CUNY School of Professional Studies

# Meetup 2

## Data 607

Peter Kowalchuk

CUNY School of Professional Studies

# Week 1 in review

## Observations

General

- Assignments must include Github code and RPubs published markdown
- Consider using the class GitHub to share your work: https://github.com/pkowalchuk/CUNY-Data_607-Spring_2025
- Data in context idea: Quatro

Week 1

- Loading data from the source is important
  - Data privacy and ownership
- Start using libraries: tidyverse
  - Dplyr already in use

Week 2

- No need to create a MySQL resource in Azure, one was provided
- Errors standing up an Azure resource: subscription regional restrictions
- Workbench is a plus, but you can also do all the work in R (or Python, other programing languages)
- Schema, database, names interchanged in R and Workbench, it can be confusing

CUNY School of Professional Studies

# This week’s demos

## Your opportunity to showcase your work

- Aaliyah
- Musrat
- Gillian
- Olivia
- Alina
- Cindy

# Data Science in Context

## Your opportunity to talk about anything data related

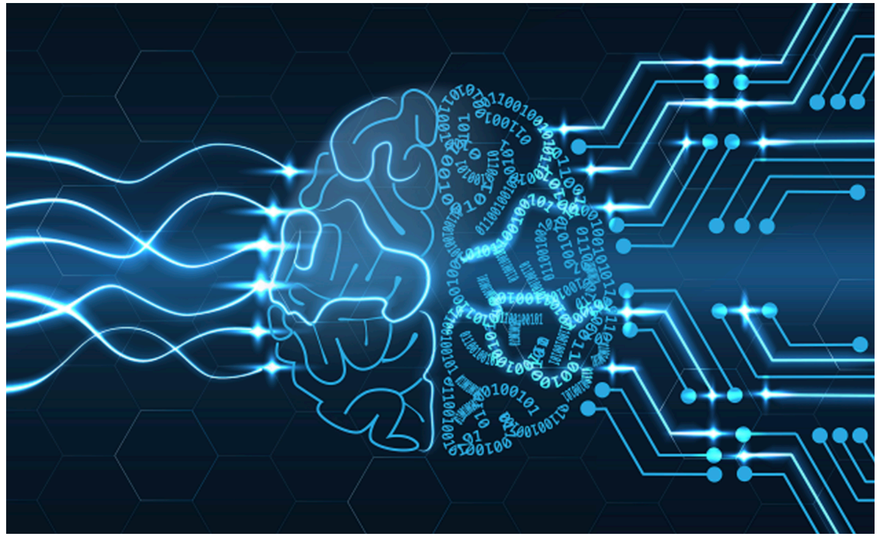
- Present any time during the semester
- 5 to 7min max
- Open content
    - There is a lot happening in Data Science, do some research, share with the class

