Lecture 1
Course Introduction

Pierre Biscaye
Université Clermont Auvergne

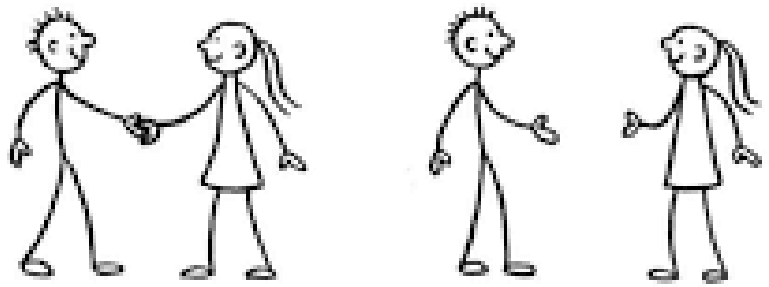
Data Science for Economics

Agenda

1. Introductions
2. Data Science in Economics
3. Syllabus
4. Introduction to Python

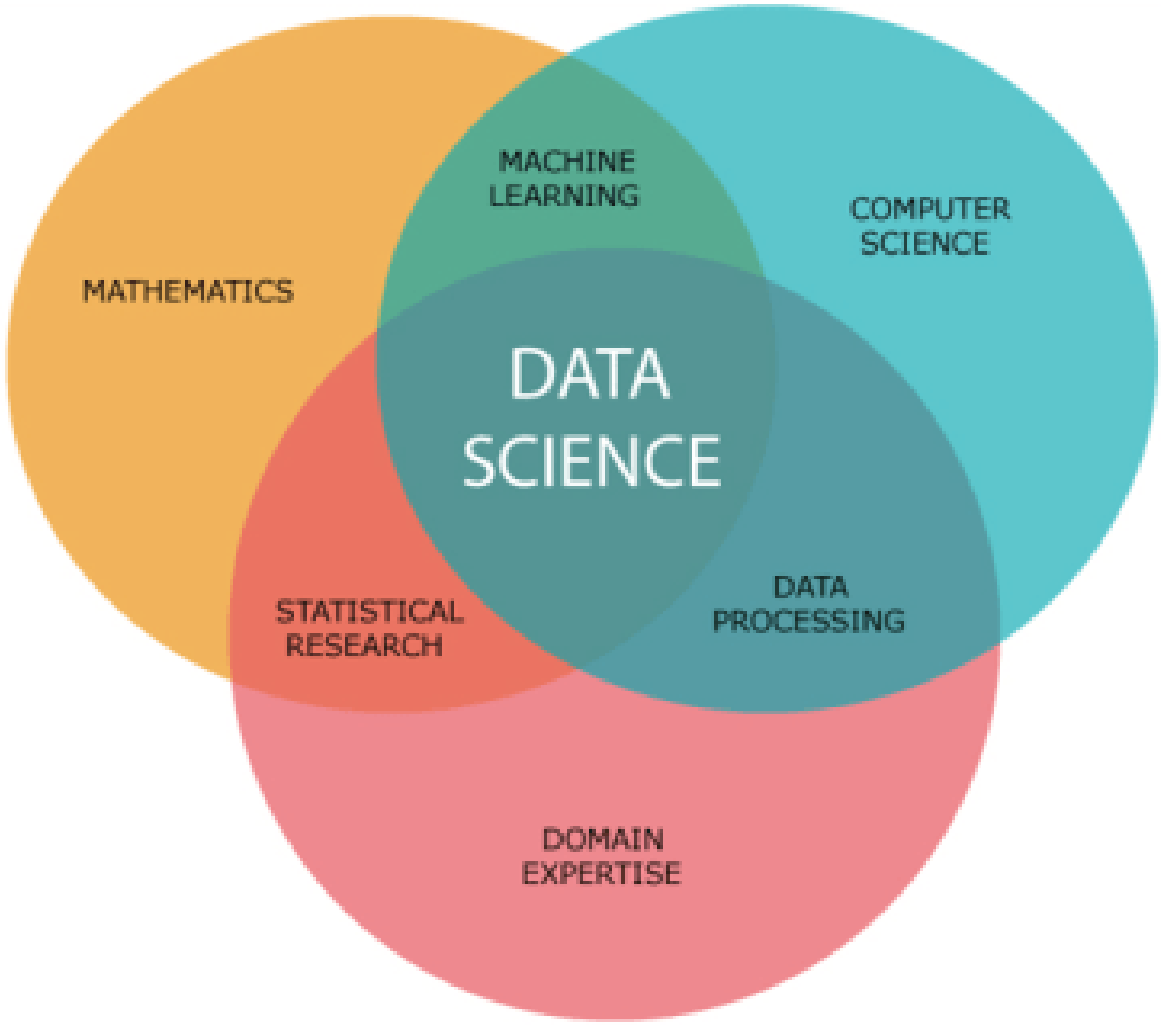
Introductions

- Name
- Where you’re from
- What you are most interested in getting from this class



What is data science?

