Machine Learning System Design

Andrei Paleyes

AI4ER, November 2023

About you!

Raise your hands if you ever:

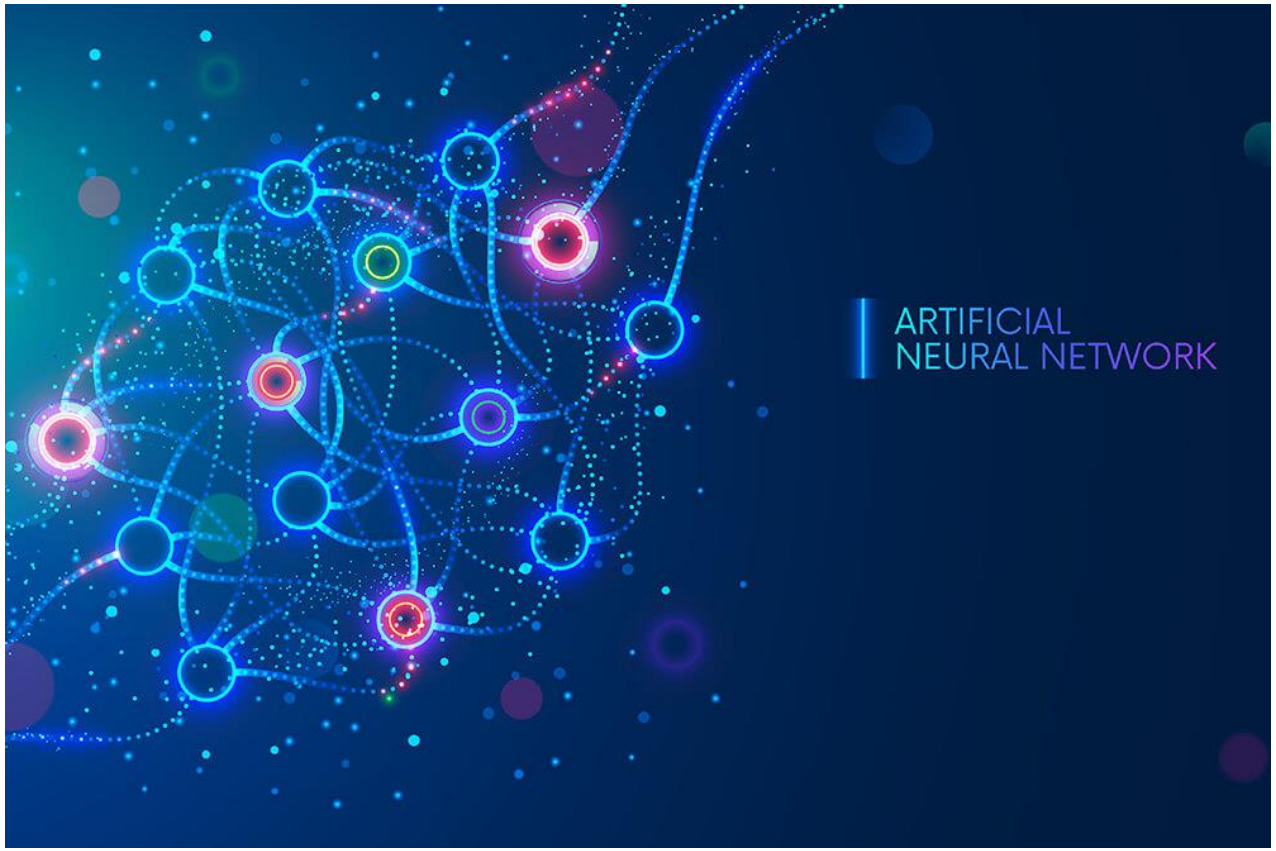
- trained a ML model?
- fixed an issue in a model?
- gave the model you worked on to anyone else?
- created an app/website/service around a model?
- updated an existing model with a new dataset?

?

What is machine learning?

What is machine learning?

Machine learning (**ML**) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.



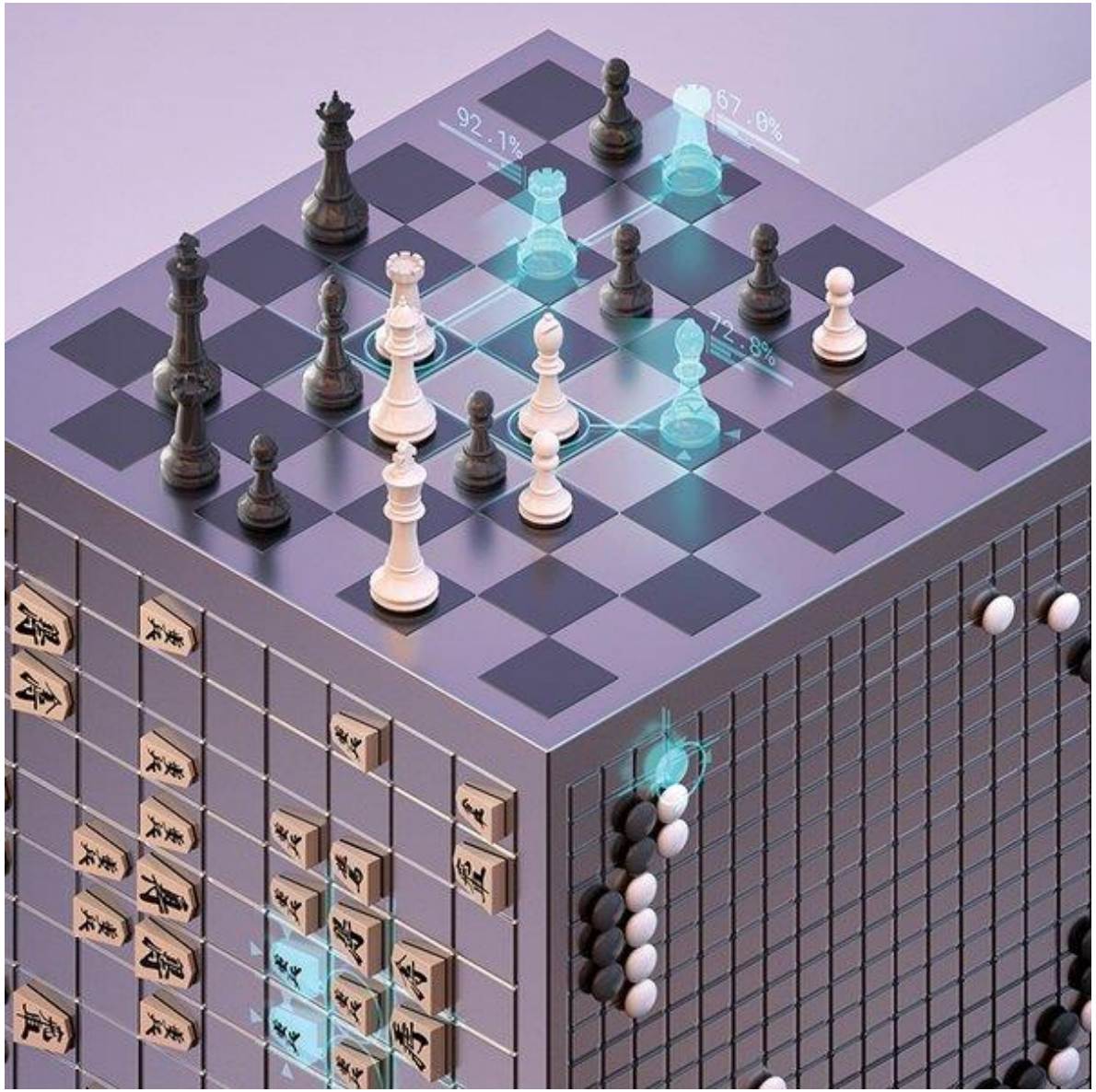
Sources: Wikipedia (text) and Bernard Marr (image)

?

What is machine learning?

What is machine learning?

Model + Data = Machine Learning


92.1%
67.8%
72.8%

What is machine learning?

Model + Data + System = Deployed Machine Learning

How to deploy ML systems

?

Step 1 – ask a question

First question you need to ask is …

“Do I really need ML?”

First question you need to ask is …

“Do I really need ML?”

because maybe you don’t

Example 1: Amazon vs Strawberries

amazon fresh

Example 1: Amazon vs Strawberries

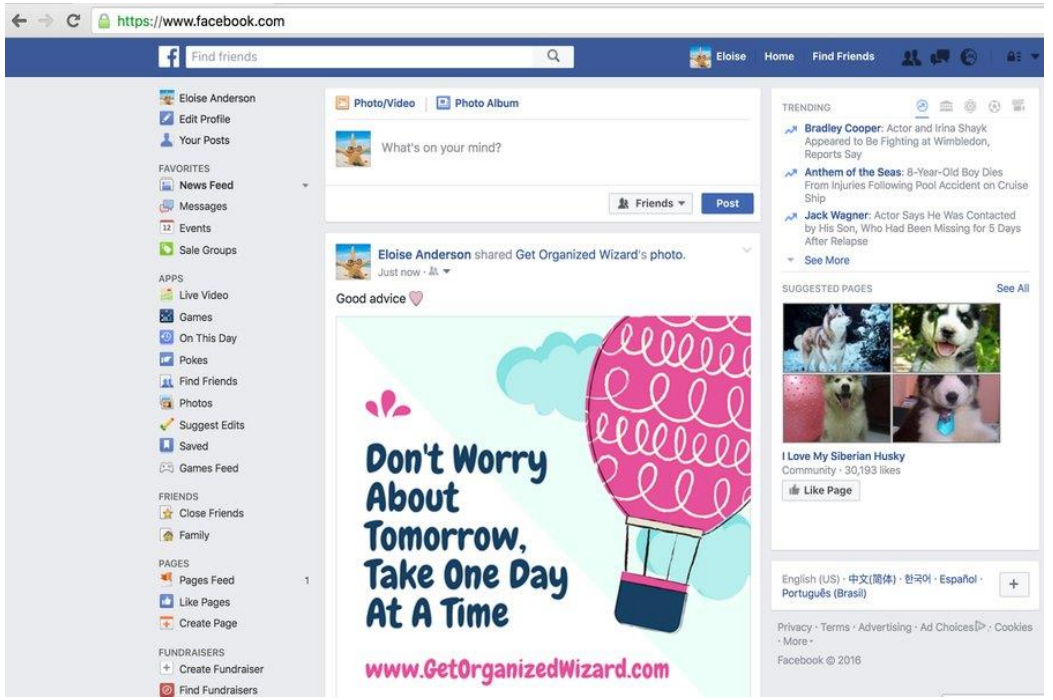
“If you think that machine learning will give you a 100% boost, then a heuristic will get you 50% of the way there.”

Martin Zinkevich, Google

?

Example 2: Facebook newsfeed

2006



2022

Take home point

The best way to design an ML system often is not to design one.

?

Step 2 – how to measure success

Ask this early!

- No, seriously
- Ask this question early!

Ask this early!

- No, seriously
- Ask this question early!

- Remember the goal
- Don’t overfocus on ML

“An interesting finding is that increasing the performance of a model does not necessarily translate into a gain in [business] value.”

150 successful machine learning models:

6 lessons learned at Booking.com

Bernadi et al., KDD’19

Other questions to consider

- Who are my end users?
- What are the biggest risks?
- Is there data? Where and in what form?

Check out ML Canvas, https://www.ownml.co/machine-learning-canvas

Step 3 – Find the data

?

