Department of COMPUTER SCIENCE

WARWICK
THE UNIVERSITY OF WARWICK

Machine Learning
CS342

Lecture 1: Introduction to ML

Dr. Theo Damoulas
T.Damoulas@warwick.ac.uk

Office hours (CS 307)

Mon 16:00-17:00 Fri 16:00-17:00

Module Organisation

- TAs/Tutors:
  - Helen McKay: H.McKay@warwick.ac.uk
  - Joe Meagher: J.Meagher@warwick.ac.uk
  - Karla M. Gomez: K.Monterrubio-Gomez@warwick.ac.uk
  - Jevgeni Gamper: J.Gamper@warwick.ac.uk
- Module website:
  - https://www2.warwick.ac.uk/fac/sci/dcs/teaching/modules/cs342/
  - Check Syllabus and Online Material
- Assessment:
  - 60% Final exam (Lecture material & PPs & Worksheets)
  - 40% Coursework:
    - 15% First assignment
    - 25% Second assignment
- Help? Questions? Use the ML Forum or contact your TAs and myself

Module Organisation

- Lectures:
  - Monday, 17:00-18:00, PLT
  - Thursday, 17:00-18:00, H0.52
  - Friday, 11:00-12:00, L3

- Labs:
  - Monday, 11:00-13:00, CS0.06
  - Thursday, 09:00-11:00, CS0.06 **& CS0.03**

  Necessary to understand material and practise ML in Python
  Lab components in assignments

- Seminars:
  - See Tabula allocation

  Worksheets build up your methodology and math understanding
  Questions (or variations of them) may be in Exam

- Background:
  - Linear Algebra, Probability Theory, Programming
  - Work with and help each other - we run Plagiarism detection.
  - Come at office hours, ask me or your TAs!

Learning Outcomes & Goals of CS342

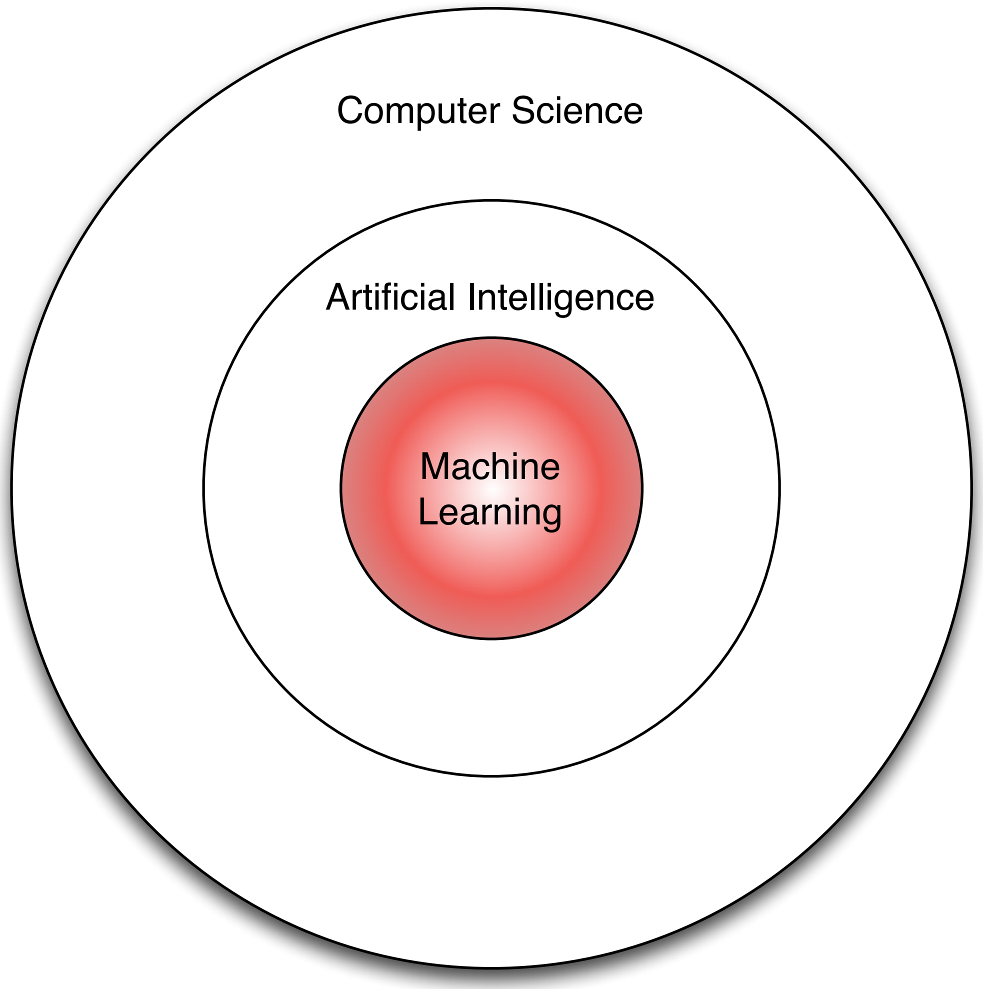
- What is Machine Learning?
  - Main areas, subfields, applications
- Learning models from data - basic principles
- Understand a wide variety of learning algorithms
- Understand how to fit and evaluate learning algorithms
- Apply various learning algorithms to real problems
- Have fun while teaching your computer how to learn from data!

