

A Short Tour of Machine Learning

Michael Mathioudakis

Covve, Athens, 2017-08-02

:HELVIA

today

- goal:
 - what is machine learning?
 - emphasis on principles – we'll start slow
 - examples of ML tasks and algorithms
 - how you would formulate a ML task
 - how you would use a ML algorithm
 - not technical details about how to develop one
- part 1: machine-learning algorithms
 - a. basic concepts and the ML pipeline
 - b. algorithms
- part 2: platforms and software
- part 3: hands-on session

part 1: machine learning algorithms

machine learning

Apple launches machine learning
TechCrunch · Jul 19, 2017

RELATED COVERAGE

Improving the Realism of Synthetic
Most Referenced · Apple Machine Learning Research

MORE ABOUT

Apple

Journal of Machine Learning Research

1-16 of 20,570 results for "machine learning"

Show results for

Books

- AI & Machine Learning
- Computers & Technology
- Artificial Intelligence & Semantics
- Data Processing
- Probability & Statistics
- Data Mining
- Computer Science
- Computer Programming
- Machine Theory
- Python Programming

See Less

Kindle Store


- Computers & Technology
- Computer Programming
- Python Computer Programming
- Computer Software
- Probability & Statistics
- Computer Databases
- Mathematics
- General Technology & Reference
- Two-Hour Computers & Technology Short Reads
- Business Software

See Less


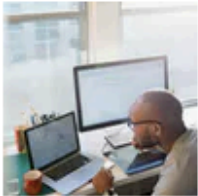

Toys & Games

- Home & Kitchen
- Clothing, Shoes & Jewelry
- Baby
- Electronics
- Office Products
- Arts, Crafts & Sewing
- Apps & Games

Degrees



Courses and Sp



Repositories 64K Code Commits 50K Issues 23K Wikis 22K Users 8K

Advanced search

64,082 repository results

Sort: Best match

wepe/MachineLearning

Basic Machine Learning and Deep Learning

Updated 2 days ago

Python

1.5k

udacity/machine-learning

Content for Udacity's Machine Learning curriculum

Updated 6 hours ago

Jupyter Notebook

1.3k

robbiebarrat/rapping-neural-network

rap-song writing recurrent neural network

neural-network songs rap-songs lyrics

Updated on Mar 25

Python

471

josephmisiti/awesome-machine-learning

A curated list of awesome Machine Learning frameworks, libraries and software.

Updated 4 days ago

Python

24.4k

hangtwenty/dive-into-machine-learning

Dive into Machine Learning with Python Jupyter notebook and scikit-learn

data-science jupyter-notebook

Updated 5 days ago

6.9k

quinnliu/machineLearning

supervised and unsupervised algorithms from

Matlab

201

Languages

Python	13,908
HTML	10,561
Jupyter Notebook	7,498
Matlab	7,341
R	2,462
Java	2,310
C++	1,210
JavaScript	954
TeX	523
Scala	393

what is machine learning?

'learning'

what do we *learn*?

a **description** of the **data**

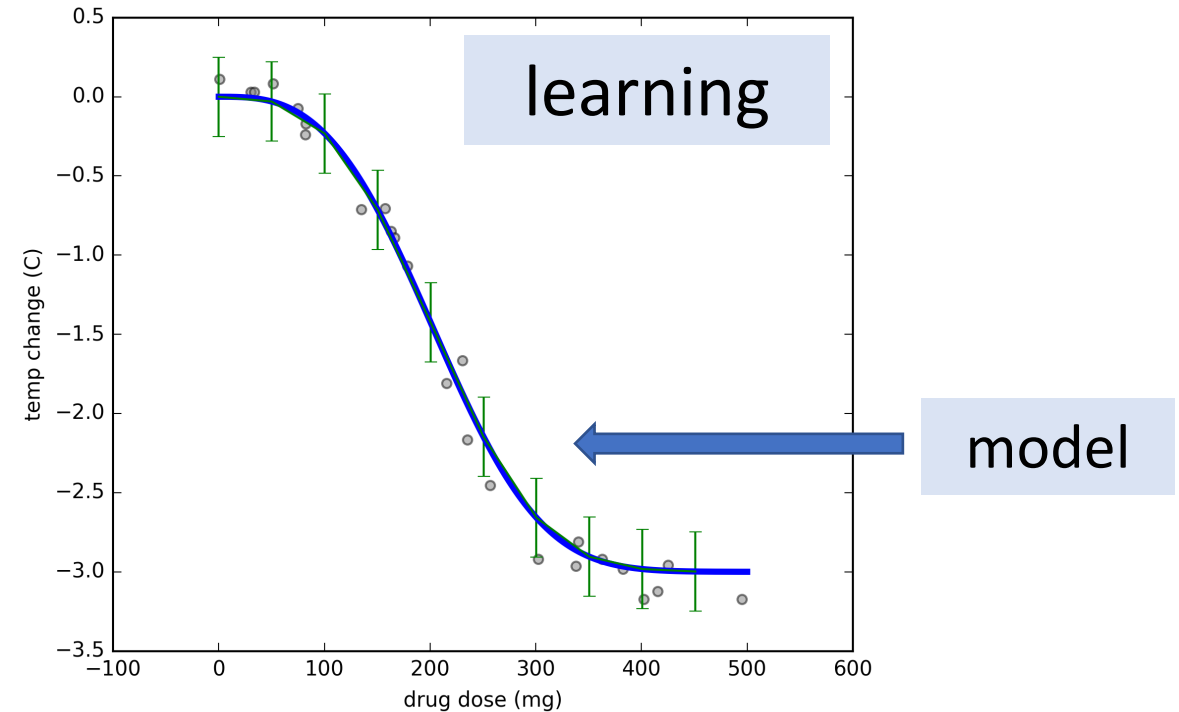
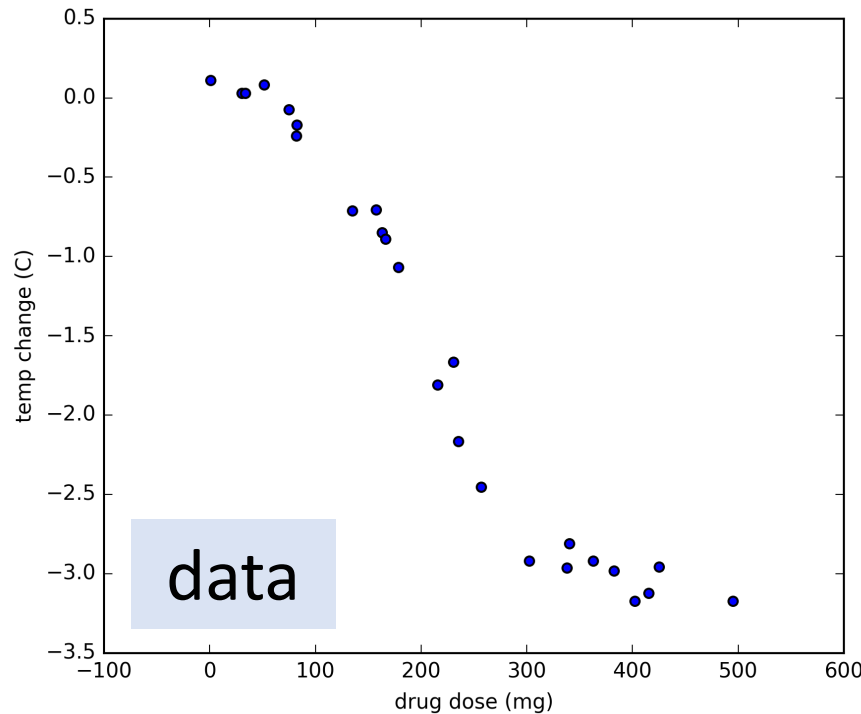
a '**model**' that tells us how the data are distributed

why?

to make *predictions (or inferences / guesses)* and *decisions*
(not only about the future)

example

the patient's temperature has just exceeded 40C
we supply the medicine and observe their temperature change after 2 hours



ok, we 'learned' - then what?

predict what happens to temperature if we supply 200mg?

decide minimum dose to be certain to achieve at least 2C temp drop?

can do with the model
without the data

example

0	1	2	3	4	5	6	7
8	9	0	1	2	3	4	5
6	7	8	9	0	1	2	3
4	5	6	7	8	9	0	9
5	5	6	5	0	9	8	9
8	4	1	7	7	3	5	1
0	0	2	2	7	8	2	0
1	2	6	3	3	7	3	3

prediction task

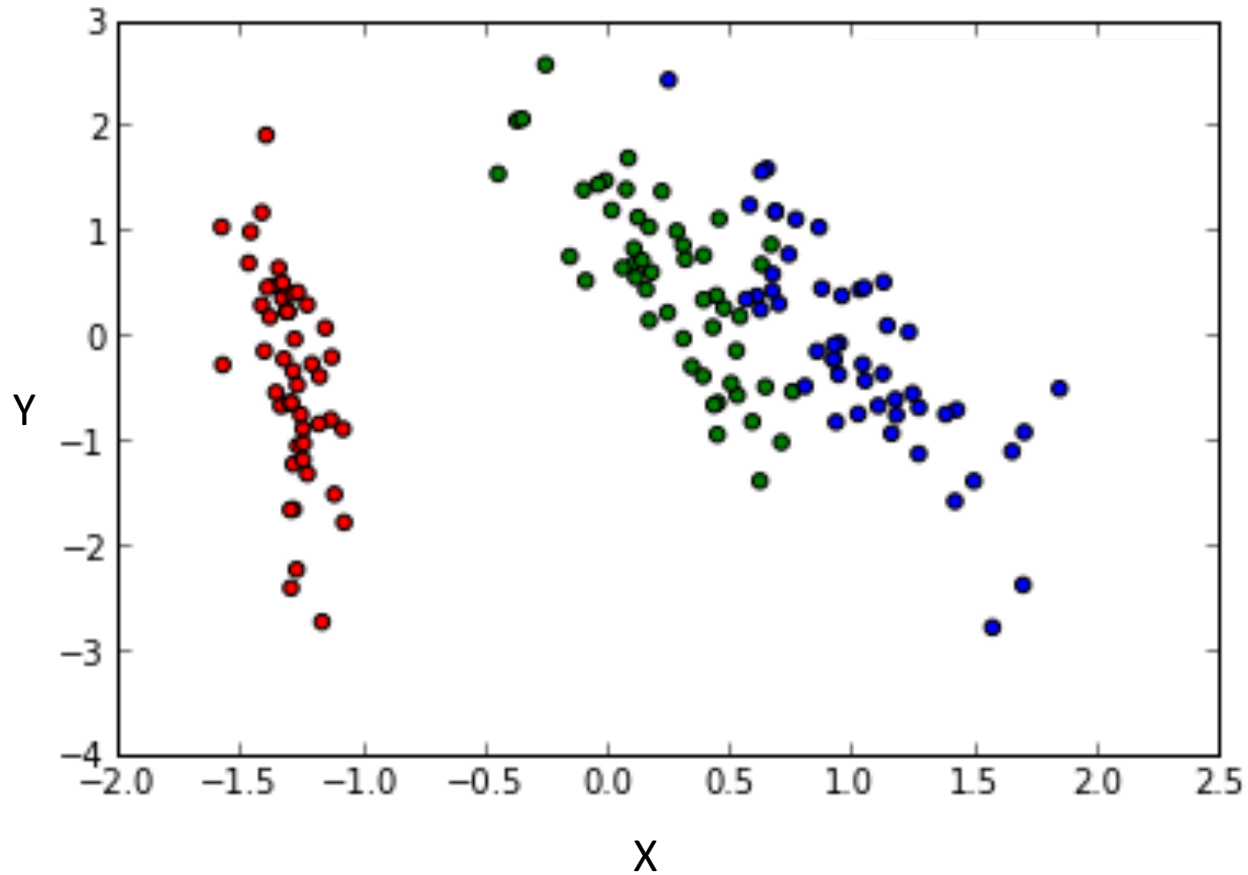
digit recognition

given a handwritten digit, predict (guess)
what number it represents

classification

let's say we are given manually labeled data
how would we approach this?
how would we use a model?

example



[prediction task](#)

what are the data?

what will the next point be?

clustering

density estimation

how would we use a model?