Find the data

- Not a trivial problem!
- Unlike in exercises or academic research

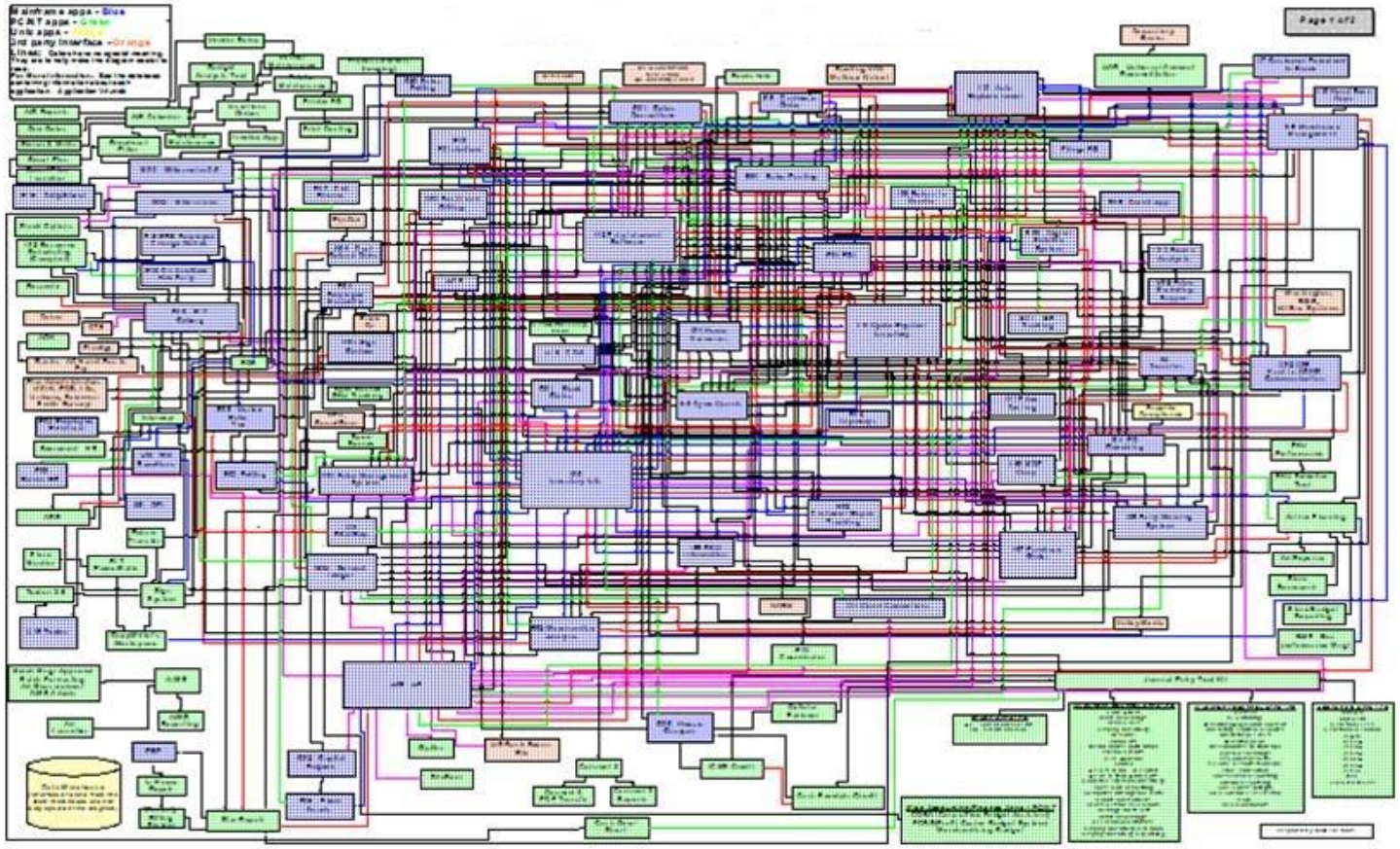
Why getting data is hard?

Data might be

- Split between sources
- Not saved
- Different between training and inference
- Unlabeled
- Not clean

Example 1: Twitter

Can you find all user data here?



Scaling Big Data Mining Infrastructure — The Twitter Experience. Jimmy Lin and Dmitriy Ryaboy, 2013

Example 2: Atlanta Fire Department, project Firebird

?

- 12 datasets
  - History of incidents
  - Business licenses
  - Households
  - etc.
- Join data on buildings by address
- Took weeks!

Example 2: Atlanta Fire Department, project Firebird

?

- Wolfson building, Madingley rise, Cambridge
- Bullard Laboratory, CB3 0EZ, Cambridge, UK

Other questions to consider

- What generates the data?
- How to access it?
- In which format is it?
- Is it complete?

Step 4 – Store the data

Storage options

- Memory
- Text file(s)
- CSV file(s)
- JSON file(s)
- SQL database
- NoSQL database
- Data stream

?

How do I choose?

Depends on your use case! Two main considerations:

- Data modality
- Purpose

Modalities

- Tabular data
- Time series
- Images
- Video
- Free text

Purpose

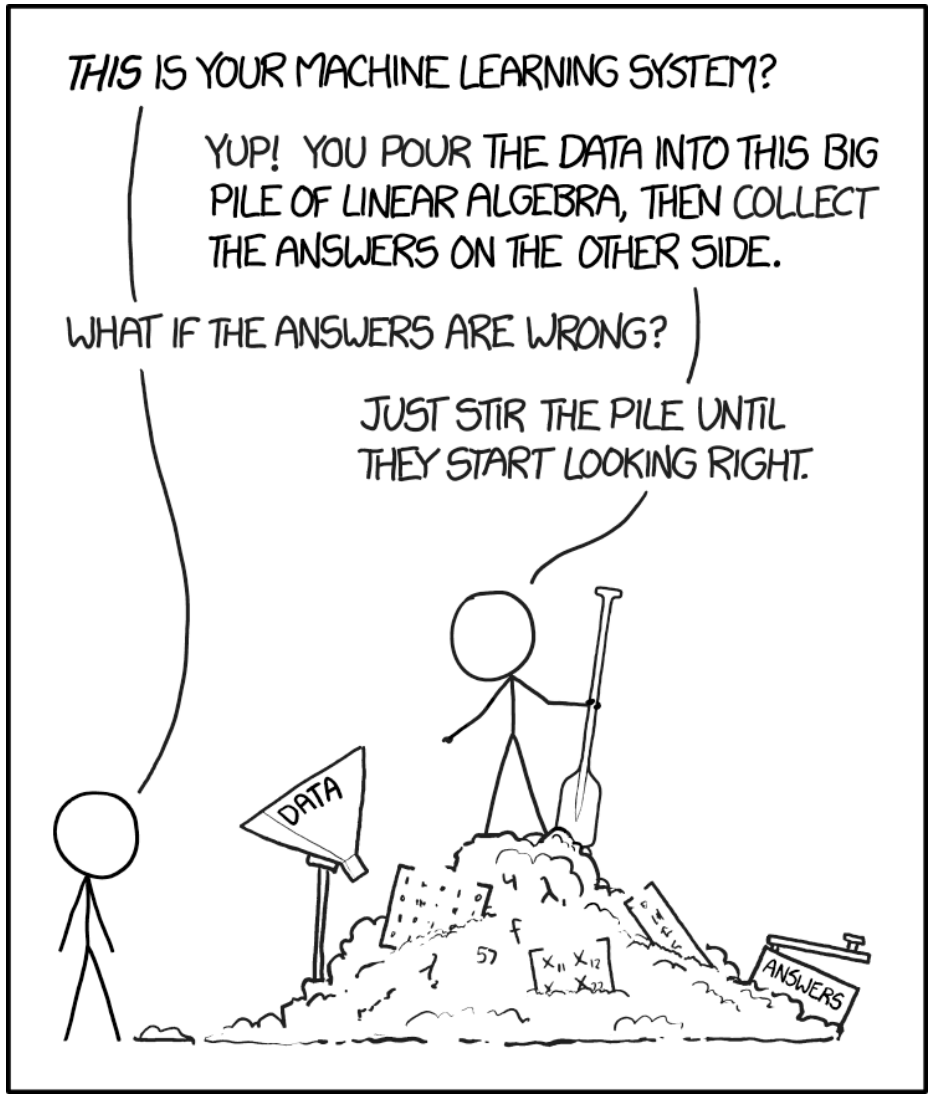
- To play with on your own
- For fellow data scientists
- For training
- For batch processing
- For online prediction

Other considerations

- Structure (e.g. nested)
- Size
- Performance

Step 5 – train a model

Model training

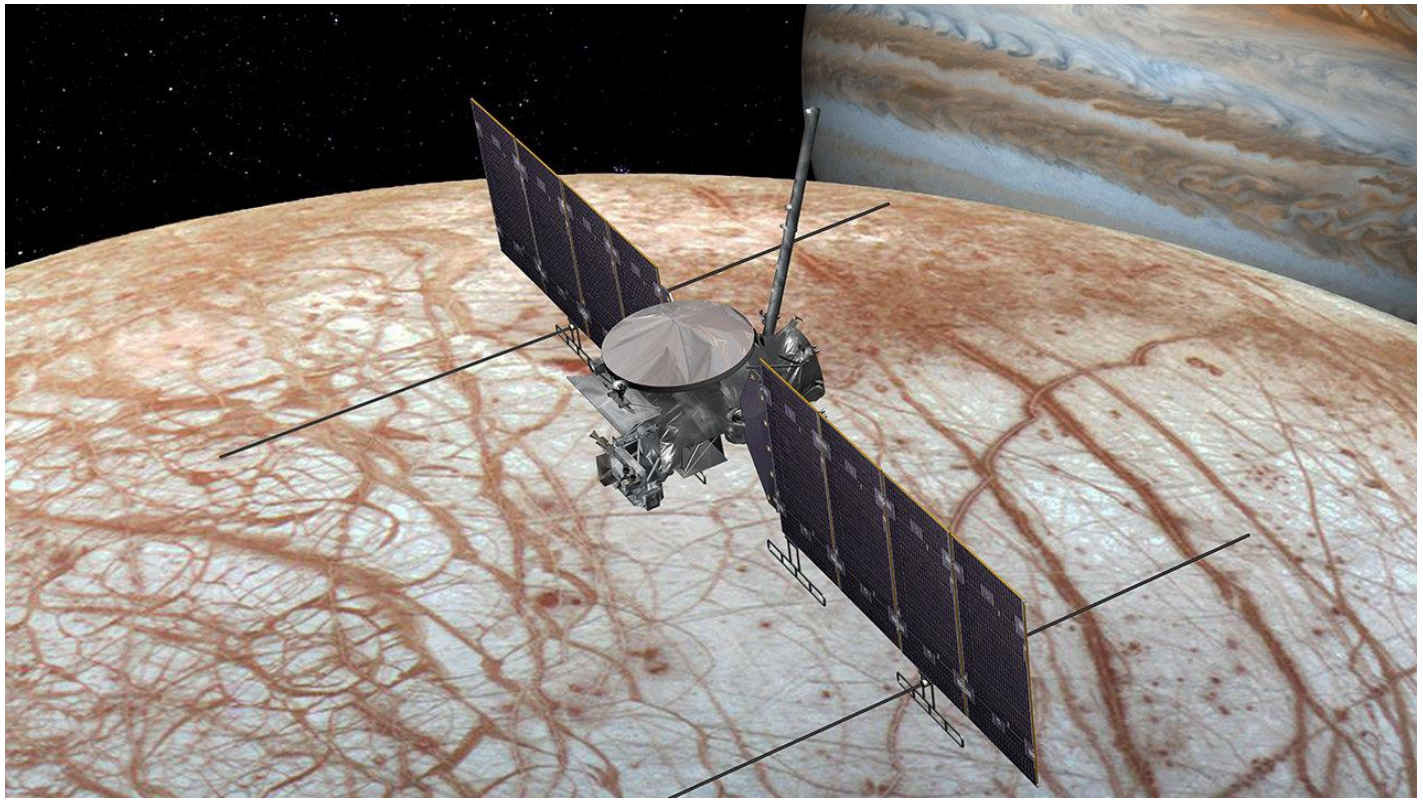


Model selection

Complex model is not necessarily the best!

Why? Let’s ask spacecrafts!

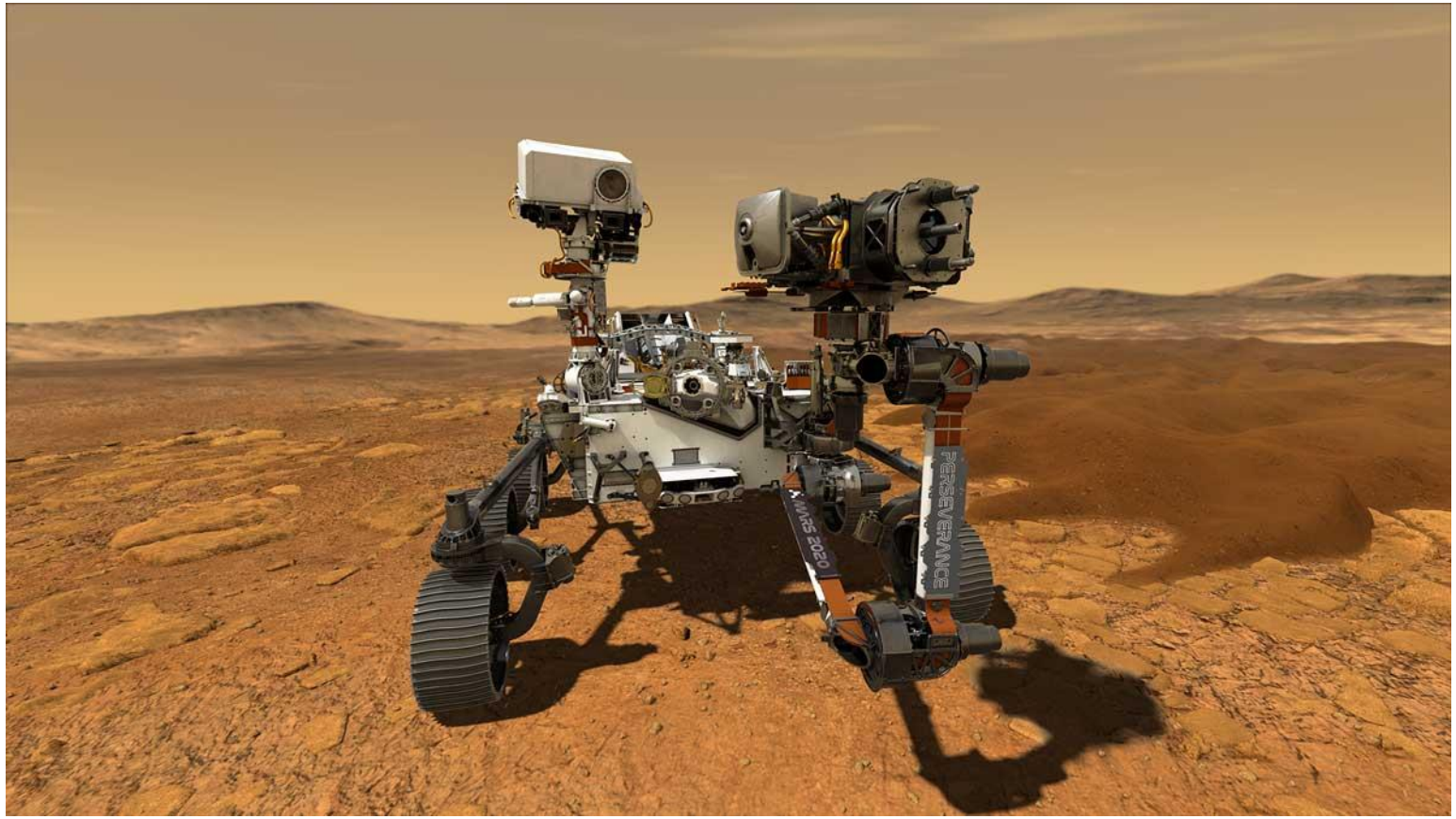
Example 1: Europa Clipper, 2024



PCA for anomaly detection!

