ROADMAP:
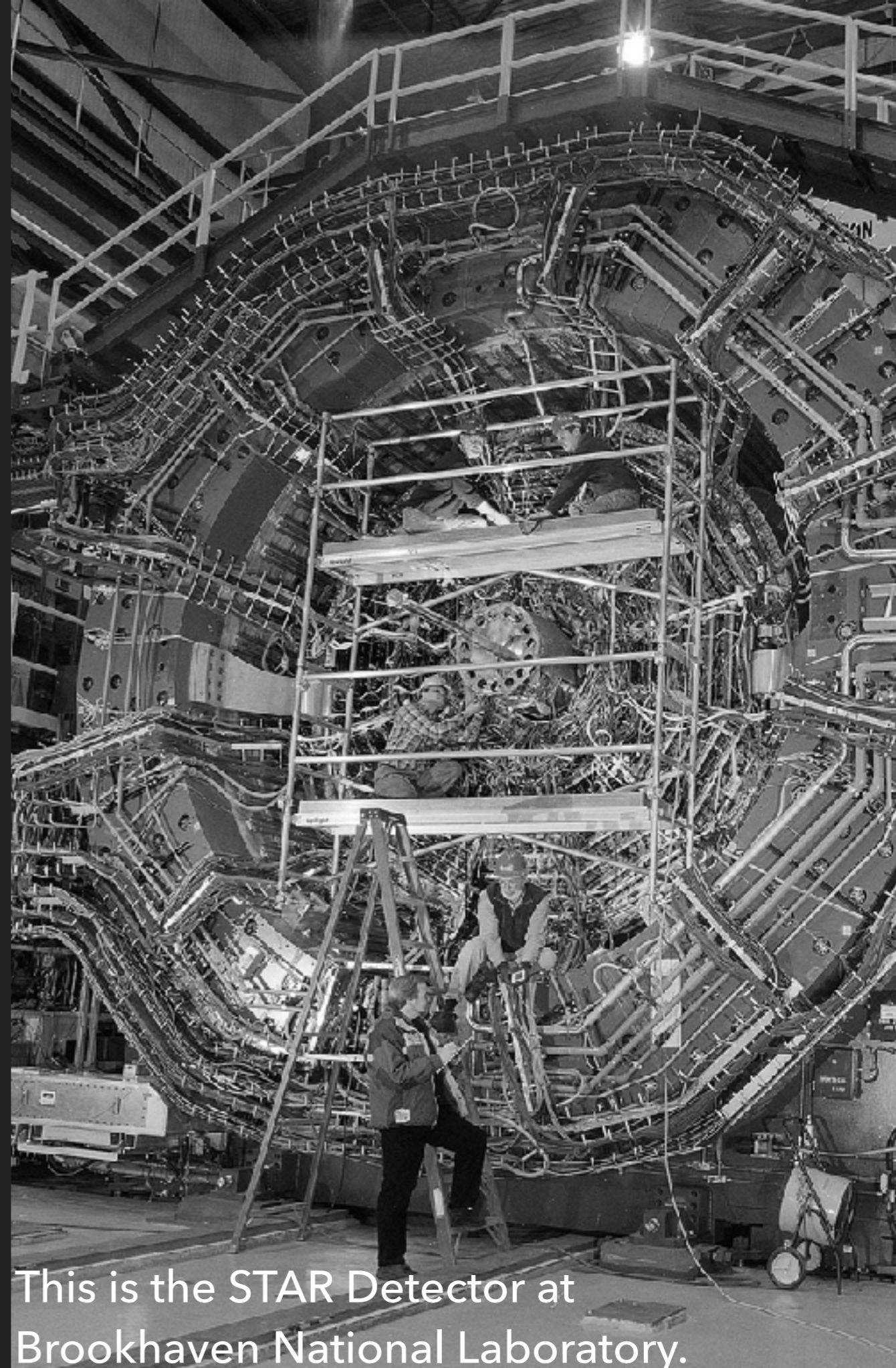## SIX MONTHS TO MACHINE LEARNING
ZACHARIAH MILLER - 5/20/17

WHO AM I?

AND WHY SHOULD YOU LISTEN TO ME?

This is the STAR Detector at Brookhaven National Laboratory.

