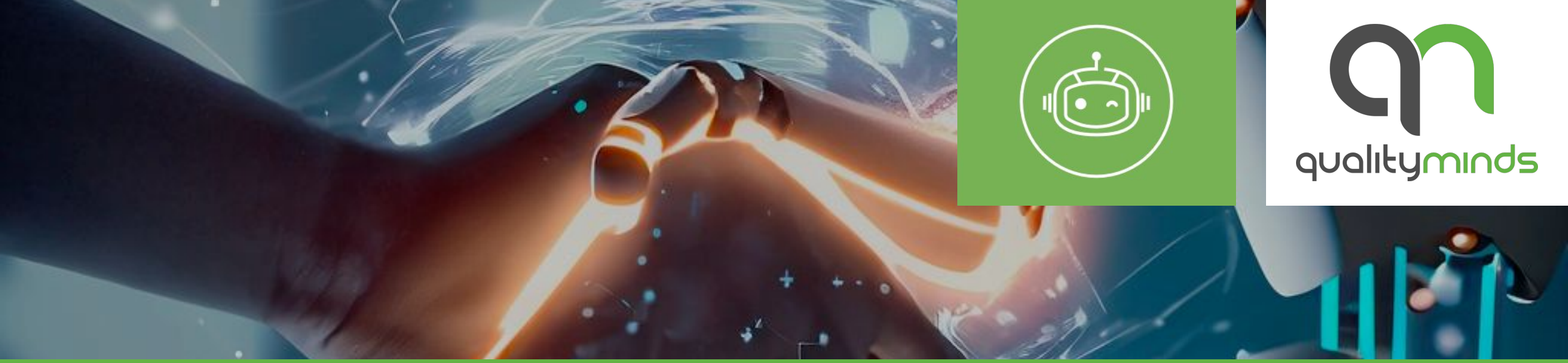


TESTING AI:
THE AI PERSPECTIVE

AGENDA

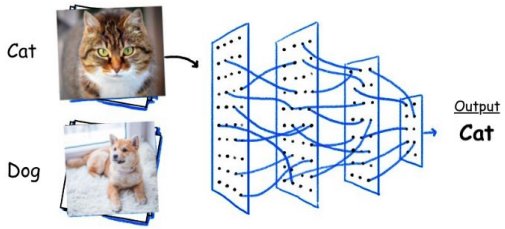
- Crash course Neural Networks, NLP, LLMs
- Testing
- Data!
- Risk Analysis
- Wrap-up

PROGRAMMING VS MACHING LEARNING

Machine Learning

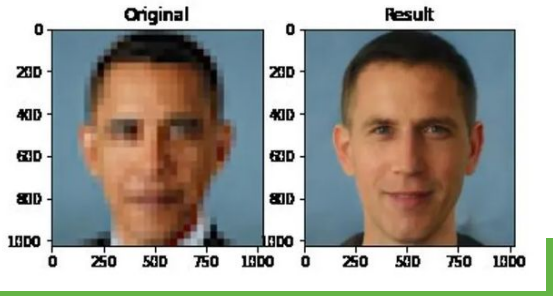
pro

Can do anything!
(general approximation theorem)



con

Can do anything!
(allignment



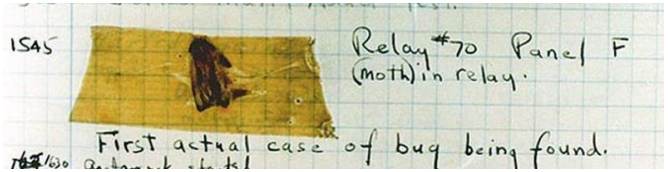
Software Development

Can do anything!
(computability)

```
sudoku(Rows) :-
        length(Rows, 9),
        maplist(same_length(Rows), Rows),
        append(Rows, Vs), Vs ins 1..9,
        maplist(all_distinct, Rows),
        transpose(Rows, Columns),
        maplist(all_distinct, Columns),
        Rows = [As,Bs,Cs,Ds,Es,Fs,Gs,Hs,Is],
        blocks(As, Bs, Cs),
        blocks(Ds, Es, Fs),
        blocks(Gs, Hs, Is).

blocks([], [], []).
blocks([N1,N2,N3|Ns1], [N4,N5,N6|Ns2], [N7,N8,N9|Ns3]) :-
        all_distinct([N1,N2,N3,N4,N5,N6,N7,N8,N9]),
        blocks(Ns1, Ns2, Ns3).
```

Can do anything!
(bugs)



