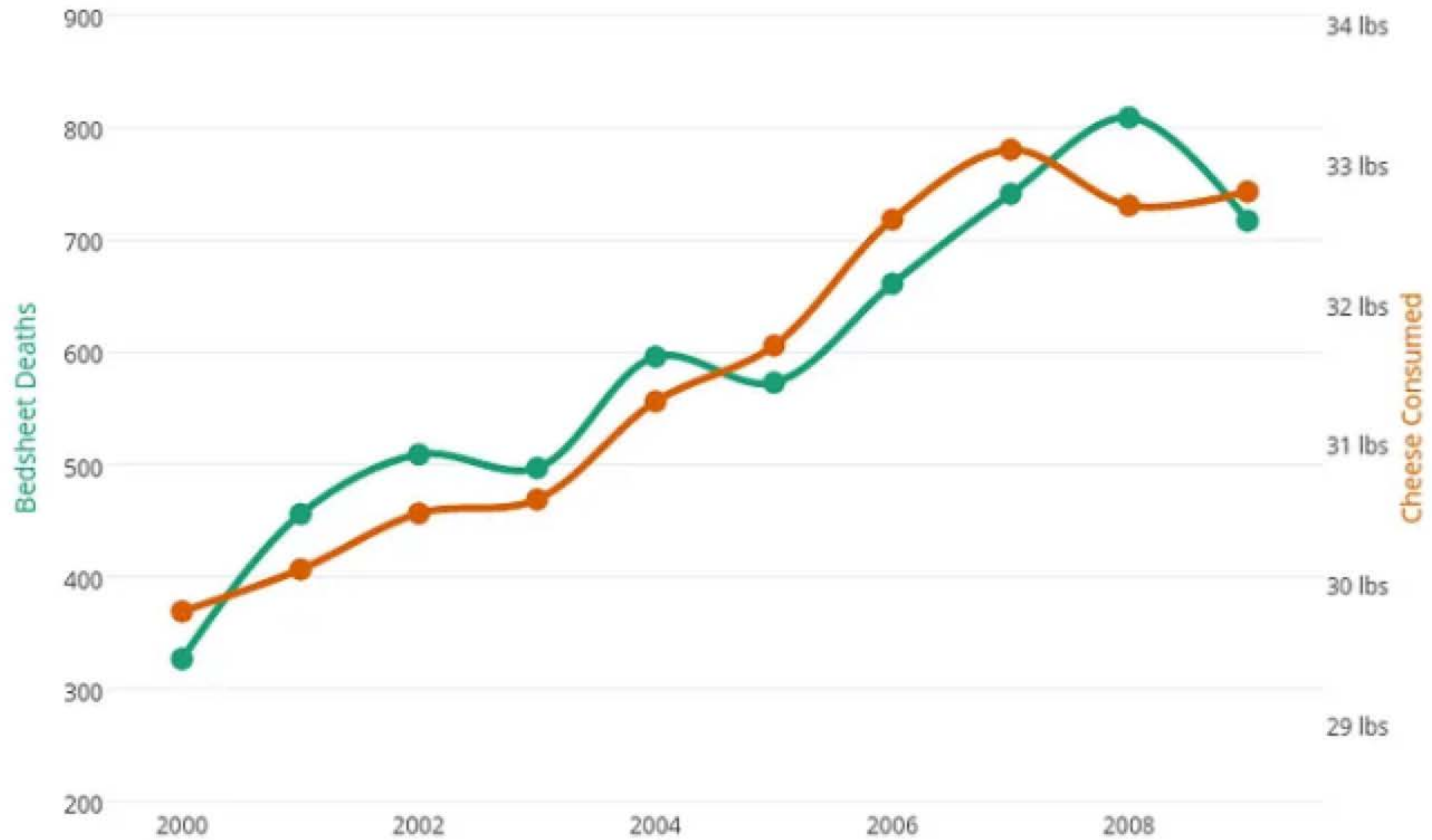


Open Data's Next Era

Opendata.ch Forum 2024

**Scott Smith — Managing
Partner, Changeist**

Deaths by Becoming Tangled in Bedsheets &
Per Capita Cheese Consumption



Working Theory:

**The history of data
correlates with the
history of Taylor Swift**



Look Back Twice as Far as You Look Forward

"The texture of past events can be used to connect the dots of present indicators and thus reliably map the future's trajectory—provided one looks back far enough."

—Paul Saffo, Harvard Business Review

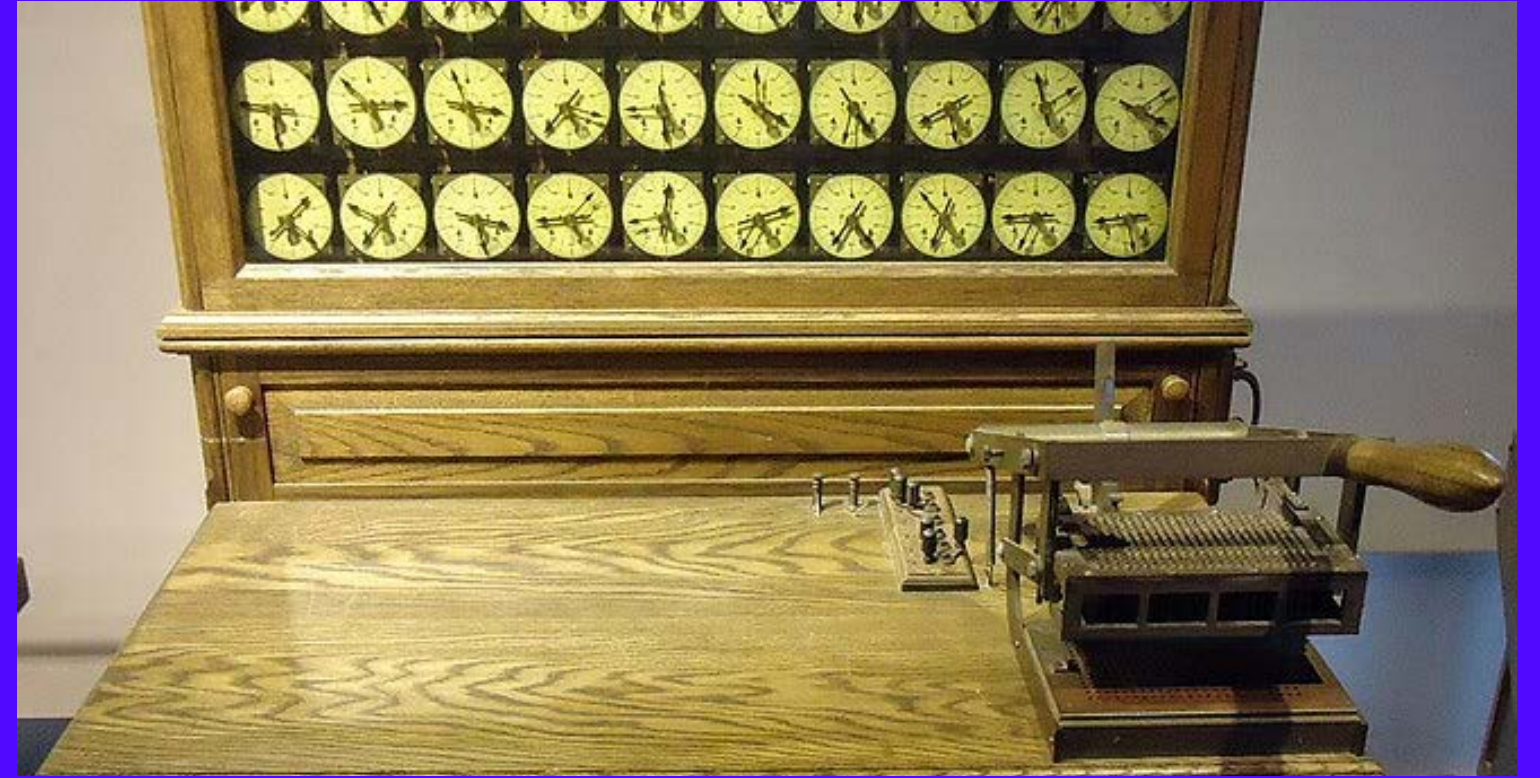
The Manual Era of Data

Pre-20th c.



