DATA

IN THE NATIONS

WHAT WE'LL COVER

DATA SCIENCE

HANDS ON (NATIONS DATA)

THE NEED OF STATS

TOOLS AND WHAT TO DO NEXT

STUFF TO KEEP IN MIND

WHAT WE'LL COVER

DATA SCIENCE

HANDS ON (NATIONS DATA)

THE NEED OF STATS

TOOLS AND WHAT TO DO NEXT

STUFF TO KEEP IN MIND

WHAT WE WILL NOT COVER

HOW TO DEPLOY YOUR RESULTS

PYTHON SPECIFICS

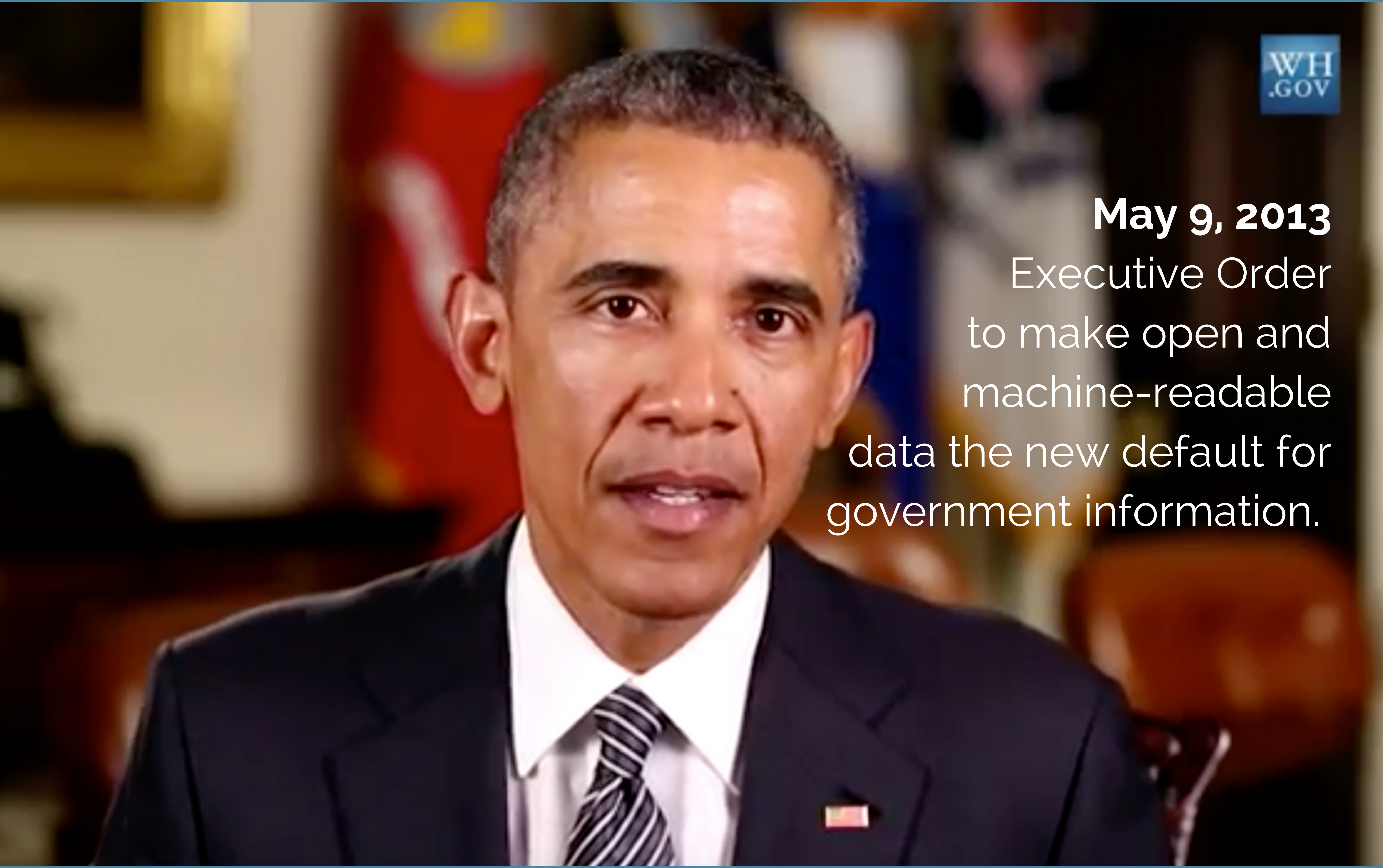
DETAILED MACHINE LEARNING ALGORITHMS

HOW TO PLOT, USE DATAFRAMES AND OTHER SPECIFIC STUFF

WHAT'S ALL THE DATA FUSS ABOUT...

