TECHNOLOGY FUNDAMENTALS FOR ANALYTICS

Jason Kuruzovich

Agenda

- Announcements
- What is in a model?
- Model Types
- Model Evaluation
- Titanic/Kaggle Models

Announcement

- Sorry. Office Hours conflict with PDW
- Office Hours: SA Lounge 9-11 Thursday

- Looking for help from someone who may want to gain knowledge of deployment
  - Dockerize the VM
  - Tech Background Required

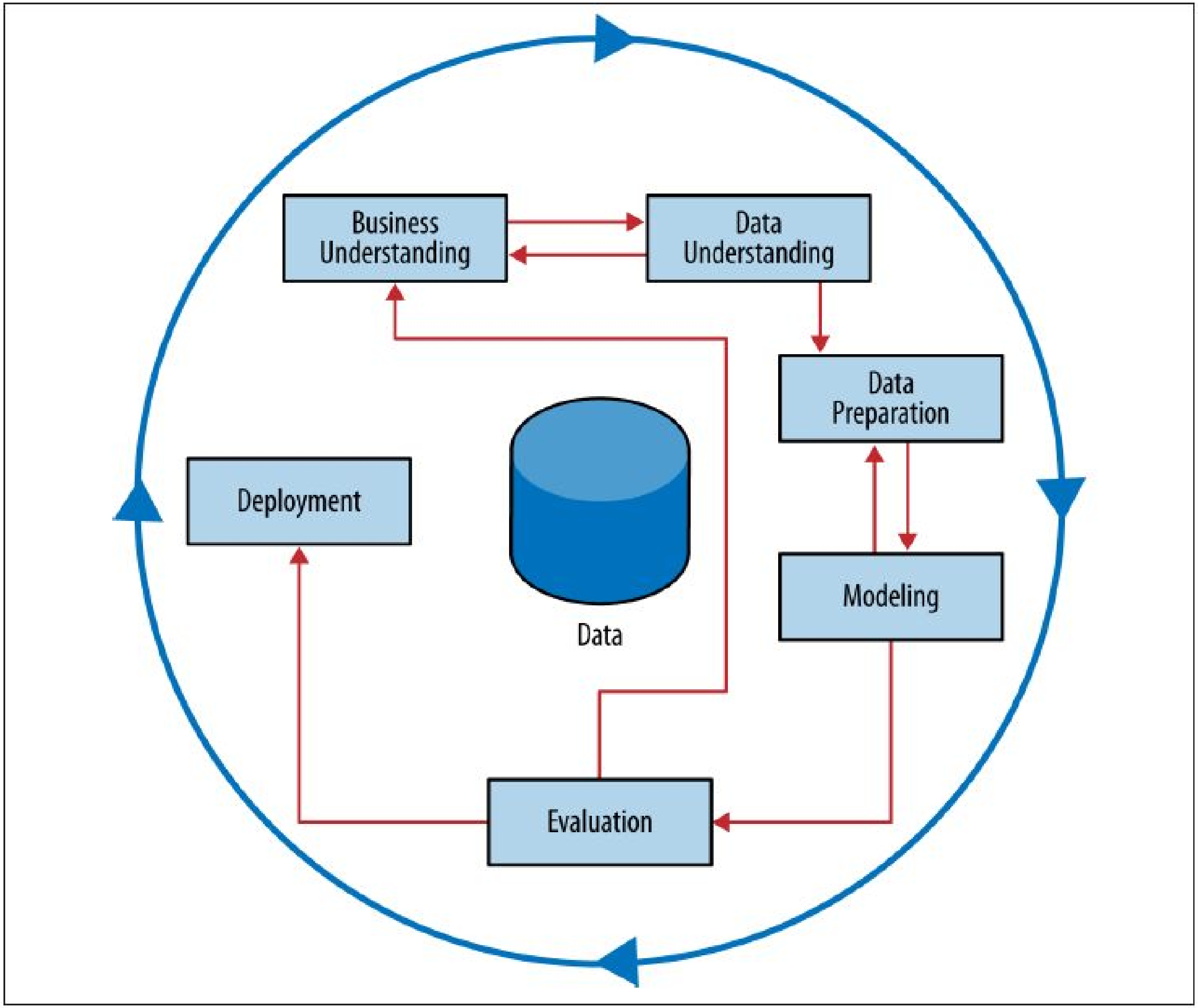


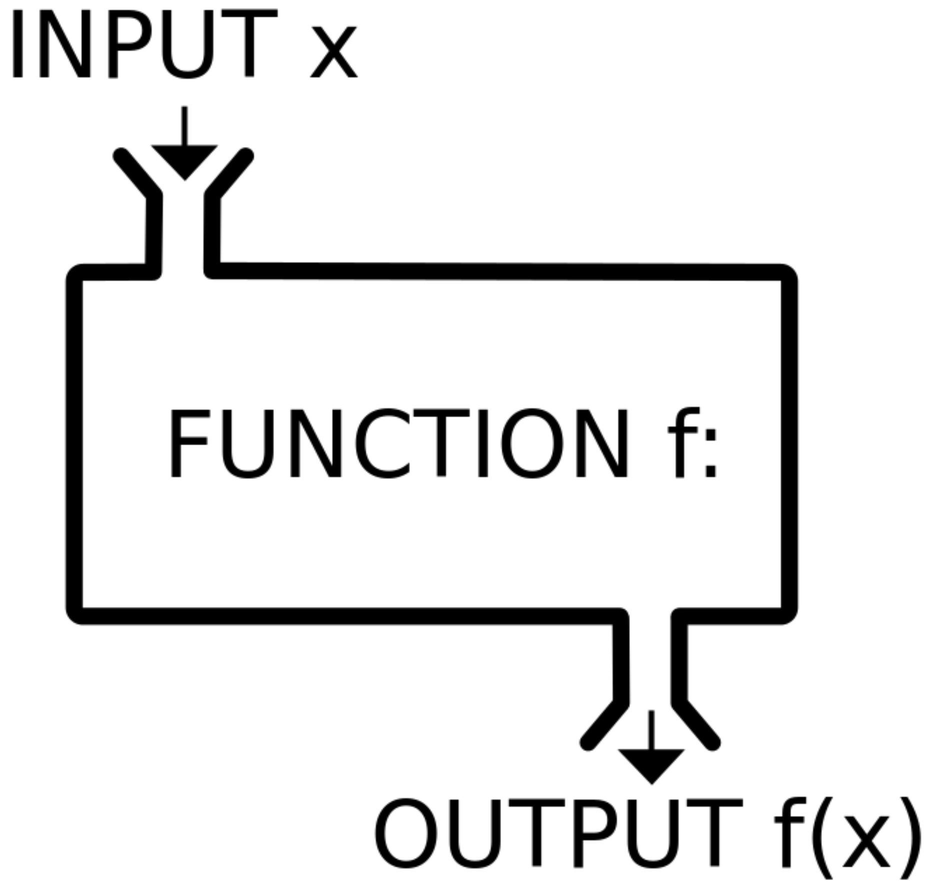
Figure 2-2. The CRISP data mining process.

