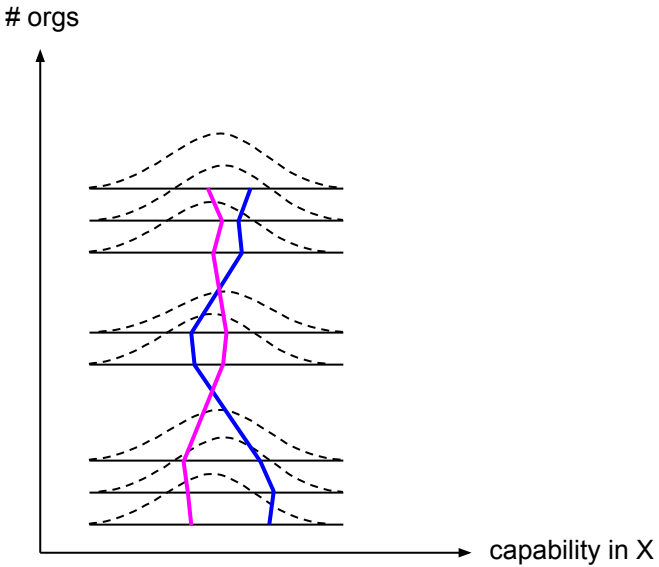
Industrialised data - the key to AI success

Lars Albertsson, Founder, Scling
2024-04-25

The great capability divide

Myth:

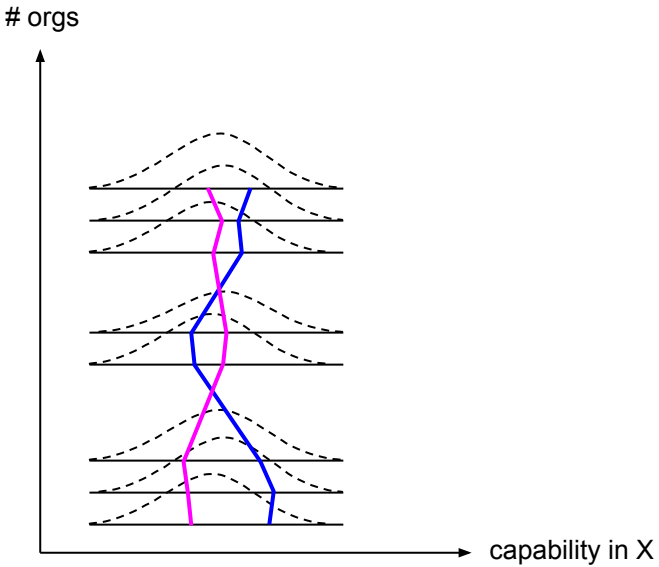
- We are all doing quite ok
- 2-10x leader-to-rear span



The great capability divide

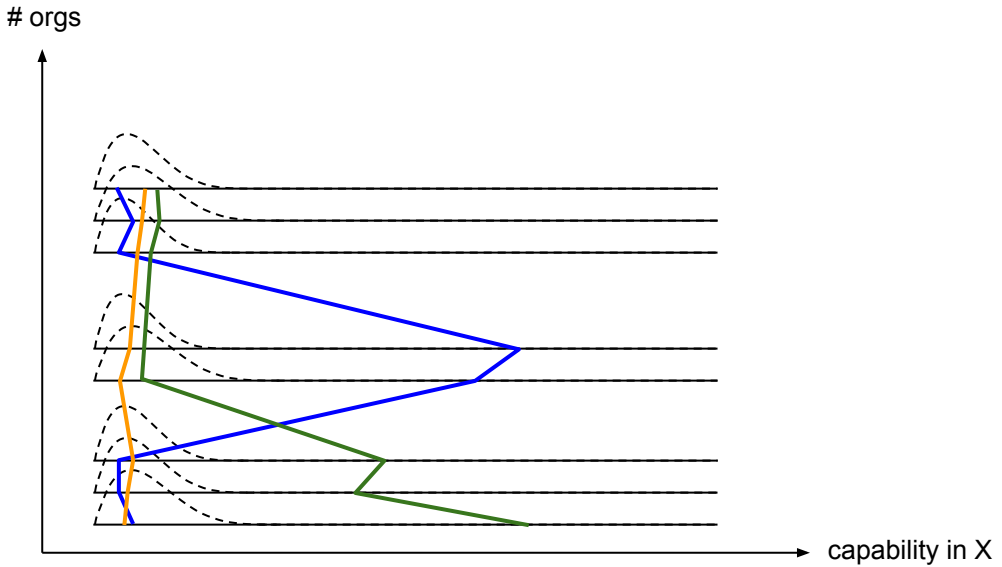
Myth:

- We are all doing quite ok
- 2-10x leader-to-rear span



Reality:

