Anti-data-science

the Taleb's barbell in the AI world

Alex Honchar | AI Ukraine | October 2021

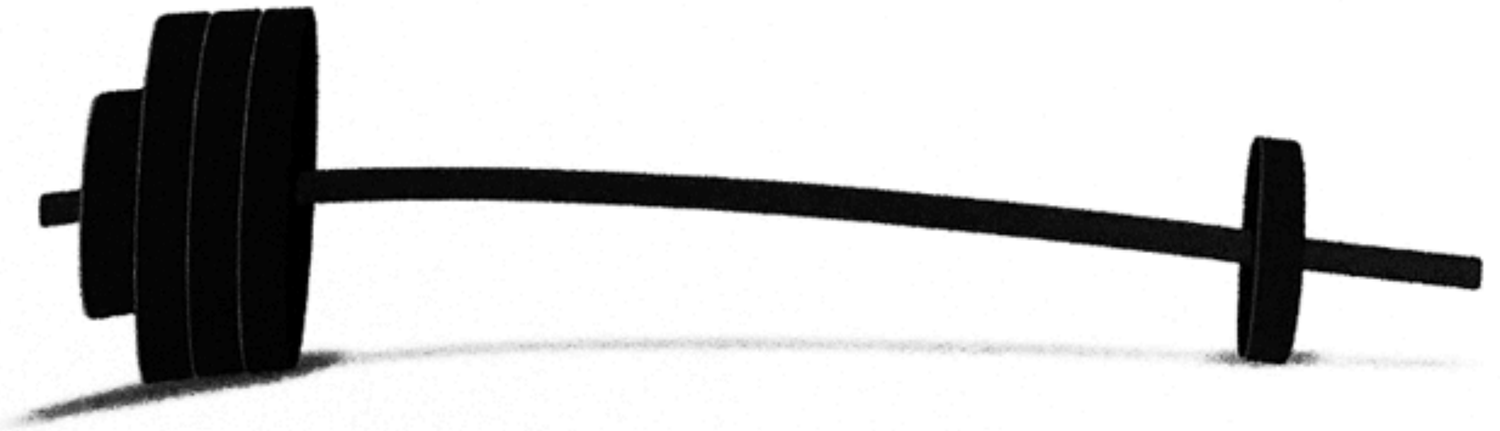
Why I am here?

- ~~Bla-bla-bla about company I've founded, universities I've taught in, systems I've built, articles published, etc~~
- What novel and contrarian I can bring to the table:
  - how to build a company with culture of independent experts who are engaged in projects with other institutions
  - **why to build data science and AI solutions where there is no data and no AI**

The Barbell Strategy (Taleb, 2007)

Low risk
(the known)

High risk
(the unknown)



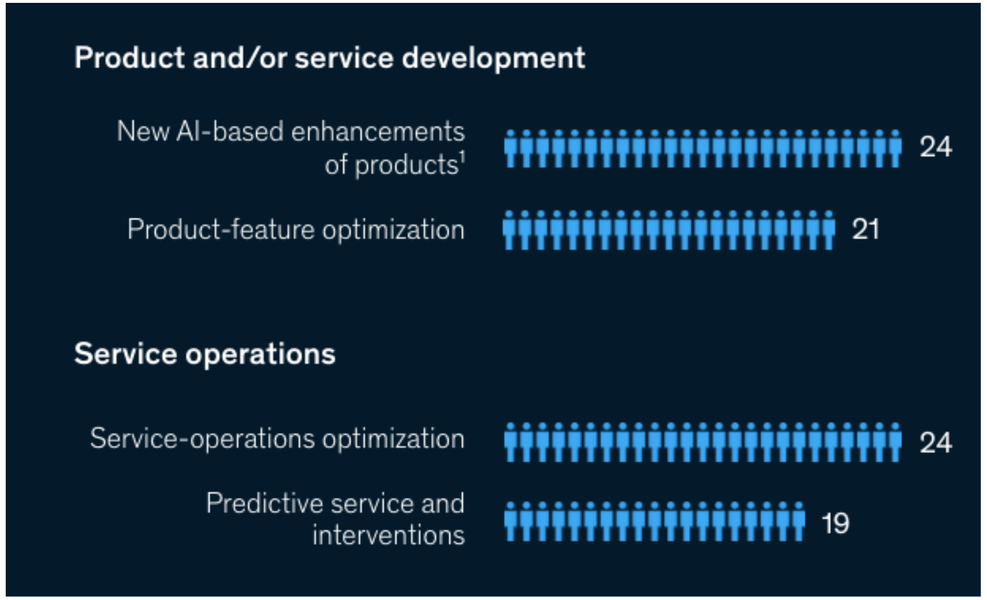
Extremely conservative investments
(90%)

Extremely aggressive investments
(10%)

Extremely conservative investments in AI

Business of AI

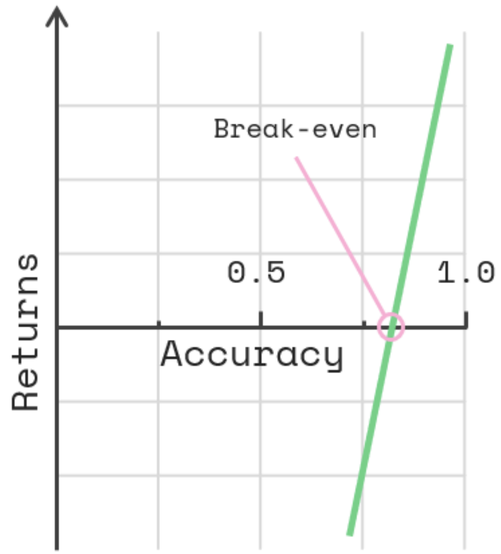
Conservative big data



https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020

$$returns = value - (1 - accuracy) * cost\ of\ a\ mistake$$

$$accuracy = 1 - \frac{value}{cost\ of\ a\ mistake}$$



https://towardsdatascience.com/return-on-investment-for-machine-learning-1a0c431509e

