ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Data Science for Business

MGT 432

Prof. Kenneth Younge, Ph.D.

Associate Professor

Chair of Technology and Innovation Strategy

Session 1

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Data Science for Business

MGT 432 → Data Science **Methods** for Business Problems

MGT ??? → Data Science **Applications** for a Business Context

Prof. Kenneth Younge, Ph.D.

Associate Professor

Chair of Technology and Innovation Strategy

Session 1

Motivation: Why study Data Science?

- "You can't manage what you ~~don't measure~~ can't **predict**."

- "Data are becoming the new raw material of business."
  ~ Craig Mundie (former chief strategy officer at Microsoft)

- "Big" data are growing in **volume**, **velocity**, & **variety**

- The scale of new data enable & require new methods

- You want a job…

Motivation: Why study Data Science?

McKinsey Global Institute says:

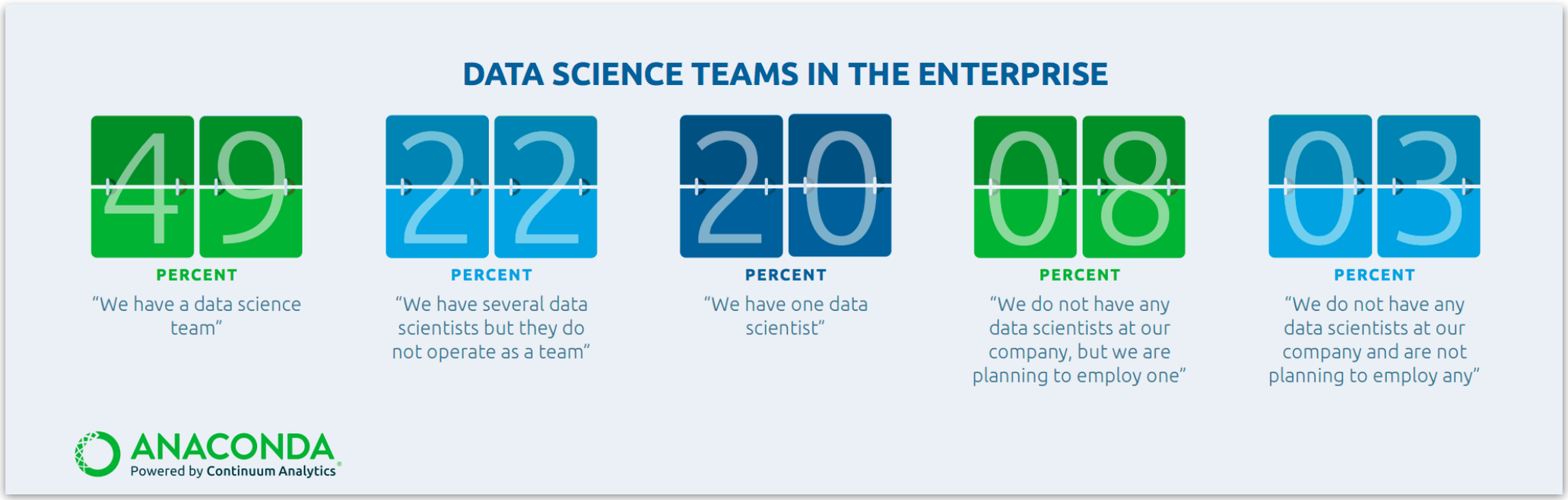
> *"There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."*

Source: "Big data: The next frontier for innovation, competition, and productivity." *McKinsey Global Institute* (2011)

That was so **2011**.
Demand for data science has exploded,
far, far, far beyond expectations.

Motivation: Why study Data Science?

- Data science requires cross-functional ***teams***



Source: "Winning at Data Science: How Teamwork Leads to Victory." Open Data Science Survey Continuum Analytics, 2017.

- To work effectively with a data science team, you need to learn the vocabulary & concepts of the field.

Motivation: Why study Data Science?

- How do you get onto a data science team?

- Questions for hiring a data scientist:
  - Explain precision and recall. How do they relate to the ROC curve?
  - What is regularization? When and why it is useful?
  - How do you know that an improvement to a model is really an improvement?
  - When is it better to have too many false positives? Or too many false negatives?
  - What is selection bias? Why is it important? How can you avoid it?
  - Give an example of using an experiment to answer a difficult-to-answer question.
  - How do you know if your data has outliers? What should you do if you find them?
  - What is machine learning? And why is it called *learning*?
  - When/why should we build a neural net?

Source: Compendium of questions/statements found online.

- You will be able to answer these questions, and more!

Objective

- This course is **not**:
  - an algorithm course — i.e., computational efficiency
  - a statistics course — i.e., inferential testing
  - a probability course — i.e., proofs for methods
  - an optimization course — i.e., solutions for methods
  - a python course — i.e., coding best-practices
  - a policy course — i.e., privacy, ethics, welfare
  - a strategy course — i.e., business goals & tradeoffs

Objective

- This course **is**:

  - a survey of data science ***methods***

  - an introduction to how data science can model and predict a broad range of problems for a business domain

  - a foundation to help you evaluate & communicate data science options with others (i.e., as part a DS team)

  - a new course - so expect adjustments along the way

Admin: Instructors

- Professor

  Kenneth A. Younge, PhD
  Odyssea 2.02 Phone: 021 693 00 09
  kenneth.younge@epfl.ch
  Office Hours: Mondays 4:30-5:30

  Co-founder / Director of Development for several internet startups; applied economist / econometrics

- Post-Doc

  Omid Shahmirzadi, PhD
  Odyssea 2.17
  omid.shahmirzadi@epfl.ch
  Office Hours: Fridays 10:00-11:00

  Computer scientist with research experience in distributed systems and bioinformatics

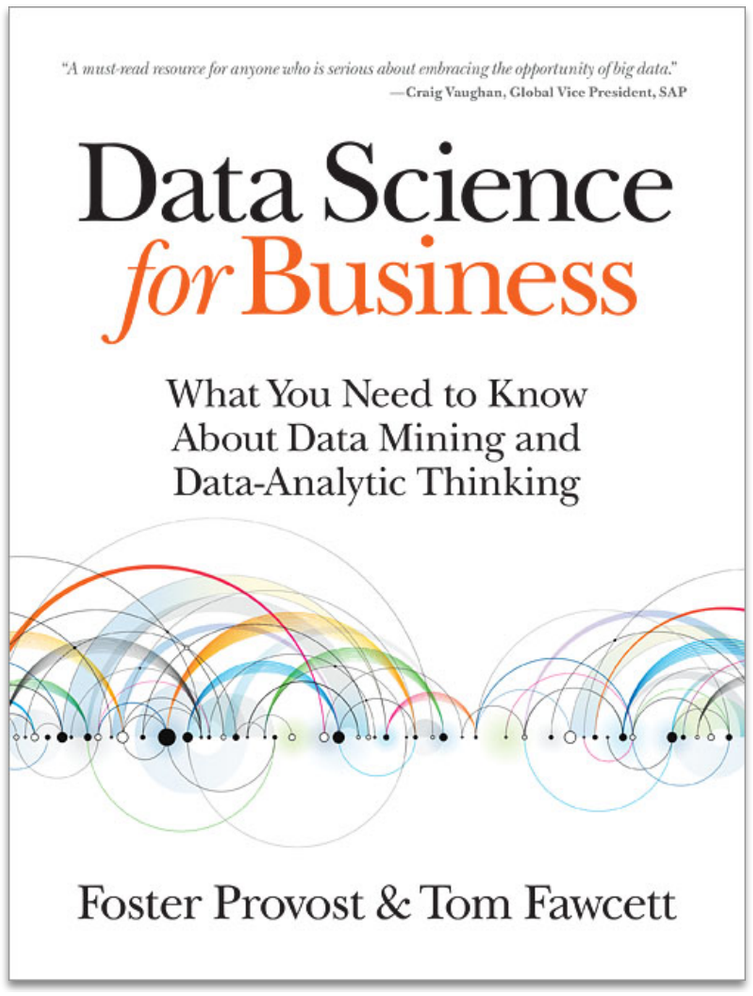
Admin: Prerequisites

- The following are required to take this course
  - Linear algebra & calculus (undergraduate)
  - Statistics and probability (undergraduate)
  - Strong computing skills
  - Python tutorials to be completed by 2nd class
  - Ability to setup Anaconda
  - Ability to use `git` and GitHub.com

Admin: Assigned Reading

- One main textbook:

Data Science for Business
(Provost & Fawcett, 2016)
Buy and read this book.



A (very) easy introduction to the basic concepts with little math and no programming code.