'machine' learning

why do we need the machines?

to make learning

automated

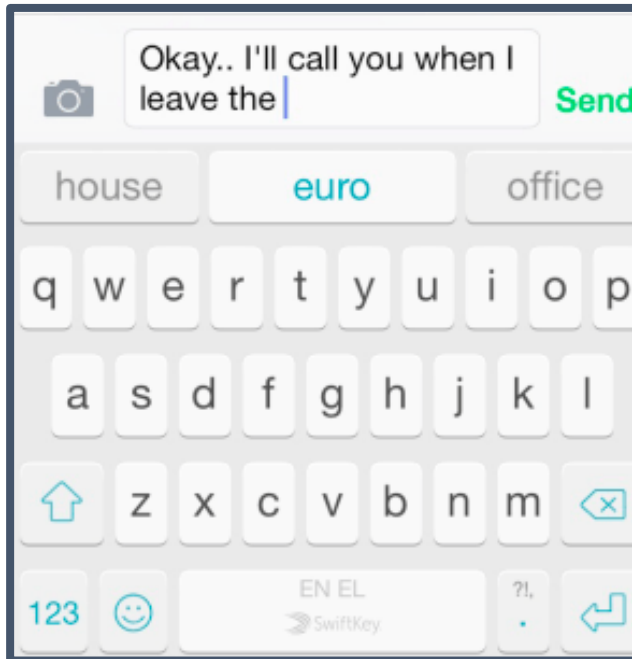
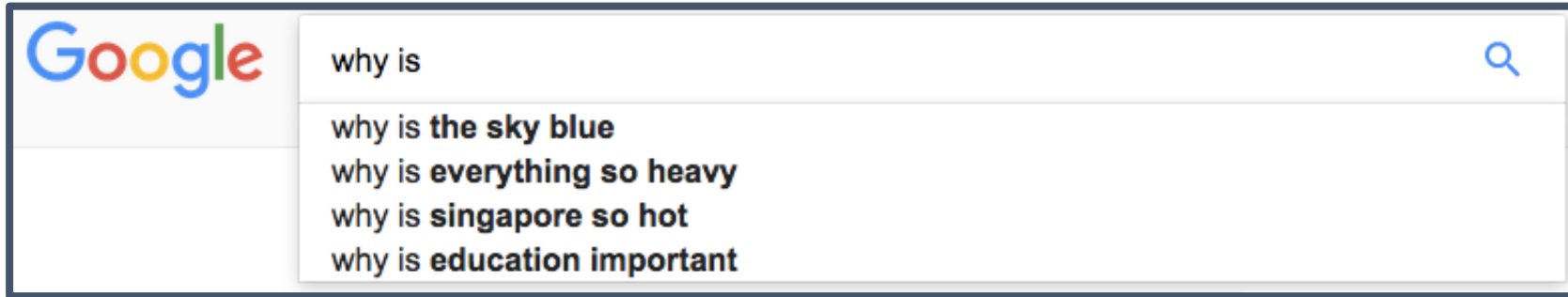
and

efficient

big data

complex models

example: language



task: complete the sentence

language is complex

basic rules (syntax and grammar)

do not suffice for good predictions

requires complex models

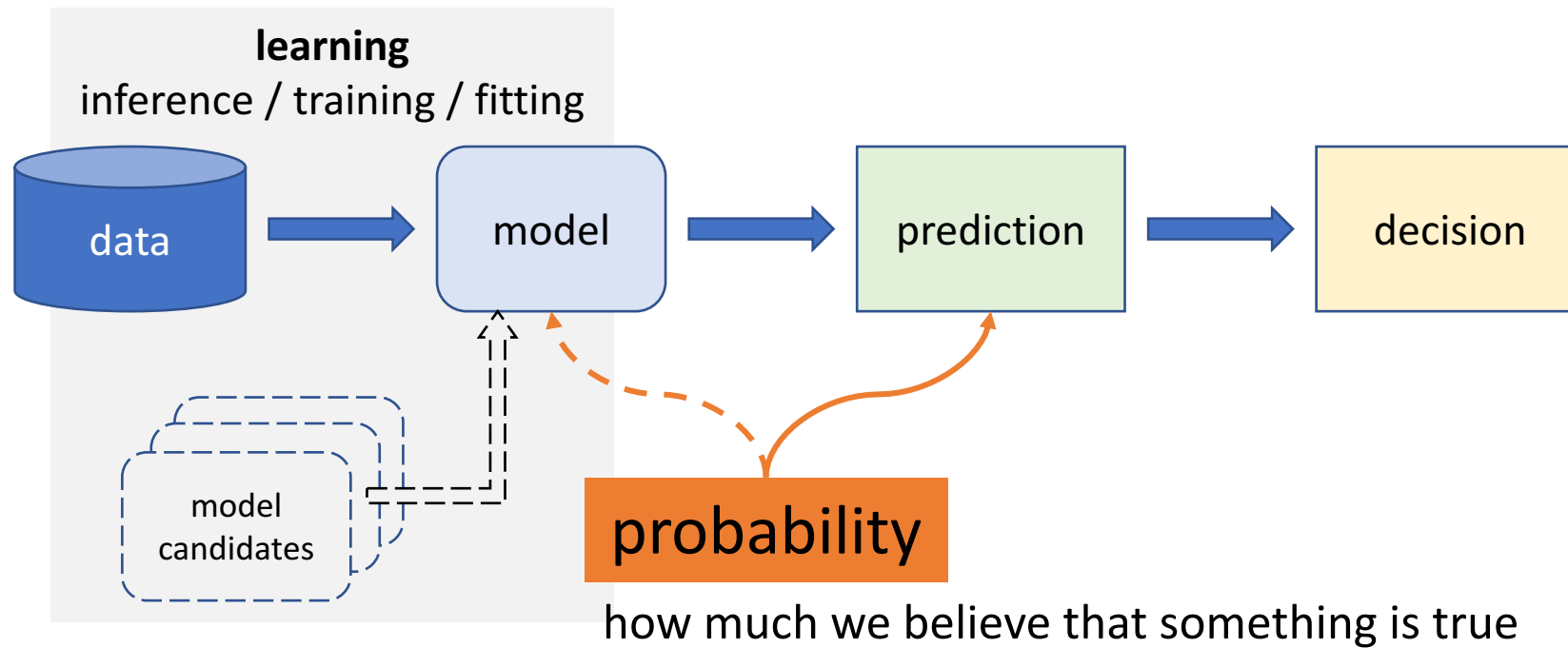
data

millions/billions of sentences/queries

user features

session attributes

ML pipeline



outline

- what is machine learning
 - examples of prediction tasks
 - data, learning, prediction, decision; probability
- probability
- algorithms
 - regression
 - classification
 - clustering
- deep learning

probability

'proposition'

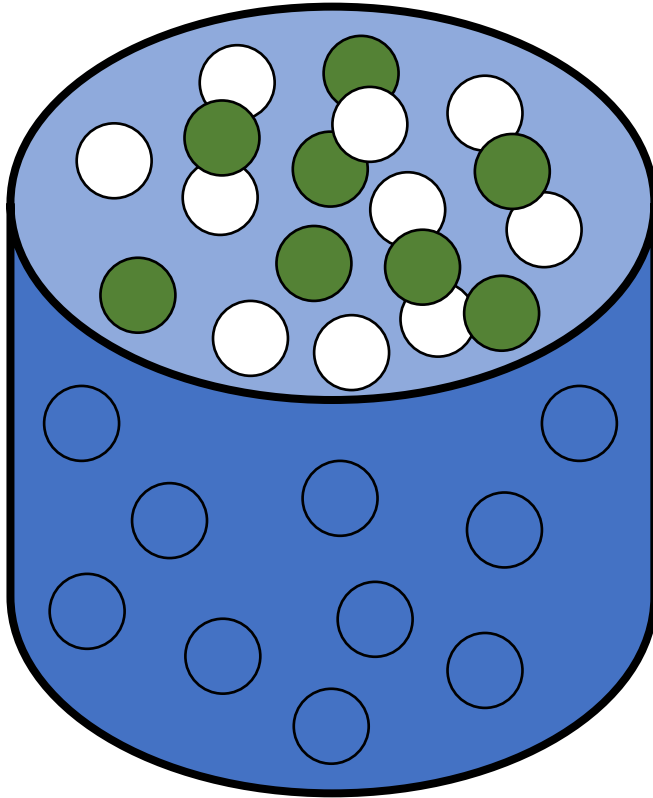
how much we believe that something is true
GIVEN the information at hand



VERY IMPORTANT!

0: no chance 1: certain

probability



a ball drops out of the box

it is green

what is the probability that the
proposition is true

GIVEN that
there are 100 balls,
40 of them green?

related term: 'frequency'

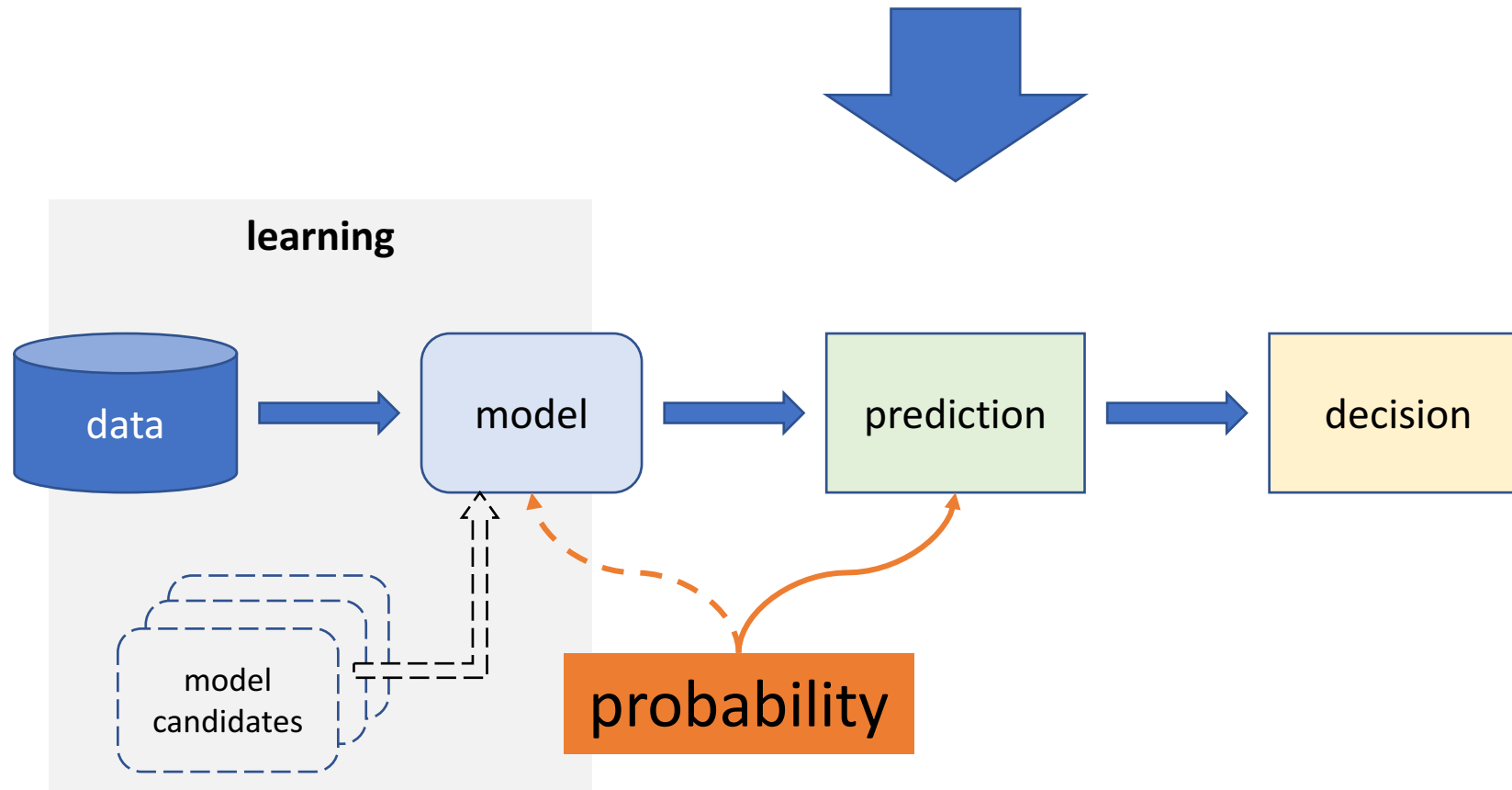
probability



this shape is '1'

'frequency'?

