Sweden AI Factory

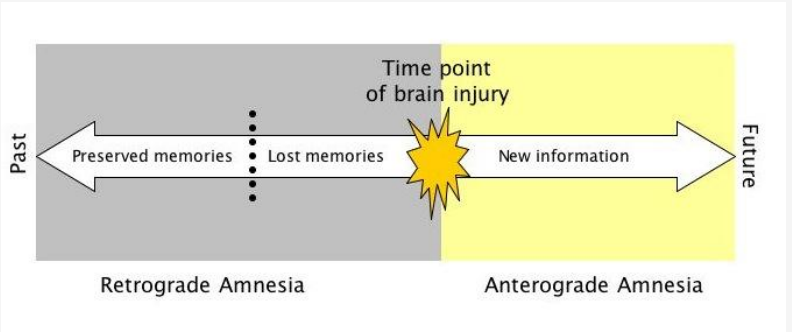
AI and processes

Erik Ylipää (erik.ylipaa@ri.se)

AI for Life Science

The Memento problem

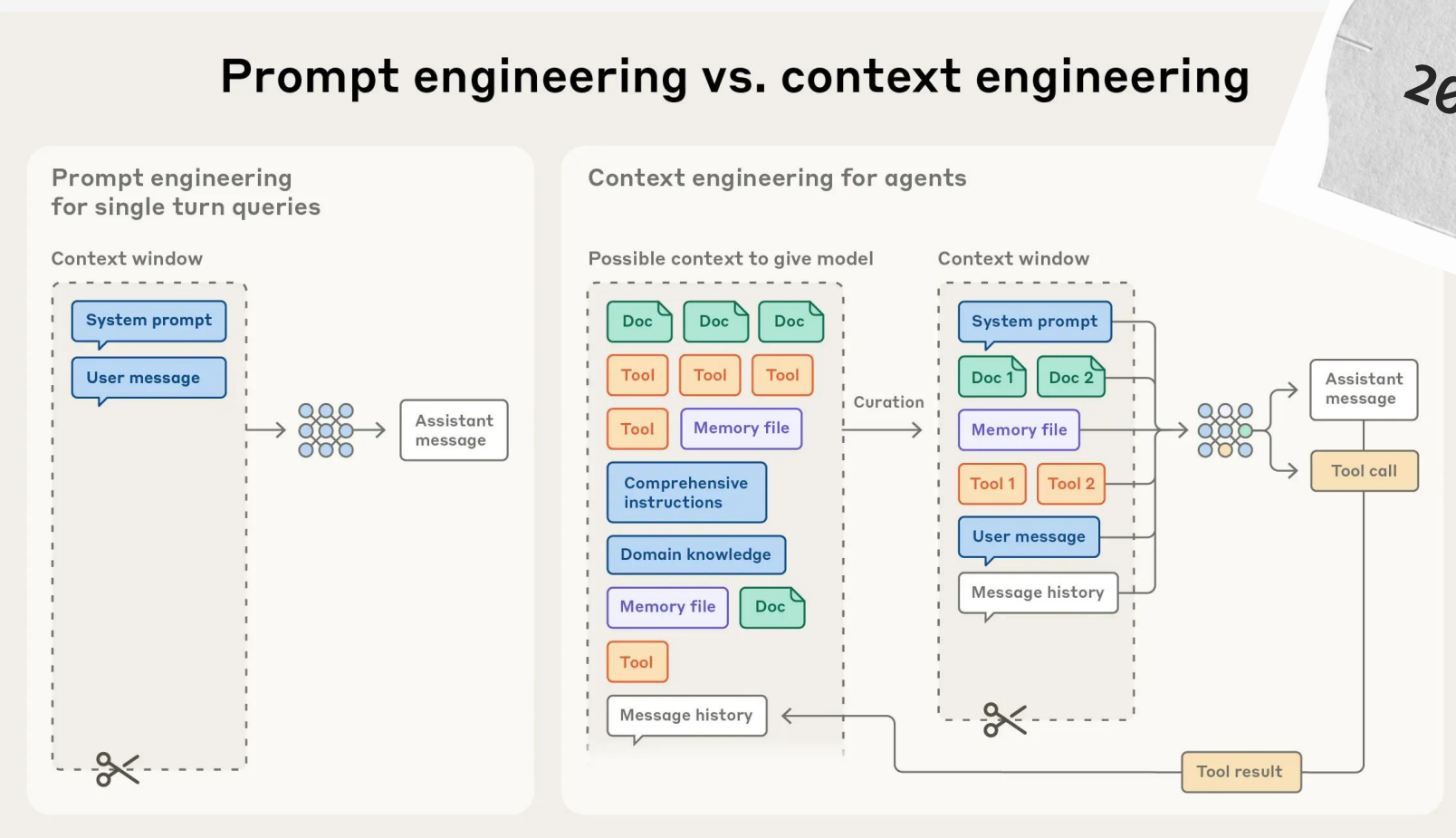
- Christopher Nolan film from 2000 about Leonard
  - Leonard suffers from *anterograde amnesia*
  - He can't form any new memories after the incident which caused his condition
  - The film centers around how he tries to understand his world based on a **context** he (and people around him) **engineers**



http://mercercognitivepsychology.pbworks.com/w/page/61206437/Anterograde%20Amnesia


Prompt engineering vs. context engineering
Prompt engineering for single turn queries
Context window
System prompt
User message
Assistant message
Context engineering for agents
Possible context to give model
Doc
Tool
Memory file
Comprehensive instructions
Domain knowledge
Curation
Context window
System prompt
Doc 1
Doc 2
Memory file
Tool 1
Tool 2
User message
Message history
Assistant message
Tool call
Tool result
Message history

” 95% of AI engineering is just Context engineering” – May 2026



https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents

Sweden AI Factory

In two years, current context engineering will be **obsolete**

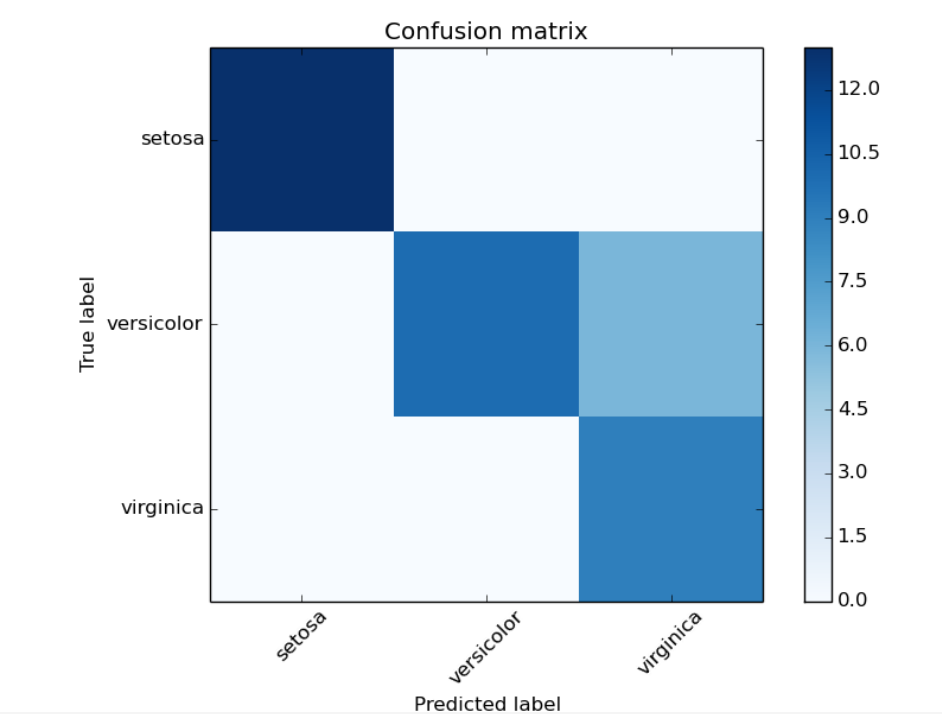
- Focus on building the foundation which will be needed regardless of AI capability
  - Processes
  - Evaluation
  - Experiment infrastructure
  - Data readiness

Sweden AI Factory

AI (based on Machine Learning) is statistical: can never be **perfect**

Data is never perfect

- Errors connecting inputs to outputs (e.g. filling in the wrong field in a form)
- Measurement noise in gathering data
- Incomplete, can't cover all possible situations
- Inherent uncertainty or unobservability in the problem



100% correct on test data does not guarantee perfect performance

Sweden AI Factory

Machine learning (AI): the high-interest rate credit card of technical debt

- Technical debt is an important concept in IT
- The system you build now will have to be able to survive and adapt
- Someone must maintain it and manage updates
- For data-driven AI, this is compounded by the coupling to data
    - What happens when data changes?