IF YOU ONLY REMEMBER ONE THING FROM THIS TALK

JUST BUILD SOMETHING WITH DATA

AND EXPECT TO SUCK AT IT, FOR A WHILE

My First
Data Project

Unedited.

```cpp
G4Box* box1 = new G4Box(-10.,-10.,-9.,-9.,g4water);
G4Box* box2 = new G4Box(-10.01,-10.,-9.,-9.,g4water);
G4Box* box3 = new G4Box(-10.02,-10.,-9.,-9.,g4water);
G4Box* box4 = new G4Box(-10.03,-10.,-9.,-9.,g4water);
G4Box* box5 = new G4Box(-10.04,-10.,-9.,-9.,g4water);
G4Box* box6 = new G4Box(-10.05,-10.,-9.,-9.,g4water);
G4Box* box7 = new G4Box(-10.06,-10.,-9.,-9.,g4water);
G4Box* box8 = new G4Box(-10.07,-10.,-9.,-9.,g4water);
G4Box* box9 = new G4Box(-10.08,-10.,-9.,-9.,g4water);
G4Box* box10 = new G4Box(-10.09,-10.,-9.,-9.,g4water);
G4Box* box11 = new G4Box(-10.10,-10.,-9.,-9.,g4water);
G4Box* box12 = new G4Box(-10.11,-10.,-9.,-9.,g4water);
G4Box* box13 = new G4Box(-10.12,-10.,-9.,-9.,g4water);
G4Box* box14 = new G4Box(-10.13,-10.,-9.,-9.,g4water);
G4Box* box15 = new G4Box(-10.14,-10.,-9.,-9.,g4water);
G4Box* box16 = new G4Box(-10.15,-10.,-9.,-9.,g4water);
G4Box* box17 = new G4Box(-10.16,-10.,-9.,-9.,g4water);
G4Box* box18 = new G4Box(-10.17,-10.,-9.,-9.,g4water);
G4Box* box19 = new G4Box(-10.18,-10.,-9.,-9.,g4water);
G4Box* box20 = new G4Box(-10.19,-10.,-9.,-9.,g4water);
G4Box* box21 = new G4Box(-10.20,-10.,-9.,-9.,g4water);
G4Box* box22 = new G4Box(-10.21,-10.,-9.,-9.,g4water);
G4Box* box23 = new G4Box(-10.22,-10.,-9.,-9.,g4water);
G4Box* box24 = new G4Box(-10.23,-10.,-9.,-9.,g4water);
G4Box* box25 = new G4Box(-10.24,-10.,-9.,-9.,g4water);
G4Box* box26 = new G4Box(-10.25,-10.,-9.,-9.,g4water);
G4Box* box27 = new G4Box(-10.26,-10.,-9.,-9.,g4water);
G4Box* box28 = new G4Box(-10.27,-10.,-9.,-9.,g4water);
G4Box* box29 = new G4Box(-10.28,-10.,-9.,-9.,g4water);
G4Box* box30 = new G4Box(-10.29,-10.,-9.,-9.,g4water);
G4Box* box31 = new G4Box(-10.30,-10.,-9.,-9.,g4water);
G4Box* box32 = new G4Box(-10.31,-10.,-9.,-9.,g4water);
G4Box* box33 = new G4Box(-10.32,-10.,-9.,-9.,g4water);
G4Box* box34 = new G4Box(-10.33,-10.,-9.,-9.,g4water);
G4Box* box35 = new G4Box(-10.34,-10.,-9.,-9.,g4water);
G4Box* box36 = new G4Box(-10.35,-10.,-9.,-9.,g4water);
G4Box* box37 = new G4Box(-10.36,-10.,-9.,-9.,g4water);
G4Box* box38 = new G4Box(-10.37,-10.,-9.,-9.,g4water);
G4Box* box39 = new G4Box(-10.38,-10.,-9.,-9.,g4water);
```

My First
Data Project

Edited.

```
float x = -10.00
for(int i = 0; i <= 100; i++){
  boxes[i] = new G4Box(x, -10., -9., -9. g4water);
  x -= 0.01
}
```

But the first one made the
plots in my thesis.

