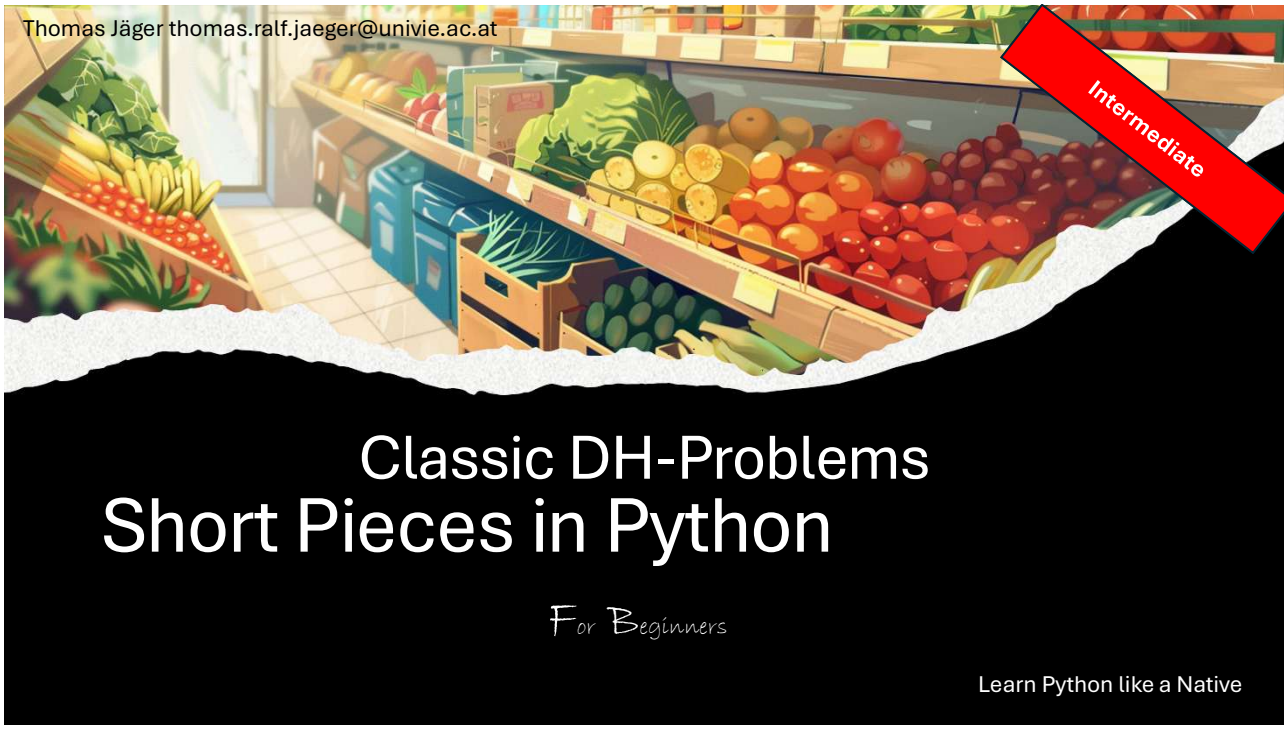
Thomas Jäger thomas.ralf.jaeger@univie.ac.at



Classic DH-Problems
Short Pieces in Python

For Beginners

Learn Python like a Native

Programming:
- Levenshtein
- Clusters
- Tree

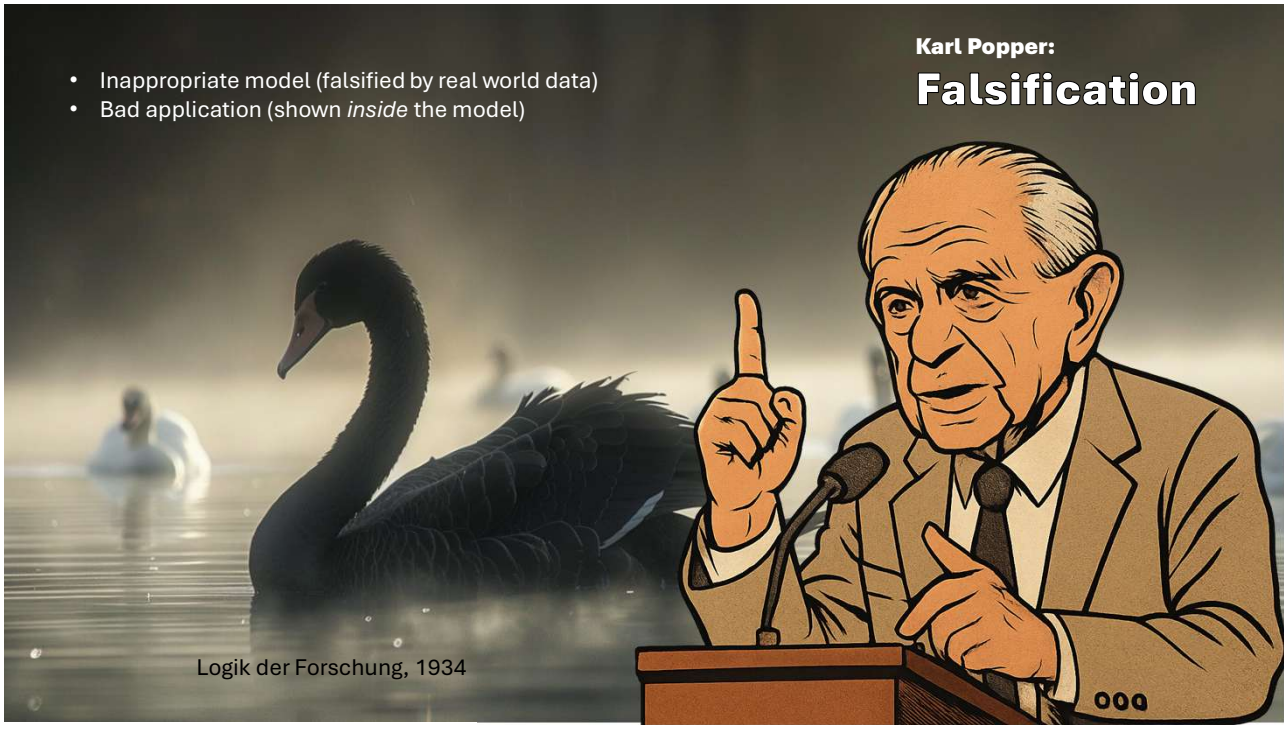
Developing:
- Pipelines

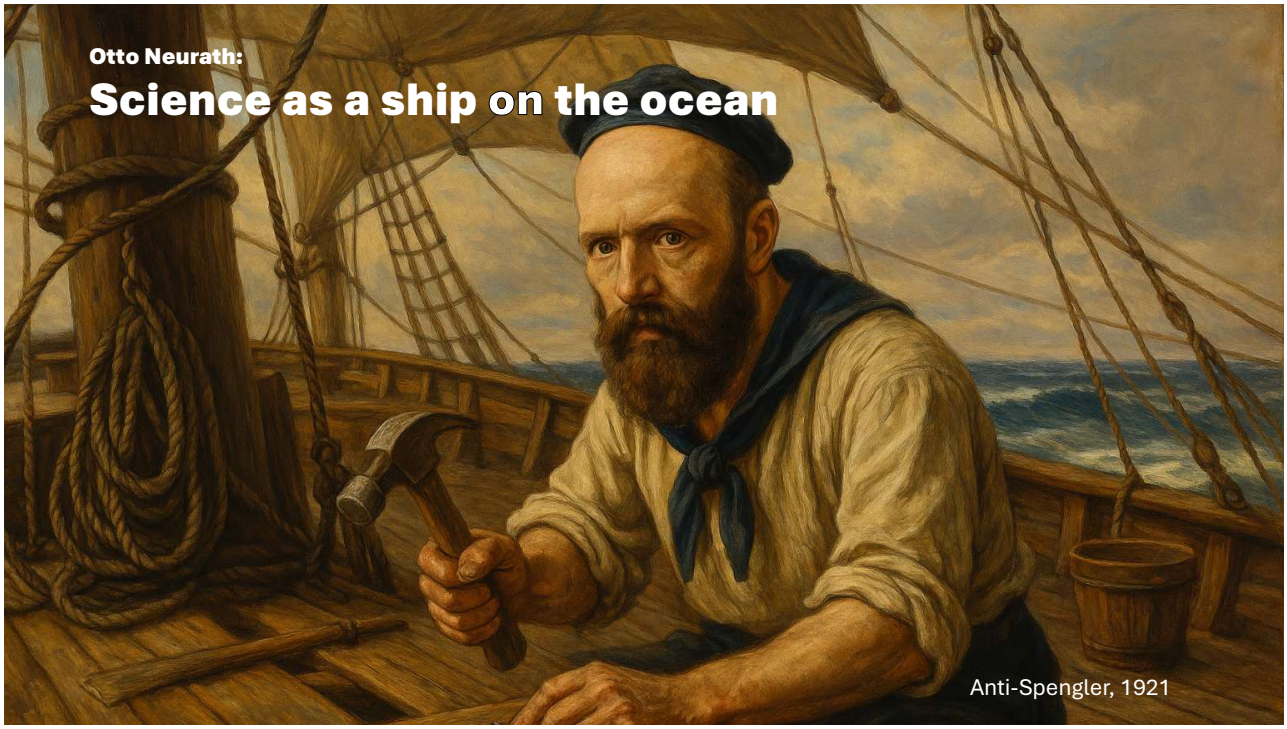
Architecture:
- Pipelines
- Folders
- Check Tasks

