MLOps: get your ML apps ready for production

Francesco Tamborra

NMLM, 12/12/2022 – Nantes

Outline

- ✦ Intro: What & why

`requirements.txt`
✦ ML development VS standard development ✦ ML lifecycle

- ✦ MLOps maturity levels
- ✦ Tools overview
- ✦ `@app.get("/predictions")`

What is MLOps ?

"The ability to apply DevOps principles to Machine Learning applications"

What is MLOps ?

"The ability to apply DevOps principles to Machine Learning applications"

MLOps is an ML engineering **culture and practice** that aims at unifying ML system development (Dev) and ML system operations (Ops).

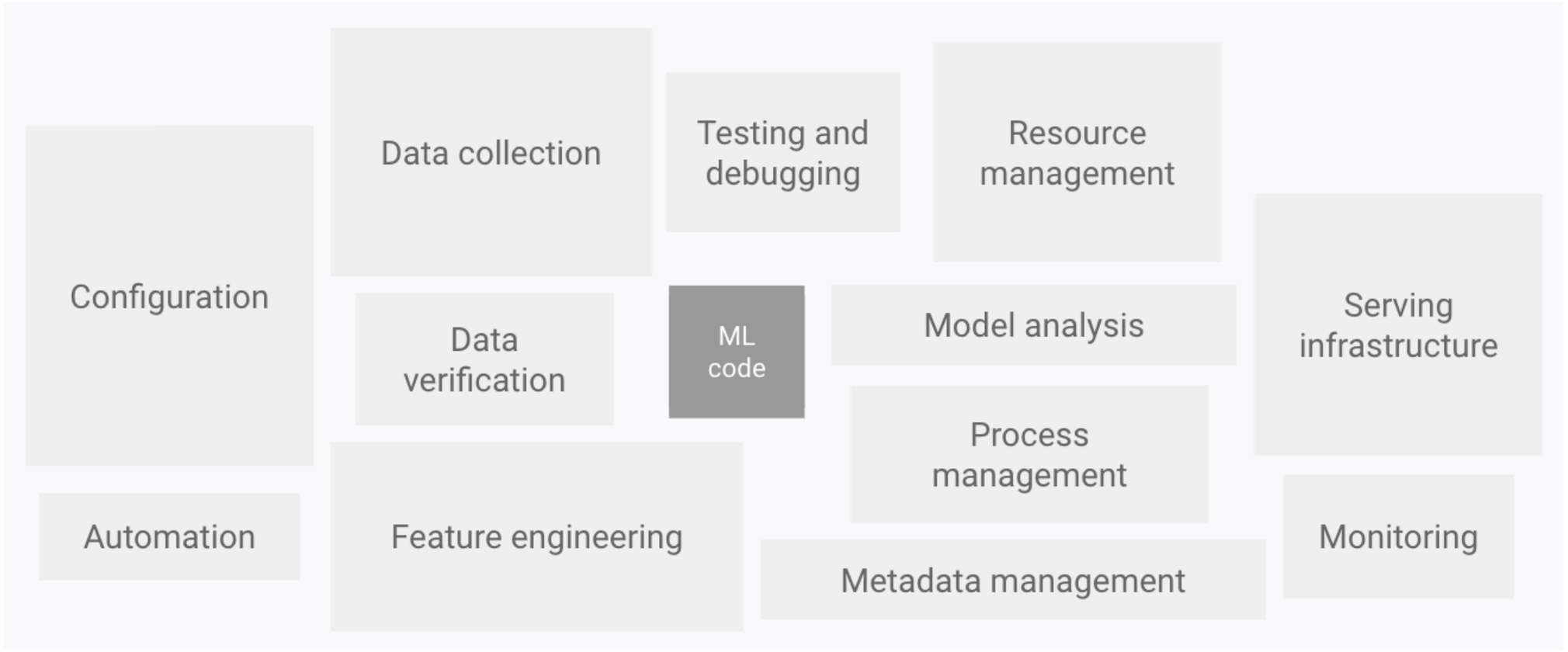
Practicing MLOps means that you advocate for automation and monitoring at all steps of ML system construction, including integration, testing, releasing, deployment and infrastructure management.

What is MLOps ?

"The ability to apply DevOps principles to Machine Learning applications"

MLOps is an ML engineering **culture and practice** that aims at unifying ML system development (Dev) and ML system operations (Ops).

Practicing MLOps means that you advocate for automation and monitoring at all steps of ML system construction, including integration, testing, releasing, deployment and infrastructure management.

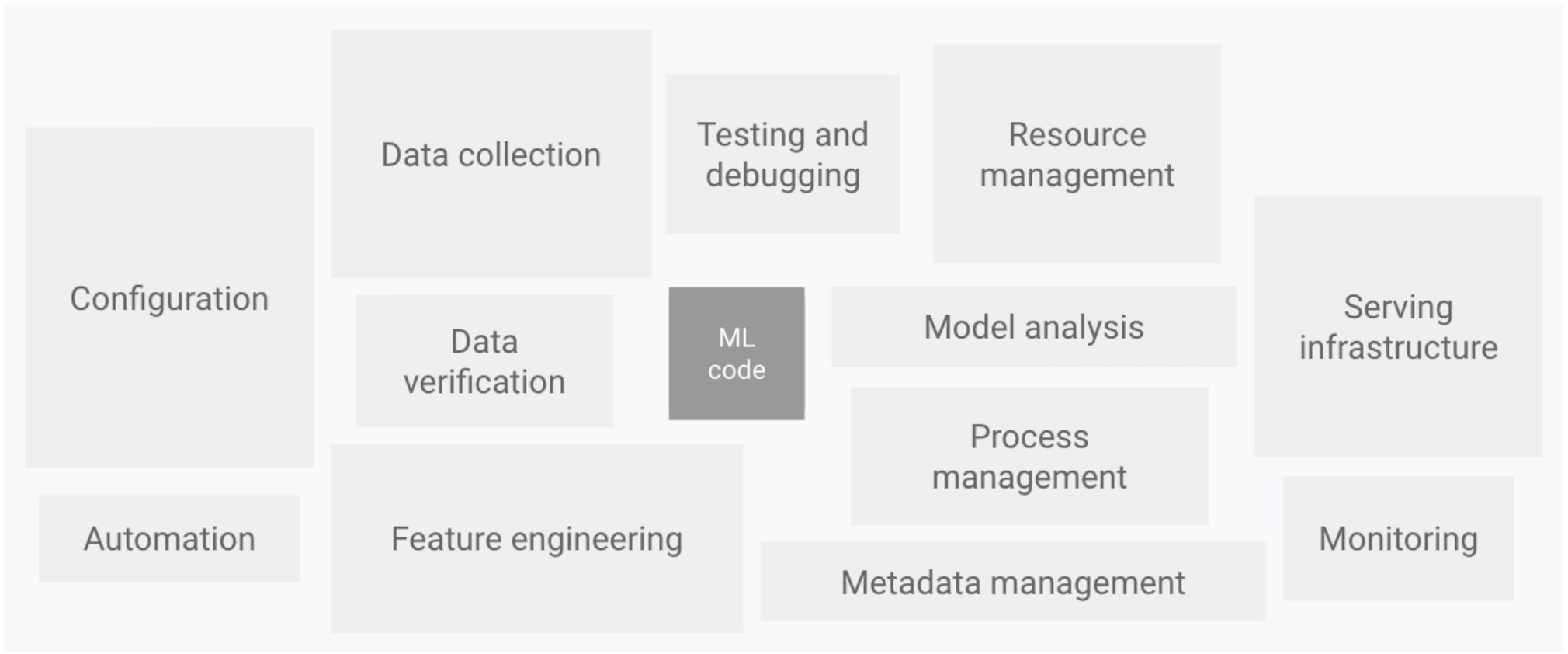


What is MLOps ?

"The ability to apply DevOps principles to Machine Learning applications"

MLOps is an ML engineering **culture and practice** that aims at unifying ML system development (Dev) and ML system operations (Ops).

Practicing MLOps means that you advocate for automation and monitoring at all steps of ML system construction, including integration, testing, releasing, deployment and infrastructure management.



Global MLOps Market Size estimated at 612 USD million in 2021 and projected to reach 6.1 USD billion by 2028

Why MLOps ?

“2020 State of Enterprise ML” by Algorithmia (survey size: 750) report results:

Why MLOps ?

"2020 State of Enterprise ML" by Algorithmia (survey size: 750) report results:

55% of companies surveyed have **not** deployed
a machine learning model

Why MLOps ?

"2020 State of Enterprise ML" by Algorithmia (survey size: 750) report results:

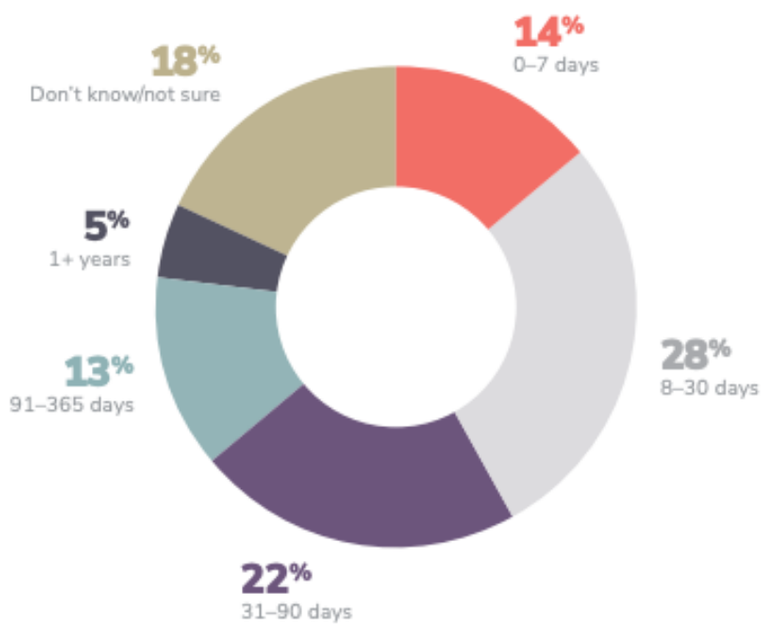
55% of companies surveyed have **not** deployed
a machine learning model

Why MLOps ?

"2020 State of Enterprise ML" by Algorithmia (survey size: 750) report results:

55% of companies surveyed have **not** deployed a machine learning model

Model deployment timeline

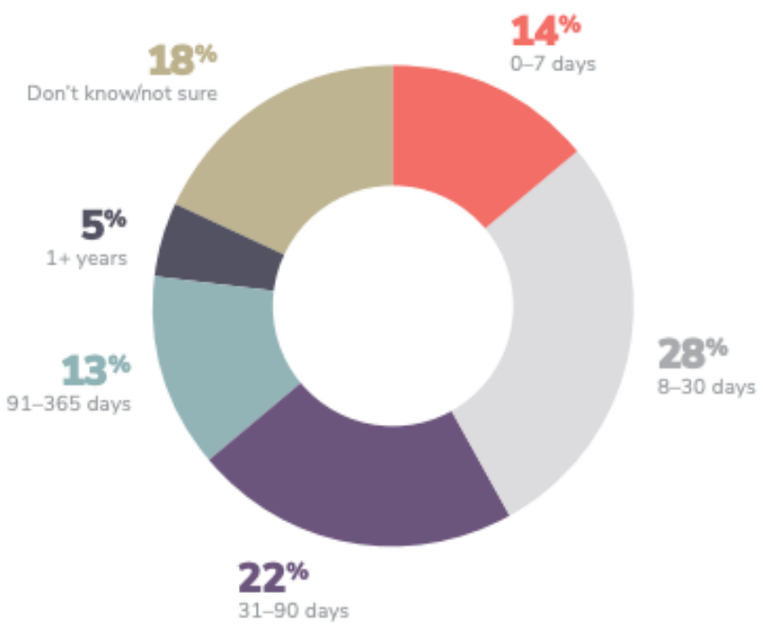


Why MLOps ?

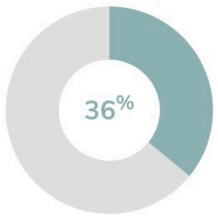
"2020 State of Enterprise ML" by Algorithmia (survey size: 750) report results:

55% of companies surveyed have **not** deployed a machine learning model

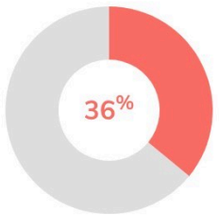
Model deployment timeline



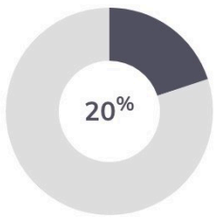
What percentage of your data scientists' time is spent deploying ML models?



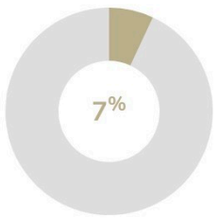
36% of survey participants said their data scientists spend **a quarter** of their time deploying ML models



36% of survey participants said their data scientists spend **a quarter to half** of their time deploying ML models



20% of survey participants said their data scientists spend **half to three-quarters** of their time deploying ML models



7% of survey participants said their data scientists spend **more than three-quarters** of their time deploying ML models

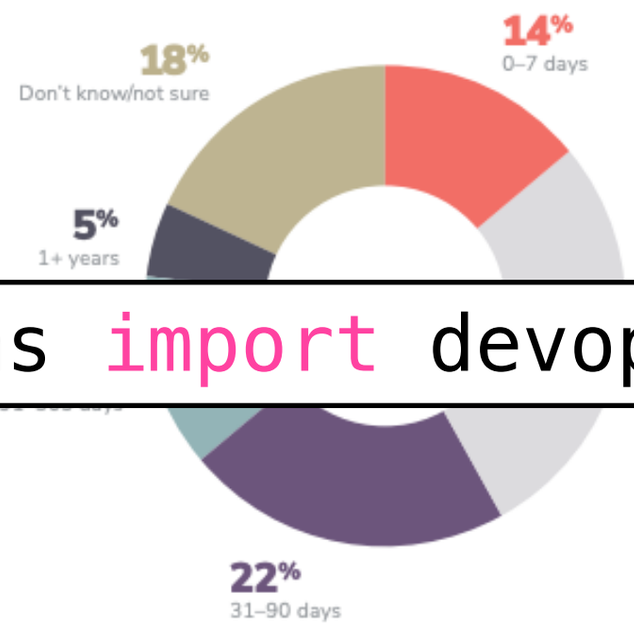
1% of respondents said they were unsure.

Why MLOps ?

"2020 State of Enterprise ML" by Algorithmia (survey size: 750) report results:

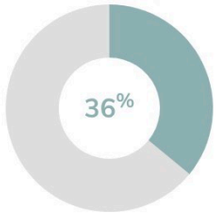
55% of companies surveyed have **not** deployed a machine learning model

Model deployment timeline

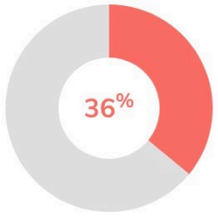



```
from DevelopmentOperations import devops
```

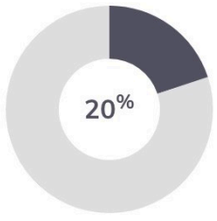
What percentage of your data scientists' time is spent deploying ML models?



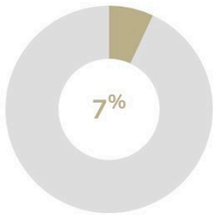
36% of survey participants said their data scientists spend **a quarter** of their time deploying ML models



36% of survey participants said their data scientists spend **a quarter to half** of their time deploying ML models



20% of survey participants said their data scientists spend **half to three-quarters** of their time deploying ML models

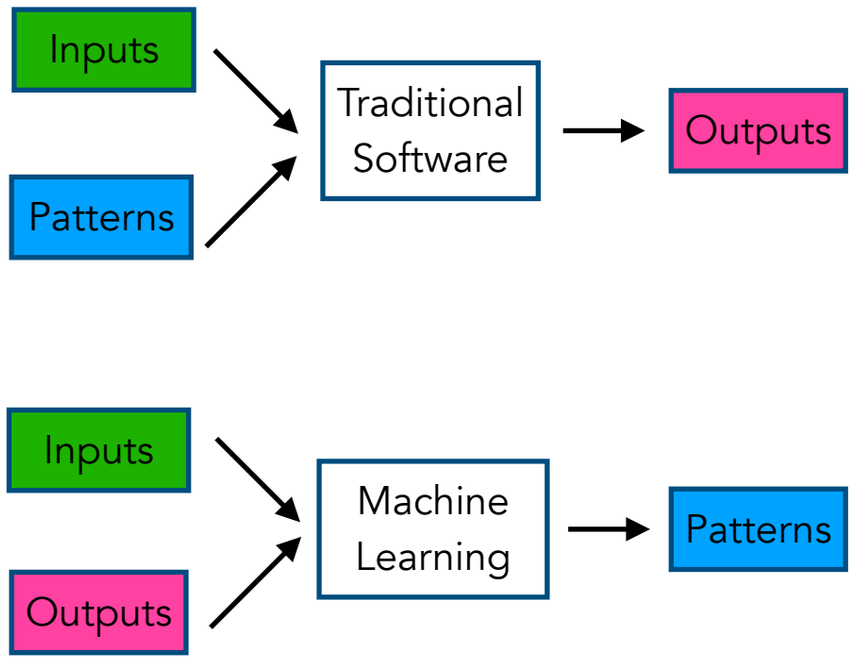


7% of survey participants said their data scientists spend **more than three-quarters** of their time deploying ML models

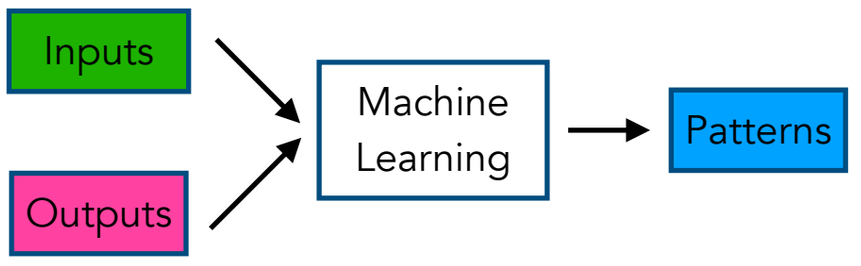
1% of respondents said they were unsure.

