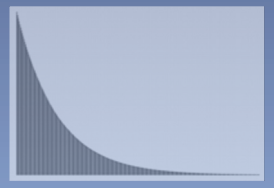


Test & Evaluation of AI Systems

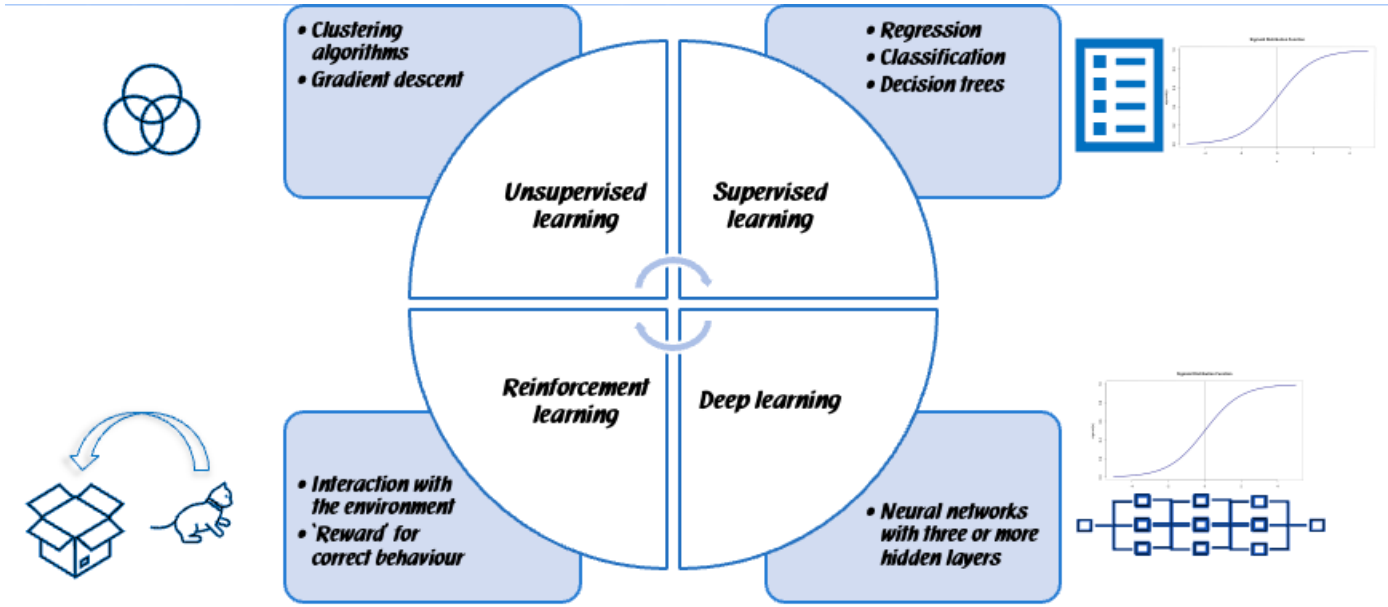
Webinar Series Process Fellows

Overview

- 1. Overview & Introduction
- 2. A governance view on V&V
- 3. Common Misconceptions
- 4. The Engineer`s View on V&V
- 5. Summary

