A Meta-Summary of Challenges in Building Products with ML Components

Collecting Experiences from 4758+ Practitioners

Nadia Nahar*

Haoran Zhang

Grace Lewis

Shurui Zhou

Christian Kästner

2nd International Conference on AI Engineering – Software Engineering for AI (CAIN 2023)

Machine Learning in Software Products

Data Science Process is Model Centric

face_detection.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

+ Code + Text Copy to Drive

```
[6] print("[INFO] loading model...")
    prototxt = 'deploy.prototxt'
    model = 'res10_300x300_ssd_iter_140000.caffemodel'
    net = cv2.dnn.readNetFromCaffe(prototxt, model)
```

```
[INFO] loading model...
```

Use the dnn.blobFromImage function to construct an input blob by resizing t

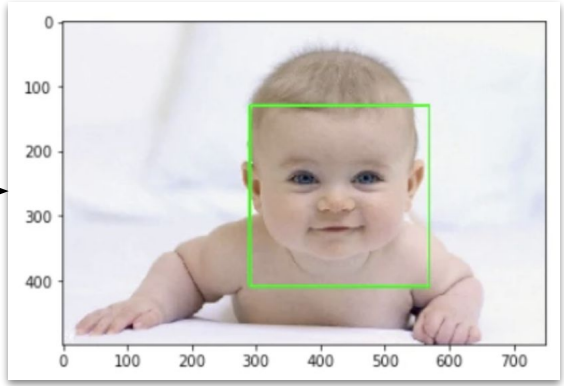
```
[7] # resize it to have a maximum width of 400 pixels
    image = imutils.resize(image, width=400)
    blob = cv2.dnn.blobFromImage(cv2.resize(image, (300, 300)),
```

Pass the blob through the neural network and obtain the detections and prec

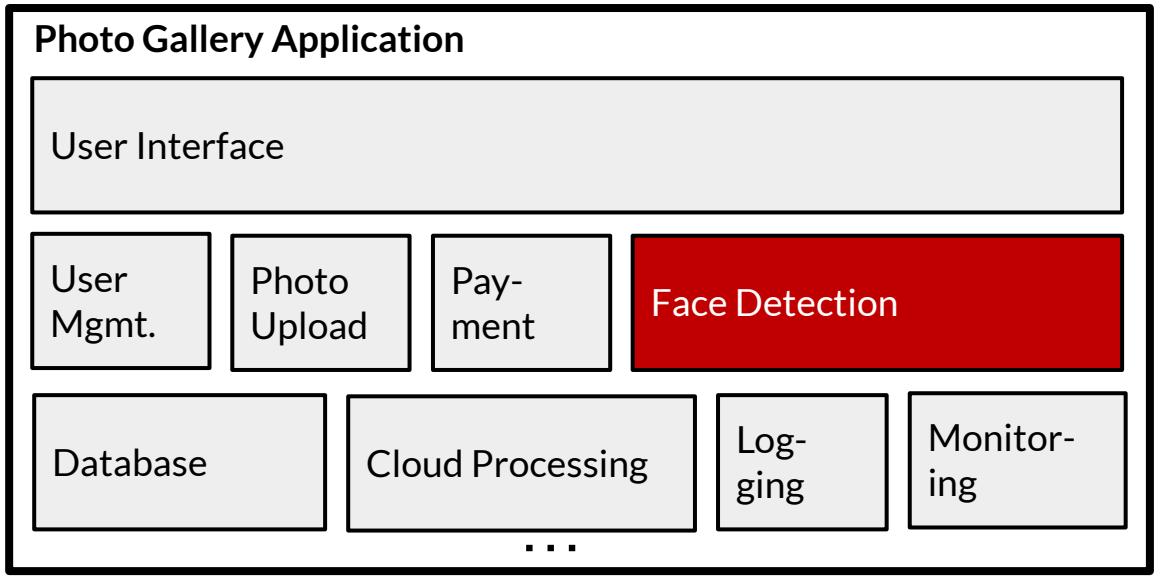
```
[8] print("[INFO] computing object detections...")
    net.setInput(blob)
    detections = net.forward()
```

```
[INFO] computing object detections...
```

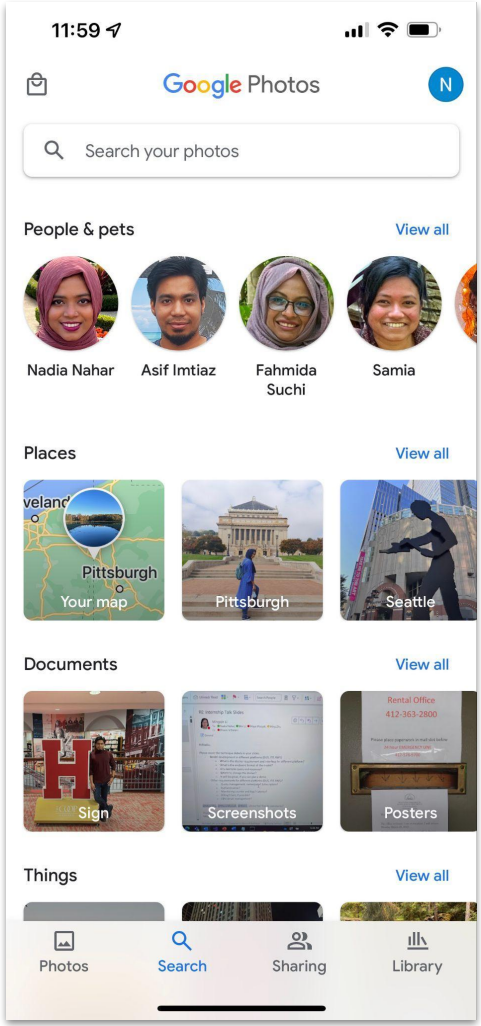
Loop over the detections and draw boxes around the detected faces

