“Big” Data

The challenges in modern data management in three numbers

CHIDATA

CHIDATA

Three Numbers

The cost per byte of storage today is **35 million** times less than it was in 1980.

The accuracy of object identification in images has improved from 50% to **90%** in since 2011.

79% of Americans are concerned about the data collected about them by internet companies.

CHIDATA

Three Numbers

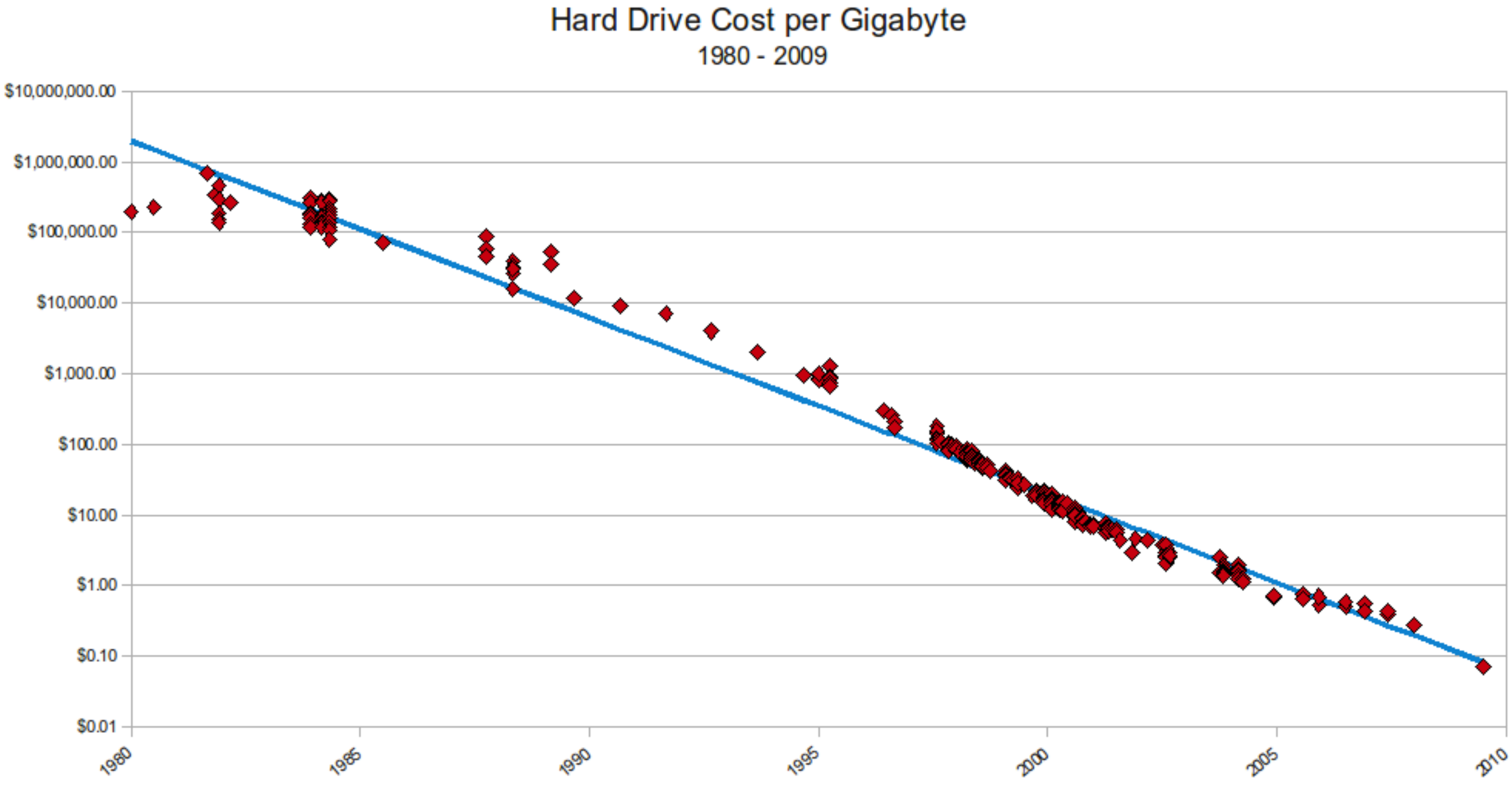
The cost per byte of storage today is **35 million** times less than it was in 1980.

The accuracy of object identification in images has improved from 50% to **90%** in since 2011.

79% of Americans are concerned about the data collected about them by internet companies.

