DATA ANALYTICS SYMPOSIUM

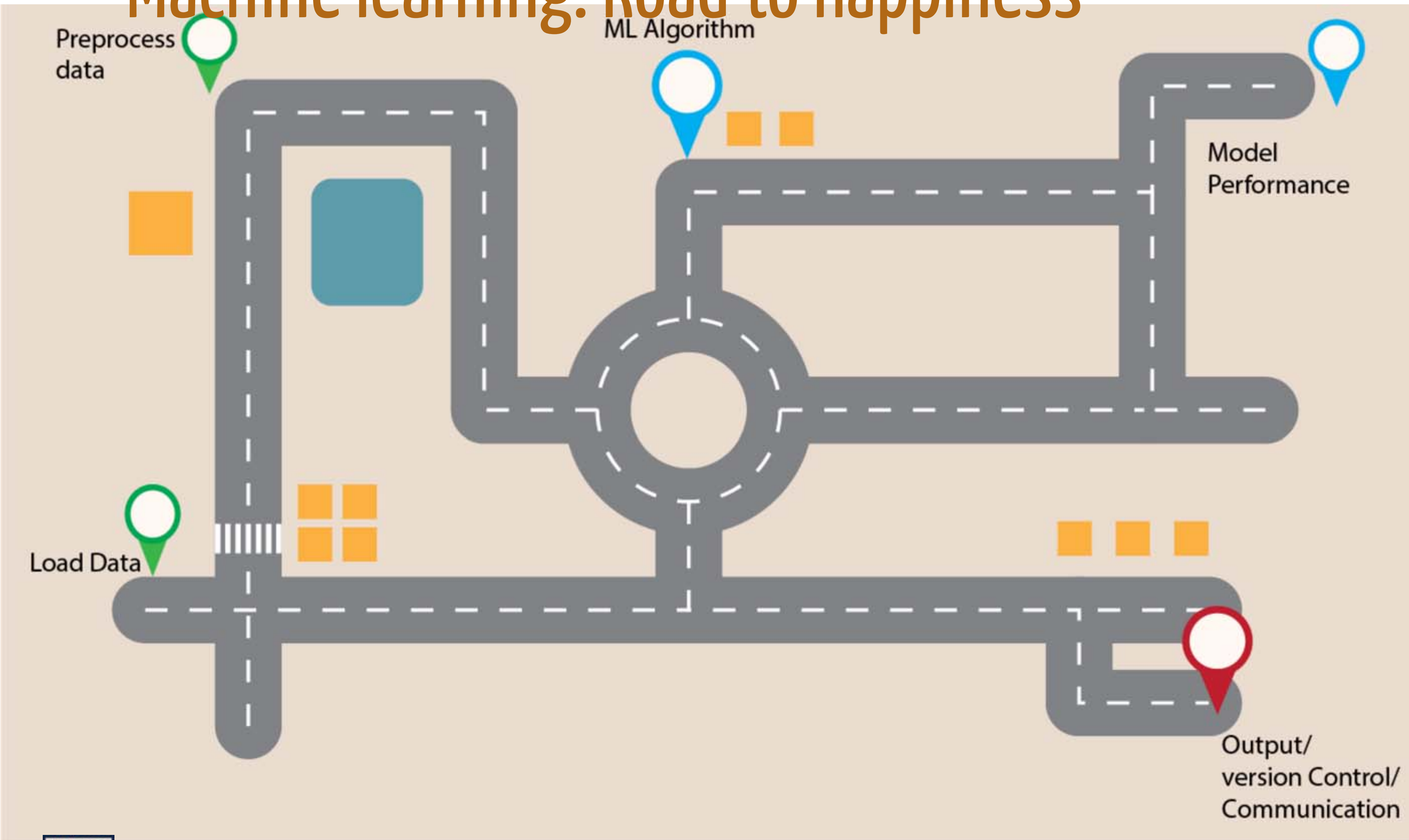
Smart Data

3 Case Studies of Machine Learning in Fundraising Analytics

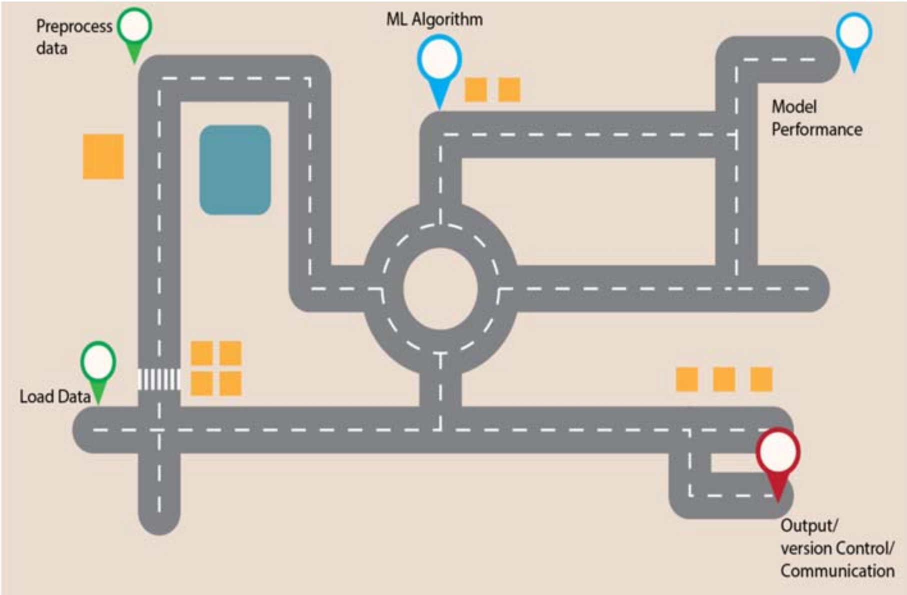
Claudia Rangel and Mai Bui