- Few leaders in each area
- 100-10000x leader-to-rear span

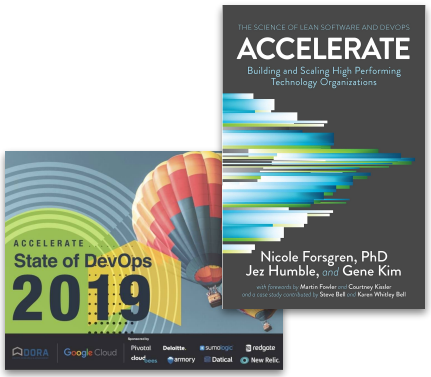


Capability KPIs

DORA research / State of DevOps report:

- Deployment frequency
- Lead time for changes
- Change failure rate
- Time to restore service

Small elite
~1000x span



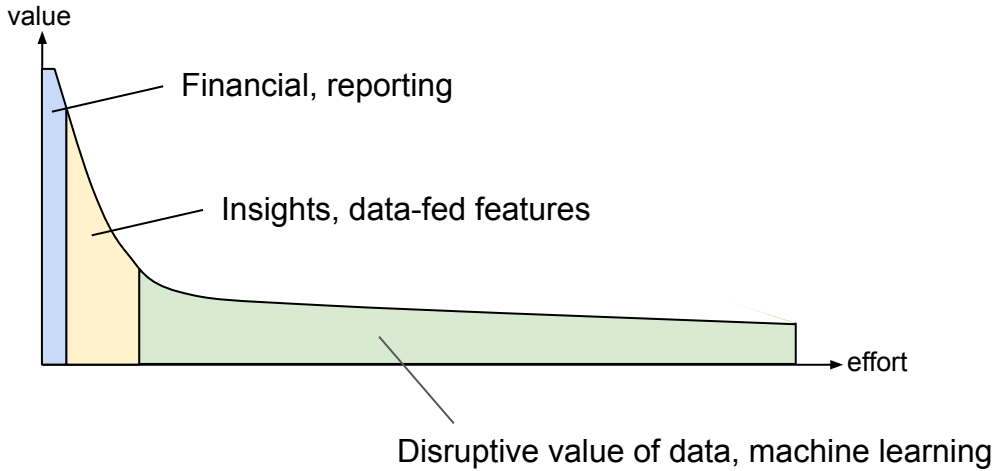
Observed differences in data organisations:

- Lead time from idea to production
- Time to mend / change pipeline
- Number of pipelines / developer
- Number of datasets / day / developer

Small elite
100 - 10000x span (or more)

Efficiency gap, data cost & value

- Data processing produces *datasets*
  - Each dataset has business value
- Proxy value/cost metric: *datasets / day*
  - S-M traditional: < 10
  - Bank, telecom, media: 100-1000



2014: 6500 datasets / day
2016: 20000 datasets / day
2018: 100000+ datasets / day,
25% of staff use BigQuery
2021: 500B events collected / day

Spotify

2016: 1600 000 000 datasets / day

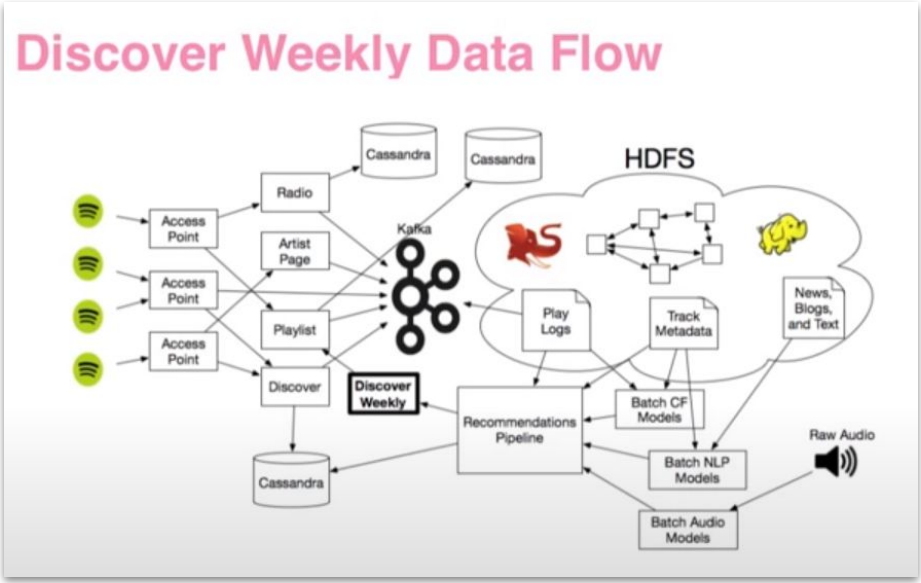
Google

Enabling innovation

"The actual work that went into Discover Weekly was very little, because we're reusing things we already had."

"Discover Weekly wasn't a great strategic plan and 100 engineers. It was 3 engineers that decided to build something."

"I would have killed it. All of a sudden, they shipped it. It's one of the most loved product features that we have."
- Daniel Ek, CEO



https://youtu.be/A259Yo8hBRs
https://youtu.be/ZcmJxli8WS8
https://musically.com/2018/08/08/daniel-ek-would-have-killed-discover-weekly-before-launch/

Data not acted upon - value left on table

The car has sensors for precipitation, temperature, and window state. I would have liked to receive a mobile app warning when all windows were down during a snowy night.