Sculley, David, et al. "Hidden technical debt in machine learning systems." *Advances in neural information processing systems* 28 (2015).

Sweden AI Factory

Traditional system vs. AI system

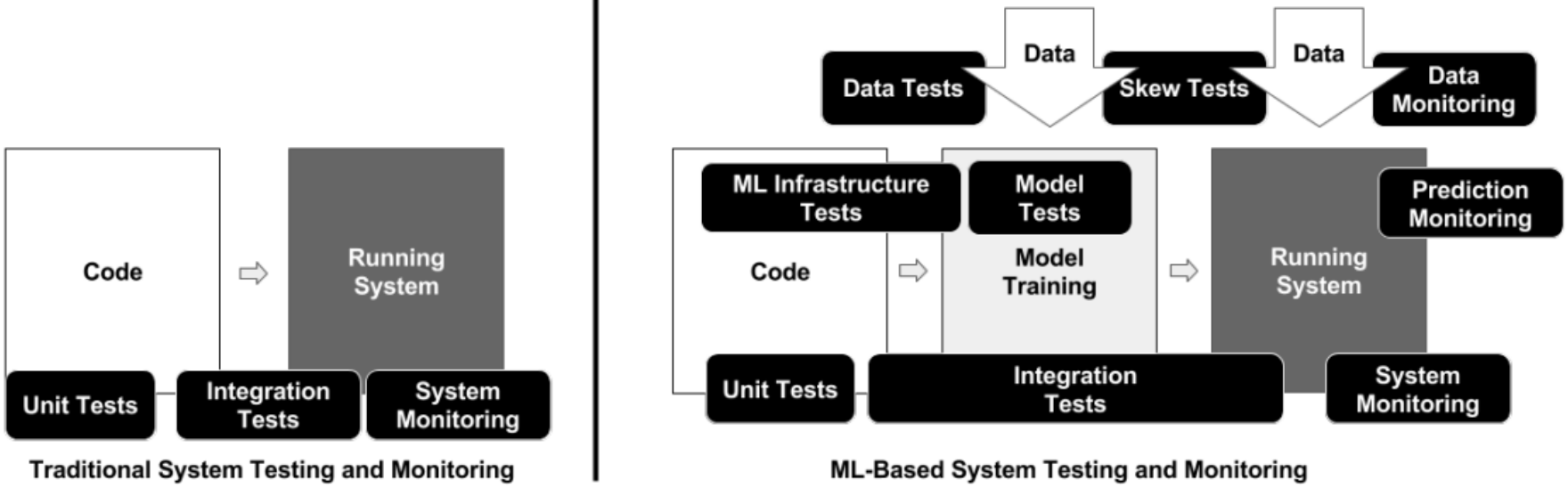


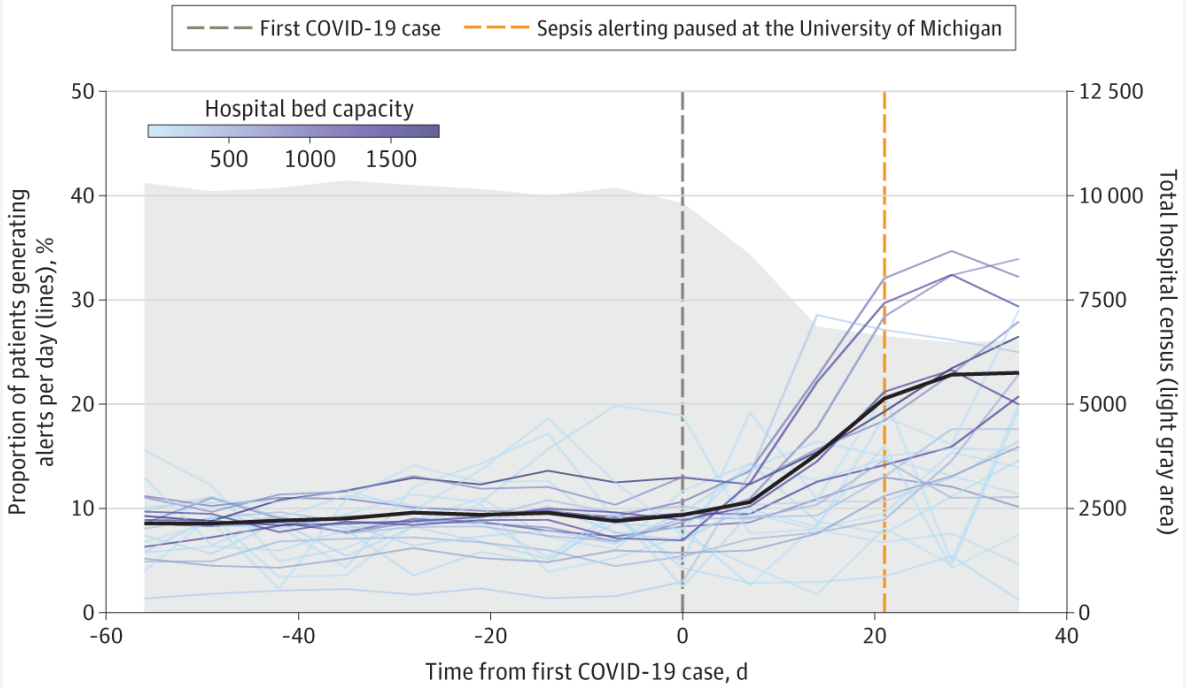
Figure 1. **ML Systems Require Extensive Testing and Monitoring.** The key consideration is that unlike a manually coded system (left), ML-based system behavior is not easily specified in advance. This behavior depends on dynamic qualities of the data, and on various model configuration choices.

Breck, Eric, et al. "The ML test score: A rubric for ML production readiness and technical debt reduction." *2017 IEEE international conference on big data (big data)*. IEEE, 2017.

Sweden AI Factory

Challenges of data-driven software

- Data is central to the model, changes in input data distribution can lead to unforeseen predictions



Wong A, Cao J, Lyons PG, et al. Quantification of Sepsis Model Alerts in 24 US Hospitals Before and During the COVID-19 Pandemic. *JAMA Netw Open.* 2021;4(11):e2135286. doi:10.1001/jamanetworkopen.2021.35286

Sweden AI Factory

API model name	Current state	Deprecated	Tentative retirement date
`claude-fable-5`	Active	N/A	Not sooner than June 9, 2027
`claude-opus-5`	Active	N/A	Not sooner than July 24, 2027
`claude-opus-4-8`	Active	N/A	Not sooner than May 28, 2027
`claude-opus-4-7`	Active	N/A	Not sooner than April 16, 2027
`claude-opus-4-6`	Active	N/A	Not sooner than February 5, 2027
`claude-opus-4-5-20251101`	Active	N/A	Not sooner than November 24, 2026
`claude-opus-4-1-20250805`	Retired	June 5, 2026	August 5, 2026
`claude-opus-4-20250514`	Retired	April 14, 2026	June 15, 2026
`claude-sonnet-5`	Active	N/A	Not sooner than June 30, 2027
`claude-sonnet-4-6`	Active	N/A	Not sooner than February 17, 2027
`claude-sonnet-4-5-20250929`	Active	N/A	Not sooner than September 29, 2026
`claude-sonnet-4-20250514`	Retired	April 14, 2026	June 15, 2026
`claude-3-7-sonnet-20250219`	Retired	October 28, 2025	February 19, 2026
`claude-haiku-4-5-20251001`	Active	N/A	Not sooner than October 15, 2026
`claude-3-5-haiku-20241022`	Retired	December 19, 2025	February 19, 2026
`claude-3-haiku-20240307`	Retired	February 19, 2026	April 20, 2026