2018/7/16 (updated: 2018-08-09)

Machine learning: Road to happiness



3 / 70

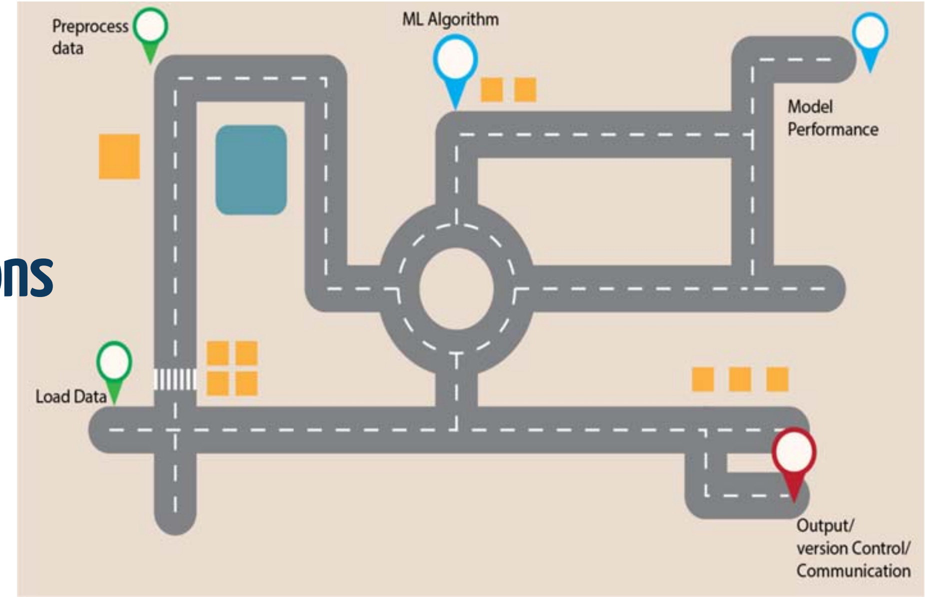

Preprocess data
ML Algorithm
Model Performance
Load Data
Output/ version Control/ Communication