Business of AI

Conservative big data

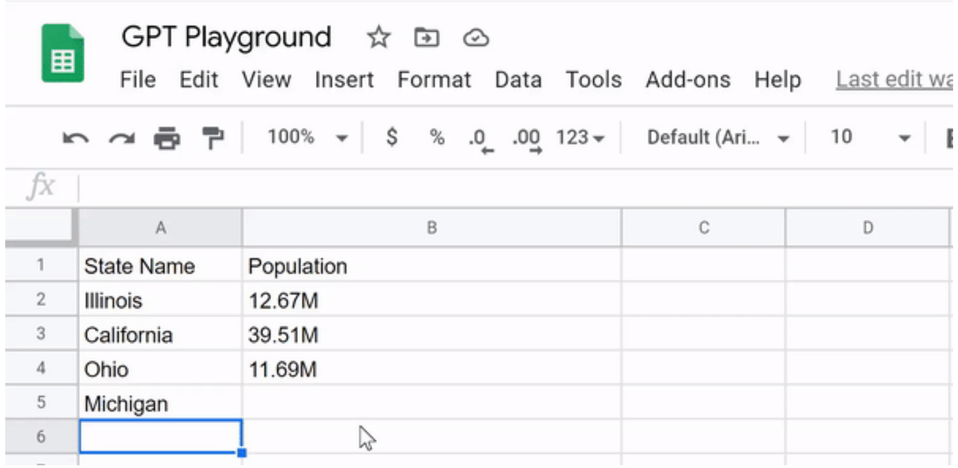
Category	Conservative big data	Aggressive zero-data
The “world”	Bits and bytes	
The process	Routine, often intuitive tasks (System 1)	
Capitalization / moats	Massive data	
Unit economics	Linear / logarithmic improvements	

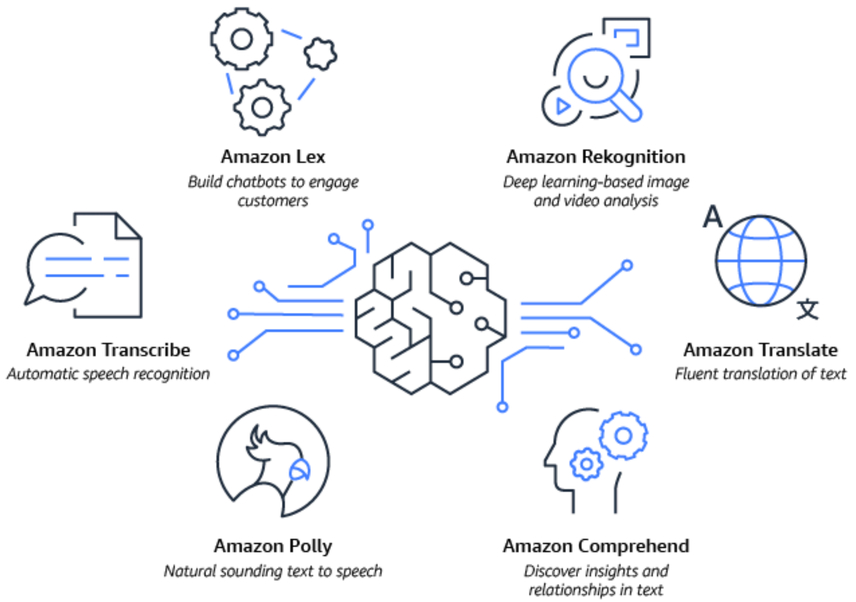
Technology of AI

Conservative big data

Dataset	**Quantity (tokens)**	**Weight in training mix**	**Epochs elapsed when training for 300B tokens**
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

https://in.springboard.com/blog/openai-gpt-3/





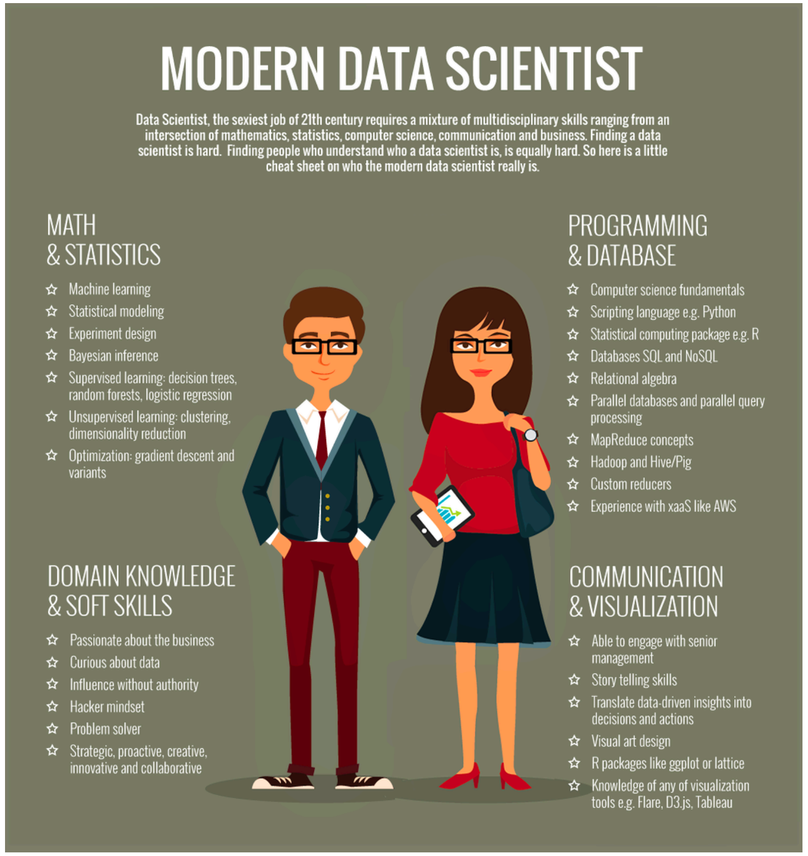
Technology of AI

Conservative big data

Category	Conservative big data	Aggressive zero-data
Datasets size	Millions, billions samples	
Targets and metrics	Single optimization criteria	
Mathematical problem	Pattern recognition	
Algorithms	Popular, commoditized, ready-to-use	

