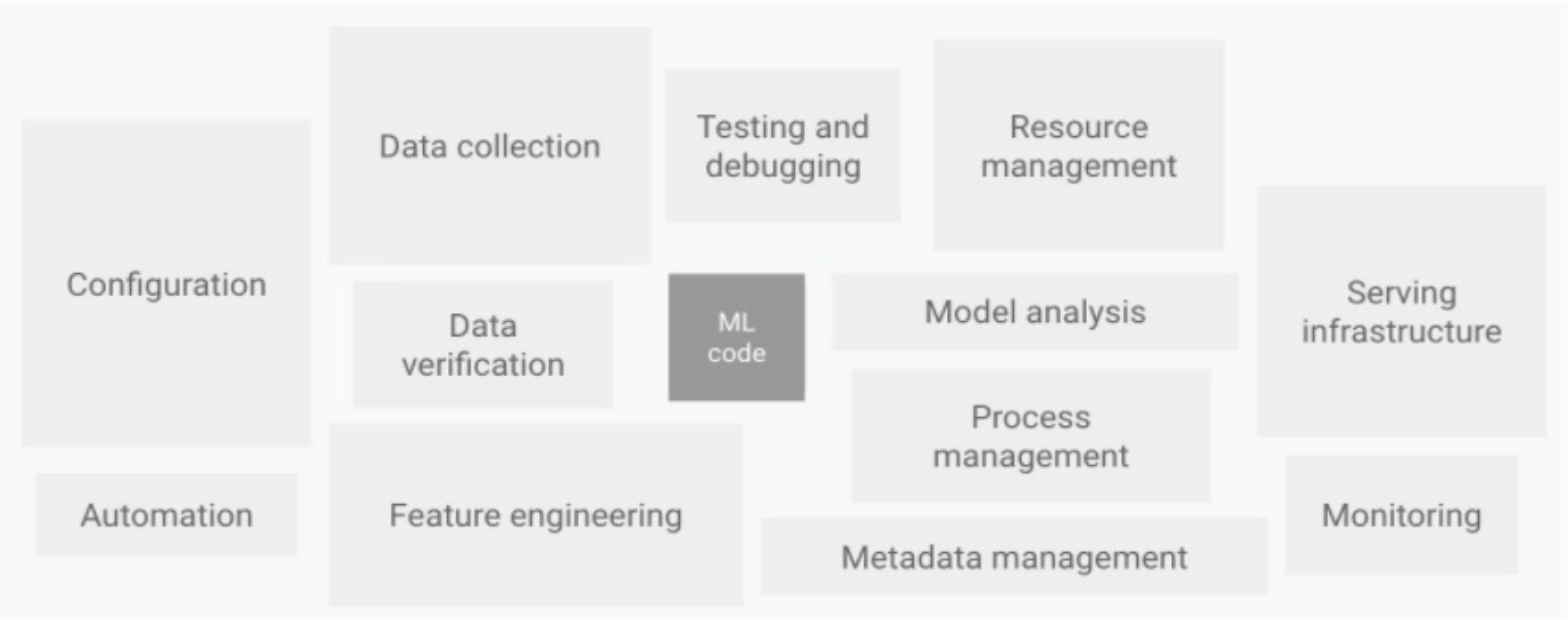
2 PoC to Production Gap

PoC versus Production

"All of AI, .., has a proof-of-concept-to-production gap. The full cycle of a machine learning project is not just modelling. It is finding the right data, deploying it, monitoring it, feeding data back [into the model], showing safety—doing all the things that need to be done [for a model] to be deployed. [That goes] beyond doing well on the test set, which fortunately or unfortunately is what we in machine learning are great at."

- Andrew Ng

The big picture



[Google - CD and automation pipelines in ML]

Basic ML building blocks

Data Management	Experimentation	Production
Process and govern the data used by models: • Usually large data sets • Should be of high quality • Should be compliant with legislation • Should be tracked	Build a model based on business requirements, after iteration of experimentation: • Workflow is iterative • Experiment should be tracked • Code should have standards • Accuracy metrics should be tracked • Retraining should be possible • Requires specific infrastructure	Integrate prediction into production and business processes: • Generate systematic predictions • Track performance across time • Follow best engineering practices

Moving to production is hard

(Not so) Fun fact

According to VentureBeat, roughly 1 out of 10 Machine Learning models actually makes it into production. But why?