Enabling onboard detection of events of scientific interest for the Europa Clipper spacecraft. Wagstaff et al, KDD 2019

Example 2: Perseverance, 2020



Random forest for landmark registration!

Precision instrument targeting via image registration for the Mars 2020 rover. Doran et al, IJCAI 2016

?

Why?

Why?

- Hardware constraints
- Reliability
- Interpretability
- Easy to use

Step 6 – Host the model

Where models can live

- Your machine
- Cloud server
- Mobile phone

Scenario 1

Learning about random forests with a Jupyter Notebook.

Where would the model live?

Scenario 1

Learning about random forests with a Jupyter Notebook.

Where would the model live?

On your own machine!

?

Scenario 2

Using simple linear regression in a mobile phone app

Where would the model live?

Scenario 2

Using simple linear regression in a mobile phone app

Where would the model live?

On the phone!

?

Scenario 3

Very deep neural net that uses a lot of memory and CPU to personalize website functions for users.

Where would the model live?

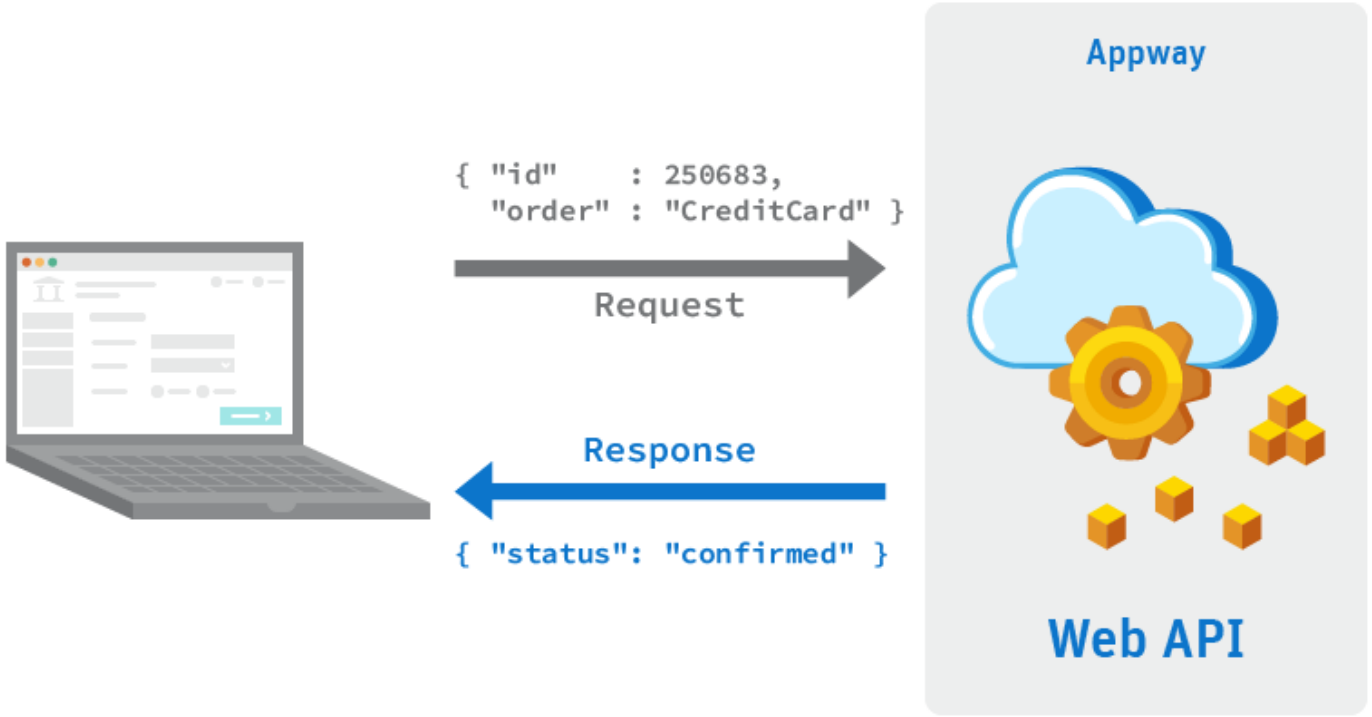
Scenario 3

Very deep neural net that uses a lot of memory and CPU to personalize website functions for users.

Where would the model live?

Dedicated host or cloud

A service!



?

Scenario 4

A model that updates book recommendations for users once a day.

Where would the model live?

Scenario 4

A model that updates book recommendations for users once a day.

Where would the model live?

Dedicated host or cloud

Step 7 - Monitoring and updating

?

A myth

“If we don’t do anything, model performance remains the same.”

Chip Huyen, Claypot AI

Data tends to drift

- Seasonality
- Change of habits
- Unforeseen factors
- Unexpected events

Can happen both in features and in labels!

Monitor for drifts

- Distribution of feature values
- Distribution of model predictions

?

Drift example

Drift example: COVID-19

- Online shopping patterns changed – Stitch Fix
- New terminology affected translation models – Facebook
- Mobility patterns changed – Google

Business metrics too!

Model improvement ≠ business improvement

Model that improves clicks ≠ better conversion

150 successful machine learning models: 6 lessons learned at Booking.com. Bernardi et al, KDD 2019

?

When to update?

When to update?

To get started

- When you feel the need (use metrics!)
- On a set schedule, e.g. once a month

Eventually this can become automated

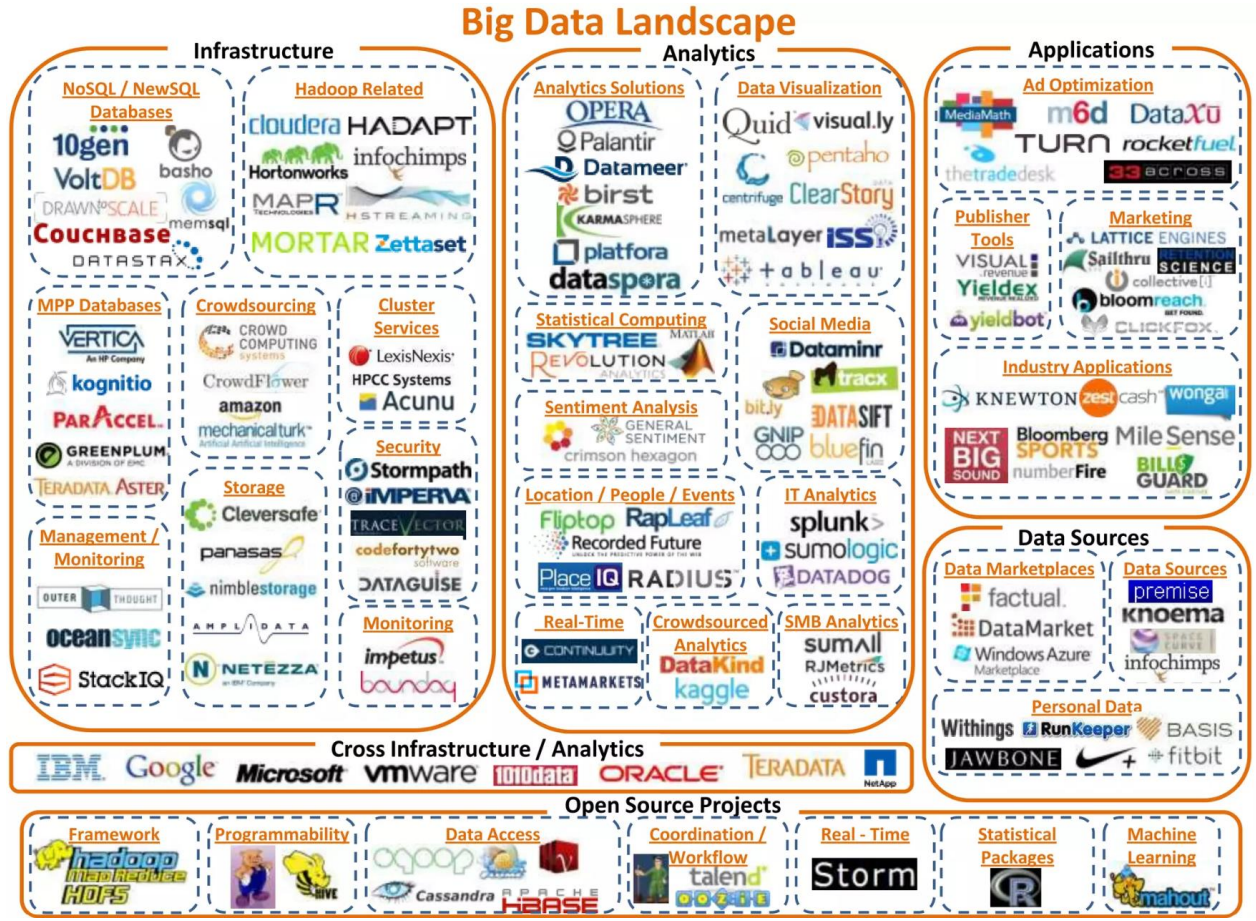
- AWS and Netflix deploy multiple times a day

Other things to consider

- Good software engineering practices
- Fairness, law and ethics
- Security
- Quality assurance
- User interface

A note on tools

- Their name is Legion
- Specific recommendations are impossible to give
- Focus on the goal and architecture
- Leverage available expertise


Big Data Landscape
Infrastructure
NoSQL / NewSQL Databases
10gen
VoltDB
basho
DRAWNtoSCALE
memsql
Couchbase
DATASTAX
Hadoop Related
cloudera
HADAPT
Hortonworks
infochimps
MAPR
HSTREAMING
MORTAR
Zettaset
MPP Databases
VERTICA
kognitio
PARACCEL
GREENPLUM
TERADATA ASTER
Crowdsourcing
CROWD COMPUTING systems
CrowdFlower
amazon mechanicalturk
Cluster Services
LexisNexis
HPCC Systems
Acunu
Security
Stormpath
IMPERVA
TRACEVECTOR
codefortytwo
DATAGUISE
Management / Monitoring
OUTER THOUGHT
oceansync
StackIQ
Storage
Cleversafe
panasas
nimblestorage
AMPLIDATA
NETEZZA
Monitoring
impetus
boundary
Analytics
Analytics Solutions
OPERA
Palantir
Datameer
birst
KARMASPHERE
platfora
dataspora
Data Visualization
Quid
visual.ly
pentaho
centrifuge
ClearStory
metaLayer
ISS
tableau
Statistical Computing
SKYTREE
MATLAB
REVOLUTION ANALYTICS
Social Media
Dataminr
tracx
bitly
DATASIFT
GNIP
bluefin
Sentiment Analysis
GENERAL SENTIMENT
crimson hexagon
Location / People / Events
Fliptop
RapLeaf
Recorded Future
PlaceIQ
RADIUS
IT Analytics
splunk>
sumologic
DATADOG
Real-Time
CONTINUITY
METAMARKETS
Crowdsourced Analytics
DataKind
kaggle
SMB Analytics
sumall
RJMetrics
custora
Applications
Ad Optimization
MediaMath
m6d
DataXu
TURN
rocketfuel
thetradedesk
33across
Publisher Tools
VISUAL revenue
Yieldex
yieldbot
Marketing
LATTICE ENGINES
Sailthru
RETENTION SCIENCE
collective[i]
bloomreach
CLICKFOX
Industry Applications
KNEWTON
zestcash
wonga
NEXT BIG SOUND
Bloomberg SPORTS
numberFire
MileSense
BILLSGUARD
Data Sources
Data Marketplaces
factual
DataMarket
Windows Azure Marketplace
Data Sources
premise
knoema
infochimps
Personal Data
Withings
RunKeeper
BASIS
JAWBONE
fitbit
Cross Infrastructure / Analytics
IBM
Google
Microsoft
vmware
1010data
ORACLE
TERADATA
NetApp
Open Source Projects
Framework
hadoop MapReduce HDFS
Programmability
HIVE
Data Access
sqoop
Cassandra
APACHE HBASE
Coordination / Workflow
talend
OOZIE
Real - Time
Storm
Statistical Packages
R
Machine Learning
mahout