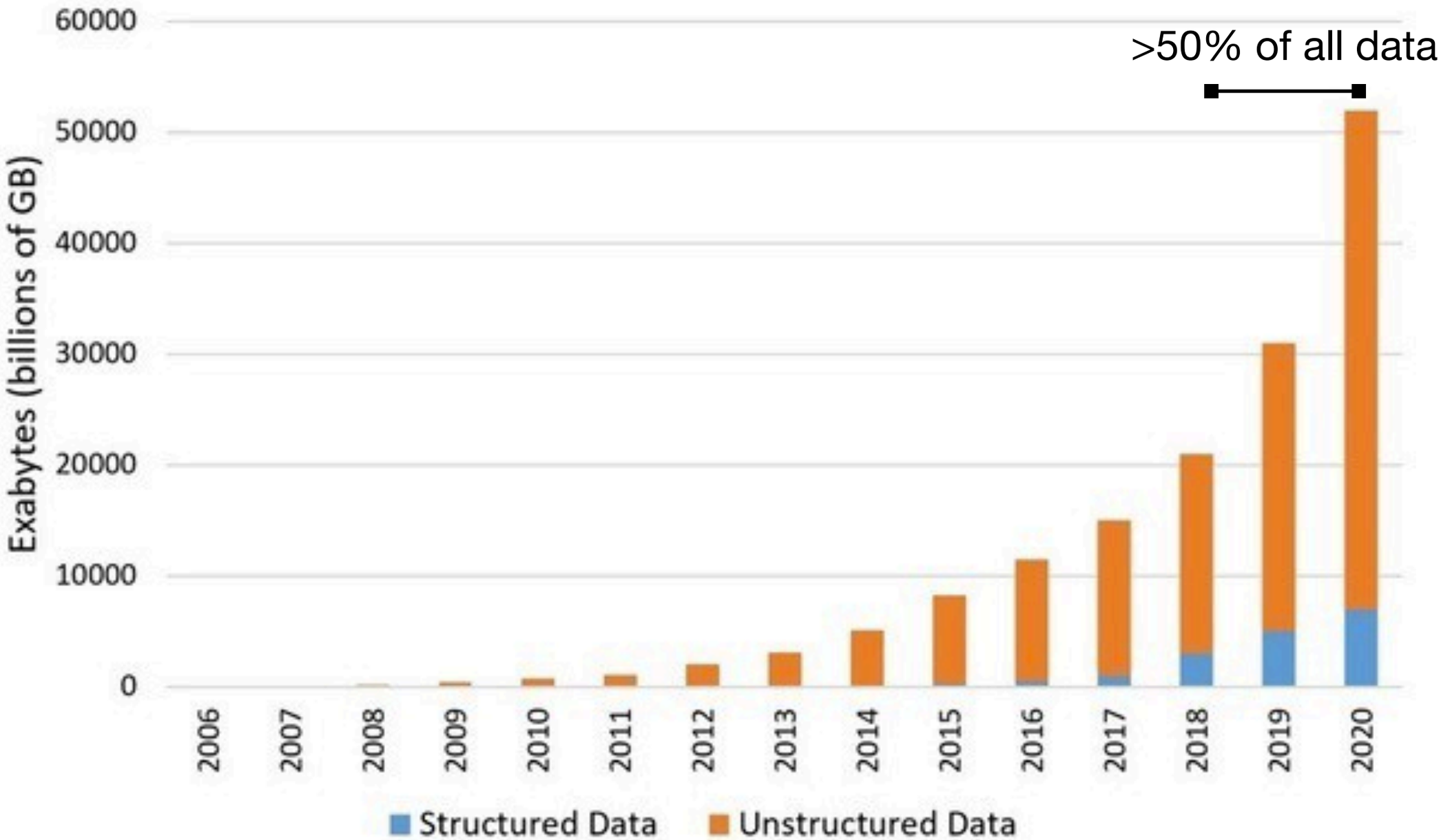
CHIDATA

Storage is Almost Free



CHIDATA

Storage is Almost Free



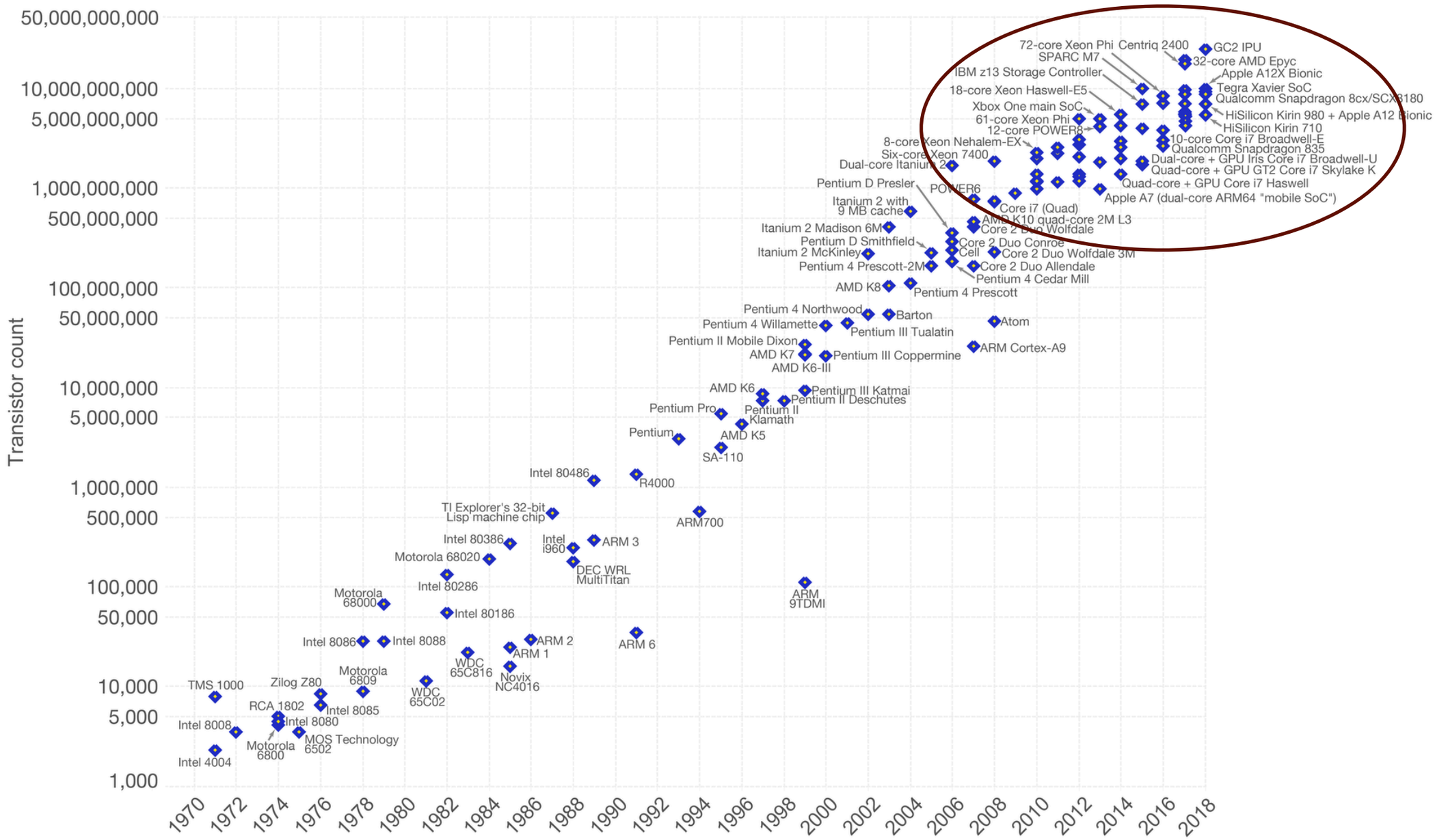
CHIDATA

Slowing Computers

Moore's Law – The number of transistors on integrated circuit chips (1971-2018)

Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)

CHIDATA

What Do You Do?

Divide-and-conquer: break a big task into smaller, more manageable tasks and distribute to multiple computers.

Dataflow systems, Parallel computing, Out-of-Core Algorithms

Smarter algorithms: approximate complex analytics with similar and more resource-efficient techniques.

Compression, Hashing, Approximate Matching, Indexing

CHIDATA

Human Effort Doesn't Scale

	A	B	C	D
1	**Date**	**Participant ID Number**	**Age**	**What parish do you live in?**
2	18-06-14	249	28	Naluwoli
3	17-06-14	2977	20	
4	17/06/2014	03500	52	Butansi
5	19/06/2014	4194	32	Naluwoli
6	17/06/2014	07420	19 1/2	Butansi
7	17/06/2014	07428	21	Naluwoli
8	17/06/2014	10011	Twenty	Butansi
9	17/06/2014	10061	30	Butansi
10	13-06-14	10431	27	Butansi
11	18/06/2014	10685	27 years	Butansi
12	19/06/2014	10920	19 years	Naluwoli
13	19/06/2014	10982	25	Naluwoli
14	13-06-14	11164	22	Naluwoli
15	17/06/2014	12138	Twenty-Two	Naluwoli

CHIDATA

What Do You Do?