Machine learning: Road to happiness

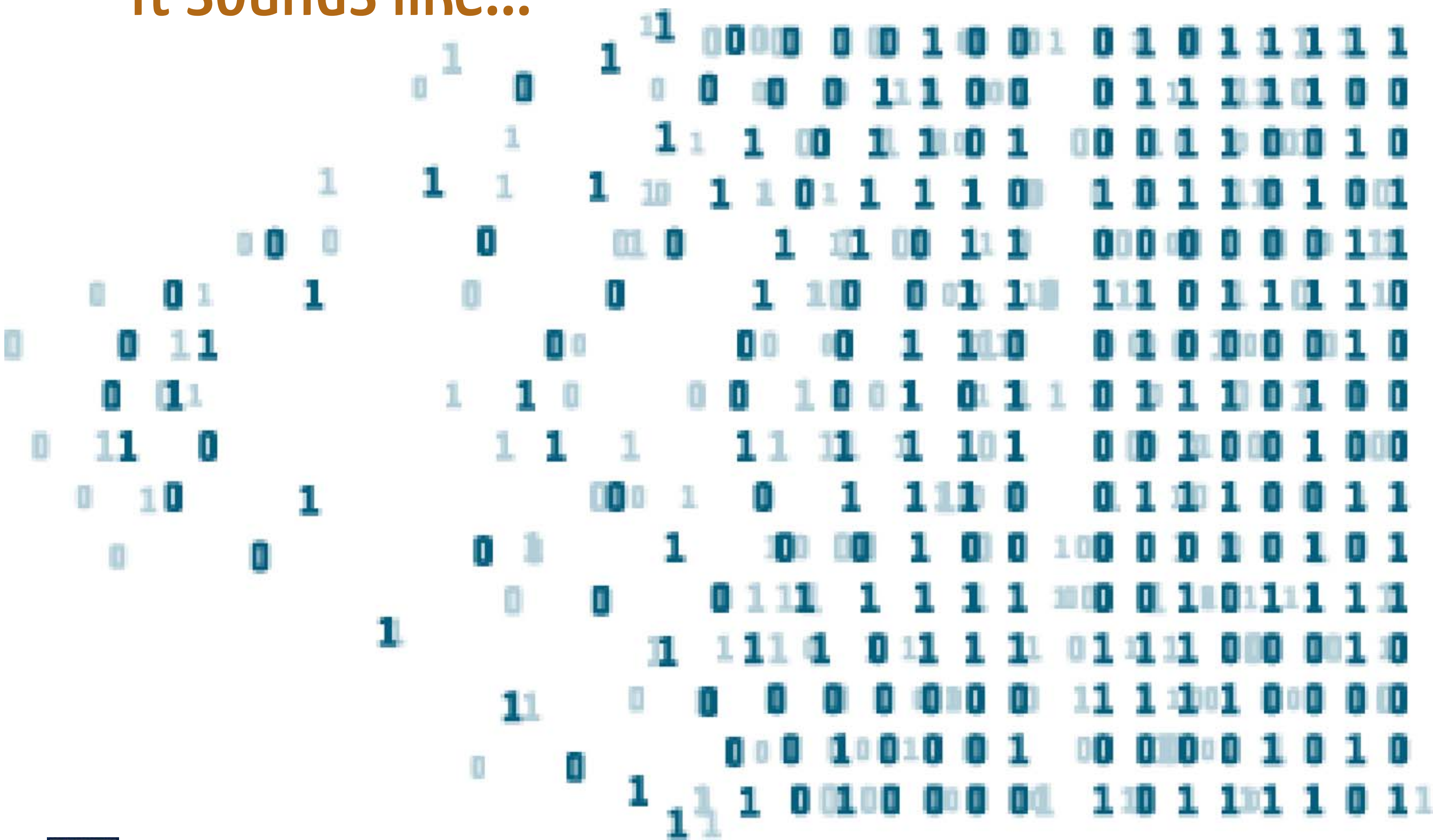
The Vehicle: Machine Learning

The Destination: Analytics questions

The route: ML analysis workflow



it sounds like...



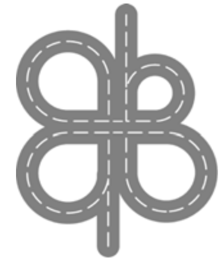
But it is actually more like...

Machine learning

Machine learning

- **Subfield of AI**

Machine learning



- **Subfield of AI**

- **automated learning approaches used to detect patterns in data *-algorithms-***

- ubiquitous: antispam software; search engines; product recommendation; website chatbots; face detection in phones and cameras...

Machine learning

- **Subfield of AI**
- **automated learning approaches used to detect patterns in data *-algorithms-***
- ubiquitous: antispam software; search engines; product recommendation; website chatbots; face detection in phones and cameras...
- **Concerned with prediction error on new data.**

--

- **Learning the patterns in data, with adjustable parameters -*tweaks*- by optimizing a performance metric -*benchmark*-**

Machine learning

- **Subfield of AI**

- **automated learning approaches used to detect patterns in data *-algorithms-***

- ubiquitous: antispam software; search engines; product recommendation; website chatbots; face detection in phones and cameras...

- **Concerned with prediction error on new data.**

--

- **Learning the patterns in data, with adjustable parameters -*tweaks*- by optimizing a performance metric -*benchmark*-**

Why Machine learning

Why Machine learning

- **Data as asset**

Why Machine learning

- **Data as asset**

- **Efficiency**

Why Machine learning

- **Data as asset**
- **Efficiency**
- **Automated workflow**

Overview of common ML Algorithm

Overview of common ML Algorithm

Unsupervised Learning

Overview of common ML Algorithm

Unsupervised Learning

- No target variable defined (identifying the different groups). Task focuses on grouping observations (donors, alumni,organizations) to maximize differences within groups.
- Clustering

Overview of common ML Algorithm

Unsupervised Learning

- No target variable defined (identifying the different groups). Task focuses on grouping observations (donors, alumni,organizations) to maximize differences within groups.
- Clustering

Supervised Learning

- We know what we are looking for (previous donor, event attendee, engaged alum). Task focuses on correctly classifying new observations.