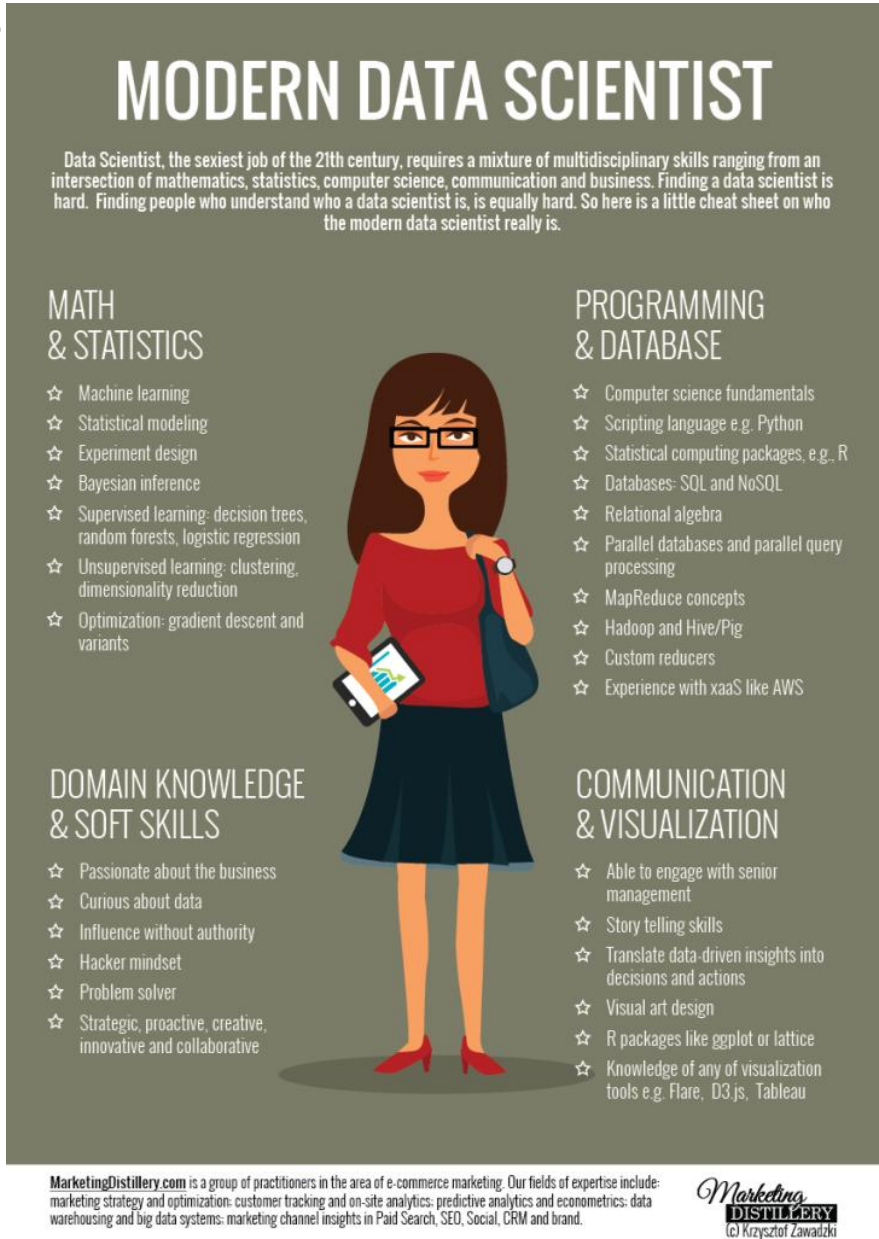
Data science is an interdisciplinary field focused on extracting insights from data using statistical, computational, and analytical techniques.



Source

Core competencies and tools

- Core competencies:
  - Data collection and preprocessing
  - Exploratory data analysis
  - Statistical modeling and machine learning
  - Visualization and reporting
- Key tools and technologies
  - Programming: Python, R
  - Data storage: SQL, NoSQL
  - Visualization: Tableau, Power BI, etc.



Why learn data science?

Source

Data science in different jobs:

- Marketing and Advertising: Customer segmentation, targeted campaigns, and A/B testing
- Healthcare: Predicting disease outbreaks, personalized medicine, and clinical trials
- Finance: Fraud detection, credit risk modeling, and algorithmic trading
- Retail and E-Commerce: Demand forecasting, pricing optimization, and recommendation systems
- Technology: Search engine optimization, natural language processing, and AI development

How is data science important for economists?

- New sources of data: Leverage NLP and web scraping to transform text (central bank speeches) and images (satellite data) into actionable economic indicators.
- Advanced prediction/forecasting: Use machine learning to capture complex, non-linear relationships that traditional linear models often miss.
- Precision causal inference: Use ML to identify relevant controls and reduce research bias.
- Scalable & reproducible workflows: Transition from Stata/Excel to Python, R, and Git to build automated, auditable pipelines.
- "Nowcasting": Utilize high-frequency data (credit card swipes, GPS mobility, satellite data) to construct real-time indicators to inform policy.

