Lecture 1: Introduction

General organization: People

- Lecturer: Mathieu Salzmann
- Teaching Assistants:
  - Sena Kiciroglu
  - Andrey Davydov
  - Krishna Nakka
  - Vidit Vidit
  - Benoît Guillard
  - Baran Ozaydin
- Student Assistants:
  - Julien Vignoud
  - Antoine Magron
  - Andrea Gilot

General organization: Schedule/Info

- Lectures: Tuesdays 8:15-10 in CE4
- Exercises: Fridays 8:15-10 in INJ218 and INM202
- Moodle: https://moodle.epfl.ch/course/view.php?id=16071
- Main references:
  - C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
  - M. Welling, A First Encounter with Machine Learning, 2011

General organization: Lectures/exercises

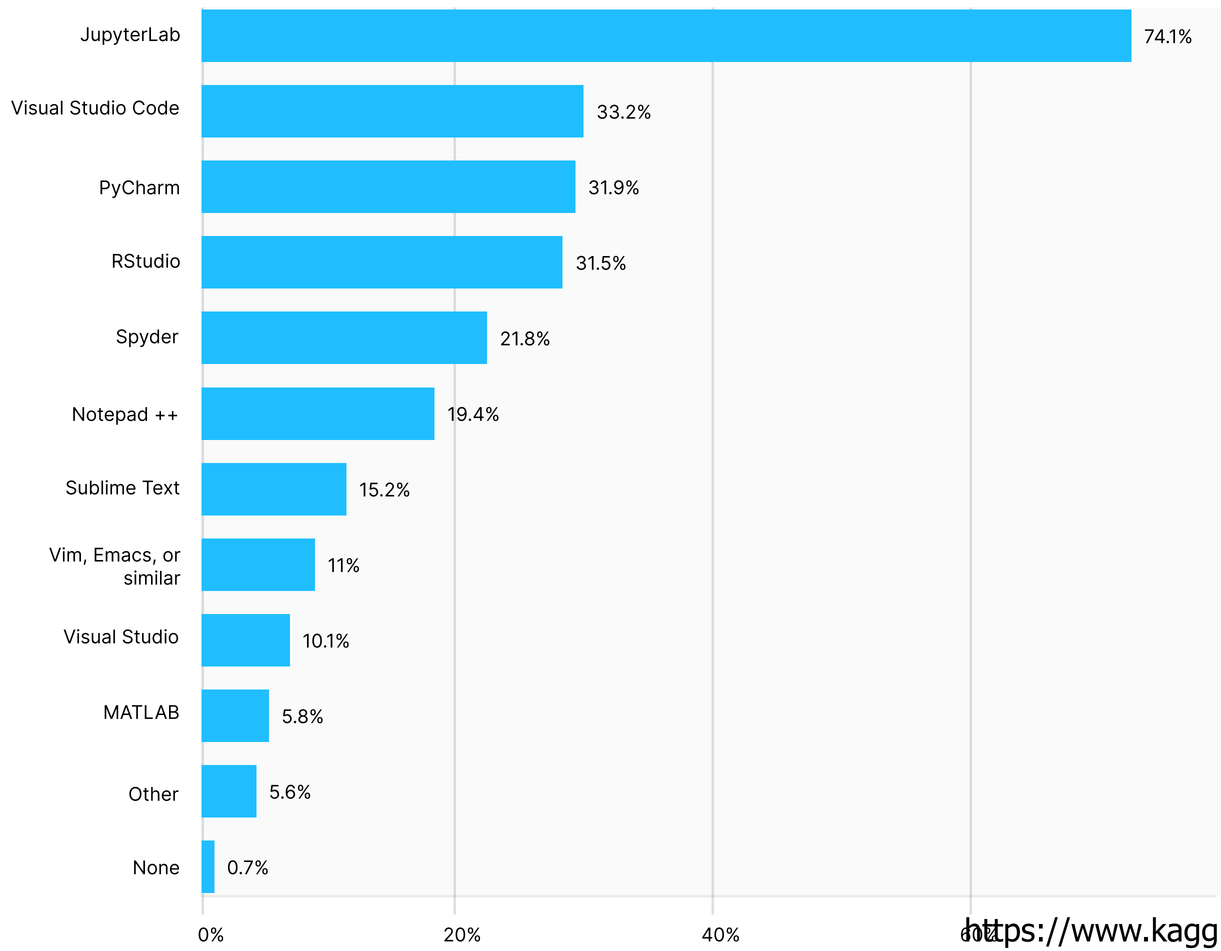
- The lectures will be taught in the classroom, streamed live and recorded

- The videos will be posted on SWITCHtube

- The exercise sessions will be managed in presence

General organization: Exercises

- The practical exercises will be done in Python
- According to a 2020 survey among data scientists regarding the most commonly used coding environments:



https://www.kaggle.com/kaggle-survey-2020

General organization: Evaluation

- Two graded exercise sessions
  - Tentative dates: 05.11.21 & 17.12.21
  - Each is worth 10% of the final grade
  - They will be done during the regular exercise session time

- Final exam: 80% of the grade

- Past years' exams and graded exercise sessions are available on Moodle

Course content

1. Introduction
    - ML Basics
2. Linear Supervised ML
    - Linear regression
    - Linear models for classification
3. Nonlinear Supervised ML
    - K-nearest neighbors
    - Feature expansion
    - Kernel methods
    - Artificial Neural Networks
4. Unsupervised ML
    - Linear dimensionality reduction
    - K-means clustering

Goals of today's lecture

- Introduce at a high level what Machine Learning is
- Introduce some basic Machine Learning concepts
  - These concepts will keep coming back throughout the semester
- Derive an initial formulation for linear regression

What is Machine Learning?

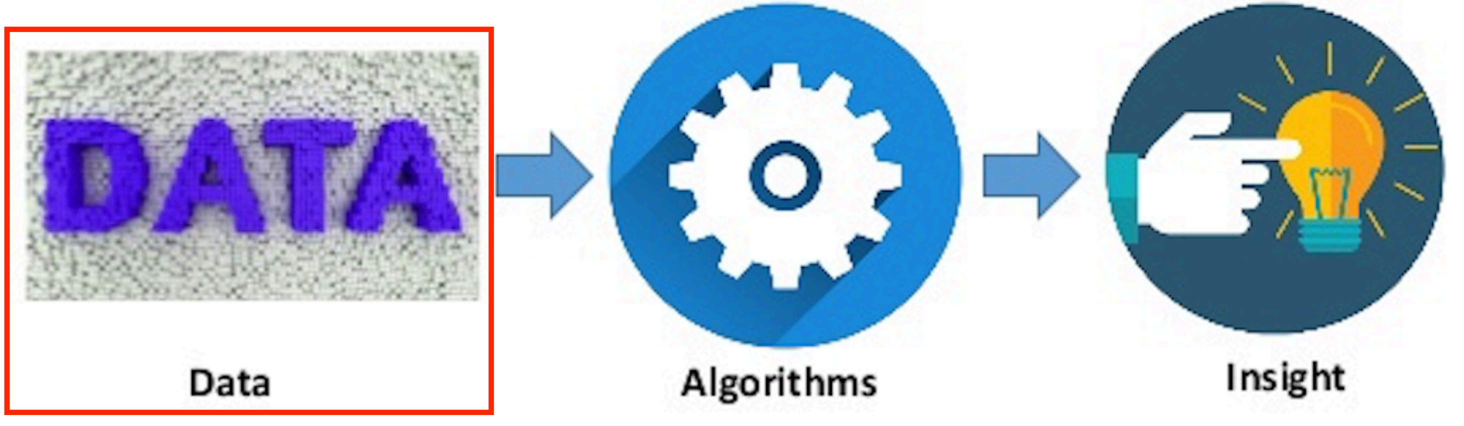
What is Machine Learning?

- Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.

- Machine learning algorithms seek to provide knowledge to computers through data, observations, and interaction with the world. It is then used to make accurate predictions given new observations.

- Machine Learning is applied statistics.

What is Machine Learning?



What data?

- Attributes: E.g., patient information related to births

	Age at delivery	Weight prior to pregnancy (pounds)	Smoker	Doctor visits during 1st trimester	Race	Birth Weight (grams)
Patient 1	29	140	Yes	2	Caucasian	2977
Patient 2	32	132	No	4	Caucasian	3080
Patient 3	36	175	No	0	African-Am	3600
*	*	*	*	*	*	*
*	*	*	*	*	*	*
Patient 189	30	95	Yes	2	Asian	3147

Image from Lumen Learning

What data?