Overview of common ML Algorithm

Unsupervised Learning

- No target variable defined (identifying the different groups). Task focuses on grouping observations (donors, alumni,organizations) to maximize differences within groups.
- Clustering

Supervised Learning

- We know what we are looking for (previous donor, event attendee, engaged alum). Task focuses on correctly classifying new observations.
- Classification

Overview of common ML Algorithm

Unsupervised Learning

- No target variable defined (identifying the different groups). Task focuses on grouping observations (donors, alumni,organizations) to maximize differences within groups.
- Clustering

Supervised Learning

- We know what we are looking for (previous donor, event attendee, engaged alum). Task focuses on correctly classifying new observations.
- Classification
- Regression

Overview of common ML Algorithm

Unsupervised Learning

- No target variable defined (identifying the different groups). Task focuses on grouping observations (donors, alumni,organizations) to maximize differences within groups.
- Clustering

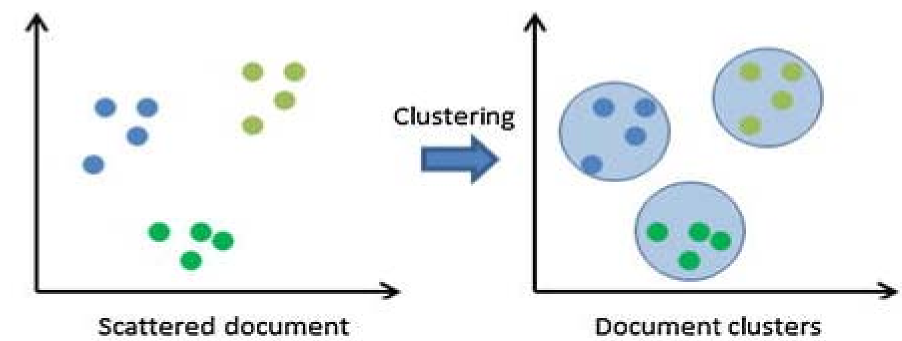
Supervised Learning

- We know what we are looking for (previous donor, event attendee, engaged alum). Task focuses on correctly classifying new observations.
- Classification
- Regression
- Many others out of intro scope: (Semi-supervised learning, Reinforcement learning, Deep Learning, Adversarial Learning)

ML: Unsupervised Learning

- K-means
- DBSCAN

Clustering



ML: Supervised Learning

ML: Supervised Learning

Rule-based approach

Decision tree, regression trees, and random forest algorithm.

ML: Supervised Learning

Rule-based approach

Decision tree, regression trees, and random forest algorithm.

Probabilistic approach

Naive Bayes algorithm.

ML: Supervised Learning

Rule-based approach

Decision tree, regression trees, and random forest algorithm.

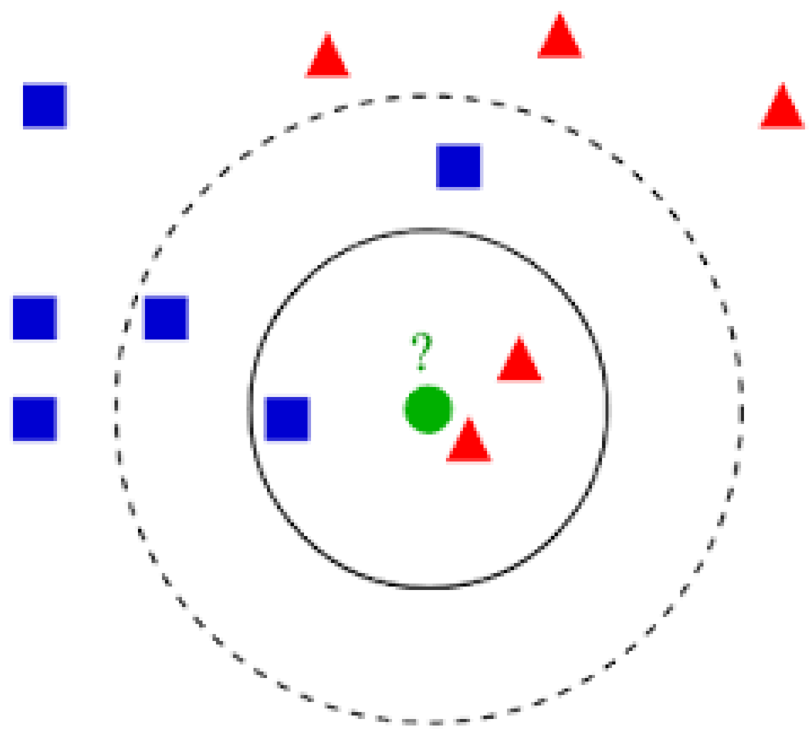
Probabilistic approach

Naive Bayes algorithm.

