Welcome to Data Science!

who am i?

From New York

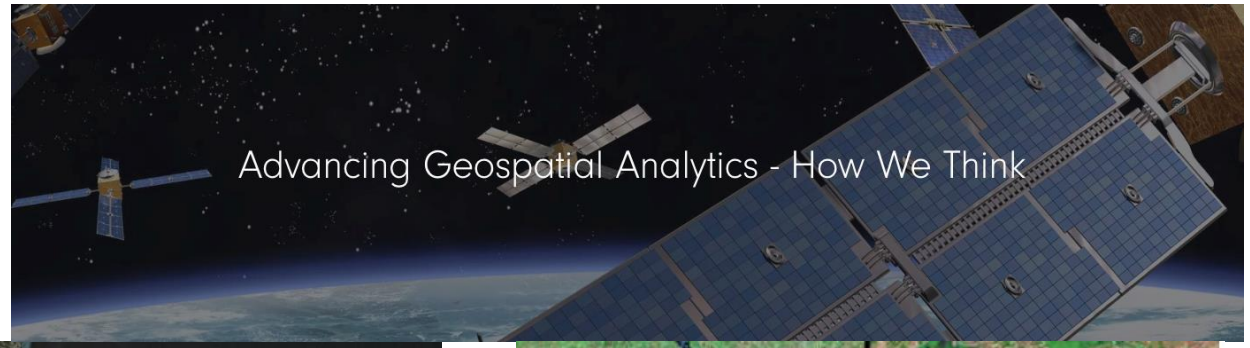
B.S. in Physics From Stanford ‘16

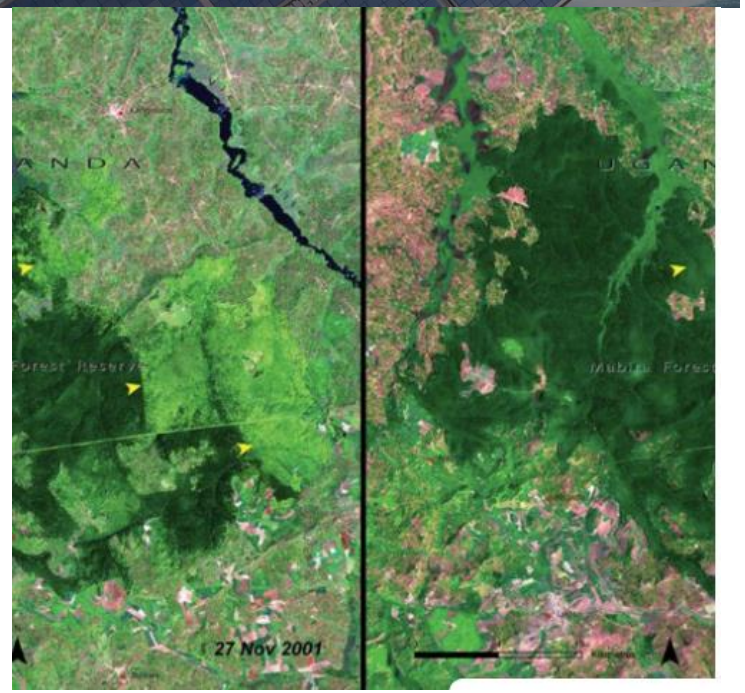
$$i\hbar \frac{\partial}{\partial t}\Psi = H\Psi$$

Started a PhD in Physics at University of Illinois

Data Scientist at Orbital Insight

Orbital Insight

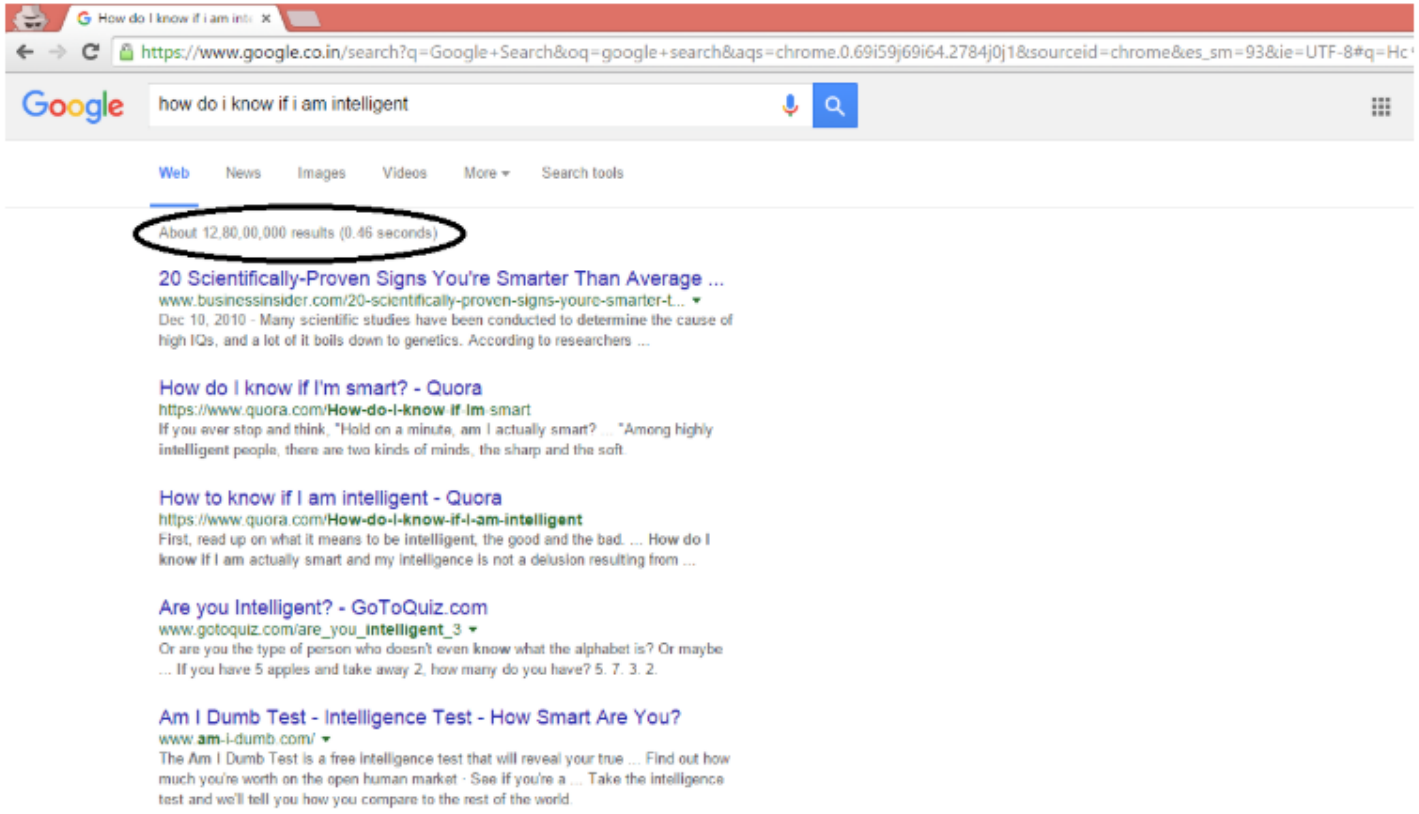




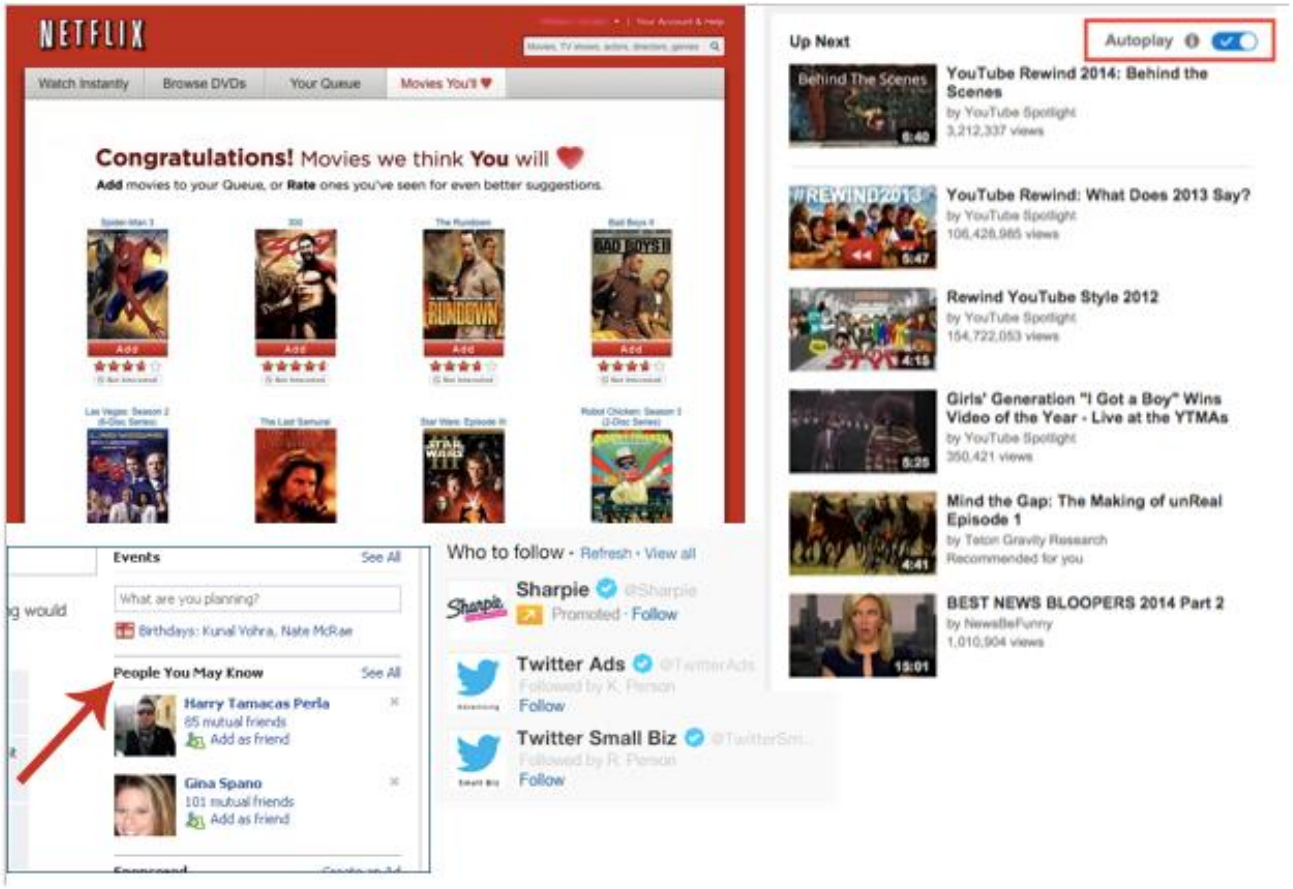
What is Data Science?

