Building A Fun-Sized MLOps Stack From Scratch

Mikiko

Head of MLOps

Powered By:

featureform

Goals

- What are the main problems MLOps tries to solve.
- What are the most common tools being used & their drawbacks.
- What are some OSS projects & tools that have been developed in the past 2-3 years and how do they solve some of the pain points of the prior tools.
- What is the realistic roadmap for companies that are forever “not-Google” scale but want to continue improving their data and ML maturity.

How?

🤷🏻‍♀️

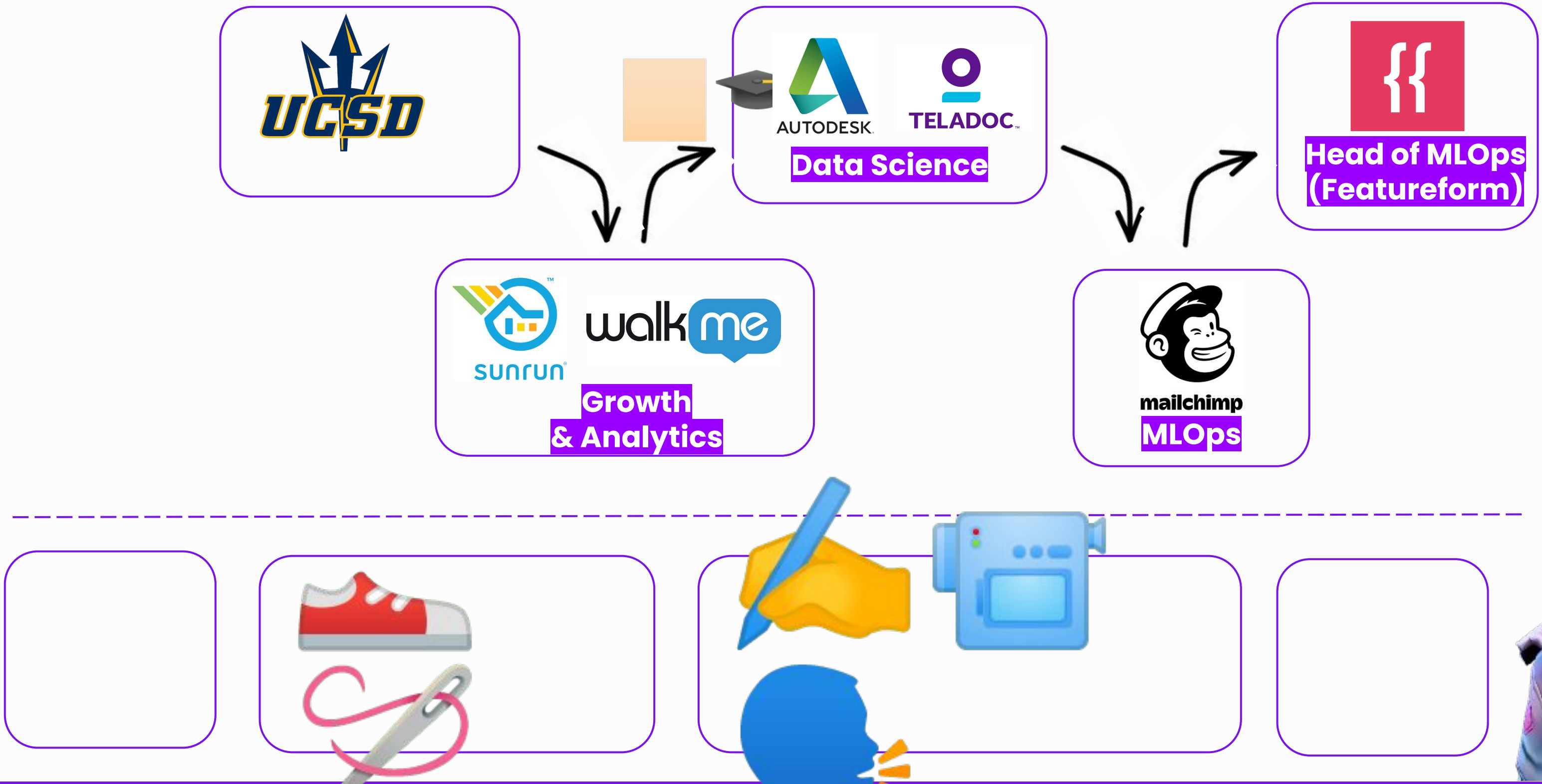
By

- Describing the original promises of MLOps (& the current shortfalls).
- Understanding the “Jobs-To-Be-Done” of Data Scientists (& how the current ecosystem supports them).
- Describe the pain-points of the Solo Data Scientist, the SMB Data Science Team, & the areas of opportunity for Enterprises.
- Propose stacks that can be easily implemented in a relatively short period (sometimes even a day!).

Who Am I?

💁🏻‍♀️

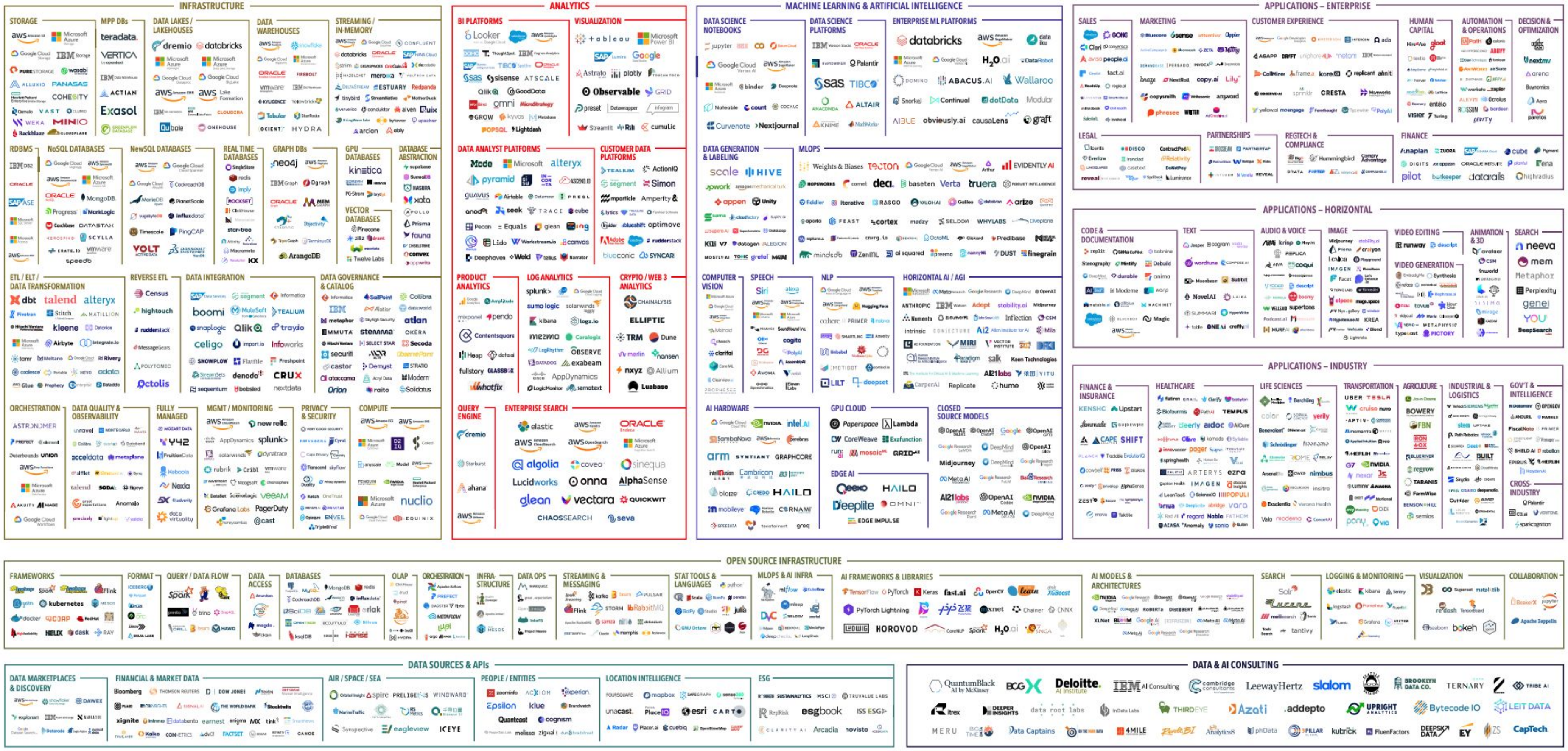
Mikiko Bazeley (But Call Me Mickey 🐭)



What’s The Current Landscape Look Like?

🗺️

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



INFRASTRUCTURE

STORAGE · MPP DBs · DATA LAKES / LAKEHOUSES · DATA WAREHOUSES · STREAMING / IN-MEMORY · RDBMS · NoSQL DATABASES · NewSQL DATABASES · REAL TIME DATABASES · GRAPH DBs · GPU DATABASES · DATABASE ABSTRACTION · VECTOR DATABASES · ETL / ELT / DATA TRANSFORMATION · REVERSE ETL · DATA INTEGRATION · DATA GOVERNANCE & CATALOG · ORCHESTRATION · DATA QUALITY & OBSERVABILITY · FULLY MANAGED · MGMT / MONITORING · PRIVACY & SECURITY · COMPUTE

ANALYTICS

BI PLATFORMS · VISUALIZATION · DATA ANALYST PLATFORMS · CUSTOMER DATA PLATFORMS · PRODUCT ANALYTICS · LOG ANALYTICS · CRYPTO / WEB 3 ANALYTICS · QUERY ENGINE · ENTERPRISE SEARCH

MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

DATA SCIENCE NOTEBOOKS · DATA SCIENCE PLATFORMS · ENTERPRISE ML PLATFORMS · DATA GENERATION & LABELING · MLOPS · COMPUTER VISION · SPEECH · NLP · HORIZONTAL AI / AGI · AI HARDWARE · GPU CLOUD · EDGE AI · CLOSED SOURCE MODELS

APPLICATIONS – ENTERPRISE

SALES · MARKETING · CUSTOMER EXPERIENCE · HUMAN CAPITAL · AUTOMATION & OPERATIONS · DECISION & OPTIMIZATION · LEGAL · PARTNERSHIPS · REGTECH & COMPLIANCE · FINANCE

APPLICATIONS – HORIZONTAL

CODE & DOCUMENTATION · TEXT · AUDIO & VOICE · IMAGE · VIDEO EDITING · VIDEO GENERATION · ANIMATION & 3D · SEARCH

APPLICATIONS – INDUSTRY

FINANCE & INSURANCE · HEALTHCARE · LIFE SCIENCES · TRANSPORTATION · AGRICULTURE · INDUSTRIAL & LOGISTICS · GOV'T & INTELLIGENCE · CROSS-INDUSTRY

OPEN SOURCE INFRASTRUCTURE

FRAMEWORKS · FORMAT · QUERY / DATA FLOW · DATA ACCESS · DATABASES · OLAP · ORCHESTRATION · INFRA-STRUCTURE · DATA OPS · STREAMING & MESSAGING · STAT TOOLS & LANGUAGES · MLOPS & AI INFRA · AI FRAMEWORKS & LIBRARIES · AI MODELS & ARCHITECTURES · SEARCH · LOGGING & MONITORING · VISUALIZATION · COLLABORATION

DATA SOURCES & APIs

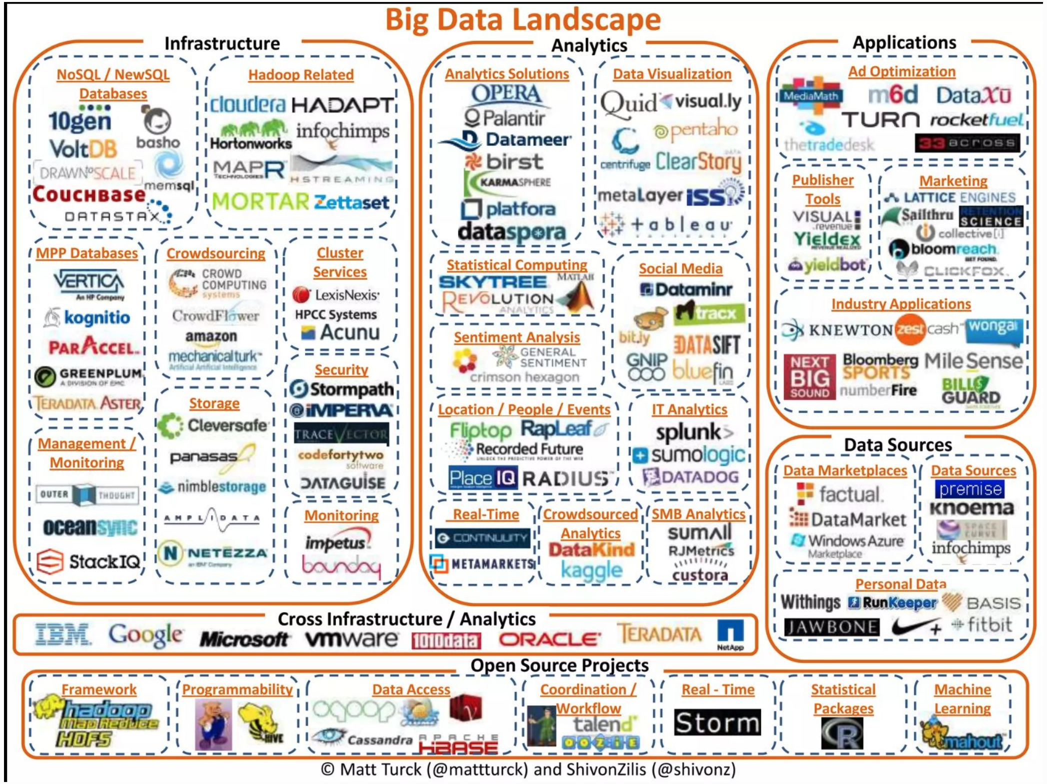
DATA MARKETPLACES & DISCOVERY · FINANCIAL & MARKET DATA · AIR / SPACE / SEA · PEOPLE / ENTITIES · LOCATION INTELLIGENCE · ESG