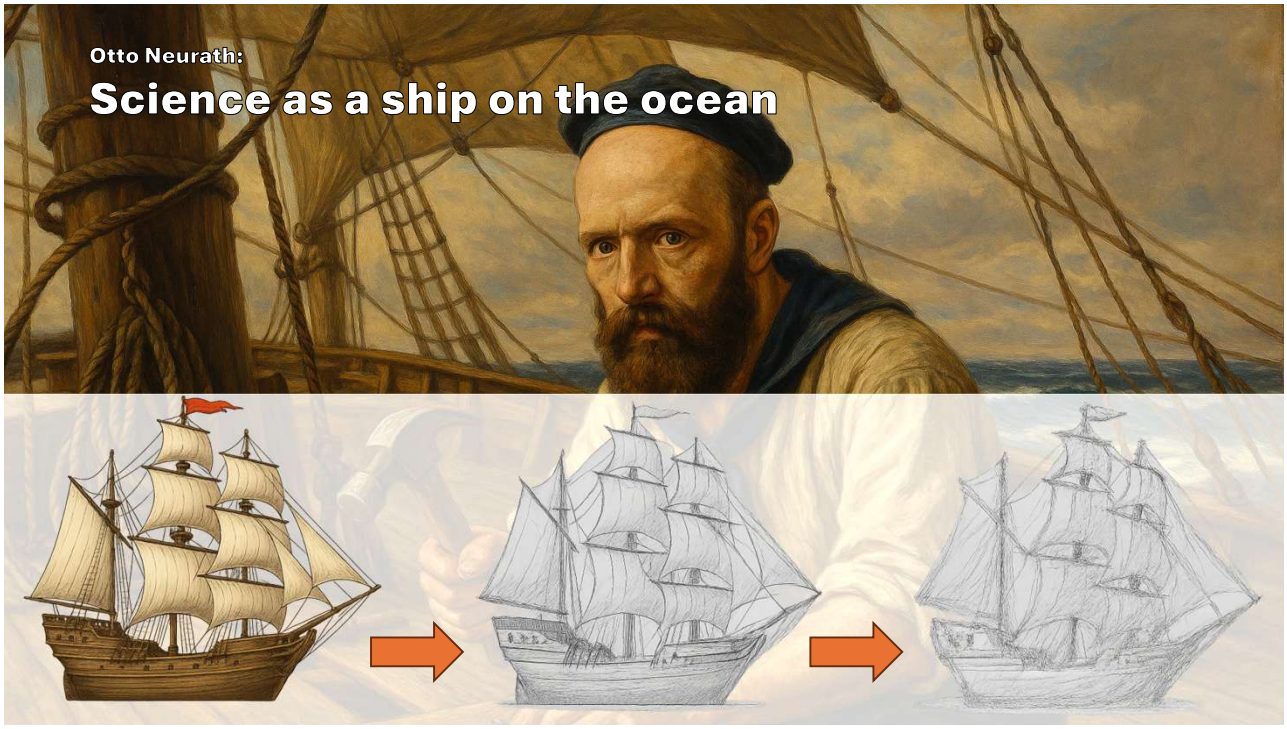
Professional Knowledge:
- Pair Programming
- Model collapse

Herbert Stachowiak: Allgemeine Modelltheorie, Wien/New York: Springer 1973

- **Abbild (image)**
  - → New creation, not identical to the original
- **Verkürzung (shortening)**
  → just the relevant parts
- **Pragmatik (obvious)**
  → needs a subjective perspective to work