- “uses **scientific** methods, processes, algorithms and systems to extract knowledge and insights from **data**”

Internet Search



Recommender Systems



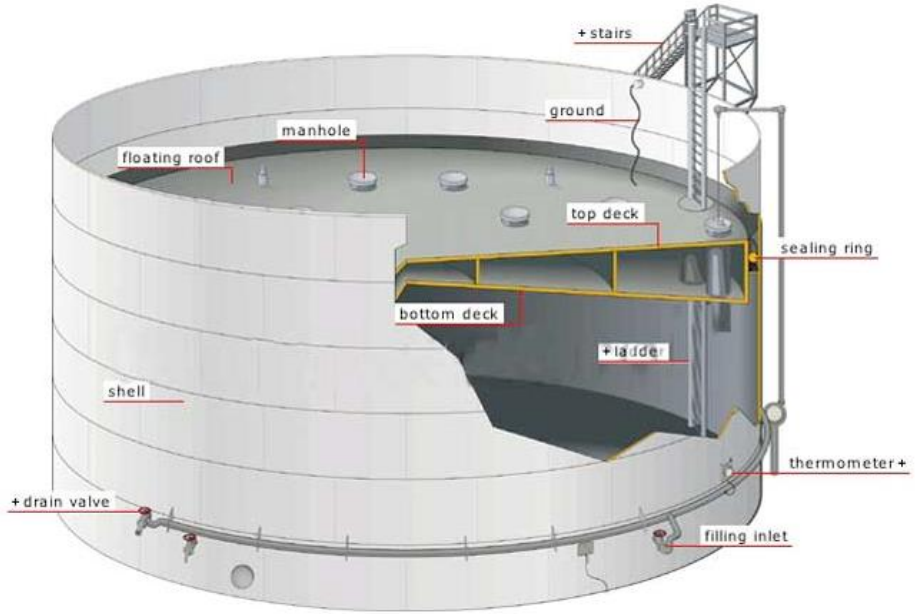
Google Translate

الحب Love	الحكمة Wisdom
السّلام Peace	الحياة Life

Self Driving Cars

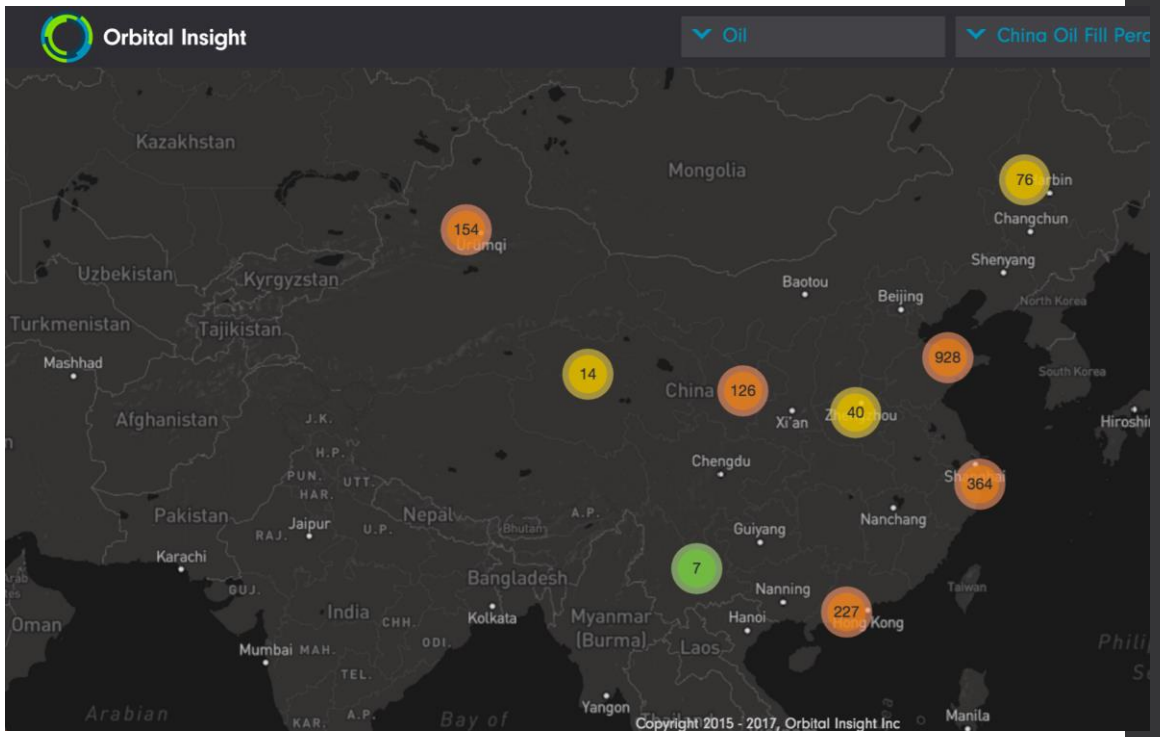
What do I do at Orbital Insight?

What We Measure



Energy Product Vision

Orbital Insight will delight customers with insights on all aspects of the global energy supply chain that can be observed from overhead data, providing an unprecedented objective view to make more informed decisions.



World Oil Storage

Dalian, China's largest oil storage hub

Monitor crude oil inventory worldwide

- Track 25k+ storage tanks in 1000+ areas around the world by satellite.
- Create continuously-updated estimate of global oil inventory
- More comprehensive and much more timely than the International Energy Administration's survey methods
- Insight about regions that do not publish storage data, e.g. China, South America

In the beginning...

...sometime in 2014, Jimi, et al said let's measure all the crude oil stored, worldwide, every day…

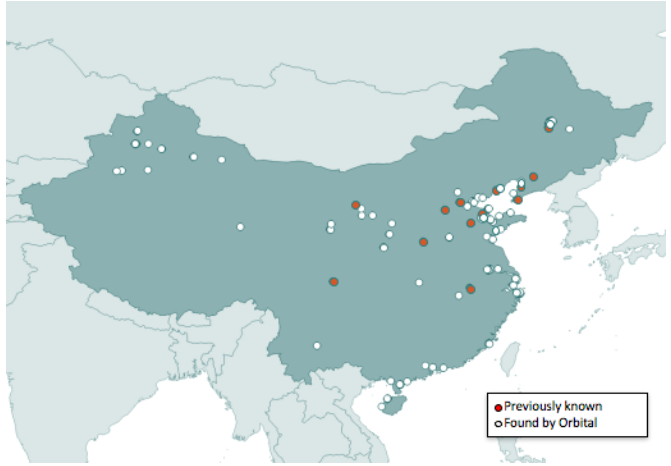
easy right?

hmmm….

