Models in Minutes not Months: Data Science as Microservices

Sarah Aerni, PhD

Einstein Platform
saerni@salesforce.com

@itweetsarah

LIVE DEMO

Agenda

BUILDING AI APPS: Perspective Of A Data Scientist

- Journey to building your first model
- Barriers to production along the way

DEPLOYING MODELS IN PRODUCTION: Built For Reuse

- Where engineering and applications meet AI
- DevOps in Data Science – monitoring, alerting and iterating

AUTO MACHINE LEARNING: Machine Learning Pipelines as a Collection of Microservices

- Create reusable ML pipeline code for multiple applications customers
- Data Scientists focus on exploration, validation and adding new apps and models

ENABLING DATA SCIENCE

A DATA SCIENTISTS VIEW OF BUILDING MODELS

salesforce

Access and Explore Data

Engineer Features and Build Models

Interpret Model Results and Accuracy

A data scientist's view of the journey to building models

Access and Explore Data

Engineer Features and Build Models

Interpret Model Results and Accuracy

A data scientist's view of the journey to building models

DATA SCIENCE IS A TEAM EFFORT

Data Engineers: Access to data

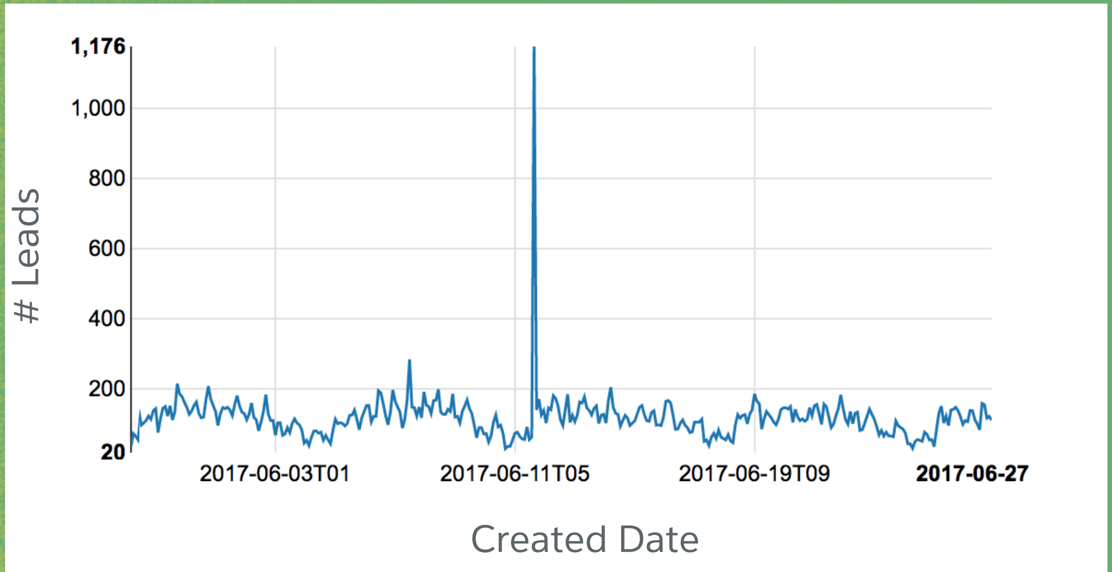
IT: Environment and tools

Domain Experts: Context and input at each step

Access and Explore Data

Engineer Features and Build Models

Interpret Model Results and Accuracy



Access and Explore Data

Engineer Features and Build Models

Interpret Model Results and Accuracy

Engineer Features

Empty fields

One-hot encoding (pivoting)