People of AI

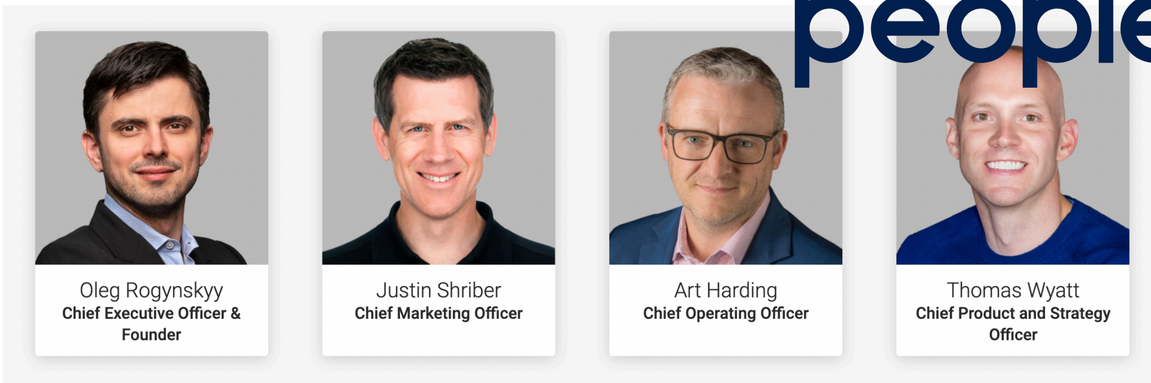
Conservative big data



https://www.kdnuggets.com/2020/01/wanna-be-data-scientist.html

Max Lytvyn | Alex Shevchenko | Dmytro Lider

https://www.grammarly.com/about



https://people.ai/about-us/

People of AI

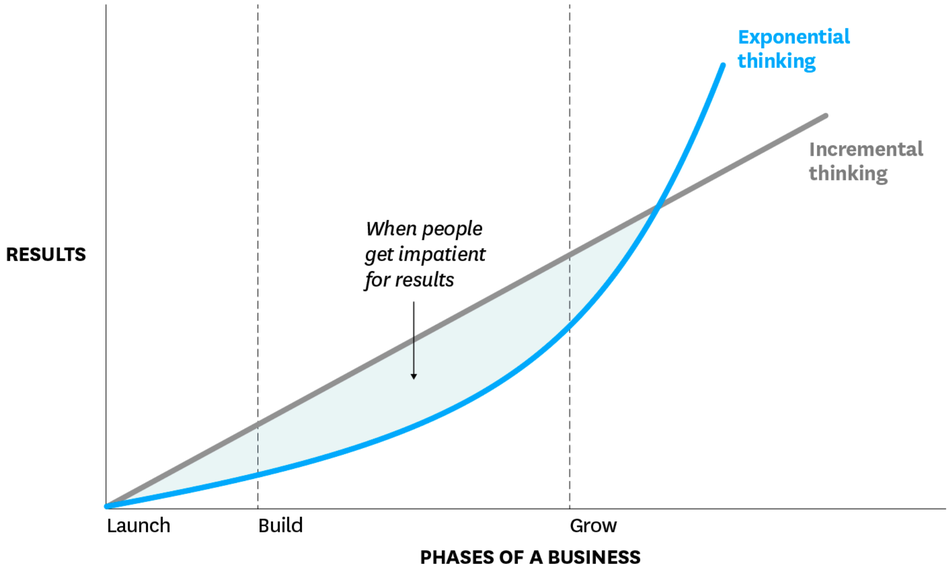
Conservative big data

Category	Conservative big data	Aggressive zero-data
Stakeholders	Business, product, marketing	
Developers	Engineers-integrators	
Customers	“Regular users”	

Extremely aggressive investments in AI

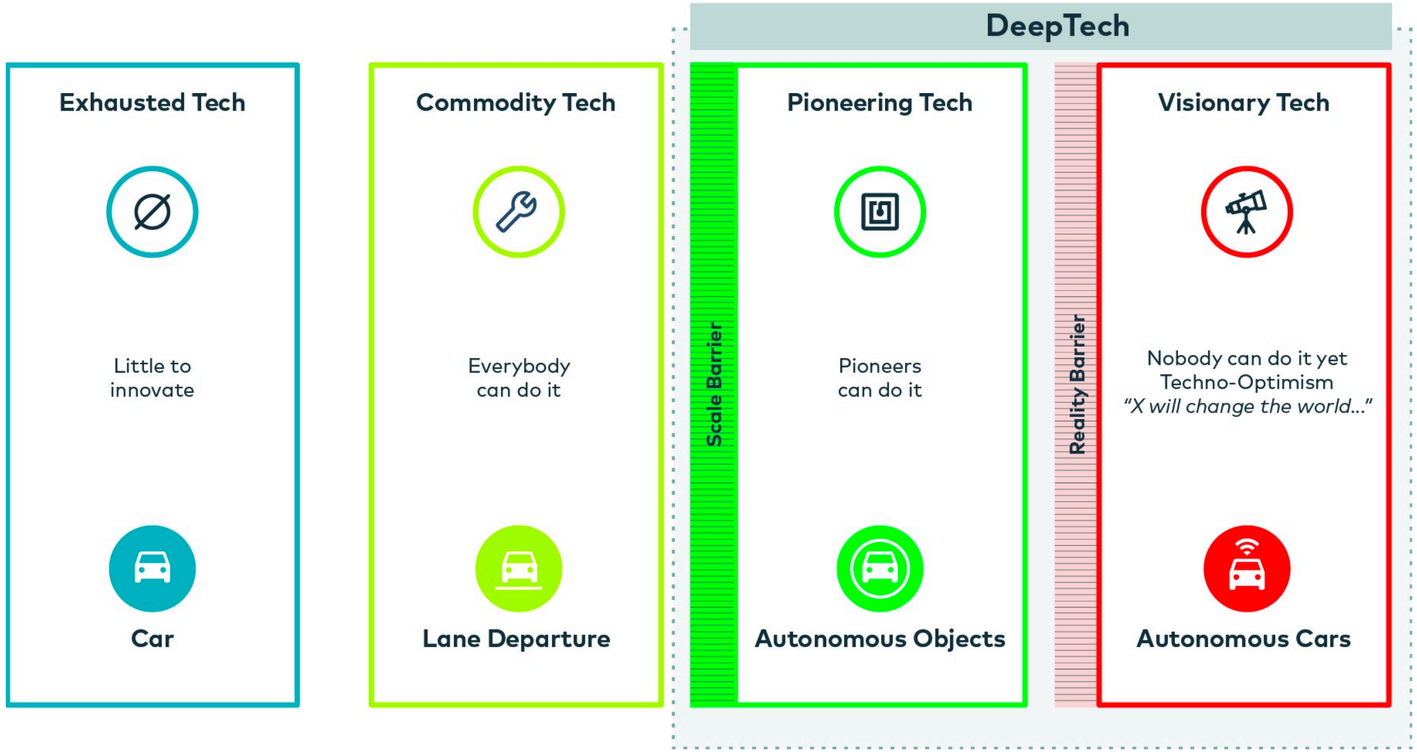
Incremental vs. Exponential Thinking When Growing a Business

