CAP 5768 – Intro to Data Science

Data Science: the big picture

Oge Marques, PhD

Professor

College of Engineering and Computer Science

College of Business

FAU FLORIDA ATLANTIC UNIVERSITY

@ProfessorOge

ProfessorOgeMarques

Questions

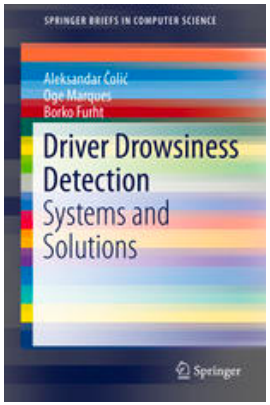
Who am I?

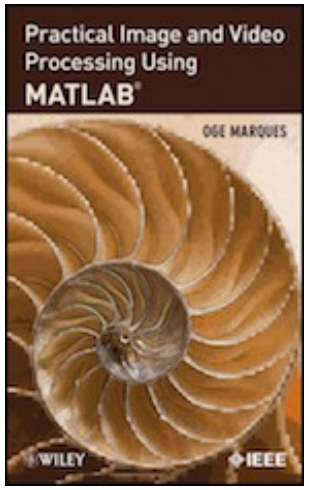
My research interests

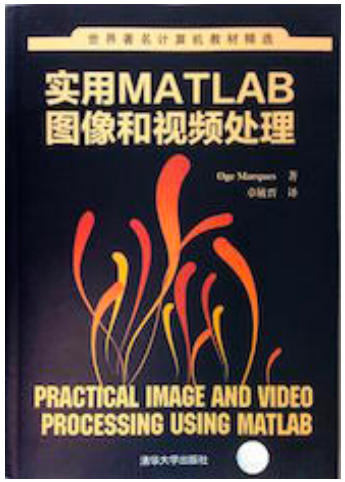
Intelligent processing of visual information

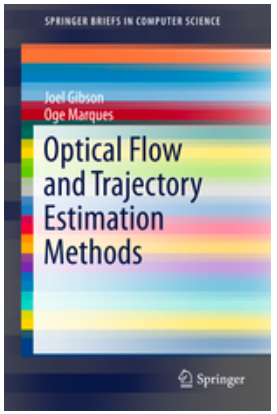
- image processing
- medical image analysis
- computer vision
- human vision
- artificial intelligence
- machine learning
- deep learning