- Text: E.g., Movie reviews

5	Column1
6	A very, very, very slow-moving, aimless movie about a distressed, drifting young man.
8	Not sure who was more lost - the flat characters or the audience, nearly half of whom walked out.
10	Attempting artiness with black & white and clever camera angles, the movie disappointed - became e
11	Very little music or anything to speak of.
13	The best scene in the movie was when Gerardo is trying to find a song that keeps running through his
15	The rest of the movie lacks art, charm, meaning... If it's about emptiness, it works I guess because it's
16	Wasted two hours.
18	Saw the movie today and thought it was a good effort, good messages for kids.
20	A bit predictable.
22	Loved the casting of Jimmy Buffet as the science teacher.
23	And those baby owls were adorable.
25	The movie showed a lot of Florida at it's best, made it look very appealing.
26	The Songs Were The Best And The Muppets Were So Hilarious.

Image from Integrated Knowledge Solutions

What data?

- Speech:

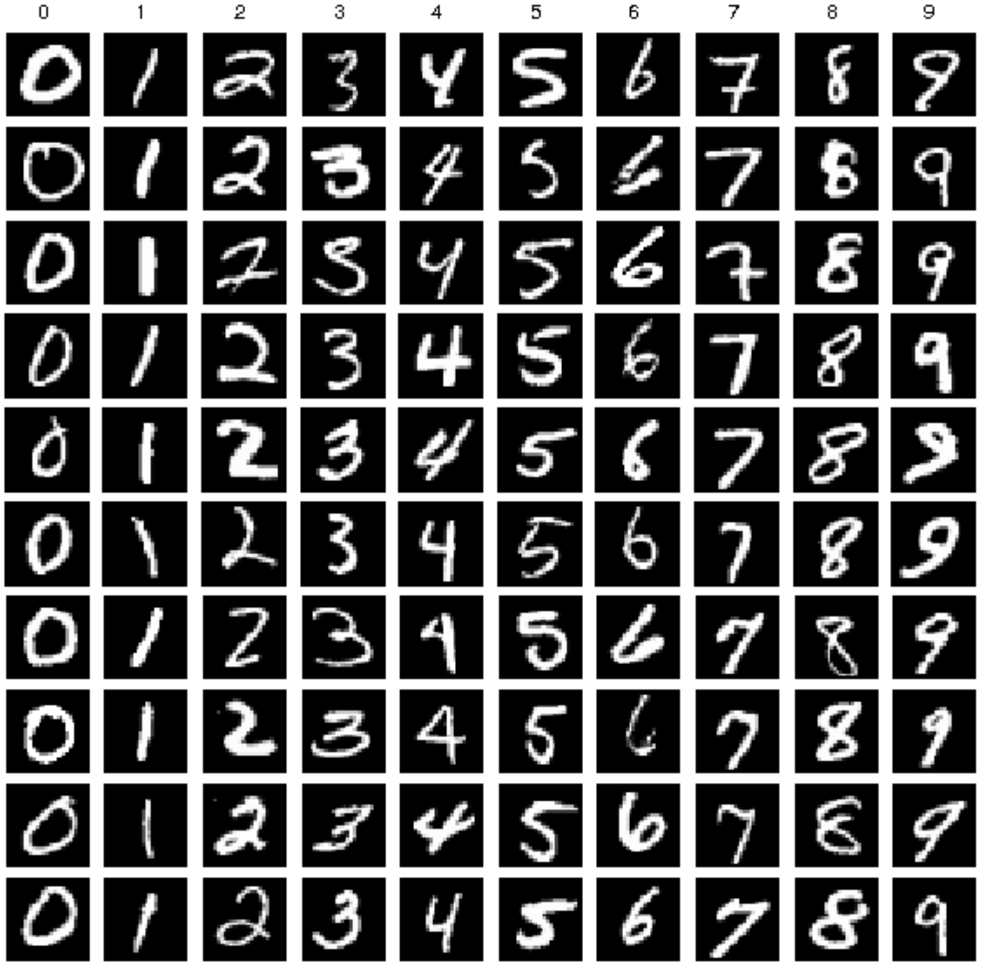
Sample 1: “without the dataset the article is useless”

Sample 2: “the boy looked out at the horizon”

Data from Nicholas Carlini

What data?

- Images: E.g., handwritten digits (MNIST dataset)



What data?

- Mixed: E.g., Collection from the Carnegie Museum of Art in Pittsburgh
  - https://github.com/cmoa/collection

Data set vs data sample

- Data sample (or example, or point): An individual observation. E.g., an image, a list of attributes for one patient

Patient 3	36	175	No	0	African-Am	3600

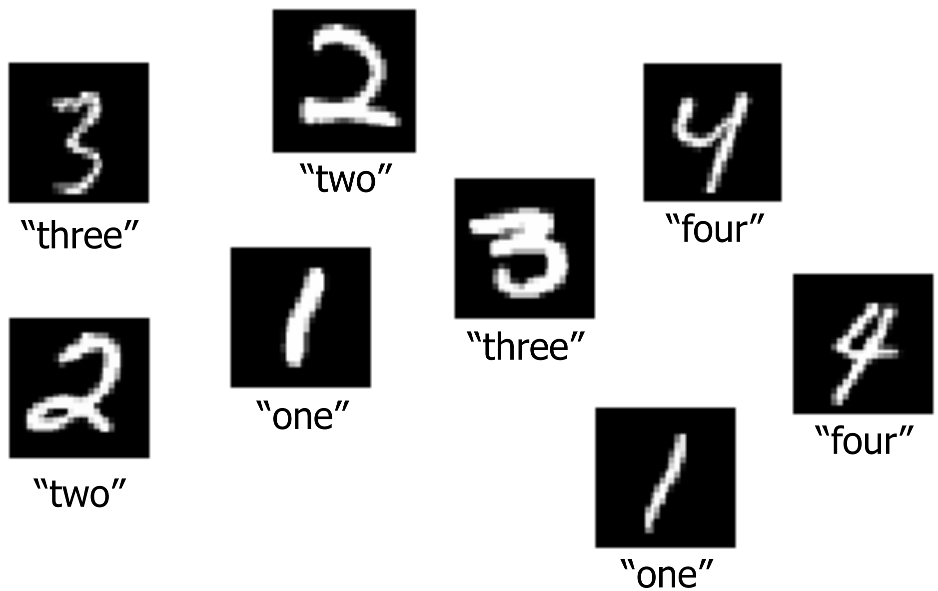
- Data set: A collection of multiple data samples. E.g., a collection of images, the attribute lists for multiple patients

Unsupervised data

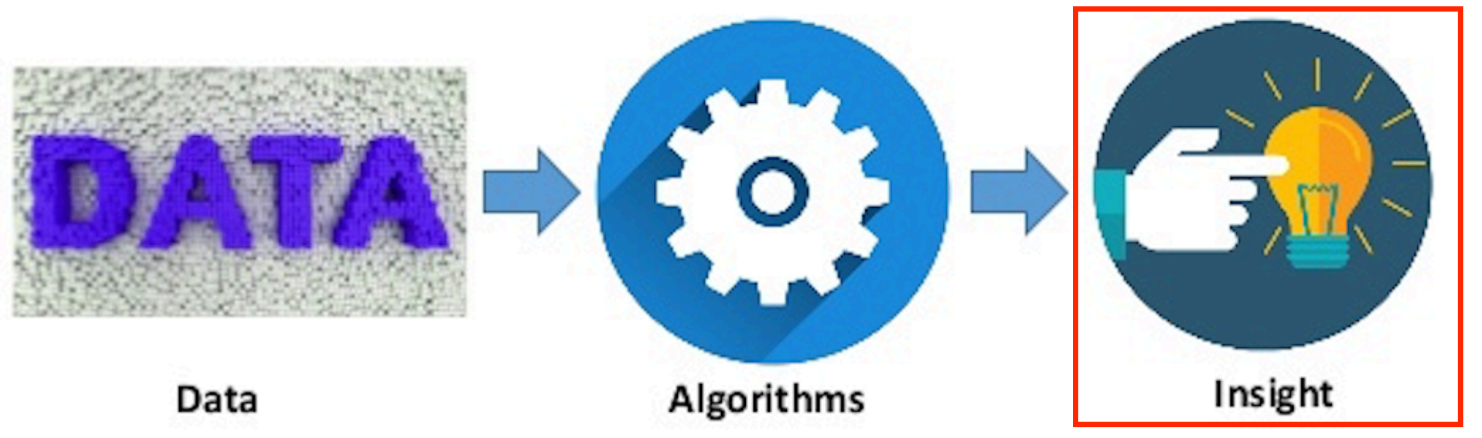
- Each sample consists only of an observation, e.g., an image (without information about its content)

Supervised data

- Each sample comes with additional annotations, e.g., category label

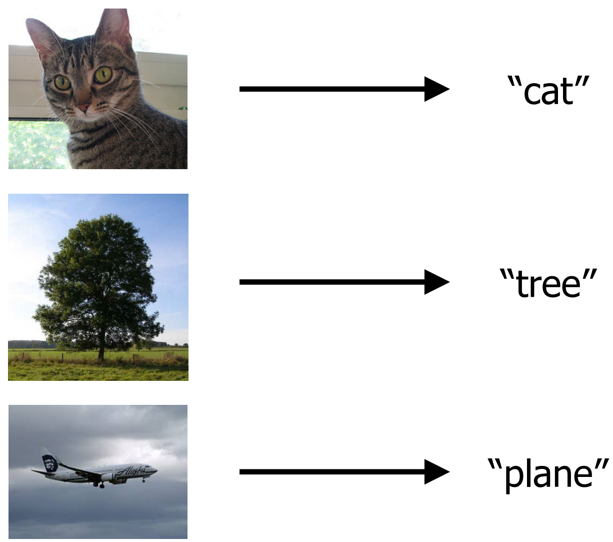


What is Machine Learning?