https://platform.claude.com/docs/en/about-claude/model-deprecations

Changing models

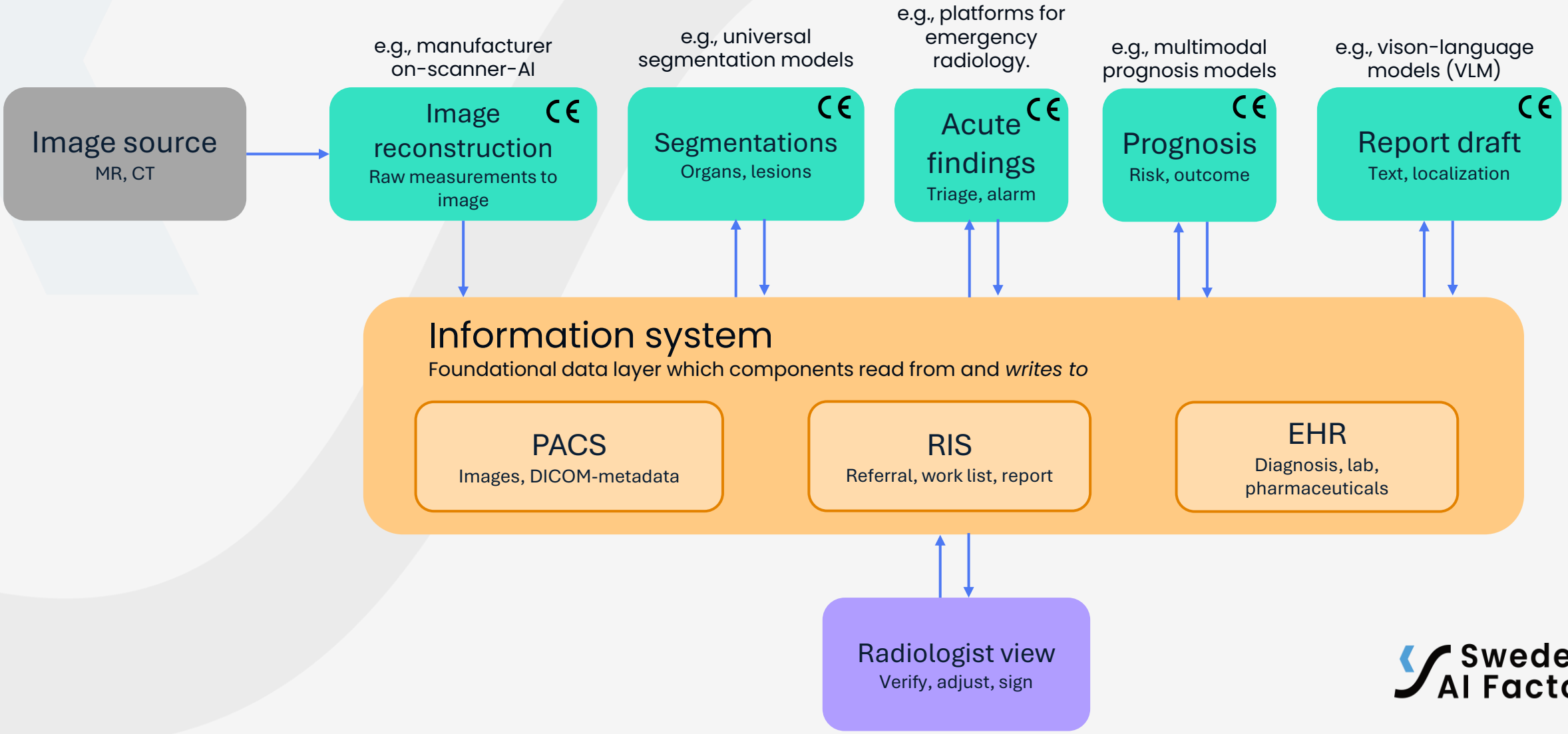
- A frontier model provider can’t keep old models running forever, they need to allocate their resources to competing at the frontier
- If the expected lifetime of a model is a year, do you have a plan to migrate?
- A year in AI is a long time, but not for most businesses

CACE - Changing Anything Changes Everything

https://www.moderndailyknitting.com/community/unraveling-yarn/

Sweden AI Factory

Near future hospital AI systems



Sweden AI Factory

Model development is the least of your concerns

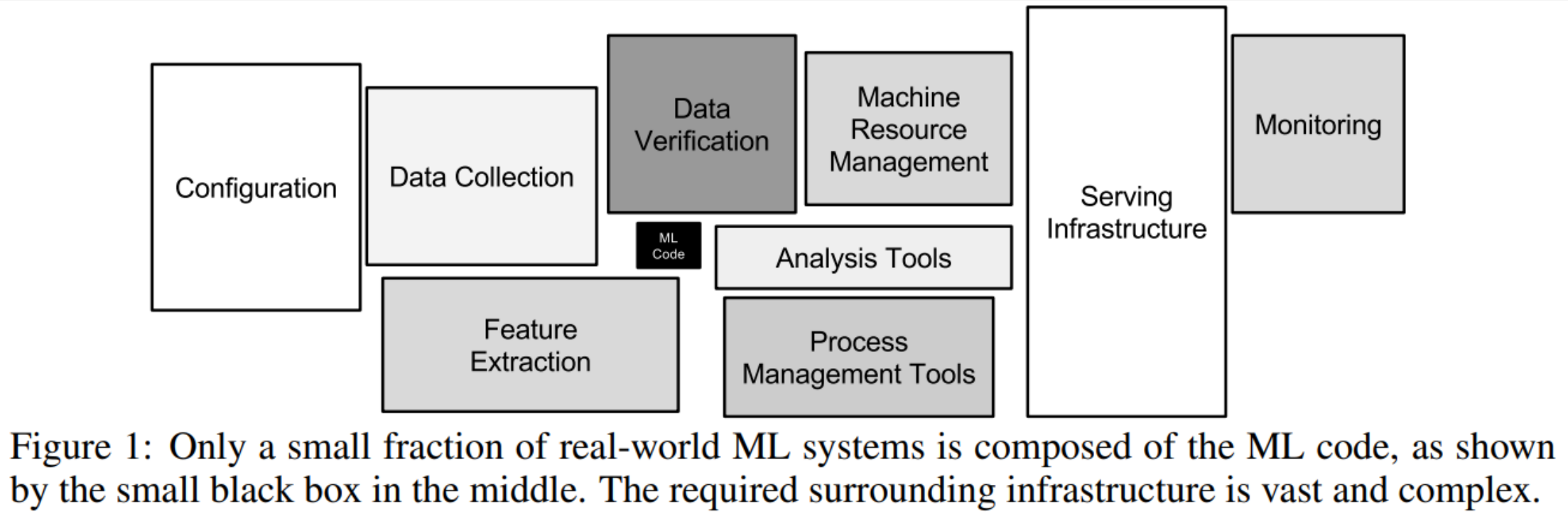
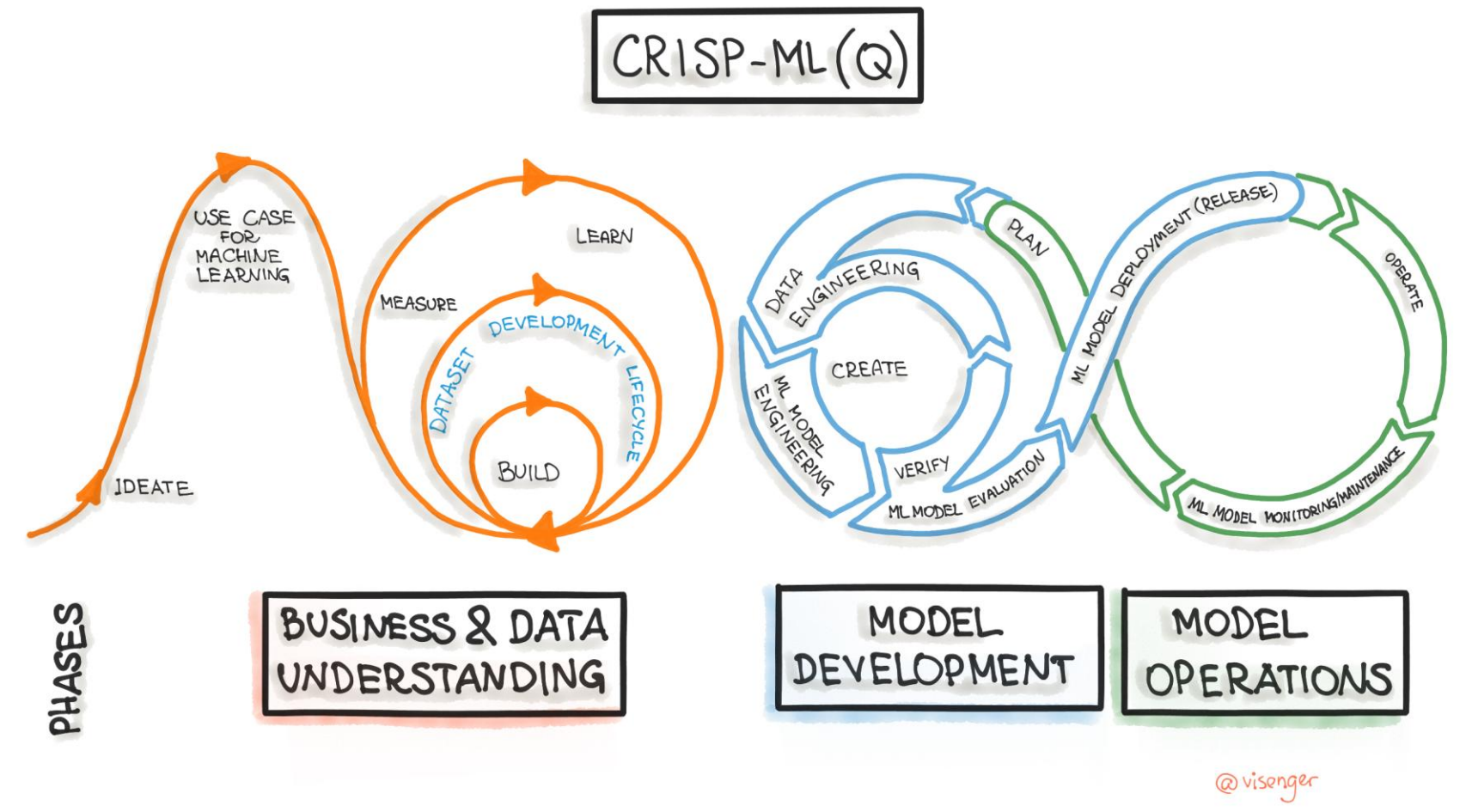