ML Dev VS standard SW Dev

ML Dev VS standard SW Dev



ML Dev VS standard SW Dev

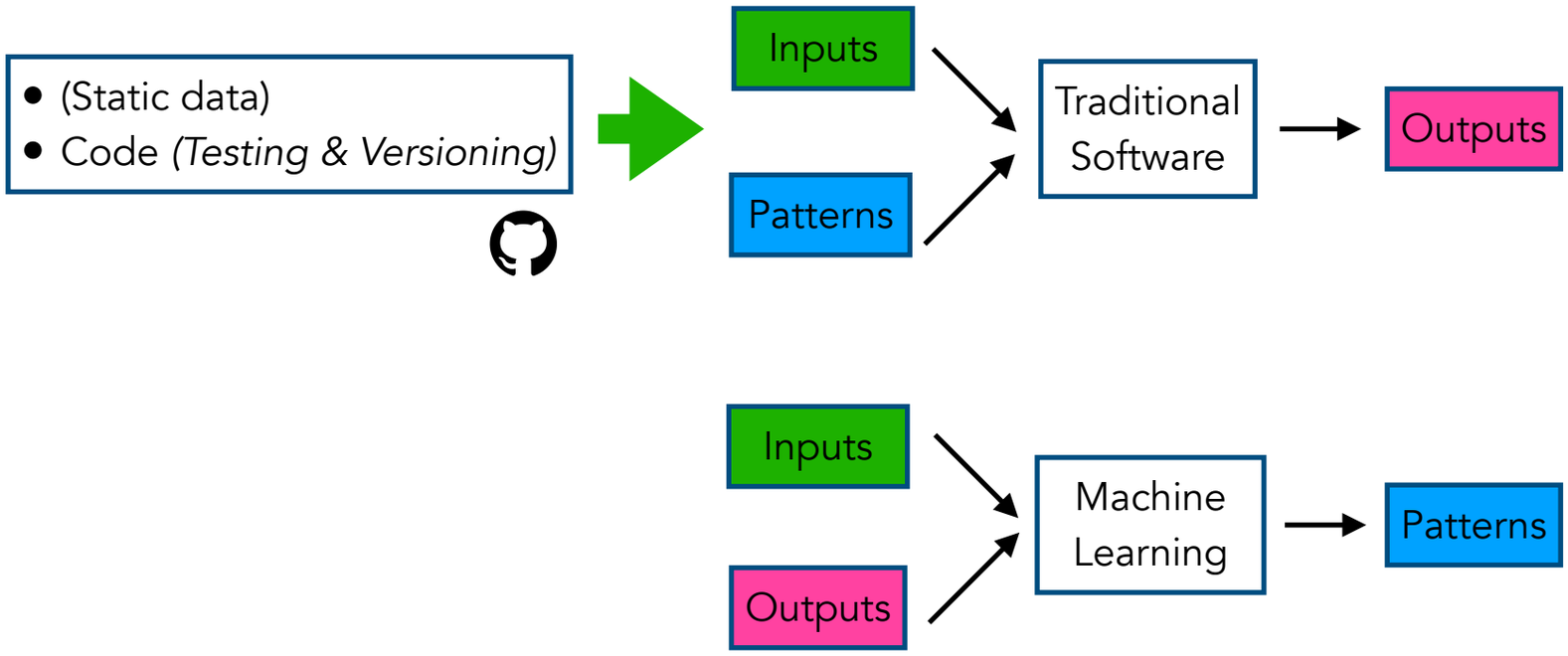


Andrej Karpathy

Nov 11, 2017 · 9 min read · Listen

Software 2.0

ML Dev VS standard SW Dev

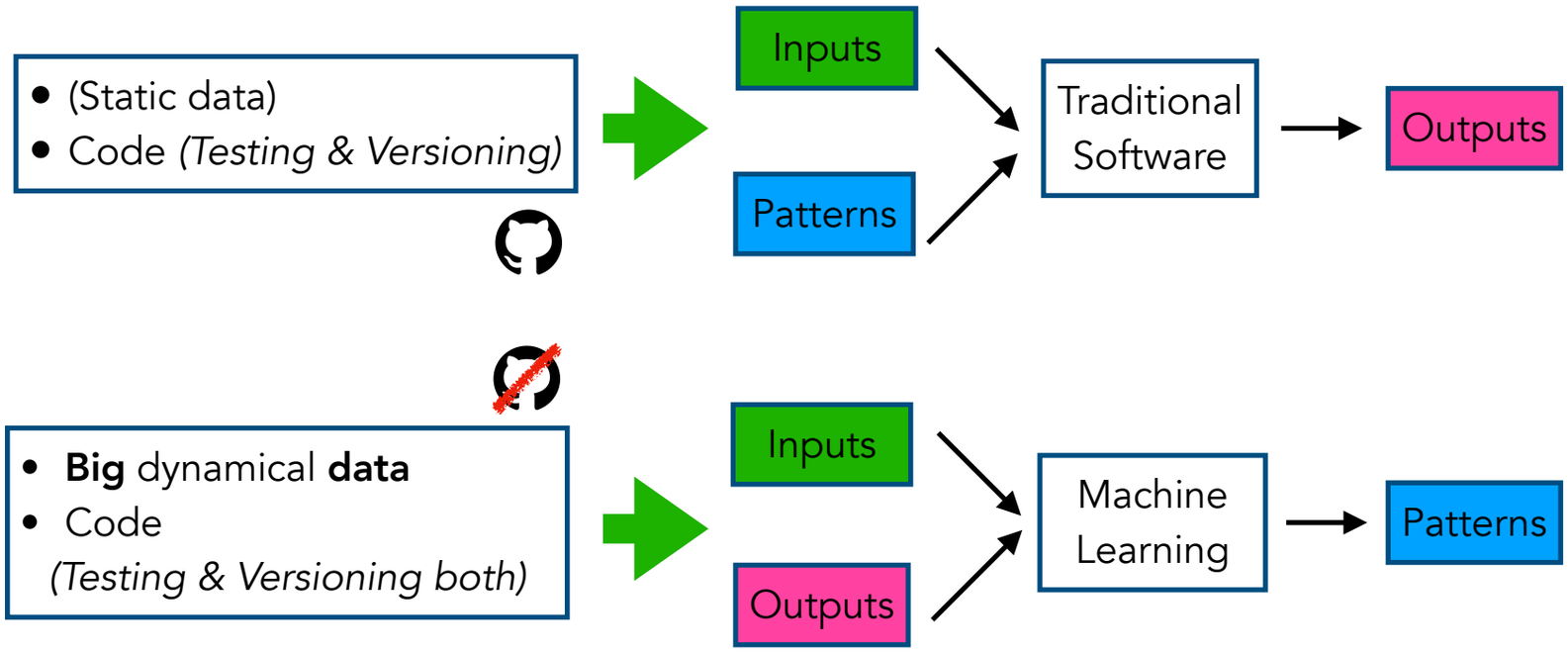


Andrej Karpathy

Nov 11, 2017 · 9 min read · Listen

Software 2.0

ML Dev VS standard SW Dev



Andrej Karpathy

Nov 11, 2017 · 9 min read · Listen

Software 2.0

ML Dev VS standard SW Dev

- (Static data)
- Code *(Testing & Versioning)*

Inputs + Patterns → Traditional Software → Outputs

- **Big** dynamical **data**
- Code

  (Testing & Versioning both)

Inputs + Outputs → Machine Learning → Patterns

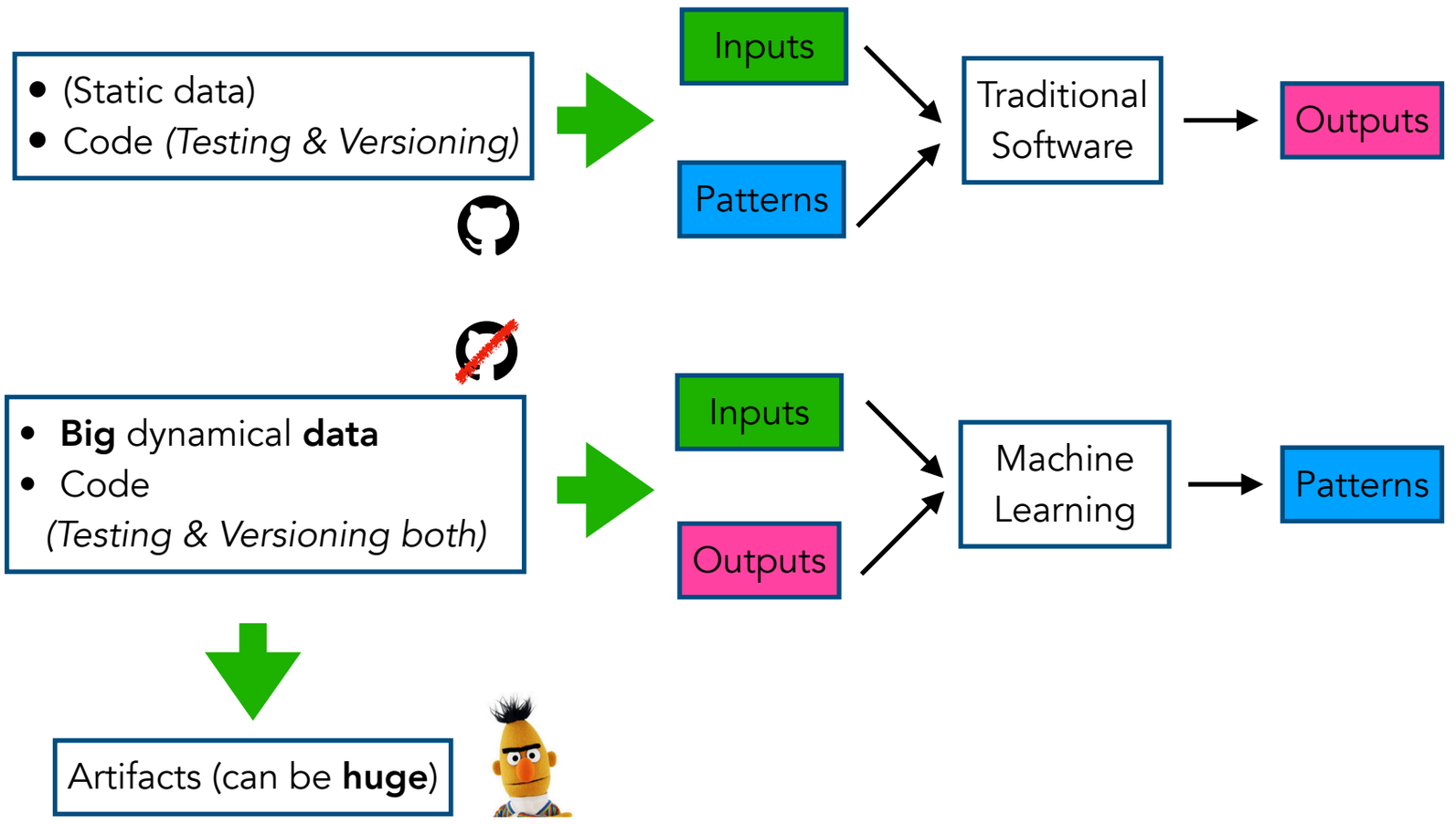
Artifacts (can be **huge**)

Andrej Karpathy

Nov 11, 2017 · 9 min read · Listen

Software 2.0

ML Dev VS standard SW Dev



Other differences:

Development Experimental in nature: **tracking** what worked and what didn't is essential to maintain **reproducibility**

ML Dev VS standard SW Dev

- (Static data)
- Code *(Testing & Versioning)*

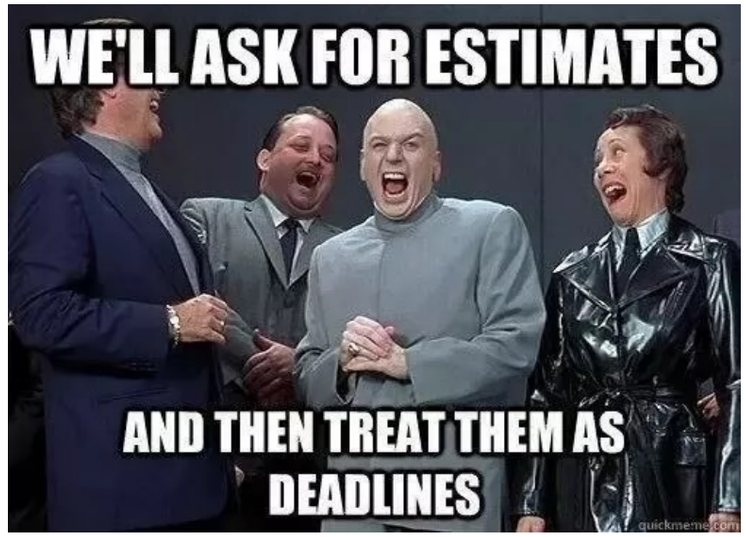
Inputs + Patterns → Traditional Software → Outputs

- **Big** dynamical **data**
- Code

(Testing & Versioning both)

Inputs + Outputs → Machine Learning → Patterns

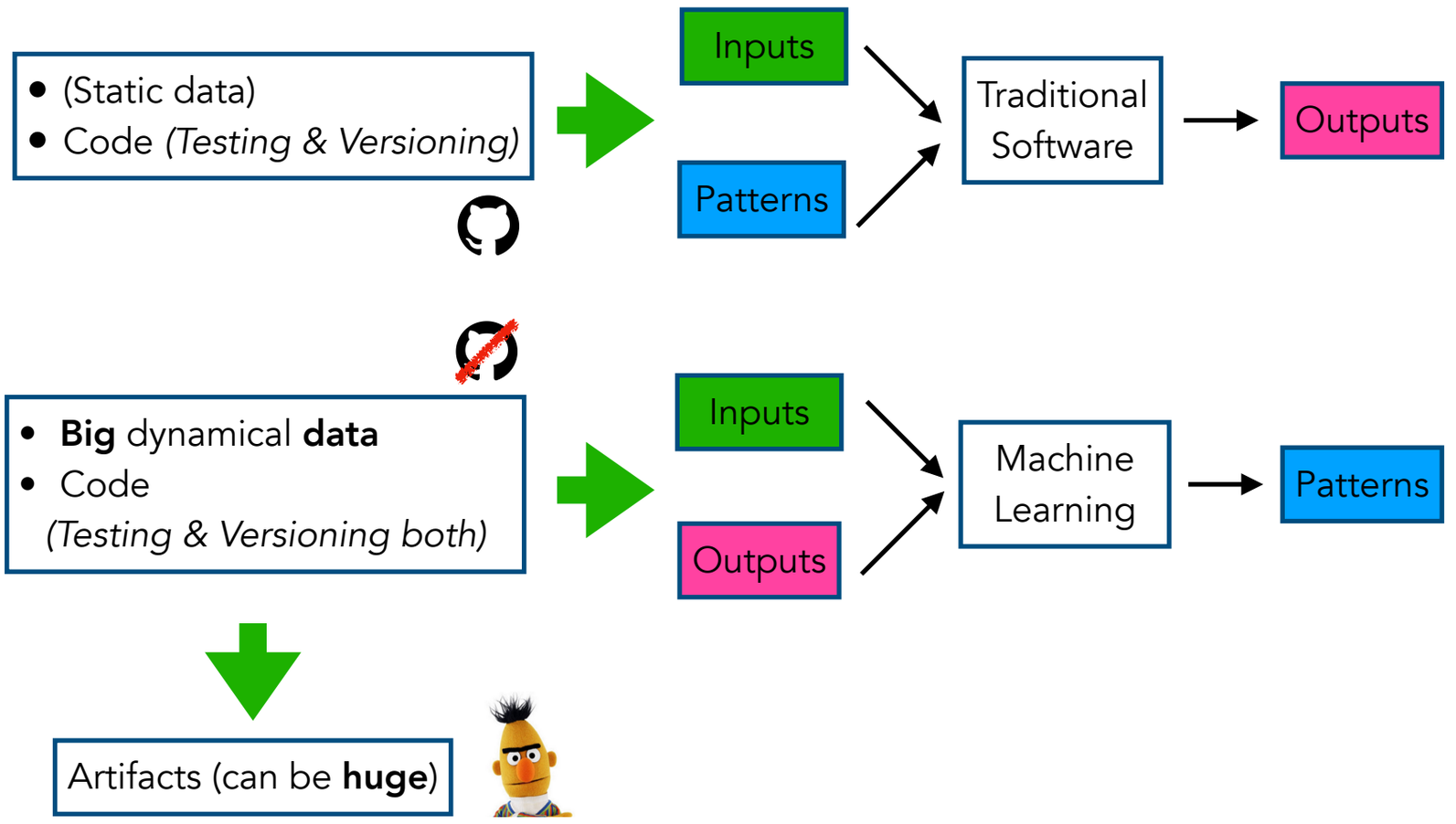
Artifacts (can be **huge**)