What insight?

- Precise, concrete prediction. E.g., the category depicted by an image



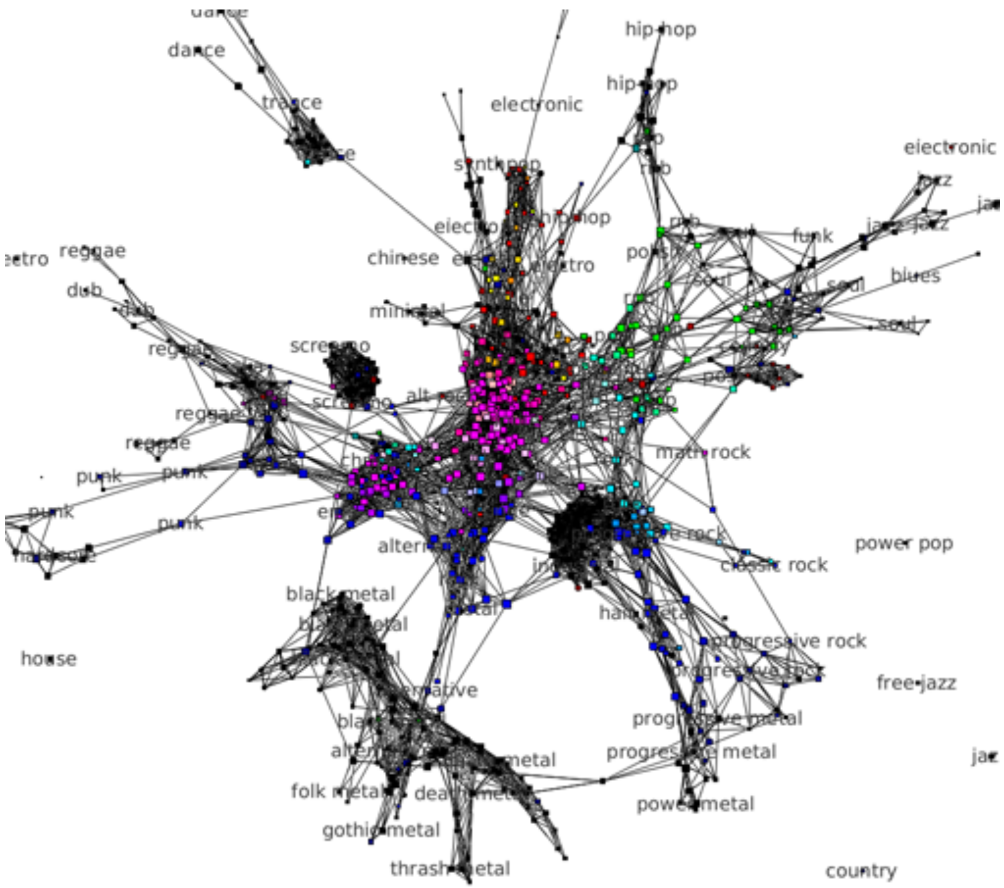
What insight?

- Better understanding of a phenomenon. E.g., identify the mother's characteristics that lead to low birth weight

	Age at delivery	Weight prior to pregnancy (pounds)	Smoker	Doctor visits during 1st trimester	Race	Birth Weight (grams)
Patient 1	29	140	Yes	2	Caucasian	2977
Patient 2	32	132	No	4	Caucasian	3080
Patient 3	36	175	No	0	African-Am	3600
*	*	*	*	*	*	*
*	*	*	*	*	*	*
Patient 189	30	95	Yes	2	Asian	3147

What insight?

- Better understanding of data. E.g., analyze the influences between different musical genres



What insight?

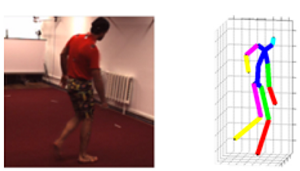
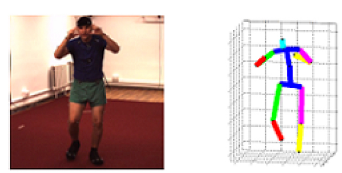
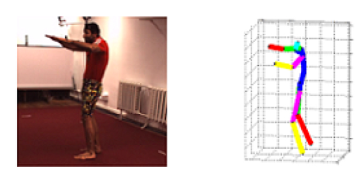
- In this course, we will mostly focus on precise insights. In particular, we will study two classes of ML problems:
  - Regression
  - Classification
- In the unsupervised learning part, we will nonetheless aim to obtain some form of data understanding

Regression vs Classification

- Regression: Predict continuous value(s) for a given sample
  - E.g., predict the birth weight of a baby (single value)

	Age at delivery	Weight prior to pregnancy (pounds)	Smoker	Doctor visits during 1^{st} trimester	Race	Birth Weight (grams)
Patient 1	29	140	Yes	2	Caucasian	2977
Patient 2	32	132	No	4	Caucasian	3080
Patient 3	36	175	No	0	African-Am	3600
*	*	*	*	*	*	*
*	*	*	*	*	*	*
Patient 189	30	95	Yes	2	Asian	3147

  - E.g., human pose estimation: predict the 3D positions of human joints (multiple values)

Regression vs Classification

- Classification: Predict one discrete label for a given sample
  - E.g., binary classification for movies (like vs not like; 1 vs 0)

5	Column1	Column2
6	A very, very, very slow-moving, aimless movie about a distressed, drifting young man.	0
8	Not sure who was more lost - the flat characters or the audience, nearly half of whom walked out.	0
10	Attempting artiness with black & white and clever camera angles, the movie disappointed - became e	0
11	Very little music or anything to speak of.	0
13	The best scene in the movie was when Gerardo is trying to find a song that keeps running through his	1
15	The rest of the movie lacks art, charm, meaning... If it's about emptiness, it works I guess because it's	0
16	Wasted two hours.	0
18	Saw the movie today and thought it was a good effort, good messages for kids.	1

  - E.g., multi-class image recognition (cat vs tree vs car; 0 vs 1 vs 2)


Cat


Tree


Car

Regression vs Classification

- In regression, the values typically follow an order
  - E.g., predicting a weight of 3002g instead of 2977g is better than predicting 2500g

	Age at delivery	Weight prior to pregnancy (pounds)	Smoker	Doctor visits during 1st trimester	Race	Birth Weight (grams)
Patient 1	29	140	Yes	2	Caucasian	2977

- In classification, the categories do not follow an order
  - E.g., predicting “tree” instead of “cat” is just as wrong as predicting “car”
  - Even when categories are represented with numbers (e.g., 1, 2, 3), predicting category 2 instead of 1 is as wrong as predicting category 3

Cat

Example

- Image recognition:
  - http://scs.ryerson.ca/%7Eaharley/vis/fc/

Example

- Text analysis:
  - https://natural-language-understanding-demo.ng.bluemix.net

Example

- Image captioning:
  - https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

Example

- Human pose estimation:
  - https://storage.googleapis.com/tfjs-models/demos/posenet/camera.html?source=post_page---------------------------
  - From EPFL: https://vitademo.epfl.ch

