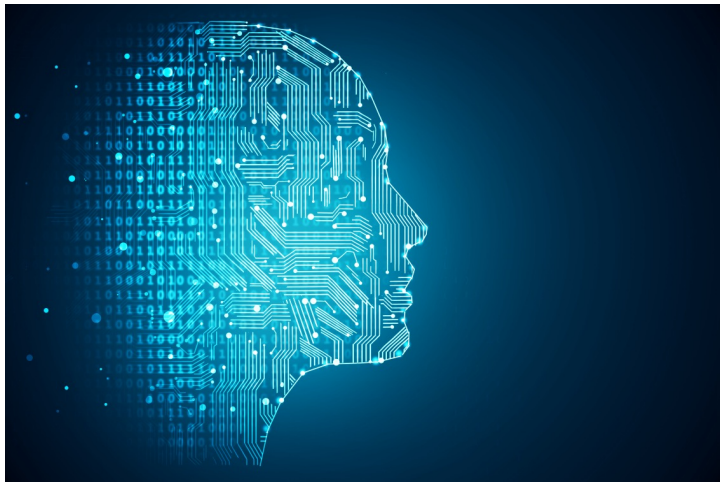
PRINCETON UNIVERSITY

CENTER FOR STATISTICS AND MACHINE LEARNING

Introduction to Machine Learning

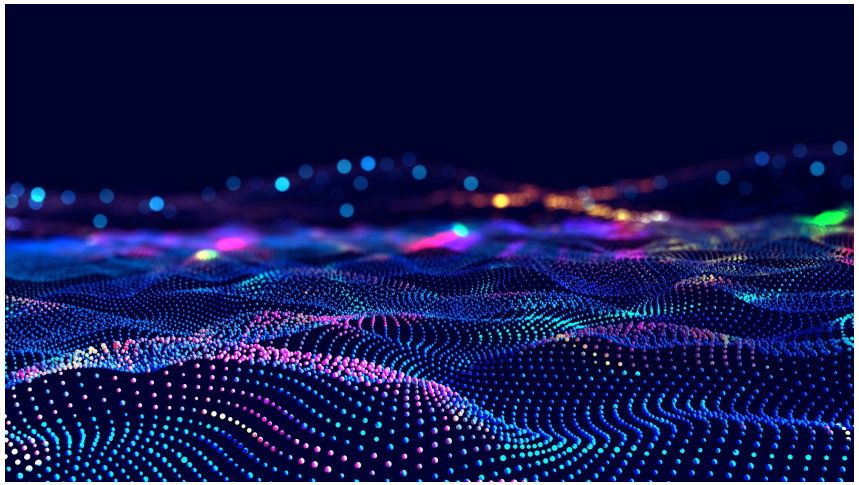


Wintersession 2024
Jan 16-18, 22-23

Gage DeZoort

Julian Gold

Jake Snell



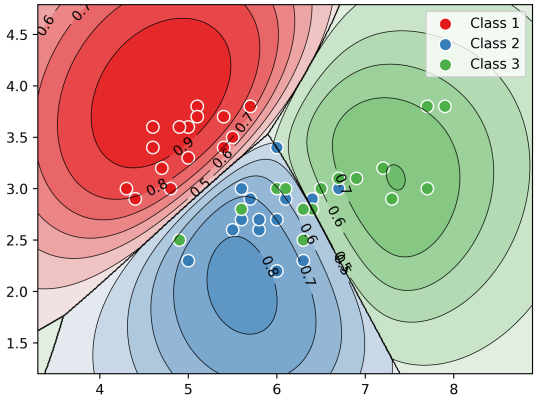
With materials from:
Brian Arnold, Jonathan Halverson, Christina Peters, Savannah Thias, and Amy Winecoff

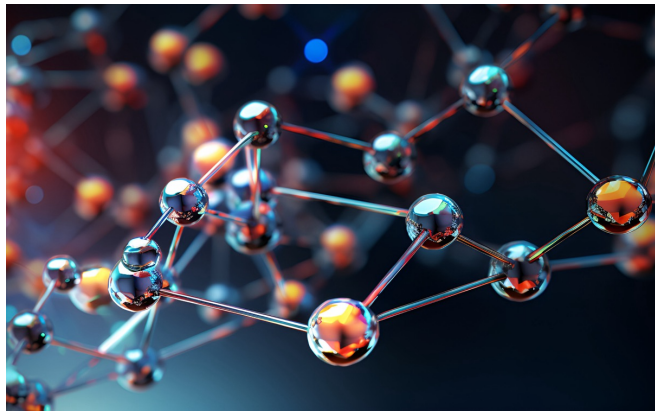
DataX
Accelerating Scientific Discovery at Princeton

Welcome!

About Me

- Background: deep learning with limited data.
- Working with Tom Griffiths in the Computational Cognitive Science lab.
- Develop algorithms for adaptable and reliable AI.
- Interested in inductive bias: what principles are important for generalizing in the “right way?”





About the Course

- Overview of conceptual foundations and model building.
- Light on math → get productive quickly with Python.
- Only cover several of the **many** ML models in existence.
- We hope this will be a useful starting point for your ML journey.