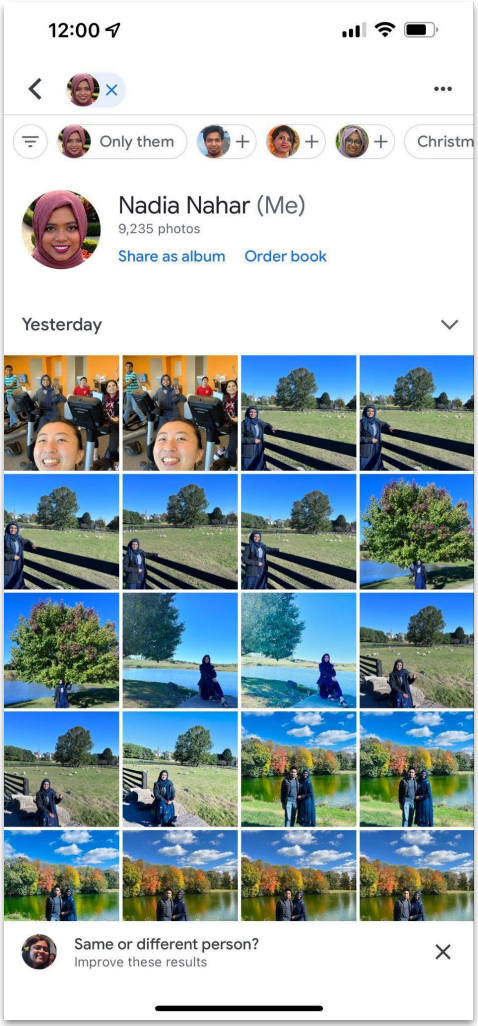


Model as a Component



Google Photos



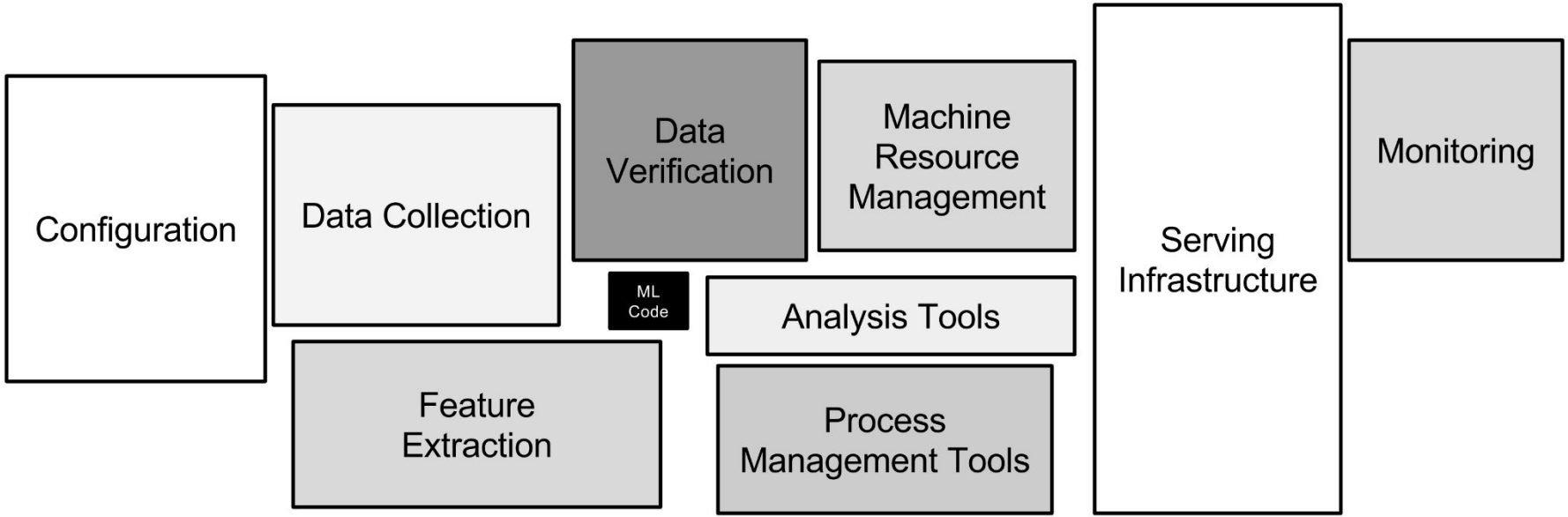


From Model to Product

Data Science Pipeline

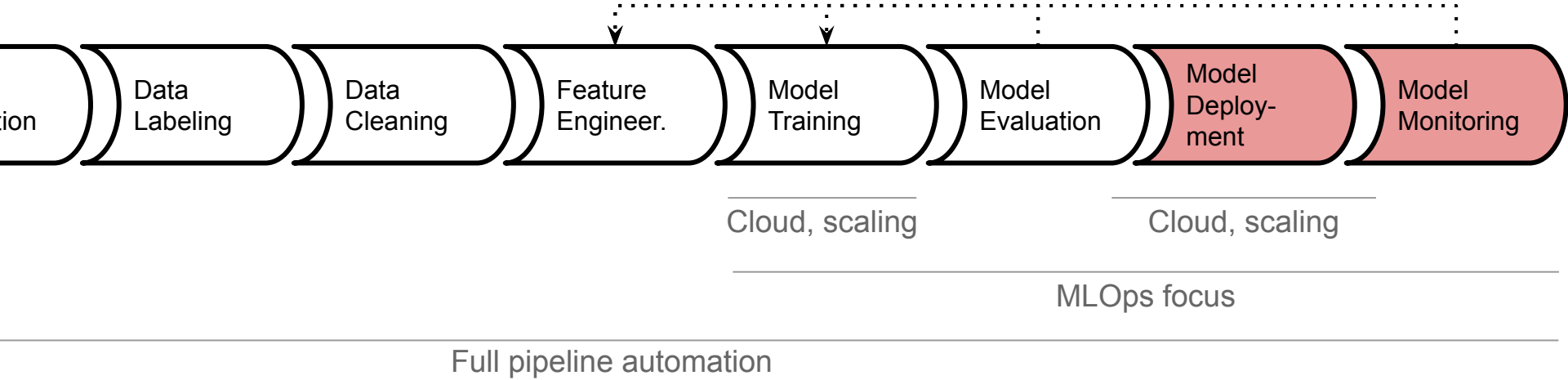
Typical Machine Learning Book / Course

Model Deployment is Complex

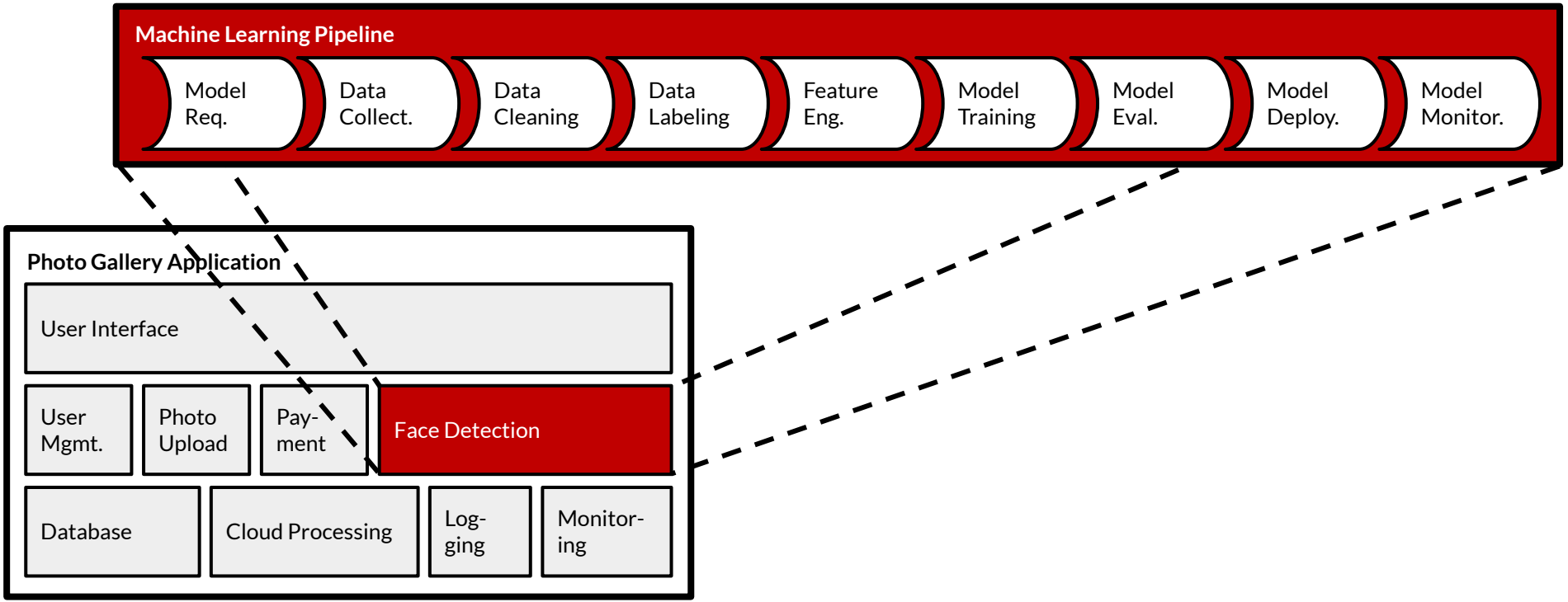


Sculley, et al. *"Hidden Technical Debt in Machine Learning Systems."* NeurIPS 28 (2015): 2503-2511.

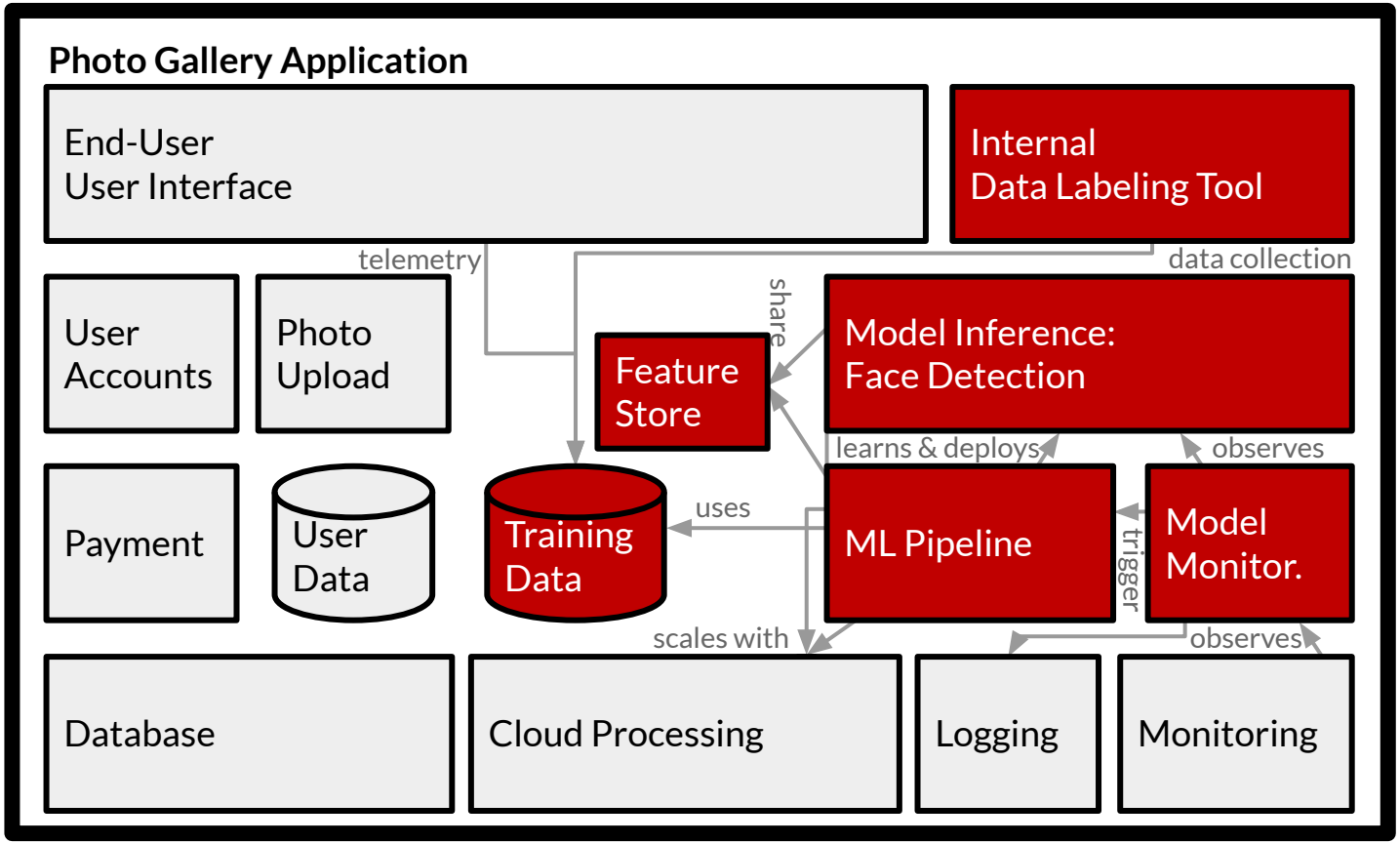
Pipeline Automation and MLOps



ML is a Component of a Product



…Or Many Components



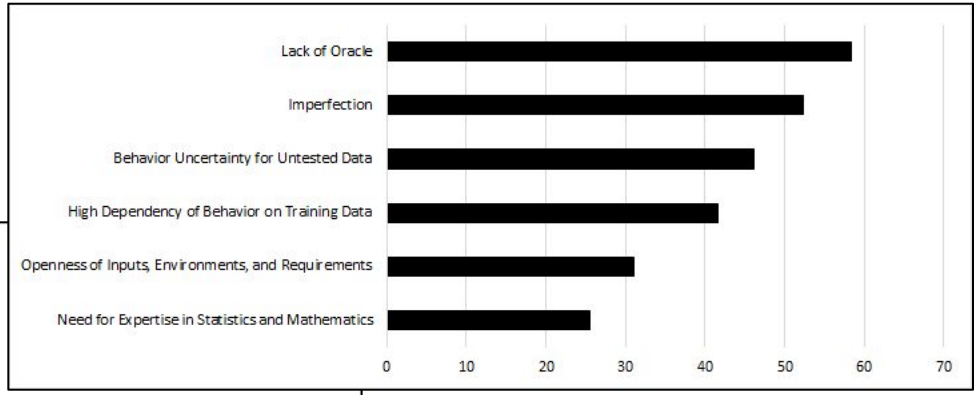
How Does Introduction of ML Impact Software Engineering?

What are the Struggles of Practitioners?

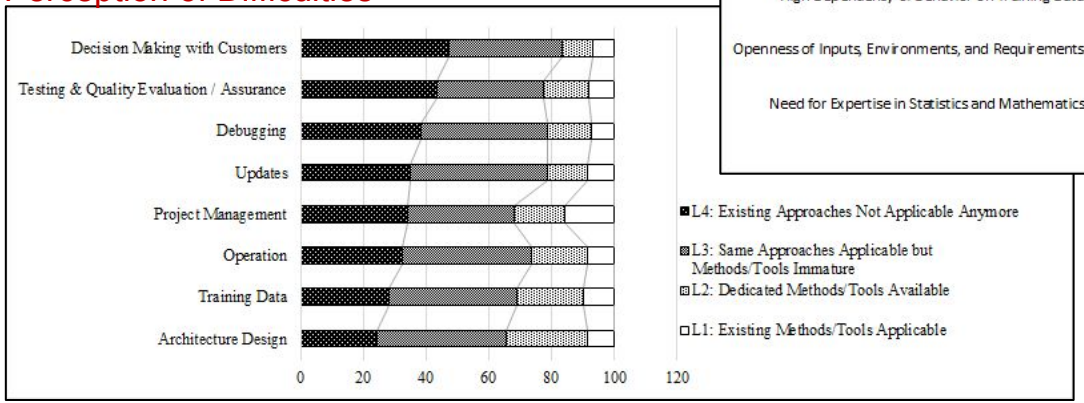
Researchers have attempted to understand the challenges of industry practitioners through interviews, surveys, case studies, ethnographic studies, and so on.

Study on Engineering of ML Systems

Causes of Difficulties

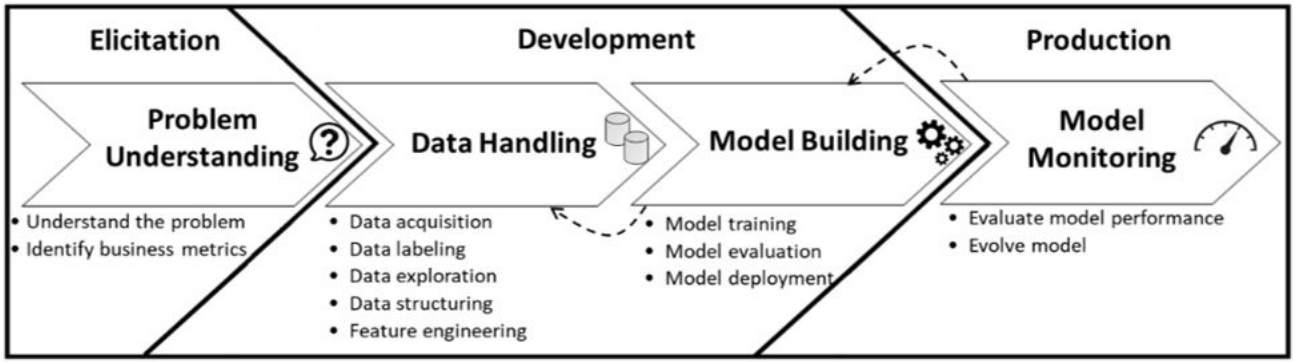


Perception of Difficulties



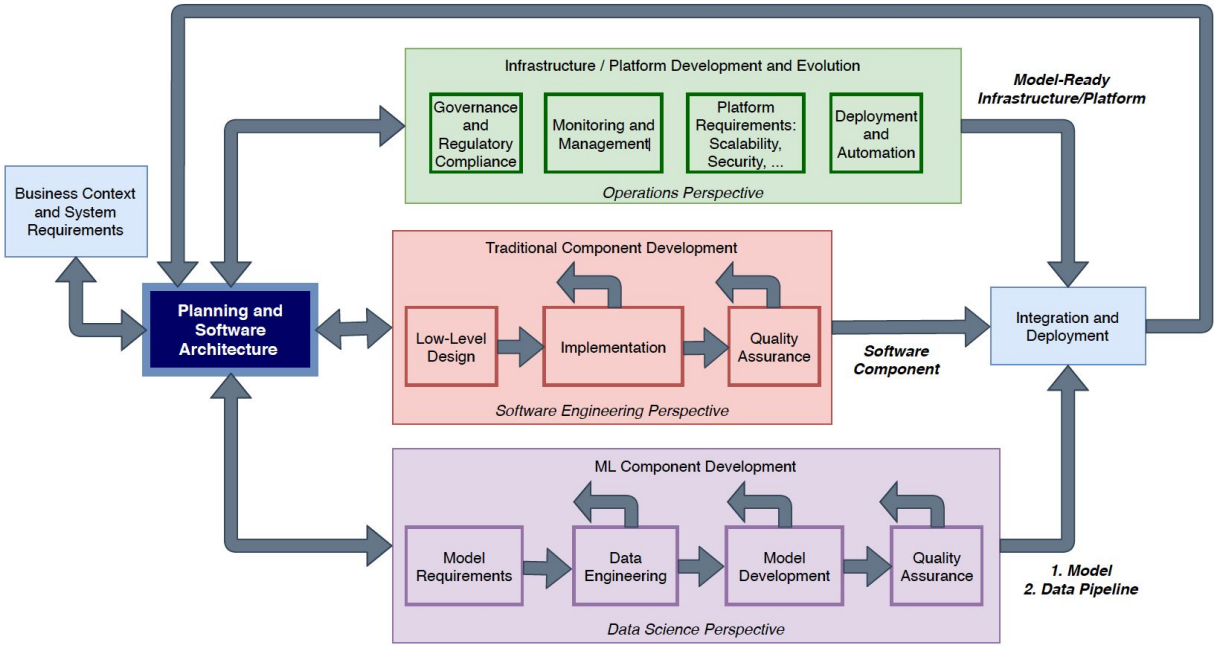
Ishikawa, Fuyuki, and Nobukazu Yoshioka. "How Do Engineers Perceive Difficulties in Engineering of Machine-Learning Systems? -Questionnaire Survey." In 2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP), pp. 2-9. IEEE, 2019.

