Machine Learning: the fly-by overview

Honey

Diana Pfeil

Machine Learning

Supervised Learning

Unsupervised Learning

Statistical Modeling

Descriptive, Predictive, and Prescriptive Analytics

AI

What about Big Data?

Supervised Learning

$\mathbf{x_i}$	features (input variables)
y_i	target (output variable)
$(x_i, y_i), i = 1, \ldots, m$	training set

Goal: learn a function

$$h : \mathcal{X} \rightarrow \mathcal{Y}$$

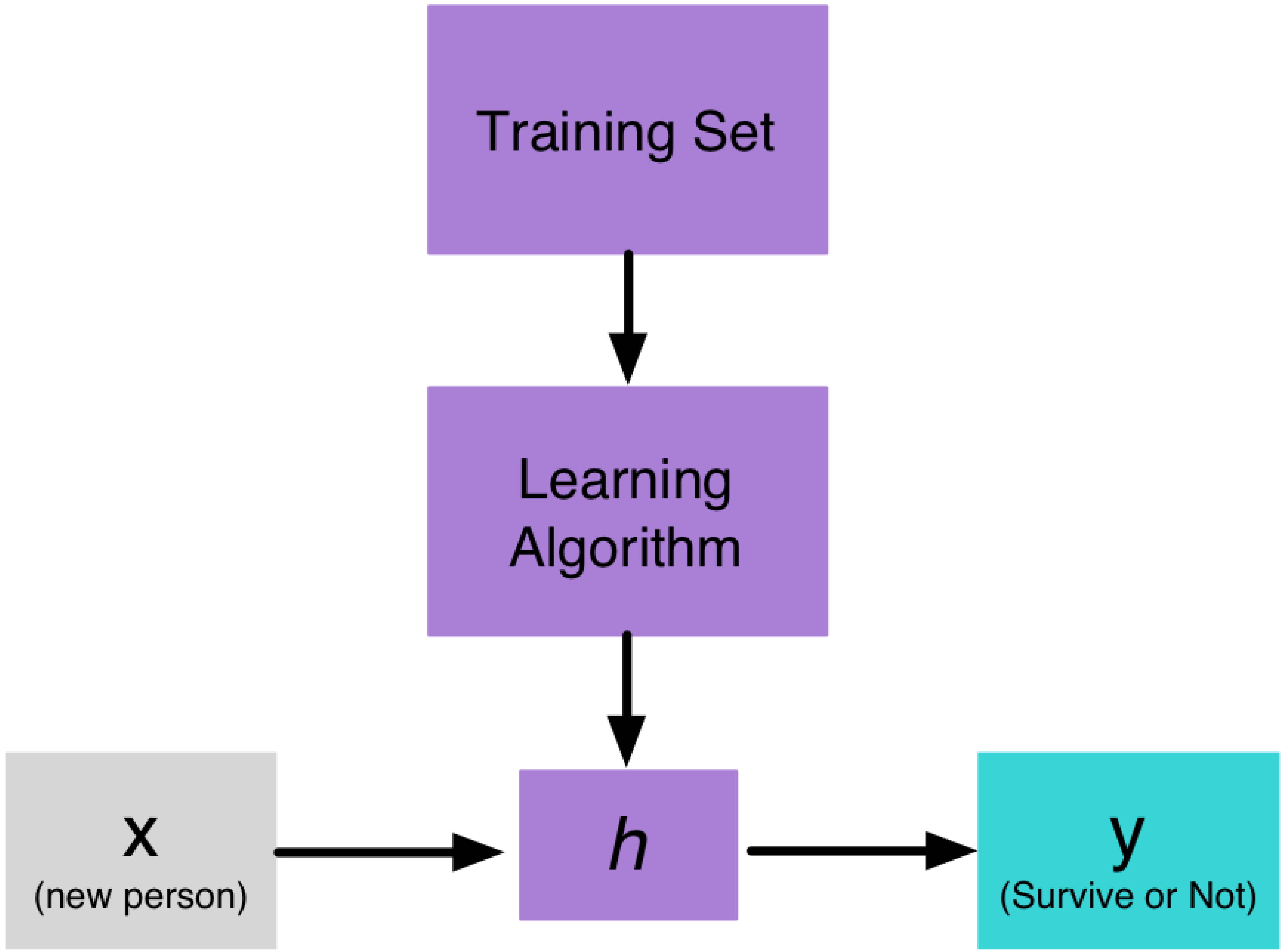
such that $h(x)$ is a good predictor of y on **new** data

features x can be

numeric/metric	Age: 14, 56, 1
ordinal	Ranking: 1st, 2nd, 3rd
categorical/nominal	Sex: male/female

target y can be

continuous (regression)	Housing Price: 500K, 150K, 2MM
categorical (classification)	Survival: Perish, Survive


Training Set
Learning Algorithm
x
(new person)
h
y
(Survive or Not)

Example Data

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2	<NA>	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.3	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9	<NA>	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.1	<NA>	S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.5	<NA>	Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.9	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.1	<NA>	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1	<NA>	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.1	<NA>	C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.6	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.1	<NA>	S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.3	<NA>	S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.9	<NA>	S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16.0	<NA>	S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.1	<NA>	Q
18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0	<NA>	S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18.0	<NA>	S
20	1	3	Masselmani, Mrs. Fatima	female	NA	0	0	2649	7.2	<NA>	C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26.0	<NA>	S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13.0	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0	<NA>	Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.1	<NA>	S

But where do we find this h?