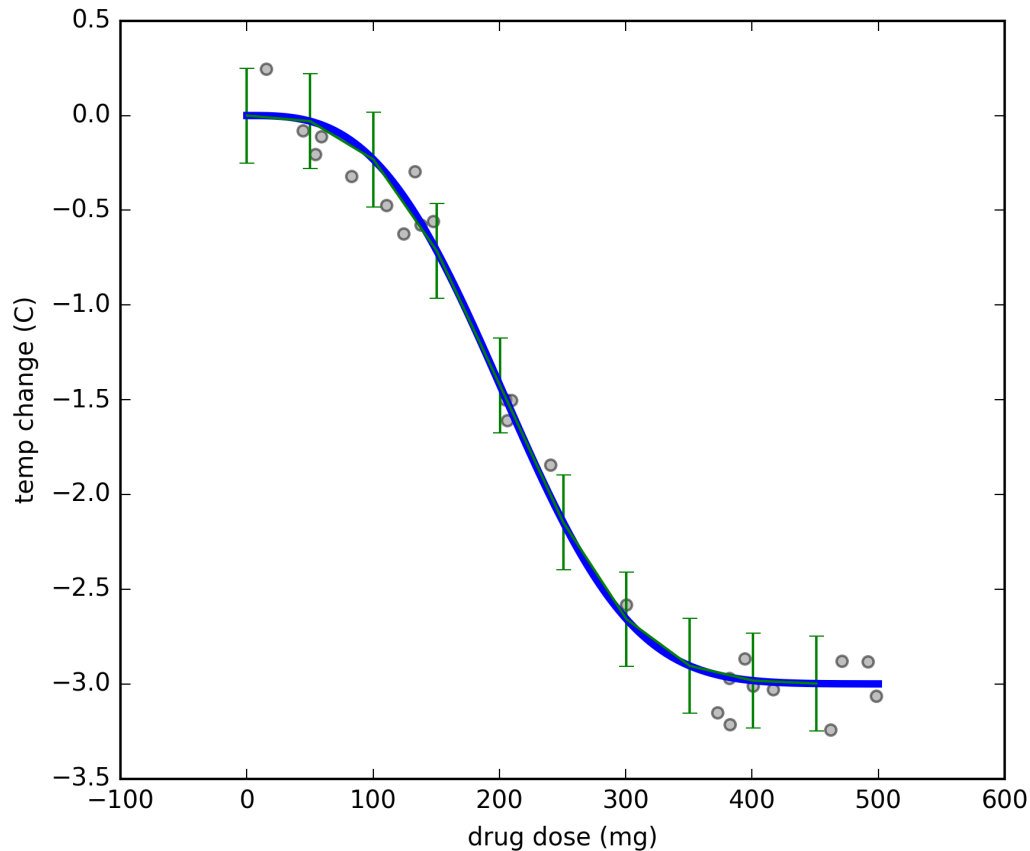
ML pipeline



assign probabilities to propositions

probability :: prediction

what is the probability that a dose of *300mg*
drops temperature more than 2C ?



probability of
temp. drop > 2C
given
dose = 300 mg and model (see figure)

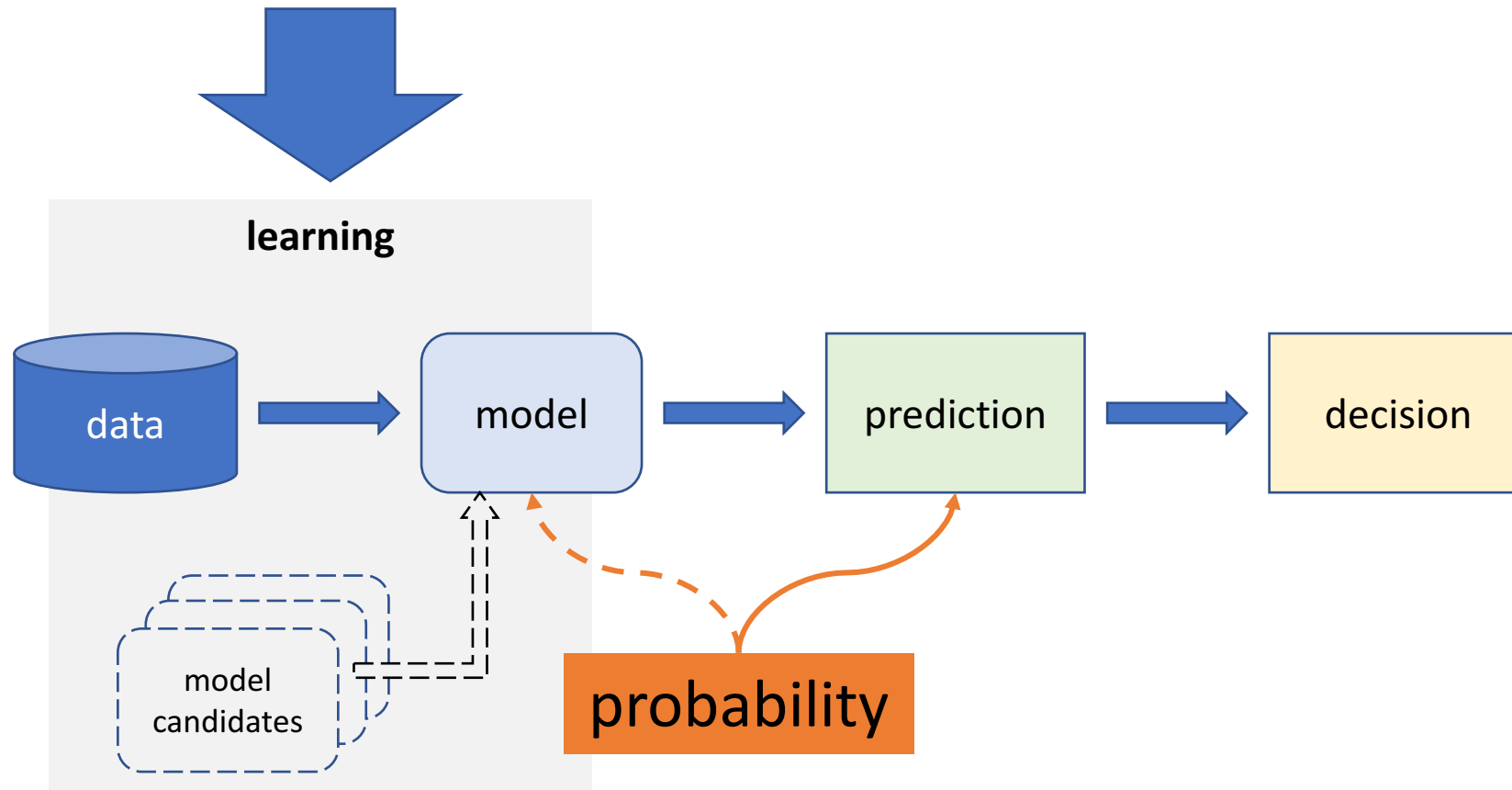
proposition

$$p(\text{temp. drop} > 2\text{C} \mid \text{dose} = 300 \text{ mg} ; \text{model})$$

given information

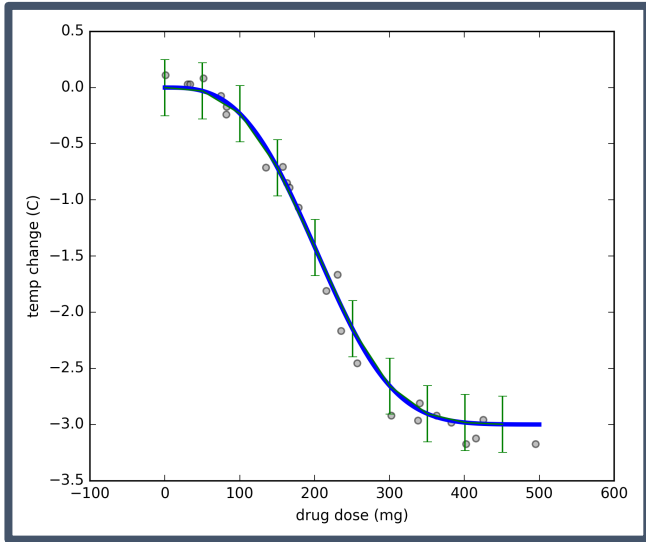
the value for this probability is provided by the model!

ML pipeline

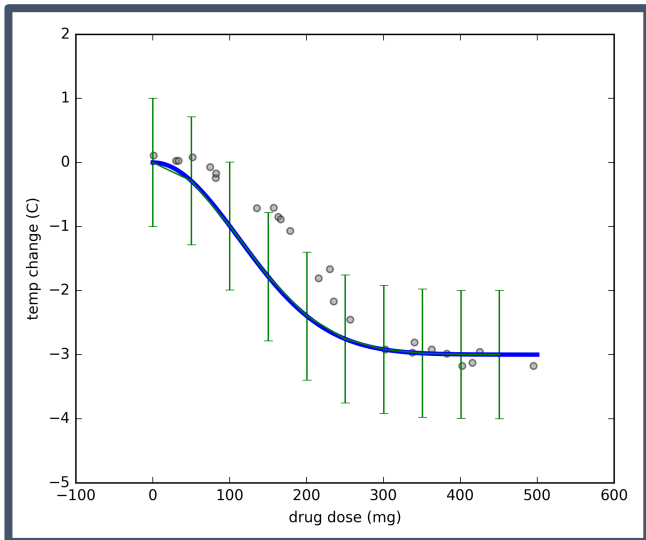


assign probabilities to models

probability :: learning



model M1



model M2

consider models where temperature drops
exponentially with dose

drop : dose^{-k} and error up to ϵ

what is the probability that the right model is M1 / M2 / ... ?

probability of

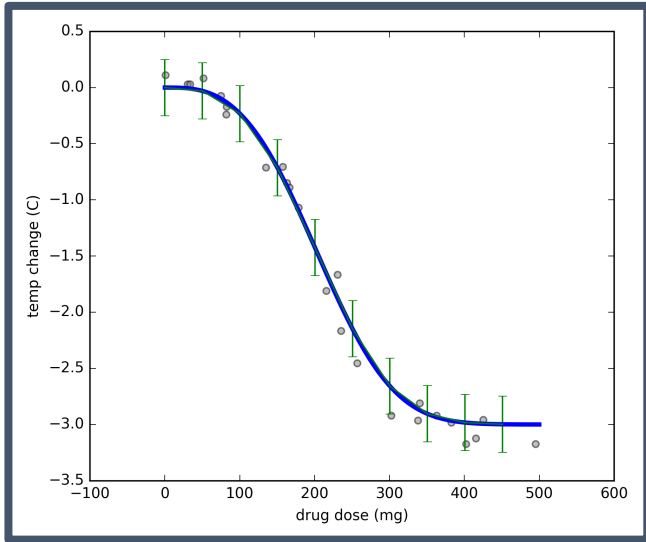
model (k, ϵ)

given

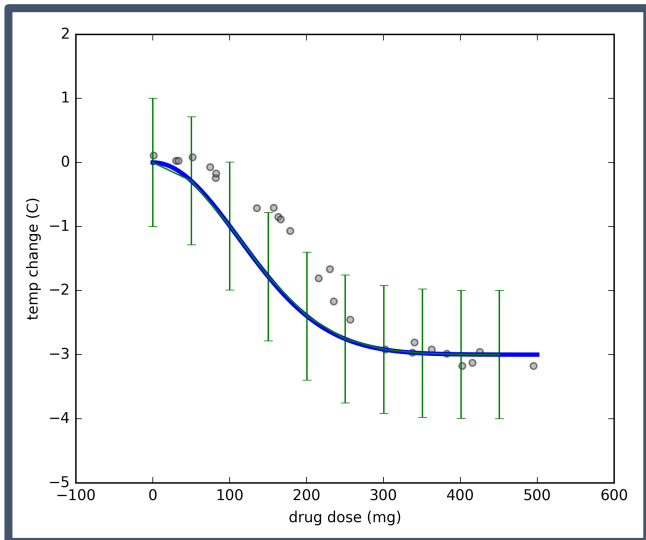
data

k from -5 to + 5 and ϵ from -2 to +2

probability :: learning



model M1



model M2

$$p(\text{model}(k, \epsilon) \mid \text{data}; k \text{ in } [-5, +5], \epsilon \text{ in } [-2, +2]) \\ p(\text{M} \mid \text{D}; \text{I})$$

