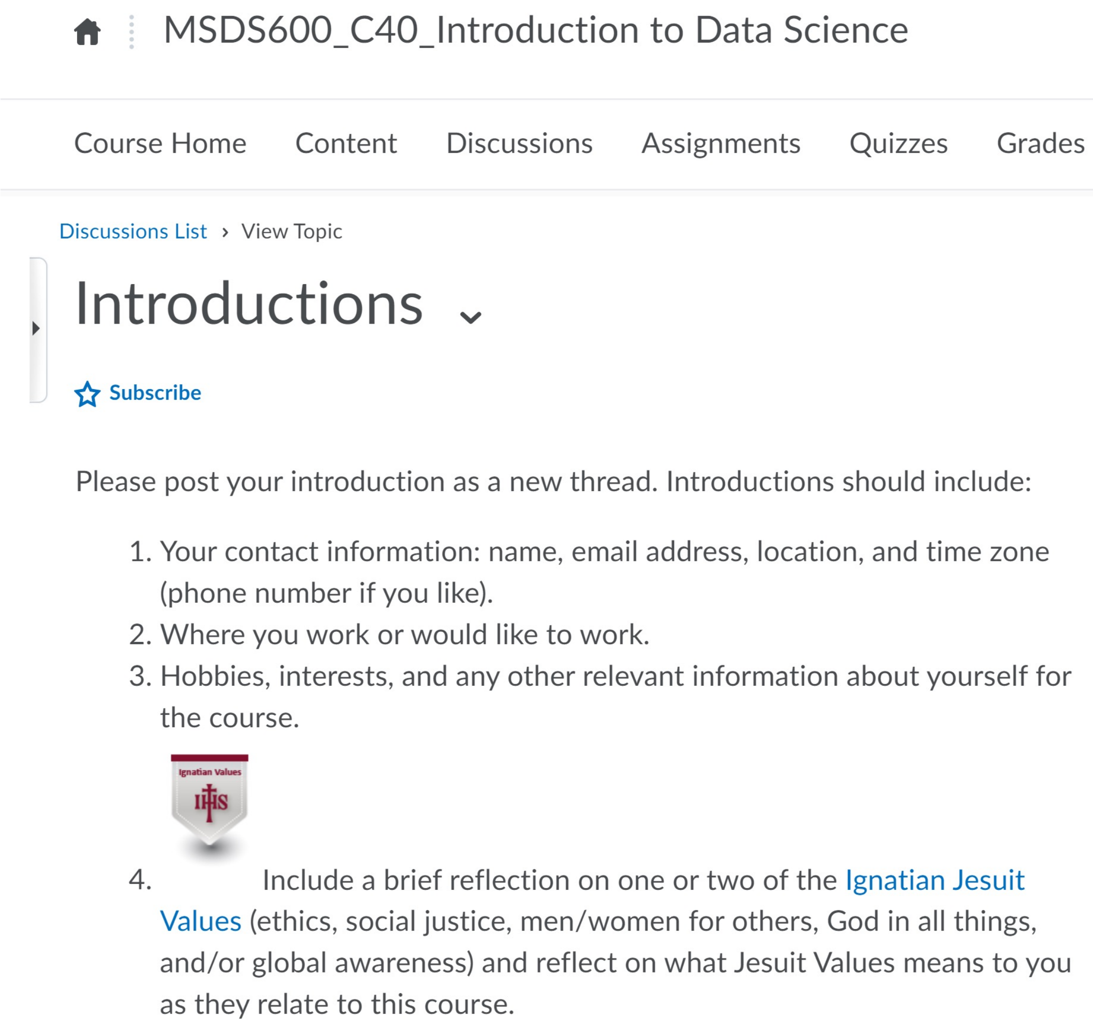
MSDS600: Introduction to Data Science

REGIS UNIVERSITY

Introductions

- In each class, we will introduce ourselves
  - If taking an online section, be sure to introduce yourself in the discussions
- Most students in MSDS course are MSDS students or certificate students
- Some other Regis master's programs allow students to take MSDS courses as electives or a specialization

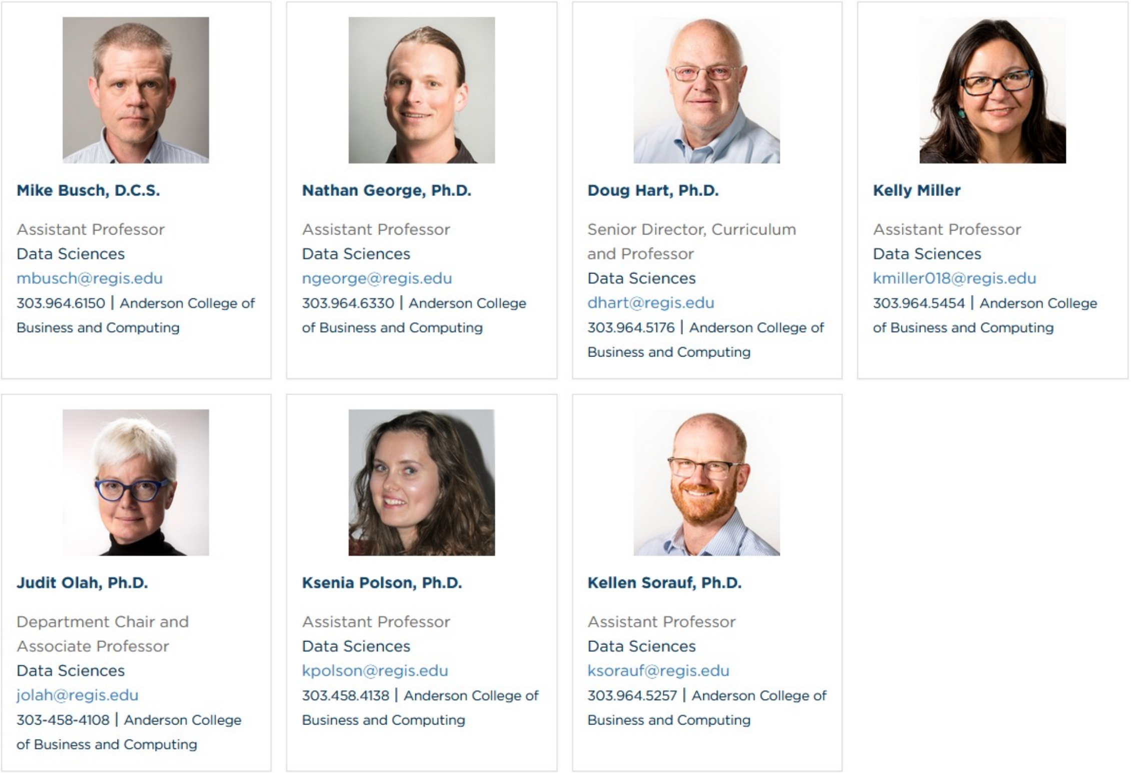


Please post your introduction as a new thread. Introductions should include:

1. Your contact information: name, email address, location, and time zone (phone number if you like).
2. Where you work or would like to work.
3. Hobbies, interests, and any other relevant information about yourself for the course.
4. Include a brief reflection on one or two of the Ignatian Jesuit Values (ethics, social justice, men/women for others, God in all things, and/or global awareness) and reflect on what Jesuit Values means to you as they relate to this course.

MSDS Full-time Faculty

- **Mike Busch** – Full-time MSDS/MSDE faculty and undergraduate data science program director
- **Nate George** – Former full-time MSDS faculty and program director. Recently a job in industry.
- **Kellen Sorauf** – Half-time professor, half-time data director for Anderson College, program director MSDS program
- **Ksenia Polson** – Full time MSDS faculty
- **Doug Hart** – Teaches some MSDS courses (mainly MSDS684 and practicum courses), created the MSDS program
- **Janet Houser** - Part-time MSDS faculty
- **Kelly Miller** – Undergrad HIM program manager
- **Judit Olah** – MSHI program director, DS dept. chair



Mike Busch, D.C.S.

