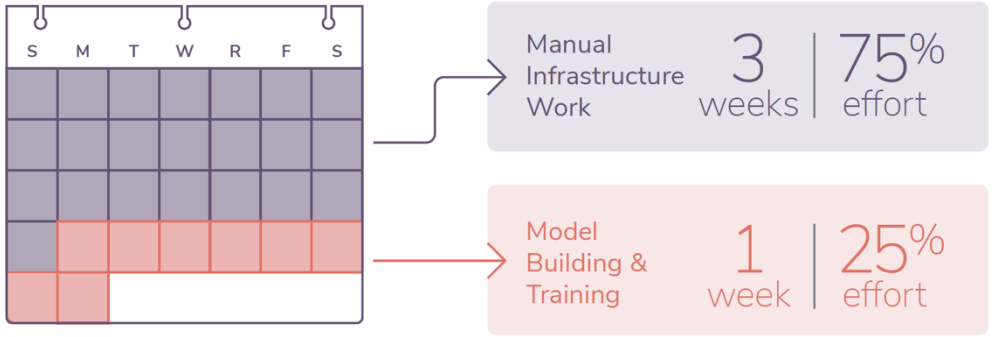
ALGORITHMIA

The Challenges and Roadblocks of Machine Learning

Hernan Alvarez
VP of Product, Algorithmia
halvarez@algorithmia.io

Survey: Teams are Capable of Much More

75% of Time Spent on Infrastructure

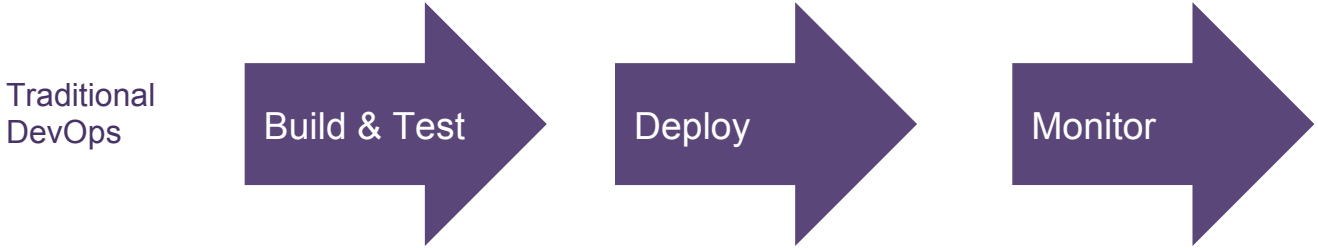


Key Challenges

- **30%:** supporting different languages and frameworks.
- **30%:** model management tasks such as versioning and reproducibility.
- **38%:** deploying models at necessary scale.

* - survey of > 500 practitioners & management in summer of 2018

But DevOps!



Machine Learning is Different...

Heterogeneity

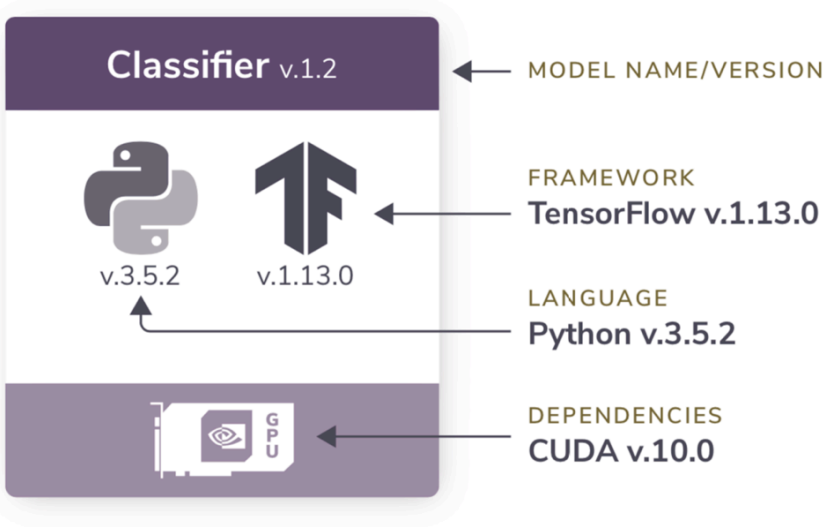
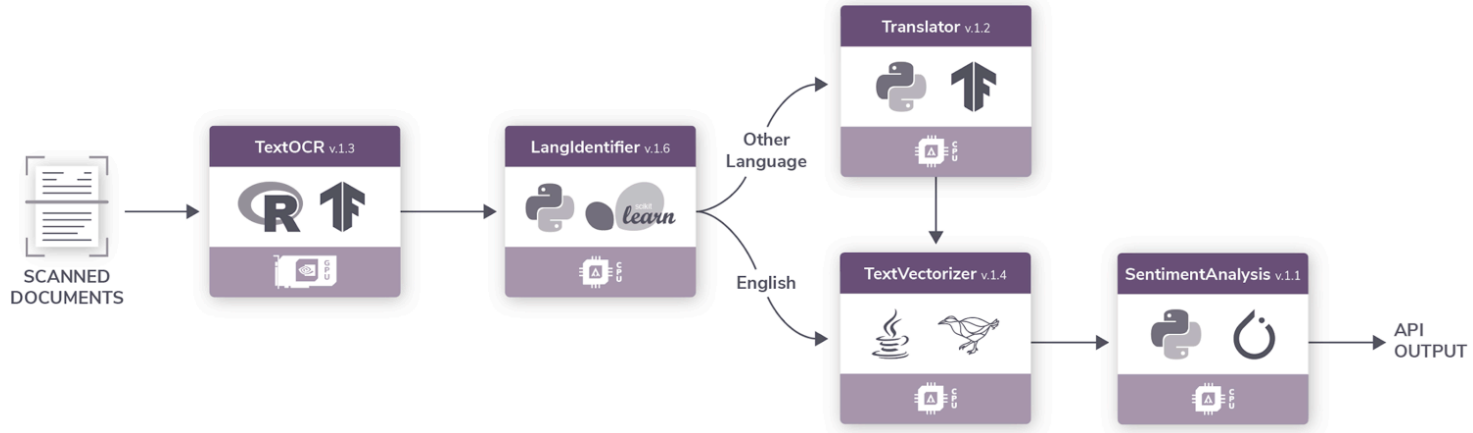