What is a model?

“A mathematical model is a description of a system using mathematical concepts and language.”
-Wikipedia

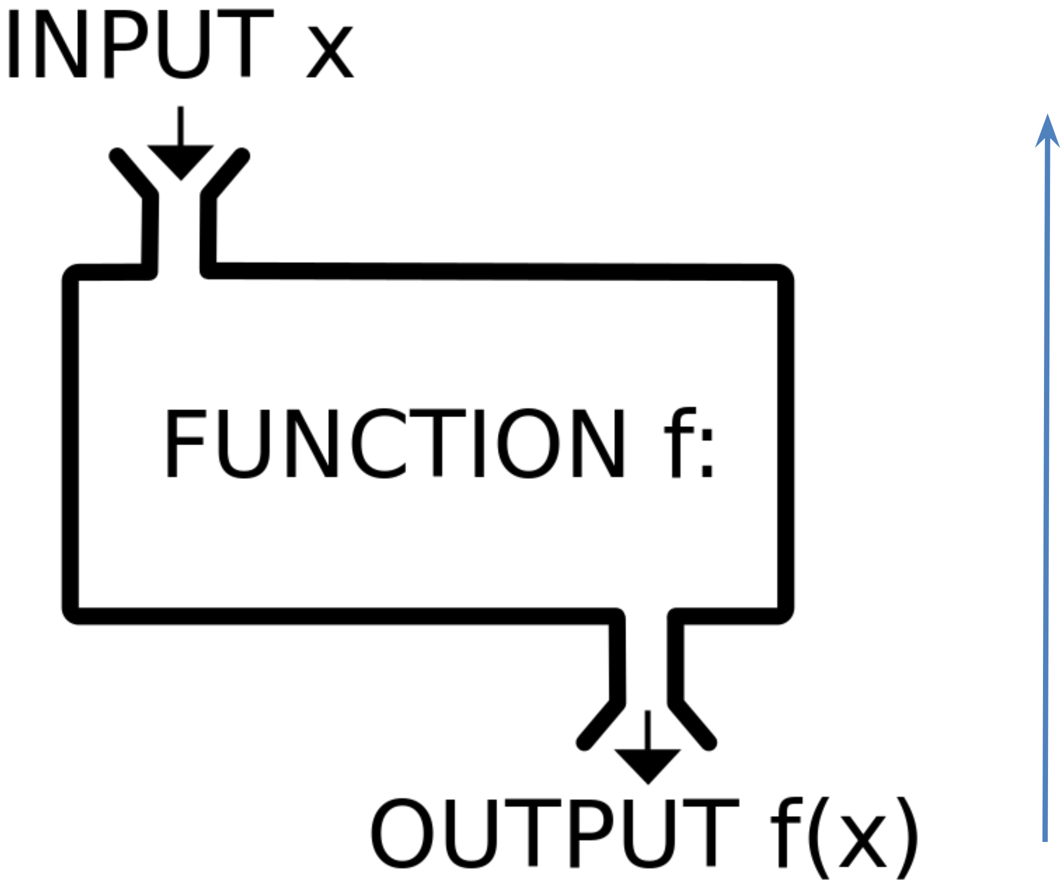
“A model is a simplified representation of reality created to serve a purpose.” - Provost & Fawcett

Independent or Explanatory Variables



Target or Dependent Variable

Independent or Explanatory Variables



INPUT x

FUNCTION f:

OUTPUT f(x)

DATA
LEAKAGE
Information
of DV moves
back to IV

Target or Dependent Variable

Models and Data Leakage

- Models for electronic commerce sales: Who is a great customer?
  - Incorporate total web page views
  - This data isn’t known until the session is over and individual has already purchased (can’t use for prediction)
  - Because they are a good customer, they have had a lot of web page views

Statistical Inference vs. Prediction

- Statistical Inference: Determine the underlying relationship for broader management issues
  - Do smaller classes lead to better student outcomes?
- Prediction: Provide a prediction of the resulting relationship
  - Which of the population of applicants is likely to be a better employee?

	Goals
Traditional Statistics	EXPLAIN the role of specific constructs
Predictive Analytics	CALCULATE an ACCURATE PREDICTION

	Variables	Model
Traditional Statistics	MEASURE VALIDATED CONSTRUCTS of interest used by OTHER RESEARCHERS	DATA REDUCTION and EASY UNDERSTAND RELATIONSHIP ANALYSIS (SEM or REGRESSION)
Predictive Analytics	INCLUDE ALL AVAILABLE DATA (with feature selection algorithems)	Complex BLACK BOX methods like NEURAL NETWORKS and SUPPORT VECTOR MACHINES

For the purposes of our discussions…
Model ~ Function ~ Algorithm
And I’ll use them interchangeably

Analytics

Model Attributes

- Supervised Learning
- Unsupervised Models

Types of Models

- Classification
- Regression
- Similarity Matching
- Clustering
- Co-Occurrence Grouping
- Profiling
- Link prediction
- Data reduction
- Causal modeling

