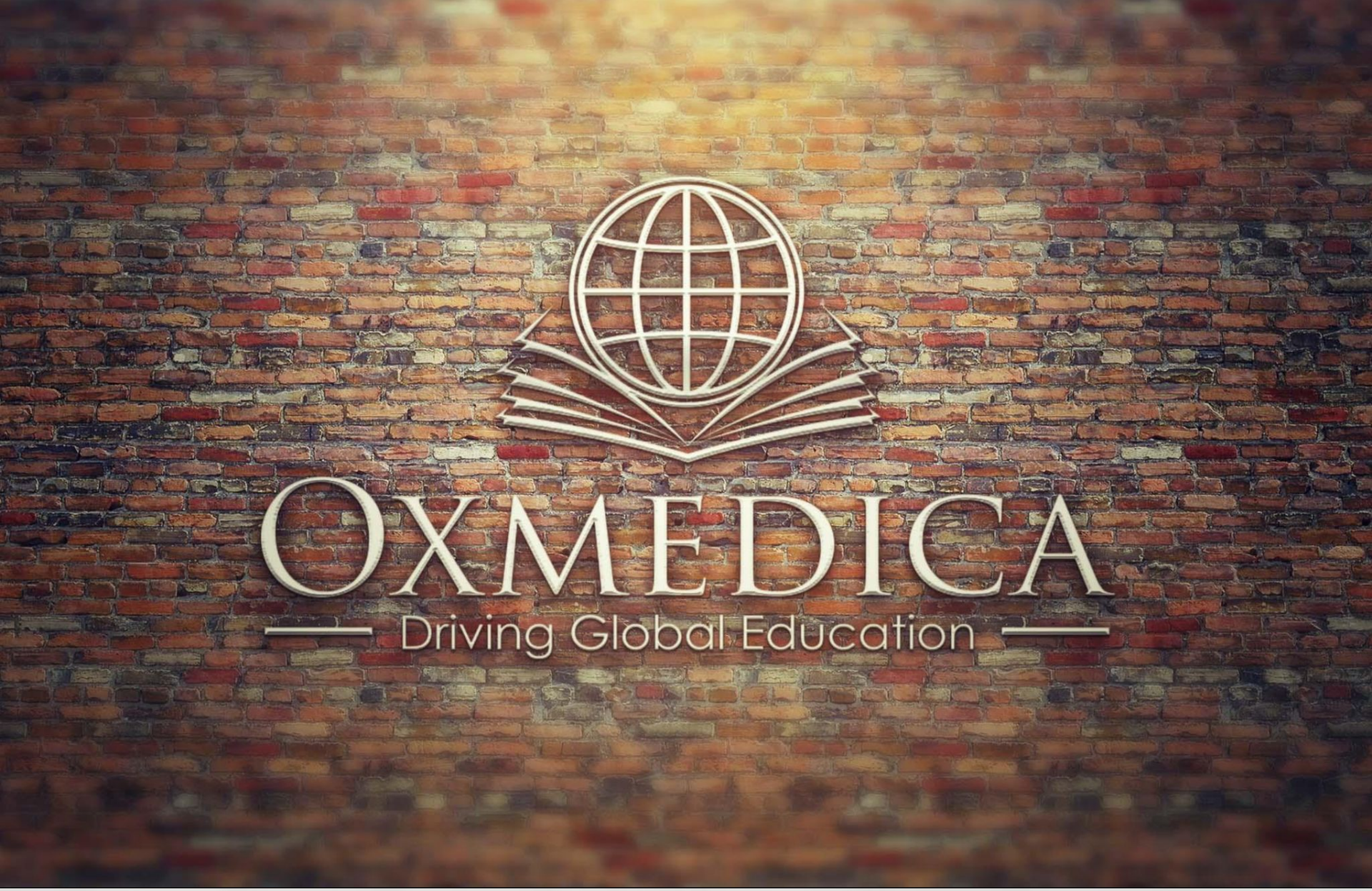


Introduction to Big Data

Overview

1. What do we mean when we talk about data?
2. What is big data and how is it different?
3. What are the unique challenges of working with big data?