Technologies and Types of ML Systems

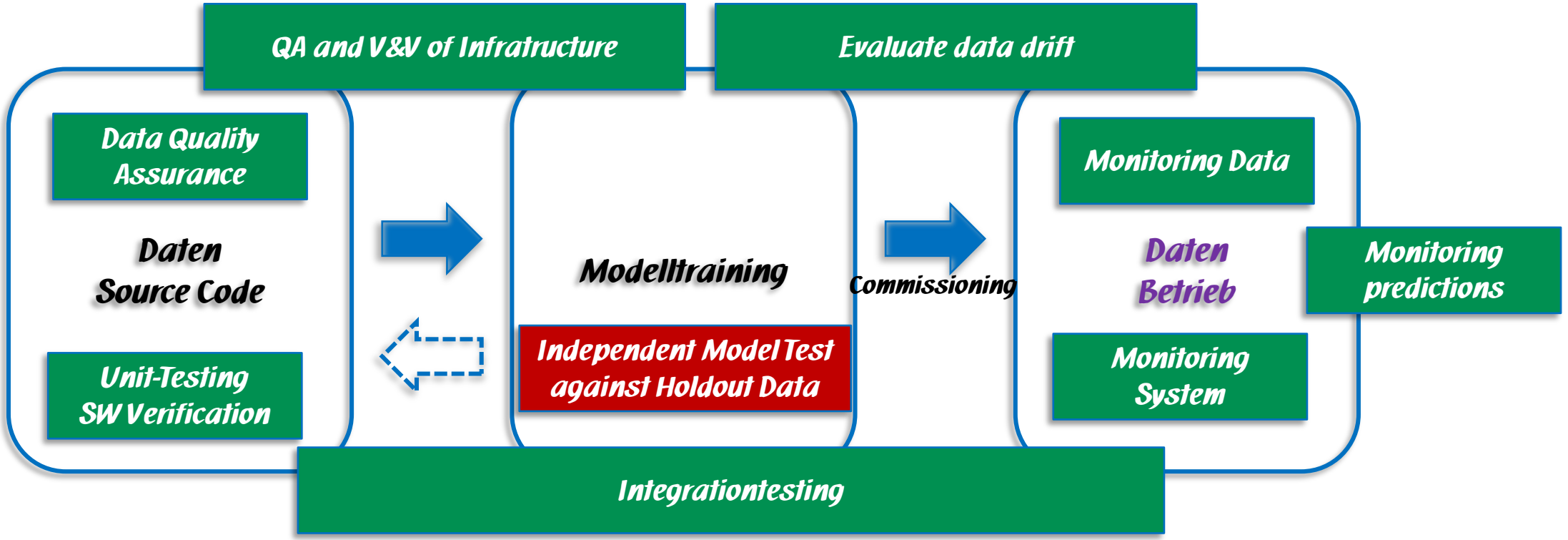
- **Large Language Models**
  - e.g. ChatGPT
- **Multisensory AI**
- **Neural networks**
- **Neuromorphic computing**
- **Evolutionary algorithms**
- **AI-generated Software**



Challenges for V&V

- Mathematical / Technical / Data-related Complexity
- Governance & Process Complexity
- Psychological Factors like Automation Bias (Machines are objective and free from bias).
- Misconceptions of how AI systems work
- Fallacy: AI research concepts == Systems Engineering concepts
- AI changes the Data from which it learns (e.g. pollution by hallucination and fraud)
- Environmental and Social Impact
- Security with genAI is a Nightmare - Multitudes of new and difficult to control attack vectors

A general Overview on how an AI System is tested



after Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, D. Sculley Google, Inc., „The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction; https://research.google/pubs/the-ml-test-score-a-rubric-for-ml-production-readiness-and-technical-debt-reduction/

A Lifecycle View on AI Verification & Validation

The Governance Perspective

Business Concept	Data Governance	Create AI Model	Model Verification and Validation	Deployment	Operation & Monitoring
• Plan & Design Project Phases	• Collect and understand Data • Set up Data governance • Training Data, Test Data, Usage Data	• Develop, fine-tune or customize model	• Verification of Model and Methods • Validation of Model and Methods	• System Integration • Manage organizational change	• Operate system according to intended use

adapted from NIST-AI 100-1

• Risk Assessment • Plan V & V • Audit plans	• Assess Data Quality • Implement Data Quality Assurance and Data Management Process Control • Supplier Controls	• Assure process quality, application of good practices, including proper model testing & evaluation	• Assure independent Model Verification and Validation • Supplier Controls	• System Integration and Acceptance Testing (e.g. Regulatory Compliance, User acceptance) • Supplier Controls (e.g. Cloud)	• Apply Statistical Process Control (monitor PEAS-performance) • Apply RCA • Monitor impacts • Perform Audits

V&V relevant activities

The Use-scenario Perspective

Use scenario 1: Built your own AI - everything from the preceding slide applies

Use scenario 2: Adapt an existing AI - everything from the preceding slide applies - with probably less complexity

- e.g. Data Governance for your own training and test data
- test against your specified requirements and performance measures

Use scenario 3: Use an existing AI system with specific agent files and context data

- No ‚Create AI Model'
- Model Verification and Validation still required

Use scenario 4: Use an existing AI system out of the box without any adaptions

- No ‚Create AI Model'
- Model Verification and Validation still required with lesser extent as in scenario 3

Some Common Misconceptions on GenAI

Simple Summary of what a Generative AI does

- A probabilistic pattern matcher, not a truth engine
  - GenAI maps relationships between concepts.
  - It doesn't "know" your business; it recognizes patterns that look like your business.
- Conceptual "Cross-Wiring":
  The AI can easily confuse two different concepts that „feel" similar mathematically. It might just confidently follow the wrong conceptual map.
- The Reliability Gap:
  The system optimizes for plausibility (matching a pattern) not accuracy (matching reality)
  That`s only one reason, why you need the expert human in the loop
- Result: It is a high-risk asset. It requires a human-led verification layer to ensure the pattern it found is actually the fact you need.