Otto Neurath:
Science as a ship on the ocean
Anti-Spengler, 1921

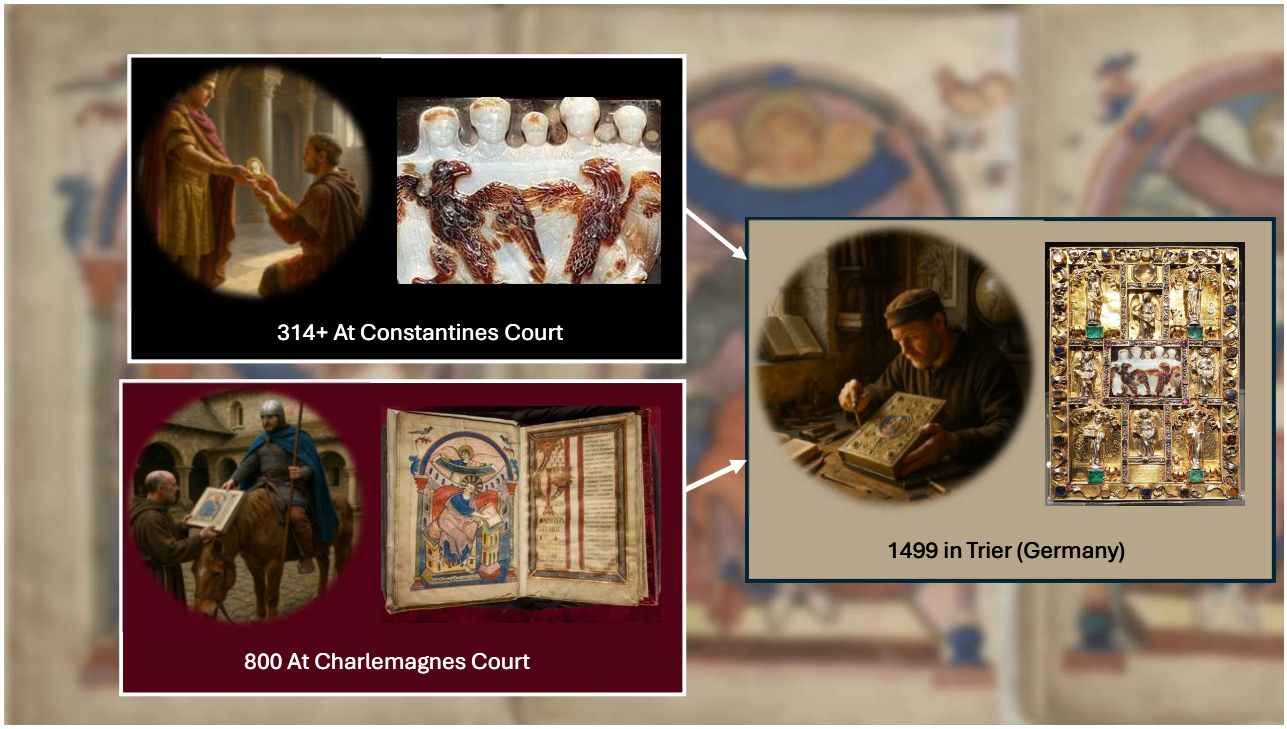

Otto Neurath:
Science as a ship on the ocean

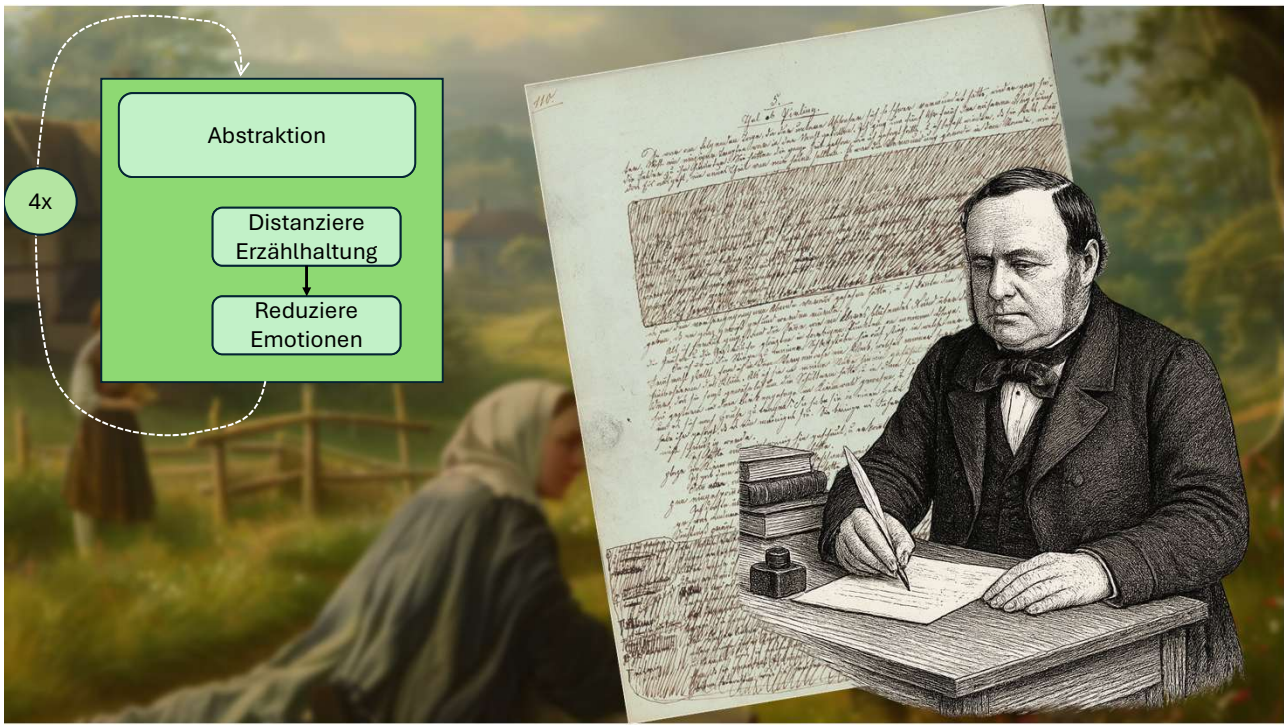

LIVE
A PARADIGM SHIFT!
Thomas S. Kuhn:
The structure of scientific revolutions, 1962
Paradigm Shifts

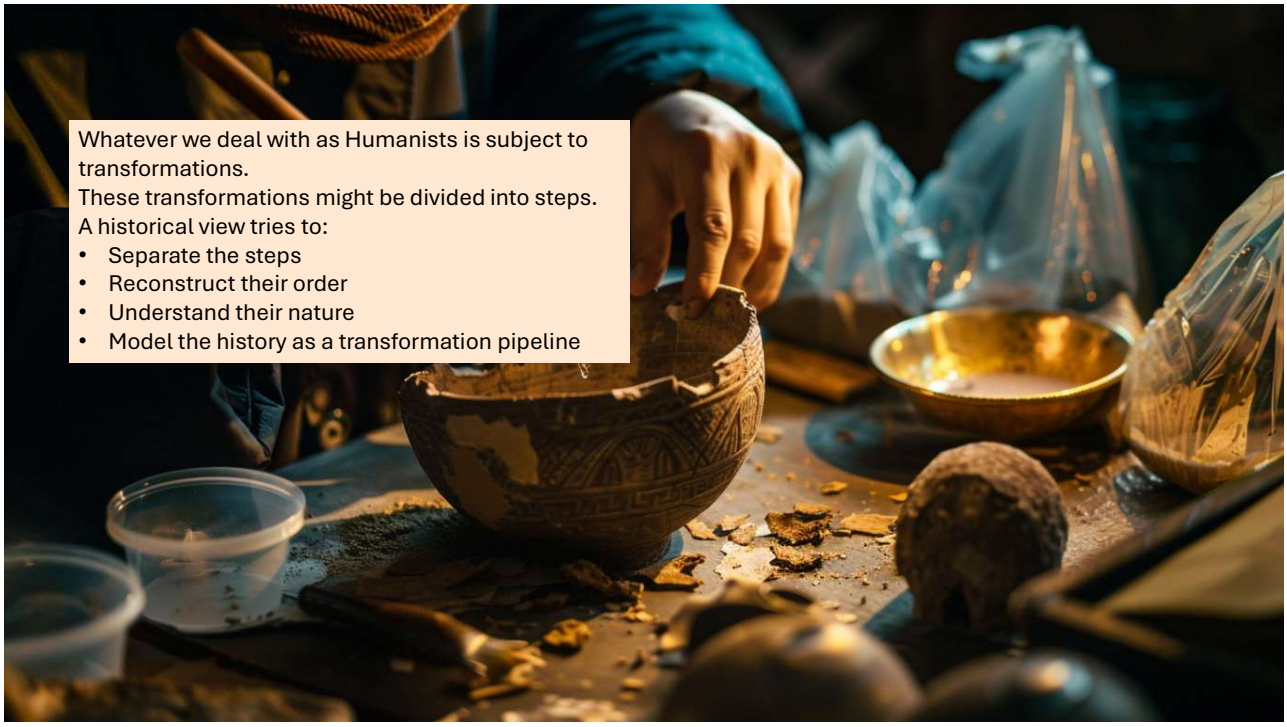

LIVE
A PARADIGM SHIFT!
Thomas S. Kuhn
• Falsification of ruling model is not accepted
• Scientists get discredited
• Science starts to zoom in on details and shift from reality
Paradigm Shifts
• Change happens when power shifts
• Change is radical


In science:
Models have to collapse, to make place for new models ...
Or ... never trust a model.


VIENNA PANNONIE
Historical Objects have been transformed over time

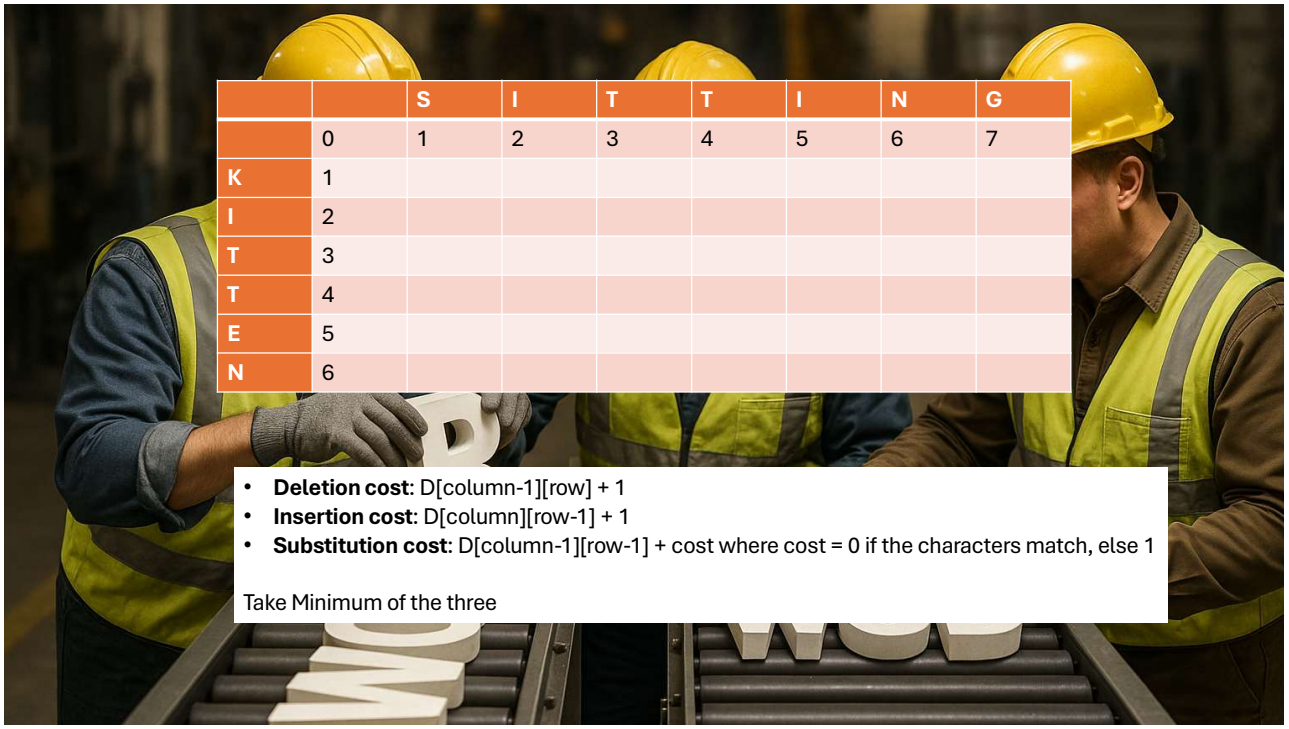

314+ At Constantines Court
800 At Charlemagnes Court
1499 in Trier (Germany)


4x
Abstraktion
Distanziere Erzählhaltung
Reduziere Emotionen



Whatever we deal with as Humanists is subject to transformations.
These transformations might be divided into steps.
A historical view tries to:

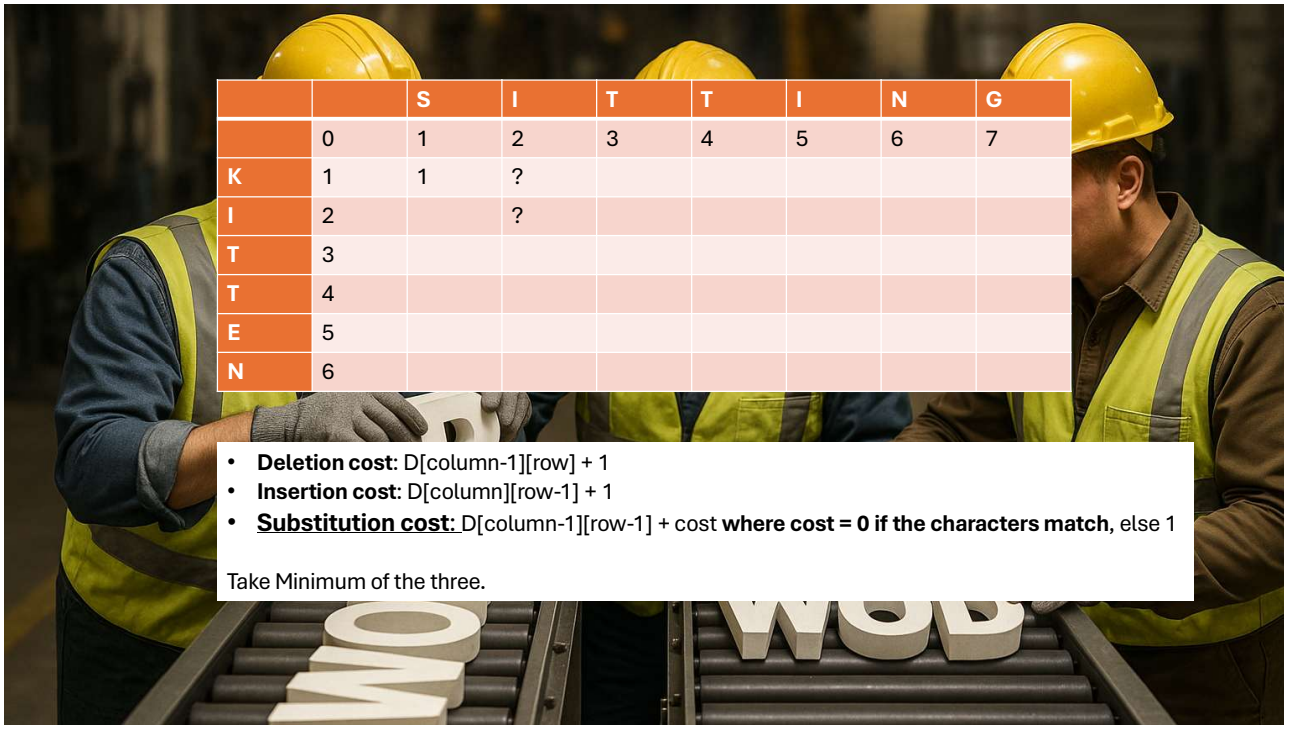
- Separate the steps
- Reconstruct their order
- Understand their nature
- Model the history as a transformation pipeline



		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1							
I	2							
T	3							
T	4							
E	5							
N	6							

- **Deletion cost**: D[column-1][row] + 1
- **Insertion cost**: D[column][row-1] + 1
- **Substitution cost**: D[column-1][row-1] + cost where cost = 0 if the characters match, else 1

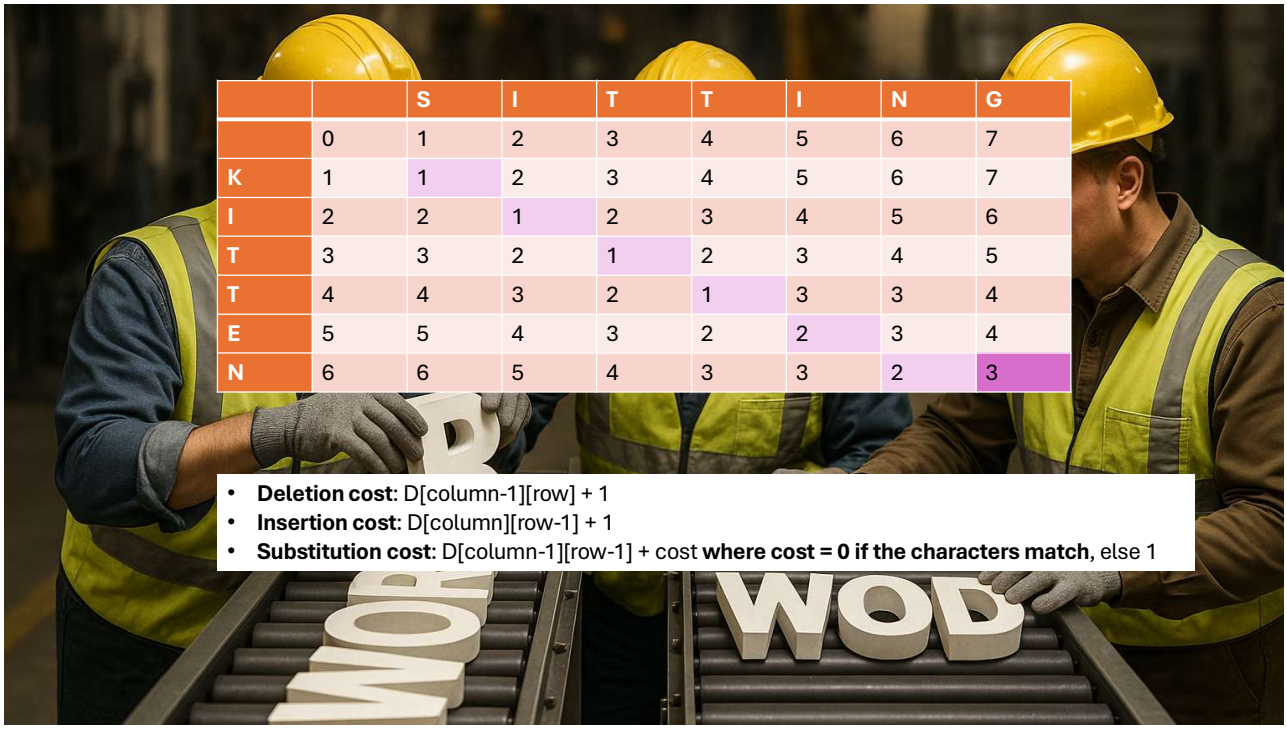
Take Minimum of the three



		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	?					
I	2		?					
T	3							
T	4							
E	5							
N	6							

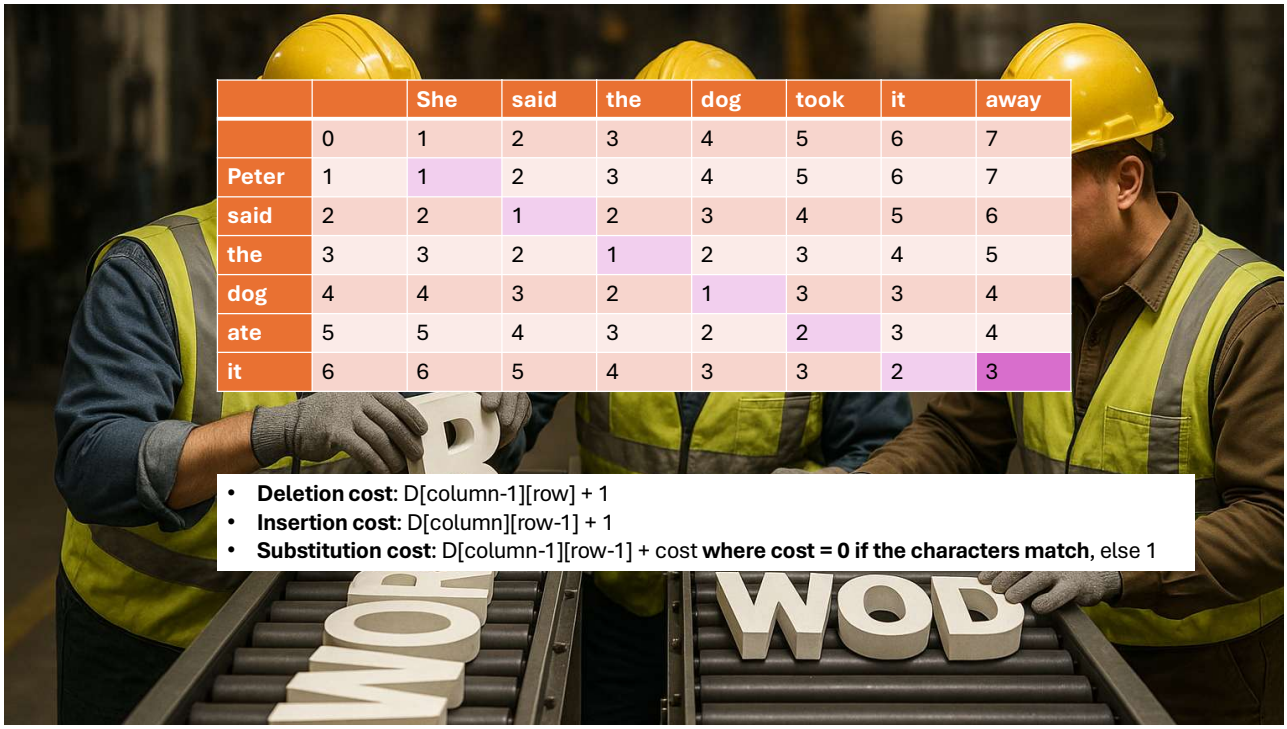
- **Deletion cost**: D[column-1][row] + 1
- **Insertion cost**: D[column][row-1] + 1
- **Substitution cost:** D[column-1][row-1] + cost **where cost = 0 if the characters match**, else 1

Take Minimum of the three.