Incremental thinking delivers immediate and steady results, while exponential thinking generates results that accelerate over time. The wrong expectations can lead teams to quit the exponential path too soon.



SOURCE MARK BONCHEK

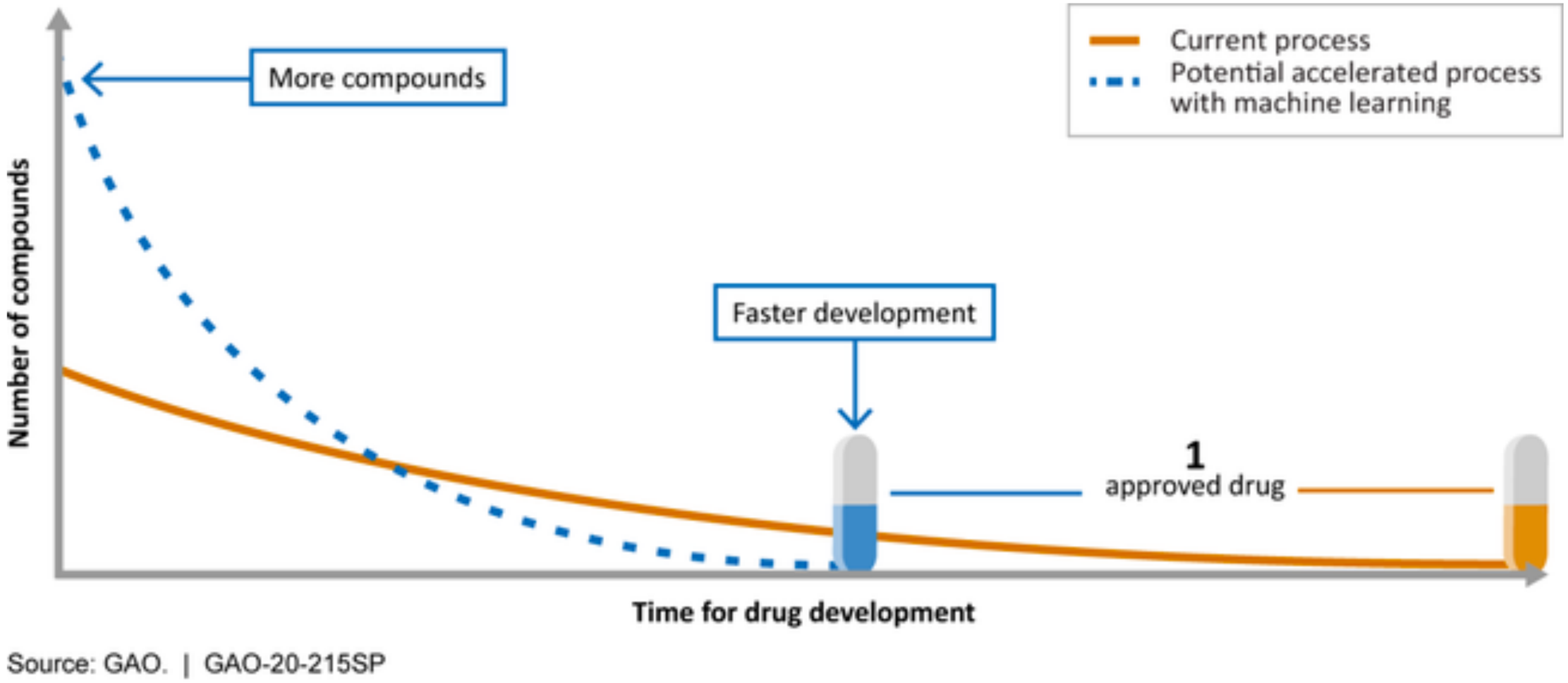
https://hbr.org/2016/07/how-to-create-an-exponential-mindset



https://medium.com/bcg-digital-ventures/the-right-time-for-deep-tech-dcb317fc3636