- Email domain of a user
- Business titles of a user

Historical spend

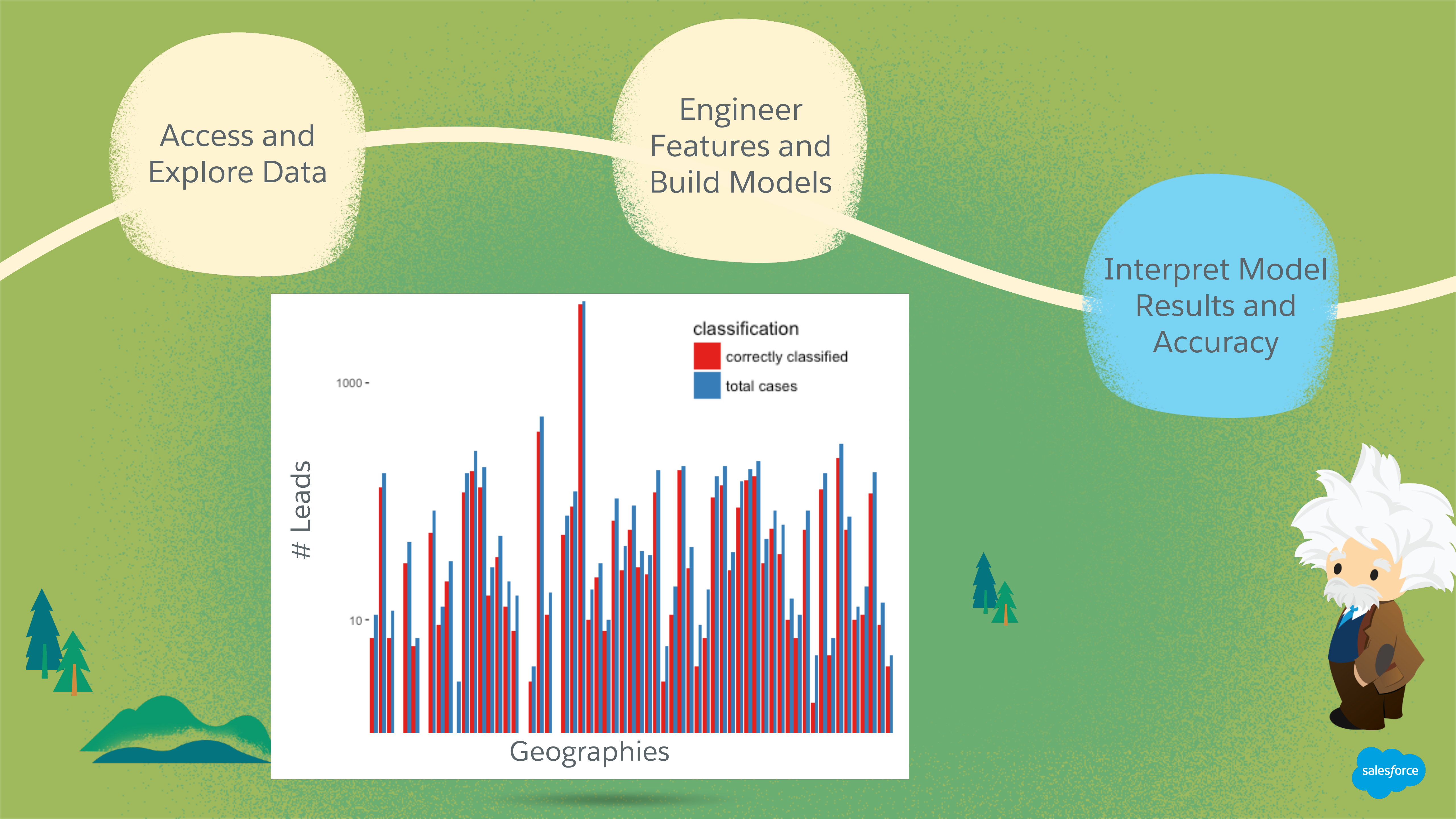
Email-Company Name Similarity

Access and Explore Data

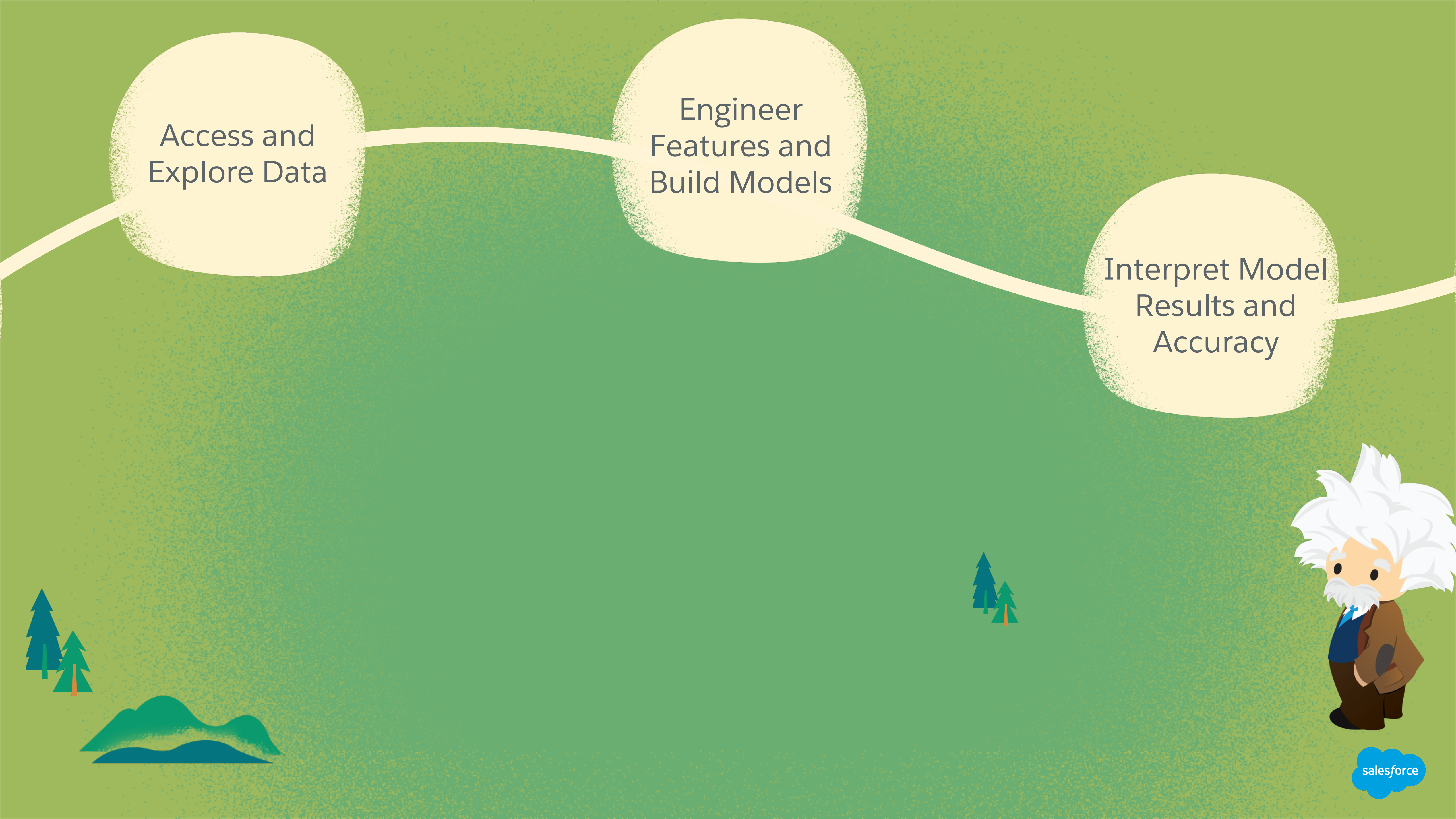
Engineer Features and Build Models

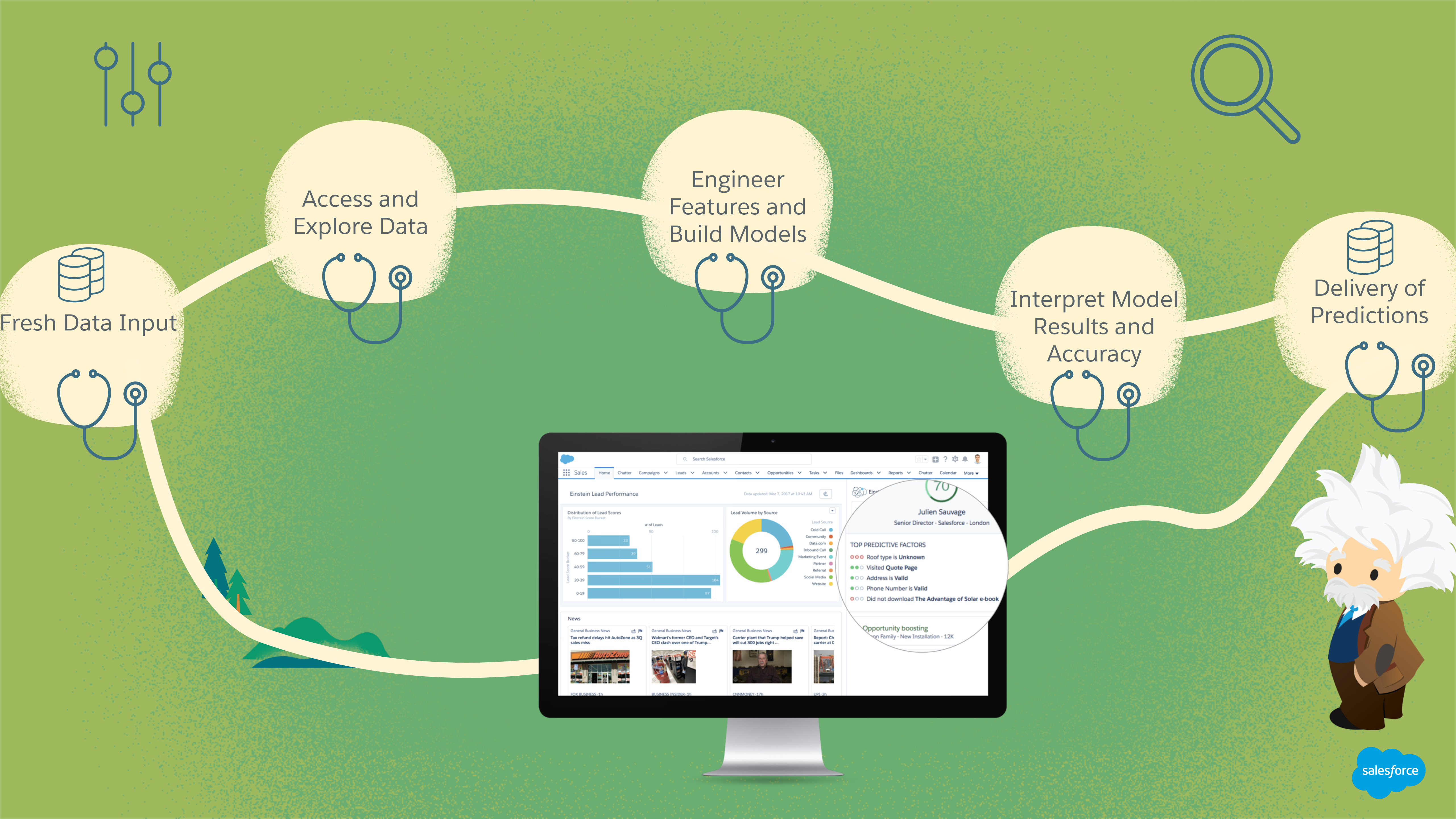
Interpret Model Results and Accuracy

```
>>> from sklearn import svm
>>> from numpy import loadtxt as l, random as r
>>> pls = numpy.loadtxt("leadFeatures.data", delimiter=",")
>>> testSet = r.choice(len(pls), int(len(pls)*.7), replace=False)
>>> X,  y = pls[-testSet,:-1], pls[-testSet:,-1]
>>> clf = svm.SVC()
>>> clf.fit(X,y)
SVC(C=1.0, cache_size=200, class_weight=None,
          coef0=0.0,decision_function_shape=None, degree=3,
          gamma='auto', kernel='rbf', max_iter=-1,
          tol=0.001, verbose=False)
>>> clf.score(pls[testSet,:-1],pls[testSet,-1])
 0.8857142857142857142568
```



Access and Explore Data
Engineer Features and Build Models
Interpret Model Results and Accuracy
classification
correctly classified
total cases
1000
10
Leads
Geographies
salesforce


Access and Explore Data
Engineer Features and Build Models
Interpret Model Results and Accuracy


Fresh Data Input
Access and Explore Data
Engineer Features and Build Models
Interpret Model Results and Accuracy
Delivery of Predictions
Einstein Lead Performance
Distribution of Lead Scores
Lead Volume by Source
299
Julien Sauvage
Senior Director - Salesforce - London
TOP PREDICTIVE FACTORS
Roof type is Unknown
Visited Quote Page
Address is Valid
Phone Number is Valid
Did not download The Advantage of Solar e-book
Opportunity boosting
News
salesforce

Bringing a Model to Production Requires a Team

Applications deliver predictions for customer consumption

Predictions are produced by the models live in production

Pipelines deliver the data for modeling and scoring at an appropriate latency

Monitoring systems allow us to check the health of the models, data, pipelines and app

salesforce

Source: Salesforce Customer Relationship Survey conducted 2014-2016 among 10,500+ customers randomly selected. Response sizes per question vary.

Bringing a Model to Production Requires a Team

Data Scientists

Continue evaluating models

Monitor for anomalies and degradation

Iteratively improve models in production

Data Engineers

Provide data access and management capabilities for data scientists

Set up and monitor data pipelines

Improve performance of data processing pipelines

Front-End Developers

Build customer-facing UI

Application instrumentation and logging

Product Managers

Gather requirements & feedback

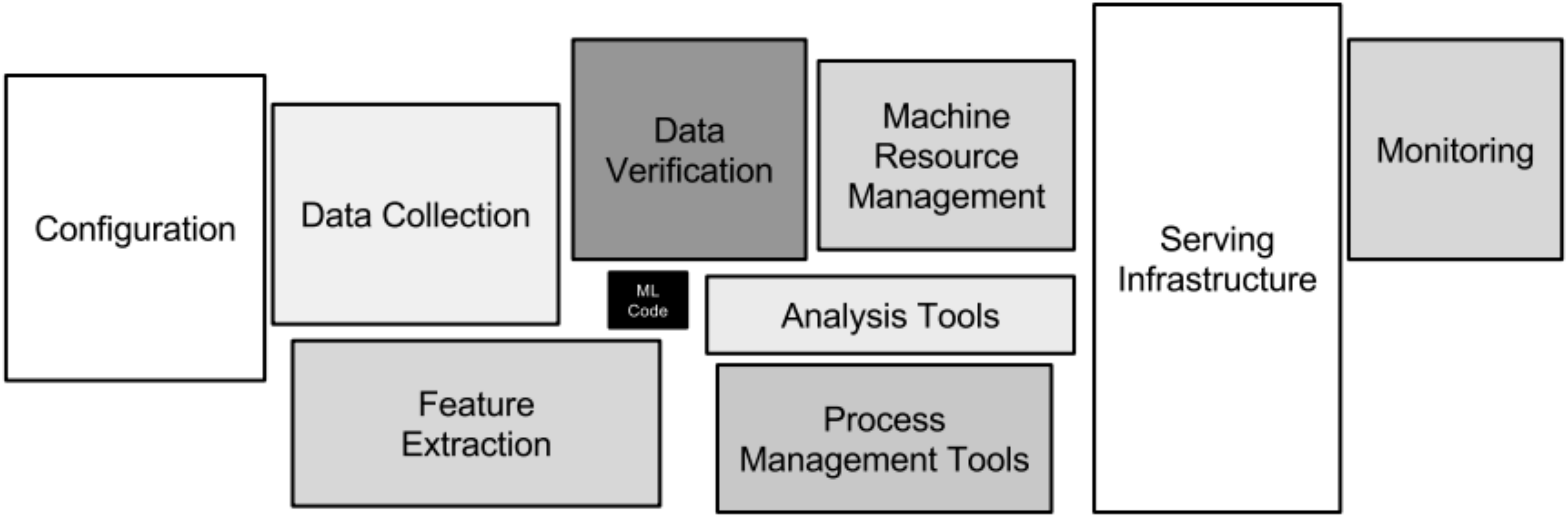
Provide business context

Platform Engineers

Machine resource management

Alerting and monitoring

Supporting a Model in Production is Complex



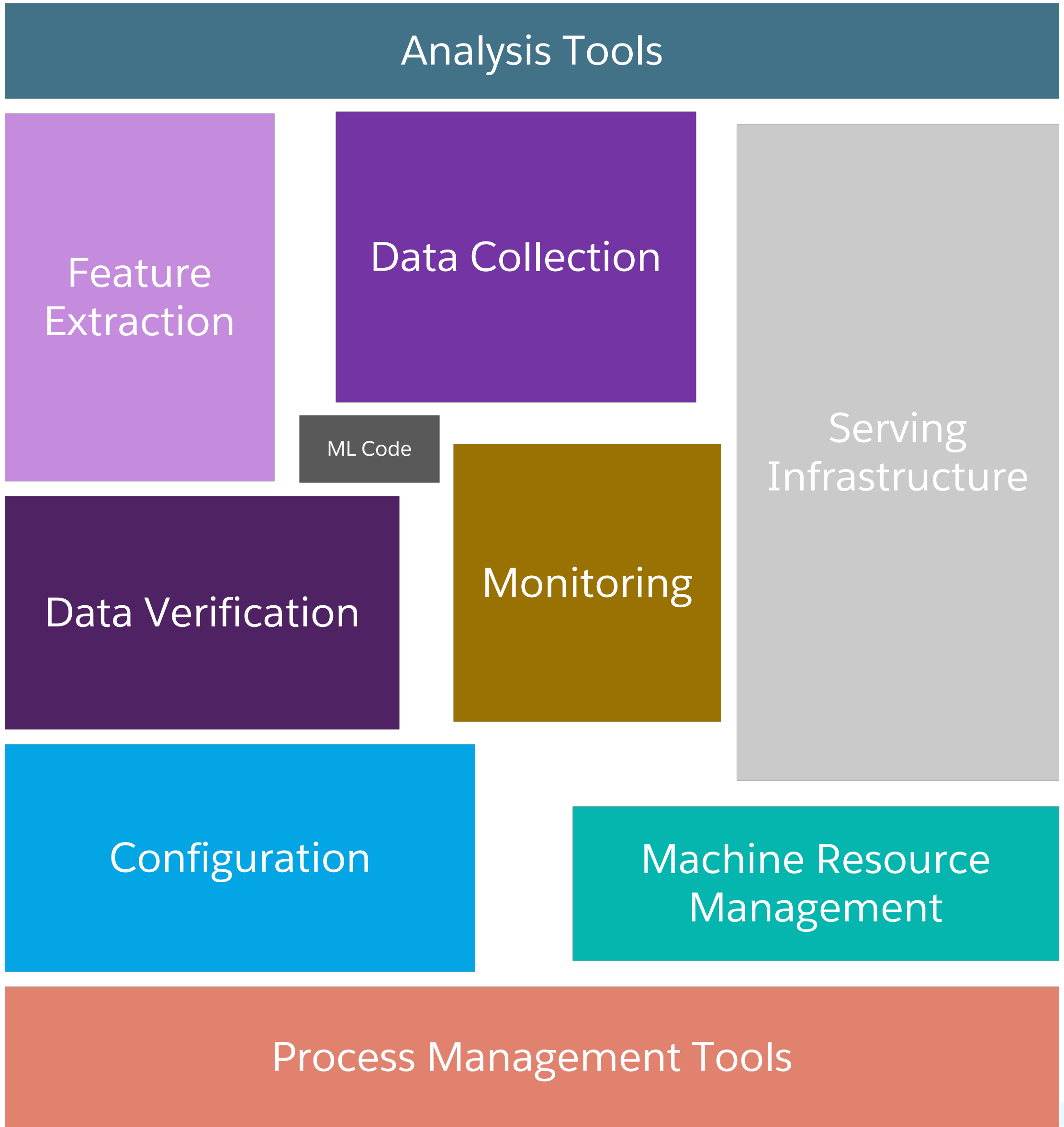
Only a small fraction of real-world ML systems is a composed of ML code, as shown by the small black box in the middle. The required surrounding infrastructure is fast and complex.