THE (SOMEWHAT) UNFORTUNATE TRUTH
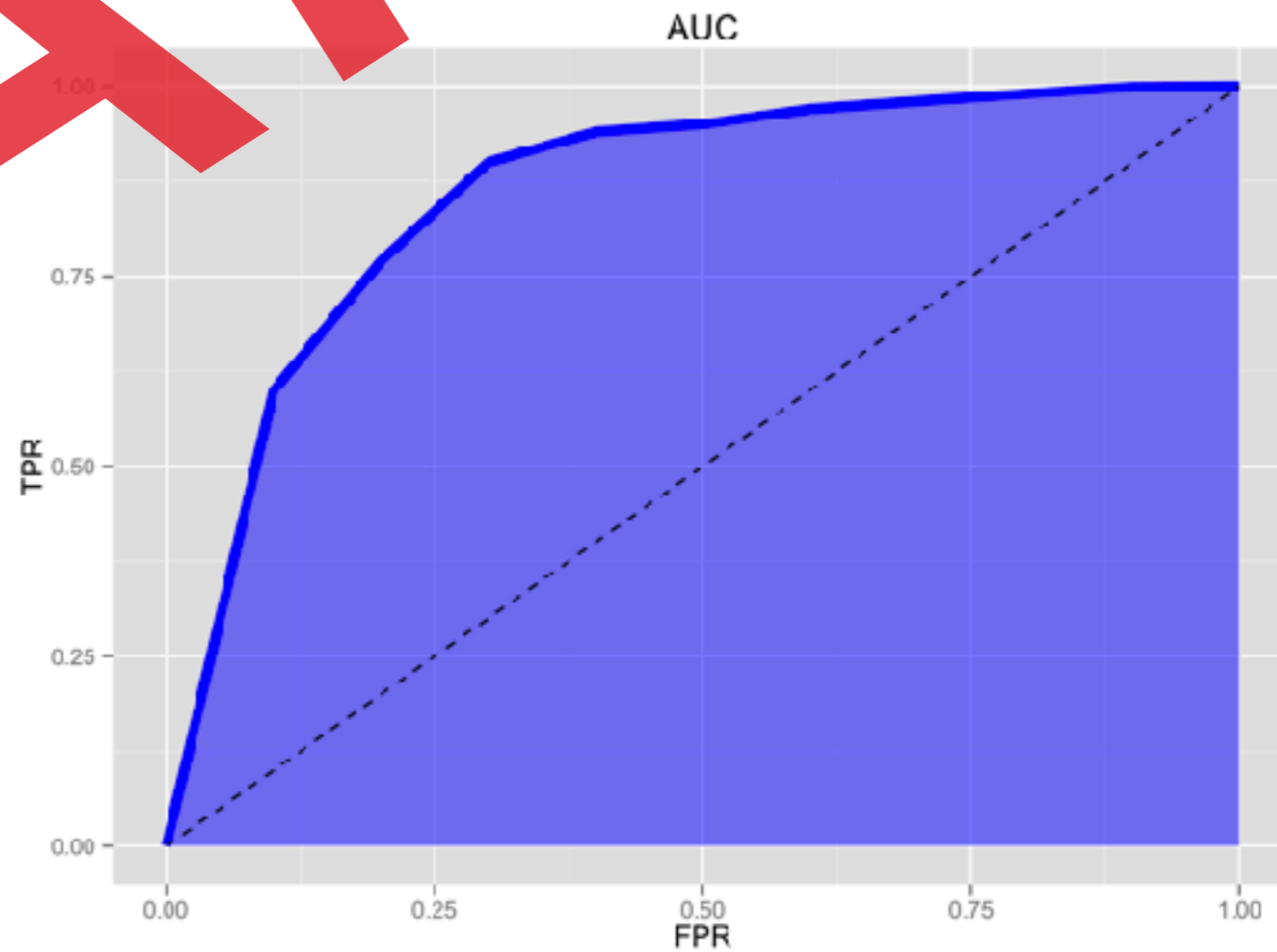