Using data science for economic development

- Poverty Reduction:
  - Identifying vulnerable populations using satellite imagery and socioeconomic data
  - Targeting aid distribution efficiently with machine learning models
- Agricultural Productivity:
  - Analyzing weather patterns and soil quality to optimize crop yields
  - Building early warning systems for droughts or pest outbreaks
- Education and Workforce Development:
  - Using data to assess school performance and target investments
  - Identifying skill gaps in the workforce and tailoring training programs
- Infrastructure Planning:
  - Mapping underserved regions to prioritize investments in transportation, energy, and internet access
- Healthcare Access:
  - Optimizing resource allocation for clinics and hospitals in remote areas
  - Predicting disease outbreaks to inform preventive measures

Measuring economic growth from outer space (Henderson et al 2011 AER)

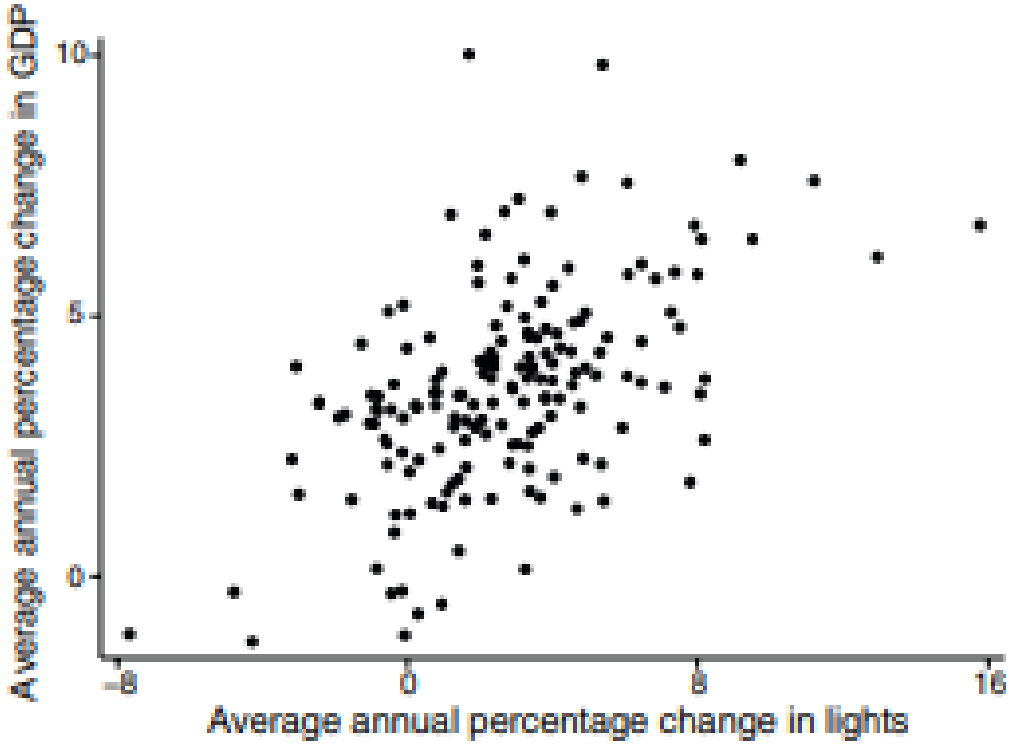


FIGURE 1. GDP VERSUS LIGHTS:
LONG DIFFERENCES 1992–1993 TO 2005–2006

Combining satellite imagery and machine learning to predict poverty (Jean et al 2016 Science)