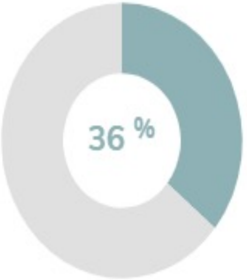
The Set up is not right

- Bad infrastructure
- Disconnect between the relevant parties
- Poor data management
- Leadership doesn’t understand

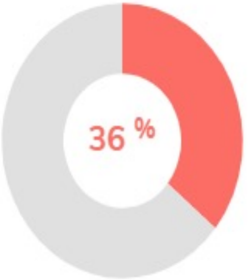
ML has its own difficulties

- Scaling is not easy
- Duplication is widespread
- Management not on board
- Lack of Reproducibility
- Support across technologies

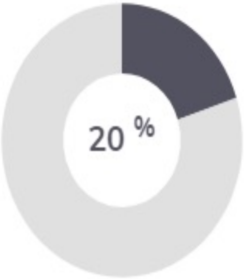
Deploying models takes time



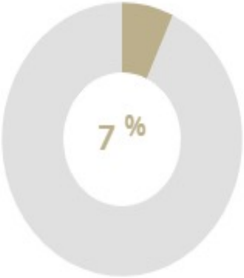
36% of survey participants said their data scientists spend **a quarter** of their time deploying ML models



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



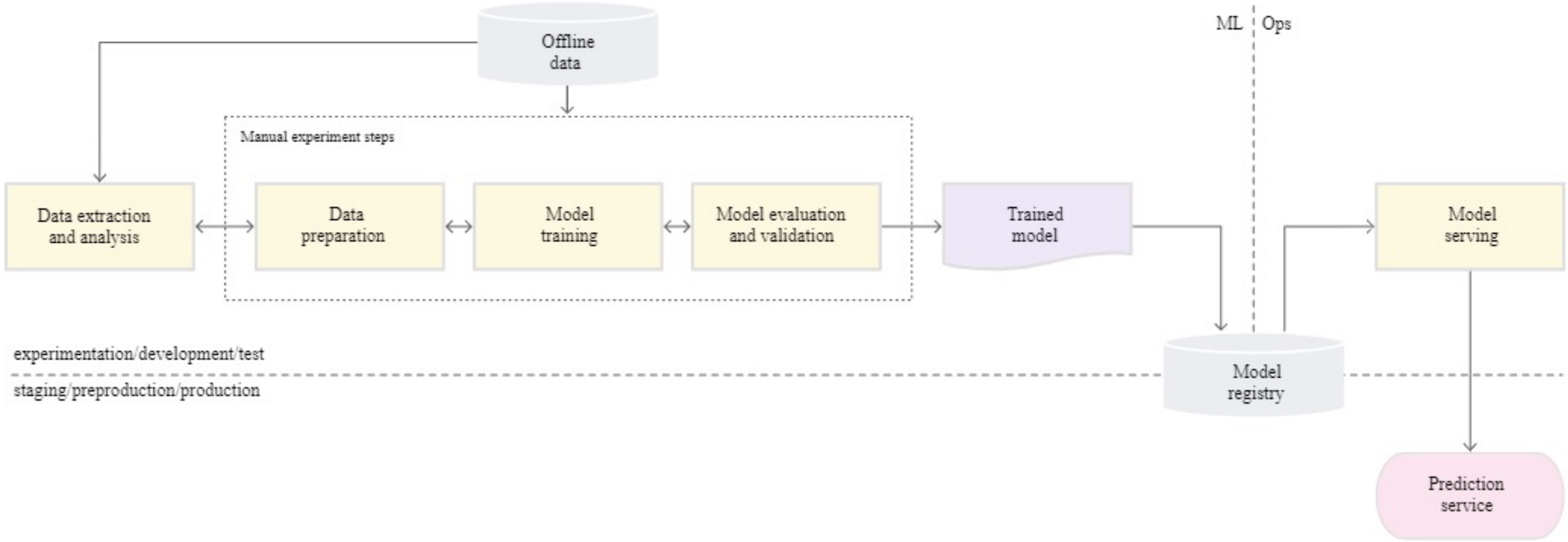
20% of survey participants said their data scientists spend **half to three-quarters** of their time deploying ML models