This is the process of doing supervised learning

Models for Supervised Learning

Classification Tree	Regression Tree
Random Forest	Linear Regression
Support Vector Machine	Logistic Regression
Boosting	K-Nearest Neighbors
Naive Bayes	Neural Networks/Deep Learning (AI)

ML Workflow for prototyping

1. Clean and explore the data (EDA)
2. Come up with new features (feature engineering)
3. Split data into training and validation
4. Tune the model and parameters using cross-validation
5. Compare model results

Key skills for *doing* data science/ML/AI

- Defining the problem and what success looks like
- Exploratory data analysis
- Machine learning
- Setting aside time to *think*
- Data communication and visualization

A Typical Toolkit

- Python with pandas, numpy, scipy, jupyter
- tensorflow/keras/pytorch
- Unix utilities

Other options

- JVM-based eco-system: Spark, Hadoop
- Vowpal wabbit
- R, RStudio, RMarkdown
- SPSS, Excel, RapidMiner

More on data cleaning and EDA

Purpose of EDA

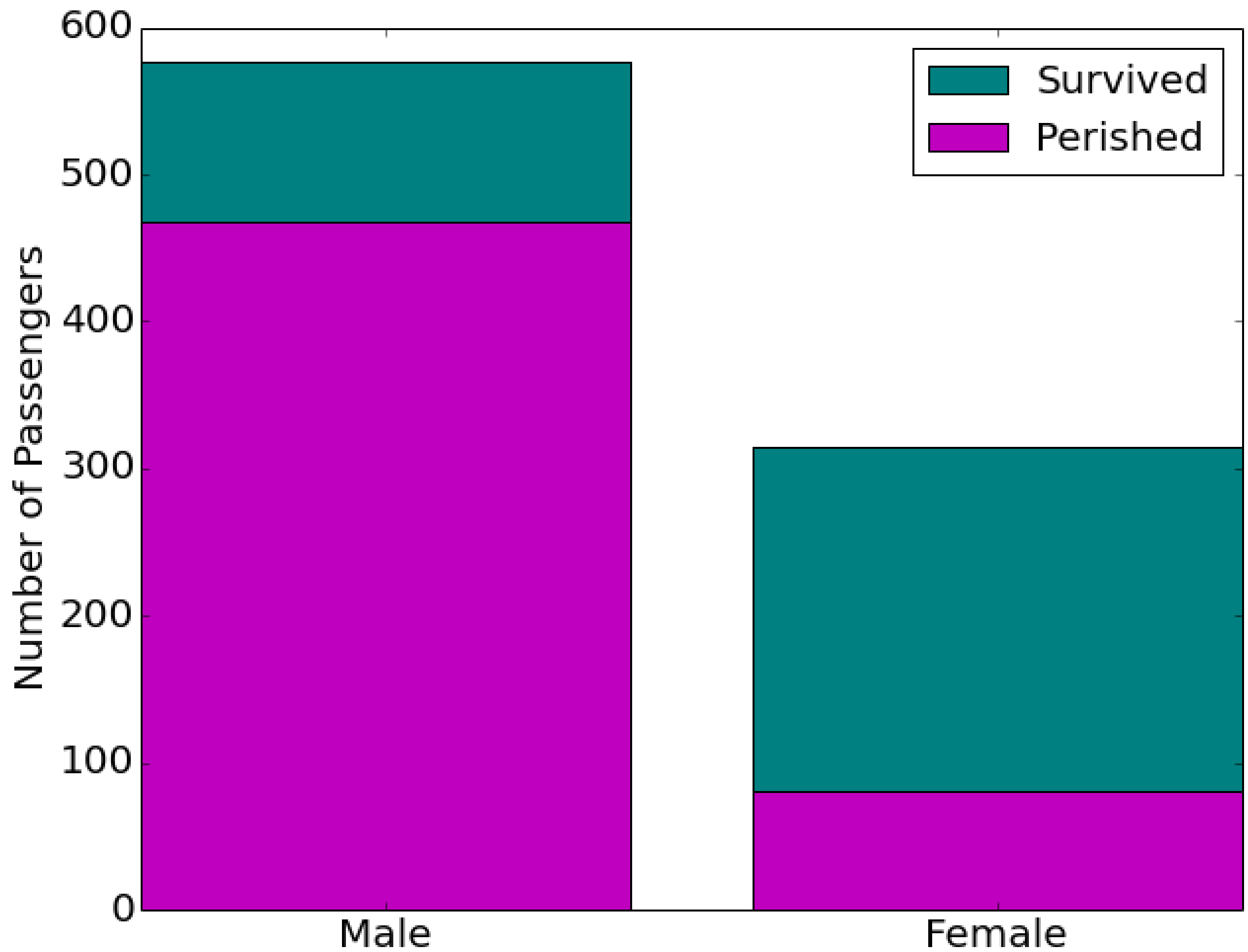
- Do you have the right data for the question you’re trying to answer?
- Check assumptions and detect mistakes
- Get a sense for the data you have, and start to understand how it can answer the question at hand