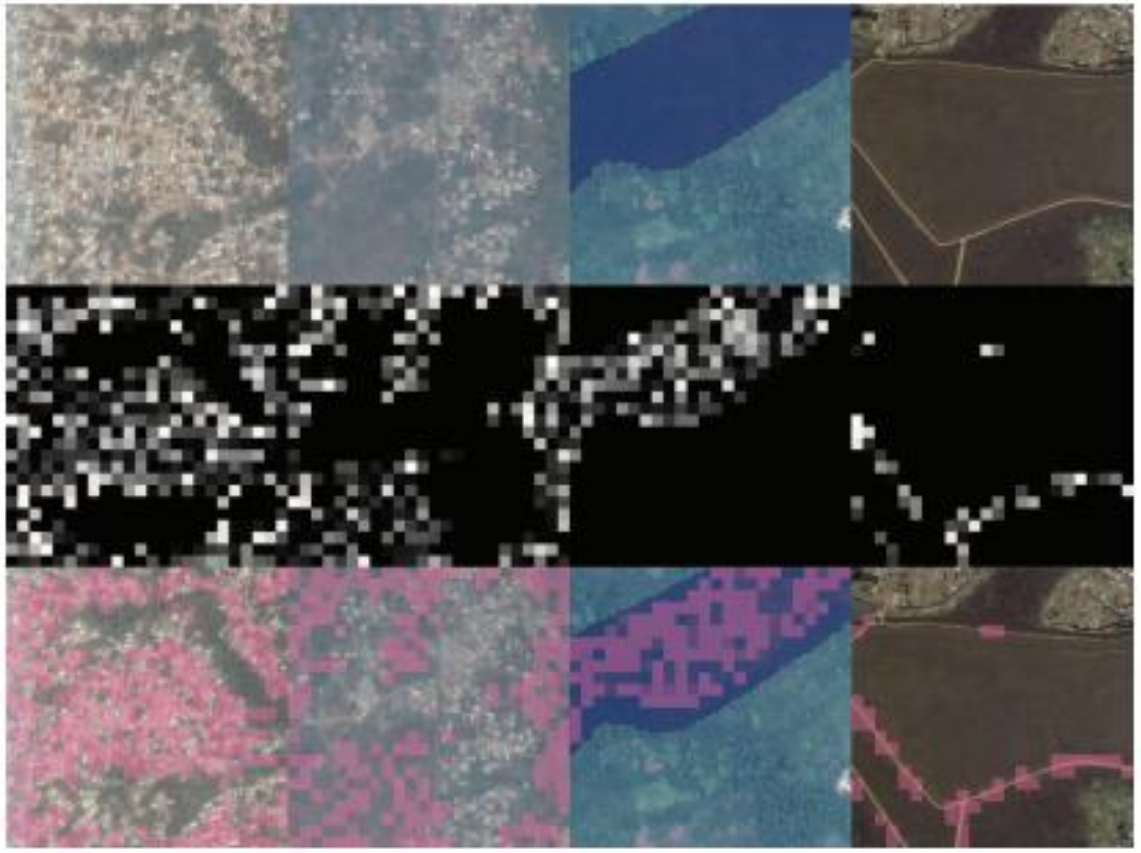
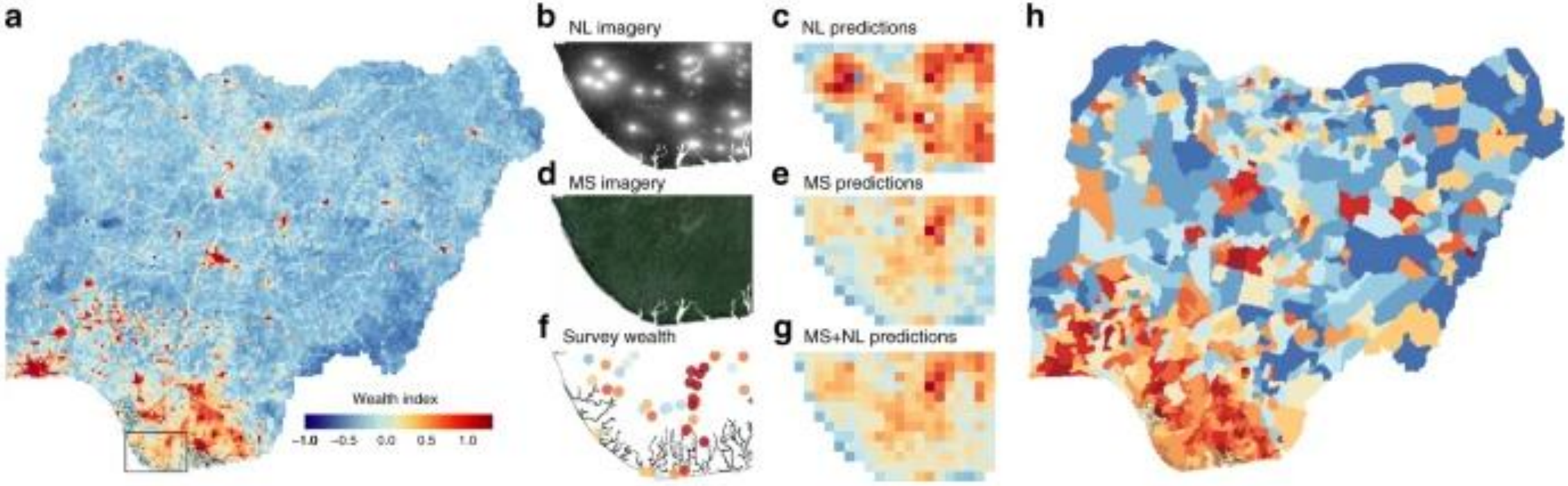


Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

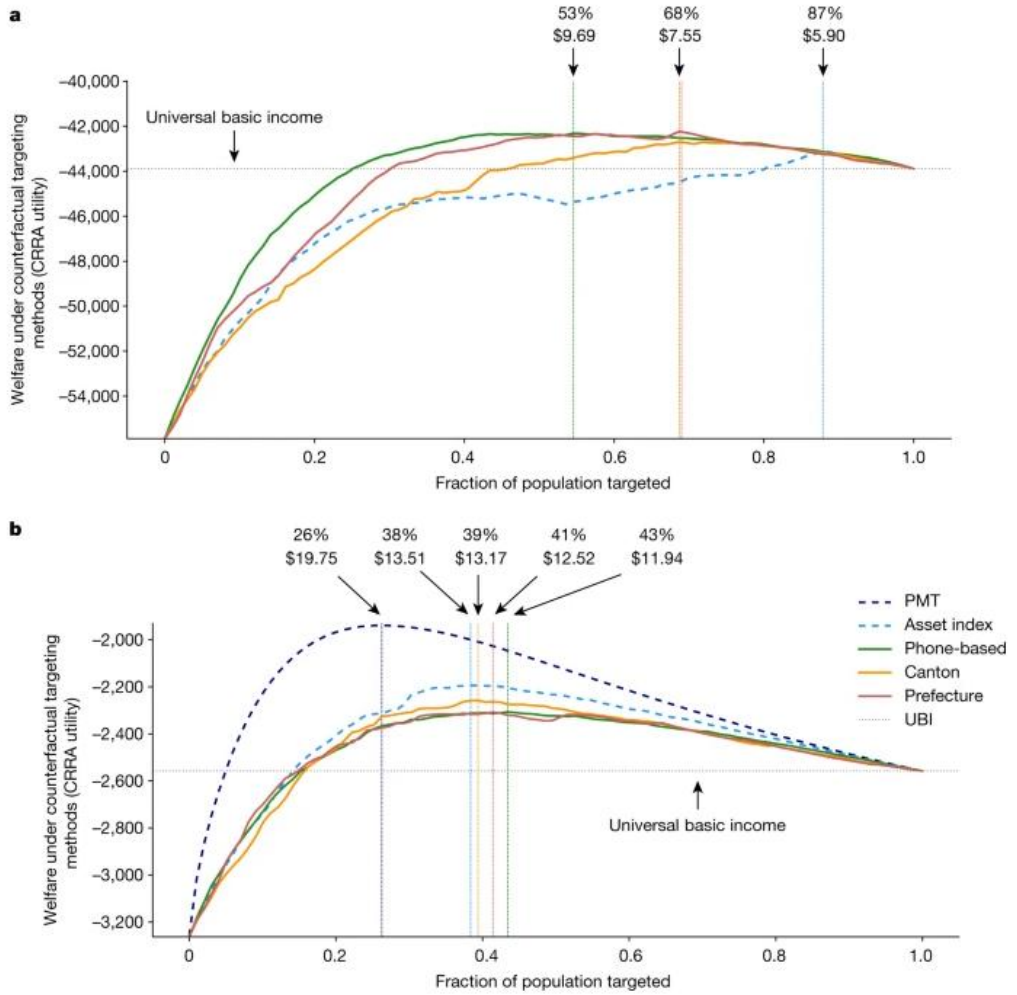
Using publicly available satellite imagery and deep learning to understand economic well-being in Africa (Yeh et al 2020 Nature Comm)

Fig. 6: Spatial extent of imagery allows wealth predictions at scale.



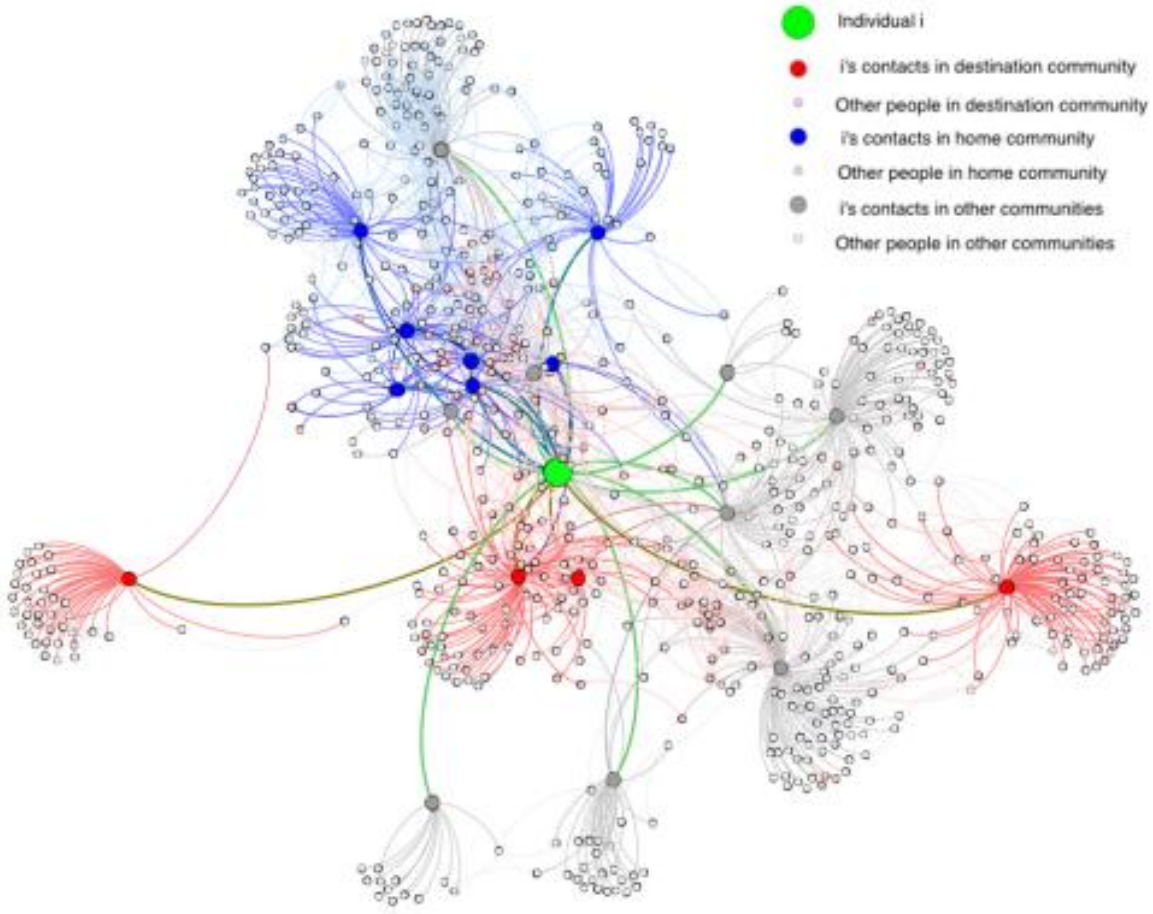
Machine Learning and Phone Data Can Improve the Targeting of Humanitarian Aid (Aiken et al 2022 Nature)