Study on Development Processes



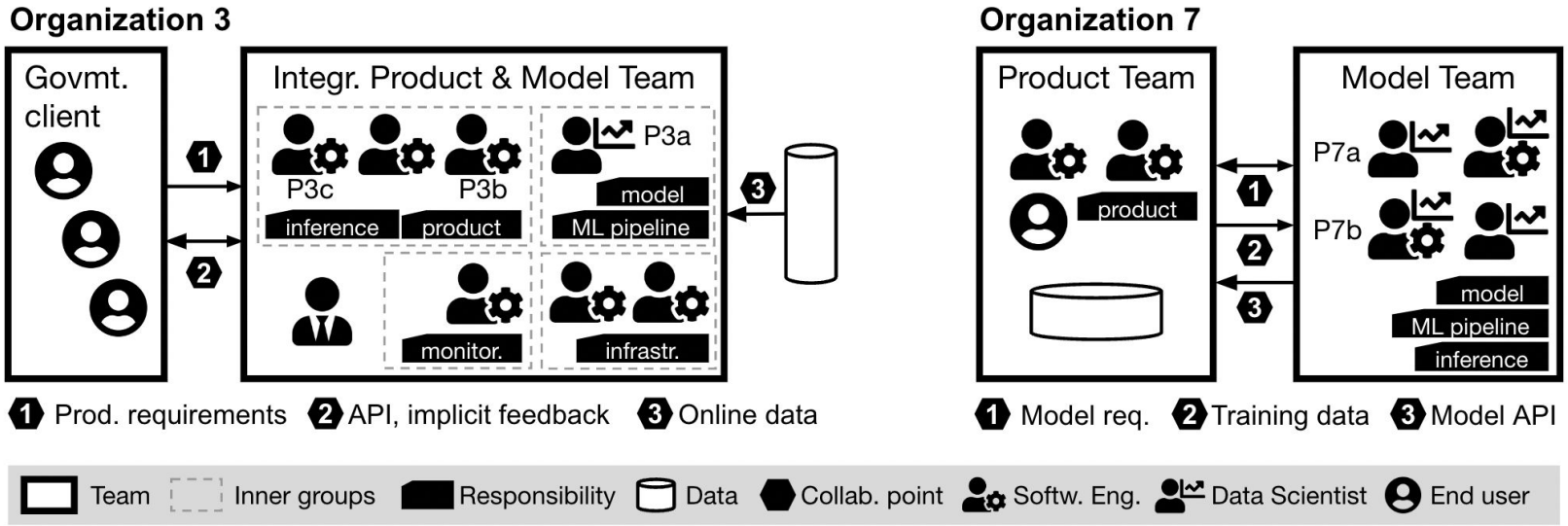
de Souza Nascimento, Elizamary, Iftekhar Ahmed, Edson Oliveira, Márcio Piedade Palheta, Igor Steinmacher, and Tayana Conte. "Understanding Development Process of Machine Learning Systems: Challenges and Solutions." *In Proceedings of International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pp. 1-6, 2019.

Study on Software Architecture



Lewis, Grace A., Ipek Ozkaya, and Xiwei Xu. "Software Architecture Challenges for ML Systems." In *Proceedings of International Conference on Software Maintenance and Evolution (ICSME)*, pp. 634-638, 2021.

Study on Collaboration and Teams



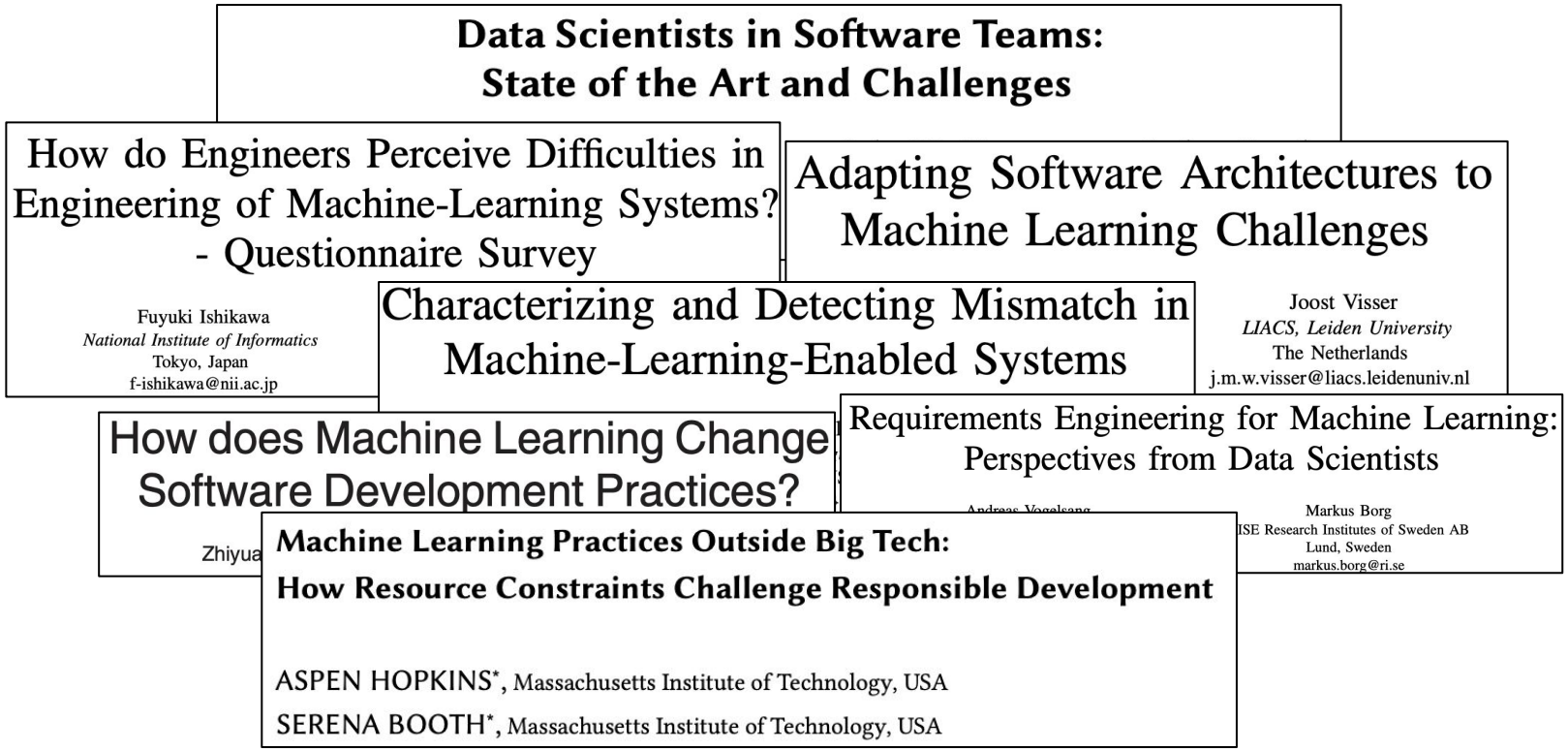
Nahar, Nadia, Shurui Zhou, Grace Lewis, and Christian Kästner. "Collaboration Challenges in Building ML-enabled Systems: Communication, Documentation, Engineering, and Process." In Proceedings of the 44th International Conference on Software Engineering, pp. 413-425. 2022.

Lots of Pain-point Papers

Researchers study different challenges in building ML Products

There are lots of studies with industry practitioners

Lots of Pain-point Papers



Lots of Pain-point Papers

Researchers study different challenges in building ML Products

There are lots of studies with industry practitioners

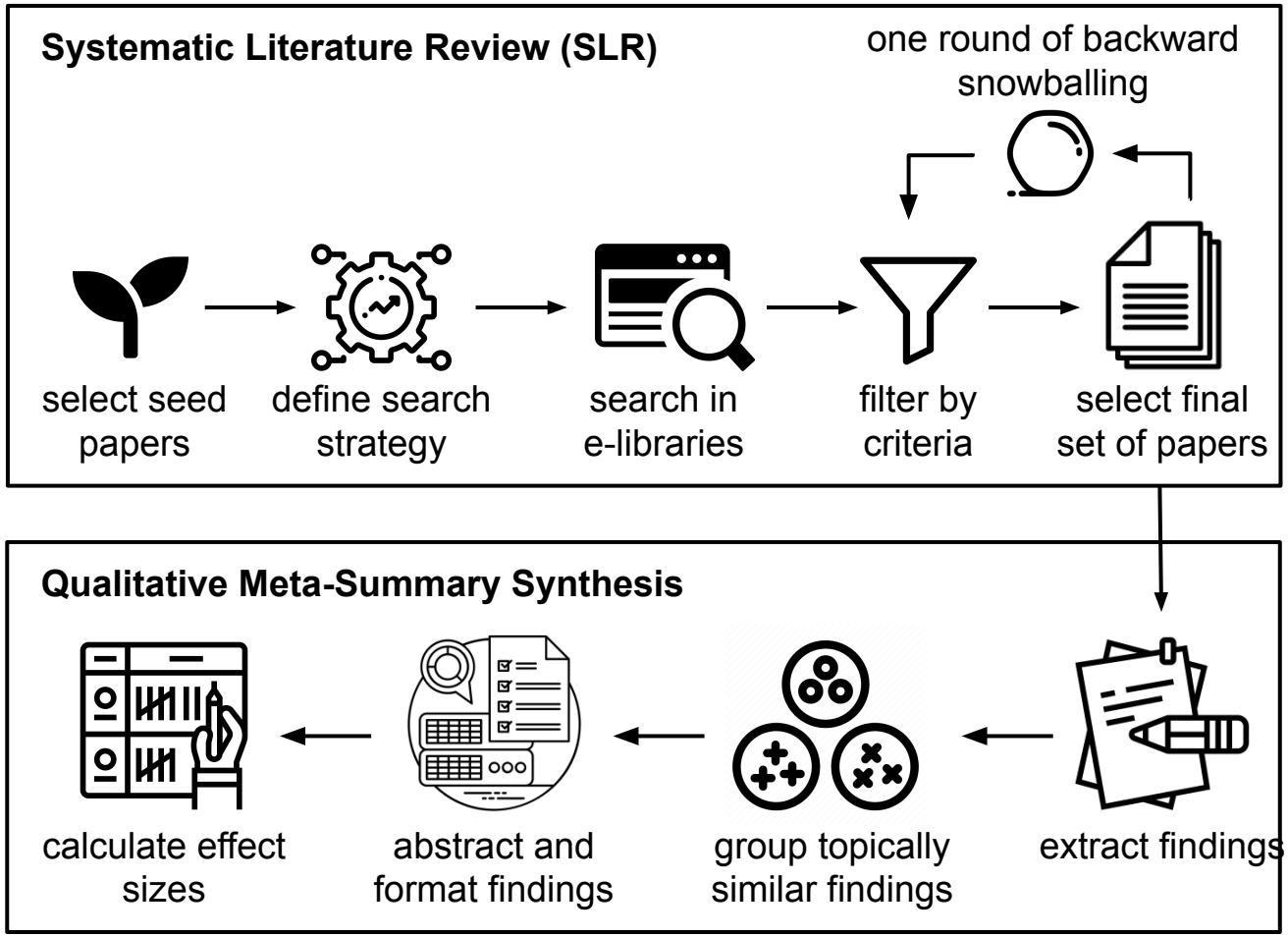
Time to explore what we know collectively…

