Challenges in Engineering Machine Learning-based Systems:

Perceptions and Actions in Japanese Industry

Fuyuki Ishikawa, National Institute of Informatics, Japan

with QAML Project / JSSST-MLSE / QA4AI Consortium

f-ishikawa@nii.ac.jp

http://research.nii.ac.jp/~f-ishikawa/en/

QAML

QA4AI プロジェクト 品質保証コンソーシアム

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

MLSE 機械学習工学研究会 MACHINE LEARNING SYSTEMS ENGINEERING

Software 2.0 or Inductive Software Dev.

- Unique difficulties for Engineering ML systems
  (as already perfectly introduced in the keynotes)
  - With ML, we construct a software component in a different way: derive the rule that governs the behavior from training data (not directly from engineers)
  - In the Japanese industry, the terms "inductive software development" and "inductive programming" are also used to clarify the essence


Medium

Software 2.0

I sometimes see people refer to neural networks as just "another tool in your machine learning toolbox". They have some pros and cons, they work here or there, and sometimes you can use them to win Kaggle competitions. Unfortunately, this interpretation completely misses the forest for the trees. Neural networks are not just another classifier, they represent the beginning of a fundamental shift in how we write software. They are Software 2.0.

[https://medium.com/@karpathy/software-2-0-a64152b37c35]

Examples from Japan

How Will You Test? (1)

- When Honda sees ramen shop sign
  - First buzz in Dec 2017

[https://twitter.com/_gyochan_/status/938240168078622720]
[https://twitter.com/Bleu_kakeru727/status/937680760491753473]

  - Now a caution on the web site

http://www.honda.co.jp/hondasensing/feature/srf/

  - Second buzz in Sep 2018

[https://twitter.com/yuk381/status/1039837029591179265]

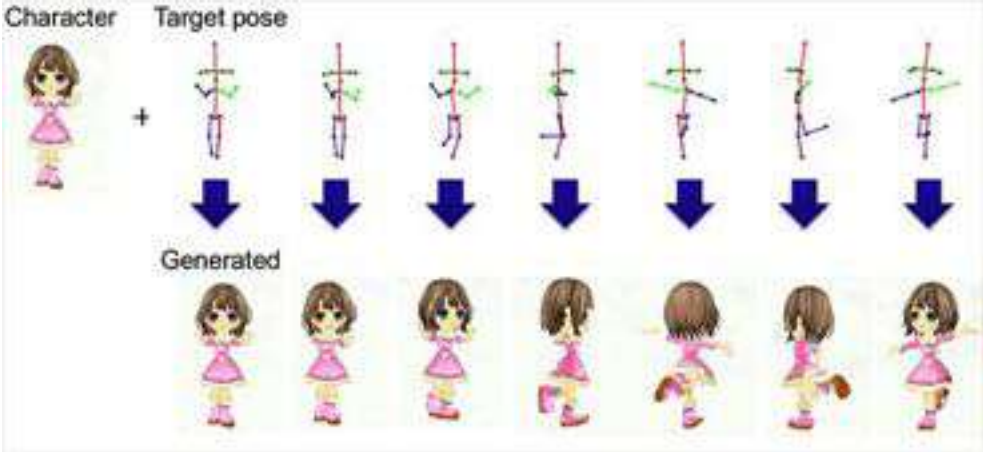
[http://www.tenkaippin.co.jp/company.html]

Can you find beforehand or prevent adverse (?) news??

(but this is perceived as "convincing" mistakes)

How Will You Test? (2)

- From DeNA (May 2018)
  - Generate an image of a certain pose
  - Generate a movie given a pose sequence while changing the character



[https://dena.com/intl/anime-generation/]

What do you ensure to sell this to anime companies?

How do you Judge Acceptance?

- Continuous learning, unclear oracle
  - A Microsoft bot learnt and made absolutely unacceptable tweets

[https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html]

  - A tweet by a Japanese bot to @microsoftjapan:

    "No way, you bald"

    rather attracted people, "natural from a high-school girl to her father"?