Analytics

Model Attributes

- Supervised Learning
- Unsupervised Models

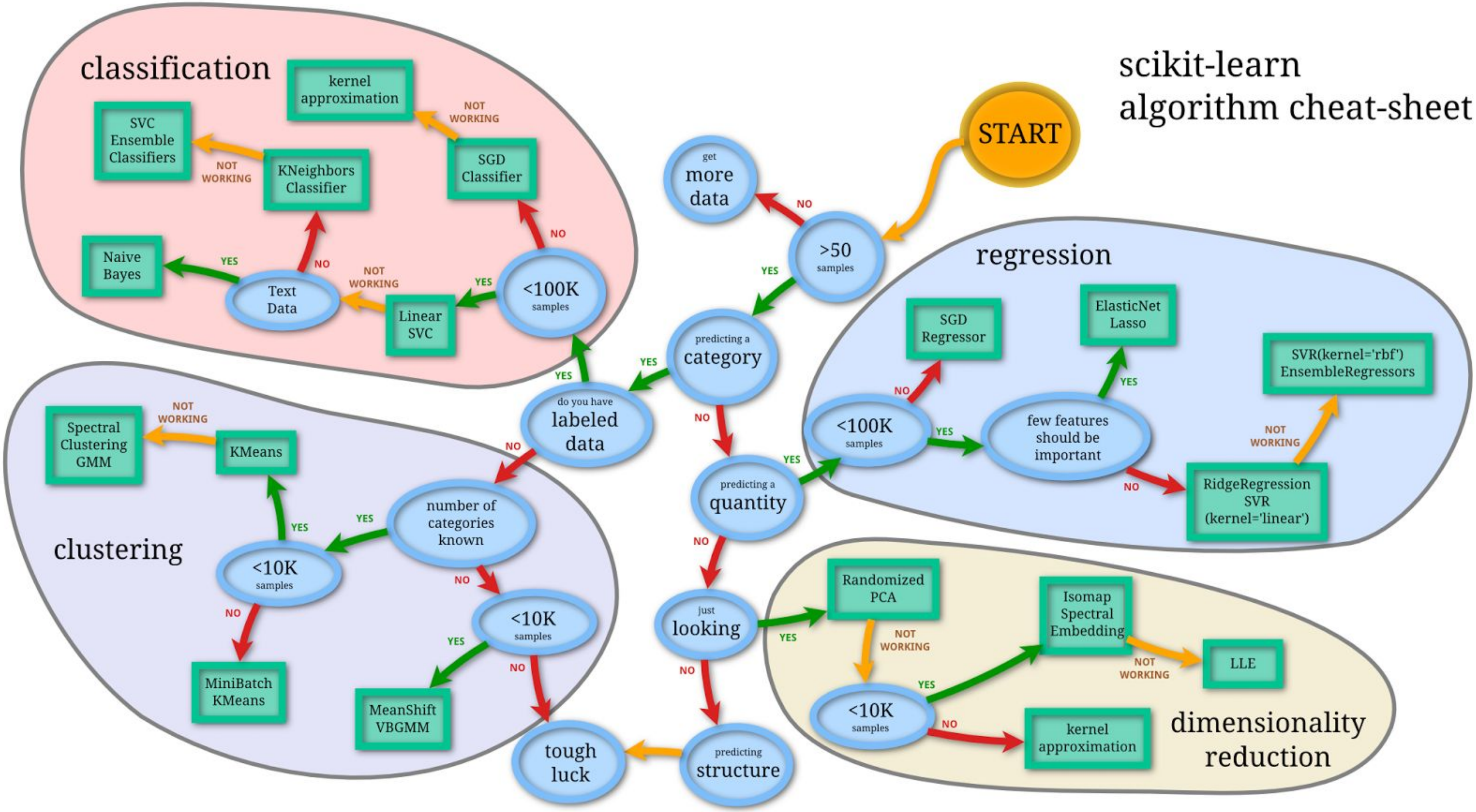
Model Types

- Classification
- Regression
- Clustering
- Similarity Matching
- Co-Occurrence Grouping
- Profiling
- Link prediction
- Data reduction
- Causal modeling

Many different algorithms for each model type

We won’t go into specific differences in this class

scikit-learn algorithm cheat-sheet



Data Science for Business

Supervised Learning

- Prediction with focused target variable
- Training data provided
- Example:
  - Most regression and classification models
  - Titanic

Unsupervised Learning

- Finding hidden structures in unlabeled data
- No target dependent variable is provided
- Example:
  - Cluster analysis
  - Can be combined with supervised learning

Kaggle Exercise

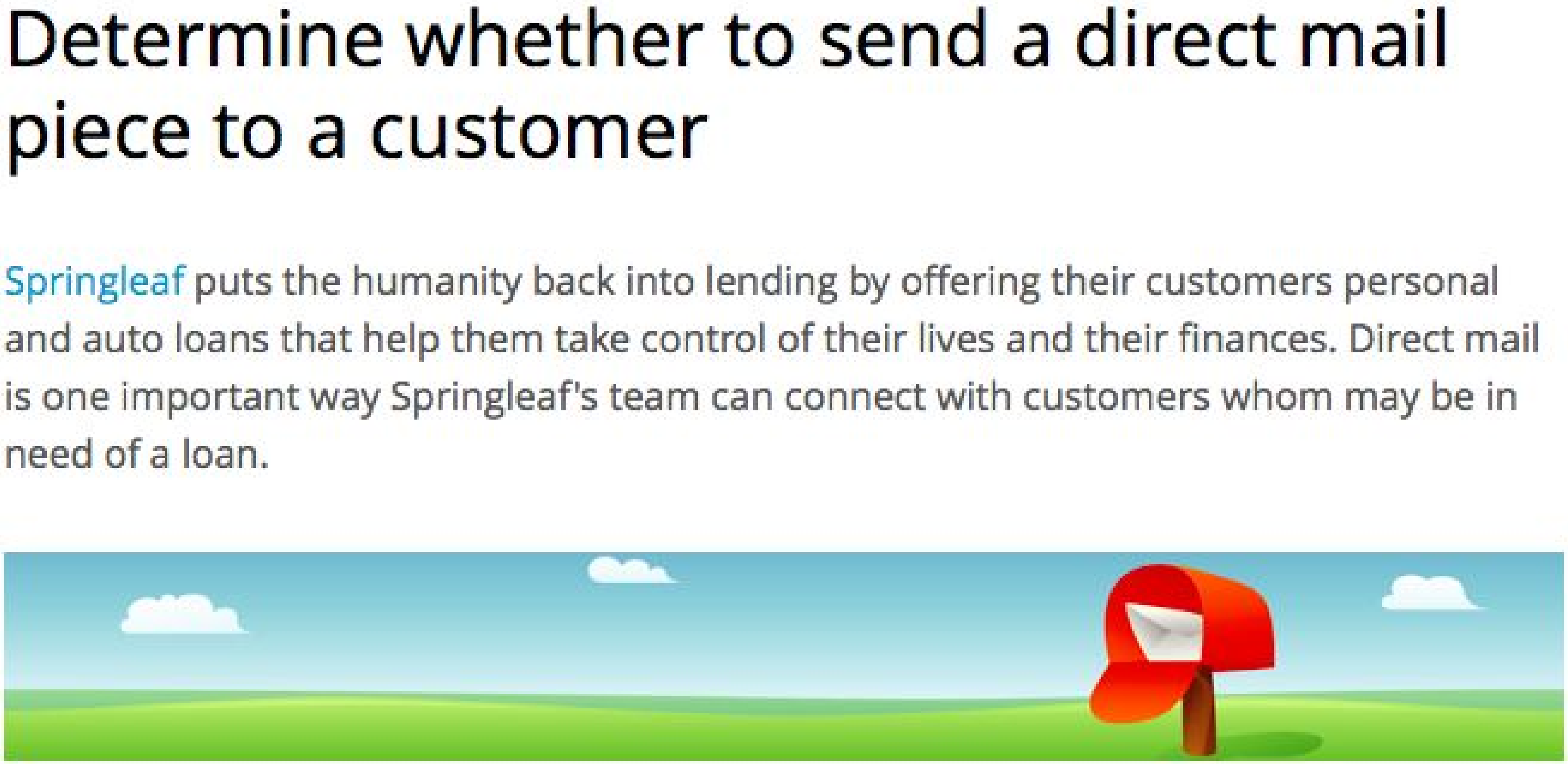
Work with someone next to you and pick 2 Kaggle competitions (don’t everyone pick same). Post a new Note (not question) to Piazza (Lab 6) with Link and type of analysis for each and why.

