NTNU

AIA SCIENCE
ANALYTICS OF THINGS

Data Science: Industry Challenges and Expectations

Thiago G. Martins, PhD

Principal Data Scientist @ AIA Science
Associate Professor II @ NTNU
Trondheim, Oct. 2017

About me

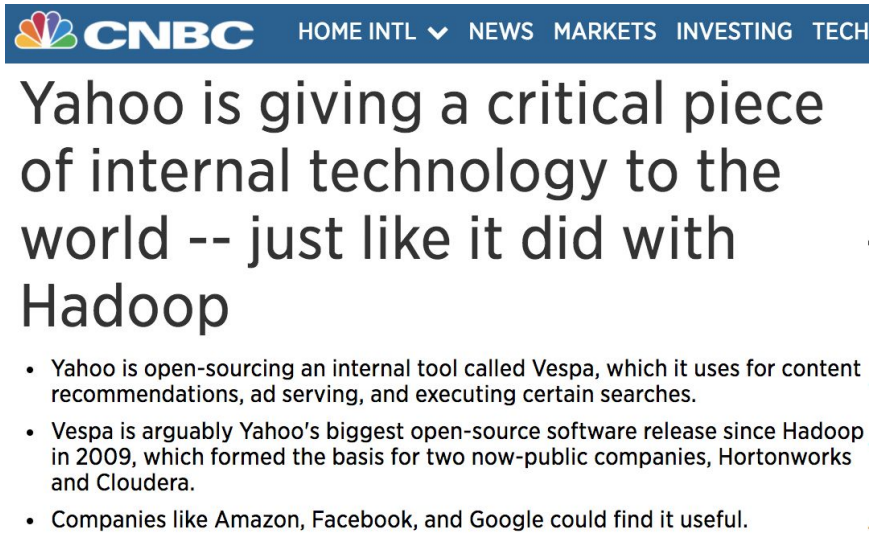
- Started in Statistics before it was “sexy”
- Stat PhD from NTNU
- Previously Data Scientist at Yahoo!
- Now Principal Data Scientist at AIA Science
- Part-time Associate Professor at NTNU

Stat PhD @ NTNU

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int \pi(\theta)f(y|\theta)d\theta}$$

- Practical challenges
  - Prior specification
  - Computation of the normalizing constant
- Focus
  - How to properly design complex models
  - How to approximate their posteriors fast enough

Yahoo!



- Huge volumes of data
- Big Data technology
- Batch training
- Serving predictions requirements
- Integrating models with applications
- Scalability before model design

AIA Science

- In general, sub-optimal solutions used.
  - Scientific
  - Engineering

- First employee in TRD.
- Data - Valuable problems - Solution that works in production.