What do we mean when we talk about data?

Tabular Data

X1	X2	X3	X4
0.1	1	45	0
-0.7	1	0.2	0
-0.2	0	3	2
0.9	0	24	2
-0.2	0	3	1
-0.4	1	1	3
0.3	0	1	0

Text Data

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache.

→ Harry Potter

Image Data

Source

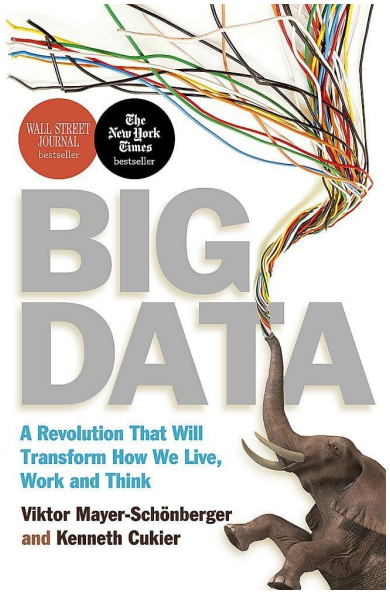
What do we mean when we talk about data?

- These are all different data ‘modalities’
- What other modalities might we find?
- Modern AI models are **‘multimodal’** in the sense that they reason over data from different modalities.

Big data: the three Vs

Source

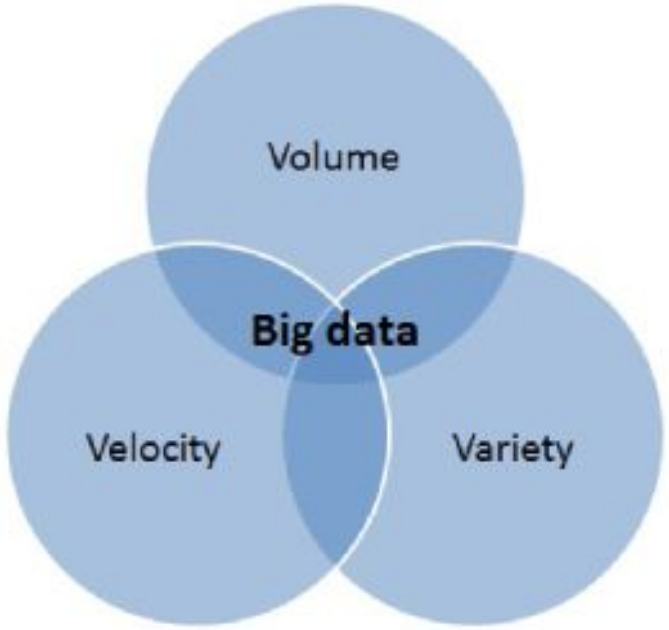
- The term was coined by Professor Viktor Mayer-Schönberger, who is at my department!
- Proposed “the 3 Vs to define big data” in his book “Big Data”



Source

Big data: the three Vs

1. Volume
2. Velocity
3. Variety



Source

1. Volume

- Just lots and lots of it… large storage cost!

2. Velocity

- Velocity = The speed at which data is generated
- Data is constantly being generated and must be used on-the-fly. This is in contrast to classic datasets which might have been generated a long time ago.
- Think about social media data. Constantly collecting likes, views, scrolling patterns…etc. Millions of data points!

3. Variety

- **Everything** can be collected and used these days and it takes multiple forms.
- Consider which types of data are collected by instagram?
- Not obvious how to use all of this data?

Big data in the past vs now (the age of large language models)

- Not really that big by modern standards!
- Maybe a couple of GBs (fraction of the storage of a modern iPhone)
- Modern large language models like ChatGPT are built on lots of data
- LOTS of data
- More like TB of data (many, many, many times your computer's storage)
- Think the **entire internet.** Everything that exists…

The unique challenges of working with big data

1. Storage infrastructure and costs
2. A lot more difficult to build models that can accommodate all of this data
3. Not immediately obvious how to use some types of data
4. Compute!

Compute!