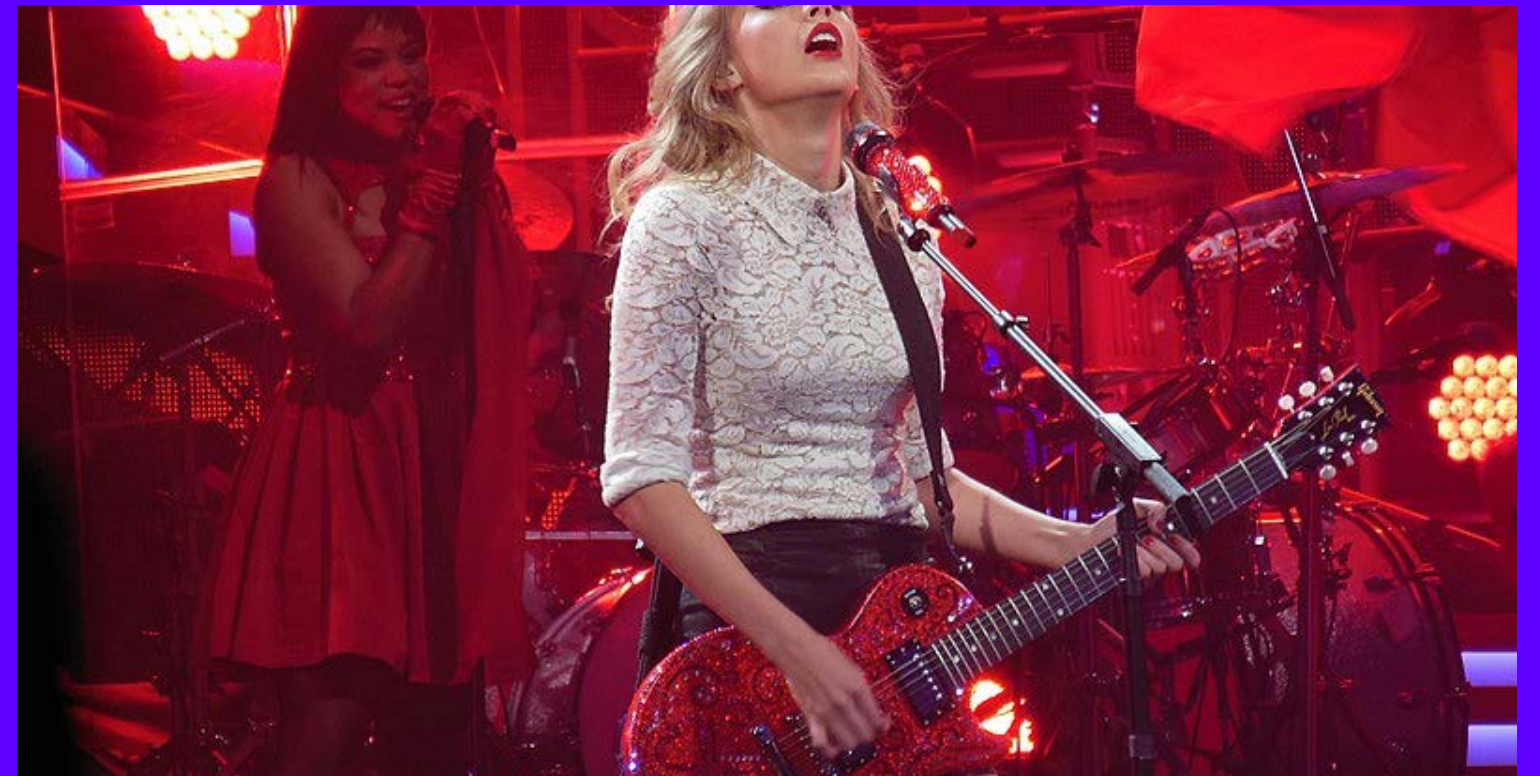
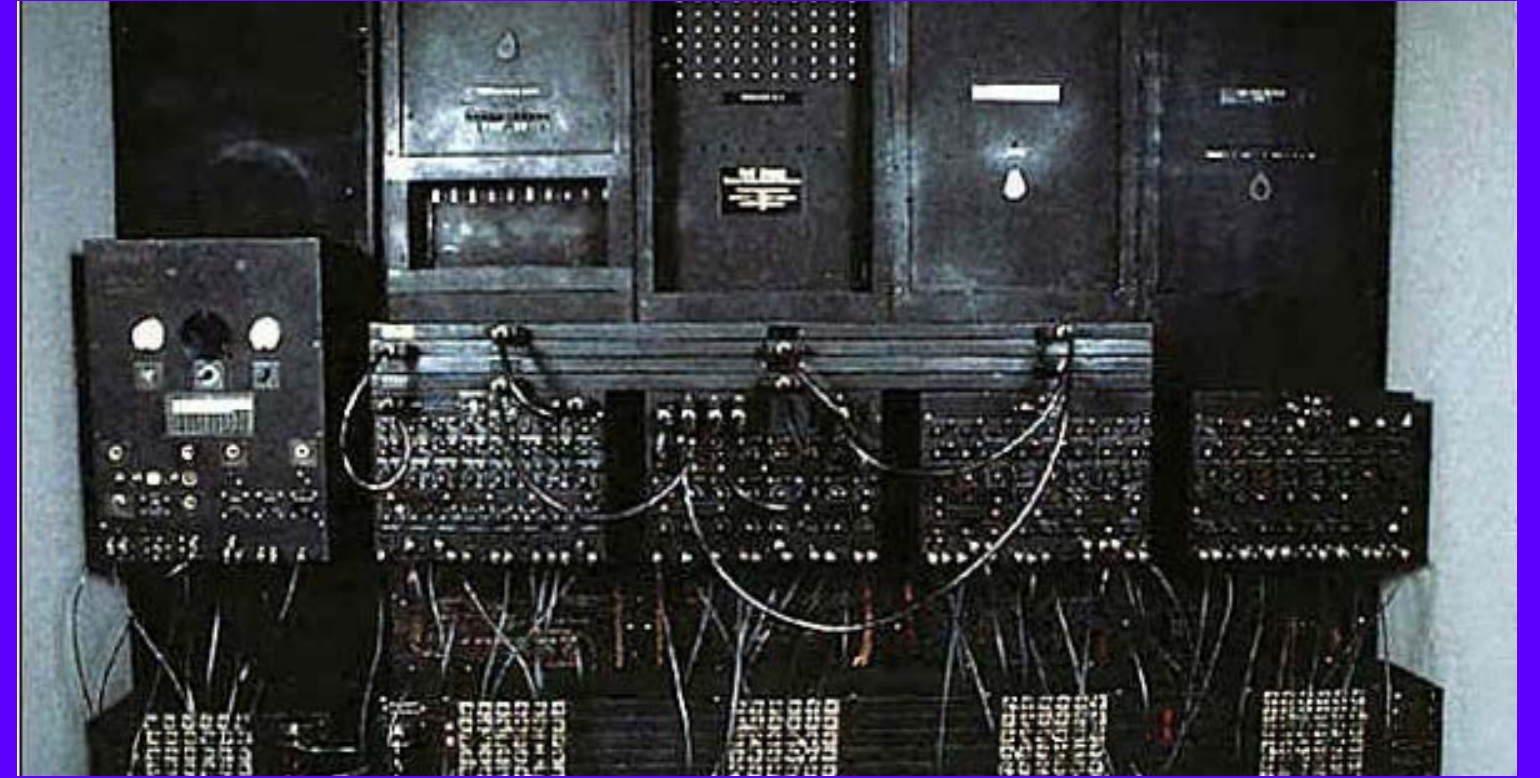
The Mechanical Era of Data

Late 19th - Early 20th
C.



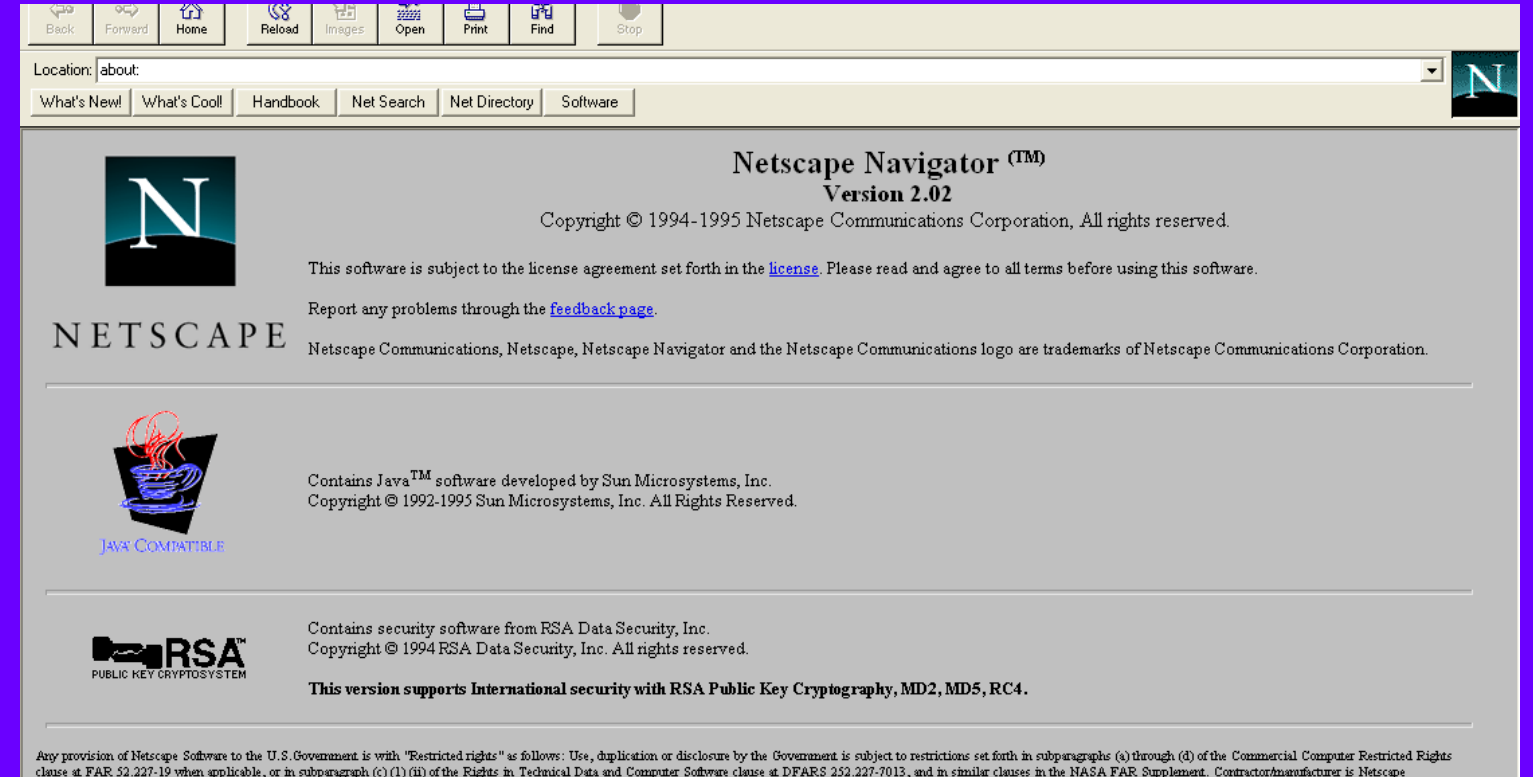
The Electronic Era of Data

Mid-Late 20th C.