D. Sculley, et al. Hidden technical debt in machine learning systems. In Neural Information Processing Systems (NIPS). 2015

MODELS IN PRODUDCTION

WHAT IT TAKES TO DEPLOY AN AI-POWERED APPLICATION

Supporting Models in Production is Mostly NOT AI

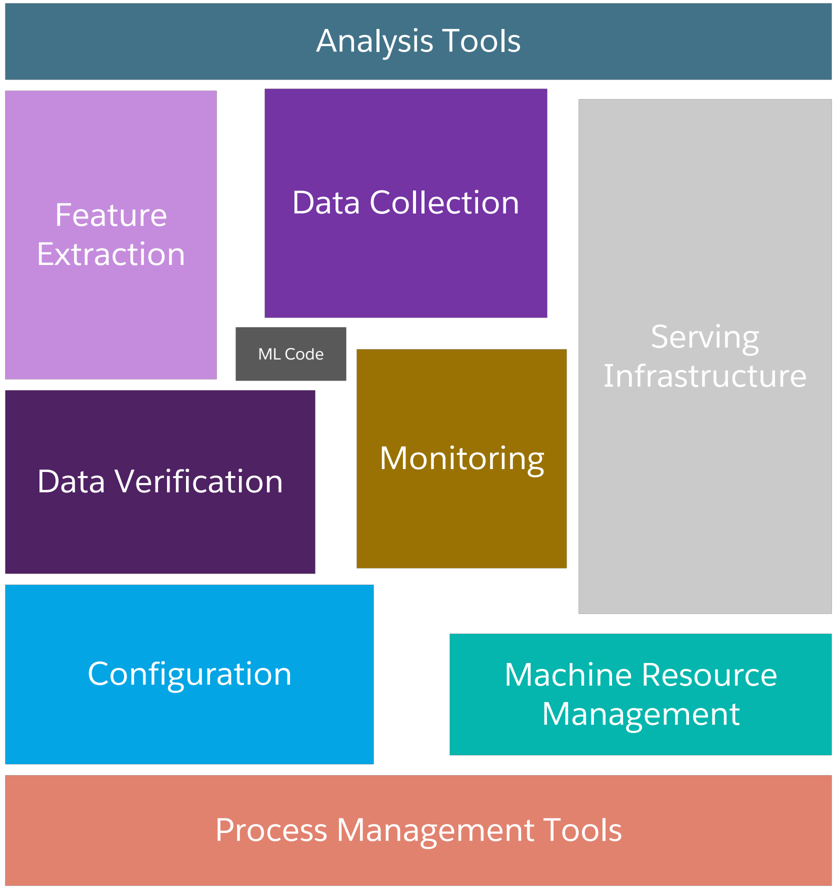


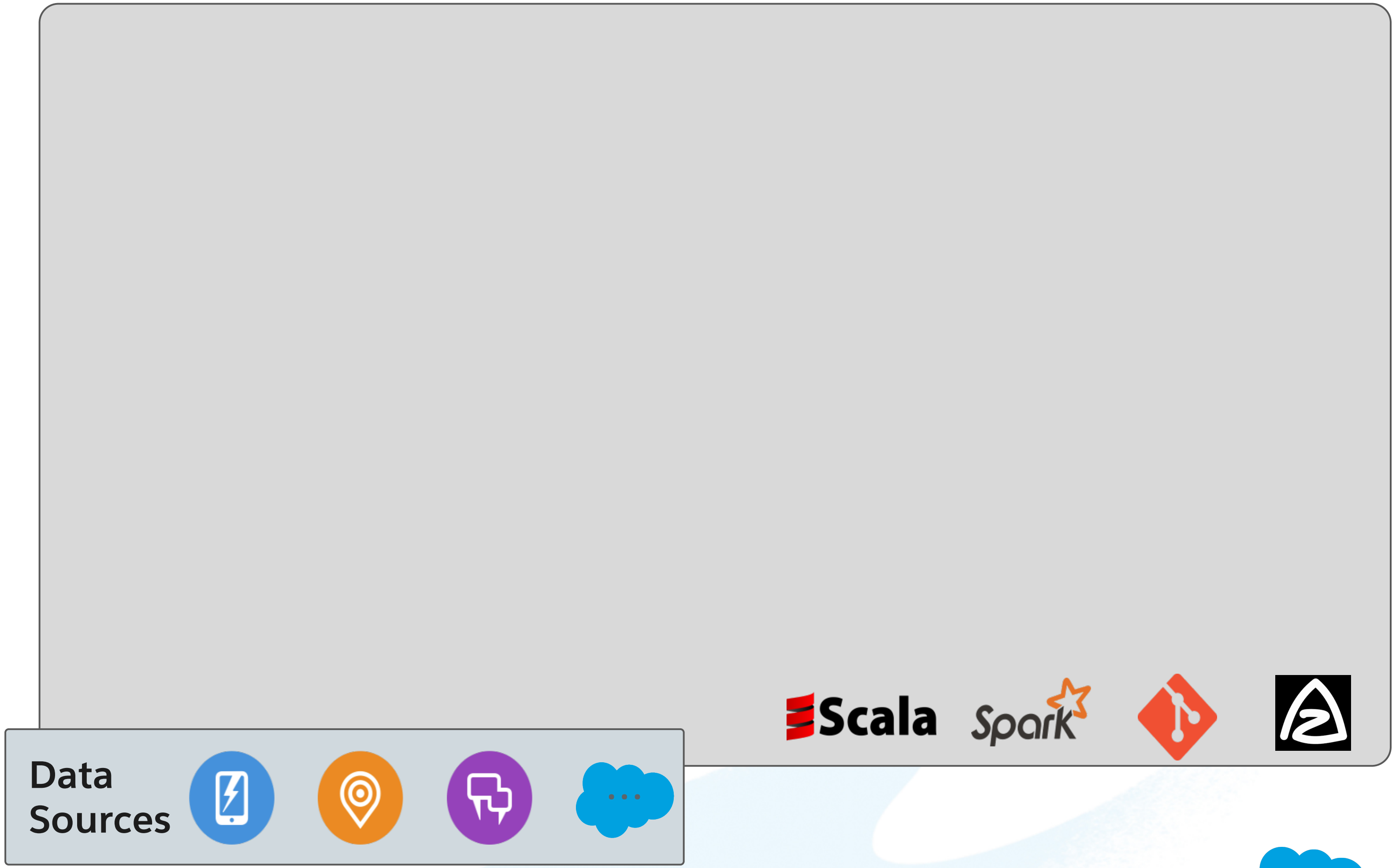
Only a small fraction of real-world ML systems is a composed of ML code, as shown by the small black box in the middle. The required surrounding infrastructure is fast and complex.

Adapted from D. Sculley, et al. Hidden technical debt in machine learning systems. In Neural Information Processing Systems (NIPS). 2015