Research Question

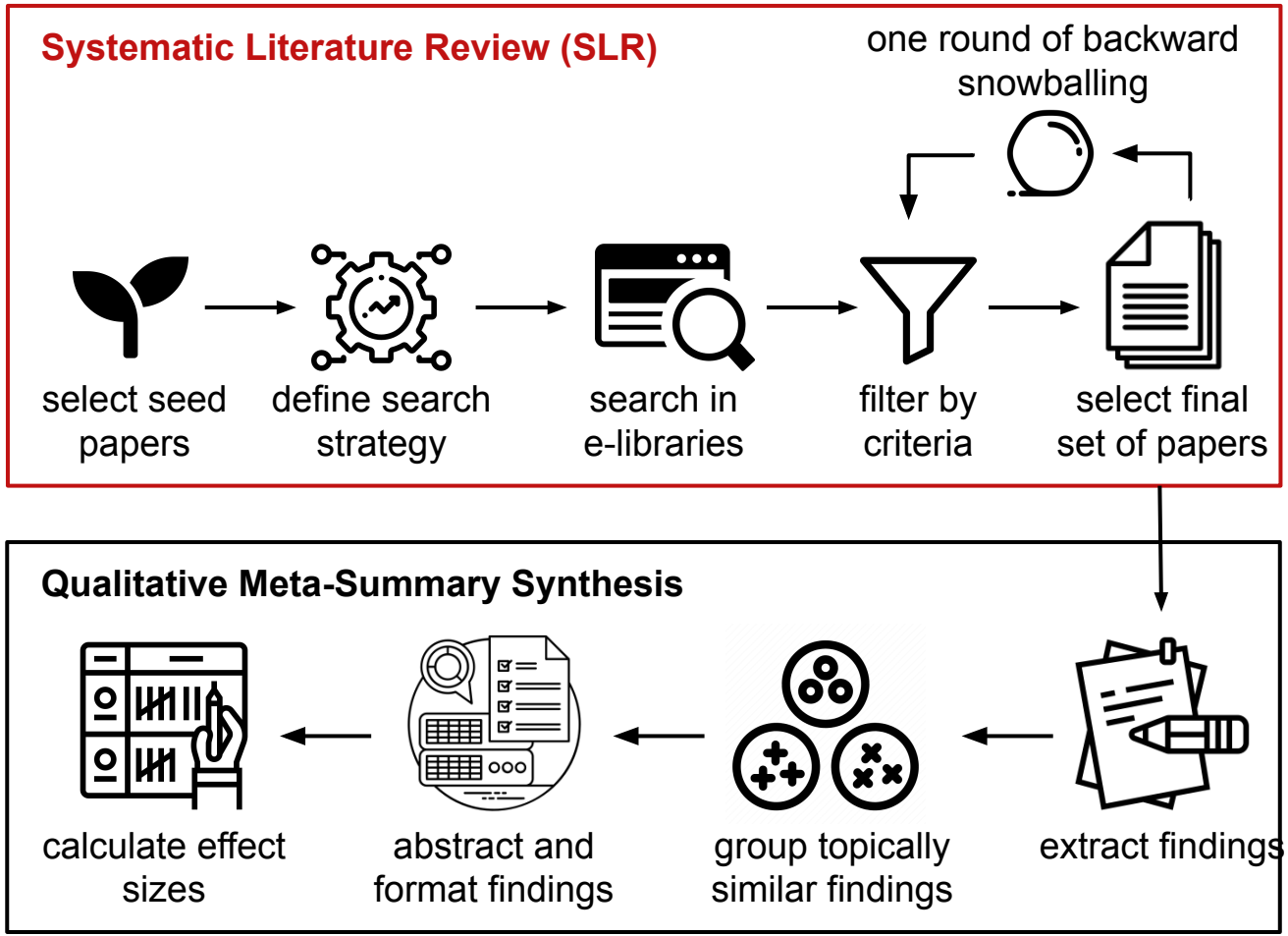
“What are the challenges experienced by industry practitioners in building software products with ML components? ”

Our Contribution

Overview


Systematic Literature Review (SLR)
one round of backward snowballing
select seed papers
define search strategy
search in e-libraries
filter by criteria
select final set of papers
Qualitative Meta-Summary Synthesis
calculate effect sizes
abstract and format findings
group topically similar findings
extract findings

Overview


Systematic Literature Review (SLR)
one round of backward snowballing
select seed papers
define search strategy
search in e-libraries
filter by criteria
select final set of papers
Qualitative Meta-Summary Synthesis
calculate effect sizes
abstract and format findings
group topically similar findings
extract findings

Paper Selection Criteria

Paper includes software engineering challenges for ML systems.

Paper uses interview or survey with industry practitioners (software engineers, data scientists, etc.) to identify the challenges.

Excluded model centric papers.

Excluded single case study papers.

Excluded interviews with non-technical people only.

Start with Seed Papers

	A	B	D	E
1		**Papers**	**Abstracts**	**Where**
2	1	Software engineering for machine learning: A case study	Recent advances in machine learning have stimulated widespread interest within the	IEEE
3	2	Requirements Engineering for Machine Learning: Perspectives from Data Scientists	Machine learning (ML) is used increasingly in real-world applications. In this paper, we	IEEE
4	3	Improving fairness in machine learning systems: What do industry practitioners need?	The potential for machine learning (ML) systems to amplify social inequities and unfai	ACM
5	4	Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices	Large and ever-evolving technology companies continue to invest more time and resc	ACM
6	5	Adapting Software Architectures to Machine Learning Challenges	Unique developmental and operational characteristics of ML components as well as th	IEEE
7		Data Scientists in Software Teams: State of the Art and		

Seed Paper Analysis for Defining Query

	A	B	C	D	E	F	G	H	I
1		Papers	Why missed initially?	Abstracts	Where	SE search keyword	ML search keyword	Method search keyword	Other
2	1	Software engineering for machine learning: A case study	no mention of interview in abstract	Recent advances in machine le	IEEE	software engineering	machine learning, data science		challenge
3	2	Requirements Engineering for Machine Learning: Perspectives from Data Scientists	no mention of "Software Engineering" in abstract, but have engineering	Machine learning (ML) is used	IEEE	ML systems	machine learning	interview	challenge
4	3	Improving fairness in machine learning systems: What do industry practitioners need?	no mention of "Software Engineering" or engineering in abstract	The potential for machine learn	ACM	ML systems	machine learning	interview, survey	challenge
5	4	Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices	no mention of "Software Engineering" or engineering in abstract	Large and ever-evolving techn	ACM	production-read y systems	artificial intelligence	interview	challenge
			no mention of						

Search Query

A: need an **ML-related keyword**

- “machine learning” OR “artificial intelligence” OR “deep learning” OR “ML component” OR “data science”

B: need a **software engineering or ML deployment-related keyword**

- “software engineering” OR “software systems” OR “production-ready systems” OR “ML systems” OR “deploying ML” OR “ML deployment”

C: need to mention **surveys or interviews**

- “interview” OR “survey” OR “questionnaire”

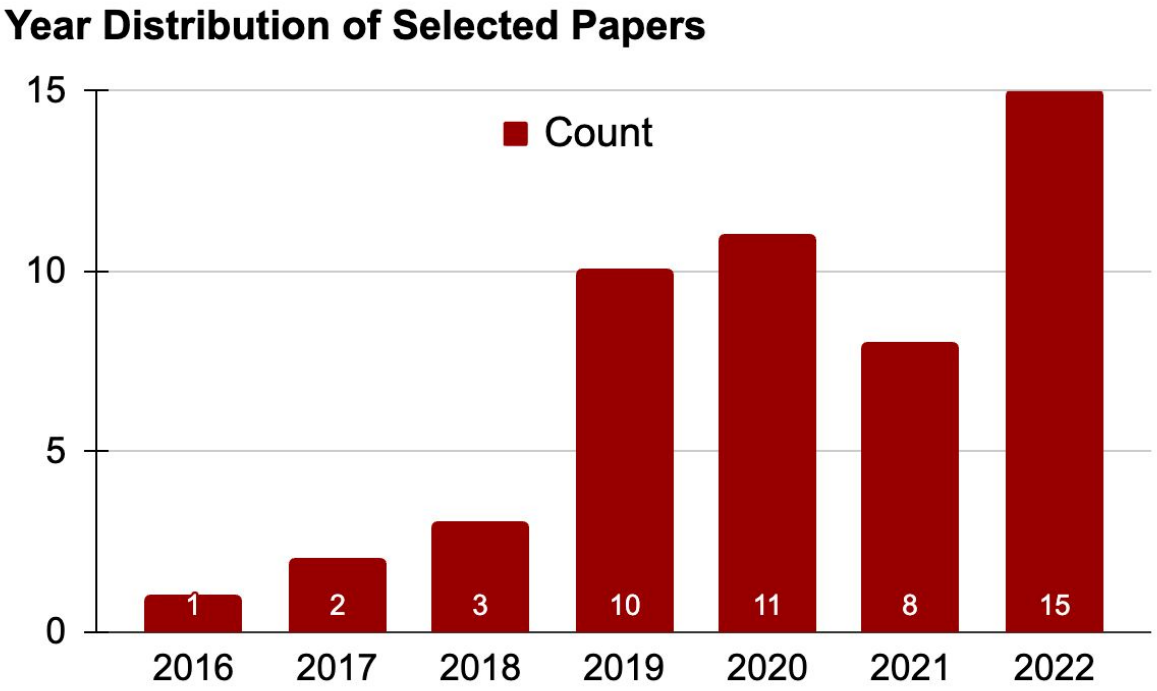
“A AND B AND C”

50 Papers Qualified

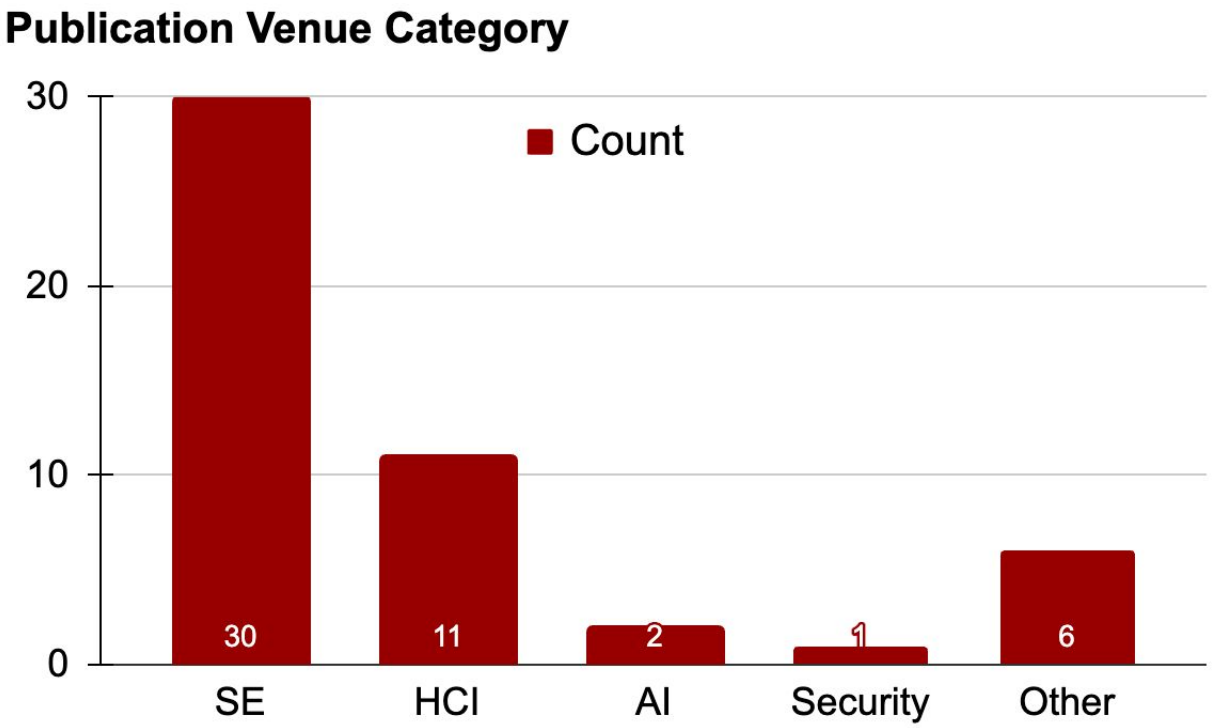
Data Source	Initial Search Result	After Filtering by Title/Abstract	Final Selection
IEEE	69	30	19
ACM	48	11	10
Willey	6	0	0
ScienceDirect	32	5	3
Engineer Village	101	3	0
Springer	6*	3	2
arXiv	79	8	5
Snowballing	-	26	11
Total	**341**	**86**	**50**