DATA & AI CONSULTING

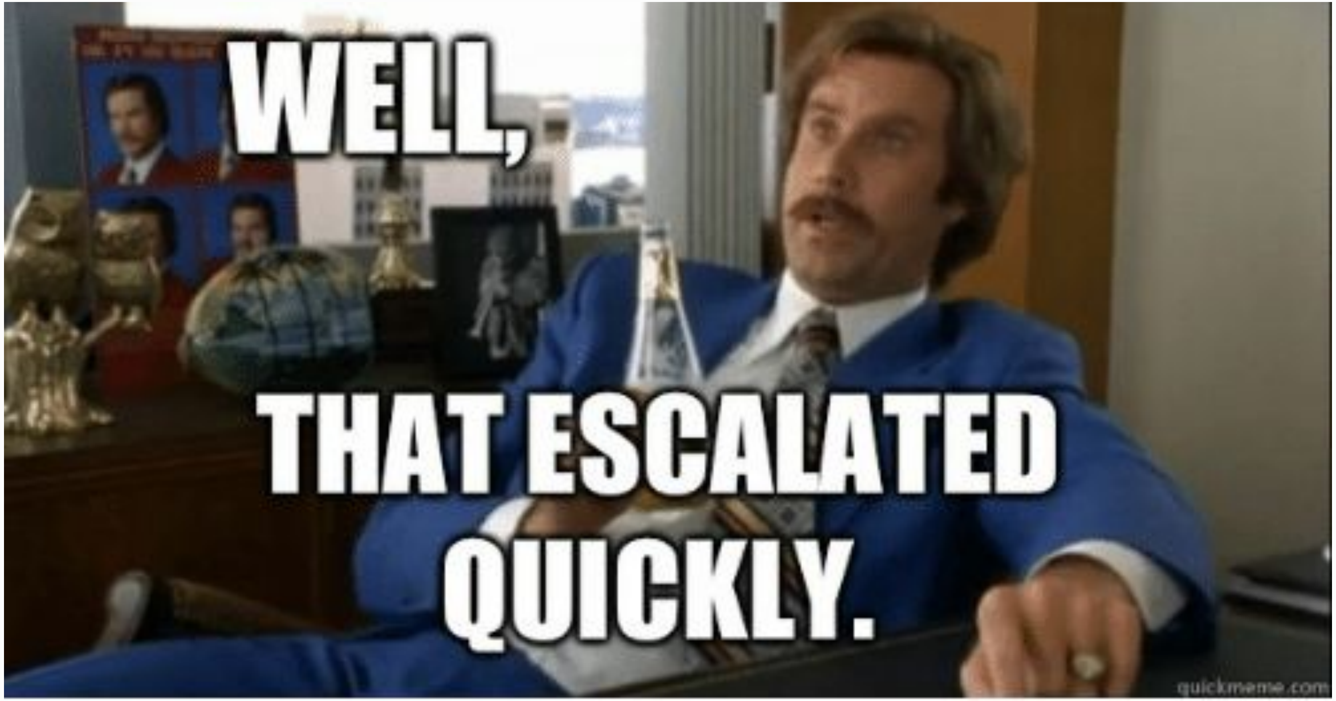
Version 1.0 - Feb 2023 © Matt Turck (@mattturck), Kevin Zhang (@ykevinzhang) & FirstMark (@firstmarkcap) Blog post: mattturck.com/MAD2023 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2023@firstmarkcap.com

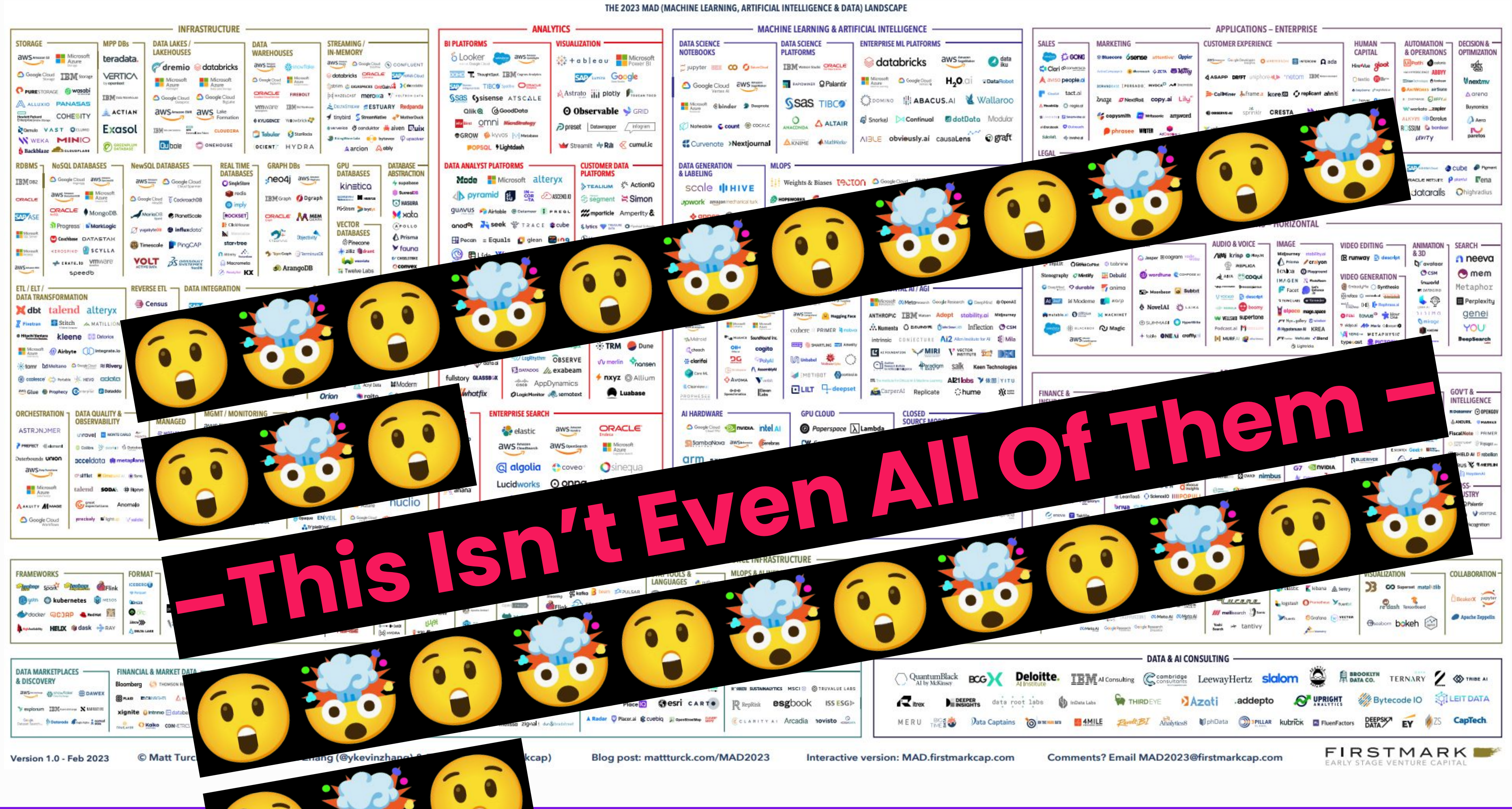
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

2012



© Matt Turck (@mattturck) and ShivonZilis (@shivonz)


WELL,
THAT ESCALATED
QUICKLY.
quickmeme.com



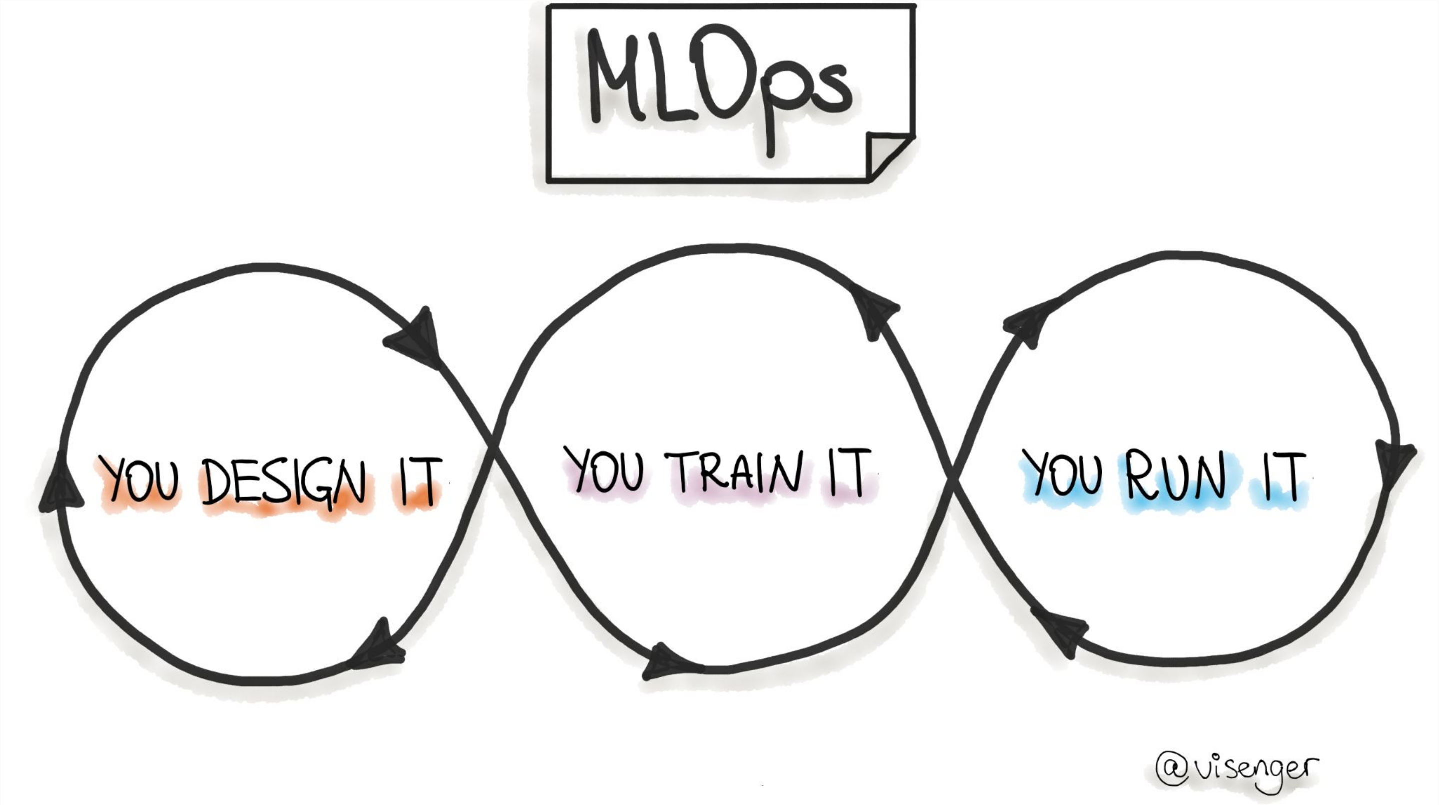
—This Isn't Even All Of Them —

featureform

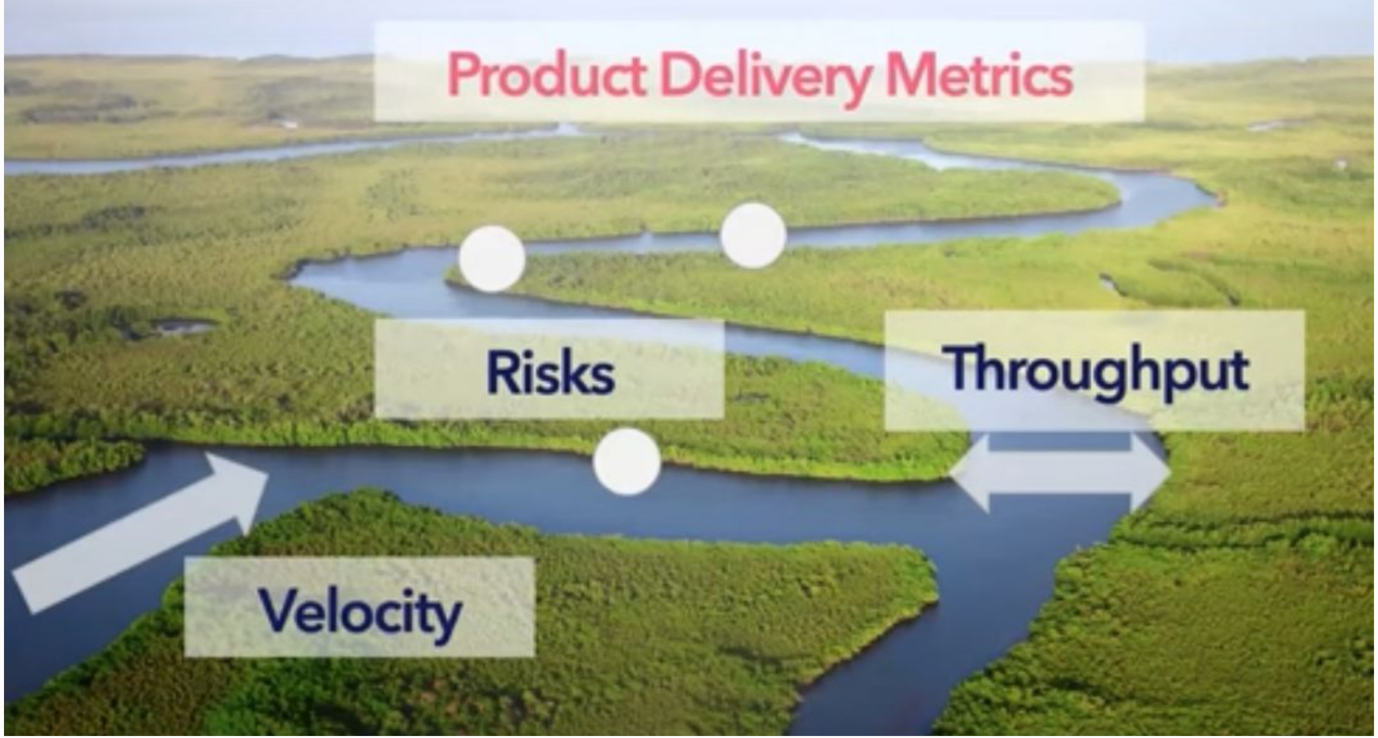
What Was The Original Promise of MLOps?