Data not acted upon - value left on table

The car has sensors for precipitation, temperature, and window state. I would have liked to receive a mobile app warning when all windows were down during a snowy night.

The car keeps informing me that there is a software upgrade, but no matter what I do, nothing gets upgraded.

I send a design suggestion to Volvo with pics of a hacked solution for a luggage problem. A simple hook could solve it. I do not receive "We have assigned your suggestion id X, and will get back if it is implemented." I will not send another, and feel less connected to the brand.

Our Volvo incorrectly activates the automatic safety brake again. This is a known problem since years back, which affects our car occasionally. I have no efficient channel to report time and circumstances for this occasion, allowing Volvo to get more data on the problem, and me to feel confident that the issue is worked on.

I make the effort to submit a detailed bug report regarding Volvo's known over-aggressive automatic safety brake problem, in order to provide more data for debugging. I receive no feedback from the process, and lose interest.

The Volvo phone app warns about many things: interrupted charging, car parked but not being charged, etc. But not about the roof and windows left open in the rain. Rear view mirror electronics now need a repair.

After a repair at Bilia of the rear view mirror, the car does not properly detect cars in front. Camera near mirror is suspected. Bilia cannot obtain camera measurements or car detection statistics from Volvo, so a cycle of trial and error repairs follows, taking me to the mechanic multiple times.

Neighbour jumps in their Volvo, talks to husband through window, and drives off. On arrival she discovers that she has no car key - husband's key was close enough to start car. An early warning that key is no longer in proximity would have been valuable.

Map updates fail to auto install. Support cannot obtain diagnose information. Manual installation attempts fail without error messages.

When seat heating was automatically activated, "Increase seat heating" audio command turns it off. I make additional commands to get to the desired state. Measurements could identify repeated audio or UX commands.

Wife unlocks Volvo. Driver seat switches to her position. I take the driver's seat, pull it back, and change to my profile. Car switches seat back to her position, and I have to pull it back again.

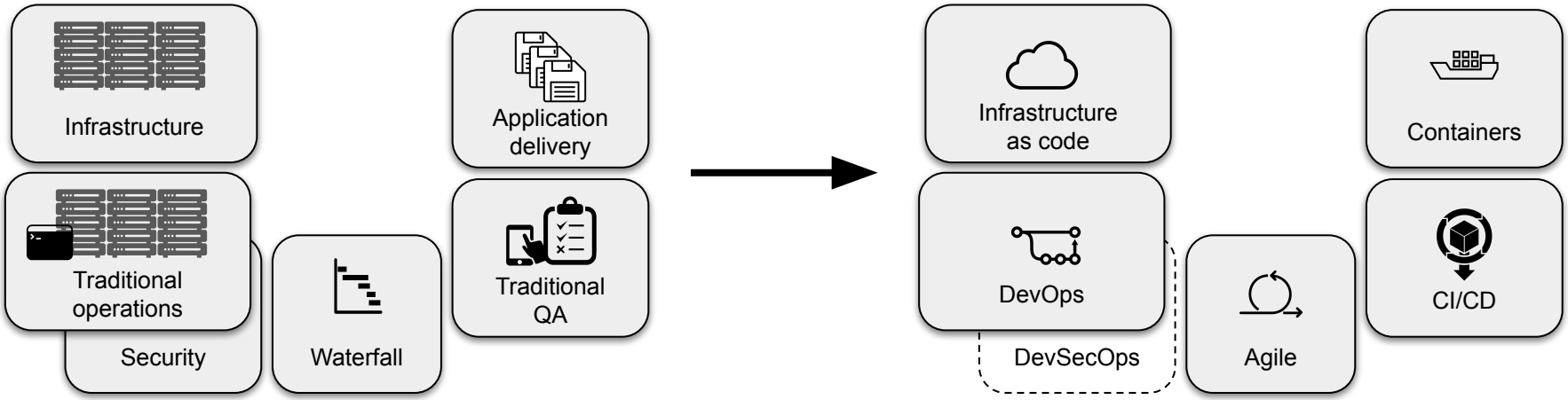
Changing to winter tyres at Bilia. When I arrived, the staff could see that mechanics were currently not fully booked, and offered me a discounted front wheel adjustment, which I accepted.

Data value taken from table.