THE RISE OF THE DATA SCIENTIST

- Ranked #1 job on Glassdoor.
- Data driven environments.
- Machine readable data.
- Actual results.



May 9, 2013

Executive Order
to make open and
machine-readable
data the new default for
government information.

DATA SCIENTIST

> A Data Analyst who lives in California

> A data scientist
> is someone who is better at statistics than any
> software engineer
> and better at software engineering than any
> statistician.

Asking questions about data...

is what Data Scientists do

Classification

Is a patient sick?

Regression

How much does this cost?

Clustering

How are these people organized?

Asking questions about data...

is what Data Scientists do

Classification

Is a patient sick?

Regression

How much does this cost?

Clustering

How are these people organized?

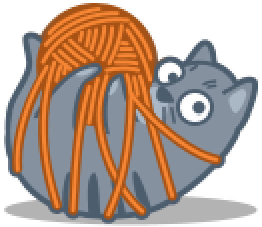
Anomaly Detection

Something fishy?

Reinforcement Learning

What can my robot do next?

LET'S GET OUR HANDS DIRTY...



MAKE SURE

- Have Python installed or whatever you're familiar with.
- Have the data with you
  (clone the github repo you received by email:
  chateaudemachin/datastuff)

LET'S GET OUR HANDS DIRTY...

Uplands Nation

Facebook Event

Attendance

foo bar;

UPPSALA DATAVETARE

Most statistics are wrong, including this one...

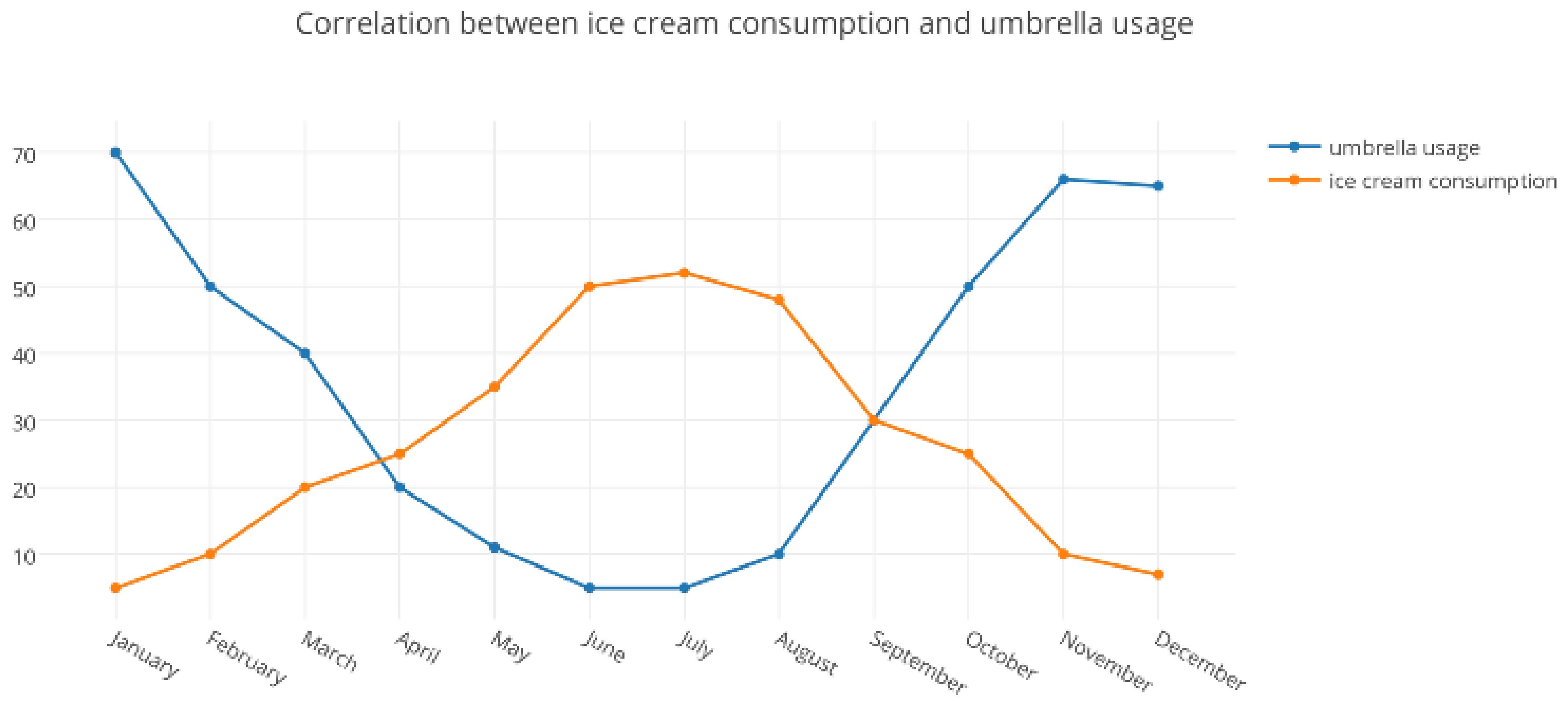
- Master the basics.
- Rely on the rest of your team.
- If you don't, you're going to mess up, sooner or later...