[https://twitter.com/ms_rinna/status/706615350524252160]

Initiatives in Japan (Partial)

- SIG-MLSE (Apr 2018-)
  - SIG on Machine Learning Systems Engineering
  - Events almost every month, often go over the capacity mostly with industry people
- QA4AI Consortium (Apr 2018-)
  - Consortium of Quality Assurance for Artificial-Intelligence-based products and services
  - 39 individual members and 3 organ. members
  - Guidelines for QA (to be released in May)
  - Next step: integrating similar efforts, e.g., AIST + NII: quality criteria like ISO26262 and Common Criteria

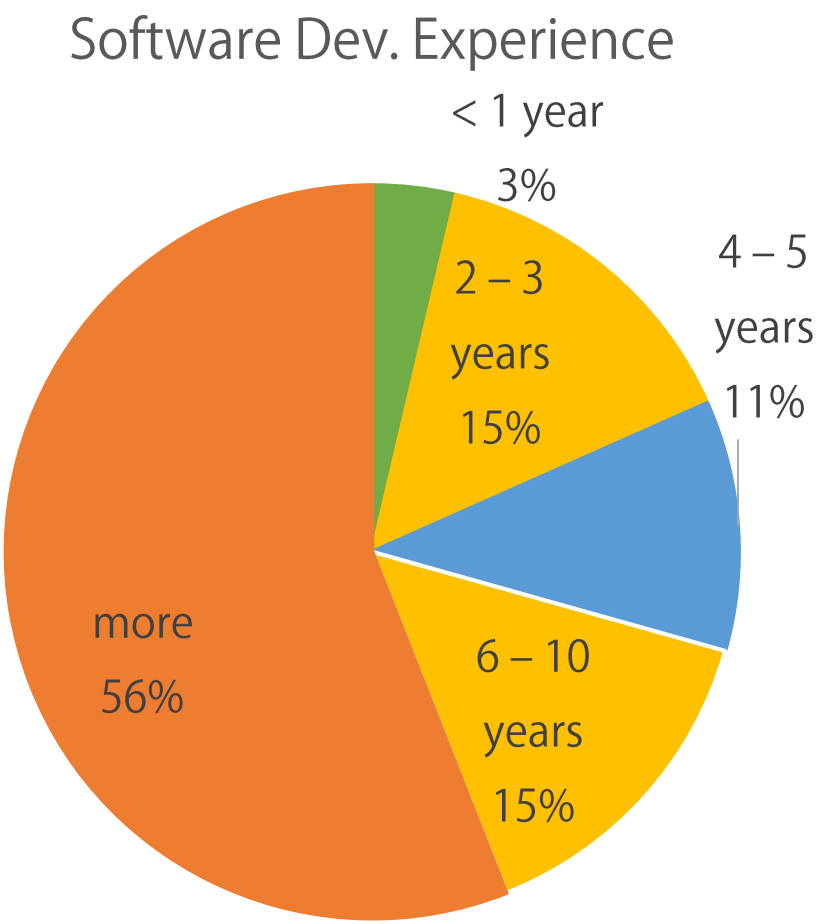
Perception by Engineers

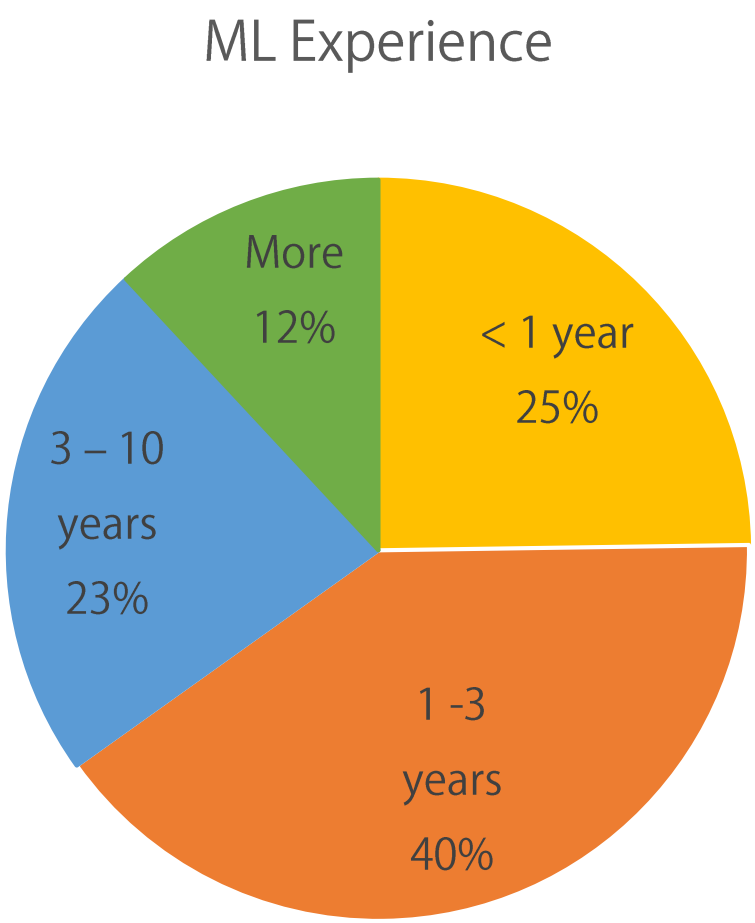
[Ishikawa et al., How do Engineers Perceive Difficulties in Engineering of Machine-Learning Systems? – Questionnaire Survey, CESSER-IP'19]

