An introduction to Machine Learning

Louis Kirsch
Former HPI Bachelor Student
Deep Learning Researcher

mail@louiskirsch.com

What is Machine Learning?

Picture of a kitten

Machine Learning

Kitten ✓

Is it a kitten?

Picture of a pig

Machine Learning

Kitten ✗

Is it a kitten?

“My grandma’s cookies are the best”

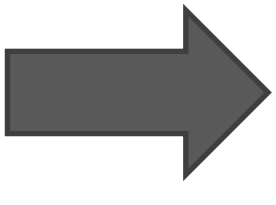
“Die Kekse meiner Oma sind die besten”

English sentence

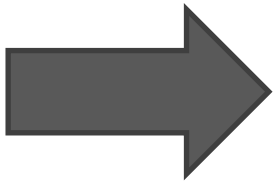
Machine Learning

German sentence

Audio wave of speech



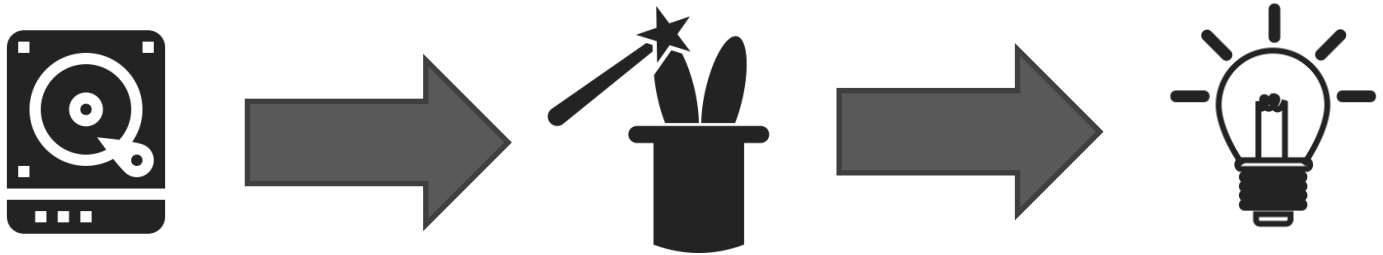
Machine Learning



“Speech to text is another application example. In recent months reportedly …”

Spoken text

What is Machine Learning?



Data

Machine Learning

Knowledge (Result)

Example: The titanic disaster

Predict for each passenger

Will he or she survive?

The titanic disaster data

Data

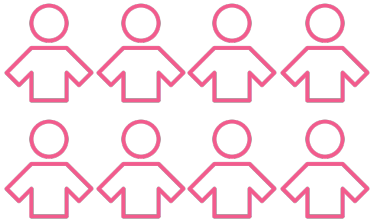
A set of N records describing each passenger

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Splitting the data in training and test set

Data

Avoiding memorization



Training set 80%
N_{train}

Test set 20%
N_{test}

A set of N records
describing each passenger

$N = N_{train} + N_{test}$

The common data format: CSV

Data

train.csv (~/Downloads) - VIM

```
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 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,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,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,27,0,2,347742,11.1333,,S
10,1,2,"Nasser, Mrs. Nicholas (Adele Achem)",female,14,1,0,237736,30.0708,,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,58,0,0,113783,26.55,C103,S
13,0,3,"Saundercock, Mr. William Henry",male,20,0,0,A/5. 2151,8.05,,S
14,0,3,"Andersson, Mr. Anders Johan",male,39,1,5,347082,31.275,,S
15,0,3,"Vestrom, Miss. Hulda Amanda Adolfina",female,14,0,0,350406,7.8542,,S
16,1,2,"Hewlett, Mrs. (Mary D Kingcome) ",female,55,0,0,248706,16,,S
17,0,3,"Rice, Master. Eugene",male,2,4,1,382652,29.125,,Q
18,1,2,"Williams, Mr. Charles Eugene",male,,0,0,244373,13,,S
19,0,3,"Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)",female,31,1,0,345763,18,,S
20,1,3,"Masselmani, Mrs. Fatima",female,,0,0,2649,7.225,,C
21,0,2,"Fynney, Mr. Joseph J",male,35,0,0,239865,26,,S
```

Samples and label

Data

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S

Samples and label

Data

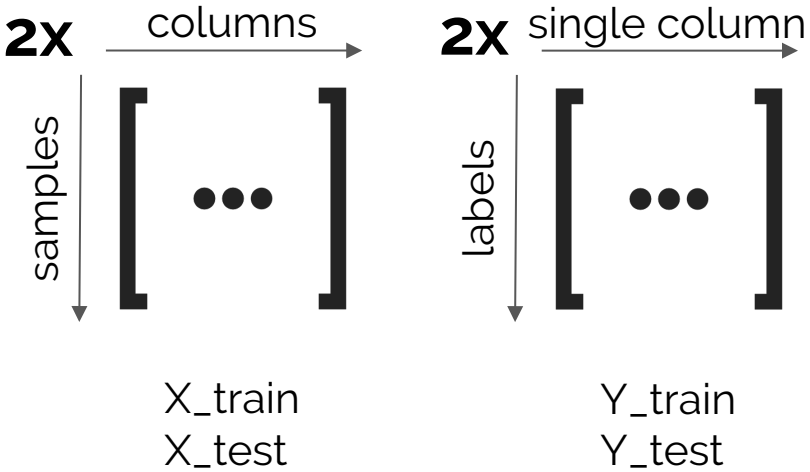
Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S

Label
Y

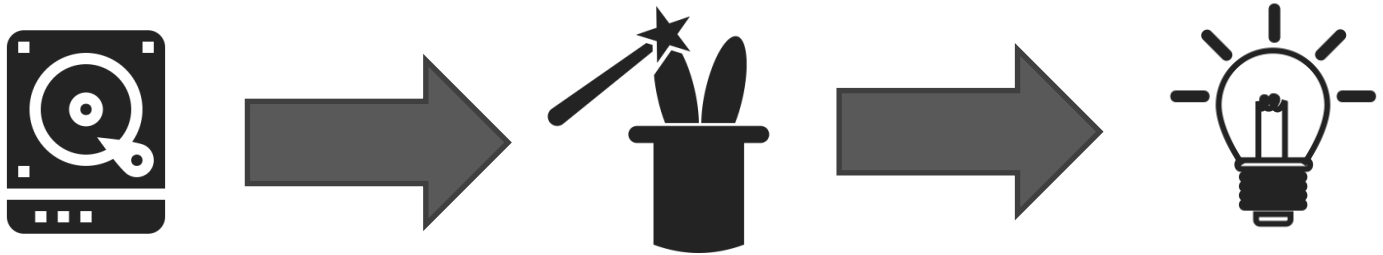
Samples
X

Load the data into 4 design matrices

Data



What is Machine Learning?

