AI Saturdays Lagos

Cohort 7

Introduction to Data Science

Slide Credits

1. CMU Practical Data science course by Prof. J. Zico Kolter (http://www.datasciencecourse.org)
2.

Today

- What is Data Science?
- What is Data Science not?
- (A few) Data Science Examples
- Discussions

What is Data Science?

Some possible definitions

Data science is the application of computational and statistical techniques to address or gain insight into some problem in the real world

Some possible definitions

Data science is the application of **computational** and **statistical** techniques to address or gain insight into some problem in the **real world**

Some possible definitions

Data science = statistics +
data processing +
machine learning +
scientific inquiry +
visualization +
business analytics +
big data + …

myjobmag.com/blog/highest-salaries-in-nigeria

In spite of Nigeria's disconcerting poverty rate, a significant amount of Nigerians earn decent remuneration from their chosen fields. Some of these professions earn in excess of millions and as a result, they are pretty competitive to access.

These high-paying jobs are influenced by education (course of study, certification etc.), skills and experience. The jobs cut across multinationals, public and private companies as well as businesses.

In addition to offering exciting perks and benefits, they also pay their workers attractive wages.

In this post, you would learn about the most financially rewarding jobs in Nigeria, how much they pay and how you can secure them.

OPTIMIZE YOUR CV TO LAND YOUR DREAM JOB

Top 25 Highest Paying Jobs in Nigeria Ranked:

1. Surgeon - **N420,000**
2. Aeronautical Engineer - **N410,000**
3. Project Manager - **N321,000**
4. Petroleum Engineer - **N294,000**
5. Sailor - **N263,000**
6. Pilot - **N210,000**
7. Investment Bankers- **N194,000**
8. Dentist- **N193,000**
9. Software Developer- **N153,000**
10. Accountant - **N126,000**

Google

25 best job in america

All Images News Videos Maps More Tools

About 2.320.000.000 results (0,81 seconds)

The Best Jobs of 2021 include:

- Physician assistant.
- Software developer.
- Nurse practitioner.
- Medical and health services manager.
- Physician.
- Statistician.
- Speech-language pathologist.
- Data scientist.

More items... • 12 Jan 2021

https://money.usnews.com › Money › Careers

The Best Jobs in America in 2021 - US News Money

What is Data Science NOT?

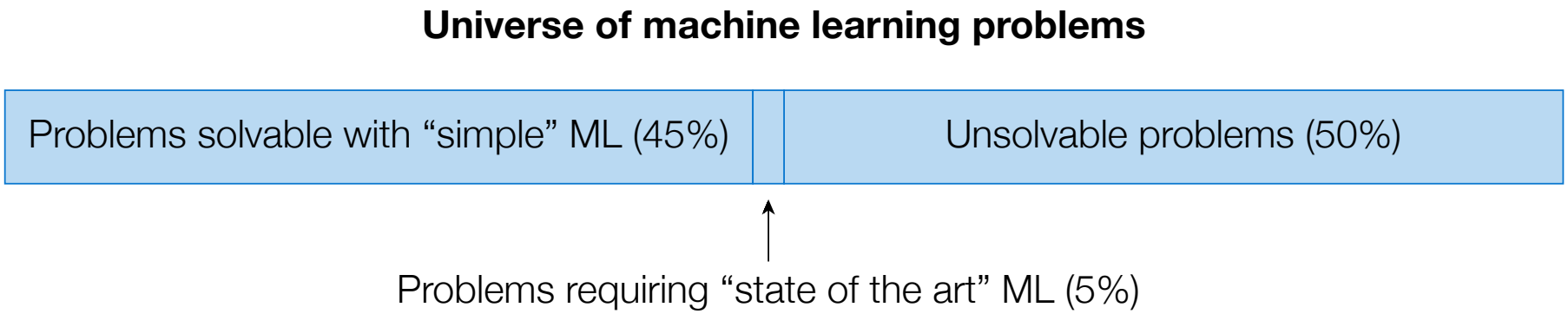
Data science is not machine learning

Machine learning involves computation and statistics, but has not (traditionally) been very concerned about answering *scientific questions*