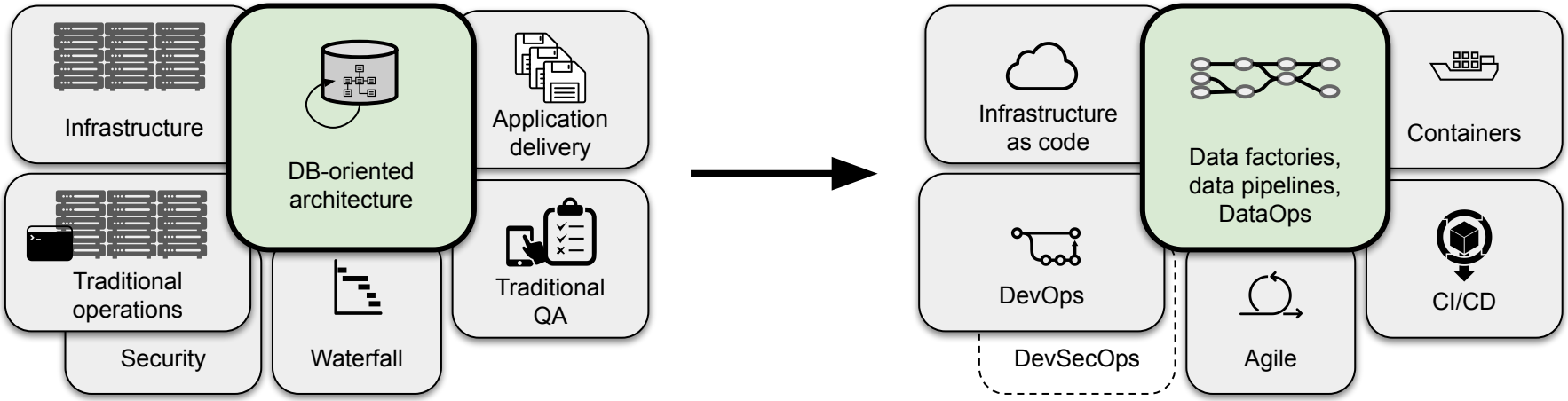
I change the audio balance in the Volvo with six screen clicks. Measurements could indicate that long click sequences at high speed indicates poor user interface.

Wife has driven our Volvo. Mirrors and seats automatically move to my position, except for the right mirror, which I need to correct again, as I have for years. Volvo could have measured and detected irregular changes of settings, and found this bug.