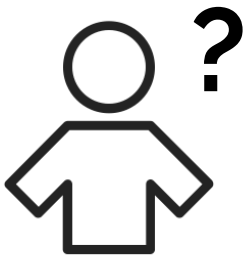


Data

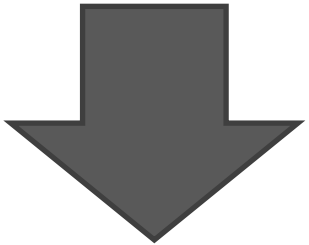
Machine Learning

Knowledge (Result)

Obtained knowledge



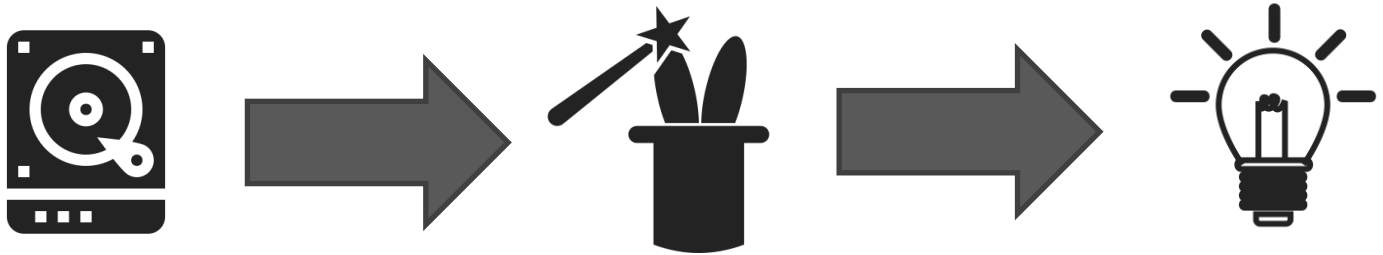
Passenger that was never seen before



Will he or she survive?

Knowledge (Result)

What is Machine Learning?



Data

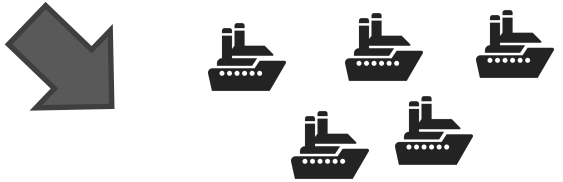
Machine Learning

Knowledge (Result)

Classification and Regression

Will he or she survive?

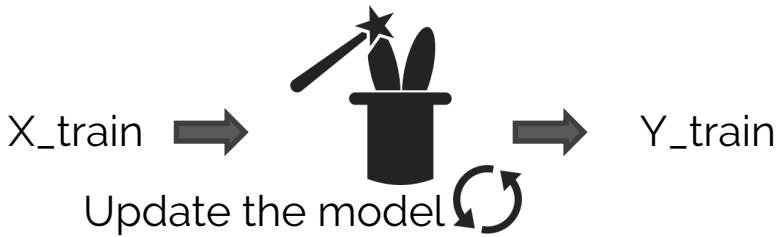
Machine Learning



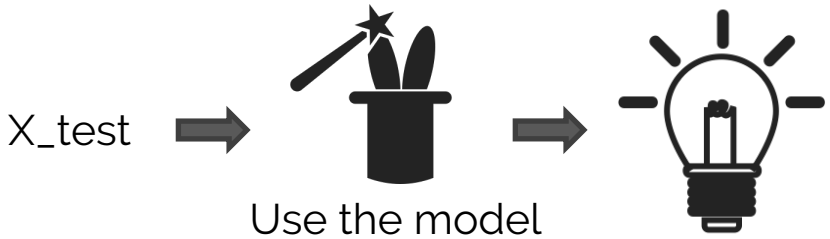
How many people will survive on a particular ship?