Figure 1.1: A *model card* showing ML model and version, the language version in which it's written, the framework version on which it's built, and any dependencies.

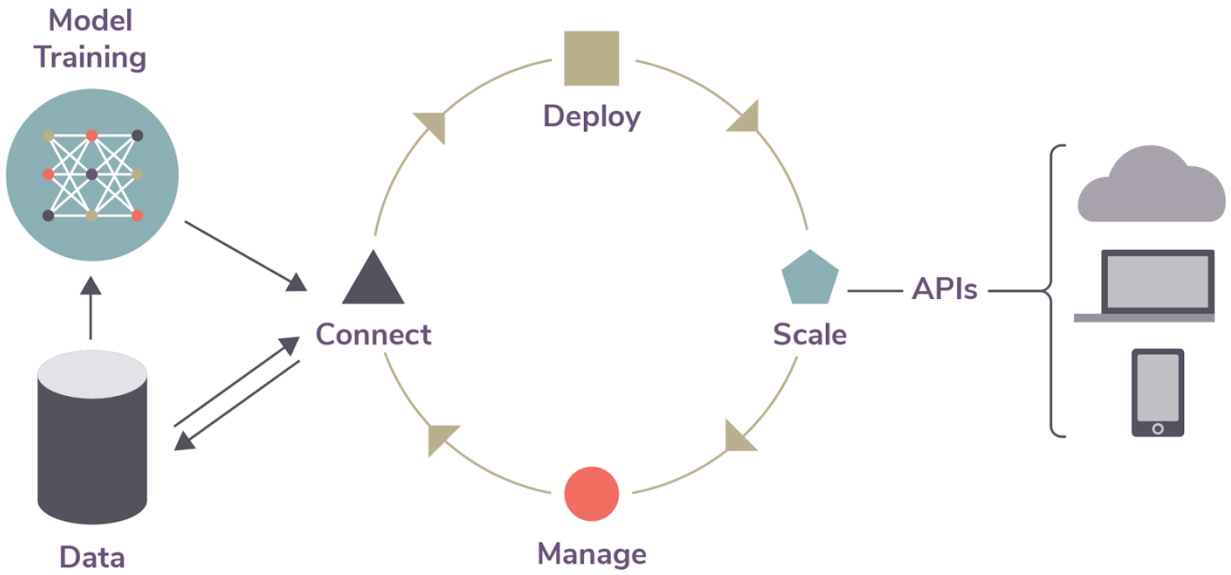
Composability



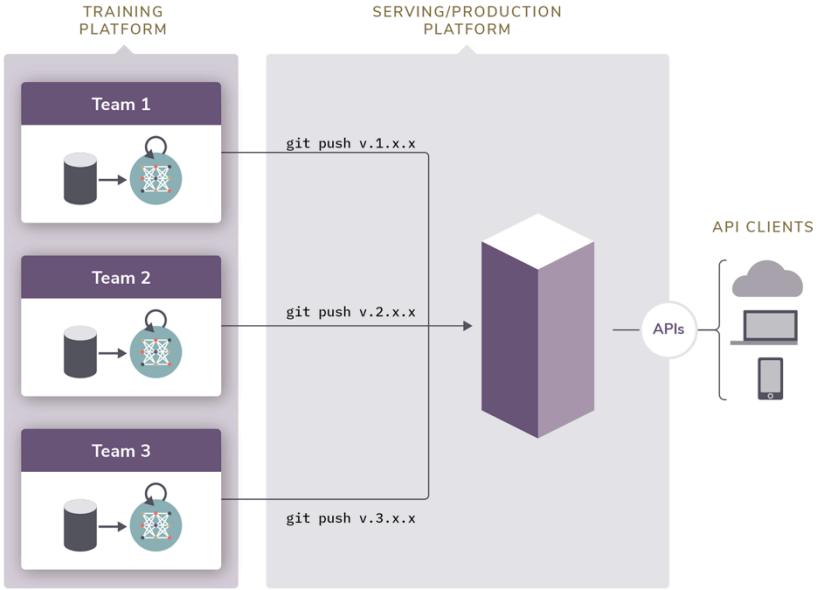
Performance: Fast, Reliable, and...

ACCURACY (TP+TN)/(TP+FP+FN+TN)	How many subjects did the model classify correctly?
SENSITIVITY/RECALL TP/(TP+FN)	How many of the actual positives did the model identify correctly?
SPECIFICITY TN/(TN+FP)	How many of the actual negatives did the model correctly identify?
PRECISION TP/(TP+FP)	How many identified as positives were actual positives?