👐

Design => Train => Run



The Three Dimensions: **Velocity, Throughput, Risk**



Source: "Humanitec -- Key DevOps metrics to improve your engineering setup"

We Failed

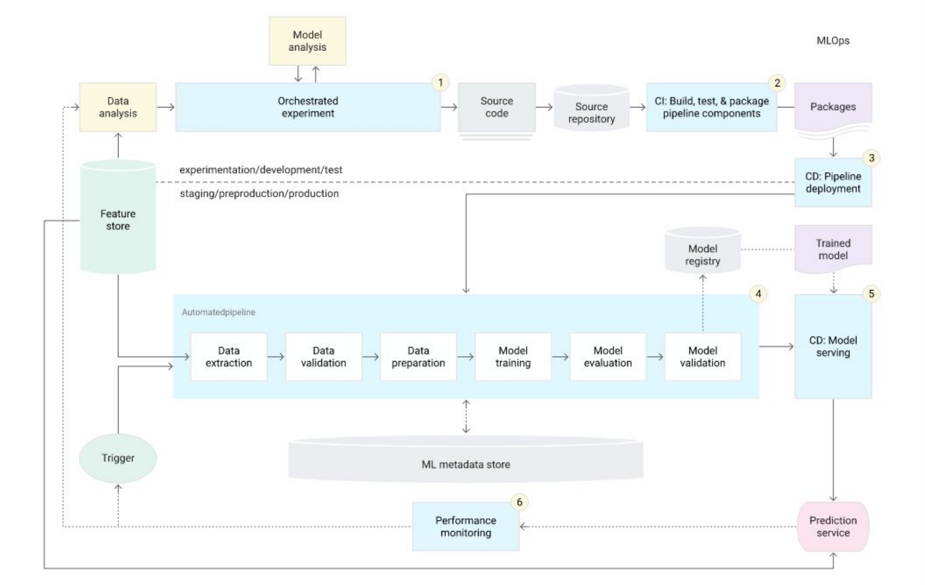
The OG Users of

MLOps:

👩🏻‍💻 Data Scientists

👩🏻‍🔬

Trying To Mimic The Heavyweights



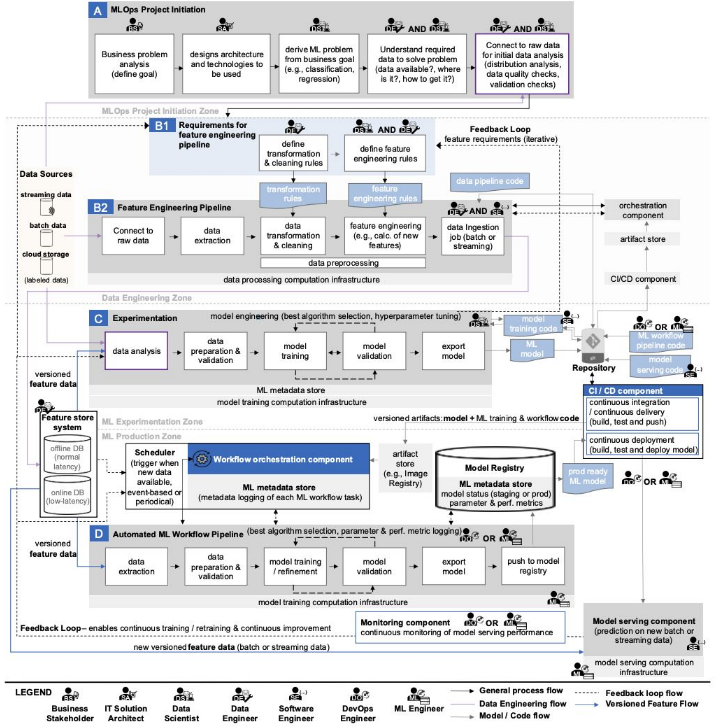
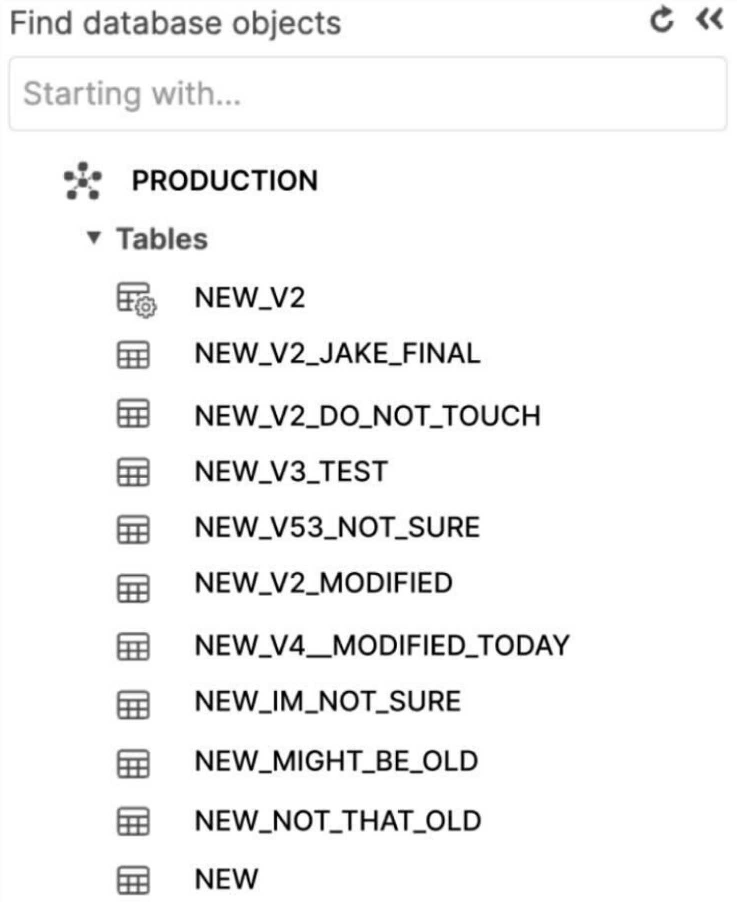

A
MLOps Project Initiation
BS
SA
DS
DE AND DS
DE AND DS
Business problem analysis (define goal)
designs architecture and technologies to be used
derive ML problem from business goal (e.g., classification, regression)
Understand required data to solve problem (data available?, where is it?, how to get it?)
Connect to raw data for initial data analysis (distribution analysis, data quality checks, validation checks)
MLOps Project Initiation Zone
B1
Requirements for feature engineering pipeline
DE
DS AND DE
define transformation & cleaning rules
define feature engineering rules
Feedback Loop
feature requirements (iterative)
Data Sources
streaming data
batch data
cloud storage
(labeled data)
transformation rules
feature engineering rules
data pipeline code
B2
Feature Engineering Pipeline
DE AND SE
Connect to raw data
data extraction
data transformation & cleaning
feature engineering (e.g., calc. of new features)
data Ingestion job (batch or streaming)
data preprocessing
data processing computation infrastructure
orchestration component
artifact store
CI/CD component
Data Engineering Zone
C
Experimentation
model engineering (best algorithm selection, hyperparameter tuning)
DS
model training code
SE
ML model
DO OR ML
ML workflow pipeline code
model serving code
SE
Repository
data analysis
data preparation & validation
model training
model validation
export model
versioned feature data
ML metadata store
model training computation infrastructure
DE
Feature store system
offline DB (normal latency)
online DB (low-latency)
ML Experimentation Zone
ML Production Zone
versioned artifacts: model + ML training & workflow code
CI / CD component
continuous integration / continuous delivery (build, test and push)
continuous deployment (build, test and deploy model)
Scheduler (trigger when new data available, event-based or periodical)
Workflow orchestration component
ML metadata store (metadata logging of each ML workflow task)
artifact store (e.g., Image Registry)
Model Registry
ML metadata store
model status (staging or prod)
parameter & perf. metrics
prod ready ML model
DO OR ML
D
Automated ML Workflow Pipeline
(best algorithm selection, parameter & perf. metric logging)
DO OR ML
versioned feature data
data extraction
data preparation & validation
model training / refinement
model validation
export model
push to model registry
model training computation infrastructure
ML
Feedback Loop – enables continuous training / retraining & continuous improvement
Monitoring component
DO OR ML
continuous monitoring of model serving performance
Model serving component (prediction on new batch or streaming data)
SE
new versioned feature data (batch or streaming data)
model serving computation infrastructure
ML
LEGEND
BS Business Stakeholder
SA IT Solution Architect
DS Data Scientist
DE Data Engineer
SE Software Engineer
DO DevOps Engineer
ML ML Engineer
General process flow
Data Engineering flow
Model / Code flow
Feedback loop flow
Versioned Feature Flow

Figure 4. End-to-end MLOps architecture and workflow with functional components and roles



Find database objects

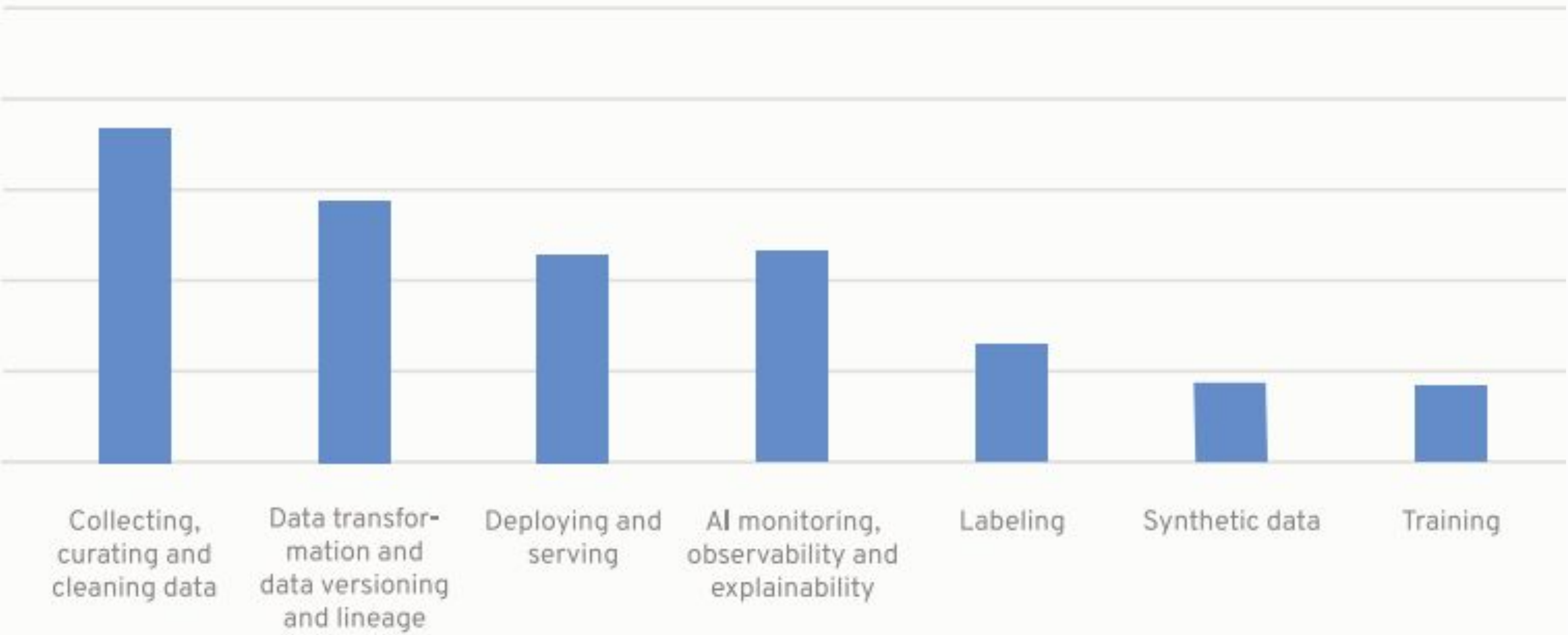
Starting with...

PRODUCTION

▼ Tables

- NEW_V2
- NEW_V2_JAKE_FINAL
- NEW_V2_DO_NOT_TOUCH
- NEW_V3_TEST
- NEW_V53_NOT_SURE
- NEW_V2_MODIFIED
- NEW_V4__MODIFIED_TODAY
- NEW_IM_NOT_SURE
- NEW_MIGHT_BE_OLD
- NEW_NOT_THAT_OLD
- NEW

Where have you faced the biggest challenges in productionizing models?