Big Data Borat
@BigDataBorat

Data Science is 99% preparation, 1% misinterpretation.

71 8:31 AM - Apr 18, 2013

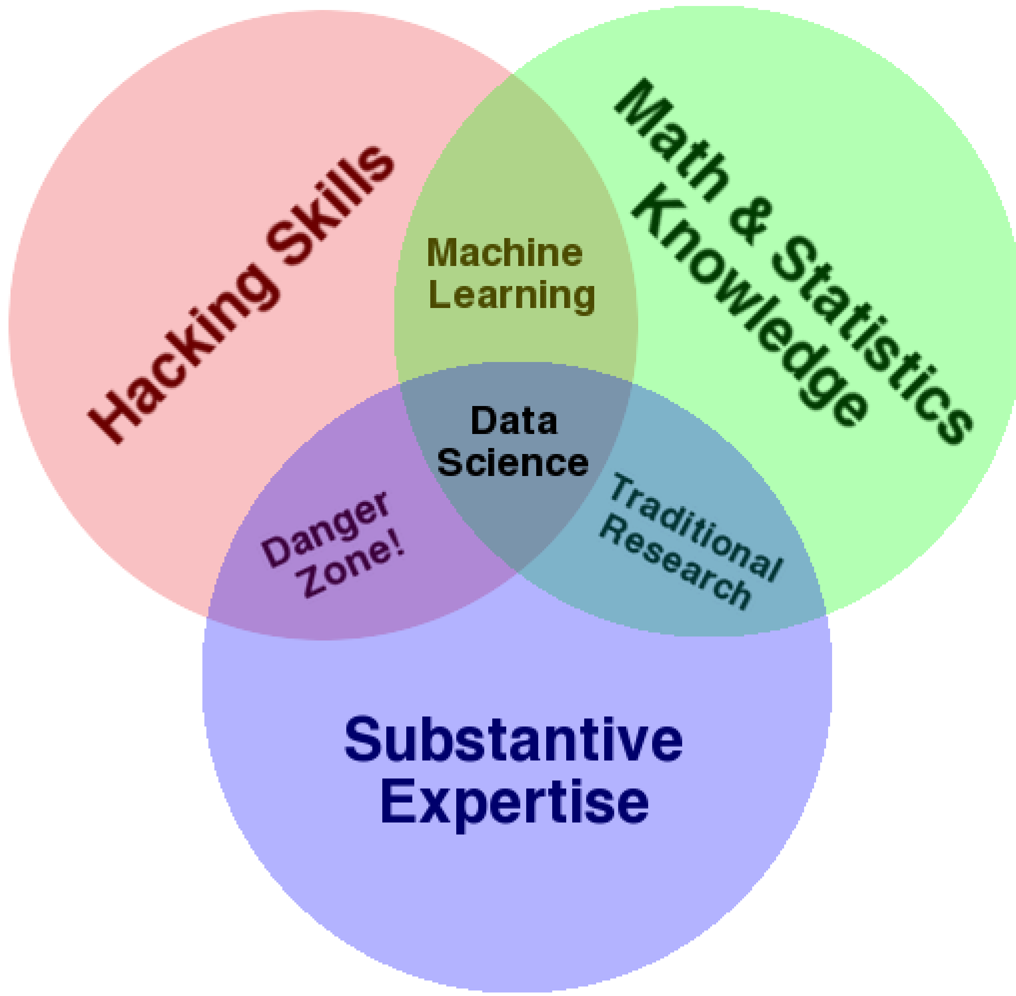
187 people are talking about this


Survived
Perished
Number of Passengers
600
500
400
300
200
100
0
Male
Female

I'm a data janitor

—Josh Wills, head of Data Engineering at Slack

Feature engineering



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

A little exercise

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2	<NA>	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.3	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9	<NA>	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.1	<NA>	S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.5	<NA>	Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.9	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.1	<NA>	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1	<NA>	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.1	<NA>	C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.6	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.1	<NA>	S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.3	<NA>	S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.9	<NA>	S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16.0	<NA>	S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.1	<NA>	Q
18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0	<NA>	S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18.0	<NA>	S
20	1	3	Masselmani, Mrs. Fatima	female	NA	0	0	2649	7.2	<NA>	C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26.0	<NA>	S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13.0	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0	<NA>	Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.1	<NA>	S