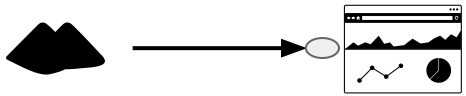
IT craft to factory



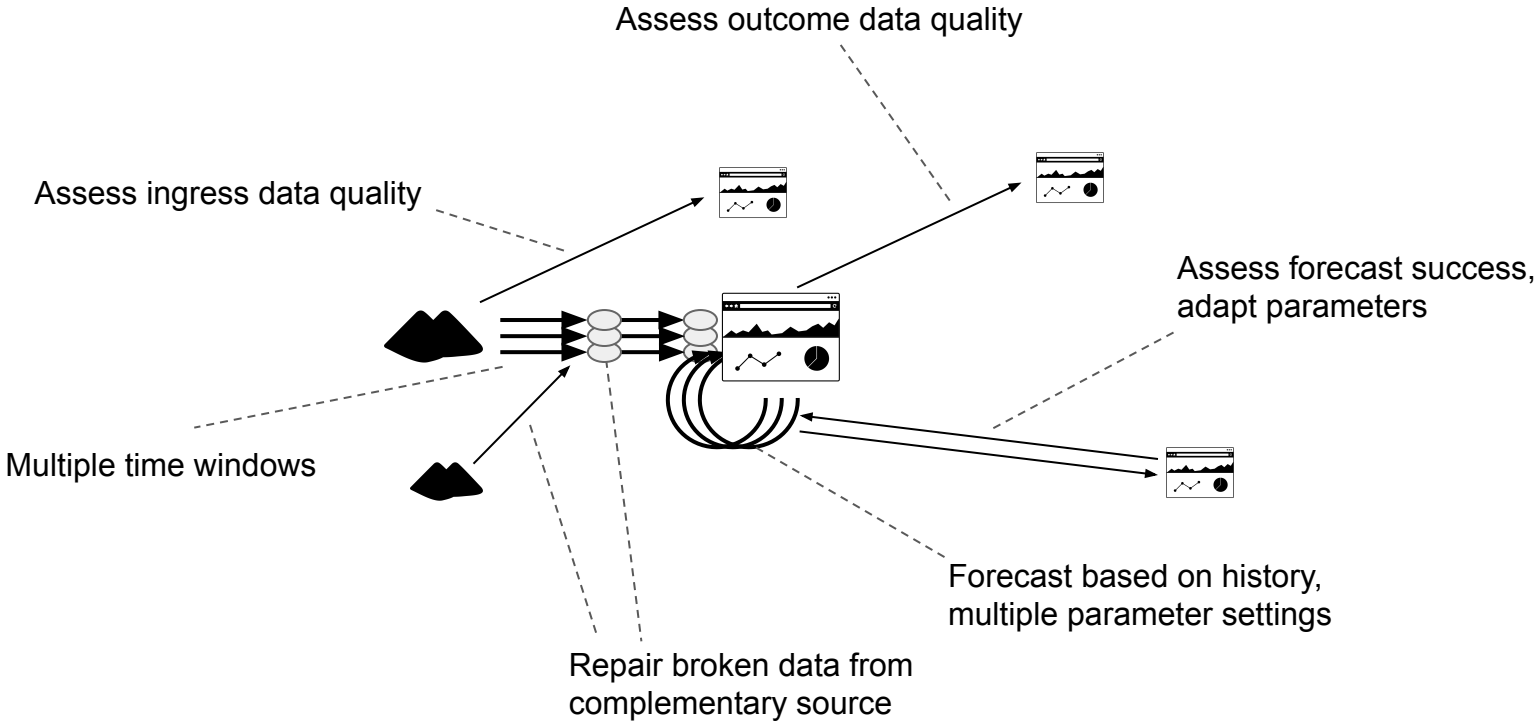
Data factories



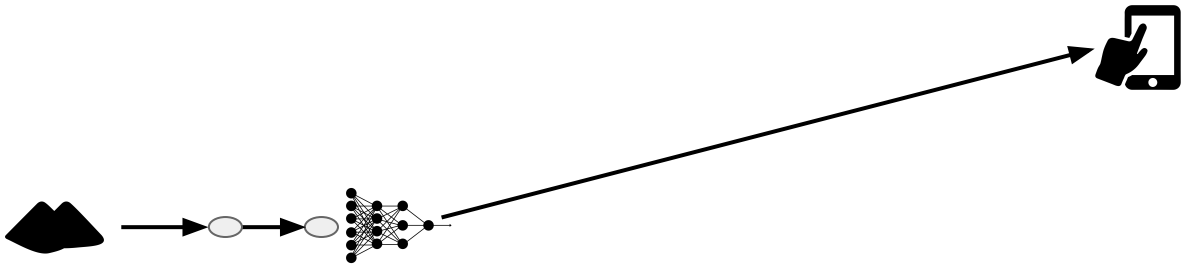
From craft to process



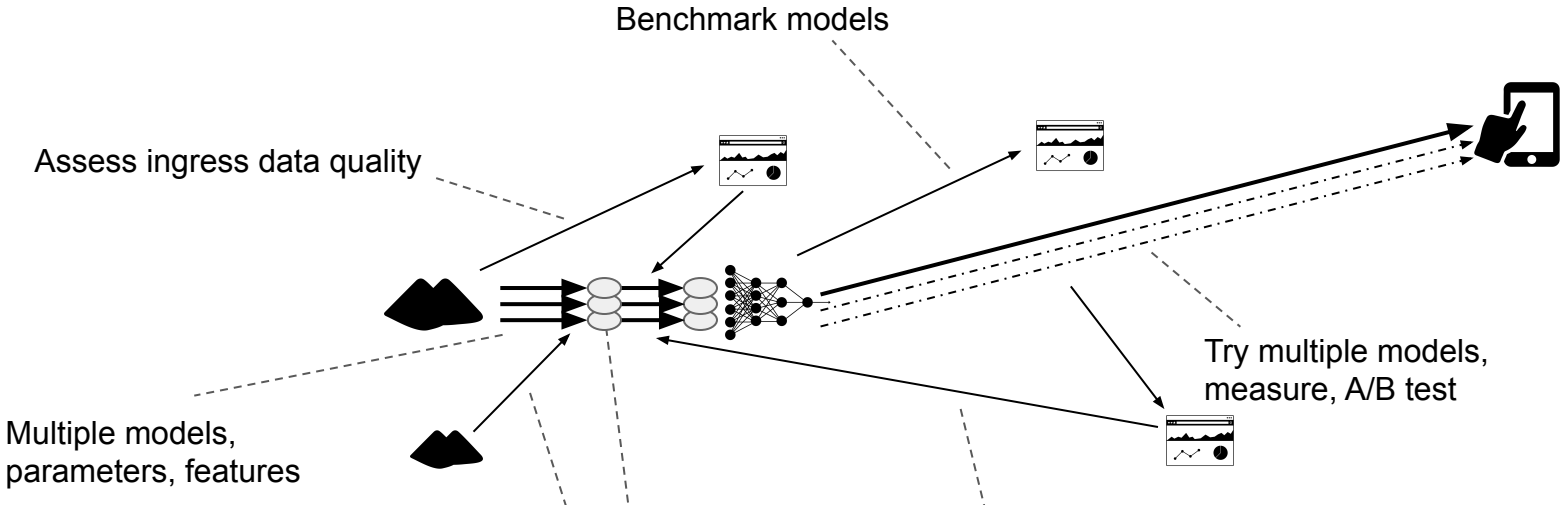
From craft to process

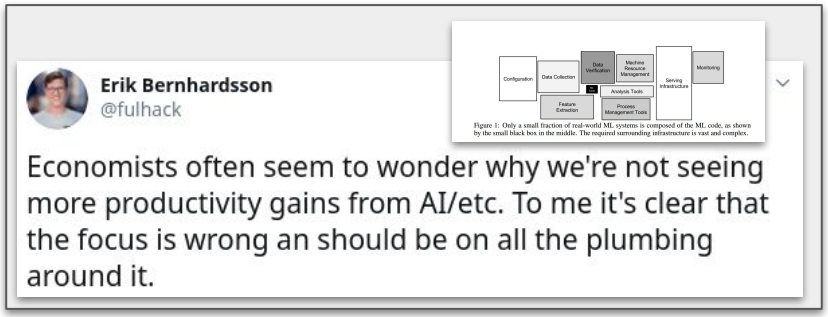


Naive machine learning

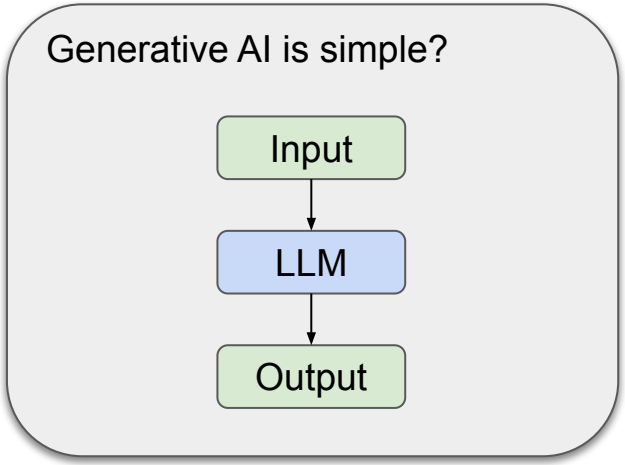


Sustainable production machine learning





Generative AI → more complex data engineering



Unless you want

- Relevance
- Correct facts
- Multiple modalities
- Prevent undesirable outcomes

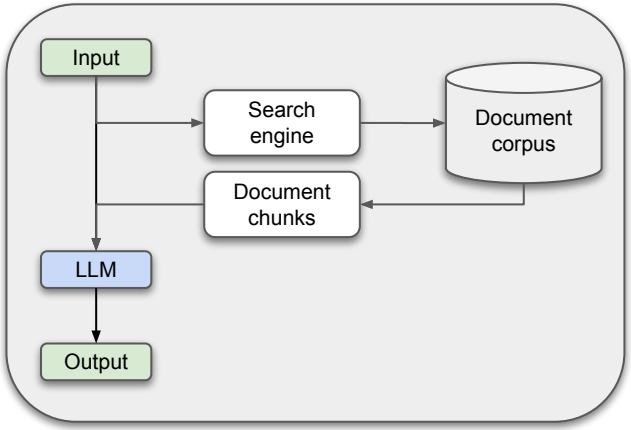
Retrieval-augmented generation (RAG)

- Training an adapted large language model is expensive and difficult

- RAG hack:
  - Find relevant document chunks
  - Concatenate with prompt (context)
  - Combined general and specific information!

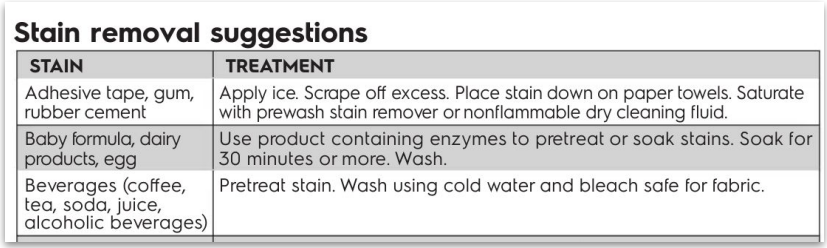
- How to search?
  - Keywords / embeddings / neural?
  - Rerank chunks?

- How to chunk documents?
  - Boundaries
  - Chunks need higher context
    - E.g. "These uses of X may cause fire: <snip> …"



RAG experiment: Household appliance interface

- Products that will be disrupted by generative AI?