The Digital Era of Data

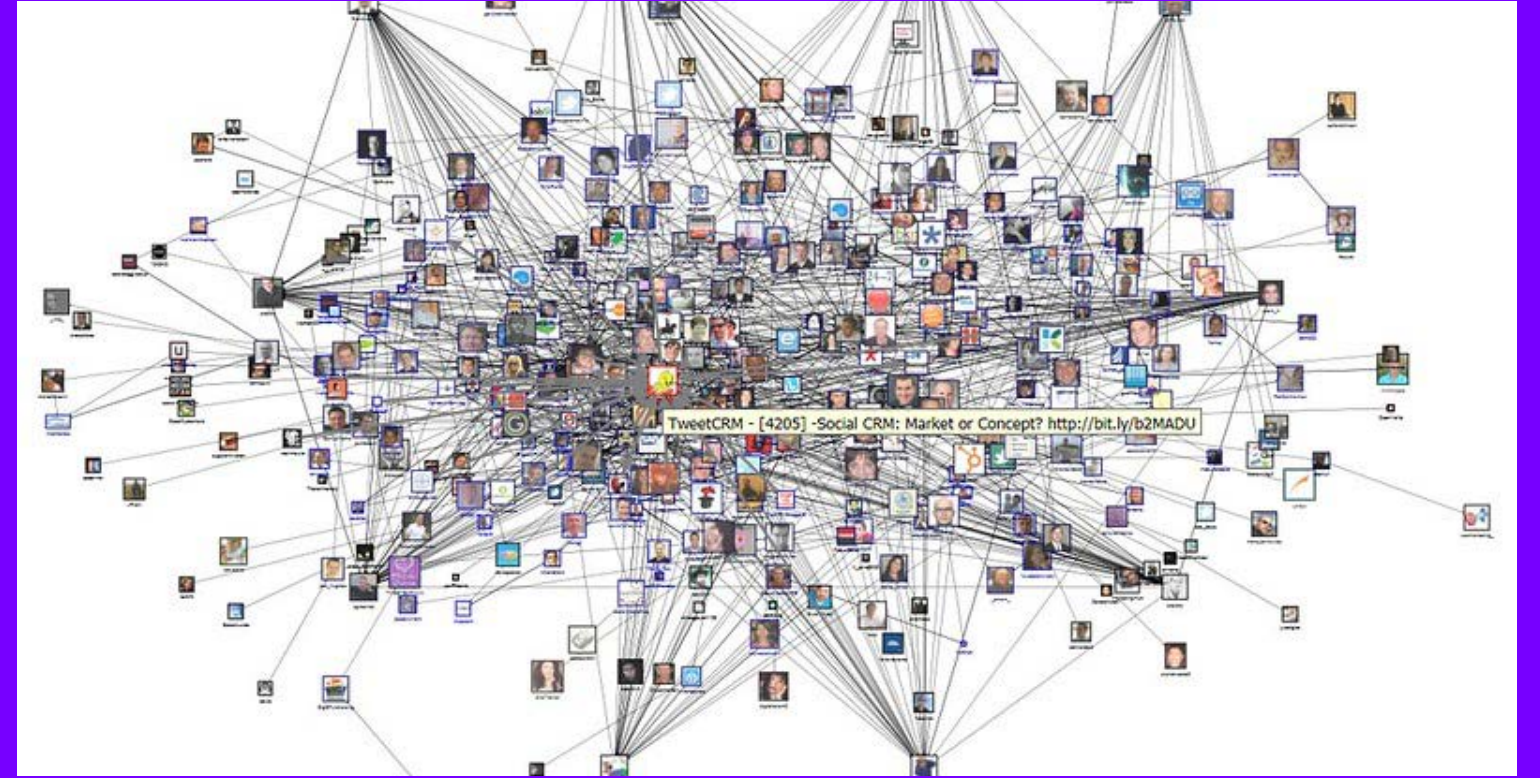
Late 20th - Early 21st C.

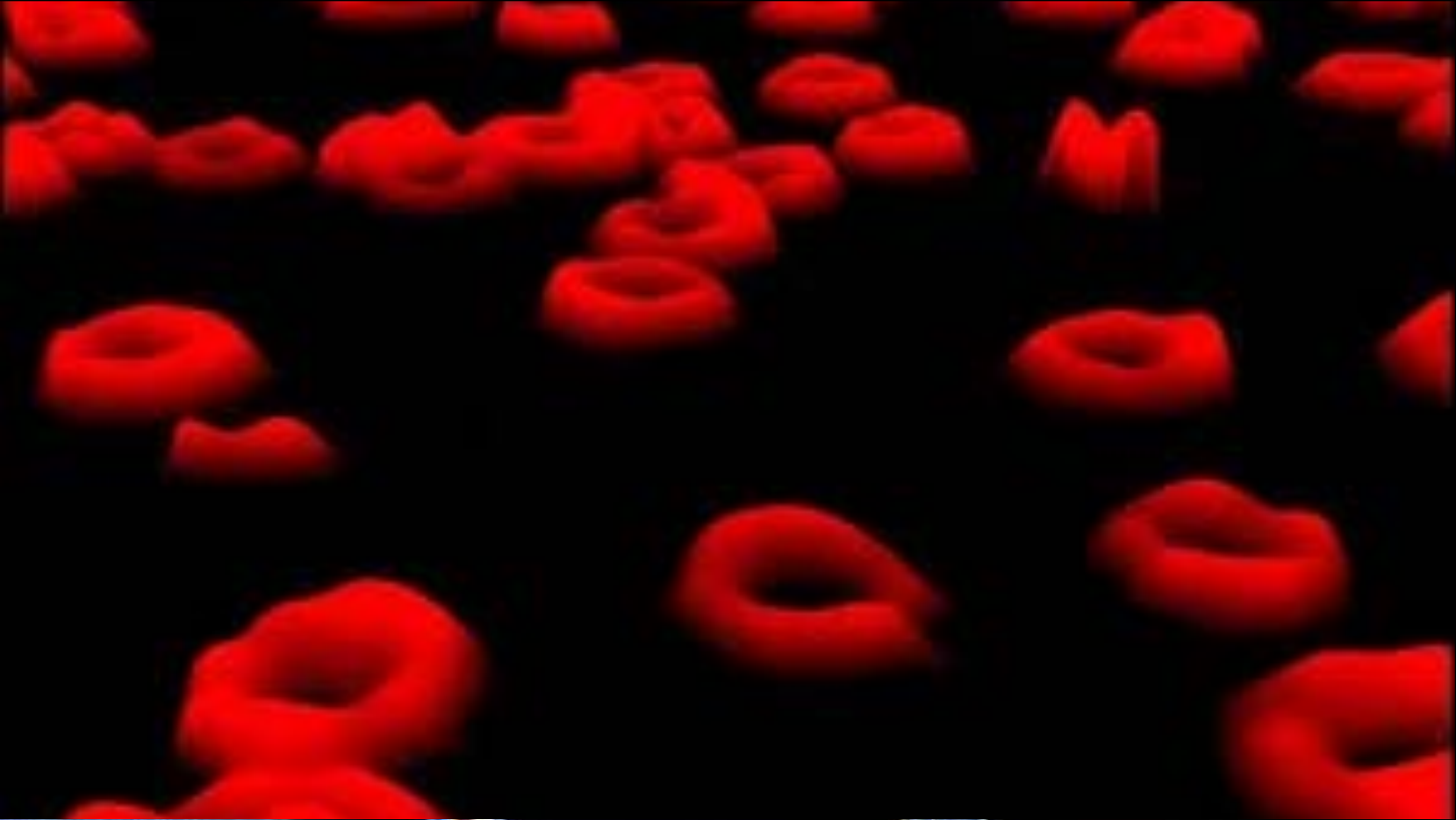
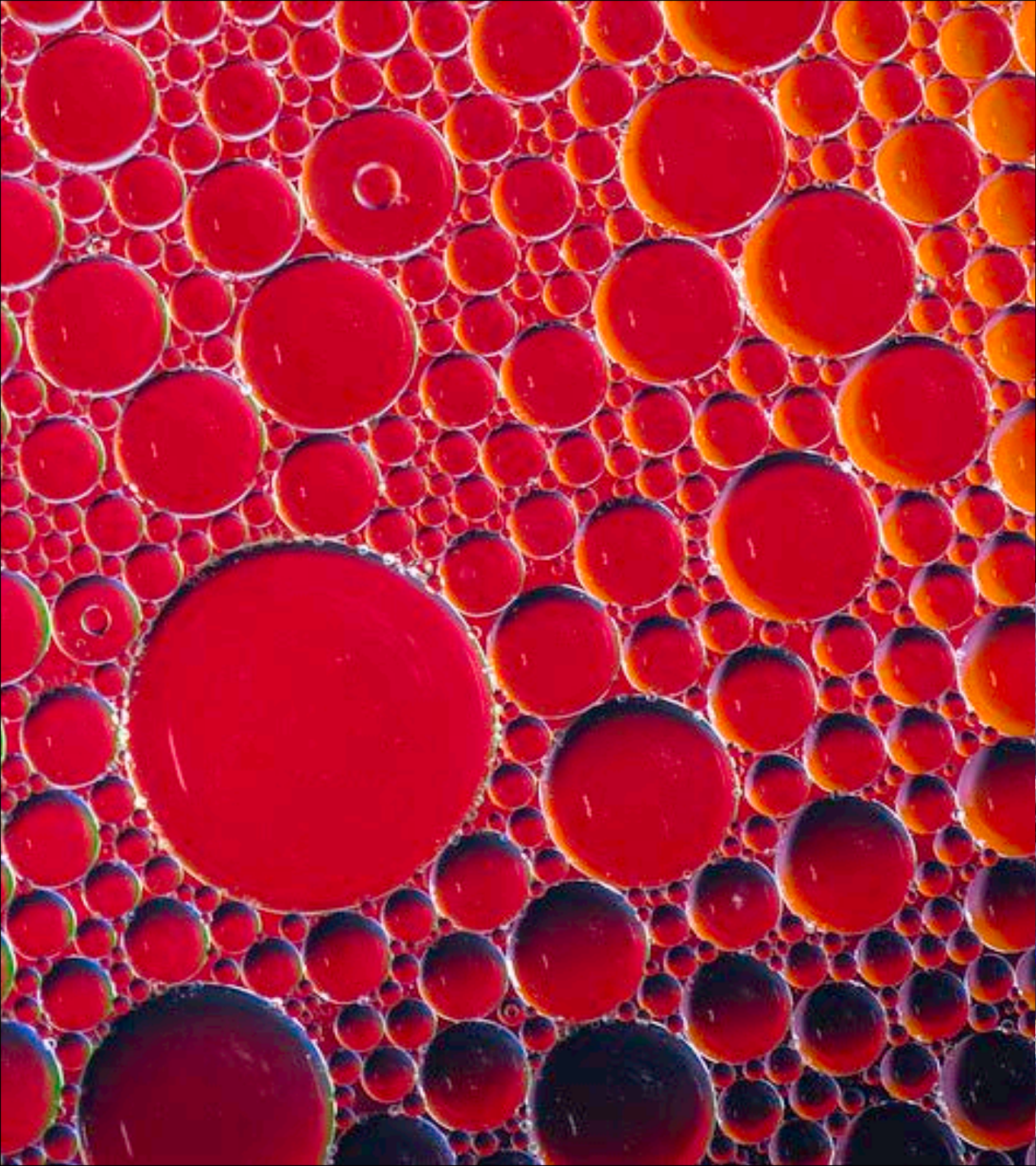