Training and Testing

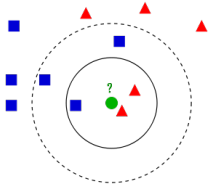
Phase 1 – Training



Phase 2 – Testing

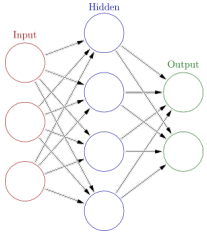


Different classifiers

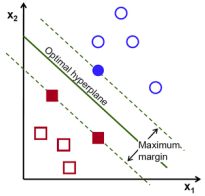


Nearest neighbors

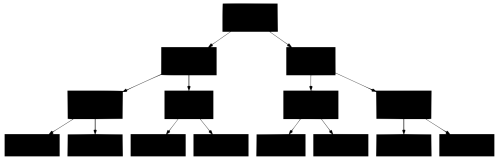
Machine Learning



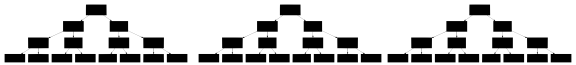
Neural network



SVM



Decision tree



Random forest

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Naïve Bayes

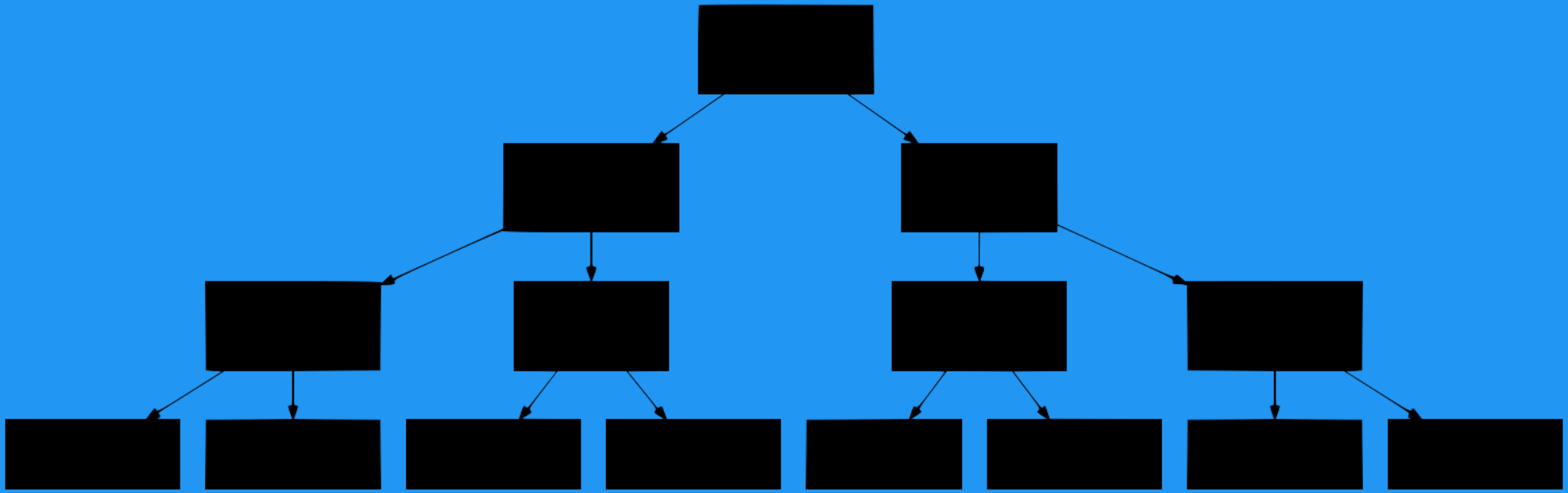
Question session

1. Ask your neighbor a hard question about **something you understood**
2. Answer your neighbor's question

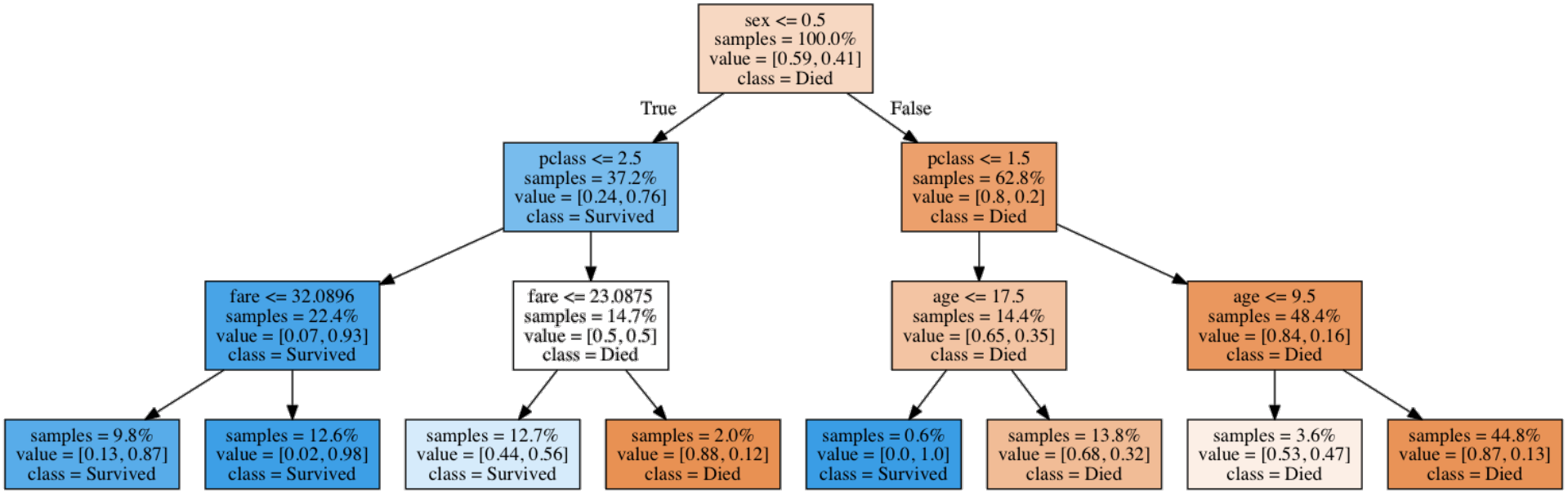
3. Ask your neighbor a question about **something you did not understand**
4. Answer your neighbor's question