Fig. 2: Welfare analysis of different targeting mechanisms.

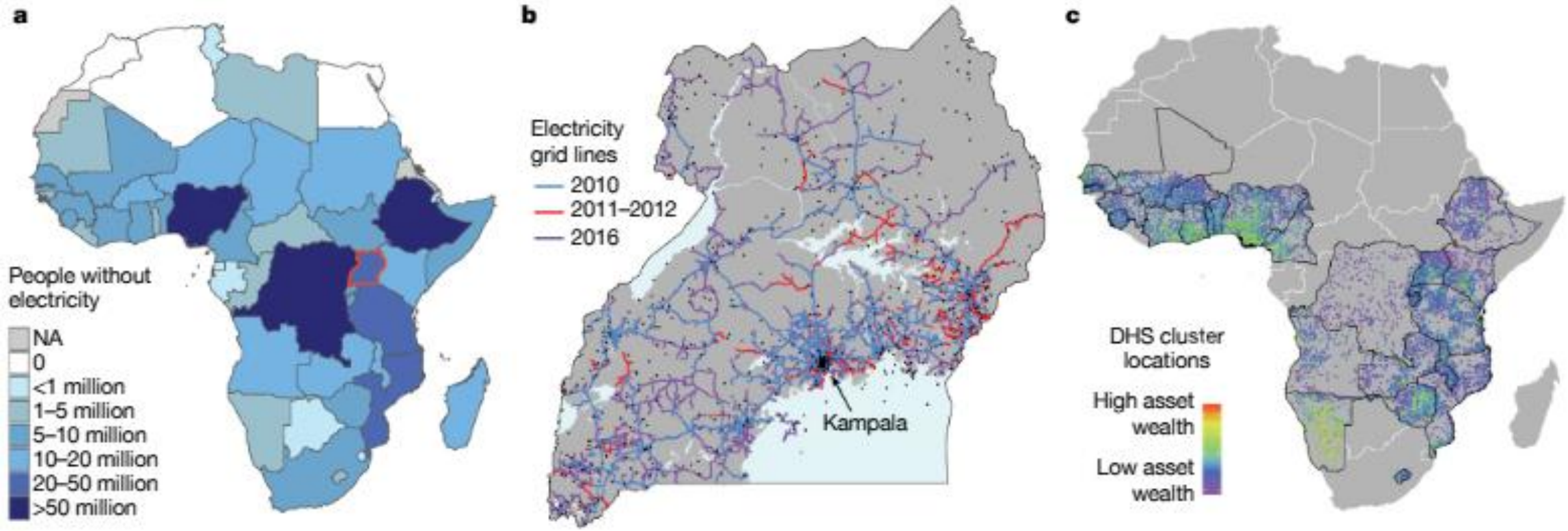


Migration and the Value of Social Networks (Blumenstock et al 2023 Restud)

Figure 2: The social network of a single migrant



Using machine learning to assess the livelihood impact of electricity access (Ratledge et al 2022 Nature)