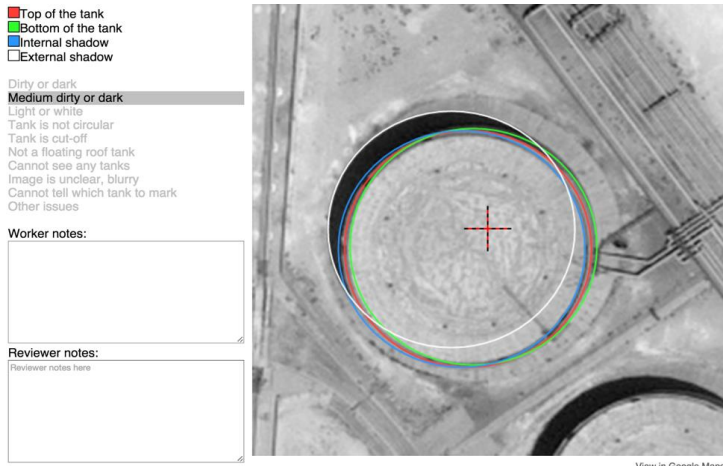
imaged 2x/month

How It Works at Scale

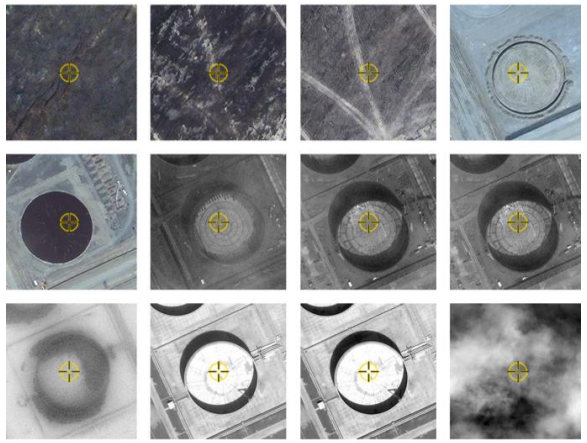
Wide Area Search



Tank Dimensioning

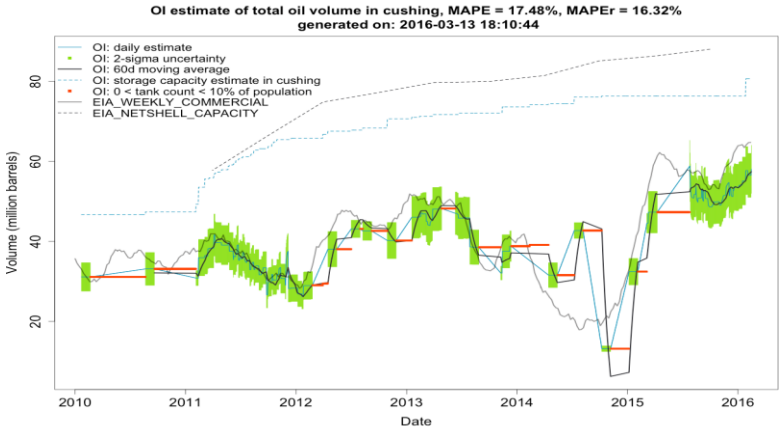


Tank Lifecycle

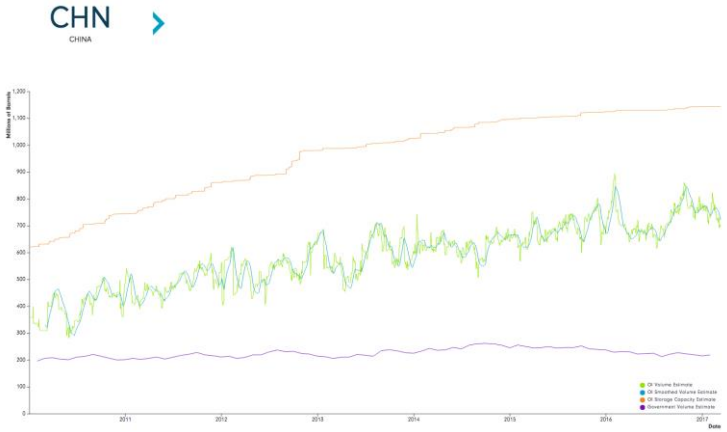


Tank Measurements

Estimates



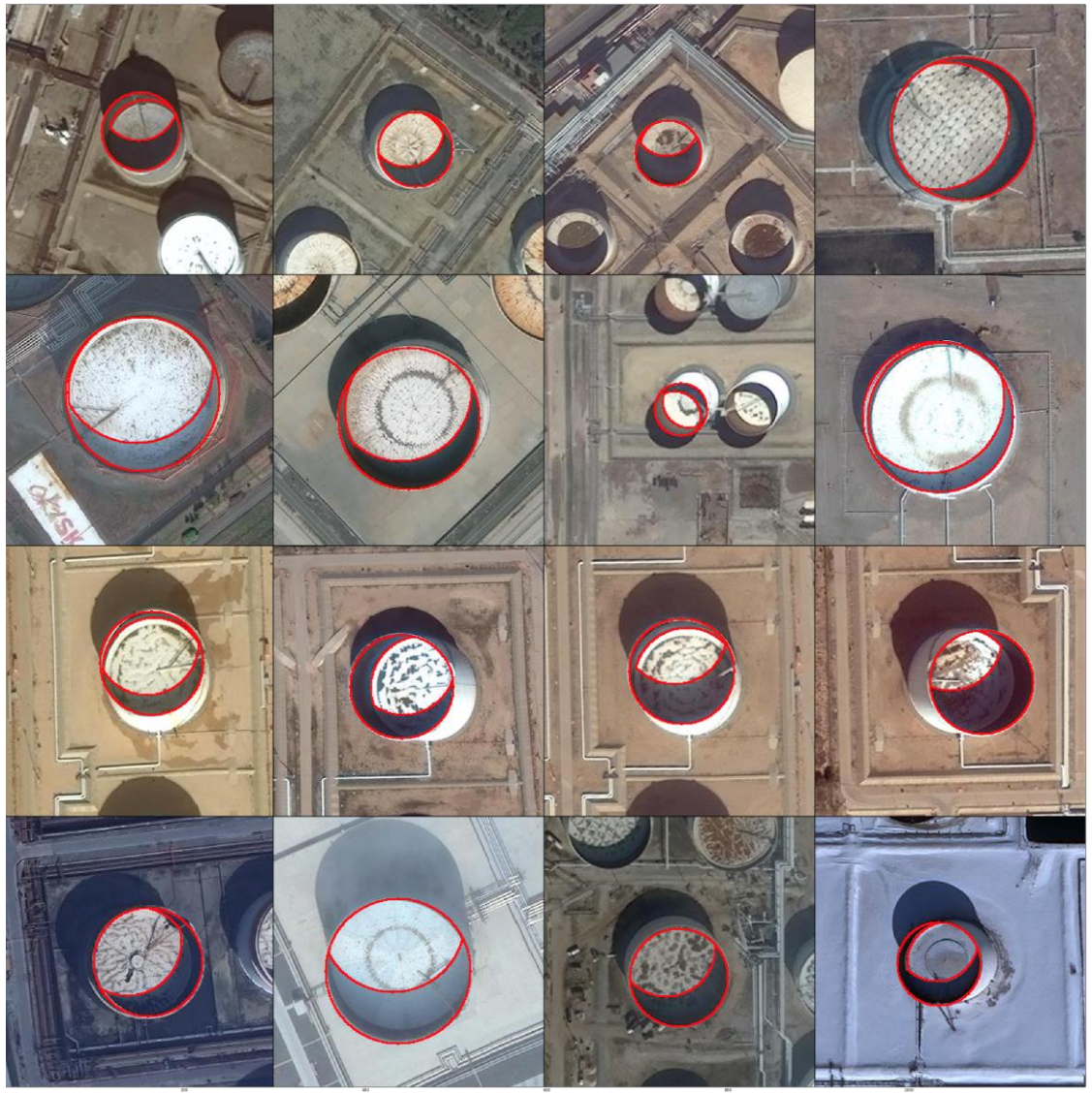
Time Series

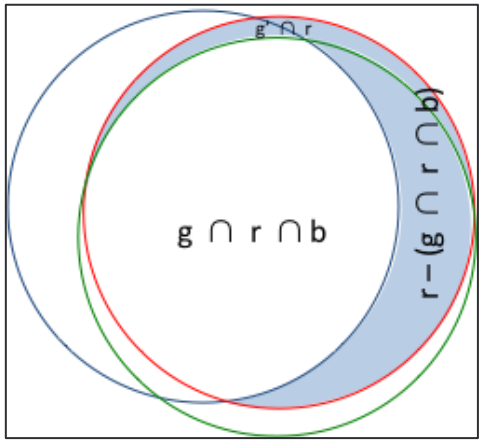


Finding “Unknown” Oil Tanks in China

- Over 2,000 tanks discovered by OI scann algorithm

Computer Vision of Oil Tanks



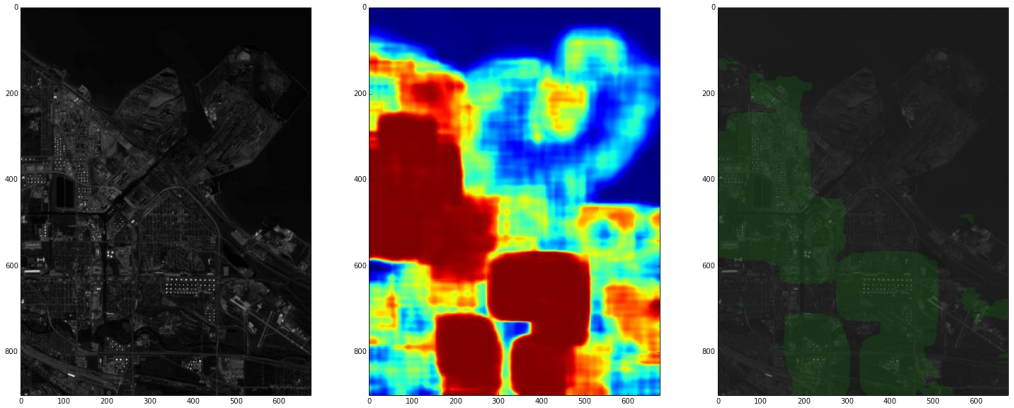


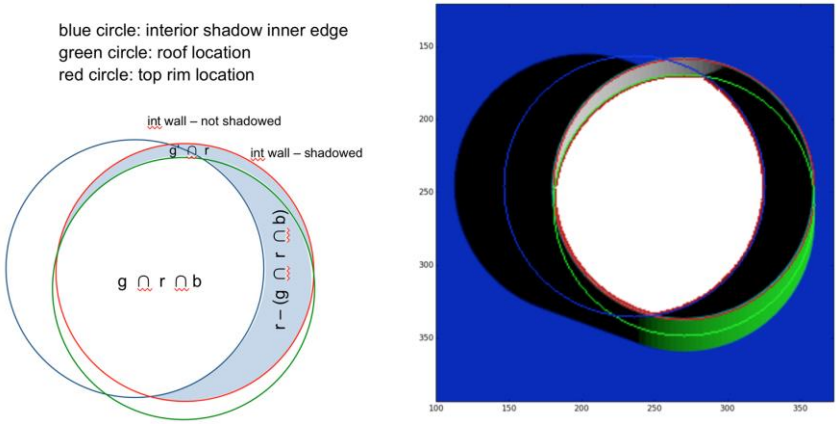
$$\Delta l_s = \frac{TH - RH}{\tan(sun\_elev)}$$

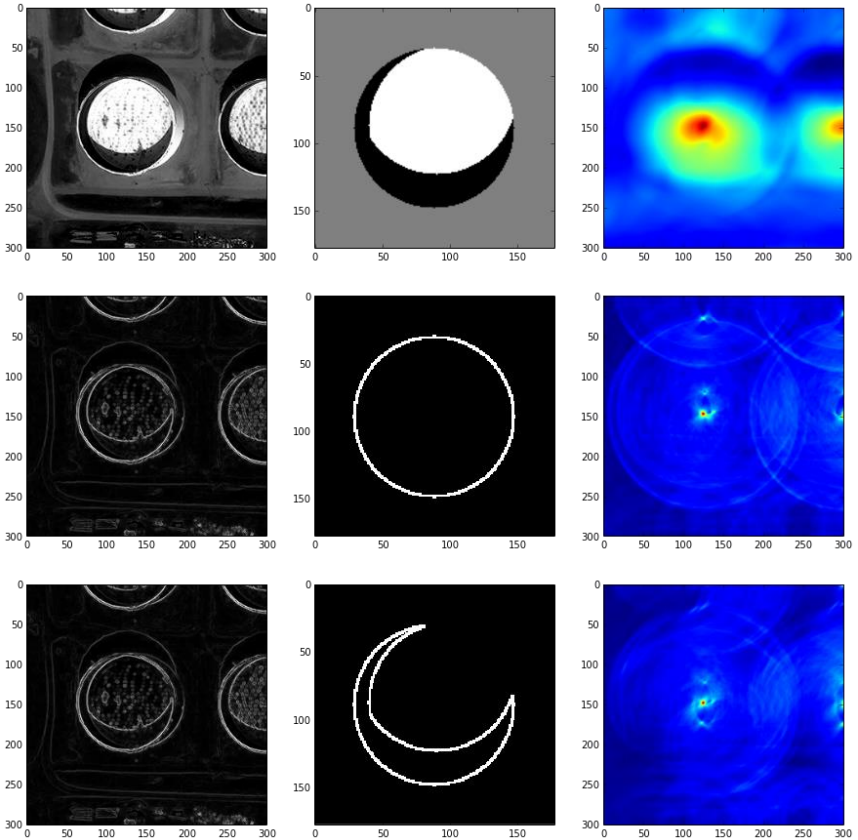
$$IS_x = BC_x + \Delta l_v \sin(sat\_az) - \Delta l_s \sin(sun\_az)$$

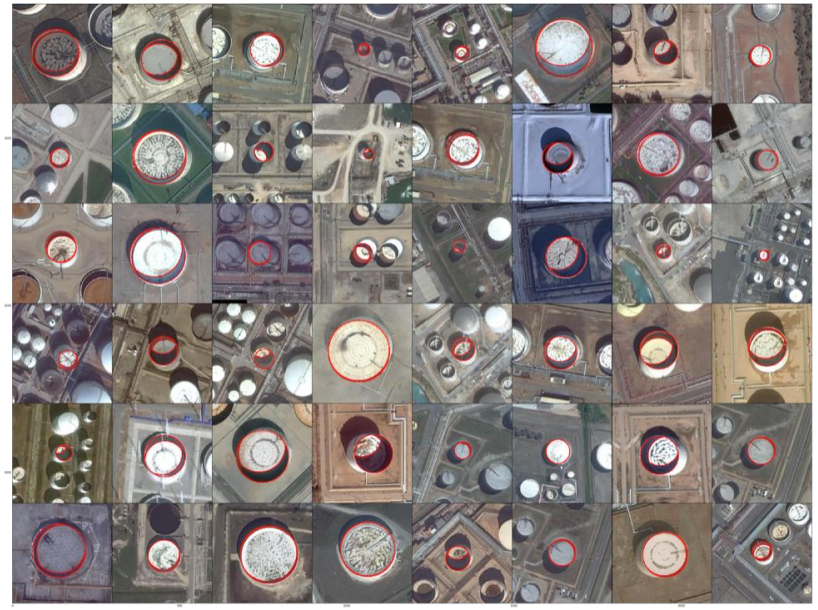
$$IS_y = BC_y + \Delta l_v \cos(sat\_az) + \Delta l_s \cos(sun\_az)$$

CV: TFF & Eyeball









Fill Percent marking

INSTRUCTIONS:
Some instructions

Tank outline

```
A, Z, a, z:   Change diameter
S, X, s, x:   Change height
D, C, d, c:   Change fill ratio
i,j,k,l:      Fine-tune position
], [:         Zoom in/out
SPACE, e:     Toggle center mark
w:            Toggle tank marks
Q:            Clear marks
click/ENTER:  Apply mark
```

Diameter:	Height:	Fill ratio:
36.7	18.2	0.46

Owner:

Content: oil

Description:

OK

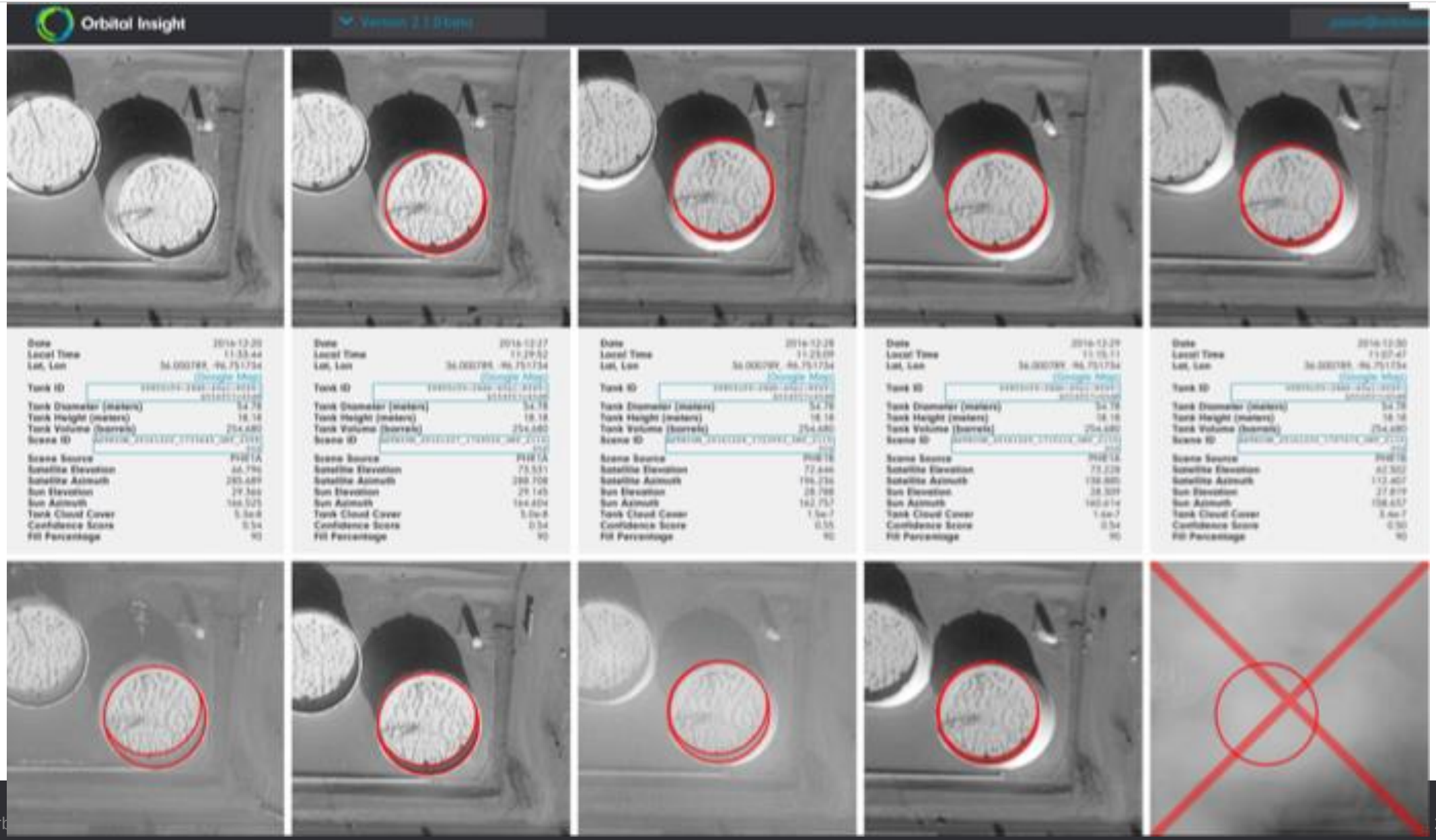
Not OK

Cannot see any tanks

Image is unclear, blurry

Other issues

Tank QA Tool



Orbital Insight

by the numbers...

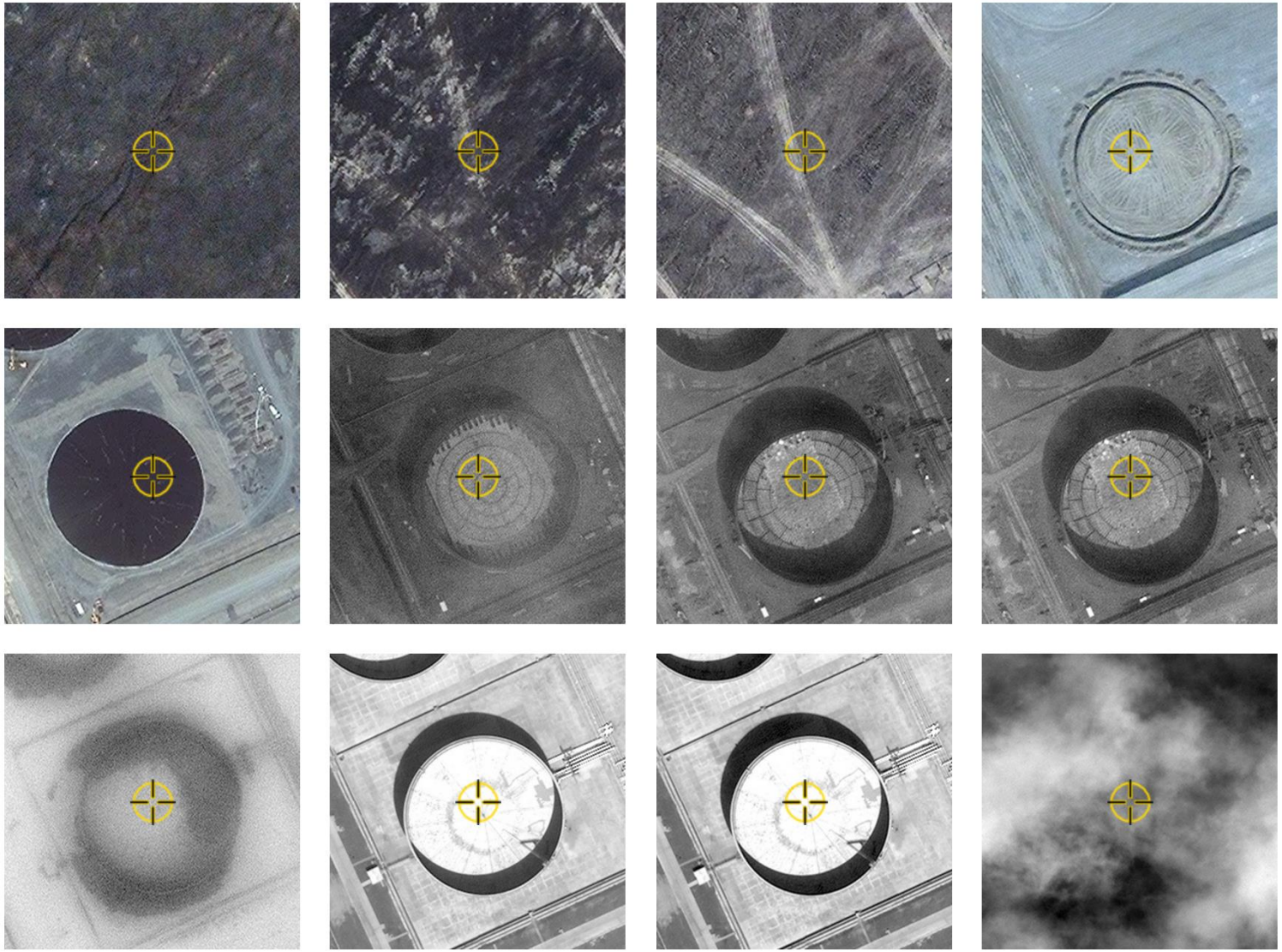
24,000	oil tanks
276,000	scenes ingested total
700	scenes ingested nightly
200	gigabytes of oil imagery
6,300,000	tank images
$1,200	monthly AWS compute cost
$10	monthly AWS storage cost
10+	press articles
1	patent pending

coming soon in oil...

Planet

SkySat

Oil Tank Lifecycle



Oil Product

Orbital Insight | Version 2.0.0 | jason@orbitalinsight.com

Region	Volume Stored	Fill Percentage	30-day Change	Tanks Observed	Latest Update
China	739.494 mb	64.6 %	-5.2 %	1472 of 2349 tanks	04-Oct-2017
OECD Americas	481.891 mb	53.8 %	-0.1 %	3606 of 6029 tanks	04-Oct-2017
OECD Asia	206.169 mb	60.9 %	-1.1 %	771 of 1428 tanks	04-Oct-2017
OECD Europe	535.093 mb	63.7 %	6.6 %	2043 of 3978 tanks	04-Oct-2017
OPEC	278.568 mb	49.3 %	-1.7 %	1207 of 2038 tanks	04-Oct-2017
US Total	416.170 mb	53.3 %	-2.8 %	3196 of 5424 tanks	04-Oct-2017
Cushing CUSHING	54.260 mb	66.9 %	1.3 %	311 of 311 tanks	04-Oct-2017
US PADD1 PADD1	26.551 mb	48.9 %	-16.7 %	254 of 450 tanks	04-Oct-2017
US PADD2 PADD2	119.302 mb	59.9 %	-3.7 %	797 of 1239 tanks	04-Oct-2017
US PADD3 PADD3	214.116 mb	51.9 %	-1.1 %	1554 of 2671 tanks	04-Oct-2017
US PADD4 PADD4	9.274 mb	50.7 %	6.7 %	152 of 273 tanks	04-Oct-2017
US PADD5 PADD5	46.844 mb	48.2 %	-0.2 %	439 of 791 tanks	04-Oct-2017
Other Regions					