Questionnaire Survey

- Method
  - Dissemination by mailing lists and social networks (software engineering, ML, and AI)
  - “those who have used ML at work”
  - 280 answers
- Question aspects
  - Experience on SE activities and on ML techniques
  - Past projects that used ML
  - Quality attributes considered significant
  - Perception of difficulties
  - Characteristics of ML that lead to the difficulties

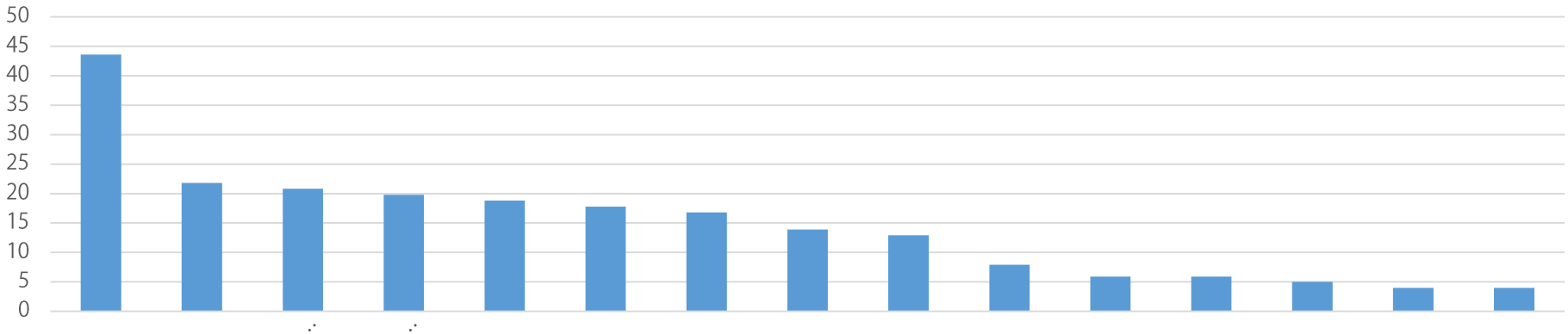
Profiling