Other differences:

Development Experimental in nature: **tracking** what worked and what didn't is essential to maintain **reproducibility**

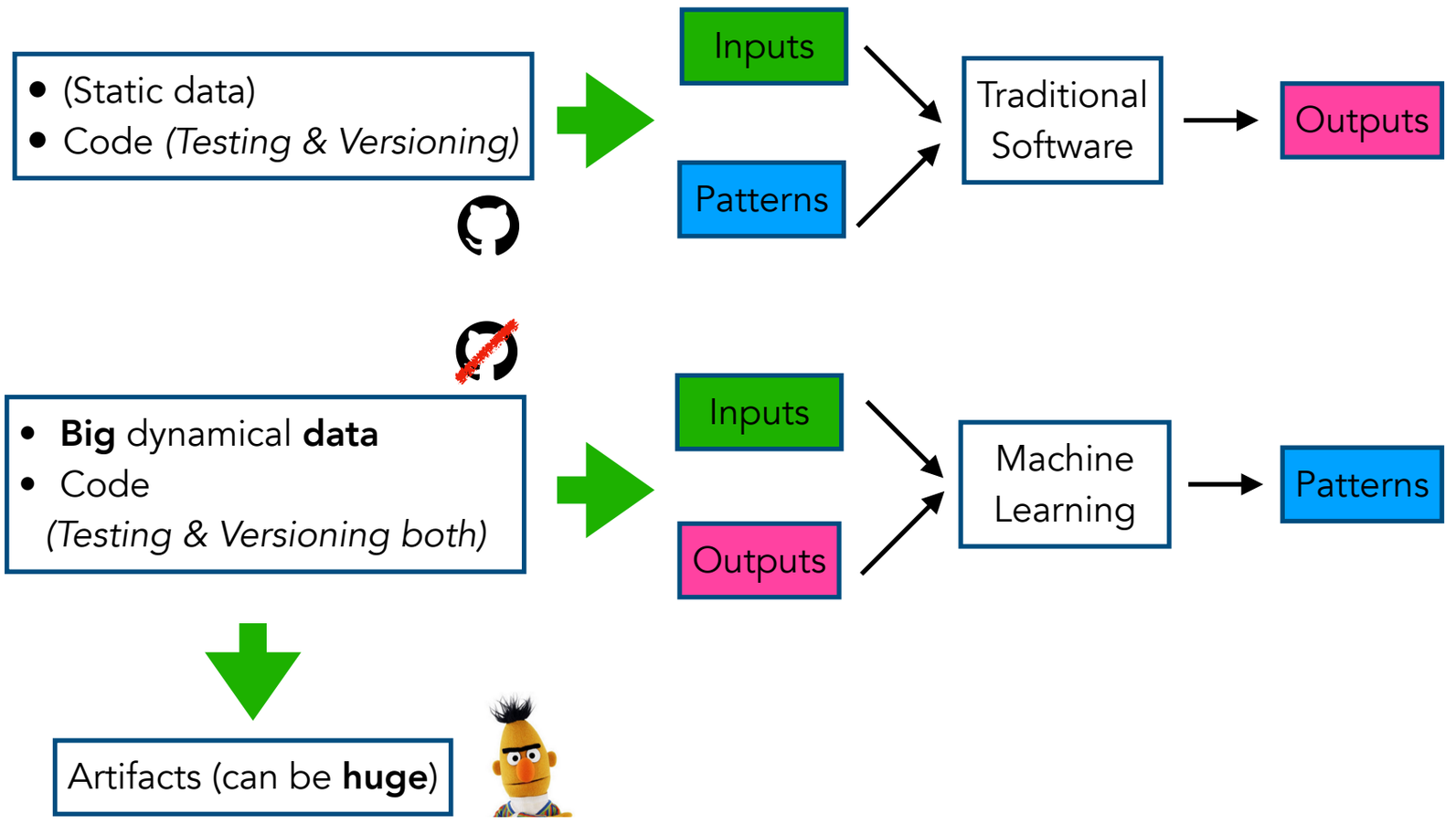
ML Dev VS standard SW Dev



Other differences:

Development	Experimental in nature: **tracking** what worked and what didn't is essential to maintain **reproducibility**
Testing	In addition to typical unit and integration tests, you need **data validation**, **trained model quality evaluation**, and **model validation**

ML Dev VS standard SW Dev



Other differences:

Development	Experimental in nature: **tracking** what worked and what didn't is essential to maintain **reproducibility**
Testing	In addition to typical unit and integration tests, you need **data validation**, **trained model quality evaluation**, and **model validation**
Production	**Models can decay** in more ways than conventional software systems, and you need to consider this degradation by **tracking summary statistics** of your data and **monitor the online performance**

DevOps VS MLOps in a nutshell

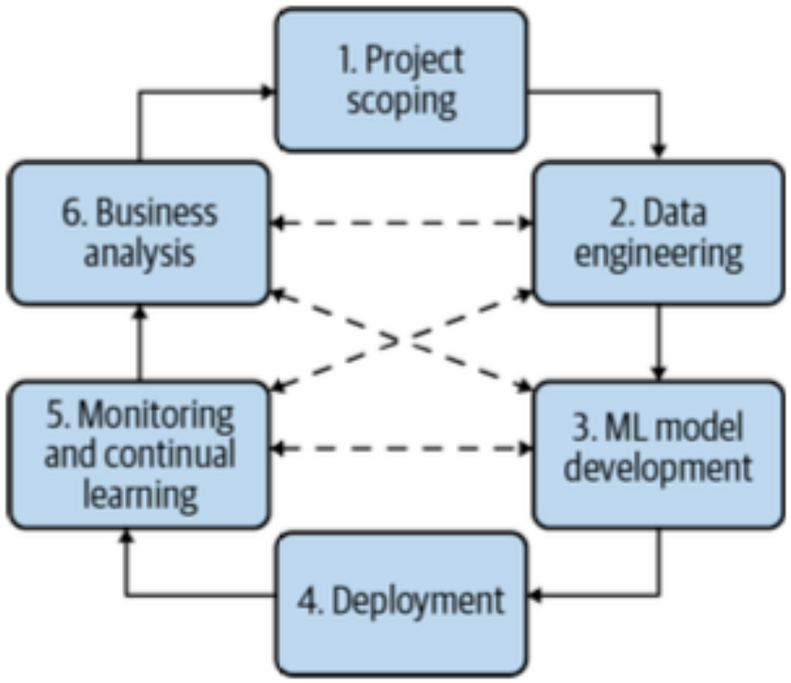
DevOps VS MLOps in a nutshell

CI is no longer only about testing and validating code and components, but also testing and validating data, data schemas, and models.

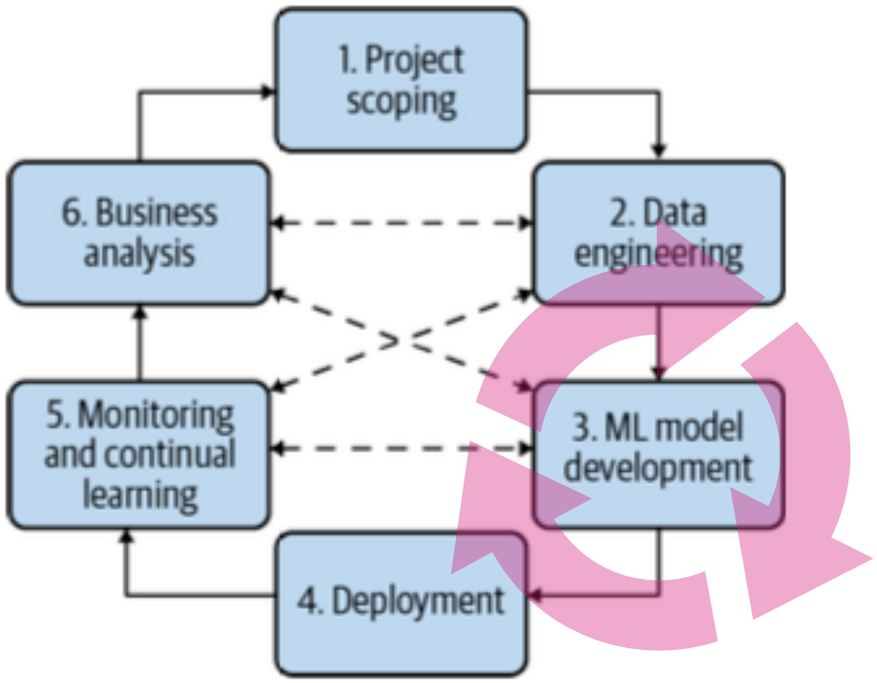
CD is no longer about a single software package or a service, but a system (an ML training pipeline) that should automatically deploy another service (model prediction service).

CT is a new property, unique to ML systems, that's concerned with automatically retraining and serving the models.