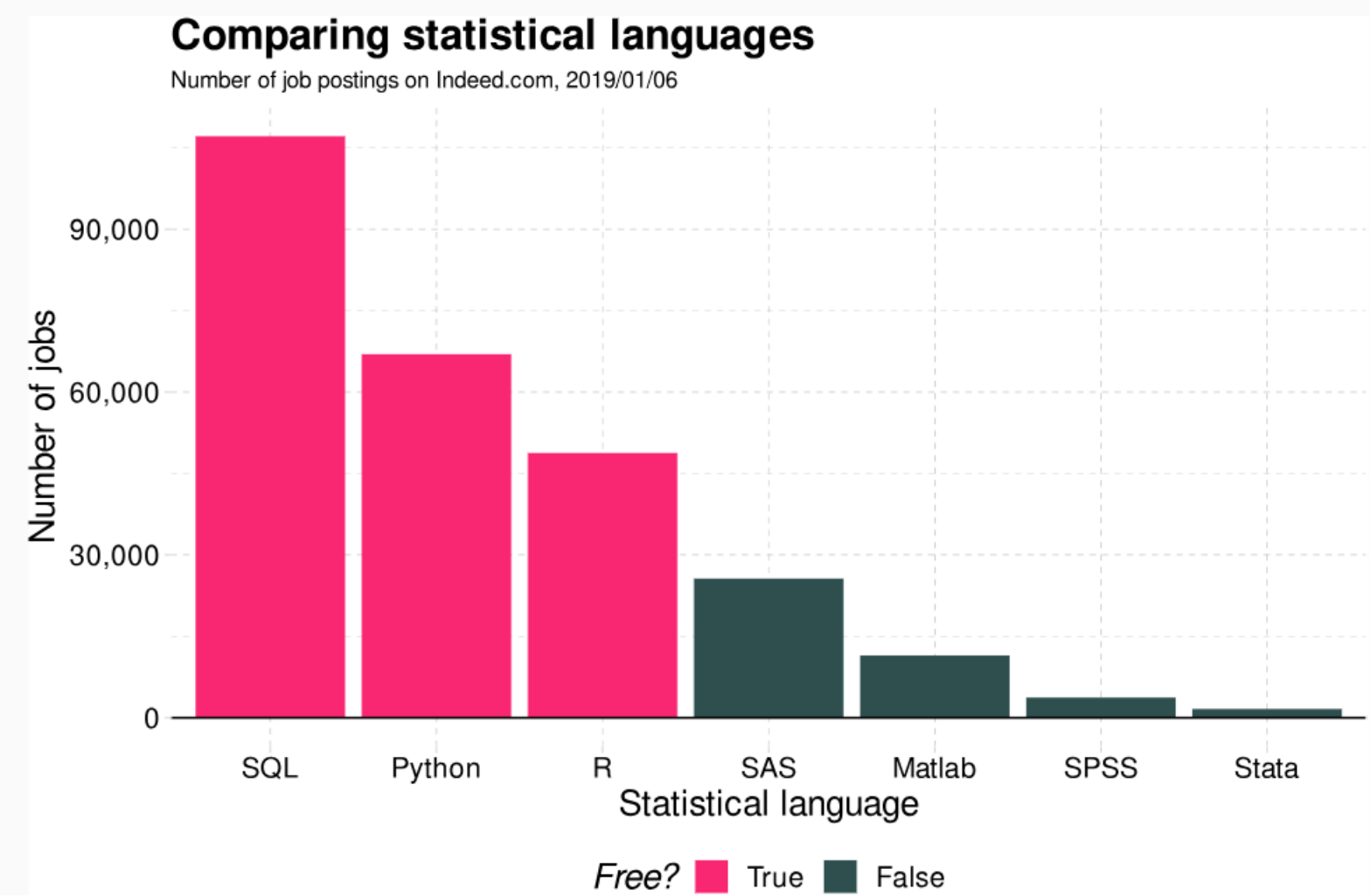


This course

- Introduction to data science tools and technologies
- Focus on using data in applied economic research
- Develop foundational skills for working with different types of data in Python
- Complement your other training in economic theory and econometrics/research methods

We will cover topics I wish had been part of my core graduate curriculum!

Why use Python?