Syllabus

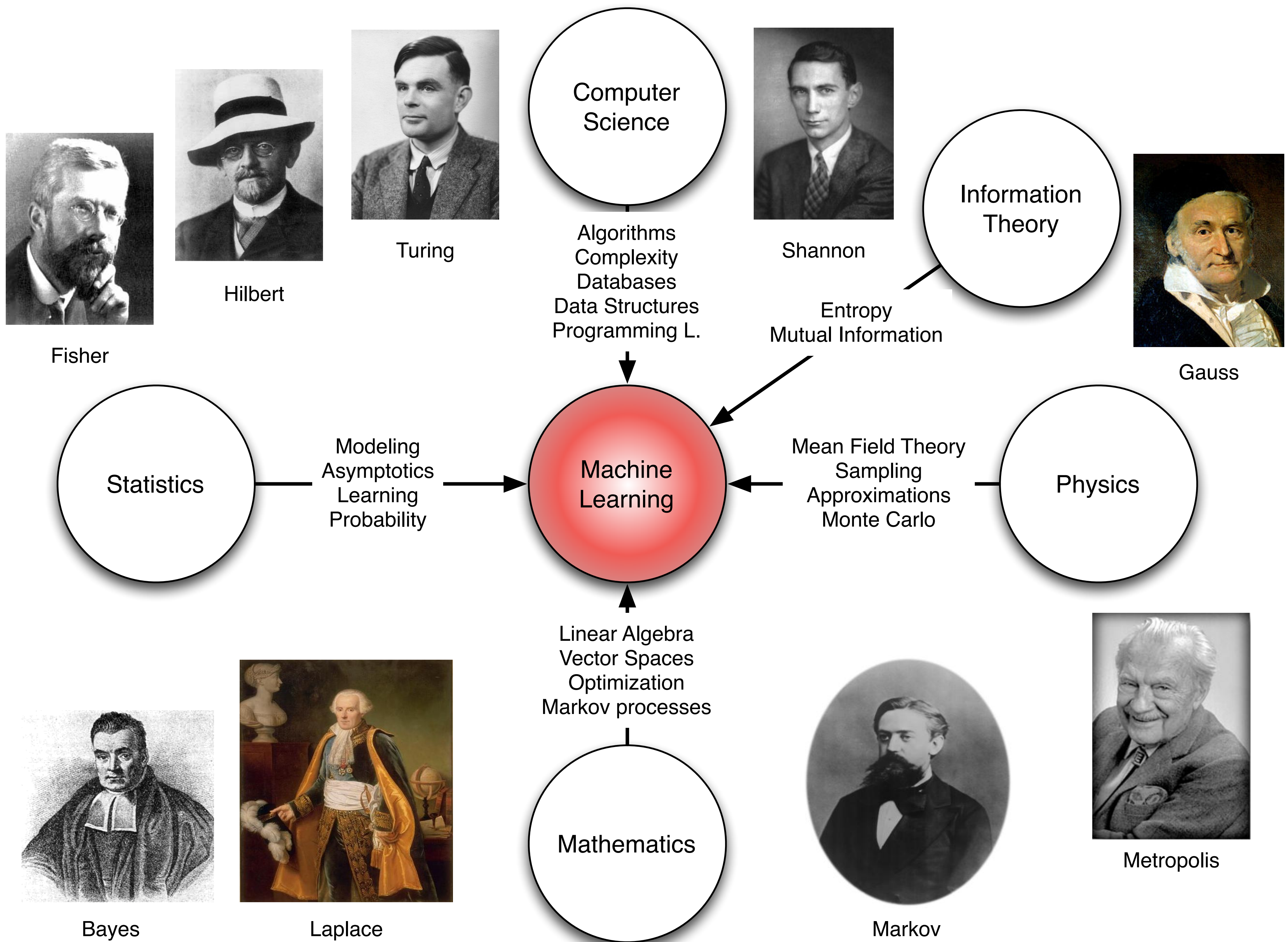
- ***A First Course in Machine Learning***, 2nd Edition, S. Rogers & M. Girolami [1st Chapter] http://www.dcs.gla.ac.uk/~srogers/firstcourseml/
- ***Machine Learning***, T. Mitchell
- ***Pattern Recognition and Machine Learning***, C. Bishop
- ***Pattern Classification***, Duda, Hart and Stork, Wiley-interscience
- *Machine Learning: A Probabilistic Perspective*, K. P. Murphy
- *Bayesian Reasoning and Machine Learning*, D. Barber

What fields make up Machine Learning?

- Computer Science?
- Statistics?
- Mathematics?
- Physics?
- Computational Neuroscience?
- Computational Psychology?



What fields make up Machine Learning?



Department of COMPUTER SCIENCE

Some AI/ML History

(ENIAC/Enigma) 1940 - 1950

(First learning system) 1952

Connectionism

(Perceptron) 1957

(Machine Translation) 1966

(k-NN) 1967

Symbolic

(Lighthill report) 1973

Expert Systems

(LISP machines) 1987

Statistical ML

SVM 1993

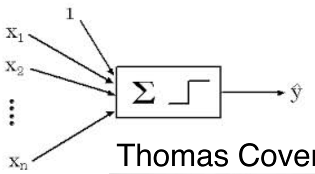
Probabilistic ML

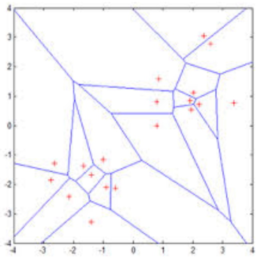
OR/OPT ML

Connectionism #2 (Deep L.)

Arthur Samuels

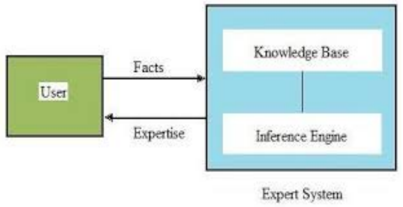
Frank Rosenblatt

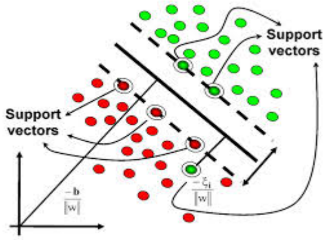




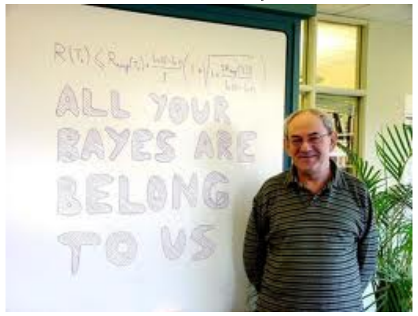
Thomas Cover

Edward Feigenbaum

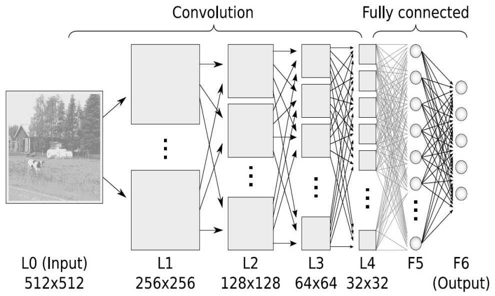




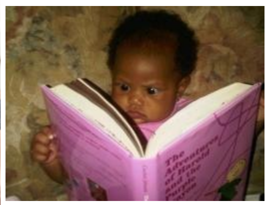
Vladimir Vapnik

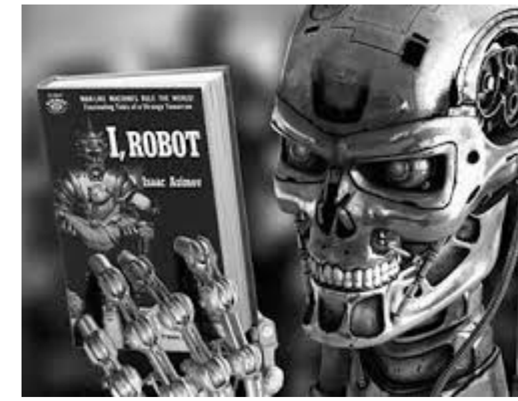


Geoff Hinton



Human vs Machine Learning





How do humans learn…?