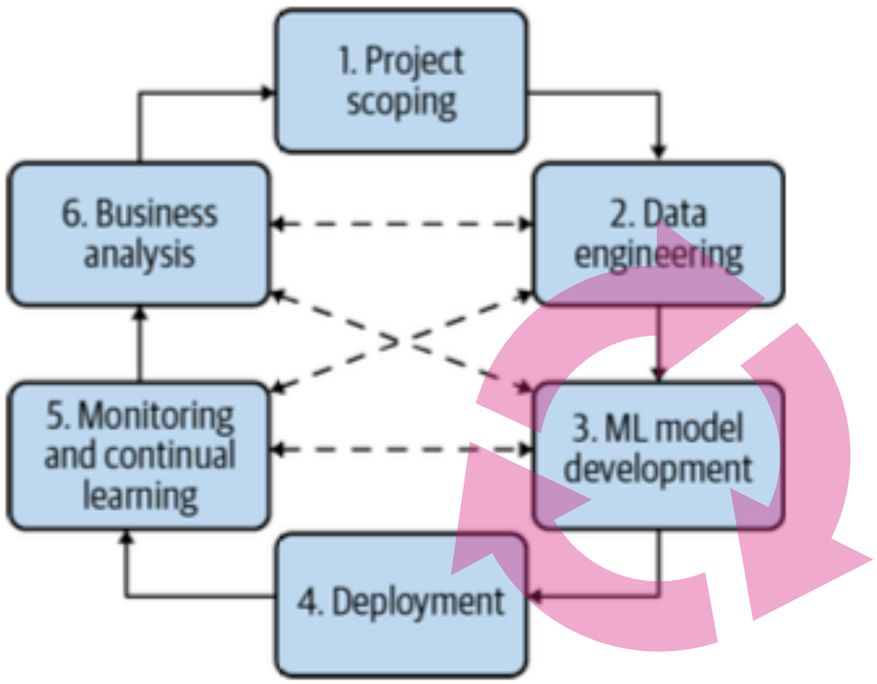
ML Lifecycle

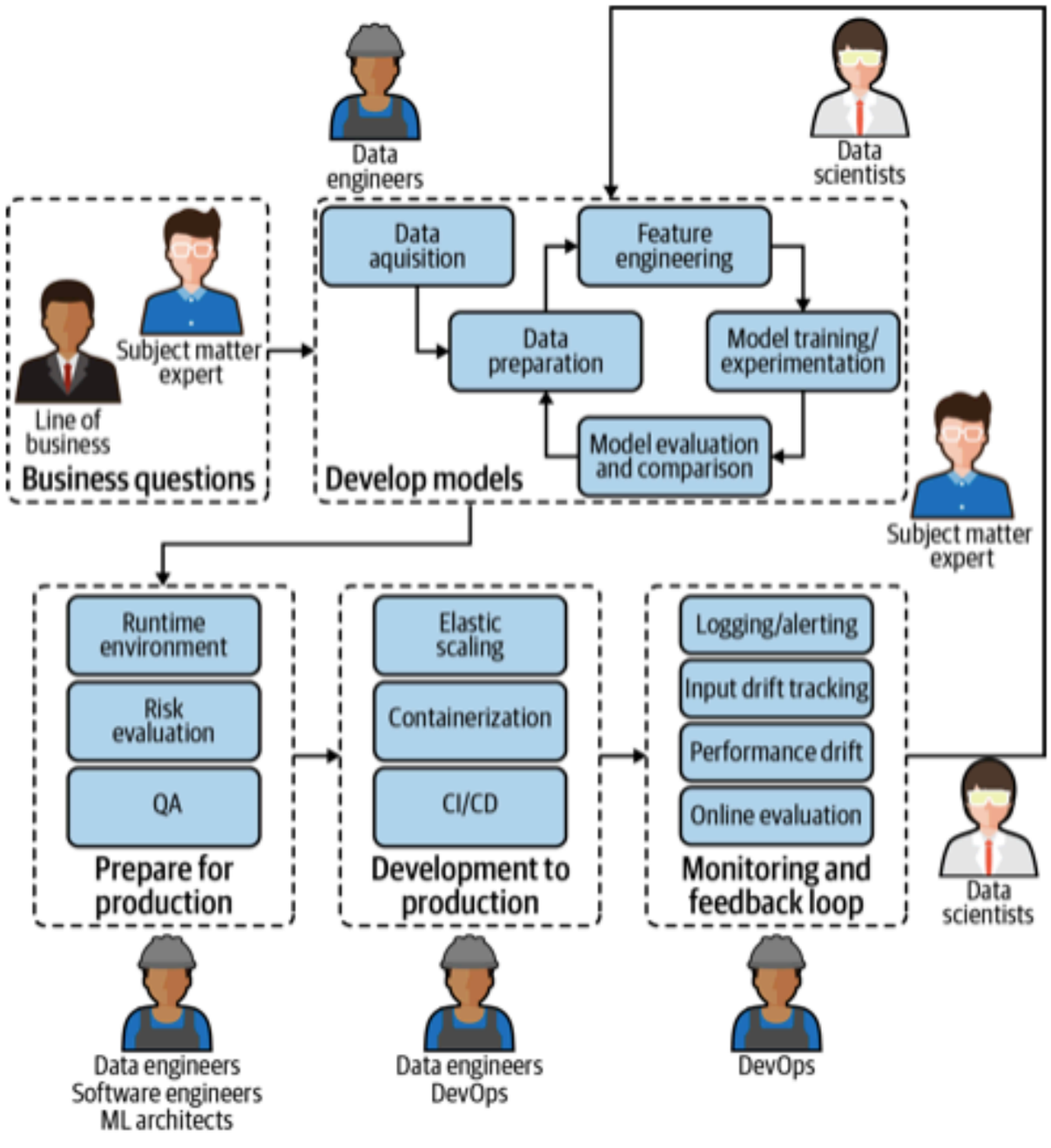


ML Lifecycle

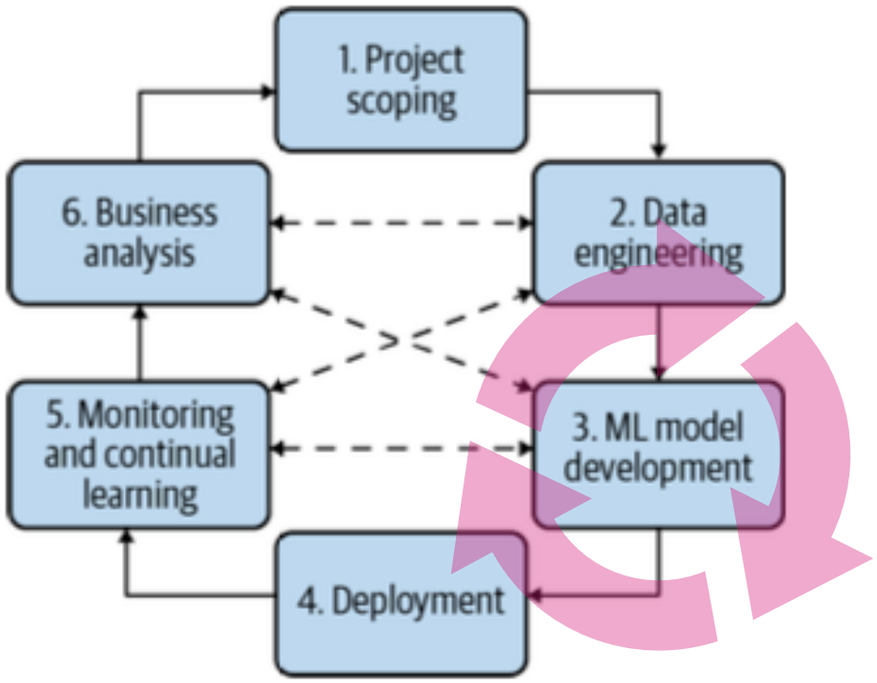


ML Lifecycle

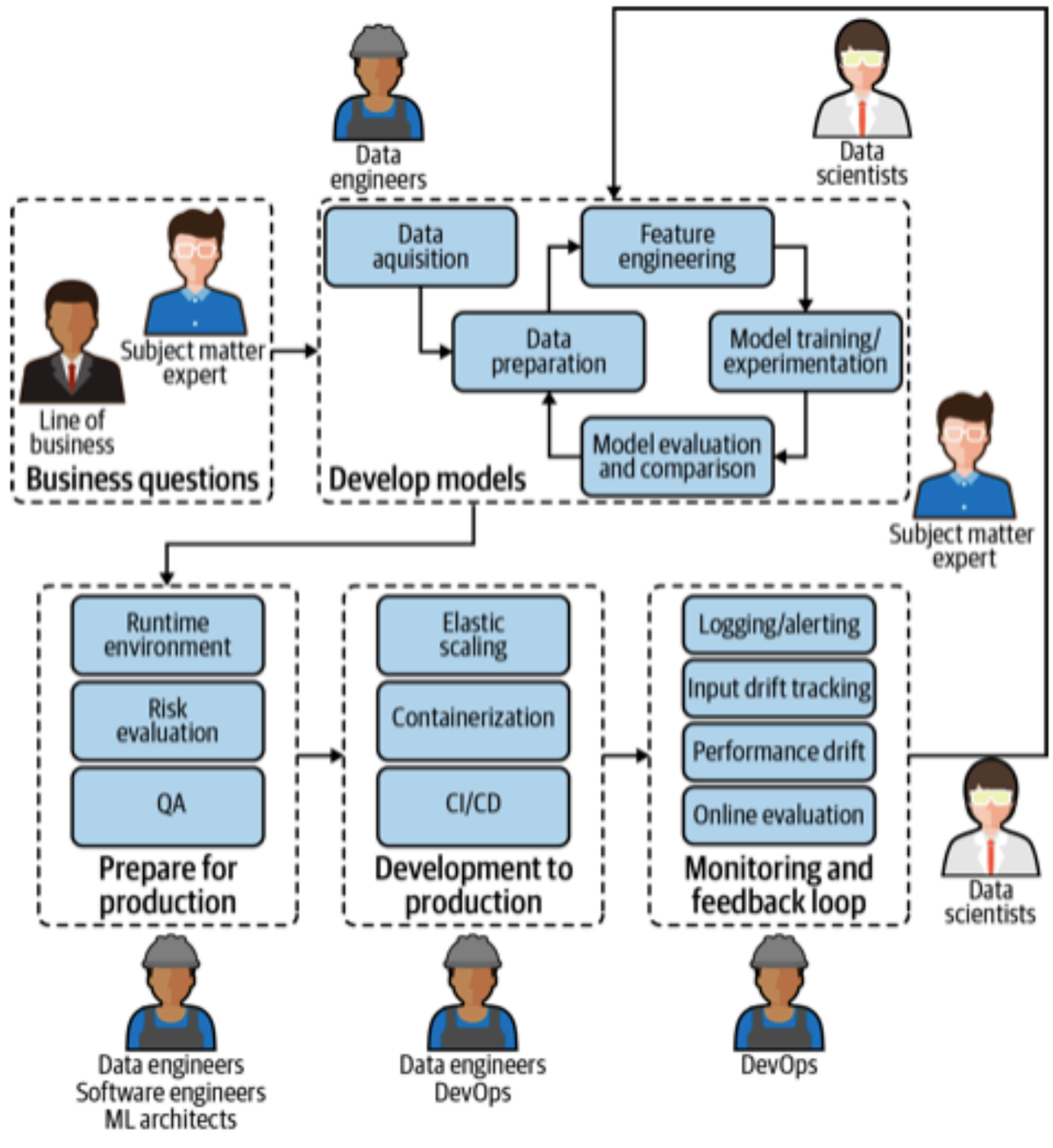




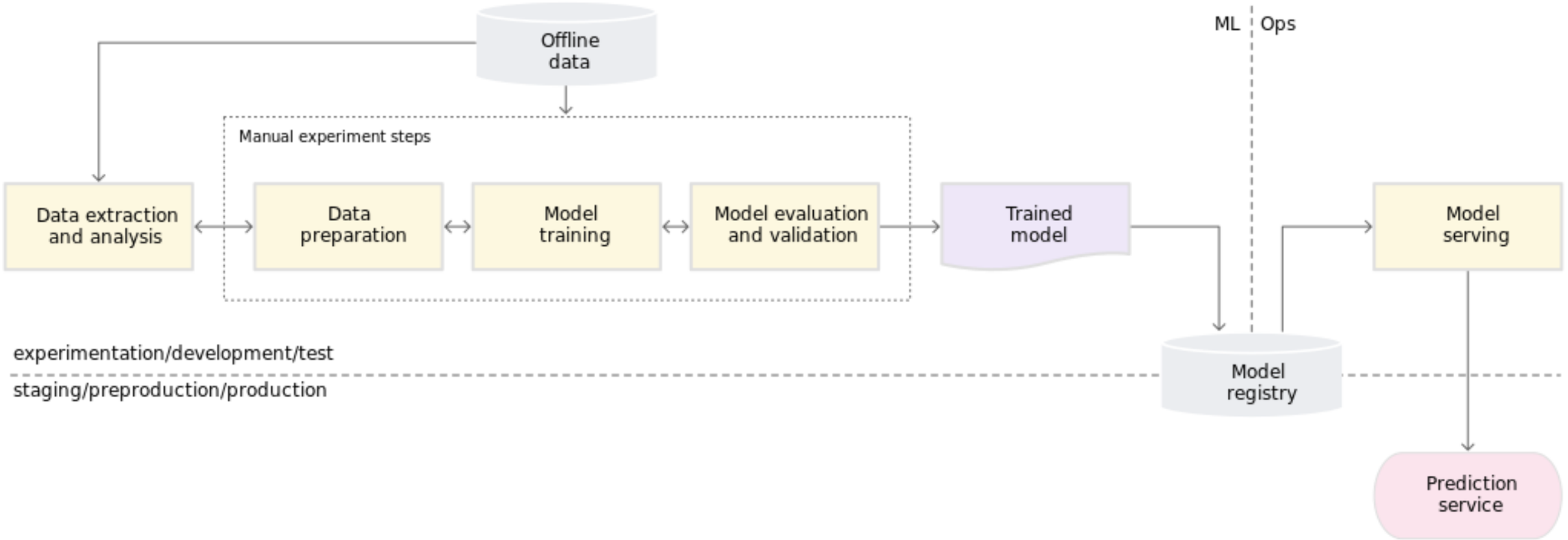
ML Lifecycle



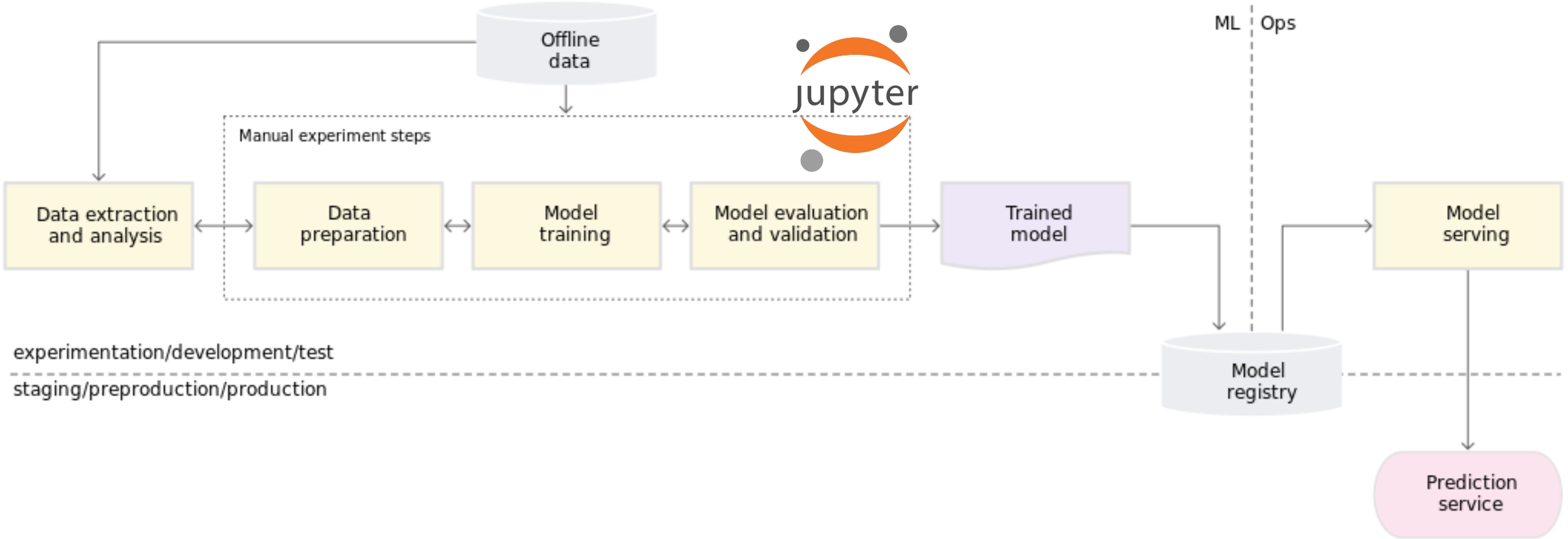
The **level of automation** of these steps defines the **maturity** of the ML process, which reflects the **velocity** of training new models given new data or new implementations



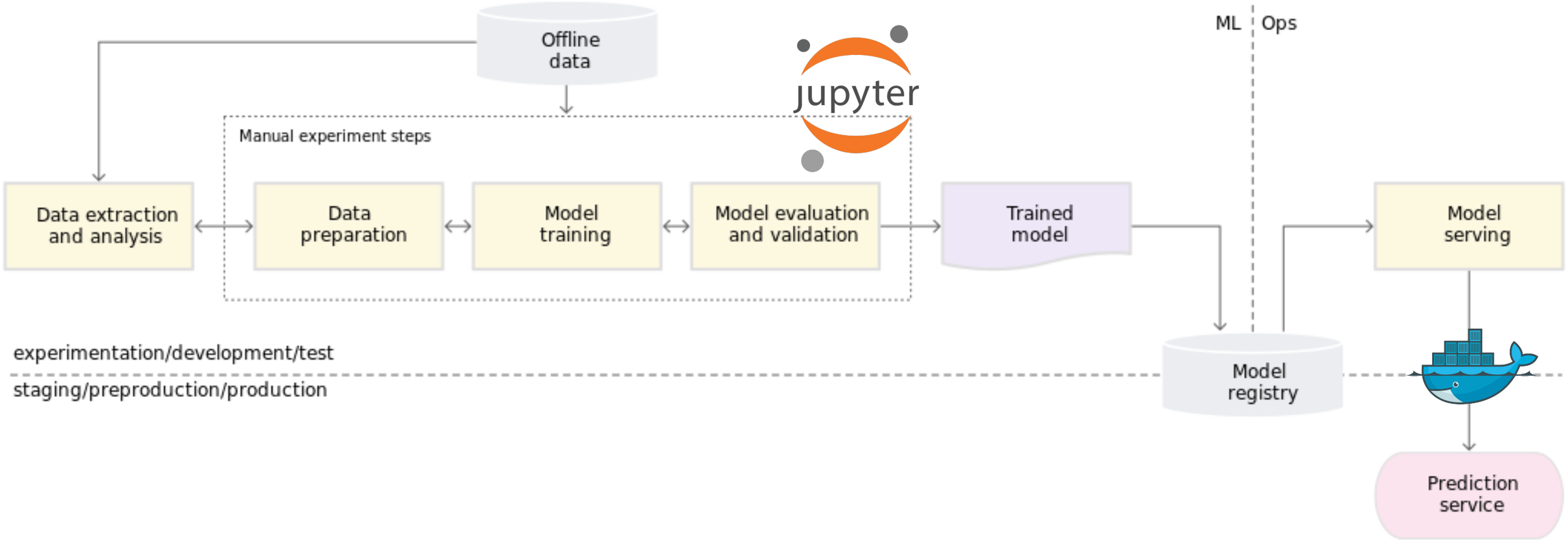
MLOps level 0



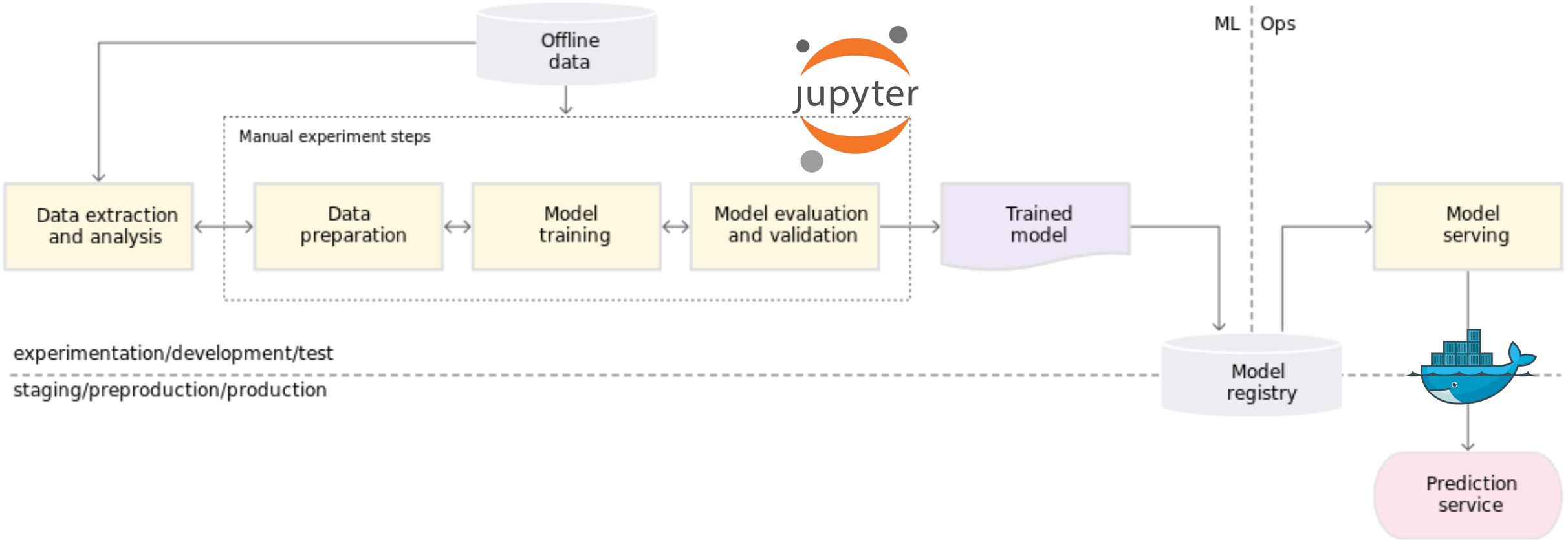
MLOps level 0



MLOps level 0

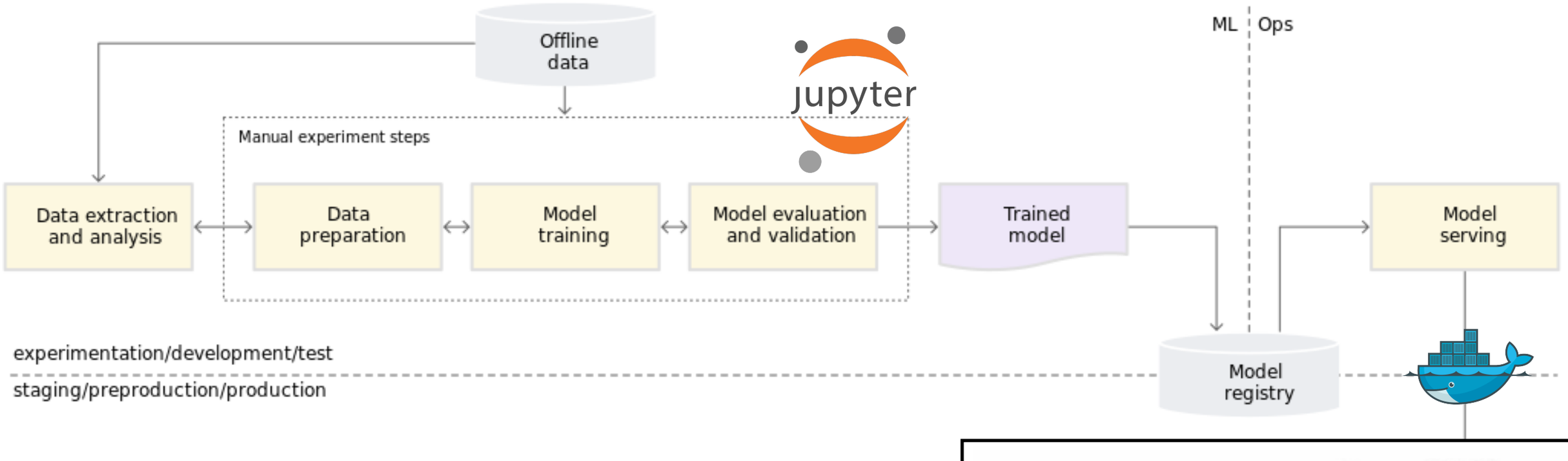


MLOps level 0

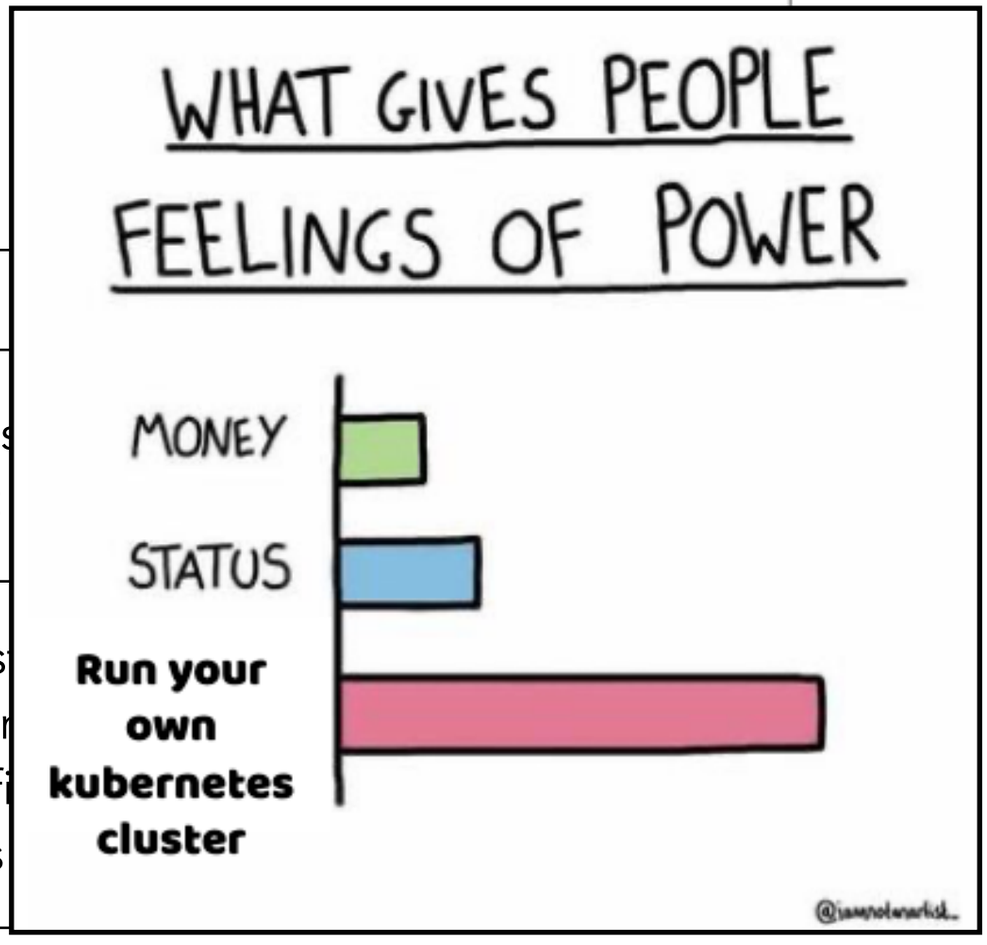


	Dev	Prod
Scale	• usually one instance • no need to worry about autoscaling	• multiple istances/nodes • you most certainly want autoscaling
State	stateful: reproducible but inflexible • install dependencies once and forget • can persist data in dedicated storage	inherently stateless: flexible but hard to reproduce • need to install dependencies on any new instance • need to figure out how to persist data/state across instances