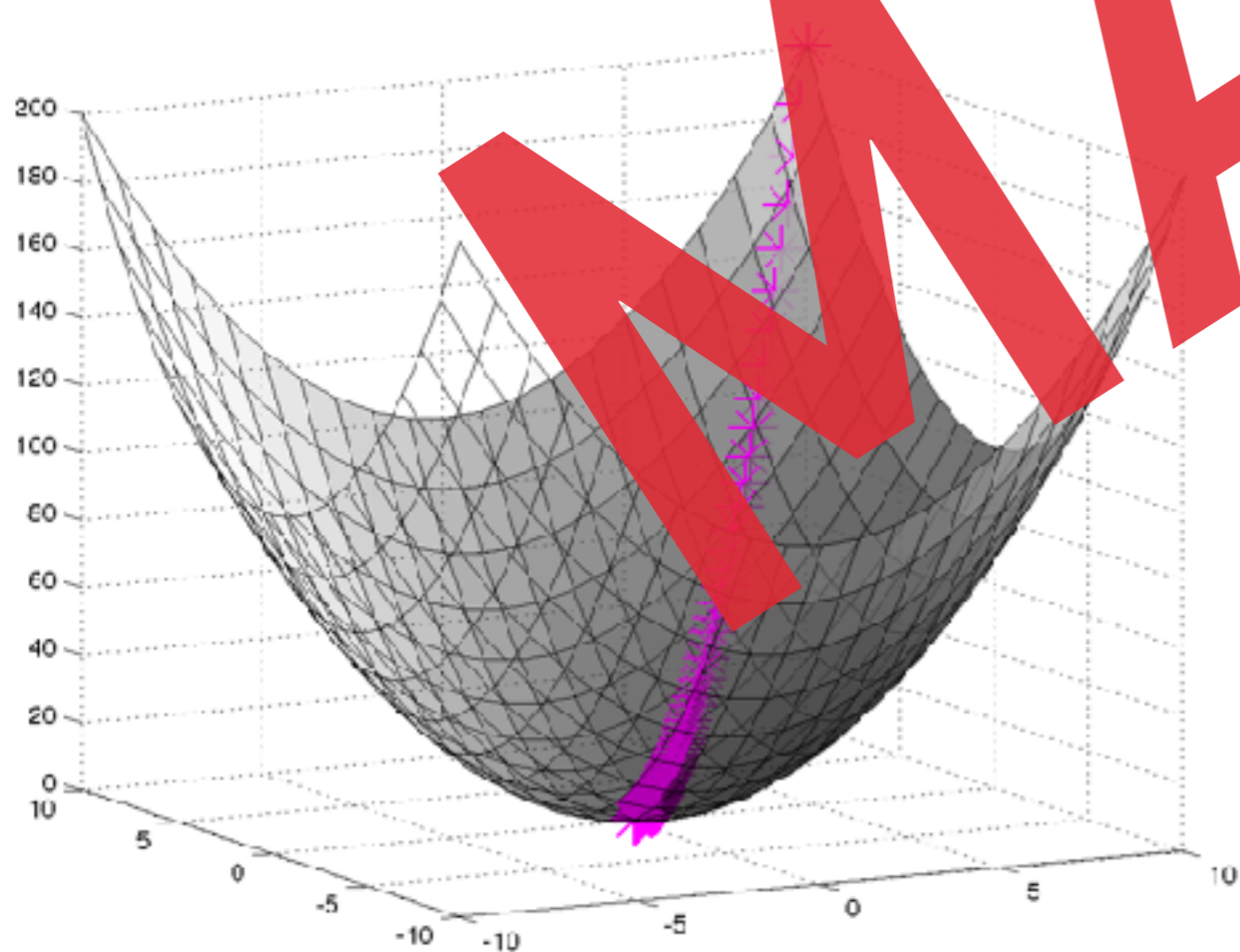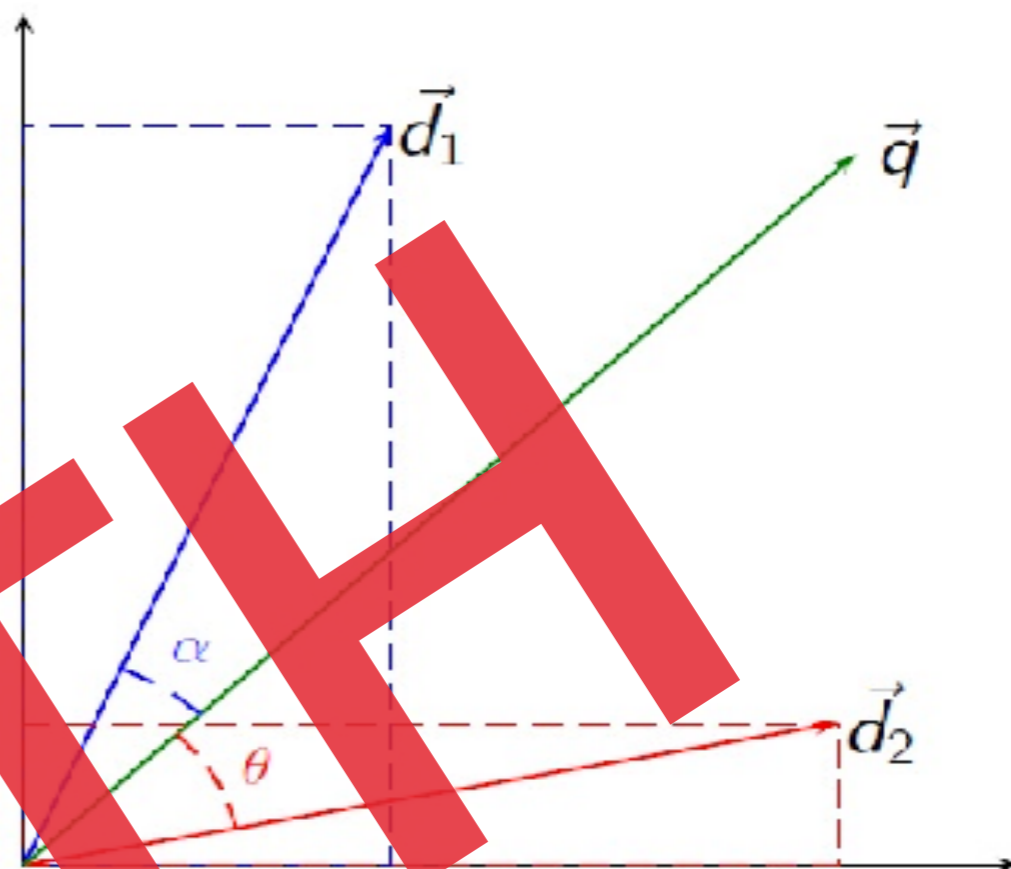