- Supervisory role?
- Unsupervised [grouping, similarity, patterns]
- Internal reward system [dopamine]
- Neural structure [Hebbian learning]
- Classical Conditioning (Pavlov's dog)

What is the common basis/goal?

- Study of systems and algorithms that “learn from data”

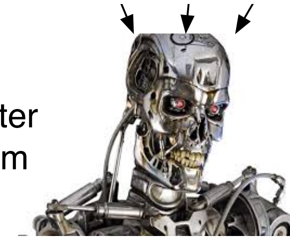
- *A computer program is said to learn from experience* E (data) *with respect to some class of tasks* T *and performance measure* P, *if its performance at tasks in* T, *as measured by* P, *improves with experience* E. [Tom Mitchell, 1998]

E: Data on Sarah




Learning Process

Computer Program



T: Find Sarah

P: % Correct recognitions

Department of COMPUTER SCIENCE

Lets try

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

You are given an algorithm that watches which phone-calls you mark as "noise complaints" or not and learns how to automatically mark calls. What is T?

A. *Watching you mark calls*

B. *Classifying phone-calls as noise complaints or not*

C. *Number of calls correctly classified as noise complaints or not*

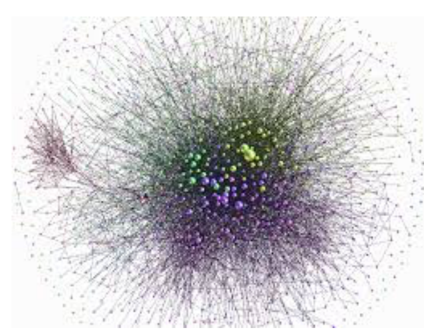
D. *This is not a ML problem*

Examples of deployed ML systems

Recommendation Systems

Recognition Systems

Pattern Analysis



and many more.. pretty much whenever we want to **learn from data**

Main subfields of Machine Learning

Machine Learning

- **Supervised Learning**
  - Regression ← **We will start here**
  - Classification
  - Ranking
  - Structured Prediction

  $$\mathcal{D} = \{\mathbf{X}, \mathbf{t}\}, \quad e.g. \quad \mathbf{X} \in \Re^{N \times D}, \mathbf{t} \in \Re^{N}$$

- **Unsupervised Learning**
  - Clustering
  - Dimensionality Reduction
  - Manifold Learning

  $$\mathcal{D} = \{\mathbf{X}\}, \quad \mathbf{X} \in \Re^{N \times D}$$

- **Reinforcement Learning**
  - Markov Decision Process (e.g. Robotics)
  - Multi-Agent Systems
  - TD/Q-Learning (State-Action-Reward)

  $$\mathcal{D} = \{\mathbf{S}, \mathbf{A}, \mathbf{R}, \pi^*, \mathbf{V}\}$$

What are these symbols??

$$\mathcal{D} = \{\mathbf{X}, \mathbf{t}\}$$

Here is a dataset D

Student reg. no.	ML grade	P. Skills grade	final degree
1	92%	84%	78%
2	54%	100%	62%
3	58%	50%	52%
4	85%	96%	72%
5	67%	98%	68%
6	75%	86%	72%
7	52%	100%	61%
8	82%	90%	85%

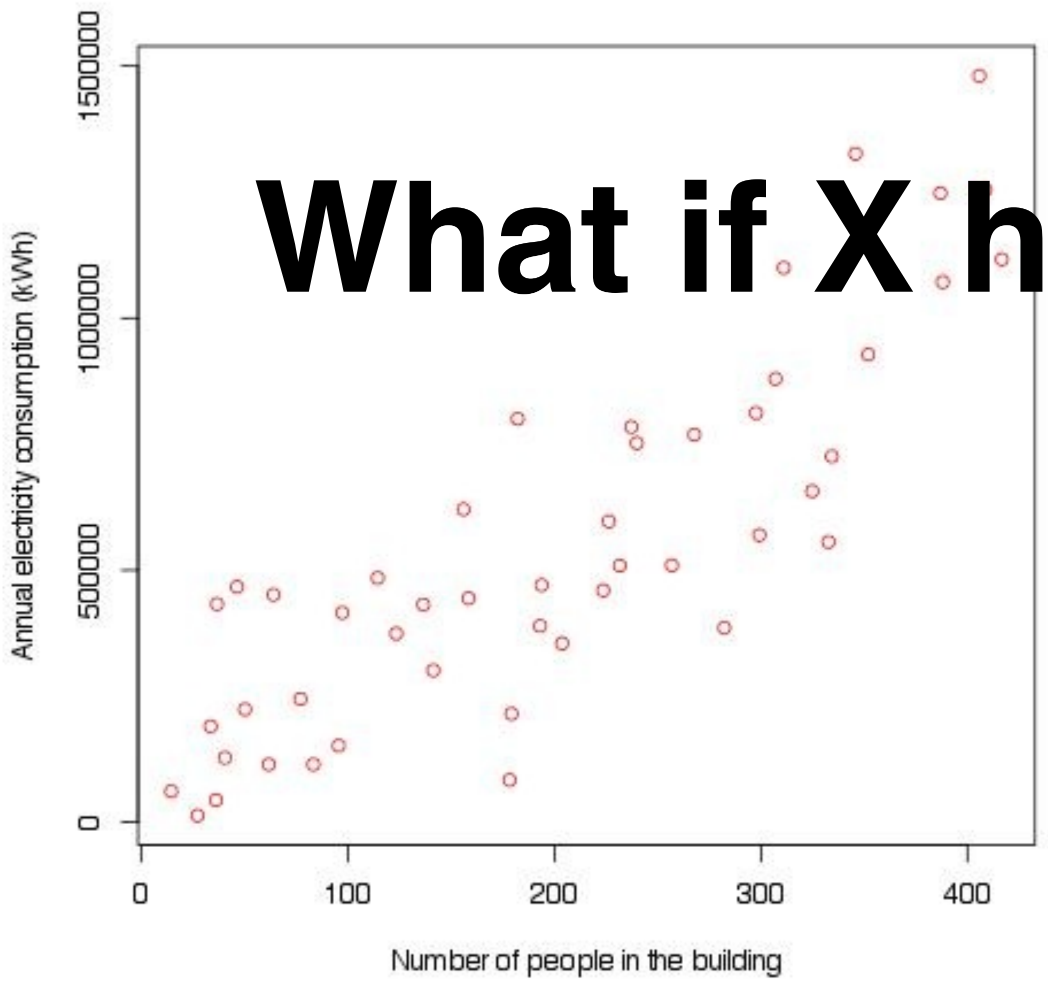
Choose a task T. What is the data/experience E you will use to predict T? Why? What is the input **X** and what is the target **t**? their dimensions?