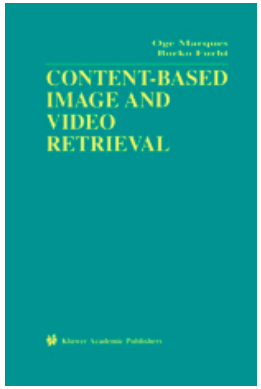
My work: selected books

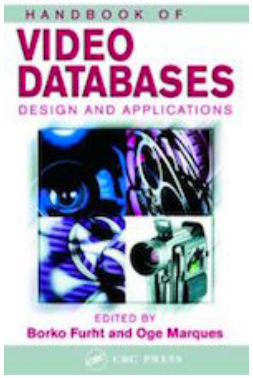


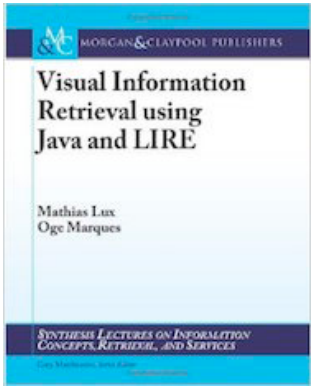




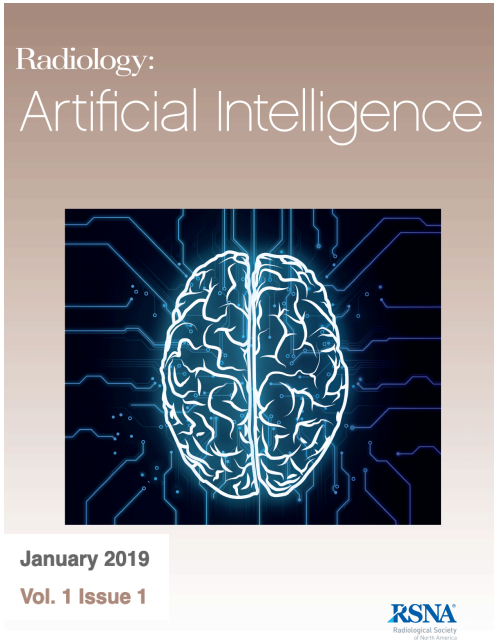


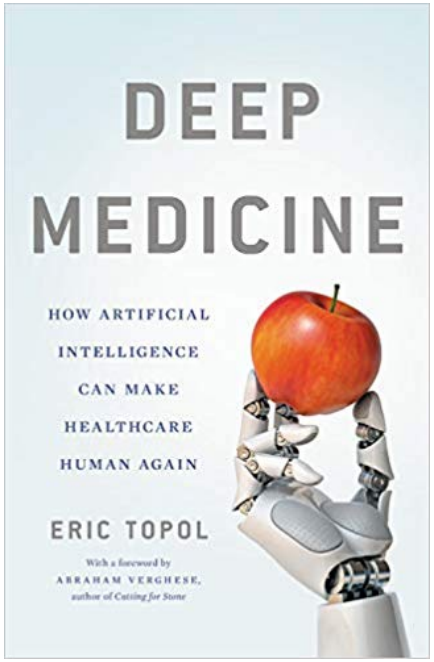






My work: current focus (AI and Medicine)



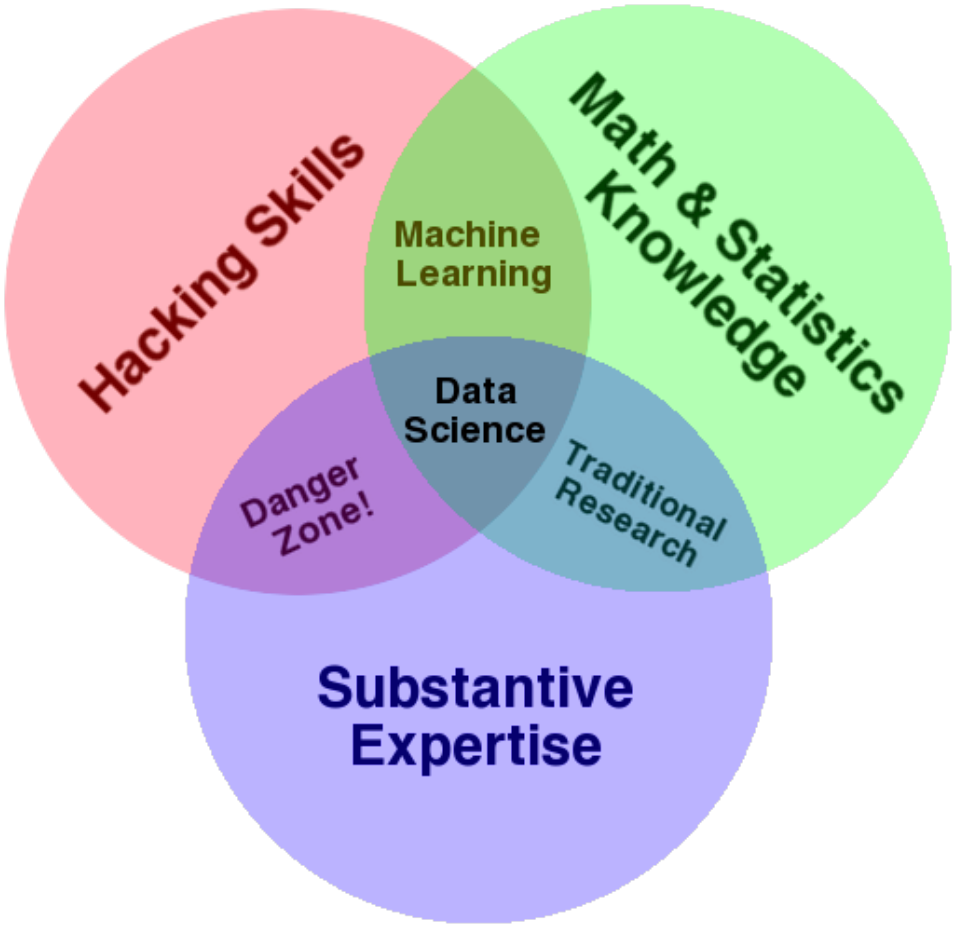


What is Data Science?