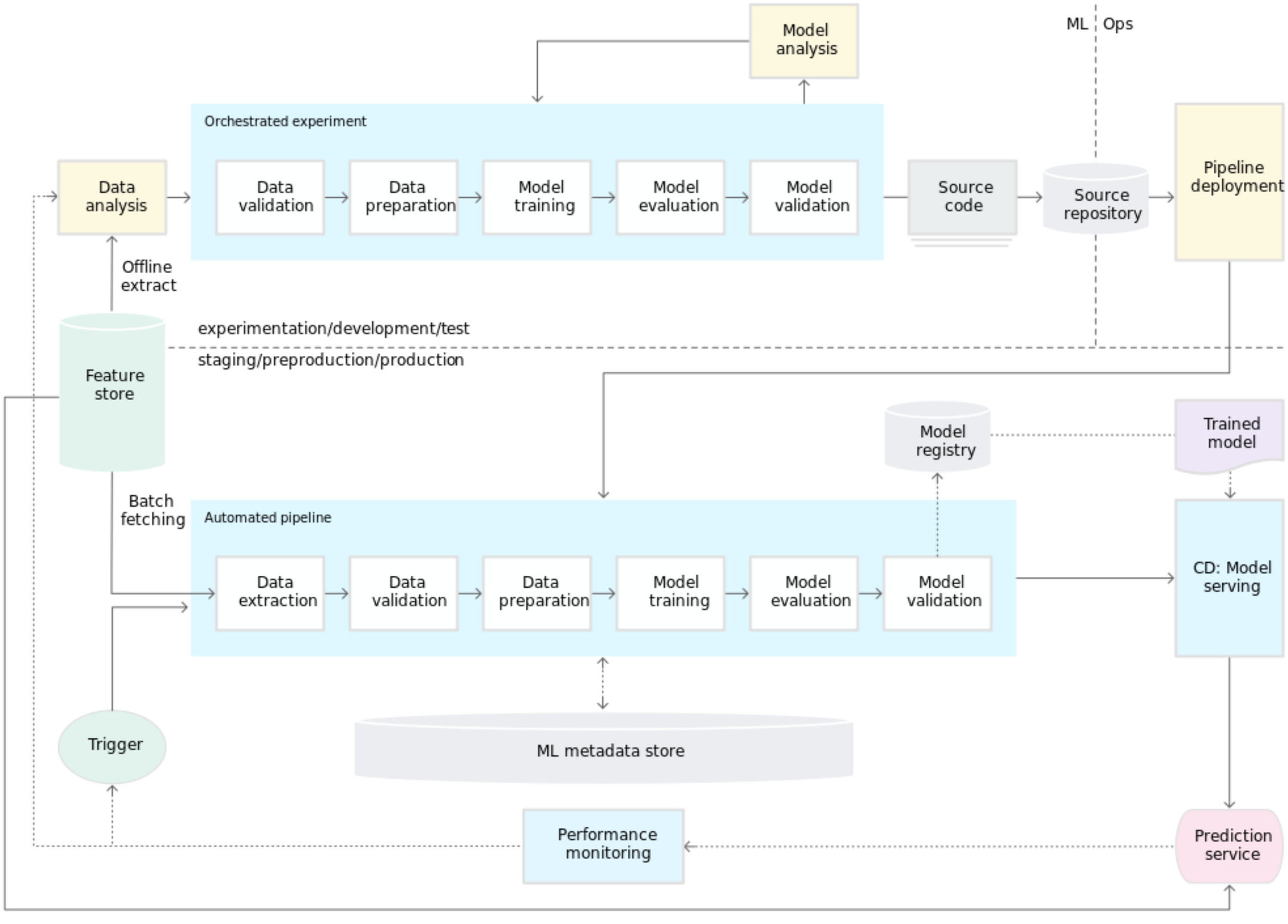
MLOps level 0



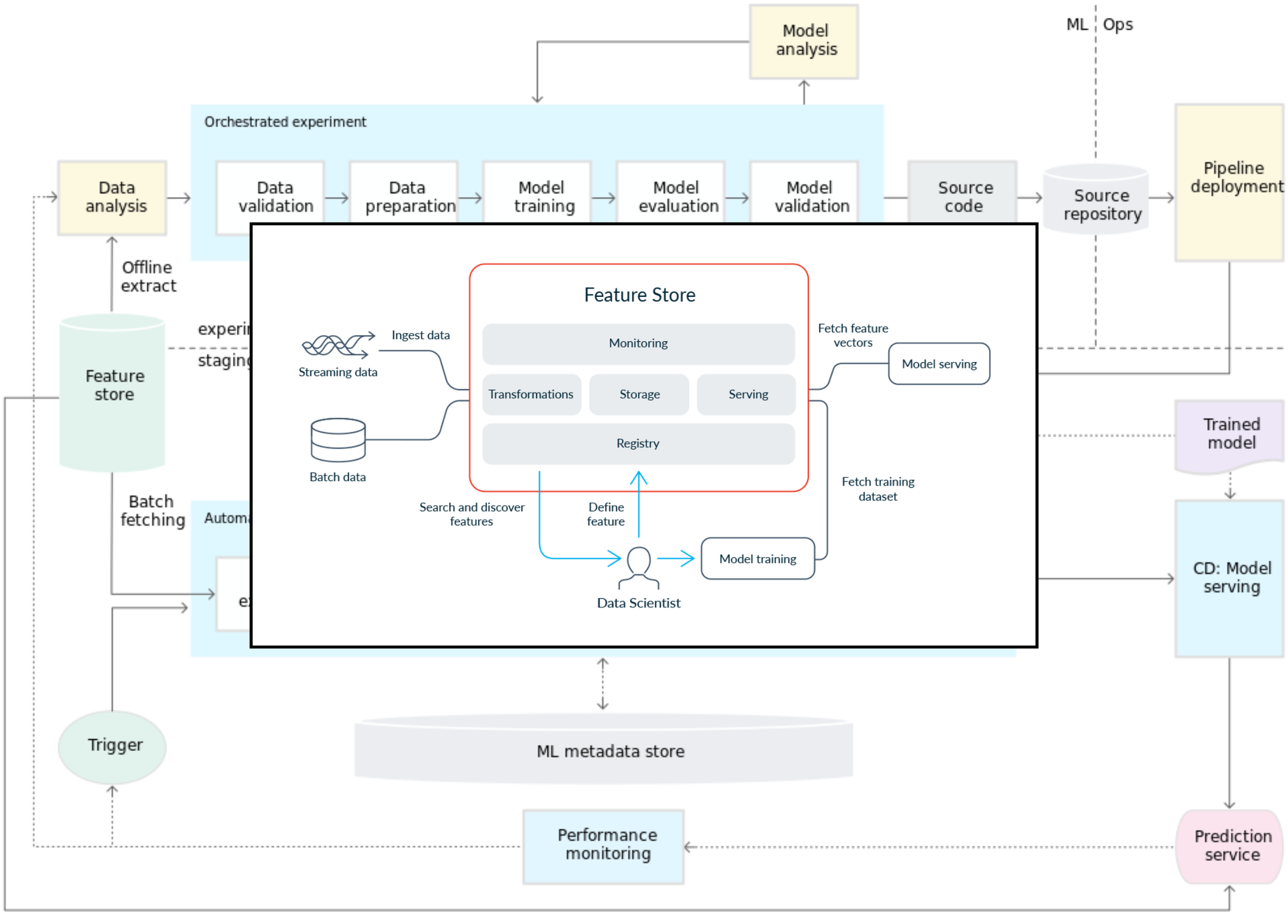
	Dev	
Scale	• usually one instance • no need to worry about autoscaling	• multiple is • you most
State	stateful: reproducible but inflexible • install dependencies once and forget • can persist data in dedicated storage	inherently s • need to ir • need to fi instances