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Sculley, David, et al. "Hidden technical debt in machine learning systems." *Advances in neural information processing systems* 28 (2015).

Sweden AI Factory

Processes - MLOps

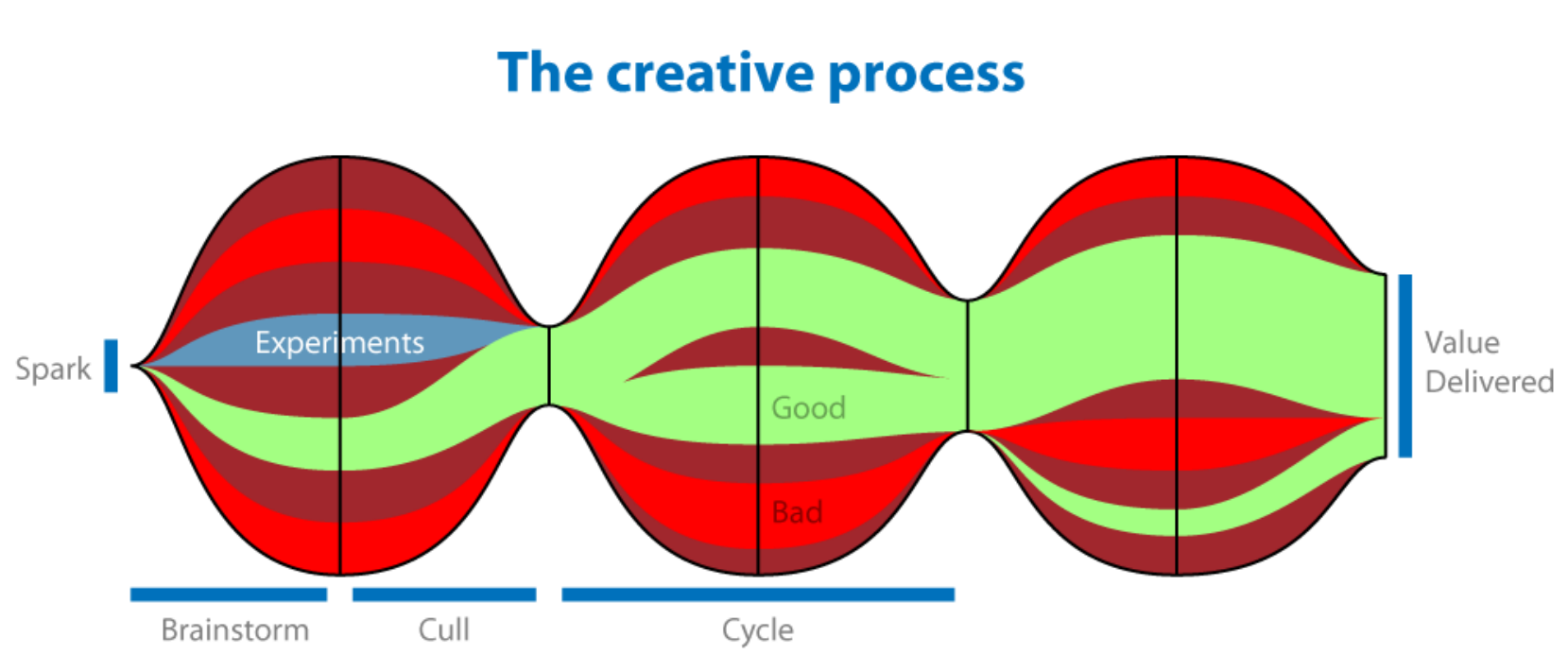


https://ml-ops.org/content/crisp-ml

Sweden AI Factory

The Creative process

Amateurs works with ideas – Professionals works with processes
- Sören Hellqvist

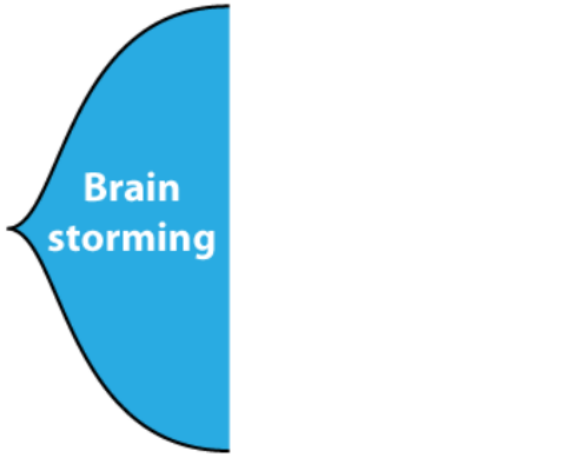


Daniel Cook, *Visualizing the creative process*,
https://lostgarden.com/2010/08/17/visualizing-the-creative-process/

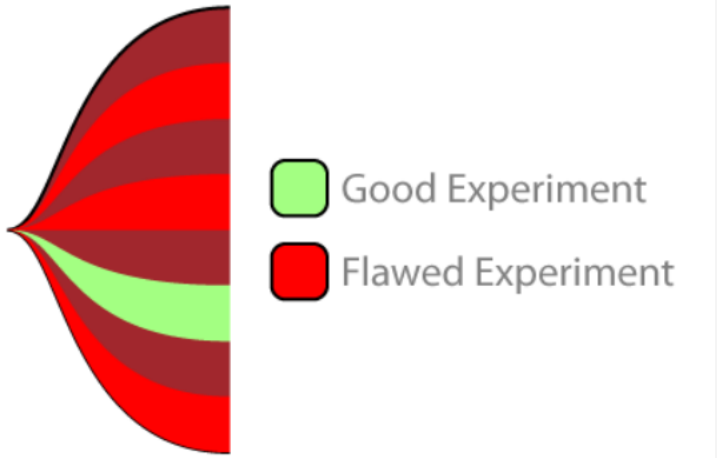
Sweden AI Factory

Brainstorming, sketching, expanding

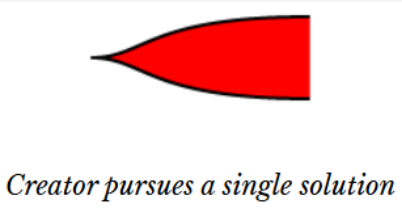
Creatives love the brainstorming phase!