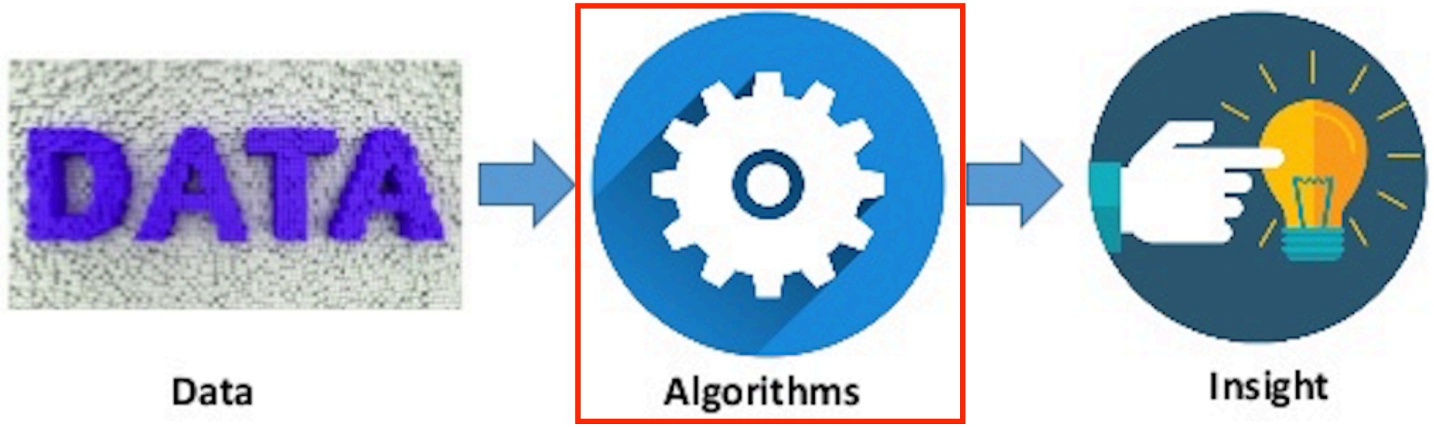
Example

- Image to image translation:
  - https://affinelayer.com/pixsrv/

Other fun demos (for you to play)

- AI Pictionary:
  - https://quickdraw.withgoogle.com
- Semantris: ML-based word association games
  - https://research.google.com/semantris/
- Talk to books:
  - https://books.google.com/talktobooks/

What is Machine Learning?

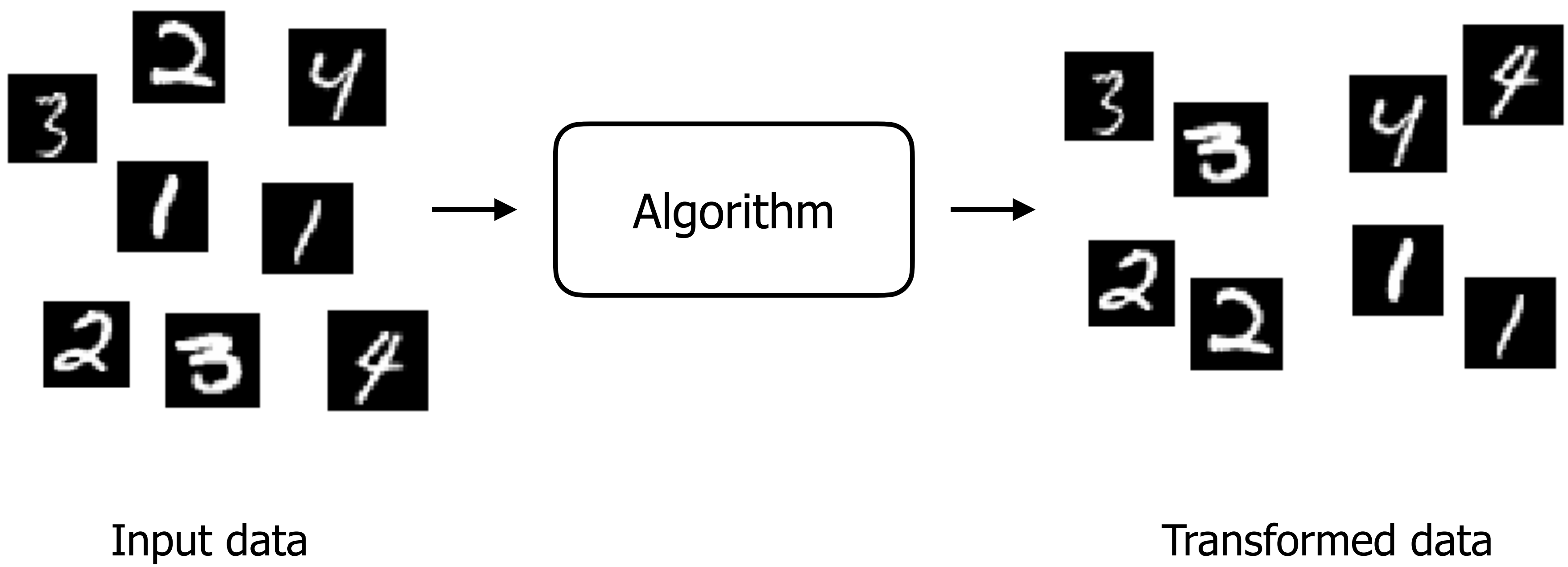


What (classes of) algorithms?

- Supervised learning
  - Relies on supervised data
  - The annotations typically correspond to the desired insight
- Unsupervised learning
  - Relies on unsupervised data
  - The goal is rather to analyze the observed data set
- (Reinforcement learning)
  - Learn to react to the environment
  - Not covered in this course

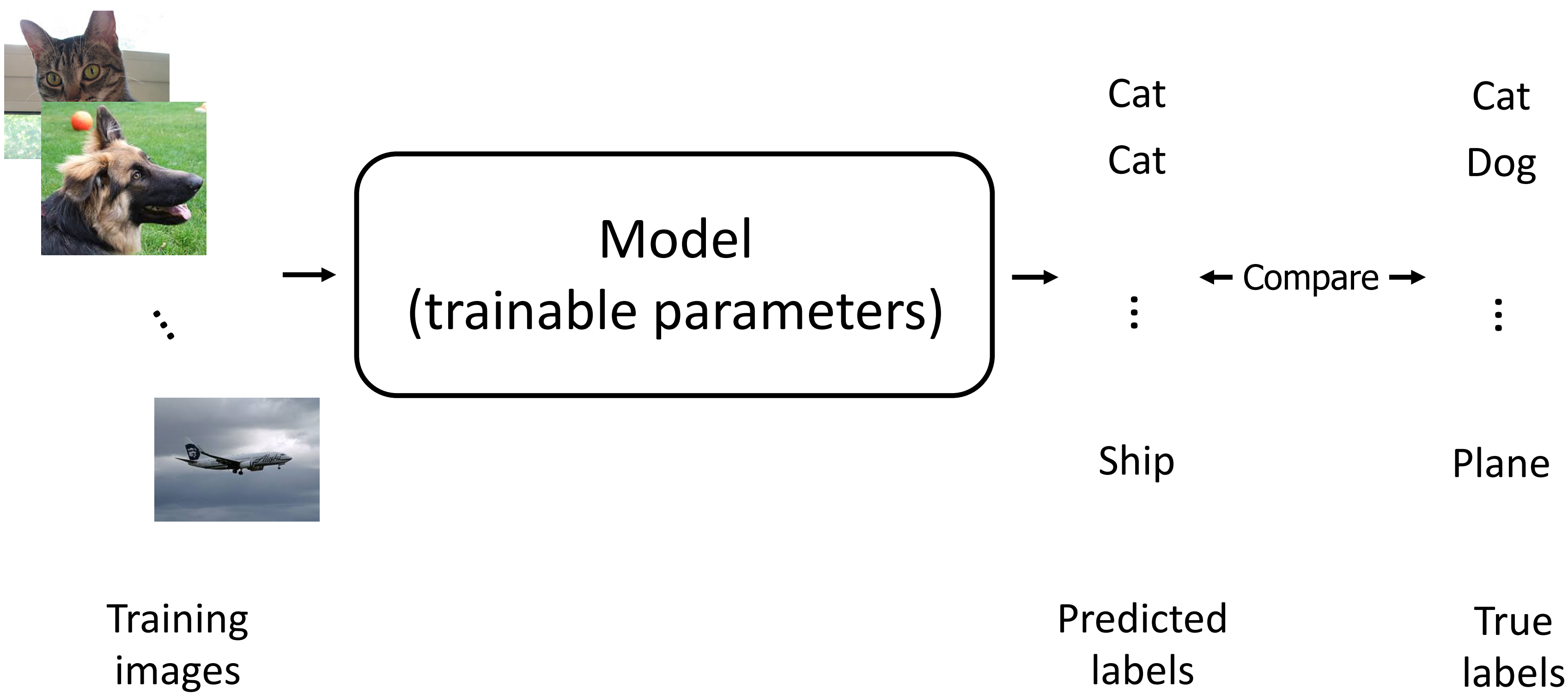
Unsupervised learning

- A single stage: Transform the data for further analysis



Supervised learning

- Stage 1: Training: Use data with ground-truth labels to optimize model parameters



Supervised learning

- Stage 2: Testing: Predict the output for a new data sample

Test image

Model
(fixed parameters)

