Chapter BD+LT:I

I. Introduction

- ❑ Language Technologies
- ❑ Big Data Processing Architectures
- ❑ ML/AI/Data Tools Landscape

Goals of Language Technology

1. Aid humans in writing.
   Correcting mistakes, formulating and paraphrasing text, transcription.

2. Identify texts related to spoken or written requests.
   Text information retrieval, semantic text similarity, question answering.

3. Make sense of texts without reading the originals.
   Categorization, information extraction, summarization, translation.

4. Instruct, and be advised by a computer.
   Audio interfaces (e.g., dialog systems, robotics), learning and assessment.

↓

5. Converse with computers as if they were human.
   Turing test, conversational AI and chatbots, computational humor.
   What is the nature of language and its relation to (artificial) intelligence?

Language Technologies

A Brief History

1940s	Machine Translation (Shannon & Weaver, 1949)
1950s	Turing Test (Turing, 1950)
	Generative Grammars (Chomsky, 1957 & 1965)
1960s	Chatbots: ELIZA (Weizenbaum, 1966), SHRDLU (Winograd, 1968)
1970s–80s	Ontologies & Symbolic AI
1990s	Statistical NLP
2000s	Neural Language Models (Bengio et al., 2003)
	IBM Watson (2006–2011)

Language Technologies

A Brief History (continued)

2010s Word Embeddings (Mikolov et al., 2013; Pennington et al., 2014)
Deep Recurrent & Convolutional Nets in NLP
(e.g. LSTM; Hochreiter & Schmidhuber, 1997)
Sequence-to-sequence models (Sutskever et al., 2014)
Attention (Bahdanau et al., 2015)

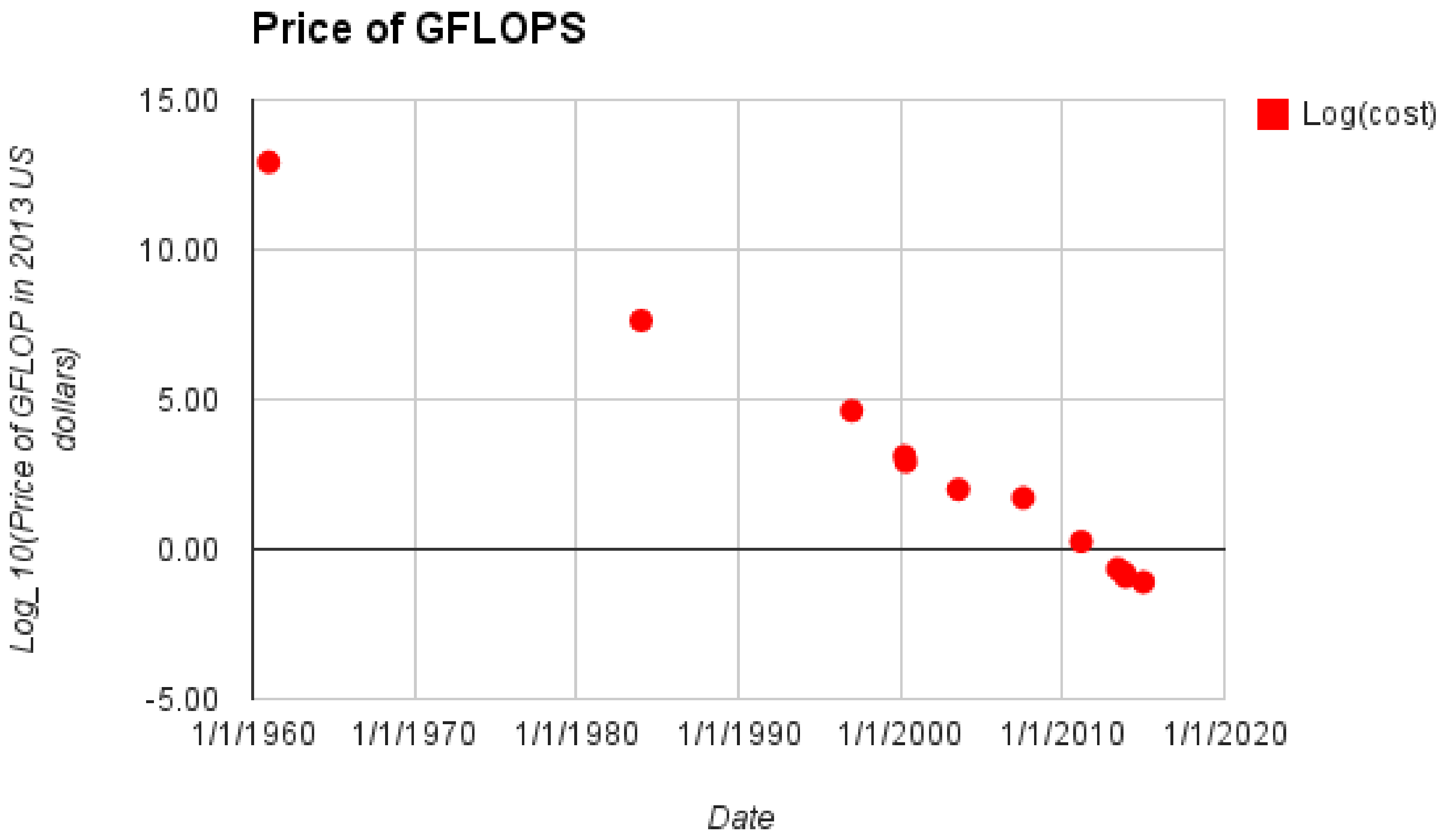
2017 Self-attention; Transformer (Vaswani et al., 2017)