Infrastructure

Storage

- aws Amazon S3
- Microsoft Azure Storage
- Google Cloud Storage
- IBM Storage
- PURESTORAGE
- wasabi
- ALLUXIO
- PANASAS
- Hewlett Packard Enterprise Nimble Storage
- COHESITY
- Qumulo
- VAST
- CLUMIO
- WEKA
- MINIO
- Backblaze
- CLOUDFLARE

MPP DBs

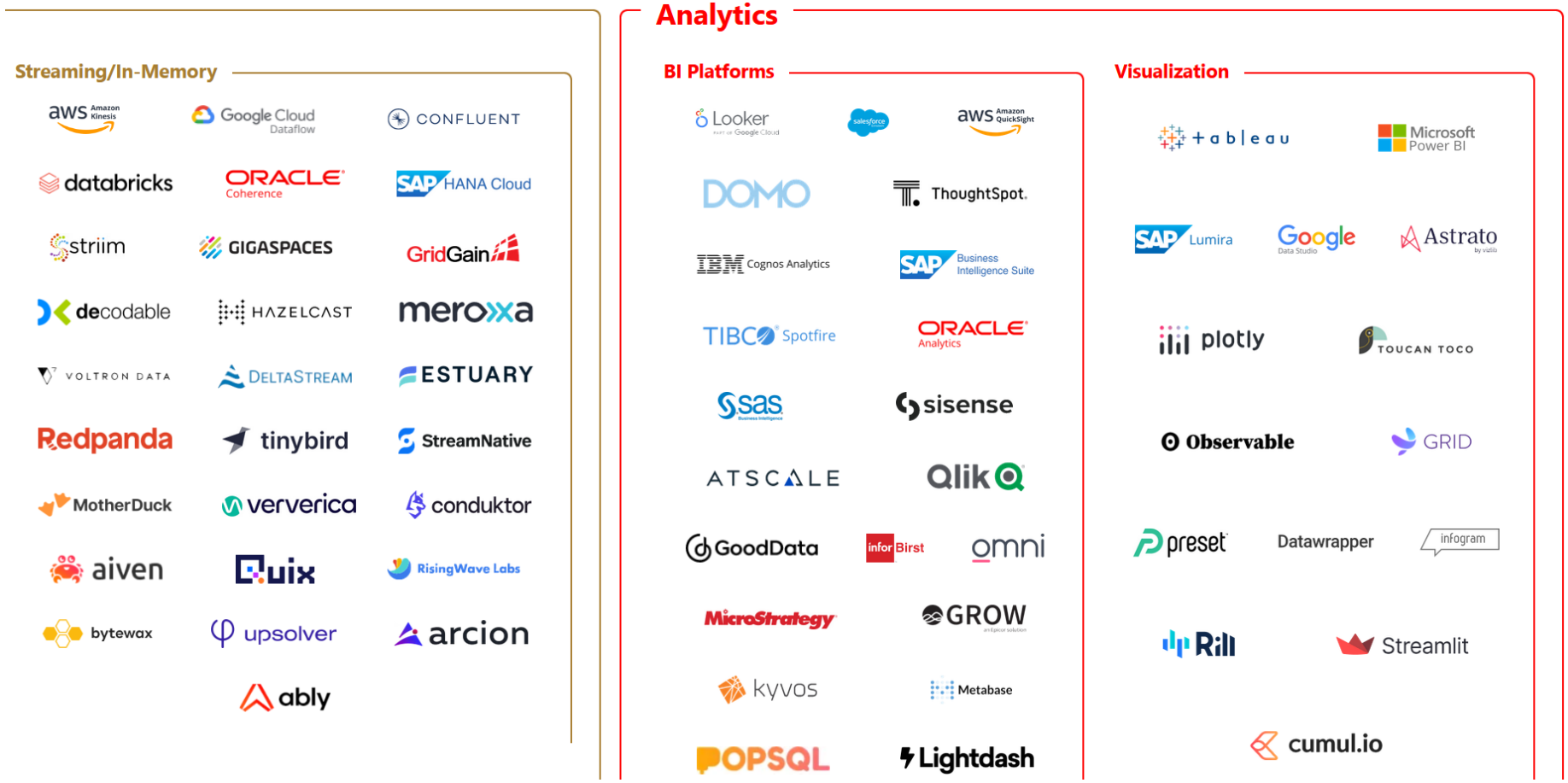
- teradata.
- VERTICA by opentext
- IBM Data Warehouse
- ACTIAN
- Exasol
- GREENPLUM DATABASE

Data Lakes/Lakehouses

- dremio
- databricks
- Microsoft Azure HDInsight
- Microsoft Azure Data Lake Storage
- Google Cloud Dataproc
- Google Cloud BigLake
- aws Amazon EMR
- aws Lake Formation
- IBM Data Lake Solutions
- Hewlett Packard Enterprise Ezmeral Data Fabric
- CLOUDERA
- Qubole
- ONEHOUSE

Data Warehouses

- aws Amazon RedShift
- snowflake
- Google Cloud BigQuery
- Microsoft Azure Synapse Analytics
- ORACLE Exadata Cloud Service
- FIREBOLT
- vmware Greenplum
- IBM Db2 Warehouse
- KYLIGENCE
- Yellowbrick
- Tabular
- StarRocks
- (OCIENT)™
- HYDRA


Streaming/In-Memory
aws Amazon Kinesis
Google Cloud Dataflow
CONFLUENT
databricks
ORACLE Coherence
SAP HANA Cloud
striim
GIGASPACES
GridGain
decodable
HAZELCAST
meroxa
VOLTRON DATA
DELTASTREAM
ESTUARY
Redpanda
tinybird
StreamNative
MotherDuck
ververica
conduktor
aiven
Quix
RisingWave Labs
bytewax
upsolver
arcion
ably
Analytics
BI Platforms
Looker
salesforce
aws Amazon QuickSight
DOMO
ThoughtSpot
IBM Cognos Analytics
SAP Business Intelligence Suite
TIBCO Spotfire
ORACLE Analytics
sas
sisense
ATSCALE
Qlik
GoodData
infor Birst
omni
MicroStrategy
GROW
kyvos
Metabase
POPSQL
Lightdash
Visualization
+ableau
Microsoft Power BI
SAP Lumira
Google Data Studio
Astrato
plotly
TOUCAN TOCO
Observable
GRID
preset
Datawrapper
infogram
Rill
Streamlit
cumul.io

Machine Learning & Artificial Intelligence

Data Science Notebooks

jupyter
HEX
CO
SaturnCloud
aws Amazon SageMaker
Microsoft Azure Notebooks
binder
Deepnote
Noteable
count
COCALC
Curvenote
Nextjournal

Data Science Platforms

IBM Watson Studio
ORACLE OCI Data Science
RAPIDMINER
Palantir
sas
TIBCO
ANACONDA.
ALTAIR
Open for Innovation KNIME
MathWorks

Enterprise ML Platforms

databricks
aws Amazon SageMaker
data iku
Microsoft Azure Machine Learning
Google Cloud Vertex AI
H2O.ai
DataRobot
DOMINO
ABACUS.AI
Wallaroo
Snorkel
Continual
dotData
Modular
AIBLE
obviously.ai
causaLens
graft

Data Generation & Labelling

scale
HIVE
upwork
appen
cloudfactory
super.ai
SuperAnnotate
Dataloop
datagen
ALEGION
TONIC THE FAKE DATA COMPANY
gretel

Data Generation & Labelling

scale · HIVE · amazon mechanical turk · upwork · appen · Unity · sama · cloudfactory · super.ai · Superb AI · SuperAnnotate · Dataloop · Kili · V7 · datagen · ALEGION · MOSTLY AI · TONIC · gretel · M4AI

Applications - Enterprise

Sales

salesforce · GONG · Clari · conversica · aviso · people.ai · Clearbit · tact.ai · HeadsUp · regie.ai · LAVENDER · Smartwriter.ai · Databook · Outreach · Salesloft · involve.ai

Marketing

Bluecore · 6sense · atte · ActiveCampaign · bloomreach · Demandbase · PERSADO · INVO · braze · NextRoll · copy.ai · Writesonic · anyword · phrasee

Automation & Operations

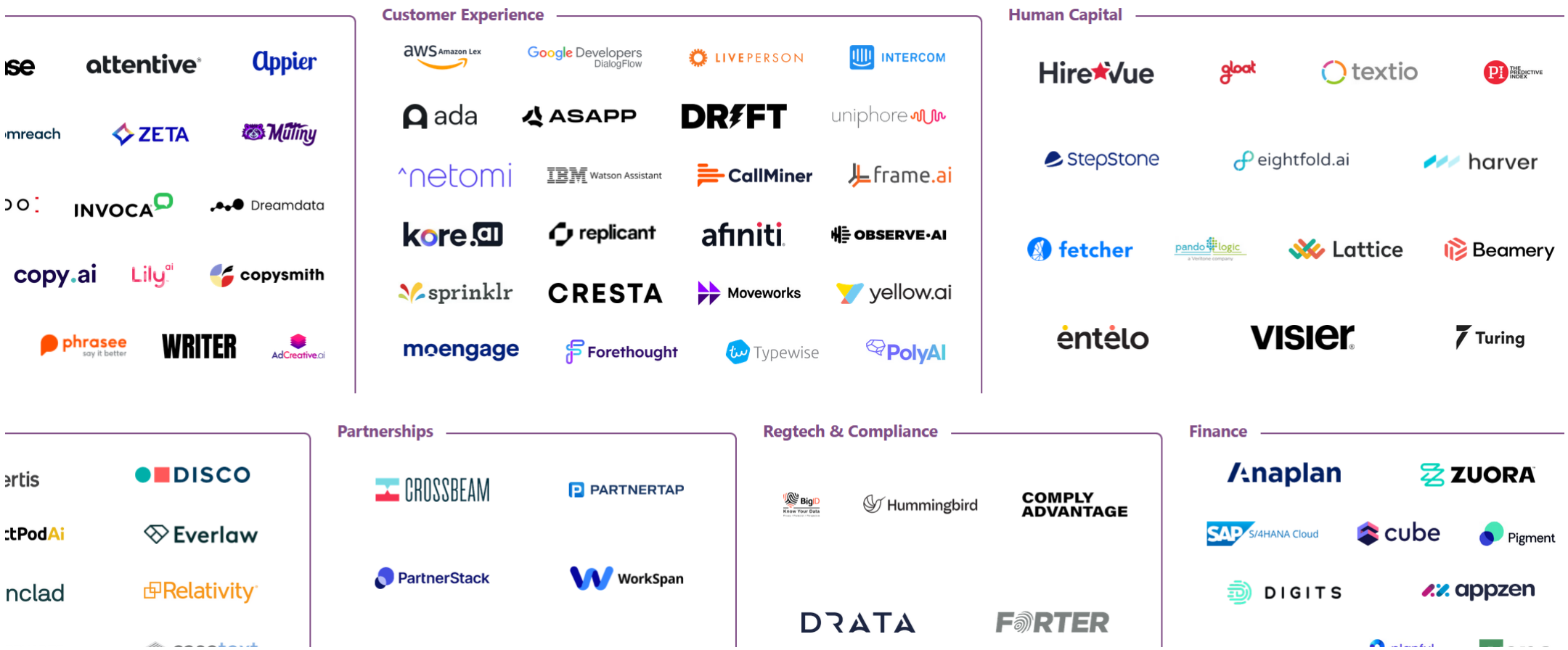
UiPath · celonis · HYPERSCIENCE · ABBYY · EigenTechnologies · tonkean · ANTWORKS · airSlate · INSTABASE · JIFFY.ai · workato · zapier · ALKYMI

