PSC NCSA NSF Award 2417789

Carnegie Mellon University

University of Pittsburgh

CI Pathways

Leading the Way to Effective Cyberinfrastructure Use

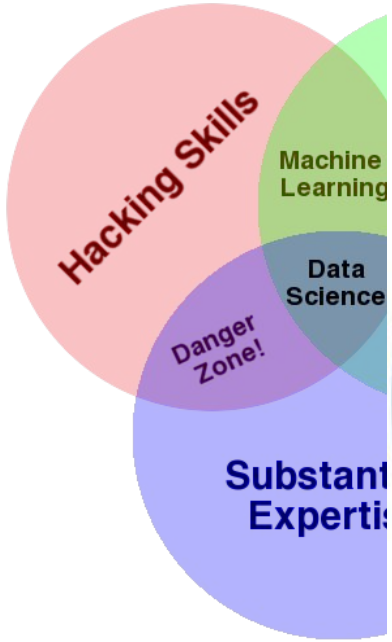
CI Awareness: Introduction to Data Science & Machine Learning

Soham Pal - Research Computing & Data Facilitator, NCSA

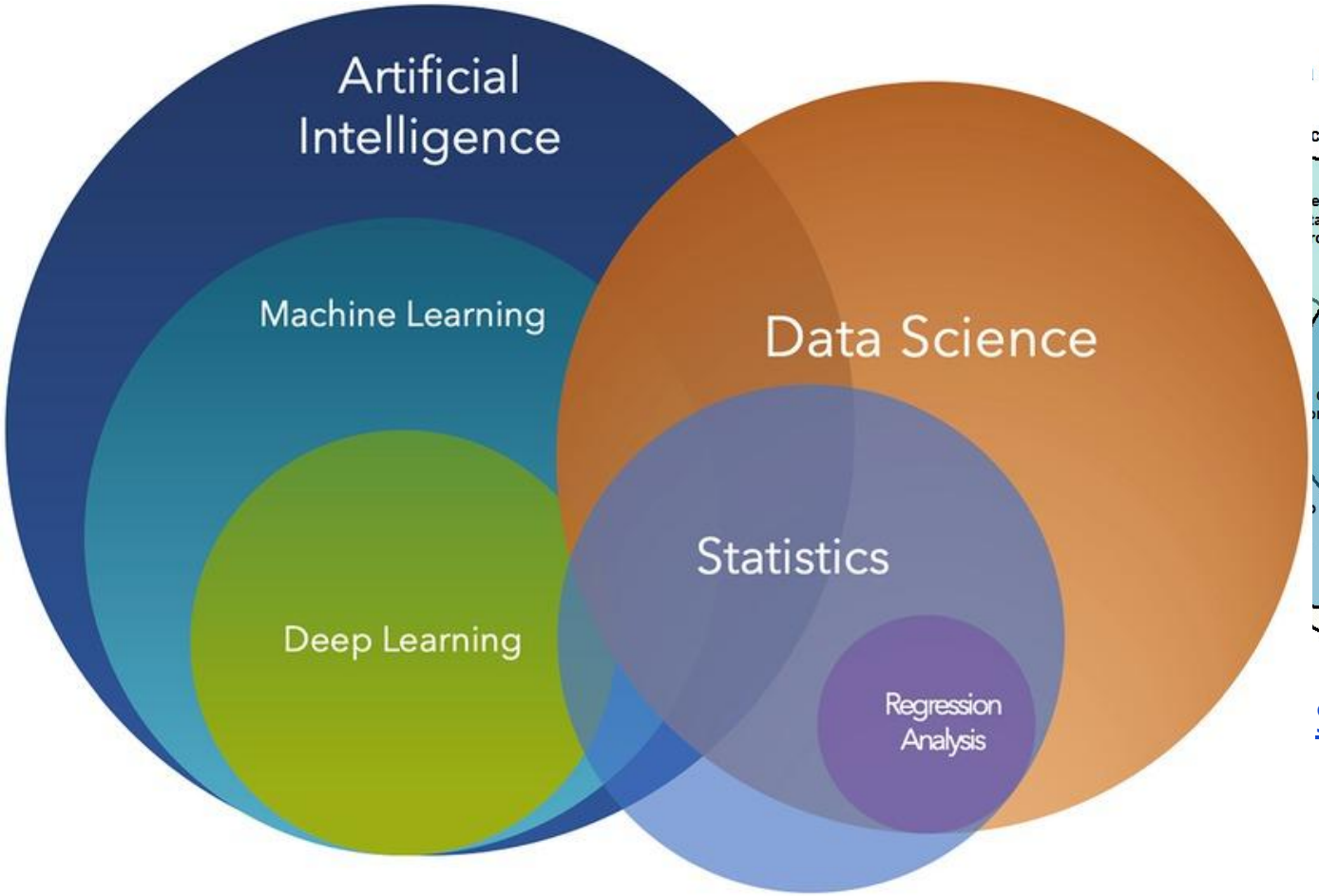
Overview

1. What is data science? What is machine learning?