2018 Pre-training: ELMo (Peters et al., 2018), BERT (Devlin et al., 2018),
GPT (Radford et al., 2018)

2020 Large language models, zero-shot transfer:
GPT-3 (Brown et al., 2020), T-Zero (Sanh et al., 2021),
GPT-NeoX-20B (Black et al., 2022)

Language Technologies

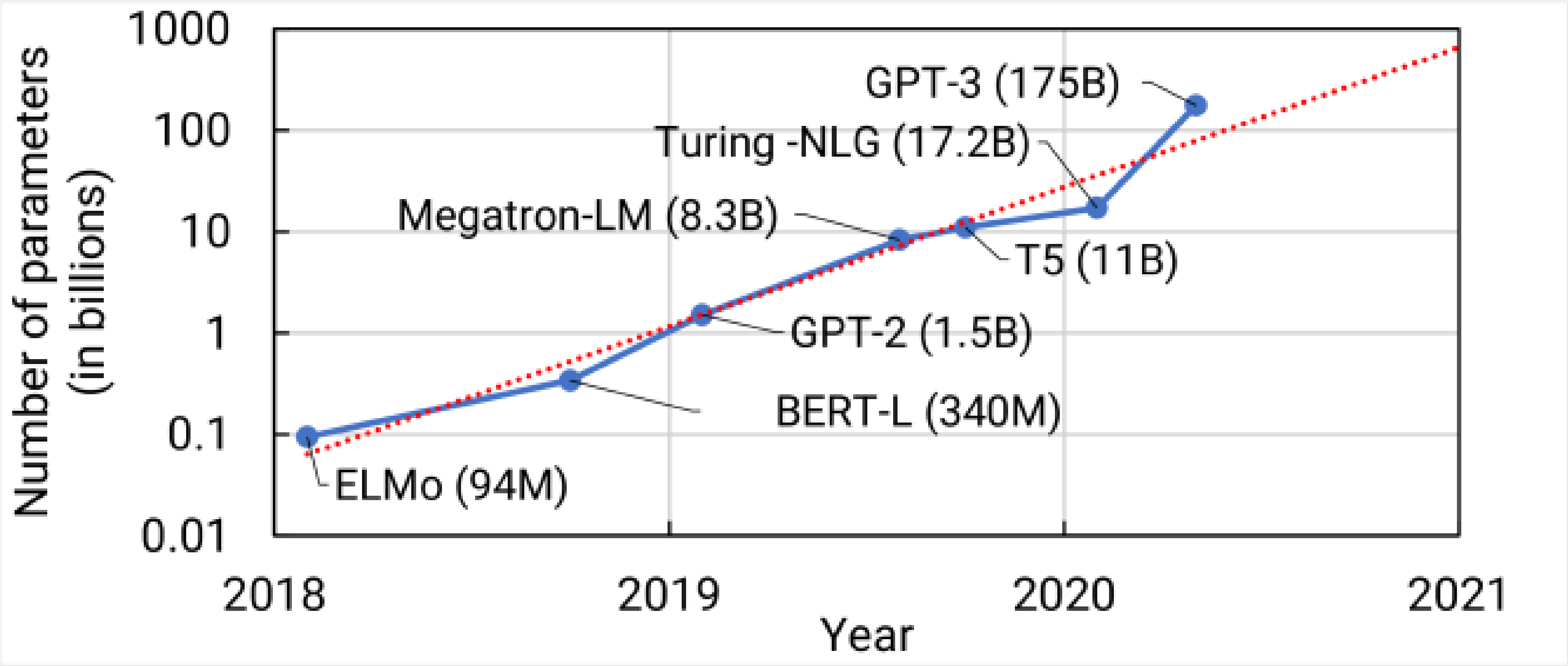
Cost of Compute



Source: [aiimpacts.org], pricing data from Wikipedia

Language Technologies

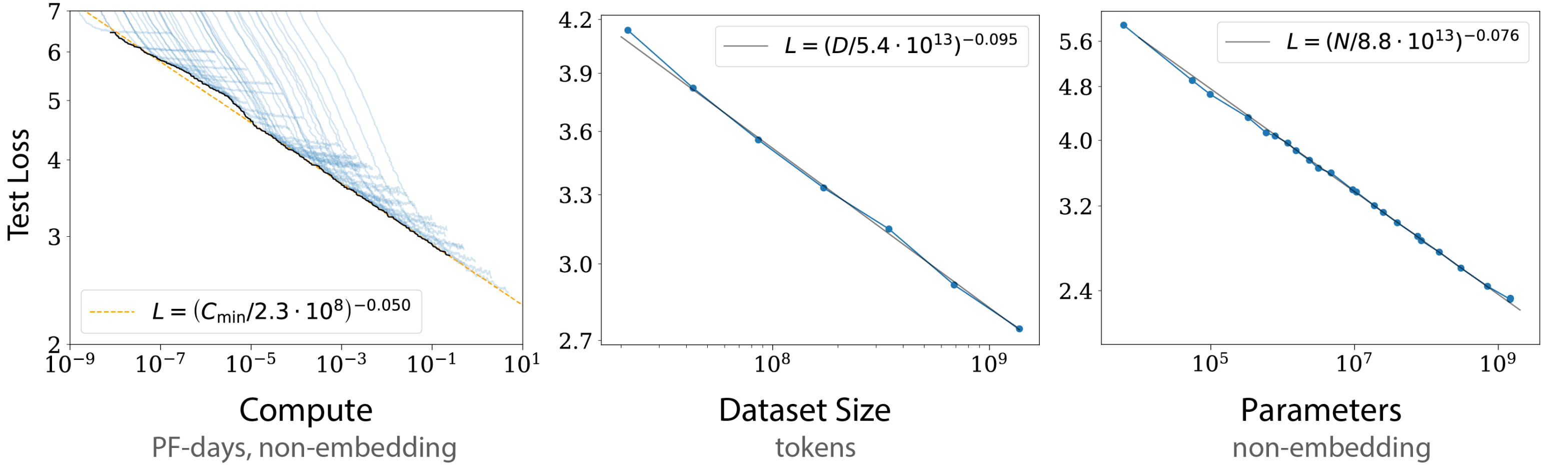
Language Model Size



(Narayanan, 2021)