We will supplement the textbook on a case-by-case basis with readings from other books, lecture notes, and programming examples from the instructors.

Admin: Other reading

- We will point you to the following texts for more detail

  - An Introduction to Statistical Learning
    (James, Witten, Hastie, and Tibshiran, 2013)

  - *Data Mining: Concepts and Techniques*
    (Han, Pei, & Kamber, 2012)

  - *Data Preprocessing in Data Mining*
    (Garcia, Luengo, & Herrera, 2015)

  - *Evaluation Learning Algorithms*
    (Japkowicz & Shah, 2011)

  - *Deep Learning*
    (Goodfellow, Bengio, & Courville, 2016)

Admin: Schedule

1.	Sept. 25	Welcome		
2.	Oct. 2	Data	Assignment 1	
3.	Oct. 9	Linear Models	Assignment 2	
4.	Oct. 16	Practical 1		Dr. Shahmirzadi
5.	Oct. 23	Model Evaluation	Assignment 3	
6.	Oct. 30	Similarity Models	Assignment 4	
7.	Nov. 6	Practical 2		Dr. Shahmirzadi
8.	Nov. 13	Midterm Exam		
9.	Nov. 20	Decision Trees	Assignment 5	
10.	Nov. 27	Support Vector Machines	Assignment 6	
11.	Dec. 4	Neural Nets		
12.	Dec. 11	Big Data		Dr. Shahmirzadi
13.	Dec. 18	Final Presentations		
--	TBD	FINAL EXAM		

Models with more Assumptions

Models with fewer Assumptions

Admin: How to learn “the material”

- Instructor-directed learning
  - Lectures
  - Examples
  - Demos

- Self-directed learning
  - Readings
  - Assignments
  - Project
  - Presentation

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

9:15 - 10:45	lecture	90 min
10:45 - 11:00	break	15 min/Q&A
11:00 - 11:30	lecture	30 min
11:30 - 12:00	examples	30 min

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

Slides, examples, etc. for each session will be available in git repo at start of class.

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

Each session includes real, running code so you can see how concepts are applied in practice

Be sure to go through the examples, play with them, and figure them out

One example is just that - one example

You need to see a range of how results change as parameters/coding change

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

Three demo / lab sessions

Walk through a complete solution.

Time for personalized Q&A

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

1.	Sept. 25	Welcome		
2.	Oct. 2	Data	Assignment 1	
3.	Oct. 9	Linear Models	Assignment 2	
4.	Oct. 16	Practical 1		Dr. Shahmirzadi
5.	Oct. 23	Model Evaluation	Assignment 3	
6.	Oct. 30	Similarity Models	Assignment 4	
7.	Nov. 6	Practical 2		Dr. Shahmirzadi
8.	Nov. 13	Midterm Exam		
9.	Nov. 20	Decision Trees	Assignment 5	
10.	Nov. 27	Support Vector Machines	Assignment 6	
11.	Dec. 4	Neural Nets		
12.	Dec. 11	Big Data		Dr. Shahmirzadi
13.	Dec. 18	Final Presentations		
--	TBD	FINAL EXAM		

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

6 Assignments

OK to discuss in groups of 2 to 4,
but you must write up your own answers.

You may NOT copy text and/or code.

Assignments are **very** important!

You won't do well in the course if you skip them.

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

Team Project on Product Reviews

A real-world problem from my lab

You will be assigned into groups of 2 after the midterm, based on your midterm grade.

Admin: Pedagogy

- How will you learn?
  - Lectures
  - Examples
  - Demos
  - Assignments
  - Project
  - Presentation

Learn to communicate results

Be prepared for Q&A !

Admin: Evaluation

- How will you be evaluated?

  - Assignments 12%

  - Midterm Exam 20%

  - Semester Project 25%

  - Final Presentation 8%

  - Final Exam 35%

Admin: Evaluation

- **Assignments** (12%)

  Submit individual git repos. It is OK to collaborate (i.e. discuss problems and solutions) in groups of 2 to 4, but you must write up your own answers (do not copy text/code)

  Graded on a question-by-question basis, based on a **Check+** (6), **Check** (5), **Check-** (4), or **No Answer** (3) basis, provided you submit at least some answer(s) for assignment.

  Question-by-question scores are averaged up to a grade for the assignment overall.

  Due two weeks after assigned (Sunday at midnight).
  No late assignments accepted.

Warning! Assignments are only 12% of your grade, but it will be very difficult to do well on tests if you do not take them seriously.

Admin: Evaluation

- **Midterm Exam** (20%)

  Mostly conceptual questions (1 or 2 sentence answers)

  Some fill in the blank, multiple choice, etc.

  Show code & ask about implications

  Change working code to do something else

  1 or 2 written problems ("Explain why...")

Admin: Evaluation

- **Semester Project** (25%)

  Paired into groups of 2 based on the midterm

  Work on a git repo for the project as a team

  Due at midnight one week before your presentation

  Lose 1% on day 1, 2% on day 2, 3% on day 3, …
  (each penalty is in addition the the penalty of the day before)

Admin: Evaluation

- **Final Presentation** (8%)

  What ?!?!? Why present in a coding/methods course?

  Because Data Science is about ***communicating*** findings

  Presentations made during the final class session

  You must be prepared to field Q&A from the instructors
  (as such, the presentation is also an oral examination)

Admin: Evaluation

- **Final Exam** (35%)

  Same format as the midterm

  Cumulative over the entire semester

Admin: Evaluation

- How will you be evaluated?

  - Assignments 12%

  - Midterm Exam 20%

  - Semester Project 25%

  - Final Presentation 8%

  - Final Exam 35%

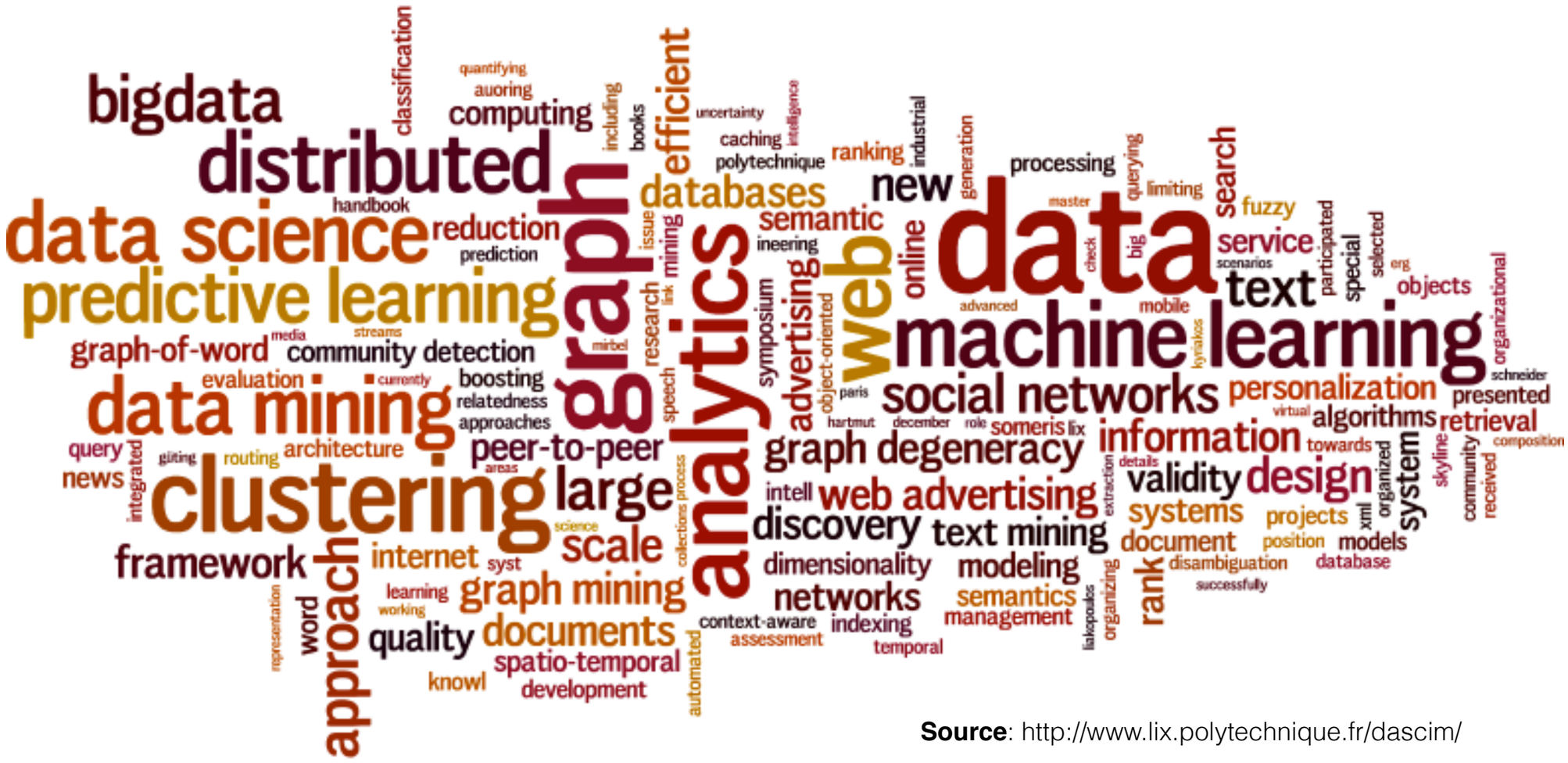
Course Admin:

Final questions about the Syllabus?