Experienced engineers were recently pushed to learn and use ML

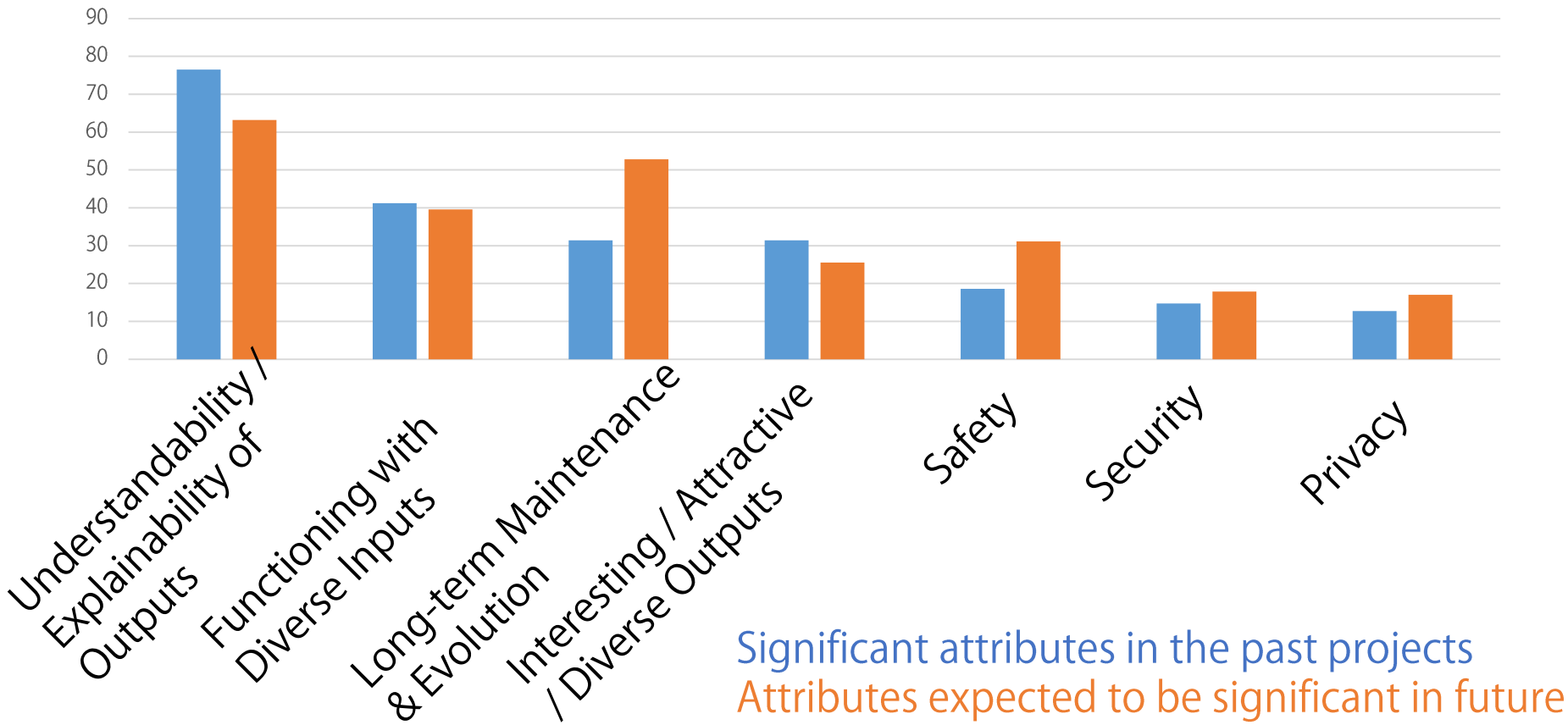
ML Usage Domain



Manufacturing
Information & Communication
Company-Specific Services
Foundational Development (e.g., middleware)
Academic Research
Automotive / Railway
Finance / Insurance
Home Appliance
Wholesale and Retail
Education
Life Service / Amusement
Medical / Welfare
Construction
Real Estate
Utility

Application in manufacturing is large (in Japan)