How the Salesforce Einstein Platform Enables Data Scientists

Deploy, monitor and iterate on models in one location





salesforce

How the Salesforce Einstein Platform Enables Data Scientists

Deploy, monitor and iterate on models in one location

Analysis Tools

Feature Extraction

Data Collection

Serving Infrastructure

ML Code

Data Verification

Monitoring

Configuration

Machine Resource Management

Process Management Tools

CLI

Web UI

Scala

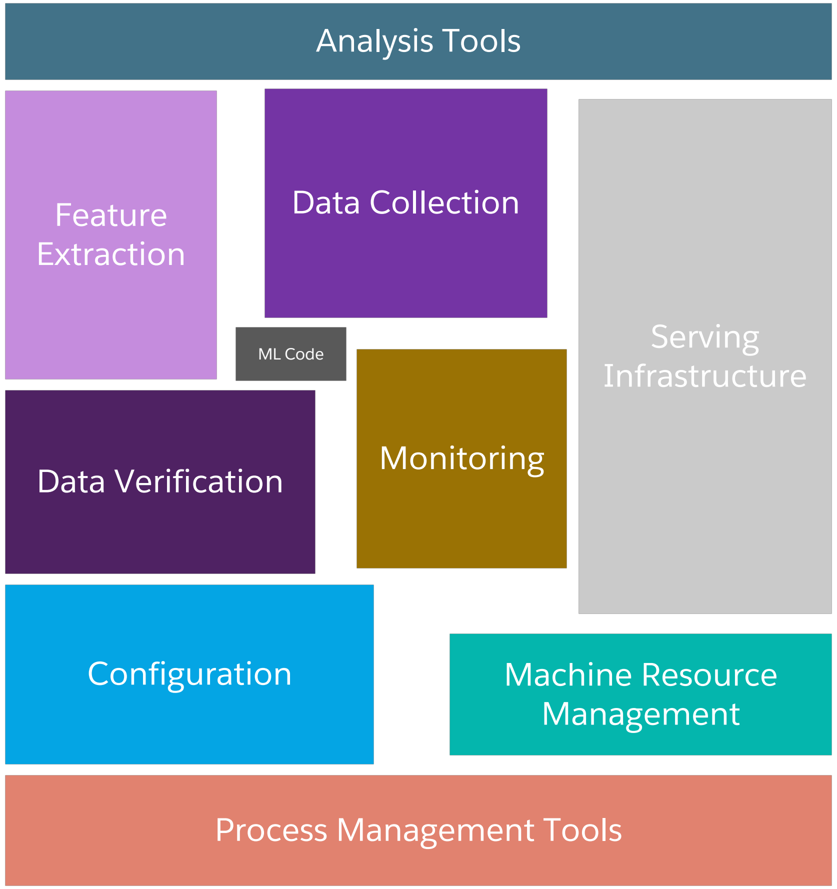
Spark

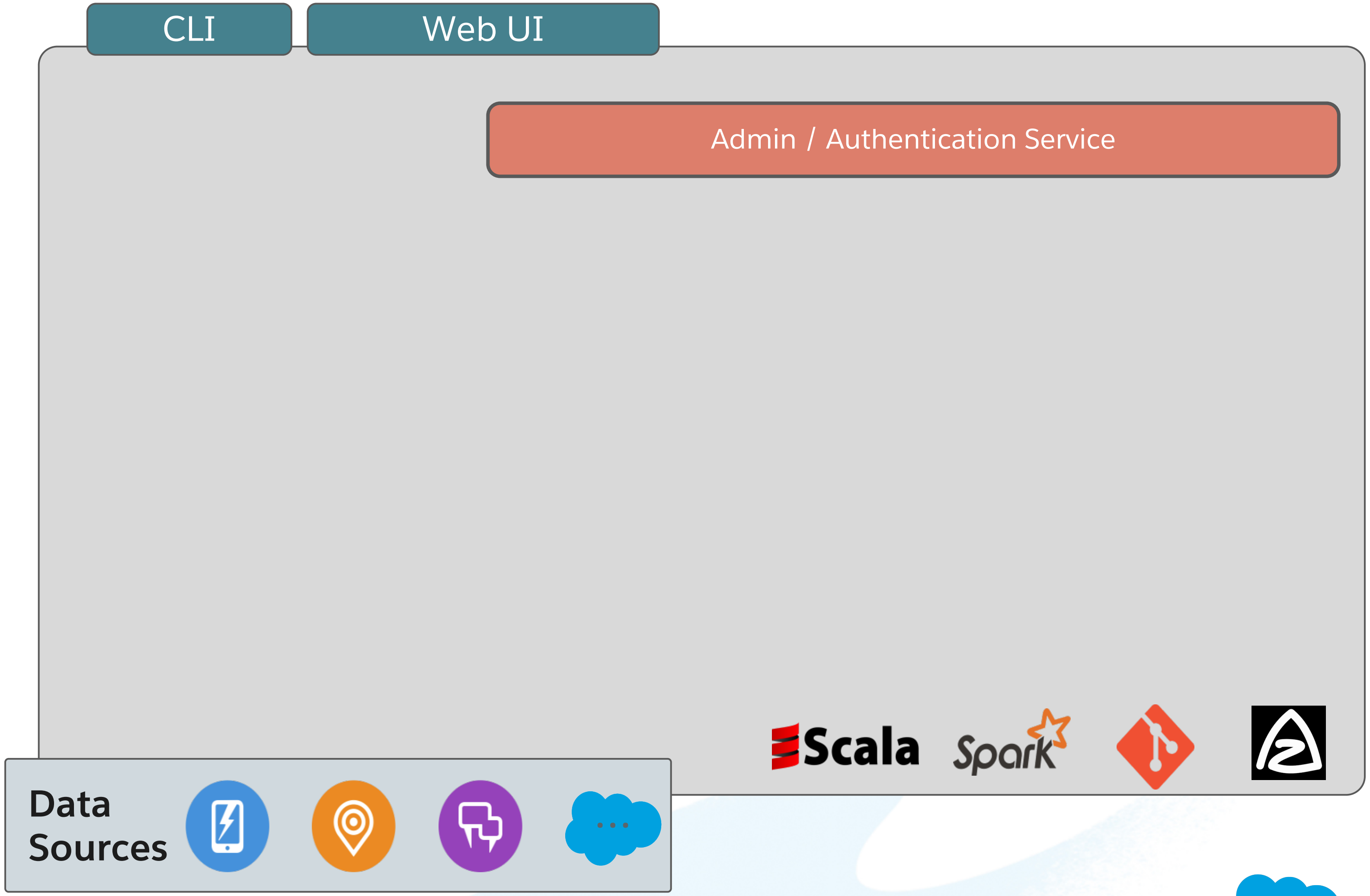
Data Sources

salesforce

How the Salesforce Einstein Platform Enables Data Scientists

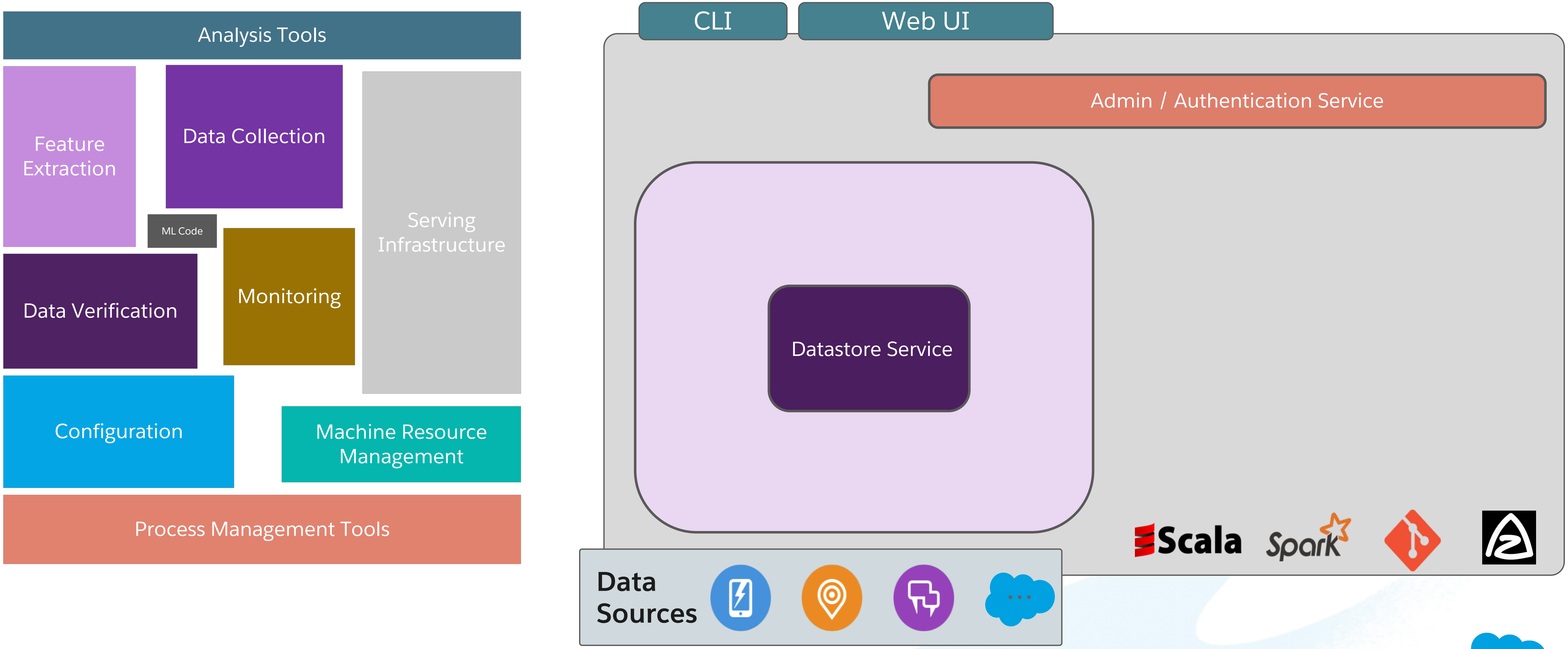
Deploy, monitor and iterate on models in one location