Lecture:

What is Data Science

What is Data Science?



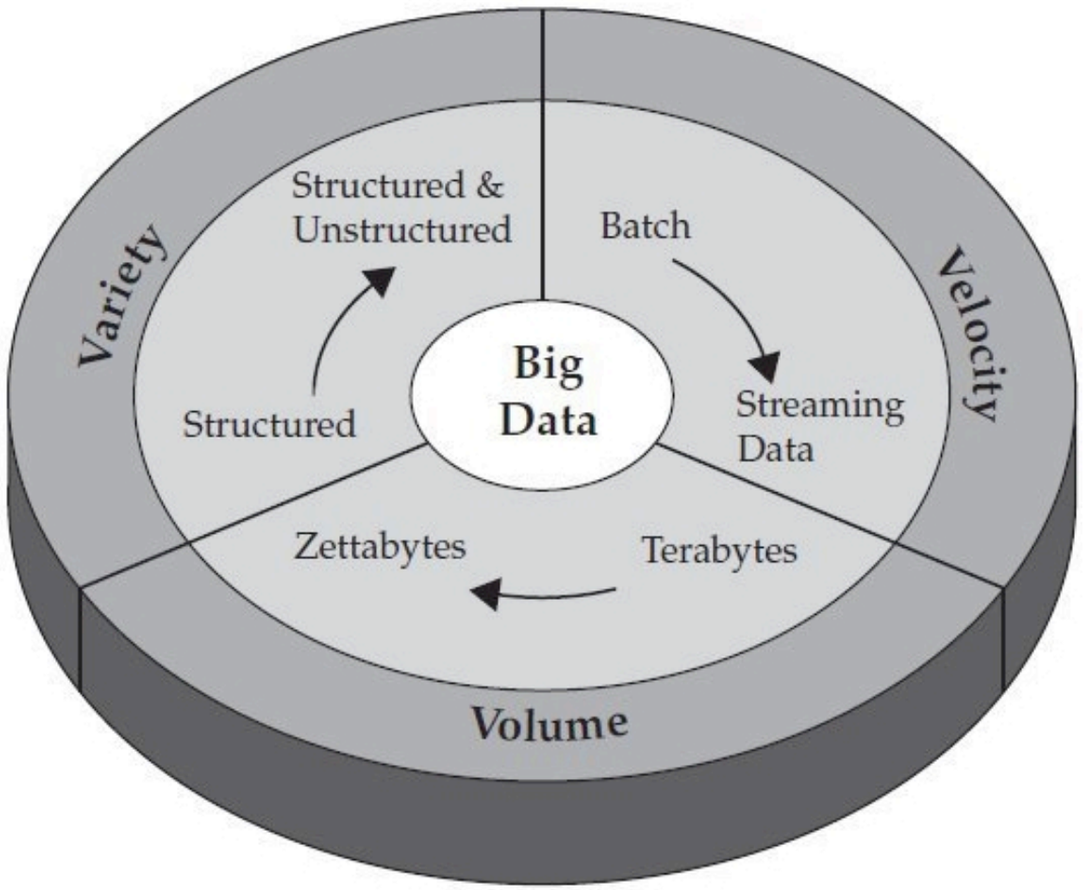
Source: http://www.lix.polytechnique.fr/dascim/

What is Data Science?

- Data science is the process of ***learning*** from data
  - typically in a computationally-enabled way
- This type of "learning" is different from other approaches that *theorize* and then *test* against data
  - much of statistics is concerned with *inferential testing*
  - "null hypothesis testing" asks how likely would we observe **y** given **x,** if **x** and **y** were actually unrelated.
- Data science is about ***predictive analytics***
  - i.e., how to make the "best" possible **prediction**

What is Data Science?

- Data Science takes a computational perspective
  - Driven by a new abundance of data



Source: https://apandre.wordpress.com/2013/11/19/datawatch/

What is Data Science?

- Data Science takes a computational perspective
  - Driven by a new abundance of data
  - But it is also driven by a new abundance of computing power
    - more powerful computers (CPUs)
    - cloud computing (often networked systems of cheap, commodity computers, GPUs, storage arrays, etc…)
- Data Science has also been shaped by the cash and culture of the internet revolution
  - Open-source ecosystems, agile dev, data-driven managers…

What is Data Science?

- Data Science relies on mathematics of optimization

  - Algorithms lie at the heart of data science

  - Many machine learning problems can be reduced to a mathematical optimization problem

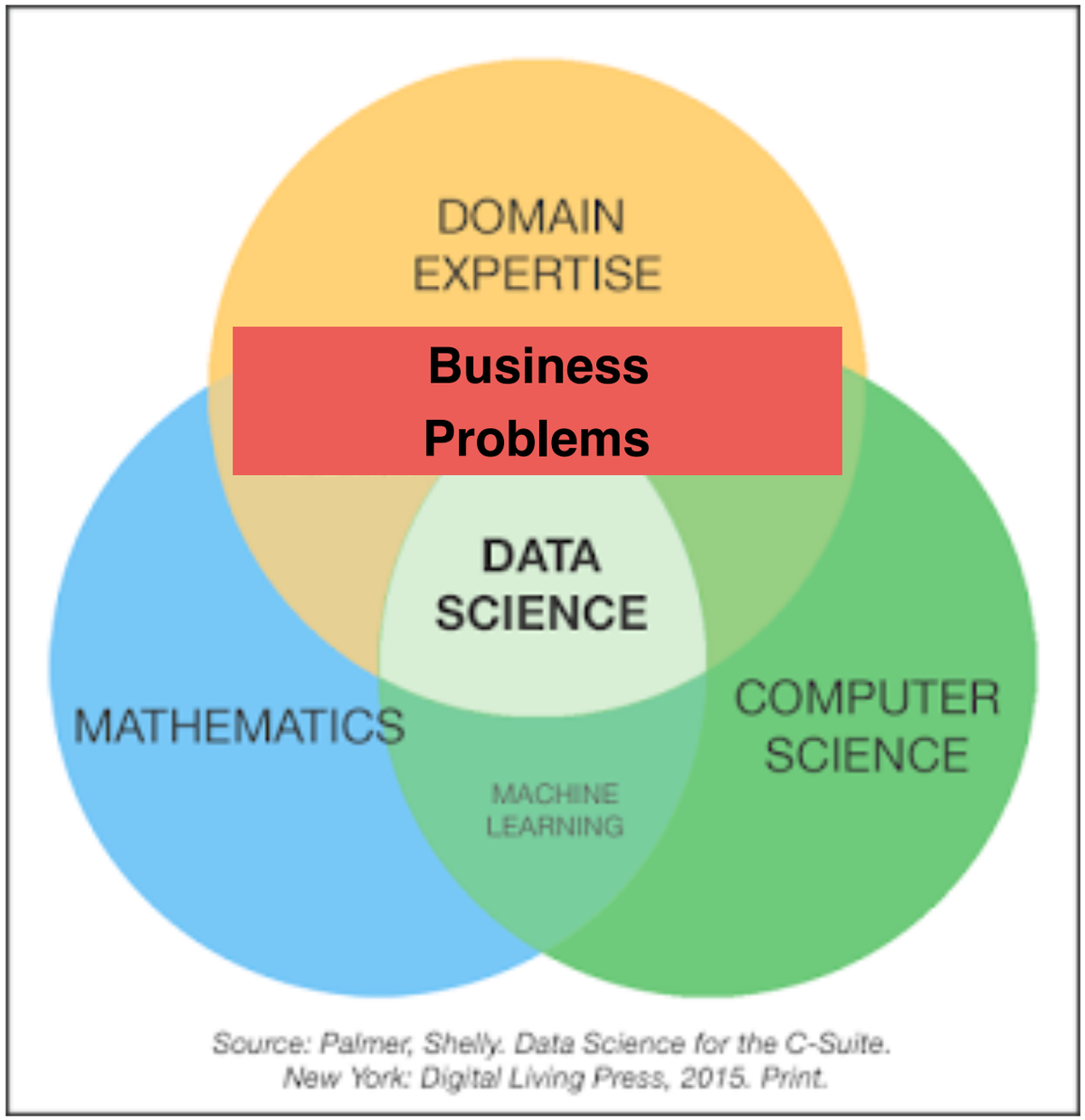
  - Advances in data science usually relate to some underlying optimization technique

  - But for this course, that is too advanced - we take use optimization as a “black box” oracle.