Orbital Insight | Version 2.0.0 | jason@orbitalinsight.com

US PADD3

PADD3

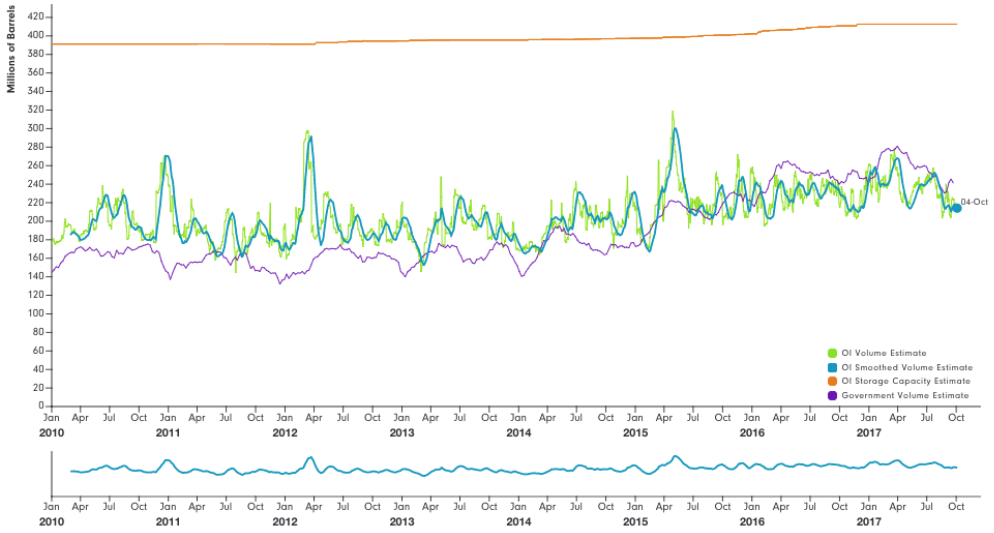
04-Oct-2017 Latest Update

214.116 mb Crude Oil Stored

-1.1% 30-Day Change

51.9% of Capacity

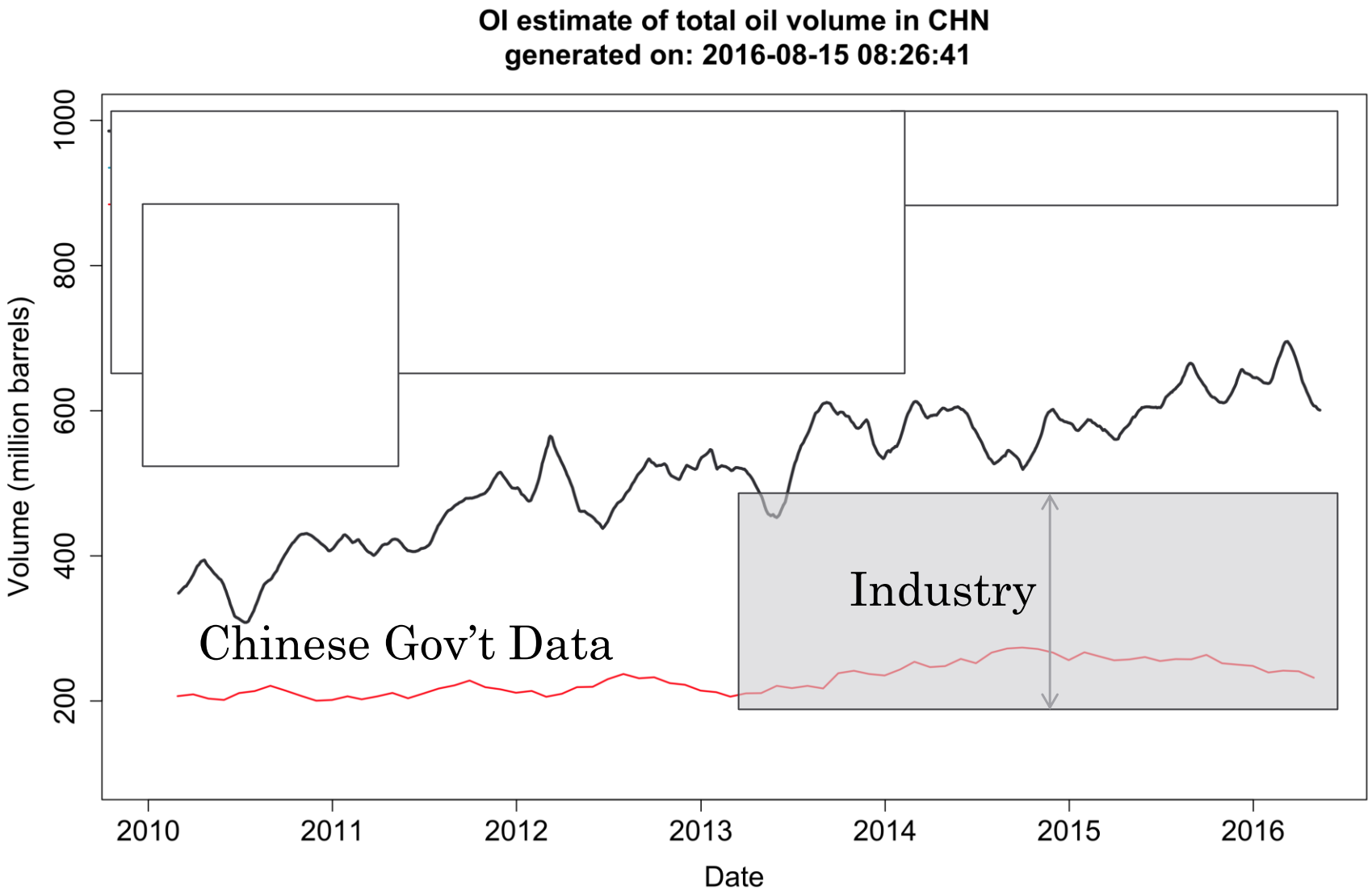
1554 of 2671 Tanks Observed



China data

Take-aways:

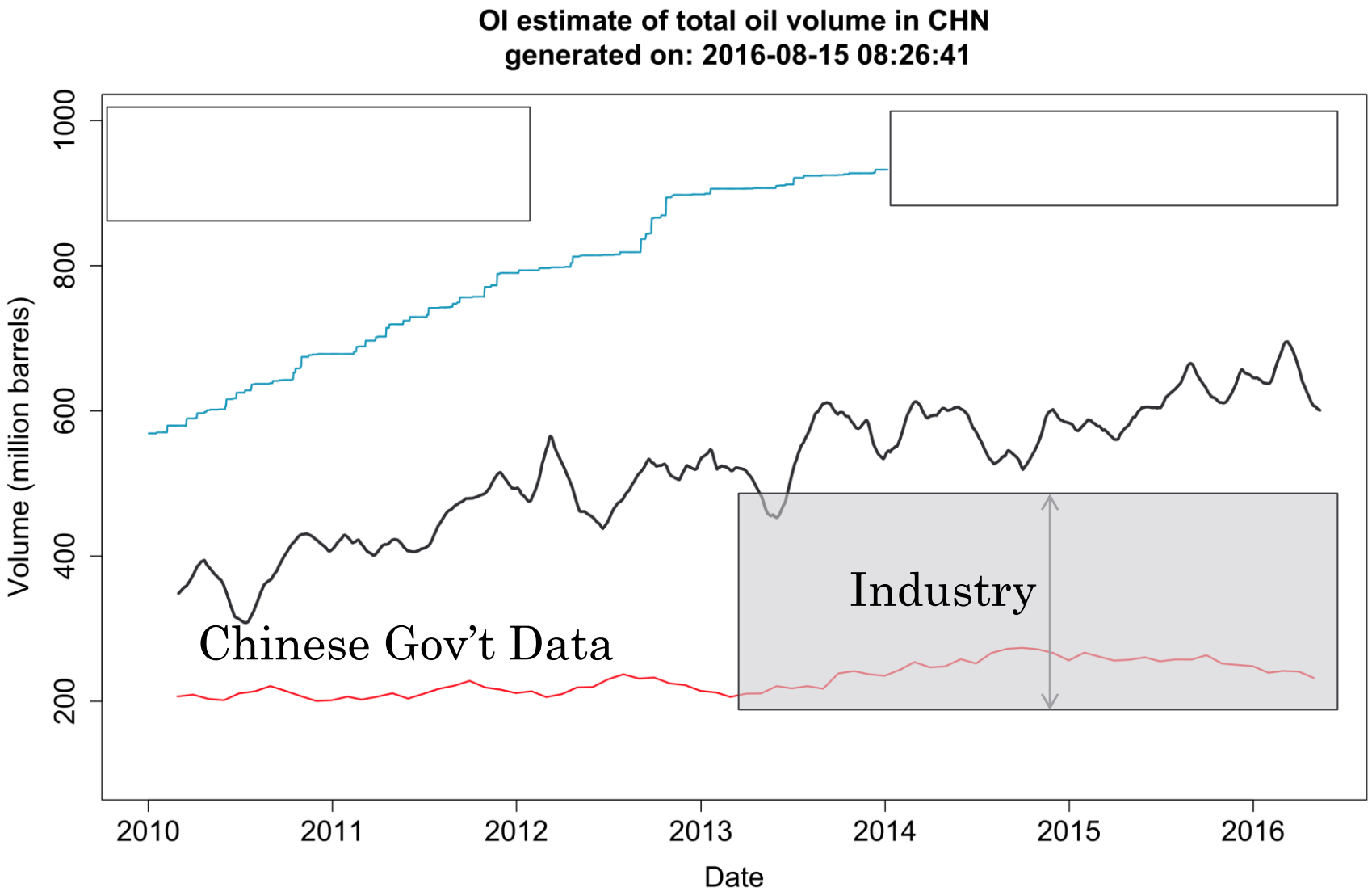
- Chinese gov't data wrong
- we show 500-600M BBL
- industry says 200-500M
- industry underestimating
- QA processes ongoing:
  - crude vs non-crude
  - adding new tanks
  - improved processes



China data

Take-aways:

- Chinese gov't data wrong
- we show 500-600M BBL
- industry says 200-500M
- industry underestimating
- QA processes ongoing:
  - crude vs non-crude
  - adding new tanks
  - improved processes



Let’s get started!

Assignment 1: Mbappe's Restaurant

- Kylian Mpabbe has quit soccer and decided to open up a restaurant! Turns out he's also a world class chef and can make any type of cuisine (French, Chinese, Palestinian, you name it). He doesn't know what kind of restaurant he wants to start, or in which city (but he doesn't want to leave Europe). He's hired you to use Trip Advisor ratings to help him understand the restaurant industry and ultimately make this decision.

- Before we look at the dataset, what would be some information that you would like to have?

Your Job

- Come up with **one** question that could help Mbappe. Here are some ideas, but note **these are mine**:

- Which cities have the most $$$$ restaurants?
- What words do people most commonly use to describe their experiences at expensive restaurants? Are these different than the words people use to describe cheap restaurants?
- What fraction of restaurants in each city has the word ‘Café’ in it.

Day 2: Intro to Machine Learning

- The Problem:
- You are in charge of collecting taxes, and you want to know if someone is likely to cheat on their taxes or not. You have a bunch of old examples with someone's **marital status**, **income**, whether they got a **tax refund**, and **whether or not they cheated** on their taxes. Your job is to **build a model** that uses someone's profile to predict whether or not they will cheat.
- Thoughts?

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data

Decision Trees

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data

Model: Decision Tree

ID3 Algorithm:

- 1.) Choose an attribute that best differentiates the instances contained in dataset
- 2.) Create a tree node whose value is the chosen attribute
- 3.) Create child links from this node where each link represents a unique value for the chosen attribute
- Repeat!

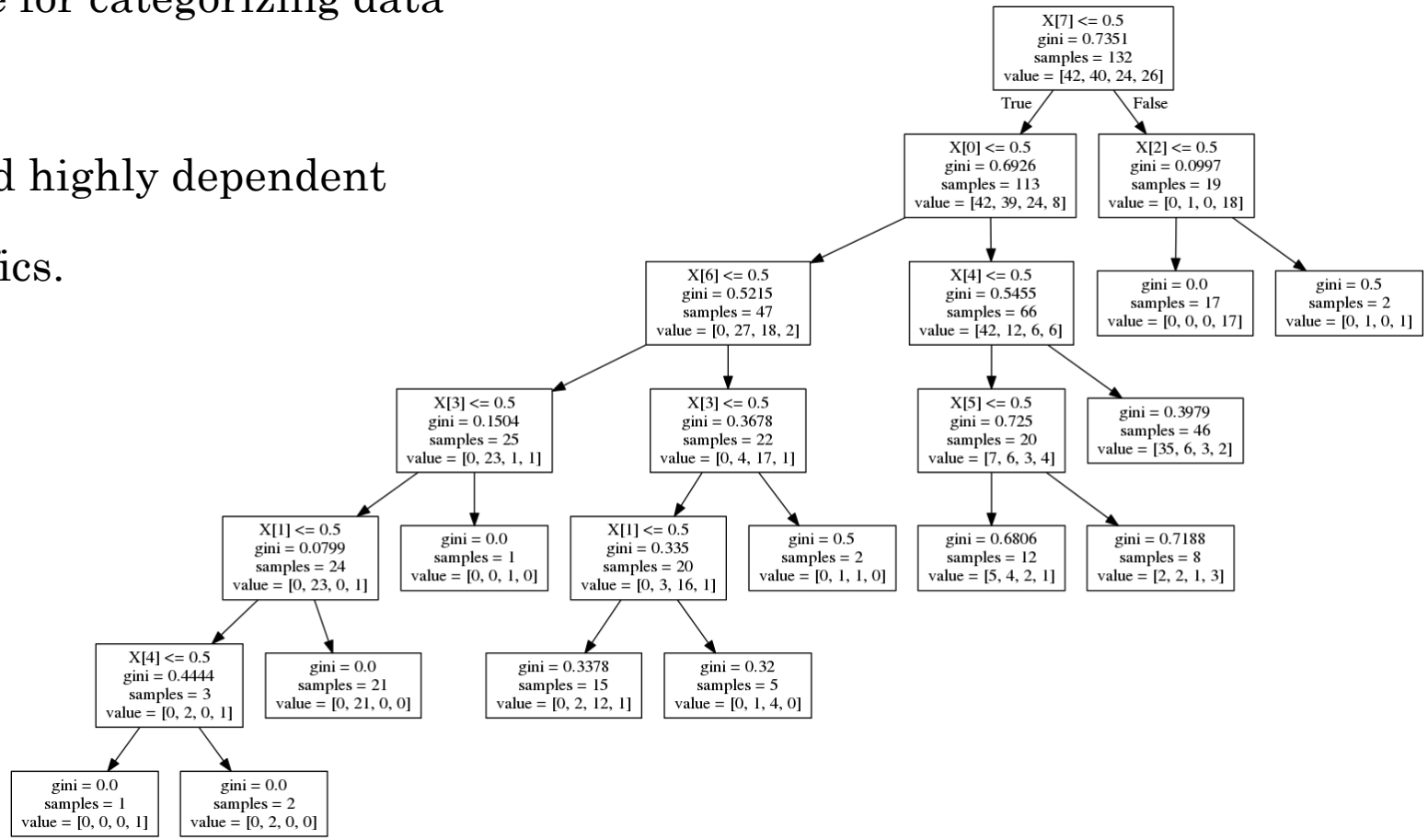
Titanic Dataset!

```
import pandas as pd
import numpy as np
import re
import matplotlib.pyplot as plt
from sklearn import tree
%matplotlib inline
```

from sklearn.cross_validation import Kfold
from sklearn.model_selection import train_test_split

Decision Tree

- Pro: Is the best possible tree for categorizing data given a training set
- Con: Overly complicated and highly dependent on training set specifics.