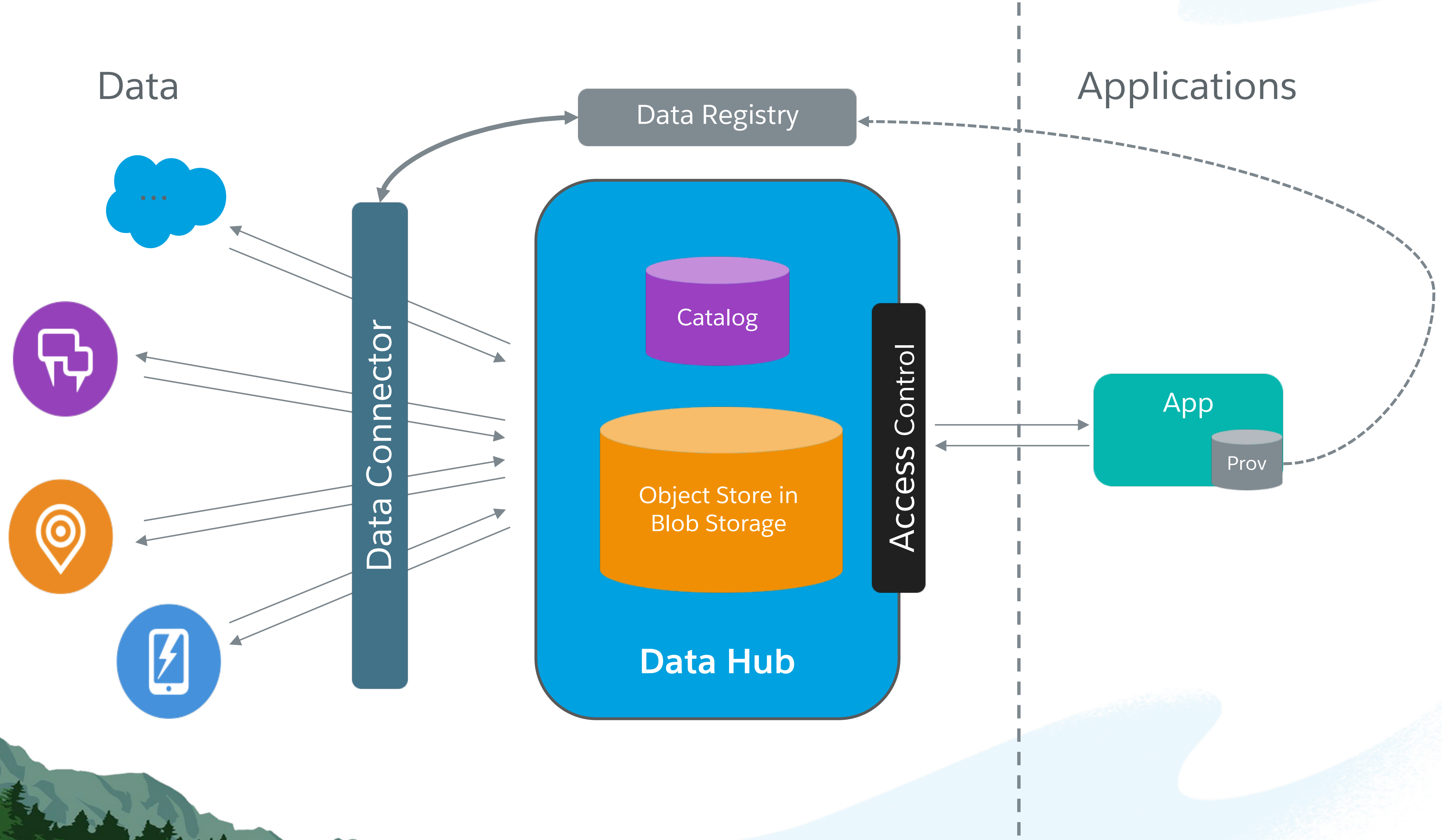


How the Salesforce Einstein Platform Enables Data Scientists

Deploy, monitor and iterate on models in one location

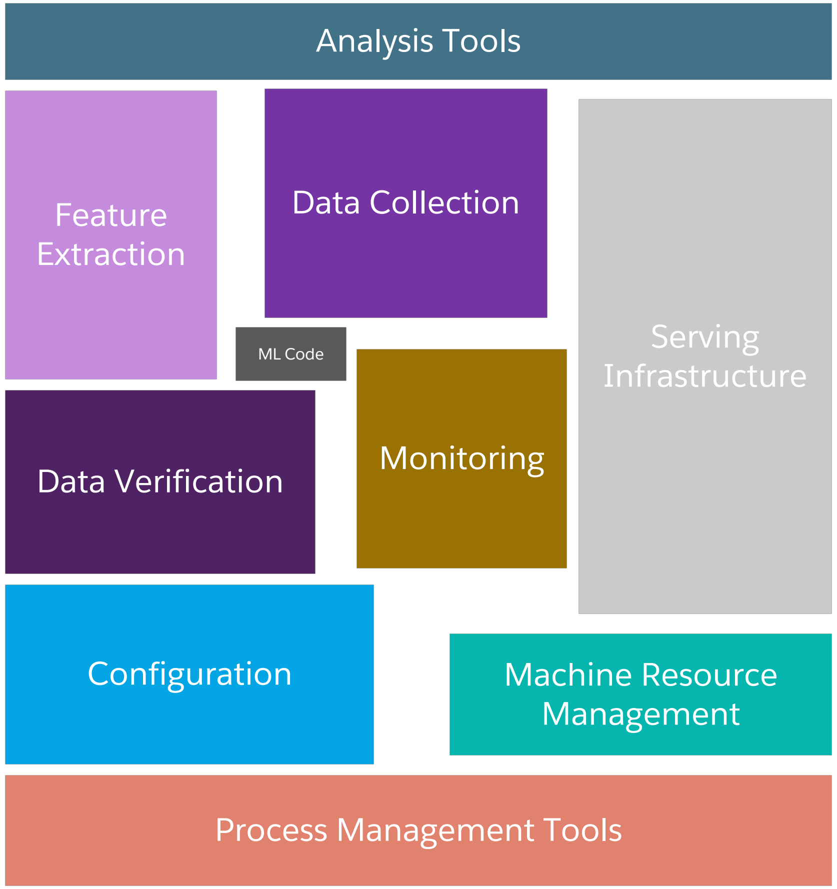


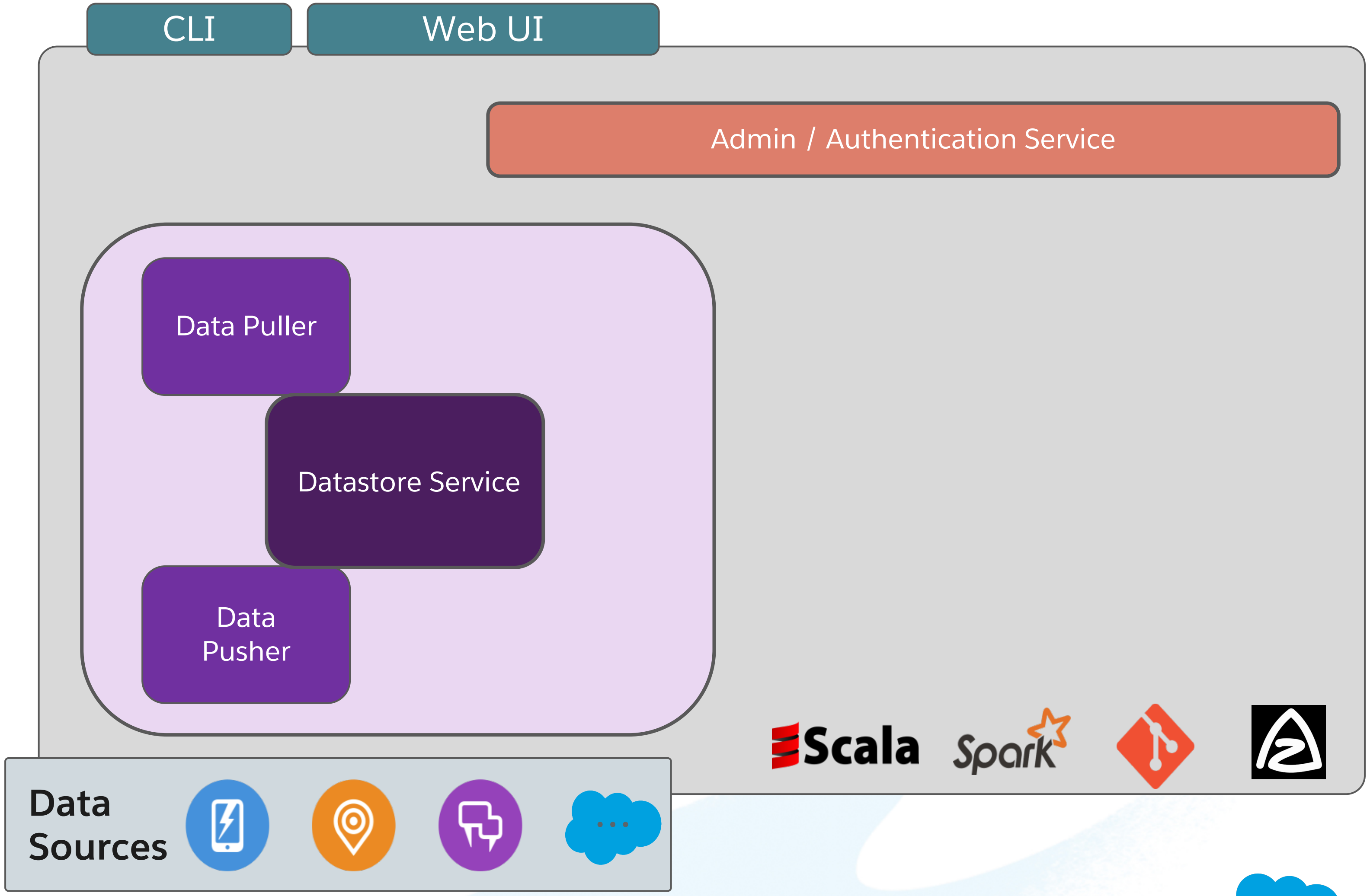
Why Data Services are Critical



How the Salesforce Einstein Platform Enables Data Scientists

Deploy, monitor and iterate on models in one location

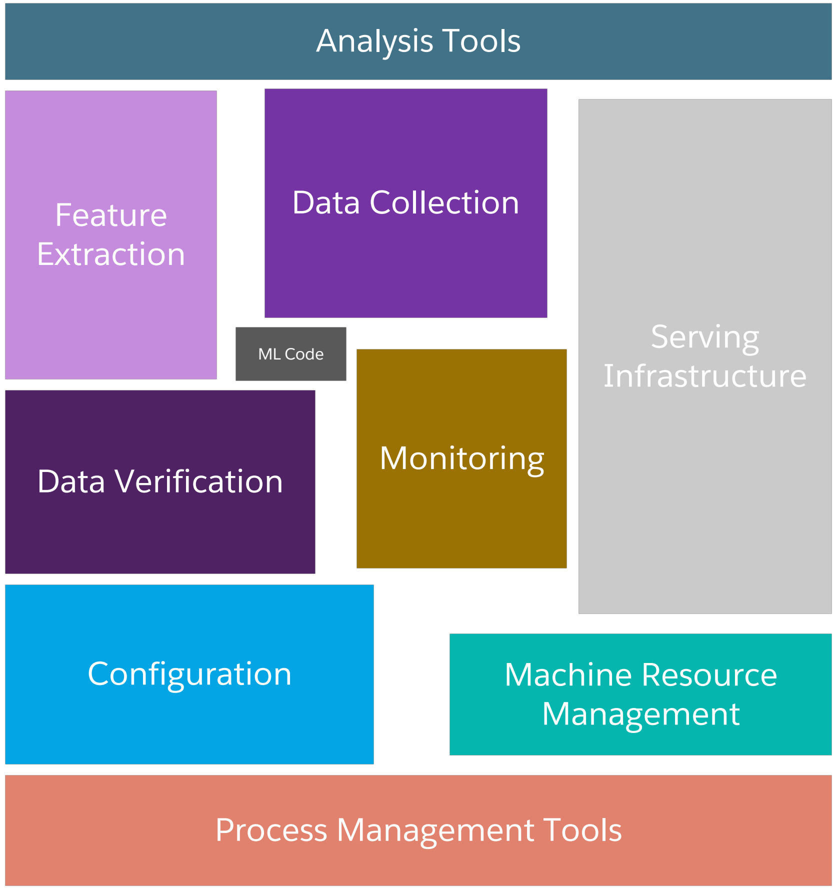


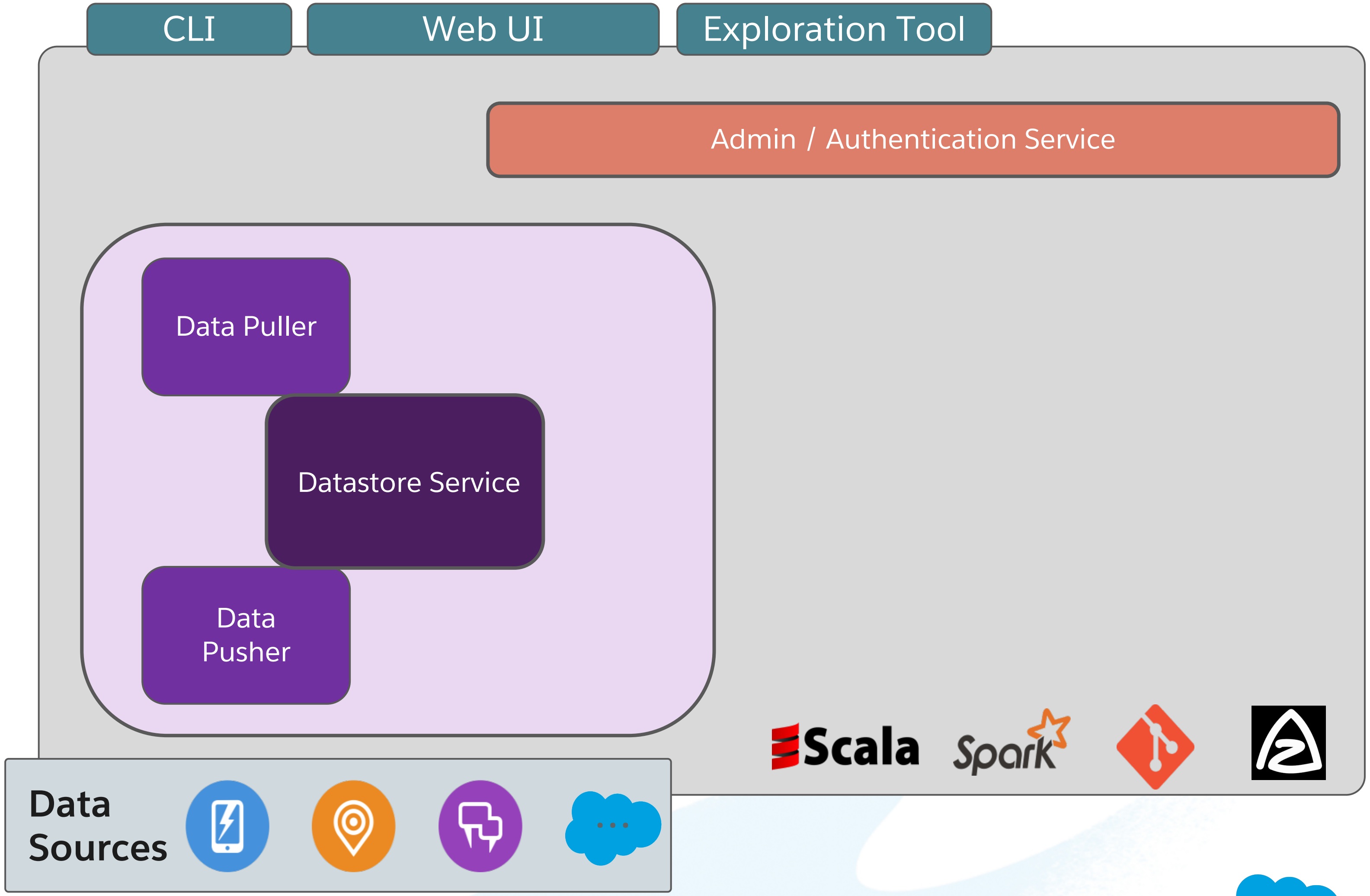


salesforce

How the Salesforce Einstein Platform Enables Data Scientists

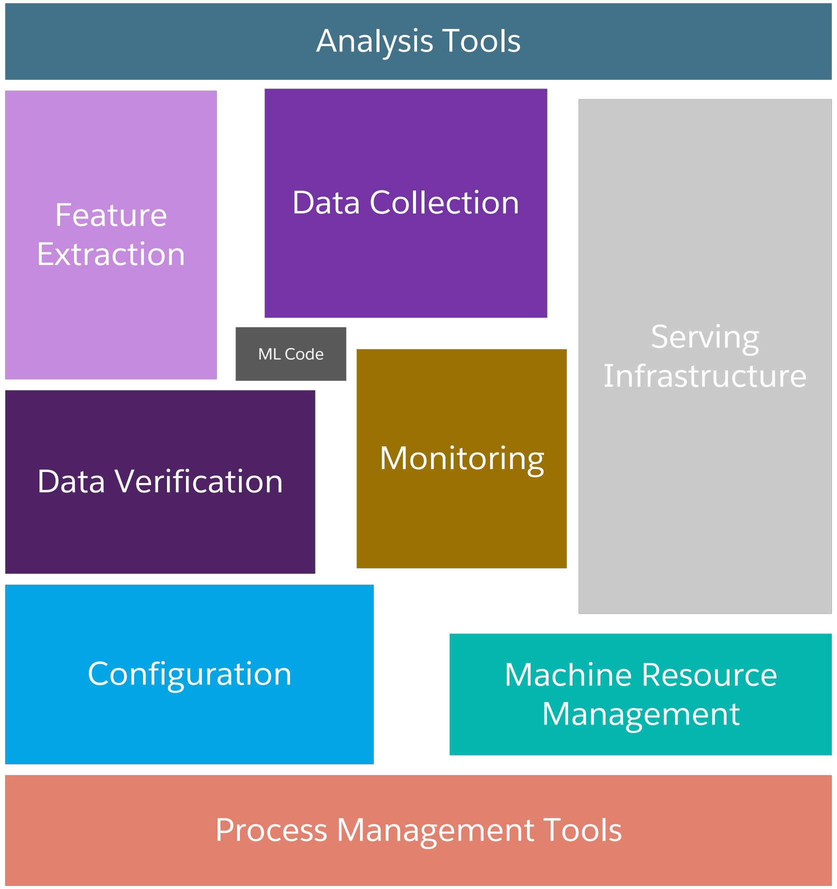
Deploy, monitor and iterate on models in one location