  - As a consequence, we require very little math.

What is Data Science?

- Data science spans many fields



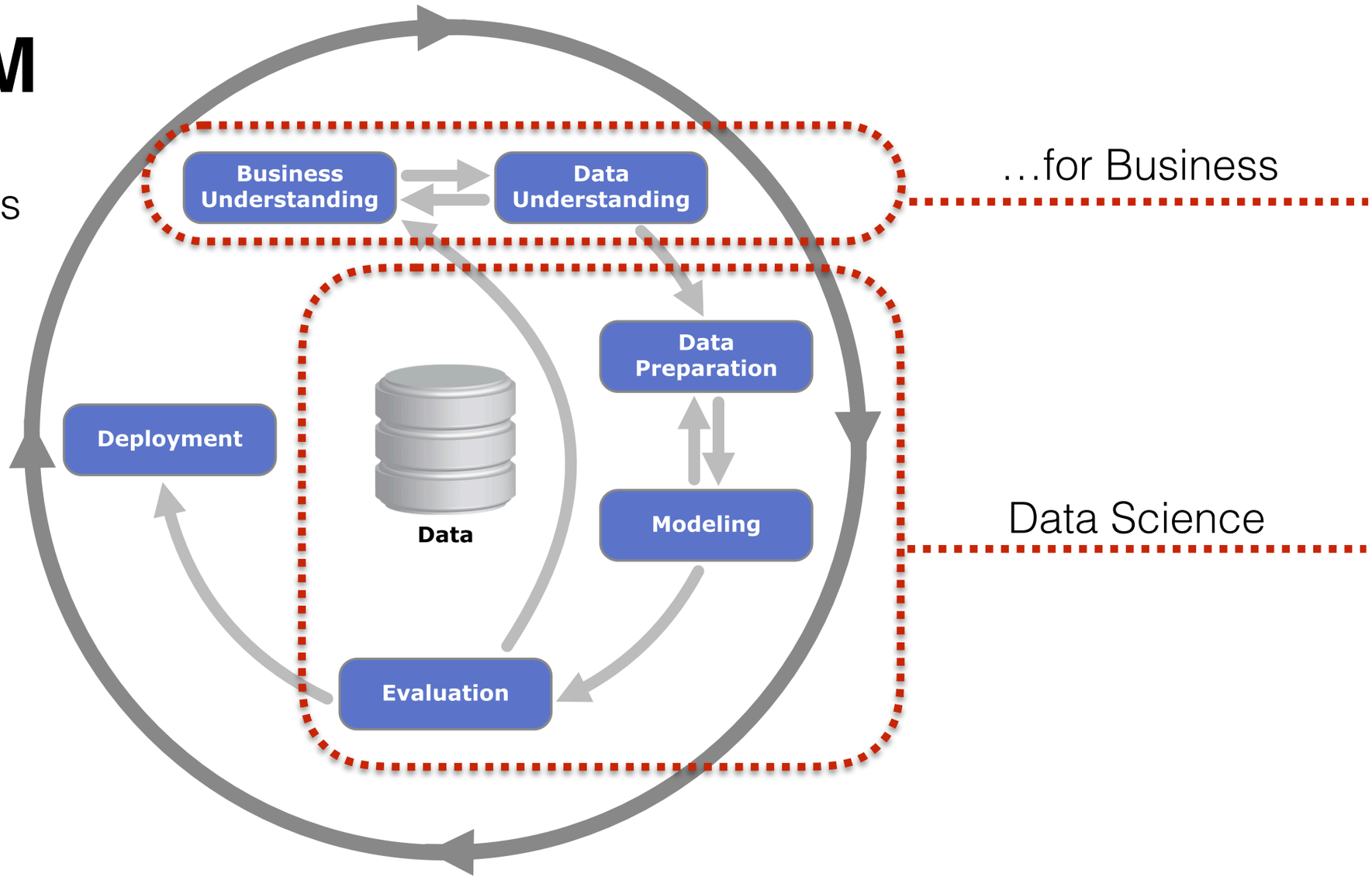
Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

What is Data Science?

- Data Science builds on prior work from **Data Mining**, Business Analytics, Management and Information Systems (MIS), and Marketing.

CRISP-DM

Cross Industry Standard Process for Data Mining

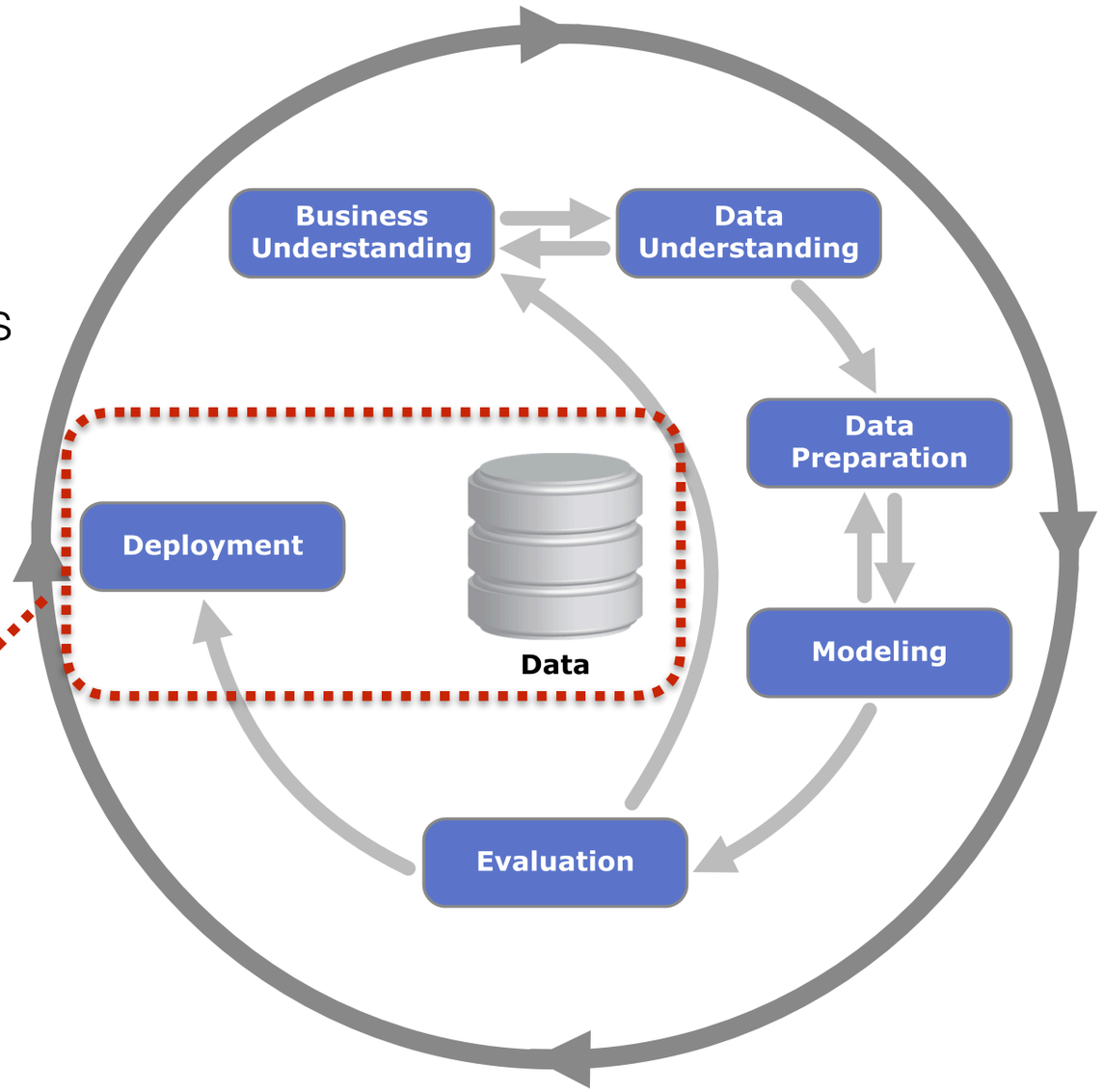


What is Data Science?