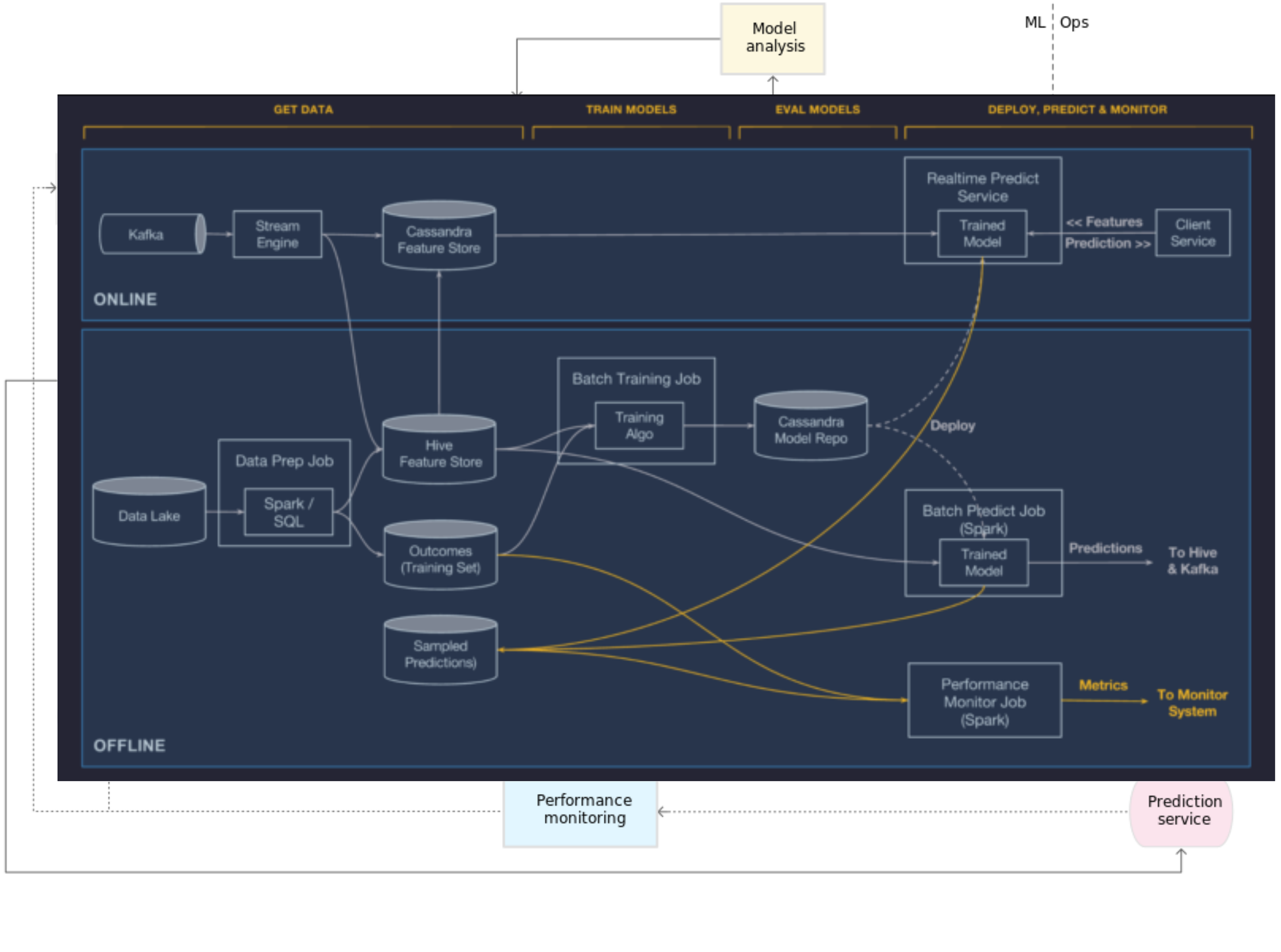
MLOps level 1



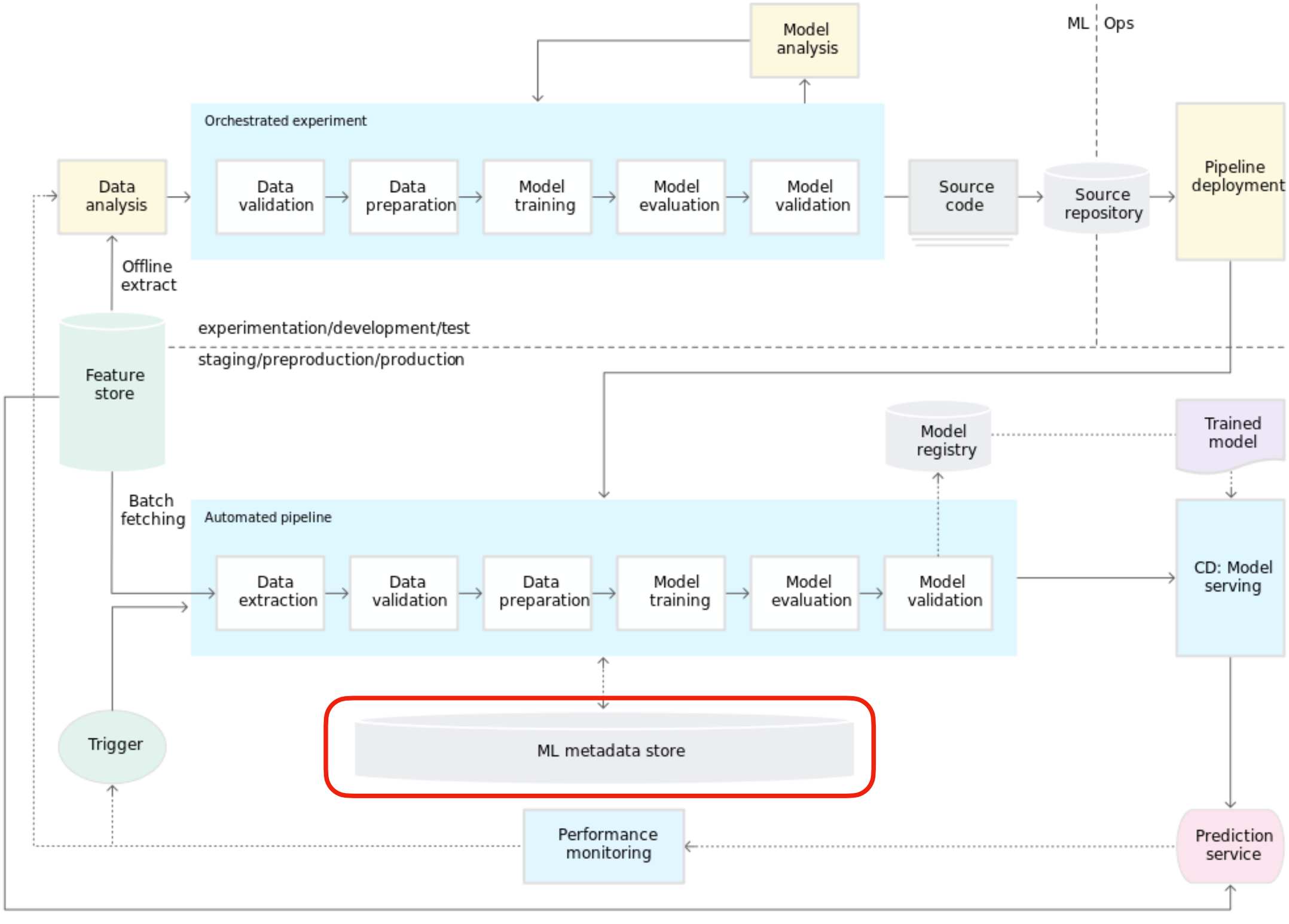
MLOps level 1



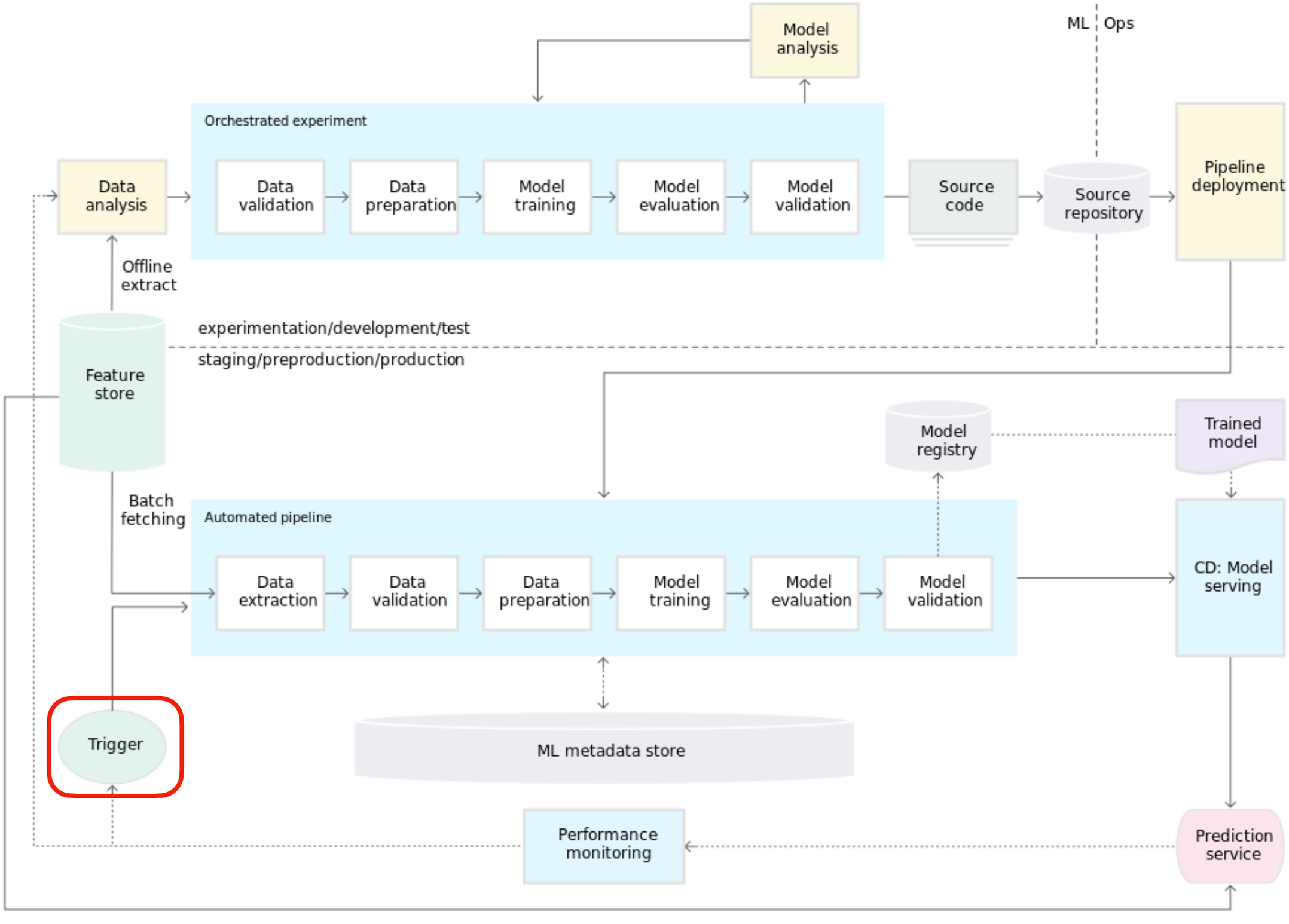
MLOps level 1



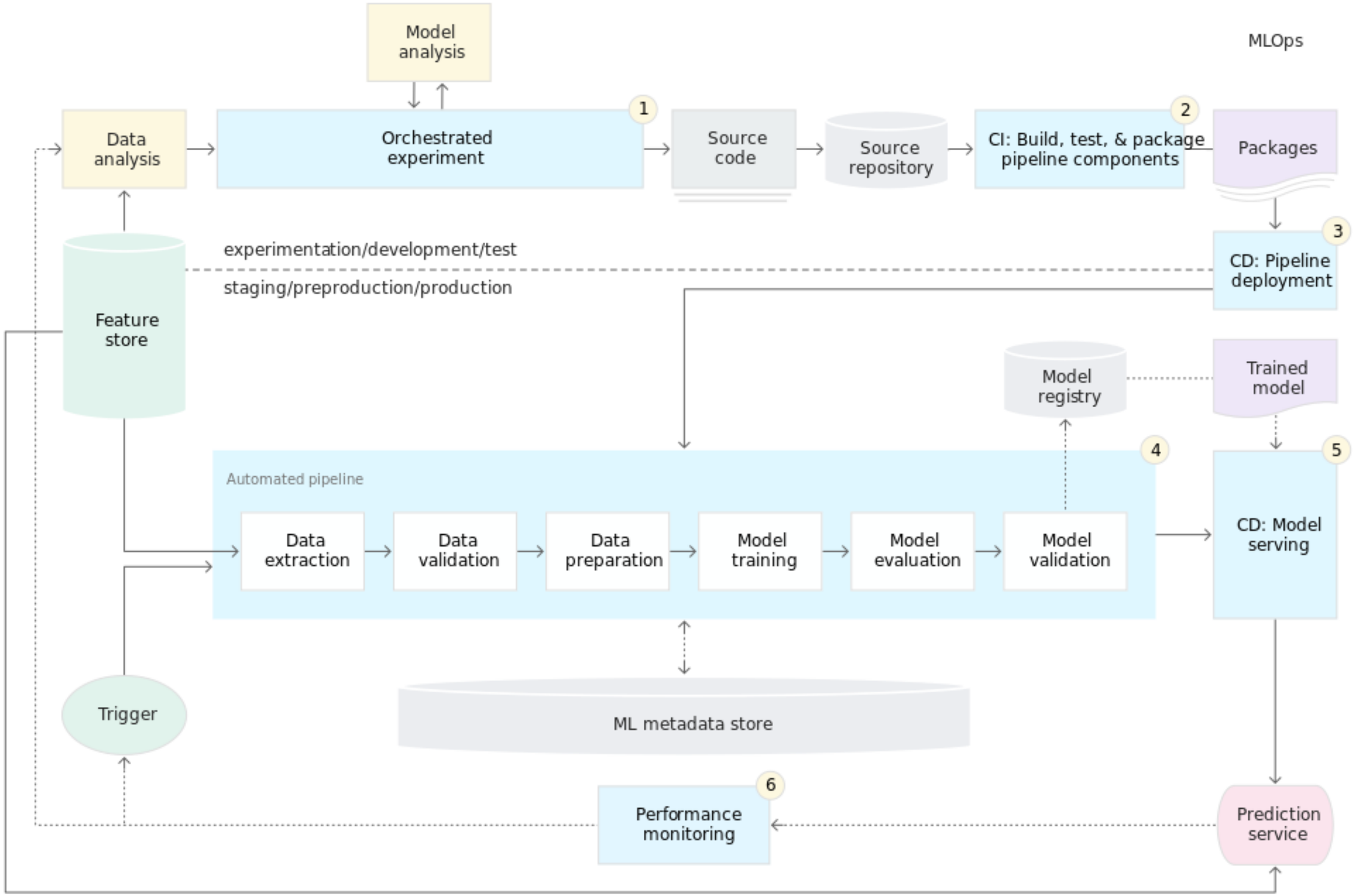
MLOps level 1



MLOps level 1



MLOps level 2



Summary

Summary

Implementing ML in a production environment doesn't **only** mean deploying your model as an API **for prediction** but rather deploying an ML pipeline that can **automate the retraining and deployment** of new models.

Setting up a **CI/CD** system enables you to **automatically test and deploy** new pipeline implementations.

This system lets you **cope with rapid changes** in your data and business environment.

You don't have to immediately move all of your processes from **one level to another**.

MLOps tools

MLOps tools

BIG DATA & AI LANDSCAPE 2018

INFRASTRUCTURE

HADOOP ON-PREMISE
cloudera · Hortonworks · MAPR · Pivotal · IBM InfoSphere · bluedata · jethro

HADOOP IN THE CLOUD
aws · Microsoft Azure · Google Cloud · IBM InfoSphere BigInsights · TREASURE DATA · Qubole · altiscale · CAZENA · CenturyLink

STREAMING / IN-MEMORY
aws · databricks · striim · confluent · GridGain · ORACLE Coherence · dataArtisans · hazelcast · TERRACOTTA · kx · FASTDATA · Wallaroo LABS

NoSQL DATABASES
Google Cloud · aws · ORACLE · Microsoft Azure · mongoDB · MarkLogic · AEROSPIKE · DATASTAX · ArangoDB · Couchbase · redislabs · SCYLLA

NewSQL DATABASES
SAP HANA · Clustrix · nuoDB · Pivotal · Cloud Spanner · Cockroach LABS · MEMSQL · influxdata · MariaDB · TIMESCALE · VOLTDB · citusdata · splice MACHINE · paradigm4 · Trafodion · TiDB

GRAPH DBs
neo4j · Amazon Neptune · IBM · ORACLE · OrientDB · GIRAPH · InfiniteGraph · Objectivity

MPP DBs
TERADATA · VERTICA · IBM Data Warehouse Systems · Actian · Kognitio · Exasol · dremio

CLOUD EDW
aws · Google Cloud · Microsoft Azure · Pivotal · snowflake

DATA TRANSFORMATION
talend · pentaho · alteryx · TRIFACTA · tamr · Paxata · StreamSets · UNIFI

DATA INTEGRATION
SAP Data Services · Informatica · MuleSoft · TEALIUM · snapLogic · enigma · podium data · Segment · alooma · xplenty · ZALONI · Stitch · import.io · Infoworks · ATTUNITY

DATA GOVERNANCE
Informatica · SailPoint · IBM · McAfee Skyhigh Security Cloud · collibra · Alation · Waterline Data · IMMUTA · OKERA

MGMT / MONITORING
aws · New Relic · actifio · rubrik · APPDYNAMICS · DATADOG · WAVEFRONT by vmware · dynatrace · splunk> · SignalFx · druva · Moogsoft · unravel · pagerduty · Numerify · Anodot

STORAGE
aws · Google Cloud · Microsoft Azure · IBM Storage · PURESTORAGE · ALLUXIO · nimblestorage · Qumulo · panasas · COHESITY

CLUSTER SVCS
aws · kubernetes · docker · MESOSPHERE · HPCC SYSTEMS · CoreOS · pepperdata

APP DEV
Lightbend · Keen IO · rainforest · CASK

CROWD-SOURCING
amazon mechanicalturk · Upwork · appen · figure eight · scale · HIVE

HARDWARE
Google TPU · arm · intel AI · GRAPHCORE · IBM Power Systems · MYTHIC · NVIDIA · cerebras · Movidius · WAVE COMPUTING · HAILO

GPU DBs
kinetica · MAPD · SQREAM TECHNOLOGIES · blazegraph · BLAZINGDB · brytlyt · PG-Strom

ANALYTICS

DATA ANALYST PLATFORMS
Microsoft · pentaho · alteryx · Digital Reasoning · GUAVUS · AYASDI · ATTIVIO · Datameer · Quid · incorta · interana · ClearStory · Origami logic · ASCEND.IO · ENDOR · MODE · Bottlenose · switchboard

DATA SCIENCE PLATFORMS
IBM · KNIME · data iku · DOMINO · rapidminer · CONTINUUM ANALYTICS · ALGORITHMIA · DATAWATCH · ANGOSS · SAS

BI PLATFORMS
Microsoft · aws · DOMO · Wave Analytics · looker · THOUGHTSPOT · ATSCALE · ARCADIA DATA · Information Builders · SISENSE · GoodData · birst · MicroStrategy

VISUALIZATION
+tableau · SAP Lumira · Google Cloud · celonis · Qlik · Periscope Data · ZEPL · GOODATA · plotly · CHARTIO · TOUCAN TOCO

MACHINE LEARNING
Azure Machine Learning · aws · Google Cloud · H2O.ai · DataRobot · gamalon · ELEMENT AI · VERSIVE · ViSENZE · deepsense.io · bonsai

COMPUTER VISION
Microsoft Azure · Amazon Rekognition · Cloud Vision API · clarifai · 商汤 · EVER AI · deepomatic · twenty bn · neurala

HORIZONTAL AI
IBM Watson · Cortana · Face++ 旷视 · sentient technologies · Voyager Labs · vicarious · :) Affectiva · PROPHESEE · CognitiveScale · Numenta · PETUUM · SCALED INFERENCE · narrative logics · CURIOUS AI · OSARO · BLUE VISION

SPEECH & NLP
Google Cloud · twilio · amazon alexa · narrative science · WolframAlpha · semantic machines · Mobvoi · EigenTechnologies · SoundHound Inc. · voicera · PRIMER · MindMeld · cortical.io · NUANCE · Gridspace · snips · yseop

SEARCH
elasticsearch · ORACLE ENDECA · EXALEAD · coveo · Lucidworks · ATTIVIO · swiftype · algolia · alphasense · MAANA · omni:us · SINEQUA

LOG ANALYTICS
splunk> · sumologic · LOGGLY · TIMBER · kibana · logz.io

SOCIAL ANALYTICS
Hootsuite · sprinklr · NETBASE · synthesio · tracx · simplereach · bitly · predata · SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics · mixpanel · AMPLITUDE · sumall · Airtable · RESCI · SIGOPT · granify · custora

APPLICATIONS – ENTERPRISE

SALES
einstein · CHORUS · INSIDESALES.COM · conversica · GONG · clari · aviso · tact.ai · fuse|machines · TROOPS

MARKETING - B2B
RADIUS · App Annie · EVERSTRING · Lattice · MI · MINTIGO · 6sense · RK · Reflektion · tubular · DataFox · ENGAGIO · mrp

MARKETING - B2C
zeta · bloomreach · SendGrid · BlueYonder · [PERSADO] · kahuna · ACTIONIQ · SAILTHRU · BLUECORE · QUANTIFIND · mparticle · Amplero · amperity · TEALIUM · Simon · Lytics

CUSTOMER SERVICE
MEDALLIA · zendesk · CLARABRIDGE · Gainsight · NG DATA · DigitalGenius · afiniti · AUTOMAT · frame.ai · msg.ai · INTERCOM · CallDesk

HUMAN CAPITAL
HireVue · entelo · hiQ · GIGSTER · textio · RESTLESS BANDIT · Wade&Wendy · Stella · Clustree · pymetrics · mya · uncommon

LEGAL
RAVEL LAW · Seal · Everlaw · JUDICATA · EBREVIA · IRONCLAD · PREMONITION · ROSS · casetext

FINANCE
Anaplan · zuora · S/4 HANA · TRADESHIFT

ENTERPRISE PRODUCTIVITY
slack · X. · ORACLE · lumiata · DIFFBOT · clara · talla · butter.ai · Kasisto

BACK OFFICE AUTOMATION
UiPath · HyperScience · blueprism · Automation Anywhere · cognicor · AppZen · WorkFusion

SECURITY
TANIUM · CYLANCE · zscaler · StackPath · illumio · CODE42 · CipherCloud · DARKTRACE · ANOMALI · ThreatMetrix · VECTRA · cybereason · Guardian Analytics · DATAVISOR · siftscience · SIGNIFYD · SentinelOne · SecurityScorecard · SOCURE · BlueTalon · Recorded Future · feedzai · CyberX · AREA 1 SECURITY · sparkcognition · IronNet Cybersecurity

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus · MediaMath · criteo · xAd · Integral Ad Science · ORACLE MOAT · OpenX · datorama · theTradeDesk · Adgorithms · dstillery · LiveIntent · dataxu · gumgum · TAPAD · DYNAMIC YIELD · teemo · yieldmo

EDUCATION
Liulishuo · KNEWTON · Clever · Declara · kidaptive · PANORAMA · knowre · gradescope

GOVERNMENT
OPENGOV · mark43 · FiscalNote · GRIDSMART · LiveStories · Passport · SmartProcure · STREETLIGHTDATA · OpenDataSoft

REAL ESTATE
REDFIN · Opendoor · VTS · CREDIFI · reonomy · COMPSTAK · CAPE ANALYTICS

FINANCE - INVESTING
KENSHO · Dataminr · Quantopian · ADDEPAR · NUMERAI · iSENTIUM · ALGORIZ · RavenPack · PAGAYA

FINANCE - LENDING
ondeck · Affirm · 拍拍贷 · JIANPU.AI · Kreditech · AVANT · TALA · zestfinance · Upstart · INSIKT · upgrade · 100Credit · WeLab · Wecash闪现普惠 · TrueAccord · MoneyLion · Active.Ai · aire · cignifi

