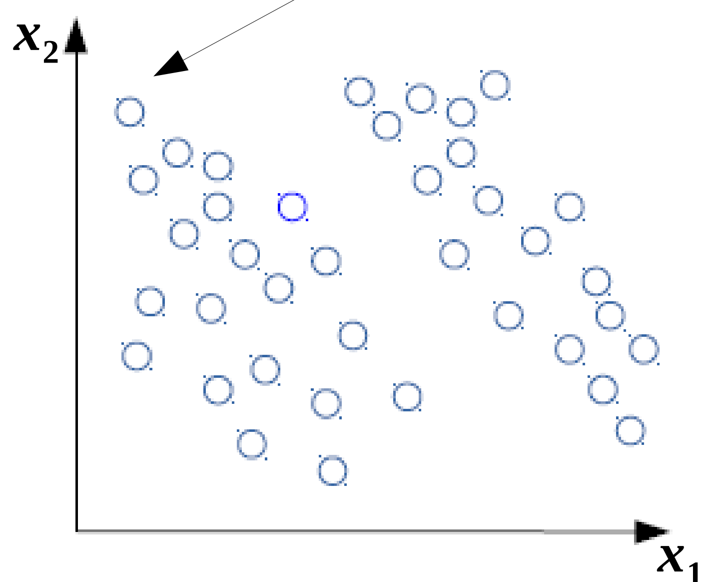
Supervised learning vs. Unsupervised learning

Dataset $\left\{x=[x_1, x_2], \mathbf{y=\{}\textcolor{red}{\mathbf{red}}, \textcolor{blue}{\mathbf{blue}}\mathbf{\}}\right\}$