Ouch!

Creating Meaningful Data Alerts is Challenging

Need to handle evolving data & Data sources can suddenly change without announcement

Insufficient data documentation

Training data is often insufficient and incomplete

“A Hotbed of Bias.” Efforts to Assess, Prevent, & Mitigate Bias

Industry-Classroom Mismatch.

“Experiment, Iterate, See We’re Getting Closer.” A Model Is Never Finished

“Data Literacy Is Not a Silver Bullet.” On Communication & Collaboration

Dev-Prod Env Mismatch

Undocumented Tribal Knowledge

Product requirements require input from the model team & Lack of ML knowledge in managers

Industry-Classroom Mismatch.

What We

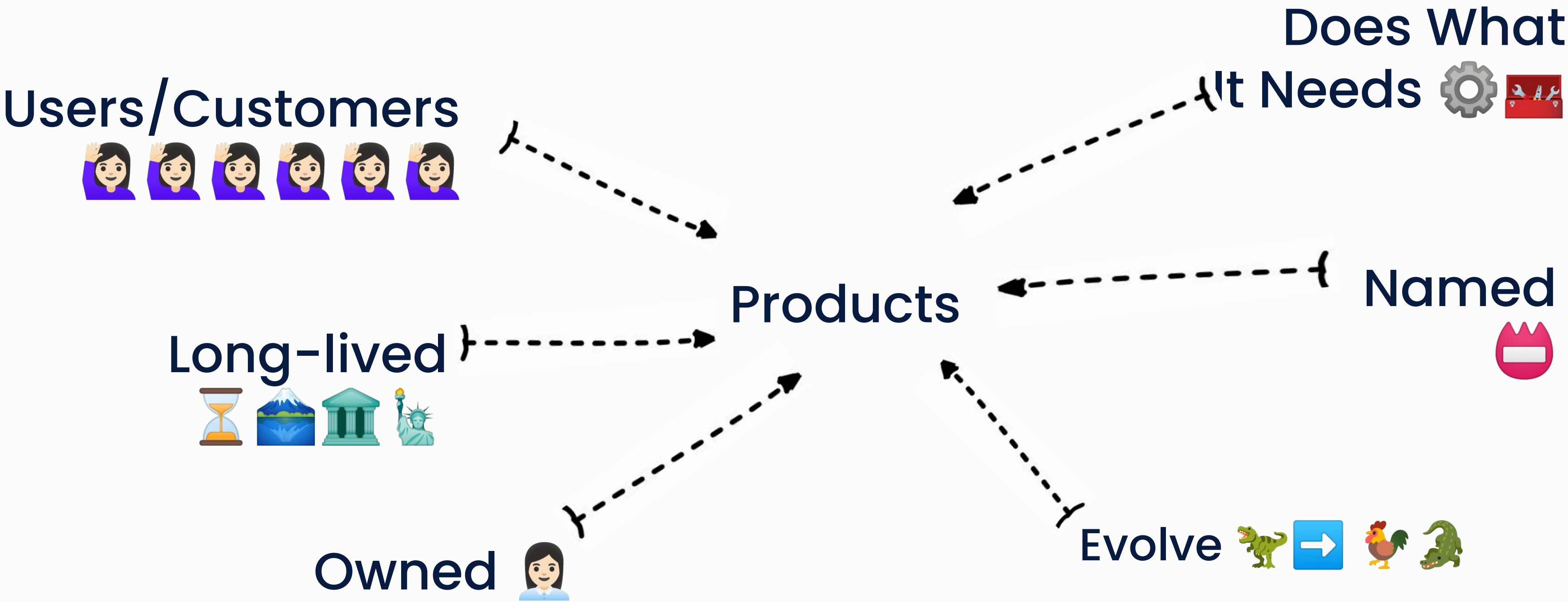
Should Be Doing

💪

📌 Best Practice #1

Treating The ML Platform As A Product

Characteristics of a Product



Users/Customers
🙋🏻‍♀️🙋🏻‍♀️🙋🏻‍♀️🙋🏻‍♀️🙋🏻‍♀️🙋🏻‍♀️

Does What It Needs ⚙️🧰

Products

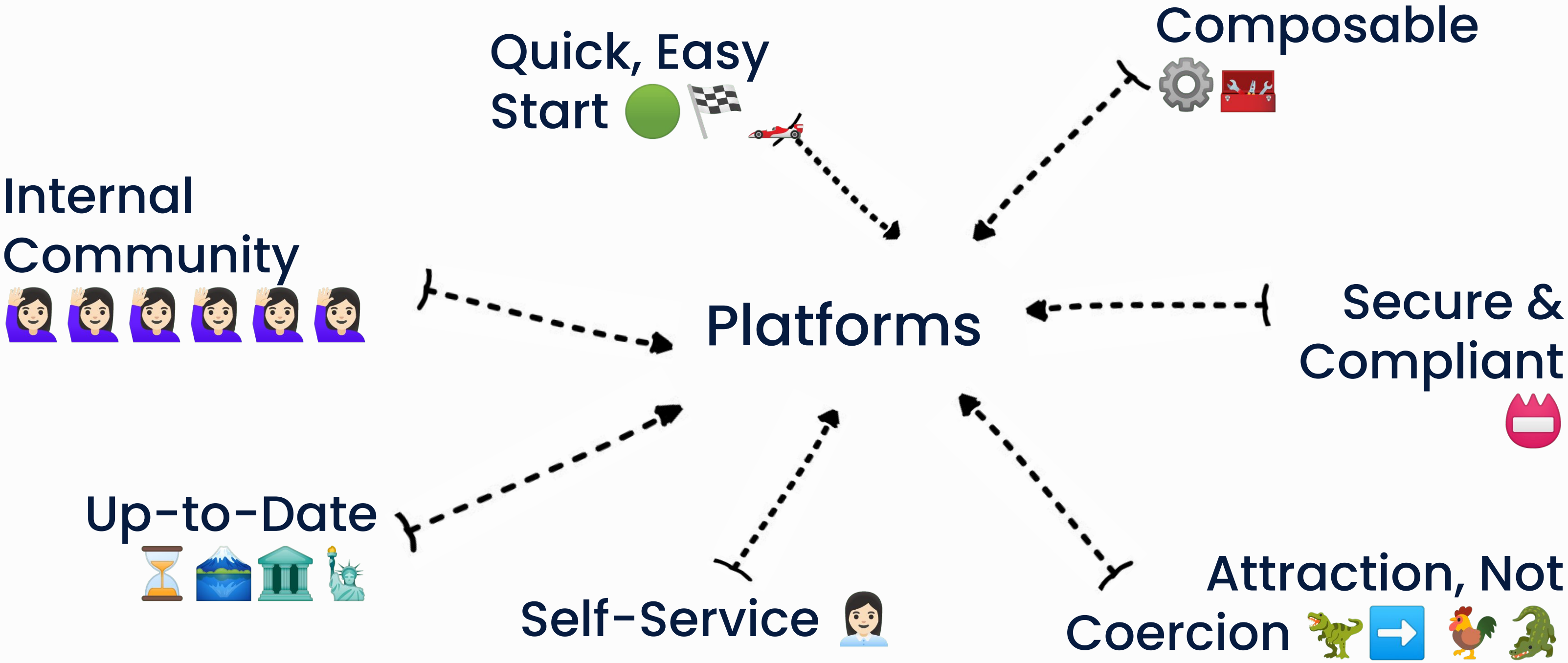
Named
📛

Long-lived
⏳🗻🏛️🗽

Owned 👩🏻‍💼

Evolve 🦖➡️ 🐓🐊

Applied to Platforms



Quick, Easy Start 🟢🏁🏎️

Composable ⚙️🧰

Internal Community 🙋🏻‍♀️🙋🏻‍♀️🙋🏻‍♀️🙋🏻‍♀️🙋🏻‍♀️🙋🏻‍♀️

Platforms

Secure & Compliant 📛

Up-to-Date ⌛🗻🏛️🗽

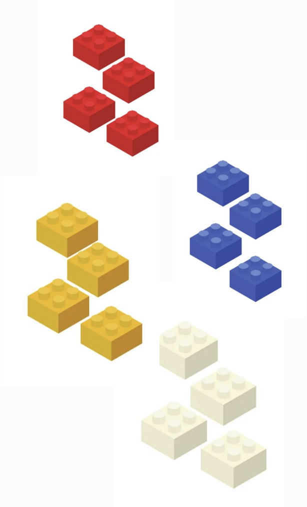
Self-Service 👩🏻‍💼

Attraction, Not Coercion 🦖➡️🐓🐊

Measuring Data Science DevExp

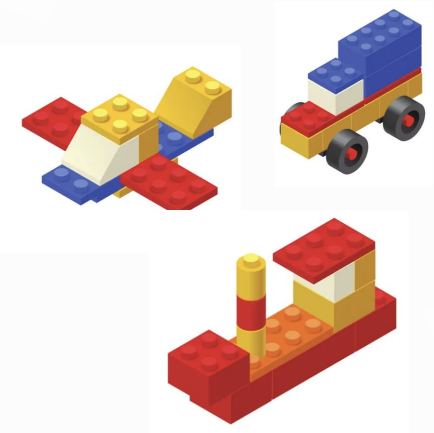
What We Should Care About	What We Should Measure (Examples)
Product (Activation, Engagement, Adoption)	Weekly Active Users, Engagement, Adoption Rate, etc
User Satisfaction	NPS, WAU Retention, Sessions Per User, # Tickets, etc
Platform Performance (Reliability, Availability, Scalability)	SLOs, Latency, # Incidents, etc

Empire State of Mind



APIs, tools, services, knowledge & support

Different Types of Platform Teams

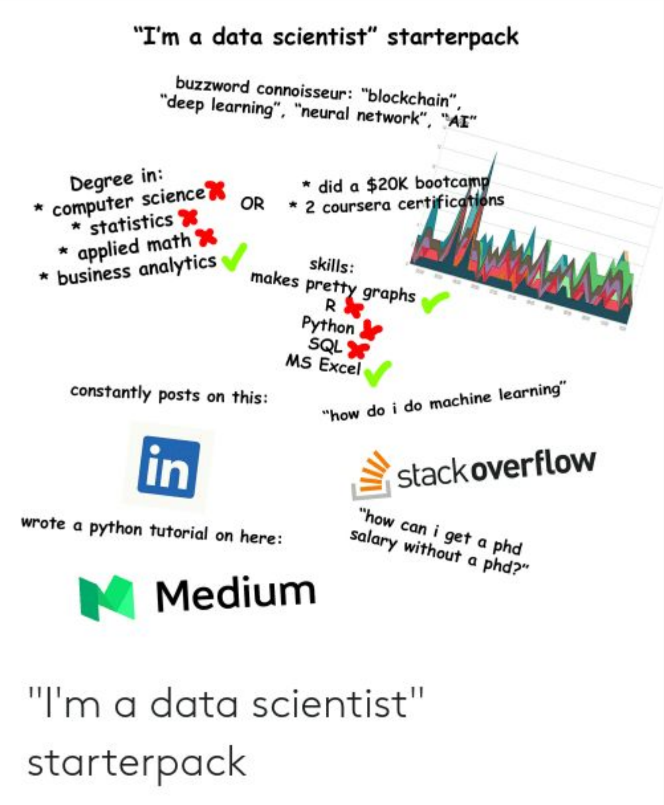


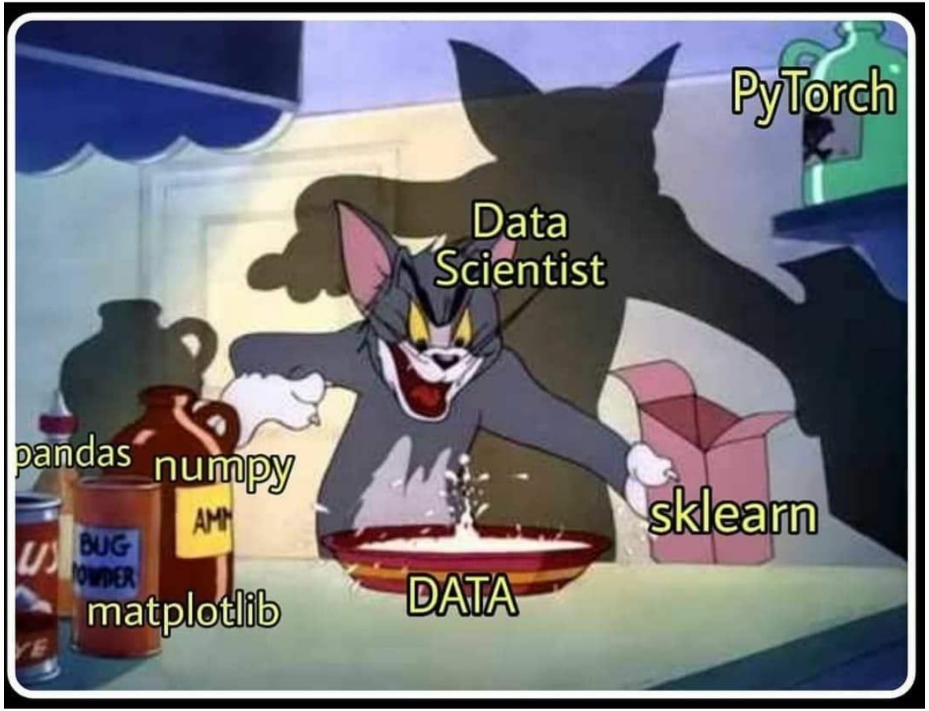
Unified Developer Experience Characterized By Ease of Use + High Bar