- Data Science builds on prior work from **Data Mining**, Business Analytics, Management and Information Systems (MIS), and Marketing.

CRISP-DM

Cross Industry Standard Process for Data Mining

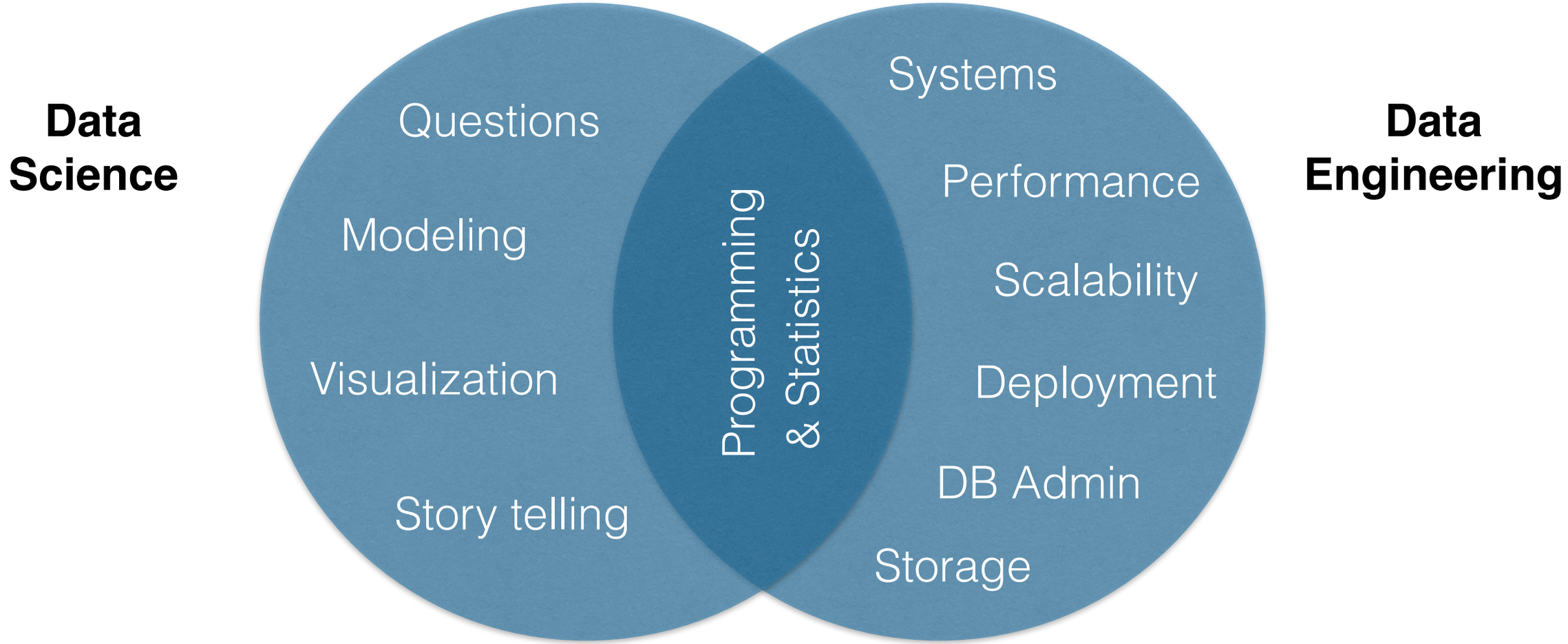


Data Engineering

- CRISP-DM assumes data is available
- Designing data collection is an important topic!
- But it isn't part of this course.

What is Data Science?

- Data science is different than data ***engineering***
  - Data Science = ***learning*** *from data*
  - Data Engineering = ***implementing*** *for data solutions*



What is Data Science?

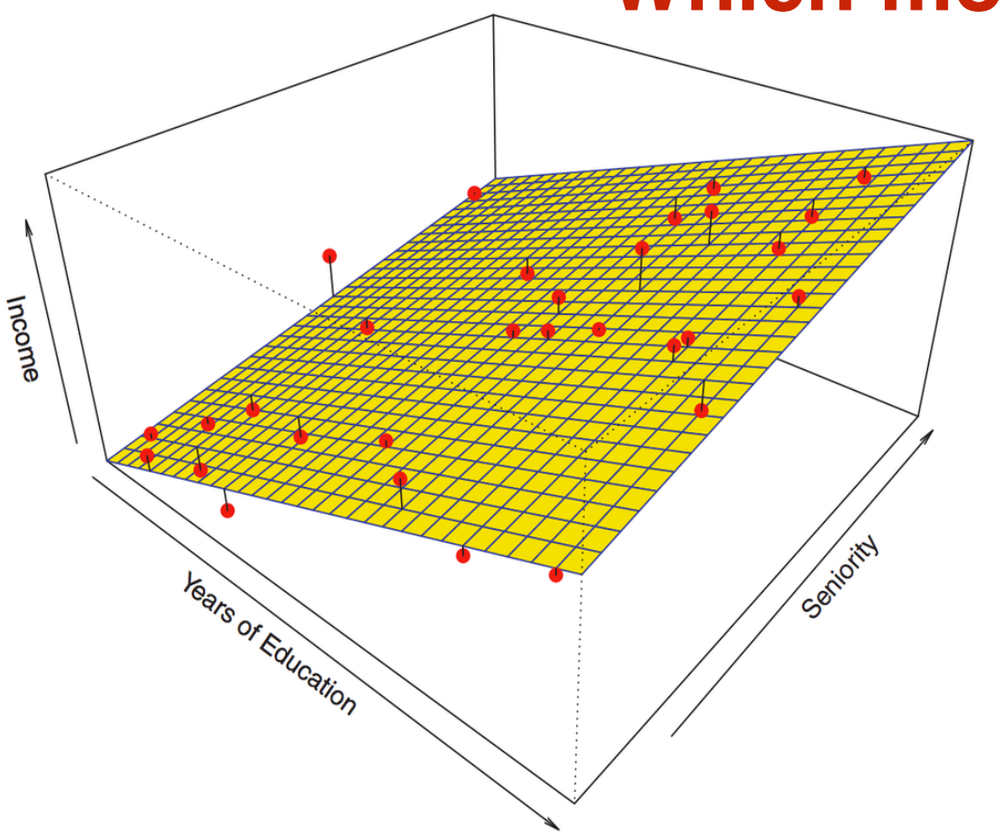
- Data science is different than data ***engineering***
  - Data Science = ***learning*** *from data*
  - Data Engineering = ***implementing*** *for data solutions*
- Data science is about figuring out which methods/models make the best predictions from data… **and why**.
- Data engineering is about collecting the right information and delivering timely predictions for decision making
- We don't worry about data engineering in this course - for that we encourage you to take more courses

What is Data Science?

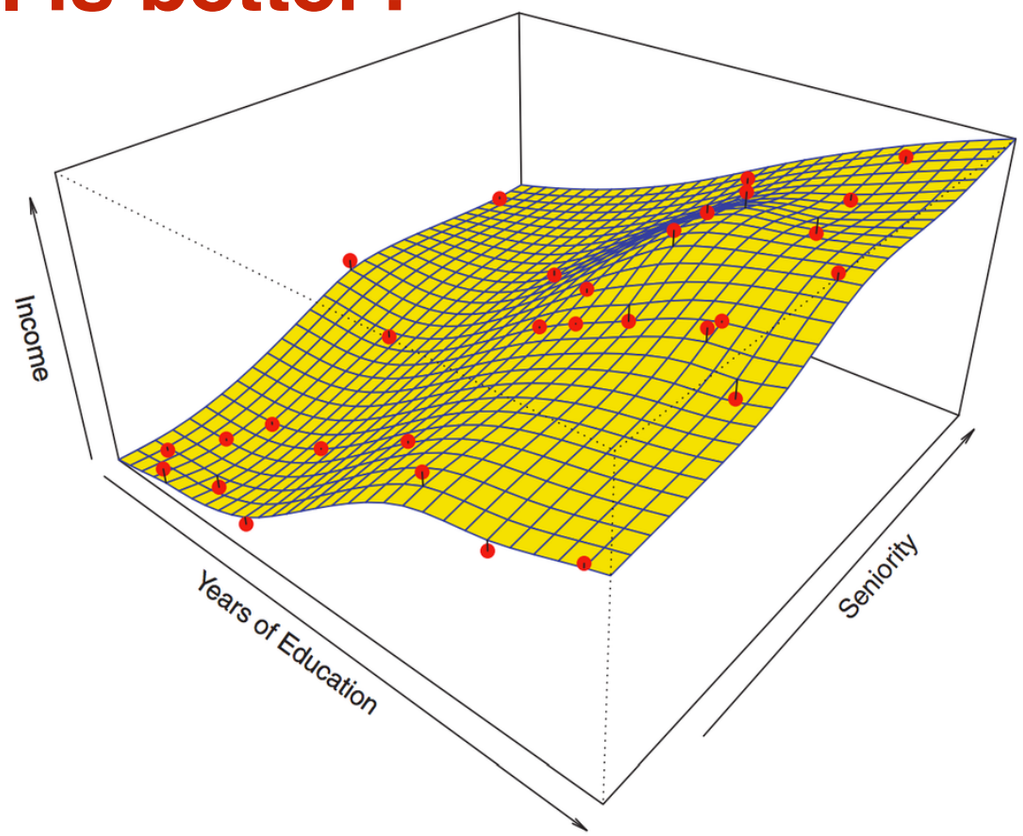
- We won't worry about data engineering considerations.
- We will focus mostly on data modeling and evaluation.