DataX
Accelerating Scientific Discovery at Princeton

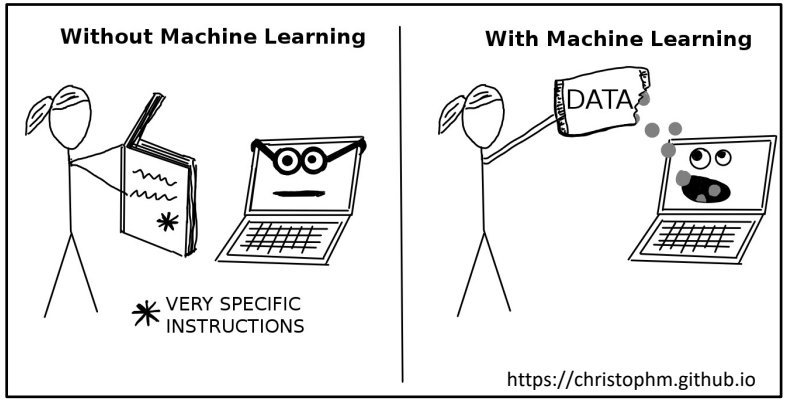
Course Outline

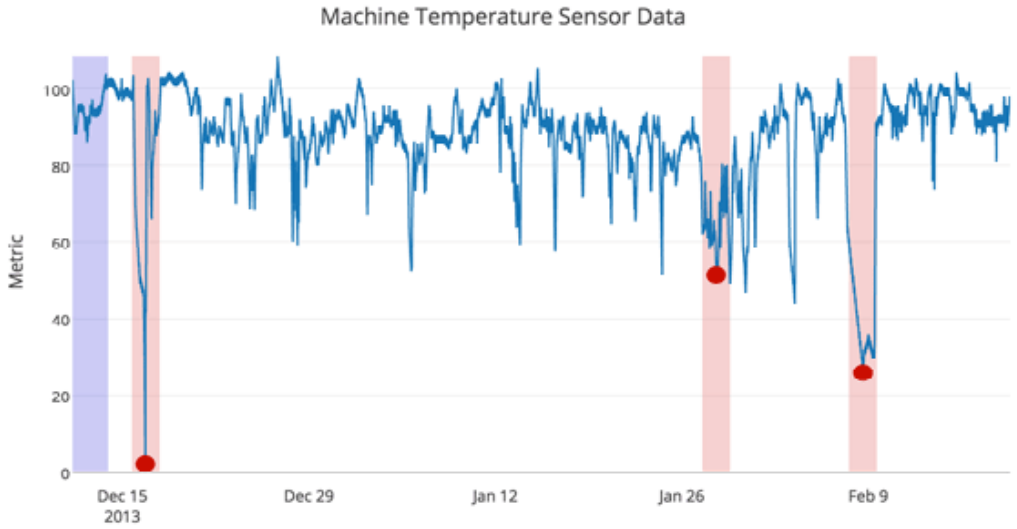
Date	Instructor	Topic
Day 1 Tue. 1/16	Jake Snell	Machine Learning Overview and Simple Models
Day 2 Wed. 1/17	Jake Snell	Model Evaluation and Improving Performance
Day 3 Thu. 1/18	Julian Gold	Introduction to Neural Networks
Day 4 Mon. 1/22	Gage DeZoort	Survey of Neural Network Architectures
Day 5 Tue. 1/23	Gage DeZoort Julian Gold Jake Snell	Hackathon! • Computer Vision • Diffusion Models • Large Language Models

DataX
Accelerating Scientific Discovery at Princeton

What is machine learning?

1. building and understanding methods that '**learn**' by using **data** to improve performance on some set of tasks
2. using and developing computer systems that can **learn** and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in **data**.





Also known as:

- pattern recognition
- artificial intelligence
- data mining
- predictive analytics

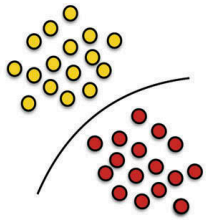
Goal is often to use data to create an algorithm/model that

- makes accurate predictions
- is interpretable, revealing (previously unknown) patterns in data

ML tasks

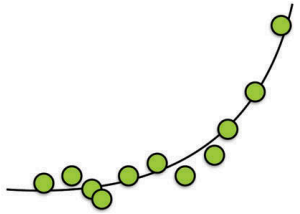
1. building and understanding methods that '**learn**' by using **data** to improve performance on some set of tasks