- Many challenges ahead.
  - Technical
  - Business model

- **Lack of qualified professionals**
  - Position @ NTNU

Basics

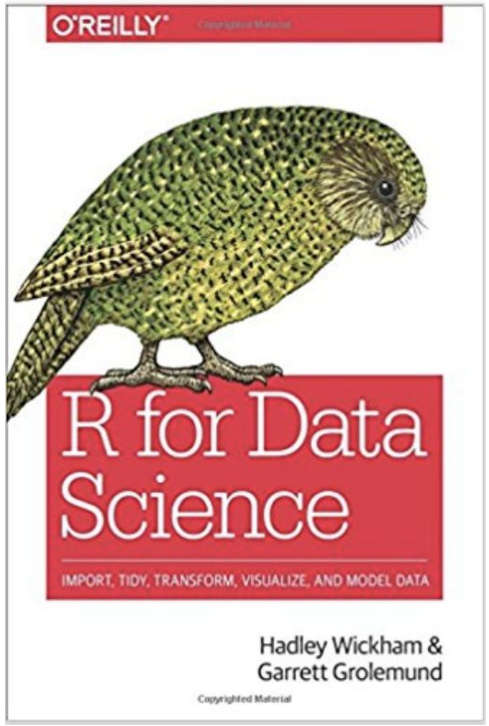
Basic CS Knowledge

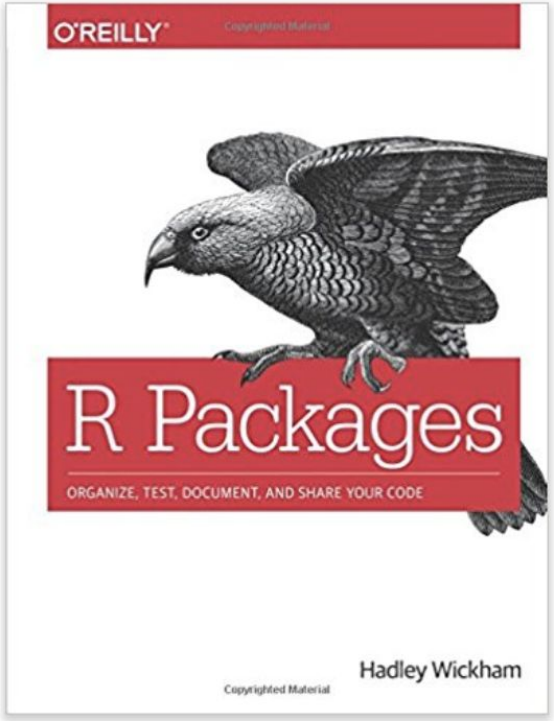
- Broad spectrum of professionals (pure CS - pure Stat)
- Diversity is extremely important. **But this talk focus on Statisticians.**
- Better code organization
  - Library/packages
  - Unit tests
  - Version control (git - Github/Bitbucket)
- Priority to R and Python

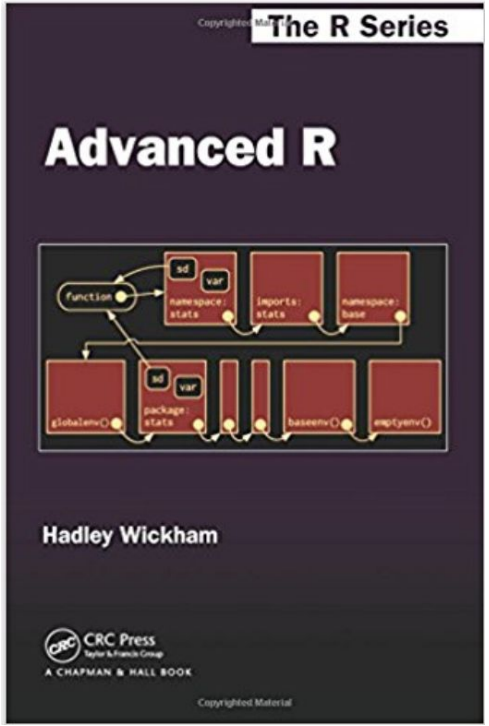
R

- Exploration and visualization
- Hadley Wickham and Tidyverse
- Tidyverse: collection of R packages designed for data science
  - tidyr: organize data
  - dplyr: data manipulation (filter, select, summarise)
  - ggplot2: grammar of graphics
- RStudio
- R notebook

R resources


O'REILLY
R for Data Science
IMPORT, TIDY, TRANSFORM, VISUALIZE, AND MODEL DATA
Hadley Wickham &
Garrett Grolemund