The Big Data Era

with added
Algorithms

Mid 2000s - mid
2020s





Attention Is All You Need

Ashish Vaswani*

Google Brain

avaswani@google.com

Noam Shazeer*

Google Brain

noam@google.com

Niki Parmar*

Google Research

nikip@google.com

Jakob Uszkoreit*

Google Research

usz@google.com

Llion Jones*

Google Research

llion@google.com

Aidan N. Gomez*[†]

University of Toronto

aidan@cs.toronto.edu

Łukasz Kaiser*

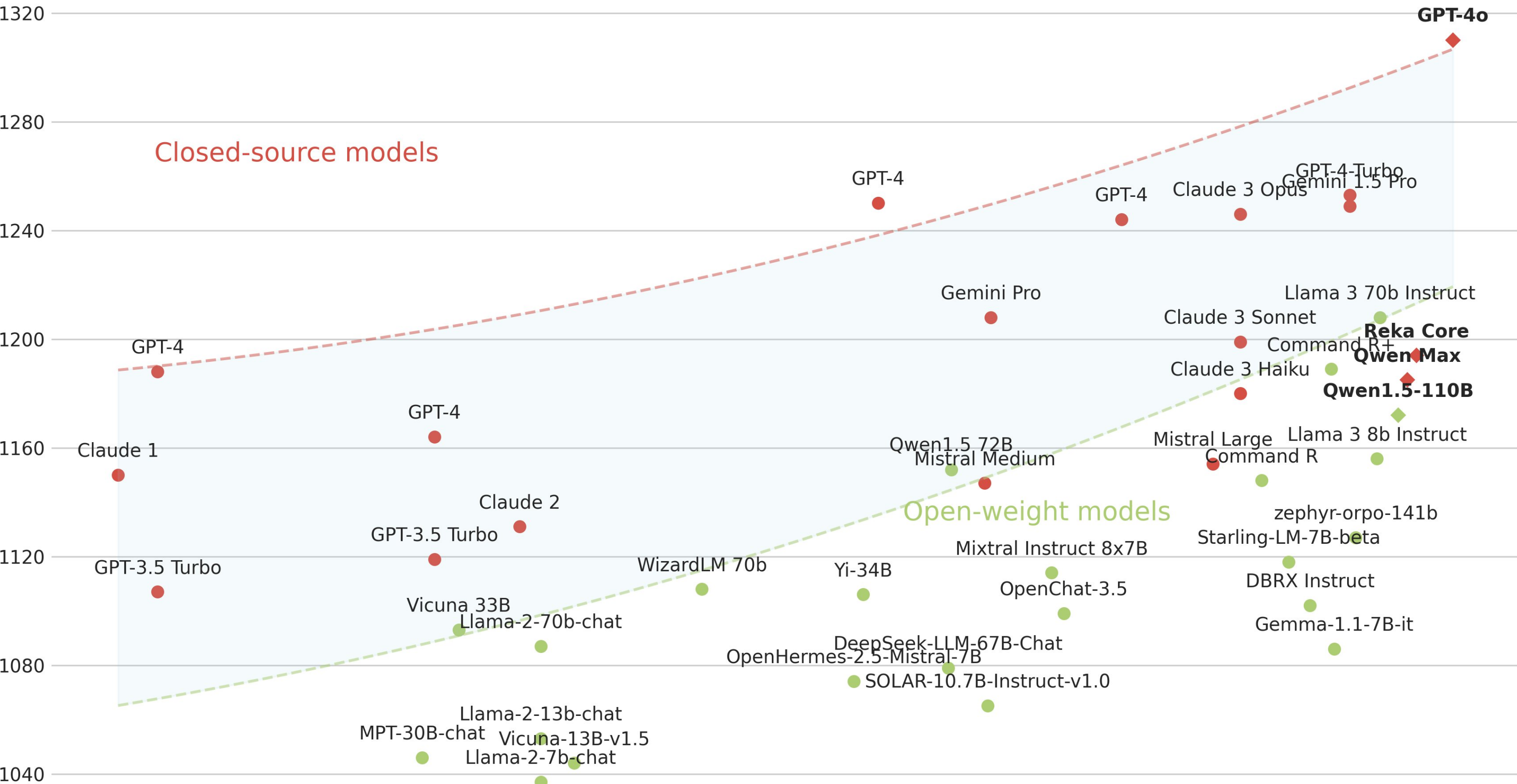
Google Brain

lukaszkaizer@google.com

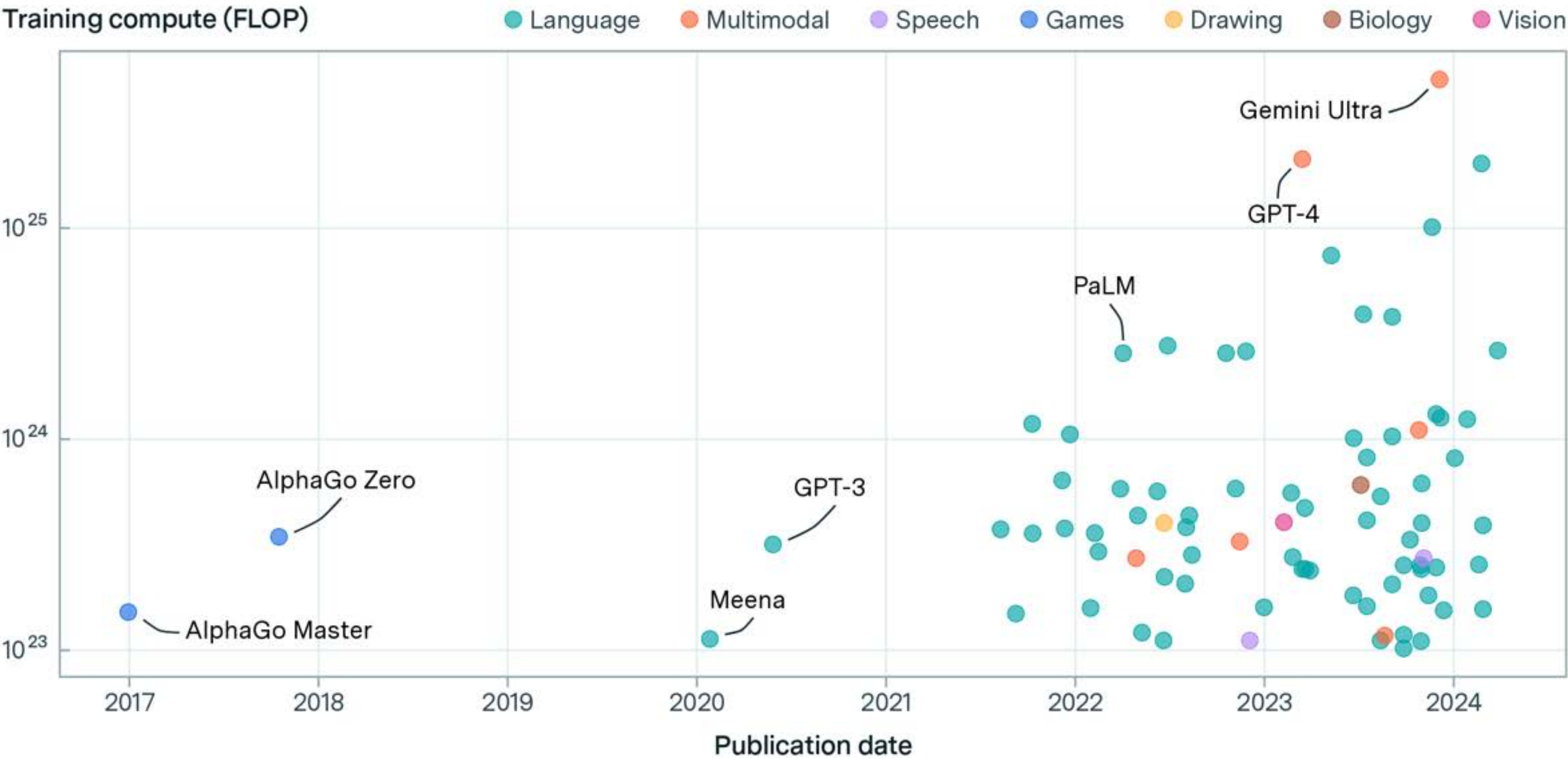
Illia Polosukhin*[‡]

illia.polosukhin@gmail.com

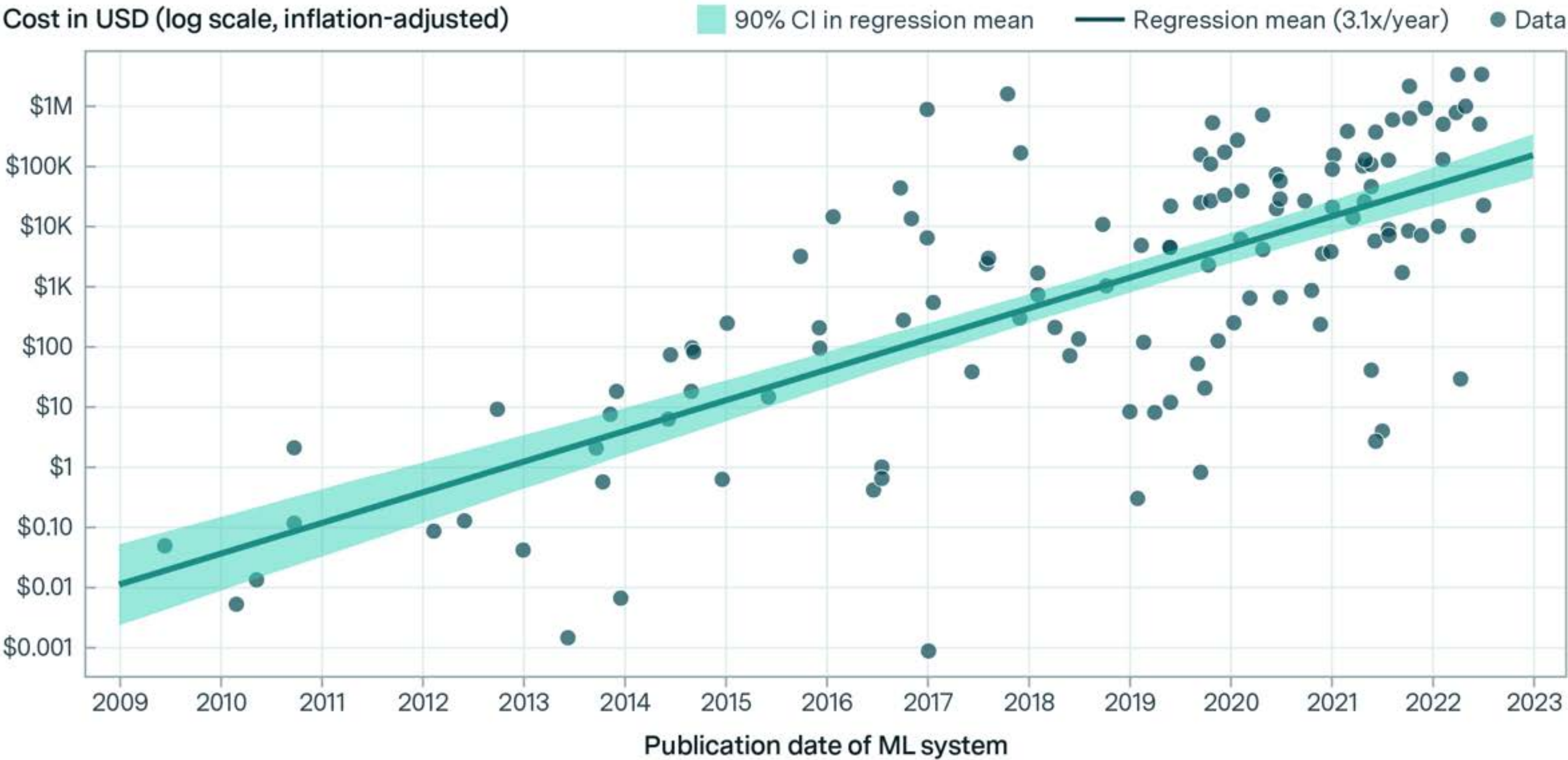
Arena ELO



Compute-intensive models by domain and publication date



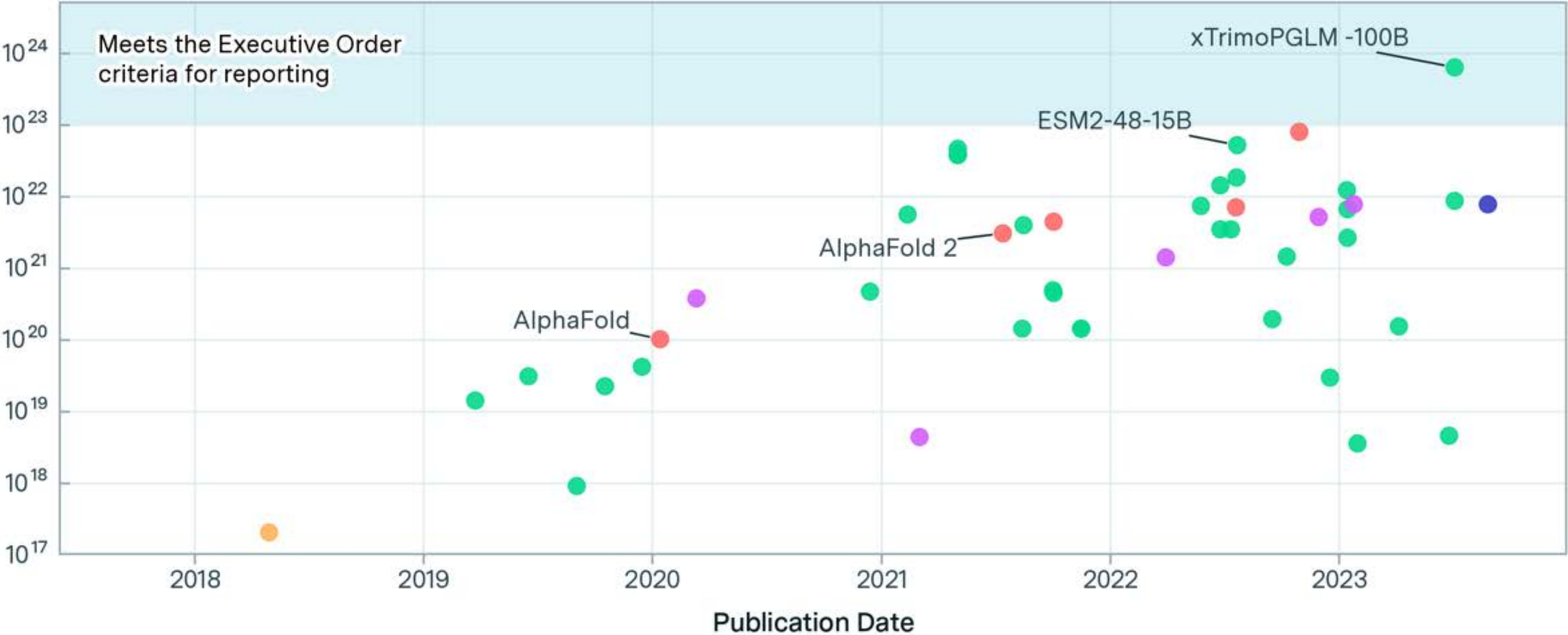
Cost of training compute for notable ML systems