classification



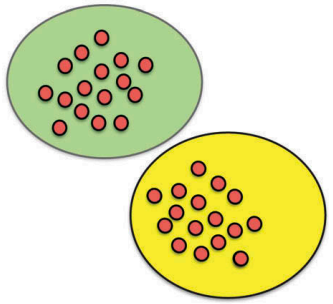
predict input data labels

regression

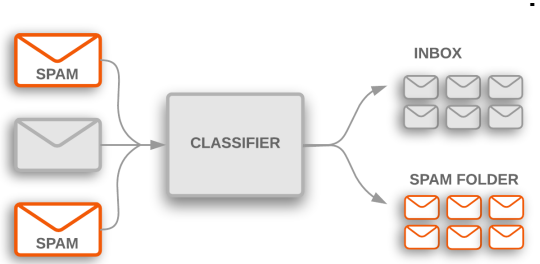


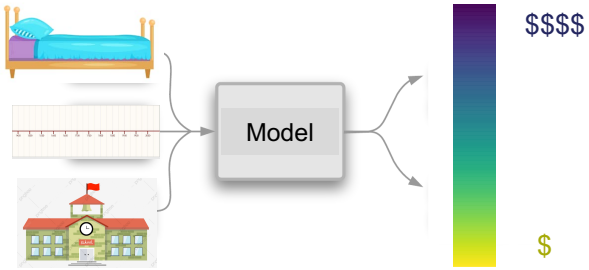
predict input data values

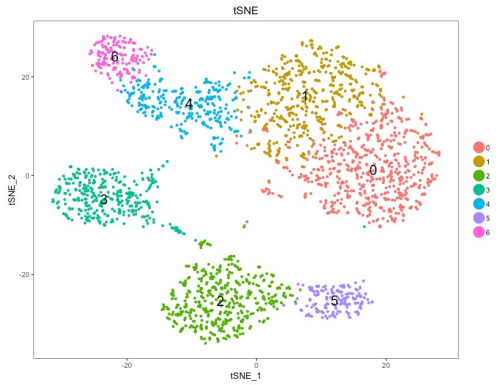
clustering



group input data



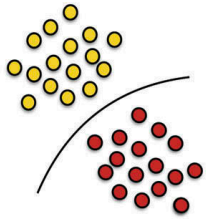




task images from from Carrasquilla 2020

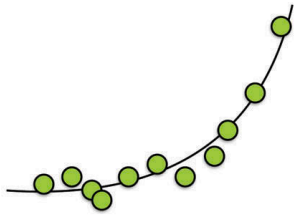
ML tasks

classification



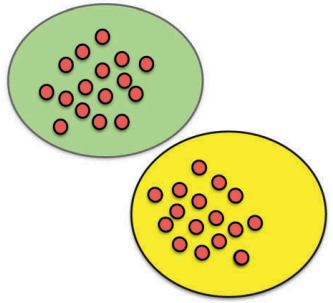
predict input data labels

regression



predict input data values

clustering



group input data

supervised

response (label/value) y_i

features $x_{i,1}, x_{i,2} \dots x_{i,n}$

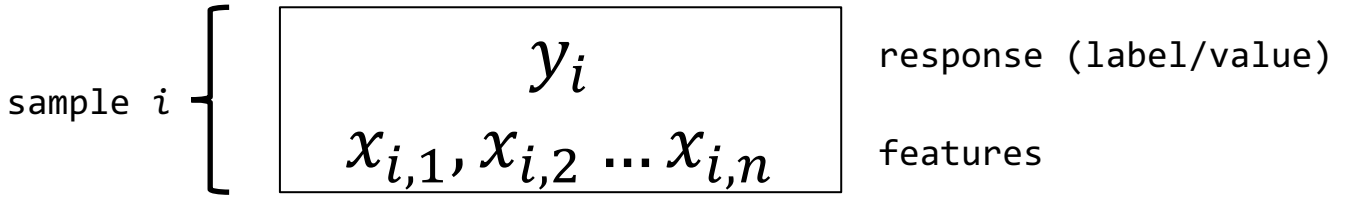
ML learns the relationship between the features and response

unsupervised

sample i $x_{i,1}, x_{i,2} \dots x_{i,n}$

ML learns patterns/groupings

Terminology



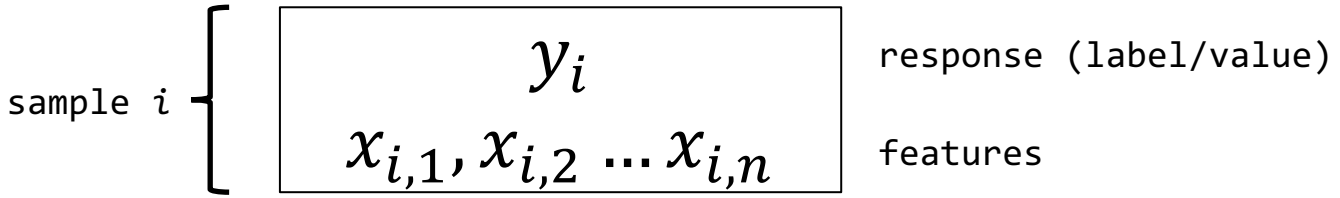
sample i
- sample
- data point
- observation

y_i
- response
- target
- class (if categorical)
- outcome
- dependent variable

$x_{i,1}, x_{i,2} \dots x_{i,n}$
- features
- predictors
- descriptors
- attributes
- covariates
- independent variables

many terms in English, but the **math** is always the same!

Discrete/categorical or continuous values!



examples

response y_i
- a sample's disease status (discrete)
- a sample's height/length (continuous)
- a house's market value (continuous)

features $(x_{i,1}, x_{i,2} \ldots x_{i,n})$
- the presence of a mutation in genome (discrete)
- cigarettes smoked per week (continuous)
- the age of a house (continuous)

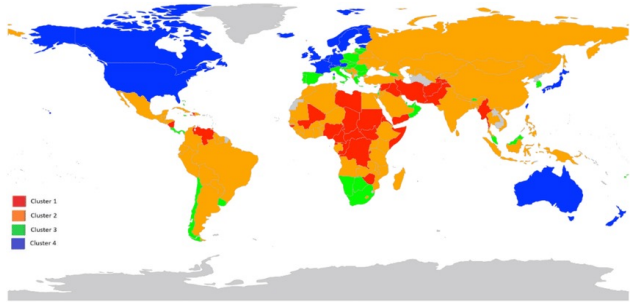
Why use machine learning?

we want to

- know a future event
- make decision based on information
- look for useful patterns in data

examples

- supervised
  - should I sell this stock?
  - how many copies will this book sell?
  - will this customer move their business to a different company?
  - how much will my house sell for in the current market?
  - does a patient have a specific disease?
  - based on past choices, which movies will interest this viewer?
  - which people should we match in our online dating service?
  - will this patient respond to this therapy?
- unsupervised
  - how do customers differ from one another?
  - how are countries different in terms of socio-economic/health?
  - how many cell types are in my sample?