Let's say we have two models

Which model is better?



Less accurate, more interpretable



More accurate, less interpretable

Source for Figures: Introduction to Statistical Learning, James, Witten, Hastie, and Tibshiran, 2013.

What is Data Science?

- Why is data science a hot topic for business?
  - Data can be a critical asset for many firms
    - Data can help businesses make better decisions
    - Data can also isolate firms from competition by others
  - Leveraging data is no longer just for “high-tech” or “information technology” (IT) companies
  - To get value from data, though, firms need to have the competencies required to take advantage of that data

Course Setup

Course Setup: Overview

git

Python

Anaconda

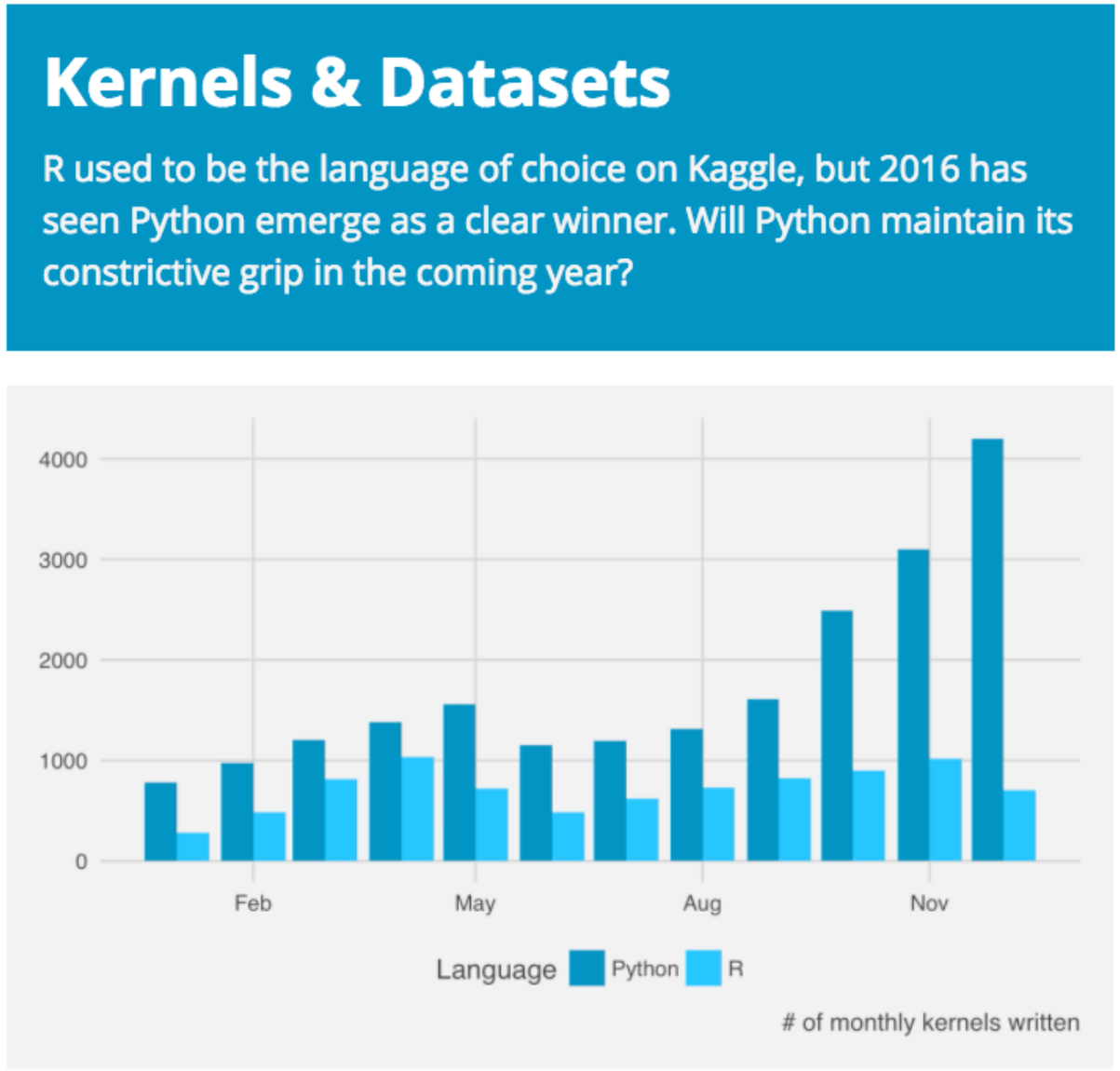
Jupyter

Course Setup: git

- **git**
  - Install **git**: https://git-scm.com/
  - Learn the basic commands: https://git-scm.com/documentation
- **github.com**
  - Create an account: https://github.com
  - Email to course assistant your GitHub **username** (NOT GitHub email)
    - We will assign each student a private repo for your assignments
    - We will assign each team a private repo for your projects

Course Setup: Python

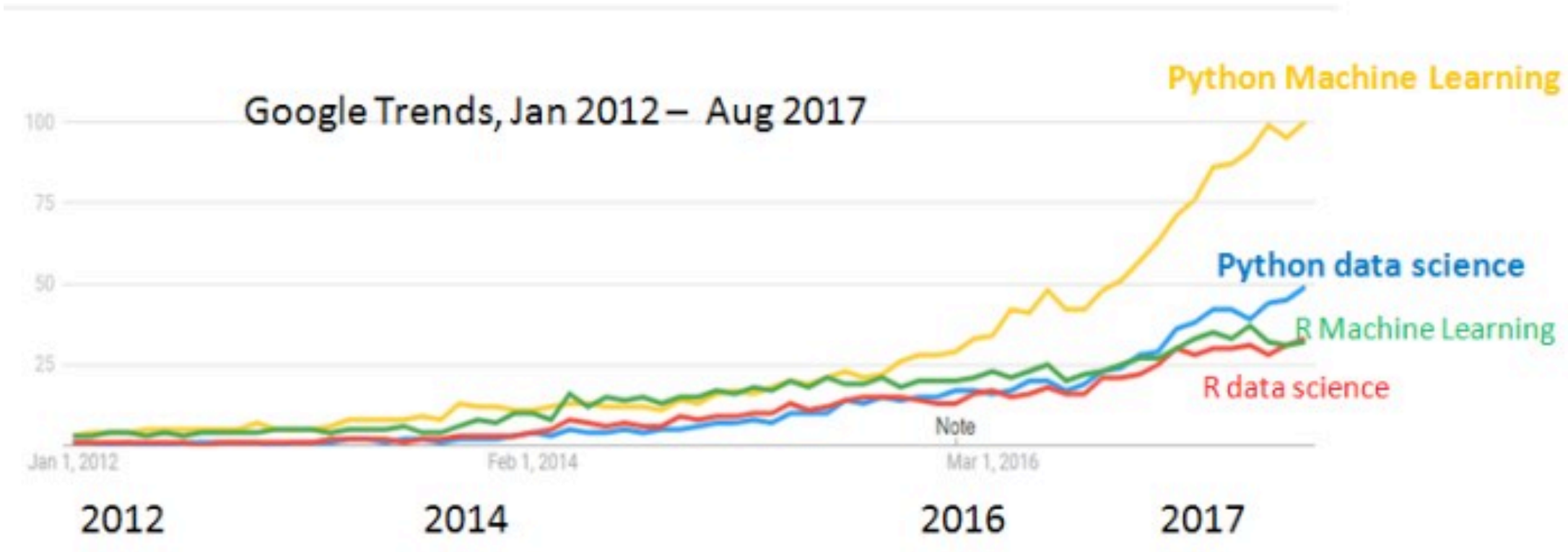
- Why Python? The Python ecosystem is beating out **R** (a competition option)