Supervised learning intro: when t is continuous

From Rogers & Girolami book

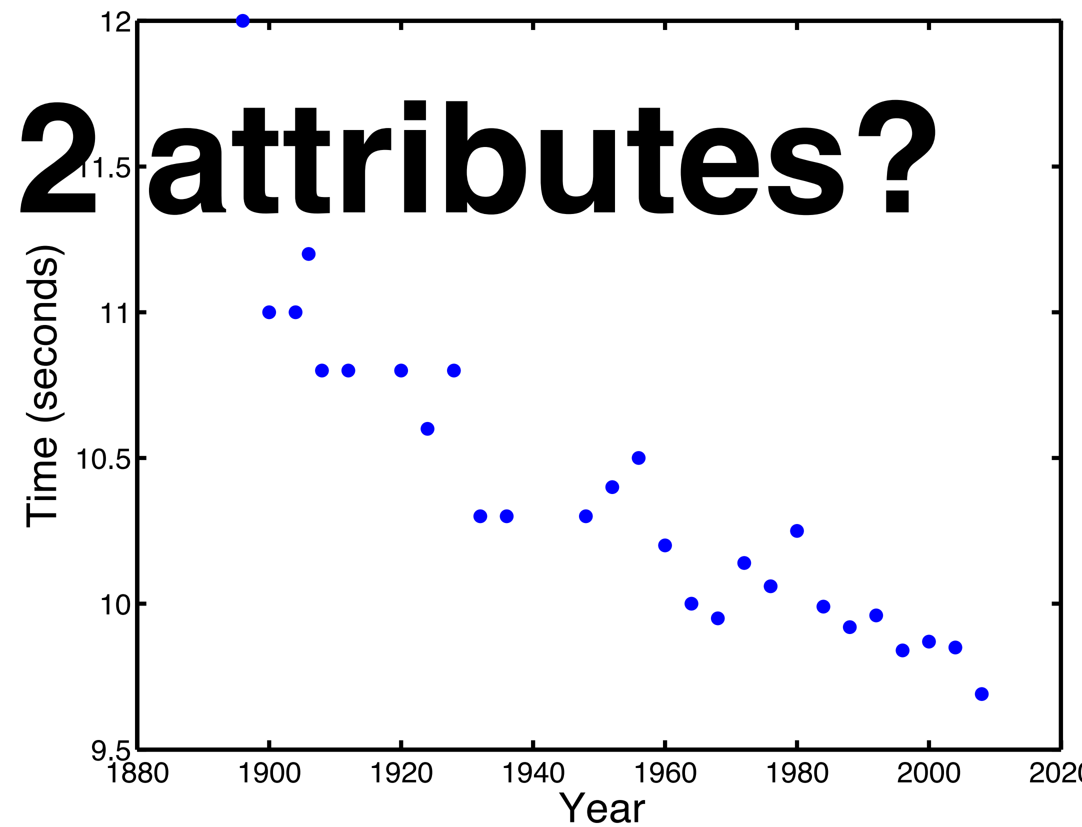
What if X has 2 attributes?

t



X

t



X

Predict electricity consumption given number of people

Predict winning men's 100m time given the year

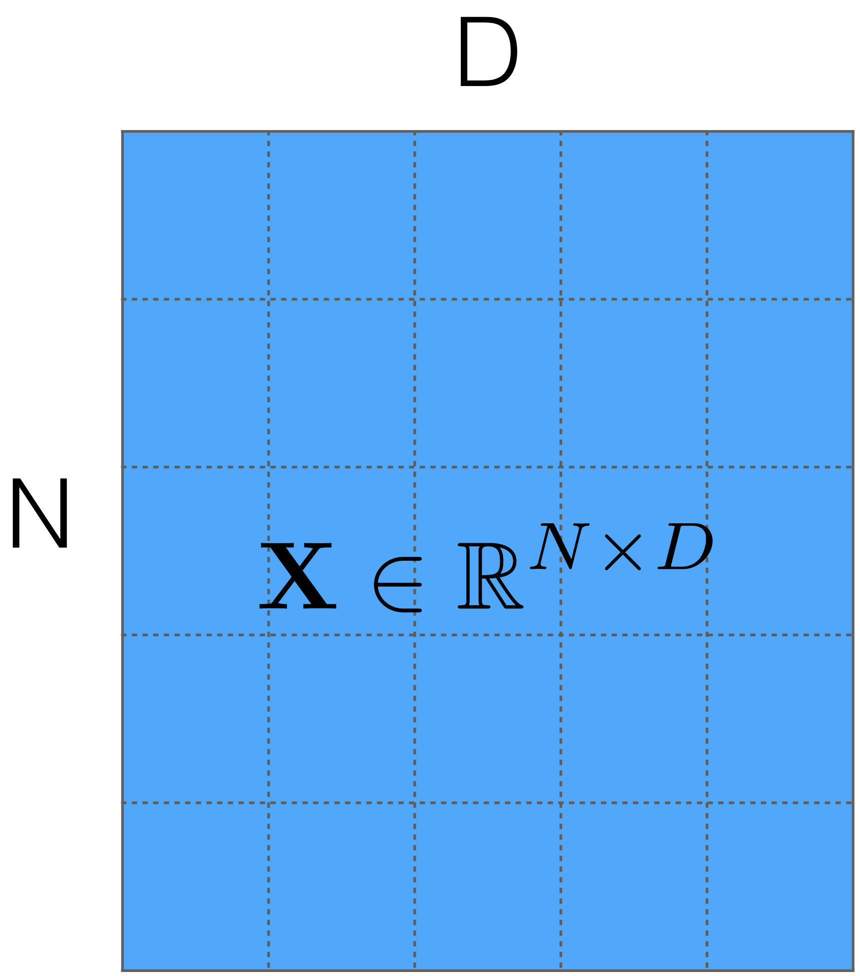
Data and terminology (hell!)