from Bayes' Rule, this is proportional to

$$p(\text{data} \mid \text{M}; \text{I}) \times p(\text{M} \mid \text{I})$$

likelihood

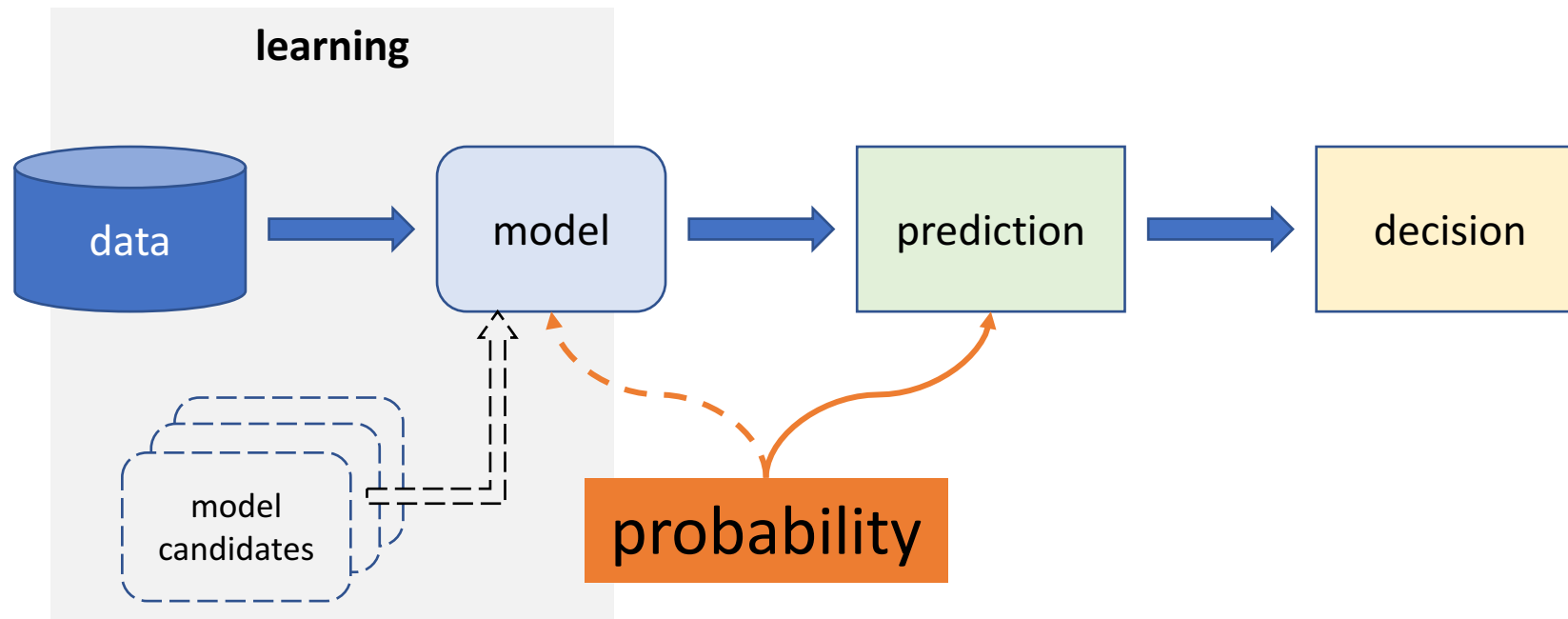
prior

we choose the model of maximum probability

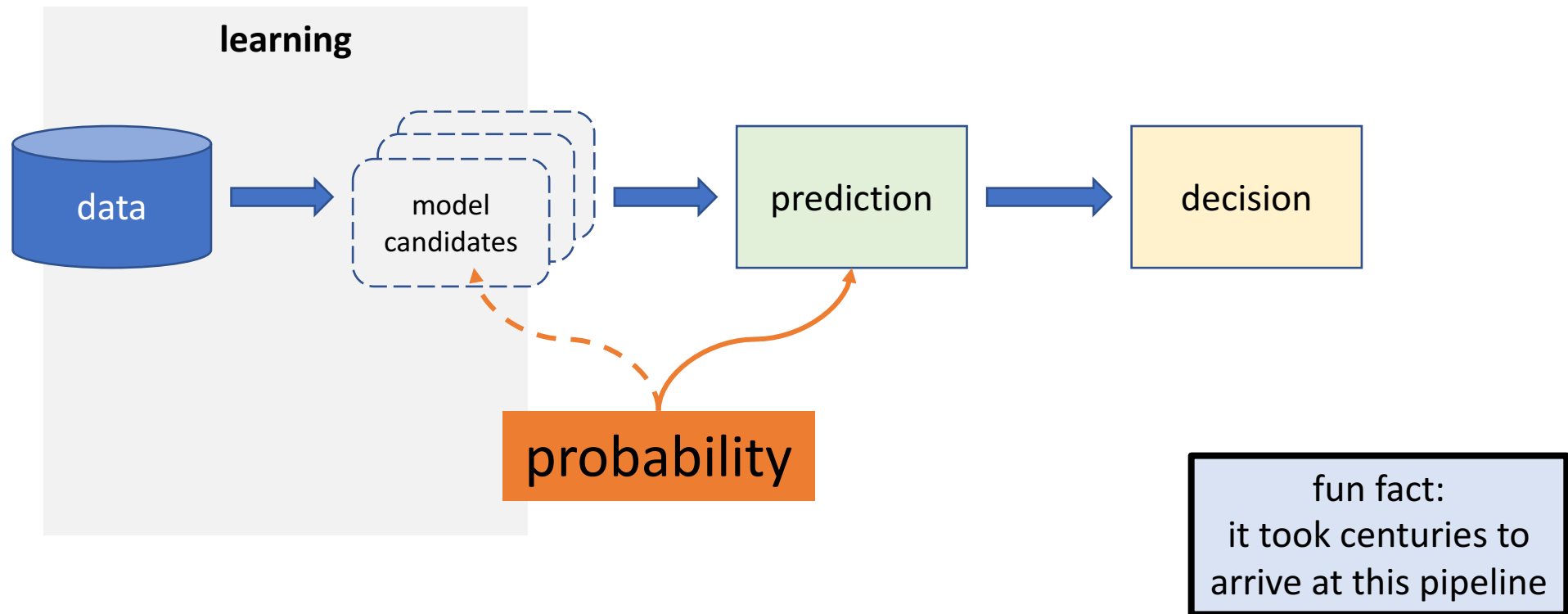
$$p(\text{M} \mid \text{D}; \text{I})$$

(do we have to?)