Classification

Types of Models: Classification

- Attempts to predict which class an individual within a population will belong
- Usually an individual must be in only on class

Types of Models: Classification



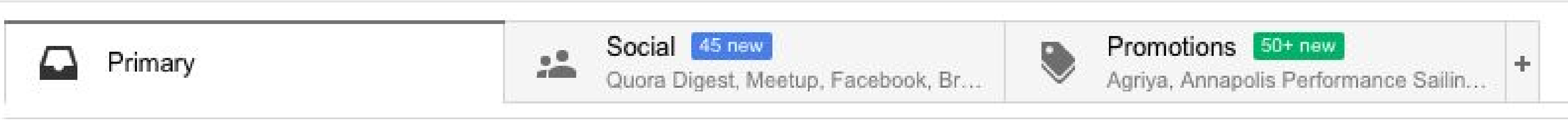
Determine whether to send a direct mail piece to a customer

Springleaf puts the humanity back into lending by offering their customers personal and auto loans that help them take control of their lives and their finances. Direct mail is one important way Springleaf's team can connect with customers whom may be in need of a loan.

https://www.kaggle.com/c/springleaf-marketing-response

Types of Models: Classification

SPAM vs Categories


Primary
Social
45 new
Quora Digest, Meetup, Facebook, Br…
Promotions
50+ new
Agriya, Annapolis Performance Sailin…
+