INSURANCE
Metromile · Lemonade · CYENCE · Shift Technology · TRACTABLE

HEALTHCARE
flatiron · Clover · KYRUUS · HealthTap · METABIOTA · Ginger.io · Glow · babylon · 3DMed · zebra · PathAI · ovia · TEMPUS · patientslikeme · AiCure · RECURSION · prognos · enlitic · imagia · Qventus · BAYLABS · ARTERYS · CLOUD MEDX · IMAGEN · Kang Health · PAIGE · DATAVANT · HUMAN DIAGNOSIS PROJECT · innovaccer · LeanTaaS

LIFE SCIENCES
23andMe · color · iCarbonX · BenevolentAI · verily · WuXi NextCODE · ZEPHYR HEALTH · Zymergen · PATHWAY GENOMICS · SCRN · freenome · Clear Labs · DNAnexus · NANOPORE · CITRINE · Phosphorus · twoXAR · Atomwise · deep genomics · SOPHIA GENETICS · OWKIN

TRANSPORTATION
UBER · TESLA · ZOOX · CLEARPATH · nuTonomy · drive.ai · nauto · AIMOTIVE · PILOT.AI · NIO · OPTIMUS RIDE · moovit · nexar · comma.ai · netradyne · Civil Maps · German Autolabs

AGRICULTURE
FARMERS · Granular · JOHN DEERE · BLUE RIVER TECHNOLOGY · FarmersEdge · FarmLogs · TARANIS · GAMAYA · TerrAvion · prospera

COMMERCE
Instacart · STITCH FIX · Dia & Co · RetailNext · HowGood · heuritech

INDUSTRIAL
AVEVA · SIEMENS · Schneider Electric · PREDIX · GE Digital · UPTAKE · TACHYUS · SIGHT MACHINE · Alluvium · SCORTEX

OTHER
eharmony · stem · rethink robotics · Amper · ByteDance · hopper · celect · BOXEVER · VERDIGRIS · duetto · Unbabel · Jukedeck · Second Spectrum · remesh · ASAPP

CROSS-INFRASTRUCTURE/ANALYTICS
aws · Google Cloud · Microsoft · IBM · SAP · Hewlett Packard Enterprise · SAS · 1010DATA · vmware · TIBCO · TERADATA · ORACLE · NetApp · syncsort · MAPR · cloudera

OPEN SOURCE

FRAMEWORK
hadoop HDFS · Apache Kylin · hadoop MapReduce · YARN · TEZ · Flink · MESOS · Spark · CDAP

QUERY / DATA FLOW
Spark SQL · HIVE · presto · SLAMDATA · APACHE DRILL · Google Cloud Dataflow · Flink

DATA ACCESS
cassandra · nifi · mongoDB · CouchDB · SciDB · riak · OPENTSDB · APACHE HBASE · Cloud Spanner · accumulo

COORDINATION
talend · Apache Zookeeper · Apache Ambari · Apache Airflow

STREAMING
Spark · APEX · Flink · beam · kafka · druid · STORM

STAT TOOLS
python · R · ScalaLab · NumPy · SciPy · julia

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow · theano · torch · MADlib · learn · Caffe · Microsoft Cognitive Toolkit · OpenAI · DMTK · Keras · PaddlePaddle · Apache SINGA · FeatureFu · mxnet · Chainer · VELES · DIMSUM · WEKA · neon™ · DSSTNE · mllib · DL4J · MAHOUT · Aerosolve

SEARCH
elasticsearch · Solr · Lucene

LOGGING & MONITORING
elasticsearch · kibana · SENTRY · logstash · Prometheus

VISUALIZATION
BeakerX · Rodeo

COLLABORATION
jupyter · Zeppelin · ANACONDA

SECURITY
Apache Ranger · KNOX · Sentry

DATA SOURCES & APIs

HEALTH
VALIDIC · practice fusion · fitbit · GARMIN · HUMAN API · kinsa

IOT
GE Digital · UPTAKE · thingworx · helium · samsara · AUGURY · estimote

FINANCIAL & ECONOMIC DATA
Bloomberg · THOMSON REUTERS · DOW JONES · S&P CAPITAL IQ · CB INSIGHTS · xignite · Quandl · ENVESTNET YODLEE · PREMISE · estimize · SECOND MEASURE · Eagle Alpha · StockTwits · PLAID · Thinknum · earnest

AIR / SPACE / SEA
Orbital Insight · planet. · SKYCATCH · Airware · AIROBOTICS · spire · kespry · PRECISIONHAWK · UNDERSTORY · Descartes Labs · WINDWARD · telluslabs · DroneDeploy · MarineTraffic

PEOPLE / ENTITIES
acxiom · experian · EPSILON · InsideView · Crimson Hexagon · BASIS TECHNOLOGY · Quantcast · SAFEGRAPH

LOCATION INTELLIGENCE
FOURSQUARE · MapAnything · mapbox · sense360 · pitney bowes · HEXAGON GEOSPATIAL · PlaceIQ · esri · factual · CARTO · Mapillary · Streetline · cuebiq · Radar

OTHER
qualtrics · DATA.GOV · data.world · enigma · mobilewalla · CRUX · HIGHCHARTS

DATA RESOURCES

DATA SERVICES
Palantir · Mu Sigma · OPERA · SILICON VALLEY DATA SCIENCE · fractal · kaggle · DataKind · EXL · INNOPLEXUS

INCUBATORS & SCHOOLS
PLURALSIGHT · GA · galvanize · DataCamp · DataElite · INSIGHT · The Data Incubator · METIS

RESEARCH
OpenAI · facebook research · MILA · MIRI · VECTOR INSTITUTE · CSAIL · DFKI · AI2 · ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Final 2018 version, updated 07/15/2018

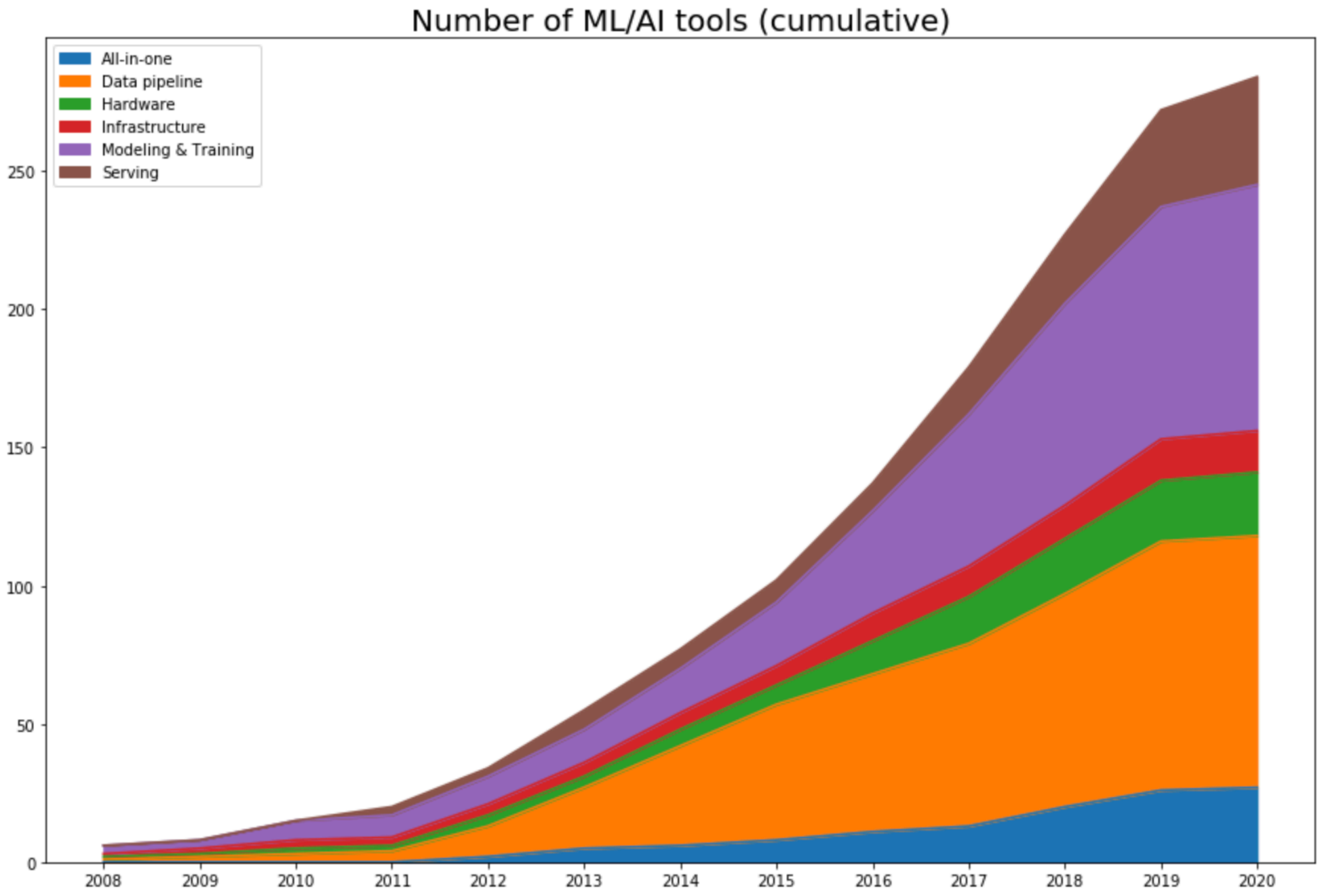
mattturck.com/bigdata2018

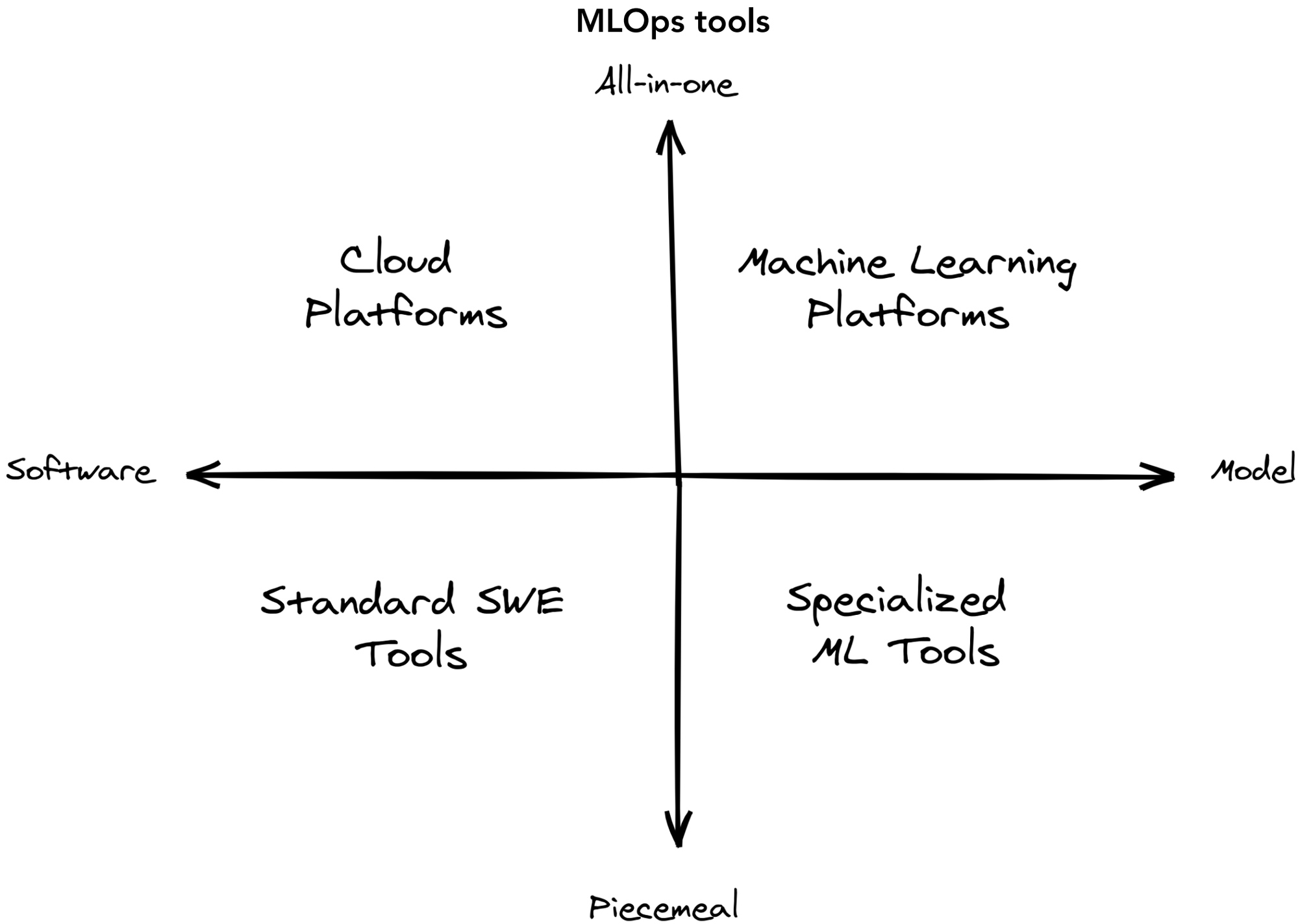
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

MLOps tools

MLOps Principles	Data	ML Model	Code
Versioning	1) Data preparation pipelines 2) Features store 3) Datasets 4) Metadata	1) ML model training pipeline 2) ML model (object) 3) Hyperparameters 4) Experiment tracking	1) Application code 2) Configurations
Testing	1) Data Validation (error detection) 2) Feature creation unit testing	1) Model specification is unit tested 2) ML model training pipeline is integration tested 3) ML model is validated before being operationalized 4) ML model staleness test (in production) 5) Testing ML model relevance and correctness 6) Testing non-functional requirements (security, fairness, interpretability)	1) Unit testing 2) Integration testing for the end-to-end pipeline
Automation	1) Data transformation 2) Feature creation and manipulation	1) Data engineering pipeline 2) ML model training pipeline 3) Hyperparameter/Parameter selection	1) ML model deployment with CI/CD 2) Application build
Reproducibility	1) Backup data 2) Data versioning 3) Extract metadata 4) Versioning of feature engineering	1) Hyperparameter tuning is identical between dev and prod 2) The order of features is the same 3) Ensemble learning: the combination of ML models is same 4)The model pseudo-code is documented	1) Versions of all dependencies in dev and prod are identical 2) Same technical stack for dev and production environments 3) Reproducing results by providing container images or virtual machines
Deployment	1) Feature store is used in dev and prod environments	1) Containerization of the ML stack 2) REST API 3) On-premise, cloud, or edge	1) On-premise, cloud, or edge
Monitoring	1) Data distribution changes (training vs. serving data) 2) Training vs serving features	1) ML model decay 2) Numerical stability 3) Computational performance of the ML model	1) Predictive quality of the application on serving data

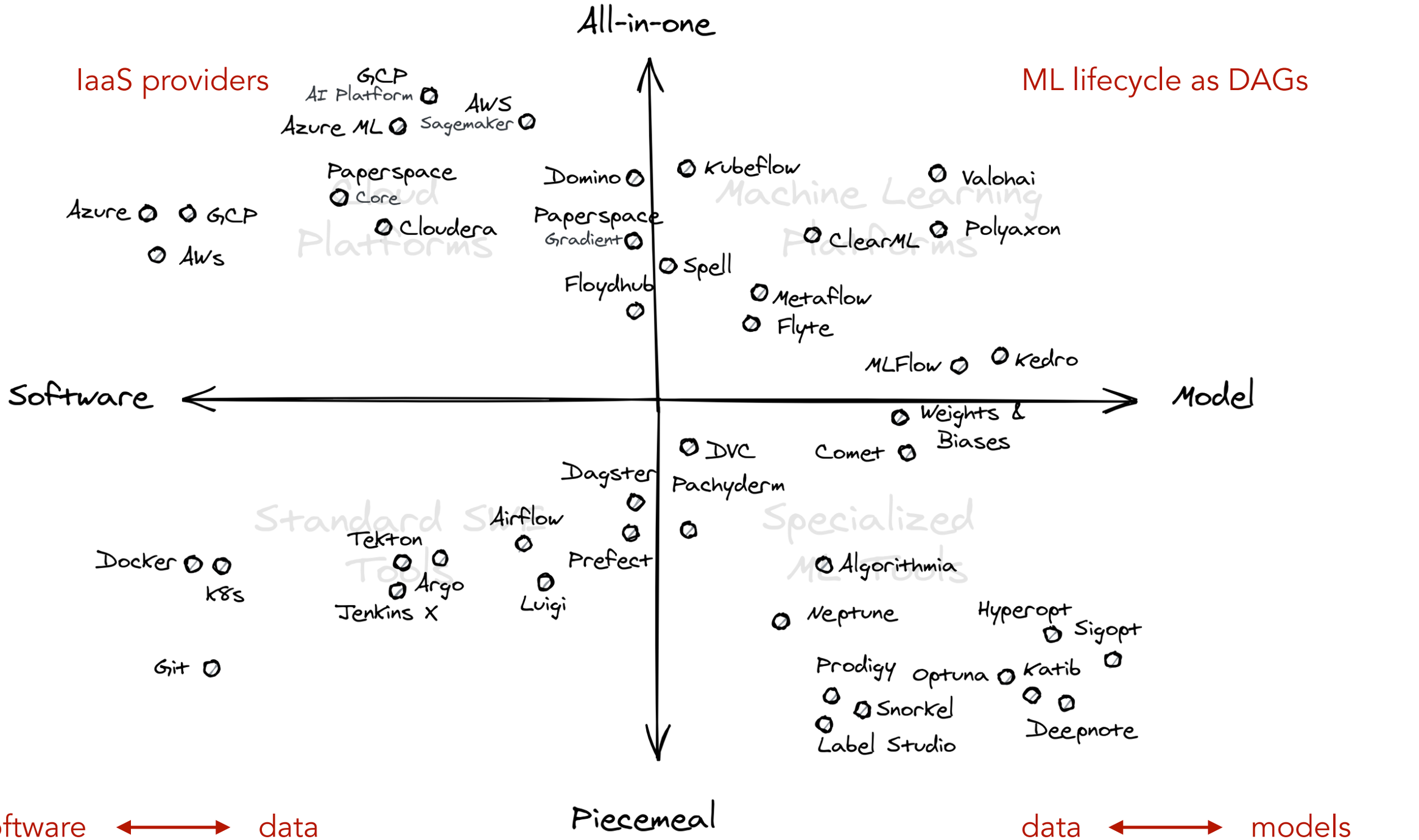
MLOps tools


Number of ML/AI tools (cumulative)
All-in-one
Data pipeline
Hardware
Infrastructure
Modeling & Training
Serving
0
50
100
150
200
250
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020