Language Technologies

Scaling Laws



- More parameters as compute budget increases (Kaplan et al., 2020)
- Amount of training data needs to grow accordingly (Hoffmann et al., 2022)

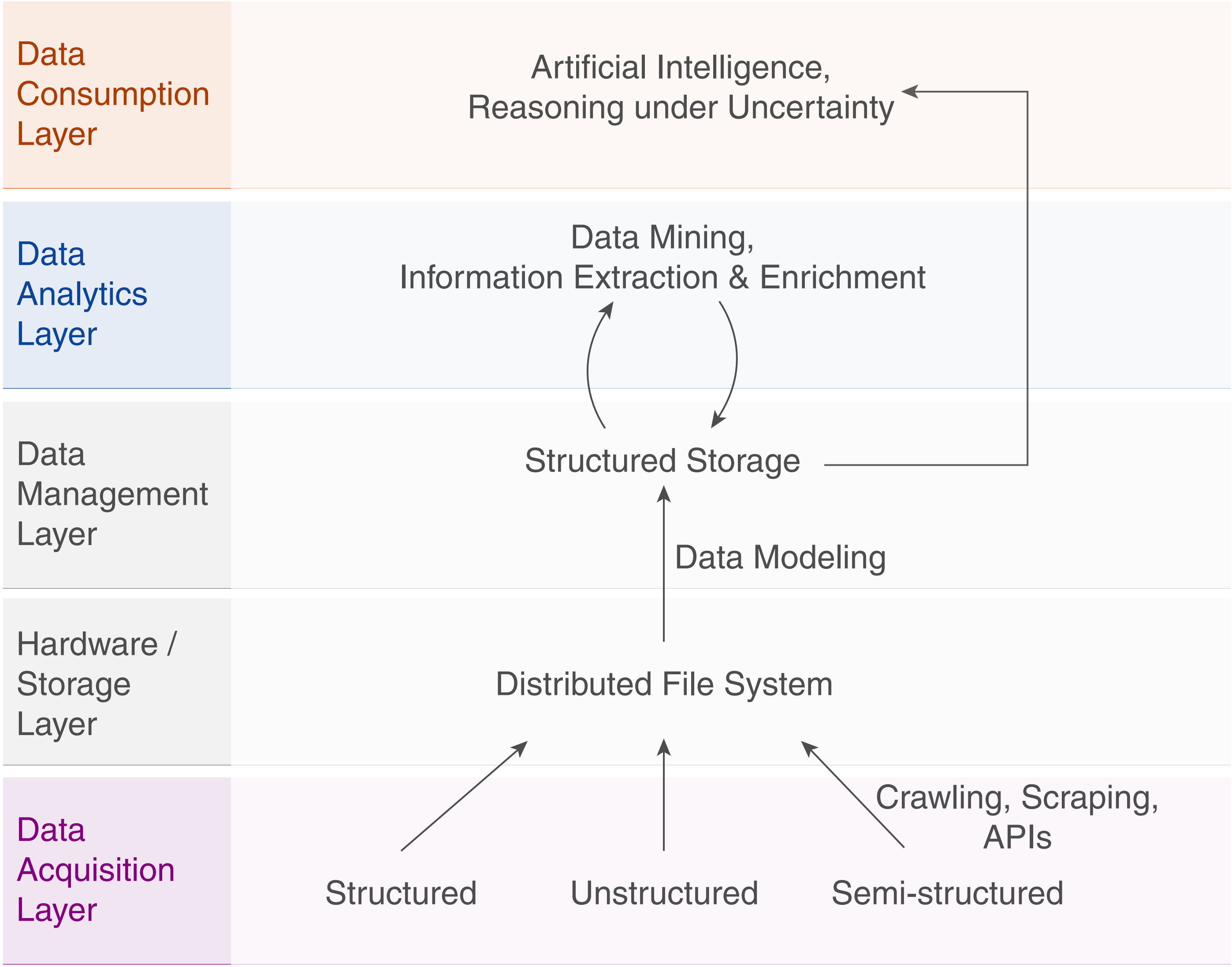
Language Technologies

Compute & Data Requirements

- ❑ GPT-NeoX-20B: Open-source autoregressive language model, trained for 2.5 months on 96 Nvidia A-100 GPUs (Black et al., 2022)