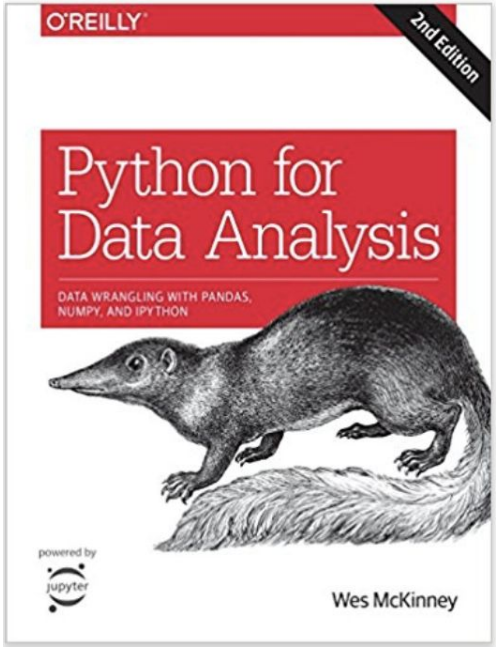

O'REILLY
R Packages
ORGANIZE, TEST, DOCUMENT, AND SHARE YOUR CODE
Hadley Wickham

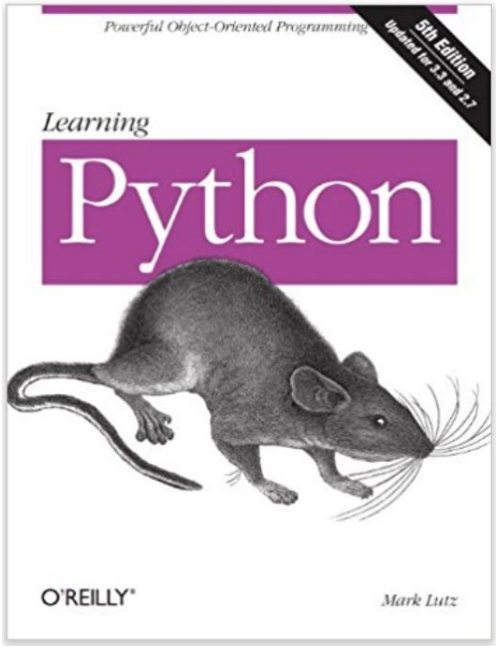

The R Series
Advanced R
Hadley Wickham
CRC Press
A CHAPMAN & HALL BOOK

Python

- General-purpose programming language
  - Widely used in industry
- Most interesting open-source libraries have Python APIs
  - TensorFlow
  - Spark
  - +++
- Scientific Computing/Data Science with Python
  - Numpy (N-dimensional array, linear algebra, rng)
  - Pandas (data frame and data manipulation functionality)
  - Matplotlib (graphics)
- I use PyCharm as IDE
- Jupyter notebook

Python Resources





Data Storage, Data Exchange and APIs

- Databases
  - Relational databases
  - NoSQL
- Data Exchange
  - JSON
  - XML
  - +++
- Webserver
- Application Programming Interface

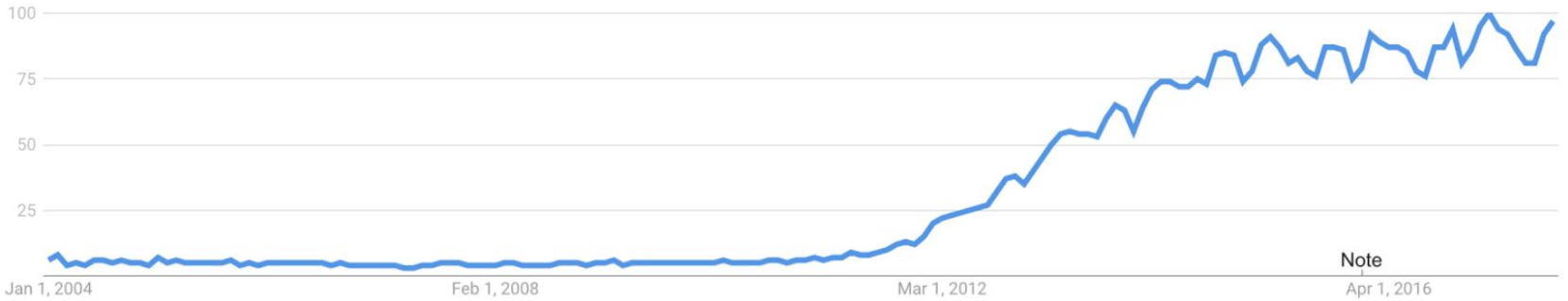
Call to action

- What the students can do:
  - Self-educate (books + MOOCs)
  - Build re-usable libraries/packages instead of scripts.
  - Hobby projects

- How the university can help:
  - “R for Data Science” and “Python for Data Analysis” courses
  - Those skills are as important as any when in industry
  - More meaningful projects, with better evaluation
    - Correctness of the solution
    - Reproducible
    - Easy of use by third-parties
    - Properly tested

Big Data

Big Data Boom



- MapReduce: Simplified Data Processing on Large Clusters (2004)
- Hadoop open sourced by Yahoo! in 2006
- Spark open-sourced in 2010

Big Data Ecosystem

- Development driven by the needs of the Tech Giants

- Open-sourced most of the interesting technologies
  - Engagement from the community (development, support)
  - New employees already familiar with their tools
  - Marketing

- Everything available for everyone. Software (open-source) + Hardware (cloud-providers)

- Overwhelming. Important to understand benefits and limitations.

- Extremely overused nowadays
- **Simply scales what you have always been able to do in your laptop**

Call to action

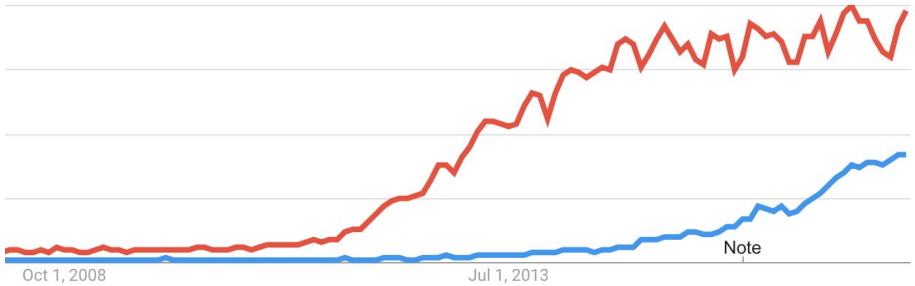
- Students
  - It is easy to install Hadoop and Spark in your laptop
  - Jobs written locally can easily scale to huge volumes of data
  - Try their Quick Start guides

- University
  - When solving problems/giving courses with R and Python, we should ask:
    - What if I had 10x, 100x, 1000x more data?
    - What if data were streaming with increasing speed, X rows per hour/minute/second?