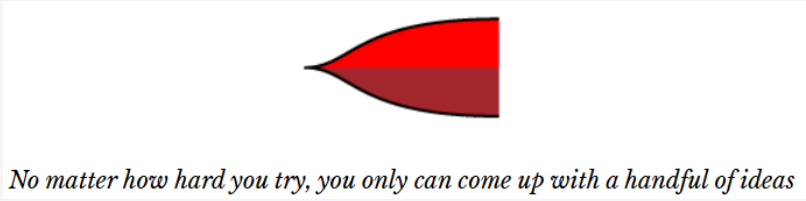
Brainstorming starts out small and expands over time



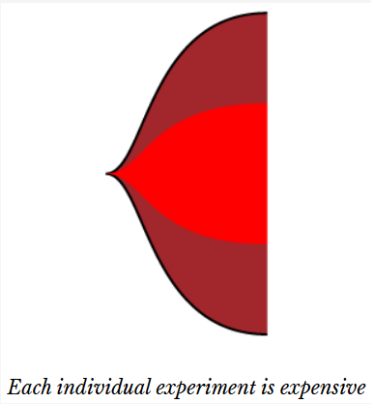
A multitude of experiments arise during brainstorming



Creator pursues a single solution



No matter how hard you try, you only can come up with a handful of ideas



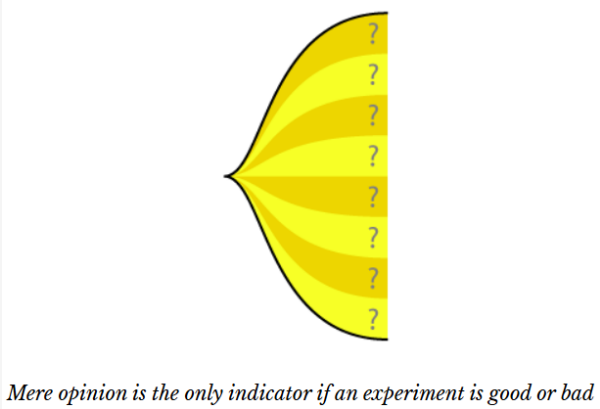
Each individual experiment is expensive

Daniel Cook, *Visualizing the creative process*,
https://lostgarden.com/2010/08/17/visualizing-the-creative-process/

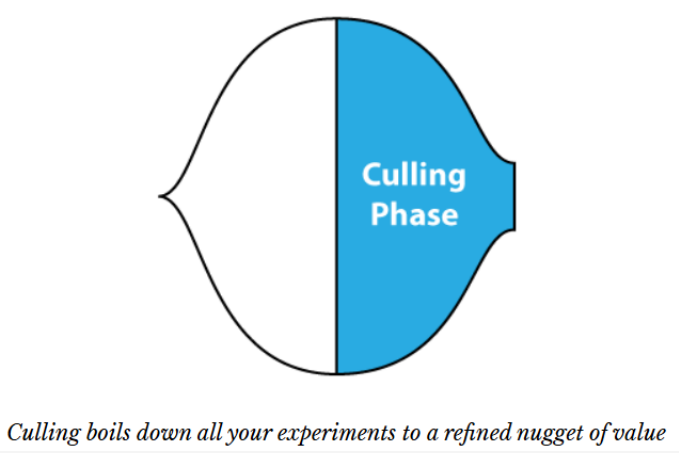
Sweden AI Factory

Culling, critiquing, editing, reviewing, contracting

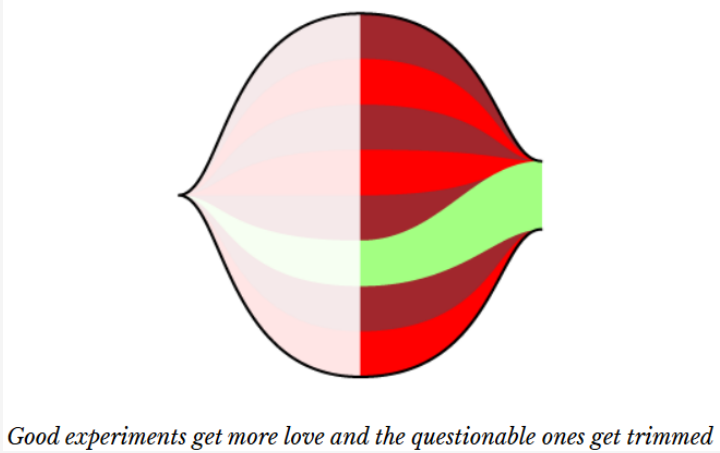
Not as loved



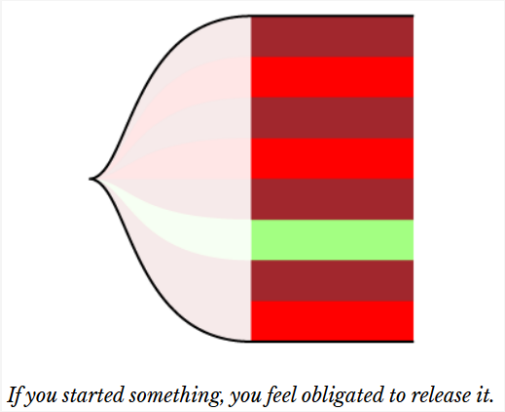
Mere opinion is the only indicator if an experiment is good or bad



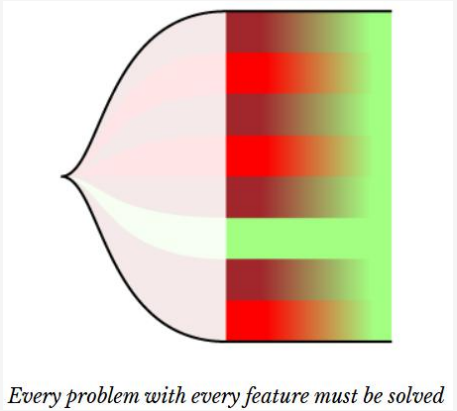
Culling boils down all your experiments to a refined nugget of value



Good experiments get more love and the questionable ones get trimmed



If you started something, you feel obligated to release it.

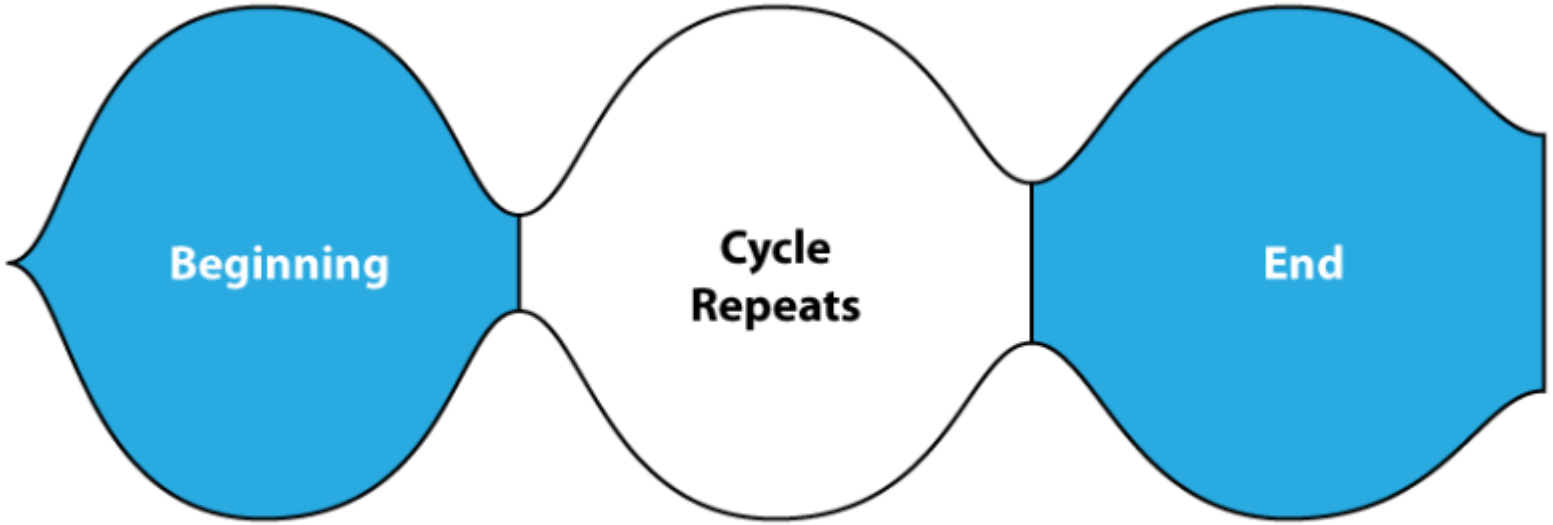


Every problem with every feature must be solved

Daniel Cook, *Visualizing the creative process*,
https://lostgarden.com/2010/08/17/visualizing-the-creative-process/