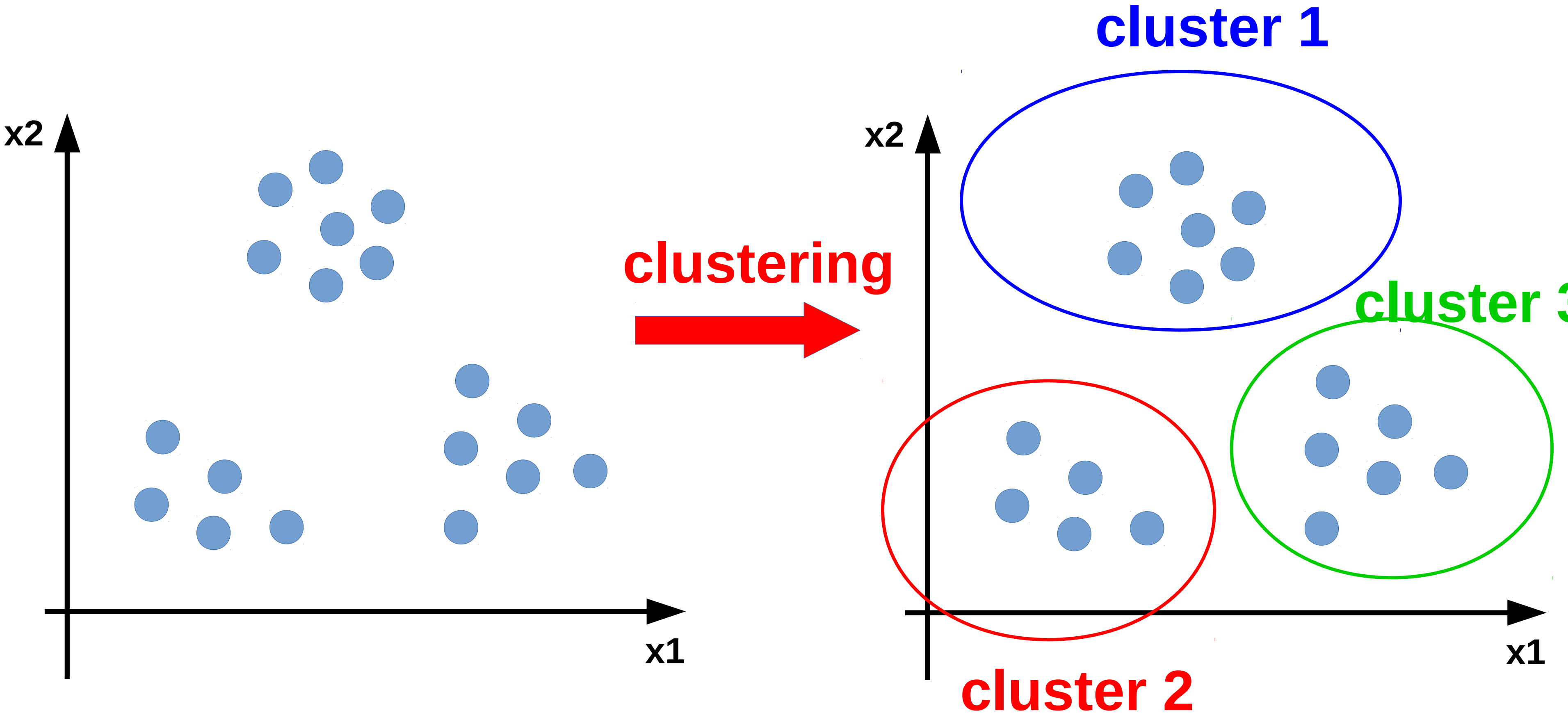
Supervised learning

Dataset $\left\{x=[x_1, x_2]\right\}$



Unsupervised learning

Unsupervised Learning: Clustering

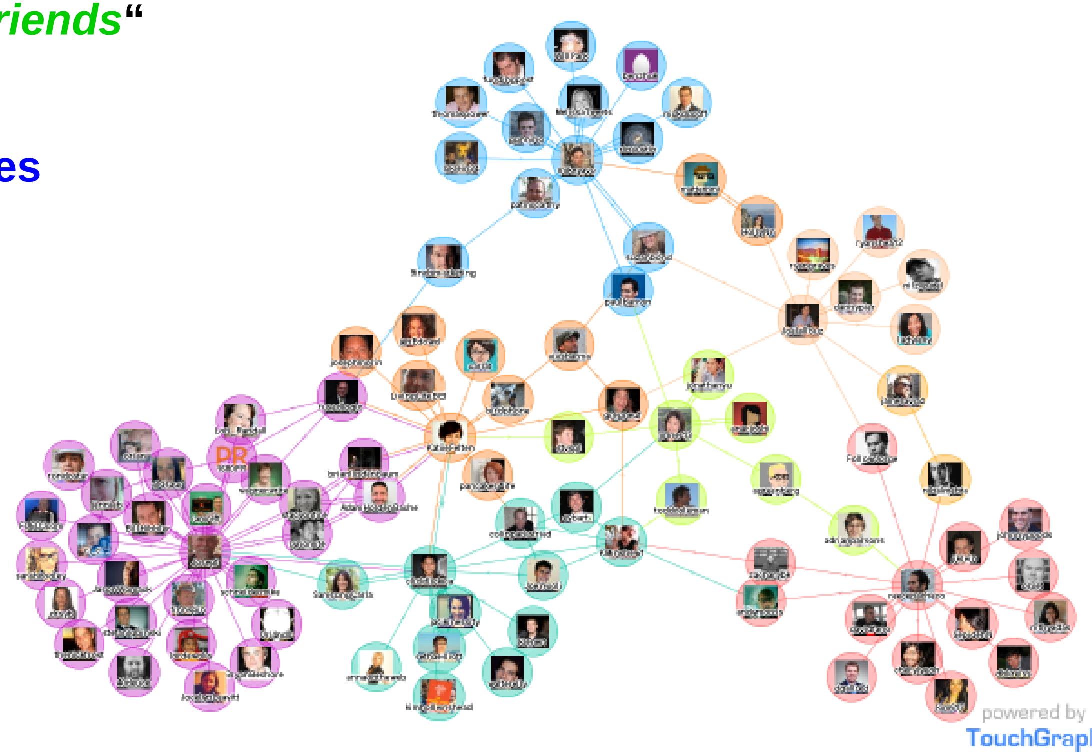


Clustering: Examples

- **Social network analysis**

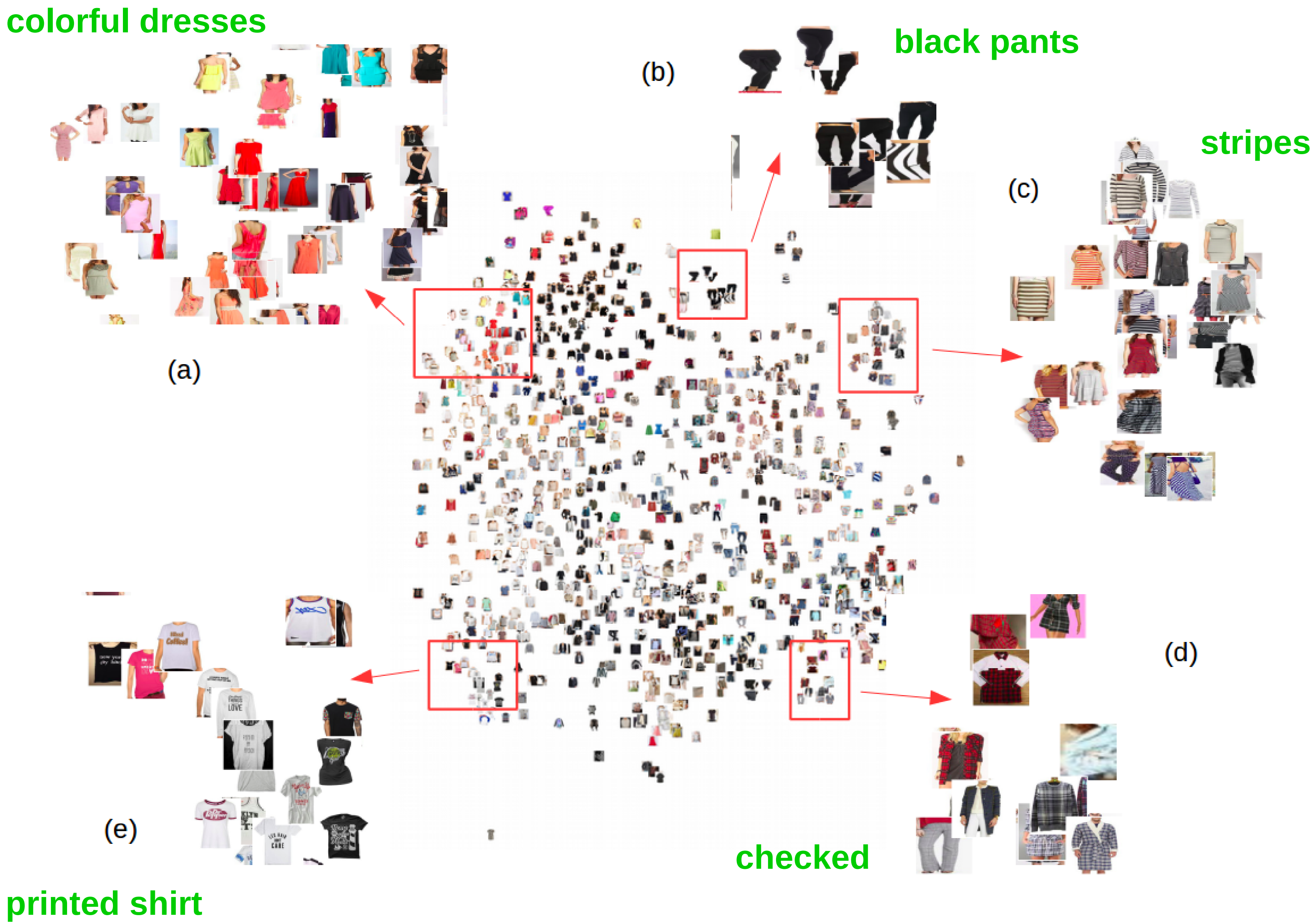
Input: {person, his interest, his likes, his groups, his friends links, etc.}

Clustering: into groups of “*friends*“
Facebook uses this to**:**
– recommend **“Add Friend”**
– recommend to **like new pages**
– put **new ads**



by TouchGraph

Clustering: Data analytics



Outline

- Outline of the module
- Introduction to machine learning
- Components of learning?
- Introduction to supervised learning
- Introduction to unsupervised learning
- **Feature representation**

Feature representation

- Features?
  - Example: what features used to predict the student's performance at the start of CSC4007.
    - data: examples/students studied in the last 2 years (~ 100 students)
    - **Helpful features**? GPA, performances in programming modules, CSC3060, CSC3061, etc.
    - Other features might cause **overfit**? Birth month/days, hair colors, etc.
  - Features never fully describe the situation

    "**All models are wrong, but some are useful**" – George Box

Features: example

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes

Initial model:

- Not enough information to generalize

Features: example

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes

Initial model:

- Egg laying
- Has scales
- Is poisonous
- Cold blooded
- No legs

Features: example

	Features					Label
Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes

Current model:

- Has scales
- Cold blooded
- No legs

Boa doesn't fit model, but is labeled as reptile.
Need to refine model