GENERATIVE ML MODELS

Generative AI

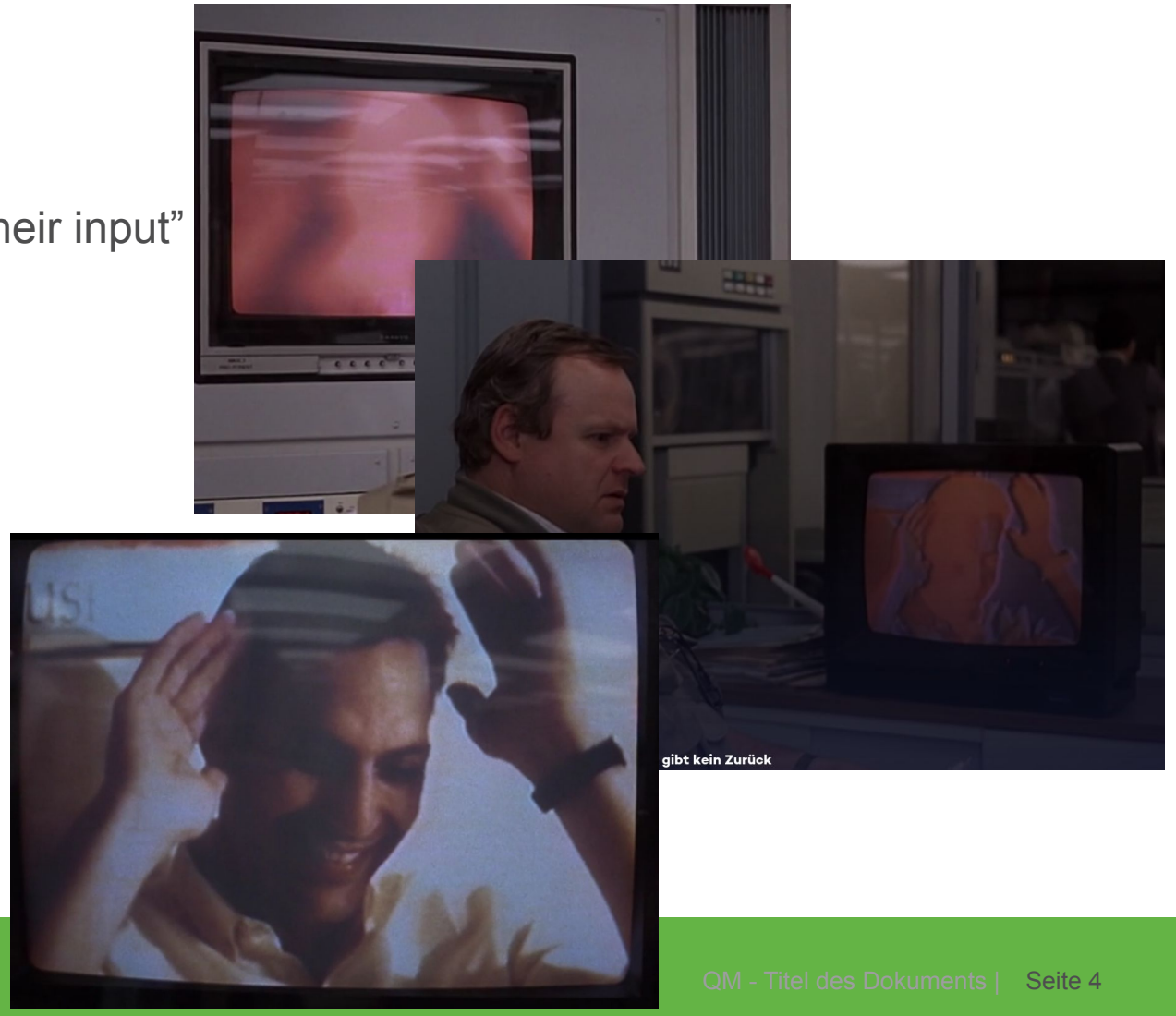
- ML models “that have deep knowledge about their input”
- These models can generate new data.

Regression and Generation

- Regression reduces the dimensionalty
- Generation increases dimensionality
- Generative models "seem" creative

Examples:

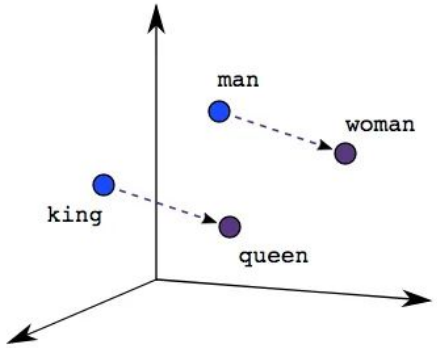
- Spam filter vs spam bot
- Celebrity detection vs deep fake
- Music genre detection vs music generation