Assistant Professor
Data Sciences
mbusch@regis.edu
303.964.6150 | Anderson College of Business and Computing

Nathan George, Ph.D.

Assistant Professor
Data Sciences
ngeorge@regis.edu
303.964.6330 | Anderson College of Business and Computing

Doug Hart, Ph.D.

Senior Director, Curriculum and Professor
Data Sciences
dhart@regis.edu
303.964.5176 | Anderson College of Business and Computing

Kelly Miller

Assistant Professor
Data Sciences
kmiller018@regis.edu
303.964.5454 | Anderson College of Business and Computing

Judit Olah, Ph.D.

Department Chair and Associate Professor
Data Sciences
jolah@regis.edu
303-458-4108 | Anderson College of Business and Computing

Ksenia Polson, Ph.D.

Assistant Professor
Data Sciences
kpolson@regis.edu
303.458.4138 | Anderson College of Business and Computing

Kellen Sorauf, Ph.D.

Assistant Professor
Data Sciences
ksorauf@regis.edu
303.964.5257 | Anderson College of Business and Computing

From https://www.regis.edu/academics/faculty-finder/index
Choose the ‘data sciences’ department

MSDS Affiliate faculty

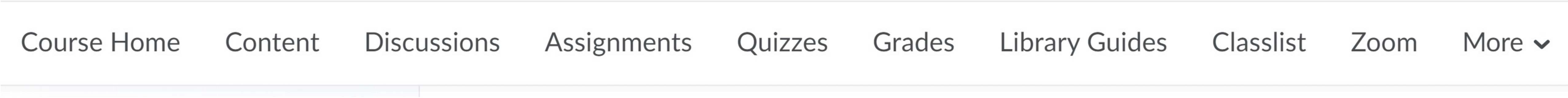
Many faculty from a variety of backgrounds

Name	Expertise	Profession & Workplace	Highest Degree	typical classes taught
Genie Hays	SQL, GIS, database management, business intelligence reporting	Department of Interior contractor	MS CIS	MSDS600, 650
Robin Kurth	Business intelligence, introductory Python	Affiliate Faculty, Regis, Johnson & Wales	MBA in Finance	MSDS600, 650
Kal Rosa	Databases, SQL, Python	Systems Analyst/Administrator at Intermountain Healthcare	6 MS degrees, including MSDS from Regis	MSC575, MSDS650
Paul Andrus	SQL, Python	IT Security Data Analyst at Oracle	MSIA and MSDS from Regis	MSDS610, 692/696, most classes
Rob Osterburg	AI, software engineering	Affiliate Faculty, Regis	MS Computer Science	MSDS688
Siri Sanguansintukul	classic statistics, R	Affiliate Faculty, Regis	PhD Information Science	MSDS660, 664, 680
Mike Prasad	APA writing, databases, Java, software development	Lead Developer/Architect at CO Dept of State	MS Computer Science	MSDE electives
Aiman Gannous	machine learning, computer science	Affiliate Faculty at Regis, CU, DU	PhD Computer Science	MSDS662, 631, 692/696
Christy Pearson	Python, machine learning	Machine Learning Engineering at DHI Group	MSDS from Regis	MSDS680, MSDS692/696, all classes
Donnie Kirk	GIS	GIS Analyst at Denver Regional Council of Governments	MS Geomorphology	MSDS674, MSDS655
Kamga Ngameni	Petroleum engineering, data science	Affiliate Faculty, Regis	PhD Petroleum Engineering	MSDS600, MSDE631
Kevin McBeth	data engineering	Senior Machine Learning Engineer at Seagate	MSDS from Regis	MSDS610, MSDS692/696
Aiman Darwiche	software engineering	Chief Data Scientist/Researcher at Compu-House	PhD Computer Science	MSDS640
John Koenig	data visualization, healthcare, VR	Data Science consultante	MBA	MSDS670
Ernest Green	data science	Data Scientist consultant (MITRE)	MS Predictive Analytics	MSDS696
Don Dalton	data science, deep learning	Principle Engineer, Atlantic Tele-network intl	MS Analytics	MSDS686, MSDS696

Navigating worldclass.regis.edu

The parts of worldclass we will use most are:

- Content – usually has assignments and reading materials
- Discussions – notes from instructor, weekly discussion topics
- Assignments – the ‘dropbox’ where you turn in assignments
- Quizzes – a few courses have quizzes (e.g. MSDE631)

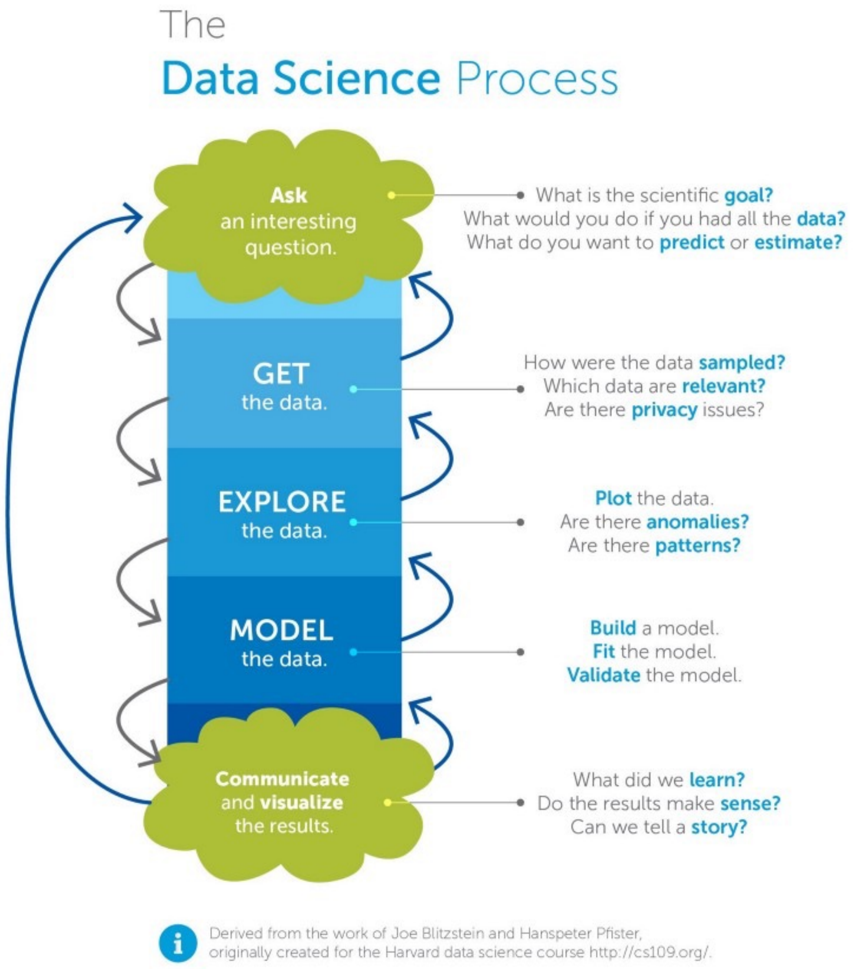


Regis MSDS benefits and resources

- Networking
  - Faculty, students, alumni, career center, regis.joinhandshake.com
- University resources
  - Library (O'Reilly/Safari link has many DS books), learning center (includes writing center with writing help), career, center, physical campus to study at and enjoy (the main Lowell campus is an arboretum), beautiful chapel with weekly masses
- Accountability
- Jesuit education – ethics, social good, whole education.
- Project-based learning

This week's content

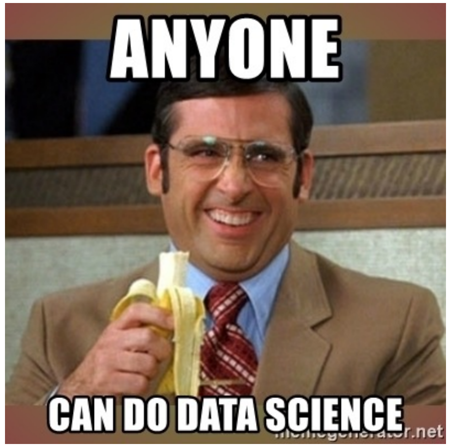
- Discuss applications of data science.
- Describe the data science project lifecycle.
- Distinguish the role of data scientists in different domains.
- Characterize big data using volume, variety, and velocity.
- Ethical concerns in big data & data science.
- Review / crash course on Python basics.