Why use machine learning?

<u>supervised</u>

- prediction
  - predict response value for new samples
- inference
  - understand *how* and *why* a model works

<u>unsupervised</u>

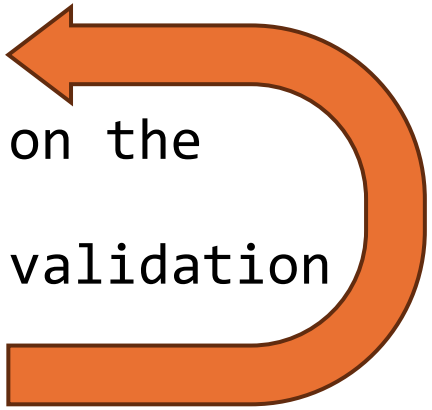
- learn underlying structure of data

Overview of Machine Learning Process

1. Define the problem to be solved.

Datasets? Input features?
Targets? Evaluation metrics?

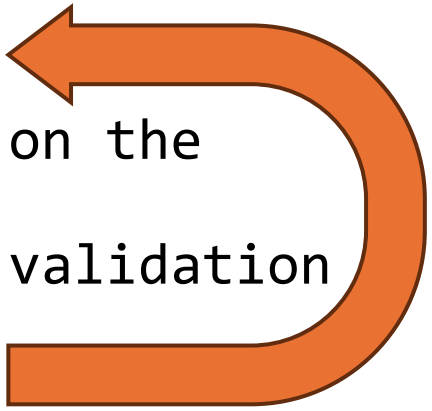
2. Split the data into train / validation / test.
3. Run the validation loop:
   a. Choose a set of models.
   b. Train each model by optimizing its parameters on the training set.
   c. Evaluate the performance of each model on the validation set.
   d. Repeat until performance is satisfactory.

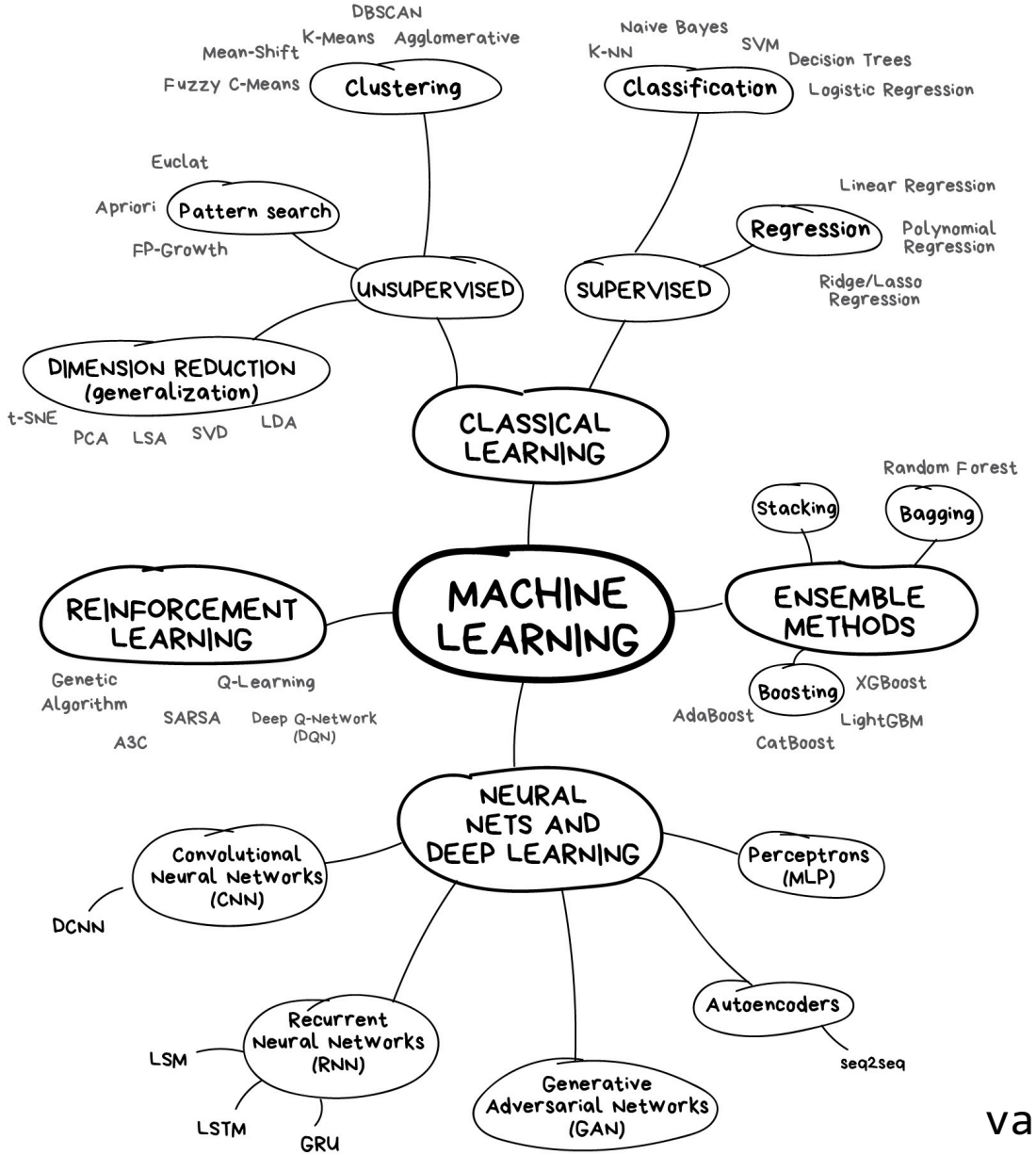


4. Evaluate final performance on the test set.

Overview of Machine Learning Process

1. Define the problem to be solved.
2. Split the data into train / validation / test.
3. Run the validation loop:
   a. Choose a set of models.
   b. Train each model by optimizing its parameters on the training set.
   c. Evaluate the performance of each model on the validation set.
   d. Repeat until performance is satisfactory.
4. Evaluate final performance on the test set.