Iteration



Infrastructure Demands: Training vs Production



Training

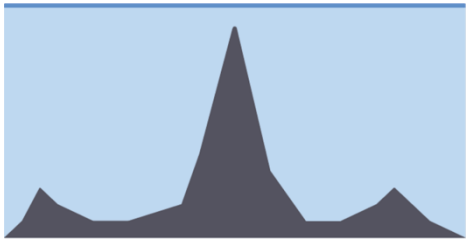
- Long compute cycle
- Fixed load
- Stateful
- Single user

Production

- Short compute bursts
- Elastic
- Stateless
- Many users

Deploying Models: Three Approaches

Traditional Capacity Planning

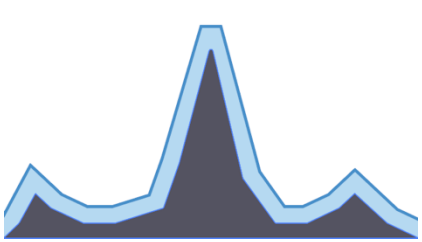


- Reserved capacity always available
- Extremely wasteful and expensive
- Hard limits: unanticipated demand can exceed capacity

Autoscaling

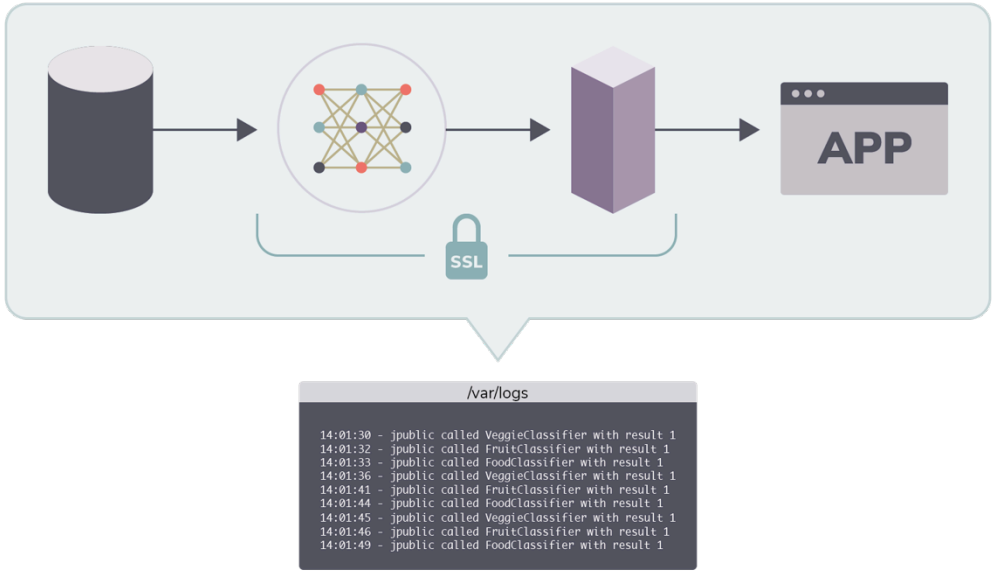
- Huge cost improvements
- Still inefficient hardware use
- Difficult to manage heterogeneous workloads on the same hardware