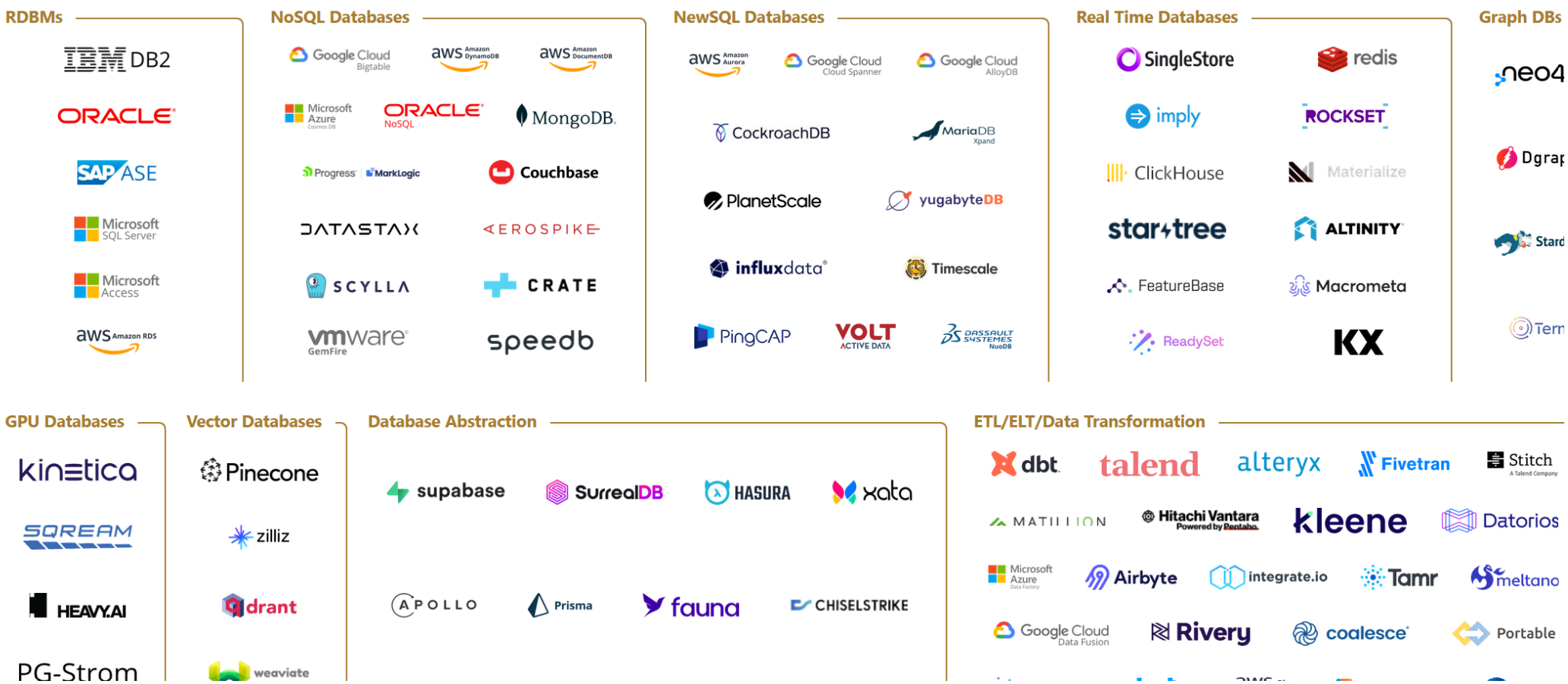
Decision & Optimization

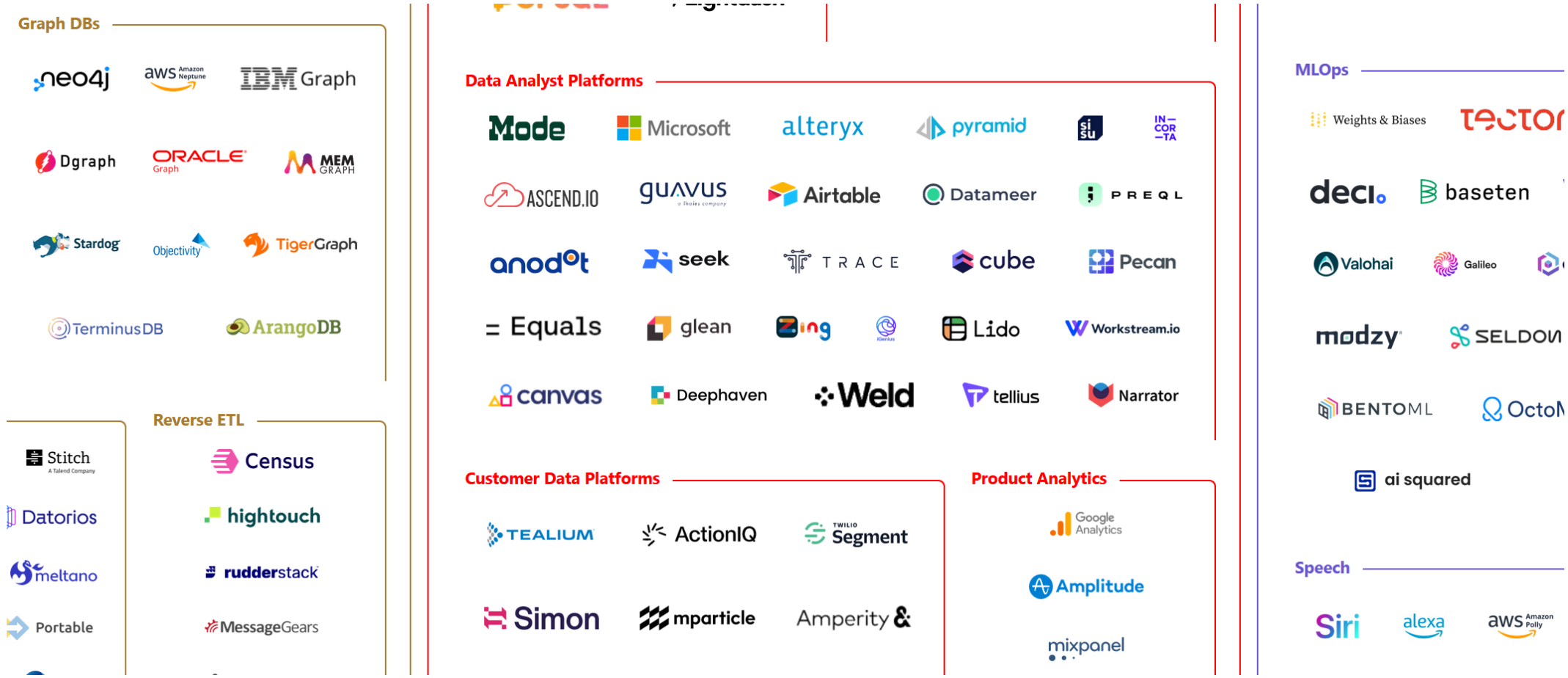
PEAK · nextmv · arena · Buynomics

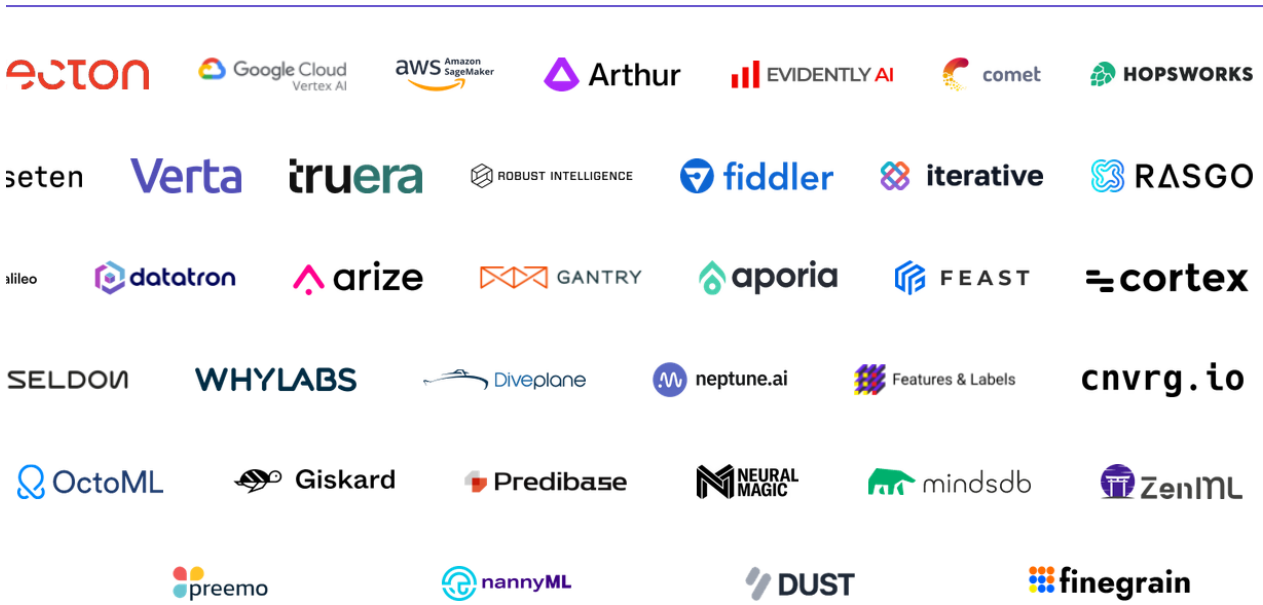
Legal

Icertis · ContractPodAi · Ironclad · LINKSQUARES

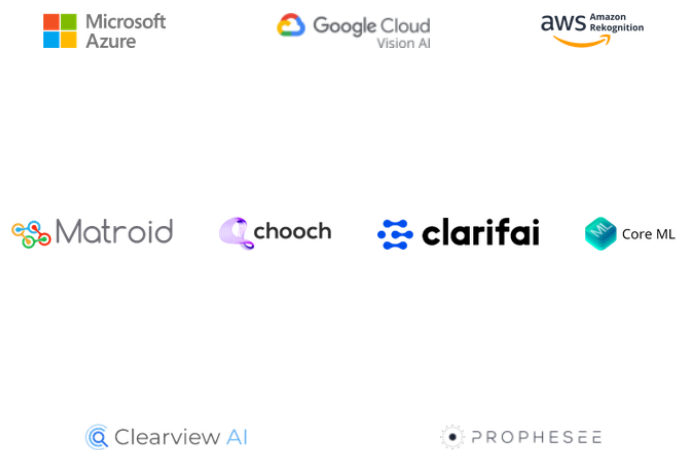

Customer Experience
attentive
Qppier
ZETA
Dreamdata
INVOCA
copy.ai
copysmith
phrasee
WRITER
AdCreative.ai
aws Amazon Lex
Google Developers DialogFlow
LIVEPERSON
INTERCOM
ada
ASAPP
DRIFT
uniphore
netomi
IBM Watson Assistant
CallMiner
frame.ai
kore.ai
replicant
afiniti
OBSERVE·AI
sprinklr
CRESTA
Moveworks
yellow.ai
moengage
Forethought
Typewise
PolyAI
Human Capital
HireVue
gloat
textio
StepStone
eightfold.ai
harver
fetcher
Lattice
Beamery
entelo
VISIER
Turing
Partnerships
DISCO
Everlaw
Relativity
CROSSBEAM
PARTNERTAP
PartnerStack
WorkSpan
Regtech & Compliance
Hummingbird
COMPLY ADVANTAGE
DRATA
FORTER
Finance
Anaplan
ZUORA
SAP S/4HANA Cloud
cube
Pigment
DIGITS
appzen


RDBMs
IBM DB2
ORACLE
SAP ASE
Microsoft SQL Server
Microsoft Access
aws Amazon RDS
NoSQL Databases
Google Cloud Bigtable
aws Amazon DynamoDB
aws Amazon DocumentDB
Microsoft Azure Cosmos DB
ORACLE NoSQL
MongoDB.
Progress | MarkLogic
Couchbase
DATASTAX
AEROSPIKE
SCYLLA
CRATE
vmware GemFire
speedb
NewSQL Databases
aws Amazon Aurora
Google Cloud Cloud Spanner
Google Cloud AlloyDB
CockroachDB
MariaDB Xpand
PlanetScale
yugabyteDB
influxdata
Timescale
PingCAP
VOLT ACTIVE DATA
DASSAULT SYSTEMES NuoDB
Real Time Databases
SingleStore
redis
imply
ROCKSET
ClickHouse
Materialize
startree
ALTINITY
FeatureBase
Macrometa
ReadySet
KX
Graph DBs
neo4
Dgraf
Stard
Tern
GPU Databases
kinetica
SQREAM
HEAVY.AI
PG-Strom
Vector Databases
Pinecone
zilliz
drant
weaviate
Database Abstraction
supabase
SurrealDB
HASURA
xata
APOLLO
Prisma
fauna
CHISELSTRIKE
ETL/ELT/Data Transformation
dbt
talend
alteryx
Fivetran
Stitch A Talend Company
MATILLION
Hitachi Vantara Powered by Pentaho
kleene
Datorios
Microsoft Azure Data Factory
Airbyte
integrate.io
Tamr
meltano
Google Cloud Data Fusion
Rivery
coalesce
Portable


Graph DBs
neo4j
aws Amazon Neptune
IBM Graph
Dgraph
ORACLE Graph
MEM GRAPH
Stardog
Objectivity
TigerGraph
TerminusDB
ArangoDB
Reverse ETL
Stitch
A Talend Company
Census
Datorios
hightouch
meltano
rudderstack
Portable
MessageGears
Data Analyst Platforms
Mode
Microsoft
alteryx
pyramid
sisu
INCORTA
ASCEND.IO
guavus
Airtable
Datameer
PREQL
anodot
seek
TRACE
cube
Pecan
Equals
glean
Zing
Lido
Workstream.io
canvas
Deephaven
Weld
tellius
Narrator
Customer Data Platforms
TEALIUM
ActionIQ
TWILIO Segment
Simon
mparticle
Amperity
Product Analytics
Google Analytics
Amplitude
mixpanel
MLOps
Weights & Biases
tecton
deci
baseten
Valohai
Galileo
modzy
SELDON
BENTOML
ai squared
Speech
Siri
alexa
aws Amazon Polly