Sweden AI Factory

Cycles, Iterations

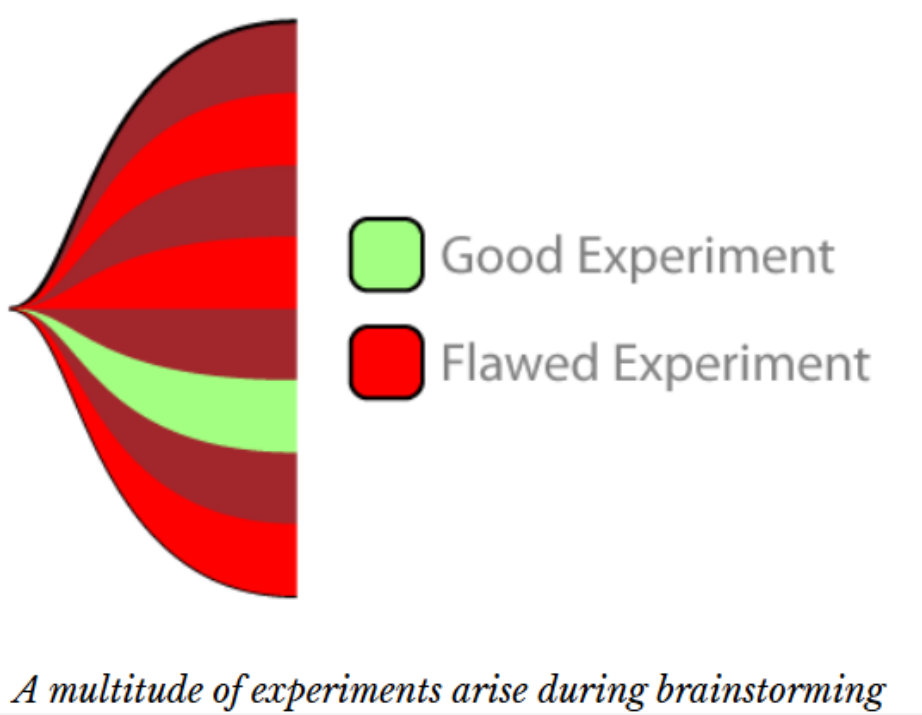


Each cycle results in accumulating more value for the customers of your labor. When you've generated enough value, you stop.

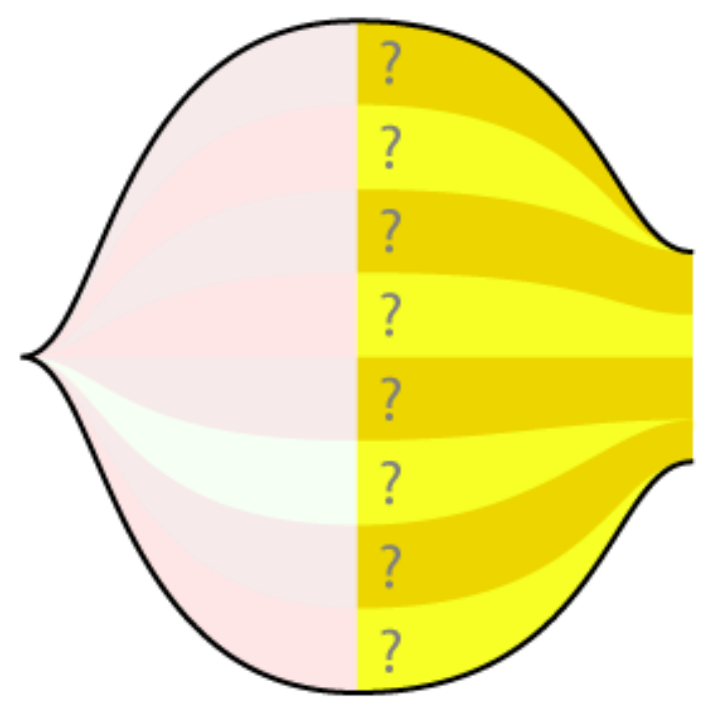
Daniel Cook, *Visualizing the creative process*,
https://lostgarden.com/2010/08/17/visualizing-the-creative-process/

Sweden AI Factory

The allure of AI projects



A multitude of experiments arise during brainstorming



Creating experiments is faster than ever before!

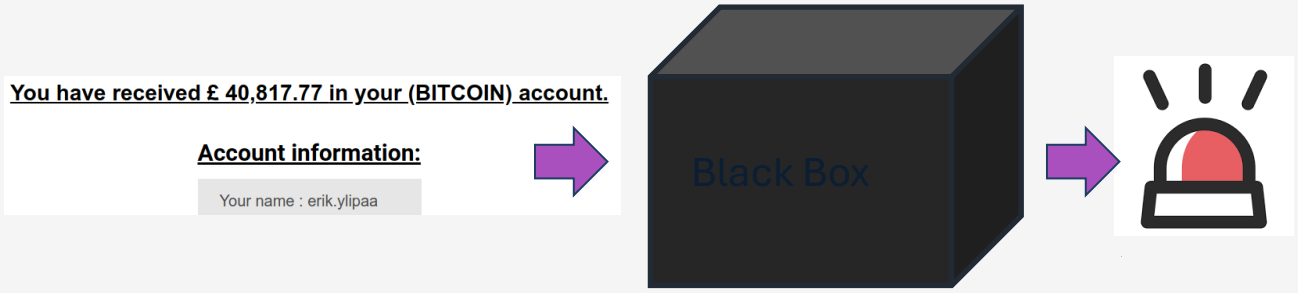
But do you have a clear idea of what is a Good or Flawed experiment?

Daniel Cook, *Visualizing the creative process*,
https://lostgarden.com/2010/08/17/visualizing-the-creative-process/

Sweden AI Factory

Evaluating AI – how to trust an alien?

Would you trust the medical diagnosis an extraterrestrial gave you?

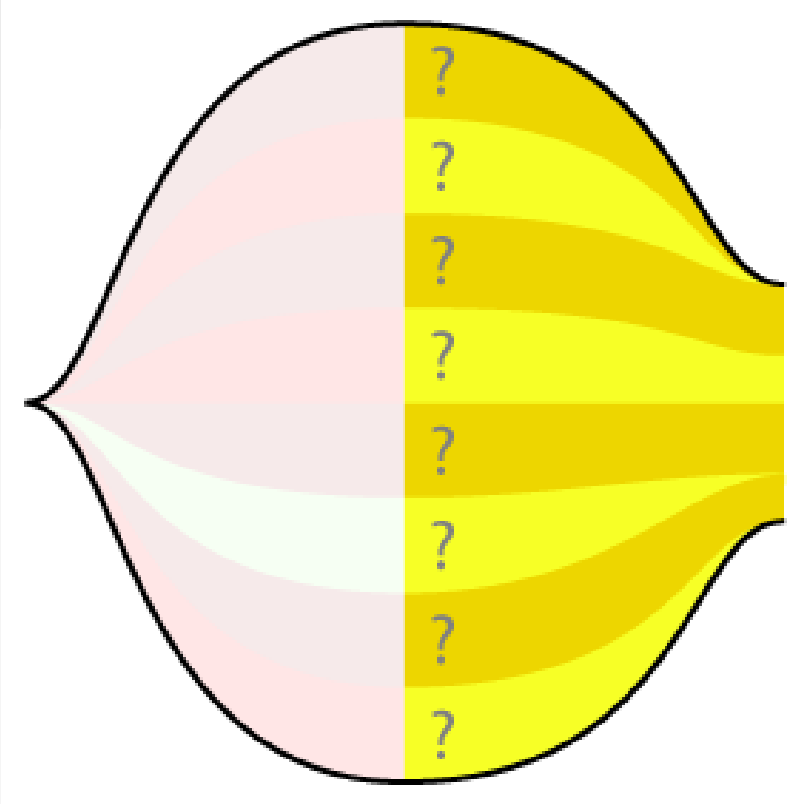


Sweden AI Factory

What does correct look like?

For evaluation, you need some kind of measuring instrument

- Rule-based metric (e.g. accuracy)
- Judgement based evaluation, human or AI-based
- Should be **calibrated**
  - **Repeatability**: Does repeated measurements with an instrument give similar results?
  - **Reproducibility**: Do two different instruments (judges) agree on the same data point?



Sweden AI Factory

Intrinsic vs. Extrinsics evaluation

Intrinsic evaluation

- **Intrinsic evaluation** focus on evaluating the model outputs directly without consideration to an end task
  - Often what we practically build our evaluation around

Extrinsic evaluation

- **Extrinsic evaluation** instead looks at how the AI affect a downstream task like reducing support tickets or improved customer satisfaction
  - Often the problem we would really like to solve

Aim for extrinsic evaluation, start with intrinsic.
Can you find an evaluation task which correlates with the intrinsic (perhaps precision and recall in retrieval with user satisfaction)?

Sweden AI Factory

Evaluation strategies

Automatic vs. Human Evaluation. *Automatic* evaluation uses deterministic functions (BLEU, exact match) or learned models (BERTScore, LLM-as-judge) to score outputs without human involvement. *Human* evaluation involves annotators rating or ranking model outputs. Table 14.1 summarises the trade-offs.

Table 14.1: Taxonomy of evaluation approaches with key trade-offs.