		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	2	3	4	5	6	7
I	2	2	1	2	3	4	5	6
T	3	3	2	1	2	3	4	5
T	4	4	3	2	1	3	3	4
E	5	5	4	3	2	2	3	4
N	6	6	5	4	3	3	2	3

- **Deletion cost**: D[column-1][row] + 1
- **Insertion cost**: D[column][row-1] + 1
- **Substitution cost**: D[column-1][row-1] + cost **where cost = 0 if the characters match**, else 1



		She	said	the	dog	took	it	away
	0	1	2	3	4	5	6	7
Peter	1	1	2	3	4	5	6	7
said	2	2	1	2	3	4	5	6
the	3	3	2	1	2	3	4	5
dog	4	4	3	2	1	3	3	4
ate	5	5	4	3	2	2	3	4
it	6	6	5	4	3	3	2	3

- **Deletion cost**: D[column-1][row] + 1
- **Insertion cost**: D[column][row-1] + 1
- **Substitution cost**: D[column-1][row-1] + cost **where cost = 0 if the characters match**, else 1

TF-IDF

- **Term Frequency**: How often is a word in a particular document?
  The more – the more important.
- **Inverse Document Frequency**: How often is the word in each document?
  Words that are common in every document, are not very meaningful.
- Multiply both!

Term frequency:
$\frac{7\ mentions\ of\ coffee}{500\ words} = 0{,}014$

Inverse document frequency:
$\log_{10}\left(\frac{10\ emails}{1+4\ with\ coffee}\right) = 0{,}693$

TF*IDF = 0,0097

In **four** of the last **ten** emails the word „coffee“ was used. The last of those emails was ***500*** words long! It contained the word „coffee **seven** times.

Jaccard

Menge	Elemente
A	{1,2,3,4}
B	{3,4,5,6}

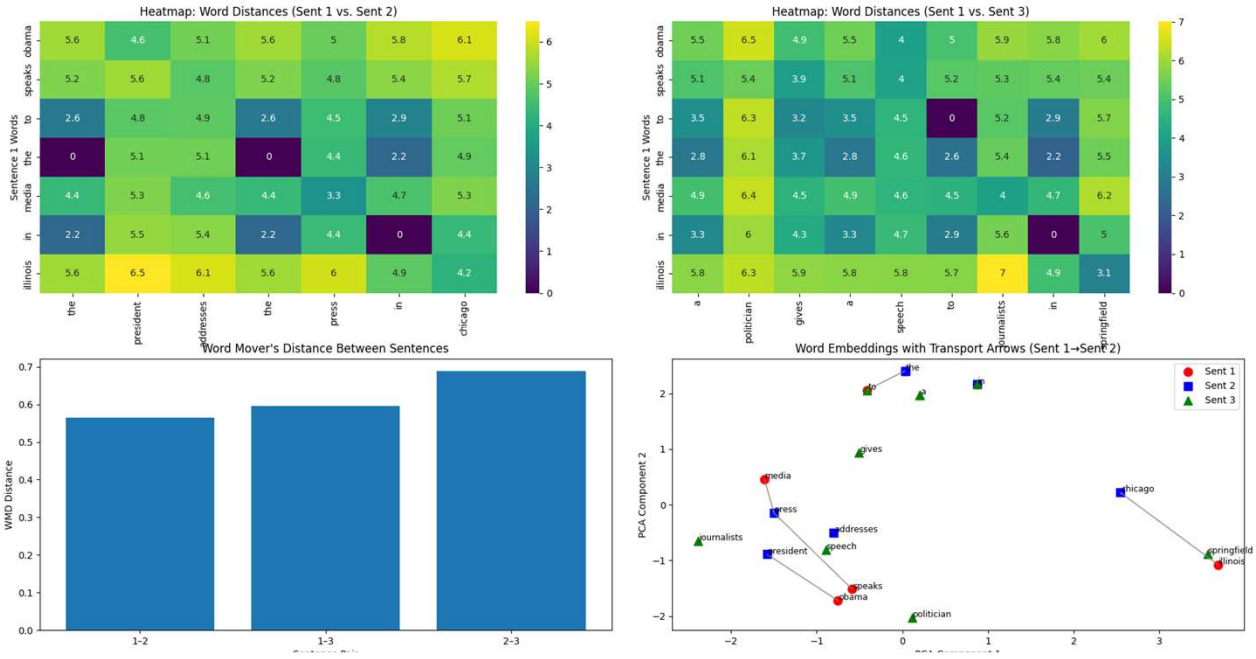
|A∩B|=2,|A∪B|=6

$$1 - \frac{2}{6}$$

Gestalt Pattern Matching

S1	K	A	T	Z	E
S2	T	A	N	Z	E

$$\frac{2\ (|"A"| + |"ZE"|)}{|S1| + |S2|} = \frac{2(1+2)}{5+5} = \frac{6}{10} = 0{,}6$$

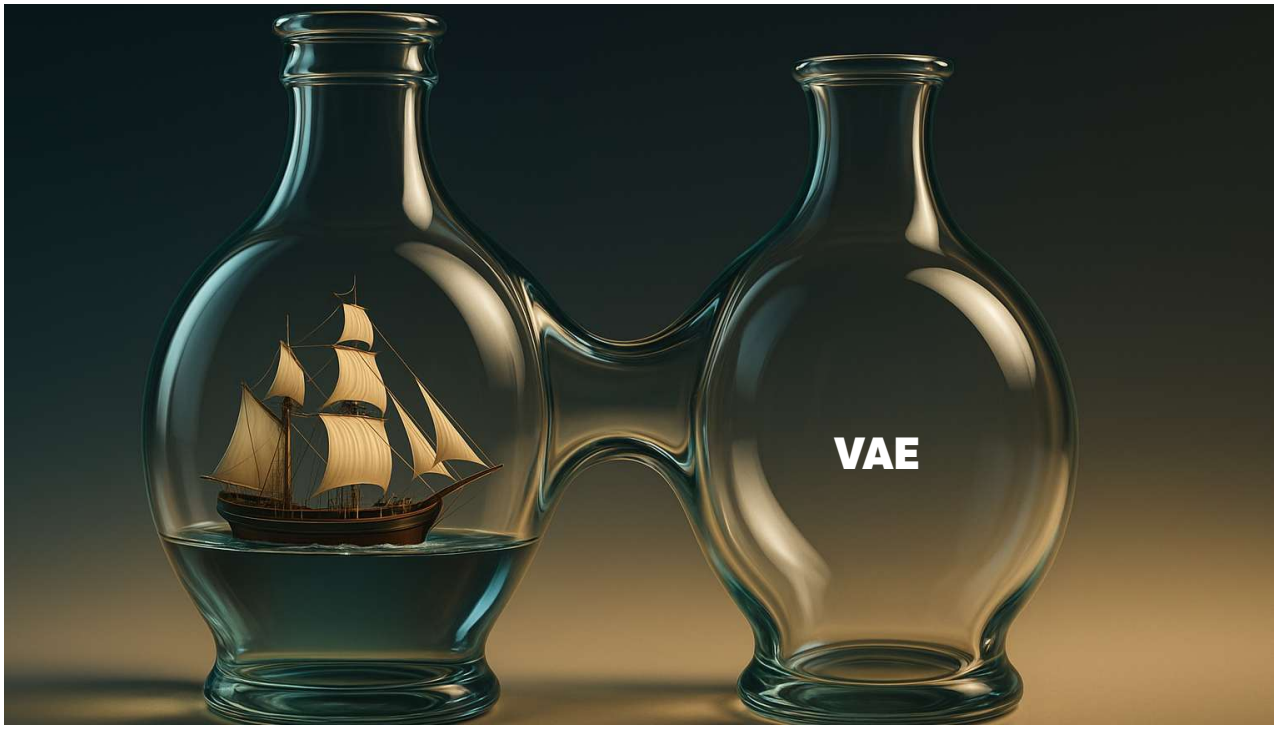


Genetic growth – getting more complex.