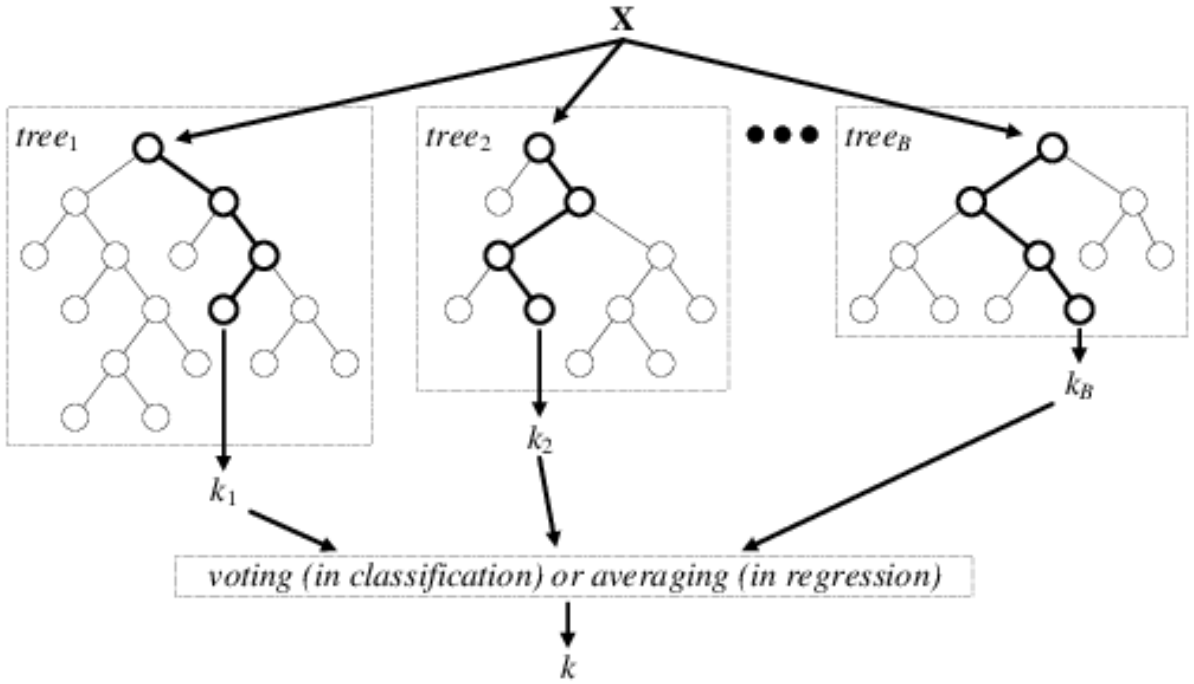
Any Ideas to Fix this Problem?

- Think for 5 minutes with a partner to try to come up with a way to fix this algorithm so that it **is less likely to overfit the data.**

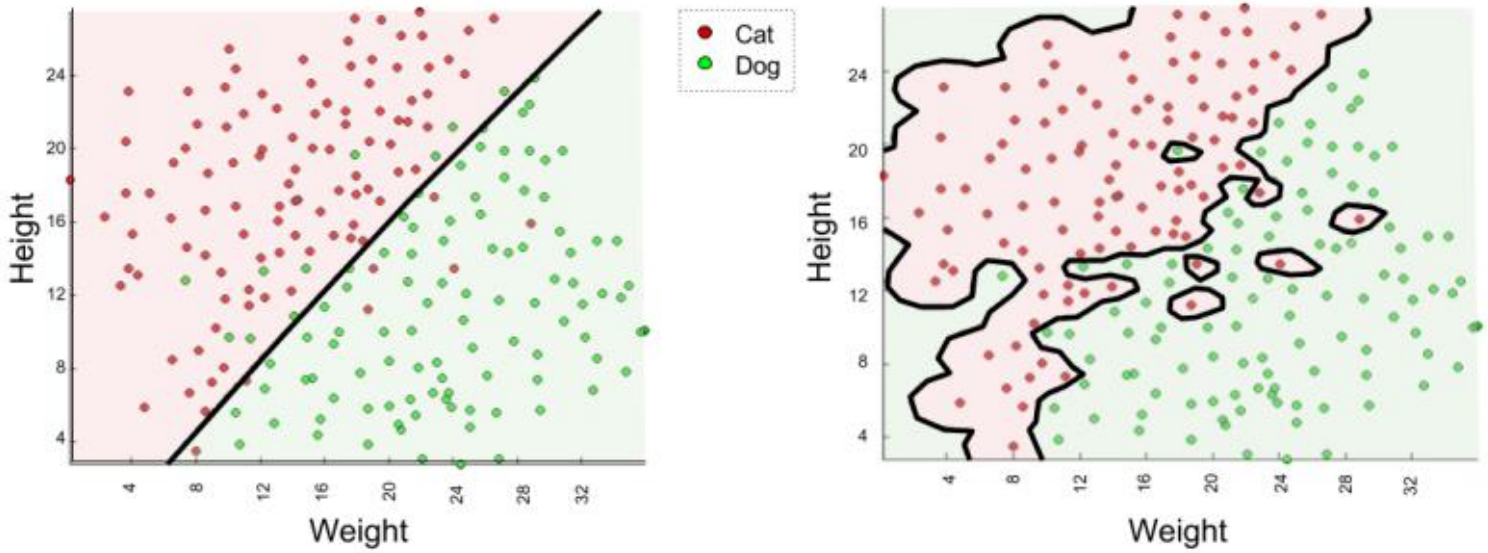
Random Forest

Basic Idea: Split the Dataset up into groups and make a smaller tree for each group. Then, the categorizat
is given by **majority** vote for all the trees

Pros: Less likely to overfit to data
Cons: Nothing too obvious, but it is very simple. Any ideas to improve it?



Bias Vs. Variance



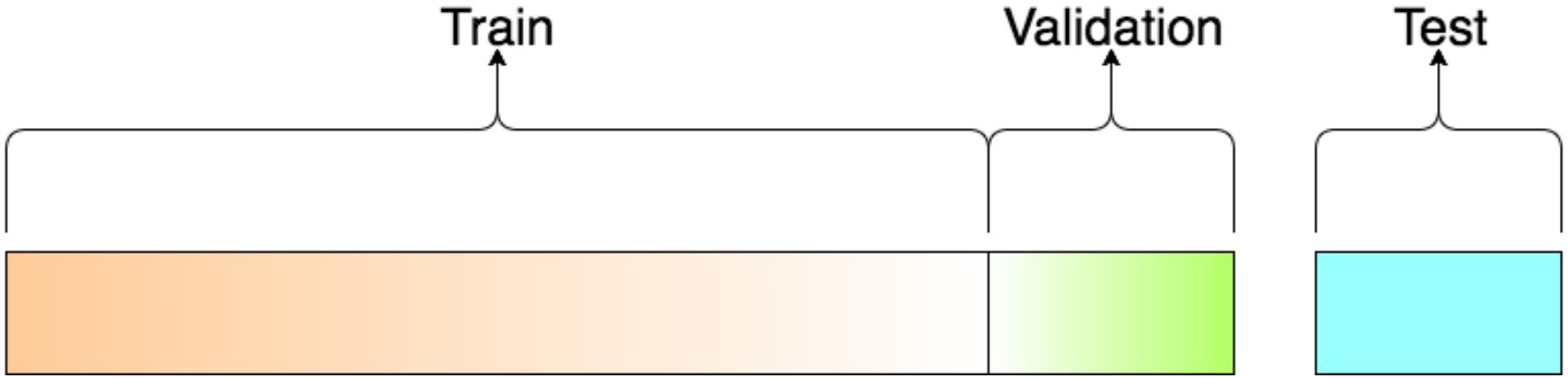
Left: High Bias and Low Variance

- Pros: Will Train the Same way on different samples of dataset
- Cons: Will Always Classify some points incorrectly

Right: Low Bias and High Variance

- Pros: Able to capture very complex relationships in data
- Cons: Will Train differently on different samples and is likely **overfit**

Cross Validation



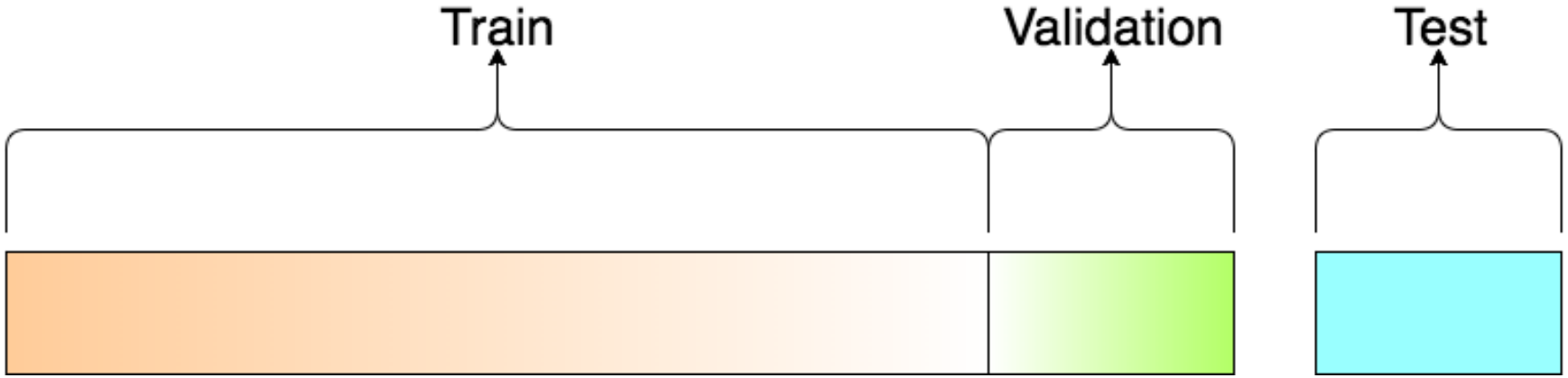
Basic Idea: Split Data into three groups:

Train: Used for training any model. Where the heart of Machine Learning Algorithms do their magic

Validation: Used to understand how good any given model is and to compare different models

Test: Used when you are done with **everything** (training models as well as choosing the best model) to see how well you've done.