Business of AI

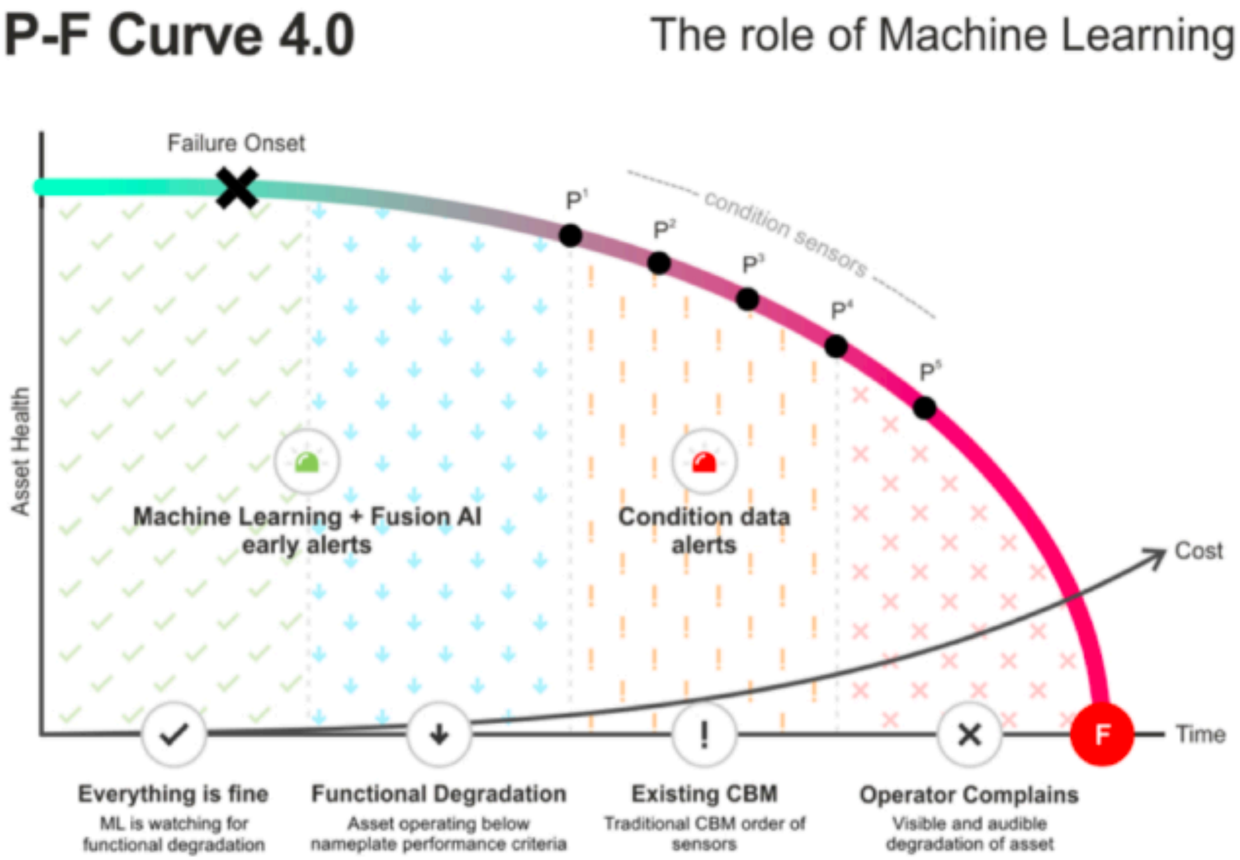
Aggressive zero-data



https://www.gao.gov/products/gao-20-215sp

Business of AI

Aggressive zero-data



https://www.quartic.ai/predictive-maintenance/

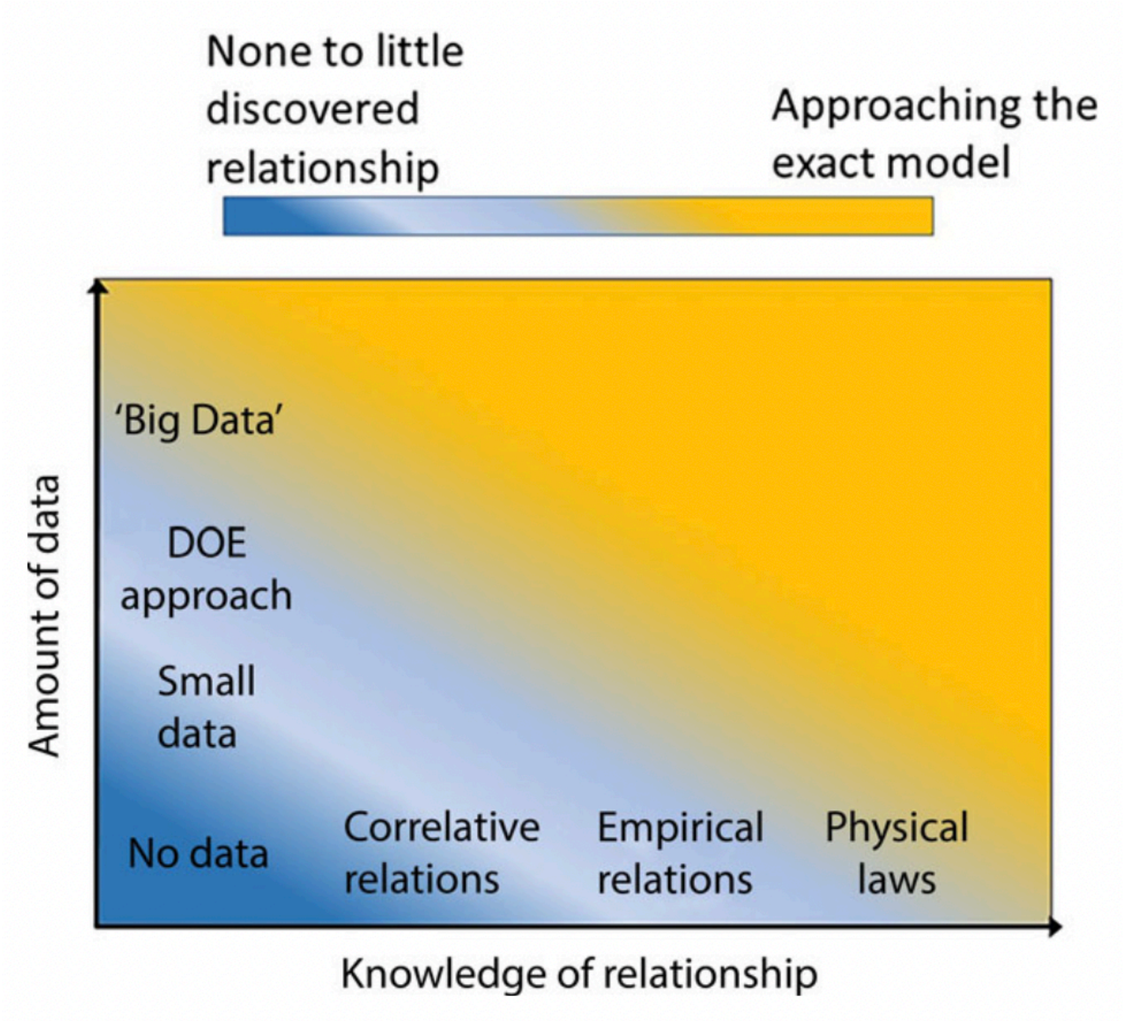
Business of AI

Aggressive zero-data

Category	Conservative big data	Aggressive zero-data
The “world”	Bits and bytes	Atoms and molecules
The process	Routine, often intuitive tasks (System 1)	Creative and scientific tasks (System 2)
Capitalization / moats	Massive data	Scientific insights
Unit economics	Linear / logarithmic improvements	Exponential improvements