📌 Best Practice #2

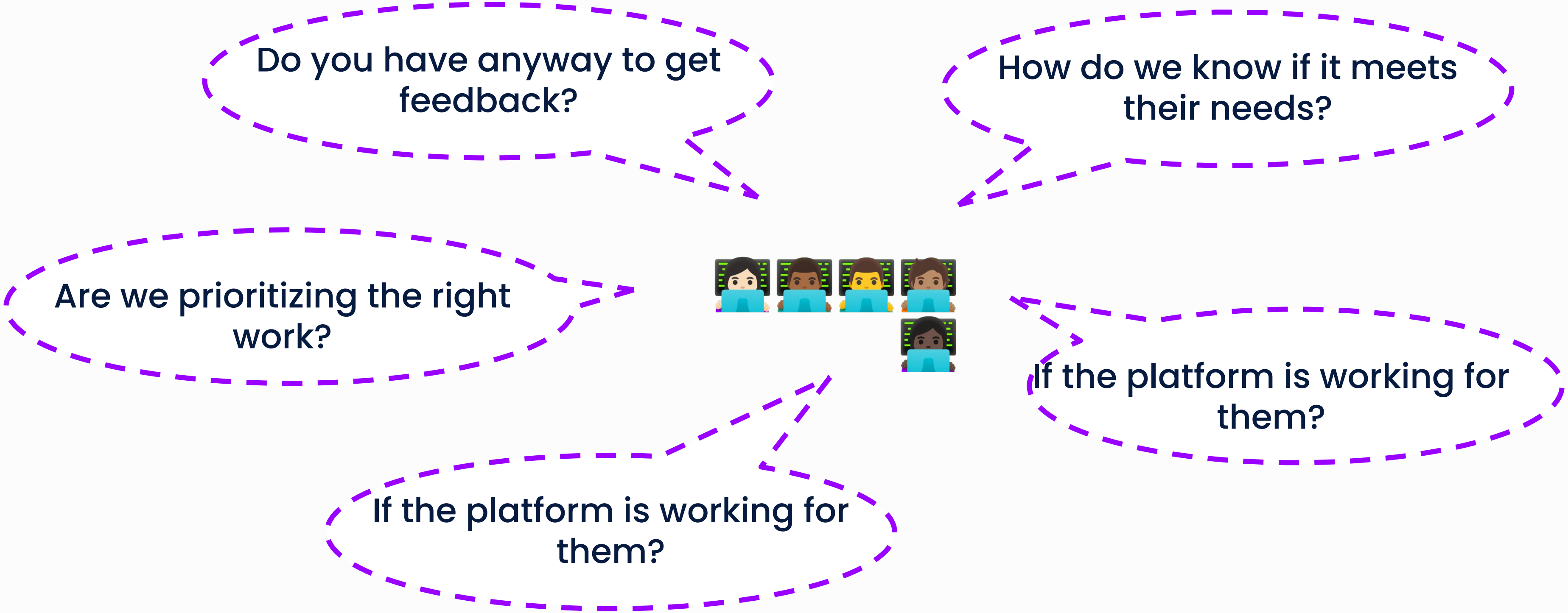
Prioritizing Enablement, The Last Mile of True Platform Adoption

The Stereotypes & Tropes


"I'm a data scientist" starterpack
buzzword connoisseur: "blockchain", "deep learning", "neural network", "AI"
Degree in:
* computer science ✗
* statistics ✗
* applied math ✗
* business analytics ✓
OR
* did a $20K bootcamp
* 2 coursera certifications
skills:
makes pretty graphs ✓
R ✗
Python ✗
SQL ✗
MS Excel ✓
constantly posts on this:
in
"how do i do machine learning"
stackoverflow
wrote a python tutorial on here:
"how can i get a phd salary without a phd?"
Medium
"I'm a data scientist" starterpack



Being User Centric



Do you have anyway to get feedback?

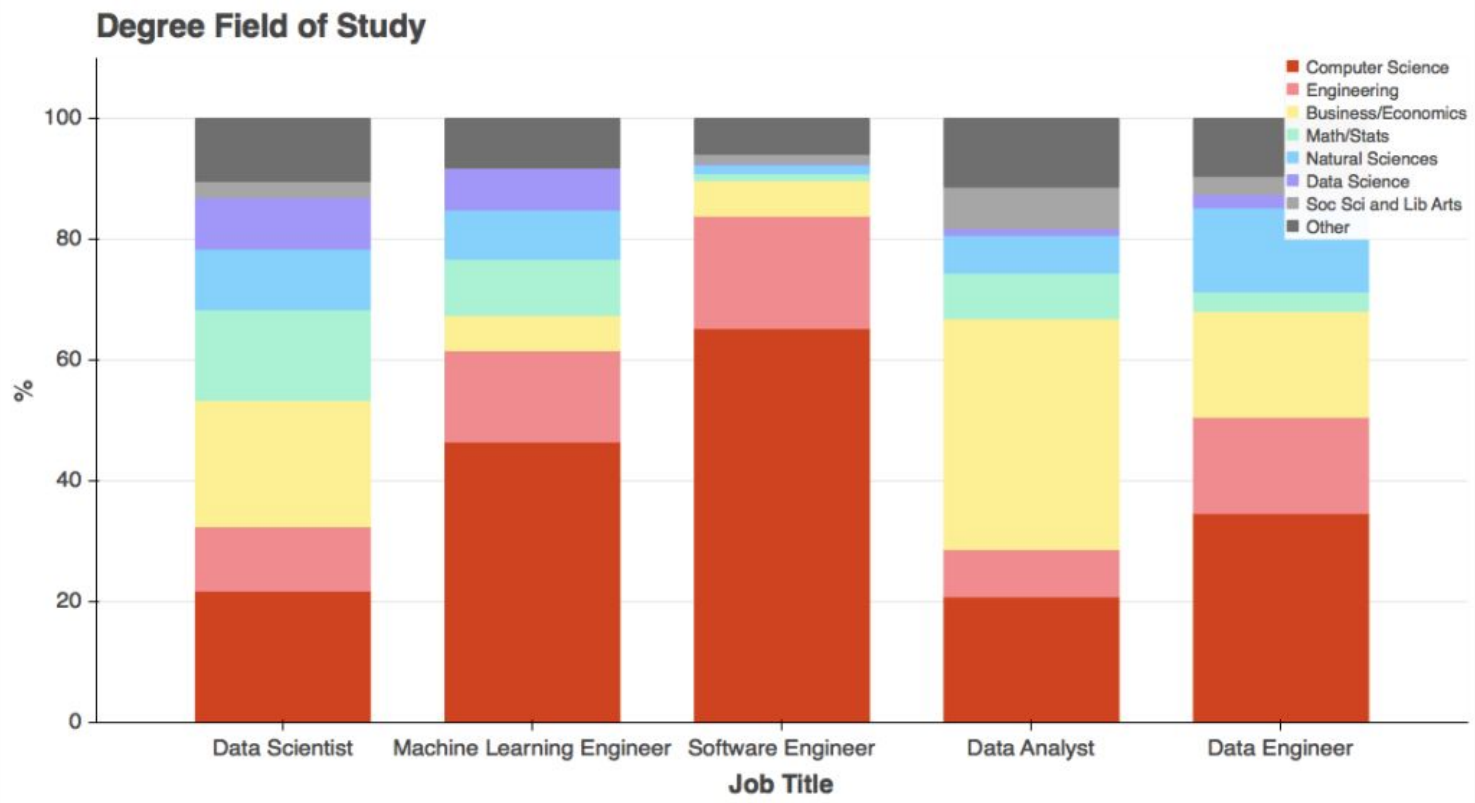
How do we know if it meets their needs?

Are we prioritizing the right work?

If the platform is working for them?

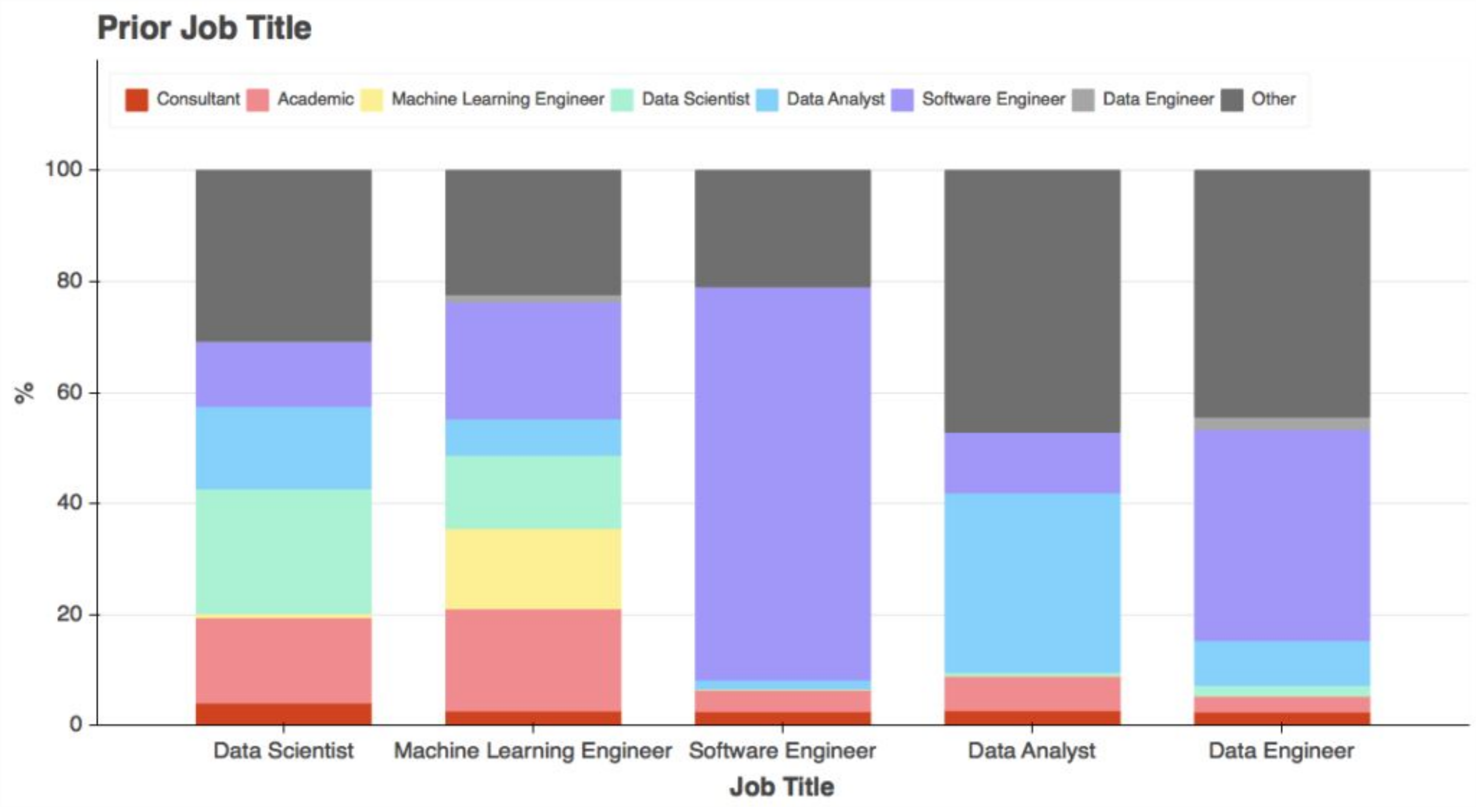
If the platform is working for them?

Do You Know Who Your Customer Is?



Indeed: Where Do Data Scientists Come From?

Do You Know Who Your Customer Is?



Indeed: Where Do Data Scientists Come From?