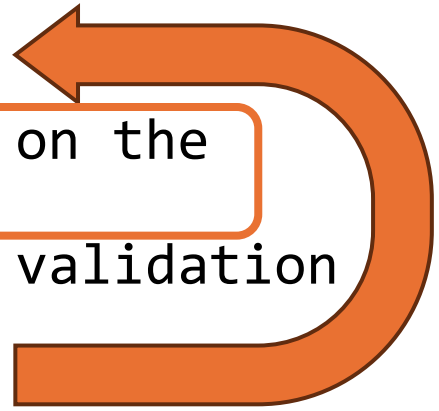




vas3k.com

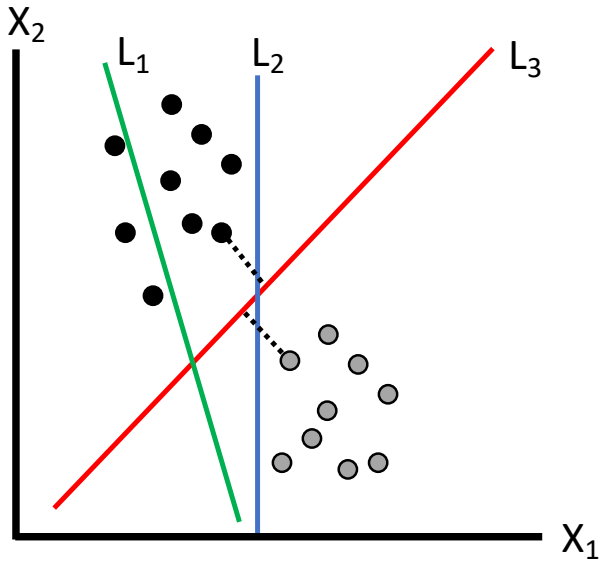
Overview of Machine Learning Process

1. Define the problem to be solved.
2. Split the data into train / validation / test.
3. Run the validation loop:
   a. Choose a set of models.
   b. Train each model by optimizing its parameters on the training set.
   c. Evaluate the performance of each model on the validation set.
   d. Repeat until performance is satisfactory.
4. Evaluate final performance on the test set.