Technology of AI

Aggressive zero-data



Embedding domain knowledge for machine learning of complex material systems

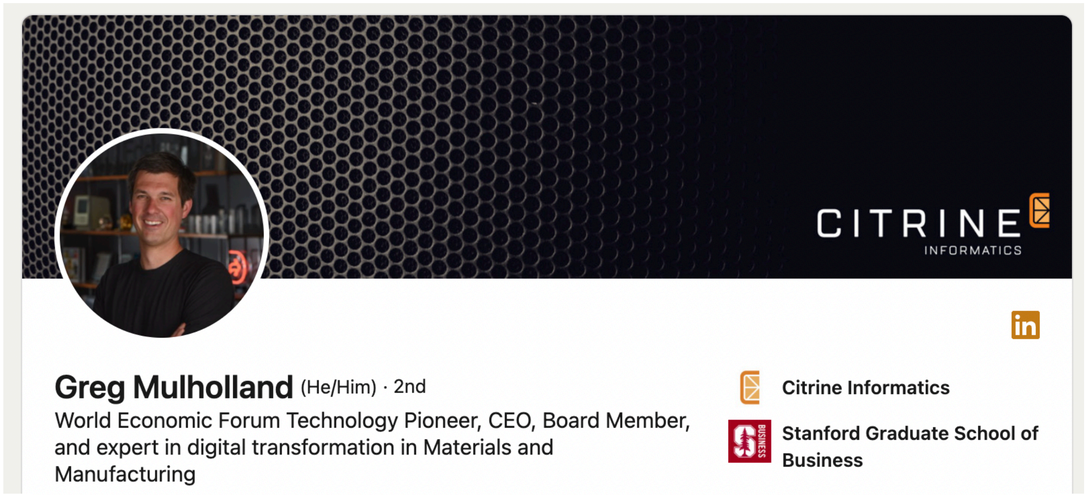
Technology of AI

Aggressive zero-data

Category	Conservative big data	Aggressive zero-data
Datasets size	Millions, billions samples	Hundreds, thousands samples
Targets and metrics	Single optimization criteria	Multi-criteria optimization
Mathematical problem	Pattern recognition	Extremum identification
Algorithms	Popular, commoditized, embedded in products (AWS, DataRobot, etc)	Physics-aware, extremely customized

People of AI

Aggressive zero-data



Greg Mulholland (He/Him) · 2nd

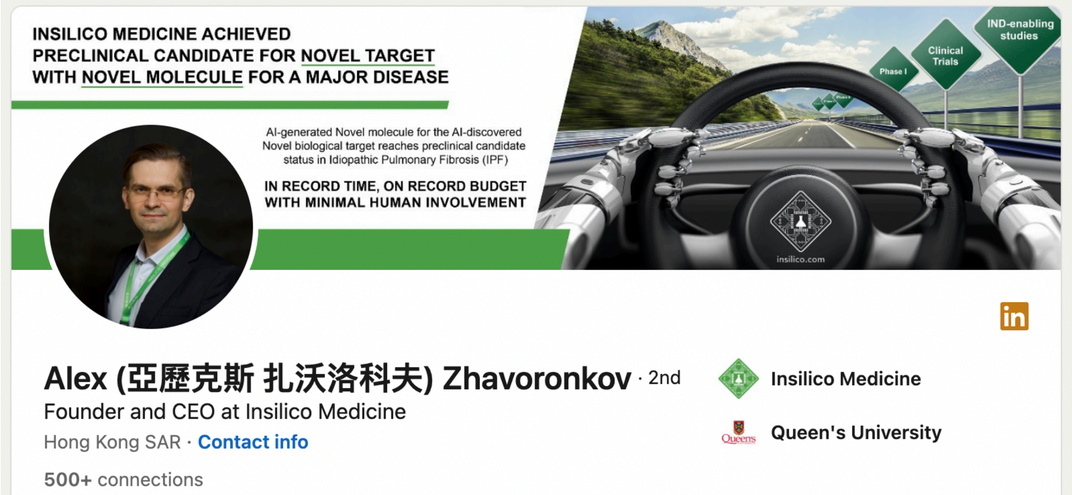
World Economic Forum Technology Pioneer, CEO, Board Member, and expert in digital transformation in Materials and Manufacturing

Citrine Informatics

Stanford Graduate School of Business

University of Cambridge
MPhil, Physics (Materials Science)
2007 – 2008
Activities and Societies: Trinity Hall, Trinity Hall Boat Club

– Thesis: Large-Scale Metal Oxide Nanowire Arrays for Hybrid Solar Cell Applications
– Research-based Master of Philosophy
– Supervised in Device Materials Group by Dr. Lukas Schmidt-Mende and Dr. Judith Driscoll
– Created anodized aluminum oxide templates for nanowire growth
– Fabricating Tin Dioxide and Titanium Dioxide nanorods using electrodeposition into template
– Nanostructures for use in hybrid organicinorganic solar cell
– Attended courses in TEM, SEM, XRD, EDX, and electrochemical methods
– Regularly perform X-Ray Diffraction, Raman Spectroscopy, and Scanning Electron Microscopy for characterization of thin films and nanostructures
– Finalist, Gates--Cambridge Scholarship



Alex (亞歷克斯 扎沃洛科夫) Zhavoronkov · 2nd

Founder and CEO at Insilico Medicine

Hong Kong SAR · Contact info

500+ connections

Insilico Medicine

Queen's University

Lomonosov Moscow State University (MSU)
PhD, Physics

- Interaction of ions and chiral compounds in biological systems.
- Racemization of amino acids in proteins in the course of aging.