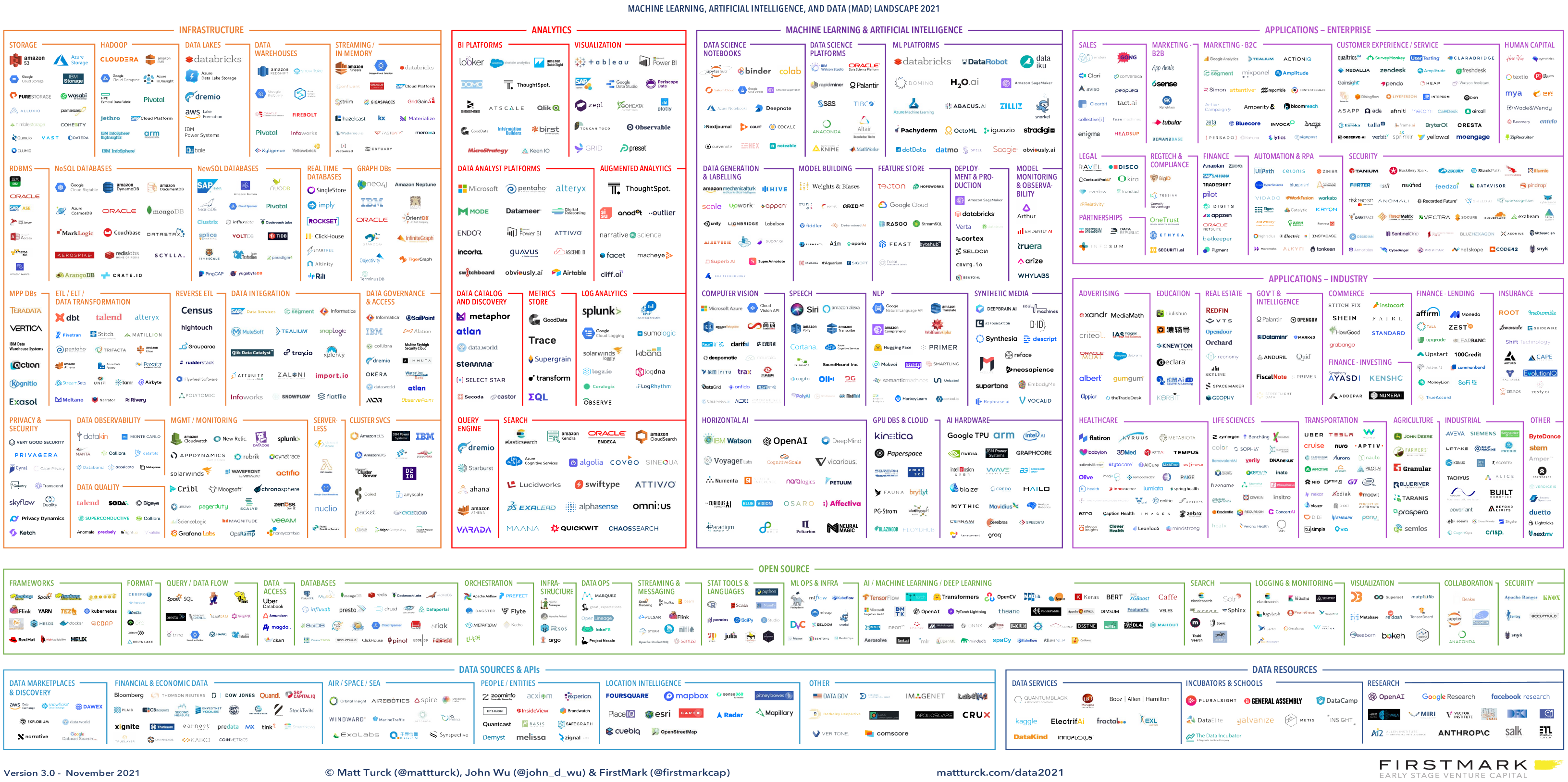
- ❑ The Pile: 825GiB open-source curated text dataset (Gao et al., 2020)