INPUTS

Attributes, Dimensions, Independent variables,

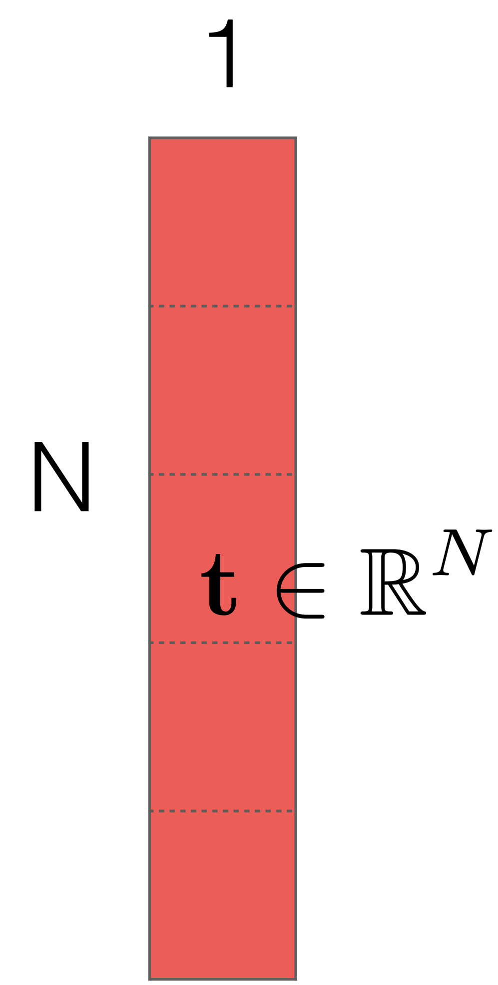
Predictor variables, Covariates, Features

Observations
Samples
Objects
Instances



OUTPUTS

Target, Response, Label
Dependent variable



Convention of notation

x is a scalar = small letter and regular font

$\mathbf{x}$ is a vector = small letter and **bold** font

$\mathbf{X}$ is a matrix = large letter and **bold** font

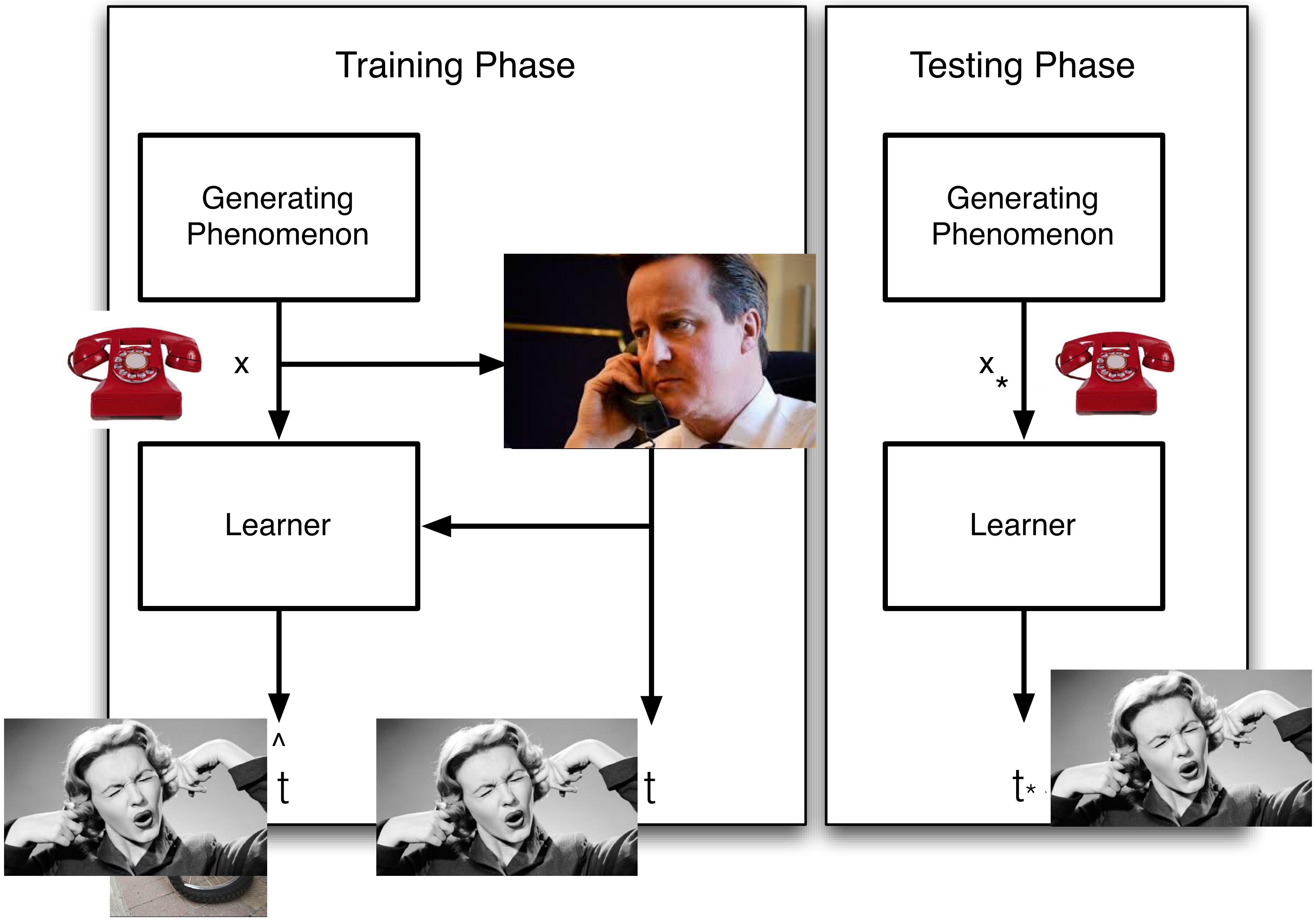
$\mathbf{x}_n$ is a row vector of **X** = small letter, index, and **bold** font