Source: https://www.kaggle.com/kernels

Course Setup: Python

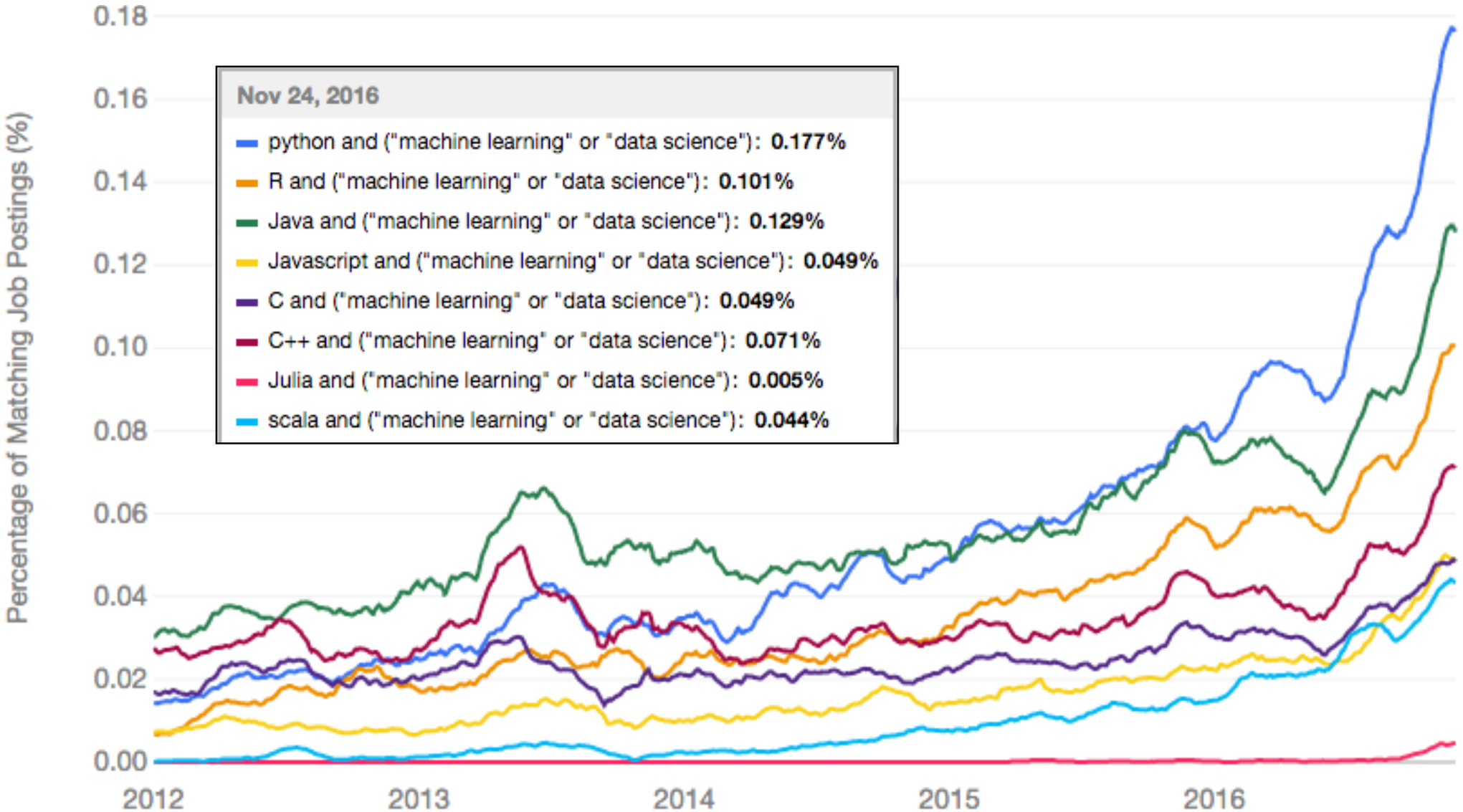
- Why Python? The Python ecosystem is beating out **R** (a competition option)



Source: Google Trends, Jan 2012 - Aug 2017, "Python Machine Learning", "R Machine Learning", "Python data science", and "R data science".

Course Setup: Python

- More Data Science jobs recruit for Python



Source: Job postings on www.indeed.com based on the following combinations of terms: python and ("machine learning" or "data science"), R and ("machine learning" or "data science"), Java and ("machine learning" or "data science"), Javascript and ("machine learning" or "data science"), C and ("machine learning" or "data science"), C++ and ("machine learning" or "data science"), Julia and ("machine learning" or "data science"), and scala and ("machine learning" or "data science"). Also see article by Jean-Francois Puget, IBM on http://www.kdnuggets.com/.

Course Setup: Python

- Why is Python winning? Python works well for both…

  - **Computation**

```
def bootstrap_sterr(data, n_samples, stat):
    n = len(data)
    ix = np.random.randint(0, n, (n_samples, n))
    samples = data[ix]
    sterr = np.std(stat(samples, 1))
    return sterr
```

    - it is a full-featured language with object-oriented classes
    - there are mature libraries for numerical calculation
    - there is a robust ecosystem for specialised data science applications

  - **Glue**

```
slope, intercept, r_value, p_value, std_err = stats.linregress(x,y)
```

    - easy syntax (can work as a scripting language with advanced features hidden)
    - intuitive data structures and manipulation
    - runs anywhere

Course Setup: Python

- We will use Python mostly as **glue**

  This is not a computer science course, don't expect best coding practices!

pandas

matplotlib

NumPy

scikit learn

StatsModels
Statistics in Python

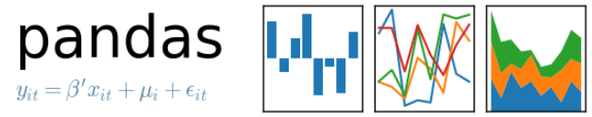
NLTK

SciPy

gensim

Google's Cloud ML Platform will extend the *scikit-learn* setup

Course Setup: pandas



- A popular library for data munging and analysis
- DataFrame object is at the core (a concept borrowed from R)
- Rich set of tools to manipulate DataFrames
  (aligning, shaping, slicing, subsetting, transforming, merging, joining, …)
- Reading and writing data from/to different sources
- Optimised for performance, with critical parts written in C

Course Setup: matplotlib

- A widely used visualisation library for high quality 2D plotting
- Support for many different types of plots
- Integrated with Pandas DataFrames
- Easy to use (well… not really)

- Can be used in python scripts as well as jupyter notebooks
- There are other visualisation libraries in Python for specialised purposes (e.g. Seaborn, Bokeh, …)

Course Setup: StatsModels

- Functions for estimating statistical models and performing statistical tests
- A general purpose statistics library to replace **R**
- Many of the advantages of **R**, but with all the advantages of a true programming language (uniform syntax, debugging, object oriented, fast)

Source: https://www.kaggle.com/kernels

Course Setup: NLTK and gensim

- NLTK
  - A “Natural Language Tool Kit” for computational linguistics
  - Used for parsing, stemming, semantics, etc. - but can do much more
- Gensim
  - A library for computing “vector spaces” and “topic models” for text
  - Used for Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) - more on that later