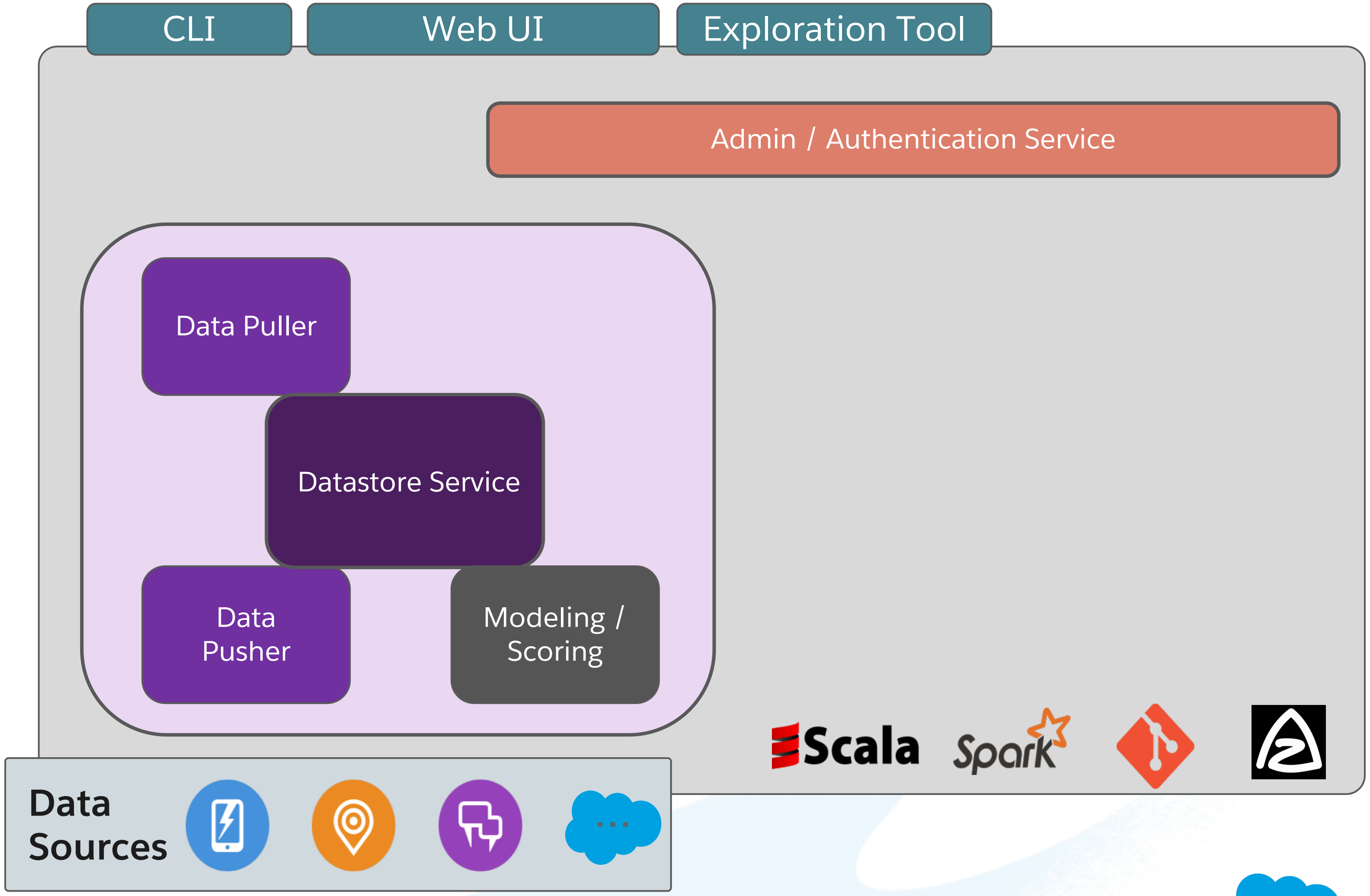




How the Salesforce Einstein Platform Enables Data Scientists

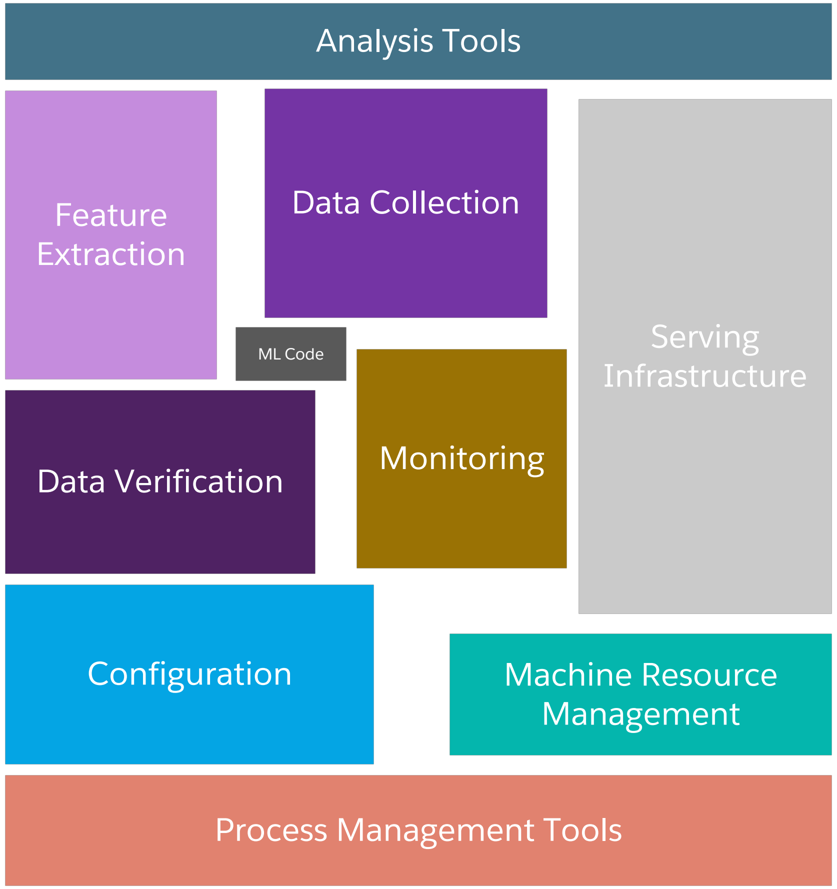
Deploy, monitor and iterate on models in one location

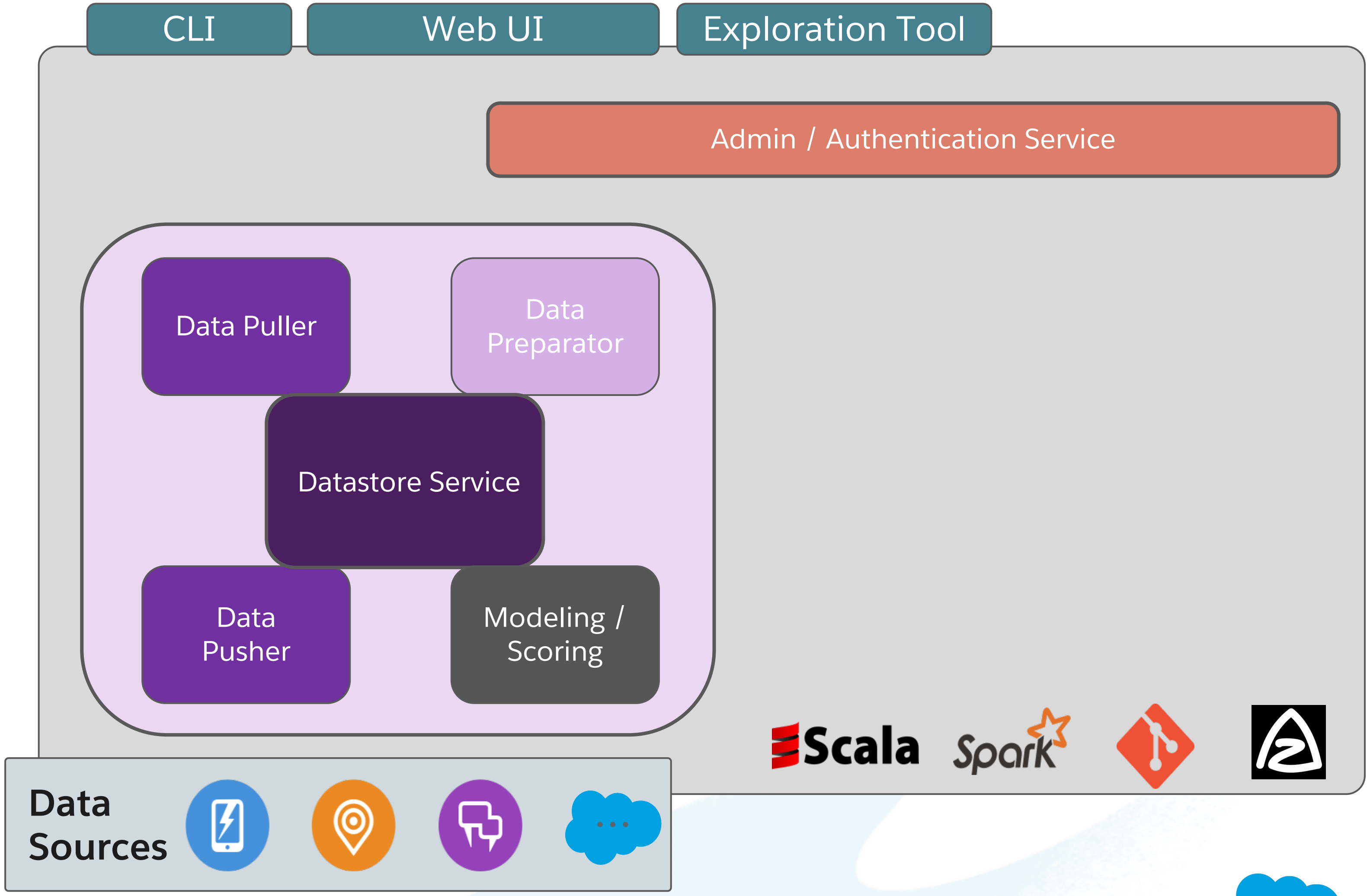




How the Salesforce Einstein Platform Enables Data Scientists

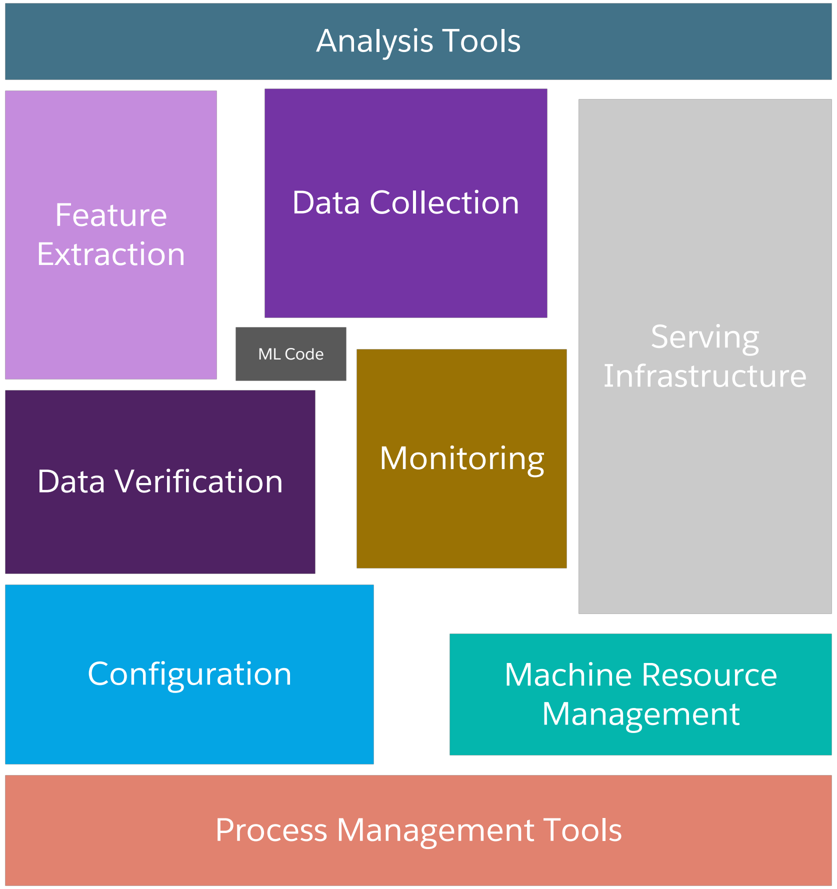
Deploy, monitor and iterate on models in one location