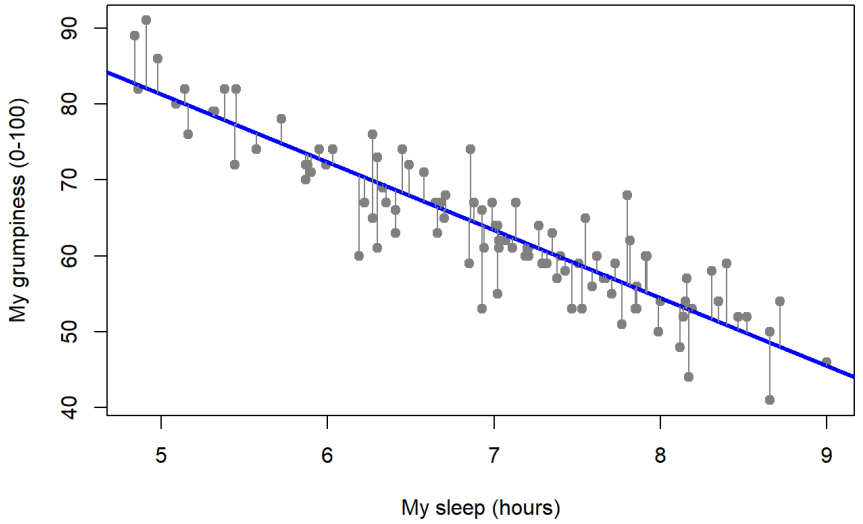


Model Training

e.g. find slope of line that **best** separates training labels

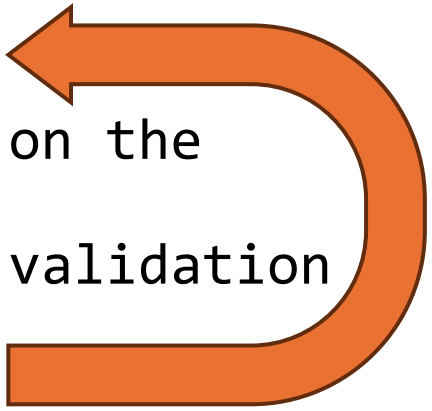


e.g. find slope of line that **best** predicts training values



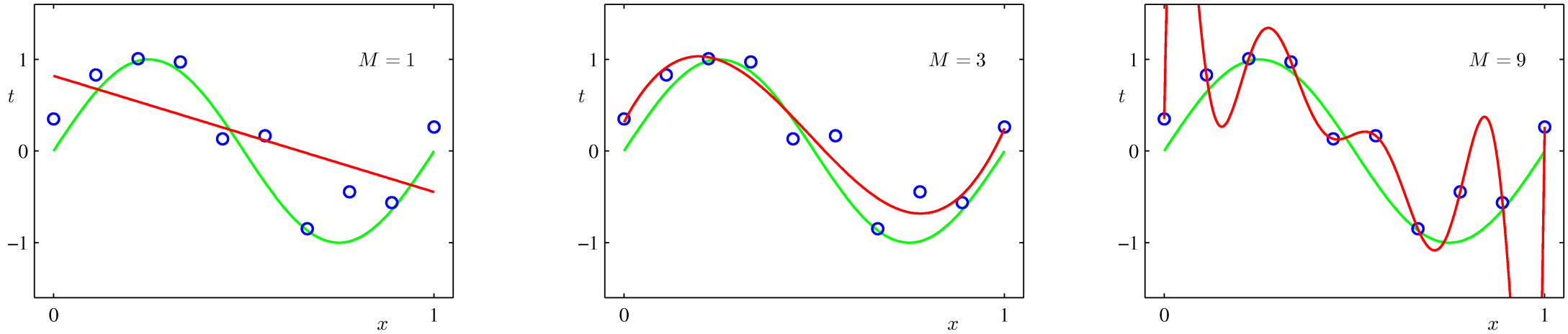
Overview of Machine Learning Process

1. Define the problem to be solved.
2. Split the data into train / validation / test.
3. Run the validation loop:
   a. Choose a set of models.
   b. Train each model by optimizing its parameters on the training set.
   c. Evaluate the performance of each model on the validation set.
   d. Repeat until performance is satisfactory.



4. Evaluate final performance on the test set.

Model Complexity

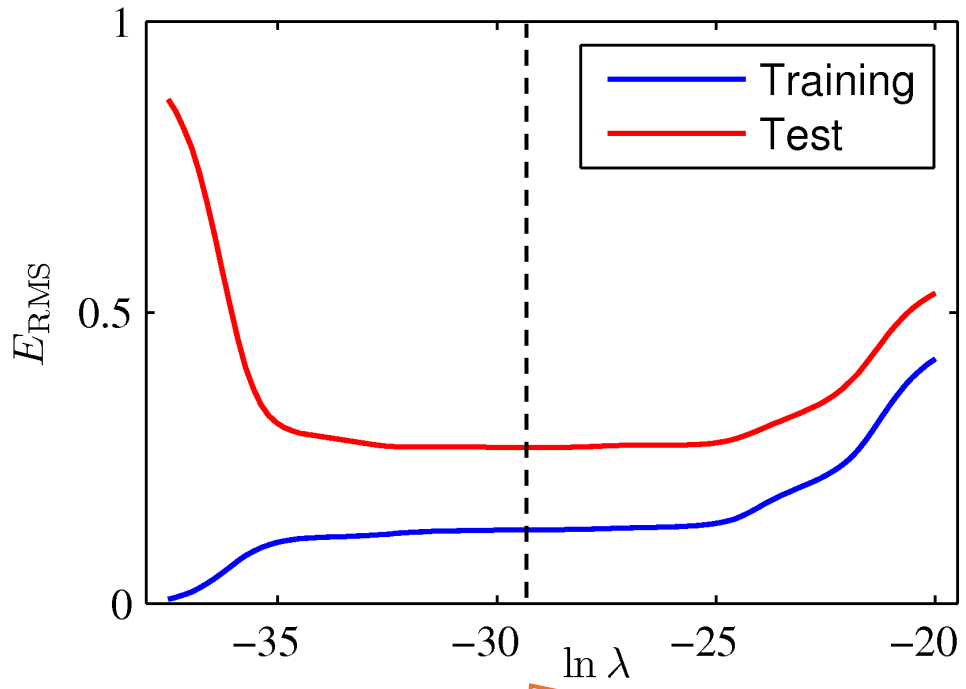


Underfitting
(misses general trend)

Overfitting
(captures noise)

Figure credit: Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*.

Effect of Complexity on Test Performance

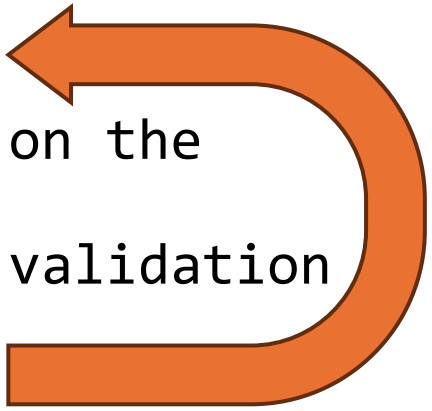


Best model has this value!