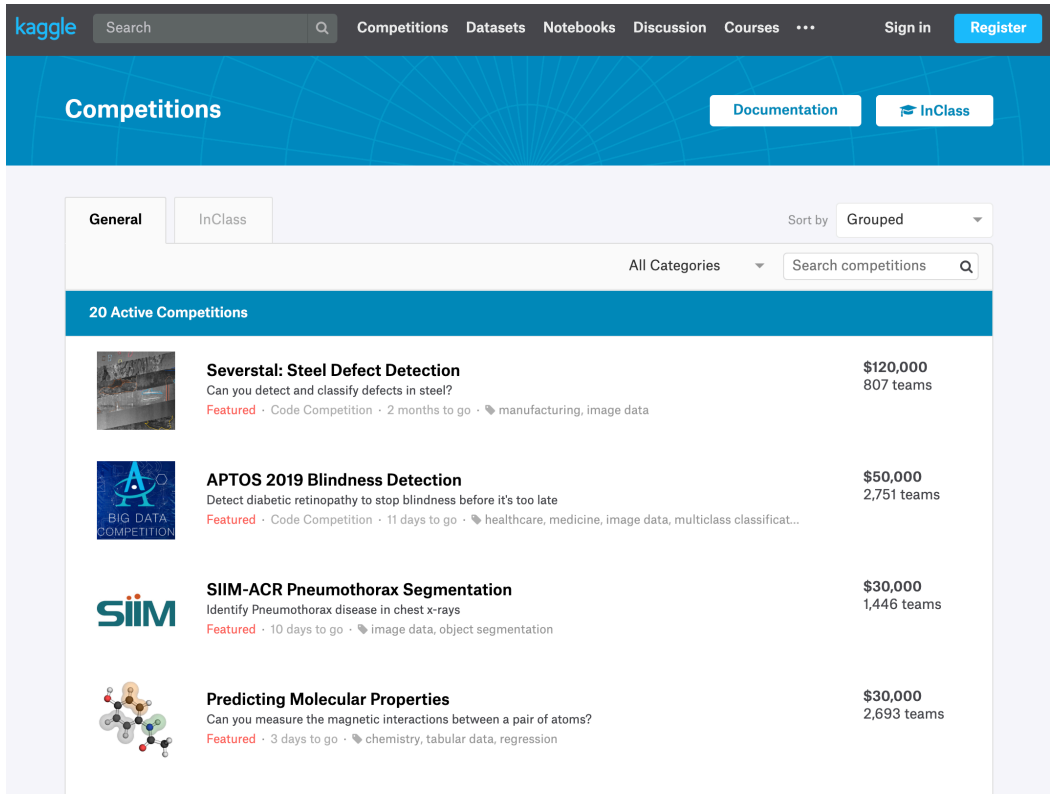
Machine learning has a heavy focus on fancy algorithms…

... but sometimes the best way to solve a problem is just by visualizing the data, for instance

Data science is not machine learning



Data science is not machine learning competitions



Data science competitions like Kaggle ask you to optimize a metric on a fixed data set

This may or may not ultimately solve the desired business/scientific problem

Data science is the iterative cycle of designing a concrete problem, building an algorithm to solve it (or determining that this is not possible), and evaluating what insights this provides for the real underlying question

Data science is not statistics

“Analyzing data computationally, to understand some phenomenon in the real world, you say? … that sounds an awful lot like statistics”

Statistics (at least the academic type) has evolved a lot more along the mathematical/theoretical frontier

Not many statistics courses have a lecture on e.g. web scraping, or a lot of data processing more generally

Plus, statisticians use R, while data scientists use Python … clearly these are completely different fields