LARGE LANGUAGE MODELS

Embeddings

- Map a word to a vector
- Encapsulates "ssssss"
- LLM: embed whole paragraphs

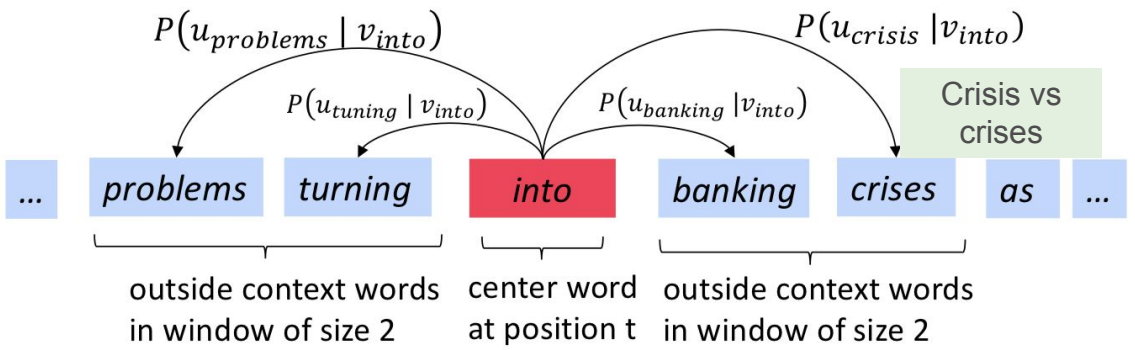


Text completion on steroids

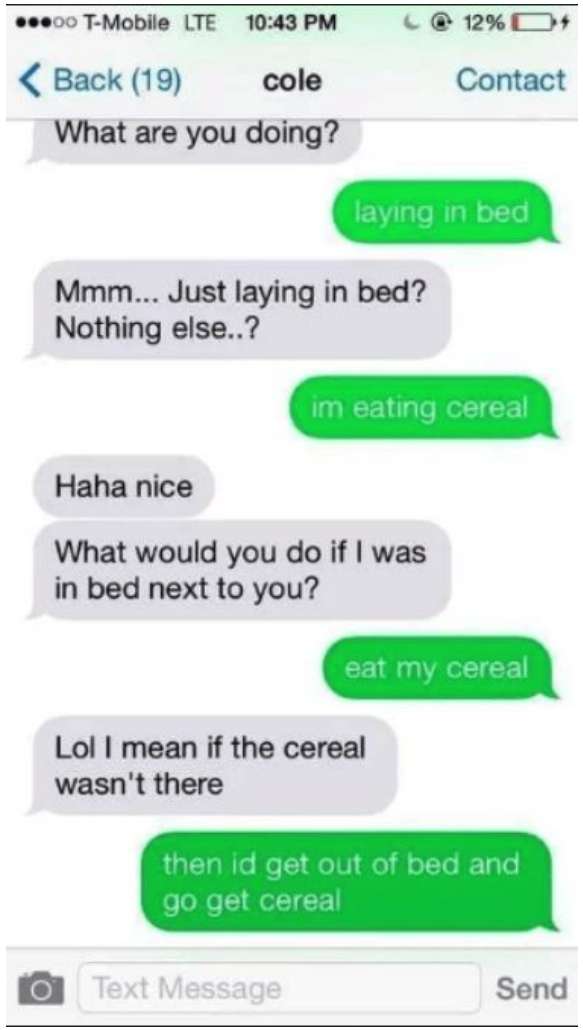
- Attention matrix: what relates to what?
- Gives a probability Distribution of words

You are the LLM now!

- A cow gives ...



PROMPTING



If alex was a chat bot (LLM), it would see:

LLM input

```
1 The following is a conversation between cole and alex.
2 While cole is making advances, alex just wants to eat
3 cereal.
4
5 cole: What are you doing?
6 alex: laying in bed
7 cole: Mmm... Just laying in bed? Nothing else.. ?
8 alex: im eating cereal
9 cole: Haha nice
10 cole: what would you di if I was in bed next to you?
11 alex: eat my cereal
12 cole: Lol I mean if the cereal wasn't there
13 alex:
```

Prompt

Conversation input

AGENDA

- Crash course of Neural Networks, NLP, LLMs
- Testing
- Data!
- Risk Analysis
- Wrap-up