Data Design Principles: How to design data models that are robust and easy to use for the desired application.

Structured/Semi-structured data sources, Privacy, Data Provenance

Automatic Checks: techniques that automatically detect faults or anomalies in data.

Integrity constraints, Functional Dependencies, Data Integration

CHIDATA

Three Numbers

The cost per byte of storage today is **35 million** times less than it was in 1980.

The accuracy of object identification in images has improved from 50% to **90%** in since 2011.

79% of Americans are concerned about the data collected about them by internet companies.

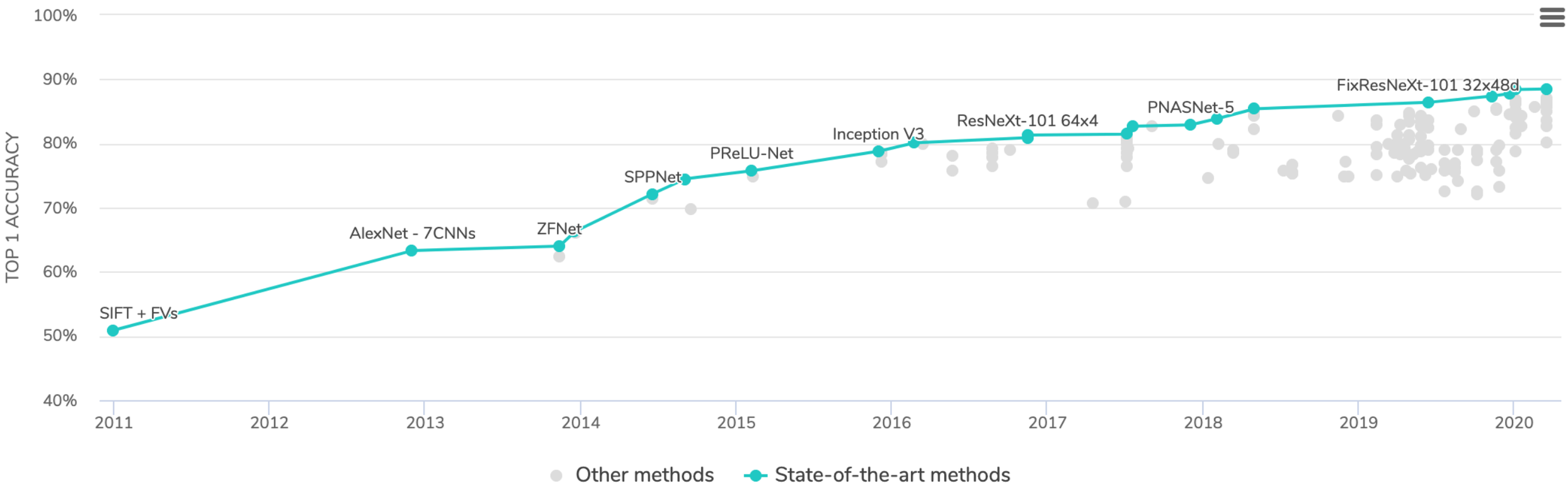
CHIDATA

Image Classification

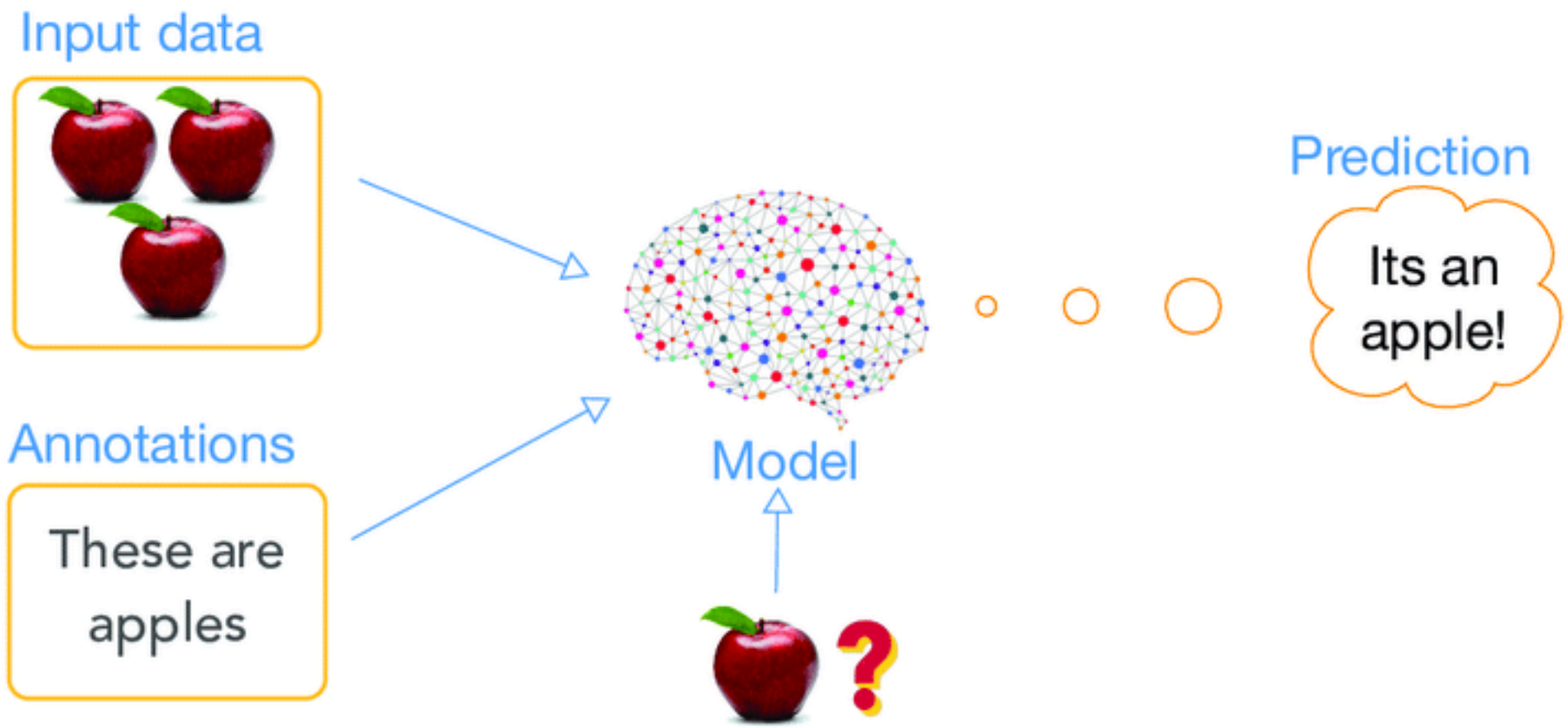
CAT

CHIDATA

Progress Since 2011



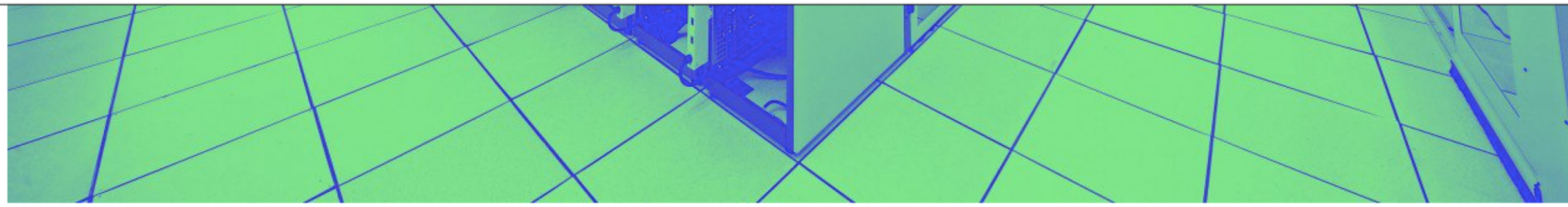
Machine Learning



Machine Learning Outsources Modeling To Data

CHIDATA

Expensive!