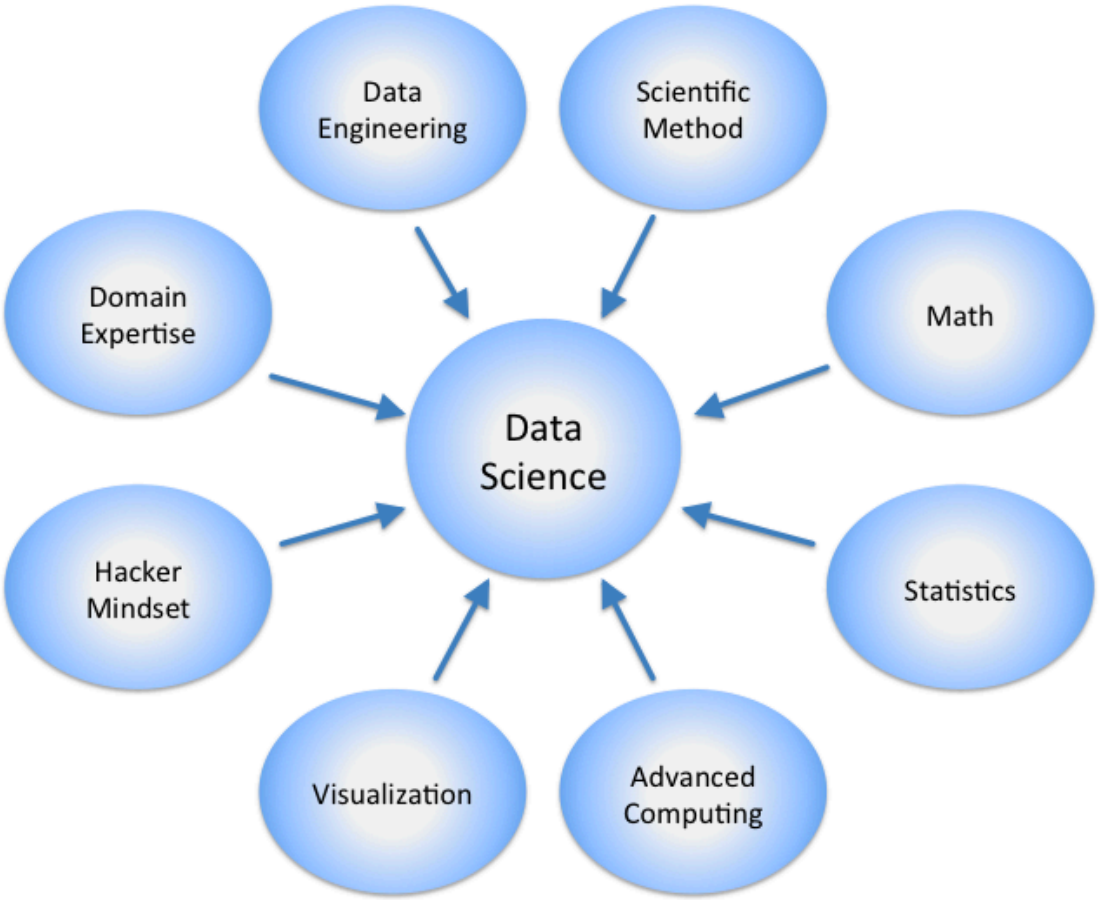
What is Data Science?

- “Data science […] is perhaps the best label we have for the ***cross-disciplinary set of skills*** that are becoming increasingly important in many applications across industry and academia.”

Source: Python Data Science Handbook by Jake VanderPlas (O’Reilly). Copyright 2017 Jake VanderPlas, 978-1-491-91205-8.