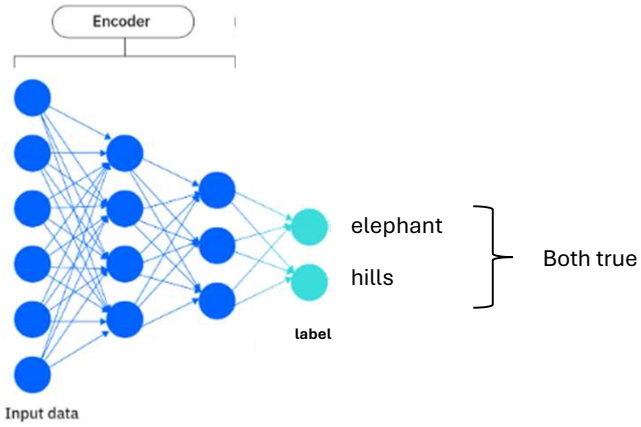

Degeneration – loosing cohesion,


Worms, 435

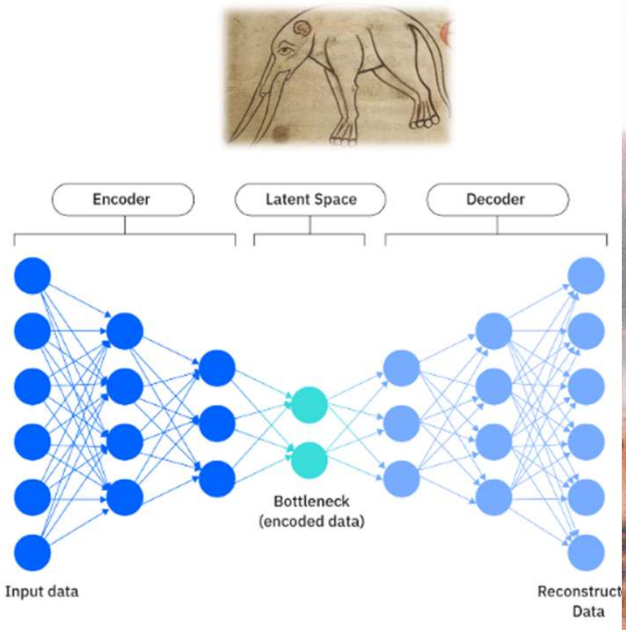

Big cauldron of myth

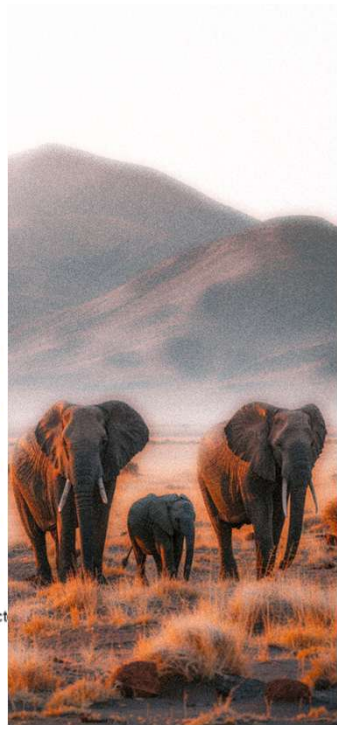

VAE

What is in this image?



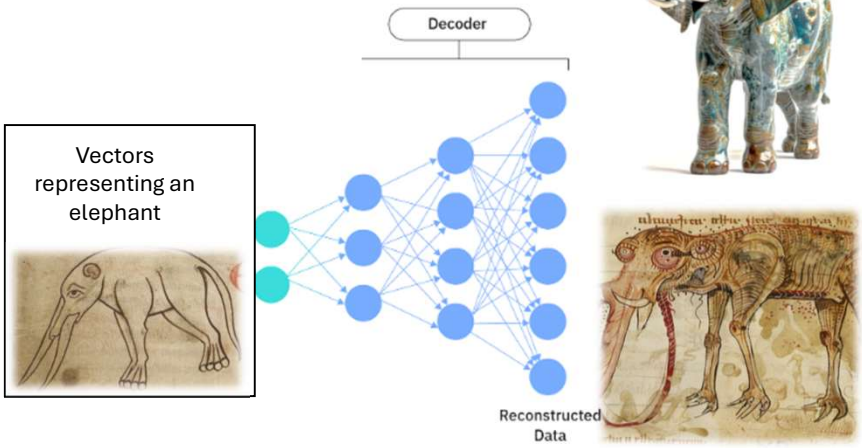
Labeling

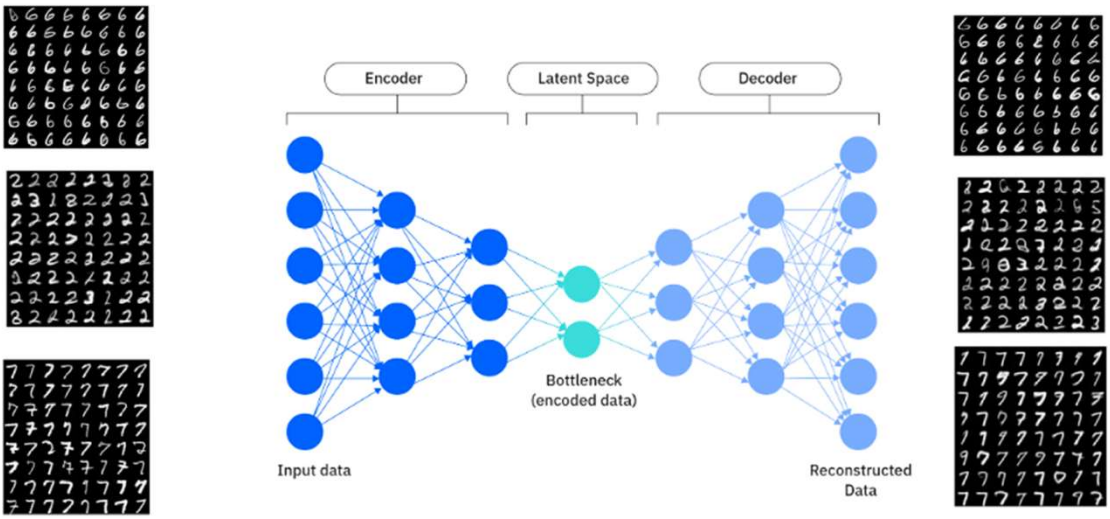




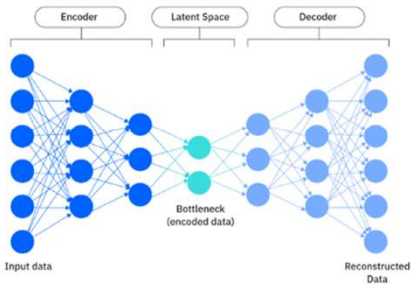
Variational Auto Encoder

Decoder

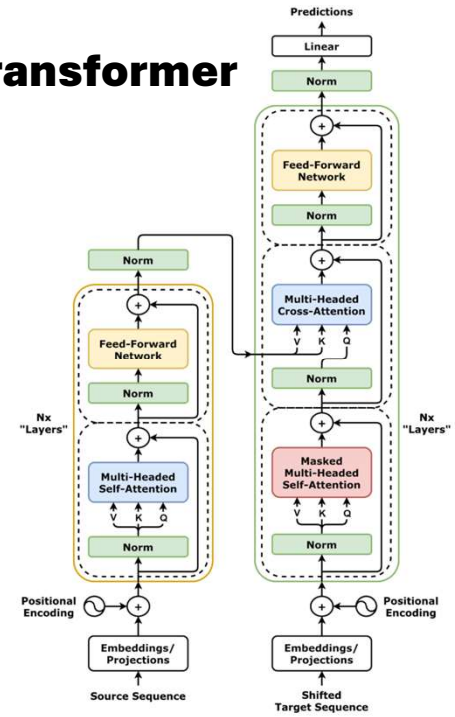

Vectors representing an elephant
Decoder
Reconstructed Data

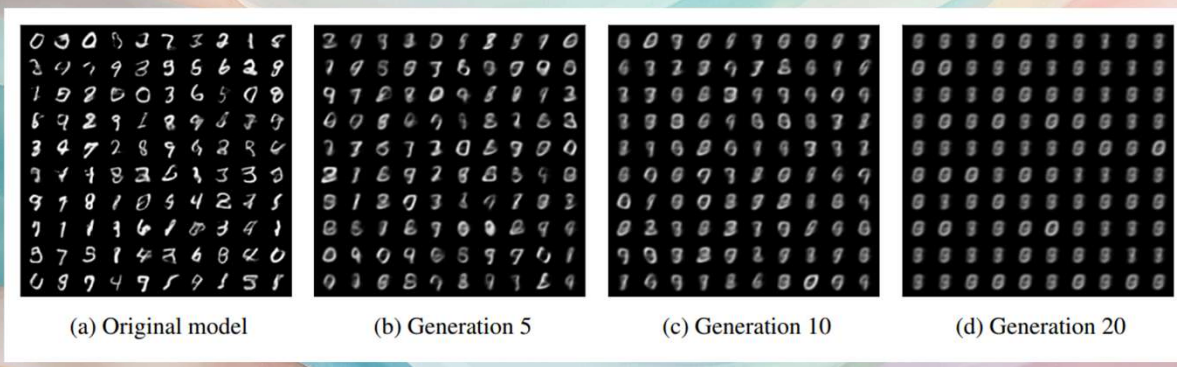

Encoder
Latent Space
Decoder
Bottleneck (encoded data)
Input data
Reconstructed Data