Type	Cost	Speed	Reproducibility	Validity
Automatic (rule-based)	Very low	Very fast	Perfect	Low–Medium
Automatic (model-based)	Low	Fast	High	Medium–High
Crowdsourced human	Medium	Days	Medium	Medium
Expert human	High	Weeks	Low–Medium	High
Extrinsic / A/B test	Very high	Months	Low	Very high

Roitman, Haggai. "The Hitchhiker's Guide to Agentic AI: From Foundations to Systems." *arXiv preprint arXiv:2606.24937* (2026).

Sweden AI Factory

Reference-based vs. Reference-free evaluation

Reference-based

- **Reference-based** evaluation has some gold standard to compare to. Here the space of agreeable answers is typically small (e.g., yes/no answers, distance metrics etc.)

Reference-free

- **Reference-free** evaluation doesn't have a reference datapoint, instead relies on something giving a judgement like human raters or LLM-as-a-judge.

If you can formulate your problem as **reference-based**, it will often make development easier

Sweden AI Factory

LLM-as-a-judge

- We can use AI to evaluate another AI system, often when the output of the system is open ended (many ways of being correct)
- If using LLM-as-a-judge, is it **calibrated** to the goal?
- By solving the AI evaluation problem using AI, you now have another AI evaluation problem
  - It's typically a smaller evaluation-problem which can greatly help in evaluation of the primary syste
  - Shouldn't replace human evaluation, run them side by side and compare human vs. LLM-as-a-judge evaluations for miscalibration
- For evaluating agent **trajectories**, it's common to use LLM-as-a-judge (difficult for humans to evaluate in volume)

Sweden AI Factory

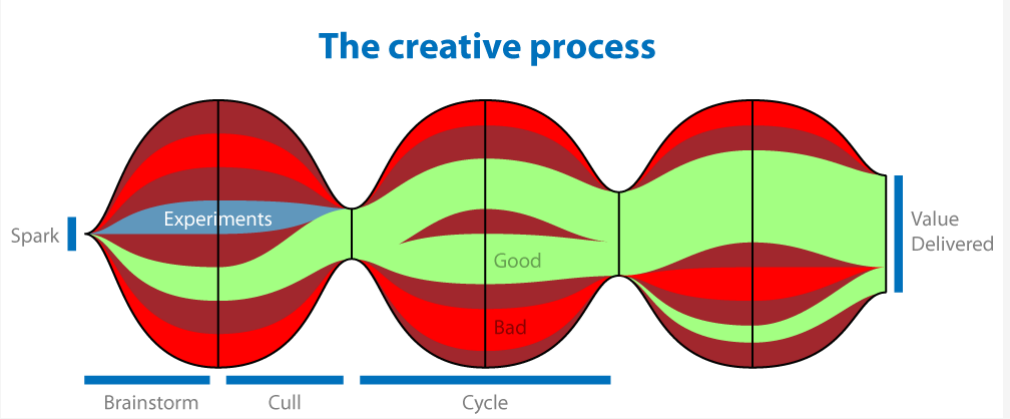
Versioning and Roll-back

- You should be able to **roll back** the system to an earlier version
  - Version control data, code, configuration and the memory system if you use something like RAG or LLM-Wiki
- **Git** for code and configurations (with LFS for large files)
- Data Version Control (**DVC**) for data stored elsewhere

Sweden AI Factory

Experimentation framework

- To automate things, adopt an experiment framework
  - It captures **versioned** experiments, from configuration, data and model, connecting them to evaluation results
  - Makes it easier to iterate on the fruitful experiments
- Tools like **MLFlow** and **Label Studio** can get you a long way



mlflow™

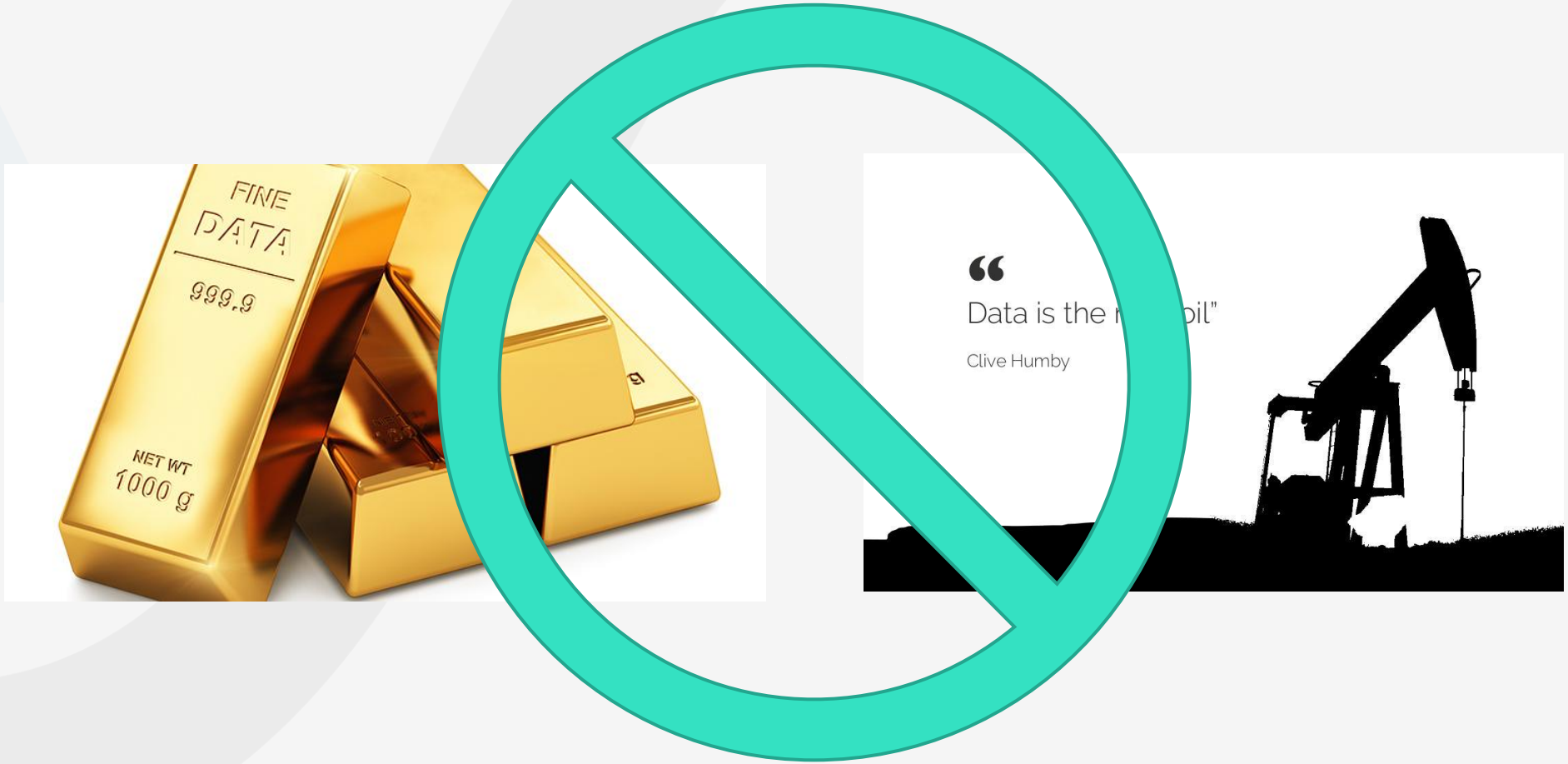
Label Studio

Sweden AI Factory

Data is the new gold/oil?

Sweden AI Factory

Data is the new gold/oil?