Plane

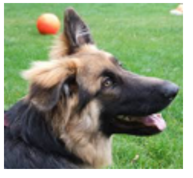
Predicted label

Supervised learning: Assumption

- The annotations (supervisory signal) are the insight that we seek to obtain from the data
  - Example: Category label

Cat

Car



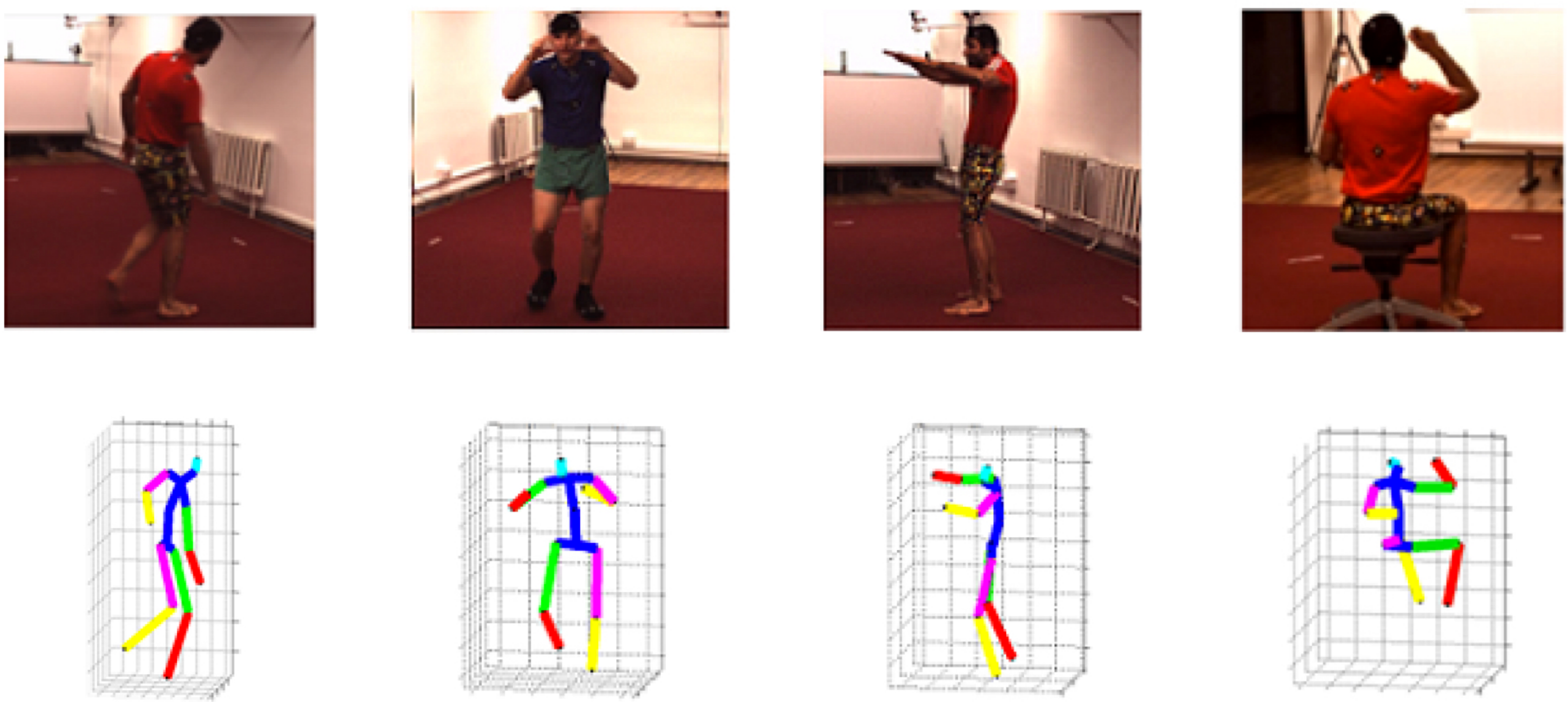
Dog

Plane

Chair

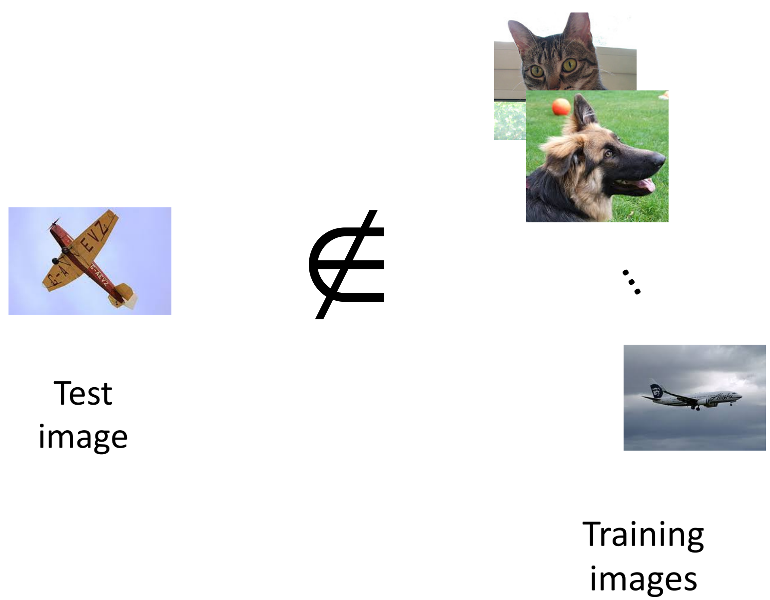
Supervised learning: Assumption

- The annotations (supervisory signal) are the insight that we seek to obtain from the data
  - Example: Human pose



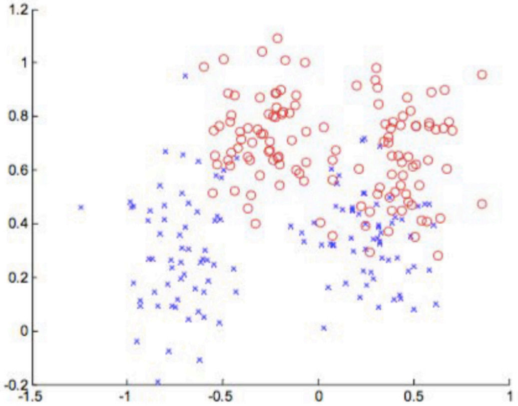
Training set vs test set

- **The training set and the test set should always be completely separate!**
- Never use the test annotations during training
  - Using the test observations (inputs) is occasionally possible and referred to as transductive learning (we will not do it in this course)

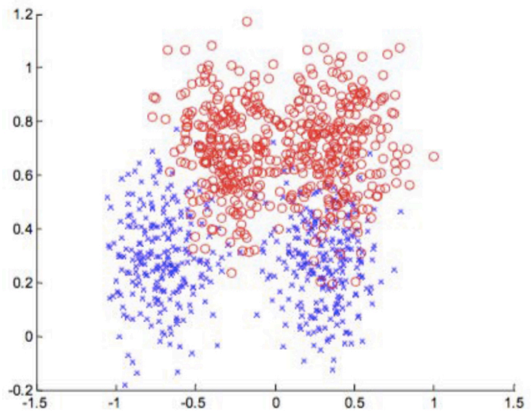


Training set vs test set

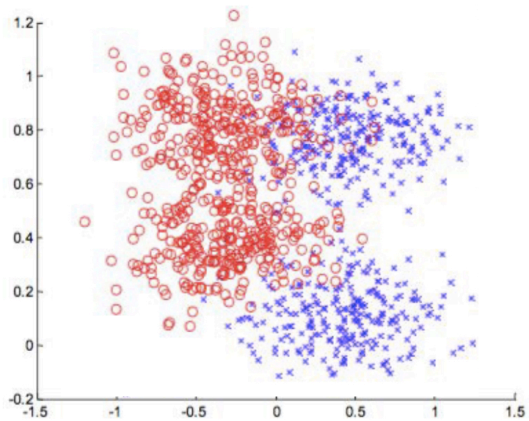
- Assumption: Training and test samples are drawn from the same statistical distribution
  - E.g., synthetic data: 2D inputs with 2 classes (colors)



Training data



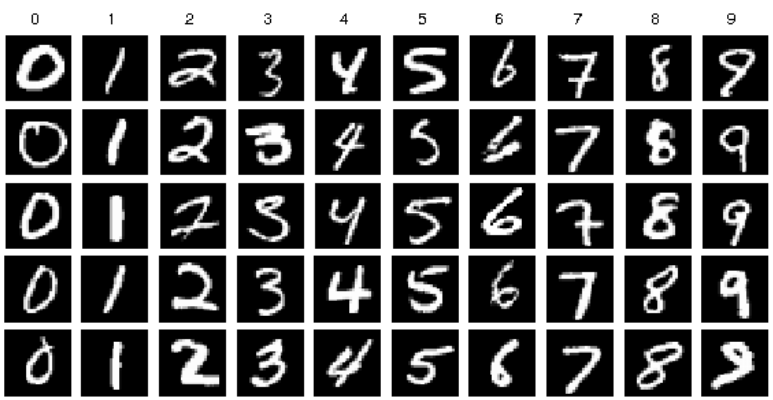
Test data from same distribution



Test data from different distribution

Training set vs test set

- Training and test samples are drawn from the same statistical distribution
  - E.g., digit recognition: A model trained on MNIST will work poorly on Street View House Numbers



Training data

Test data

Exercises

- Given a dataset where each sample is represented with a list of meteorological measurements, the goal is to predict the burned area of the corresponding forest fire
    - If you are given the ground-truth burned areas for a set of training samples, what general class of algorithms would you use to solve this problem?
    - What type of Machine Learning problem is this?