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



Source: https://upload.wikimedia.org/wikipedia/commons/4/44/DataScienceDisciplines.png

The goal of data science is to improve decision making by basing decisions on insights extracted from large data sets. As a field of activity, data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from large data sets. It is closely related to the fields of data mining and machine learning, but it is broader in scope. Today, data science drives decision making in nearly all parts of modern societies. Some of the ways that data science may affect your daily life include determining which advertisements are presented to you online; which movies, books, and friend connections are recommended to you; which emails are filtered into your spam folder; what offers you receive when you renew your cell phone service; the cost of your health insurance premium; the sequencing and timing of traffic lights in your area; how the drugs you may need were designed; and which locations in your city the police are targeting.

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

What is a Data Scientist?

What is a data scientist?

Searching the web for more information about the emerging term "data science," we encounter the following definitions from the Data Science Association's "Professional Code of Conduct"[6]

"Data Scientist" means a professional who uses scientific methods to liberate and create meaning from raw data.

50 Years of Data Science

David Donoho

To cite this article: David Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734

To link to this article: https://doi.org/10.1080/10618600.2017.1384734

What is a data scientist?

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

50 Years of Data Science

David Donoho

To cite this article: David Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734

To link to this article: https://doi.org/10.1080/10618600.2017.1384734

Fill in the blanks

Data scientists use data to…

- …
- …
- …
- …

Data scientists use data to…

Test hypotheses

Model processes

Predict outcomes

Detect anomalies

Train machine learning models

Extract information (and eventually knowledge)

Explain the world

What is the recipe for success in data science?

What is the recipe for success in data science?

Math / statistics knowledge

Hacking / coding skills

Subject-matter expertise

What is driving Data Science?

What is driving Data Science?

Enormous availability of (raw) data

Open source tools

Easy access to code and datasets

Numerous use cases

Faster / ubiquitous computing platforms

Lower barriers to enter

Why did you sign up for this course?


Data Science is a
hot field!

LinkedIn Workforce Report | United States | August 2018

Published on Aug 10, 2018

Skills Gaps | Demand for data scientists is off the charts

In 2015, there was a national surplus of people with data science skills. An employer in **Dallas** or **Atlanta** who wanted to hire data scientists had plenty of options; aside from in a few tech or finance-heavy cities like **San Francisco**, **New York City** and **Boston**, there weren't many local shortages.

But today, 3 years later, the picture has changed markedly: data science skills shortages are present in almost every large U.S. city. Nationally, we have a shortage of 151,717 people with data science skills, with particularly acute shortages in **New York City** (34,032 people), the **San Francisco Bay Area** (31,798 people), and **Los Angeles** (12,251 people). As more industries rely on big data to make decisions, data science has become increasingly important across all industries, not just tech and finance. In that sense, it's a good proxy for how today's cutting-edge skills like AI & machine learning may spread to other industries and geographies in the future.