Computer Vision



NLP

airSlate INSTABASE JIFFY.ai
workato zapier ALKYMI
Ocrolus ROSSUM bardeen
LEVITY

arena
Buynomics
Aera paretos

Ironclad Relativity
LINKSQUARES casetext
DoNotPay reveal brainspace TERM SCOUT
Spellbook by Rally Luminance

PartnerStack Wo
Habu INFO
Vendia REVE

Applications - Horizontal

Code & Documentation

replit GitHub CoPilot tabnine
Stenography Mintlify Debuild
DeepMind AlphaCode durable anima AI 2sql
Moderne warp mutable.ai
diffblue AI for Code MACHINET salesforce BLACKBOX
Magic aws Amazon CodeWhisperer

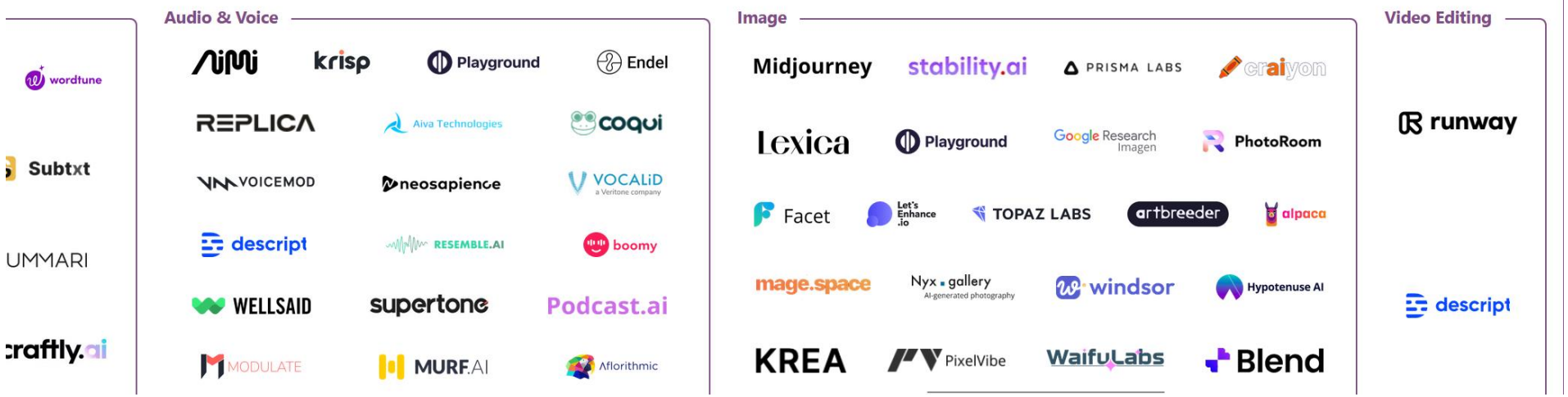
Text

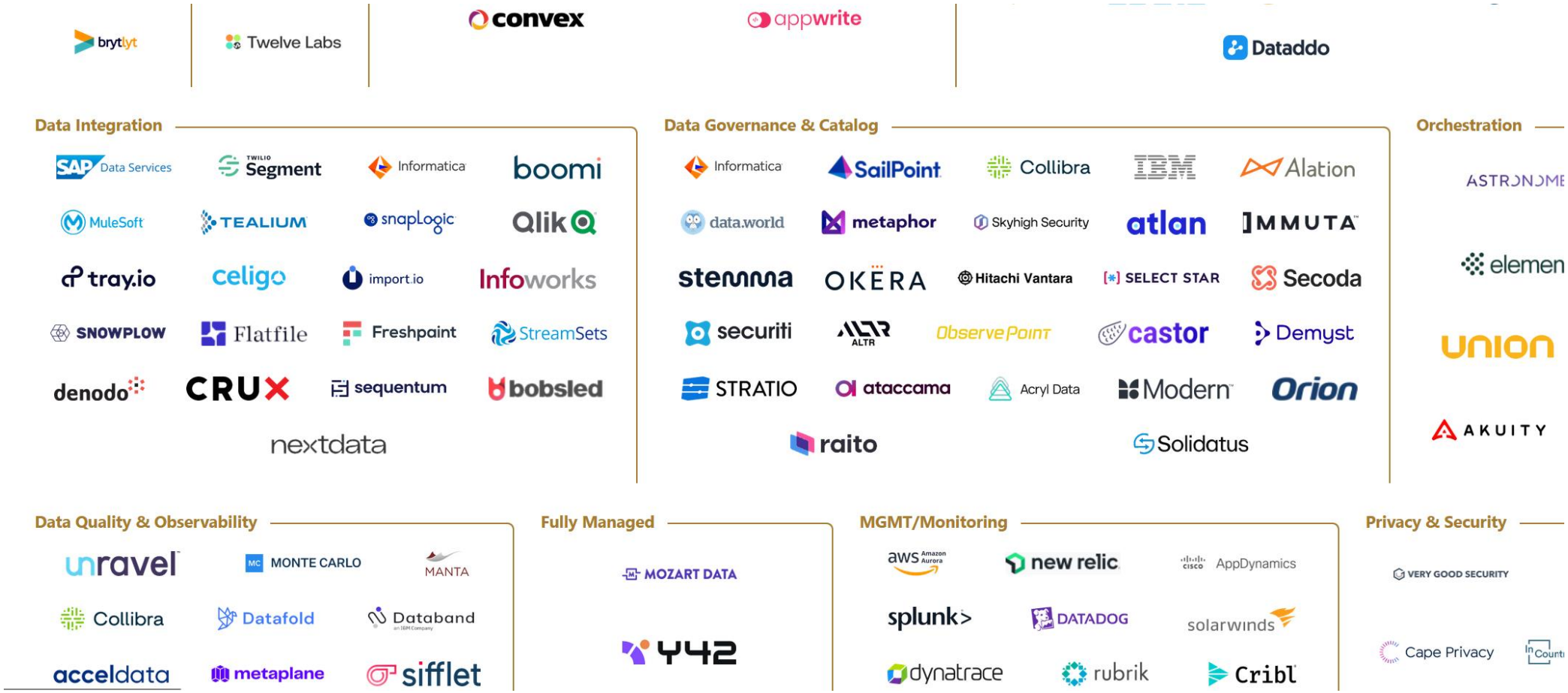
Jasper cogram sudo write wordtune
COMPOSE AI Moonbeam Subtxt
NovelAI LAIKA SUMMARI
HyperWrite fable ONE AI craftly.ai

Audio & Voice

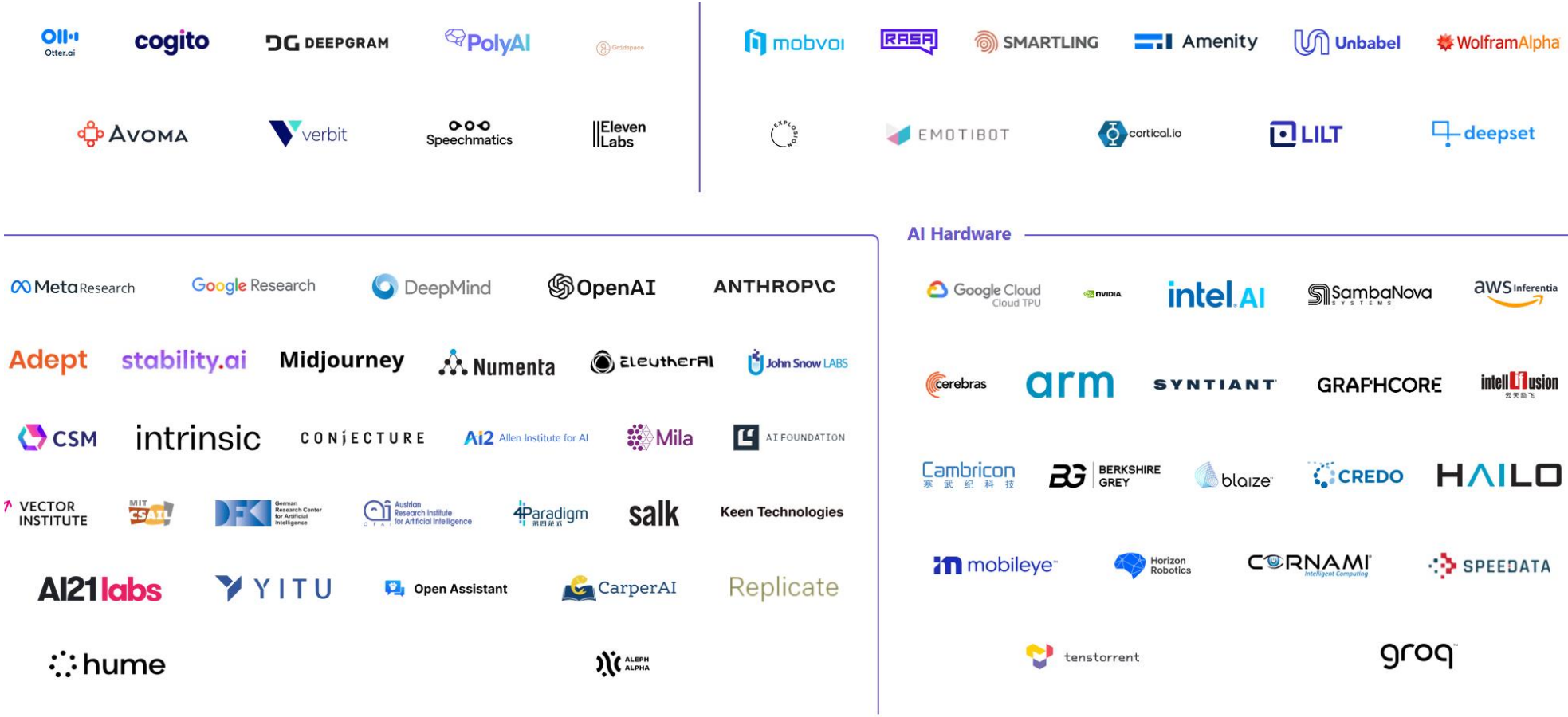
AiMi krisp Playground
REPLICA Aiva Technologies
VOICEMOD neosapience
descript RESEMBLE.AI
WELLSAID supertone
MODULATE MURF.AI


Relativity
casetext
brainspace
TERM SCOUT
Luminance
PartnerStack
WorkSpan
Habu
INFOSUM
Vendia
REVEAL
DRATA
FORTER
MinervaAI
COMPLIANCE.AI
DIGITS
appzen
ORACLE NETSUITE
planful
Vena
pilot
botkeeper
datarails
highradius


wordtune
Subtxt
UMMARI
craftly.ai
Audio & Voice
krisp
Playground
Endel
REPLICA
Aiva Technologies
coqui
VOICEMOD
neosapience
VOCALiD
a Veritone company
descript
RESEMBLE.AI
boomy
WELLSAID
supertone
Podcast.ai
MODULATE
MURF.AI
Aflorithmic
Image
Midjourney
stability.ai
PRISMA LABS
craiyon
Lexica
Playground
Google Research Imagen
PhotoRoom
Facet
Let's Enhance .io
TOPAZ LABS
artbreeder
alpaca
mage.space
Nyx gallery
AI-generated photography
windsor
Hypotenuse AI
KREA
PixelVibe
WaifuLabs
Blend
Video Editing
runway
descript


brytlyt
Twelve Labs
convex
appwrite
Dataddo
Data Integration
SAP Data Services
Twilio Segment
Informatica
boomi
MuleSoft
TEALIUM
snapLogic
Qlik
tray.io
celigo
import.io
Infoworks
SNOWPLOW
Flatfile
Freshpaint
StreamSets
denodo
CRUX
sequentum
bobsled
nextdata
Data Governance & Catalog
Informatica
SailPoint
Collibra
IBM
Alation
data.world
metaphor
Skyhigh Security
atlan
IMMUTA
stemma
OKERA
Hitachi Vantara
SELECT STAR
Secoda
securiti
ALTR
ObservePoint
castor
Demyst
STRATIO
ataccama
Acryl Data
Modern
Orion
raito
Solidatus
Orchestration
ASTRONOME
elemen
UNION
AKUITY
Data Quality & Observability
unravel
MONTE CARLO
MANTA
Collibra
Datafold
Databand
acceldata
metaplane
sifflet
Fully Managed
MOZART DATA
Y42
MGMT/Monitoring
aws Amazon Aurora
new relic
AppDynamics
splunk>
DATADOG
solarwinds
dynatrace
rubrik
Cribl
Privacy & Security
VERY GOOD SECURITY
Cape Privacy