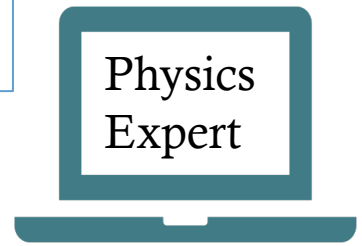
Example (using deepseek-r1:8b)

Note: Due to probabilistic nature of LLMs you may observe different behaviour. Restart and try the LLM several times.

deepseek-r1:8b

>You are an **expert for probability theory** and stochastic processes. You will answer any question based on your knowledge of probability theory and statistics.
…
> Give a short explanation of Bernoullis law.

Okay, let's clarify Bernoulli's Law. The term "Bernoulli's Law" isn't the standard name for the fundamental probability concept. **The core idea**
is actually described by the **Law of Large Numbers.**
…



deepseek-r1:8b

>You are an **expert for fluid dynamics.** You will answer any question based on your knowledge of fluid dynamics.
…
> Give a short explanation of Bernoullis law.

Bernoulli's Principle **describes the relationship between the speed of a moving fluid (like a liquid or gas) and its pressure.**

…

Common Misconceptions & Biases

- The genAI is good at X , so it must be good at Y. (Halo-effect)
  Big Tech built it, so it must work for me. (also Halo-effect)
  -> be careful with non-fact-based belief-systems

- „This sounds professional and is easy to read!"
  -> Processing Fluency of a text is why you`ll always find the next April Fool ...
  -> LLMs are really good at it

- Automation bias (Trust in the machine)

- Halucinations are removable errors
  -> No - they are byproduct on how LLMs and generative AIs work.

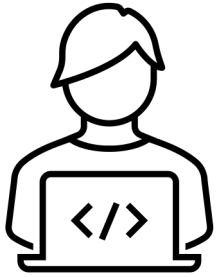
- Verification in the traditional sense is no longer possible, as no clear target result can be defined
  -> See following slides
  -> btw. - if you couldn`t verify results, you couldn`t train the models ...

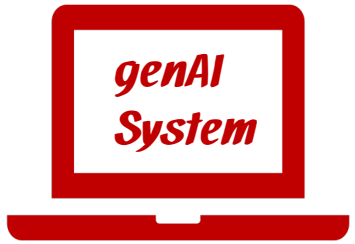
The Engineering View on V&V

The Software Testing Mindset (simplified)

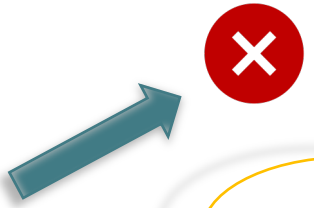
- *low-dimensioned input-output data relations*
- *defined and easy to understand complexity*
- *deterministic behaviour*

- *test against well defined acceptance criteria*
- *run the test cases once, verify the result*

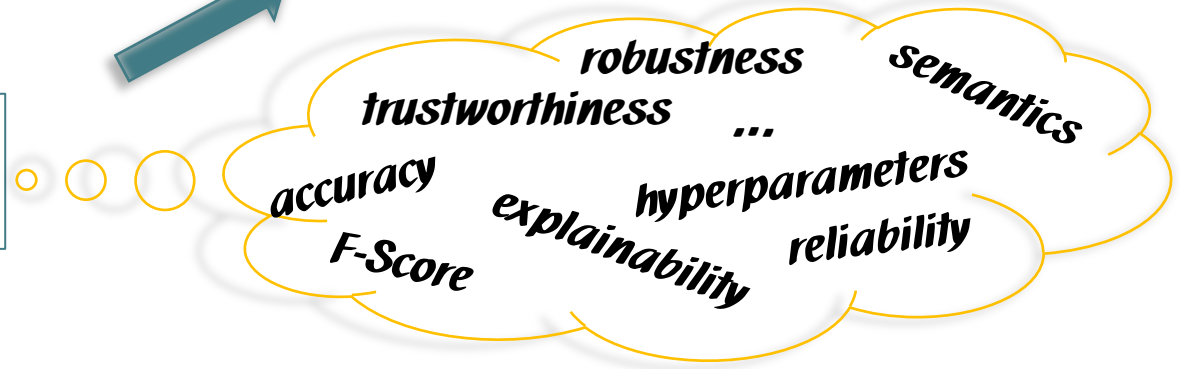




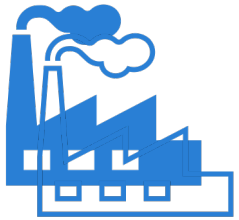
- *high-dimensional input-output data relations*
- *extremly high complexity*
- *probabilistic behaviour*