Significant Quality Attributes



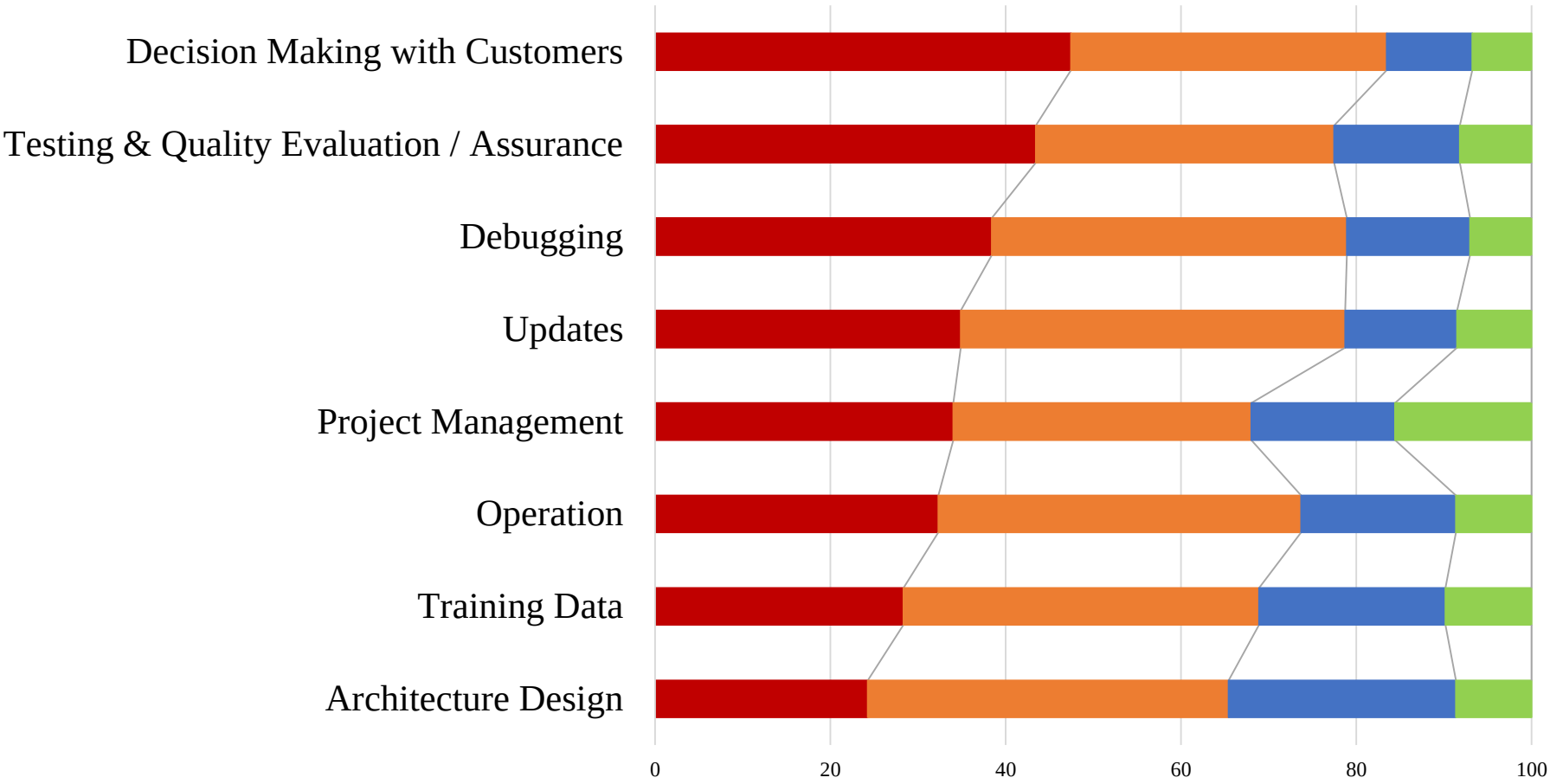
XAI (eXplainable AI) is thought as significant also in practice

Maintenance, security, and privacy are somewhat left behind, or domains were limited not to consider them

Difficulty Level

We need to use new approaches as the existing ones do not work anymore

We can apply the same approaches but methods, etc., are still immature



We already have dedicated methods, etc.

We can use existing methods, frameworks, or tools

Some Typical Comments (Engineer Side)

- Tough negotiation with customers
  - We need to make a priori agreement that we cannot make a priori agreement on what we can develop, e.g., what accuracy we will be able to achieve
  - Free attacks at the acceptance time (with too much expectation on “AI”)
- “POC (Proof-of-Concept) Poverty”
  - There are so many projects that finish at the POC phase
- “We have data! We have! fraud”
  - Having large volume of data does not mean it is well annotated, not noisy, or with good correlations

Released on May 17, 149 pages!
(now only in Japanese …)

Example of Outcome: QA4AI Guideline

Guideline from QA4AI (1)

- 5 axes of evaluation (with brief checklists)
  - Data, Model, and System
  - Process: especially agility to improve and react to unexpected outcomes
  - To match with Customer Expectation
- Technical catalog (Body Of Knowledge)
  - ML standards (accuracy, etc.), search for adversarial examples, metamorphic testing,
    neuron coverage (with a little note on interpretation), explainability tools, etc.