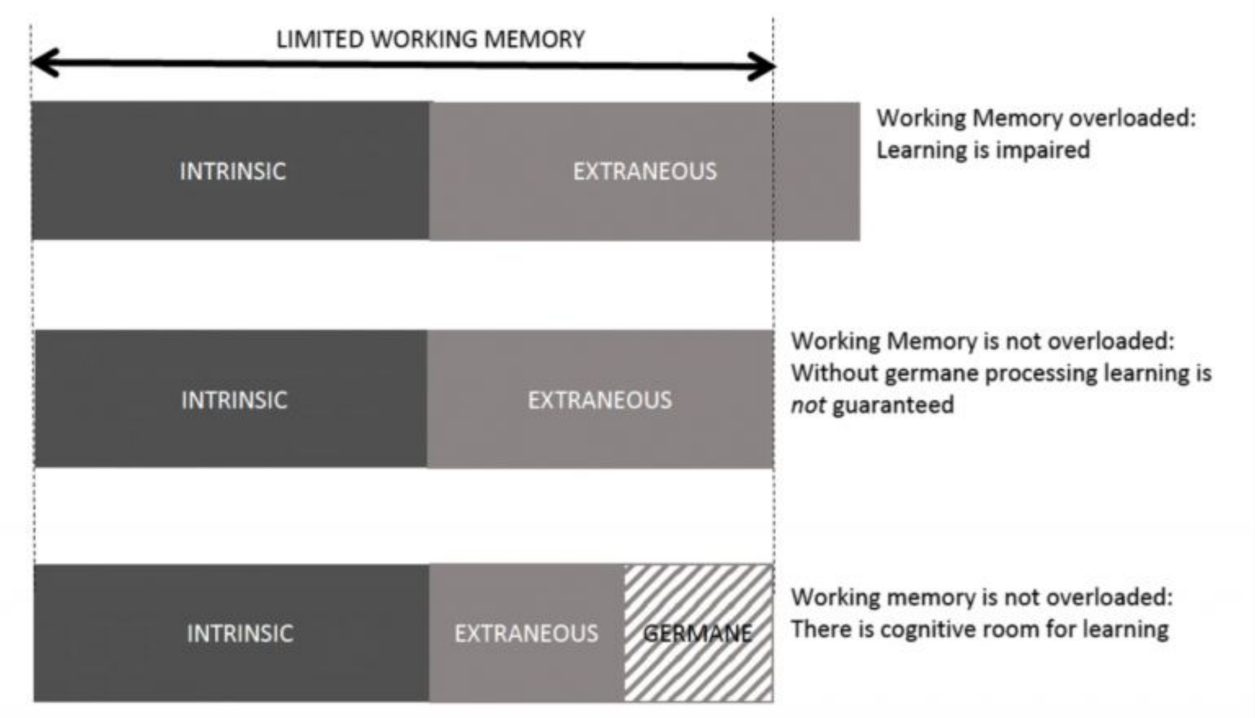
📌 Best Practice 3

Seamless Iteration

& Making Every Data Scientist

A 10X Data Scientist

1. Solving Cognitive Overload



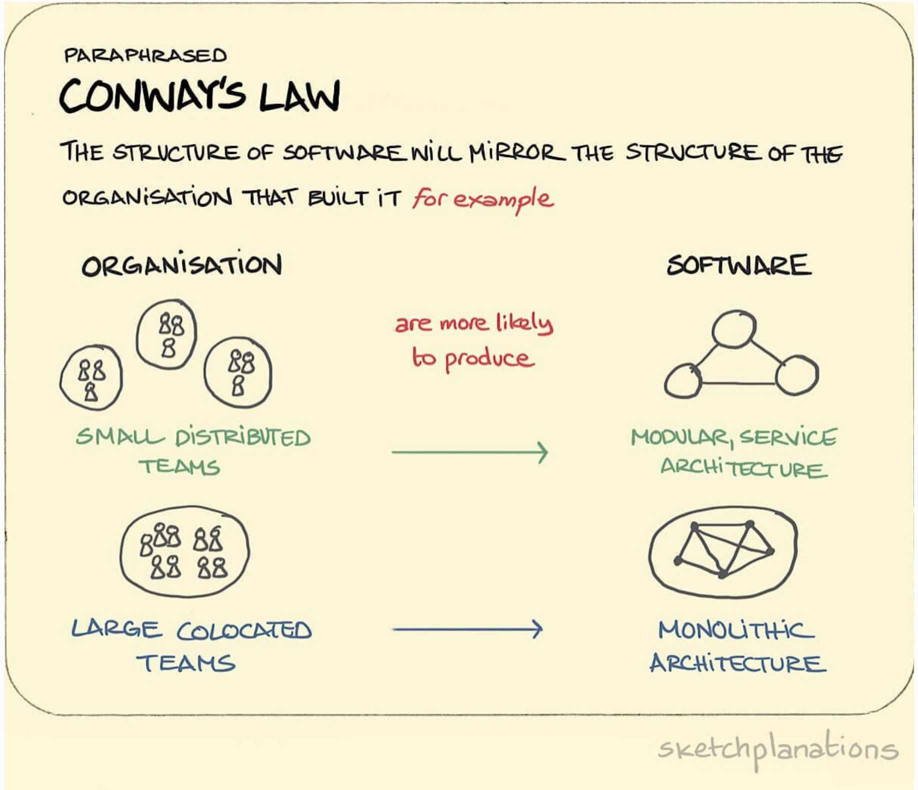
Working Memory overloaded: Learning is impaired

Working Memory is not overloaded: Without germane processing learning is *not* guaranteed

Working memory is not overloaded: There is cognitive room for learning

Tom Geraghty: "Platform as a Product"

2. Cut Friction & Increase Flow

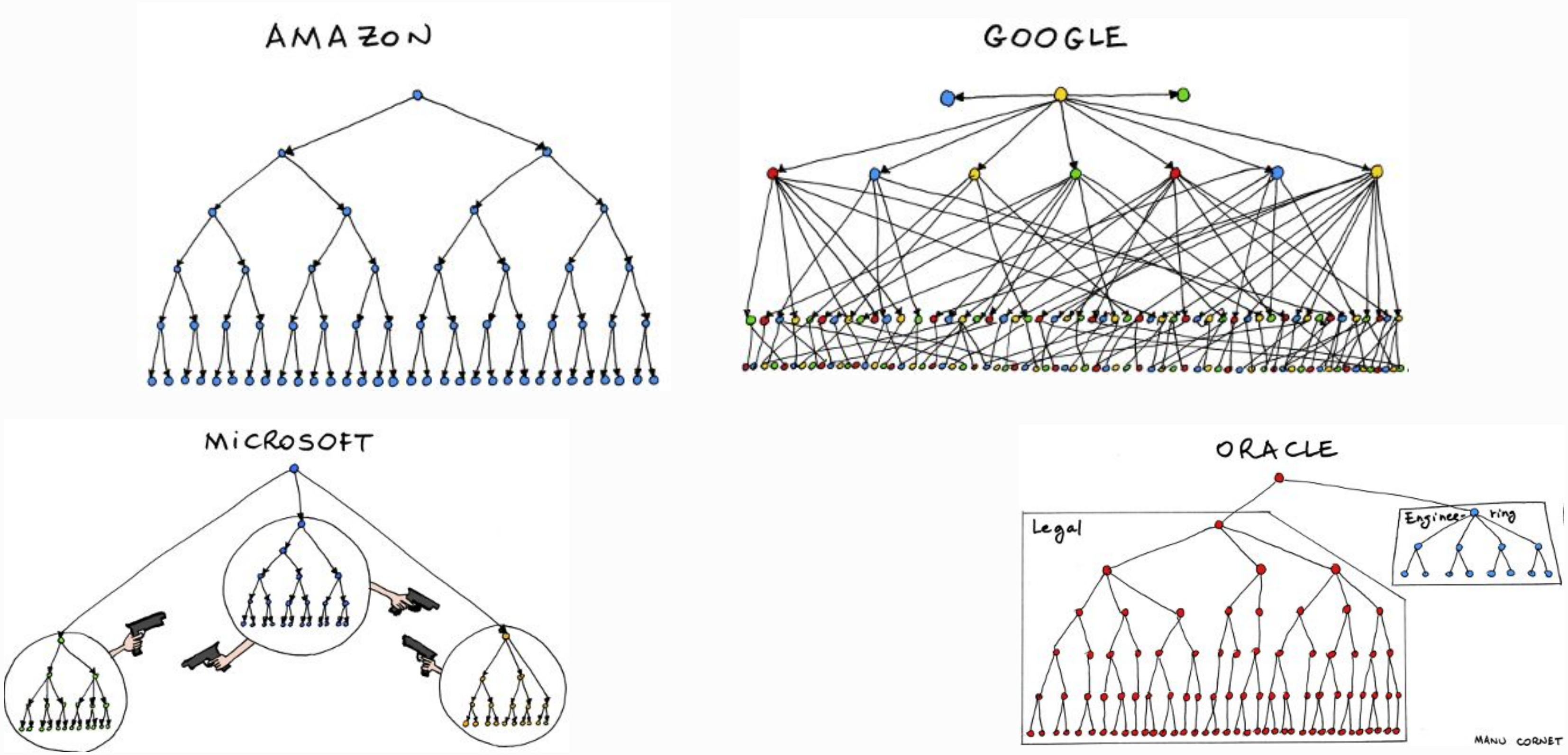


PARAPHRASED

CONWAY'S LAW

THE STRUCTURE OF SOFTWARE WILL MIRROR THE STRUCTURE OF THE ORGANISATION THAT BUILT IT *for example*

Source: Sketchplanations -- Conway's Law



Manu Cornet: Conway's Law

🤯 So Many Tools & Services Out There, And Yet… 🤯

… Raise Your Hand ✋

If Your Platform Is Basically

S3 +Spark + Redis

(And Bash Scripts)

ML PLATFORM
WAIT, IT'S ALL
JUST BASH SCRIPTS,
S3, REDIS, AND SPARK?
ALWAYS HAS BEEN.
imgflip.com

And While Those Are Great Tools...