  - How many different washing machine programs do you use?
  - What if you could talk to the machine? Show your stained shirt?



Stain removal suggestions

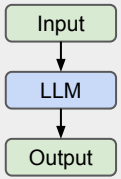
STAIN	TREATMENT
Adhesive tape, gum, rubber cement	Apply ice. Scrape off excess. Place stain down on paper towels. Saturate with prewash stain remover or nonflammable dry cleaning fluid.
Baby formula, dairy products, egg	Use product containing enzymes to pretreat or soak stains. Soak for 30 minutes or more. Wash.
Beverages (coffee, tea, soda, juice, alcoholic beverages)	Pretreat stain. Wash using cold water and bleach safe for fabric.

- Naive chunks:
  "Saturate with prewash stain remover or nonflammable dry cleaning fluid. Baby formula, dairy products, egg."

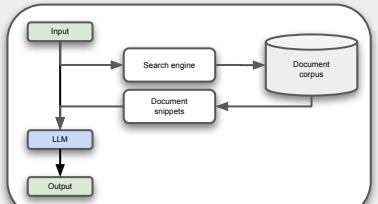
- *Domain-specific data engineering divides generative AI products from demos.*

Generative AI → more complex data engineering

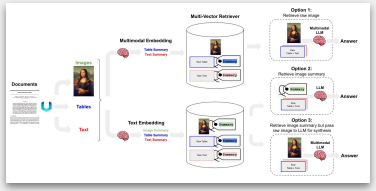
Generative AI is simple?



Add relevance? → Retrieval augmented generation (RAG)



Want multi-modal?



Langchain: Multi-modal RAG

Safety mechanisms to avoid undesirable use?