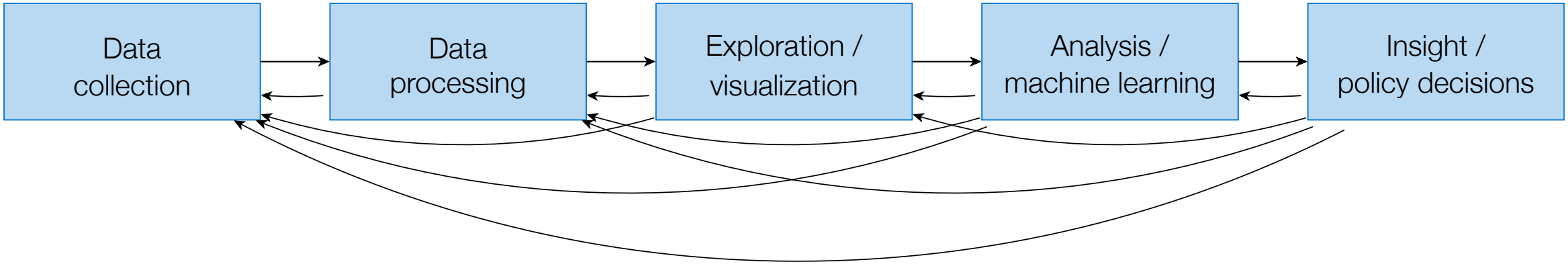
Data science is not big data

Sometimes, in order to truly understand and answer your question, you need massive amounts of data…

…But sometimes you don't

Don't create more work for yourself than you need to

Back to what data science is



(A few) Data Science Examples

Gendered language in professor reviews

Gendered Language in Teacher Reviews

This interactive chart lets you explore the words used to describe male and female teachers in about 14 million reviews from RateMyProfessor.com.

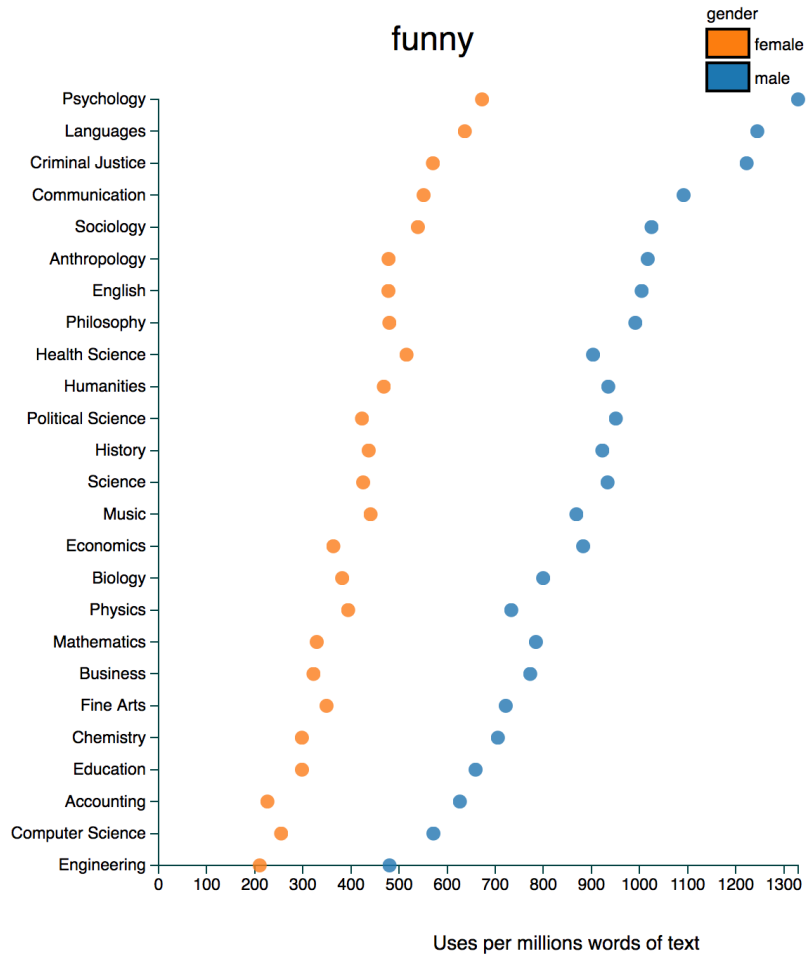
You can enter any other word (or two-word phrase) into the box below to see how it is split across gender and discipline: the x-axis gives how many times your term is used per million words of text (normalized against gender and field). You can also limit to just negative or positive reviews (based on the numeric ratings on the site). For some more background, see here.