The pyramids

The pyramids

ethics

Application

Algorithm

Infrastructure

Data

Non-self-Actualization

Basic needs

The pyramids

ethics

Application

Algorithm

Infrastructure

Data

Non-self-Actualization

Basic needs

The pyramids

ethics

Application

Algorithm

Infrastructure

Data

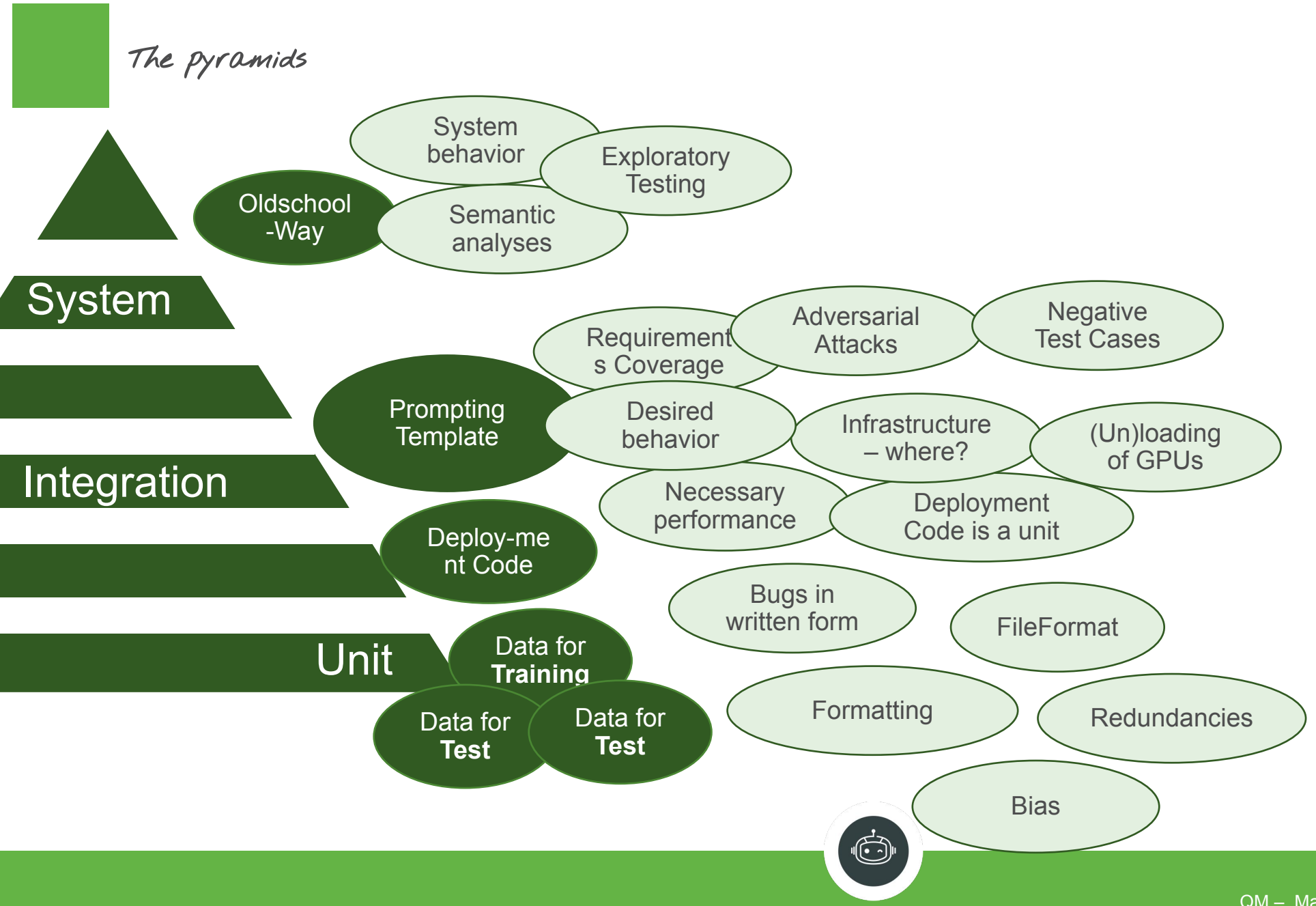
Non-self-Actualization

Basic needs

The pyramids

The pyramids

The pyramids



Testing von LLMs – Herausforderungen

There is not one ground truth to rule them all

- Context
- Use Case
- Target Audience
- …

Insufficient Metrics to assess semantic quality

- **BLEU**; statistical method based on the "word-error rate" / n-grams; originally for the evaluation of translations
- **BERTScore**; statistical method based on pairwise similarity;

https://aclanthology.org/P02-1040.pdf
https://arxiv.org/pdf/1904.09675.pdf

Prompt-Testing von LLMs – Probabilität als Ground Truth

- Wie geht Probabilität mit zugekauftem LLM à la OpenAI? --> Gar nicht, da kein Zugriff; braucht andere Näherungswerte z.B. über Testautomatisierung und wieder-Ausführung des gleichen Testfalls

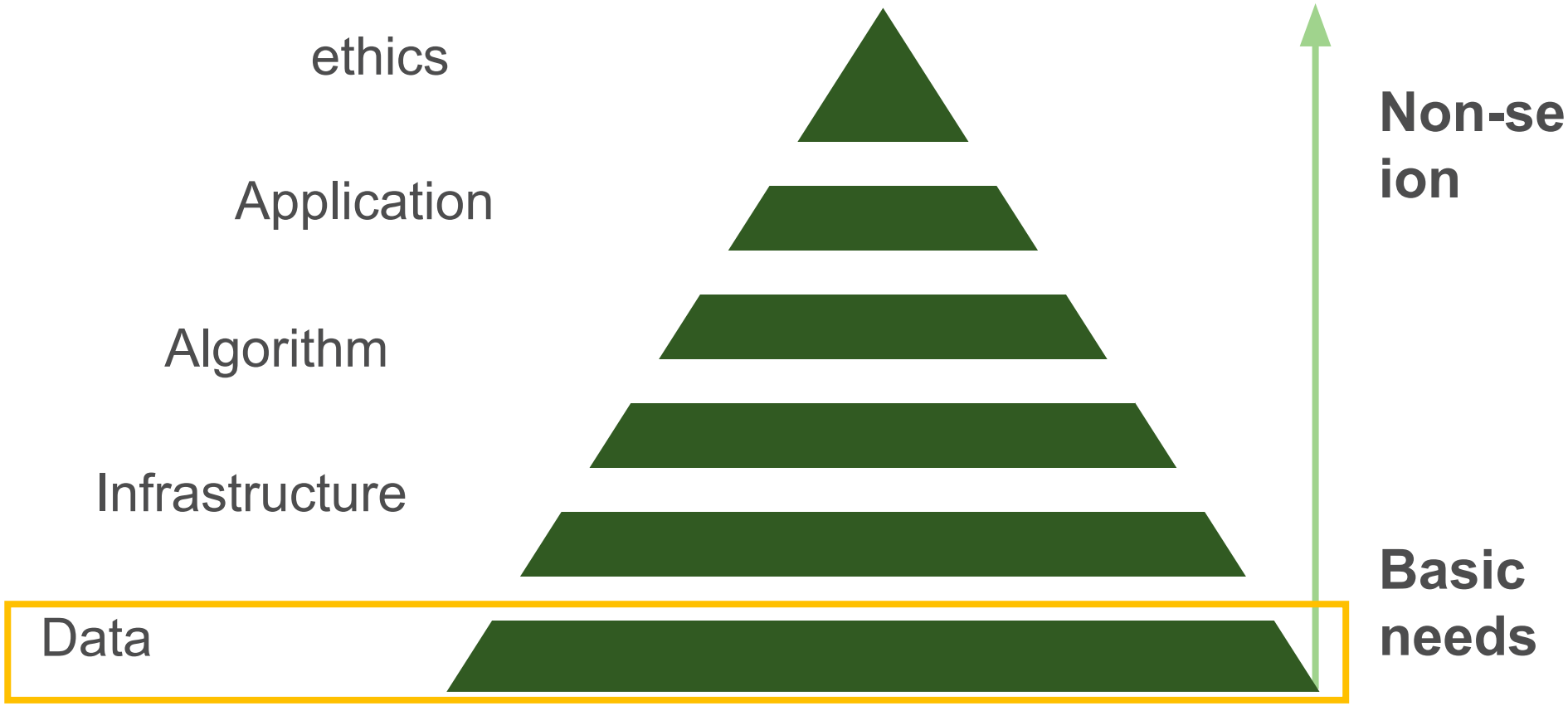
- Prompts so gestalten, dass sie möglichst einheitliche (oder: für euch passende) Befehle weitergibt
- Prompt Template als Teil der Qualitätsstrategie, s. Mural (Prompt Testing als eigene Slide

https://aclanthology.org/P02-1040.pdf
https://arxiv.org/pdf/1904.09675.pdf

AGENDA

- Crash course of Neural Networks, NLP, LLMs
- Testing
- Data!
- Risk Analysis
- Wrap-up