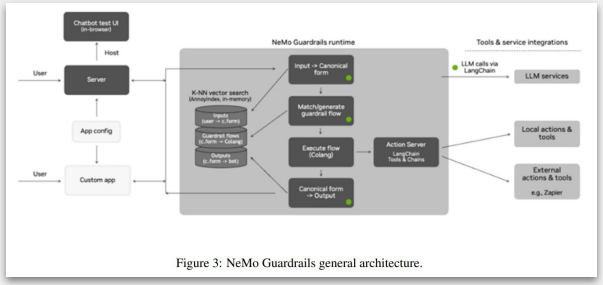
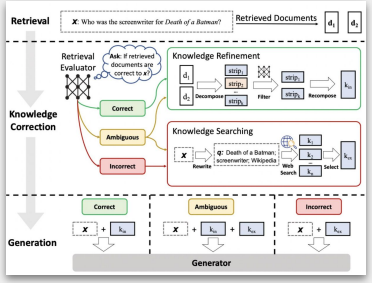


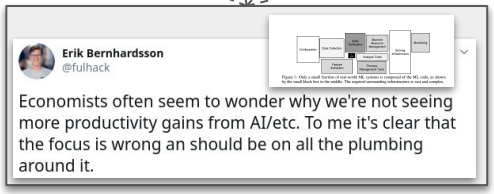
Figure 3: NeMo Guardrails general architecture.

Nvidia: Nemo Guardrails

Correct facts too?

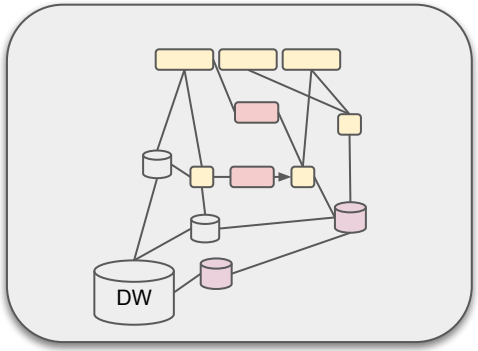


USTC + UCLA + Google: Corrective RAG

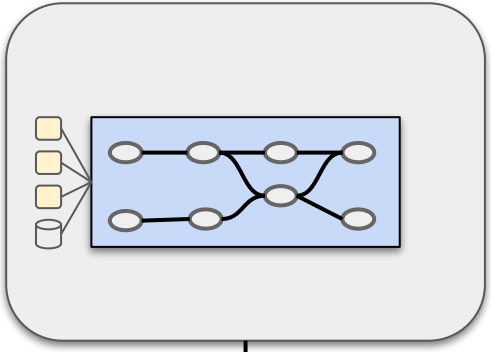


Data engineering in the future

"data factory engineering"