Your first machine learning model

- Let's start with a simple supervised learning model:

The Linear Model

- We will spend a few weeks on this
  - This will allow us to also cover general ML topics

Notation

We denote the i^{th} data sample (input) in the collection of N samples as

$$\mathbf{x}_i \in \mathbb{R}^D,$$

a vector of dimension D

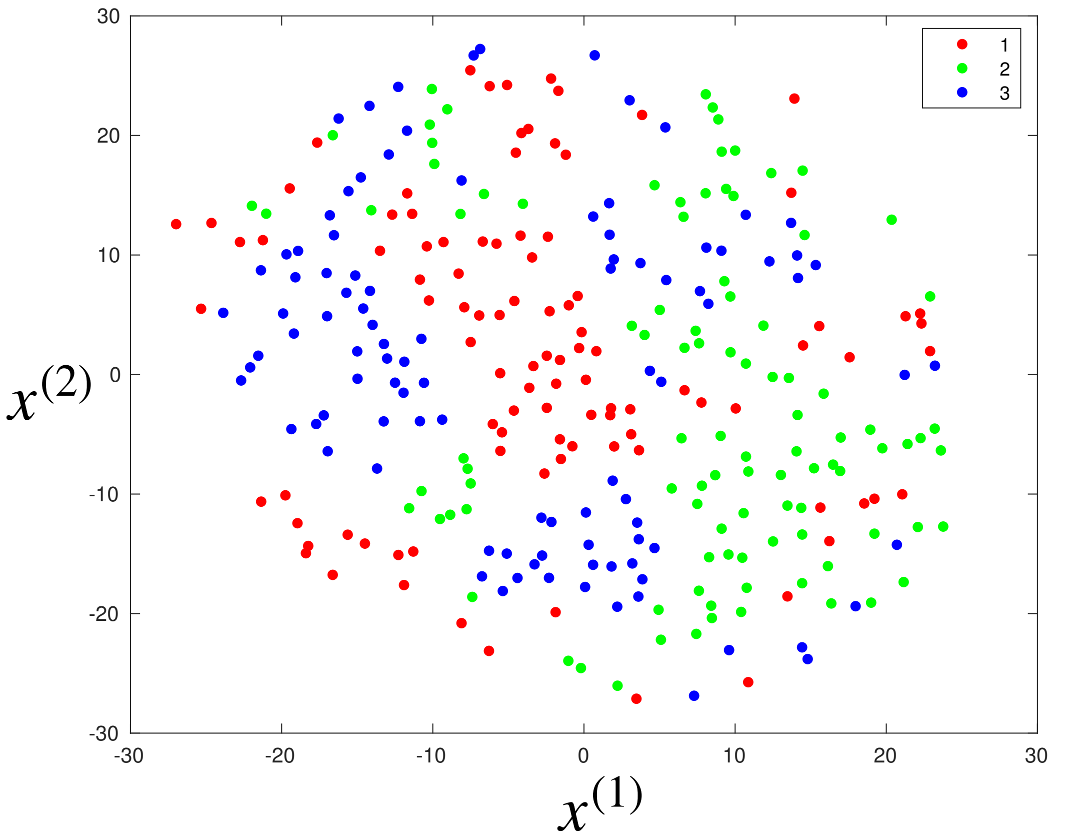
We denote the i^{th} label (output) in the collection of N samples as $\mathbf{y}_i$

For classification, $\mathbf{y}_i$ is represents a single discrete value

For regression, $\mathbf{y}_i$ can be a single continuous value, or a vector of dimension C

Notation made concrete

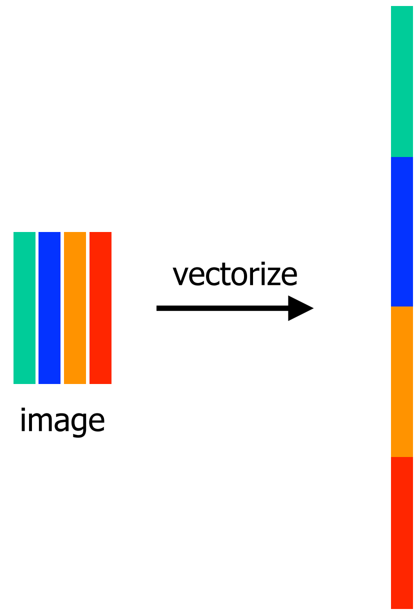
- In the following toy example, $\mathbf{x}_i$ is one 2D point, and thus $D = 2$
- $\mathbf{y}_i$ is a discrete value indicating the class (color), i.e., 1, 2, or 3



Notation made concrete

- In the digit recognition example, $\mathbf{x}_i$ is a grayscale image. If it has height H and width W, then $D = H \cdot W$

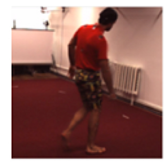
$\mathbf{x}_i = \text{vectorize}($  $)$



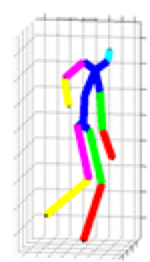
- $\mathbf{y}_i$ is a single discrete value indicating the digit, e.g., $\mathbf{y}_i = 2$

Notation made concrete

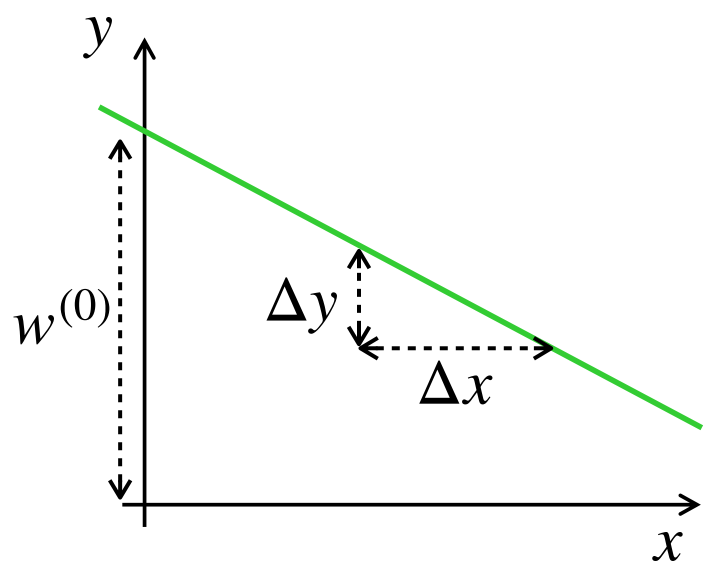
- In the human pose example, $\mathbf{x}_i$ is a color image. If it has height H and width W, then $D = 3 \cdot H \cdot W$

$\mathbf{x}_i =$ vectorize()

- Then, $\mathbf{y}_i$ is a human pose. If a human pose is defined as a skeleton with 12 joints (wrists, elbows,…), and each joint is a 3D point, then $C = 3 \cdot 12 = 36$