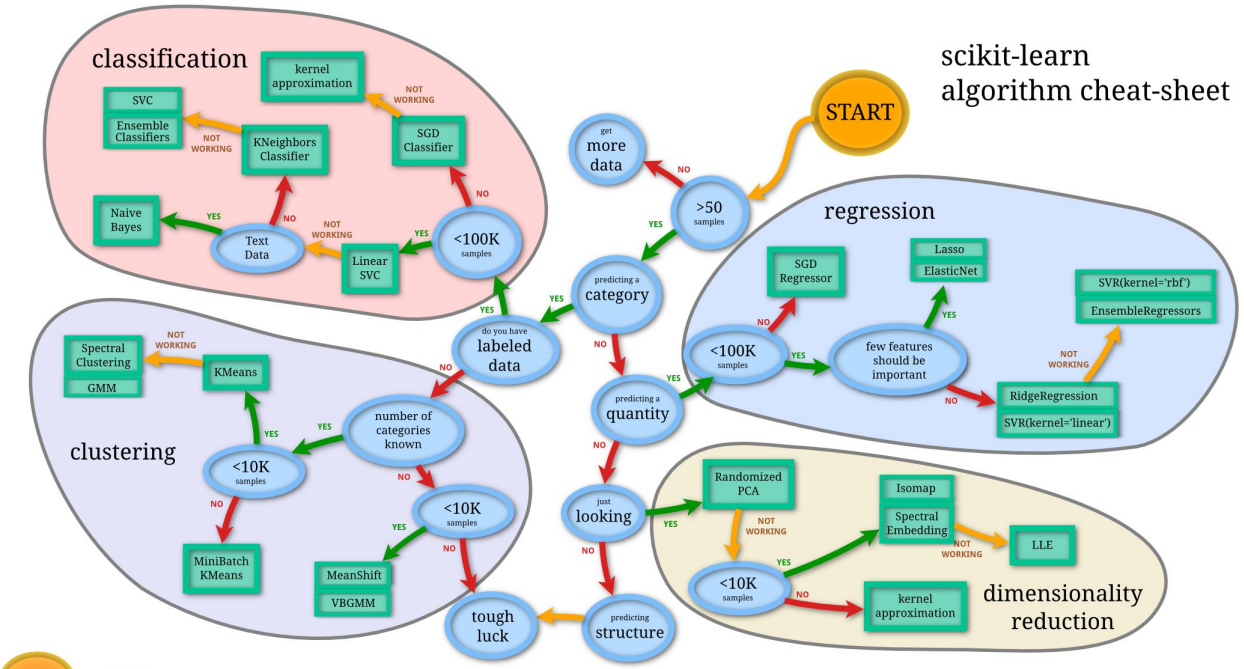
Testing von LLMs – eine neue Testpyramide



What we talk about now

HUNGER FOR DATA VS DATA QUALITY

Hunger for Data



Back

Data Quality

Training a Learning Assitant

- Collect custom training data
- Data contains typo: „mach success“
- LLM reproduces that typo consistently
- LLM invents signatures

"….

Lots of love and many greetings, Your friend in learning,

Lena"

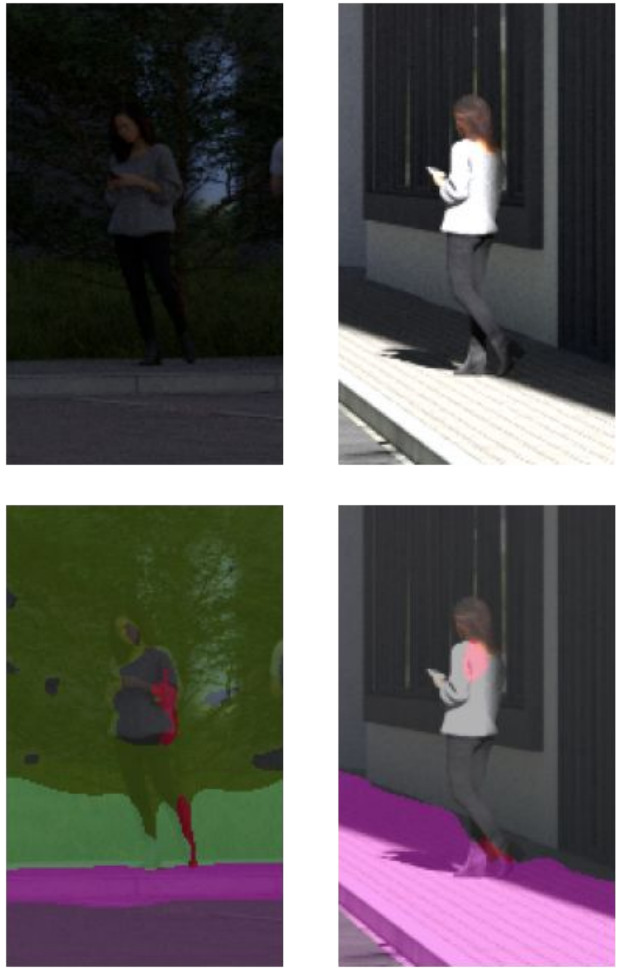
SHIFT LEFT AND DATA PRODUCTION

KIAB (AI-Assurance)

- ensure the safety of AI models for autonomous driving
- produce test data
- requirement -> data production process

Data Deliveries

- 8 tranches with ~1.5TB each
- Data quality: what is a bug?