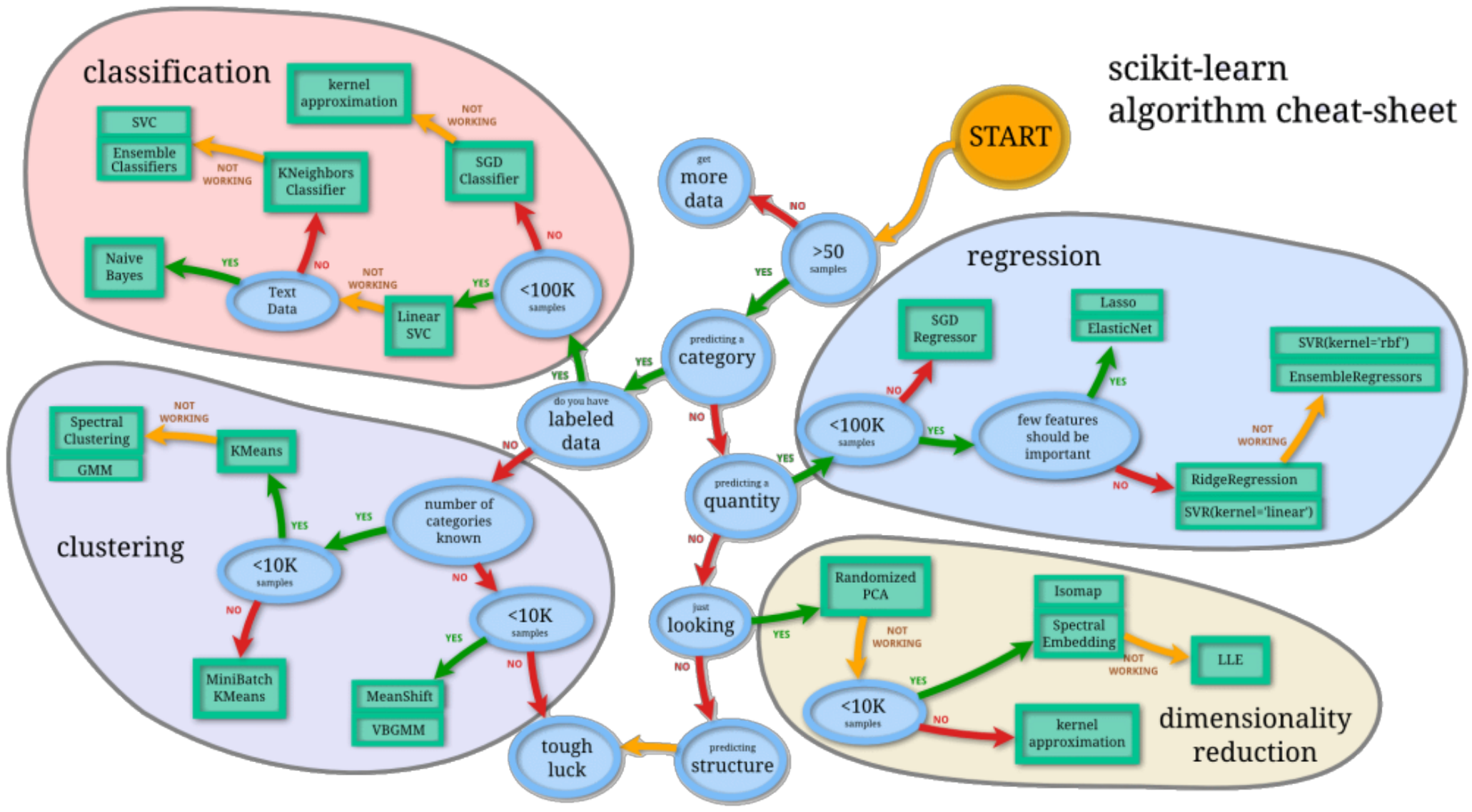
Course Setup: Numpy & SciPy

- Numpy
  - Built on a powerful N-dimensional array object
  - Useful for many basic linear algebra operations
  - The underlying linear algebra library for scipy

- SciPy
  - Python library for scientific computation
  - Higher level numerical routines such as integration and optimisation
  - Underlying computation library for sklearn data science algorithms

Course Setup: scikit-learn

Simple and efficient implementation of the main data science algorithms



Course Setup: Anaconda

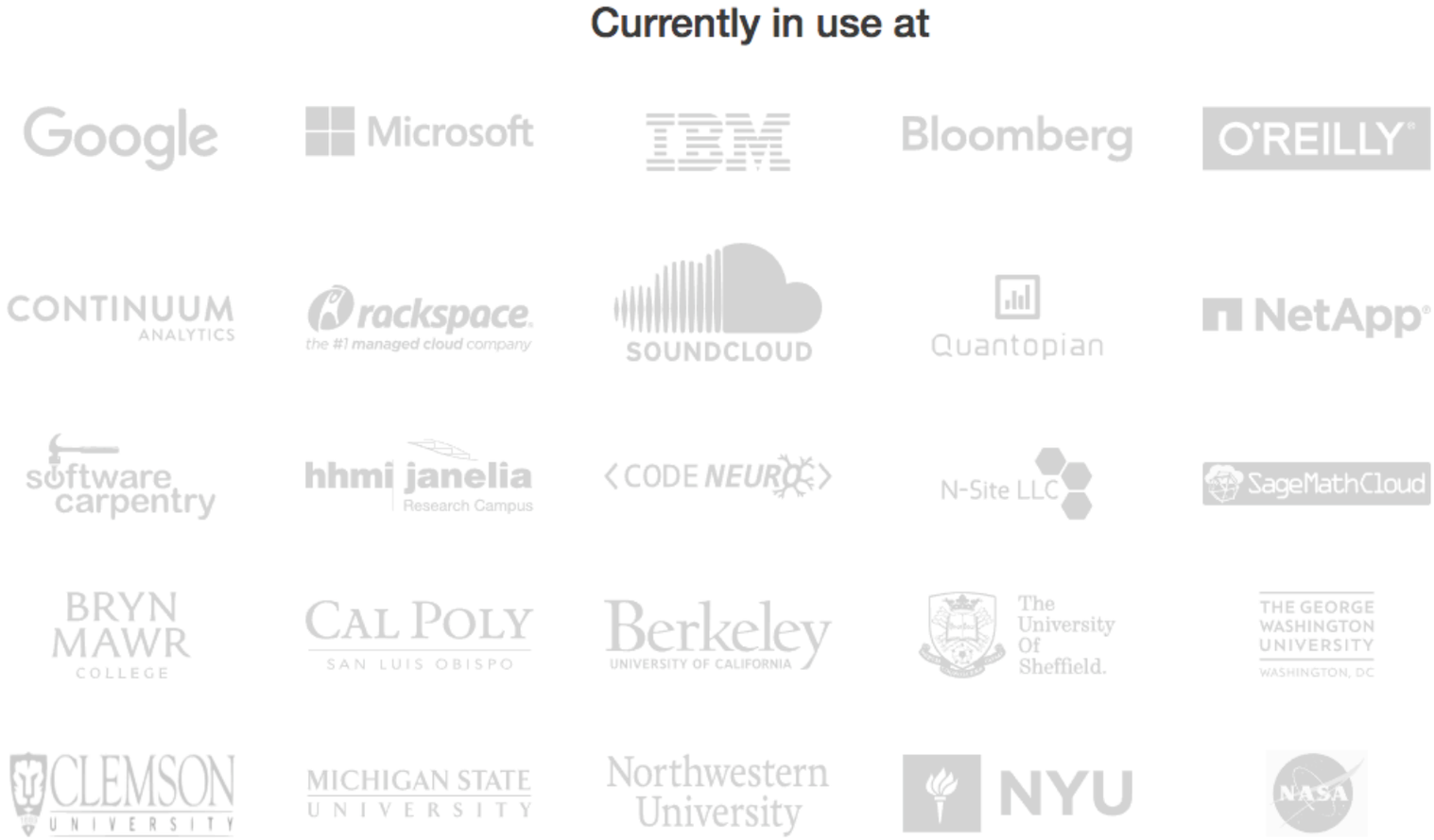
- The Anaconda distribution of Python includes:
  - Python interpreter
  - Jupyter notebook
  - Spyder IDE
  - Python libraries for data science
  - both versions 2.7 and 3.6. — we will use version 3.6
- Download and install from:

https://www.continuum.io/

Course Setup: Jupyter notebooks

- To glue components together and track results, we will execute all Python code in **jupyter notebooks**

- Why Jupyter?



Course Setup: Python

- Python is a ***prerequisite.***
  - Take the following 4-hour tutorial by DataCamp:

    https://www.datacamp.com/courses/intro-to-python-for-data-science

  - Familiarise yourself with the official Python 3 documentation:

    https://docs.python.org/3/

- Learn/review Python 3.x **this week**.
  - You'll be OK if you don't know Python - but do the bootcamp!
  - By next week it will be too late!

Conclusion

Should you take this course?

Do you really ***want*** to take this course?

Do you really ***want*** to take this course?

- This course will be hard.
- For some, this course will be ***very*** hard.
- I guarantee that the course will be worth it.

Do this **TODAY!**

- Prepare git
  - Install **git** (https://git-scm.com/) and learn basic commands
  - Signup at GitHub.com & email course assistant your GitHub **username**
- Review the course repository (in particular read **`setup.md`**):
  - Browse repo at https://github.com/epfl-tis-class/mgt-432-fall-2017
  - Clone repo to your local machine

```
git clone https://github.com/epfl-tis-class/mgt-432-fall-2017
```

- Install Anaconda
  - We do not support Windows (actually, in general we don't really do OS/system admin)
- Start on the Python MOOC & other tutorials
  - https://www.datacamp.com/courses/intro-to-python-for-data-science