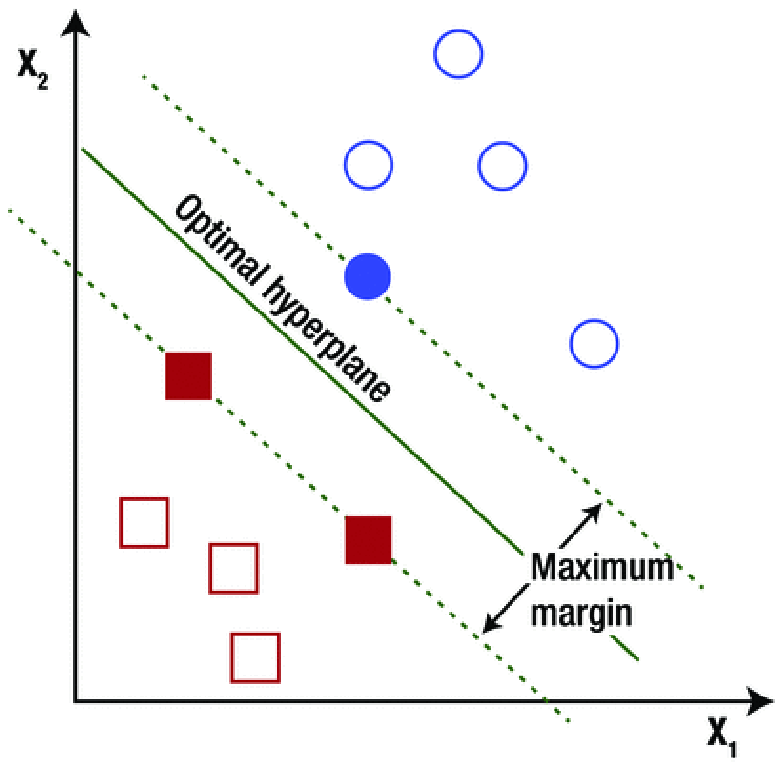
Distance-based approach

I Support vector machine, KNN

KNN



SVM

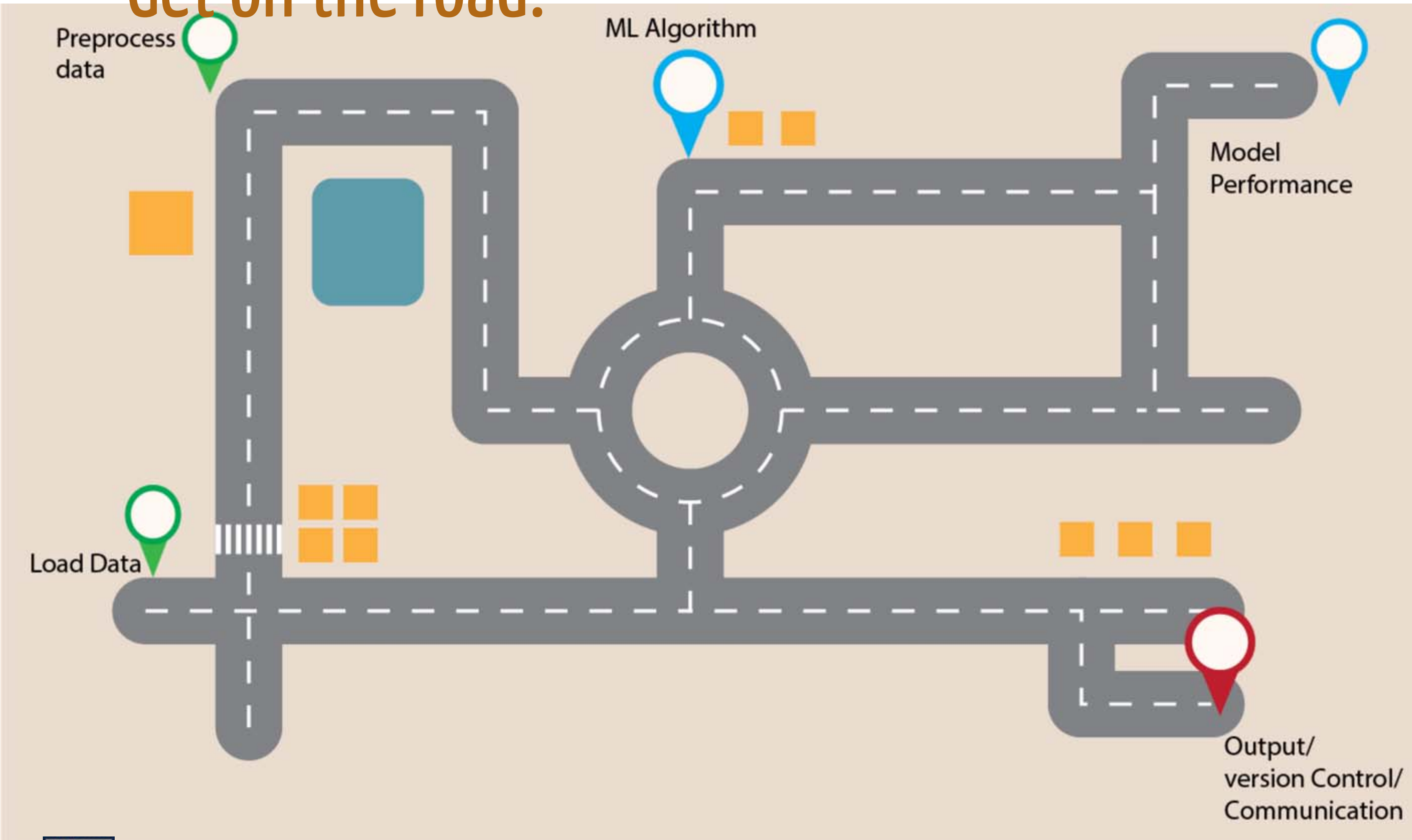


ML: Classification Algorithms

- **Ensembles**

- Common ML: **Random Forest, bagging boosting**
- Great: **Performance**
- Not so great: **Interpretability**

Get on the road!



Choose and setup your stack

Choose and setup your stack

- **Python**

Choose and setup your stack

- **Python**
- Conda/miniconda

Choose and setup your stack

- **Python**
- Conda/miniconda
- IDE, For instance: Jupyter notebook

Choose and setup your stack

- ## Python
- Conda/miniconda
- IDE, For instance: Jupyter notebook
- ## R

Choose and setup your stack

- ## Python
- Conda/miniconda
- IDE, For instance: Jupyter notebook
- ## R
- CRAN
- IDE, for instance: Rstudio

libraries code snippet:

Both RStudio and Jupyter notebooks can handle R and Python code= Great Integration

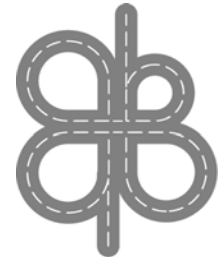
R

```
pacman:: p_load(rattle,RColorBrewer,rpart.plot,rpart,caret,DMwR,randomForest,e10