The
Data Science Process
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
GET the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
EXPLORE the data.
Plot the data.
Are there anomalies?
Are there patterns?
MODEL the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course http://cs109.org/.

https://web.archive.org/web/20190313043022/http://blog.operasolutions.com/bid/384900/what-is-data-science

What is data science?

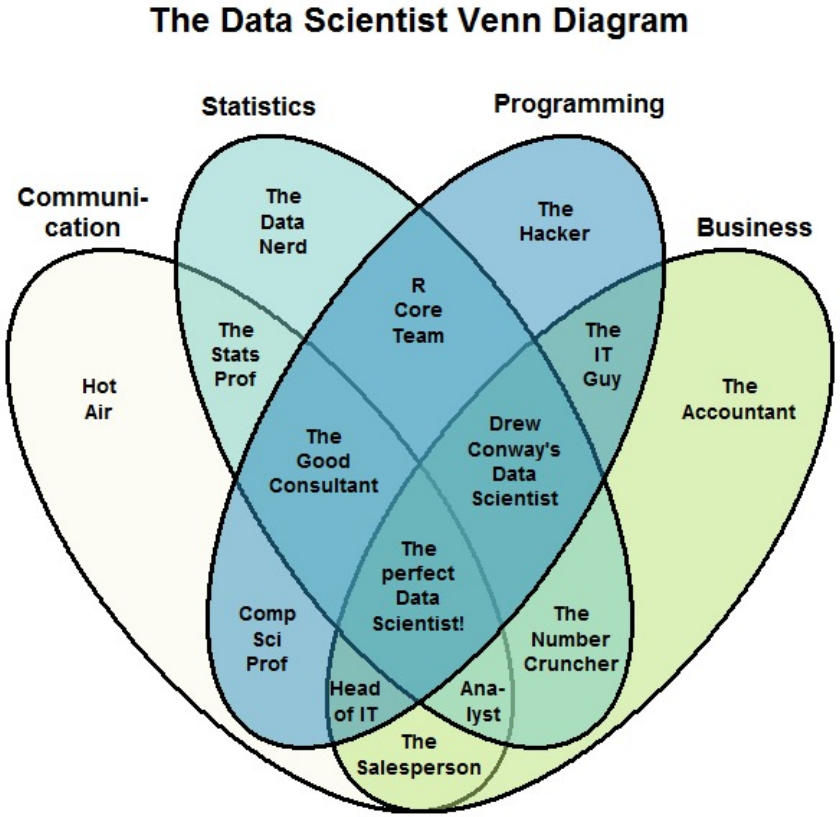


Too many Venn diagrams and opinions out there, but this one captures most of it. “Business” can also be thought of a “domain expertise” – that is, understanding of your specific use cases for DS.

Why use programming?

It’s hard to scale machine learning solutions without knowing some programming. Though it can be increasingly done, especially as cloud tools like Azure, GCP, and AWS evolve. State-of-the-art tools and techniques are usually more custom and done with programming.

Some say a key difference between an analyst and data scientist is the DS can program/code.



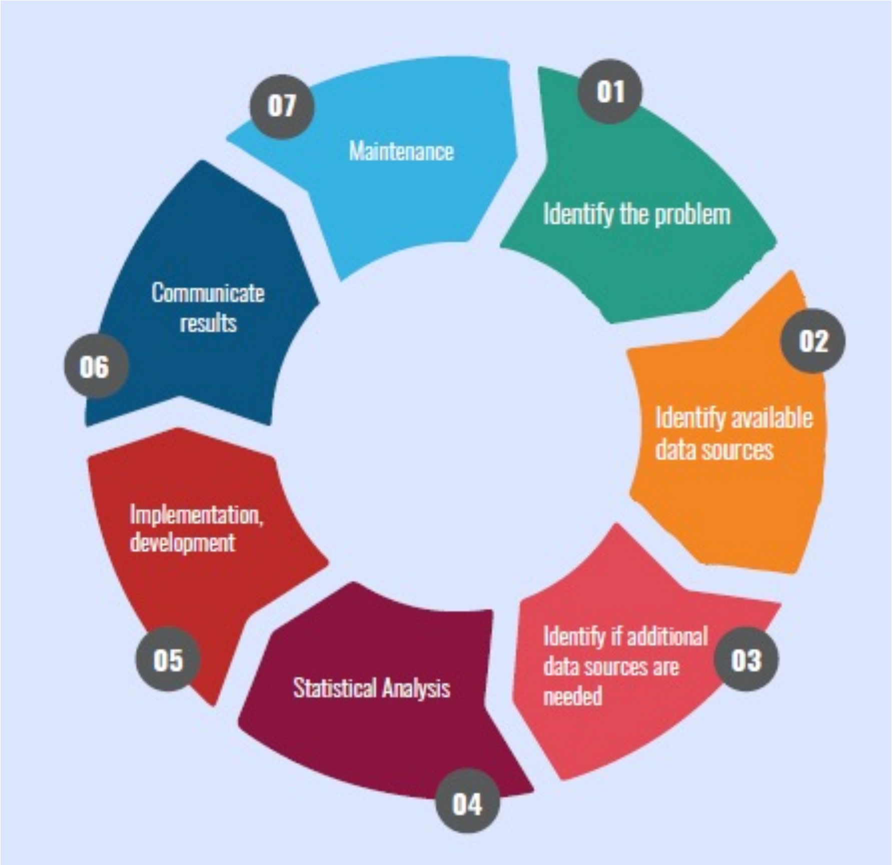
https://whatsthebigdata.com/2016/07/08/the-new-data-scientist-venn-diagram/

Data science process

DS includes work with machine learning much of the time. However, some work may involve other statistics or vizualizations.

Examples are:

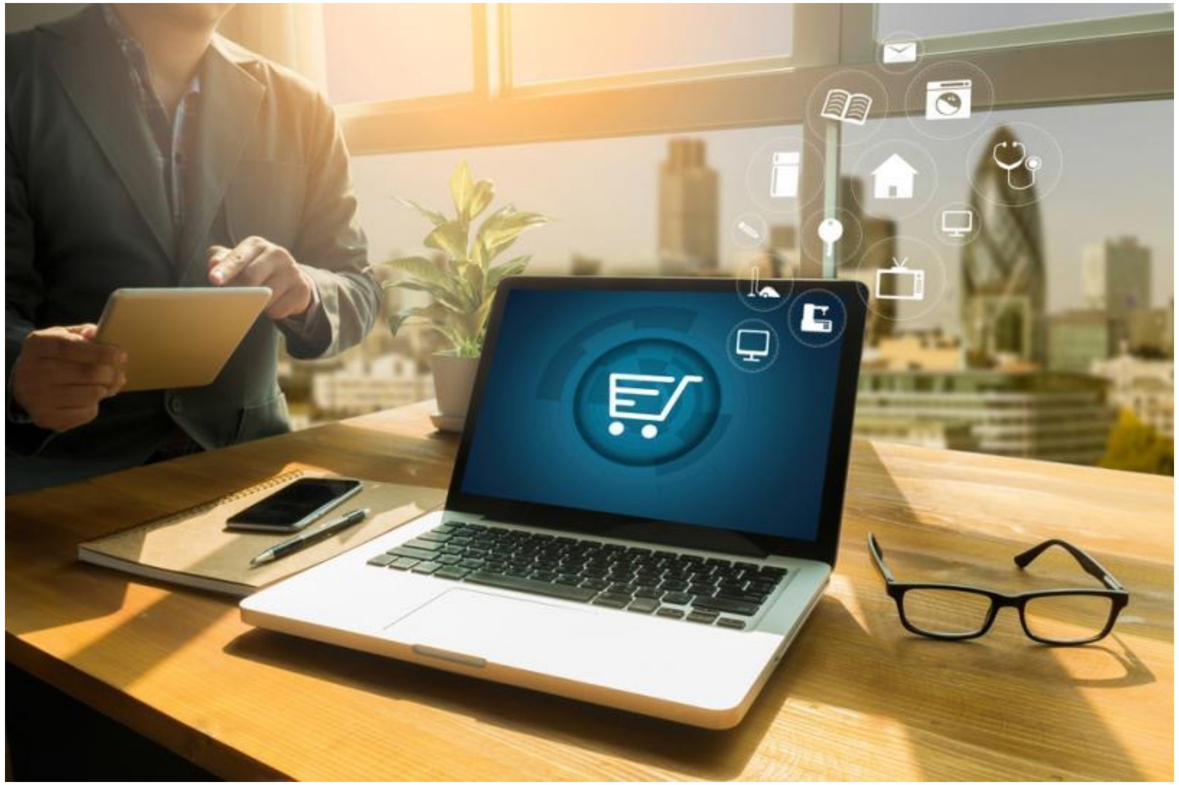
- Cluster analysis or clustering (e.g. customer segmentation)
- Anomaly detection (e.g. cybersecurity)
- Predictive analytics (e.g. predicting health outcomes)