$f(\mathbf{x}_n; \mathbf{w})$ implies function f is acting on $\mathbf{x}_n$ with parameters **w**

bold symbols indicate **multiple dimensions**

Supervised learning

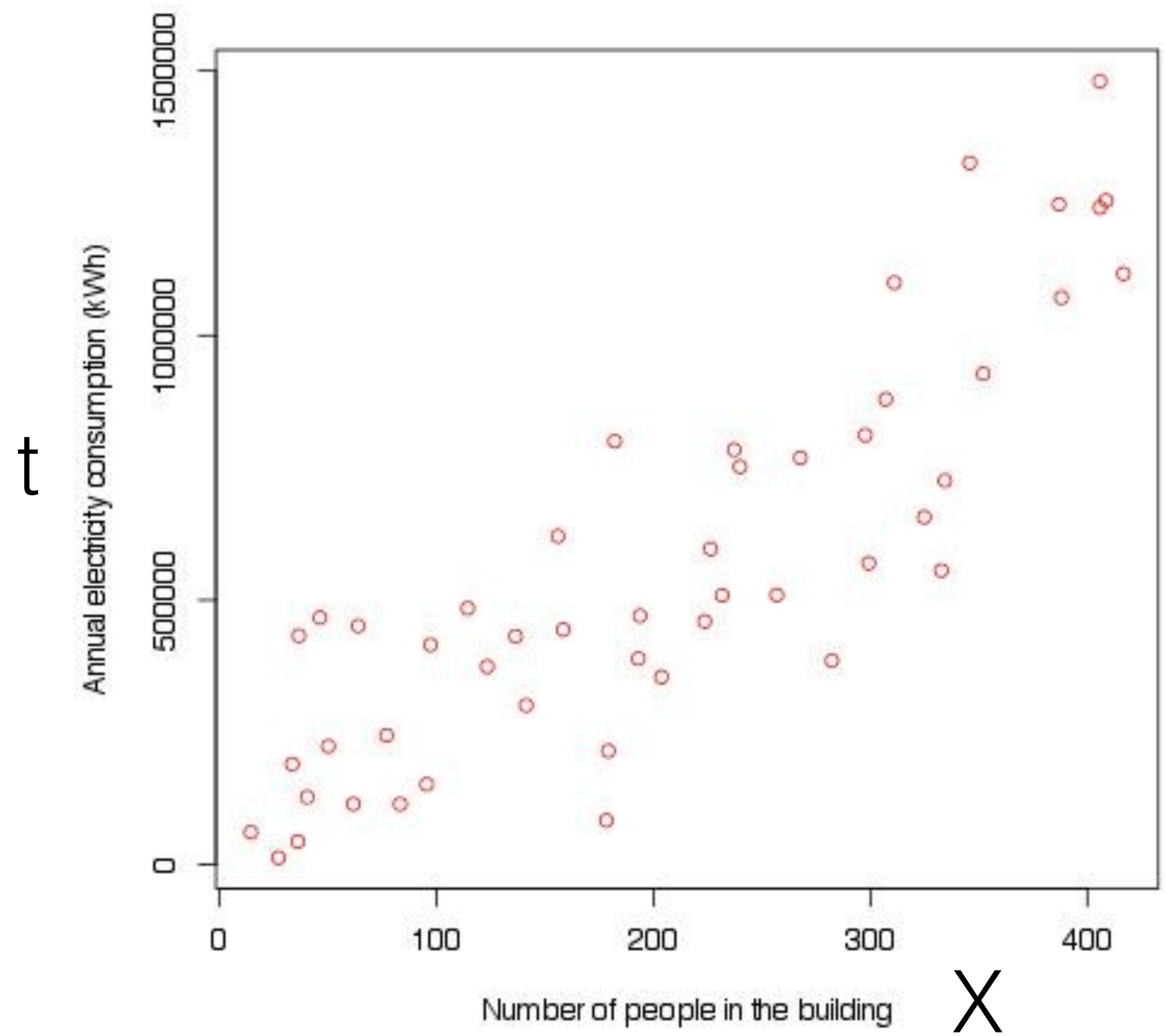
Learning Prediction



$$D_{\text{training}} = \{\mathbf{X}, \mathbf{t}\} = \{\mathbf{x}_n, t_n\}_{n=1}^{N} \qquad D_{\text{testing}} = \{\mathbf{X}_*\}$$