ML Models

Supervised Learning

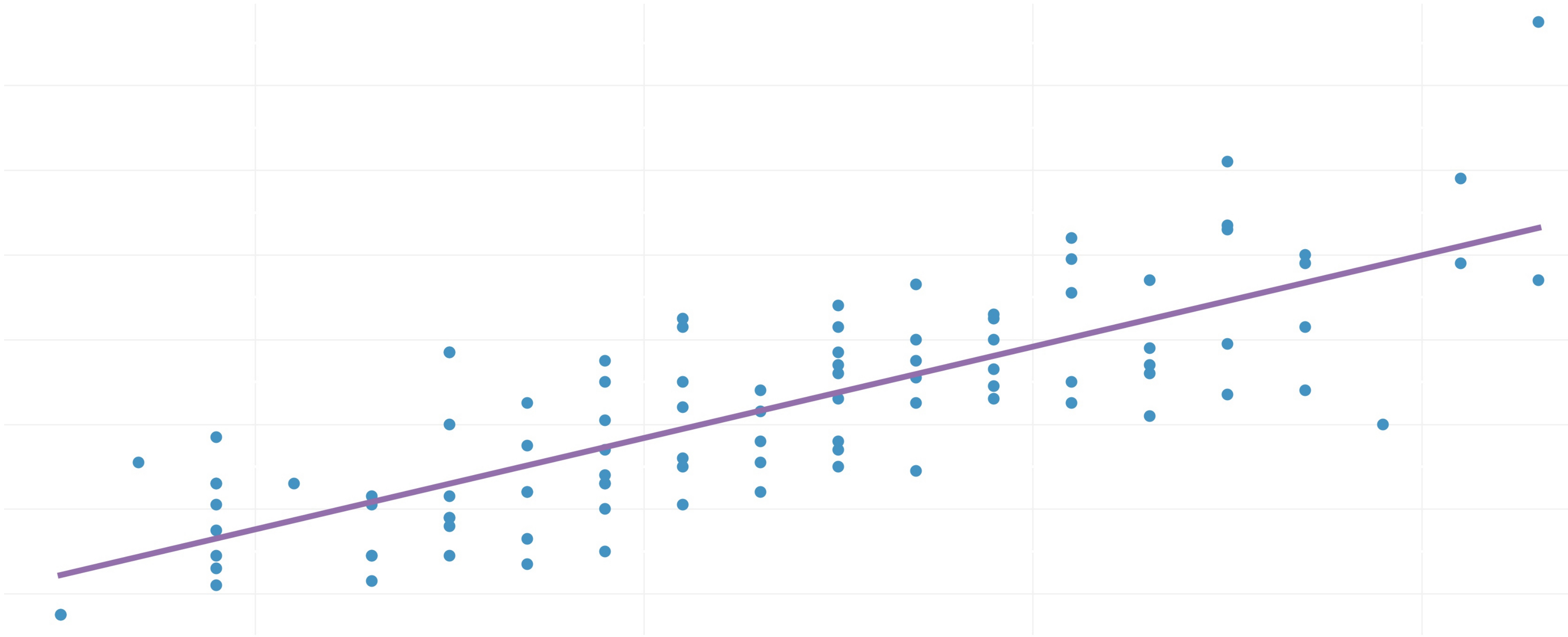
$\mathbf{x_i}$	features (input variables)
y_i	target (output variable)
$(x_i, y_i), i = 1, \ldots, m$	training set