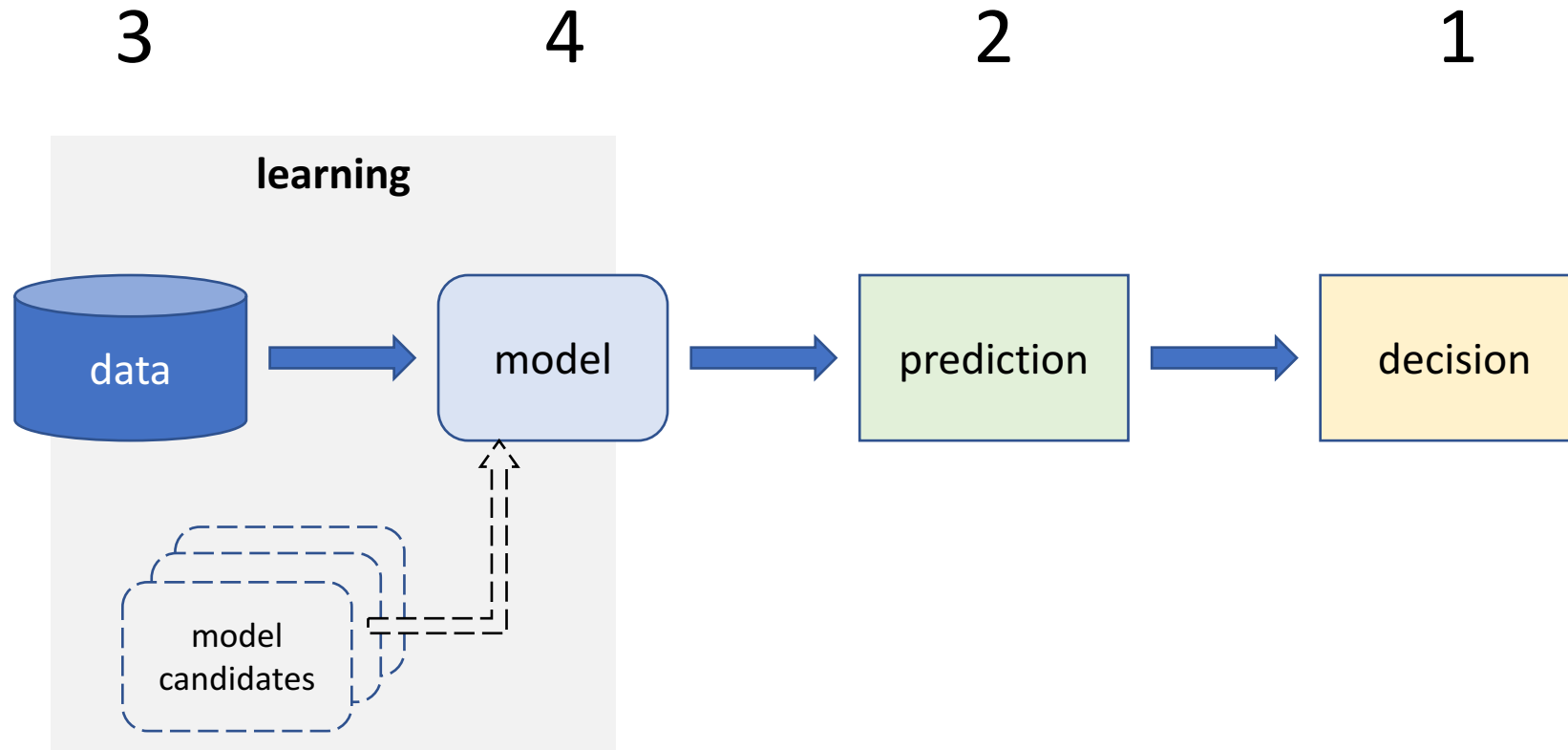
ML pipeline



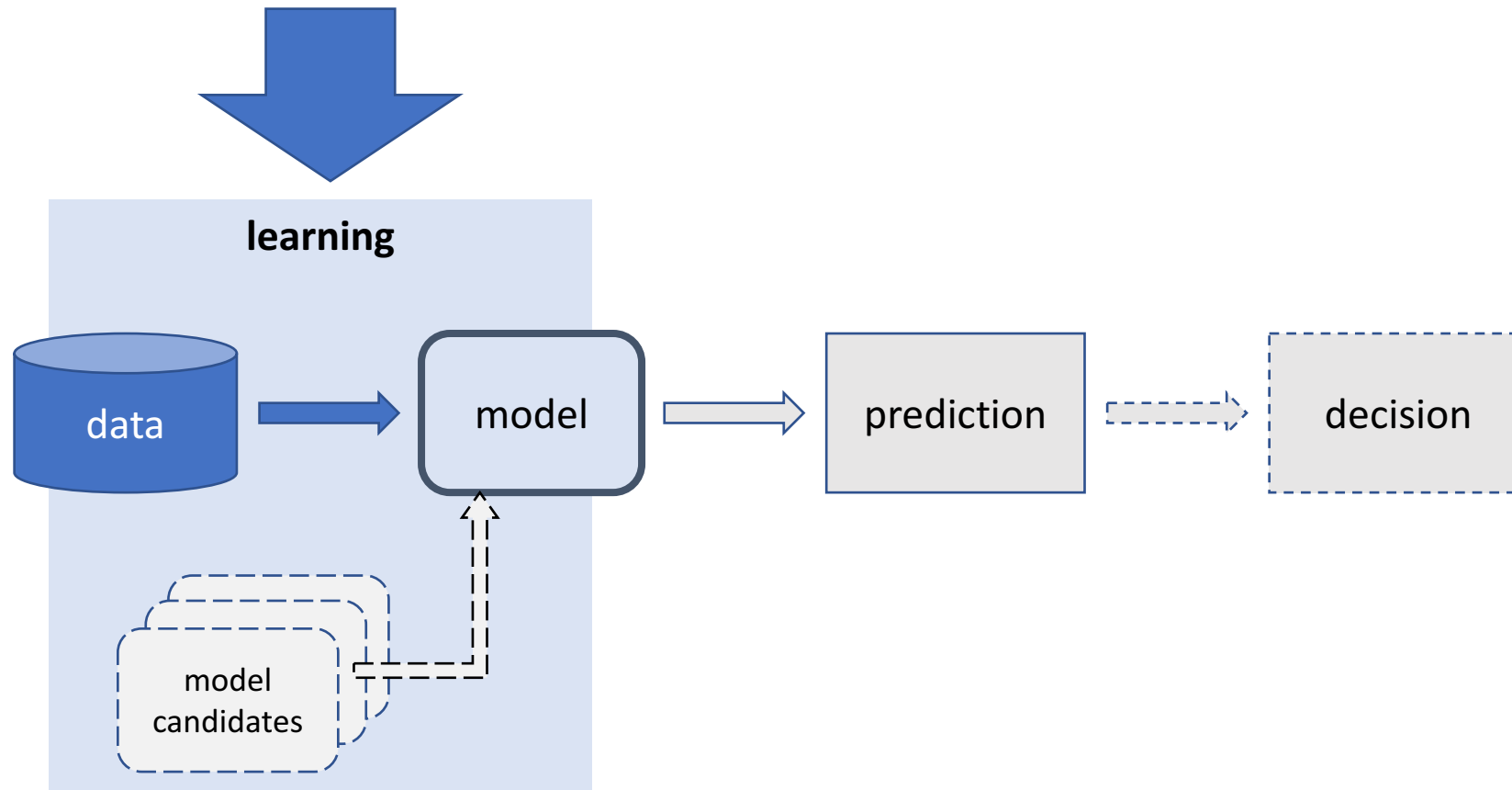
ML pipeline – the Bayesian way



ML pipeline – in practice



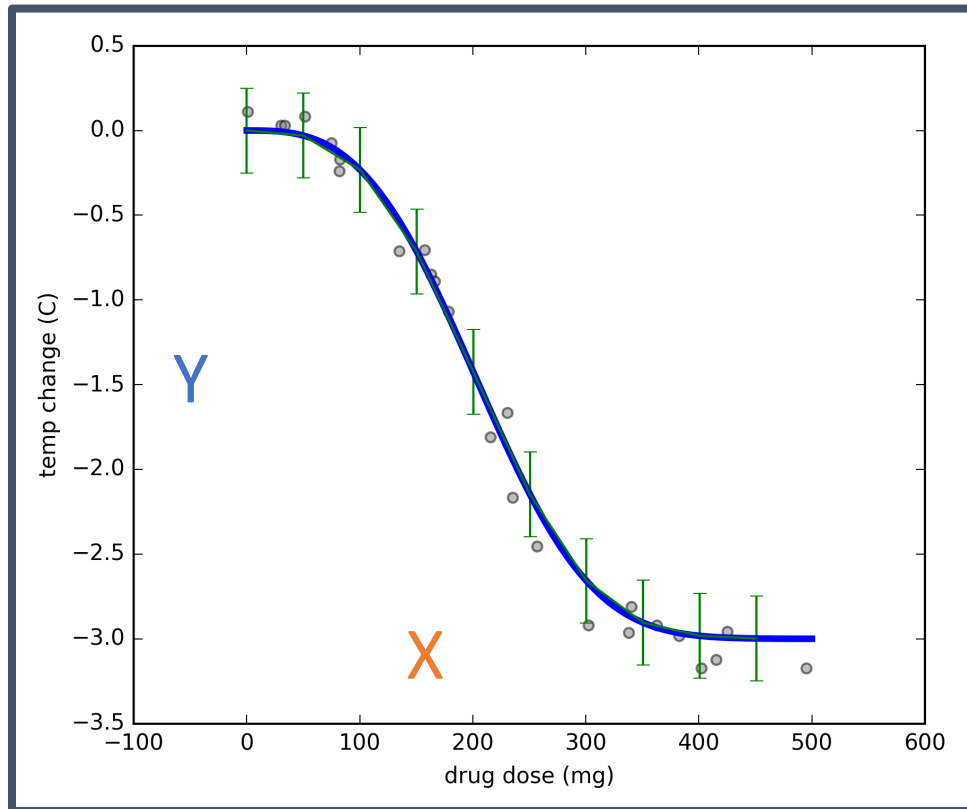
in what follows...



outline

- what is machine learning
 - examples of prediction tasks
 - data, learning, prediction, decision; probability
- probability
- algorithms
 - regression
 - classification
 - clustering
- deep learning

regression



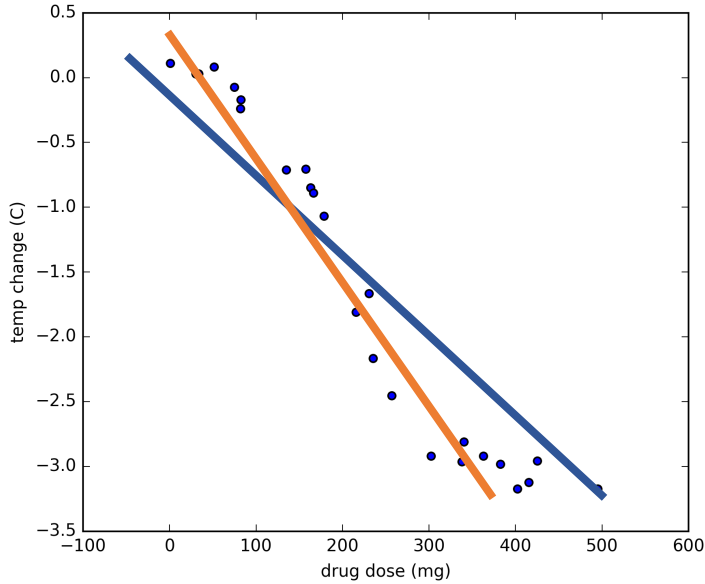
parts of the data
'features'

build model that provides
 $\text{dp}(Y = y \mid X = x; \text{Model } M)$
for real-valued Y

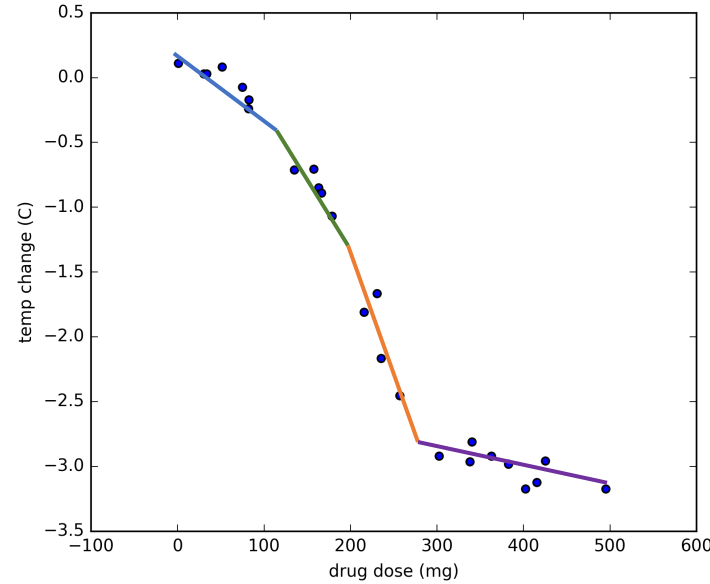
regression methods differ in
the set of model candidates
they consider

each method has corresponding
algorithm(s)
to search for best model

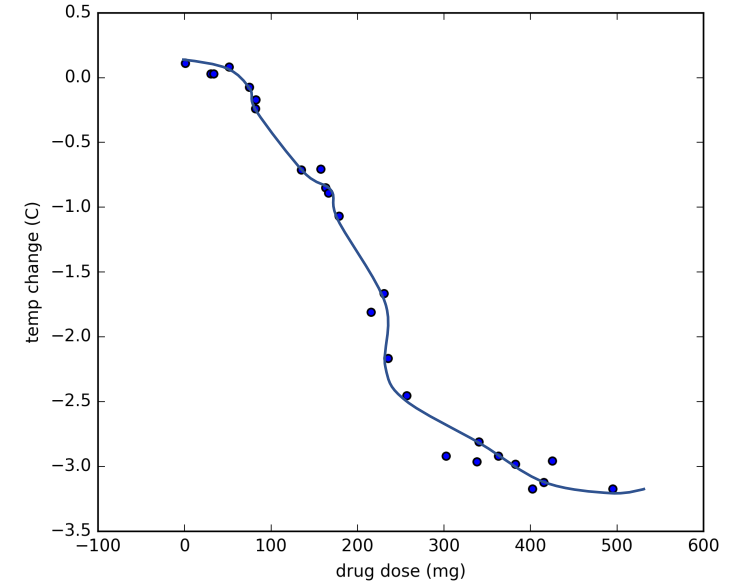
some regression methods



linear regression
line + error



segmented regression
k segments + errors



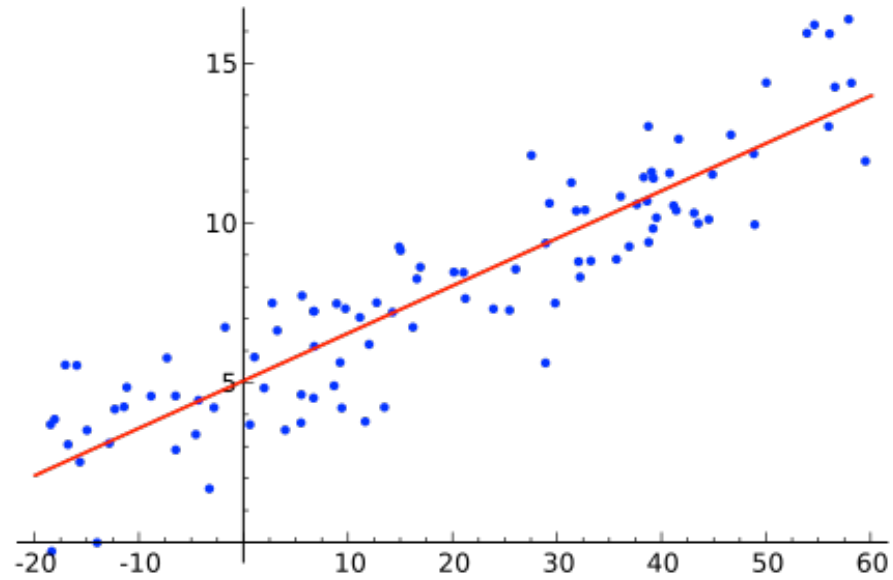
multinomial regression
curve + error

$$p(M | \text{data}; I) \propto p(\text{data} | M; I) \times p(M | I)$$

this is where methods differ

each model comes with its own

linear regression



$$\begin{aligned} Y &= E(Y|X_1, \dots, X_p) + \varepsilon \\ &= \beta_0 + \sum_{j=1}^p X_j \beta_j + \varepsilon, \end{aligned}$$

solved with linear algebra if the data points are more than the dimensions

$$\text{RSS}(\beta) = \sum_{i=1}^N (y_i - f(x_i))^2$$

ridge regression

ridge regression

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}.$$

$$\begin{aligned} \hat{\beta}^{\text{ridge}} &= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \\ &\text{subject to } \sum_{j=1}^p \beta_j^2 \leq t, \end{aligned}$$

lasso


$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

$$\begin{aligned} \hat{\beta}^{\text{lasso}} &= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \\ &\text{subject to } \sum_{j=1}^p |\beta_j| \leq t. \end{aligned}$$