Not all words have gender splits, but a surprising number do. Even things like pronouns are used quite differently by gender.

Search term(s) (case-insensitive):
use commas to aggregate multiple terms

funny

All ratings | Only positive | Only negative

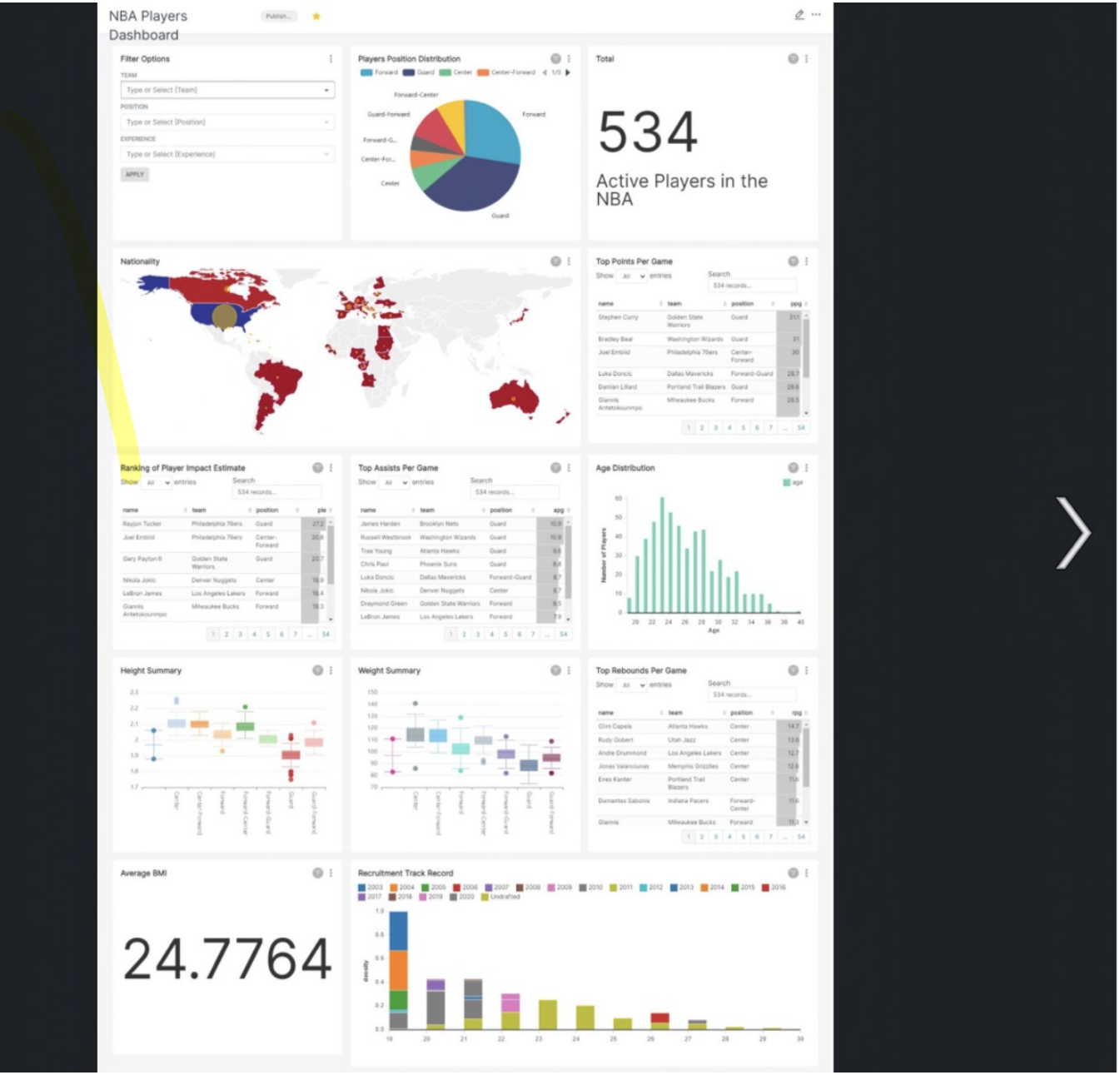


http://benschmidt.org/profGender/

Obligatory quote

The greatest value of a picture is when it forces us to notice what we never expected to see.

-John Tukey



Chizurum Olorondu • 1st
Full Stack Data Scientist
3mo • Edited •

I recently completed a full-stack, end to end data science project called NBA Players. The aim of this was to simplify and limit the time it takes for an NBA fan or a curious individual to find out useful facts about the current active players in the NBA.

The data pipeline for this process was built and scheduled (weekly) using Apache Airflow (https://lnkd.in/gpBZj77). The first component of the workflow was data collection. I used the NBA website as my data source and proceeded to gather data on all active players. This data is inserted into a sqlite3 database file and stored in a data lake; which in this case is a simple file folder (Second Component).