https://www.youtube.com/watch?v=anwy2MPT5RE

Types of Models: Classification

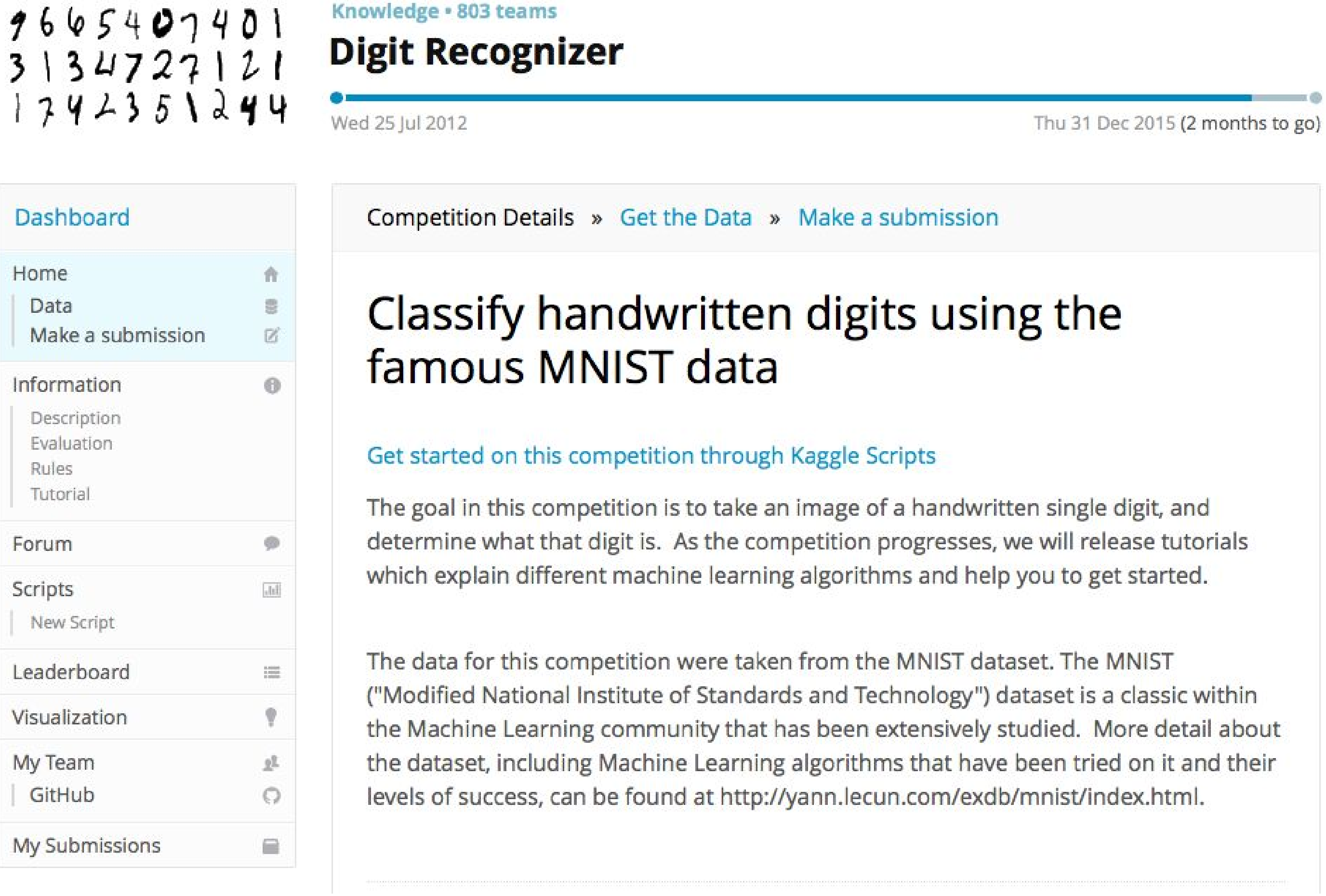
Iris setosa

Iris versicolor

Iris virginica

https://en.wikipedia.org/wiki/Iris_flower_data_set

Types of Models: Classification



Knowledge • 803 teams

Digit Recognizer

Wed 25 Jul 2012

Thu 31 Dec 2015 (2 months to go)

Classify handwritten digits using the famous MNIST data

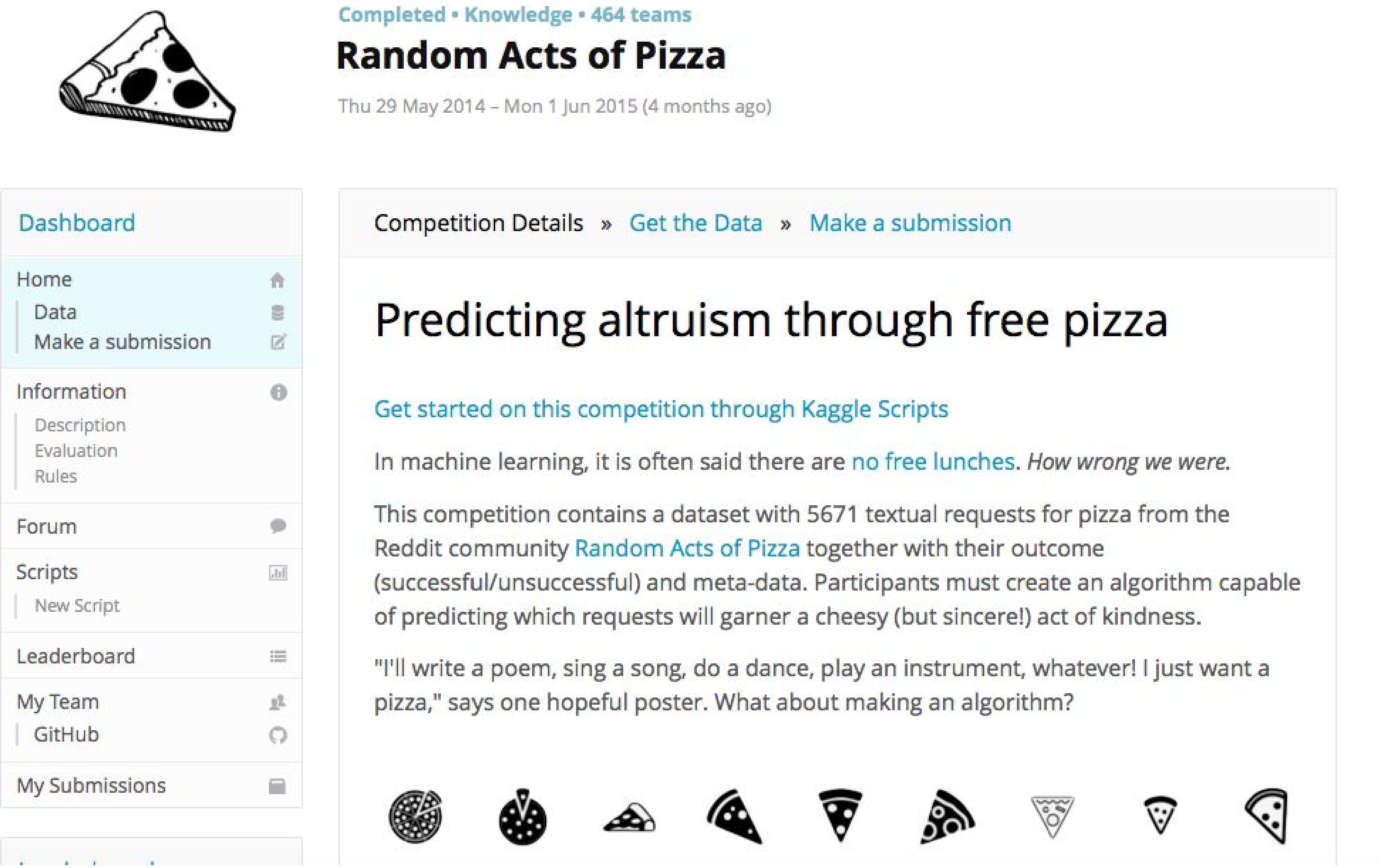
Get started on this competition through Kaggle Scripts

The goal in this competition is to take an image of a handwritten single digit, and determine what that digit is. As the competition progresses, we will release tutorials which explain different machine learning algorithms and help you to get started.

The data for this competition were taken from the MNIST dataset. The MNIST ("Modified National Institute of Standards and Technology") dataset is a classic within the Machine Learning community that has been extensively studied. More detail about the dataset, including Machine Learning algorithms that have been tried on it and their levels of success, can be found at http://yann.lecun.com/exdb/mnist/index.html.

https://www.kaggle.com/c/digit-recognizer

Types of Models: Classification



Completed • Knowledge • 464 teams

Random Acts of Pizza

Thu 29 May 2014 – Mon 1 Jun 2015 (4 months ago)

Predicting altruism through free pizza

Get started on this competition through Kaggle Scripts

In machine learning, it is often said there are no free lunches. *How wrong we were.*

This competition contains a dataset with 5671 textual requests for pizza from the Reddit community Random Acts of Pizza together with their outcome (successful/unsuccessful) and meta-data. Participants must create an algorithm capable of predicting which requests will garner a cheesy (but sincere!) act of kindness.