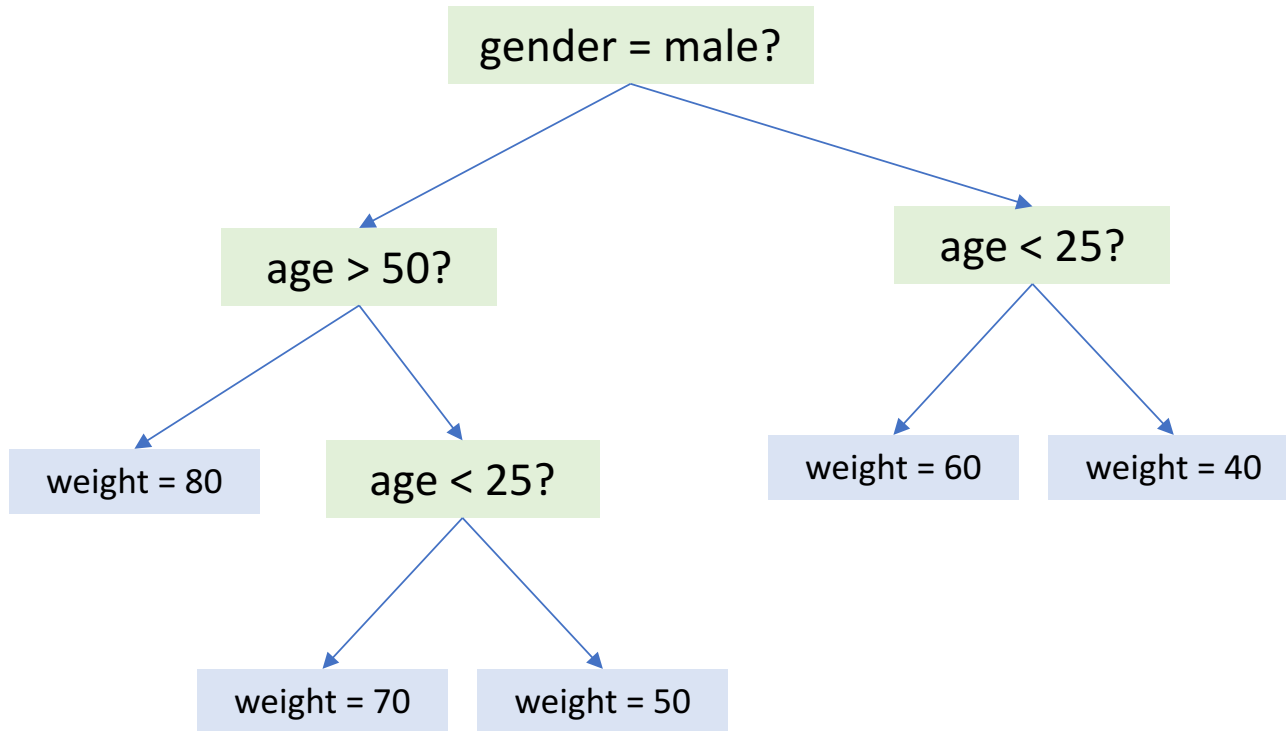
linear regression with shrinkage

the penalty on the volume of β expresses a prior

$$p(M | \text{data}; I) \propto p(\text{data} | M; I) \times p(M | I)$$

this is where methods differ 

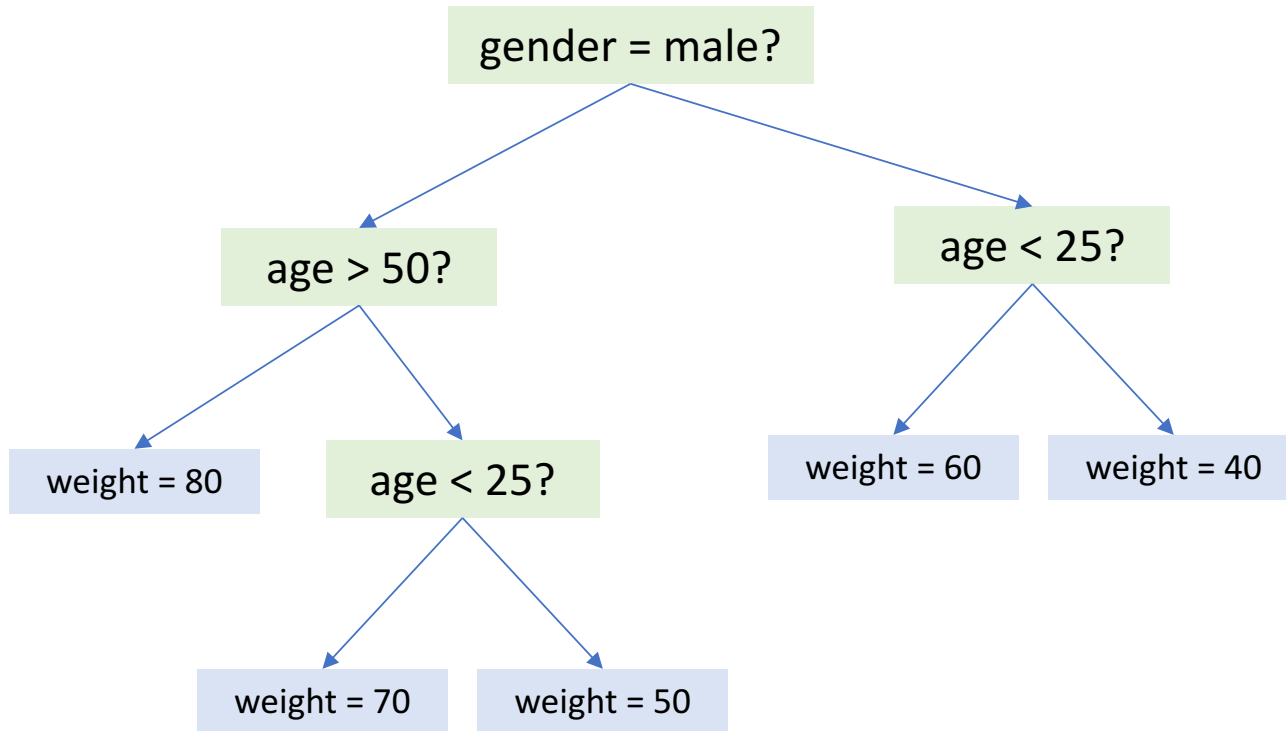
decision trees



predict the value of the leaf

build tree to minimize error
(e.g., square error)
subject to restrictions
(e.g., height of tree)

random forests



build many decision trees on
random subset of data
random subset of features

combine predictions

neural networks

idea

apply a linear model on a non-linear transformation of the input

$$h_i = g(\mathbf{x}^\top \mathbf{W}_{:,i} + c_i)$$

$$g(z) = \max\{0, z\}$$

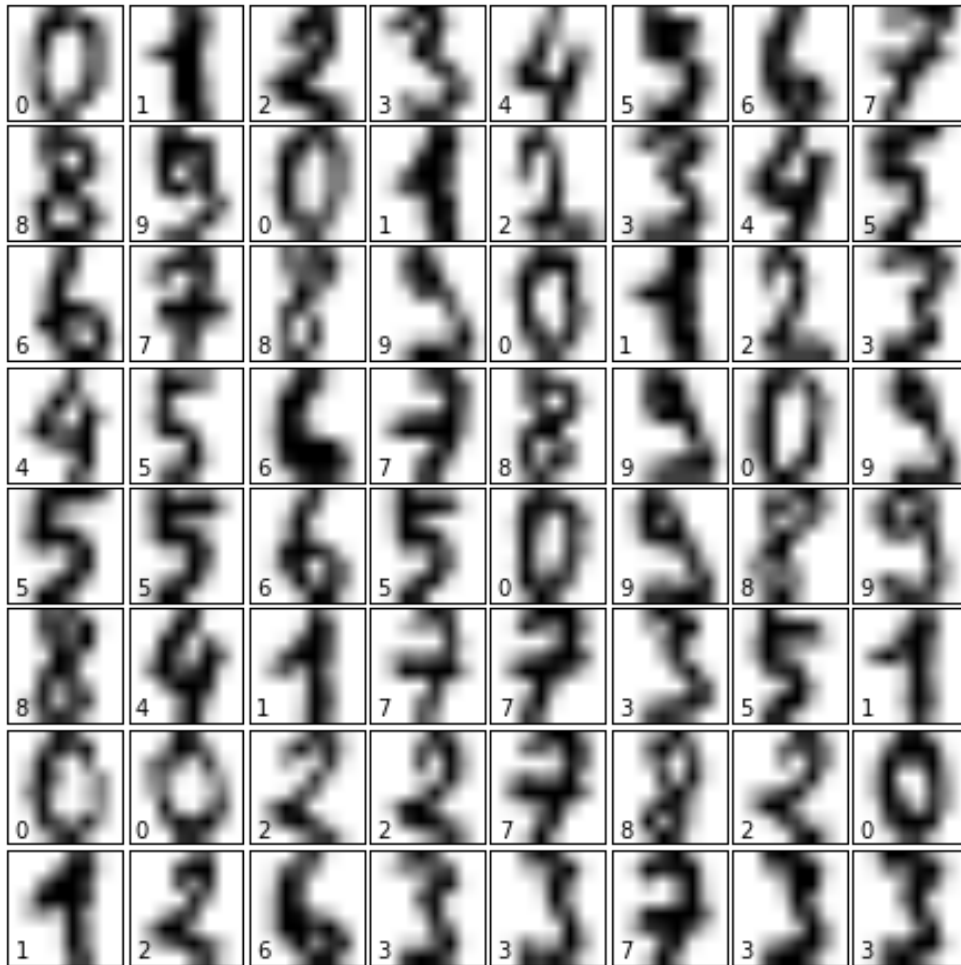
relu

$$f(\mathbf{x}; \mathbf{W}, \mathbf{c}, \mathbf{w}, b) = \mathbf{w}^\top \max\{0, \mathbf{W}^\top \mathbf{x} + \mathbf{c}\} + b.$$

outline

- what is machine learning
 - examples of prediction tasks
 - data, learning, prediction, decision; probability
- probability
- algorithms
 - regression
 - classification
 - clustering
- deep learning

classification



build model that provides
 $p(Y = y \mid X = x; \text{Model } M)$
for categorically-valued Y

classification methods differ in
the set of model candidates
they consider

each method has corresponding
algorithm(s)
to search for best model

what is X and Y for digit recognition?

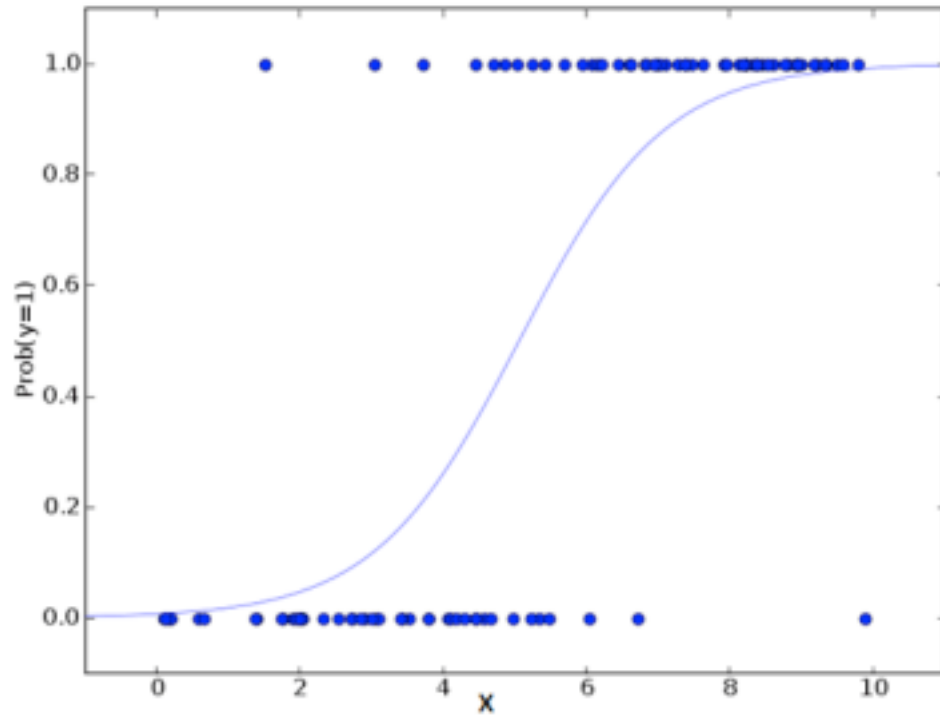
naïve-bayes

within each class, features are
distributed independently

$$p(X \mid G = j) = f_j(X) = \prod_{k=1}^p f_{jk}(X_k).$$

$$\begin{aligned} \text{logit} \frac{\Pr(G = \ell | X)}{\Pr(G = J | X)} &= \log \frac{\pi_\ell f_\ell(X)}{\pi_J f_J(X)} \\ &= \log \frac{\pi_\ell \prod_{k=1}^p f_{\ell k}(X_k)}{\pi_J \prod_{k=1}^p f_{Jk}(X_k)} \\ &= \log \frac{\pi_\ell}{\pi_J} + \sum_{k=1}^p \log \frac{f_{\ell k}(X_k)}{f_{Jk}(X_k)} \\ &= \alpha_\ell + \sum_{k=1}^p g_{\ell k}(X_k). \end{aligned}$$