Goal: learn a function

$$h : \mathcal{X} \to \mathcal{Y}$$

such that $h(x)$ is a good predictor of y on **new** data

Linear Regression



Goal: find the best line

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$$

Sum of squared error loss: $(y - \hat{y})^2$

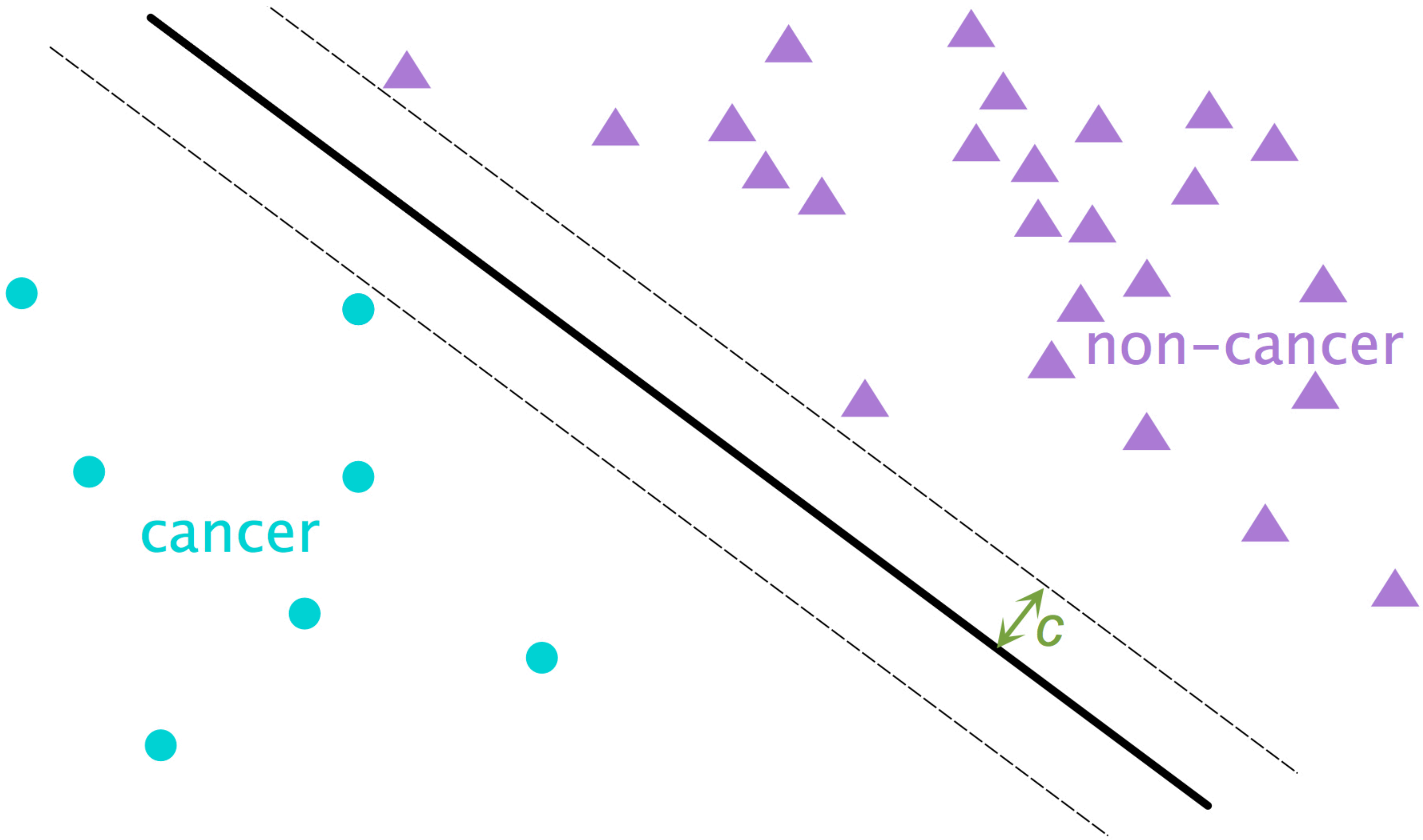
Linear Regression Advantages

- Highly interpretable
- Can assess statistical significance of each predictor

Disadvantages

- Limiting: only works for a linear relationship between features x and y
- Requires strong assumptions: no collinearity, homoscedasticity, normally distributed errors
- With collinearity, the regression is unstable (high variance)
- Sensitive to outliers