Compute used to train biological sequence models

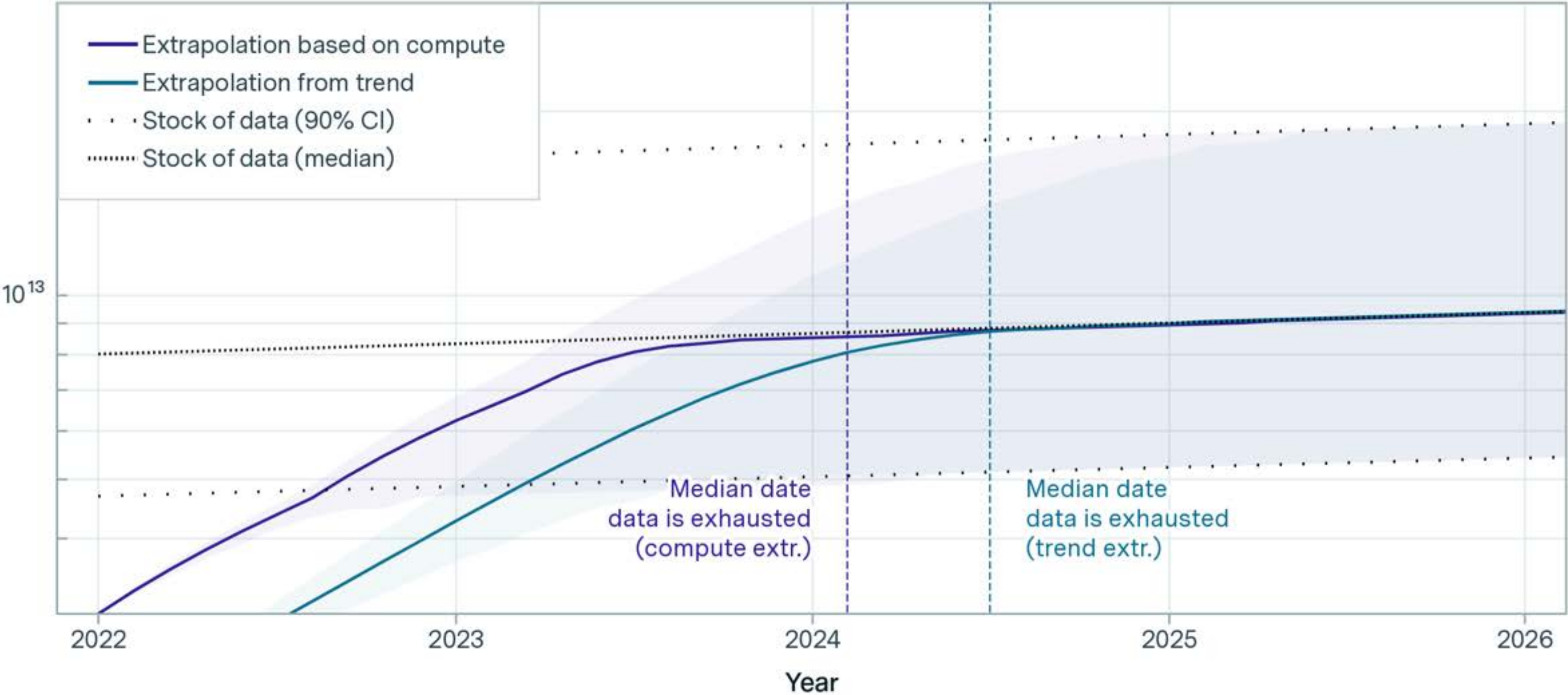
- General-purpose biological sequence model
- Protein design model
- Contact and distance prediction model
- Protein-ligand interaction prediction model
- Protein structure prediction model

Training compute (FLOP)

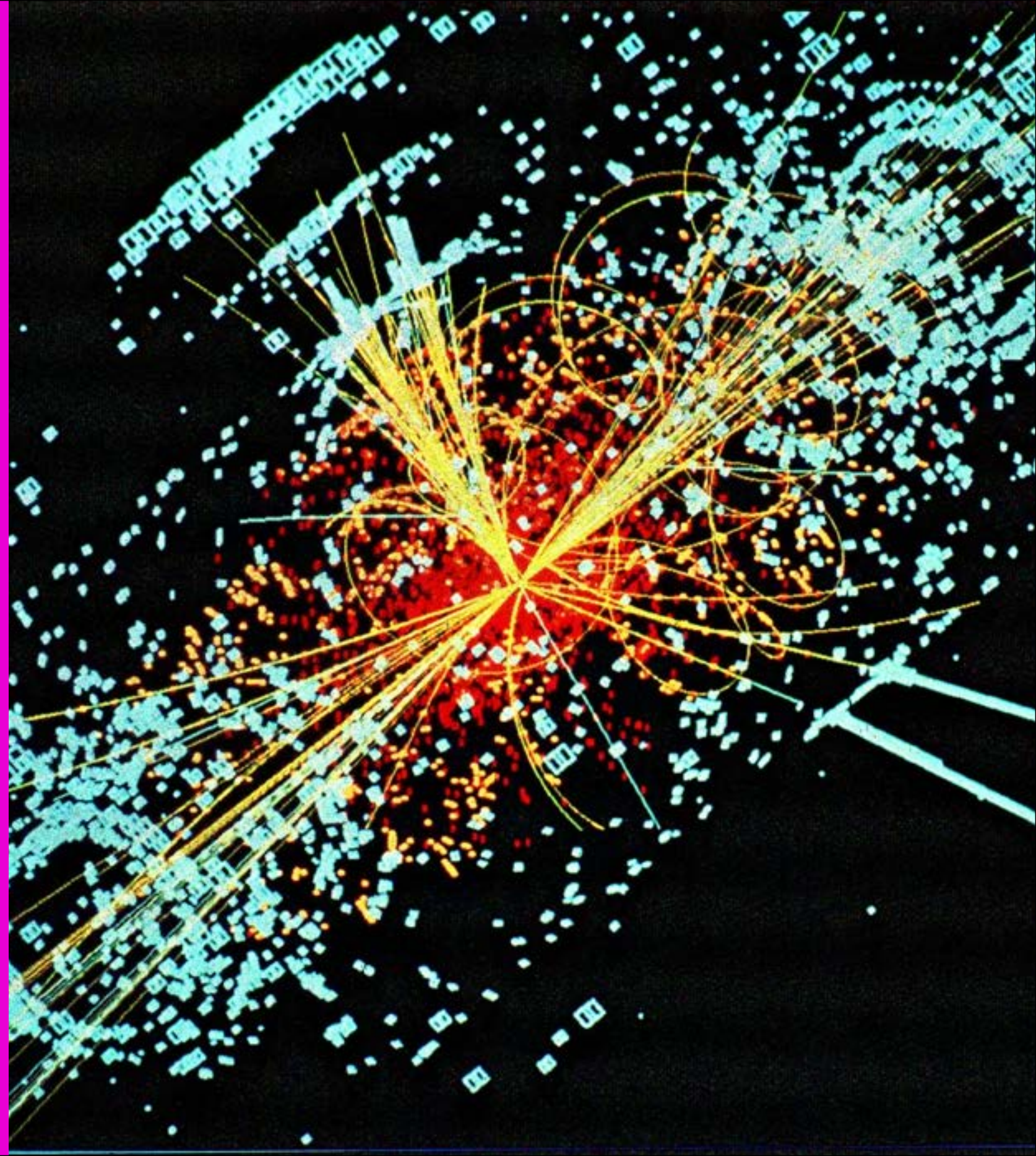


Projections of data usage (for high-quality language data)

Number of words (log)



**Everything
everywhere
is data, all at
once!**



"Yeah, I mean look, the conversation around data is really important. We obviously need to know where that data comes from. We just put out a post this week actually about this exact topic which is basically that there needs to be like a content ID system for AI that lets creators understand as they create stuff where it's going, who's training on it, being able to opt in and out of training, being able to opt in and out of use. Also like, on the other side of that, being able to actively allow your content to be put into a model or to be accessed by a model because there may be this other economic opportunity on the other end of this. And that's something we're exploring too, how do you actually go create an entirely different social contract with the web, with creators, with publishers, where as these models go off in the world and do things that are useful, create value, to the extent they're able to like, reference and incorporated content from the web, there should be some sort of way that people can kind of get a benefit from that. So, yeah, we're looking at this problem, it's really hard. We don't have all the answers yet."