```

Python

```
import pandas
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt
import sklearn as sk
```

Why R/Python?



- Not GUI: code once, reuse many -automation, efficiency
- Code IS documentation -reproducibility
- Open source -free!!! Easy to share
- Extensible: commited people creating - ready-to-use libraries for data analysis/ML
- Cross-platform

Read Data

--

Import

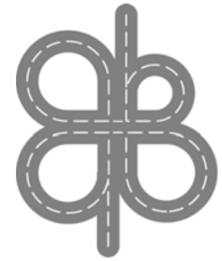
--

Merge

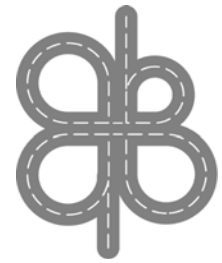
--

Fuzzy match

Data Prep/Data Wrangling



Data Prep/Data Wrangling



- **Remove blanks**

Data Prep/Data Wrangling

- Remove blanks
- Convert to numeric

Data Prep/Data Wrangling

- **Remove blanks**
- **Convert to numeric**
- **Check imbalanced target variable** ***-rare outcome-***

Data Prep/Data Wrangling

- Remove blanks
- Convert to numeric
- Check imbalanced target variable *-rare outcome-*
- Run descriptives

--

- Centered/Scale/ Deal with outlier. When appropriate!