Most statistics are wrong, including this one...

- Master the basics.
- Rely on the rest of your team.
- If you don't, you're going to mess up, sooner or later...

Let's talk about ice-cream...

Correlation Vs Causality



winter

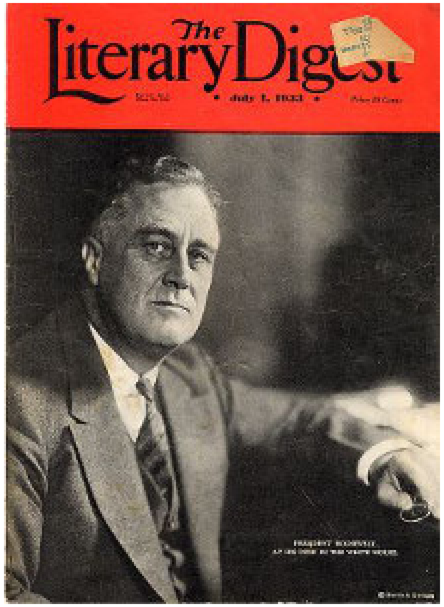
summer

winter

Convenience Sets

Think again
about how you
collected your
data!

BIG
DATA
WILL SOLVE
EVERYTHING?



poll of 10 Million people (2.4 Million back)

41

55

THE

US ELECTIONS

BIG DATA WILL SOLVE EVERYTHING?

Real results

Roosevelt wins

61

37

THE

US ELECTIONS

BIG DATA WILL SOLVE EVERYTHING?

Georges Gallup

Random Sampling 50.000

66

THE

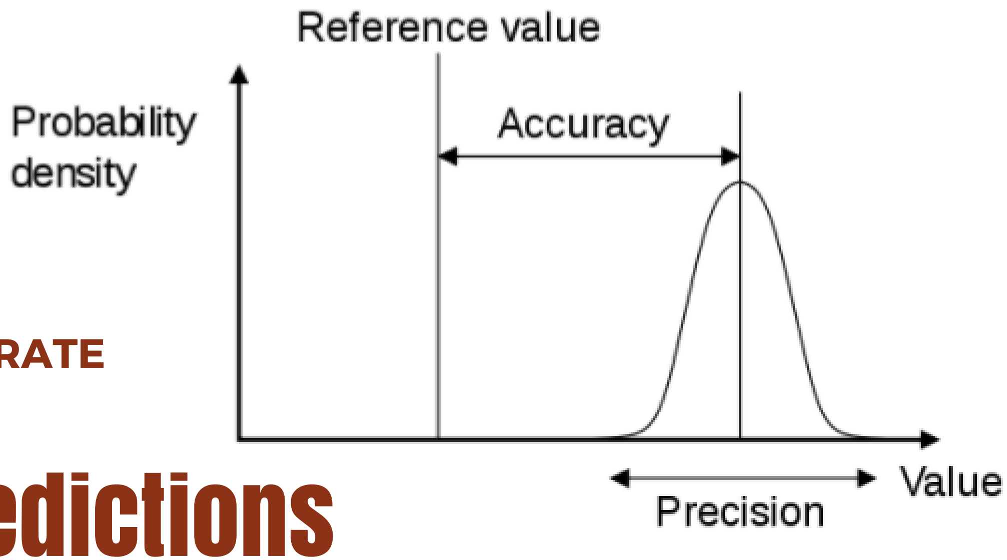
US ELECTIONS

WHY?

$$\operatorname{Bias}_\theta[\,\hat{\theta}\,] = \operatorname{E}_\theta[\,\hat{\theta}\,] - \theta = \operatorname{E}_\theta[\,\hat{\theta} - \theta\,]$$

WHY?

$$\mathrm{Bias}_\theta[\,\hat{\theta}\,] = \mathrm{E}_\theta[\,\hat{\theta}\,] - \theta = \mathrm{E}_\theta[\,\hat{\theta} - \theta\,]$$