2. How are they used in research?
3. What tools are typically used?
4. Examples

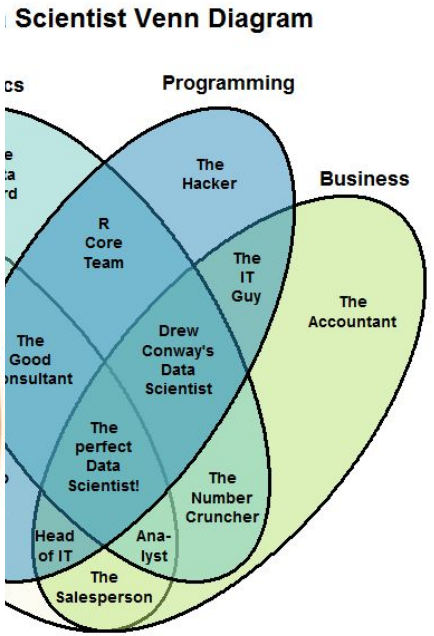
What is Data Science? What is Machine Learning?



Souce: [Drew Co



Souce: [Morris, et. al.](), 2024



[Stephan Kolassa](), 2021

What really are these?

It’s hard to pin down a precise meaning of these terms – the meanings are continuously evolving with the technologies and methodologies used in these fields.