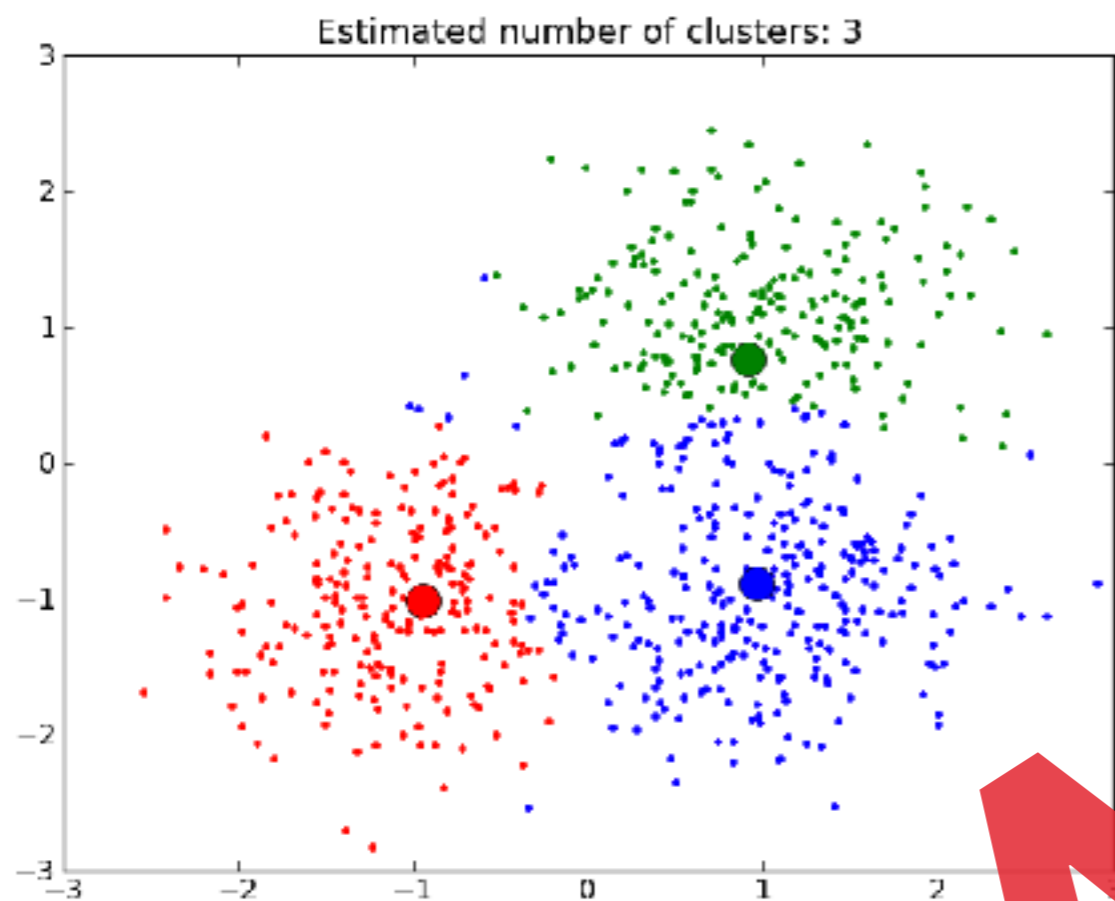
▸ Linear Algebra