A supervised learning example: Regression

Regression:
t is continuous



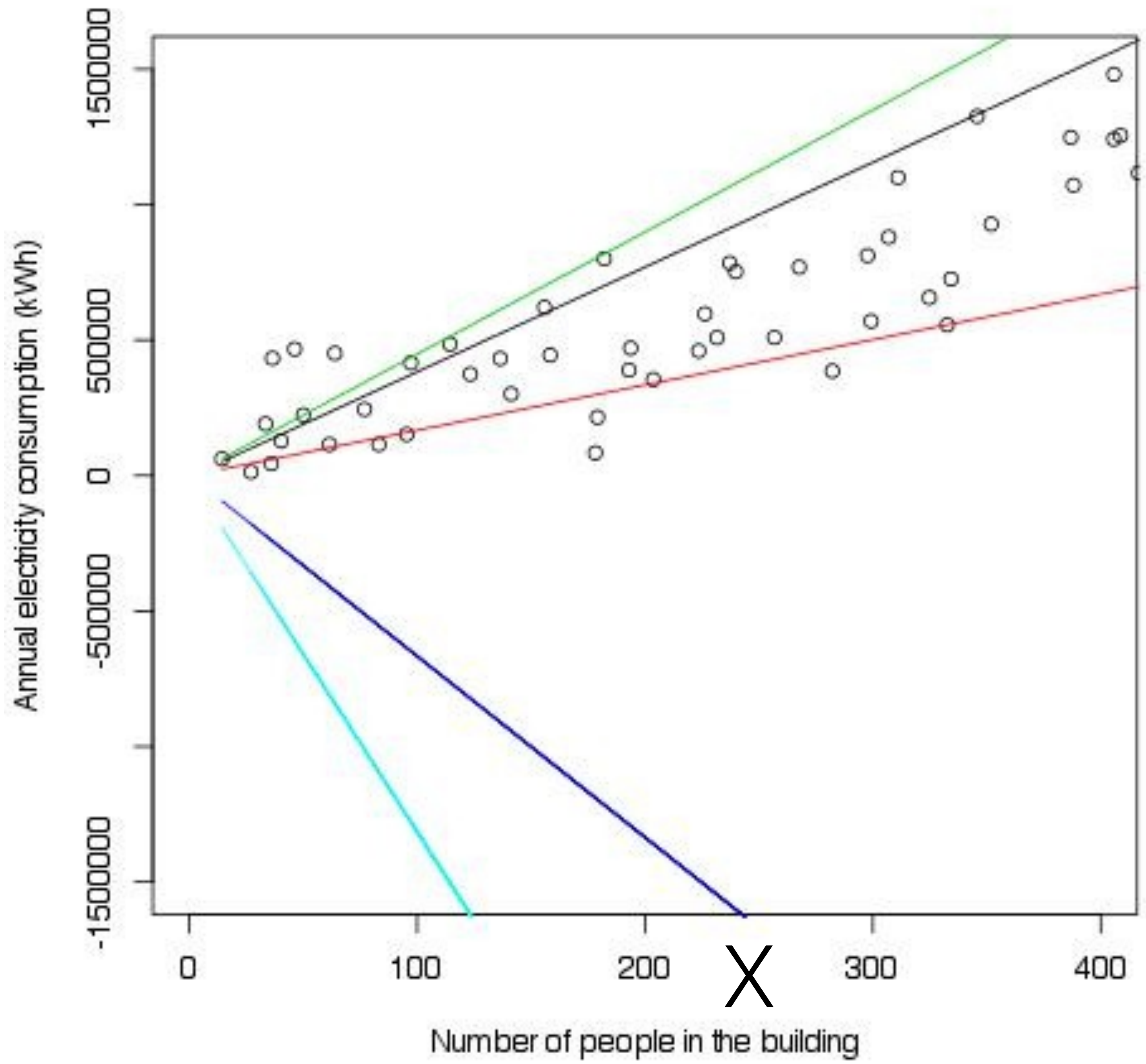
Hypothesis: The relationship between X and t is **linear**

$$\hat{t} = f(x; w_0, w_1) = w_0 + w_1 x$$

Which line is "better"?