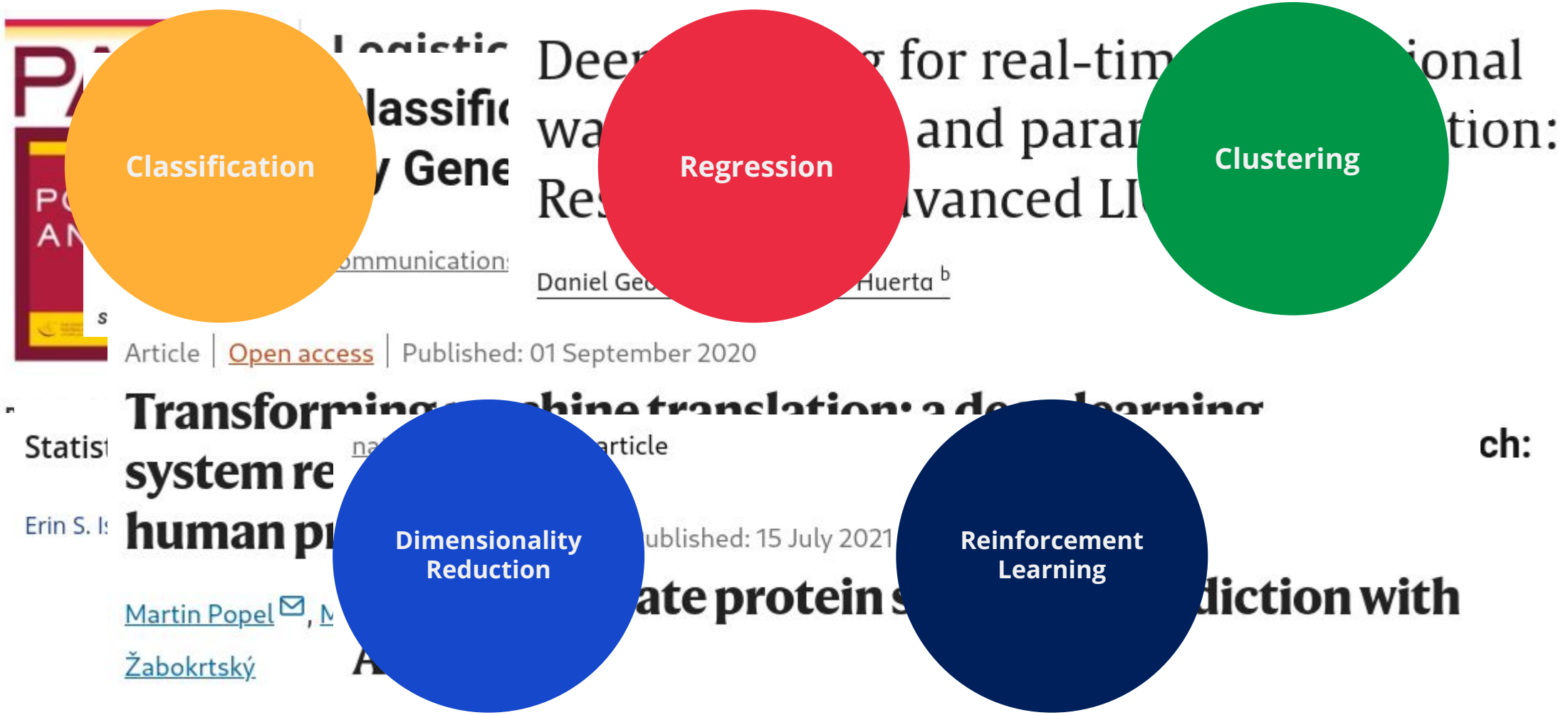
Typically data science can be thought of as “**obtaining, curating, analysing data to make decisions or predictions**”, and machine learning can be thought of as “**designing and using *self-updating* algorithms that can learn from data instead of manually coding all possible intermediate steps**”.

These fields naturally intersect.

Why are these useful for research?

1. Scientific models are abstractions of the world. Their validity are determined by how well they explain and predict real world data. **Data is inherently important to the scientific process.**
2. We now have lots of data, particularly digitized data.
3. It is impossible to manually code every possible scenario when dealing with big data. An algorithm that can learn from data can be used to extract the useful information from data.
4. Can generate *new* data.

What can you do with data science / machine learning?



Typical data science / machine learning workflow

1. Define your research problem.
2. **Obtain data.** If you are gathering new data, make sure that you follow the data safety and privacy guidelines as advised by your university/department.
3. Process and visualize the data. Raw data often cannot be used as is. Visualizing the data will help you to get a sense of what you are working with and identify patterns in the data.
4. Train machine learning model(s) on the processed data or make statistical analysis of data.
5. Evaluate model performance.
6. Publish?

Tools used for data science/machine learning

In scientific research the following languages are the most popular for data science/machine learning:

1. Python (popular libraries: Pandas, Scikit-Learn, PyTorch, etc.)
2. R (popular libraries: Tidyverse ecosystem)

Python provides broad support for most data science and machine learning tasks. It is by far the most popular language for scientific research involving data science and machine learning.

R is very popular in social sciences and bioinformatics. It has some of the best data visualization capabilities.

Other languages like SQL, Matlab, etc have more niche applications. **Use the language that will help you solve your research problem the fastest.**

Resources for learning data science/machine learning

1. An Introduction to Statistical Learning by James, *et. al.*
2. R for Data Science by Wickham and Grolemund
3. Tidy Modeling with R by Kuhn and Silge
4. Understanding Deep Learning by Prince
5. Physics based Deep Learning by Thuerey, *et. al.*
6. CI Pathways - Machine Learning and Data Science tracks

Demonstrations