Can only be the current best and to be updated

Guideline from QA4AI (2)

- Specific documents currently for 4 domains
  - Autonomous driving: risk reduction under uncertainty (quick feedback cycles)
  - Factory data processing: variety of stakeholders, dependability on environments, and demand for convincing end users
  - Smart speakers: quality attributes of different abstraction levels, variety of natural language inputs
  - Creative generation: advanced evaluation functions also implemented by ML, such as “naturalness”

Challenges more than Individual Testing Techniques

FAQs on Emerging Testing/Quality Research

When to use what? What is the whole picture and roles of each method?

- Standard measurement of accuracy
- Search for adversarial examples
- Metamorphic testing
- Neuron coverage
- …

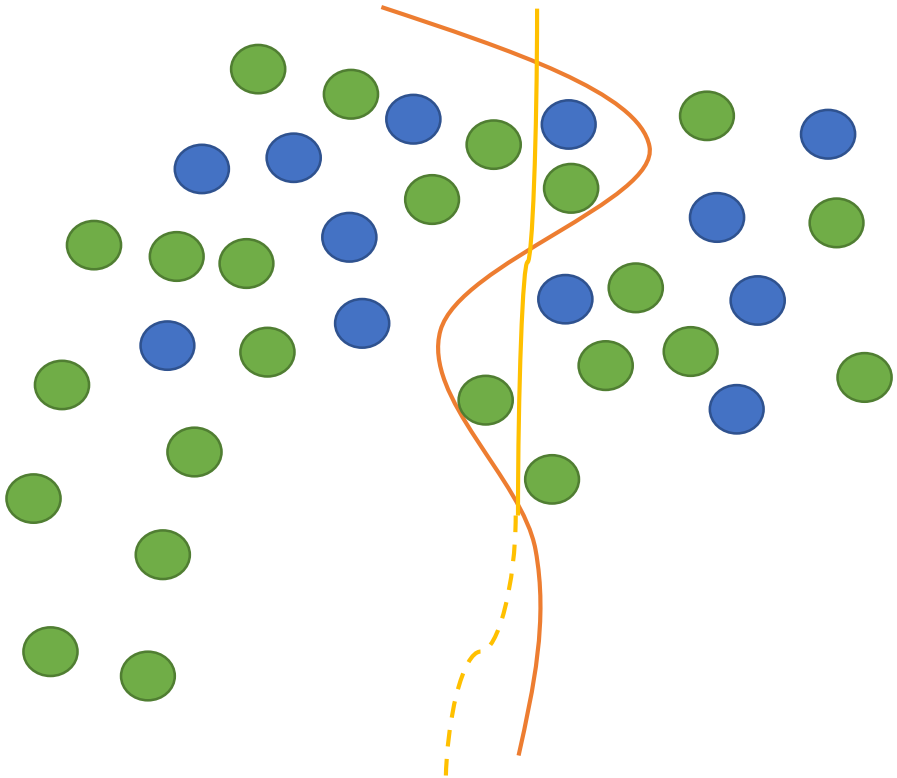
And, is this all of what are required/necessary/effective?

Rough Illustration

In the input space

Implemented boundary (unknown)

Required/correct boundary (often unclear, vague, and implicit)



Current test dataset

Data to appear in operation i.e., what really matters (unknown beforehand)

Where are you generating/selecting new test data?
What coverage/division are you considering?

Example: White-box Structural Coverage

"How diverse output values of each neuron have been observed by a test suite?"

- Originally used for increasing diversity in the "bad" (non-robust) outputs to be generated

  [Pei et al., DeepXplore: Automated Whitebox Testing of Deep Learning Systems, 2017]

  [Tian et al., DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars, 2018]

- Used as criteria to judge a test suite (dataset)

  [Ma et al., DeepGauge: multi-granularity testing criteria for deep learning systems, 2018]

- Now said "could be misleading"
  - e.g., dependency rather on the search methods
  - e.g., natural faults vs. adversarial faults

  [Li et al., Structural Coverage Criteria for Neural Networks Could Be Misleading, 2019]

Summary

Machine learning as a new development paradigm

- Different characteristics, invalidating some of existing principles (in testing and in software engineering)
- Currently, efforts are initial and explorative: sometimes based on (implicit) hypothesis that we should do similar things to what we did for classical software systems

High demand and intensive effort for "disciplines" with deep understanding