SHIFT LEFT AND DATA PRODUCTION

KIAB (AI-Assurance)

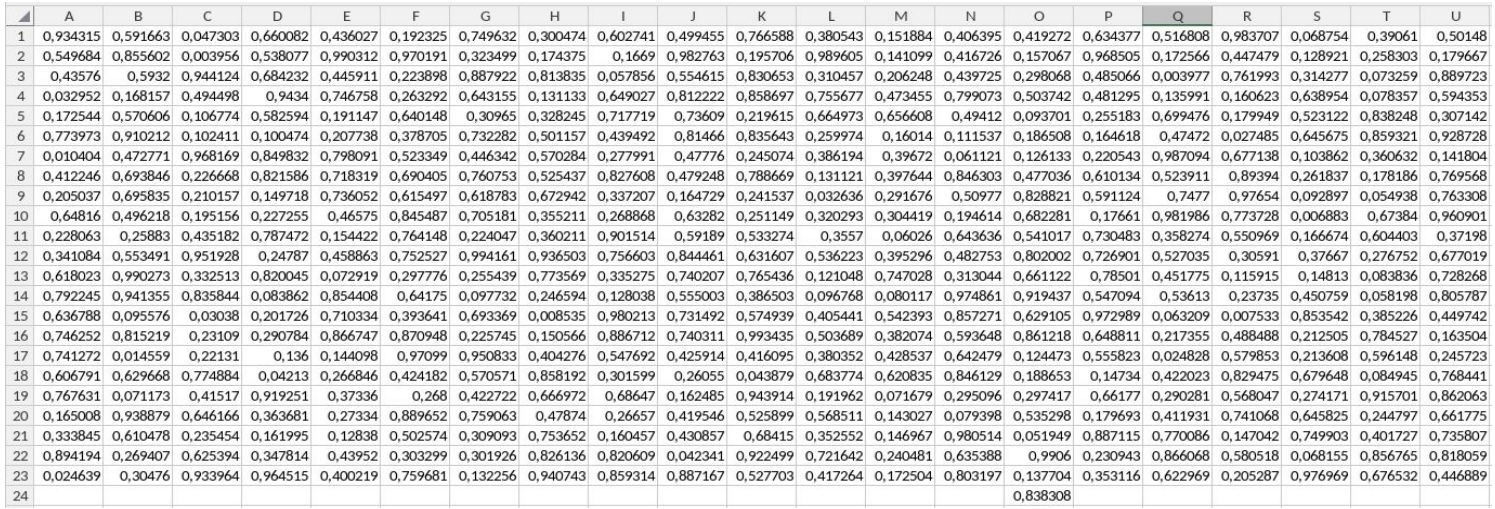
- ensure the safety of AI models for autonomous driving
- produce test data
- requirement -> data production process

Data Deliveries

- 8 tranches with ~1.5TB each
- Data quality: what is a bug?

Shift Left

- **Small** data deliveries
- Set up checks **with the producer**
- **Quick validation** cycles