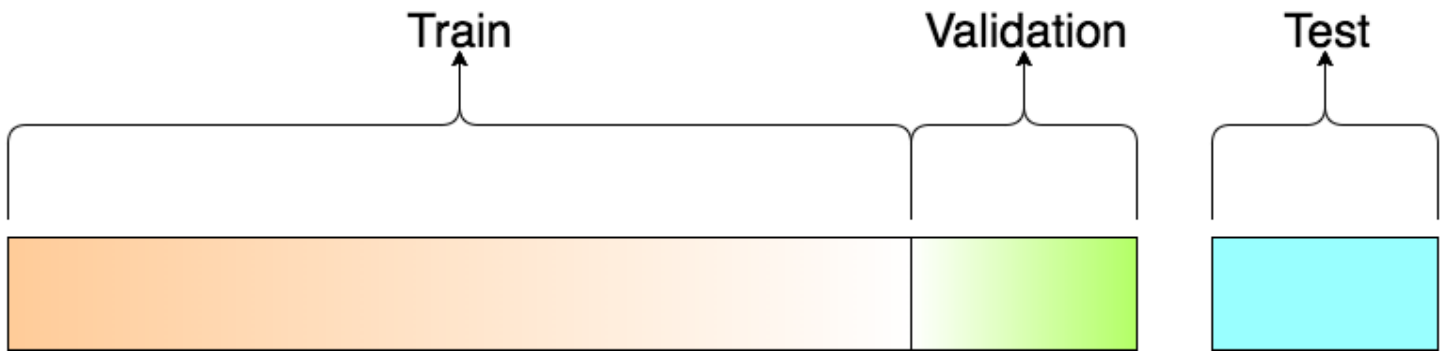
Cross Validation Part 2: Questions



In your own words answer:

1.) Why is it necessary to have a separate training set and validation set?
2.) Why is it necessary to have a separate validation set and test set?
3.) Can you think of a downside of this method?

Cross Validation Part 3: Answers



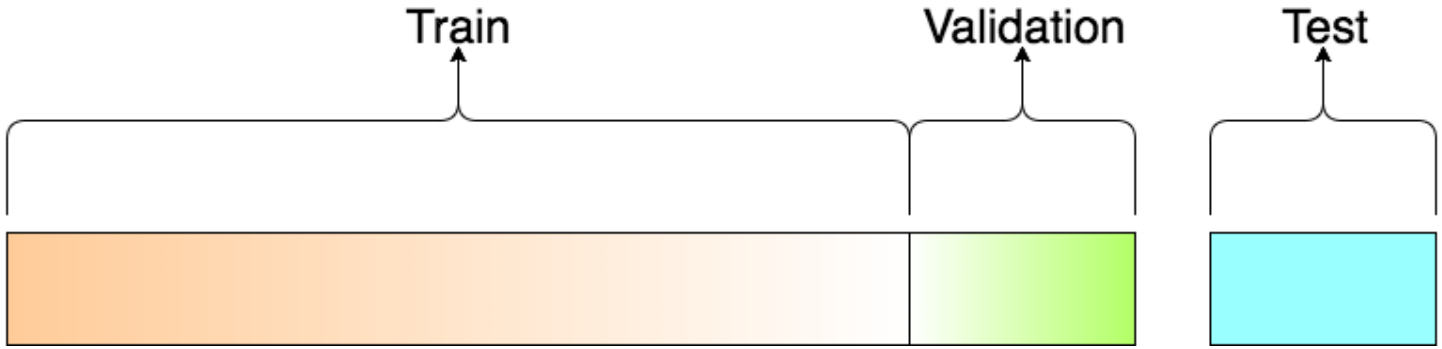
In your own words answer:

1.) **Q:** Why is it necessary to have a separate training set and validation set?
A: If we only look at accuracy on the test set, we will choose very **overfit** models

2.) **Q:** Why is it necessary to have a separate validation set and test set?
A: We need to validation set to compare different models. This way we know how good our final model is. If we only used the validation set, we wouldn't know if the model we chose was just really good with that specific dataset.

3.) **Q:** Can you think of a downside of this method?
A: 1.) We lose a lot of data that we'd like for training
2.) Our choices might be influenced by peculiarities of the

Cross Validation Part 4:

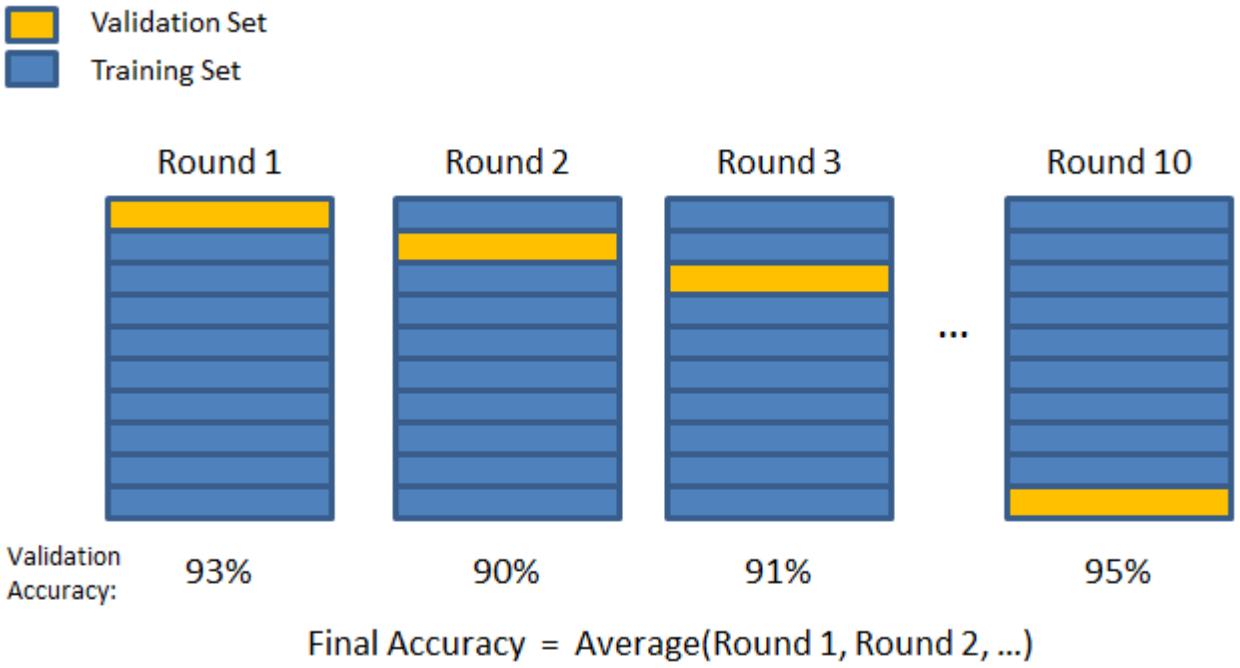


- Problem: Our training and model choice might be influenced by peculiarities of the **train** and **validation** set.
- Solution: **???**

KFold Cross Validation

Basic Idea:

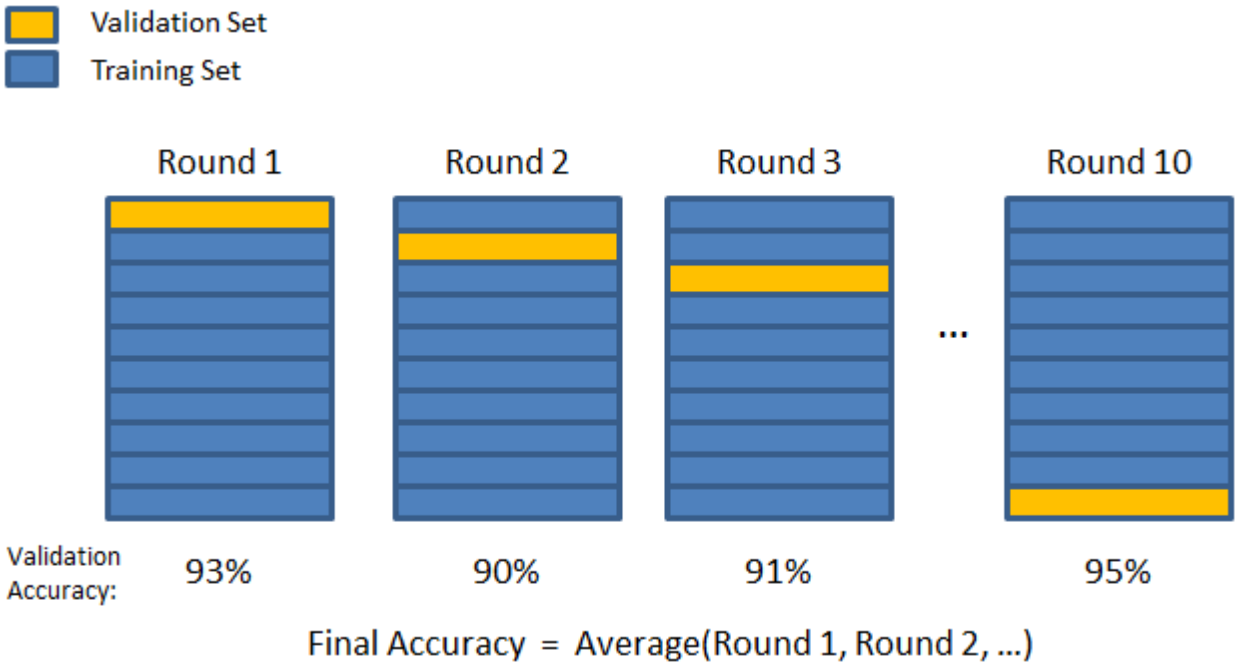
1.) Split data into K (let's say 10) groups
2.) Hold out one group and train on the rest
3.) Repeat for all 10 different groups
4.) The final score is the **average** of all these rounds



KFold Cross Validation Part 2

Questions:

1.) What is something **good** about Kfold?
2.) What is something **bad** about Kfold?



KFold Cross Validation Part 3

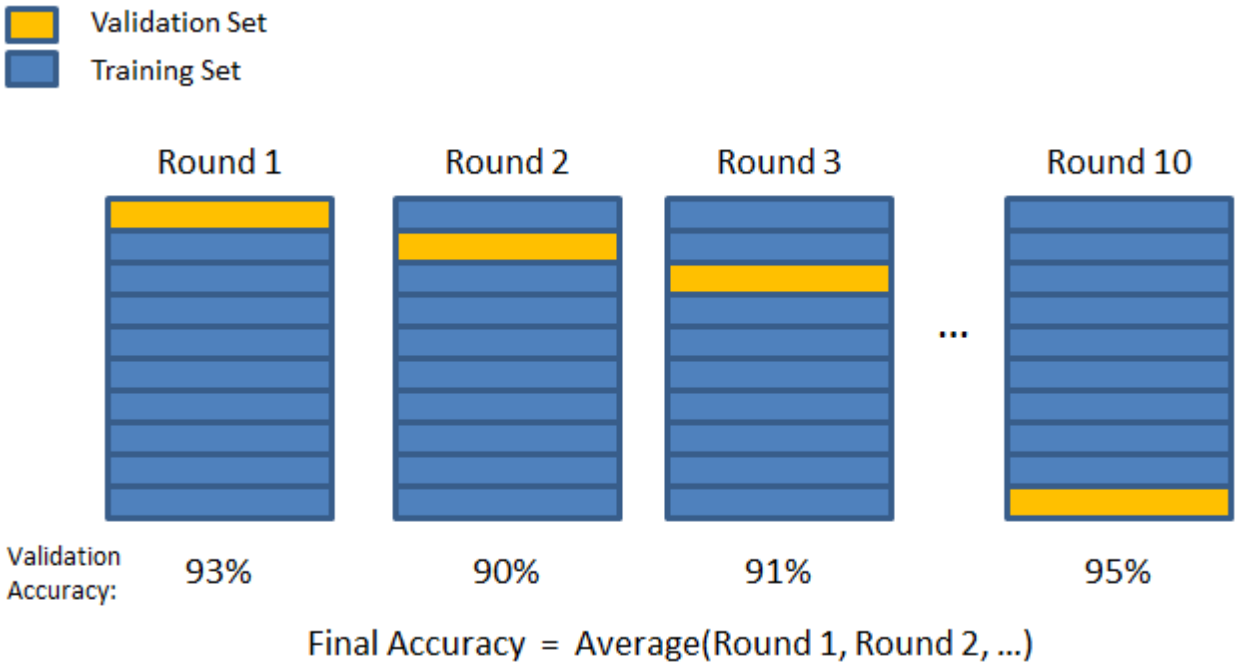
Questions:

1.) **Q:** What is something **good** about Kfold?