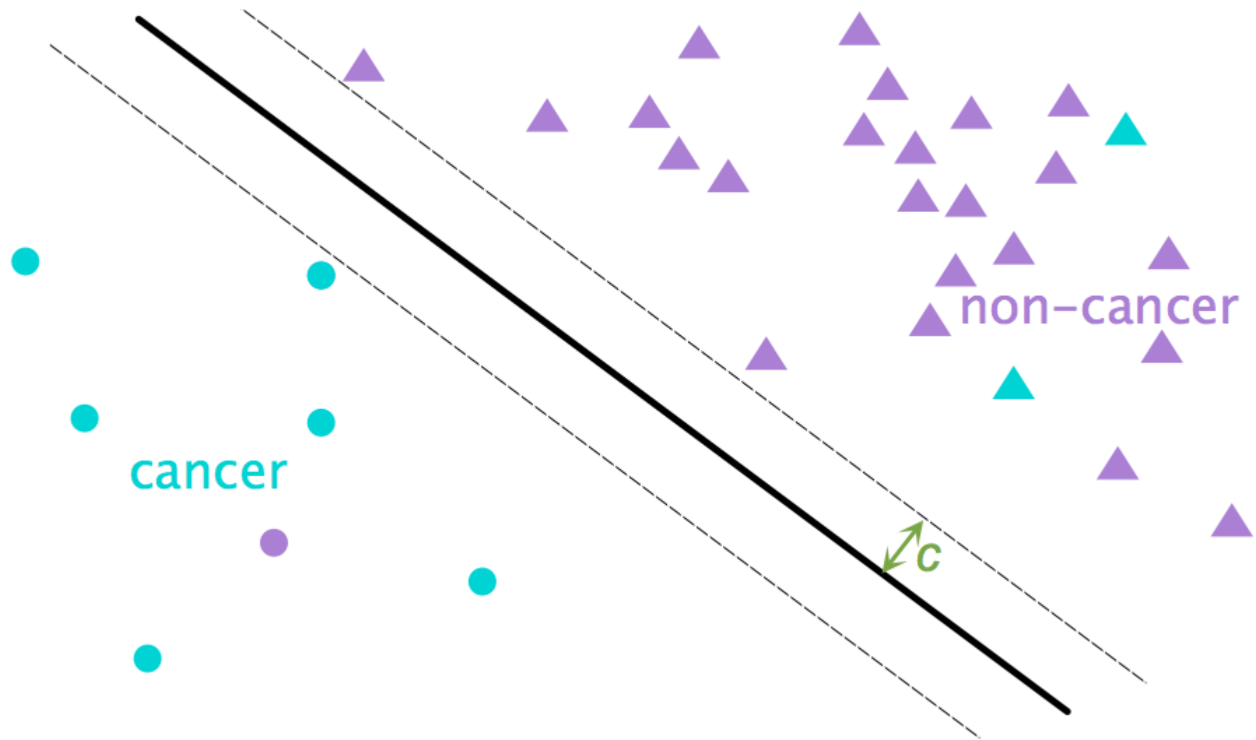
Support Vector Machine



$$\max c$$

s.t. $\|\beta\| = 1$ and $y_i(\beta^T x_i) \geq c \quad i = 1, \ldots, n$

SVM: Robust Classification



$$\max c - p$$

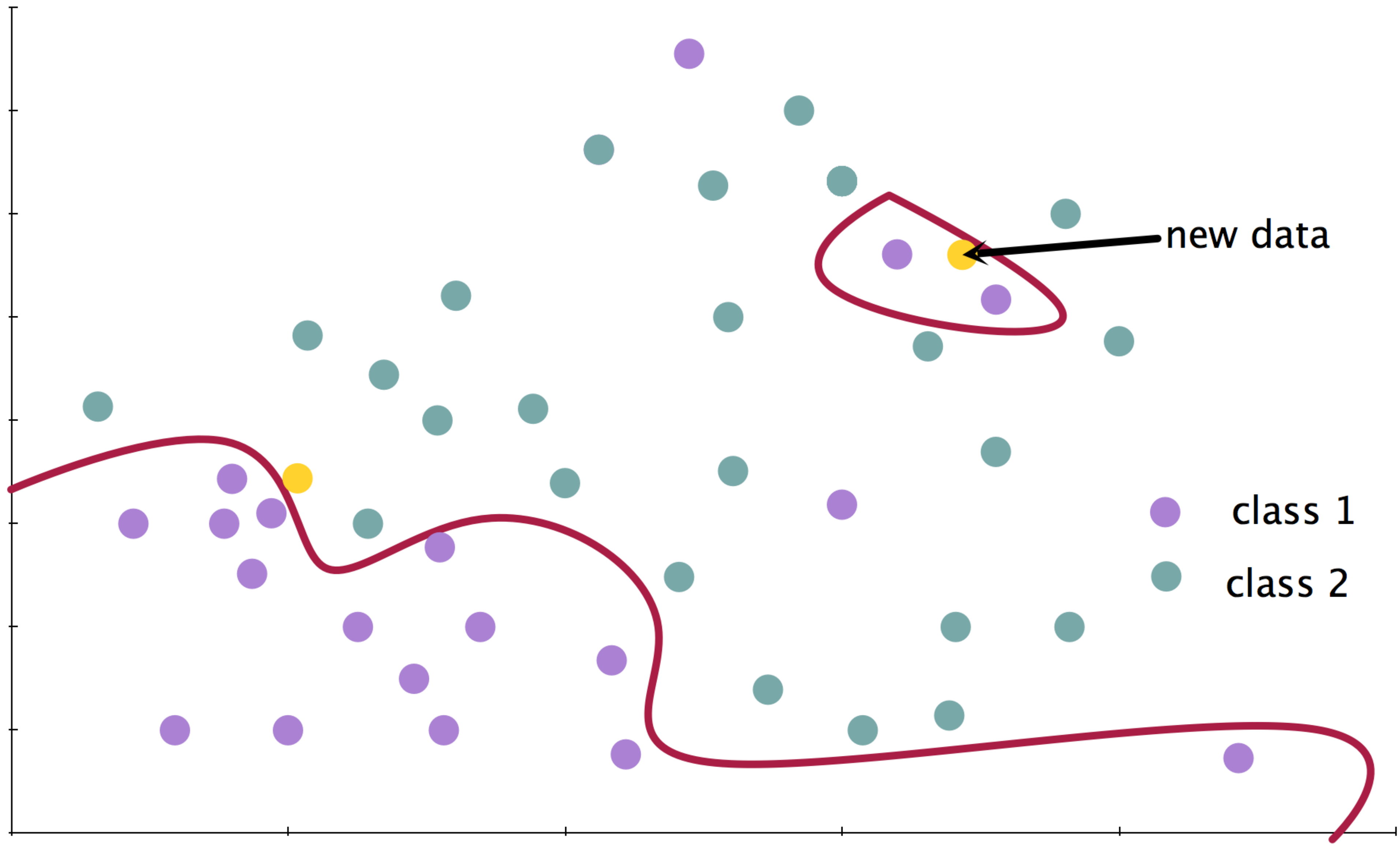
balance separation between classes against penalty for outliers