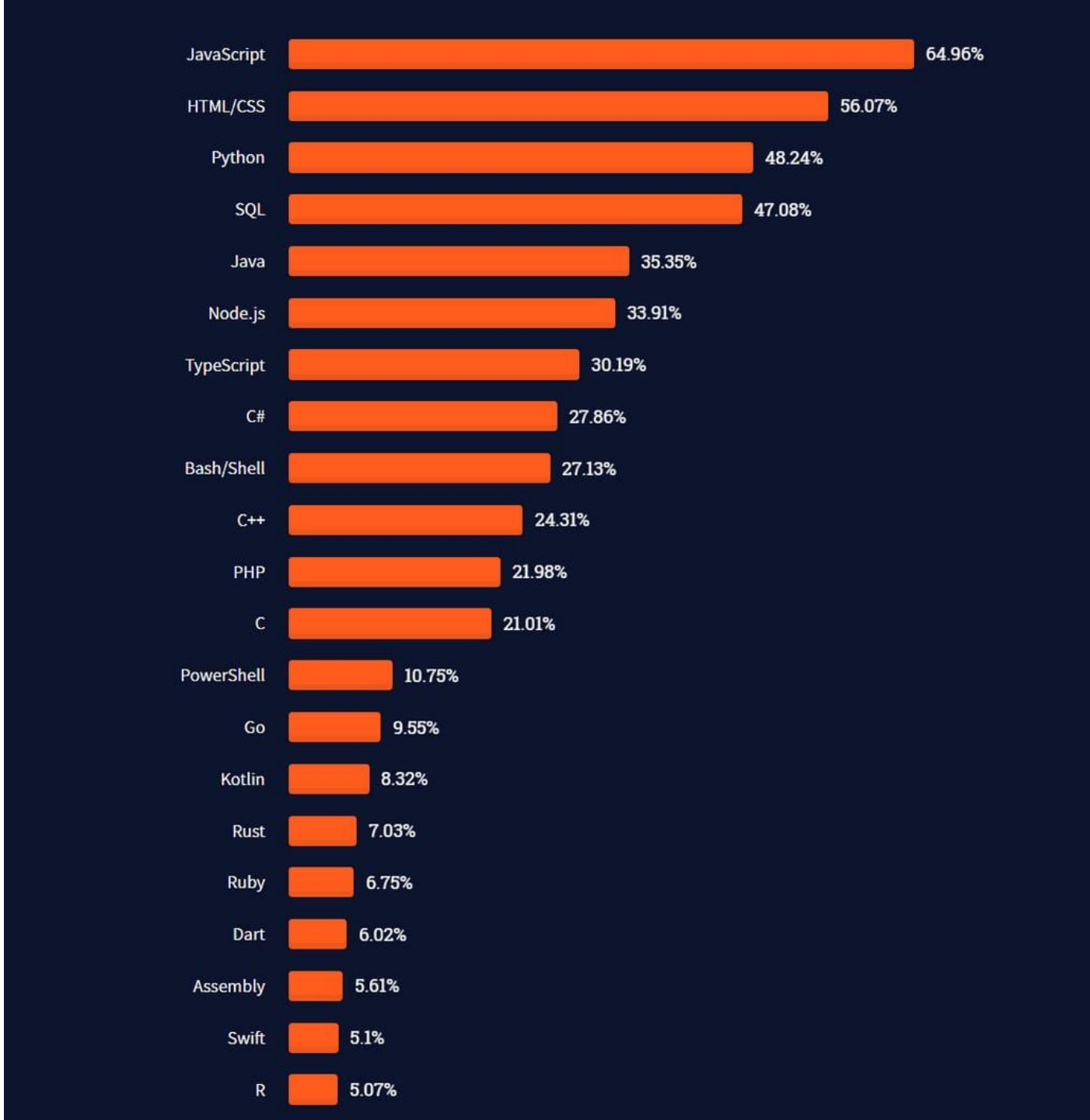


Source

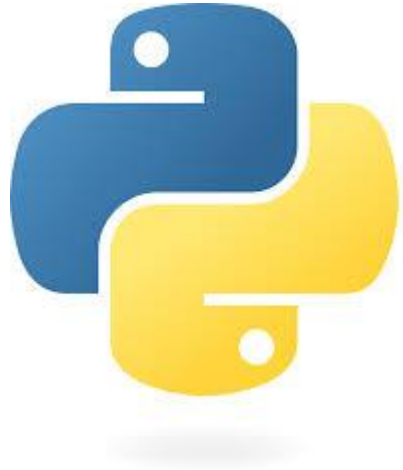
Why use Python?

- 2021 Stack Overflow survey of the most popular programming languages in world placed Python third, and it has been rising over time
- R: created by statisticians for statisticians
  - Outstanding visualization tools
- Python: multi-paradigm, also used for software development
  - Outstanding ML and AI tools

Source



What is Python?



Python is a versatile, high-level programming language known for its simplicity and readability.

- Key Features:
  - Easy-to-learn syntax suitable for beginners and experts
  - Extensive library support for various tasks (e.g., data analysis, machine learning, web development)
  - Cross-platform compatibility (works on Windows, macOS, Linux)
- Applications:
  - Web development (Django, Flask)
  - Data science and machine learning (Pandas, NumPy, Scikit-learn)
  - Automation and scripting
  - Game development and more

Why Python for this course?

- Beginner-Friendly:
  - Intuitive syntax and large online community support
  - Ideal for learners with little programming experience
- Comprehensive Libraries:
  - Pandas for data manipulation and analysis
  - NumPy for numerical computations
  - Matplotlib and Seaborn for data visualization
  - Scikit-learn for machine learning
- Versatility:
  - Supports everything from data cleaning to building machine learning models
  - Integrates well with other tools like SQL and APIs
- Widely Used:
  - One of the most popular programming languages in the data science industry
  - Used by individuals for learning and by companies for large-scale applications

Getting set up with Python

- Anaconda
- Using the command interface
- Managing the package environment
- Jupyter Notebook

Python and data science resources

- Data science resources index
- UC Berkeley D-Lab
- Stack exchange
- Google
- LLMs

Syllabus

- Course objectives
- Course logistics
- Course outline

Class structure

- Lectures:
  - Big picture
  - Applications in development economics
  - Slides posted before class
- Python workshops:
  - Guided orientation using Jupyter Notebooks
  - Posted before class
- On your own:
  - Practice notebooks with challenge tasks applying tools and methods

Poll

To help me understand background and experience of students in the class, please fill in this Google Form

BREAK

- Then, Python basics with Jupyter Notebook