POLYTOMIC
Octolis
PREFECT
Outerbounds
aws Step Functions
Microsoft Azure
MAGE
Google Cloud Workflows
privacera
Cyral
Transcend
skyflow
lytics
TREASURE DATA
Flywheel Software
Insider
blueshift
optimove
Adobe Real-Time CDP
rudderstack
blueconic
SYNCARI
Log Analytics
splunk>
Microsoft Azure
Google Cloud Cloud Logging
sumo logic
solarwinds loggly
kibana
logz.io
mezmo
Coralogix
LogRhythm
OBSERVE
DATADOG
exabeam
AppDynamics
LogicMonitor
sematext
pendo
Contentsquare
Heap
data.ai
fullstory
GLASSBOX
whatfix
Crypto & Web 3 Analytics
Chainalysis
Eleven Labs
TRM
Dune
merlin
nansen
nxyz
Allium
Luabase
SoundHound
Otter.ai
cogito
AssemblyAI
AVOMA
Horizontal AI/AGI
Microsoft Research
Meta Research
IBM Watson
Adept
stability.
Inflection
CSM
MIRI
VECTOR INSTITUTE
AI21labs
GPU Cloud


Otter.ai
cogito
DEEPGRAM
PolyAI
Girdspace
mobvoi
RASA
SMARTLING
Amenity
Unbabel
WolframAlpha
AVOMA
verbit
Speechmatics
Eleven Labs
EMOTIBOT
cortical.io
LILT
deepset
Meta Research
Google Research
DeepMind
OpenAI
ANTHROP\C
Adept
stability.ai
Midjourney
Numenta
EleutherAI
John Snow LABS
CSM
intrinsic
CONJECTURE
Ai2 Allen Institute for AI
Mila
AI FOUNDATION
VECTOR INSTITUTE
MIT CSAIL
German Research Center for Artificial Intelligence
Austrian Research Institute for Artificial Intelligence
4Paradigm
salk
Keen Technologies
AI21 labs
YITU
Open Assistant
CarperAI
Replicate
hume
ALEPH ALPHA
AI Hardware
Google Cloud Cloud TPU
NVIDIA
intel.AI
SambaNova SYSTEMS
aws Inferentia
cerebras
arm
SYNTIANT
GRAPHCORE
intellifusion
Cambricon
BERKSHIRE GREY
blaize
CREDO
HAILO
mobileye
Horizon Robotics
CORNAMI
SPEEDATA
tenstorrent
groq

Video Generation

EmbodyMe · synthesia · reface · DEEPBRAIN AI · WOMBO · soul machines · D-ID · Rephrase.ai · Fliki · tavus · Hour One · vidyo.ai · Movio · Colossyan · YEPIC AI · METAPHYSIC · typecast · PICTORY

Animation & 3D

avataar · CSM · inworld · DATAGRID · LUMA AI · mirage · KAEDIM

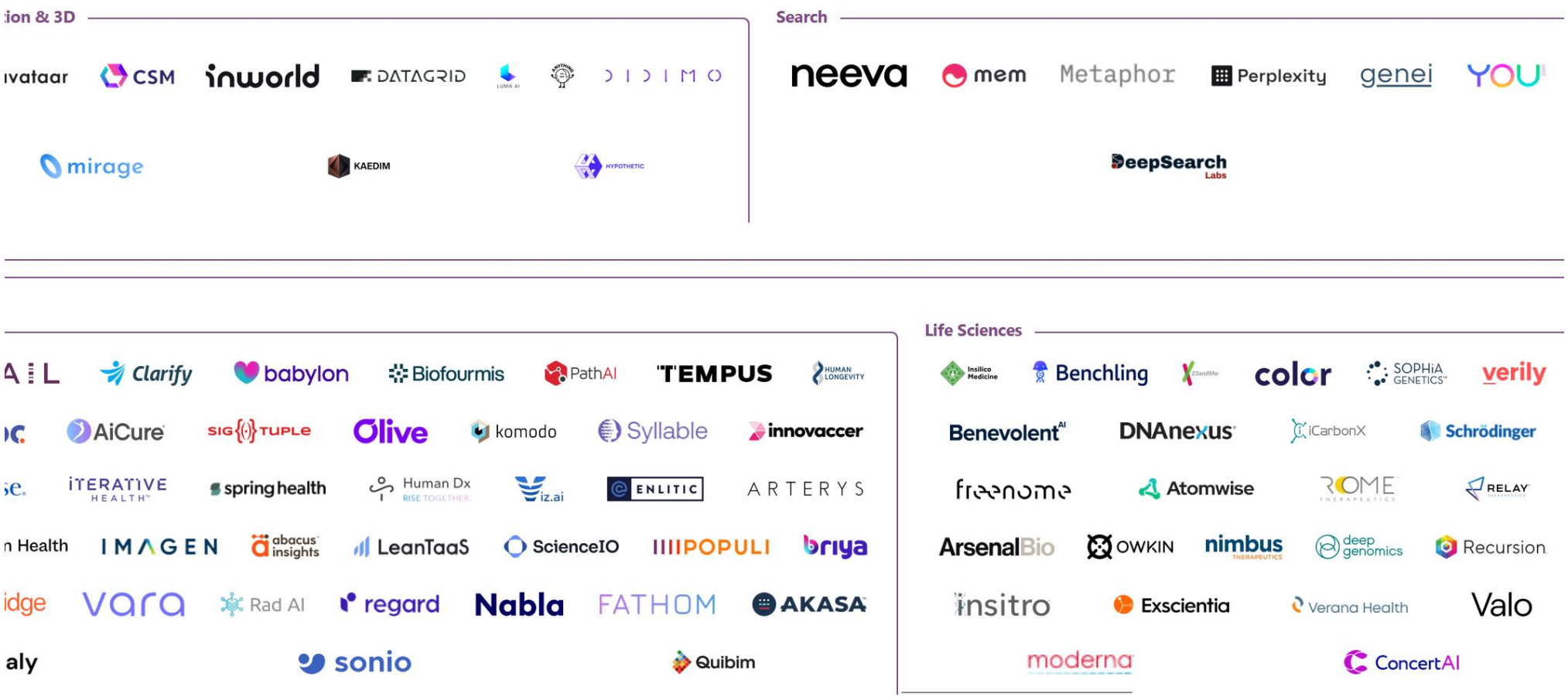
Applications - Industry

Finance & Insurance

KENSHO · Upstart · Lemonade · GUIDEWIRE · ARTURO · CAPE ANALYTICS · SHIFT · PLANCK · Tractable · EvolutionIQ · cowbell · FRISS · ZELROS · zesty AI · envelop · AlphaSense · ZEST AI · Socure · SymphonyAI SENSA · enova · Taktile

Healthcare

flatiron · GRAIL · Clarify · babylon · Biofourmis · PathAI · cleerly · aidoc · AiCure · SIGTUPLE · Olive · komodo · Syllab · Syapse · ITERATIVE HEALTH · springhealth · Human Dx · iz.ai · ENLITIC · ARTERY · IMAGEN · abacus insights · LeanTaaS · ScienceIO · POPULI · briya · vara · Rad AI · regard · Nabla · FATHOM · AKASA · Quibim


ion & 3D
ivataar
CSM
inworld
DATAGRID
LUMA AI
DIDIMO
mirage
KAEDIM
HYPOTHETIC
Search
neeva
mem
Metaphor
Perplexity
genei
YOU
DeepSearch
Labs
AIL
Clarify
babylon
Biofourmis
PathAI
TEMPUS
HUMAN LONGEVITY
AiCure
SIGTUPLE
Olive
komodo
Syllable
innovaccer
ITERATIVE HEALTH
spring health
Human Dx
RISE TOGETHER
iz.ai
ENLITIC
ARTERYS
Health
IMAGEN
abacus insights
LeanTaaS
ScienceIO
POPULI
briya
vara
Rad AI
regard
Nabla
FATHOM
AKASA
sonio
Quibim
Life Sciences
Insilico Medicine
Benchling
23andMe
color
SOPHiA GENETICS
verily
Benevolent AI
DNAnexus
iCarbonX
Schrödinger
freenome
Atomwise
ROME THERAPEUTICS
RELAY THERAPEUTICS
ArsenalBio
OWKIN
nimbus THERAPEUTICS
deep genomics
Recursion
insitro
Exscientia
Verana Health
Valo
moderna
ConcertAI