Big Data Architecture Stack



Big Data Tools

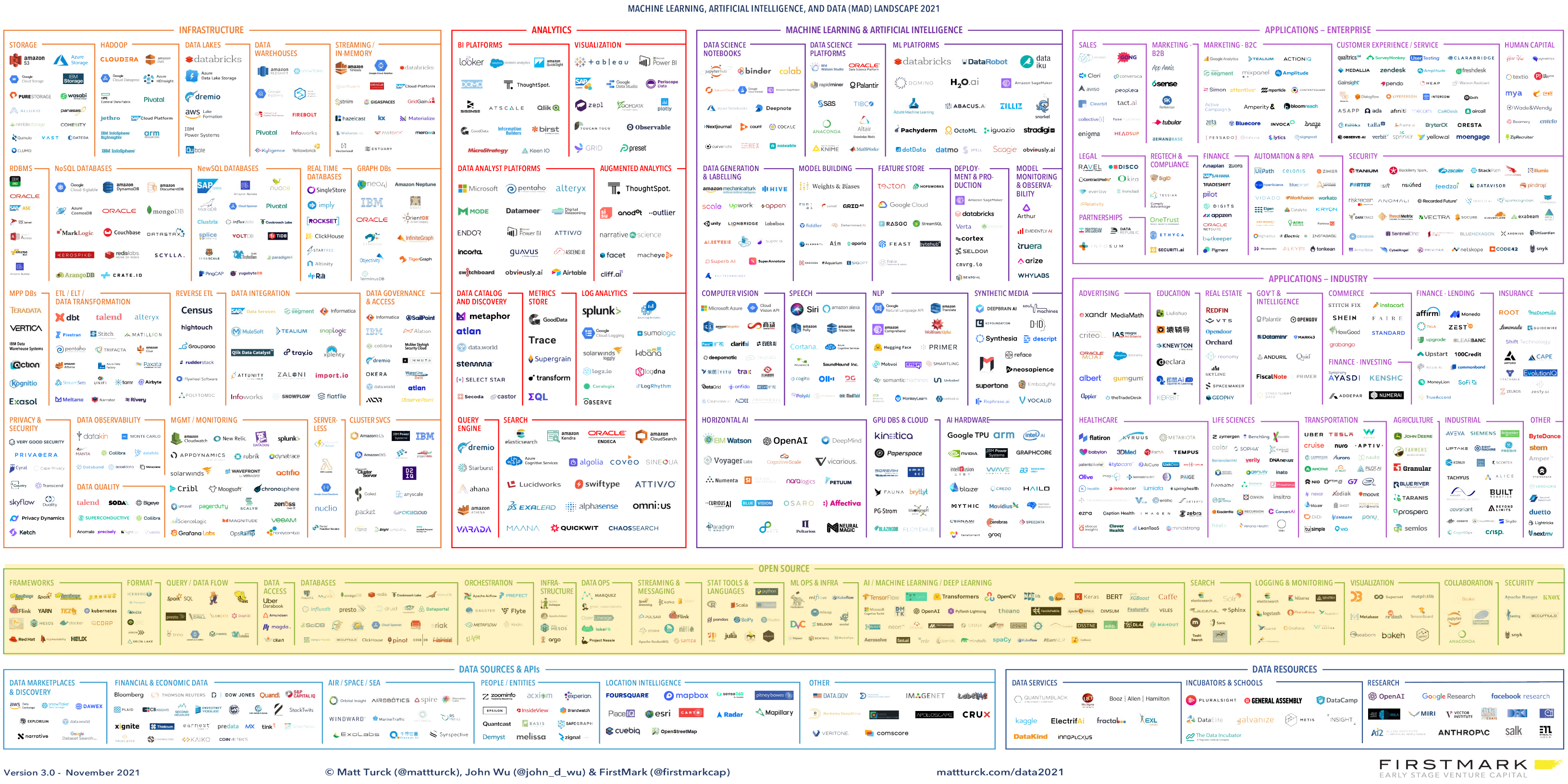
ML, AI, and Data (“MAD”) Landscape



[mattturck.com/data2021]