01
Identify the problem
02
Identify available data sources
03
Identify if additional data sources are needed
04
Statistical Analysis
05
Implementation, development
06
Communicate results
07
Maintenance

Data science applications

Anything with data.

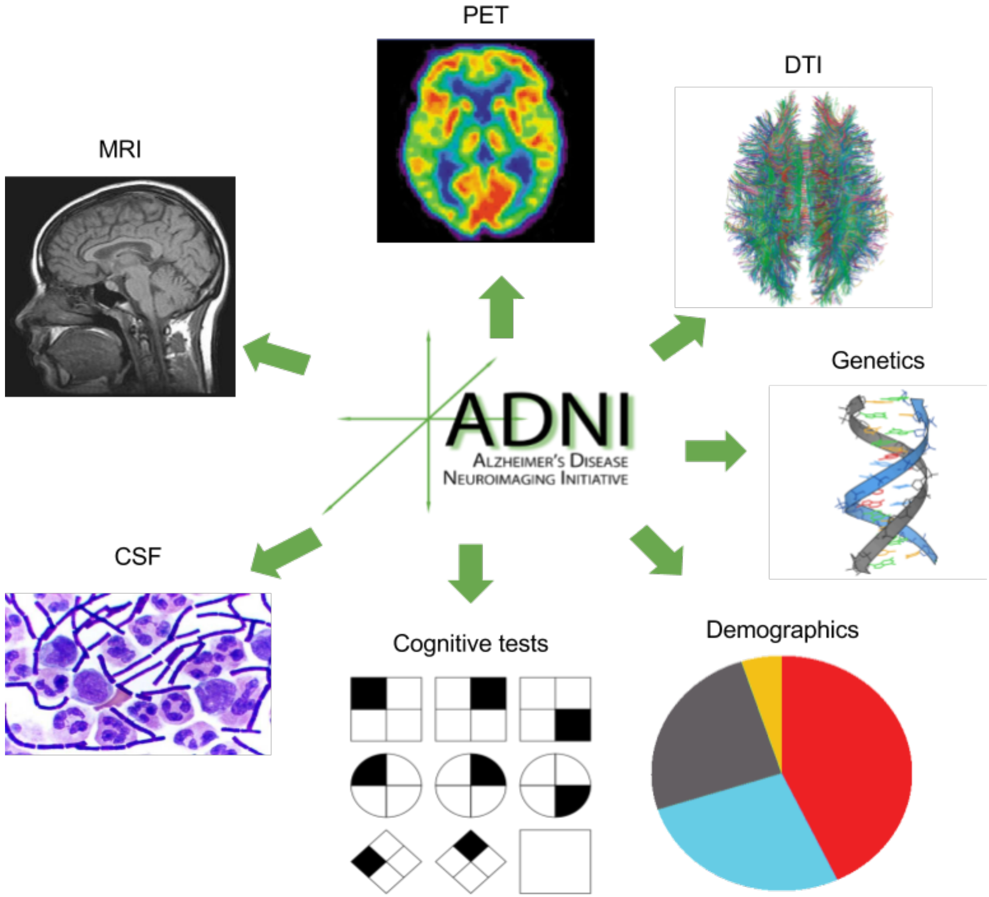
- Businesses: Sales, marketing, optimization, customer relations
- Manufacturing: Quality control, robotics
- The hard sciences: Materials science, (bio)chemistry, neuroscience, etc
- Tech: Fintech, web/phone apps, self-driving cars, NetFlix, etc
- Healthcare: Radiology, patient records, pharmaceuticals, etc
- Government: Taxes, security, military



https://builtin.com/data-science/data-science-applications-examples

Example: Predicting Alzheimer's Disease

Combines many different datasets to predict biomarkers and Alzheimer's disease onset



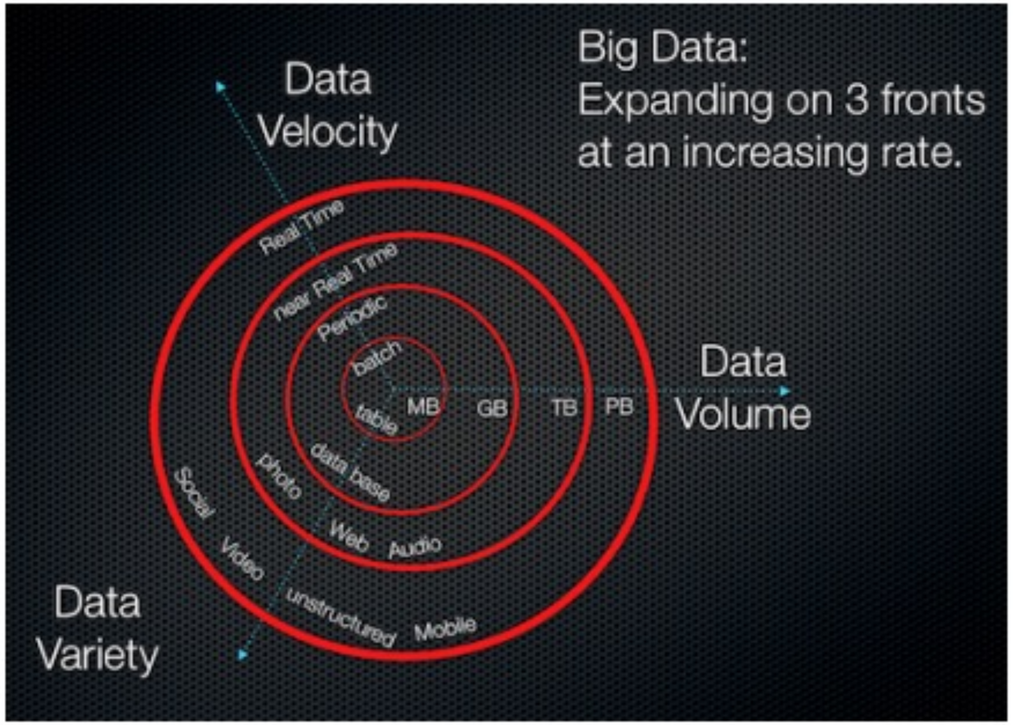
https://tadpole.grand-challenge.org/Background/

Big Data

- Volume: Size
- Variety: Type of data
- Velocity: How fast is it updated

Tools:

- Spark (pyspark, sparklr)
- Hadoop
- Cloud (AWS, GCP)
- Python big data packages (dask, H2O, more [1] [2])
- R also has big data libraries

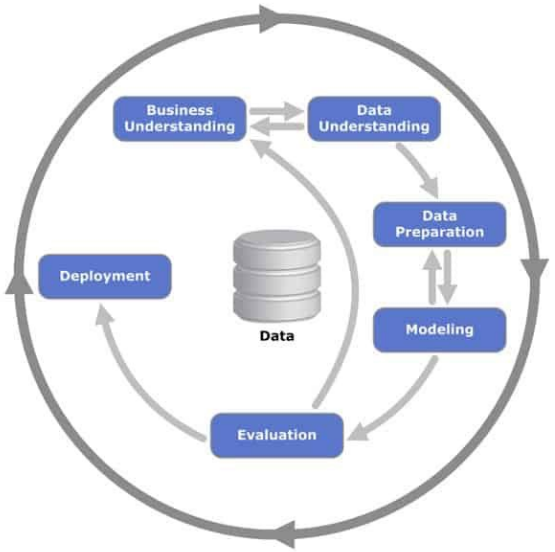


http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

The data science process

CRISP-DM, TDSP, others

Polls show 25-75% of time is spent cleaning and preparing data. Many data scientists report they even spend up to 90% of their time cleaning/preparing data. Some of this work is being moved into data engineering jobs.