VAE

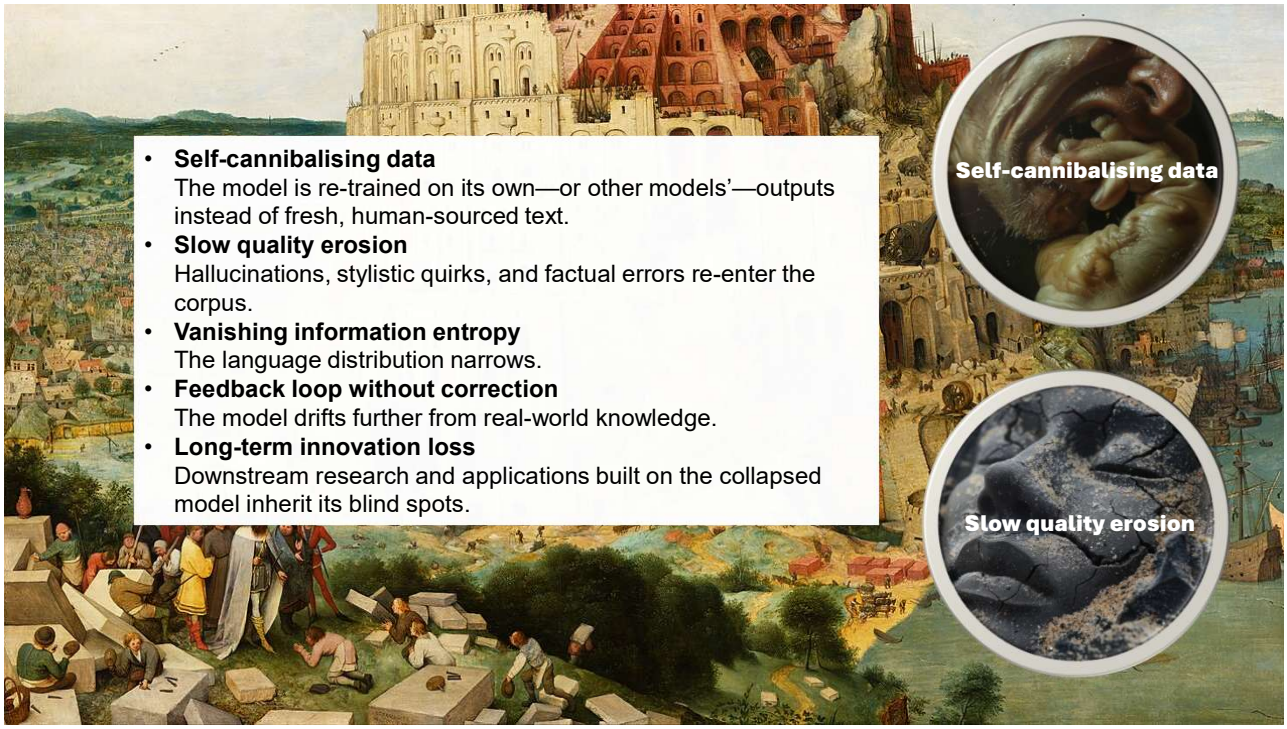


Transformer

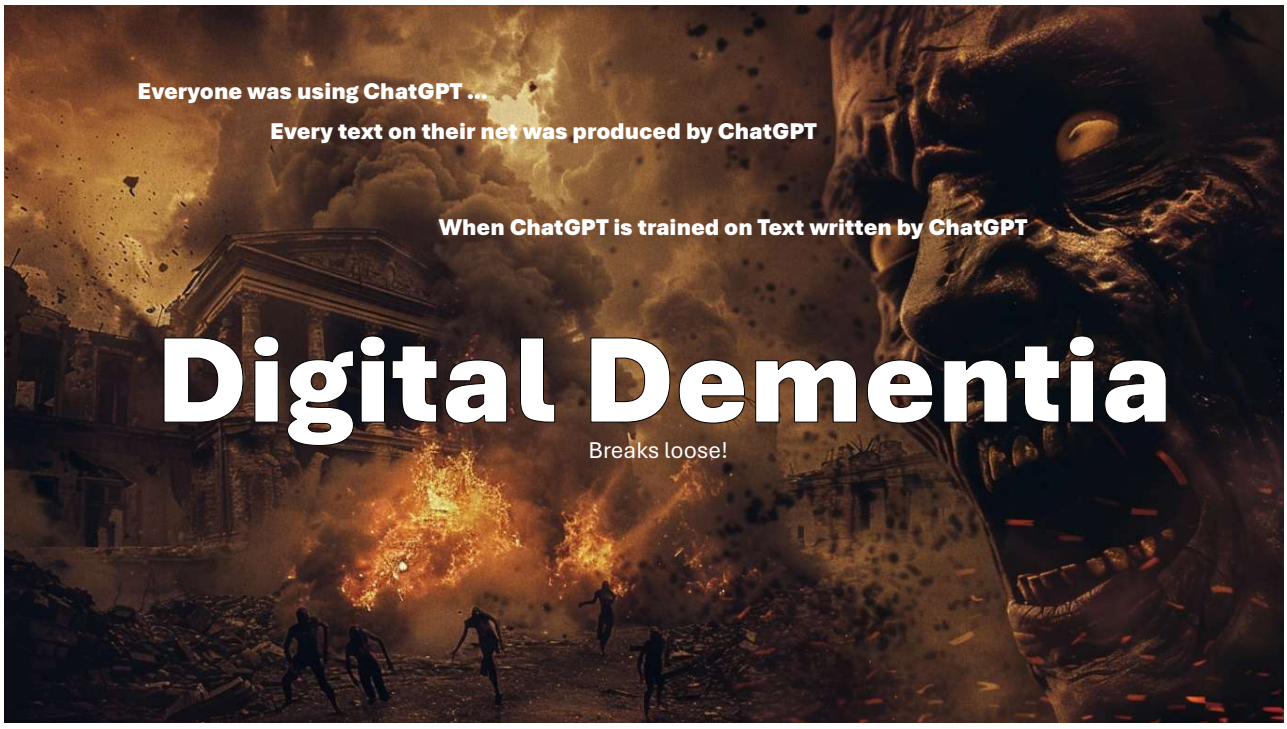




(a) Original model (b) Generation 5 (c) Generation 10 (d) Generation 20



- **Self-cannibalising data**
  The model is re-trained on its own—or other models'—outputs instead of fresh, human-sourced text.
- **Slow quality erosion**
  Hallucinations, stylistic quirks, and factual errors re-enter the corpus.
- **Vanishing information entropy**
  The language distribution narrows.
- **Feedback loop without correction**
  The model drifts further from real-world knowledge.
- **Long-term innovation loss**
  Downstream research and applications built on the collapsed model inherit its blind spots.




AI generated waste

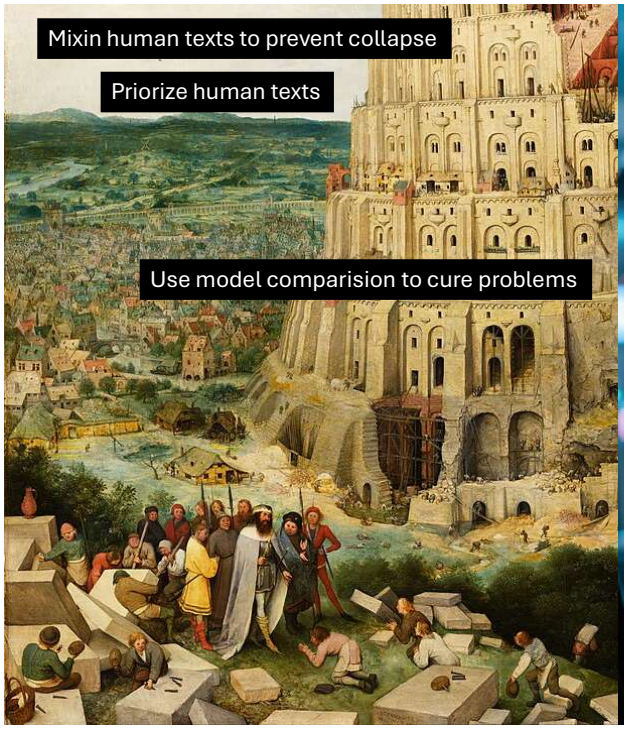

Found data:
• „All the internet and most of human literature“

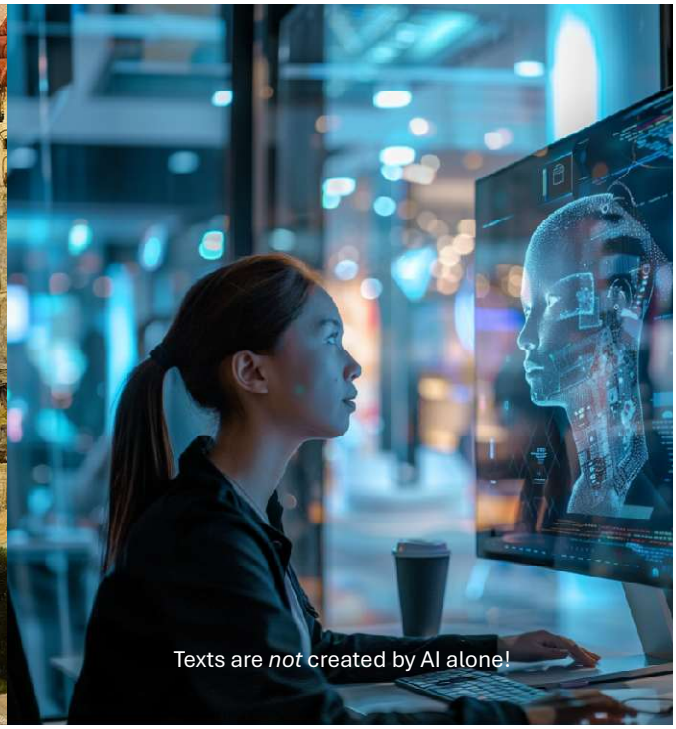
Found data:
- „All the internet and most of human literature“