… There Are Certain Classes of Problems They Don't Solve

🙅🏻‍♀️

MLOps problems fall into two categories

Ops
Platform

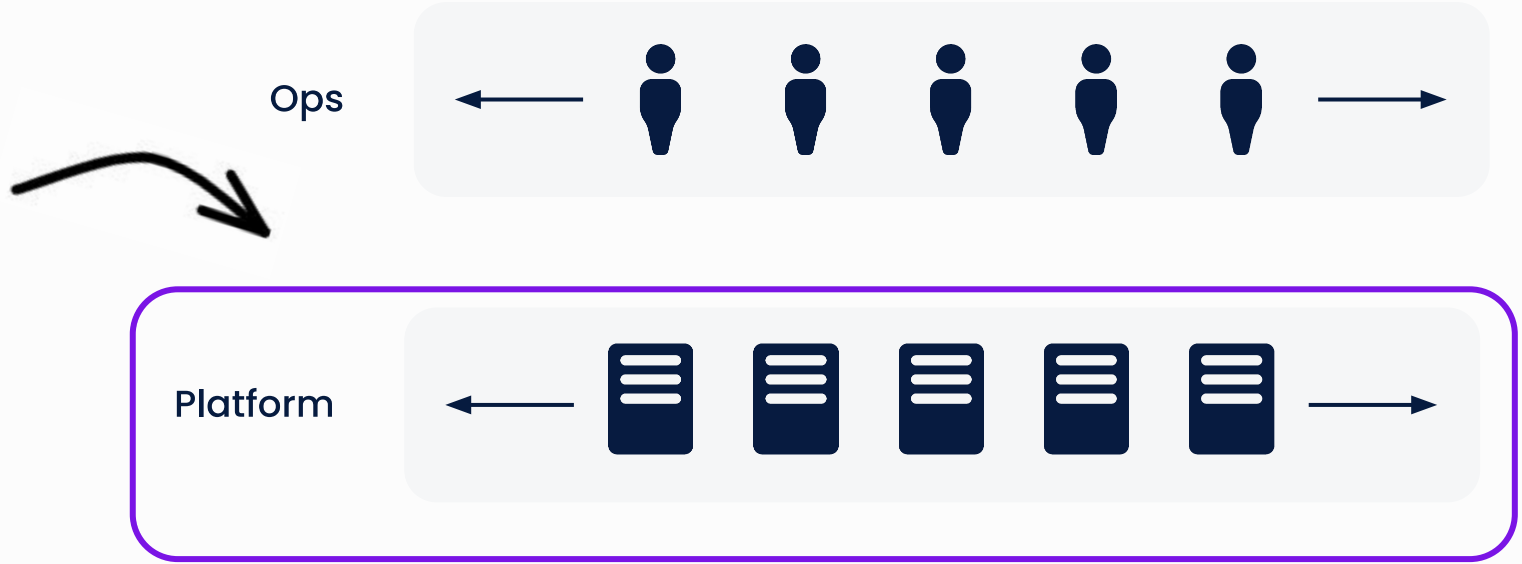
Specifically,

Problems Around

👉 Workflows & The Data

Science “Jobs-To-Be-Done”👈

MLOps problems fall into two categories



Let’s Describe The

Data Science

✅ “Jobs-To-Be-Done” ✅

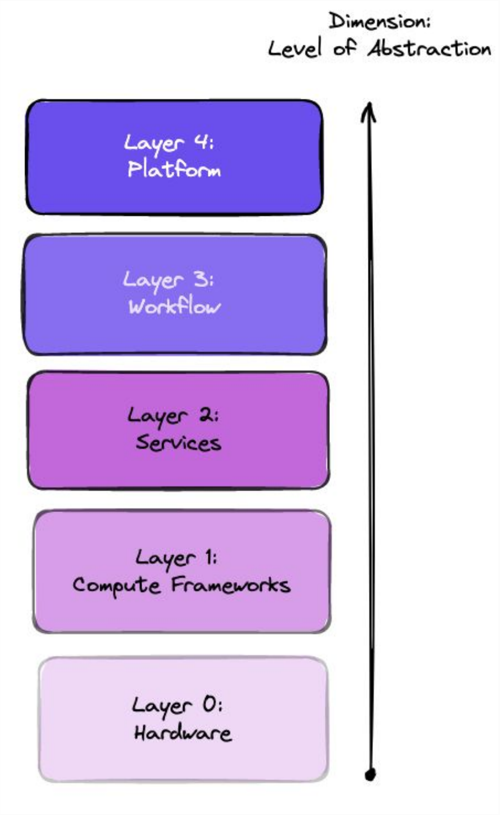
&

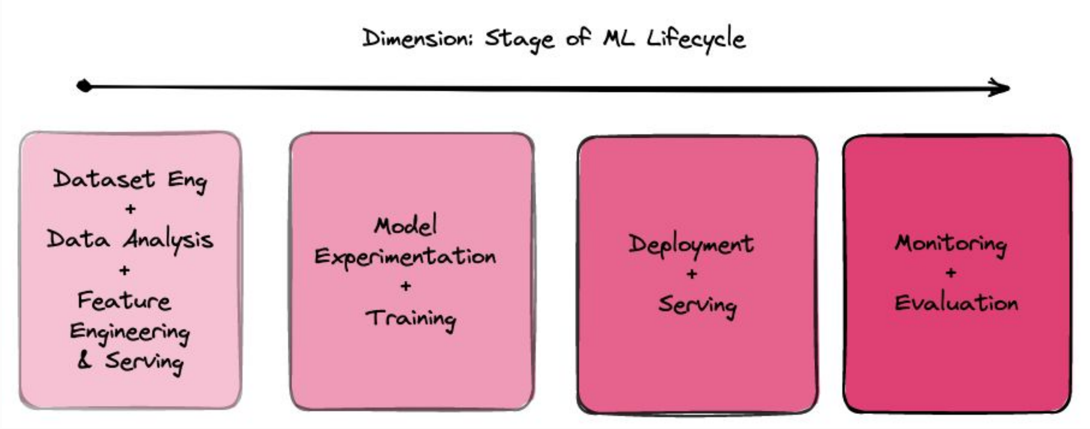
Propose A Framework

For How We Can Effectively Map Tools

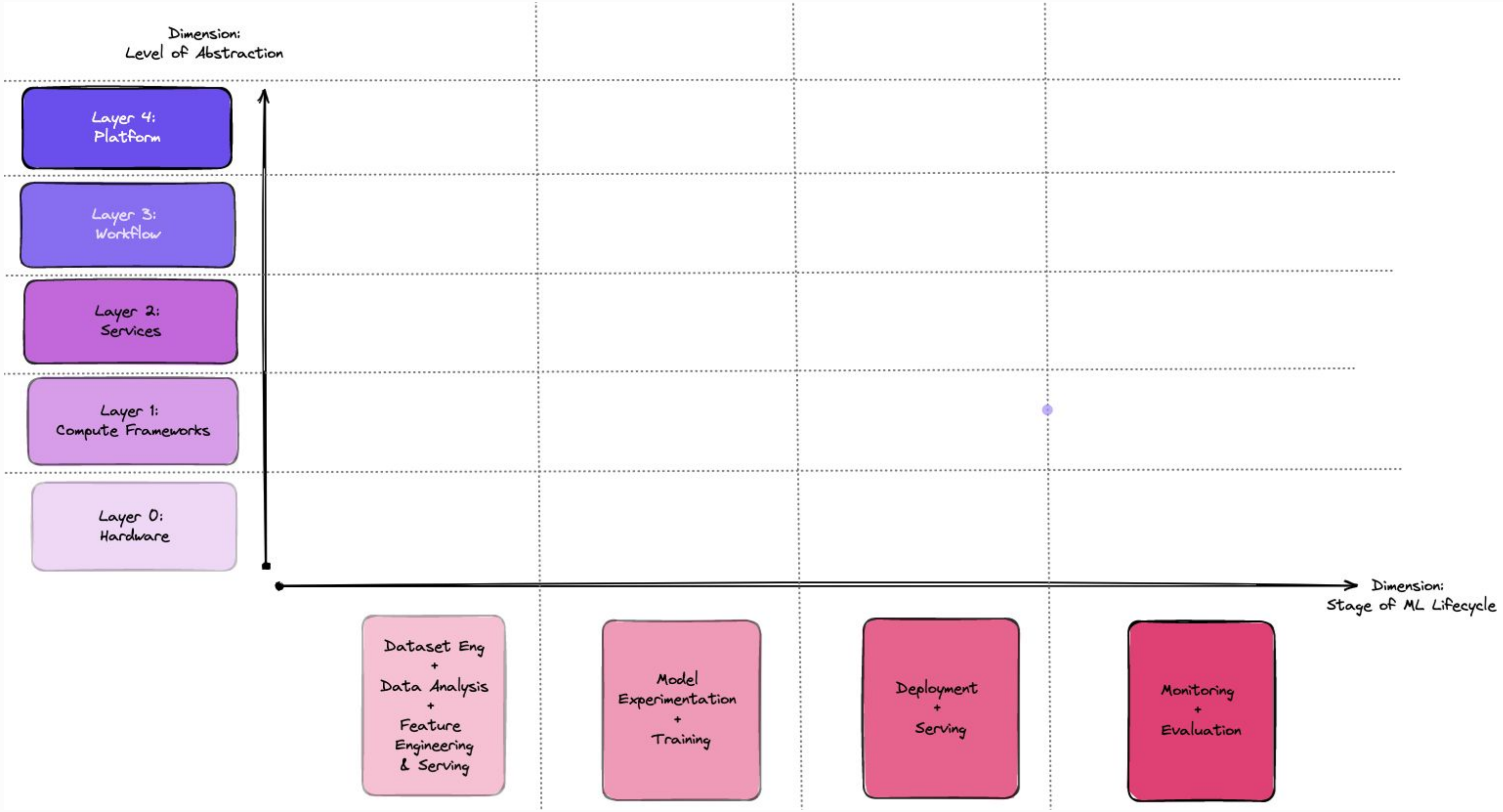
To Create An Effective Stack

DS Jobs To Be Done By:
{Level of Abstraction} VS {Stage of Lifecycle}

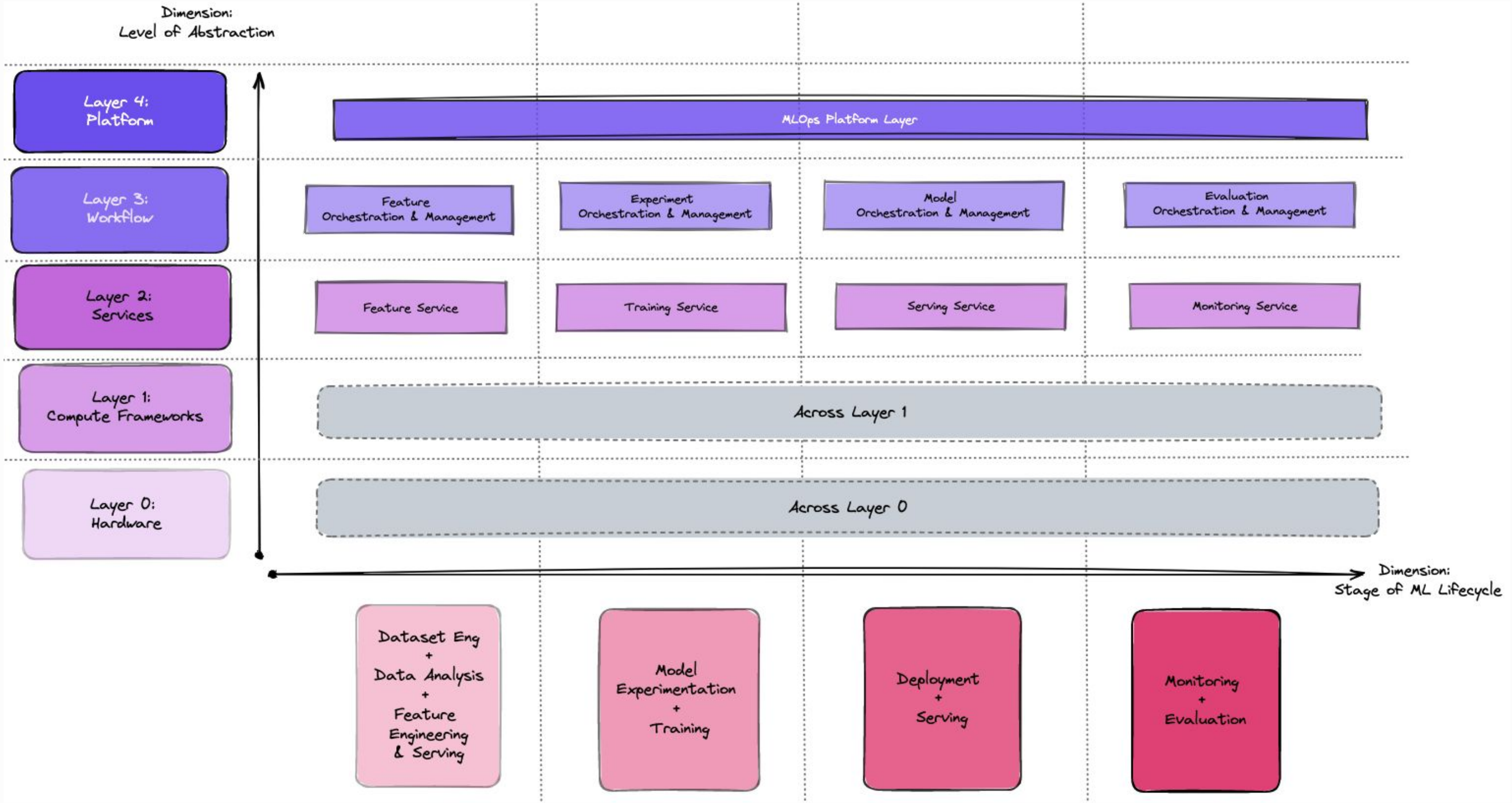



Dimension: Stage of ML Lifecycle
Dataset Eng + Data Analysis + Feature Engineering & Serving
Model Experimentation + Training
Deployment + Serving
Monitoring + Evaluation

{Level of Abstraction} VS {Stage of Lifecycle}



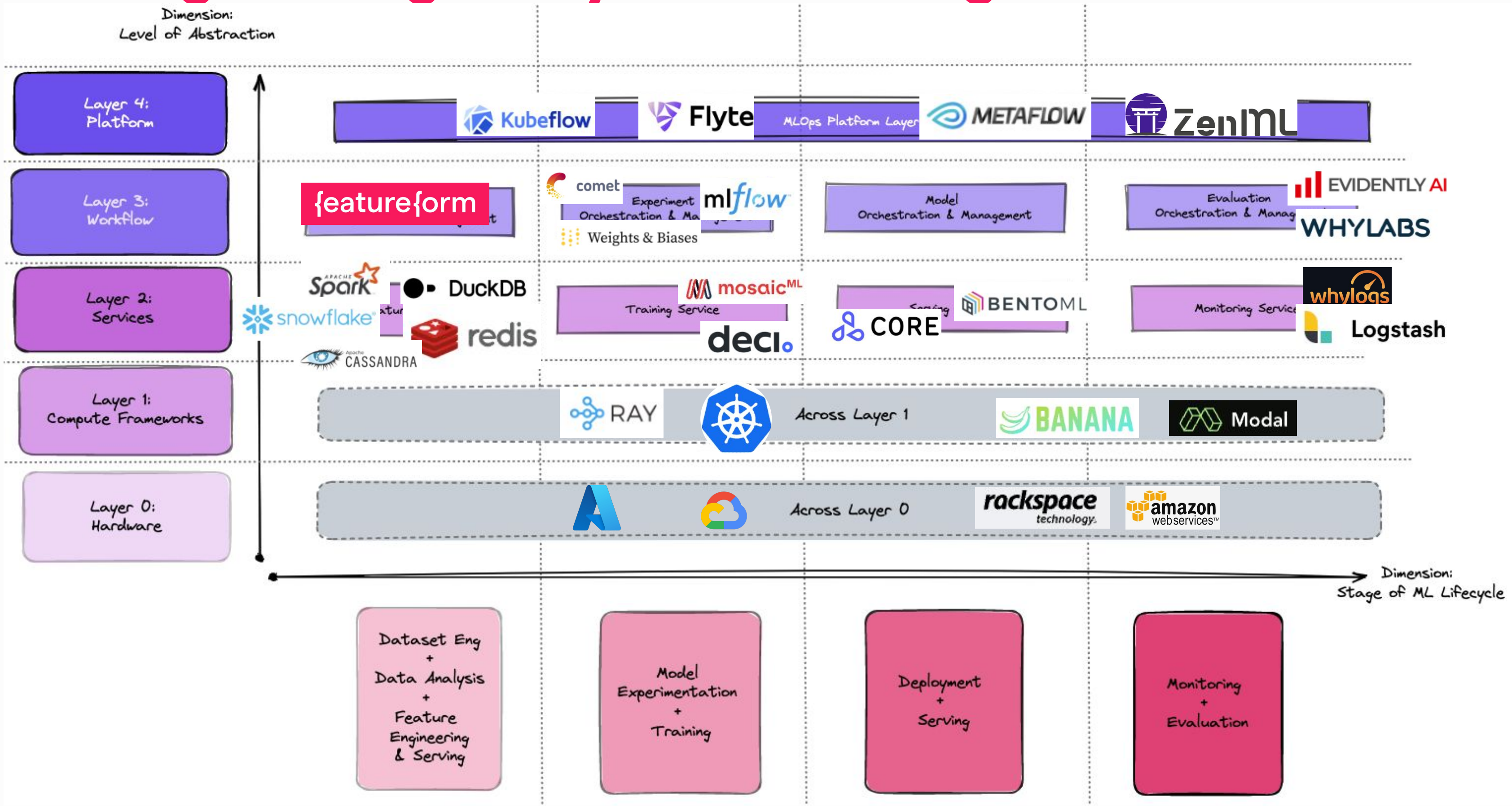
Choosing The Right Layer For the Right Job To be Done



Who Is Doing What*

*(As can be determined by their docs)

Choosing The Right Layer For the Right Job To be Done



Let's Apply The Framework For:

✅ The Solo Data Scientist

✅ The SMB Data Science Team

✅ The Enterprise DS Org

The Needs of

“The Solo Data Scientist”