The third component extracts the most recent entry to the data lake and performs a custom ETL (Extraction, Transformation and Loading) process on the data. The processed data is then stored or used to update my data warehouse; a remote PostgreSQL Database instance on elephantsql.com. Airflow automatically triggers the pipeline to run every week and updates the data warehouse.

I also developed an outlier detection machine learning model to identify outlier players present in the database (Jupyter Notebook link: https://lnkd.in/gNgQ-2z)

https://nba-superset.herokuapp.com/superset/dashboard/4/

Does it solve the problem?

FiveThirtyEight

ELECTION 2018
FiveThirtyEight

House forecast | Senate | Governor | Midterms coverage | More politics | abcNEWS

Search for a race or candidate

Search

How do you like your House forecast?

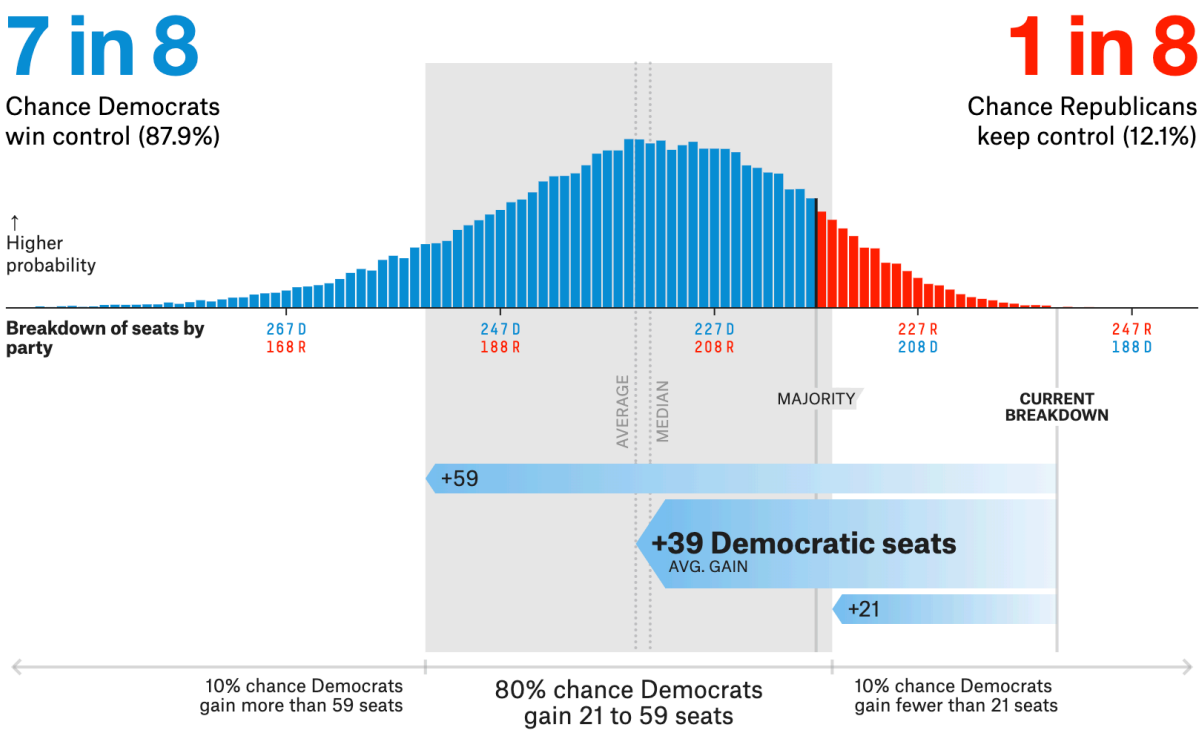
Lite
Keep it simple, please — give me the best forecast you can based on what local and national polls say