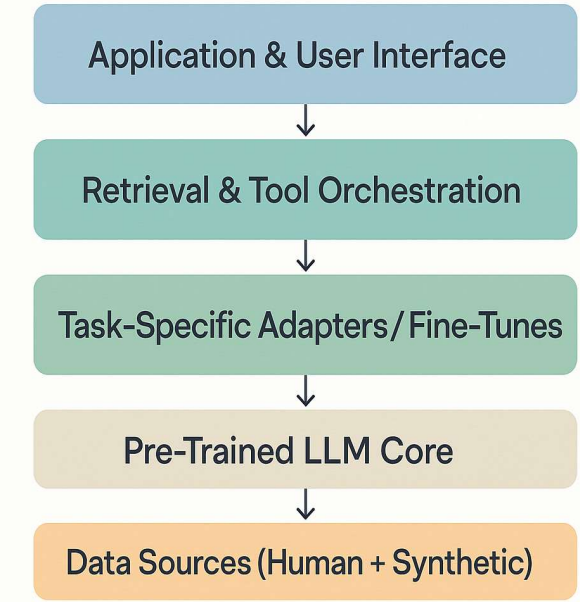
Designed data:
- Translations, Summaries, structured Data, Question-Answer-Pairs

Curated data:
- Data Samples chosen for optimal quality
- Rated, Labelled, Priorized, Categorised, Orchestrated

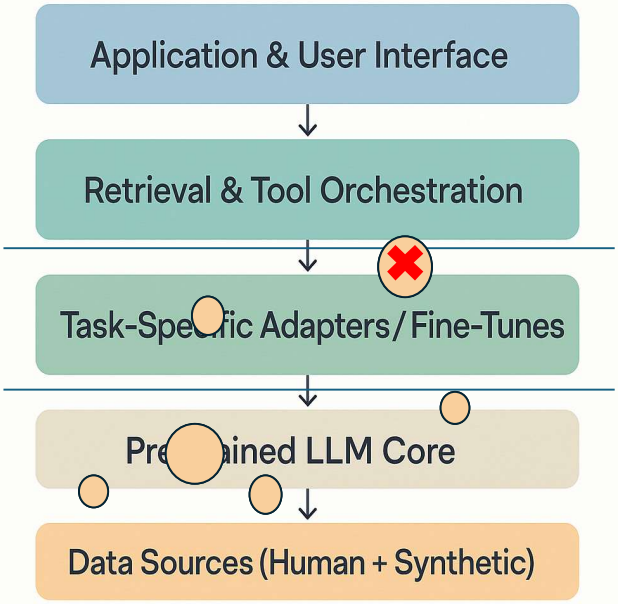
Augmented data:
- Filling blanks, adjusting proportions

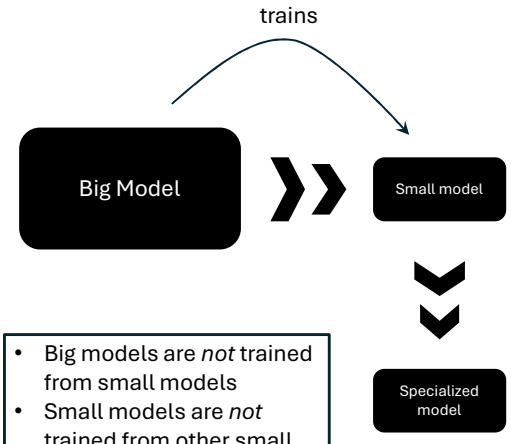





Application & User Interface
Retrieval & Tool Orchestration
Task-Specific Adapters / Fine-Tunes
Pre-Trained LLM Core
Data Sources (Human + Synthetic)

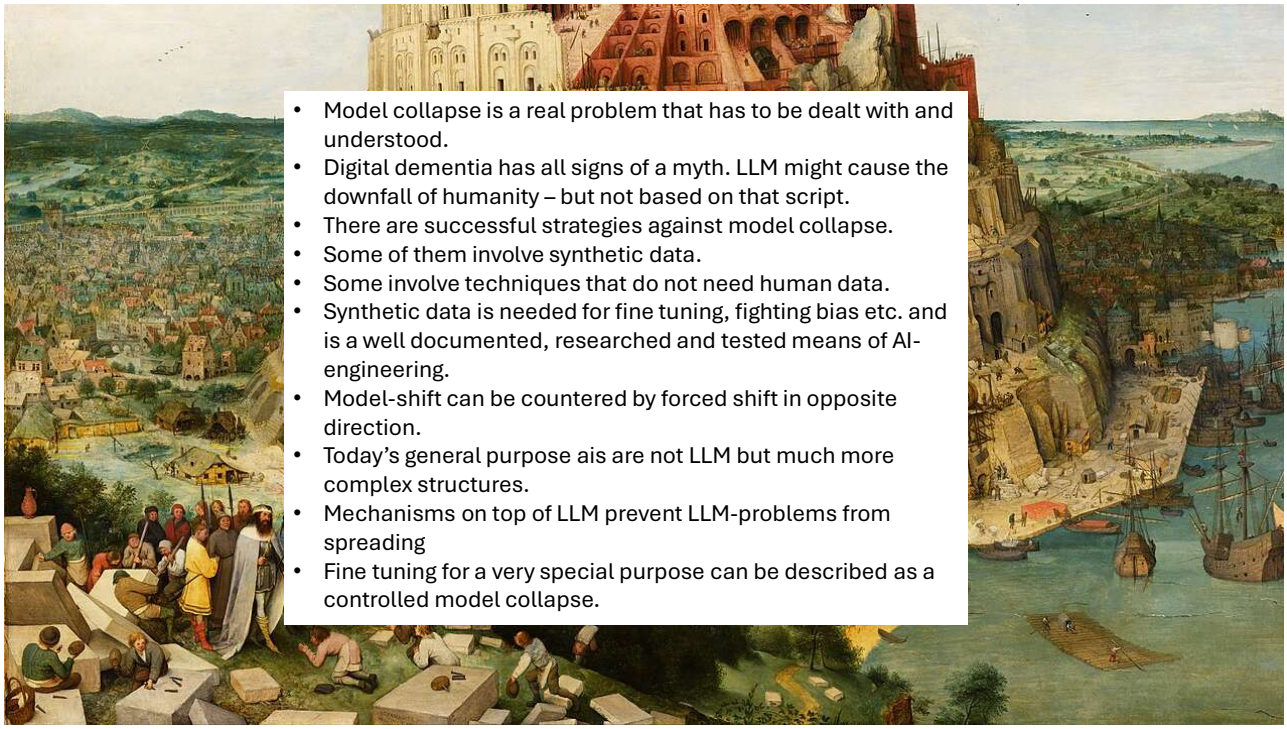

= Problems from the data bubbling up


Application & User Interface
Retrieval & Tool Orchestration
Task-Specific Adapters / Fine-Tunes
Pre-Trained LLM Core
Data Sources (Human + Synthetic)



- Big models are *not* trained from small models
- Small models are *not* trained from other small models
- Fine tuning data sets are *not* data sets for model training.

Special tasks:
- Narrowed down choices
- Clear cut decisions



- Model collapse is a real problem that has to be dealt with and understood.
- Digital dementia has all signs of a myth. LLM might cause the downfall of humanity – but not based on that script.
- There are successful strategies against model collapse.
- Some of them involve synthetic data.
- Some involve techniques that do not need human data.
- Synthetic data is needed for fine tuning, fighting bias etc. and is a well documented, researched and tested means of AI-engineering.
- Model-shift can be countered by forced shift in opposite direction.
- Today's general purpose ais are not LLM but much more complex structures.
- Mechanisms on top of LLM prevent LLM-problems from spreading
- Fine tuning for a very special purpose can be described as a controlled model collapse.