	Partition 1	Partition 2	Partition 3
Iteration 1	Train	Train	Test
Iteration 2	Train	Test	Train
Iteration 3	Test	Train	Train

Script model

	Partition 1	Partition 2	Partition 3
Iteration 1	Train	Train	Test
Iteration 2	Train	Test	Train
Iteration 3	Test	Train	Train

Script model

- **Understand how choice of parameter (model options) affect your results.**

  i.e. number of trees in classification tasks, number of neighbors in KNN task

	Partition 1	Partition 2	Partition 3
Iteration 1	Train	Train	Test
Iteration 2	Train	Test	Train
Iteration 3	Test	Train	Train

Script model

- **Understand how choice of parameter (model options) affect your results.**

  i.e. number of trees in classification tasks, number of neighbors in KNN task

- **Cross-validation is important**

	Partition 1	Partition 2	Partition 3
Iteration 1	Train	Train	Test
Iteration 2	Train	Test	Train
Iteration 3	Test	Train	Train

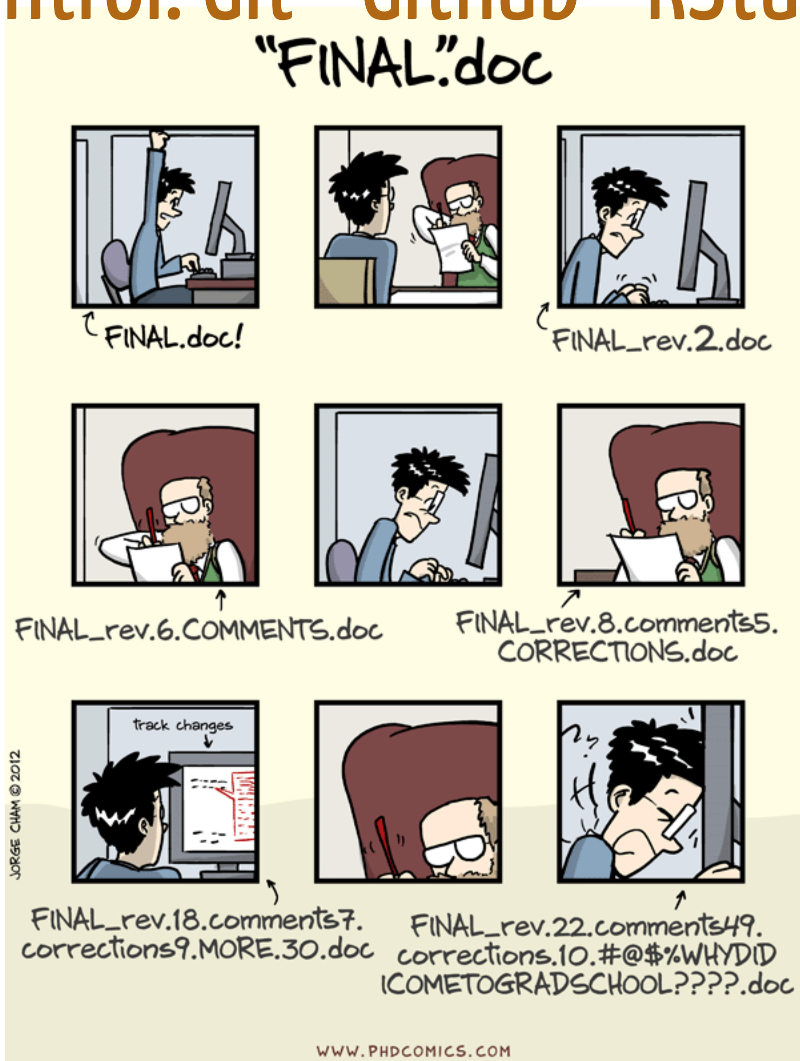
Model Performance

Many metrics. Best one depends on your goal.