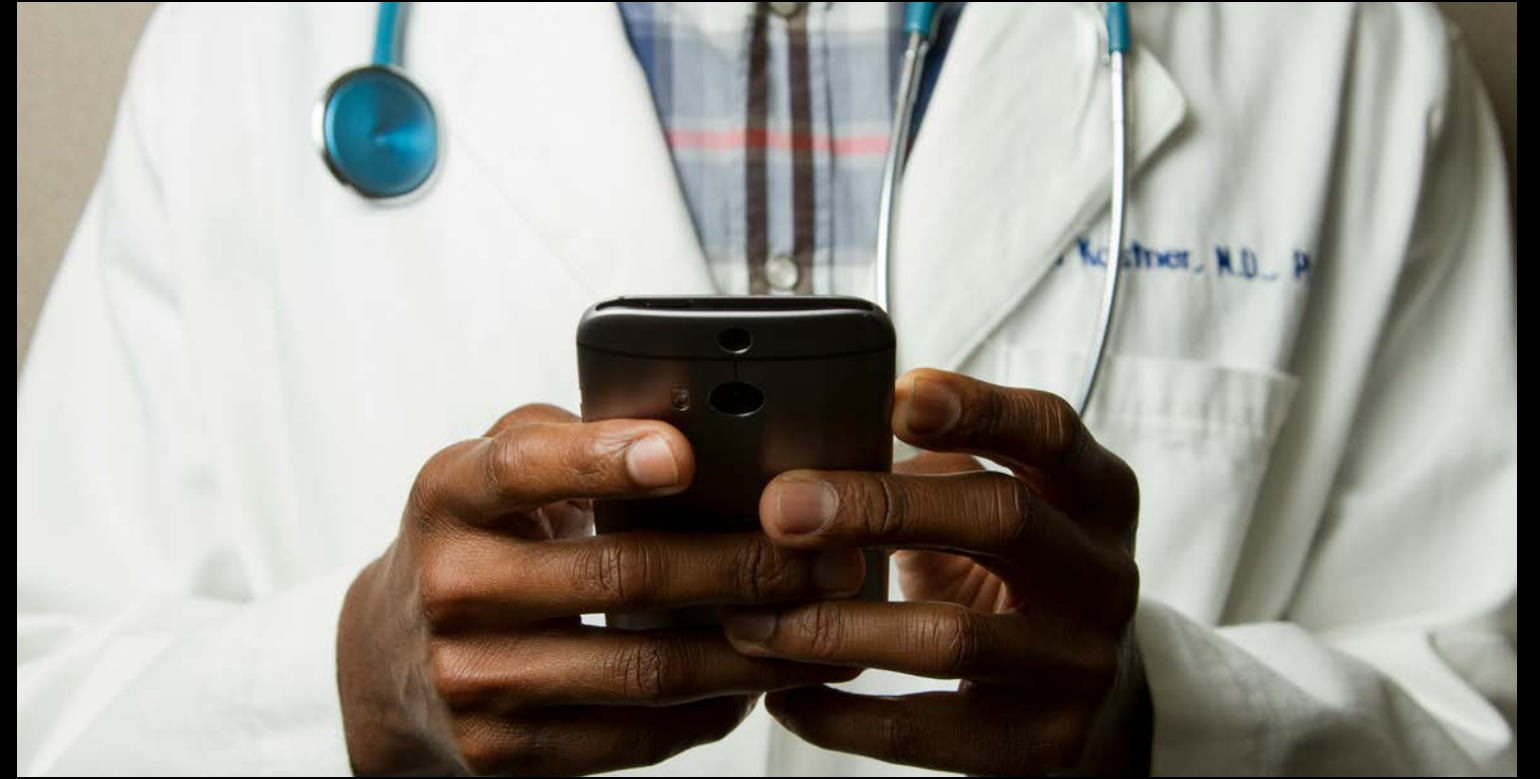
—Brad Lightcap, OpenAI COO, Bloomberg Technology Summit

Senators Propose \$32 Billion in Annual A.I. Spending but Defer Regulation

Their plan is the culmination of a yearlong listening tour on the dangers of the new technology.

 Share full article





Strong governance and regulation



I. A Guarded Commons

II. Corporate Panopticon



Predominantly
public control

Predominantly
private control



IV. Open Commons

III. Data Wild West

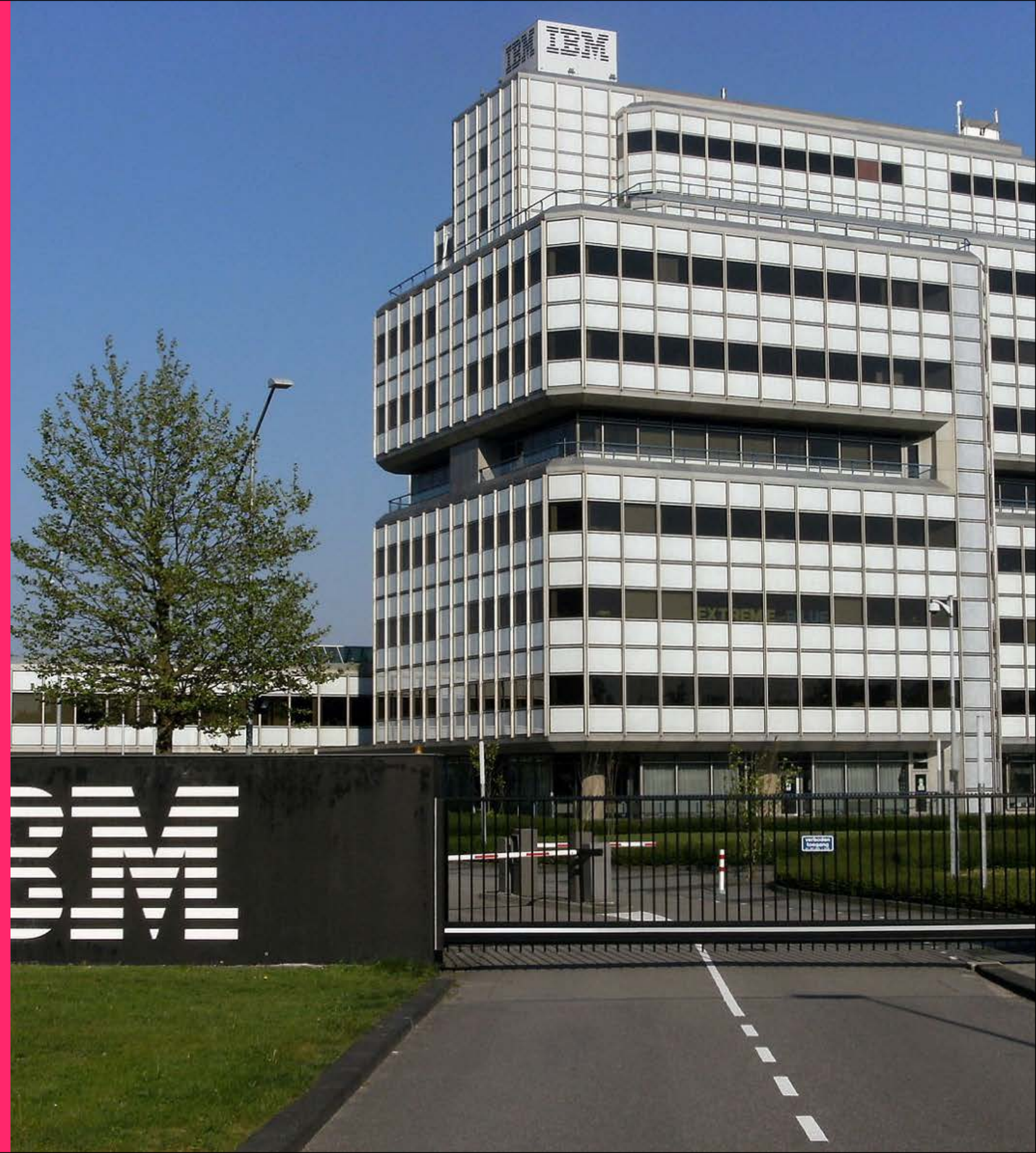


Weak governance and regulation

I. Guarded Commons

High Regulation, Public Control

Strict regulations govern data usage, but private corporations hold the majority of data control. These companies operate under rigorous oversight to ensure compliance with privacy and ethical standards.