https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018

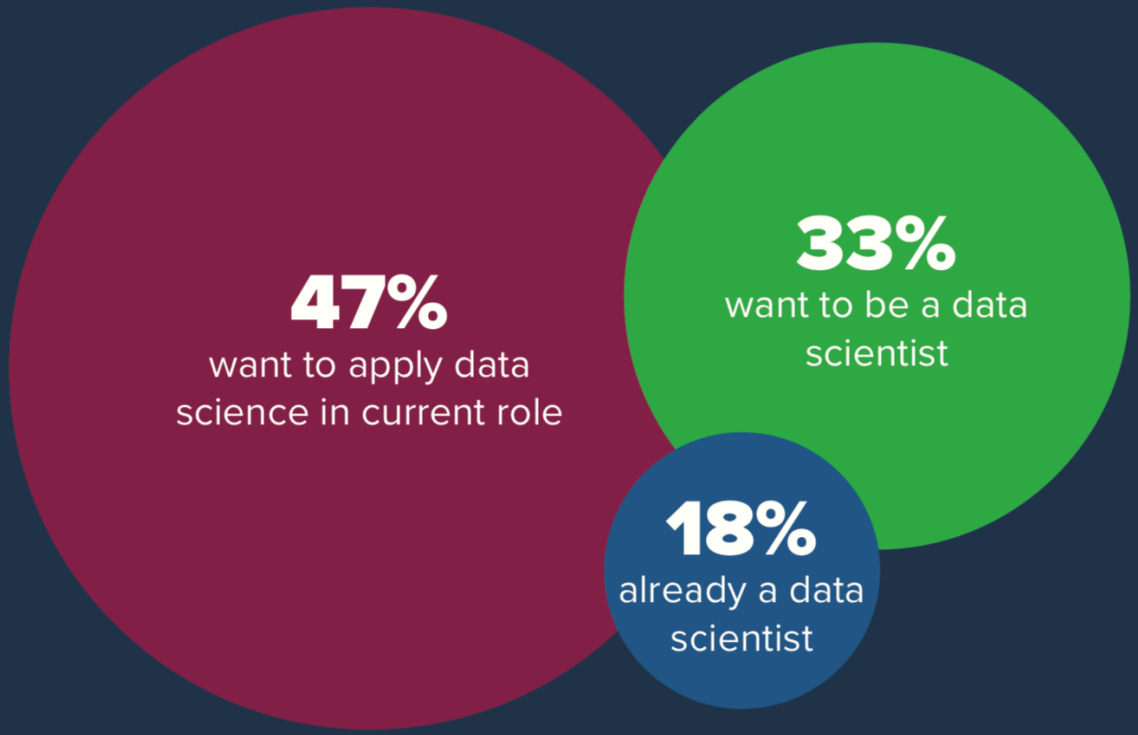
2019 STATE OF DATA SCIENCE

This spring, we surveyed nearly 5,000 members of the Anaconda community to understand current trends in data science. Here's what we found.

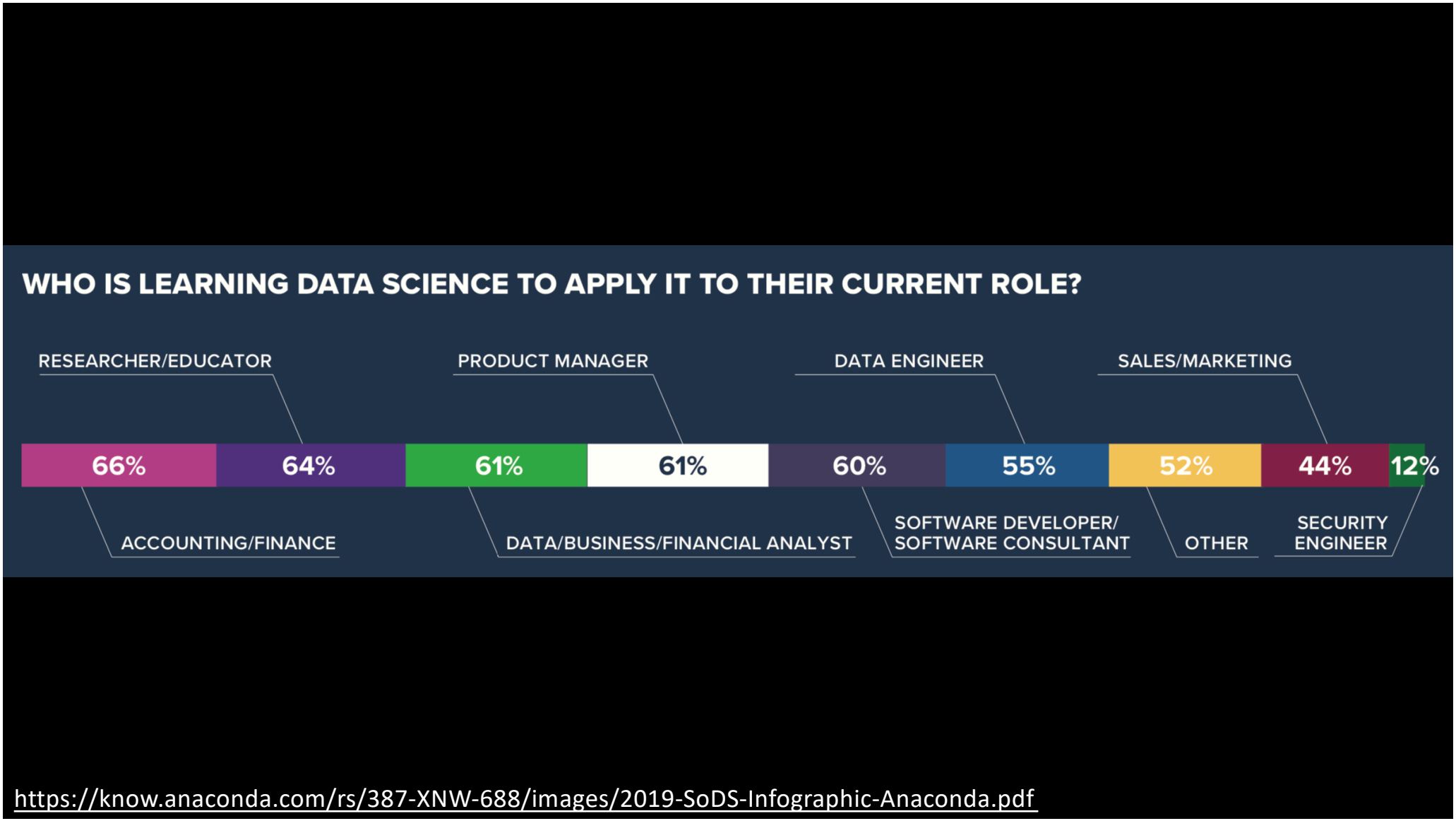
https://know.anaconda.com/rs/387-XNW-688/images/2019-SoDS-Infographic-Anaconda.pdf