Open Source Infrastructure

Frameworks

Format

Query/Dataflow

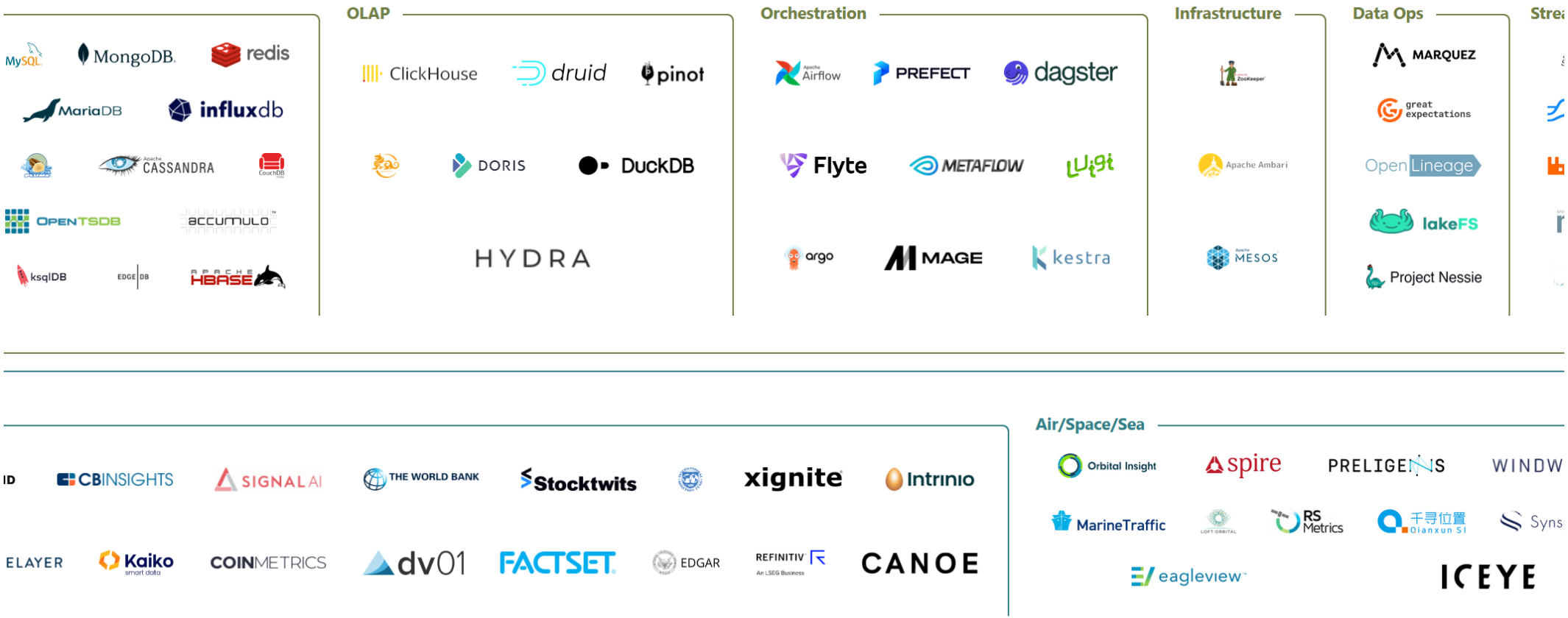
Data Access

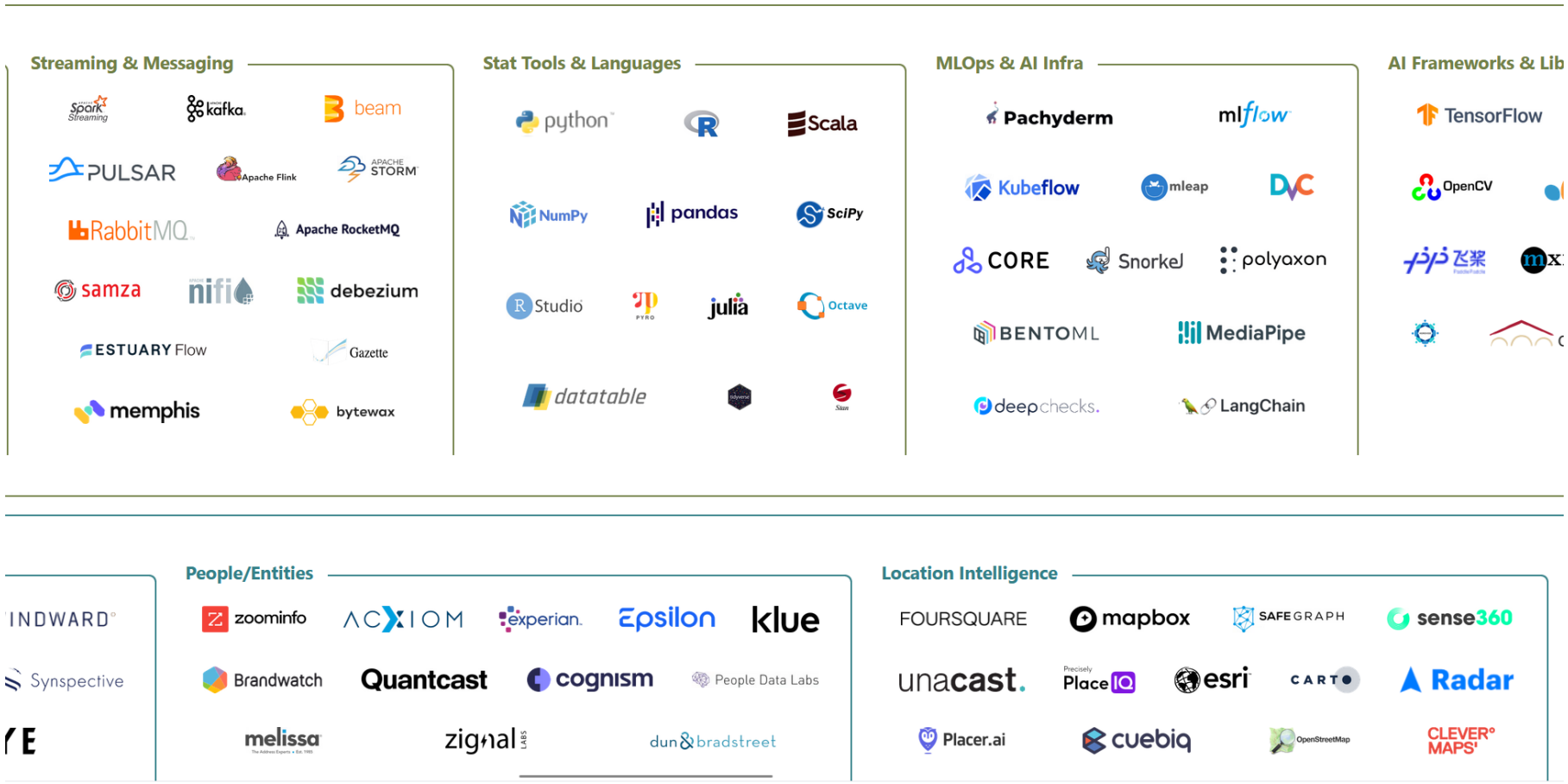
Databases

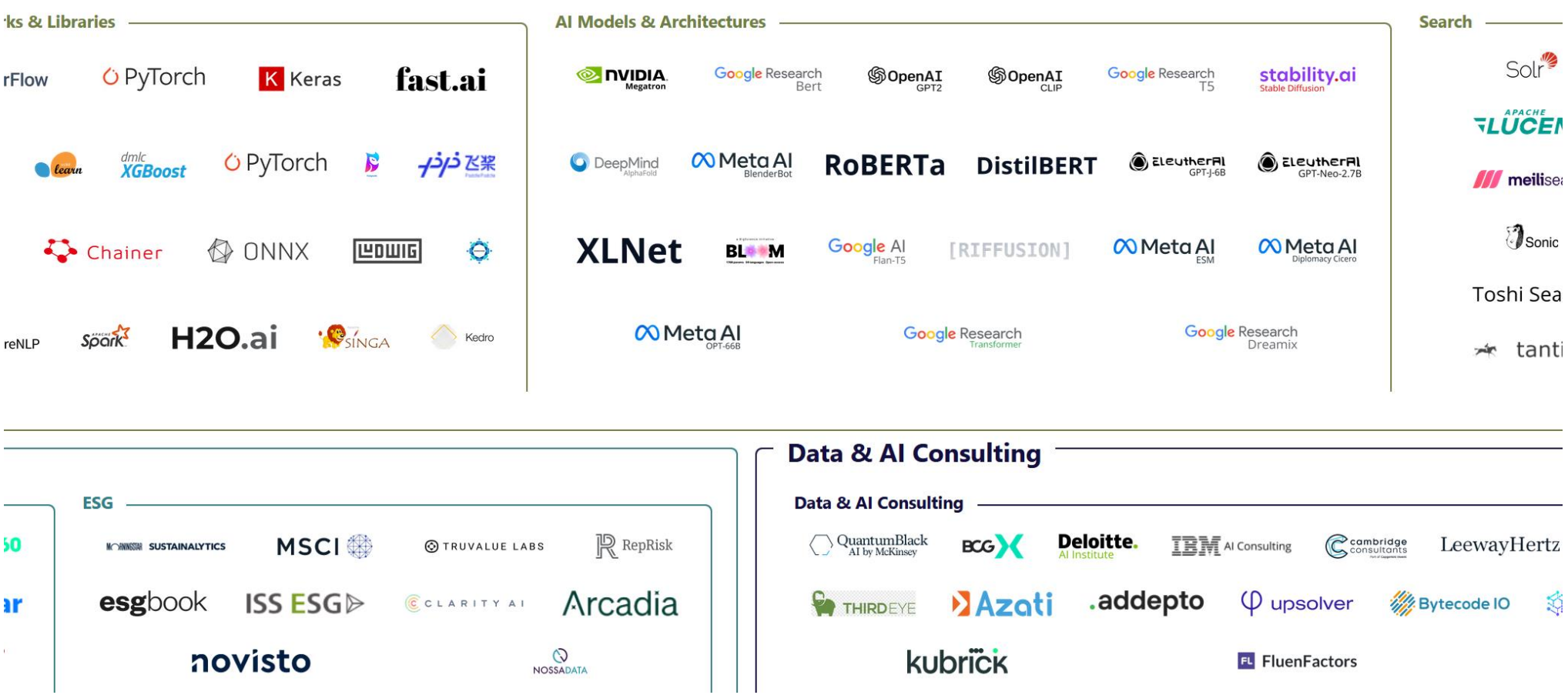
Data Sources & APIs

Data Marketplaces & Discovery

Financial & Market Data


MySQL
MongoDB.
redis
MariaDB
influxdb
Apache CASSANDRA
CouchDB
OPENTSDB
accumulo
ksqlDB
EDGE DB
APACHE HBASE
OLAP
ClickHouse
druid
pinot
DORIS
DuckDB
HYDRA
Orchestration
Apache Airflow
PREFECT
dagster
Flyte
METAFLOW
luigi
argo
MAGE
kestra
Infrastructure
Apache ZooKeeper
Apache Ambari
Apache MESOS
Data Ops
MARQUEZ
great expectations
OpenLineage
lakeFS
Project Nessie
ID
CBINSIGHTS
SIGNAL AI
THE WORLD BANK
Stocktwits
xignite
Intrinio
ELAYER
Kaiko
smart data
COINMETRICS
dv01
FACTSET
EDGAR
REFINITIV
An LSEG Business
CANOE
Air/Space/Sea
Orbital Insight
spire
PRELIGENS
WINDW
MarineTraffic
LOFT ORBITAL
RS Metrics
千寻位置
Qianxun SI
Syns
eagleview
ICEYE


Streaming & Messaging
Spark Streaming
kafka
beam
PULSAR
Apache Flink
APACHE STORM
RabbitMQ
Apache RocketMQ
samza
nifi
debezium
ESTUARY Flow
Gazette
memphis
bytewax
Stat Tools & Languages
python
R
Scala
NumPy
pandas
SciPy
Studio
PYRO
julia
Octave
datatable
Stan
MLOps & AI Infra
Pachyderm
mlflow
Kubeflow
mleap
DVC
CORE
Snorkel
polyaxon
BENTOML
MediaPipe
deepchecks.
LangChain
AI Frameworks & Lib
TensorFlow
OpenCV
飞桨
INDWARD
Synspective
People/Entities
zoominfo
ACXIOM
experian.
Epsilon
klue
Brandwatch
Quantcast
cognism
People Data Labs
melissa
zignal LABS
dun & bradstreet
Location Intelligence
FOURSQUARE
mapbox
SAFEGRAPH
sense360
unacast.
Precisely PlaceIQ
esri
CARTO
Radar
Placer.ai
cuebiq
OpenStreetMap
CLEVER° MAPS'


ks & Libraries
rFlow
PyTorch
Keras
fast.ai
dmlc XGBoost
PyTorch
Chainer
ONNX
reNLP
Spark
H2O.ai
SINGA
Kedro
AI Models & Architectures
NVIDIA Megatron
Google Research Bert
OpenAI GPT2
OpenAI CLIP
Google Research T5
stability.ai Stable Diffusion
DeepMind AlphaFold
Meta AI BlenderBot
RoBERTa
DistilBERT
EleutherAI GPT-J-6B
EleutherAI GPT-Neo-2.7B
XLNet
BLOOM
Google AI Flan-T5
[RIFFUSION]
Meta AI ESM
Meta AI Diplomacy Cicero
Meta AI OPT-66B
Google Research Transformer
Google Research Dreamix
Search
Solr
APACHE LUCENE
meilisea
Sonic
Toshi Sea
tanti
ESG
SUSTAINALYTICS
MSCI
TRUVALUE LABS
RepRisk
esgbook
ISS ESG
CLARITY AI
Arcadia
novisto
NOSSADATA
Data & AI Consulting
Data & AI Consulting
QuantumBlack AI by McKinsey
BCGX
Deloitte. AI Institute
IBM AI Consulting
cambridge consultants
LeewayHertz
THIRDEYE
Azati
.addepto
upsolver
Bytecode IO
kubrick
FL FluenFactors


Google Research T5
stability.ai Stable Diffusion
EleutherAI GPT-J-6B
EleutherAI GPT-Neo-2.7B
Meta AI ESM
Meta AI Diplomacy Cicero
Google Research Dreamix
Search
Solr
APACHE LUCENE
meilisearch
Sonic
Toshi Search
tantivy
Logging & Monitoring
elastic
kibana
Sentry
logstash
Prometheus
fluentbit
fluentd
Grafana
VECTOR
OpenTelemetry
Visualization
Apache Superset
matplotlib
re:dash
TensorFlow TensorBoard
seaborn
bokeh
Collaboration
BeakerX
jupyter
Apache Zeppelin


IBM AI Consulting
cambridge consultants
LeewayHertz
slalom
BROOKLYN DATA CO.
TERNARY DATA
TRIBE AI
itrex
DEEPER INSIGHTS
data root labs
InData Labs
upsolver
Bytecode IO
LEIT DATA
MERU
BIG TIME DATA
Data Captains
4MILE
Revolt BI
Analytics8
phData
kubrick
FluenFactors
DEEPSK DATA
EY Analytics Consulting Services
ZS
CapTech

Where to go from here?

Learning by doing

- Try doing something yourself

Where to go from here?

Good resources

- Stanford MLSys Seminar series

https://mlsys.stanford.edu/

- Chip Huyen's blog and book

https://huyenchip.com/blog/

- Rules of ML

https://developers.google.com/machine-learning/guides/rules-of-ml/

- Coursera

https://www.coursera.org/specializations/machine-learning-engineering-for-production-mlops

- ML@CL website

https://mlatcl.github.io/

Where to go from here?

Papers

- Monitoring and explainability of models in production. Klaise et al., ICML DMML workshop 2020
- Hidden Technical Debt in Machine Learning Systems. Sculley et al., NeurIPS 2015
- Challenges in Deploying Machine Learning: a Survey of Case Studies. Paleyes et al, ACM Comp. Surv. 2022
- Scaling Big Data Mining Infrastructure — The Twitter Experience. Lin and Ryaboy, KDD 2013
- MLOps: A Taxonomy and a Methodology. Testi et al., IEEE Access 2022
- 150 successful machine learning models: 6 lessons learned at Booking.com. Bernardi et al, KDD 2019
- Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. Ashmore et al., ACM Comp. Surv. 2021
- Software engineering for machine learning: A case study. Amerishi et al., ICSE 2019
- Data lifecycle challenges in production machine learning: a survey. Polyzotis et al., ACM SIGMOD Record 2018

Summary

- Sometimes ML is not needed
- Get you success metric right
- Getting data can be hard
- So can be storing
- Simple models often work best
- Models live in different places
- Monitor for drift in data and metrics

- Also consider
  - Fairness, law and ethics
  - Privacy and security
  - Quality assurance
  - Good software engineering practices
  - User interface

Questions?

Appendix

Good software engineering practices

- Version control is good (e.g. git)
- Code reviews are good
- Unit tests are good
- Separation of concerns is good
- Naming is important

Data ethics

- Who owns the data?
- How was it collected?
- Do you have explicit permission to use the data?
- Can you identify individuals from the dataset?
- Can you use model trained on this dataset commercially?
- Can you use privacy techniques?

Fairness

- Can the training data contain biases?
  - Explicit biases
  - Hidden biases aka proxies
- Are the classes balanced?
- Is there a potential to aggravate bias?

Law

- Is your area highly regulated?
  - Healthcare
  - Finance
  - Judicial
- Will you operate somewhere with country-level laws?
  - EU – GDPR
  - Canada - PIPEDA
  - Kenya – Data Protection Act (similar laws exist in Uganda, Nigeria and South Africa)

Security

- Data poisoning
- Model reverse engineering aka model stealing
- Model inversion

Quality assurance

- Do you have acceptance criteria for the system?
- How does model performance translate to business value?
- Can you test model in real life, e.g. with A/B test?
- If not, maybe you can use a simulation?

User interface

- Good UX increases trust
- Bad UX limits adoption
- ML terms are not easy to understand
- Focus on what user wants to see, not on ML