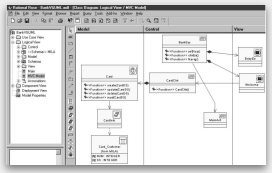
~10 year capability gap

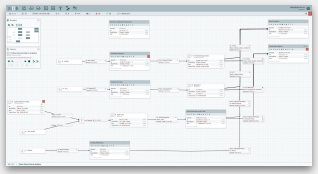


Enterprise big data failures

Data engineering education

"Modern data stack" -
traditional workflows, new technology

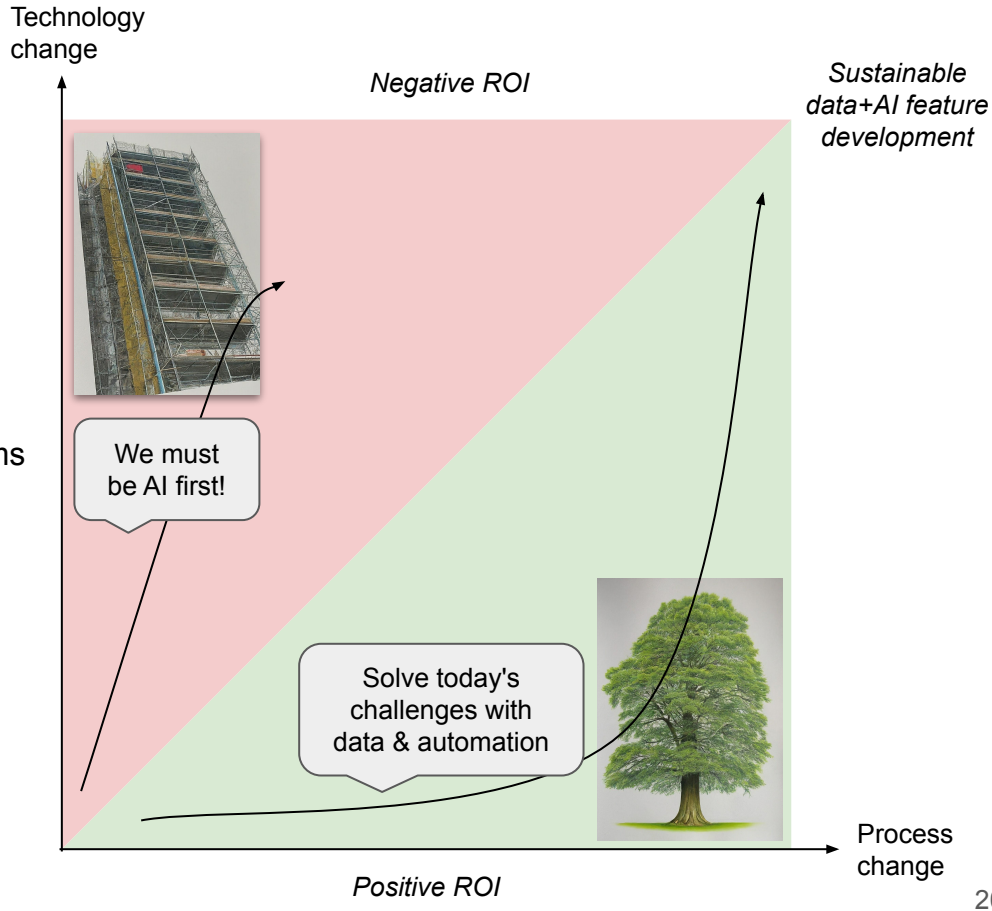




4GL / UML phase of data engineering

Don't build. Grow.

- Every data / AI project has either
  - failed
  - cost a fortune
- Every leading data company has
  - solved product challenges at hand
  - improved process / org / ways of working
  - had data / AI success through enabled teams
- Brief recipe for success
  - Align teams with value chains
  - Automation + data for current challenges
  - Centralised, immutable data in pipelines
  - It's a software engineering problem
    - QA, composability, DevOps, …
  - Feedback cycle speed < hour (day)



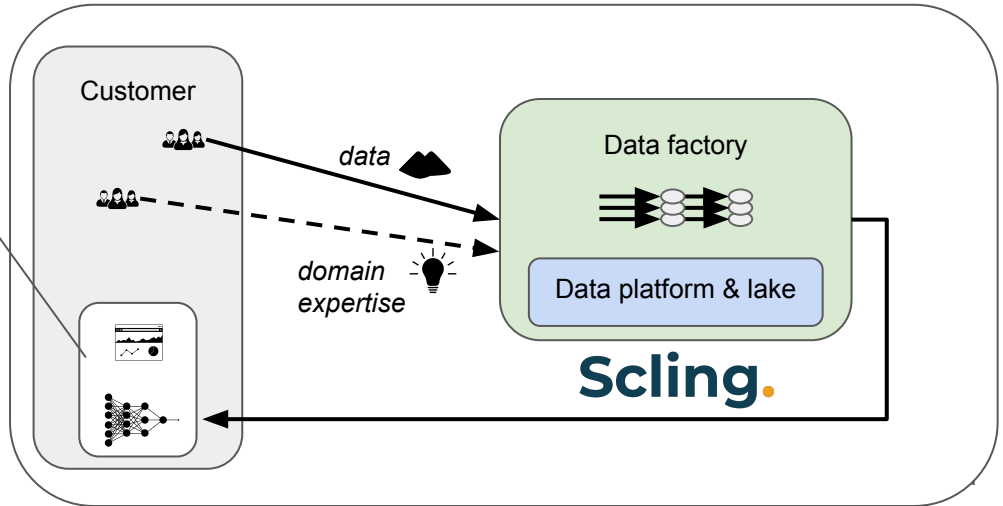
Data factory adoption

Observations:

- Industrial success requires veterans
- Not enough factory experience around
- Minimal flow of veterans to incumbents / consultants
- Artisanal tools (Data warehouse / low-code / …) inadequate for domain-specific AI
- Industrial tools not helpful without veterans

Belief:
We need new ways of collaboration.

- Other disciplines?
  - Shared innovation
  - E.g. Foxconn, Autoliv
- Tech leaders
  - Grow, don't build
  - Product-driven



Data innovation as leaders do it

Learning by doing, in collaboration

What we learnt

Success under good circumstances

- Match Spotify's numbers per developer
- 10-1000x client's own capability
- Data + AI innovation on par with leaders

Challenge: Data capability → innovation

- Domain-specific competence
- "Innovation building blocks" needed
  - Agile, pull vs push, iterative
  - Digital thinking
  - Aligned, cross-functional teams
  - Product focus
  - Value chain alignment

Greater challenge: "How hard can it be?"

- Unawareness of data divide

What we learnt, what we need

Success under good circumstances

- Match Spotify's numbers per developer
- 10-1000x client's own capability
- Data + AI innovation on par with leaders

Challenge: Data capability → innovation

- Domain-specific competence
- "Innovation building blocks" needed
  - Agile, pull vs push, iterative
  - Digital thinking
  - Aligned, cross-functional teams
  - Product focus
  - Value chain alignment

Greater challenge: "How hard can it be?"

- Unawareness of data divide

Humble, but competent clients

- Data with value potential
- Domain experts
- Innovation building blocks
- Humble regarding data divide

Partners that bridge innovation gap

- Digitalisation in some vertical
- Interested in new business models

Other ways to slice the challenge?