*abstract filtering from 5612 papers retrieved with full text search

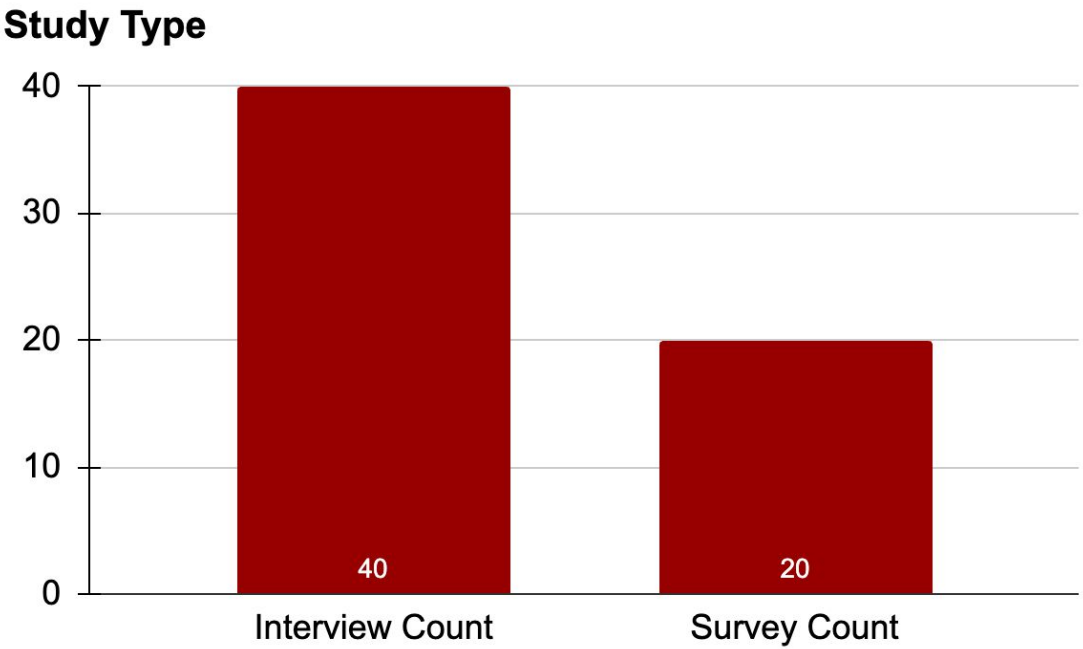
Final Set – 50 Papers



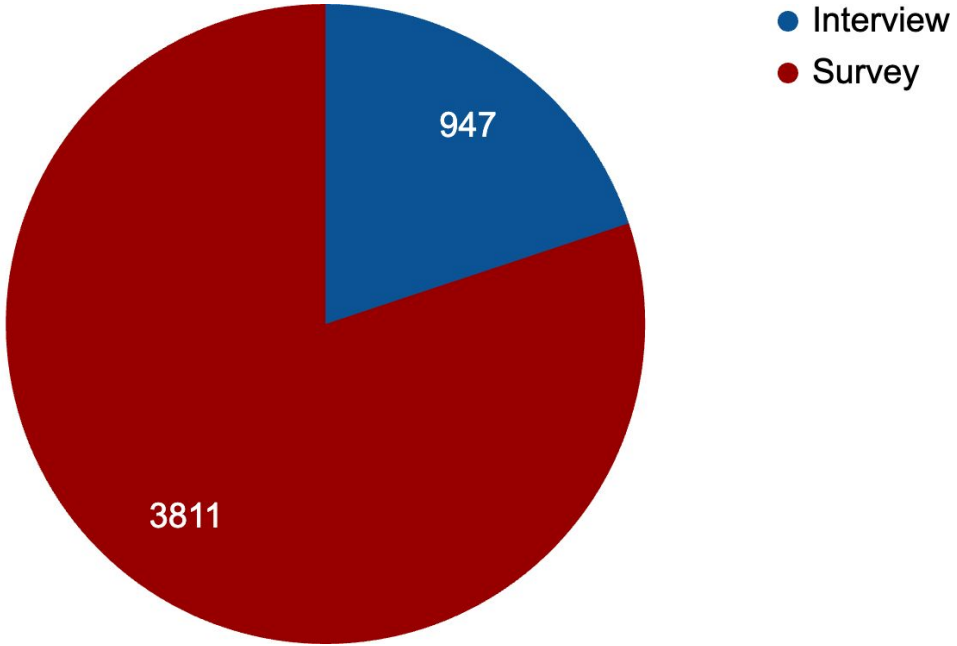
Final Set – 50 Papers



Final Set – 50 Papers

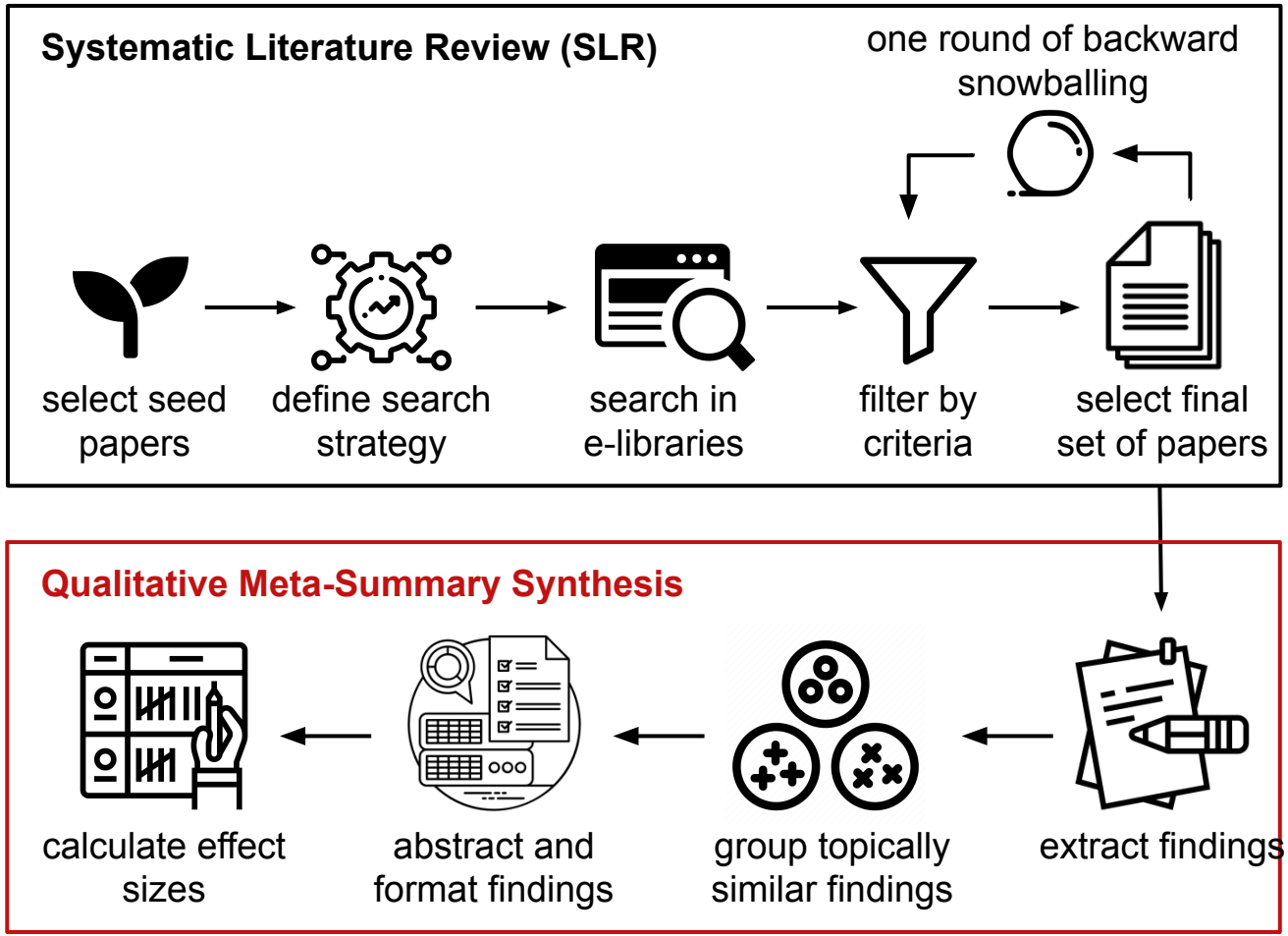


Practitioner Count – 4758+*



*count unspecified in seven papers

Overview


Systematic Literature Review (SLR)
one round of backward snowballing
select seed papers
define search strategy
search in e-libraries
filter by criteria
select final set of papers
Qualitative Meta-Summary Synthesis
calculate effect sizes
abstract and format findings
group topically similar findings
extract findings

Qualitative Synthesis Methods

Table 1: Overview of qualitative synthesis methods applied in software engineering

Synthesis method	Features	Attributes	Aim
Narrative synthesis	Narrative description and ordering of primary evidence with commentary	•Interpretive •Epistemology of idealism	An overview of the findings of primary studies is presented, summarizes the main themes, findings and related issues.
Thematic synthesis	Identifying major or recurring themes in literature and summaries of results of primary studies under the headings of these themes	•Aggregative •Epistemology of realism •Highly structured in data organizing •Outcome utilitarian	Identify, analyze, and report themes or patterns within data
Meta-ethnography	"*Interpretations and explanations in the primary studies are treated as data, and are translated across several studies to produce a synthesis*"	•Interpretive •Epistemology of realism	The integration of data from the primary study by means of induction, interpretation, translation, helps to understand and transfer ideas and concepts
Meta-summary	Quantitative oriented aggregation of qualitative findings. Identify the frequency of each discovery, as well as the discovery of high frequency findings	•Aggregative •Epistemology of realism •Outcome theoretical	Discover a pattern or theme in qualitative research based on the higher frequency of findings
Content analysis	The evidence for each of the primary study is used under a wide range of thematic headings, designed to help with repetitive extraction tools	•Aggregative •Epistemology of realism	Count and tabulate on each occurrence of the theme
Grounded theory	Identifying patterns and relationships in primary data, sampling for analysis, exploring commonalities, and generating theories or models	•Interpretive •Epistemology of realism •Iterative and circular in processes •Outcome theoretical	Generates higher-order themes and interpretations
Comparative analysis	Using Boolean logic (based on specific results of truth tables) to analyze complex causal relationships	•Aggregative •Epistemology of realism	Analyzes complex causal connections
Case survey	Making closed questions to extract data and each primary study can be seen as a specific case	•Aggregative •Epistemology of realism	Extracted data can be used for further (statistical) analysis

Huang, Xin, He Zhang, Xin Zhou, Muhammad Ali Babar, and Song Yang. "Synthesizing qualitative research in software engineering: A critical review." In Proceedings of the 40th international conference on software engineering, pp. 1207-1218. 2018.

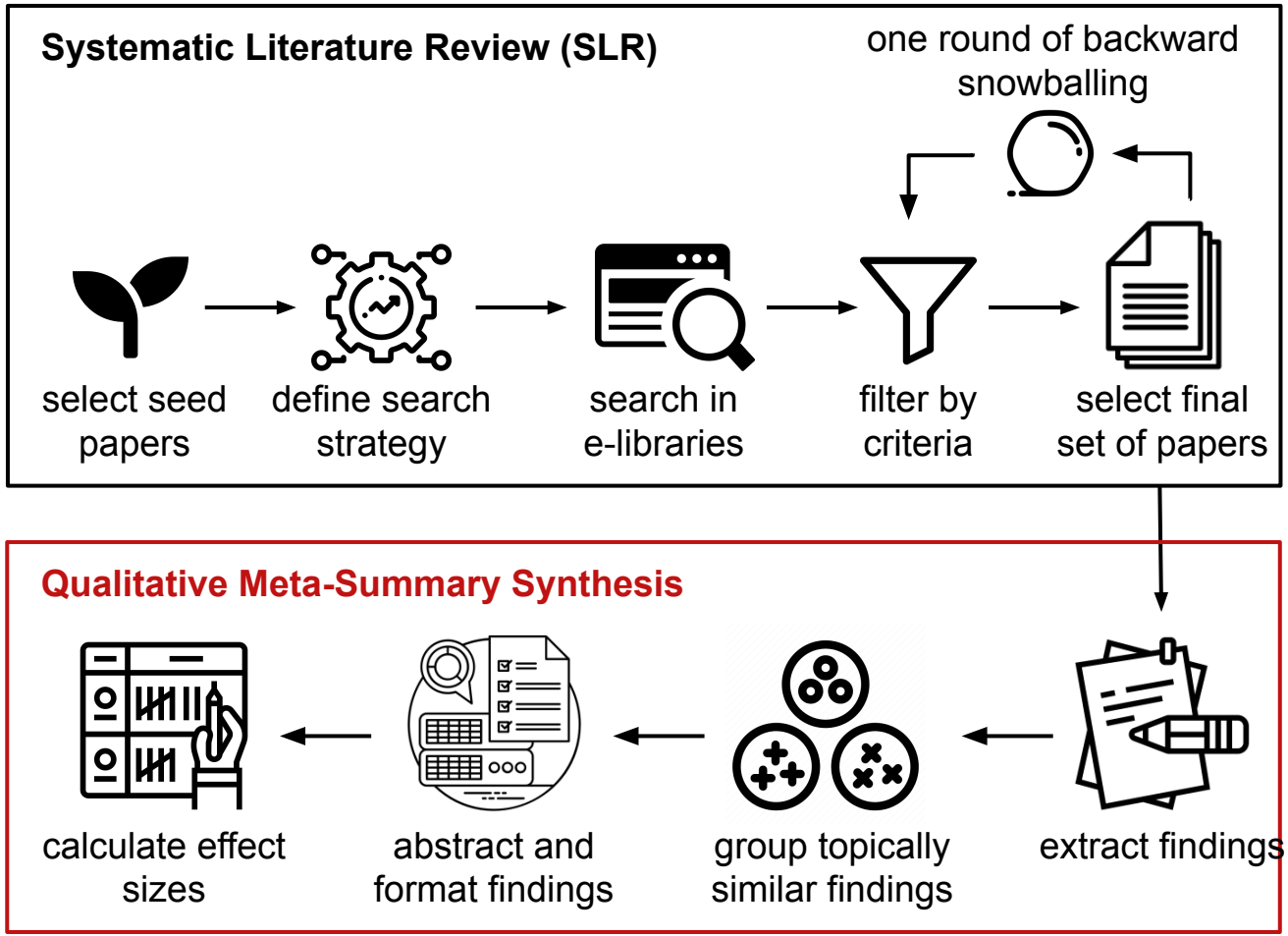
Meta-Summary Method

Well-balanced, falls between high-level analysis such as mapping studies and deeper interpretative syntheses such as meta-ethnography*