$\mathbf{y}_i =$ 

A simple parametric model: The line

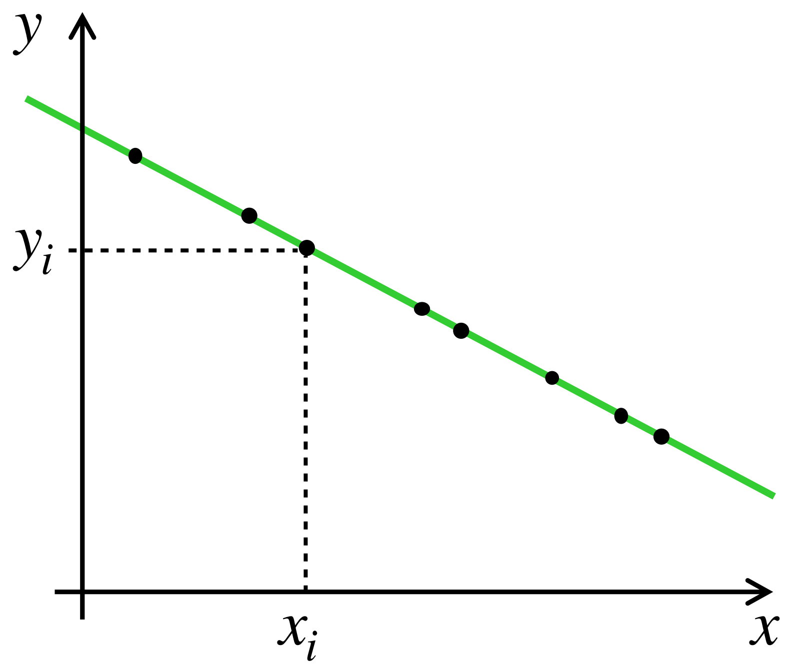


- Defined by 2 parameters
  - The y-intercept $w^{(0)}$
  - The slope $w^{(1)} = \dfrac{\Delta y}{\Delta x}$
- Mathematically, a line is expressed as

$$y = w^{(1)}x + w^{(0)}$$

Line fitting

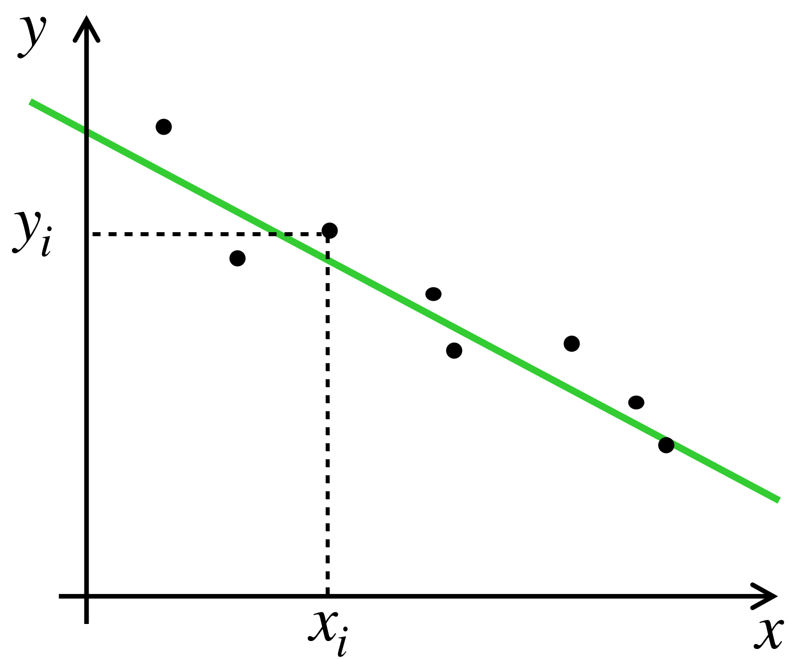
- Given N pairs $\{(x_i, y_i)\}$, find the line that passes through these observations



- This ideal case never occurs in practice

Line fitting with noise

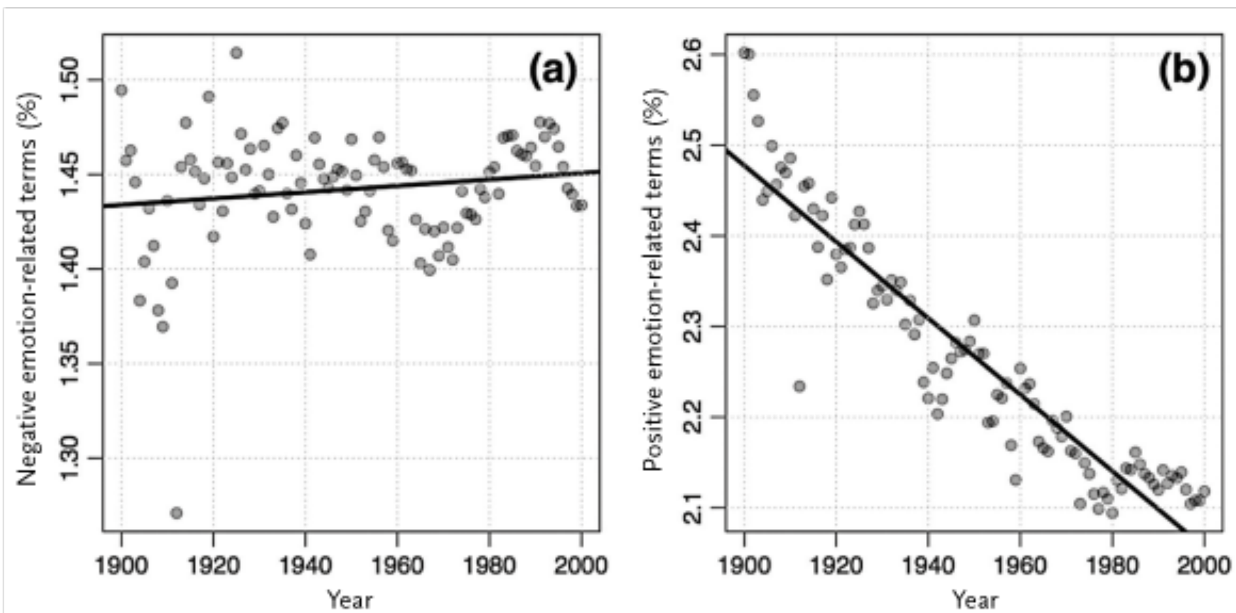
- Given N pairs $\{(x_i, y_i)\}$ of noisy measurements, find the line that best fits these observations



- This process is called *linear regression*

1D linear regression: Example

- Discover trends:
  - Example: Proportion of negative and positive emotions in anglophone fiction (Morin & Acerbi, 2016. Figure from Moretti & Sobchuk, 2019)



1D Linear regression: Training

- In essence, fitting a line consists of finding the best line parameters $w^{(0)*}$ and $w^{(1)*}$ for some given data

- This corresponds to the training stage:
    - Given N training pairs $\{(x_i, y_i)\}$, we aim to find $(w^{(0)*}, w^{(1)*})$, such that the predictions of the model

$$\hat{y}_i = w^{(1)*}x_i + w^{(0)*}$$

    are close to the true values y_i

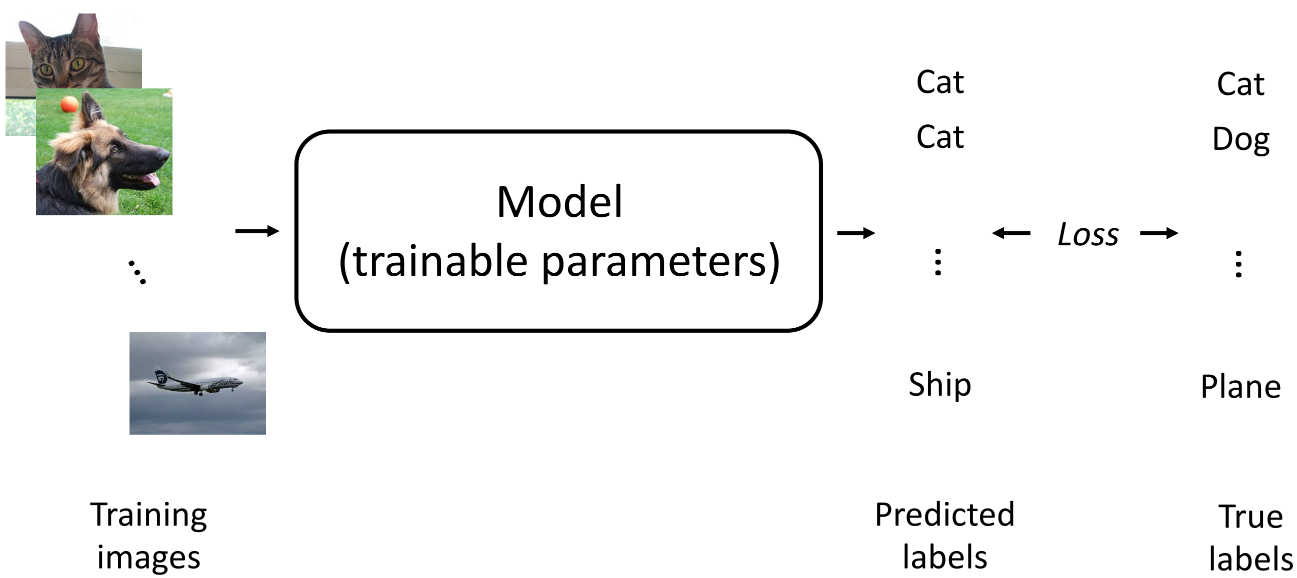
- We then need to define a measure of closeness between y_i and $\hat{y}_i$