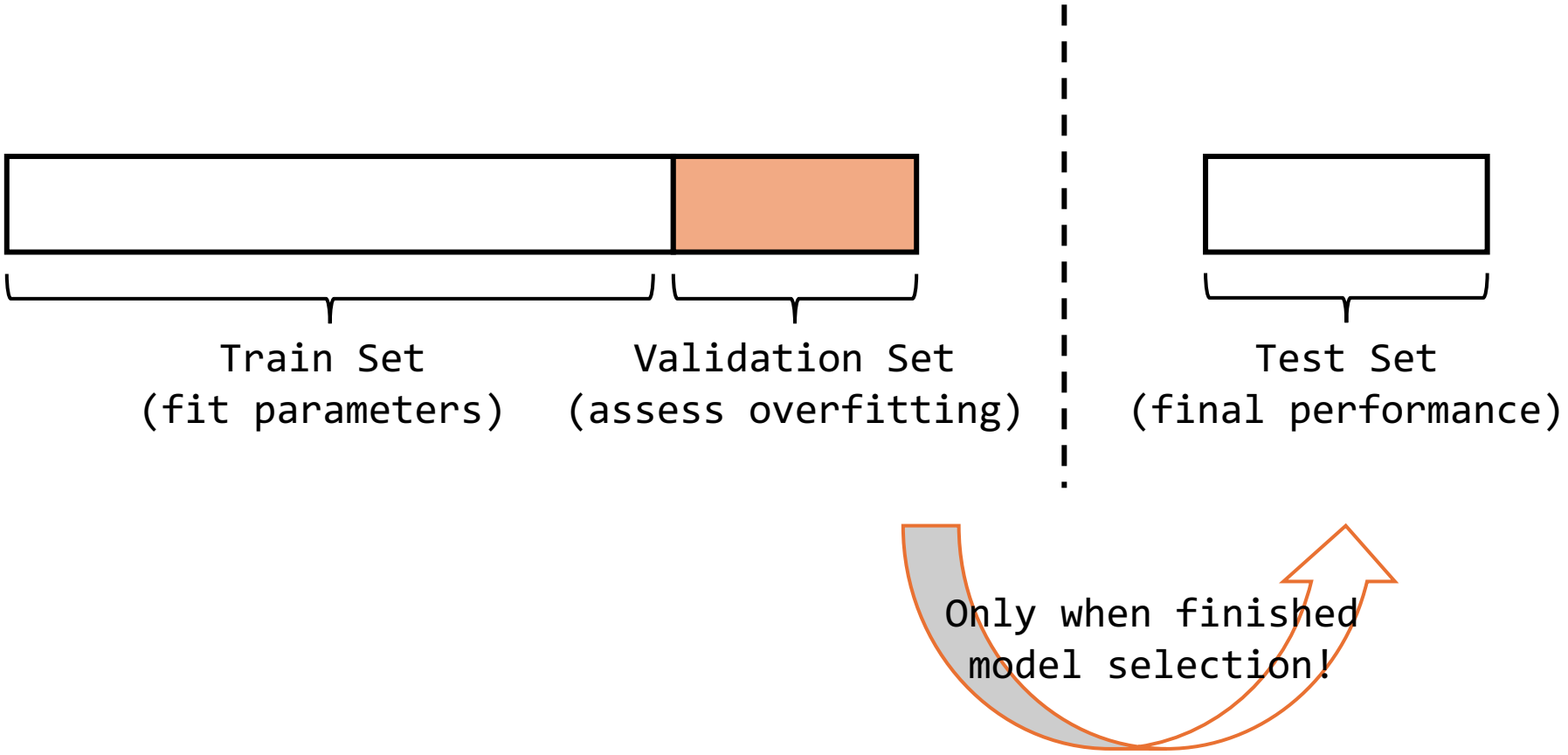
Figure credit: Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*.

Overview of Machine Learning Process

1. Define the problem to be solved.
2. Split the data into train / validation / test.
3. Run the validation loop:
   a. Choose a set of models.
   b. Train each model by optimizing its parameters on the training set.
   c. Evaluate the performance of each model on the validation set.
   d. Repeat until performance is satisfactory.
4. Evaluate final performance on the test set.

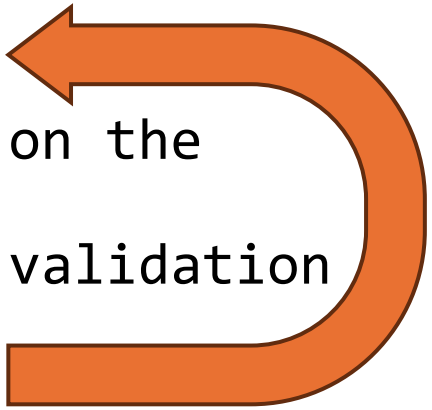


Role of the Validation Set



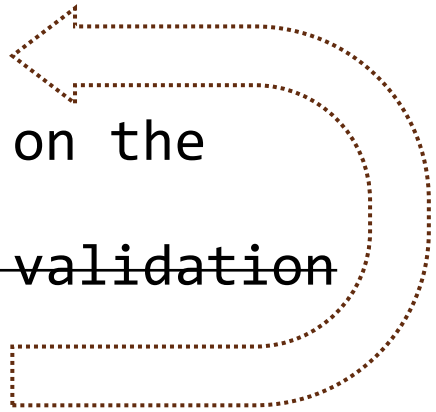
Overview of Machine Learning Process

1. Define the problem to be solved.
2. Split the data into train / validation / test.
3. Run the validation loop:
   a. Choose a set of models.
   b. Train each model by optimizing its parameters on the training set.
   c. Evaluate the performance of each model on the validation set.
   d. Repeat until performance is satisfactory.
4. Evaluate final performance on the test set.