Classic
I'll take the polls, plus all the "fundamentals": fundraising, past voting in the district, historical trends and more

Deluxe
Gimme the works — the Classic forecasts plus experts' ratings

Forecasting the race for the House

Updated Nov. 6, 2018, at 11:06 AM



https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house/

What communities in Kenya and Uganda would benefit most from charitable donations?

Poverty Mapping

Figure 2: Example of metal roof in center of satellite image.

Figure 3: Example of thatched roof in center of satellite image.

Abelson, Varshney, and Sun. “Targeting Direct Cash Transfers to the Extremely Poor,” 2012

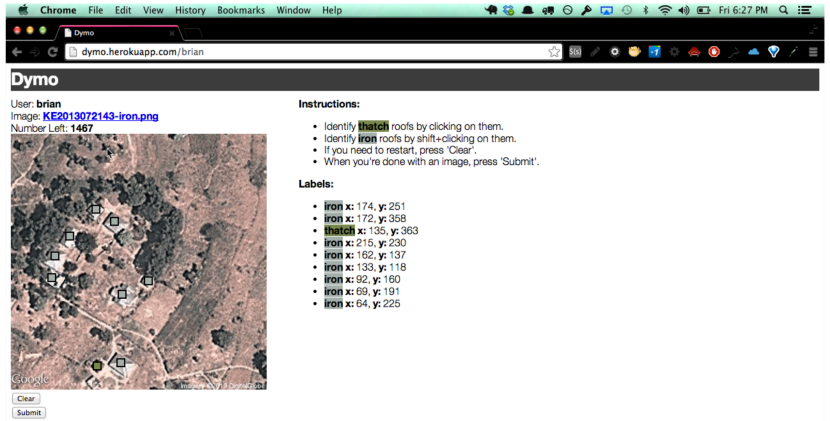


Figure 6: Screen shot of application deployed for crowdsourced labeling of roofs in satellite images.

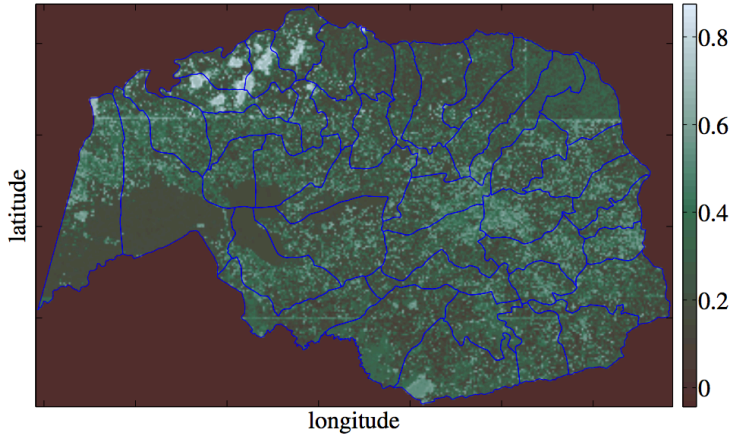


Figure 11: Heat map of proportion of roofs that are metal in the region of interest.