Aggregate and present frequencies of findings

*Ribeiro, Danilo Monteiro, Marcos Cardoso, Fabio QB da Silva, and César França. "Using Qualitative Metasummary to Synthesize Empirical Findings in Literature Reviews." In Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 1-4. 2014.

Overview


Systematic Literature Review (SLR)
one round of backward snowballing
select seed papers
define search strategy
search in e-libraries
filter by criteria
select final set of papers
Qualitative Meta-Summary Synthesis
calculate effect sizes
abstract and format findings
group topically similar findings
extract findings

Extract Challenges

mentioned by developers during our interviews.

a) Identifying business metrics is not trivial - in the initial stage of "Problem Understanding," developers need to identify what the customers' business metrics are. However, performing this task is challenging, as stated by P5, when we asked *"how do you identify customer's business metrics? That's a challenge."* Still, th
to improve their business,
metrics and data are requ
participants P4 and P6.

*"The customer wants to have a
do it, what kind of data he should
even understand the data he nee
"When the customer does not h
this is a **problem** that we still ha
academic metrics that are alread*

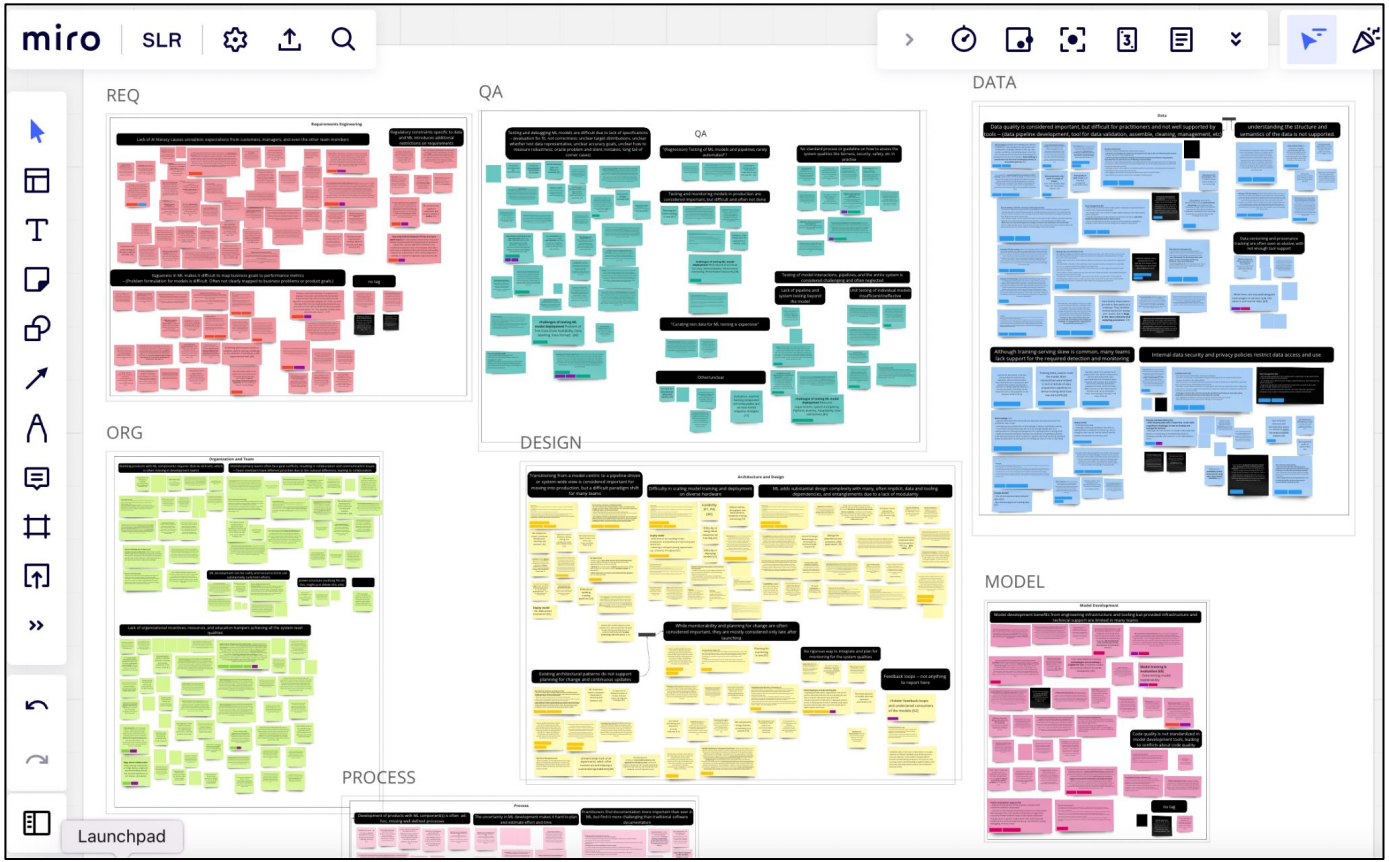
to show this to the customer is difficult, because he often does not understand it." (P4)

b) Undefined process – during the "Data Handling" stage, the developer performs various tasks including data
eprocessing, which entails checking missing data, verifying
consistencies, performing feature engineering. As stated by
: *"At Feature engineering stage, it is important to have insights.
cause we know that if we do not do anything in some attributes,
model should discard these attributes in the next stage."*

As stated by P5, failure to perform these tasks can result in
or model and performance. Since all companies do not have
defined development process for ML systems each

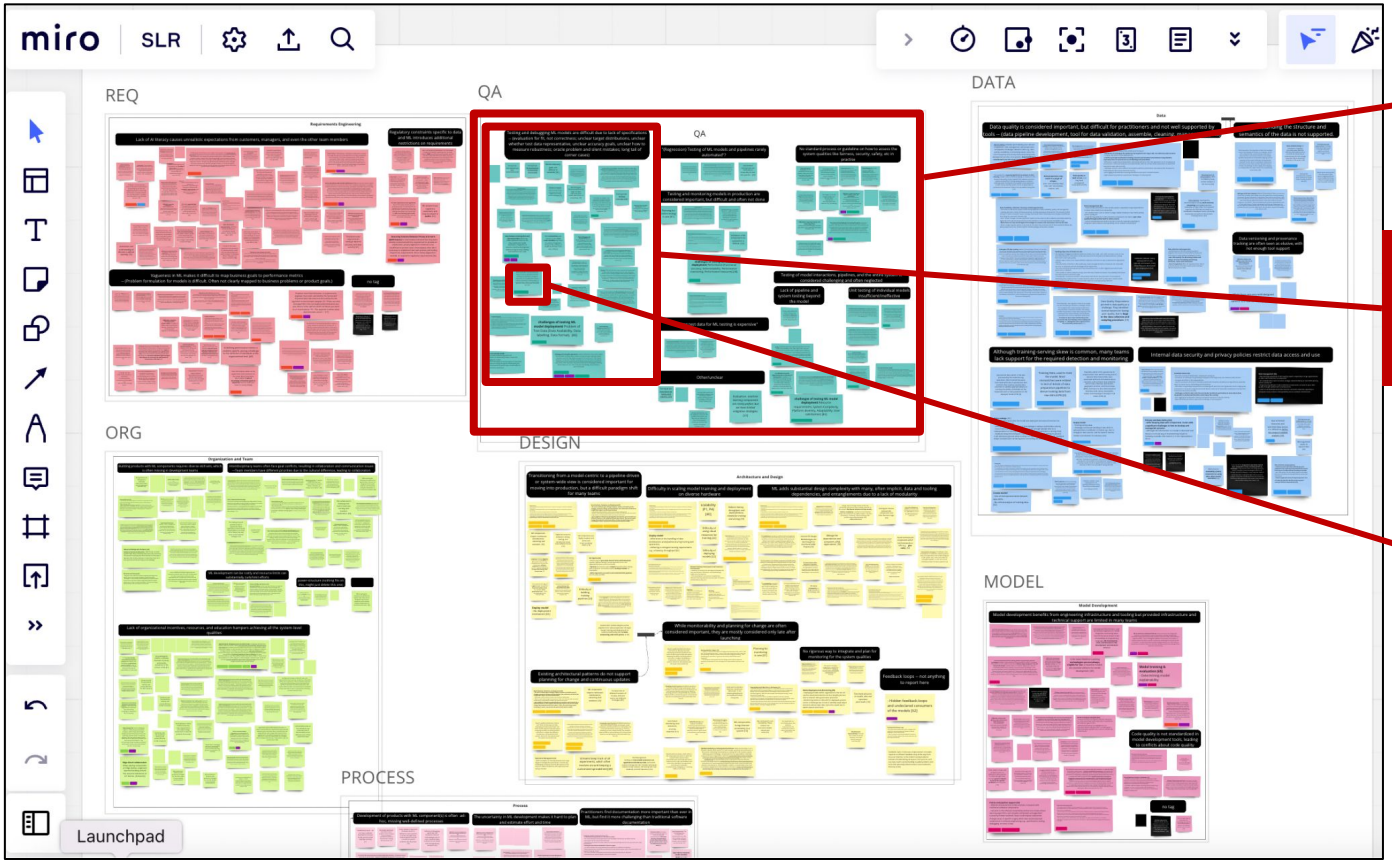
	A	B	C	D	
1	#	Title	Listed RQs	Study Method	Challenges
2	1	Understanding Development Process of Machine Learning Systems: Challenges and Solutions	RQ1: How software developers build ML systems in small companies? **RQ2: What challenges are perceived by developers during the development of ML systems in small companies?** RQ3: Is it possible to help the developers overcome these challenges?	Interviews with 7 devs from 3 small companies	Identifying business metrics is not trivial - the customer wants to h metrics and data are required to do so Undefined process - do not have a defined development process f Difficulty to design the database structure - Developers have repo technical knowledge, and it is initially a manual process.
	6	Characterizing and Detecting Mismatch in Machine-Learning-Enabled Systems	**RQ1: What are common types of mismatch that occur in the end-to-end development of ML-enabled systems?** RQ2: What are best practices for documenting data, models, and other system elements that will enable detection of ML mismatch? RQ3: What are examples of ML mismatch that could be detected in an automated way, based on the codification of best practices in machine-readable descriptors for ML system elements?	Interview with 20 practitioners, and a validation survey with 31 responses	**incorrect assumptions about the Trained Model (36%), which engineers for integration into a larger system** - Most mismatches were related to lack of test cases and data tha specifications and APIs (17%) - unawareness of decisions, assumptions, limitations, and constra **Operational Environment, which refers to the computing envi** associated with lack of runtime metrics, logs, user feedback, and troubleshooting, debugging, or retraining (54%), etc etc etc **Task and Purpose, which are the expectations and constraint** business goals or objectives that the model was meant to satisfy (