DEAN MOUHTAROPOULOS | GETTY; EDITED BY MIT TECHNOLOGY REVIEW

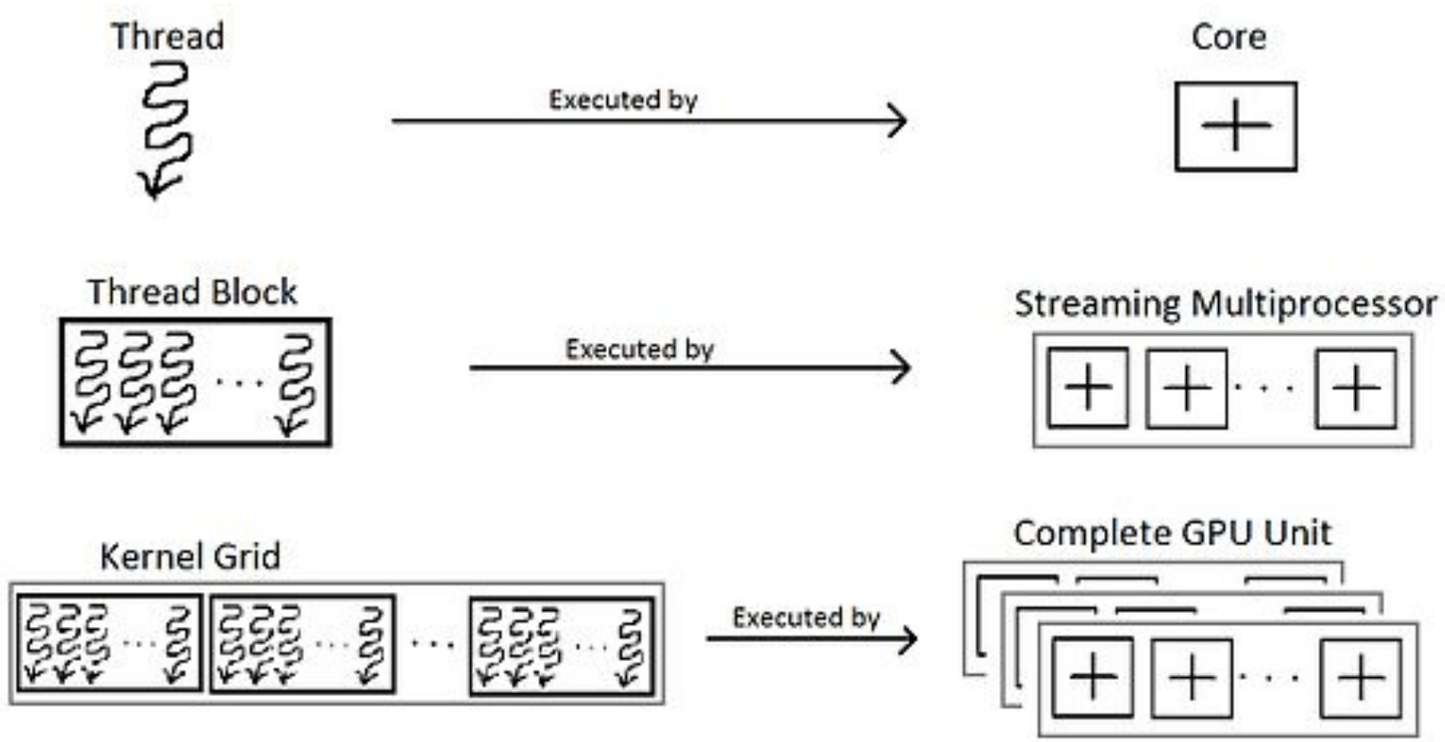
Artificial Intelligence / Machine Learning

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

CHIDATA

Specialized Hardware



CHIDATA

Challenges

Data Quality: model is only as good as the data.

Data integration, data cleaning

Scalability: systems are needed to fit such models at scale.

Systems for machine learning

CHIDATA

Three Numbers

The cost per byte of storage today is **35 million** times less than it was in 1980.

The accuracy of object identification in images has improved from 50% to **90%** in since 2011.

79% of Americans are concerned about the data collected about them by internet companies.

CHIDATA

Privacy Policy

CHIDATA

Tidbits From GDPR

"data subjects are **identifiable** if they can be directly or indirectly identified…an identification number, location data, an online identifier or one of several special characteristics, which expresses the physical, physiological, genetic, mental, commercial, cultural or social identity"

"personal data must be **erased immediately** where the data are no longer needed for their original processing purpose, or the data subject has withdrawn his consent"

"**accurate** and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay"

CHIDATA

Challenges

Data Organization: how to store and index data
Data integrity, indexes, identifiability

Data Governance: policies regarding accessing and managing data

CHIDATA

Three Numbers

The cost per byte of storage today is **35 million** times less than it was in 1980.

The accuracy of object identification in images has improved from 50% to **90%** in since 2011.

79% of Americans are concerned about the data collected about them by internet companies.