II. Corporate Panopticon

High Regulation, Private Control

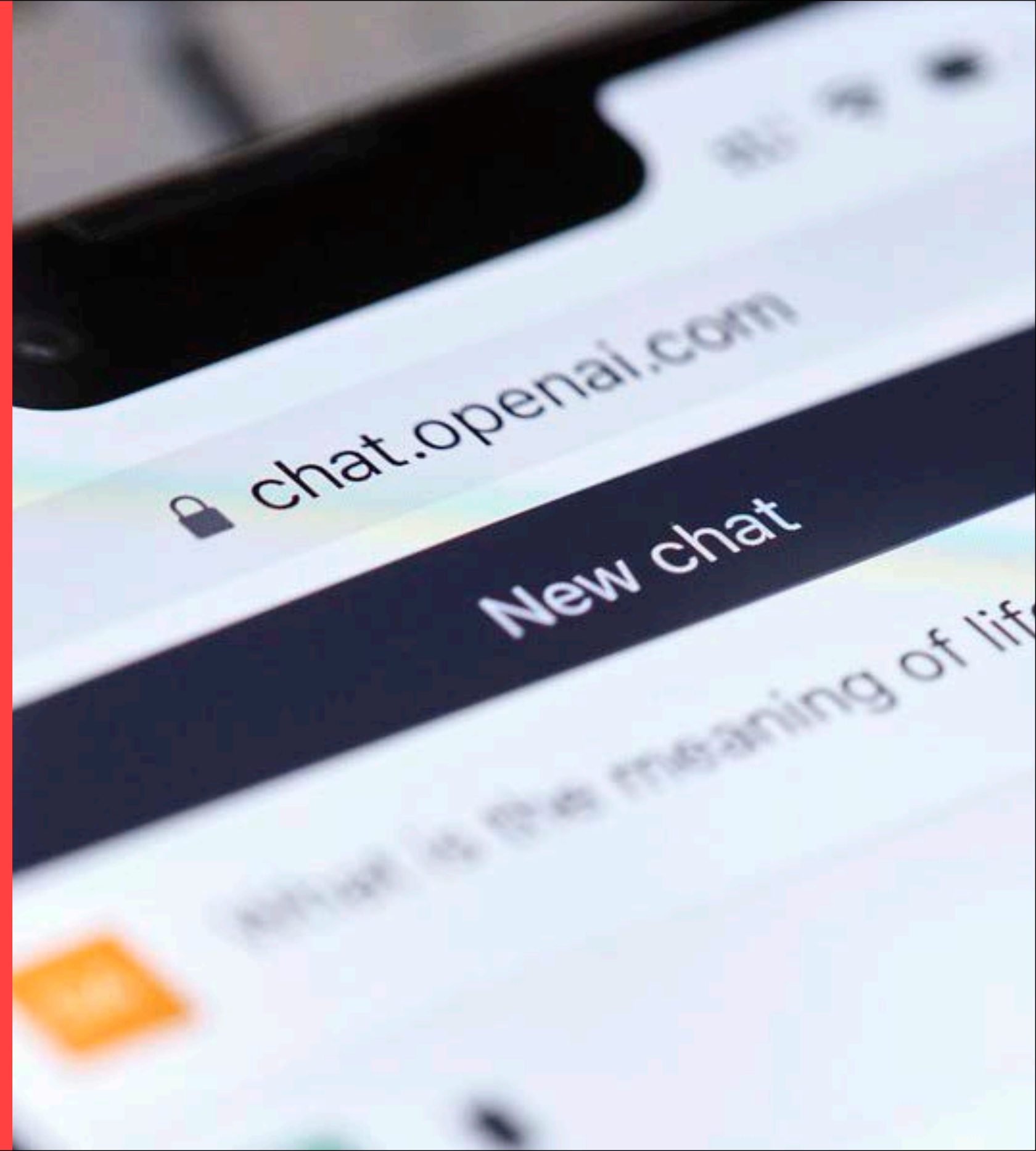
Strict regulations govern data usage, but private corporations hold the majority of data control. These companies operate under rigorous oversight to ensure compliance with privacy and ethical standards.



III. Data Wild West

Low Regulation, Private Control

Minimal regulatory oversight allows private corporations to freely collect, manage, and use data. There is little restriction on how data is used, leading to rapid innovation but also significant risks.



IV. Open Commons

Low Regulation, Private Control

Minimal regulatory oversight, but data is predominantly controlled by public entities that prioritize open access and transparency. This environment promotes widespread data sharing for public benefit.



Strong governance and regulation



I. A Guarded Commons

II. Corporate Panopticon



Predominantly
public control

Predominantly
private control



IV. Open Commons

III. Data Wild West



Weak governance and regulation

*Build Into The Technology

*Build Public Tools

*Support Public Data Infrastructure

```

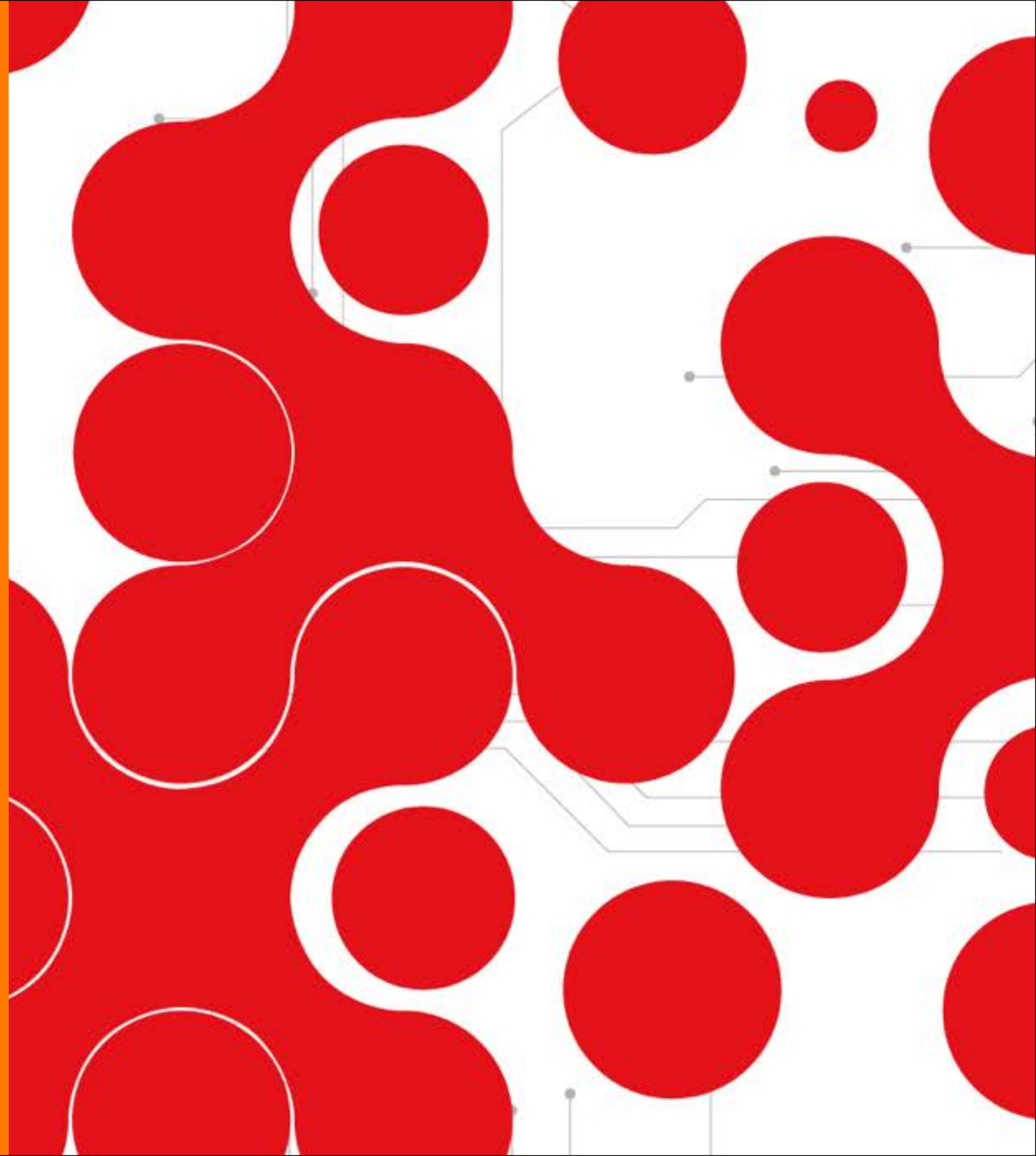
background-color: #28a745; border-color: #28a745;
.btn-outline-success: not(:disabled):not(.disabled):active:
.btn-outline-success: not(:disabled):not(.disabled):active:
background-color: #28a745; border-color: #28a745;
.btn-outline-success: not(:disabled):not(.disabled):active:
.btn-outline-success: not(:disabled):not(.disabled):active:
background-color: #17a2b8; border-color: #17a2b8;
.btn-outline-info: not(:disabled):not(.disabled):active:
.btn-outline-info: not(:disabled):not(.disabled):active:
background-color: #17a2b8; border-color: #17a2b8;
.btn-outline-info: not(:disabled):not(.disabled):active:
.btn-outline-info: not(:disabled):not(.disabled):active:
background-color: #ffc107; border-color: #ffc107;
.btn-outline-warning: not(:disabled):not(.disabled):active:
.btn-outline-warning: not(:disabled):not(.disabled):active:
background-color: #ffc107; border-color: #ffc107;
.btn-outline-warning: not(:disabled):not(.disabled):active:
.btn-outline-warning: not(:disabled):not(.disabled):active:
background-color: #212529; border-color: #212529;
.btn-outline-dark: not(:disabled):not(.disabled):active:
.btn-outline-dark: not(:disabled):not(.disabled):active:
background-color: #212529; border-color: #212529;
.btn-outline-dark: not(:disabled):not(.disabled):active:
.btn-outline-dark: not(:disabled):not(.disabled):active:
background-color: #dc3545; border-color: #dc3545;
.btn-outline-danger: not(:disabled):not(.disabled):active:
.btn-outline-danger: not(:disabled):not(.disabled):active:
background-color: #dc3545; border-color: #dc3545;
.btn-outline-danger: not(:disabled):not(.disabled):active:
.btn-outline-danger: not(:disabled):not(.disabled):active:
background-color: #6c757d; border-color: #6c757d;
.btn-outline-light: not(:disabled):not(.disabled):active:

```

- *Strengthen Public Data Policies Now (If Not Yesterday)
- *Incentivise Public Control
- *Enhance Regulatory Frameworks (New Bodies?)
- *Promote Collaborative Governance
- *Educate the Public



**This
makes
you the
future!**



"Technological tools, like continuous political participation, may help us become more responsive to fundamental societal questions as a culture, without compromising the core values that define our civilization."

**–Ben Cervený, Foundation for
Public Code, Noema Magazine**

Thank you

Merci

Danke

Email:

ssmith@changeist.com

Web: changeist.com

**Twitter/Bluesky/Masto/IG:
[@changeist](https://twitter.com/changeist)**

**Images: Smithsonian, Jana Beaver, Ronald Woan,
Computer History Museum, Wikimedia Foundation,
Marc Smith, Damien Walmsley, Cory Doctorow,
Unsplash**