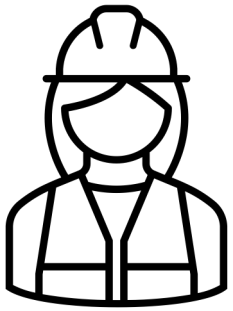
- *test against well defined acceptance criteria*
- *run the test cases once, verify the result*



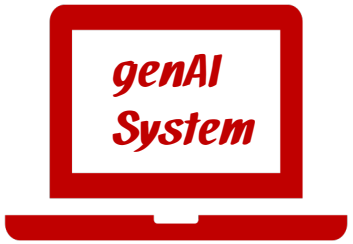
The Quality Engineers Mindset (simplified)



- physically/chemically extremly complex systems
- probabilistic random errors from a high dimensional space of potential root causes



- define objectively measurable quality characteristics
- apply empirical test design using statistical best practices
- automate tests
- Apply RCA

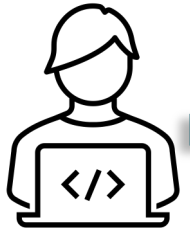


- high-dimensional input-output data relations
- extremly high complexity
- probabilistic behaviour

- define clearly measurable quality characteristics
- apply empirical test design using statistical best practices

- measure outputs „several" times against a set of acceptance criteria
- Apply RCA

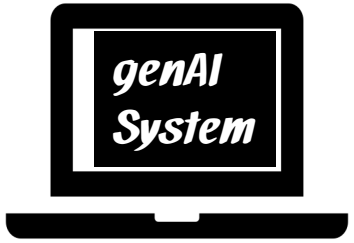
A really simple Example



- *output format must be a single integer*
- *number of characters must be correct*

- *Input: Pfeiffer, f*
- *Expected result: 3*
- *Execute 1 time*

- *Input: strawberry, r*
- *Expected result: 3*
- *Execute 1 time*



TASK: Count the number of times
the letter '{char}' appears in the word '{word}'.
RULES: Output ONLY a single whole integer.
No explanation. No text.

Pfeiffer, f

🚀 Starting simulation:
100 iterations | Model: gemma4:e4b | Temp: 1.0
⏳ Iteration 100/100...
✅ Simulation complete.
💾 Saving results to r_count_results.csv...
✅ Success!

📊 Summary: 100/100 responses were valid integers.
🎯 Accuracy: 7/100 (7.00%)

Note: Due to probabilistic nature of LLMs you may observe differing results.

Examples for Verification Methods & Validation Methods with the Engineering Mindset

Verification

- Data - verify data quality characteristics like ALCOA
- Verify generated outputs against objective criteria during testing and in operation, e.g. schema verification
- Some methods you might want to apply
  - Domain testing to identify edge cases
  - Property Based Testing
  - Prompt perturbation
- Use golden holdout sets (never to be used in training)
- Apply objective metrics not benchmark metrics

Validation

- Data - Is the training data representative with respect to the data encountered in the field?
- End-to-End Scenario Testing - including critical scenarios and edge-case scenarios
- Expert-in-the-Loop Validation
- Operational KPI Validation
- User Acceptance Testing (UAT)
- Safety & Compliance Validation

Summary

Summary

- Apply a Systems and Quality Engineering Mindset
- Develop V&V methods and tools alongside the product - incorporate that in your processes
- Implement strict Data Governance & Quality Assurance
- Generative AI - Verification
  - Apply empirically & statistically sound verification approaches - use objective measures
  - Verify Outputs against ground truths
  - Apply methods like Error Guessing, Property Based Testing and cover edge cases including data
- Generative AI - Validation
  - Evaluate the real benefits vs. the expected benefits - do not apply just benchmarks
  - Evaluate total operating costs from earliest stages on
  - Perform dynamic analysis
  - Evaluate Alternatives early on - is it really a GenAI that you need?

Thank You!

Next Training Dates - Test and Evaluation of AI Systems

May 12-13, 2026 (compact)

June 23 - 26 (extended)