📝

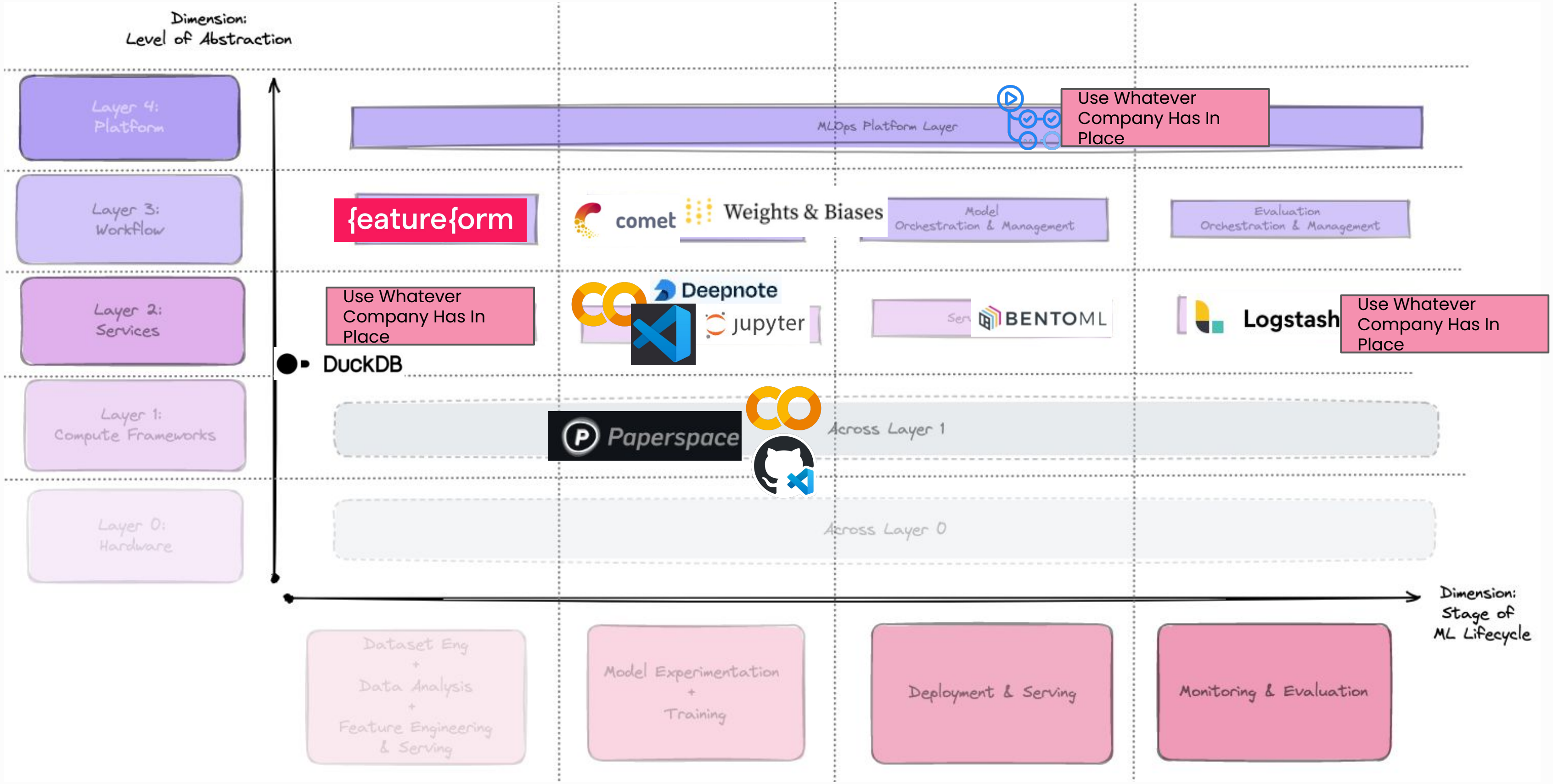
User Story: The Solo Data Scientist

I need to...	Pain-Point
Keep track of data, model, & code artifacts, including changes & experimentation runs.	**Versioning & Documentation**
Quickly iterate between features, algorithms, & hyperparameter tuning.	**Experimentation Tracking**
Train models on a "large enough" amount of data with access to GPUs.	**Serverless GPU**
Do everything with the least amount of overhead possible with the least amount of steps.	**Compatibility With Existing Product Stack**

Stack 1:

The Duke Nukem Stack

Stack 1: **The Duke Nukem Stack (Solo Data Scientist)**



Use Whatever Company Has In Place

Use Whatever Company Has In Place

Use Whatever Company Has In Place

The Needs of

“The SMB Data Science Team”

📝

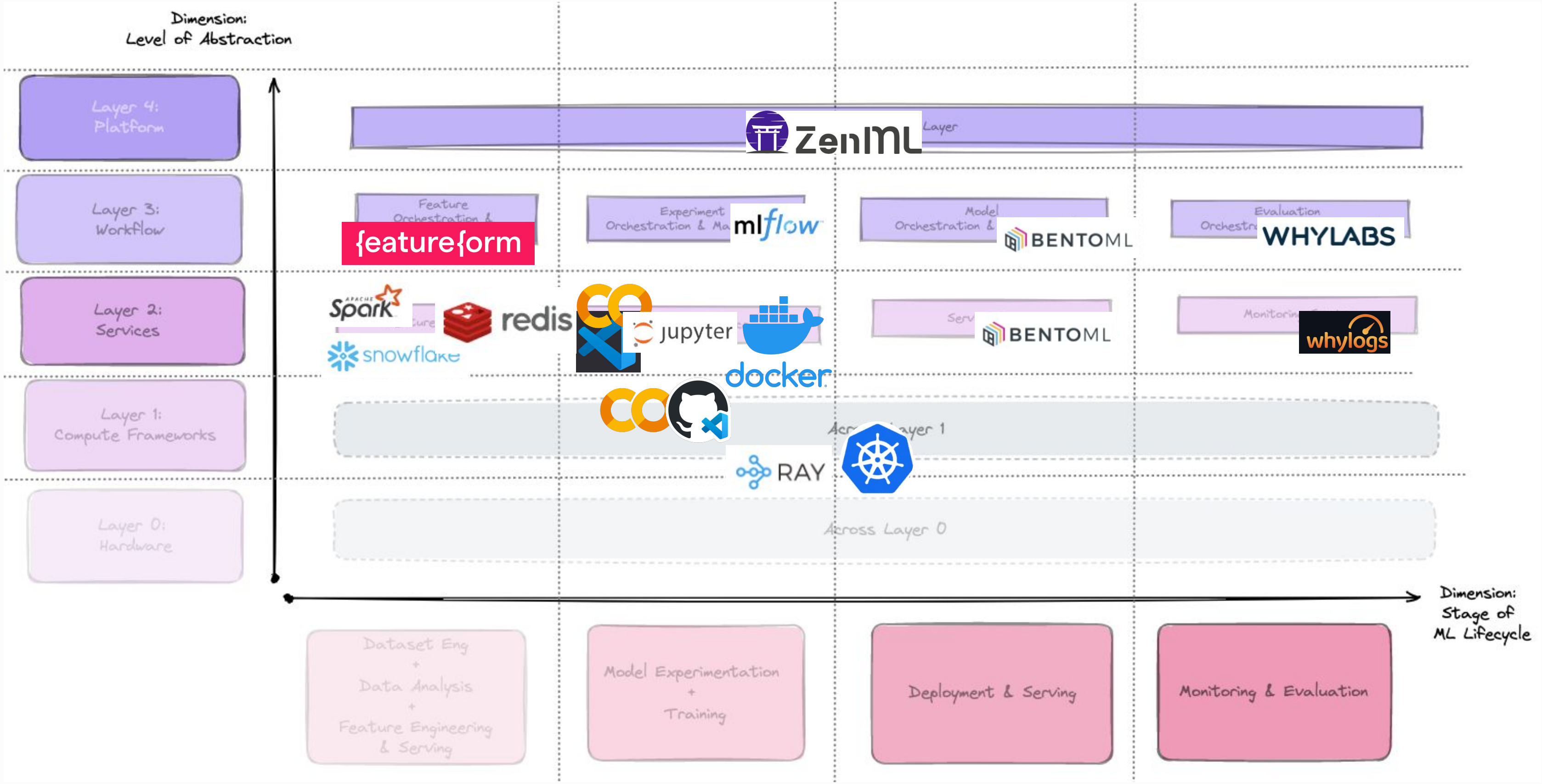
User Story: **The DS Team**

We need to…	**Pain-Point**
Collaborate with other members of the DS team (and potentially even external partners) on projects, with visibility into progress or health of data science assets.	**Collaboration**
Share & distribute knowledge asynchronously, while getting ahead of human bottlenecks & the accumulation of tribal knowledge.	**Documentation & Discoverability**
Ensure we're not "reinventing the wheel" across the organization & repeating work.	**Reuse & Resource Sharing**
Be notified when model pipelines and prediction services aren't working as expected with insight into failure conditions.	**Fine-grained Monitoring & Evaluation**

Stack 2:

The Serious Business Stack

Stack 2: **The Serious Business Stack (SMB)**



The Needs of

“The Enterprise Org”

📝

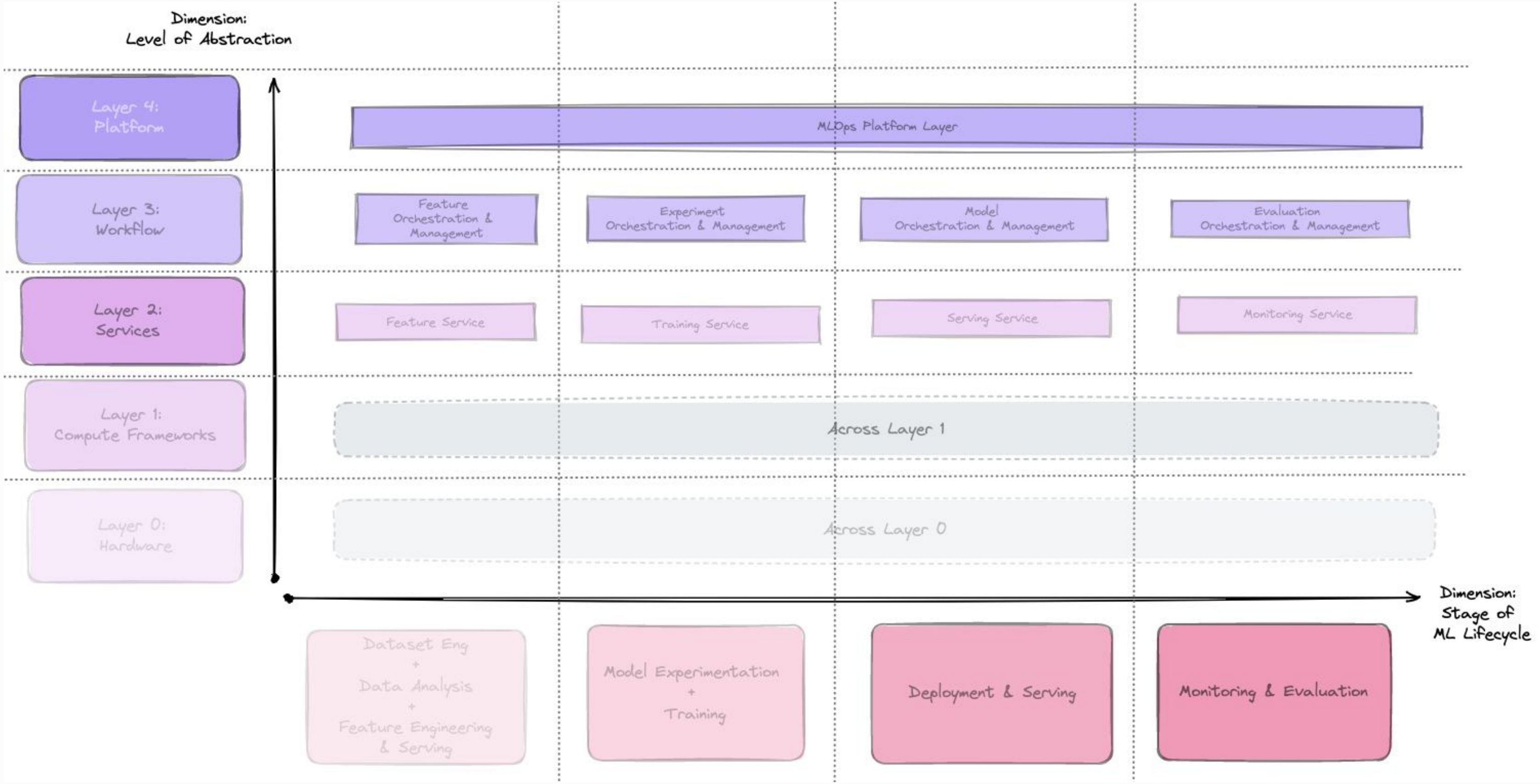
User Story: **The Organization**

We need to…	**Pain-Point**
Span and unify multiple infrastructure providers (including multi-cloud and on-prem), model deployment patterns, and model serving architectures as seamlessly as possible.	**Heterogeneous Infrastructure**
Handle a wide variety of regulation around data & models, log compliance related information & data, & streamline communication & visibility.	**Governance, Access Control, Audit Logs**
Interface with non-DS teams (including other engineering teams, as well as non-eng teams like legal & marketing).	**Cross-Functional Workflows**

Stack 3:

The Olly Olly Oxen Free Stack

Stack 3: **The Olly Olly Oxen Free Stack (Enterprise)**



There Will Be No “Modern MLOps” Stack

🙅🏻‍♀️

But Wait!

There’s Hope!

□□

ALWAYS HOPE, THERE IS
imgflip.com

In Closing

📪

Takeaways

- Describing the original promises of MLOps (& the current shortfalls).
- Understanding the "Jobs-To-Be-Done" of Data Scientists (& how the current ecosystem supports them).
- Describe the pain-points of the Solo Data Scientist, the SMB Data Science Team, & the areas of opportunity for Enterprises.
- Propose stacks that can be easily implemented in a relatively short period (sometimes even a day!).

Feel Free To Chat With Me During Office Hours

👩🏻‍🏫

Repository

bit.ly/3Yz2G95

Docs

bit.ly/423SE2W

LinkedIn

Mikiko Bazeley

Head of MLOps