5. Now **ask me** a question you **don't know the answer** to

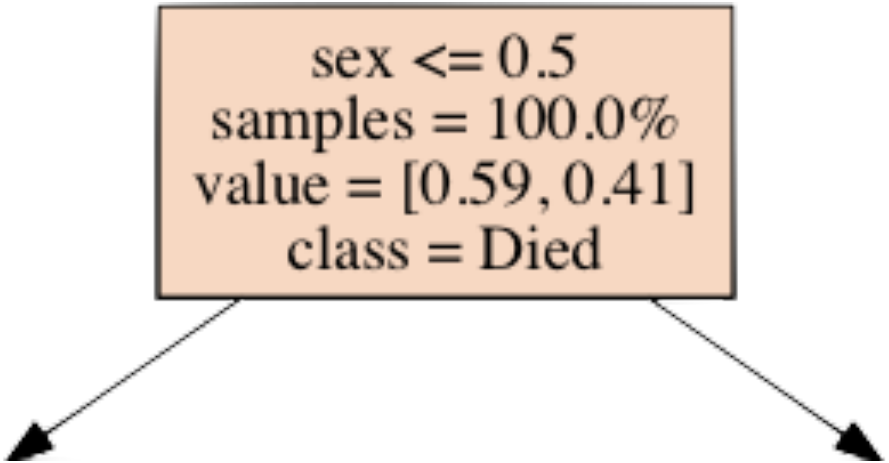
The decision tree



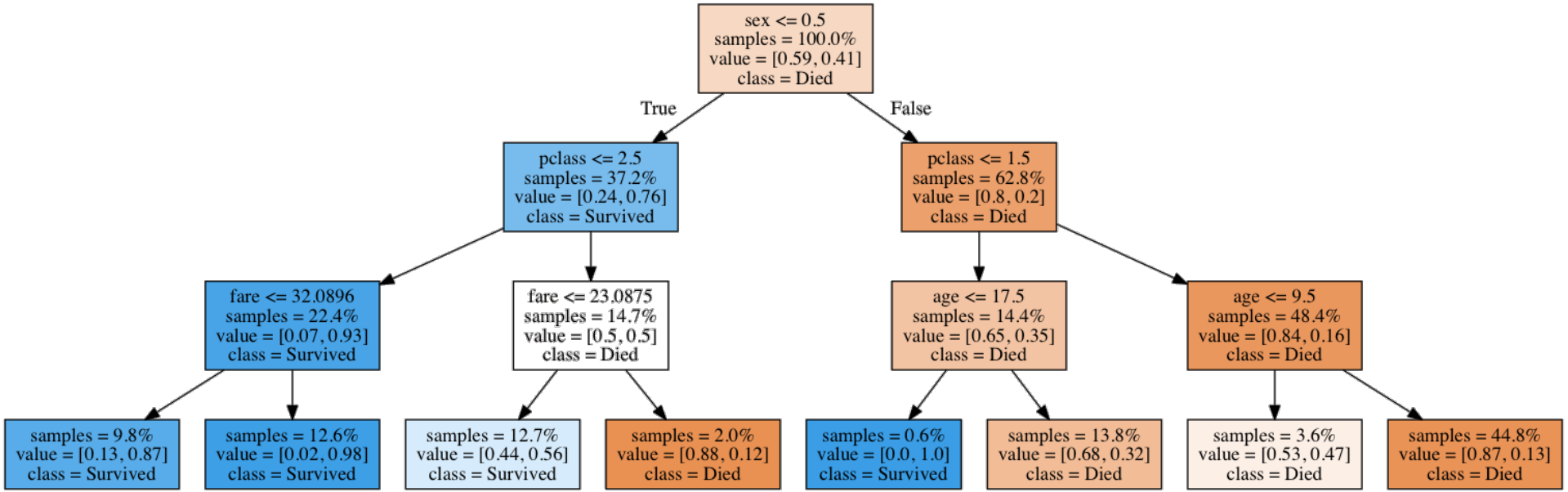
The decision tree



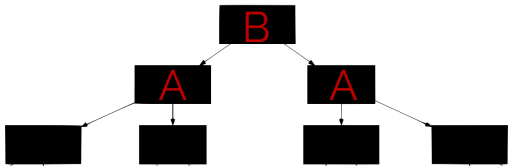
A node of a decision tree

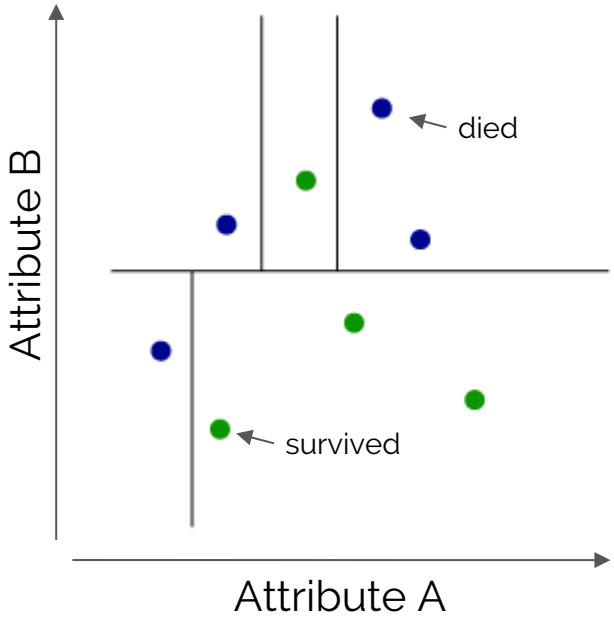


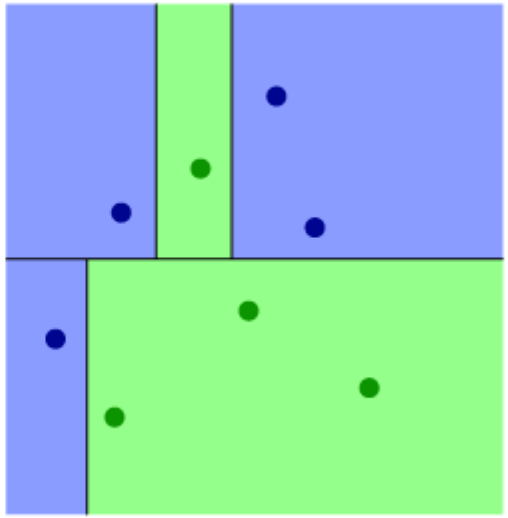
The decision tree



The decision boundary of the decision tree

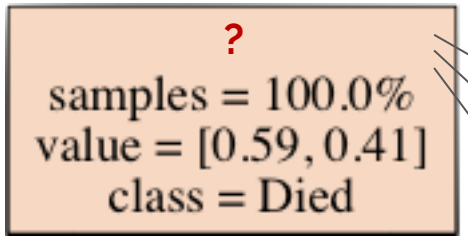






Training a decision tree

Start with root node

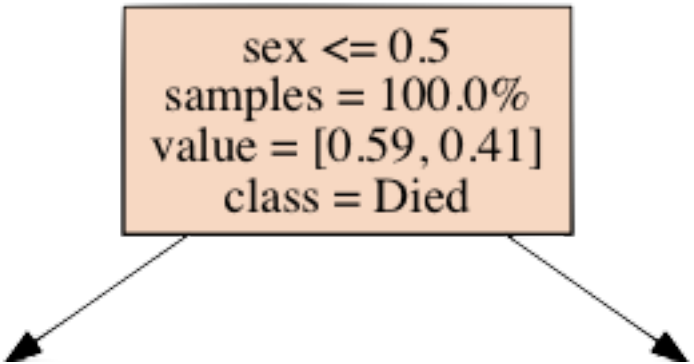


Where to split?

- sex
- age
- ...

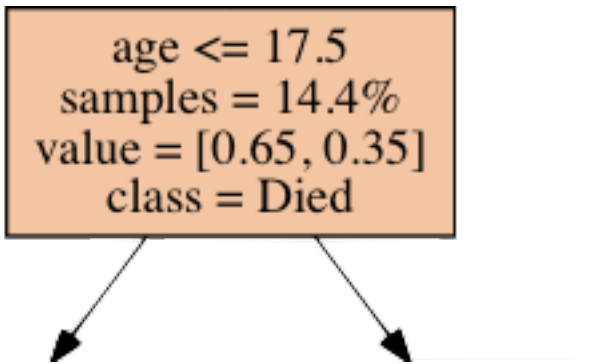
Splitting categorical and numerical attributes

Categorical



One option per category

Numerical



Sort samples by attribute age, propose split between every two values

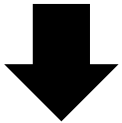
Which attribute to choose?

We now have a set S of possible splits S = {(sex, 0.5), (age, 10), (age, 12), (fare, 22), ...}
Which $s \in S$ is the best?

We employ the concept of **Information Gain**

Intuitively, how would you select the attribute and value to split?

Entropy

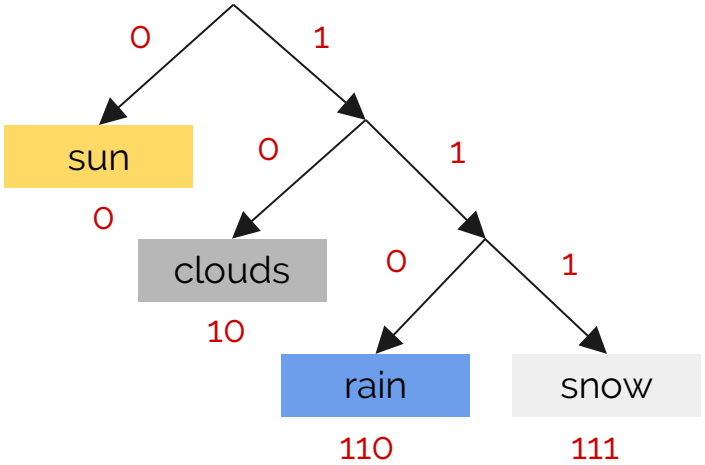


Optimal compression

00101101001101111...