Business Understanding

Data Understanding

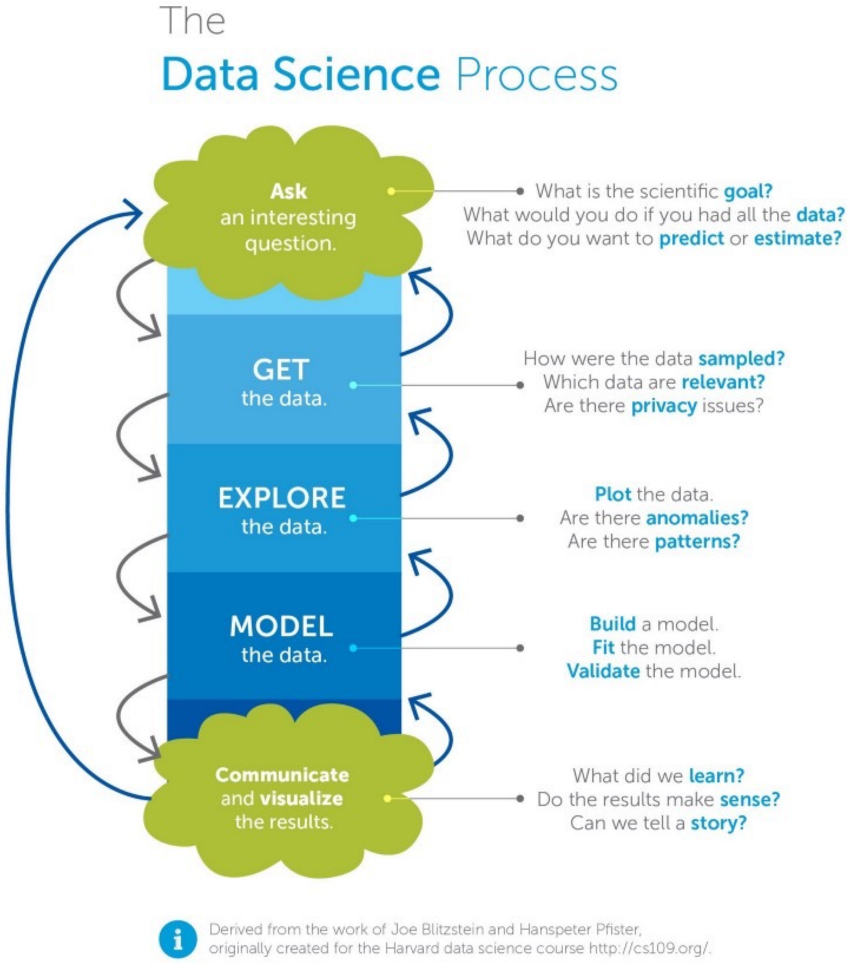
Data Preparation

Modeling

Evaluation

Deployment

Data



The
Data Science Process

Ask an interesting question.

What is the scientific **goal?**
What would you do if you had all the **data?**
What do you want to **predict** or **estimate?**

GET the data.

How were the data **sampled?**
Which data are **relevant?**
Are there **privacy** issues?

EXPLORE the data.

Plot the data.
Are there **anomalies?**
Are there **patterns?**

MODEL the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn?**
Do the results make **sense?**
Can we tell a **story?**

Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course http://cs109.org/.

http://www.datascience-pm.com/

Ethical Concerns

Privacy

- Data collection
- Data sharing
- Location services

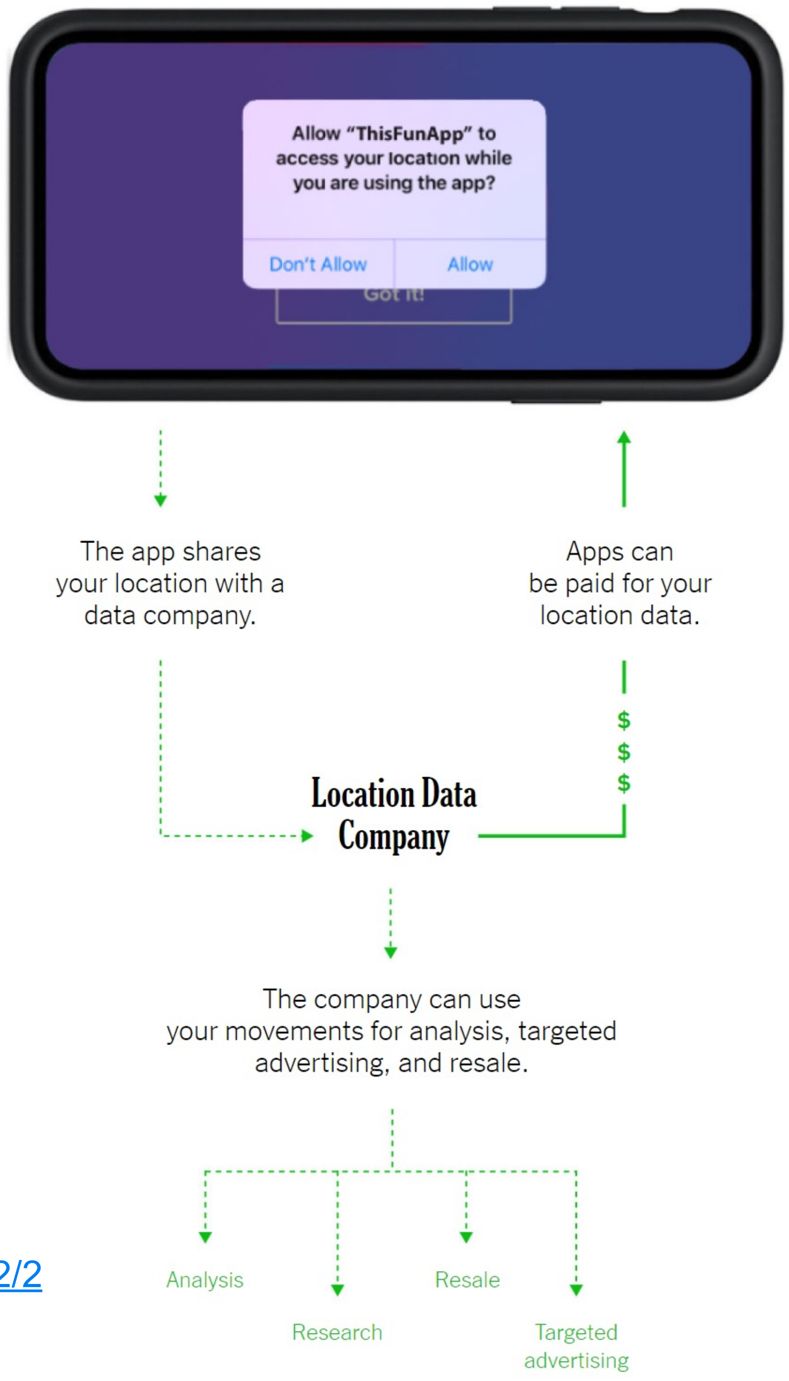
Bias in AI

- Machine learning
- Facial recognition
- Court systems
- Loan systems

Article on phone tracking and privacy.