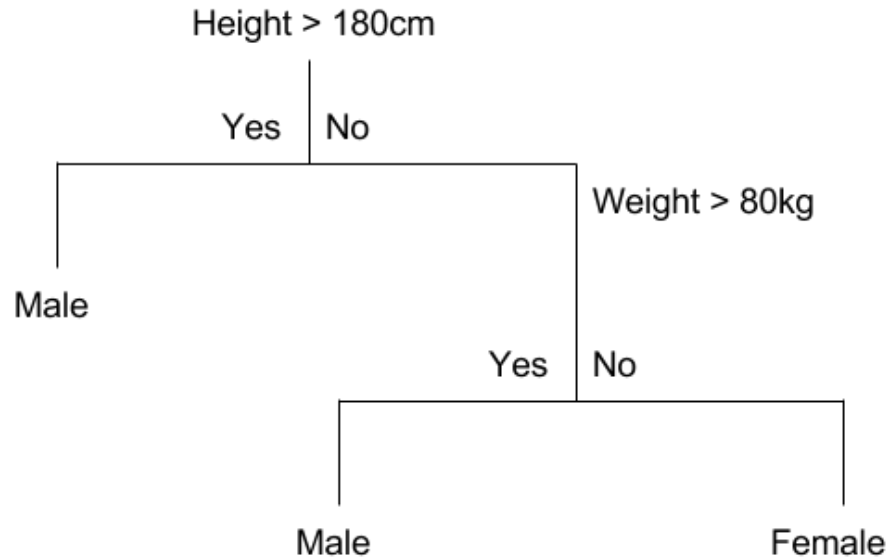
logistic regression



$$\Pr(Y_i = c) = \frac{e^{\beta_c \cdot \mathbf{X}_i}}{\sum_{k=1}^K e^{\beta_k \cdot \mathbf{X}_i}}$$

softmax

decision trees & random forests



very similar to regression methods
leafs assign probabilities to classes

neural networks

idea

apply a linear model on a non-linear transformation of the input

$$h_i = g(\mathbf{x}^\top \mathbf{W}_{:,i} + c_i)$$

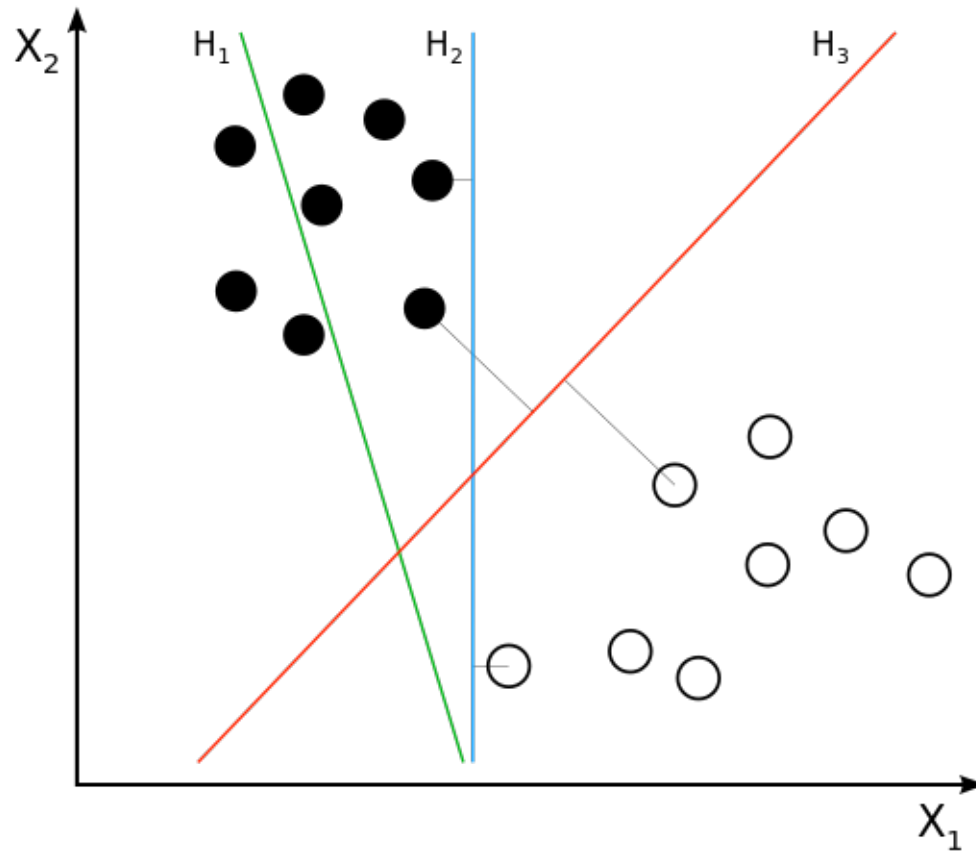
$$g(z) = \max\{0, z\}$$

$$f(\mathbf{x}; \mathbf{W}, \mathbf{c}, \mathbf{w}, b) = \mathbf{w}^\top \max\{0, \mathbf{W}^\top \mathbf{x} + \mathbf{c}\} + b.$$



evidence for one class

support-vector machines



separate the classes
with hyperplanes

outline

- what is machine learning
 - examples of prediction tasks
 - data, learning, prediction, decision; probability
- probability
- algorithms
 - regression
 - classification
 - clustering
- deep learning

supervised and unsupervised learning

the methods we saw for
regression and classification
are cases of 'supervised' learning

build model that provides

$$p(Y = y \mid X = x; \text{Model } M)$$

other data features

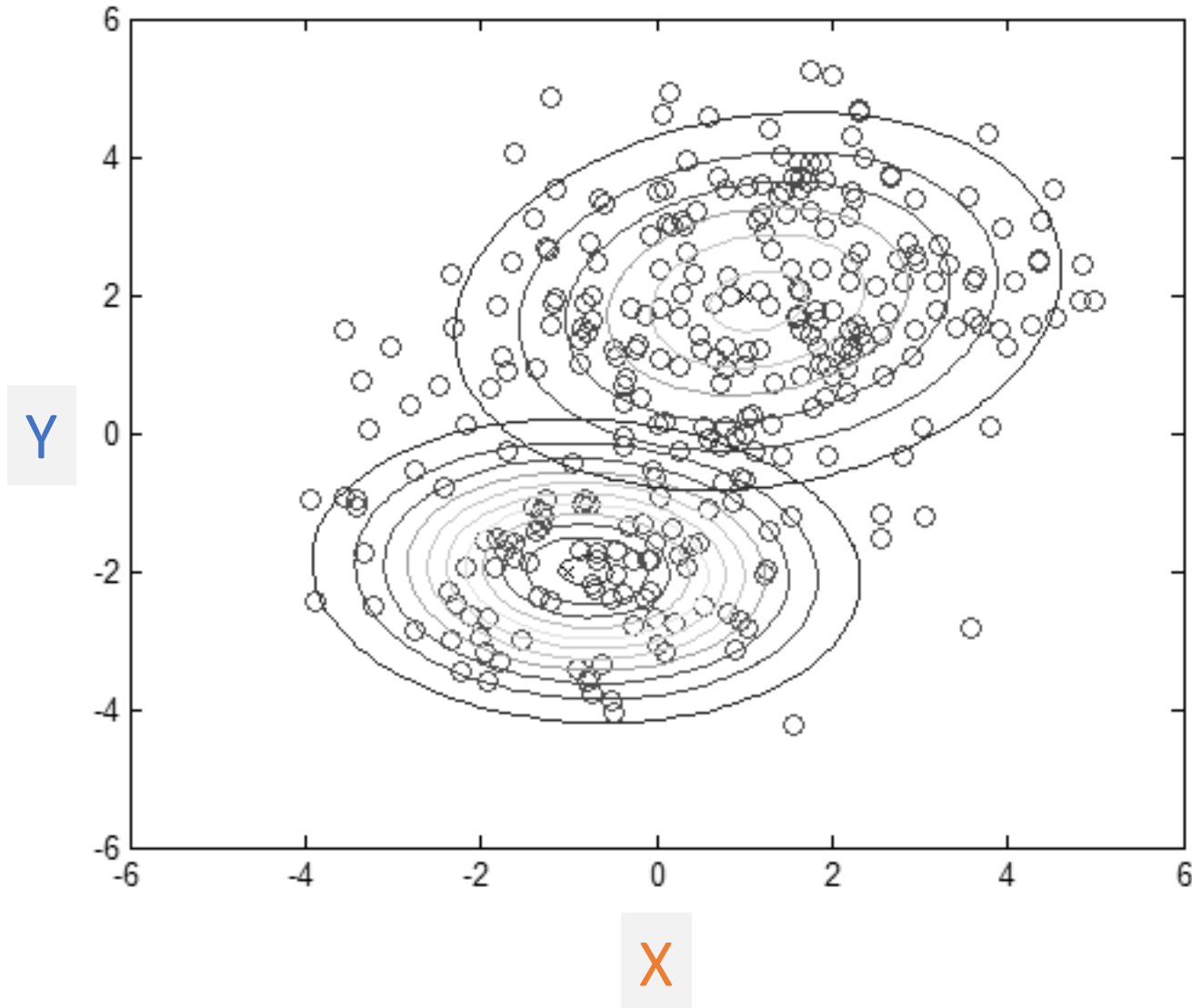
some data features

build model that provides

$$p(X = x, Y = y; \text{Model } M)$$

'unsupervised' learning

unsupervised learning

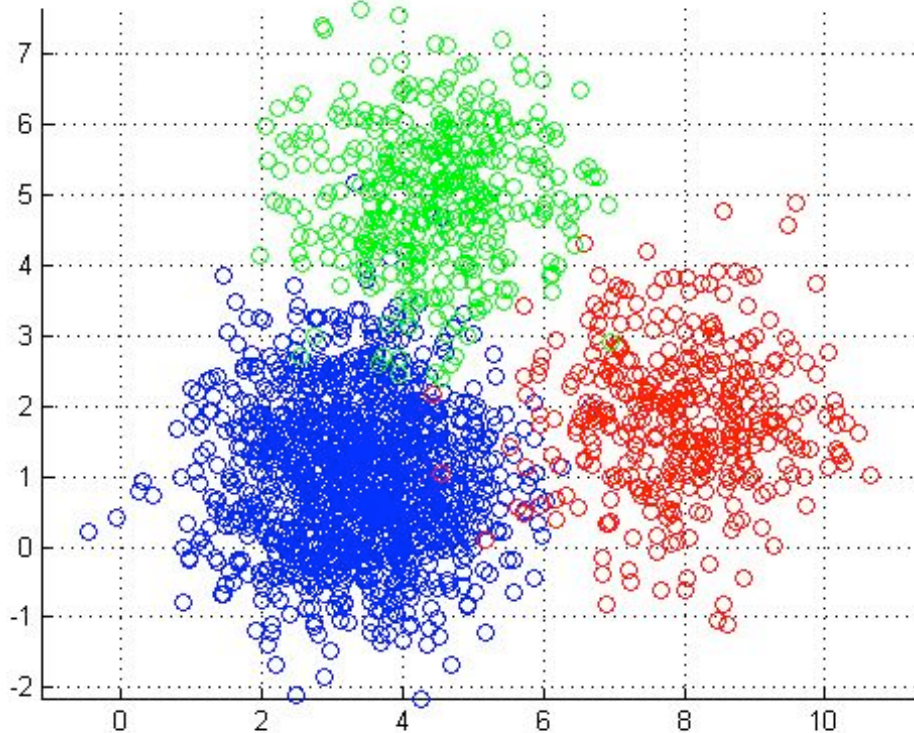


build model that provides

$\text{dp}(X = x, Y = y; \text{Model } M)$

find structure in the data

k-means clustering



X

model

gaussian mixture
with equal symmetric variance

clustering

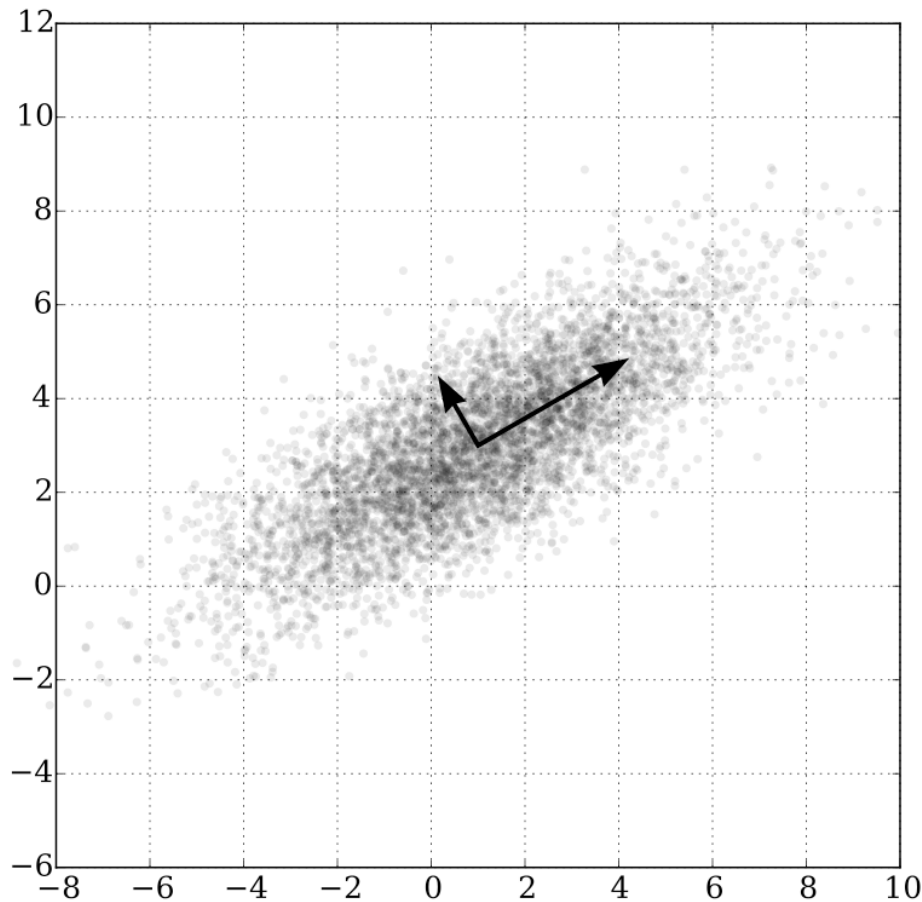
(finding the clusters)

assign points to clusters
so that total distance from cluster center is
minimized

k-means

assign points to cluster of nearest center
compute centers from assigned points
repeat