Card Sorting in Miro Board*



*miro.com

Three Layers of Clusters



top layer – development stages

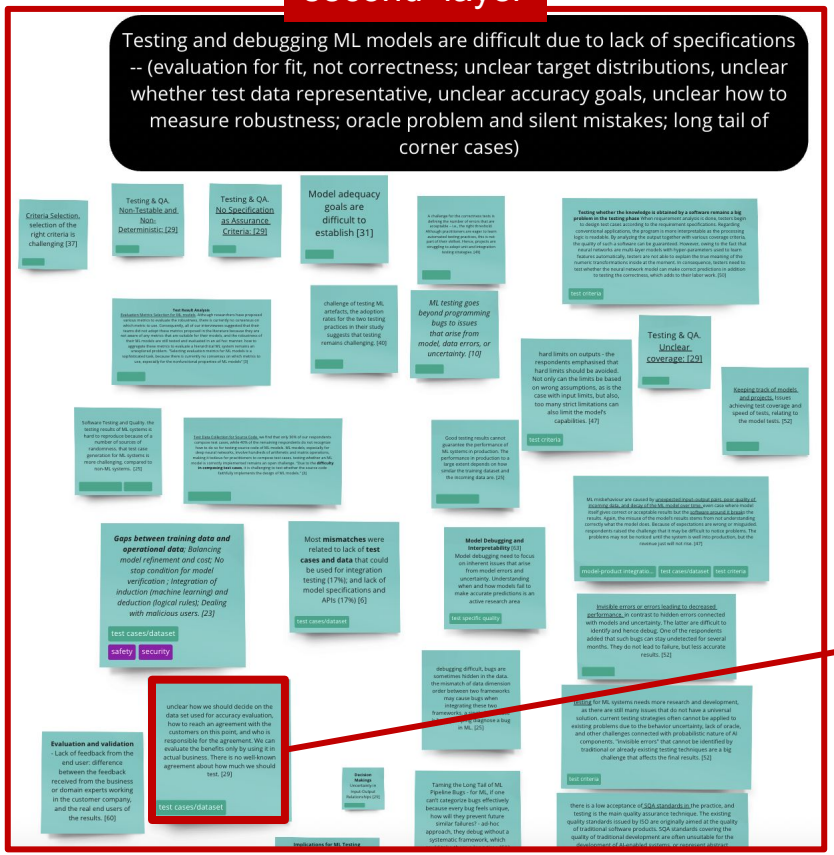
second layer – groups of common themes in the challenges

third layer — sticky notes – challenges extracted from the papers

Zooming in

second layer

Testing and debugging ML models are difficult due to lack of specifications -- (evaluation for fit, not correctness; unclear target distributions, unclear whether test data representative, unclear accuracy goals, unclear how to measure robustness; oracle problem and silent mistakes; long tail of corner cases)



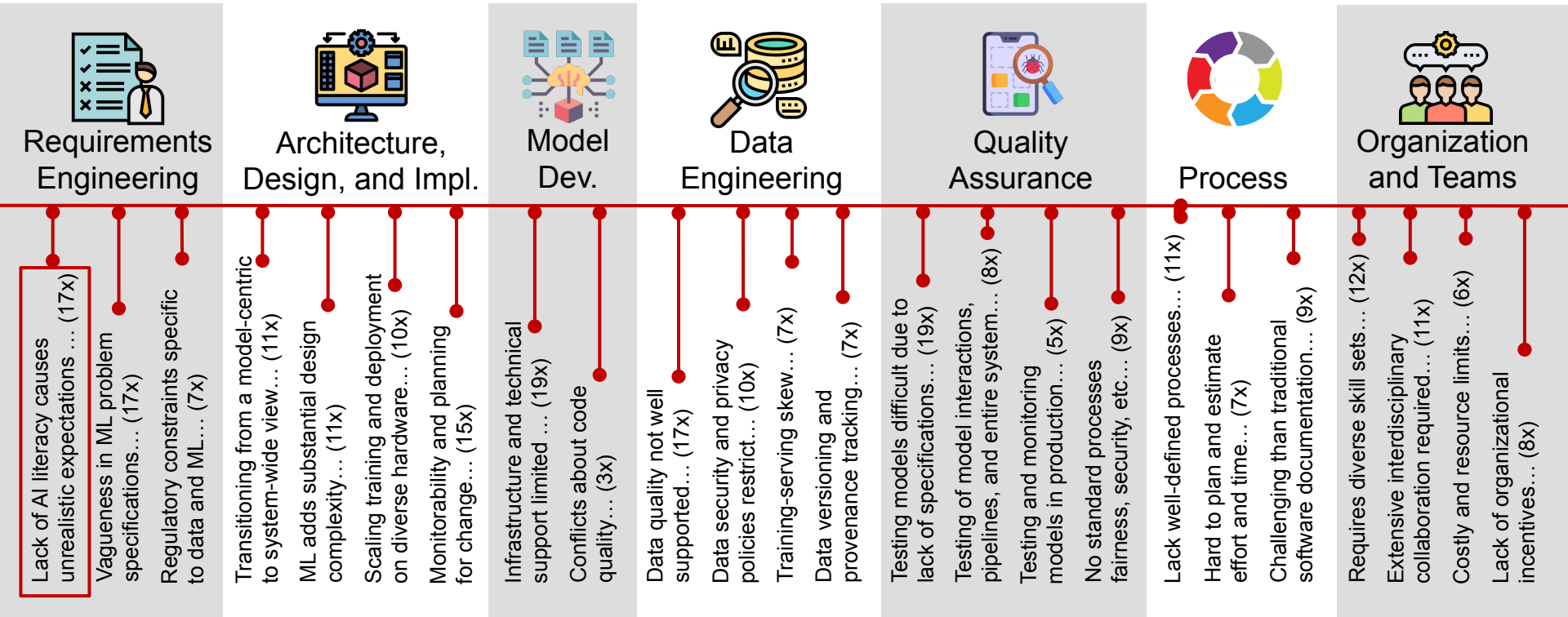
third layer: sticky note

unclear how we should decide on the data set used for accuracy evaluation, how to reach an agreement with the customers on this point, and who is responsible for the agreement. We can evaluate the benefits only by using it in actual business. There is no well-known agreement about how much we should test. [29]

test cases/dataset

Results: Challenge Themes

Overview of Challenges



Lack of AI literacy causes unrealistic expectations from customers, managers, and even other team members (17x)

Customers frequently have unrealistic expectations of ML capabilities.

- Don’t want to pay for the continuous improvement of the model
- Only consider paying for coding
- Don’t want to invest in collecting high-quality data

This is also a problem of the team members within the company itself.

Lack of AI literacy causes unrealistic expectations from customers, managers, and even other team members (17x)

[customer] believe in a “*perfect AI*” that makes no mistakes[1]

“*For this project, [the project manager] wanted to claim that we have no false positives and I was like, that’s not gonna work* [2]”

“*We designers do not understand the limits of machine learning…designers act like you can just sprinkle some data science onto a design and it will become automatically magical* [3]”

“*CEOs and executives don’t really understand what it takes [to develop and deploy ML]* [4]”

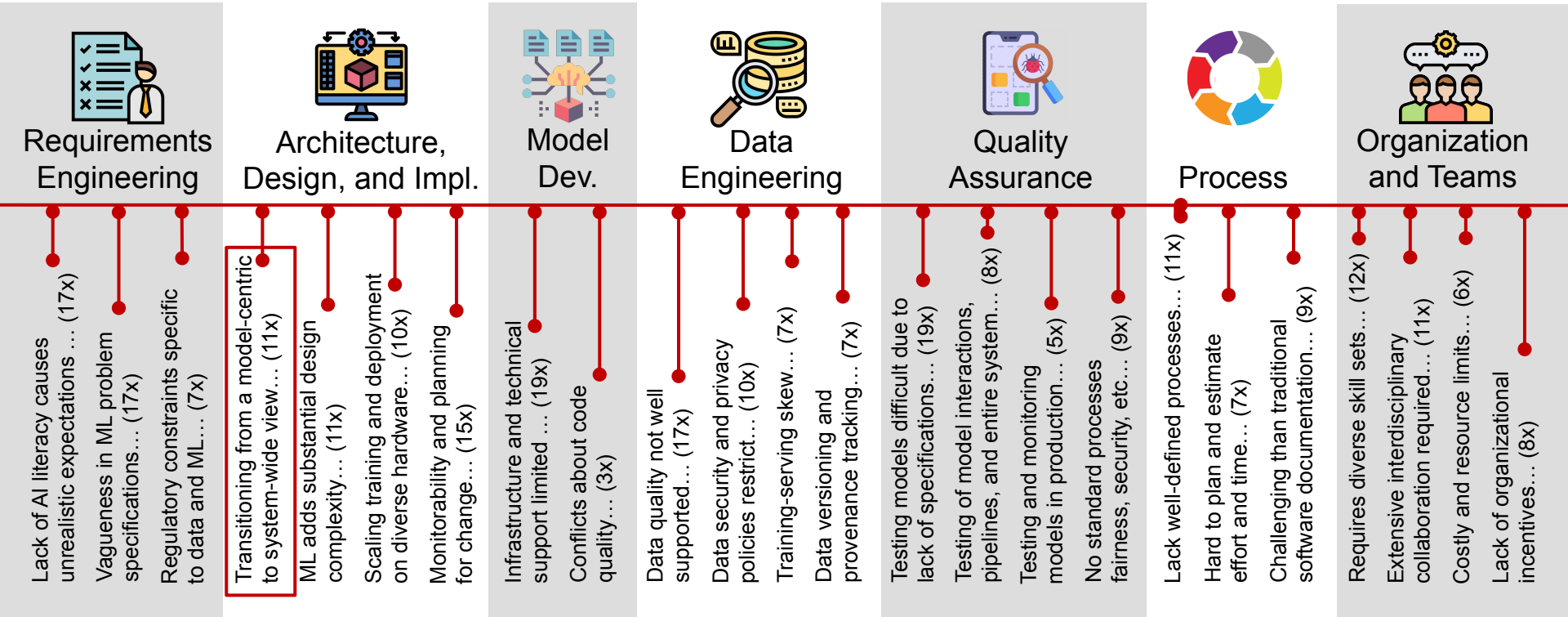
[1]Ishikawa et al. "How do engineers perceive difficulties in engineering of machine-learning systems?-questionnaire survey." In 7th CESI and 6th SER&IP, 2019.

[2]Nahar et al. "Collaboration Challenges in Building ML-enabled Systems: Communication, Documentation, Engineering, and Process." In ICSE, 2022.

[3]Dove et al. "UX design innovation: Challenges for working with machine learning as a design material." In CHI conference on human factors in computing systems, 2017.

[4]Hopkins et al. "Machine learning practices outside big tech: How resource constraints challenge responsible development." In AAAI/ACM Conference on AI, Ethics, and Society. 2021.

Overview of Challenges



Transitioning from a model-centric to a pipeline-driven or system-wide view is considered important for moving into production, but a difficult paradigm shift for many teams (11x)

Challenges in migrating from exploratory model code, often in a notebook, to deployable production-quality code in automated ML pipelines

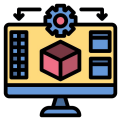
Difficulties of integrating various ML and non-ML components in a system

Overwhelming complexity of integrating many tools and frameworks

Need for engineering skills beyond the comfort zone of some data scientists

Frequent re-training and deployment of models are needed

Transitioning from a model-centric to a pipeline-driven or system-wide view is considered important for moving into production, but a difficult paradigm shift for many teams (11x)



> *"It's the difference between giving somebody a notebook…and giving a higher level tool that has a lot of built-in functionality. It's there that I see most challenges"* [1]

> *"You have this cliff edge of a gap between moving this from that Python stack, for example, into a production level ... that's the challenge for the industry"* [2]

> *"There are no tool-chains you can download in an infrastructure with deep learning like this. And we realized after the mistakes and discussions with our new IT that they didn't really have the expertise to be able to deliver this to us. So we had to create new teams, which took the responsibility of creating both the infrastructure, but also the software tool-chain to be able to train deep learning networks."* [1]

[1]Lwakatare et al. "A taxonomy of software engineering challenges for machine learning systems: An empirical investigation." In Agile Processes in Software Engineering and XP, 2019.

[2]Zdanowska et al. "A study of UX practitioners roles in designing real-world, enterprise ML systems." In CHI Conference on Human Factors in Computing Systems. 2022.