▸ Calculus

▸ Statistics

▸ Probability

▸ [MIT Linear Algebra Open Course](#)

▸ [MIT Calculus Open Course](#)

▸ [MIT Stats and Probability Course](#)

Python vs R

Data Science from Scratch

FIRST PRINCIPLES WITH PYTHON

Joel Grus

Machine Learning with R

Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications

Brett Lantz

▸ Find Data (Webscrape, APIs, CSVs)

▸ Clean the data (remove NaNs and Infinities, should that be a string? Probably not, maybe I can categorize it…)

YOU'LL SPEND A LOT OF TIME IN THESE SECTIONS. THAT'S NORMAL.

THE MACHINE LEARNING PIPELINE

▸ Find Data (Webscrape, APIs, CSVs)

▸ Clean the data (remove NaNs and Infinities, should that be a string? Probably not, maybe I can categorize it…)

▸ Choose and tune your algorithm

▸ Visualize results

GOOGLE AND STACK OVERFLOW ARE YOUR FRIENDS

▸ Find Data (Webscrape, APIs, CSVs)

▸ Clean the data (remove NaNs and Infinities, should that be a string? Probably not, maybe I can categorize it…)

▸ Choose and tune your algorithm

▸ Visualize results

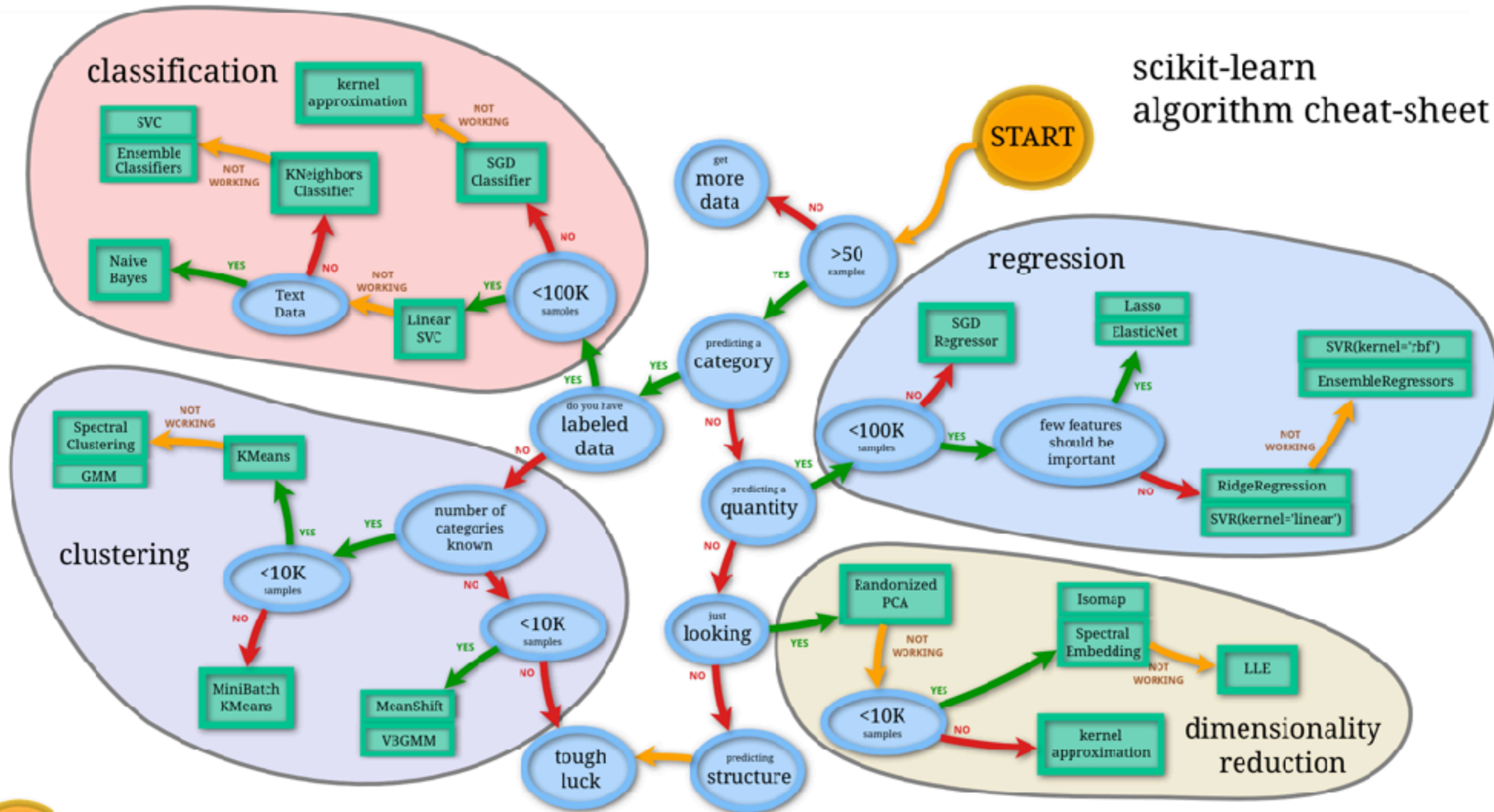## GOOGLE AND STACK OVERFLOW ARE YOUR FRIENDS

scikit-learn
algorithm cheat-sheet

classification

- kernel approximation
- SVC
- Ensemble Classifiers
- KNeighbors Classifier
- SGD Classifier
- Naive Bayes
- Linear SVC

clustering

- Spectral Clustering
- GMM
- KMeans
- MiniBatch KMeans
- MeanShift
- VBGMM

regression

- SGD Regressor
- Lasso
- ElasticNet
- SVR(kernel='rbf')
- EnsembleRegressors
- RidgeRegression