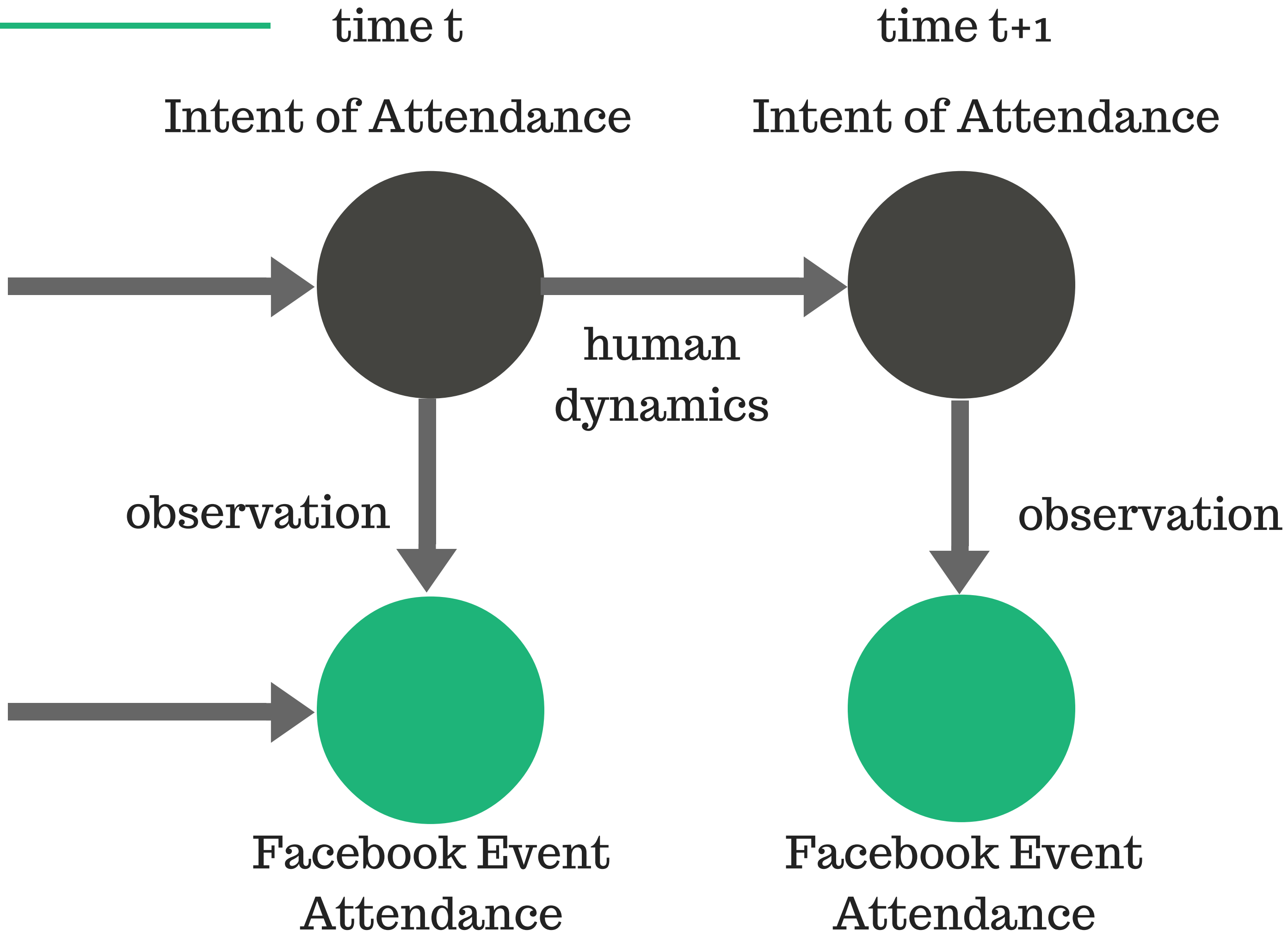
INCREDIBLY ACCURATE

Wrong Predictions

HAVE YOU HEARD OF

BIAS

DATA MODELS

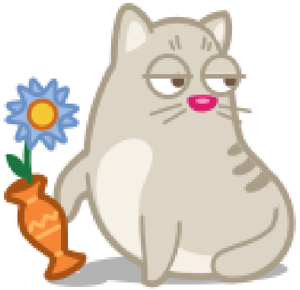


NATION EVENT

ATTENDANCE

GO PUBLIC

MESS UP
REALLY!