Simplified Machine Learning Process

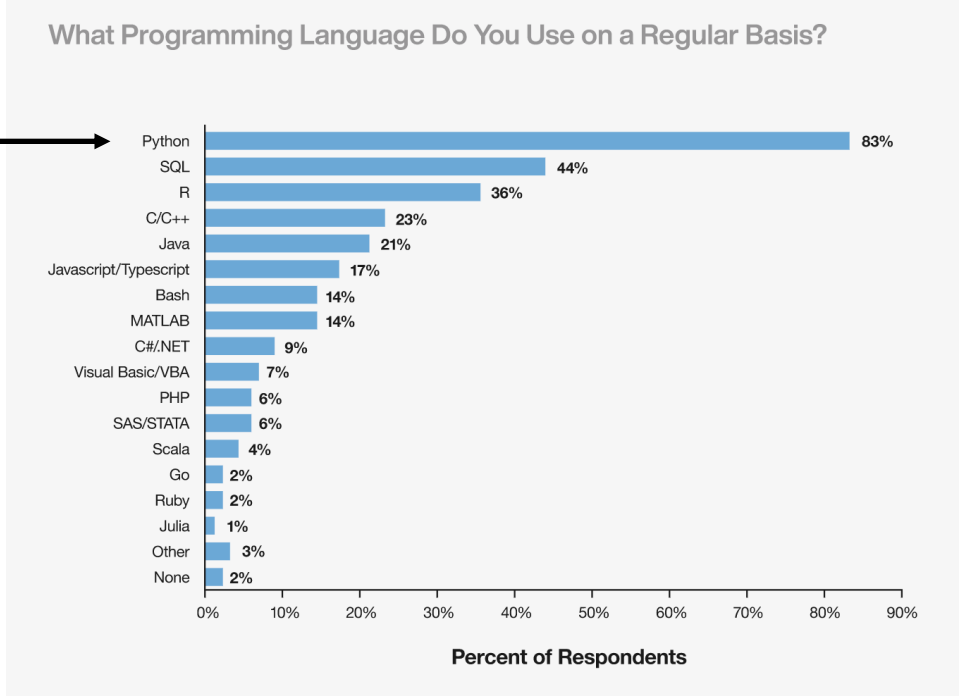
1. Download a dataset from the internet.
2. Use the predefined train / test split.
3. ~~Run the validation loop:~~
   a. Choose a ~~set of~~ model~~s~~.
   b. Train ~~each~~ model by optimizing its parameters on the training set.
   c. ~~Evaluate the performance of each model on the validation set.~~
   d. ~~Repeat until performance is satisfactory.~~
4. Evaluate final performance on the test set.

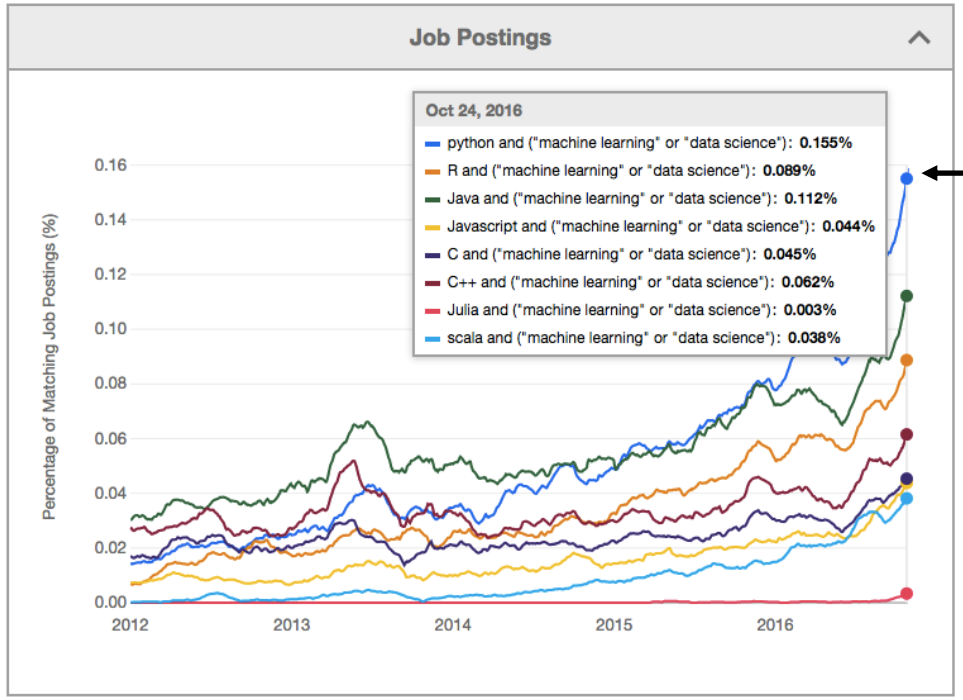


Machine learning in python™

- several options for building ML models
- Python most popular and most in demand (job postings)
- R also popular in statistics and biology communities

python →





← python

CENTER FOR STATISTICS AND MACHINE LEARNING

Fantastic Python libraries

- data analysis
  - Pandas: great for analyzing and manipulating data tables
  - Seaborn: simple functions -> detailed visualization, integrated with Pandas
  - Matplotlib: visualization

- machine learning
  - numpy: fast, powerful data structures for matrices
  - scikit-learn: simple, efficient, accessible tools for ML
  - Keras: neural networks
  - TensorFlow
  - PyTorch
  - …

today we will use **numpy** and **scikit-learn**!

ML Coding Tour in Python!

Open the iPython notebook from this link!

https://github.com/PrincetonUniversity/intro_machine_learning/tree/main/day1

DataX
Accelerating Scientific Discovery at Princeton

Intro to *K*-nearest neighbors (KNN)

- simple but powerful
- can be used for classification *or* regression!
- algorithm
  1. for a given test sample (yellow dot), find the *K* nearest training samples in feature space
  2a. for **classification**, assign label by majority vote
  2b. for **regression**, assign value by mean of neighbors

K is a tunable parameter!
- choose value that gives better predictions on test data