- SVR(kernel='linear')

dimensionality reduction

- Randomized PCA
- Isomap
- Spectral Embedding
- LLE
- kernel approximation

START

- get more data
- >50 samples
- predicting a category
- do you have labeled data
- <100K samples
- Text Data
- number of categories known
- <10K samples
- <10K samples
- predicting a quantity
- <100K samples
- few features should be important
- just looking
- predicting structure
- tough luck
- <10K samples

Back

scikit learn

SO… SIX MONTHS?

▸ Learn the math. (2 - 3 months)

▸ Learn the programming language (1 month)

▸ Machine learning tutorials and test projects (1 - 2 months)

▸ Short term passion projects (1+ month)

METIS

DATA SCIENCE AND MACHINE LEARNING CLASSES/BOOTCAMPS

SOME LAST NOTES

- You're going to fail. A lot. It's software… so who cares.

- Your models may not be predictive. That's a result. Null is just as good as non-null if you did it right.

- Track your projects on GitHub and write up your results.

THANKS!

LET'S CHAT. I'D LOVE TO TALK ABOUT PROJECTS YOU'RE CONSIDERING.

ZACH@THISISMETIS.COM

ZWMILLER.COM

PHOTO CREDITS

▸ https://c1.staticflickr.com/4/3310/3313585315_4874a81f77_b.jpg (STAR Detector)

▸ https://c1.staticflickr.com/6/5531/9638435181_7e3e44c2b8_b.jpg (Highway)

▸ https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/35389/versions/1/screenshot.png (Gradient Descent)

▸ http://blog.yhat.com/static/img/roc-auc.png (AUC)

▸ http://scikit-learn.org/stable/tutorial/machine_learning_map/ (SkLearn Map)