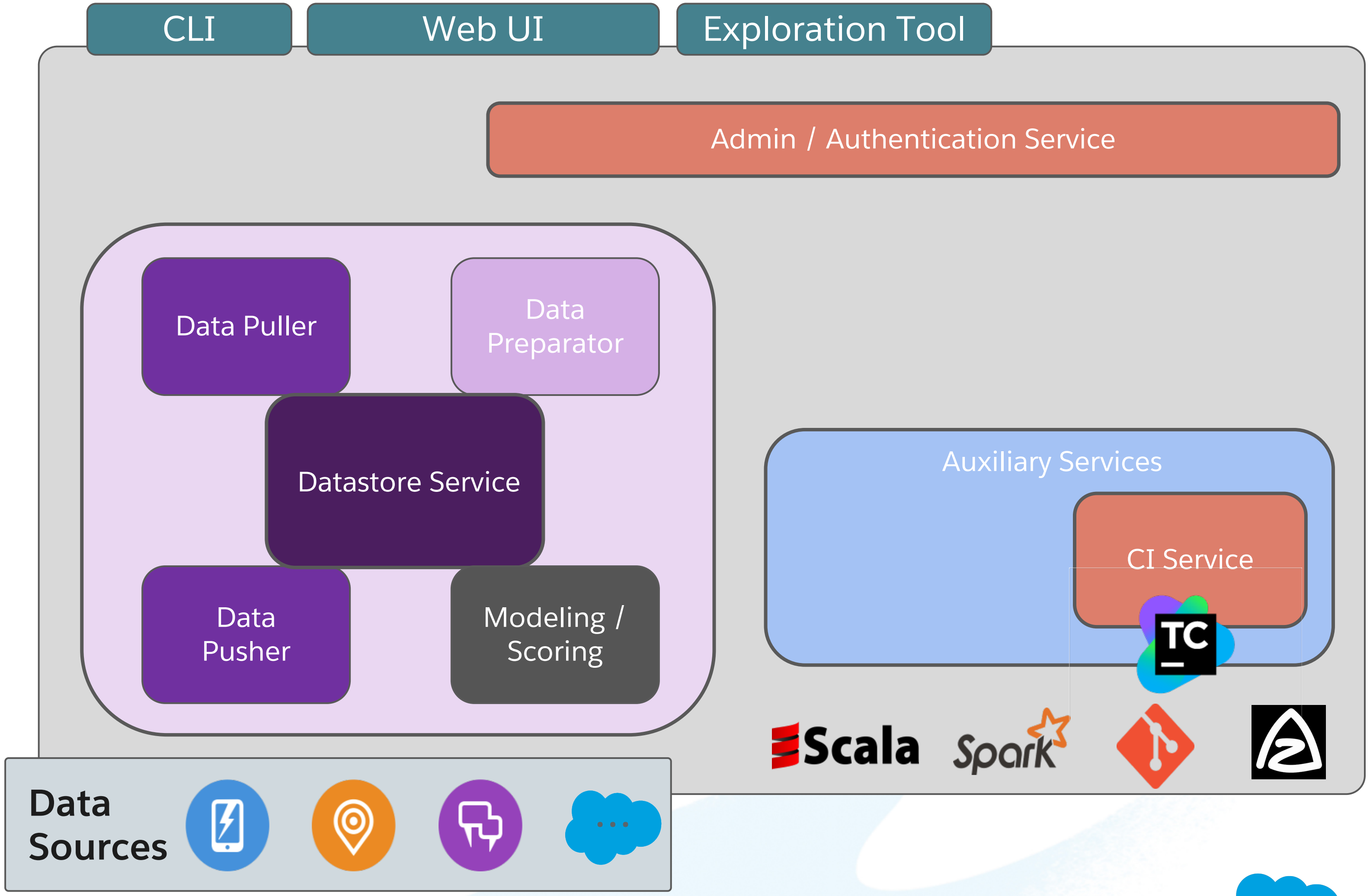




How the Salesforce Einstein Platform Enables Data Scientists

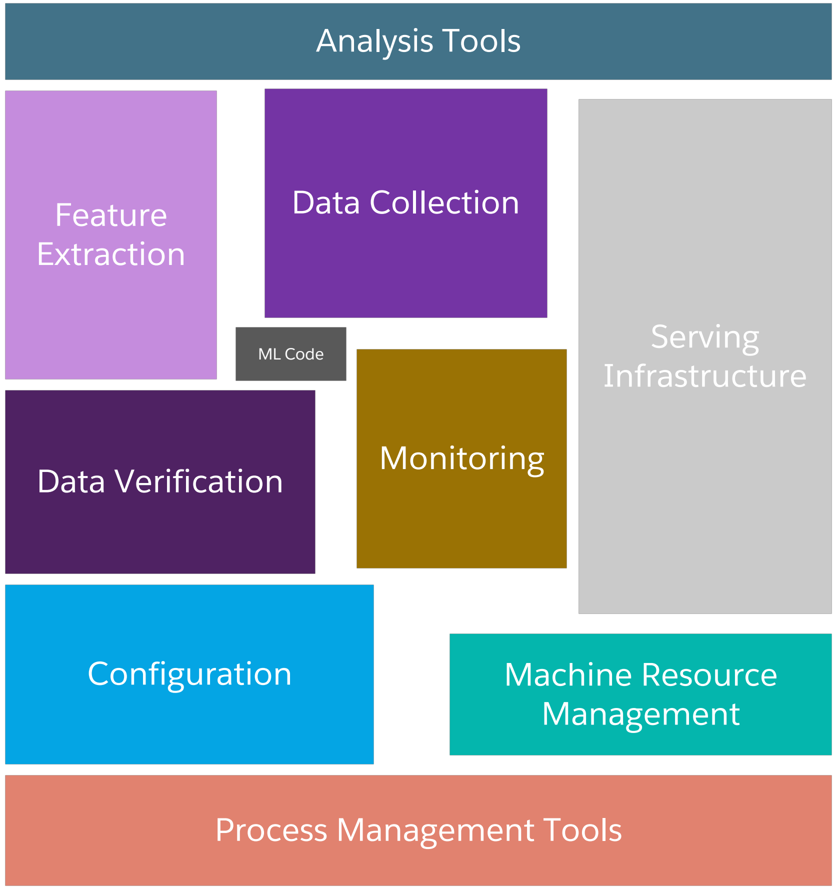
Deploy, monitor and iterate on models in one location

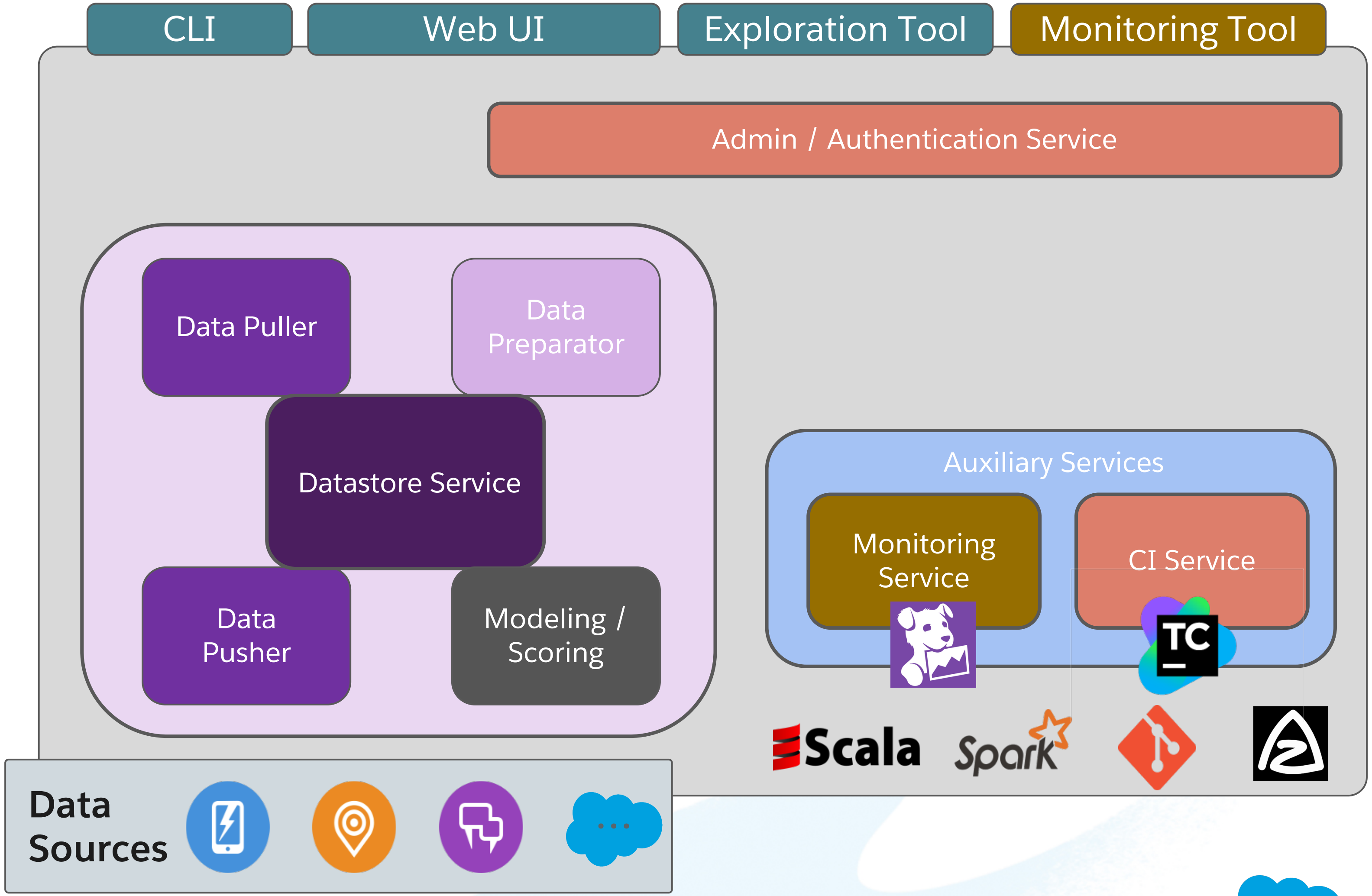




How the Salesforce Einstein Platform Enables Data Scientists

Deploy, monitor and iterate on models in one location

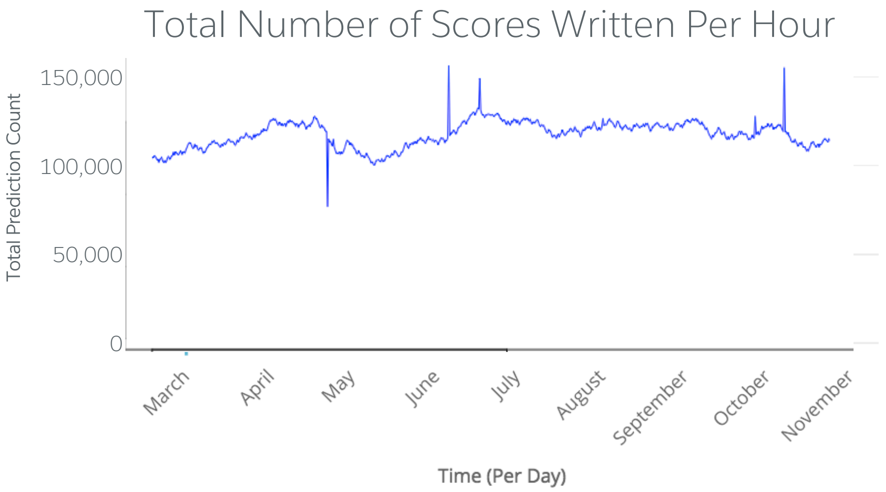


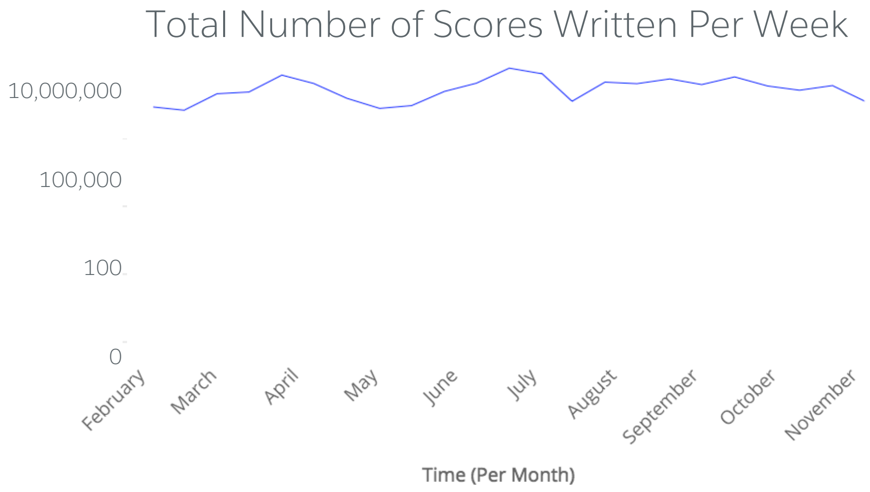


Monitoring your AI's health like any other app

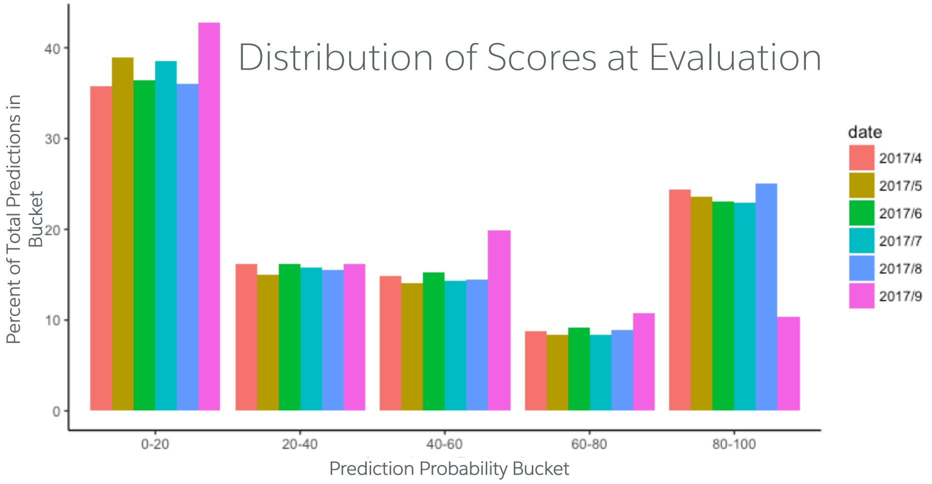
Pipelines, Model Performance, Scores – Invest your time where it is needed!