- **Our students need to understand the scalability of their solutions.**

Deep Learning

Deep Learning Boom



Top libraries by Github stars		
#1:	71627	tensorflow/tensorflow
#2:	20489	BVLC/caffe
#3:	20038	fchollet/keras
#4:	12558	Microsoft/CNTK
#5:	11369	dmlc/mxnet
#6:	7712	pytorch/pytorch
#7:	7332	torch/torch7
#8:	7297	deeplearning4j/deeplearning4j
#9:	6981	Theano/Theano
#10:	6767	tflearn/tflearn
#11:	5742	caffe2/caffe2
#12:	5544	baidu/paddle
#13:	5336	deepmind/sonnet
#14:	3242	Lasagne/Lasagne
#15:	3232	NervanaSystems/neon
#16:	2987	pfnet/chainer
#17:	2833	davisking/dlib
#18:	2525	NVIDIA/DIGITS
#19:	1775	clab/dynet

- Big Data boom around 2010/2011
- Deep Learning boom around 2014/2015
- Both still strong.
- Huge success for text and image analysis
- Many deep learning frameworks
- TensorFlow by far the most popular

Just complex models

- There is nothing inherently special about Deep Learning models

$$f(y|\theta) = \mathcal{N}(h_\theta(x), \sigma^2)$$

- Typical estimation setup: Variations of SGD
  - Forward propagation: Given x, compute h
  - Backward propagation: Efficient way to compute the gradient of the loss

- Many popular classes of models (each with many variations)
  - Convolutional Neural Networks (CNNs)
  - Recurrent Neural Networks (RNNs)
  - Generative Adversarial Networks (GANs)

TensorFlow

- TensorFlow is an open source software library for numerical computation using data flow graphs.
  - Nodes in the graph represent mathematical operations,
  - Graph edges represent the multidimensional data arrays (tensors)

- Anything that can be represented as a data flow graph can be computed using TensorFlow.

- Provides a Python API
- Relatively easy to use, after understand its “data flow graphs” philosophy

TensorFlow

```
import tensorflow as tf

# Model parameters
W = tf.Variable([.3], dtype=tf.float32)
b = tf.Variable([-.3], dtype=tf.float32)
# Model input and output
x = tf.placeholder(tf.float32)
linear_model = W * x + b
y = tf.placeholder(tf.float32)

# loss
loss = tf.reduce_sum(tf.square(linear_model - y)) # sum of the squares
# optimizer
optimizer = tf.train.GradientDescentOptimizer(0.01)
train = optimizer.minimize(loss)
```

TensorFlow

```
# training data
x_train = [1, 2, 3, 4]
y_train = [0, -1, -2, -3]
# training loop
init = tf.global_variables_initializer()
sess = tf.Session()
sess.run(init) # reset values to wrong
for i in range(1000):
  sess.run(train, {x: x_train, y: y_train})
```

Call to action

- Deep Learning models are popular and they are not going anywhere

- Students:
  - Free MOOCs
    - Deep Learning @ Coursera
    - Deep Learning @ Udacity

- University:
  - Include Neural Network/Deep Learning models into existing classes
    - Introduction to Statistical Learning
  - Full semester dedicated course to Deep Learning (Theory and Practice)
  - Very active research area
    - Attract funding and researchers

Statisticians

Statisticians and Statistics

- Hard to find in industry
  - Lack of basic CS skills is a blocker
  - It is a hard job. Not easy to follow a recipe.

- Hard for **me** to define what makes a good statistician
  - People ask me, what should I do/read to do what you do?
  - Not easy question in my opinion.