Sweden AI Factory

Data is the new e-waste

Electronic waste: When billions in gold, other precious materials are discarded
https://www.dw.com/en/un-electronic-waste-gold-silver-platinum/a-54022278

Sweden AI Factory

Data is the new e-waste

- Byproduct of some other process
- Missing the context it was produced in
- Changes over time
- We can decide what to collect and how to do it
    - Build “recycling streams”
- **Isn’t necessarily valuable!**

Sweden AI Factory

Data readiness is like building recycling

Requires real effort

Data is collected as a byproduct of some organization process

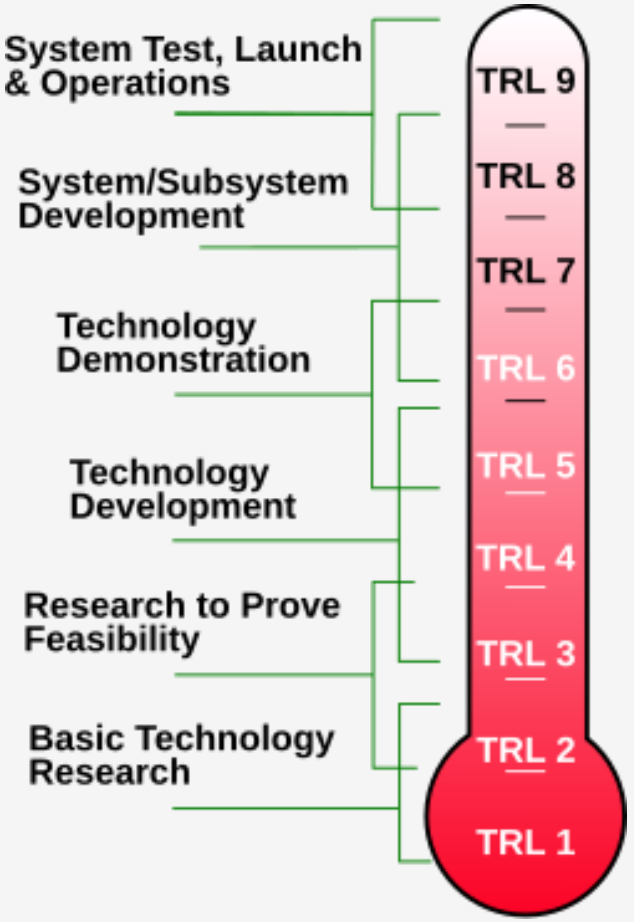
Relevant data (for a specific use) is identified by an organization domain expert

Data processing is used to refine the available data to make it useful for AI

Sweden AI Factory

Data Readiness Levels

- Conceptual framework (not really for operations)
- Makes it easier to talk about data
- Draws inspiration from **Technology Readiness Levels** which is used to conceptually talk about technologies
  - Originally defined by NASA, commonly used in the innovation system

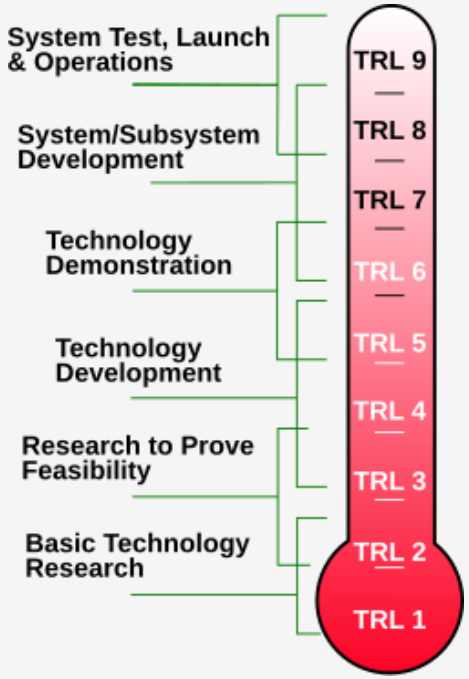


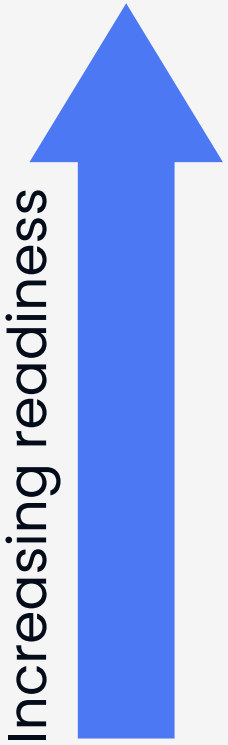
Technology Readiness Levels

Neil Lawrence, *Data Readiness Levels*
http://data-readiness.org/

Sweden AI Factory

Data Readiness Levels





Band	Level	Meaning	The state of data
A	*A-1*	***Utility***	Ready for analysis with respect to given business objectives, hypotheses, questions, or context.
	...		
B	*B-1*	***Validity***	Ready to define the candidate questions and hypotheses. Includes exploratory analysis, data characterization, entity disambiguation, and de-duplication, etc.
	...		
C	*C-1*	***Accessibility***	Ready to be loaded in analysis software. Includes programmatic access, format conversions, and legal aspects, etc.
	...		
	C-n	*Hearsay data*	*"I'm sure AI can solve this problem. We have lots of data!"*

Olsson, Fredrik, and Magnus Sahlgren. "We need to talk about data: The importance of data readiness in natural language processing." *arXiv preprint arXiv:2110.05464* (2021).

Real world example: Customer say they have thousands of text documents. When the data is received, it's thousands of PDFs containing scanned pages (images)

Sweden AI Factory

Guiding questions

Band C

1. Do you have programmatic access to the data?
2. Are your licenses in order?
3. Do you have lawful access to the data?
4. Has there been an ethics assessment of the data?
5. Is the data converted to an appropriate format?

Band B

6. Are the characteristics of the data known?
7. Is the data validated?

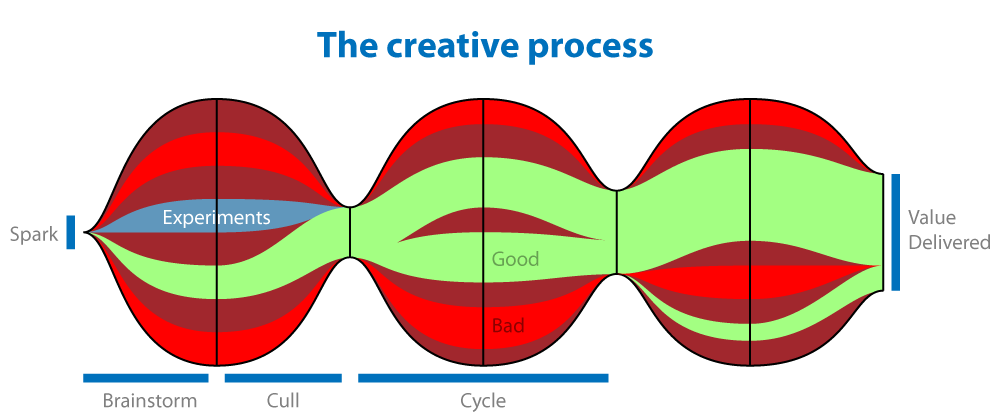
Band A

8. Do stakeholders agree on the objective of the current use case?
9. Is the purpose of using the data clear to all stakeholders?
10. Is the data sufficient for the current use case?
11. Are the steps required to evaluate a potential solution clear?
12. Is your organization prepared to handle more data like this beyond the scope of the project?
13. Is the data secured?
14. Is it safe to share the data with others?
15. Are you allowed to share the data with others?

Olsson, Fredrik, and Magnus Sahlgren. "We need to talk about data: The importance of data readiness in natural language processing." *arXiv preprint arXiv:2110.05464* (2021).

Sweden AI Factory

Summary



- Build **processes** instead of artifacts
- Spend a lot of time on designing **evaluation criteria**
  - Think about it as **quality control**
  - Before you start building, ask yourself **”how do I evaluate this”?**
- Make sure the data you use for reference and evaluation is of high **readiness**
- **Automate** processes as much as possible
  - Use an **experimentation framework** and **version control** of configuration, data and code

Sweden AI Factory

Thank you!

Erik Ylipää (erik.ylipaa@ri.se)

ML Rubric for contemporary AI – the goal is to answer yes to these questions

- **Data 5:** The data pipeline has appropriate privacy controls
- **Model 1**: Every model specification undergoes a code review and is checked in to a repository
- **Model 2**: Offline proxy metrics correlate with actual online impact metrics
- **Model 3**: All hyperparameters have been tuned
- **Model 4**: The impact of model staleness is known
- **Model 5**: A simpler model is not better
- **Model 6**: Model quality is sufficient on all important data slices
- **Model 7**: The model has been tested for considerations of inclusion
- **Infra 1**: Training is reproducible
- **Infra 4**: Model quality is validated before attempting to serve it
- **Infra 5**: The model allows debugging by observing the step-by-step computation of training or inference on a single example
- **Infra 6**: Models are tested via a canary process before they enter production serving environments
- **Infra 7**: Models can be quickly and safely rolled back to a previous serving version
- **Monitor 1**: Dependency changes result in notification
- **Monitor 4**: Models are not too stale
- **Monitor 6**: The model has not experienced a dramatic or slow-leak regressions in ~~training speed~~, serving latency, throughput, or RAM usage
- **Monitor 7**: The model has not experienced a regression in prediction quality on served data

Sweden AI Factory

Breck, Eric, et al. "The ML test score: A rubric for ML production readiness and technical debt reduction." *2017 IEEE international conference on big data (big data)*. IEEE, 2017.