Administrative

Learning objectives of this course

After taking this course, you should…

… understand the full data science pipeline, and be familiar with programming tools to accomplish the different portions

... be able to collect data from unstructured sources and store it using appropriate structure such as relational databases, graphs, matrices, etc

... know to explore and visualize your data

... be able to analyze your data rigorously using a variety of statistical and machine learning approaches

Topics covered

Week	Date	Topic	Resources	Instructors
1	4-Sep	Introduction to Data Science	Slide, Note	Tejumade Afonja
2	11-Sep	Data Collection and Scraping	Slide, Note	Akintayo Jabar
3	18-Sep	Relational Data	Slide, Note	Akintayo Jabar
4	25-Sep	Visualization and Data Exploration	Slide, Note	Ahmed Olanrewaju
5	2-Oct	Matrices, Vectors, and Linear Algebra	Slide, Note	Lawrence Francis
6	9-Oct	Data Preprocessing	Slide, Videos	Esemeje Omole
7	16-Oct	Introduction to Machine Learning	Slide, Note	Olumide Okubadejo
8	23-Oct	Linear Models	Slide, Note, Playlist	Stanley Dukor
9	30-Oct	**Break**		
10	6-Nov	Model Evaluation	Slide, Playlist	Orevaoghene Ahia
11	13-Nov	Nonlinear Models	Slide, Note	Kenechi Dukor
12	20-Nov	Probabilistic Models	Slide, Note	Tejumade Afonja

https://github.com/AISaturdaysLagos/cohort7_structure

Practicals covered

Practicals

The practicals will touch on different MLOPS and will be held alongside classes each week.

Week	Date	Topic
1	4-Sep	Data Science Notebook Frameworks
2	11-Sep	A web scraping task with basic intro first; Data Labelling Tools and Frameworks
3	18-Sep	Intro to Pandas; Optimized computational Framework part 1 (Pandas related)
4	25-Sep	Industrial Strength Visualization libraries
5	2-Oct	Intro to numpy; Optimized computational Framework part 2 (Numpy related)
6	9-Oct	Outlier and Anomaly Detection
7	16-Oct	Intro to Sklearn
8	23-Oct	No Practicals
9	30-Oct	**Break**
10	6-Nov	Model and Data Versioning

Main Resources

Practical Data Science | INFORMATION | LECTURES | ASSIGNMENTS | FORUM | STAFF | POLICIES | FAQ

This page lists the class lectures and recitations, plus additional material (slides, notes, video) associated with each lecture. Recordings of all the classes will available on the course Canvas page.

Lectures

Date	Lecture	Slides	Notes
	Data collection and management		
1-Feb	Introduction		
3-Feb	Data collection and scraping		
8-Feb	Jupyter Notebook lab		
10-Feb	Relational data		
15-Feb	Visualization and data exploration		
17-Feb	Vectors, matrices, and linear algebra		
22-Feb	Graph and network processing		
24-Feb	Free text and natural language processing		
1-Mar	(continued)		
	Statistical modeling and machine learning		

http://datasciencecourse.org

Main Resources

Machine Learning @ VU

Machine Learning at VU University Amsterdam

This page contains all public information about the course *Machine Learning* at the VU University Amsterdam. We provide the following materials:

- **Lecture slides and videos.**
- **Worksheets** These are very brief Jupyter notebooks to help you get the software installed and to show the basics. They introduce the libraries Numpy, Matplotlib, Pandas, Sklearn and Keras.
- **Homework** The homework consists of small pen-and-paper exercises to help you test that you've really understood the more technical points of the lectures. Answers are provided. If you are a registered student, please refer to the Canvas page instead. All material authored by Peter Bloem unless noted otherwise.

Reuse is allowed under a creative commons license, details below.

			homework	worksheets	previous
	1. Introduction 1.1 What is machine learning? 1.2 Classification 1.3 Other abstract tasks 1.4 Social impact 1	 playlist slides			 2020 2019 2018

https://mlvu.github.io

Questions?