Goal of Model Fitting

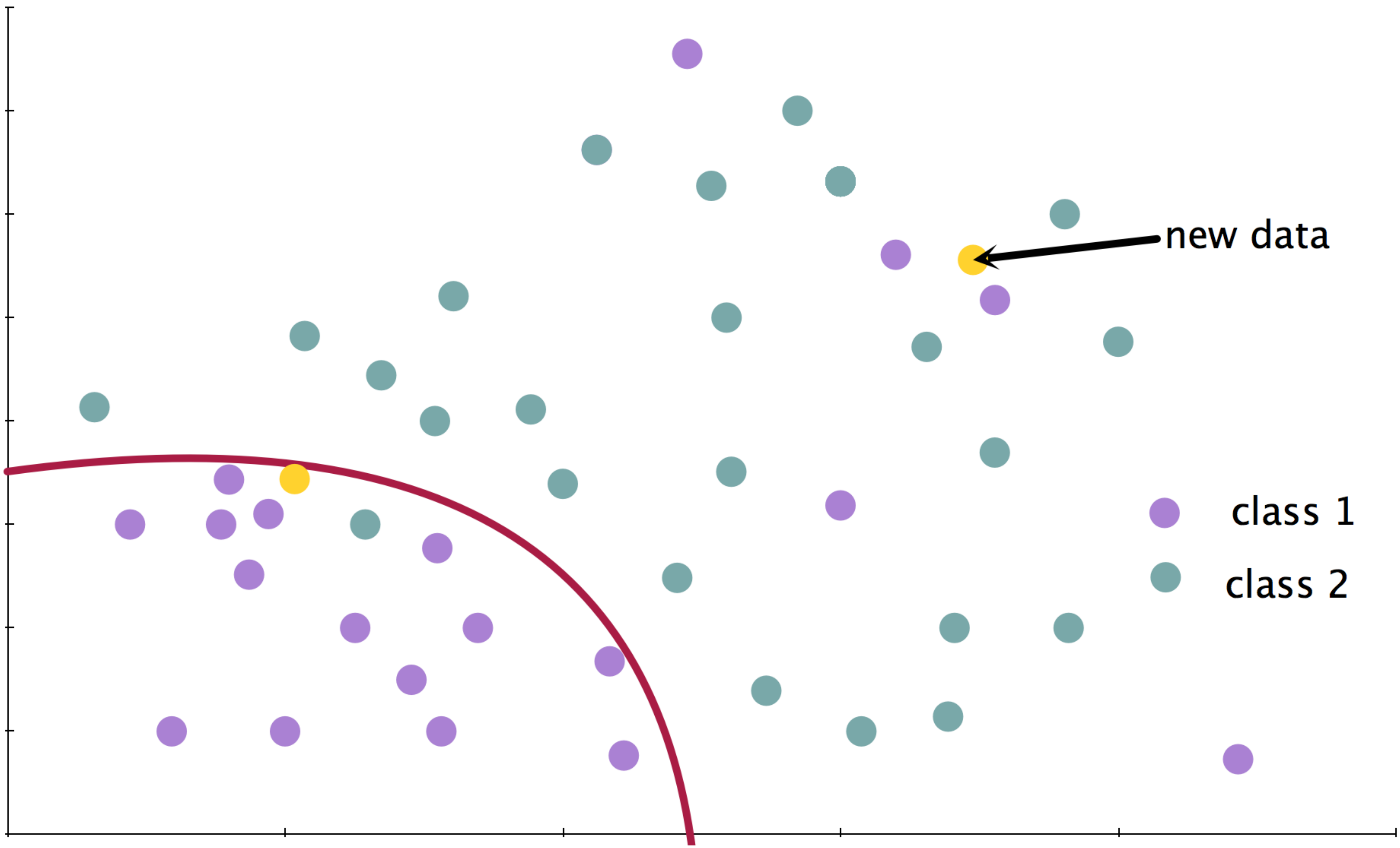
Build a model that has great accuracy on new data

- Accuracy comes from minimizing a loss function
- Typical loss function for regression: mean squared error
- Avoid overfitting!