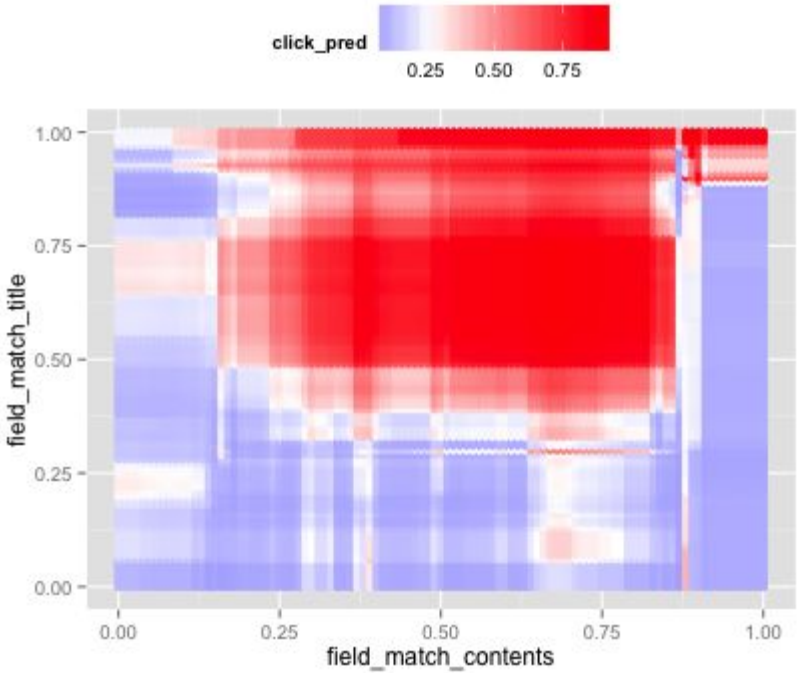
- Analyse every problem/solution I see using core stat knowledge
  - Probability, Statistical Inference, Bayesian/Classical Statistics, etc.
  - Everything is connected.
  - Simpler to judge advantages and disadvantages of different methodology.

- ML and Stat look at problems a bit differently

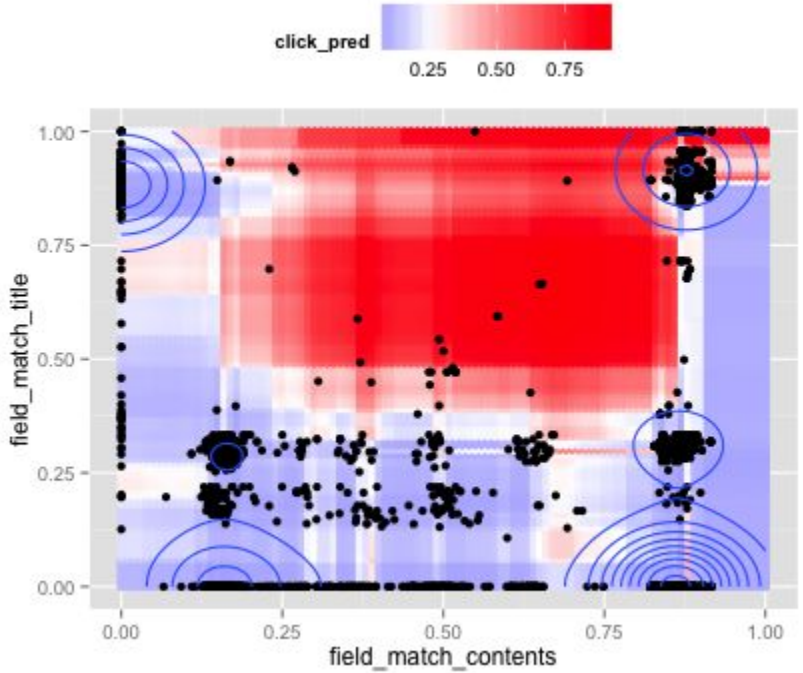
Case I: Uncertainty misconceptions

- Classification tasks are very popular in ML
- Given a set of covariates x, classify y as either being 0 or 1
- As a statistician, I look at this problem as a regression.
  - Predict the probability that y = 1
- But a popular ML book considers p(y=1) to be a measure of uncertainty of your classification.
- **p(y=1) is a point estimate, not an uncertainty measure**

Case II: GBDT


click_pred
0.25
0.50
0.75
1.00
0.75
0.50
0.25
0.00
field_match_title
0.00
0.25
0.50
0.75
1.00
field_match_contents

Case II: GBDT


click_pred
0.25
0.50
0.75
1.00
0.75
0.50
0.25
0.00
field_match_title
0.00
0.25
0.50
0.75
1.00
field_match_contents

Call to Action

- I am continuously learning how to be a better Applied Statistician
- I am now leading a team of young Data Scientists at AIA
- Main learning lesson so far:
  - **We need to explain every decision we made**
  - **Justify what we are going to do next before doing it**
- Enough with demos, we need to solve valuable problems.
  - Why not start with problems affecting the university??

Thank you!

Thiago Guerrera Martins, PhD
Principal Data Scientist, AIA Science

A better future through the use of Artificial Intelligence, Analytics and Machine Learning

AIA | SCIENCE
ANALYTICS OF THINGS