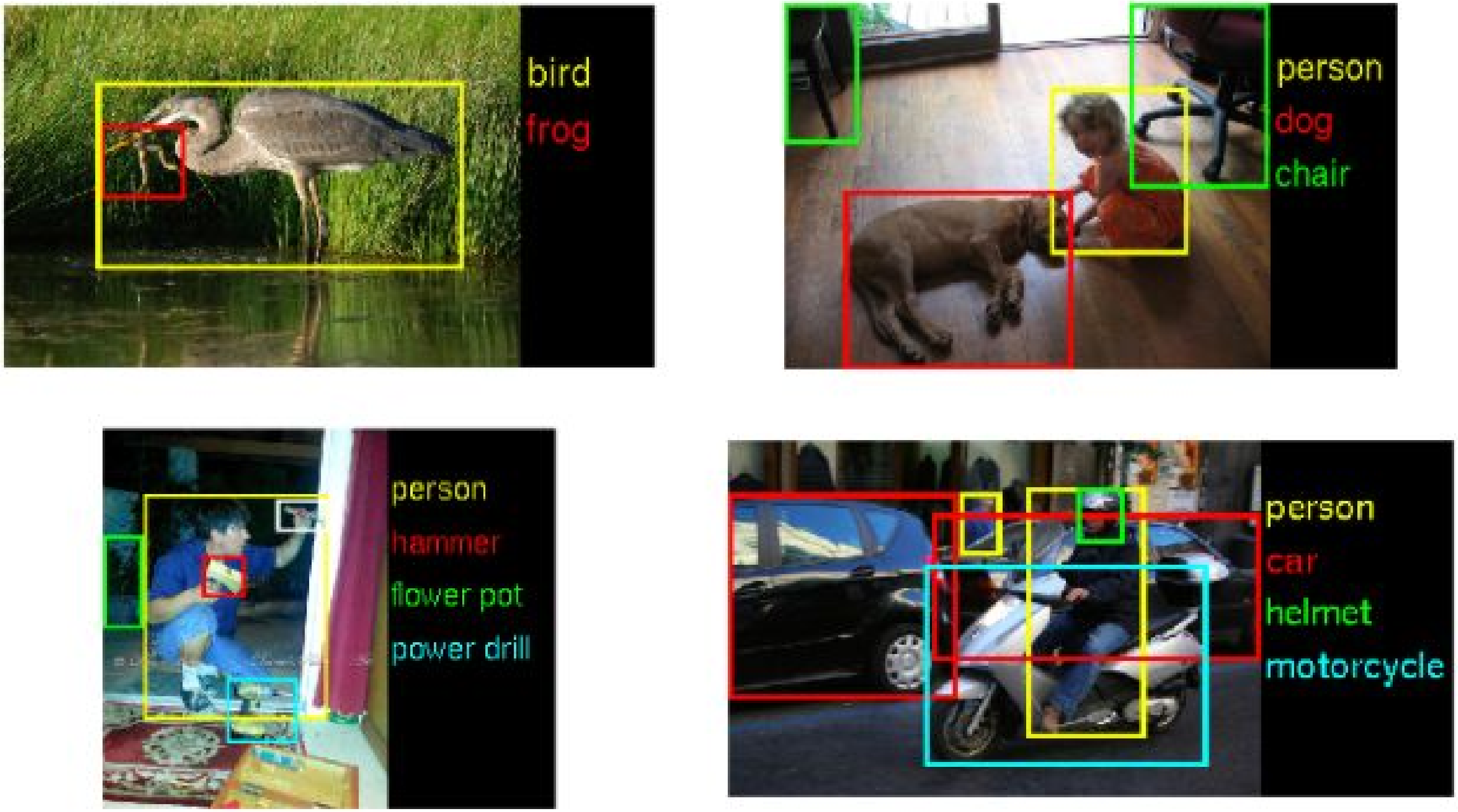
"I'll write a poem, sing a song, do a dance, play an instrument, whatever! I just want a pizza," says one hopeful poster. What about making an algorithm?

https://www.kaggle.com/c/random-acts-of-pizza

What if the category is anything and the data is real and visual?

Types of Models: Classification

Example ILSVRC2014 images:



http://image-net.org/challenges/LSVRC/2014/index

Types of Models: Classification

http://research.google.com/archive/unsupervised_icml2012.html

Types of Models: Classification

Companies

- https://www.metamind.io/vision/general
- http://www.dataversity.net/apple-buys-machine-learning-company-perceptio/
- http://www.medaware.com

In the Titanic example, what is the most simple model possible?

survived = 0

>

survived = 1

Evaluating Classification

- **Naïve rule:** classify all of the records as belonging to the most prevalent class
  - Often used as a benchmark

Evaluating Classification

CONFUSION MATRIX

	Predicted Class		
		True	False
Actual Class	True	True positive (tp)	False Negative (fn)
	False	False Positive (fp)	True Negative (tn)

Evaluating Classification

Accuracy =
(True Negative + True Positive)/Population

Accuracy is used for Titanic:
(True Survived + True Died)/Population

Would accuracy be a good metric for things like fraud?

Evaluating Classification

- When goal is to identify rare outcomes, best model may have lower accuracy

- Must ask, what is the value of a false positive, false negative, true positive, true negative

Outcomes like survival (Titanic) or differ in the level of entropy

Entropy

“In information theory, entropy (more specifically, Shannon entropy) is the expected value (average) of the information contained in each message received. 'Messages' don't have to be text; in this context a 'message' is simply any flow of information.”