~Entropy

Message x	Probability p(x)	Code length $c(x) = -\log_2 p(x)$ bits
sun	½	1 bit
clouds	¼	2 bits
rain	⅛	3 bits
snow	⅛	3 bits



Expected code length H(weather):

$= E_{x \sim weather}\, c(x) = -\Sigma_x\, p(x) \log_2 p(x)$ bits

1¾ bits = ½·1bit + ¼·2bits + ⅛·3bits + ⅛·3bits

Entropy

Entropy

used in ID3 / C4.5

Entropy

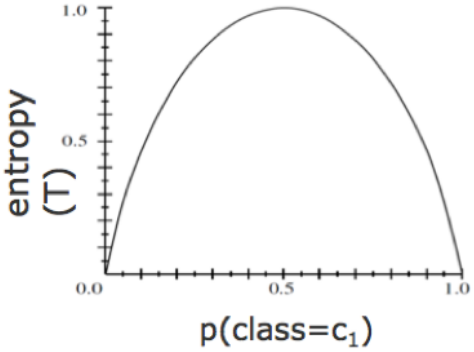
- minimum number of bits to encode a message that contains the class label of a random training object
- the entropy of a set T of training objects is defined as follows:

$$entropy(T) = -\sum_{i=1}^{k} p_i \cdot \log_2 p_i$$

for k classes c_i with frequencies p_i

- entropy(T) = 0 if p_i = 1 for any class c_i
- entropy (T) = 1 if there are k = 2 classes with p_i = ½ for each i

for two classes:

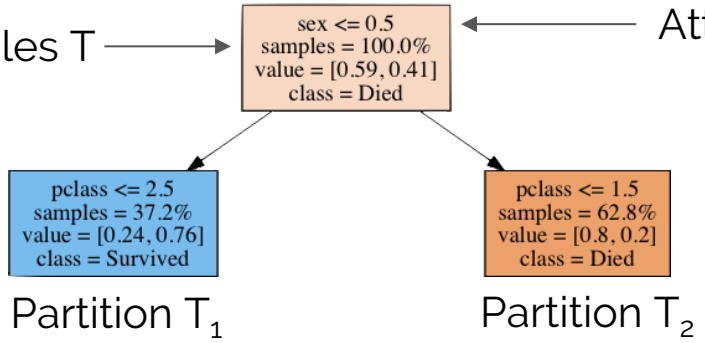


Information Gain

Let A be the attribute that induced the partitioning $T_1, T_2, \ldots, T_m$ of T. The information gain of attribute A w.r.t. T is defined as follows:

$$information\ gain(T, A) = entropy(T) - \sum_{i=1}^{m} \frac{|T_i|}{|T|} \cdot entropy(T_i)$$

Samples T



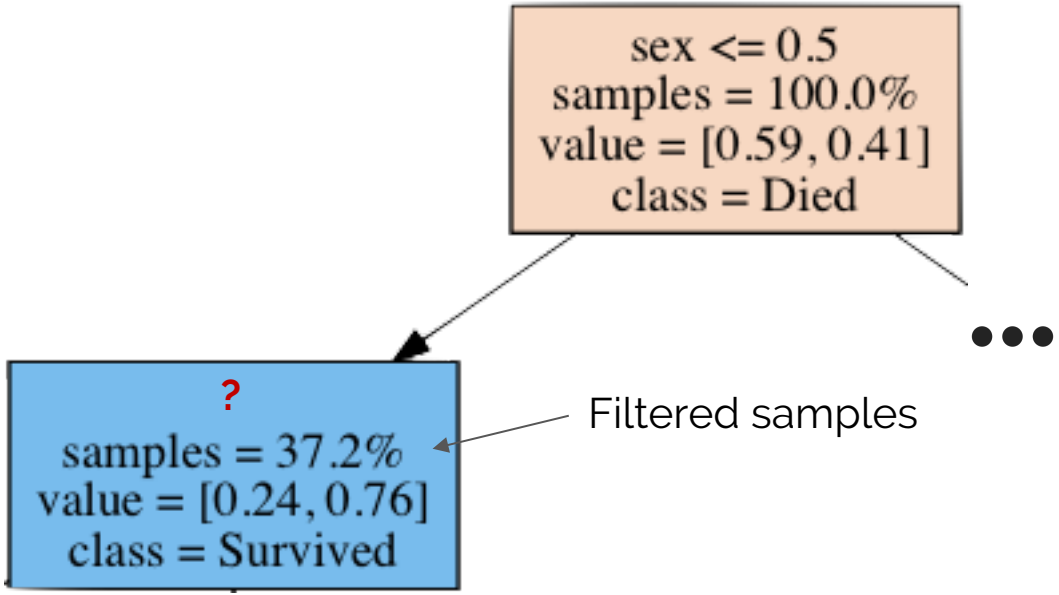
Attribute A = 'sex' split at 0.5

Partition T_1

Partition T_2

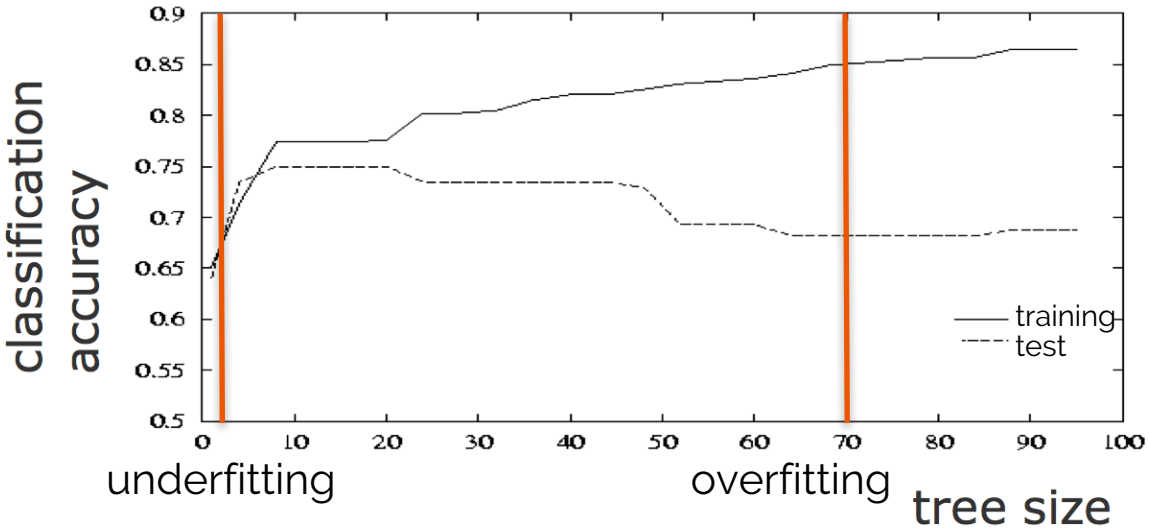
➔ We pick the attribute A that yields the highest information gain

Filter samples by split, continue recursively



Overfitting and Underfitting

How deep should the decision tree be?