Testing and debugging ML models is difficult due to lack of specifications (19x)

> *there isn't always an actual spec of exactly what data they have, what data they think they're going to have and what they want the model to do*

> *They (clients) just throw us the data and says: look at it and maybe you can find something – [participant 7][2]*

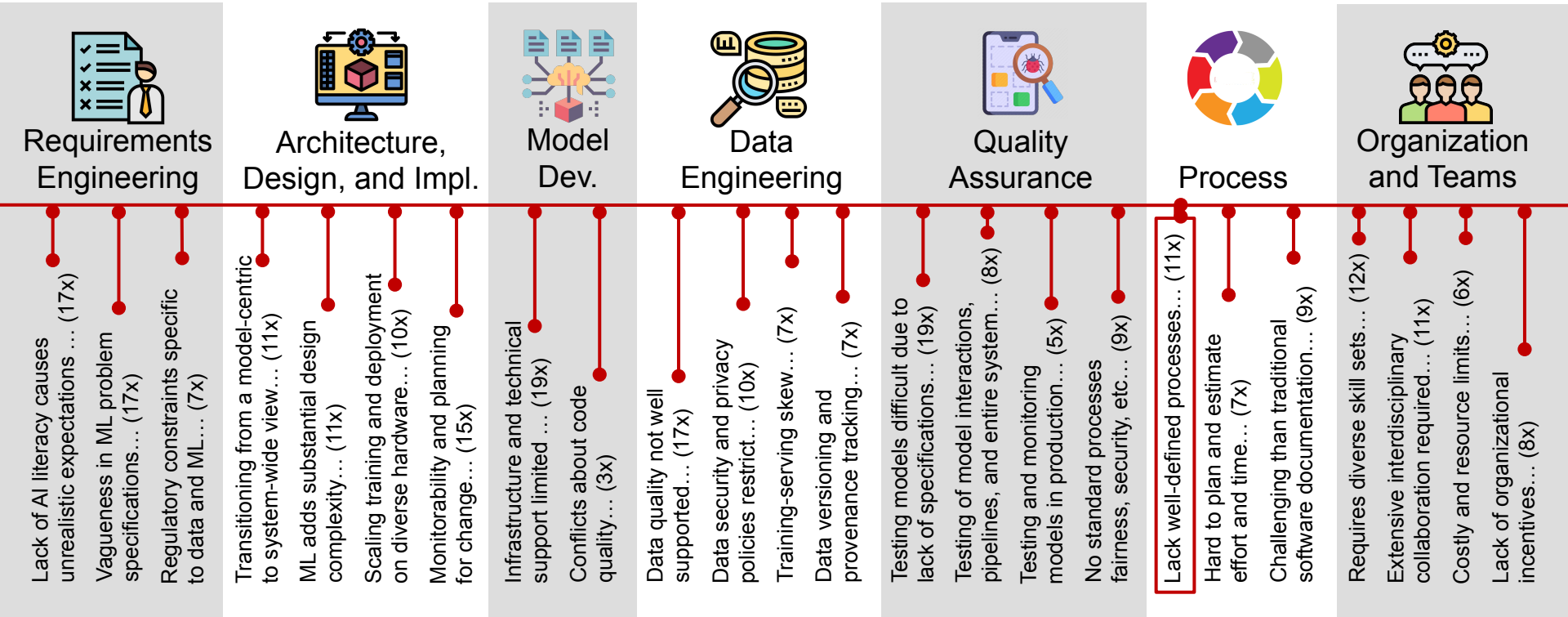
> *...there is, to my knowledge, no decisive way to ensure correctness but to leverage more data for testing predictions".*

[1]Ishikawa et al.

[2]Liu et al. "Emerging and changing tasks in the development process for machine learning systems." In Proceedings of the international conference on software and system processes, 2020.

[3]Lwakatare et al.

Overview of Challenges



Development of products with ML component(s) is often adhoc, lacking well-defined processes (11x)

Practitioners struggle finding a good process for developing ML components and products around

- ad-hoc strategies + lack of good engineering practices

> *"There are projects that I do one thing, and there are others that I do not do… there is no well-defined process here"* [1]

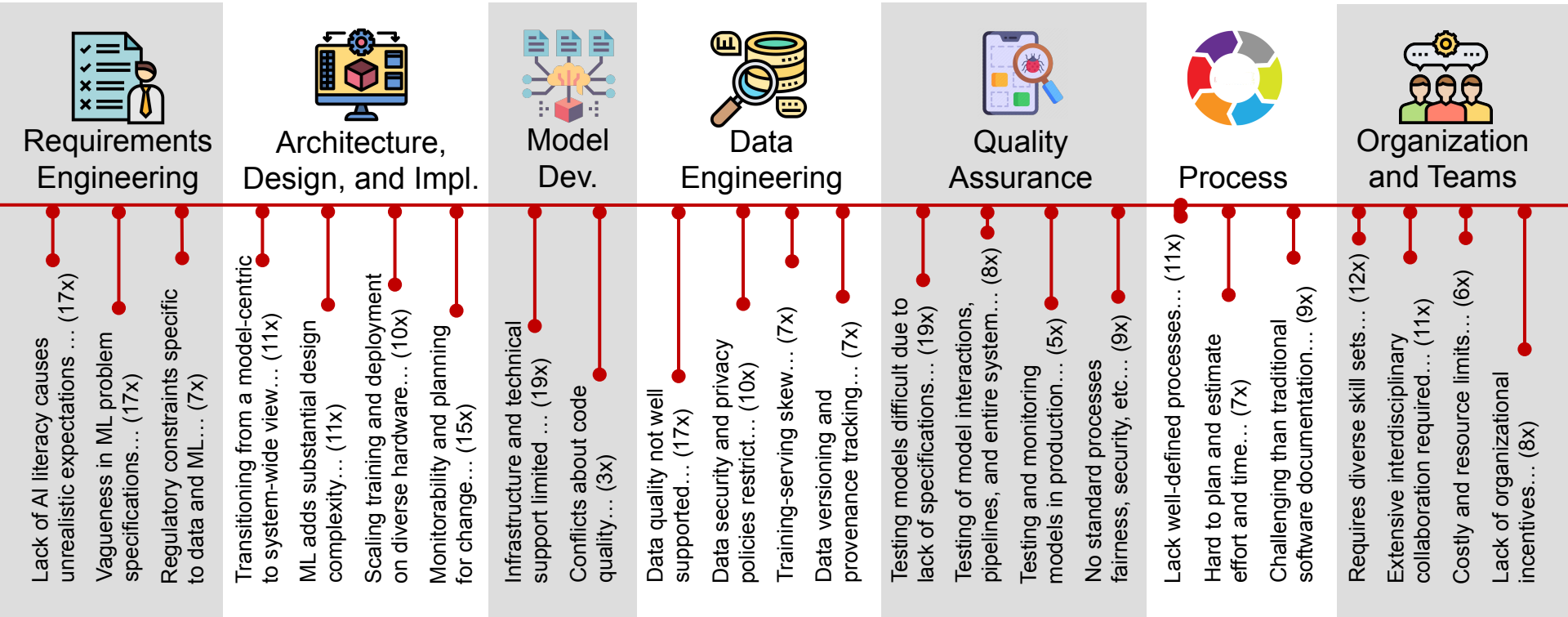
Waterfall: poor fit for exploratory development work

Agile: sprint timeline too fixed and short

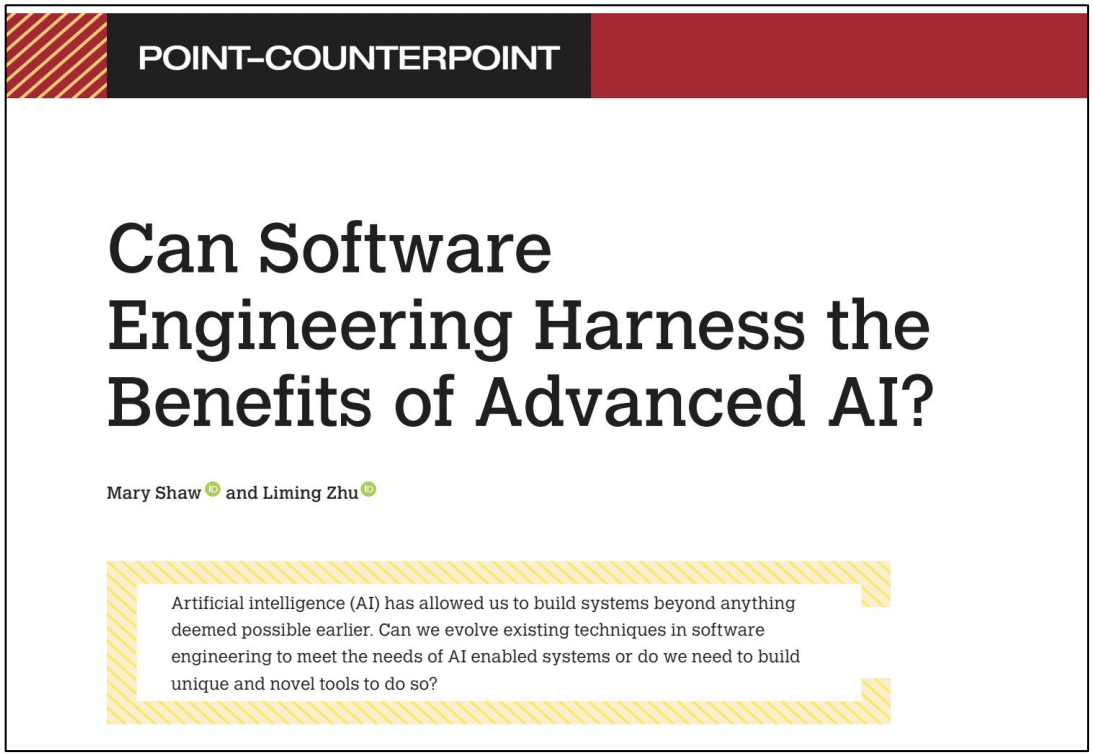
+ hard to set expectations for each sprint with unclear project objectives at the beginning

[1] de Souza et al. "Understanding Development Process of Machine Learning Systems: Challenges and Solutions." In ESEM, 2019.

Overview of Challenges



Old Or New Challenges?



POINT–COUNTERPOINT

Can Software Engineering Harness the Benefits of Advanced AI?

Mary Shaw and Liming Zhu

Artificial intelligence (AI) has allowed us to build systems beyond anything deemed possible earlier. Can we evolve existing techniques in software engineering to meet the needs of AI enabled systems or do we need to build unique and novel tools to do so?

Shaw, Mary, and Liming Zhu. "Can Software Engineering Harness the Benefits of Advanced AI?." IEEE Software 39, no. 6 (2022): 99-104.

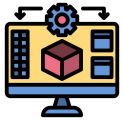
Old Or New Challenges?


Requirements


Process


Collaboration

Old, but seem to be more problematic


Architecture


Quality Assurance

Not new, but problems of different nature with ML


Model-related

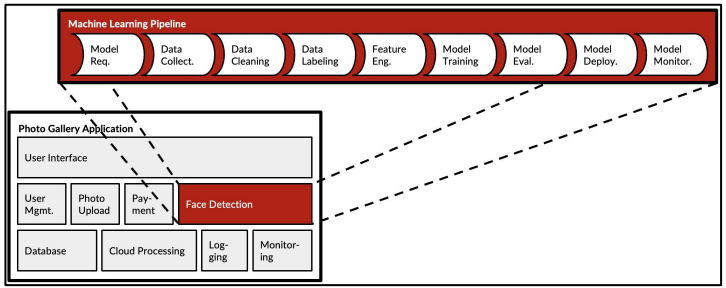

Data-related

Not new, but much more important now

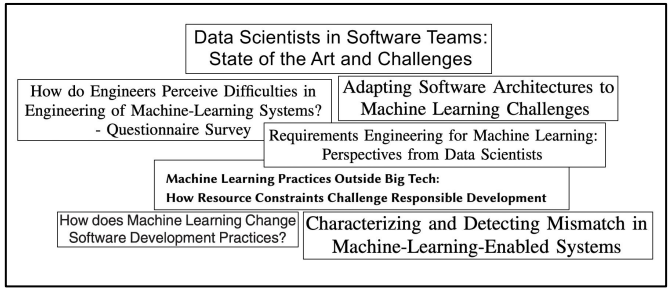
We know the challenges, time to work on solutions!

Summary

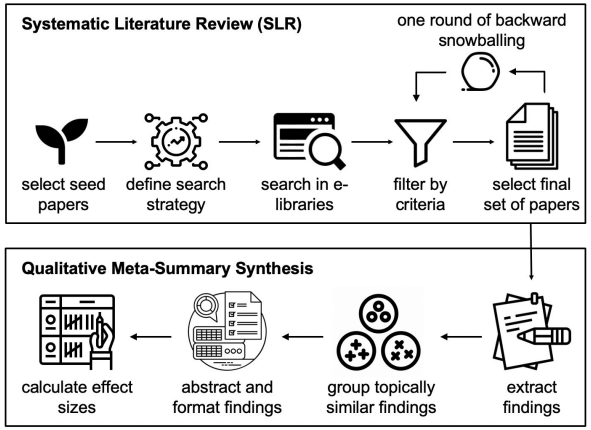
ML as a Component of a Product



Lots of Pain-point Papers for ML Products



We Conducted a Meta-summary



We Summarized and Presented the Challenges