PARADIGM SHIFT: DATA SCIENCE WILL IMPACT ALL BUSINESS ROLES

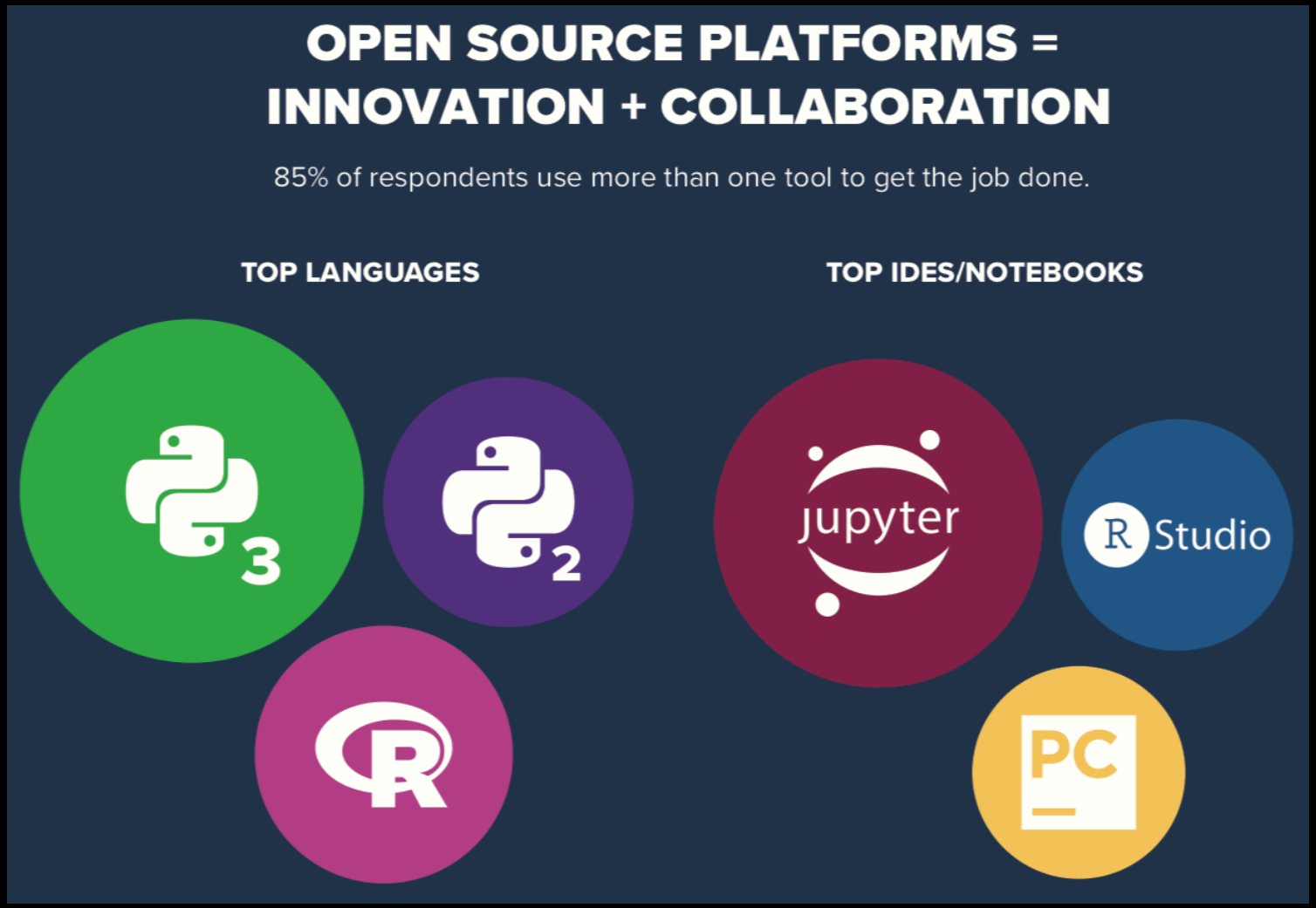
Nearly 50% of respondents are learning data science to apply it to roles in multiple fields.