- You'll hear people talking about 'GPUs' → Computers specialised for AI
- These are seriously expensive and powerful things

HOME › NEWS › IT HARDWARE & SEMICONDUCTORS

Report: Saudi Arabia acquires 3,000 Nvidia GPUs, UAE buys thousands

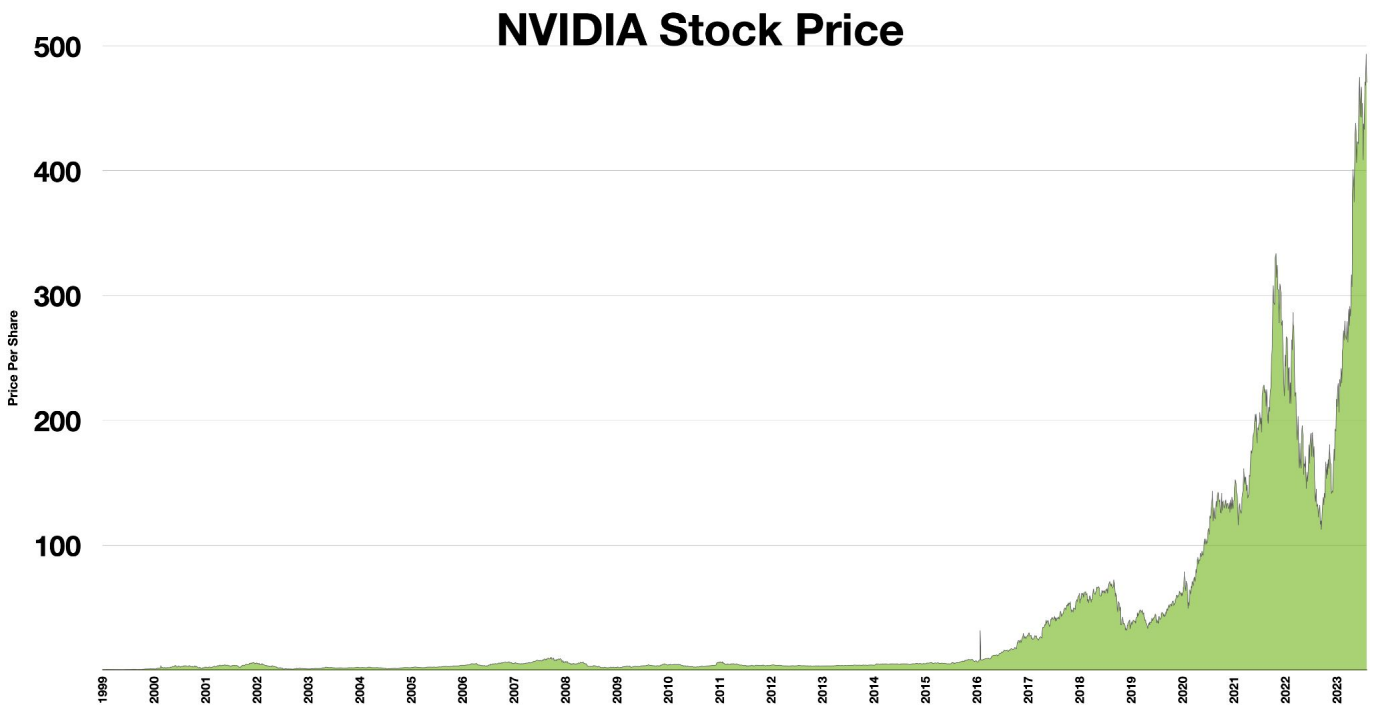
As everyone jumps into the generative AI race

August 15, 2023 By: Sebastian Moss Have your say

Source

Building GPUs is **big business**

- **NVIDIA now the most valuable company in the world**



Source

“Big data is the new oil”

Source

Open questions…?

1. Is big data a privacy issue for users?
2. How easy it is to make money from peoples' data?
3. Are we running out of human data?

Recap questions

1. What is big data?
2. What defines big data?
3. What are the unique challenges of big data?
4. What does big data look like in the era of large language models?