The Johns Hopkins University
Master of Science, Biotechnology

- Baltimore, Maryland, USA

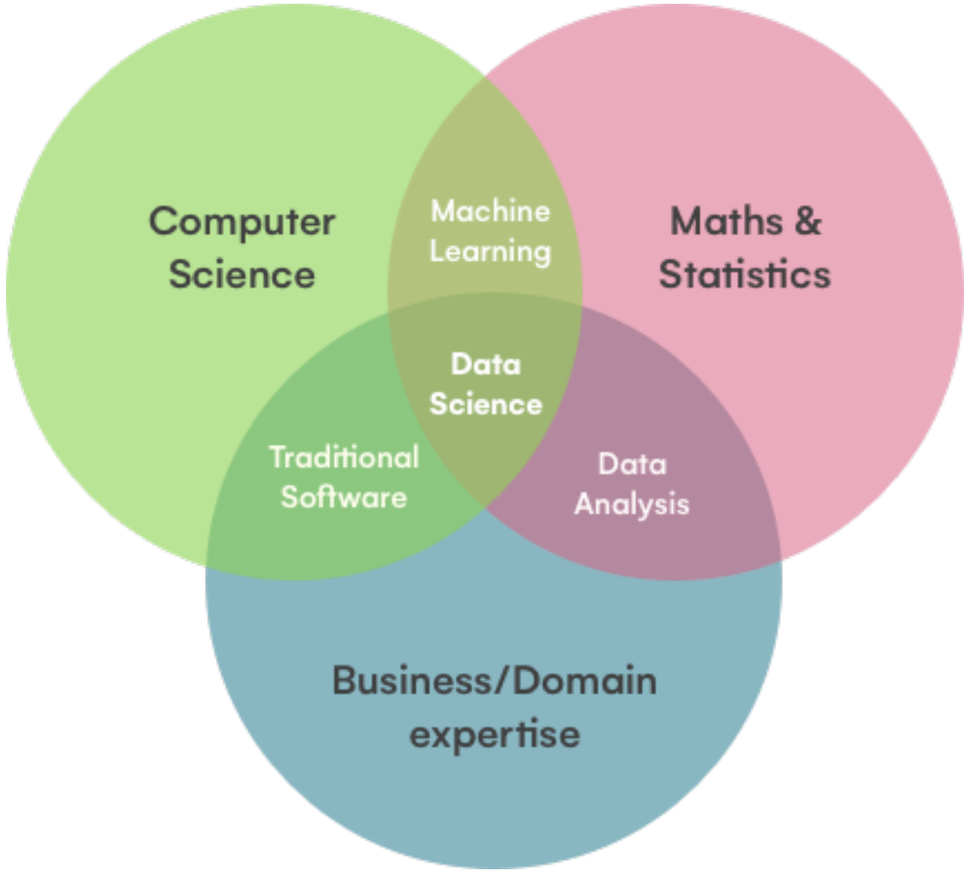
People of AI

Aggressive zero-data

Category	Conservative big data	Aggressive zero-data
Stakeholders	Business, marketing	Subject matter experts, often with scientific background
Developers	Engineers-integrators	
Customers	“Regular users”	

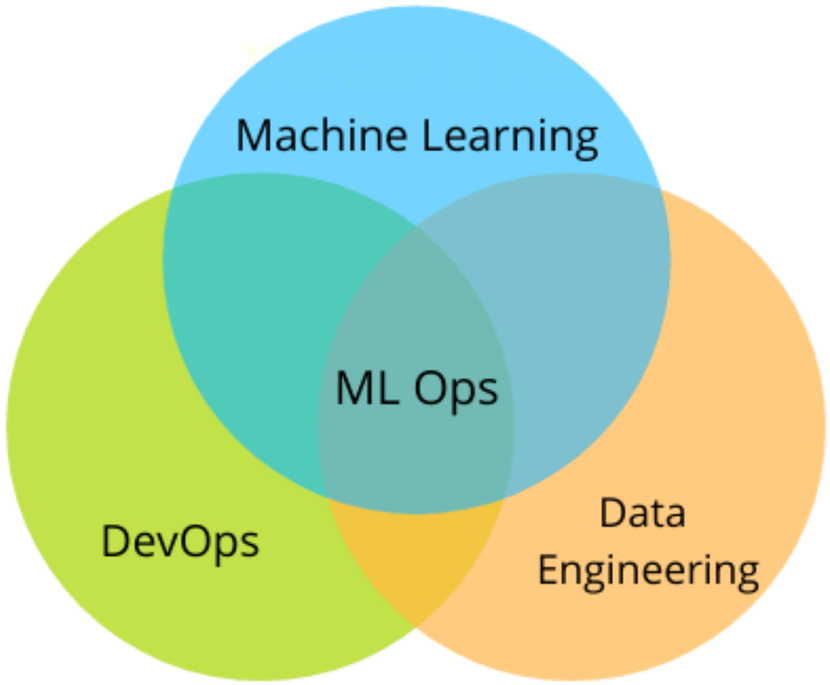
Why anti-data-science?

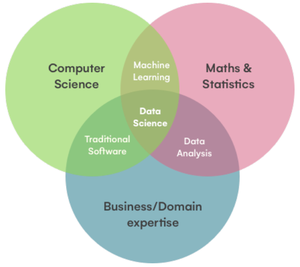
What is data science?