Interlude

Loss Function and Empirical Risk

Loss function

- The loss function $\ell(\hat{y}_i, y_i)$ computes an error value between the prediction and the true value
  - This is a general ML concept, not only for linear regression
  - We will see different loss functions in the upcoming lectures



Empirical risk

- Given N training samples $\left\{(\mathbf{x}_i, y_i)\right\}$, the empirical risk is defined as

$$R(\{\mathbf{x}_i\}, \{y_i\}, \mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} \ell(\hat{y}_i, y_i)$$

  where $\{\mathbf{x}_i\}$ and $\{y_i\}$ are the sets of training inputs and labels, respectively

- During training, our goal is to find the parameters $\mathbf{w}$ (e.g., $w^{(0)}$ and $w^{(1)}$) that minimize the empirical risk
  - Note that the risk depends on $\mathbf{w}$ via $\hat{y}_i$

Minimizing the risk

- This is expressed as the optimization problem

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \ell(\hat{y}_i, y_i)$$

- We can also write that the best parameters are the solution

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \ell(\hat{y}_i, y_i)$$

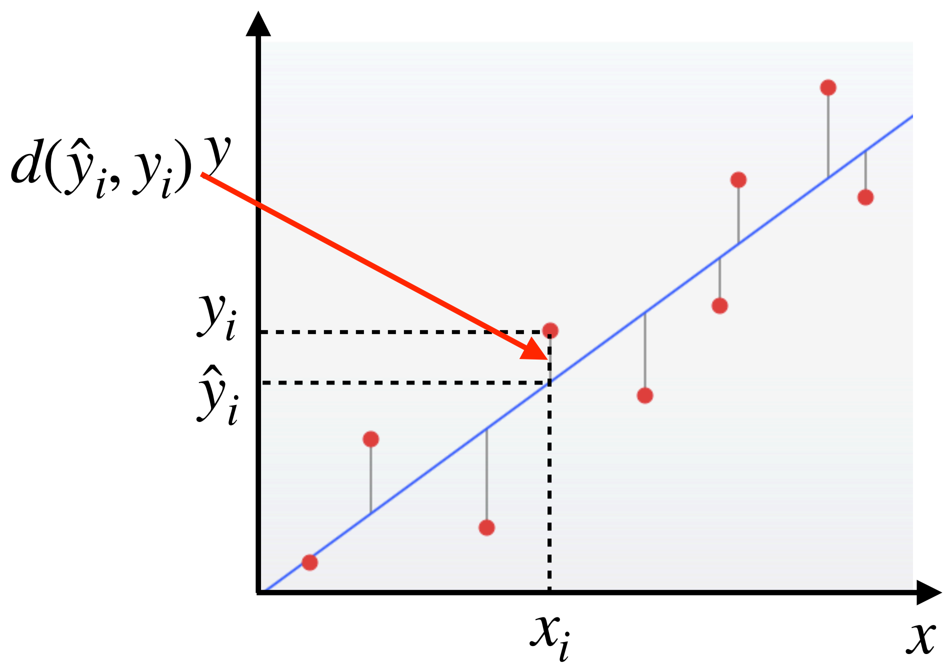
End of the interlude

Back to Linear Regression

1D Linear regression: Training

- A natural measure of closeness is the Euclidean distance

$$d(\hat{y}_i, y_i) = \sqrt{(\hat{y}_i - y_i)^2}$$



- The difference between $\hat{y}_i$ and y_i is often referred to as the *residual*

1D Linear regression: Training

- In practice, one often prefers using the squared Euclidean distance

$$d^2(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$$

- Training can then be expressed as the *least-squares* minimization problem

$$\min_{w^{(0)},w^{(1)}} \frac{1}{N} \sum_{i=1}^{N} d^2(\hat{y}_i, y_i)$$

where $\hat{y}_i$ depends on $w^{(0)}$ and $w^{(1)}$

- We will see how to solve this problem next week

1D linear regression: Demo

- http://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html

1D Linear regression: Prediction

- Once you found the best line for the given N observations, you can use it to predict a y value for a new x

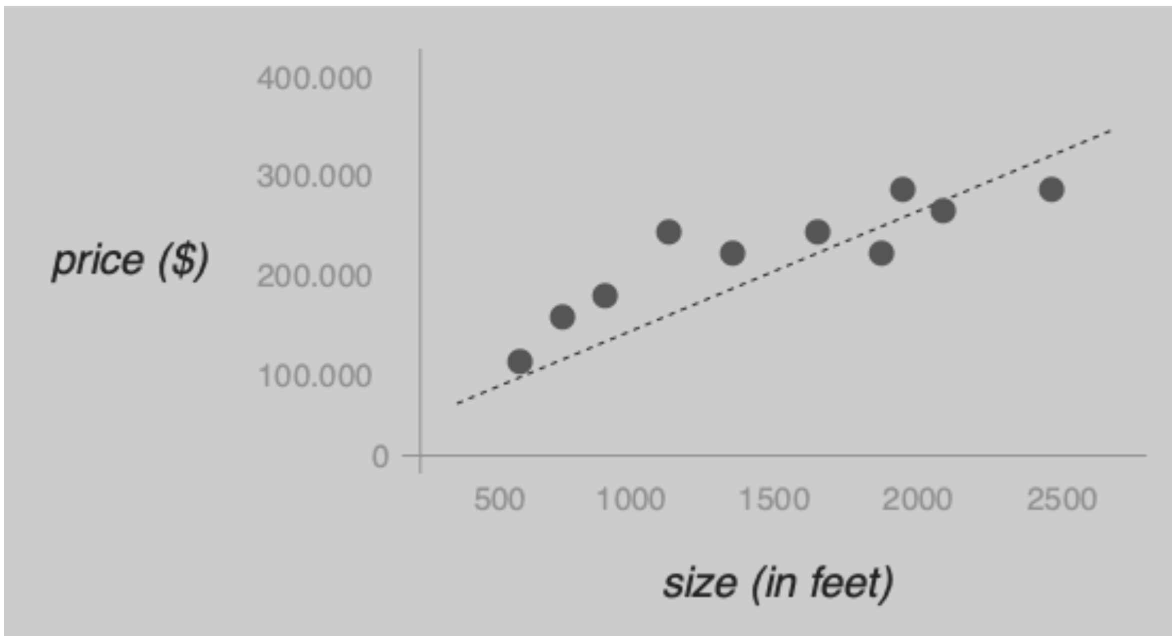
- Let $w^{(0)*}$ and $w^{(1)*}$ be the best line parameters given the observations

- Then, for any value x, you can predict an estimate of the corresponding y as

$$\hat{y} = w^{(1)*}x + w^{(0)*}$$

1D linear regression: Example

- Predict quantities
  - Predict the price of a house based on its size (example from https://www.internalpointers.com/post/linear-regression-one-variable)



  - With temporal trends, one can predict what will happen in the future

Model evaluation

- Once an ML model is trained, one would typically understand how well it performs on unseen test data
  - At this stage, the parameters of the model are fixed
  - Recall that the training and testing data must be separated!

- During this evaluation, one compares the predictions of the model with the true annotations of the test data
  - In contrast to the training stage, the model parameters are *not* updated
  - The evaluation metric may directly be the loss function, but may also differ from it

Evaluation metrics for regression

- Mean Squared Error (MSE)
  - Same as the loss function but for N_t test samples

$$MSE = \frac{1}{N_t} \sum_{i=1}^{N_t} (\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ is the prediction for test sample i and y_i the corresponding ground-truth value

- Root Mean Squared Error (RMSE)
  - Square-root of the MSE

Evaluation metrics for regression

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{N_t} \sum_{i=1}^{N_t} |\hat{y}_i - y_i|$$

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{N_t} \sum_{i=1}^{N_t} \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

Taking a percentage w.r.t. the true value might be easier to interpret

Exercises

- Given the following dataset for birth weight prediction:

	Age at delivery	Weight prior to pregnancy (pounds)	Smoker	Doctor visits during 1st trimester	Race	Birth Weight (grams)
Patient 1	29	140	Yes	2	Caucasian	2977
Patient 2	32	132	No	4	Caucasian	3080
Patient 3	36	175	No	0	African-Am	3600
*	*	*	*	*	*	*
*	*	*	*	*	*	*
Patient 189	30	95	Yes	2	Asian	3147

  - How many samples (N) can you assume this dataset to contain?
  - What is the dimensionality (D) of the input to the ML model?
  - What is the dimensionality (C) of the output of the ML model?

Survey

- Please fill in the survey on the Moodle page to comment on the pace of the lecture