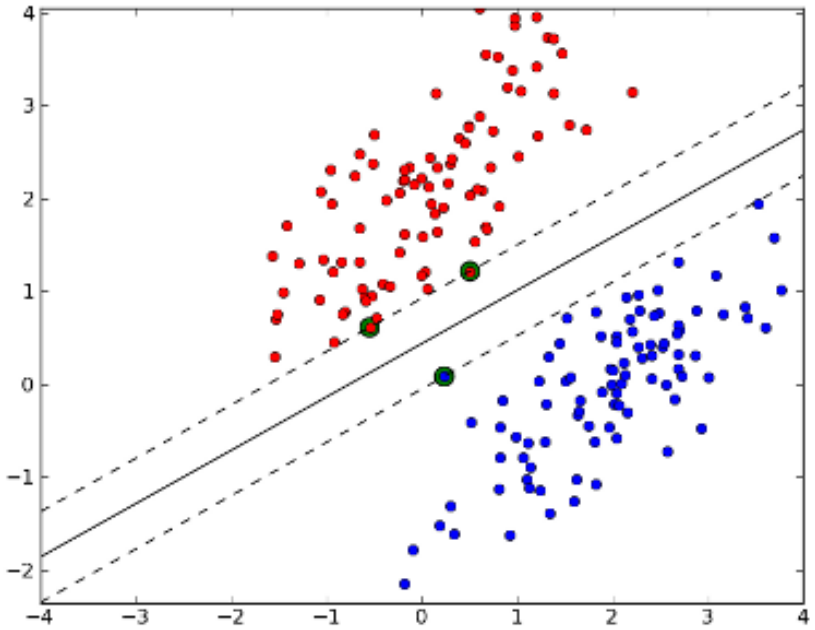
A: It treats all the data the same and isn't prone to choosing an unrepresentative validation set.

2.) **Q:** What is something **bad** about Kfold?

A: 1.) It makes training take way longer
2.) It doesn't take into account relationships between samples (i.e. timeseries)
3.) Not obvious how to choose

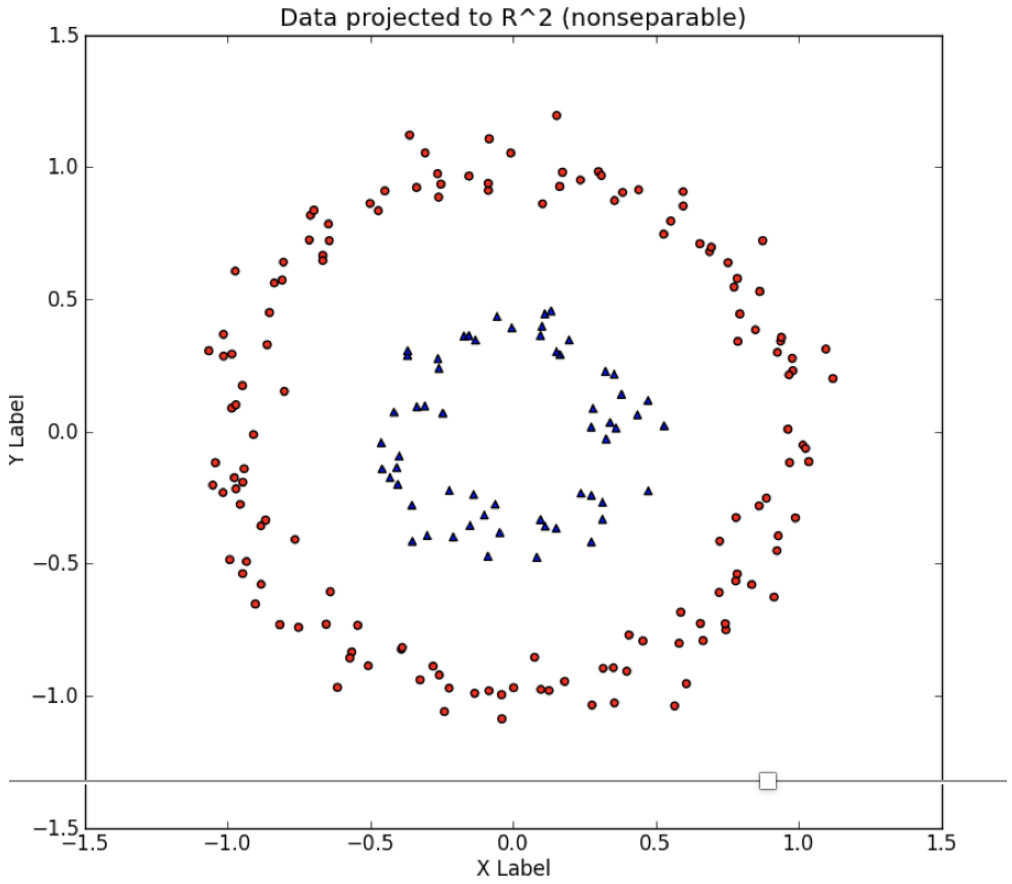


Other Techniques: SVM (Support Vector Macine)



SVM: Continued

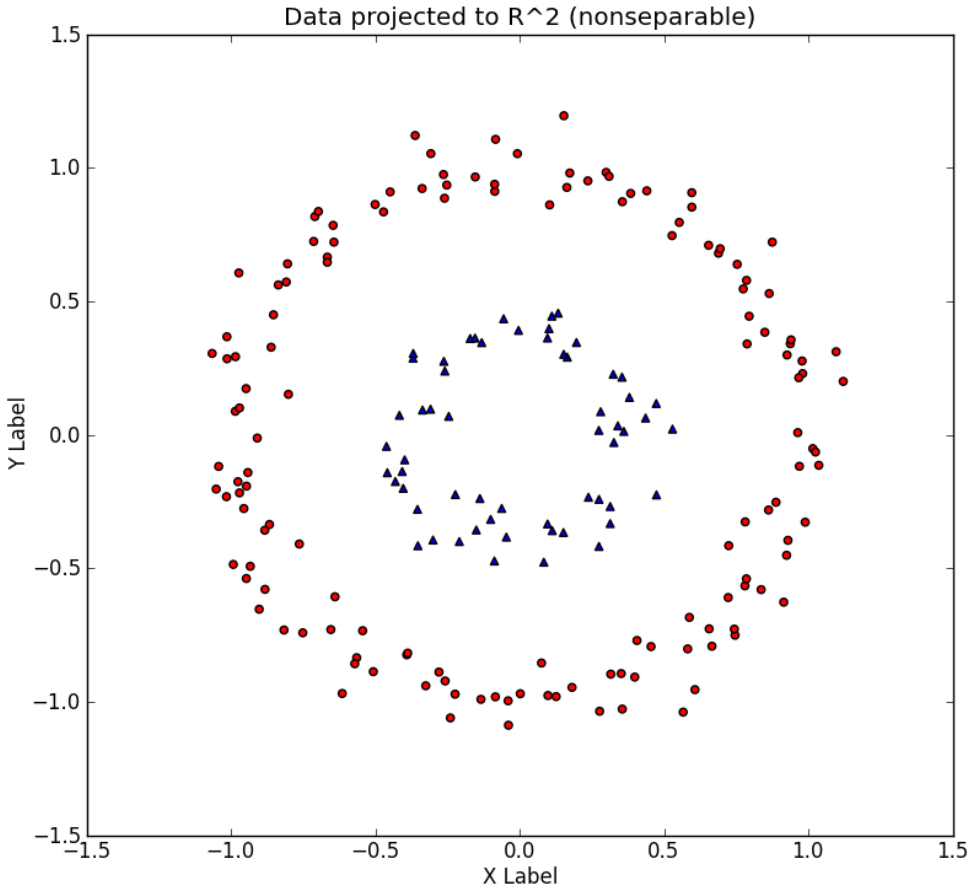
- Question: What happens if we can't draw a straight line through the points?

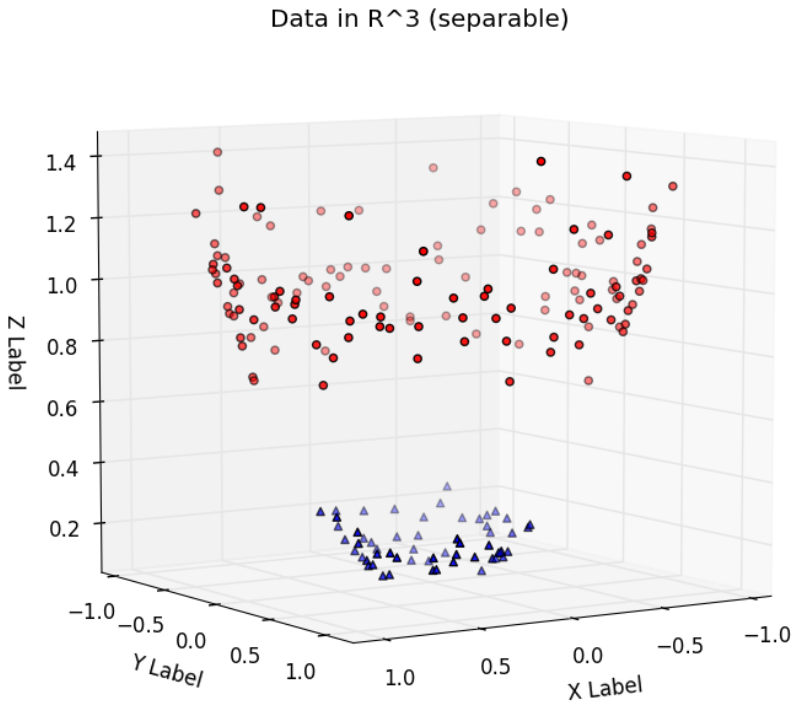


Non-separable Data Points

- Hint: Think in more dimensions

Non-separable Data Points





Linear Regression

Homework day 1:

Play around with **https://www.autodraw.com/**

Give me your best guess: come up with any ideas of how you think this works? What does it have to do with data? Note – any ideas will do! This is a hard question

Homework day 2:

- I mentioned that Random Forest is very possible, but very simple. One very popular improvement is **AdaBoost** and the distinction between **bagging** and **boosting.** Your job is to do research on these methods and come to class with a paragraph written on:

- 1.) How does Adaboost work? Why is it an improvement over Random Forest?
- 2.) What is the difference between bagging and boosting?

HW 3: Practice Reading through Kernels

- Go through the Kernels on the Titanic Dataset. Find something interesting someone else did that you can share with the class.