PCA



project the data on orthogonal system
so that successive dimensions maximize
remaining variance

useful for dimensionality reduction

model

gaussian distribution
on k orthogonal dimensions
equal noise on the others

outline

- what is machine learning
 - examples of prediction tasks
 - data, learning, prediction, decision; probability
- probability
- algorithms
 - regression
 - classification
 - clustering
- deep learning

deep learning

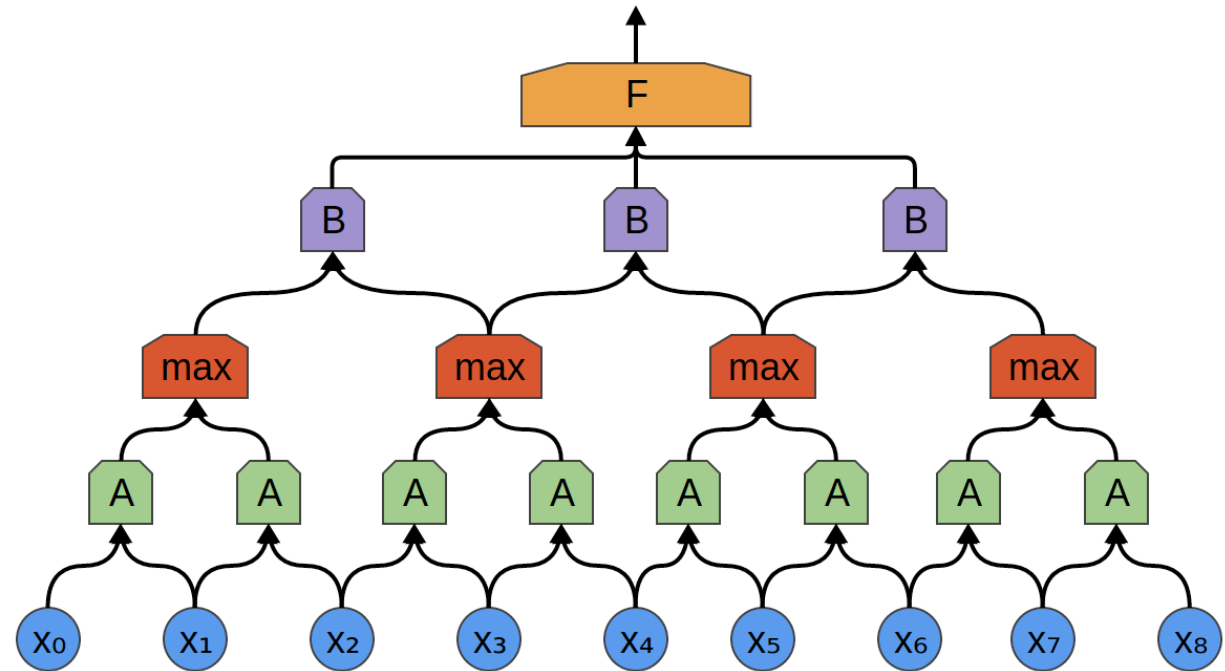
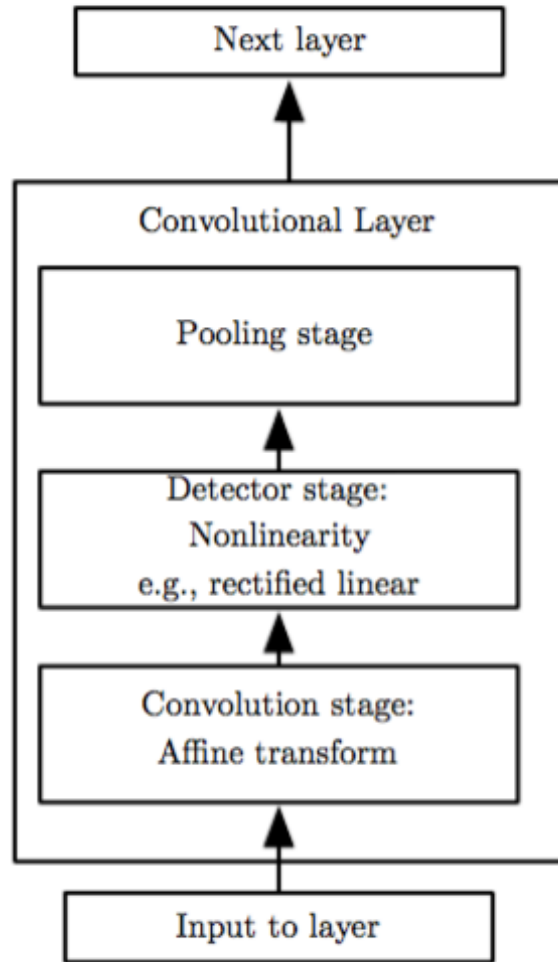
basically another name for 'neural networks'
with many layers
and generalized structure

rebranded due to
efficiency and good results on difficult tasks

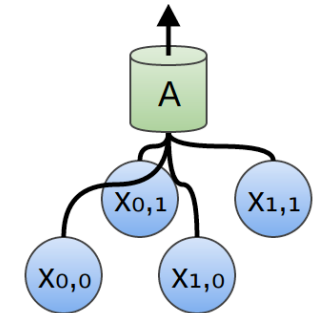
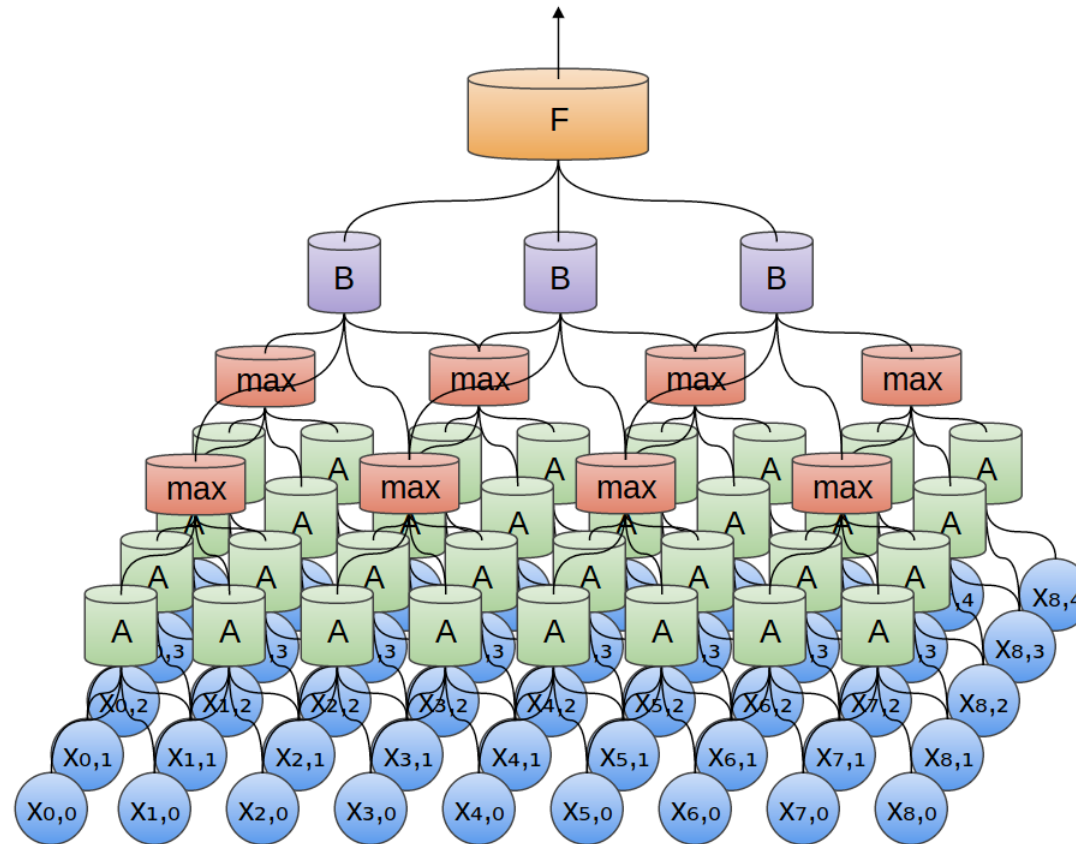
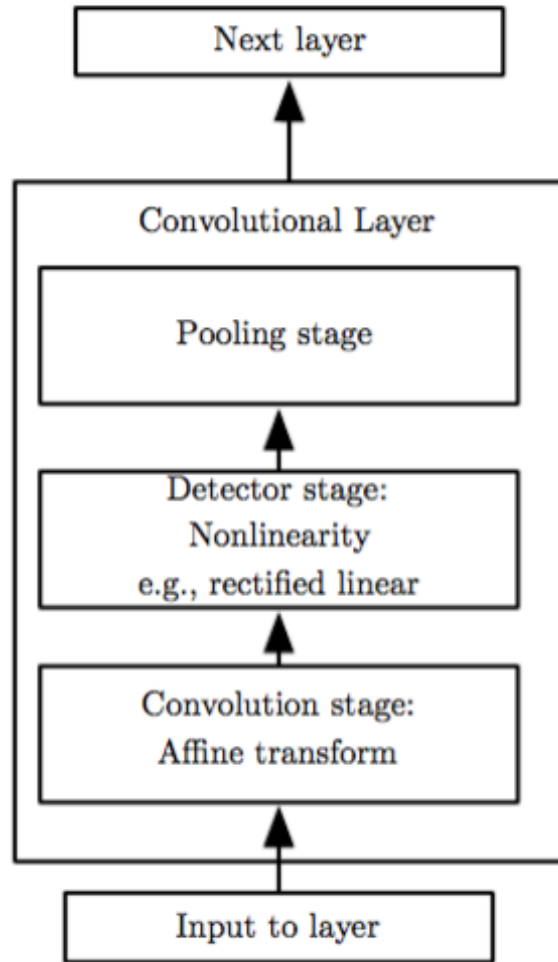
language
(translation, sentence completion, voice recognition)
image recognition

recurrent neural networks

convolutional neural networks



convolutional neural networks



part 2: platforms and software

outline

- scikit-learn
- open source deep-learning libraries
- cloud ML products
- apache spark

scikit-learn



python ML library
on top of scipy stack

many general ML algorithms
standardized pipeline
ideal for fast prototyping
on moderate datasets

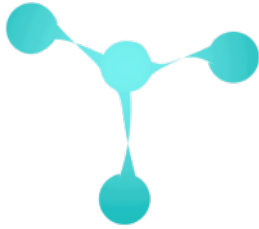
deep learning :: tensorflow



TensorFlow

deep learning library
based on user-defined
computation graphs
for out-of-python optimization

deep learning :: other



torch.ch

open source machine learning library
scientific framework, programming language (Lua)
used by Facebook Research



theano

<http://deeplearning.net/software/theano/>
deep learning with efficient numerical operations



microsoft cognitive toolkit (cntk)

<https://cntk.ai/>
tensorflow alternative



keras

simpler tensorflow, theano, cntk in python

cloud :: google



Google Cloud Platform

Cloud ML Engine

basically offers the ML pipeline
with Deep Learning models
implemented in Tensorflow

other services

trained models for other applications

speech, video or image tagging, translation

<https://cloud.google.com/products/machine-learning/>

pricing: about 0.5\$ per hour

cloud :: other



amazon aws

classification and regression
with logistic and linear regression

microsoft azure

‘cortana intelligence’

ML pipeline

apache spark



machine learning algorithms
on top of Spark

iterative optimization

part 3: hands-on session

outline

- scikit-learn
- tensorflow

scikit-learn

<http://scikit-learn.org/>

tensorflow

<https://www.tensorflow.org/>

the end...

ML pipeline