More will be covered in MSDS640 and other courses.

https://www.nytimes.com/interactive/2019/12/20/opinion/location-tracking-smartphone-marketing.html



Allow "ThisFunApp" to access your location while you are using the app?

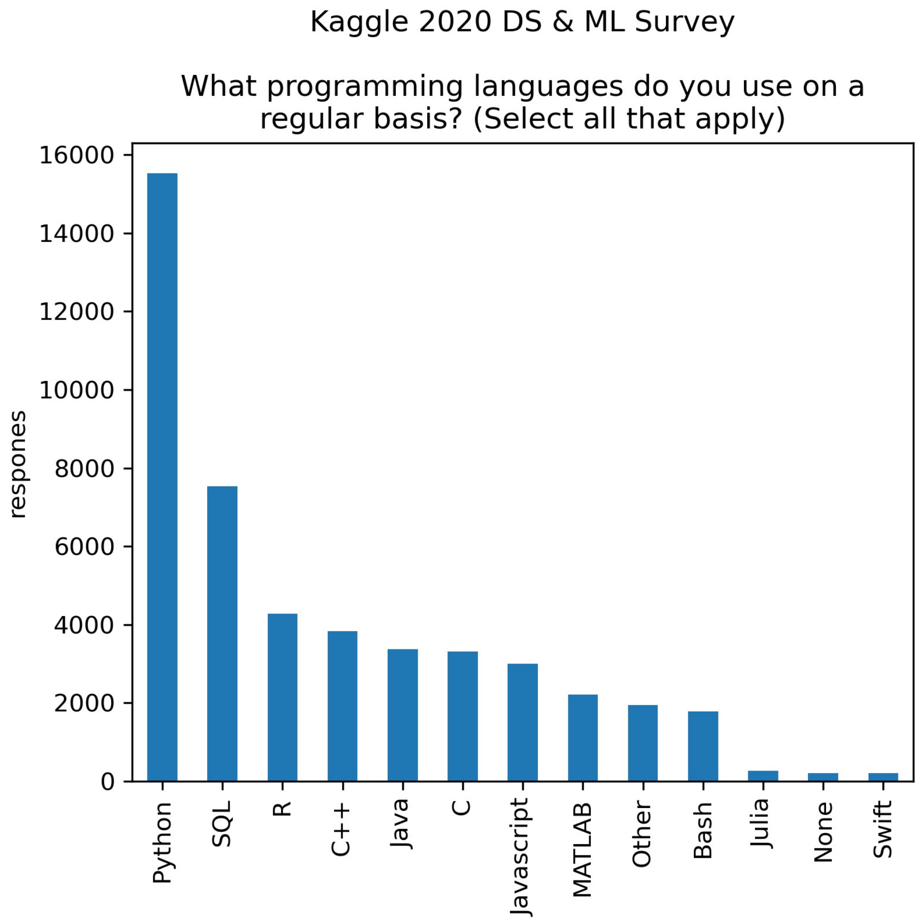
The app shares your location with a data company.

Apps can be paid for your location data.

Location Data Company

The company can use your movements for analysis, targeted advertising, and resale.

Data science tools



Python seems to be the top data science programming language.

Python is (mostly) object-oriented (OOP), while R is a functional language. Most “industrial strength” coding languages like Java are OOP, making it easier to translate Python code into Java or use them together.

Python also has many libraries and a big network effect. So does R. The two are different tools for different tasks, although both can technically be used for anything.

SQL is important to know since it’s the standard database language.

Python is better for:

Neural networks, serving websites, interacting with the cloud

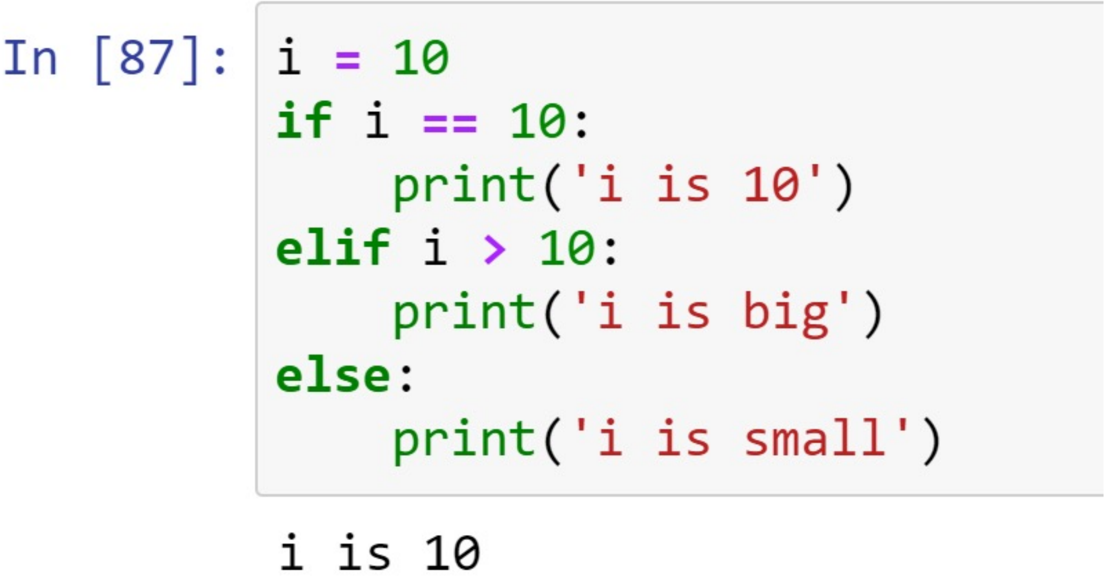
R is better for:

Classic statistics tests, certain tasks where libraries are pre-built

Why Python

Python is meant to be fun, simple, clear, and easy to use. There are some parts of “The Zen of Python” that describe this:

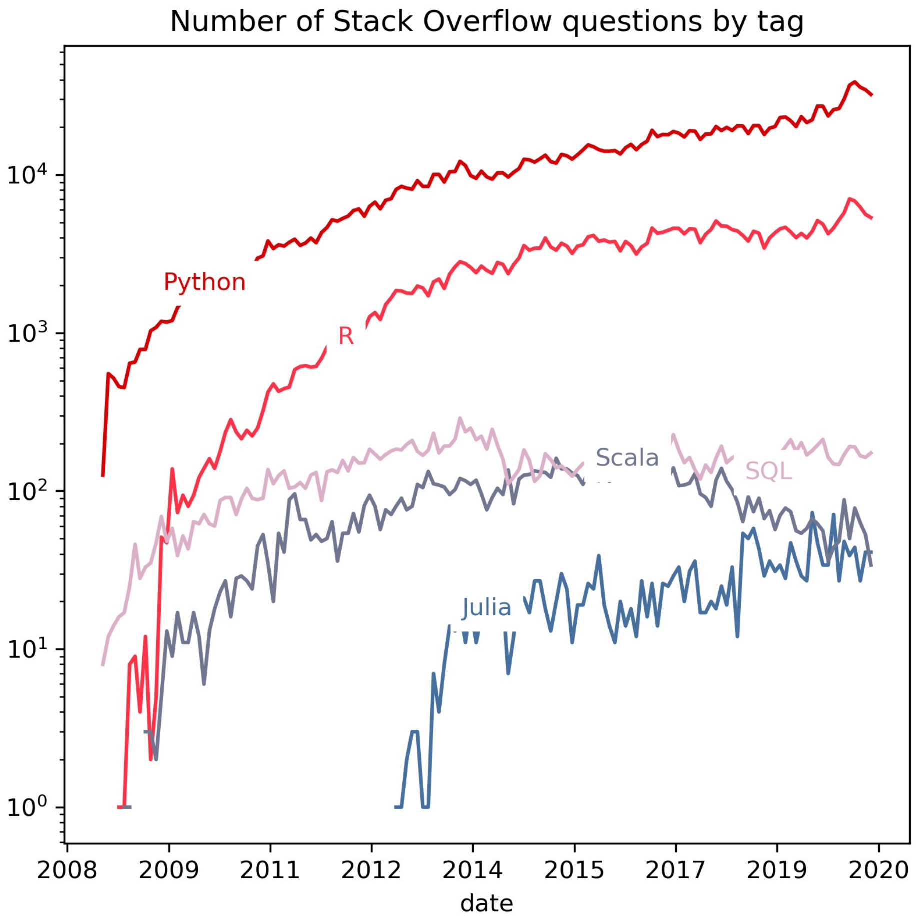
- Beautiful is better than ugly.
- Explicit is better than implicit.
- Simple is better than complex.
- Complex is better than complicated.
- Readability counts.
- "there should be one— and preferably only one — obvious way to do it"



```
In [87]: i = 10
         if i == 10:
             print('i is 10')
         elif i > 10:
             print('i is big')
         else:
             print('i is small')

         i is 10
```