Big Data Tools

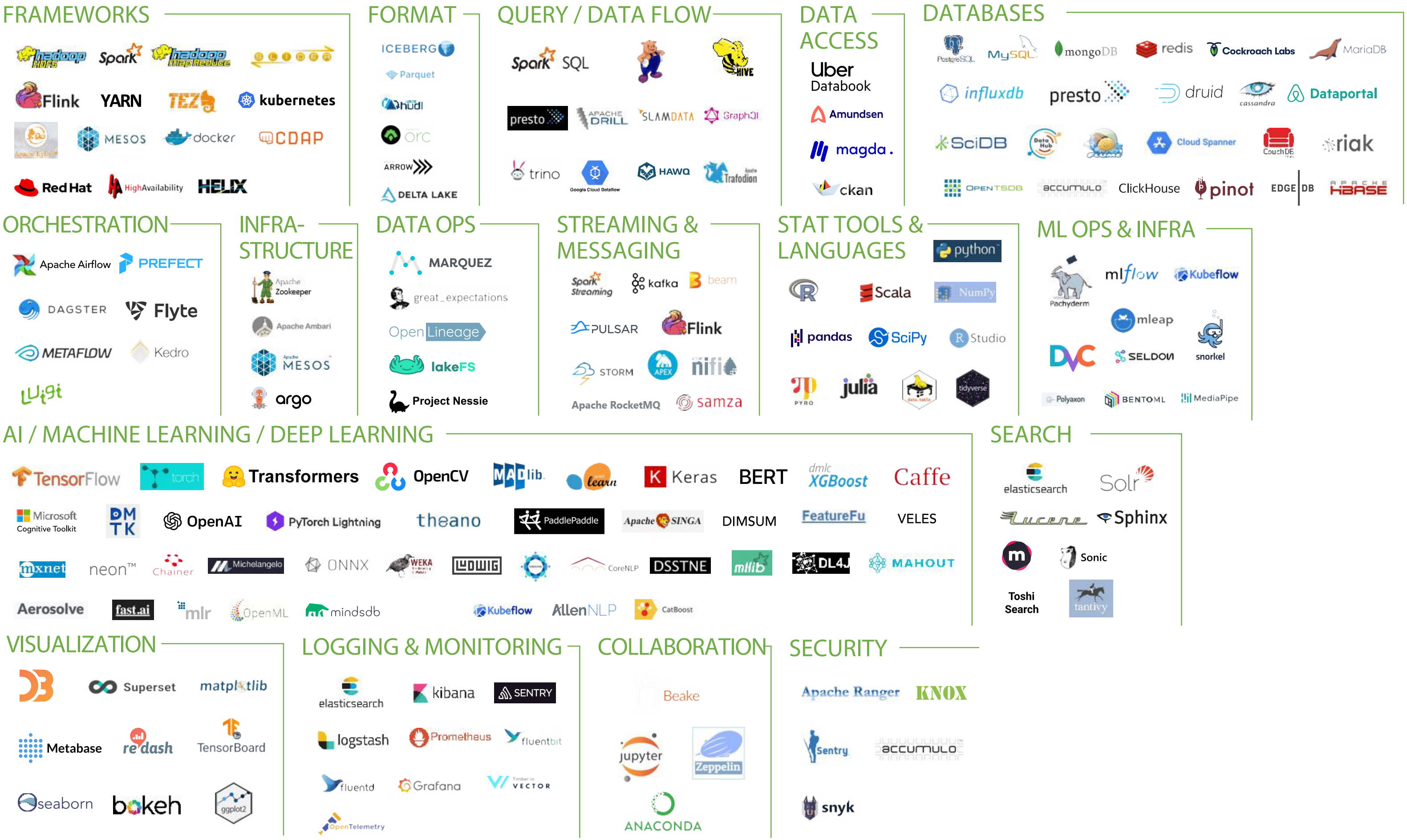
ML, AI, and Data (“MAD”) Landscape



[mattturck.com/data2021]

ML, AI, and Data (“MAD”) Landscape

Open Source Tools



FRAMEWORKS
Hadoop HDFS, Spark, Hadoop MapReduce, Flink, YARN, TEZ, kubernetes, Apache Kylin, MESOS, docker, CDAP, Red Hat, HighAvailability, HELIX

FORMAT
ICEBERG, Parquet, hudi, orc, ARROW, DELTA LAKE

QUERY / DATA FLOW
Spark SQL, HIVE, presto, APACHE DRILL, SLAMDATA, GraphQl, trino, Google Cloud Dataflow, HAWQ, Apache Trafodion

DATA ACCESS
Uber Databook, Amundsen, magda., ckan

DATABASES
PostgreSQL, MySQL, mongoDB, redis, Cockroach Labs, MariaDB, influxdb, presto, druid, cassandra, Dataportal, SciDB, Data Hub, Cloud Spanner, CouchDB, riak, OPENTSDB, accumulo, ClickHouse, pinot, EDGEDB, APACHE HBASE

ORCHESTRATION
Apache Airflow, PREFECT, DAGSTER, Flyte, METAFLOW, Kedro, Luigi

INFRASTRUCTURE
Apache Zookeeper, Apache Ambari, Apache MESOS, argo

DATA OPS
MARQUEZ, great_expectations, OpenLineage, lakeFS, Project Nessie

STREAMING & MESSAGING
Spark Streaming, kafka, beam, PULSAR, Flink, STORM, APEX, nifi, Apache RocketMQ, samza