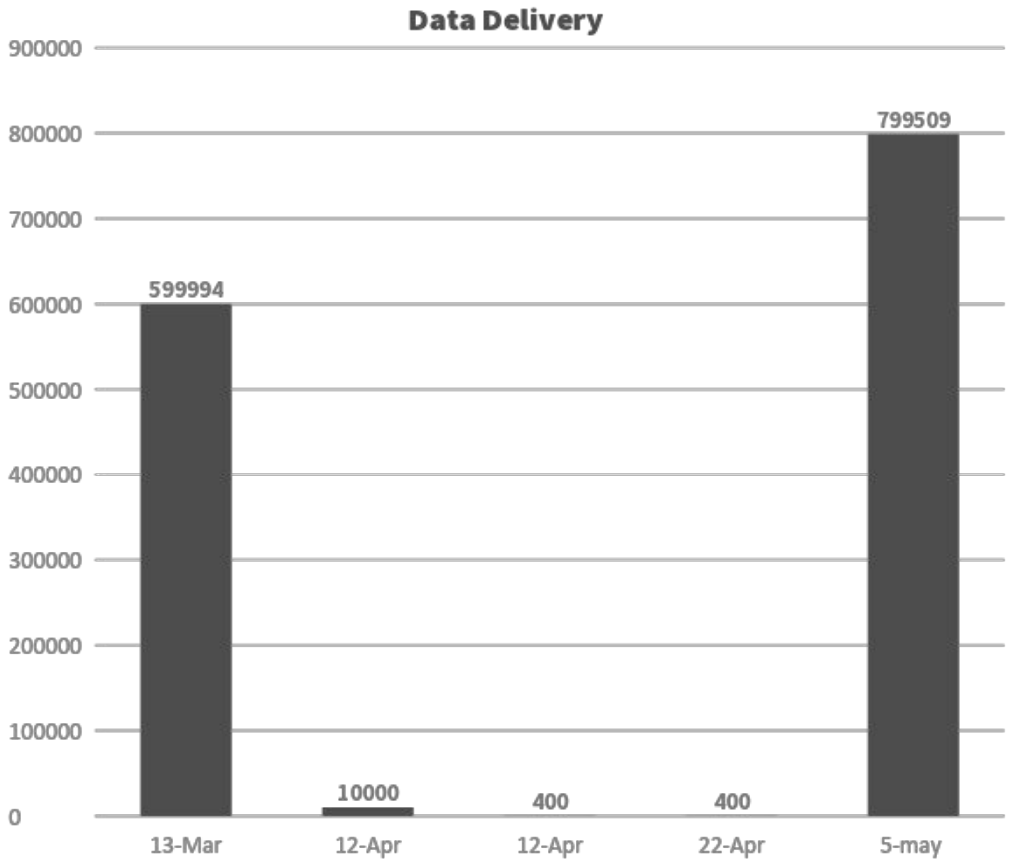


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	0,934315	0,591663	0,047303	0,660082	0,436027	0,192325	0,749632	0,300474	0,602741	0,499455	0,766588	0,380543	0,151884	0,406395	0,419272	0,634377	0,516808	0,983707	0,068754	0,39061	0,50148
2	0,549684	0,855602	0,003956	0,538077	0,990312	0,970191	0,323499	0,174375	0,1669	0,982763	0,195706	0,989605	0,141099	0,416726	0,157067	0,968505	0,172566	0,447479	0,128921	0,258303	0,179667
3	0,43576	0,5932	0,944124	0,684232	0,445911	0,223898	0,887922	0,813835	0,057856	0,554615	0,830653	0,310457	0,206248	0,439725	0,298068	0,485066	0,003977	0,761993	0,314277	0,073259	0,889723
4	0,032952	0,168157	0,494498	0,9434	0,746758	0,263292	0,643155	0,131133	0,649027	0,812222	0,858697	0,755677	0,473455	0,799073	0,503742	0,481295	0,135991	0,160623	0,638954	0,078357	0,594353
5	0,172544	0,570606	0,106774	0,582594	0,191147	0,640148	0,30965	0,328245	0,717719	0,73609	0,219615	0,664973	0,656608	0,49412	0,093701	0,255183	0,699476	0,179949	0,523122	0,838248	0,307142
6	0,773973	0,910212	0,102411	0,100474	0,207738	0,378705	0,732282	0,501157	0,439492	0,81466	0,835643	0,259974	0,16014	0,111537	0,186508	0,164618	0,47472	0,027485	0,645675	0,859321	0,928728
7	0,010404	0,472771	0,968169	0,849832	0,798091	0,523349	0,446342	0,570284	0,277991	0,47776	0,245074	0,386194	0,39672	0,061121	0,126133	0,220543	0,987094	0,677138	0,103862	0,360632	0,141804
8	0,412246	0,693846	0,226668	0,821586	0,718319	0,690405	0,760753	0,525437	0,827608	0,479248	0,788669	0,131121	0,397644	0,846303	0,477036	0,610134	0,523911	0,89394	0,261837	0,178186	0,769568
9	0,205037	0,695835	0,210157	0,149718	0,736052	0,615497	0,618783	0,672942	0,337207	0,164729	0,241537	0,032636	0,291676	0,50977	0,828821	0,591124	0,7477	0,97654	0,092897	0,054938	0,763308
10	0,64816	0,496218	0,195156	0,227255	0,46575	0,845487	0,705181	0,355211	0,268868	0,63282	0,251149	0,320293	0,304419	0,194614	0,682281	0,17661	0,981986	0,773728	0,006883	0,67384	0,960901
11	0,228063	0,25883	0,435182	0,787472	0,154422	0,764148	0,224047	0,360211	0,901514	0,59189	0,533274	0,3557	0,06026	0,643636	0,541017	0,730483	0,358274	0,550969	0,166674	0,604403	0,37198
12	0,341084	0,553491	0,951928	0,24787	0,458863	0,752527	0,994161	0,936503	0,756603	0,844461	0,631607	0,536223	0,395296	0,482753	0,802002	0,726901	0,527035	0,30591	0,37667	0,276752	0,677019
13	0,618023	0,990273	0,332513	0,820045	0,072919	0,297776	0,255439	0,773569	0,335275	0,740207	0,765436	0,121048	0,747028	0,313044	0,661122	0,78501	0,451775	0,115915	0,14813	0,083836	0,728268
14	0,792245	0,941355	0,835844	0,083862	0,854408	0,64175	0,097732	0,246594	0,128038	0,555003	0,386503	0,096768	0,080117	0,974861	0,919437	0,547094	0,53613	0,23735	0,450759	0,058198	0,805787
15	0,636788	0,095576	0,03038	0,201726	0,710334	0,393641	0,693369	0,008535	0,980213	0,731492	0,574939	0,405441	0,542393	0,857271	0,629105	0,972989	0,063209	0,007533	0,853542	0,385226	0,449742
16	0,746252	0,815219	0,23109	0,290784	0,866747	0,870948	0,225745	0,150566	0,886712	0,740311	0,993435	0,503689	0,382074	0,593648	0,861218	0,648811	0,217355	0,488488	0,212505	0,784527	0,163504
17	0,741272	0,014559	0,22131	0,136	0,144098	0,97099	0,950833	0,404276	0,547692	0,425914	0,416095	0,380352	0,428537	0,642479	0,124473	0,555823	0,024828	0,579853	0,213608	0,596148	0,245723
18	0,606791	0,629668	0,774884	0,04213	0,266846	0,424182	0,570571	0,858192	0,301599	0,26055	0,043879	0,683774	0,620835	0,846129	0,188653	0,14734	0,422023	0,829475	0,679648	0,084945	0,768441
19	0,767631	0,071173	0,41517	0,919251	0,37336	0,268	0,422722	0,666972	0,68647	0,162485	0,943914	0,191962	0,071679	0,295096	0,297417	0,66177	0,290281	0,568047	0,274171	0,915701	0,862063
20	0,165008	0,938879	0,646166	0,363681	0,27334	0,889652	0,759063	0,47874	0,26657	0,419546	0,525899	0,568511	0,143027	0,079398	0,535298	0,179693	0,411931	0,741068	0,645825	0,244797	0,661775
21	0,333845	0,610478	0,235454	0,161995	0,12838	0,502574	0,309093	0,753652	0,160457	0,430857	0,68415	0,352552	0,146967	0,980514	0,051949	0,887115	0,770086	0,147042	0,749903	0,401727	0,735807
22	0,894194	0,269407	0,625394	0,347814	0,43952	0,303299	0,301926	0,826136	0,820609	0,042341	0,922499	0,721642	0,240481	0,635388	0,9906	0,230943	0,866068	0,580518	0,068155	0,856765	0,818059
23	0,024639	0,30476	0,933964	0,964515	0,400219	0,759681	0,132256	0,940743	0,859314	0,887167	0,527703	0,417264	0,172504	0,803197	0,137704	0,353116	0,622969	0,205287	0,976969	0,676532	0,446889
24															0,838308						