Overfitting



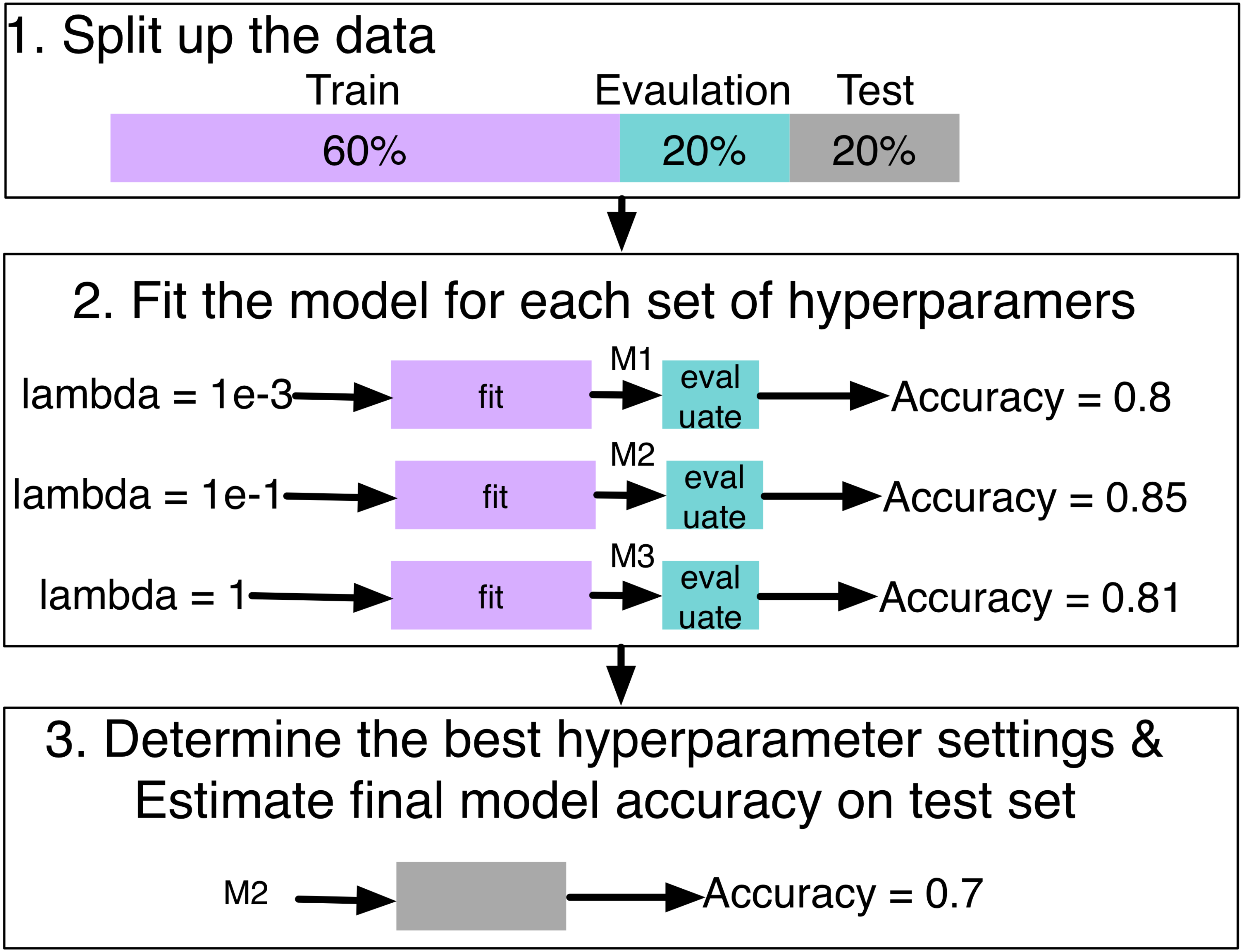
Overfitting



Model Tuning

Many models have parameters which cannot be estimated directly from the data.

These are called: hyperparameters or tuning parameters


1. Split up the data
Train
Evaulation
Test
60%
20%
20%
2. Fit the model for each set of hyperparamers
lambda = 1e-3
fit
M1
eval uate
Accuracy = 0.8
lambda = 1e-1
fit
M2
eval uate
Accuracy = 0.85
lambda = 1
fit
M3
eval uate
Accuracy = 0.81
3. Determine the best hyperparameter settings & Estimate final model accuracy on test set
M2
Accuracy = 0.7

Cross Validation

Advantages:

- usually a good estimate of model performance

Disadvantages:

- Computationally expensive for large data sets or when tuning many points

Choosing Between Models

Try many models, choose the simplest model that performs well.

Areas to explore next

- Lots more on cleaning and exploring data (EDA)
- Lots more to discuss on feature engineering
- Model fitting and evaluation: hands on, how to do this well
- How each model works, including deep learning
- Data ethics
- Models in production: testing, model decay, model maintenance
- Model Explainability
- Recommender systems and collaborative filtering
- Unsupervised learning
- A/B testing