Post-Pruning a decision tree

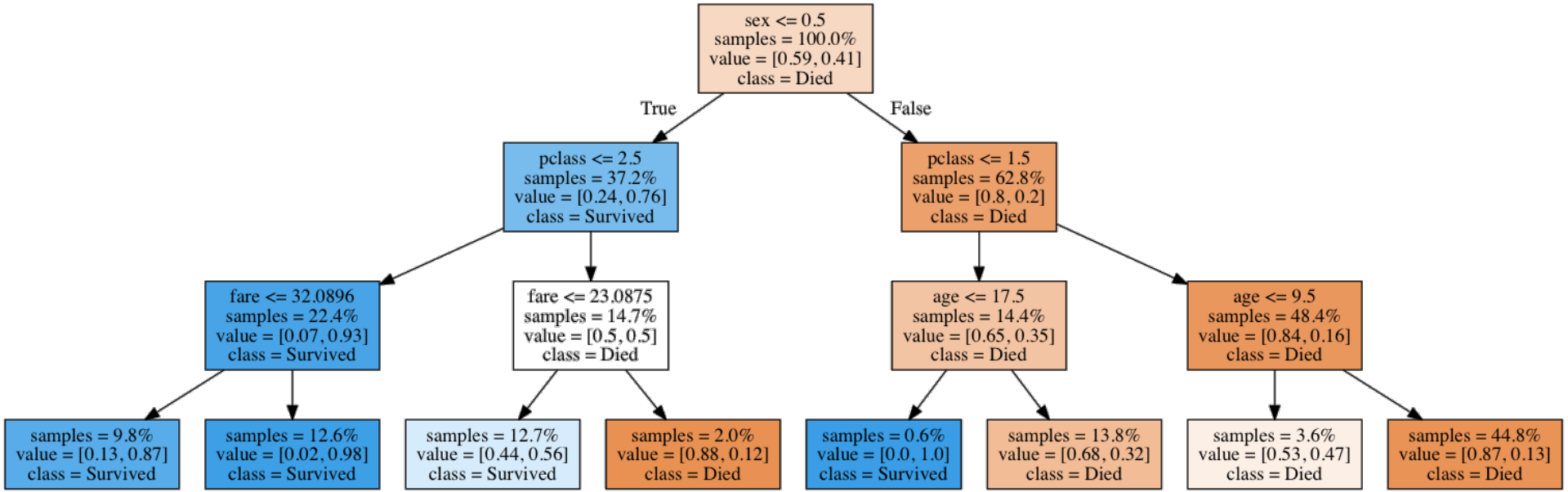
Pruning of overspecialized branches

Several algorithm options

- Reduced-Error Pruning
  - Determine all subtrees that reduce the classification error on an additional validation dataset
- Minimal Cost Complexity Pruning
  - Weighted sum of classification error and tree size

$$CC_T(E, \alpha) = F_T(E) + \alpha \cdot |E|$$

The decision tree



Question session

1. Ask your neighbor a hard question about **something you understood**
2. Answer your neighbor's question

3. Ask your neighbor a question about **something you did not understand**
4. Answer your neighbor's question

5. Now **ask me** a question you **don't know the answer** to

Are we done?

80% is understanding and preparing the data

Data exploration

Why is data exploration important

Not all attributes (features) are usable by a classifier

Missing values

Most classifiers work better with fewer, only the relevant features

Most classifiers benefit from feature engineering

Different classifiers favor different data types

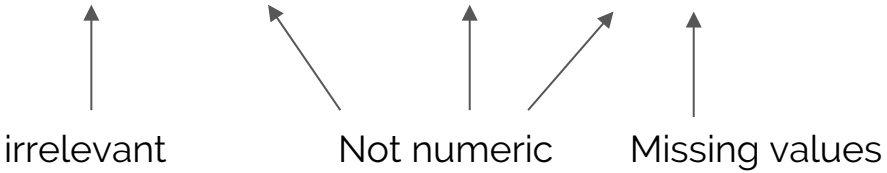
Incorrect samples

Unusable for the decision tree

```
In [33]: df[['Name', 'Sex', 'Ticket', 'Embarked', 'Cabin']].head()
```

Out[33]:

	Name	Sex	Ticket	Embarked	Cabin
0	Braund, Mr. Owen Harris	male	A/5 21171	S	NaN
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	PC 17599	C	C85
2	Heikkinen, Miss. Laina	female	STON/O2. 3101282	S	NaN
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	113803	S	C123
4	Allen, Mr. William Henry	male	373450	S	NaN