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

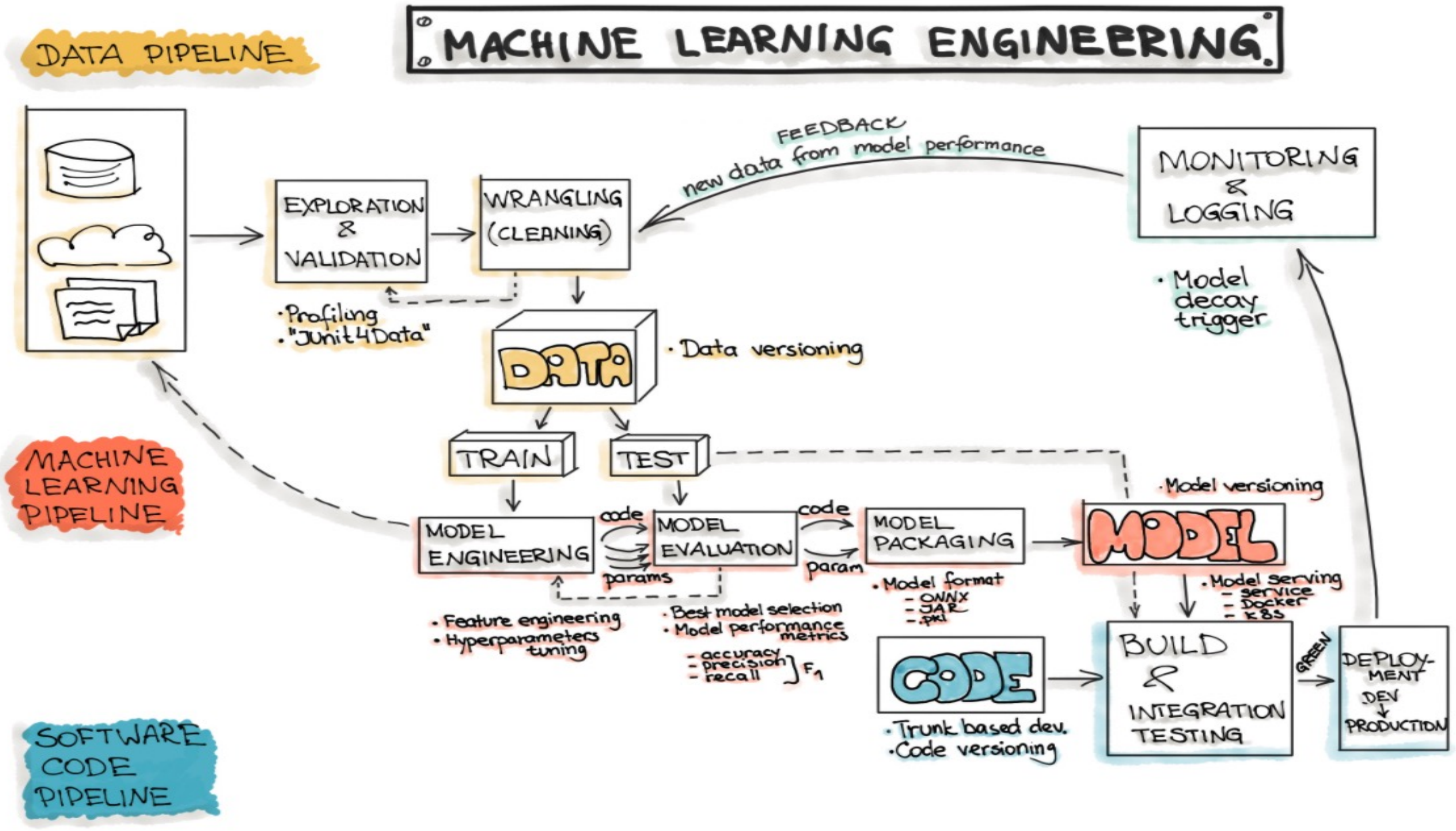
Algorithmia - State of Enterprise ML

Basic process for building a model



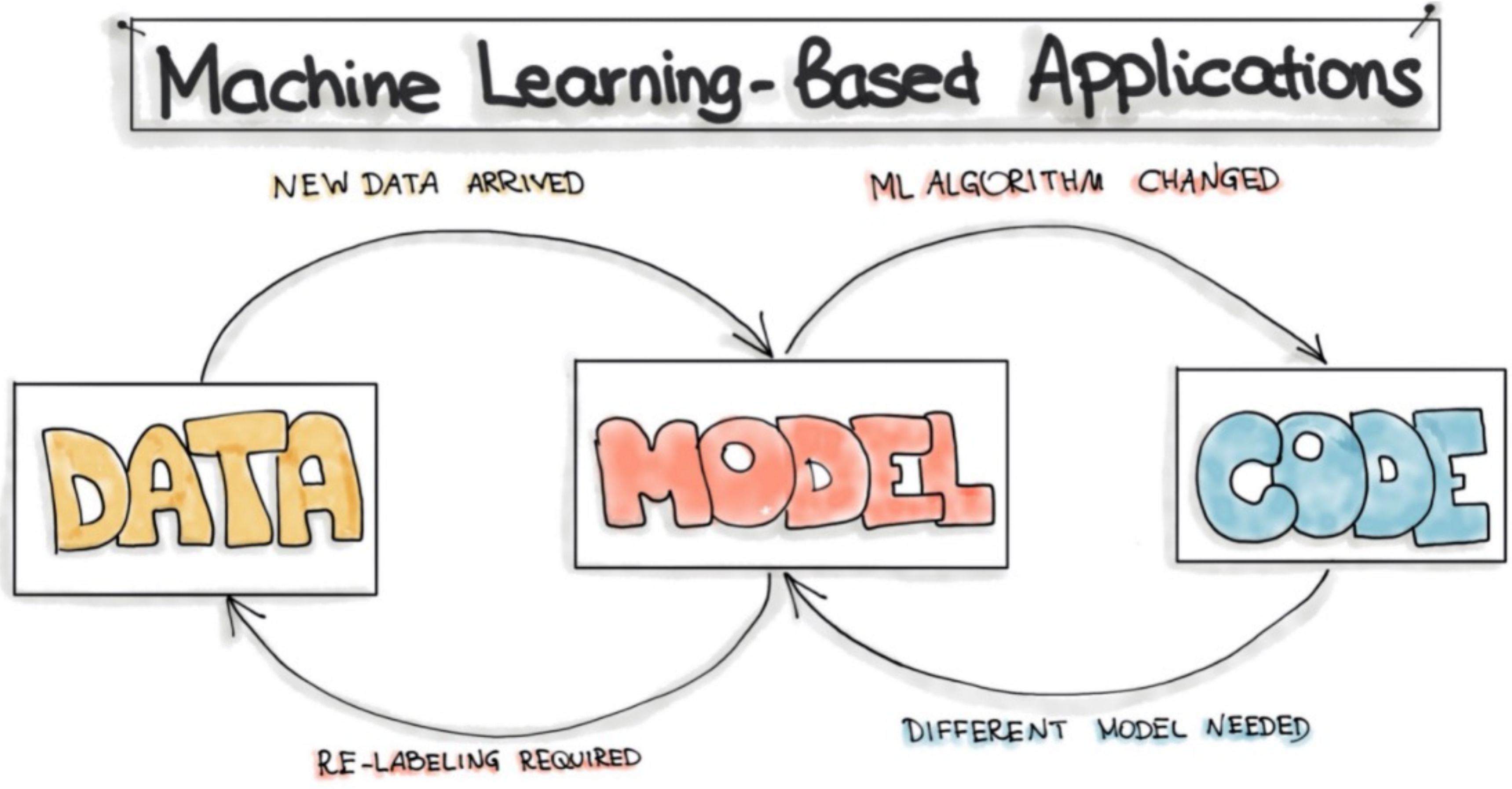
Google - CD and automation pipelines in ML

Real life is a bit more complicated



[MLOps.org - End-to-end workflow]

Changing anything changes all



MLOps.org - Motivation for MLOps

Hidden technical debt

Developing and deploying ML systems is relatively fast and cheap, but maintaining them over time is difficult and expensive. Some of the reasons for this are:

- Data dependencies cost more than code dependencies
- Feedback Loops
- ML-Systems anti-patterns
- Configuration debts
- Always changing external world
- Other ML related debt (e.g Data testing, Reproducibility debt)

Other production issues

- Data quality:
  - ML models reflect the data they are build on, so they are very dependent on its size and quality
- Model decay:
  - As times goes by, there might be changes in behavior that the original data would not necessarily reflect causing the quality of the model to drop
- Locality:
  - The quality of the performance of ML model does not always translates completely to production