- Accuracy, ROC, Entropy, Loss Function

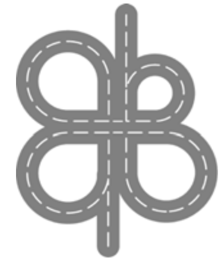
		Actual	
		1	0
Predicted	1	A	B
	0	C	D

Version Control: Git - Github - RStudio



- **Git is a version control system**

Files structured in a repository



- **Git is a version control system**

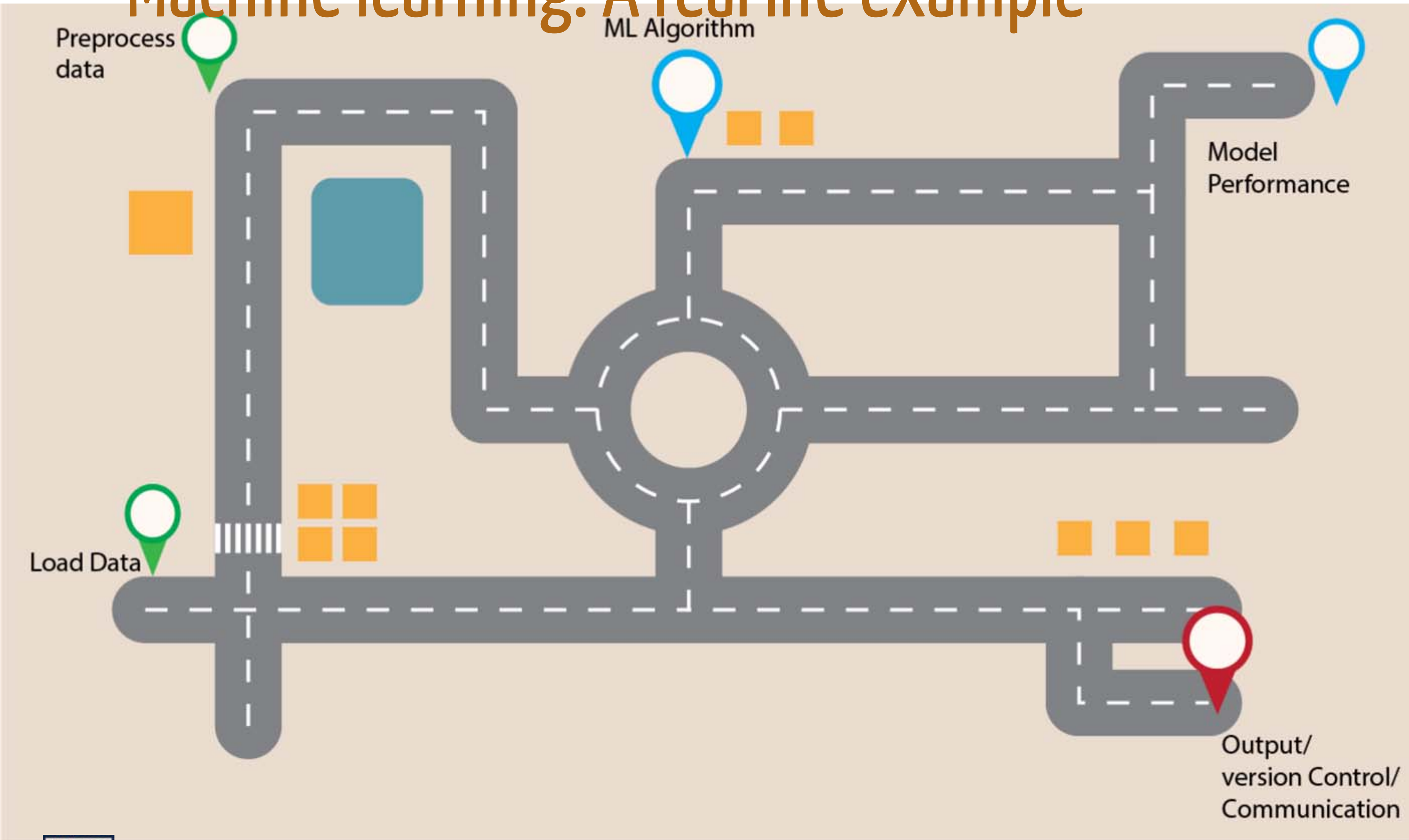
Files structured in a repository

- **Github provides a hosting service for your git repositories on the internet. (dropbox loose example).**

Functionality: share, synch, make changes

Remote repository that we can go back to, after the feared delete/replace/corrupt local files crises.

Machine learning: A real life example



Brainstorming

Brainstorming

- **Pick a problem from your *unending* analysis to-do list**

Brainstorming

- Pick a problem from your *unending* analysis to-do list
- Choose the ML algorithm that best match the question and data

Brainstorming

- Pick a problem from your *unending* analysis to-do list
- Choose the ML algorithm that best match the question and data
- Think about the nature of the data: preprocessing, imbalanced

Brainstorming

- **Pick a problem from your *unending* analysis to-do list**
- **Choose the ML algorithm that best match the question and data**
- **Think about the nature of the data: preprocessing, imbalanced**
- **What is the *first barrier* to start your road to happiness**
    - **Can we help?** DAS community / book resources / specific questions

Next Steps:

Next Steps:

- **Detailed R and Python notebooks are available in our github site**

Next Steps:

- Detailed R and Python notebooks are available in our github site

- Got more time?

Next Steps:

- Detailed R and Python notebooks are available in our github site

- Got more time?

These are some extra Resources

Next Steps:

- **Detailed R and Python notebooks are available in our github site**

- **Got more time?**

These are some extra Resources

- **Github, Git and Rstudio for version control workflows:**

  http://happygitwithr.com/

- Rstudio

- Rbloggers, useR

- Stackoverflow, Kaggle

Thanks!

This presentation used xaringan ninja style.

CSS file based on Rladies by Alison Presmanes Hill