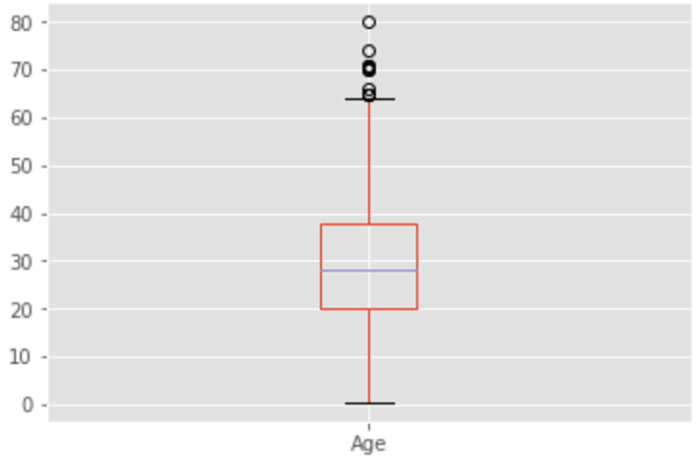


Histograms

In [29]:
```
df['Age'].plot.hist(bins=40)
plt.xlabel('Age')
plt.show()
```

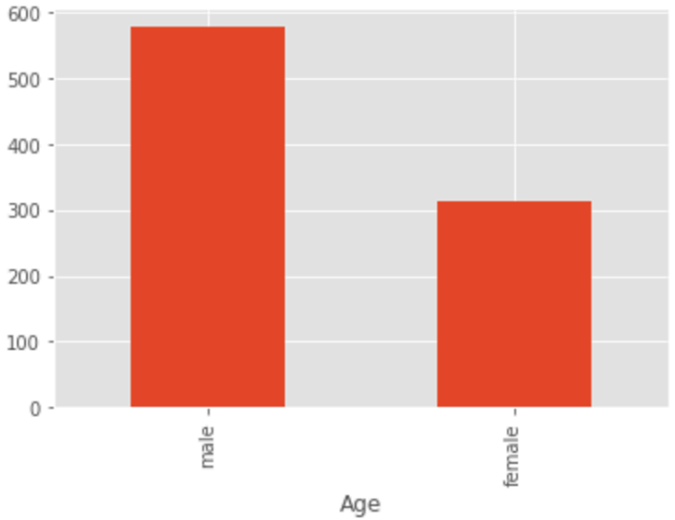
Box Plot

In [37]:
```
df['Age'].plot.box()
plt.show()
```



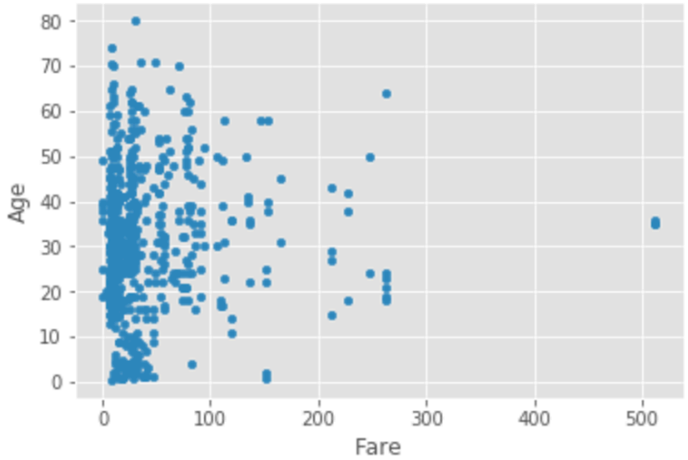
Bar plot

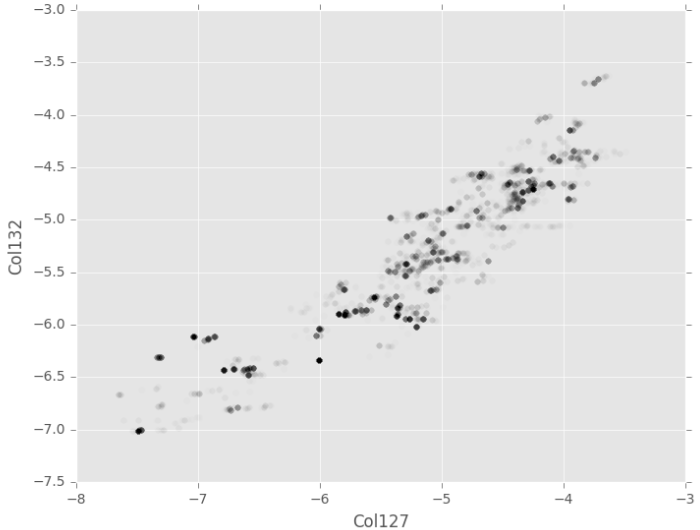
In [30]:
```
df['Sex'].value_counts().plot.bar()
plt.xlabel('Age')
plt.show()
```



Scatter plots

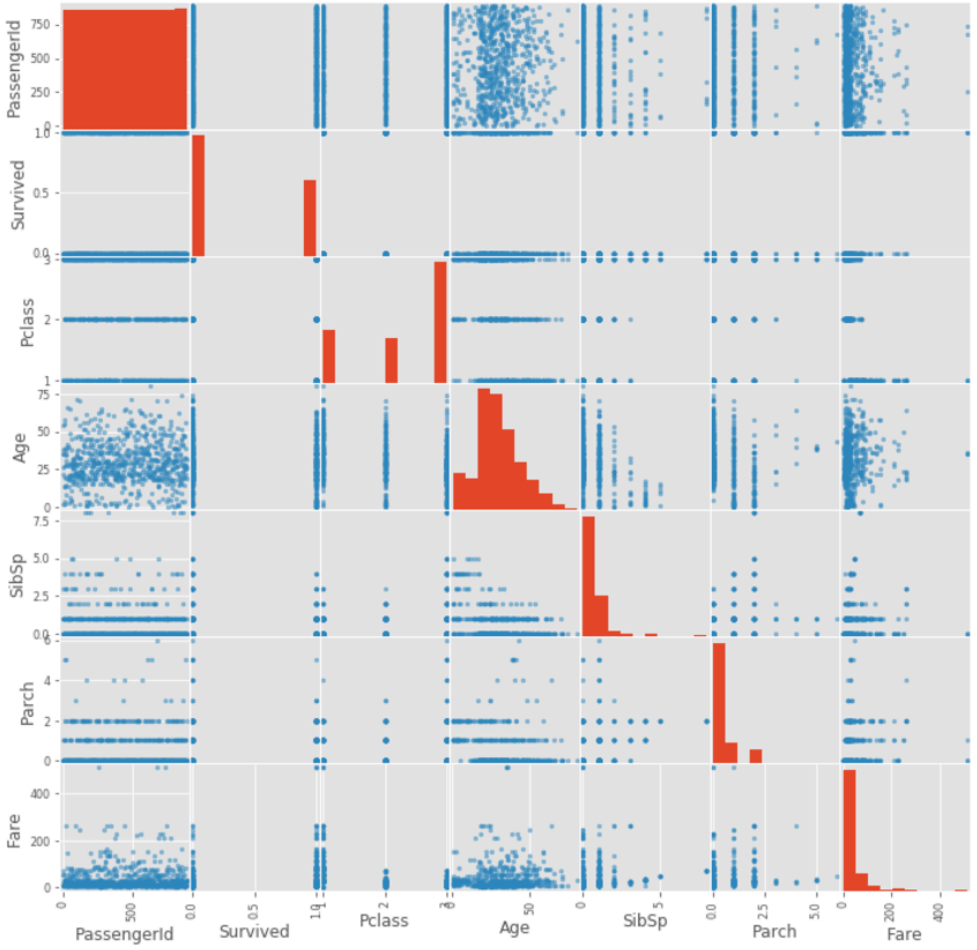
In [59]:
```
df.plot.scatter('Fare', 'Age')
plt.show()
```





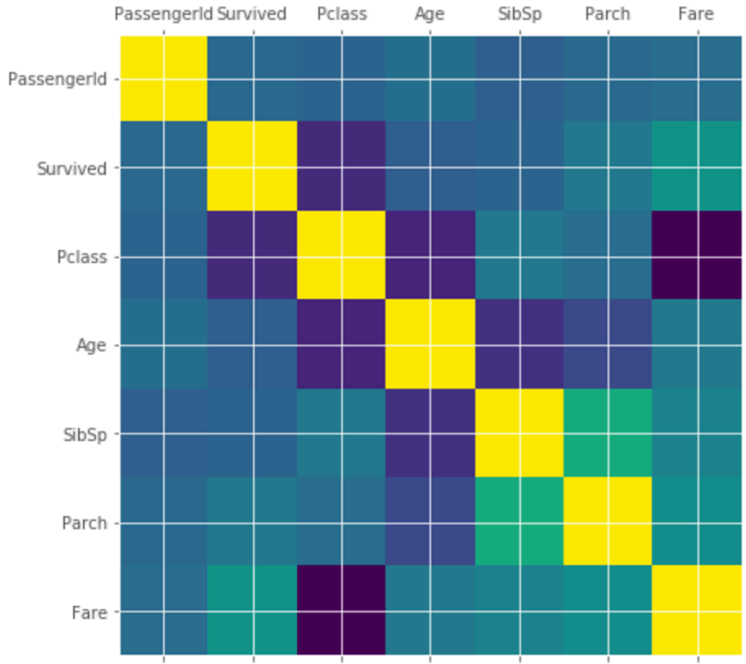
Scatter matrix

In [61]:
```
from pandas import scatter_matrix
scatter_matrix(df, figsize=(12,12))
plt.show()
```



Heatmaps

In [56]:
```
size = 7
corr = df.corr()
fig, ax = plt.subplots(figsize=(size, size))
ax.matshow(corr)
plt.xticks(range(len(corr.columns)), corr.columns)
plt.yticks(range(len(corr.columns)), corr.columns)
plt.show()
```