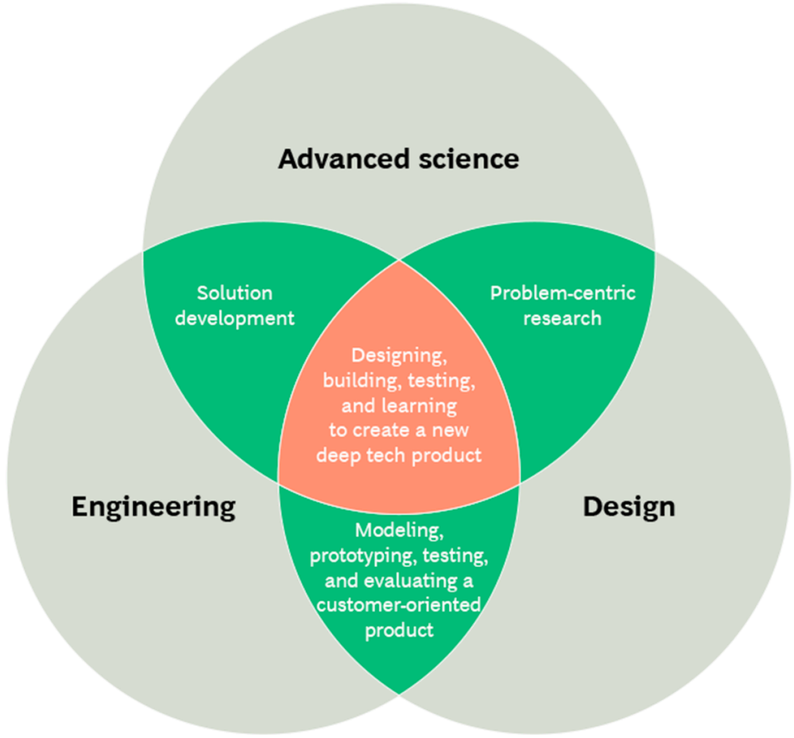


https://www.ironhack.com/es/data-analytics/data-science-vs-data-analytics

What is data science?







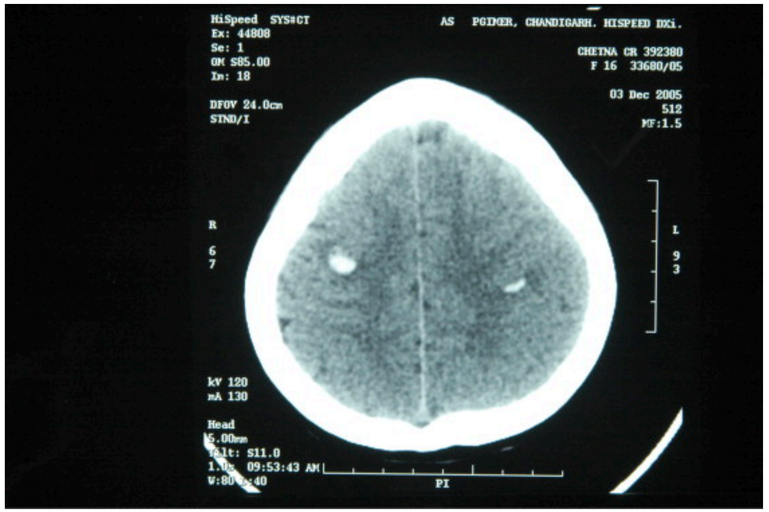
https://neptune.ai/blog/end-to-end-mlops-platforms

https://www.bcg.com/publications/2021/deep-tech-innovation

Anti-data-science

Applying “big data” tricks in biological world

⛔ Oncology, IBM Watson



- Trying to solve “routine” doctor tasks
- 53K samples, regular vision algorithms, focus on accuracy
- Organized by business people, done by engineers, doctors don’t use it

✅ AlphaFold, DeepMind

- Hard scientific task, novel finding
- 350K samples, satisfaction of chemical properties
- Done by scientists, with scientists, for scientists

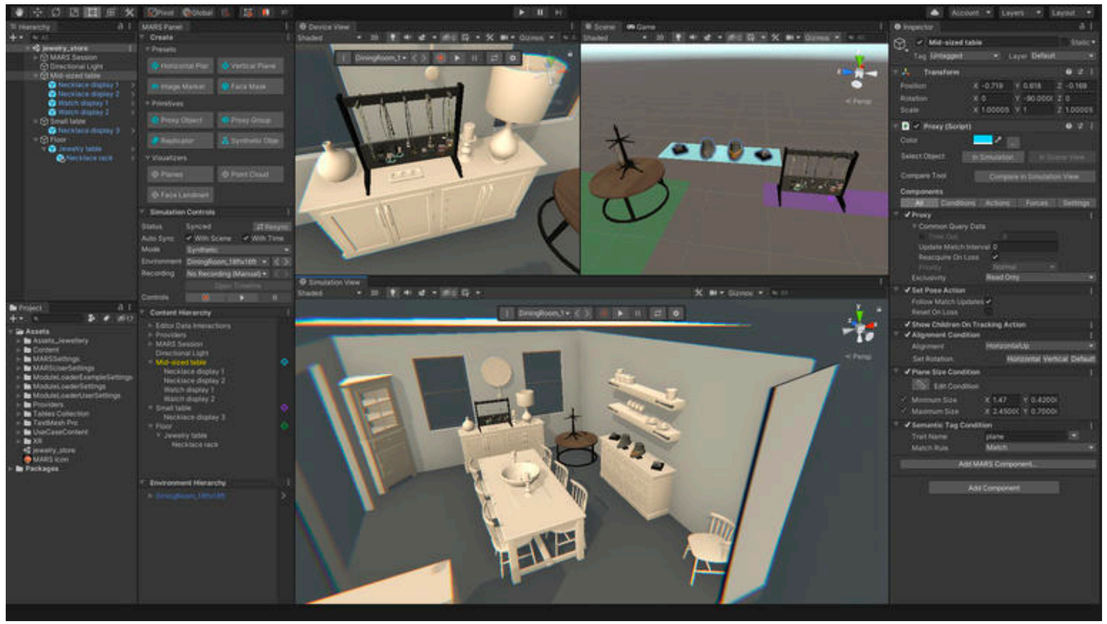
Anti-data-science

Applying “zero data” tricks in digital world

⛔ GANs for data augmentation

- Naive idea to use same old data to get “new data”
- Generating millions images with zero statistical value
- Organized by people that “just need more data”

✅ Unity, 3D modeling



- Requires expert knowledge in graphics and 3D
- You can generated limited high-quality samples
- Organized by domain experts with specific reqs

- Millions of people doing **digital routine tasks**
- Business driven, **linear** unit economics
- **Data monopoly** as a competitive edge
- Algorithms already **commoditized**

- Niche, scientific, often **physical challenges**
- Research driven, **exponential** unit economics
- **Expertise monopoly** as a competitive edge
- Algorithms **tailored** to the problem

let us help you to get on the right barbell side

find me in Linkedin or ping us at neurons-lab.com