SHIFT LEFT IN AN INTERNAL PROJECT

A colleague asks
- „can you quickly produce some data“
- „simple physics simulation“

Reply: „Sure thing! Will produce 600,000 data points over night.“



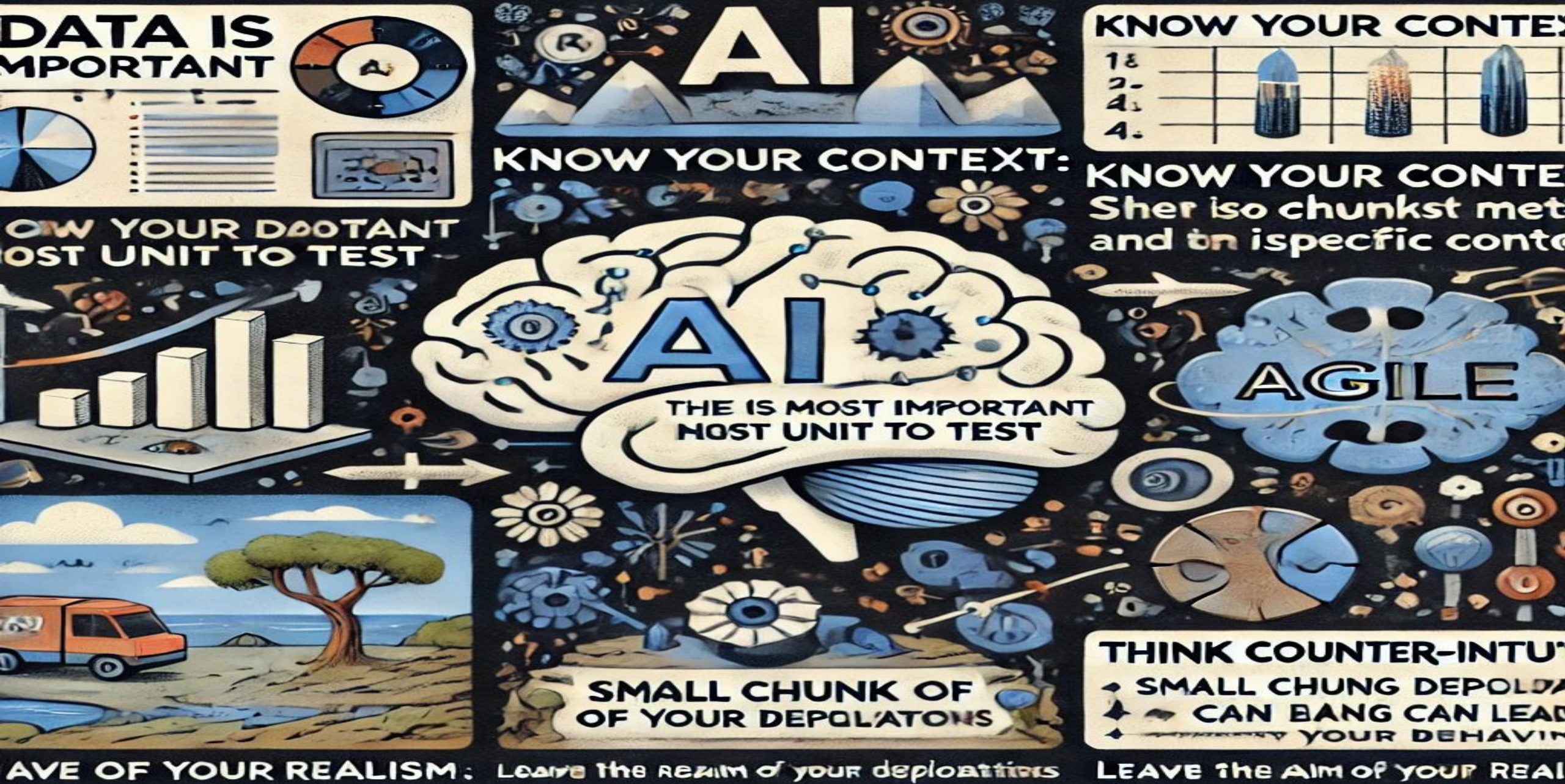
AGENDA

- Crash course of Neural Networks, NLP, LLMs
- Testing
- Data!
- Risk Analysis
- Wrap-up

AGENDA

- Crash course of Neural Networks, NLP, LLMs
- Testing
- Data!
- Risk Analysis
- Wrap-up

- Data is important: Your data is the first and most important Unit to test
- Know your context: There is no good metric for specific contexts
- Work agile: Small chunks of data is easily tested, big bang deployments can lead to irreparable behavior
- Think counter-intuitive: Leave the realm of your realism


DATA IS
MPORTANT
AI
KNOW YOUR CONTEX
KNOW YOUR CONTEXT:
KNOW YOUR CONTE
Sher iso chunkst met
and tn ispecfic conte
OW YOUR DOOTANT
OST UNIT TO TEST
AI
THE IS MOST IMPORTANT
NOST UNIT TO TEST
AGILE
SMALL CHUNK OF
OF YOUR DEPOLATONS
THINK COUNTER-INTU
SMALL CHUNG DEPOL
CAN BANG CAN LEA
YOUR DEHAVI
AVE OF YOUR REALISM:
Leave the realm of your deploattions
LEAVE the AIm of YOUr REAL

STAY IN TOUCH

Dr. Niels Heller
niels.heller@qualityminds.de

Senior Data Scientist/ AI Tech Lead

Focus: NLP, LLMs, AI Testing and AI Strategy & Consulting

Bastian Knerr
bastian.knerr@qualityminds.de

Team Lead Testing

Focus: Testprocesses, Testmanagement