Hypothesis Space: All lines in this 2-D space

t

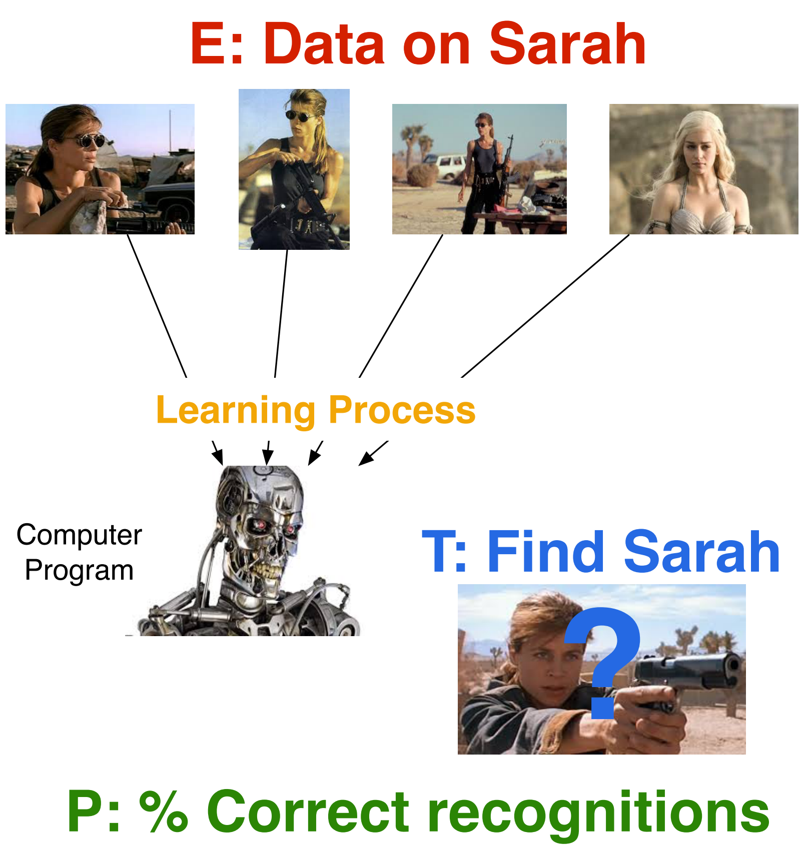


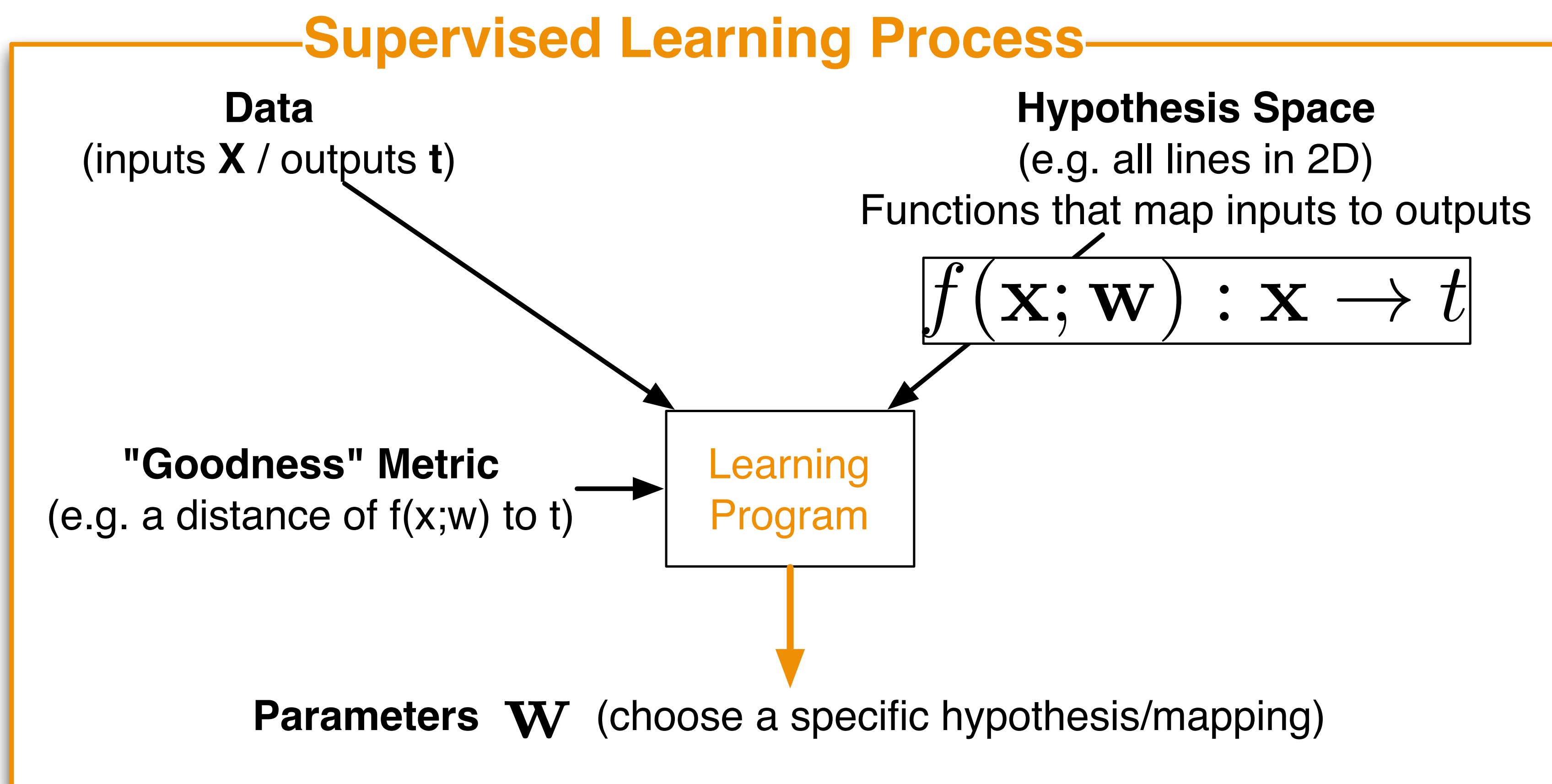
X

What can we "tune"?
What do we learn?

$$\hat{t} = f(x; w_0, w_1) = w_0 + w_1 x$$

Components of a (supervised) learning system



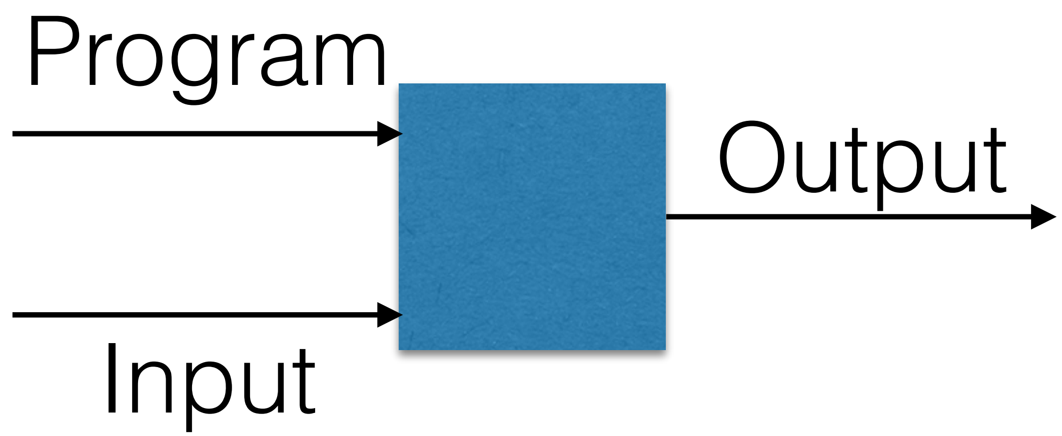


Learning as parameter (hypothesis) inference

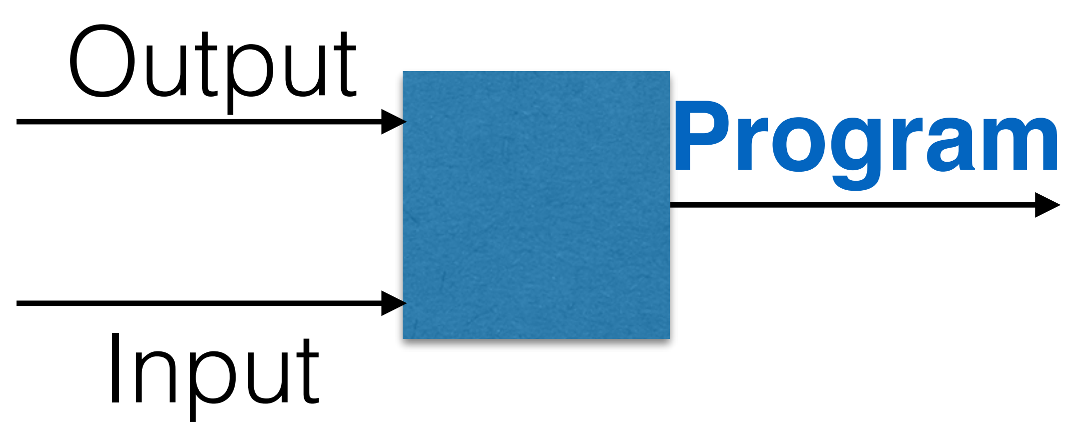
What should X be for our program to have a chance of recognising Sarah?
this is a focus area of the Computer Vision field…

ML from the perspective of CS

Classic CS



ML
(Learning/Training phase)



Some questions you should be already asking

- What is the appropriate task T we should be addressing?
- How do I choose the input data X and how do I encode it?
- What hypothesis space or function f(x;w) should I fit?
- How do I choose the model complexity / hypothesis space?
- What should be my performance metric?
- How much training data is needed?
- What do I do if training data is too small/big? ("big data")
- What prior knowledge can I exploit?

Next lecture we dive in some of these questions within a linear regression setting (R&G Ch1)