https://know.anaconda.com/rs/387-XNW-688/images/2019-SoDS-Infographic-Anaconda.pdf



https://know.anaconda.com/rs/387-XNW-688/images/2019-SoDS-Infographic-Anaconda.pdf



https://know.anaconda.com/rs/387-XNW-688/images/2019-SoDS-Infographic-Anaconda.pdf

https://know.anaconda.com/rs/387-XNW-688/images/2019-SoDS-Infographic-Anaconda.pdf

WEBINAR: The State of Data Science in 2019 and How to Overcome the Talent Shortage

Peter Wang | Co-Founder and CTO of Anaconda

Maria Khalusova | Developer Advocate at JetBrains

August 28, 2019 | 2:00pm ET/11:00am PT

Register Now

Meet Our Speakers

PETER WANG | CTO and Co-Founder at Anaconda
Peter has been developing commercial scientific computing and visualization software for over 15 years. He has extensive experience in software design and development across a broad range of areas. Peter's interests in the fundamentals of vector computing and interactive visualization led him to co-found Anaconda. As a creator of the PyData community and conferences, he devotes time and energy to growing the Python data science community and advocating and teaching Python at conferences around the world.

MARIA KHALUSOVA | Developer Advocate at JetBrains
Maria is a Developer Advocate at JetBrains. She primarily focuses on all things Data Science and Machine Learning. Maria is passionate about learning and sharing knowledge which is why she also blogs, presents at conferences and runs PyData Montreal.

https://know.anaconda.com/201908-State-of-Data-Science-Webinar_Registration.html

Syllabus and more

What will this course teach?

- How to use contemporary tools to develop a "data science workflow/pipeline"
  - Python
  - IPython / Jupyter notebooks / Google Colab
  - NumPy
  - Pandas
  - Matplotlib
  - Scikit-Learn
- The math/statistics background behind data analysis
- Basic Machine Learning algorithms for regression and classification
- Critical thinking, perspective, broad view of data science problems

What will this course not cover?

- Python programming
- Advanced Machine Learning algorithms
- Neural Networks
- Deep Learning
- “Big Data” tools, frameworks, etc.
- …

Administrivia

Live section

Video recordings

Canvas

Tentative week-by-week schedule

Assignments: hands-on

Exam: fully online

Any further questions?

Let’s get started!