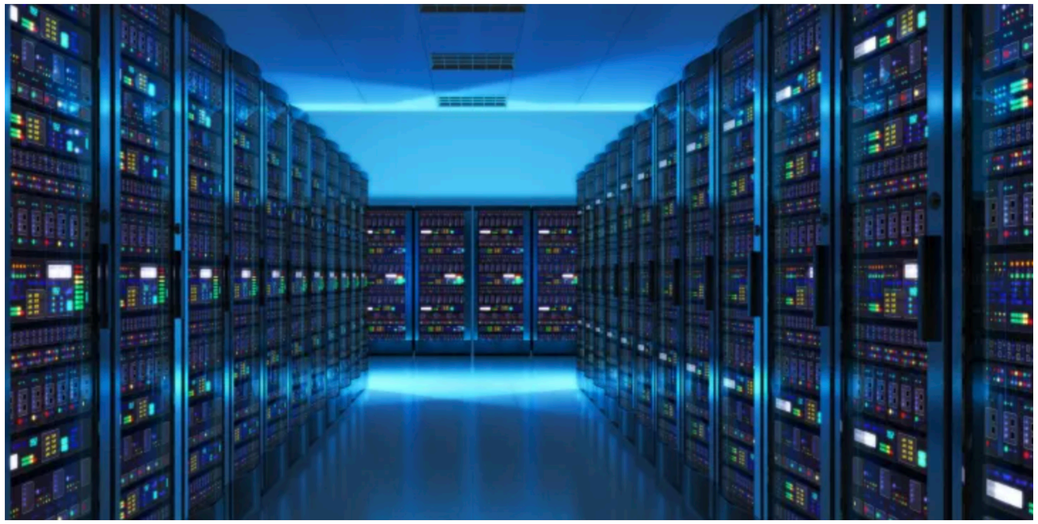
CUNY School of Professional Studies

# Data ownership, using it responsively

## A data in context example

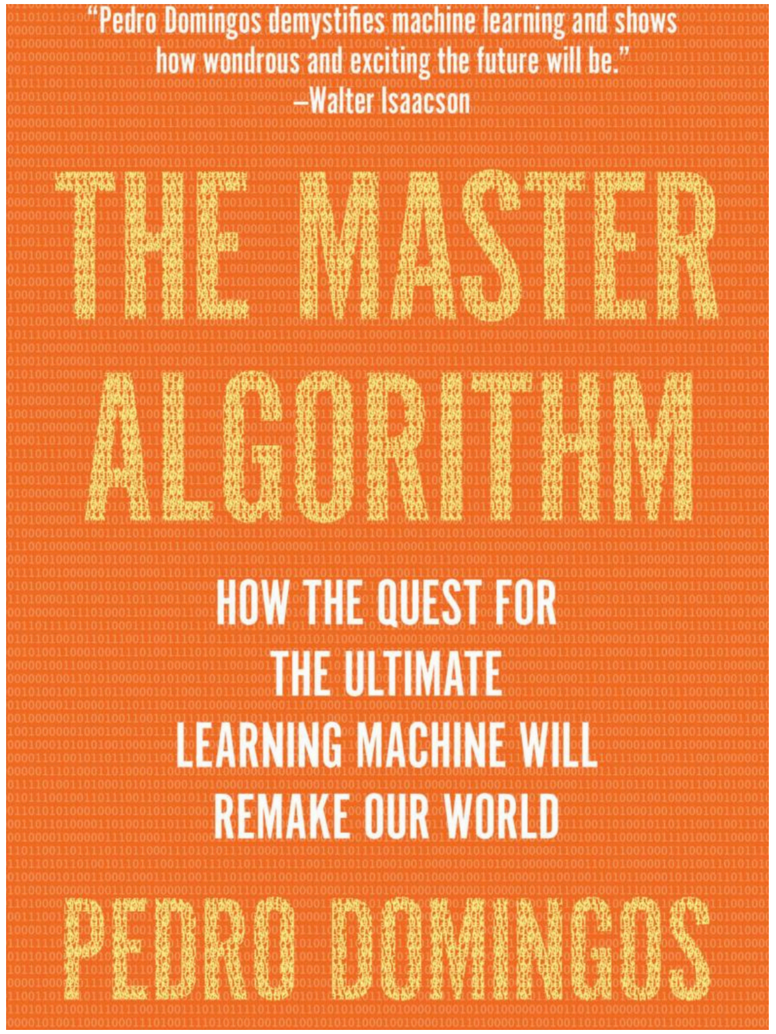
Models or data





# Models have been king

## Looking for the best model



# Large Language Models

## Did we find the Master Algorithm?

OpenAI

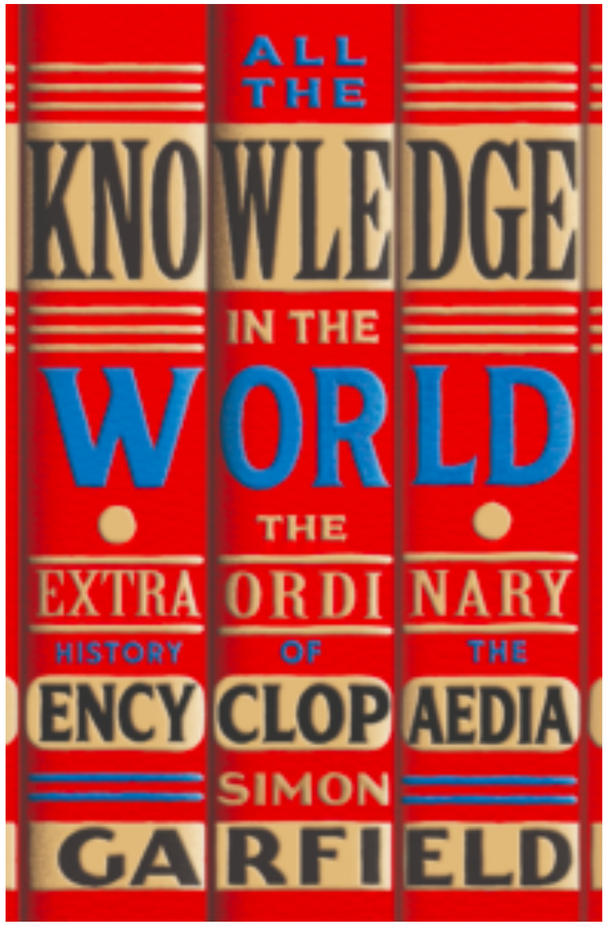
Google BERT

LLaMA by Meta

Hugging Face