Question session

1. Ask your neighbor a hard question about **something you understood**
2. Answer your neighbor's question

3. Ask your neighbor a question about **something you did not understand**
4. Answer your neighbor's question

5. Now **ask me** a question you **don't know the answer** to

Data preprocessing

Preparing the data

Convert data types

Normalize data

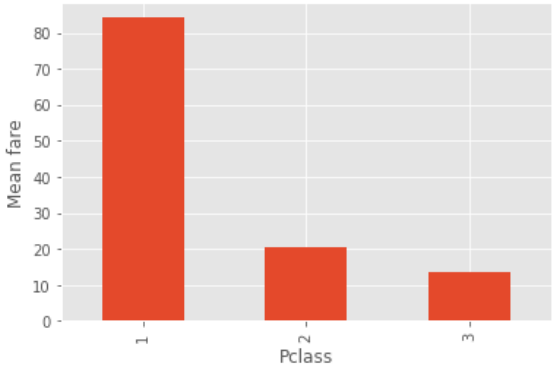
Remove features

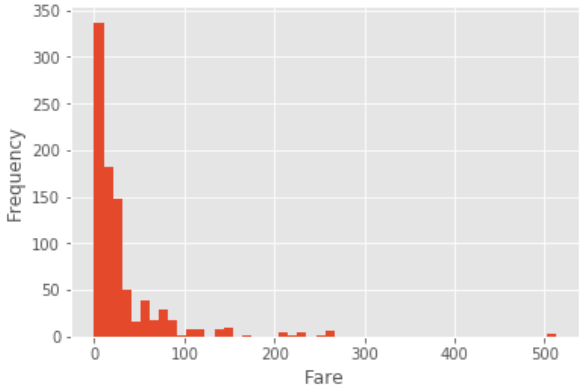
Feature engineering

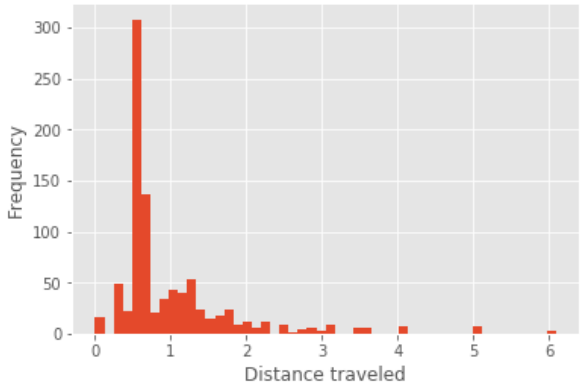
Remove rows with missing values
Or impute data

Feature engineering

In [8]:
```
mean_fare = df.groupby('Pclass', as_index=False)['Fare'].mean()
mean_fare.columns = ['Pclass', 'MeanFare']
df = df.merge(mean_fare, on='Pclass')
df['DistanceTraveled'] = df['Fare'] / df['MeanFare']
```







Feature engineering

	Pclass	Fare	MeanFare	DistanceTraveled
0	3	7.2500	13.67555	0.530143
1	3	7.9250	13.67555	0.579501
2	3	8.0500	13.67555	0.588642
3	3	8.4583	13.67555	0.618498
4	3	21.0750	13.67555	1.541071

Use scikit-learn classifiers

Scikit-Learn API

```
from sklearn import tree
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, Y)
clf.predict(X_test)
```

Import classifier class

Instantiate classifier

Fit (= training of) classifier

Predict (= testing) with classifier

Example decision tree

http://scikit-learn.org/stable/modules/tree.html

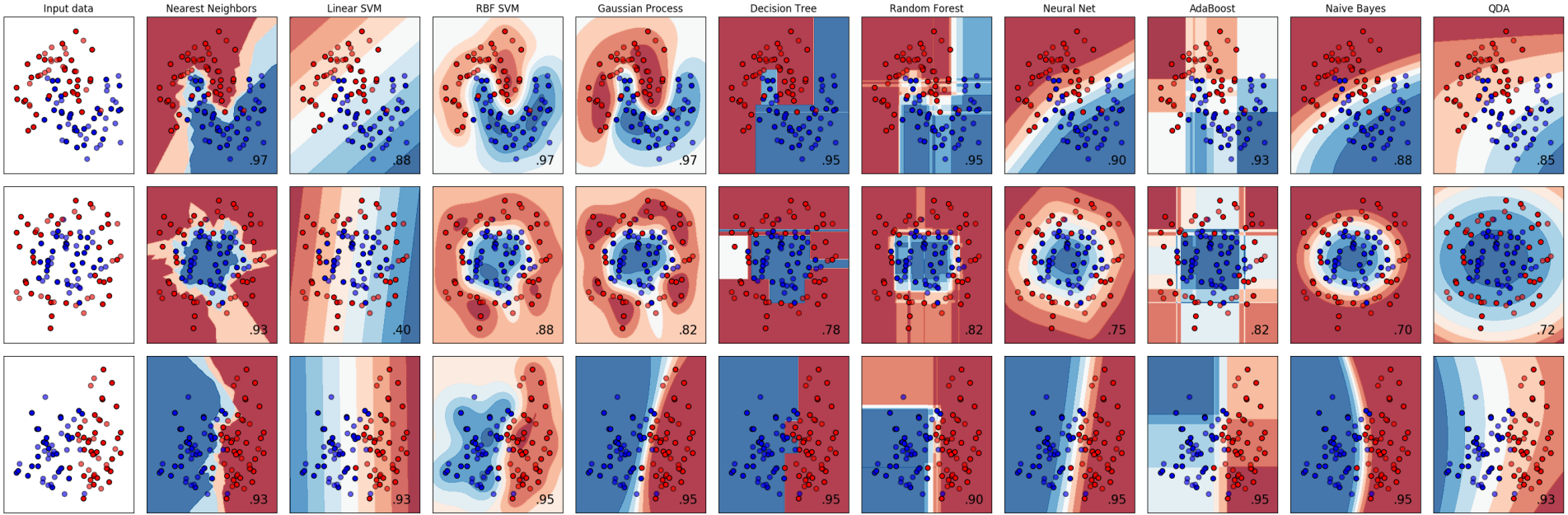
As with other classifiers, `DecisionTreeClassifier` takes as input two arrays: an array X, sparse or dense, of size `[n_samples, n_features]` holding the training samples, and an array Y of integer values, size `[n_samples]`, holding the class labels for the training samples:

```
>>> from sklearn import tree
>>> X = [[0, 0], [1, 1]]
>>> Y = [0, 1]
>>> clf = tree.DecisionTreeClassifier()
>>> clf = clf.fit(X, Y)
```

After being fitted, the model can then be used to predict the class of samples:

```
>>> clf.predict([[2., 2.]])
array([1])
```

Different decision boundaries



Final question session

Ask me only questions you **don’t know the answer** to

Live coding

- Let’s open a jupyter notebook and load the data in
- Tell me what you would like to do
  - Explore the data
  - Transform the data
  - Visualize the data
  - Run a classifier
  - Or anything else?