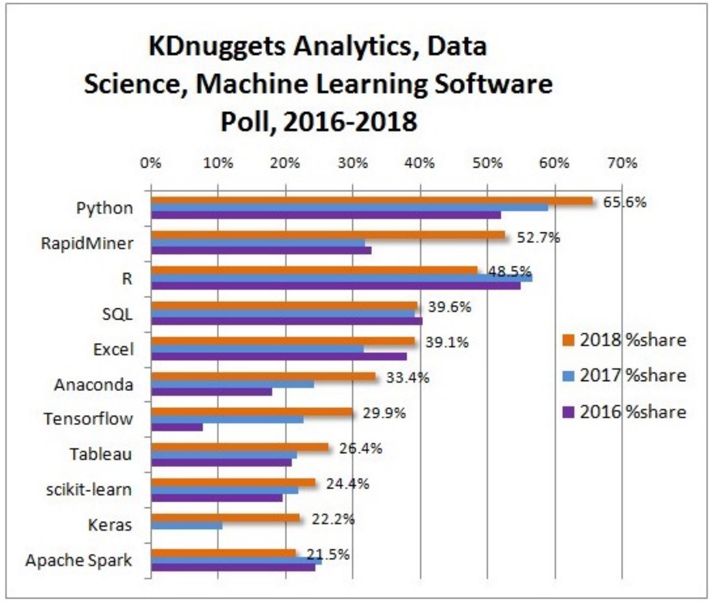
Network effects



The more users, the bigger the community and more support available. More tools are built.

Julia is a newer language, and may be more viable as a top data science language in 5-10 years.

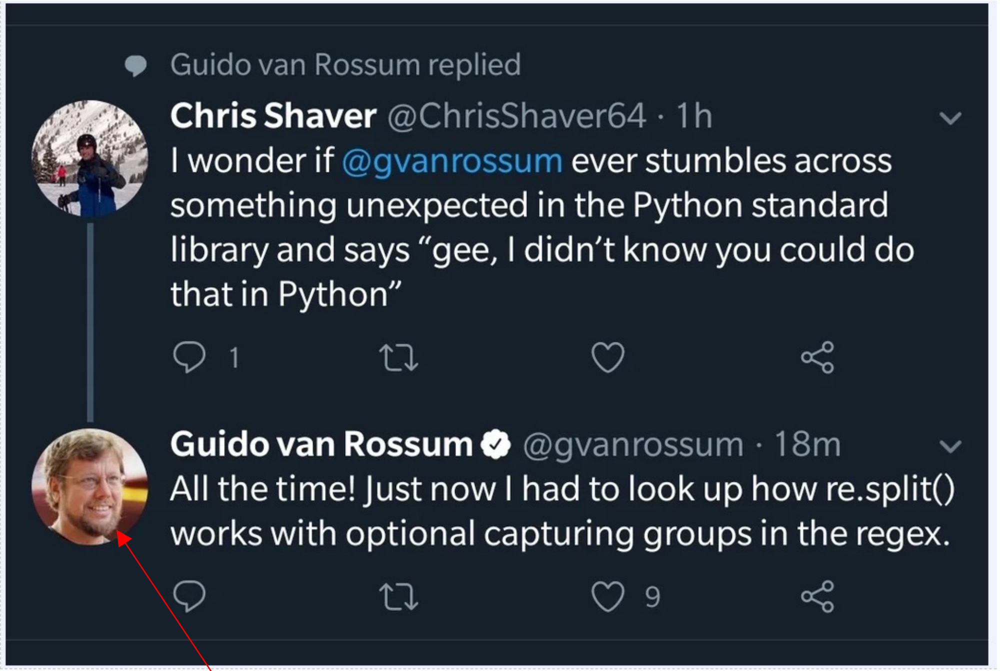
Many other tools out there to do data science with, including GUIs such as RapidMiner and Excel.



https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

Resources for the program and Python/R

- Your professor/instructor is here to help
- Your advisor and success coach
- Other students can help too (e.g. via discussions)
- Regis Learning Commons
  - Writing Center and Tutoring
- Documentation (e.g. docs.python.org)
- Stack Overflow
- Search engines (google, duckduckgo)
- Archives for broken links (archive.is, archive.org)



Guido van Rossum replied

Chris Shaver @ChrisShaver64 · 1h

I wonder if @gvanrossum ever stumbles across something unexpected in the Python standard library and says "gee, I didn't know you could do that in Python"

Guido van Rossum @gvanrossum · 18m

All the time! Just now I had to look up how re.split() works with optional capturing groups in the regex.

Python benevolent dictator for life (Resigned in July 2018)

Docs and books

Using a search engine and stackoverflow is part of writing code and doing data science!

Books (e.g. through the library and O'Reilly Safari) are a good way to understand fundamentals of statistics, machine learning, and specialized topics.

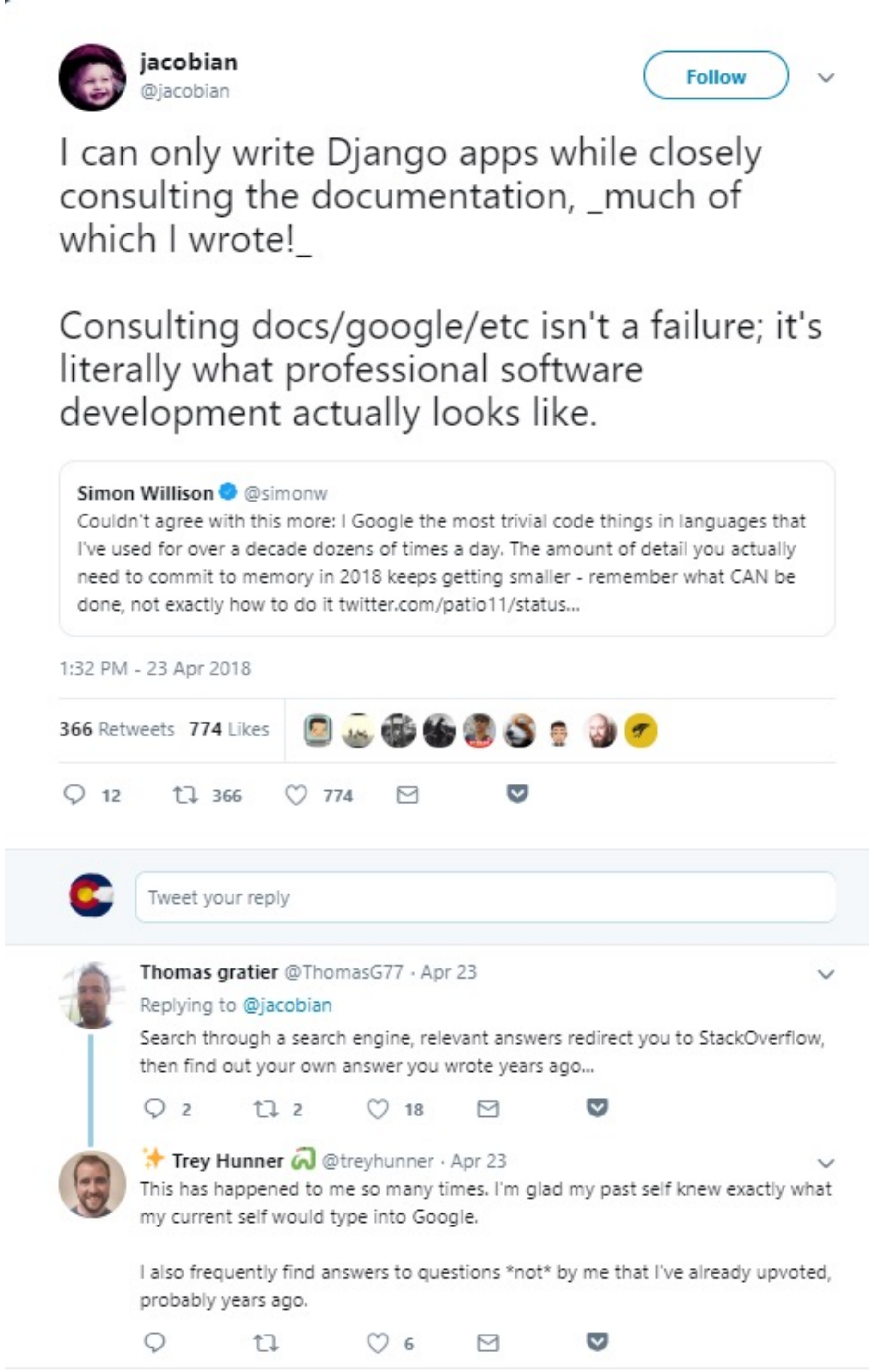
ISLR (stats in R)