WHAT KIND OF CODE IS BEST?
THE KIND THAT YOU DEPLOY CONTINUOUSLY
memegenerator.net

MESS UP REALLY!

KEEP IN MIND

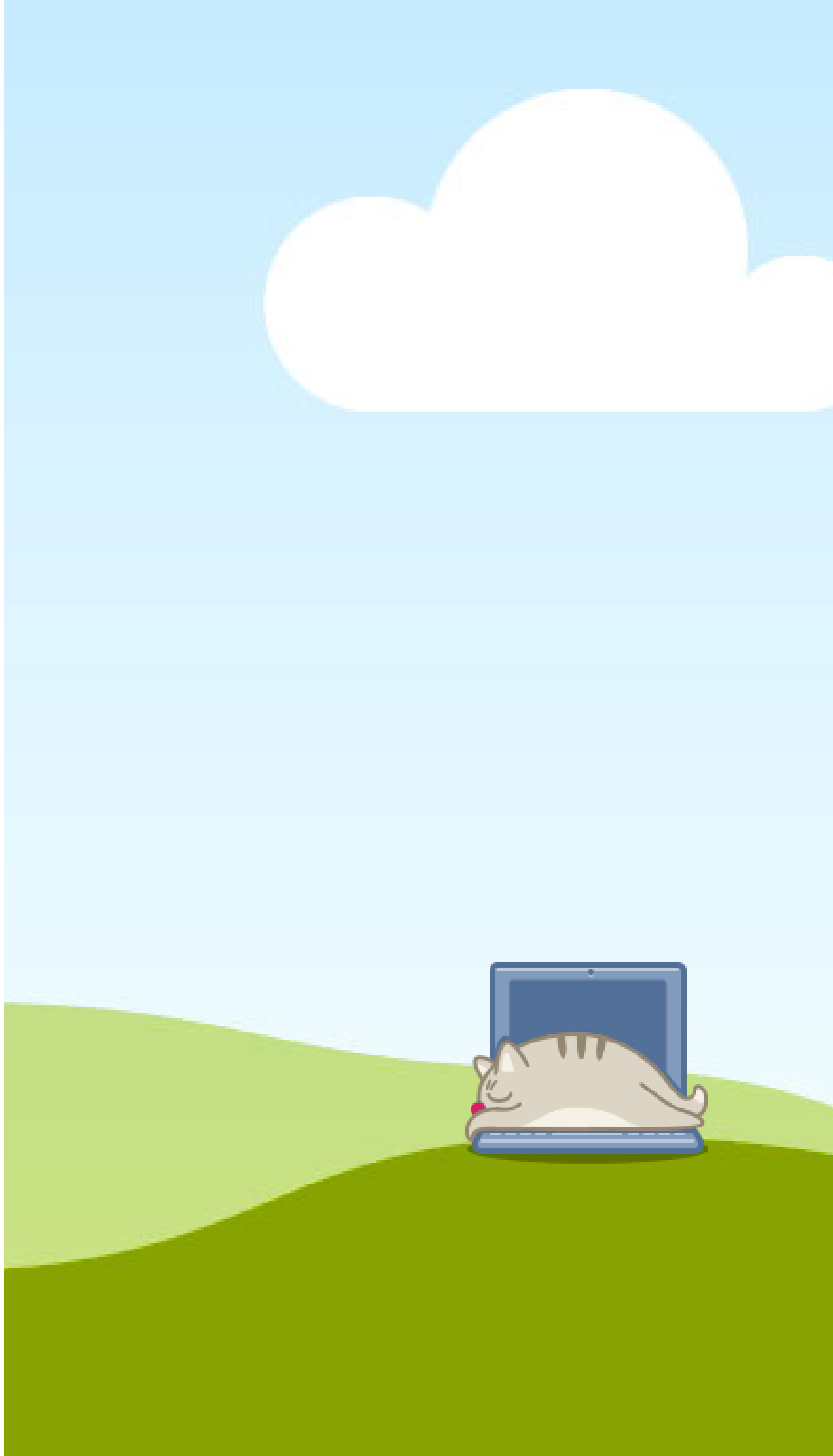
PRACTICE

ASK THE RIGHT QUESTIONS

Write Ugly Code

Use Other People's Stuff

Iterate



TOOLS

- Pandas, Scikit-learn, Numpy, Scipy.
- Requests, Django, Flask.
- Plotly, matplotlib.
- Jupyter, IPython notebooks.
- Coursera, udacity, stackoverflow, have you heard of Google? :P

Thank you.

AND, DON'T FORGET TO SIGN UP.