Entroypy = -p1(log(p1)) – p2(log(p2)) -…

Entropy

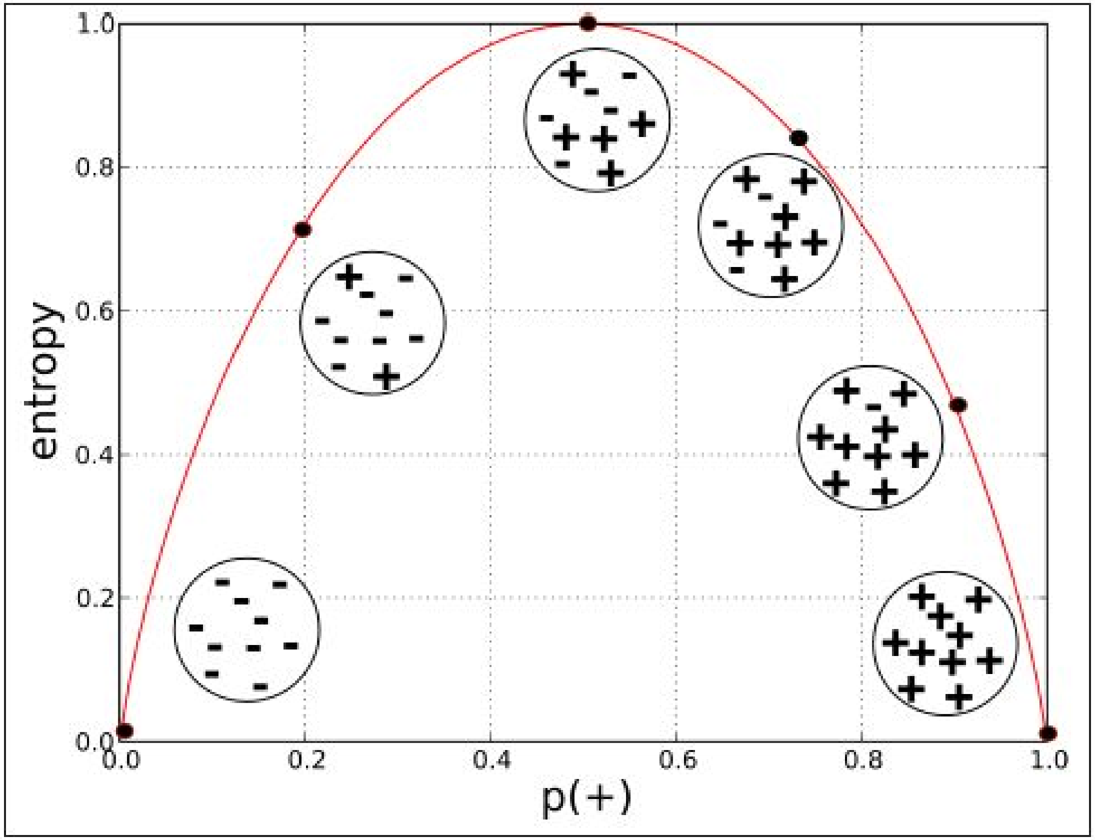


Figure 3-3. Entropy of a two-class set as a function of p(+).

Types of Models: Regression

Regression examines relationships among variables, predicting a continuous dependent variable

Example:

Weight = f(height, age, genes, eating, etc.)

[More details in different class]

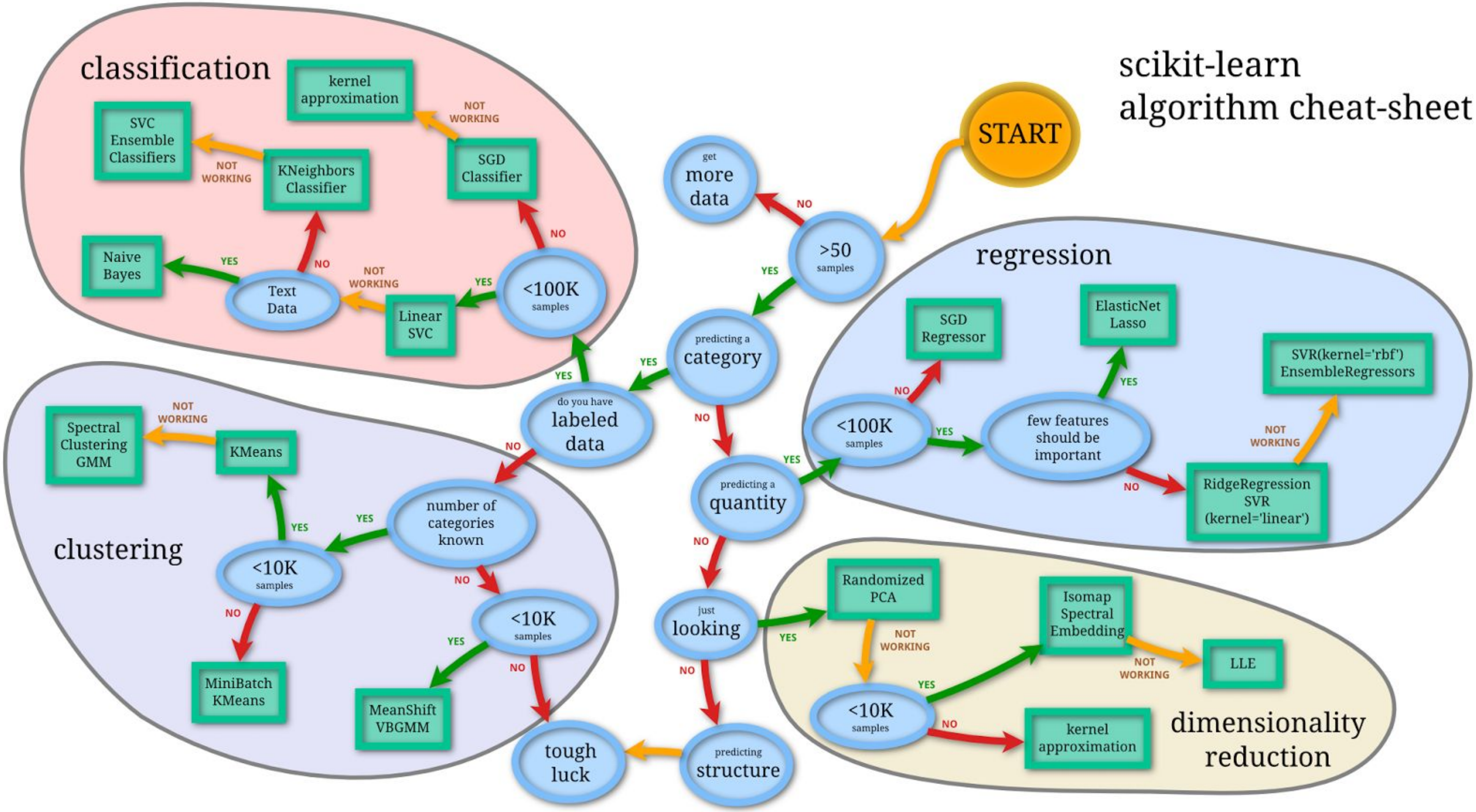
Classify individuals on whether they are likely to be survivors of the Titanic disaster