ESL (stats methods and math)

Python Machine Learning - Third Edition (Packt)

Clean code in Python – Second Edition (Packt)

R for Data Science



jacobian
@jacobian

I can only write Django apps while closely consulting the documentation, _much of which I wrote!_

Consulting docs/google/etc isn't a failure; it's literally what professional software development actually looks like.

Simon Willison @simonw
Couldn't agree with this more: I Google the most trivial code things in languages that I've used for over a decade dozens of times a day. The amount of detail you actually need to commit to memory in 2018 keeps getting smaller - remember what CAN be done, not exactly how to do it twitter.com/patio11/status...

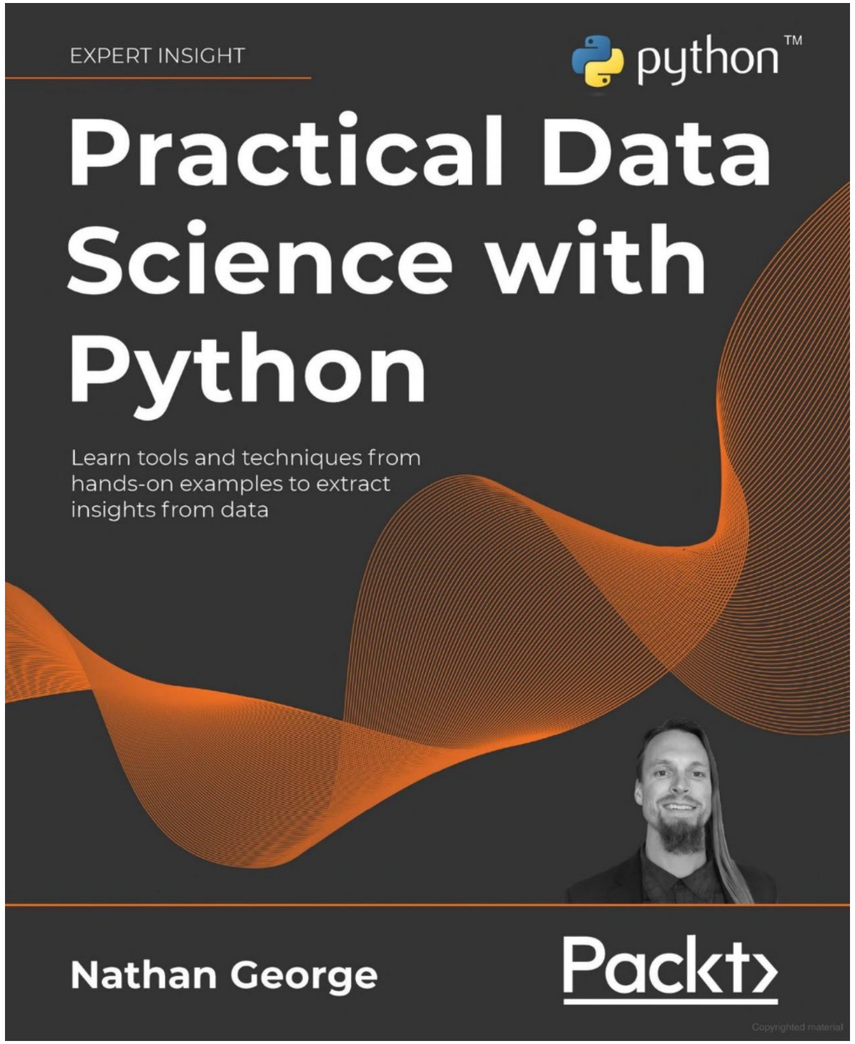
1:32 PM - 23 Apr 2018

366 Retweets 774 Likes

Thomas gratier @ThomasG77 · Apr 23
Replying to @jacobian
Search through a search engine, relevant answers redirect you to StackOverflow, then find out your own answer you wrote years ago...

Trey Hunner @treyhunner · Apr 23
This has happened to me so many times. I'm glad my past self knew exactly what my current self would type into Google.

I also frequently find answers to questions *not* by me that I've already upvoted, probably years ago.



EXPERT INSIGHT

python™

Practical Data Science with Python

Learn tools and techniques from hands-on examples to extract insights from data

Nathan George

Packt

https://www.packtpub.com/product/practical-data-science-with-python/9781801071970

Assignment for this week

The “from the expert” Jupyter Notebook covers a basic Python review and crash course. It also demos some basice EDA and data understanding.

Use this demo with your reading materials (and ask for help from the instructor if you need it) to complete the assignment, which is a similar task on different datasets.

Reading list (available through O’Reilly through the library):

[1] Python Data Science Essentials - Third Edition by Alberto Boschetti and Luca Massaron.
Sections:

- First Steps (“First Steps” through “Alternatives to Jupyter”)
- Strengthen Your Python Foundations (“Strengthen your Python Foundations” through “Don’t be shy, take a real challenge”)

[2] Python for Data Science For Dummies, 2nd Edition by John Paul Mueller and Luca Massaron.

- Chapters 1 and 2.

Extra Resources for Learning Python

Python is a foundational skill for much of modern data science. It is worth putting in some time to build solid Python skills.

If you are still learning Python or feeling shaky on the basics (even after doing the readings), consider doing the following:

Complete most of the Udacity Intro to Python Course.

- Weeks 1 and 2: Lessons 1 and 2
- Week 3 and 4: Lesson 3
- Week 5 and 6: Lesson 4
- Week 7 and 8: Lesson 5

If you are feeling relatively comfortable with Python, you might consider doing the Kaggle course on Python if you need a refresher: https://www.kaggle.com/learn/python

There are many other Python learning resources for all levels listed here: https://forums.fast.ai/t/recommended-python-learning-resources/26888

Ways to continue building your Python and R skills

- DataCamp
- DataQuest
- Hackerrank
- Reading the official Python documentation or documentation for other Python packages
- Kaggle (participating in competitions, looking at what others are doing)
- Codewars.com
- Books (e.g. Clean Code with Python 2nd Edition by Mariano Anaya), Minimal Python by Noah Gift and Alfredo Deza, Python Data Science Essentials by Alberto Boschetti, etc)