Features: example

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No

Current model:

- Has scales
- Cold blooded
- No legs

Features: example

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No
Alligator	True	True	False	True	4	Yes

Current model:

- Has scales
- Cold blooded
- Has 0 or 4 legs

Alligator doesn't fit model, but is labeled as reptile.
Need to refine model

Features: example

		Features				Label
Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No
Alligator	True	True	False	True	4	Yes
Dart frog	True	False	True	False	4	No

Current model:

- Has scales
- Cold blooded
- Has 0 or 4 legs

Features: example

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No
Alligator	True	True	False	True	4	Yes
Dart frog	True	False	True	False	4	No
Salmon	True	True	False	True	0	No
Python	True	True	False	True	0	Yes

Current model:

- Has scales
- Cold blooded
- Has 0 or 4 legs

No (easy) way to add to rule that will correctly classify salmon and python (since identical feature values)

Features: example

Features **Label**

Name	Egg-laying	Scales	Poisonous	Cold-blooded	# legs	Reptile
Cobra	True	True	True	True	0	Yes
Rattlesnake	True	True	True	True	0	Yes
Boa constrictor	False	True	False	True	0	Yes
Chicken	True	True	False	False	2	No
Alligator	True	True	False	True	4	Yes
Dart frog	True	False	True	False	4	No
Salmon	True	True	False	True	0	No
Python	True	True	False	True	0	Yes

Good model:

- Has scales
- Cold blooded

Not perfect, but no false negatives (anything classified as not reptile is correctly labeled); some false positives (may incorrectly label some animals as reptile)

Summary

- Today, we covered
  - Machine learning definition
  - Components of learning problems
  - Supervised learning vs. unsupervised learning
  - Feature representation in machine learning