Serverless



- Significantly improved hardware use dramatically lowers costs
- Easier to run and scale different types of workloads on the same machines

Auditability & Governance



Think Enterprise

CATALOG

Models in a searchable, fully permissioned enterprise repository

GOVERN

Model usage with permissioning, auditing, and logging

INTEGRATE

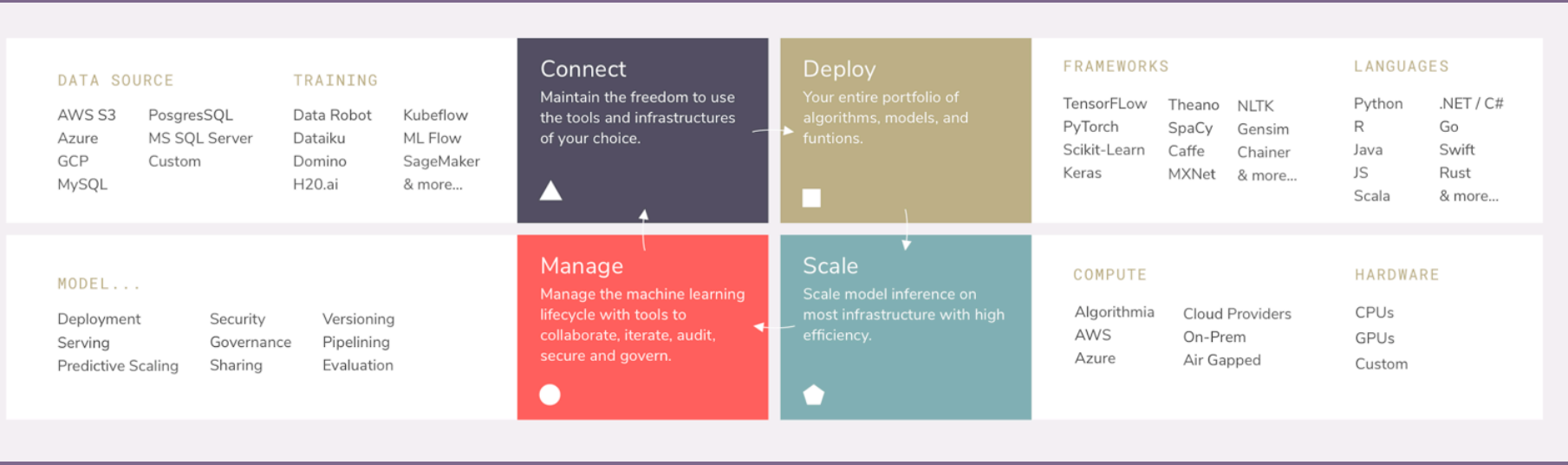
Existing tools and workflows via APIs, custom auth, and custom data sources

SECURE

Built-in security

Conclusion:

- **ML is different**: but you can still borrow from best practices in the SDLC
- **Build for repeatability:** ML iterates much faster than app development; make it a feature
- **Abstract:** infrastructure, components, data sources. Allow teams to focus on the value add
- **Profit:** Getting this right will make your data science and ML teams infinitely more productive, ultimately making your business bottom line better.


DATA SOURCE
AWS S3
Azure
GCP
MySQL
PosgresSQL
MS SQL Server
Custom
TRAINING
Data Robot
Dataiku
Domino
H20.ai
Kubeflow
ML Flow
SageMaker
& more...
Connect
Maintain the freedom to use the tools and infrastructures of your choice.
Deploy
Your entire portfolio of algorithms, models, and funtions.
FRAMEWORKS
TensorFLow
PyTorch
Scikit-Learn
Keras
Theano
SpaCy
Caffe
MXNet
NLTK
Gensim
Chainer
& more...
LANGUAGES
Python
R
Java
JS
Scala
.NET / C#
Go
Swift
Rust
& more...
MODEL...
Deployment
Serving
Predictive Scaling
Security
Governance
Sharing
Versioning
Pipelining
Evaluation
Manage
Manage the machine learning lifecycle with tools to collaborate, iterate, audit, secure and govern.
Scale
Scale model inference on most infrastructure with high efficiency.
COMPUTE
Algorithmia
AWS
Azure
Cloud Providers
On-Prem
Air Gapped
HARDWARE
CPUs
GPUs
Custom

We’re Hiring!

www.algorithmia.com/careers
www.algorithmia.com

Thank you

halvarez@algorithmia.io