STAT TOOLS & LANGUAGES
python, R, Scala, NumPy, pandas, SciPy, RStudio, PYRO, julia, data.table, tidyverse

ML OPS & INFRA
Pachyderm, mlflow, Kubeflow, mleap, DVC, SELDON, snorkel, Polyaxon, BENTOML, MediaPipe

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, torch, Transformers, OpenCV, MADlib, scikit-learn, Keras, BERT, dmlc XGBoost, Caffe, Microsoft Cognitive Toolkit, DMTK, OpenAI, PyTorch Lightning, theano, PaddlePaddle, Apache SINGA, DIMSUM, FeatureFu, VELES, mxnet, neon™, Chainer, Michelangelo, ONNX, WEKA, LUDWIG, CoreNLP, DSSTNE, mllib, DL4J, MAHOUT, Aerosolve, fast.ai, mlr, OpenML, mindsdb, Kubeflow, AllenNLP, CatBoost

SEARCH
elasticsearch, Solr, Lucene, Sphinx, Sonic, Toshi Search, tantivy

VISUALIZATION
D3, Superset, matplotlib, Metabase, redash, TensorBoard, seaborn, bokeh, ggplot2

LOGGING & MONITORING
elasticsearch, kibana, SENTRY, logstash, Prometheus, fluentbit, fluentd, Grafana, VECTOR, OpenTelemetry

COLLABORATION
Beaker, jupyter, Zeppelin, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry, accumulo, snyk

[mattturck.com/data2021]