# How are Large Language Models trained?

## They use large amounts of data



# Concerns over data used to train models

## Emphasis on the data



TECHNOLOGY | ITALY

### Italy lifts ban on ChatGPT after data privacy improvements

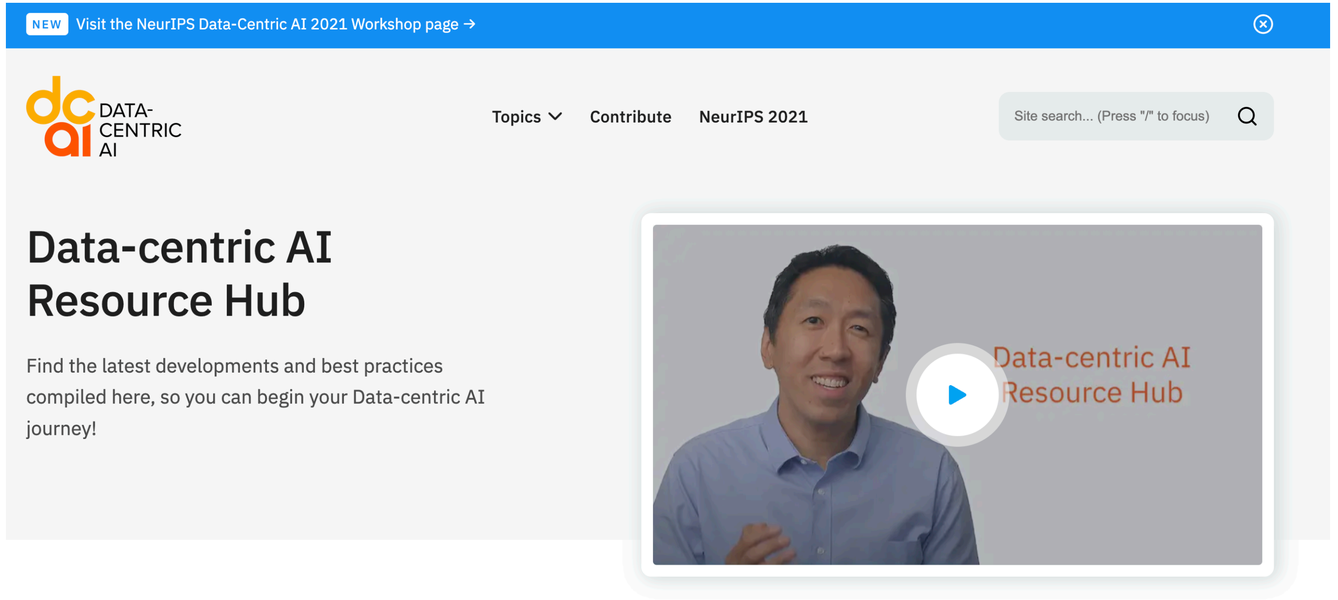
04/29/2023

**The hotly debated AI chatbot is back online in Italy after installing new warnings for users and the option to opt-out of having chats be used to train ChatGPT's algorithms.**