x-axis: artifact
y-axis: scope

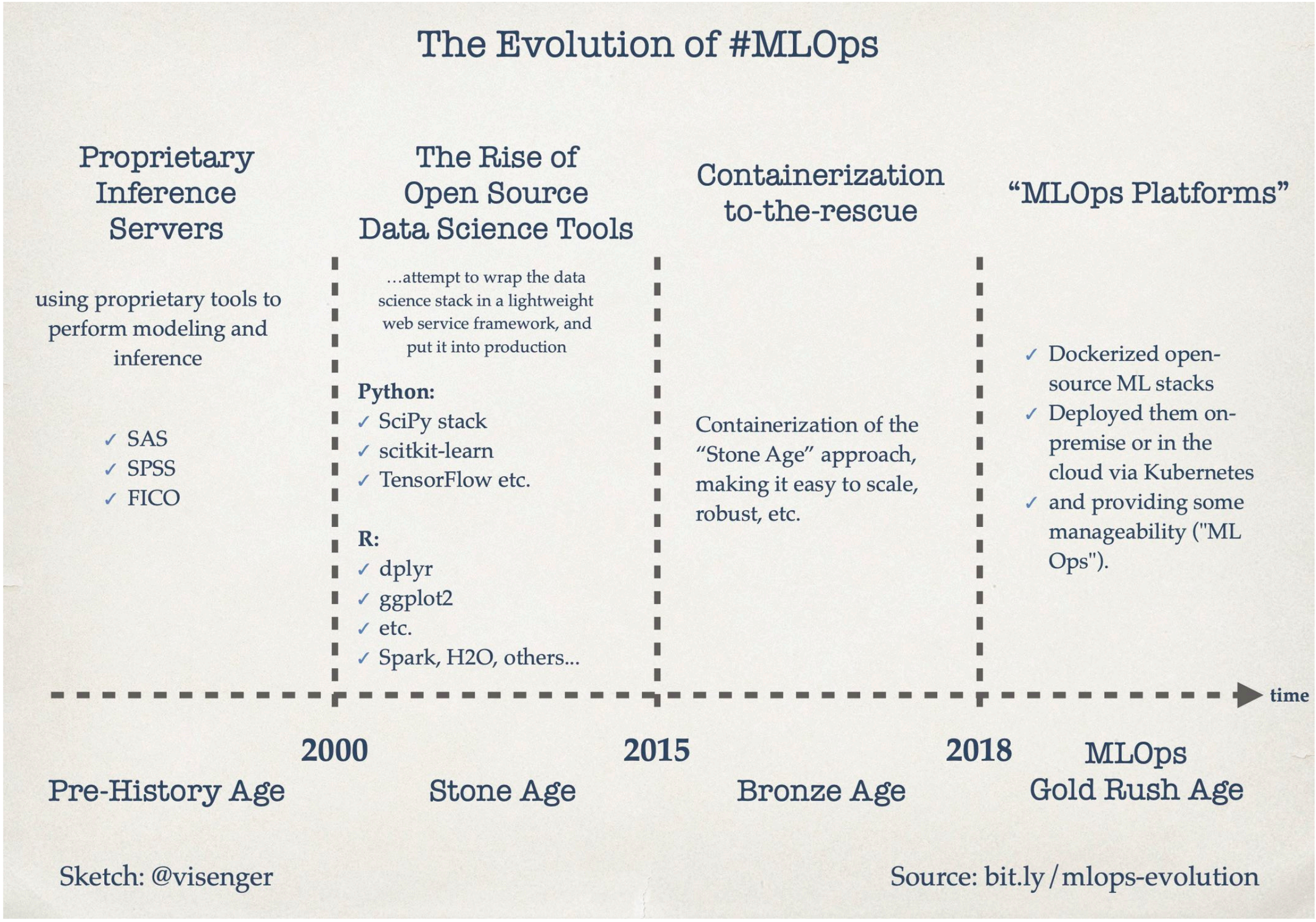
MLOps tools



x-axis: artifact

y-axis: scope

Predictions



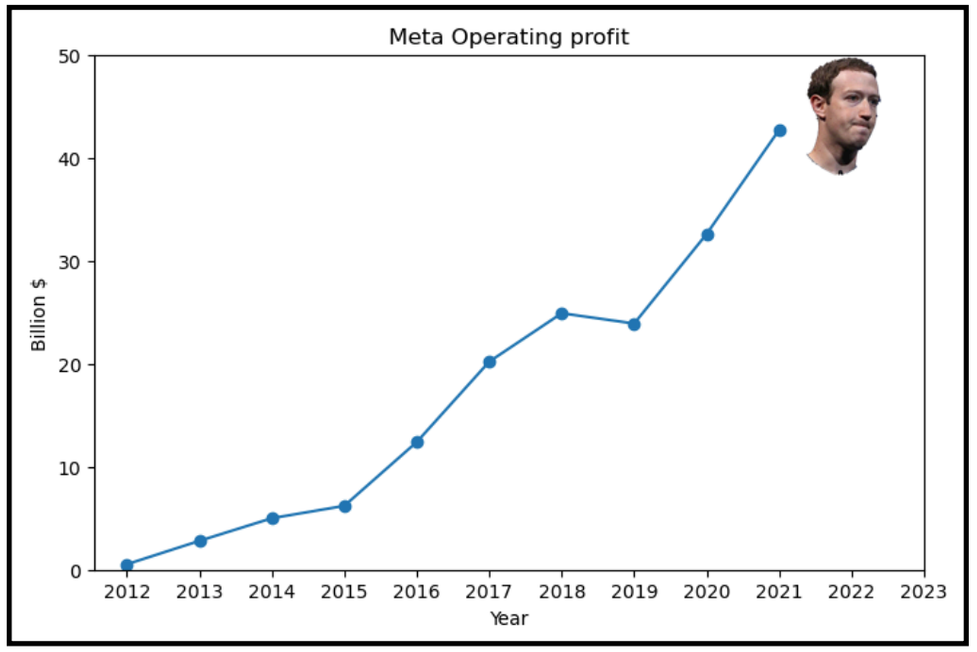
Signals from Big AI Companies

Signals from Big AI Companies

Layoffs

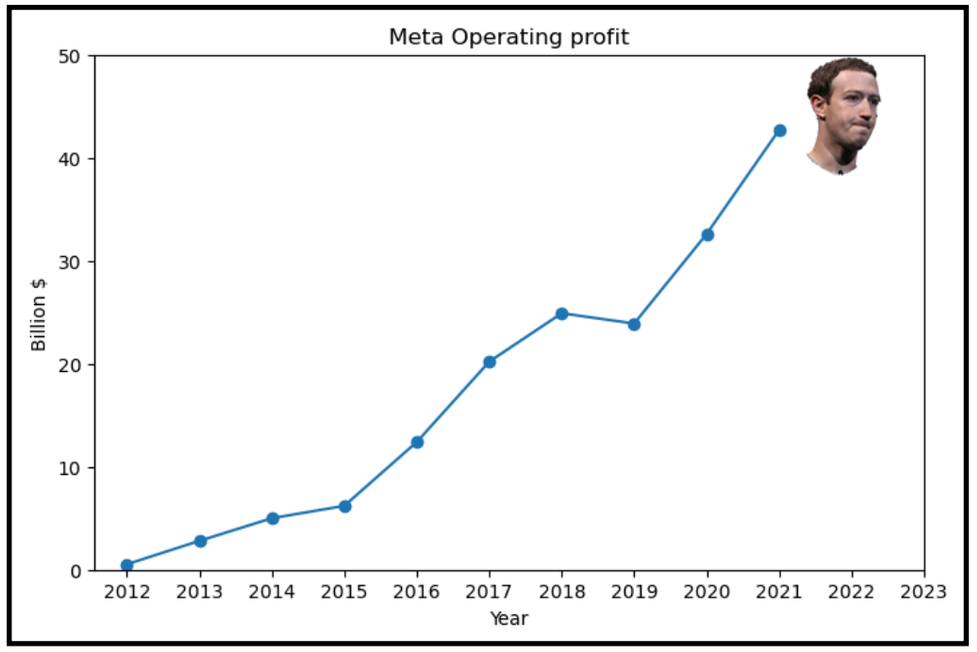
Signals from Big AI Companies

Layoffs



Signals from Big AI Companies

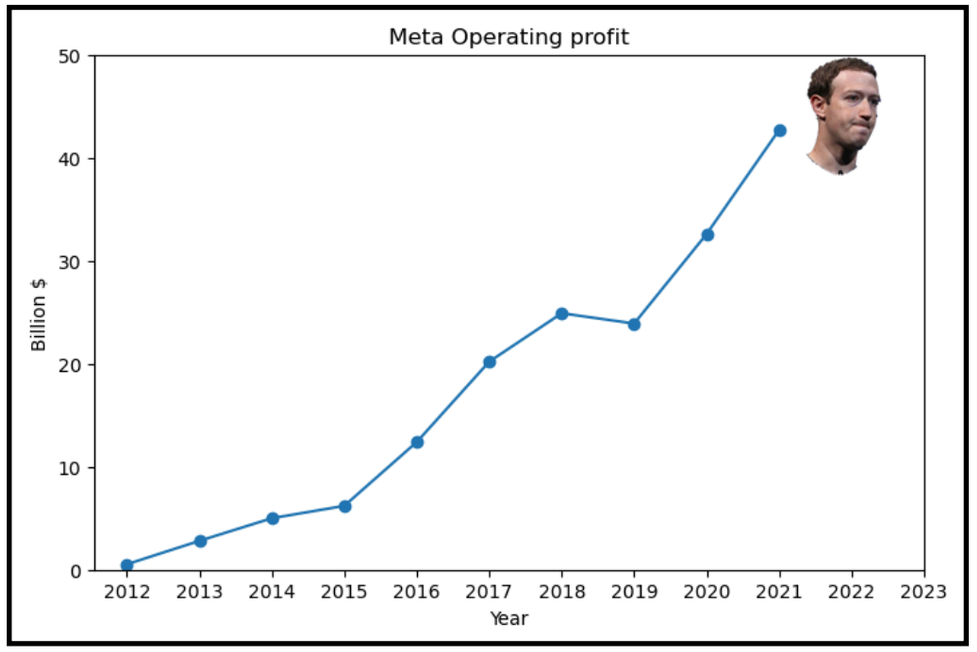
Layoffs



Physical world's problems

Signals from Big AI Companies

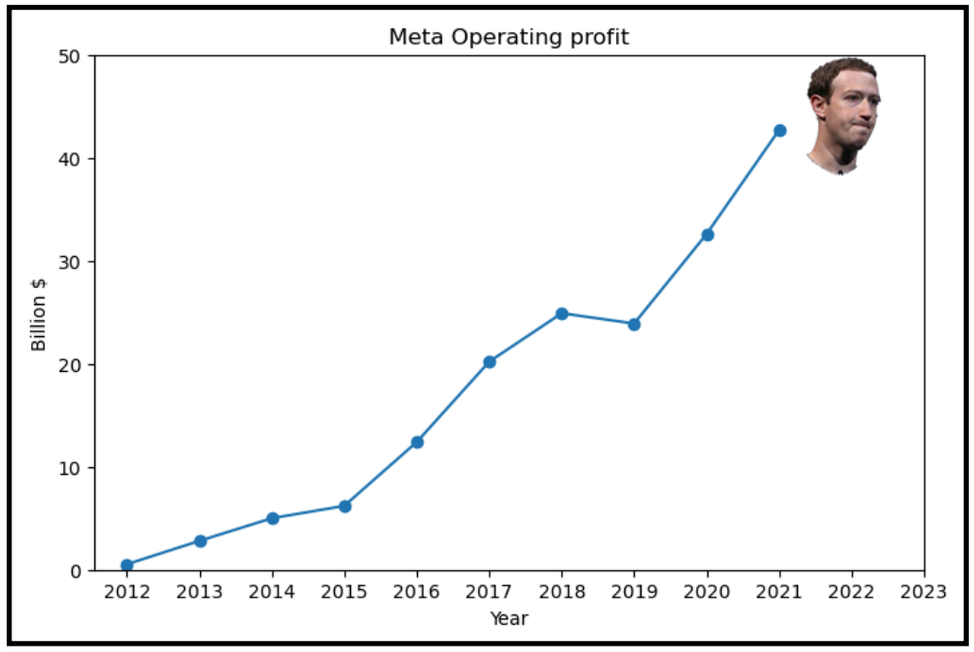
Layoffs



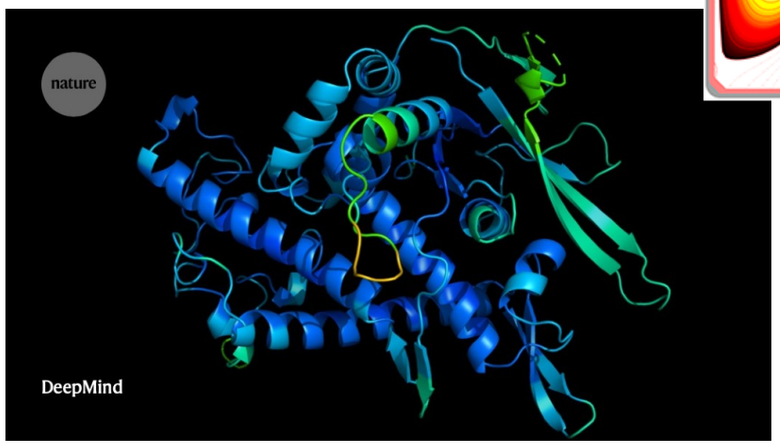
Physical world's problems

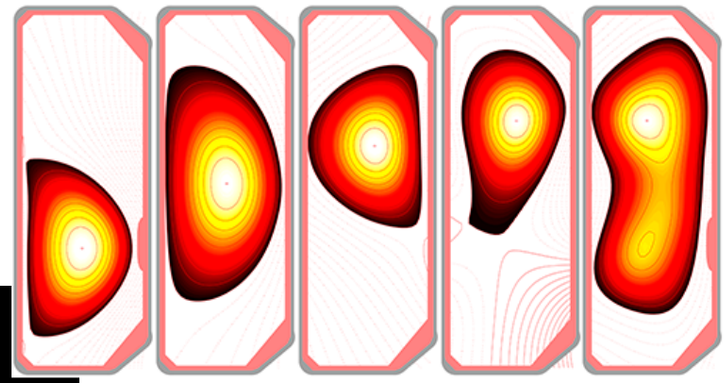
Signals from Big AI Companies

Layoffs



Physical world's problems





Prediction / Hope

Prediction / Hope

MLOps adoption and ease of use might allow different organisations (governments, research institutes, all type of companies) to deploy and use AI to **help solving or mitigating** some of the problems we're inevitably going to face in the next decades (food and energy scarcity, climate change damages)

Prediction / Hope

MLOps adoption and ease of use might allow different organisations (governments, research institutes, all type of companies) to deploy and use AI to **help solving or mitigating** some of the problems we're inevitably going to face in the next decades (food and energy scarcity, climate change damages)

And we might realise that maybe AI was never meant to be used for generating profits or controlling people using recommended ads and customer behaviour profiling **but to solve real problems in the physical world**

References

Books :

- "Designing Machine Learning Systems" - Chip Huyen, O'Reilly 2022
- "Introducing MLOps" - Mark Travel & the Dataiku team, O'Reilly 2020

Papers / Blogs / Reports / Videos :

- "2020 state of enterprise machine learning" - Algorithmia
- "Hidden Technical Debt in Machine Learning Systems" - Sculley et al. (Google Inc.), NIPS 2015
- "Practitioners guide to MLOps" - Google cloud white paper, 2021
- "MLOps: Continuous delivery and automation pipelines in machine learning" - Google Cloud Architecture Center, 2020
- "Why data scientists shouldn't need to know Kubernetes" - Chip Huyen's blog, 2021
- "Machine Learning Tools Landscape v2" - Chip Huyen's blog, 2020
- "Navigating the MLOps tooling landscape (Part I, II, III)" - Lj Miranda's blog, 2021
- "Software 2.0" - Andrej Karpathy Medium blog, 2017 (also on youtube)
- "Rules of Machine Learning: Best Practices for ML Engineering" - M. Zinkevich (Google developers guide), 2022
- "Machine Learning Operations (MLOps): Overview, Definition, and Architecture" - D. Kreuzberger et al., 2022