The model follows from the questions you want to answer

Often questions can be answered in different ways with different models

So if you can actually add VALUE in PREDICTION, customers are likely to be VERY HAPPY

scikit-learn algorithm cheat-sheet



START

>50 samples

NO

get more data

YES

predicting a category

YES

do you have labeled data

YES

classification

<100K samples

YES

Linear SVC

NOT WORKING

Text Data

YES

Naive Bayes

NO

KNeighbors Classifier

NOT WORKING

SVC Ensemble Classifiers

NO

SGD Classifier

NOT WORKING

kernel approximation

NO

clustering

number of categories known

YES

<10K samples

YES

KMeans

NOT WORKING

Spectral Clustering GMM

NO

MiniBatch KMeans

NO

<10K samples

YES

MeanShift VBGMM

NO

tough luck

NO

predicting a quantity

YES

regression

<100K samples

NO

SGD Regressor

YES

few features should be important

YES

ElasticNet Lasso

NO

RidgeRegression SVR (kernel='linear')

NOT WORKING

SVR(kernel='rbf') EnsembleRegressors

NO

just looking

YES

dimensionality reduction

Randomized PCA

NOT WORKING

<10K samples

YES

Isomap Spectral Embedding

NOT WORKING

LLE

NO

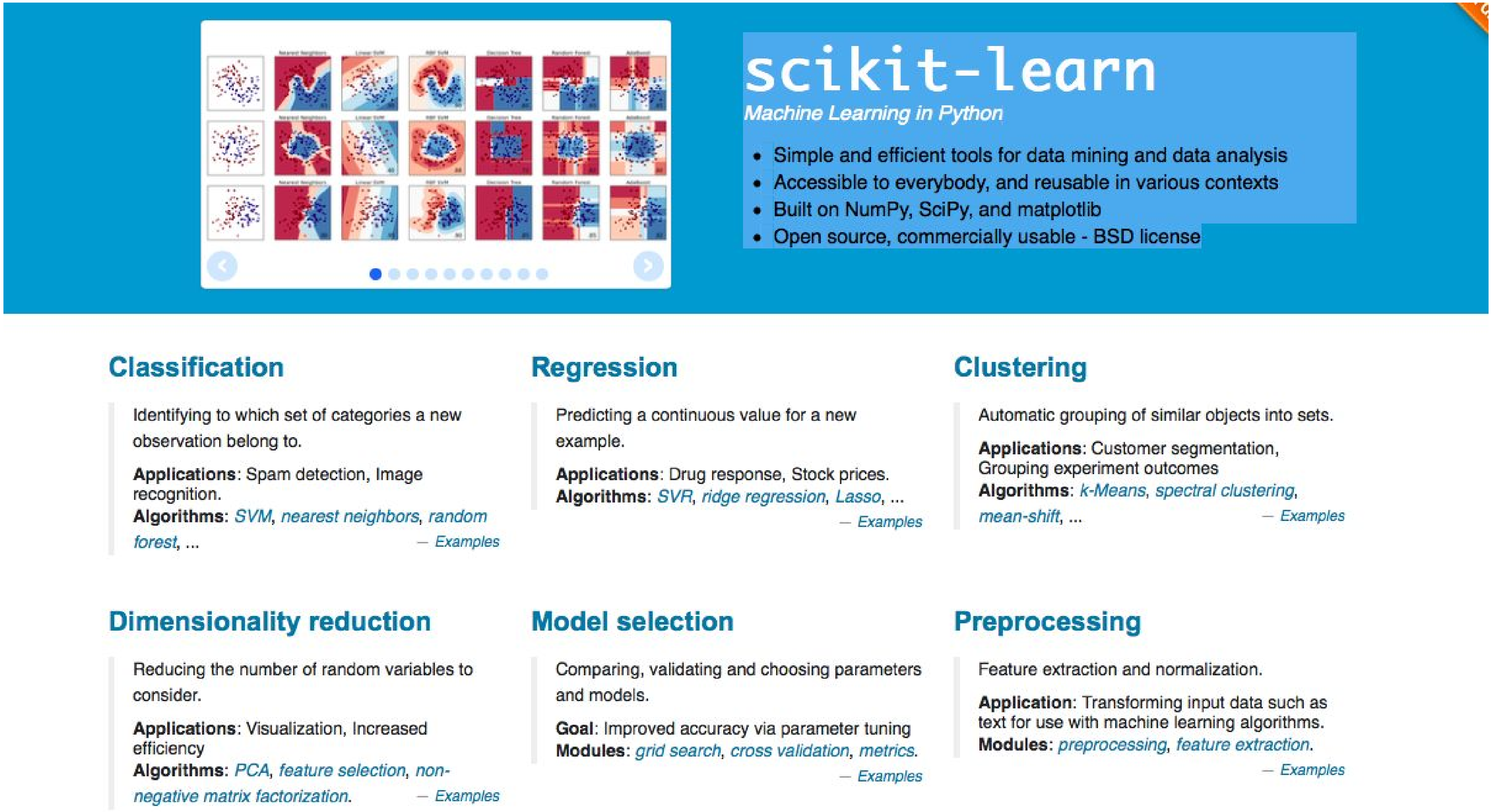
kernel approximation

NO

predicting structure

tough luck

What is scikit-learn?



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.
Algorithms: *SVM*, *nearest neighbors*, *random forest*, ...

— *Examples*

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.
Algorithms: *SVR*, *ridge regression*, *Lasso*, ...

— *Examples*

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: *k-Means*, *spectral clustering*, *mean-shift*, ...

— *Examples*

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: *PCA*, *feature selection*, *non-negative matrix factorization*.

— *Examples*

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: *grid search*, *cross validation*, *metrics*.

— *Examples*

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: *preprocessing*, *feature extraction*.

— *Examples*

What can we learn from this?

- <50 observations…get more data
  - Why? Inadequate power to effectively detect patterns or relationships. Visualization can still be very useful.
  - The power of a statistical test is the probability that it correctly rejects the null hypothesis when the null hypothesis is false.

What can we learn from this?

- Different Categories
  - Regression (today)
  - Classification (today)
  - Clustering
  - Dimension Reduction

Next Time

- Presentations
- Midterm