© MARCO BERTORELLO/AFP/Getty Images

# Data is becoming more important

## Even before LLMs



# Who owns the data?

## Data privacy regulation

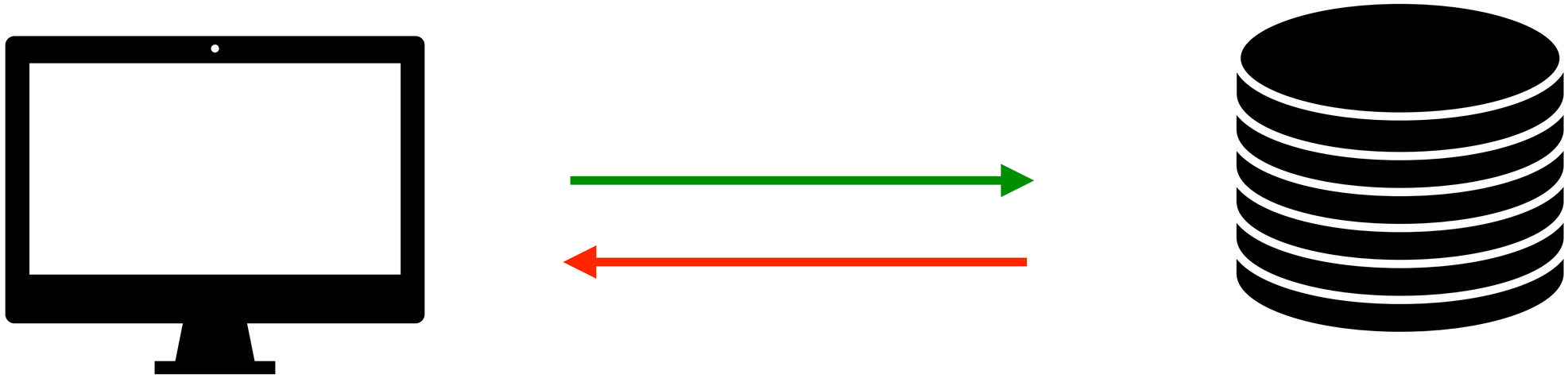

CUNY School of Professional Studies

# What does it mean for us?

## What to consider in data science

# What does it mean for our course?

## Always use data at its origin



# How to connect to external SQL data sources?

## SQL, hosted SQL, other sources

- SQL database
- Hosted SQL
  - VM
  - Serverless
  - Database as a service
- Non-relational databases
- Cloud storage

# How to connect to hosted storage

## The principles

- Secure Access
  - Firewalls
- Authentication
- API calls

CUNY School of Professional Studies

# Good data use starts with you

# Week 3 Assignment - Data Normalization

## Data engineering, but data analysts must understand

- **Organizing** the attributes of a relational database to
  - **Eliminate Data Redundancy:** Eliminate duplicate data to save storage and improve efficiency
  - **Achieve Data Integrity:** Ensure data consistency and accuracy across the database
  - **Simplify Queries:** Make it easier to manage and query the database

# Week 3 Assignment - Data Normalization

## Data engineering, but data analysts must understand

- **Normal Forms (Boyce-Codd):**
  - **First Normal Form (1NF):** Ensure each column contains atomic (indivisible) values and each record is unique.
  - **Second Normal Form (2NF):** Achieve 1NF and remove partial dependencies (no non-key attribute depends on a part of a composite key).
  - **Third Normal Form (3NF):** Achieve 2NF and remove transitive dependencies (non-key attributes do not depend on other non-key attributes).

# Week 3 Assignment - REGEX

## Still relevant

- REGEX is still in use
  - Operations detection in energy field operations: looking for literal descriptions
- LLMs provide a new option for some applications
  - Hazards detection in energy field operations: looking for semantic descriptions

# REGEX vs LLM pattern search

## Practical example

Find the answers to questions in a body of text

- REGEX

```
pattern = re.compile(re.escape(start.replace('\n', '')) + '(.*?[\s\S]*?)' + re.escape(end.replace('\n', '')), re.DOTALL)
answer = pattern.search(pdf_text.replace('\n', '')).group(1)
```

- LLM

```
actor_prompt="Actor: You are a college professor reading a text and identifying each answer to a question."
task_prompt="Task: return the text between the two defined questions in the given text:"
rubric_prompt="question 1: "+start+"\nquestion 2: "+end
key_prompt="Answer:"
```

# Enjoy REGEX

## Use the simplest match

**See you next week!**