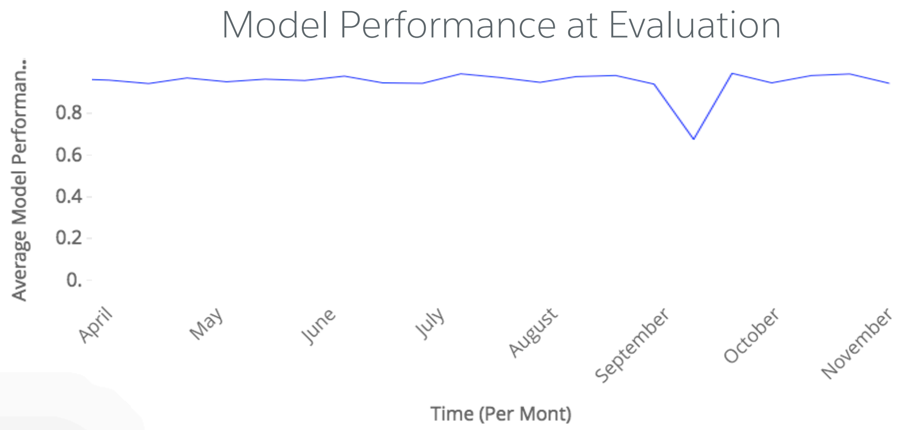
105,874
Scores Written Per Hour(1 day moving avg)





0.86
Evaluation auROC

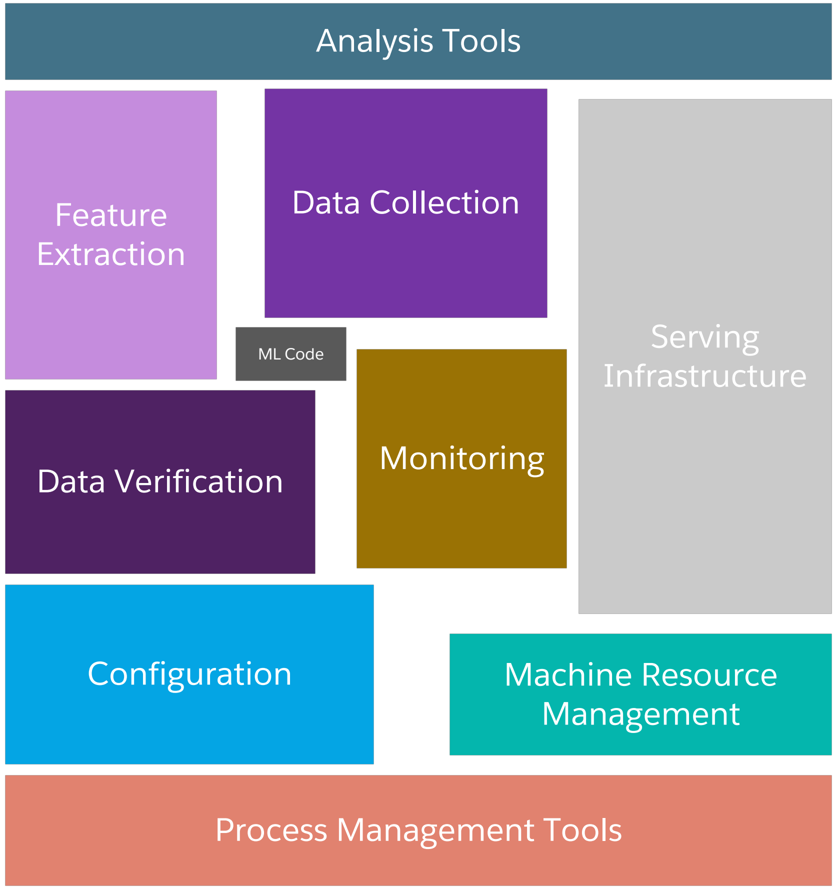


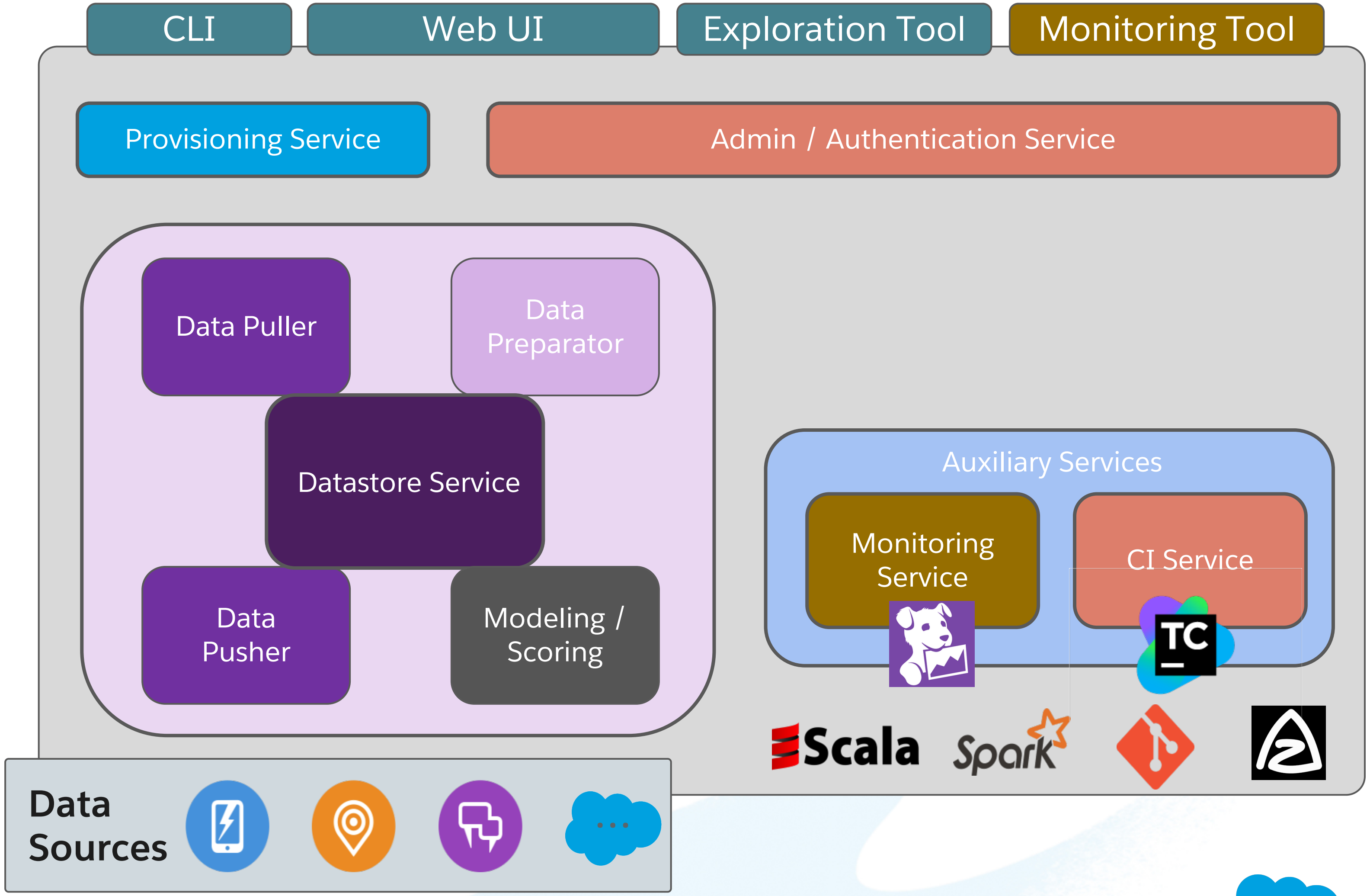


Sample Dashboard on Simulated Data

How the Salesforce Einstein Platform Enables Data Scientists

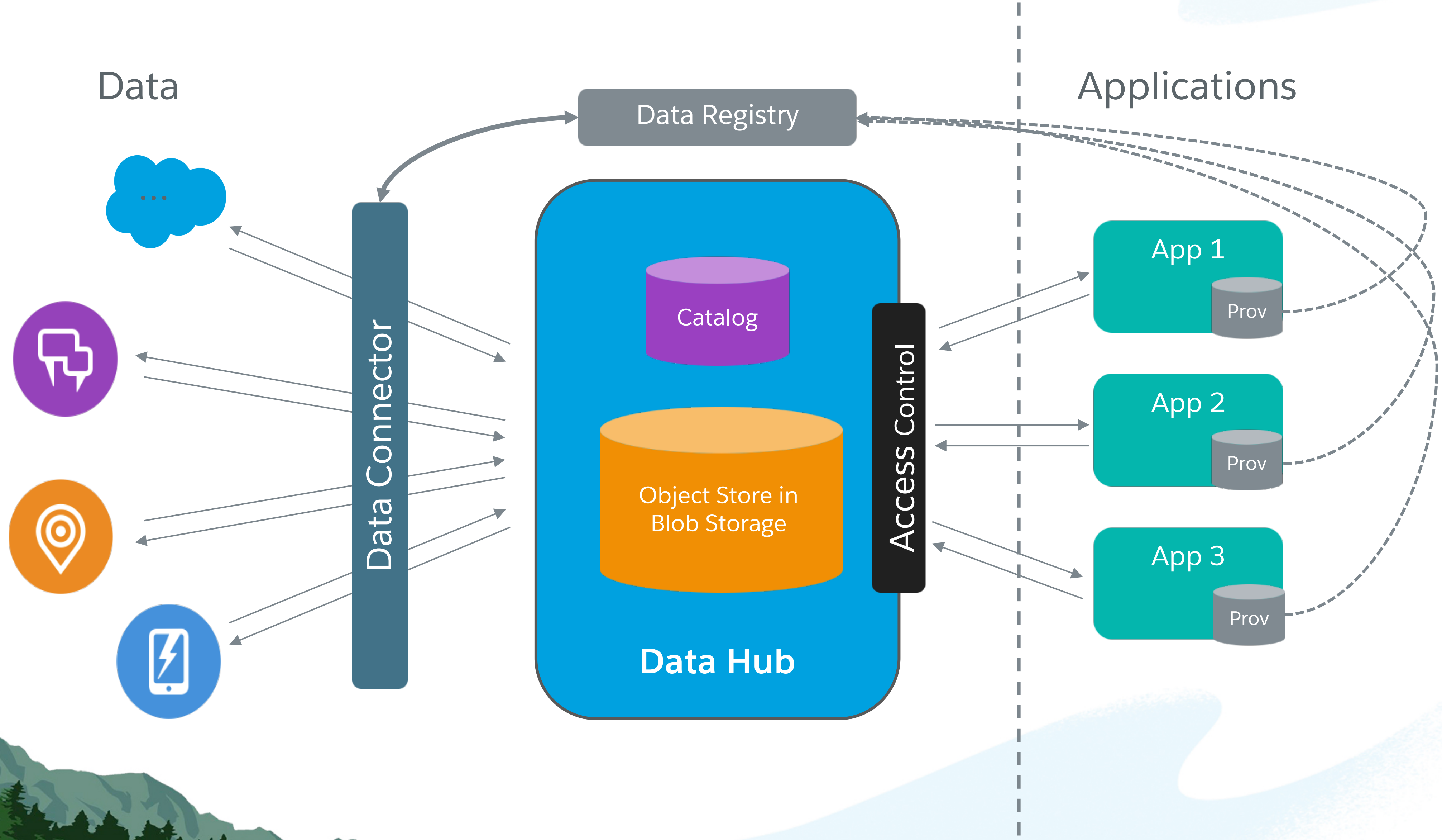
Deploy, monitor and iterate on models in one location





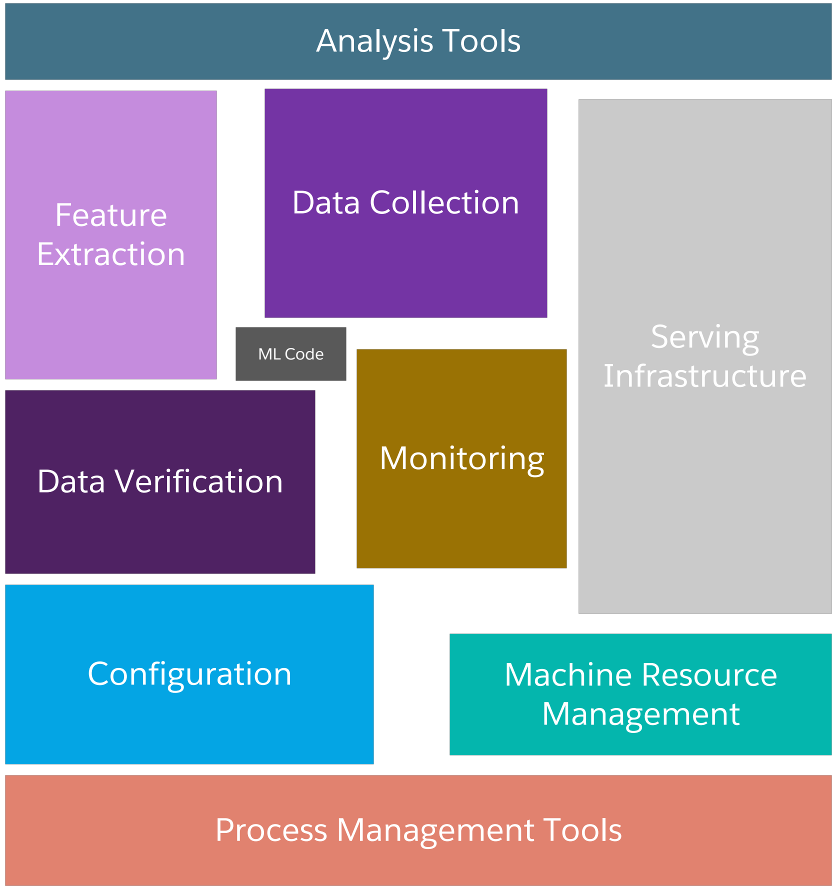
salesforce

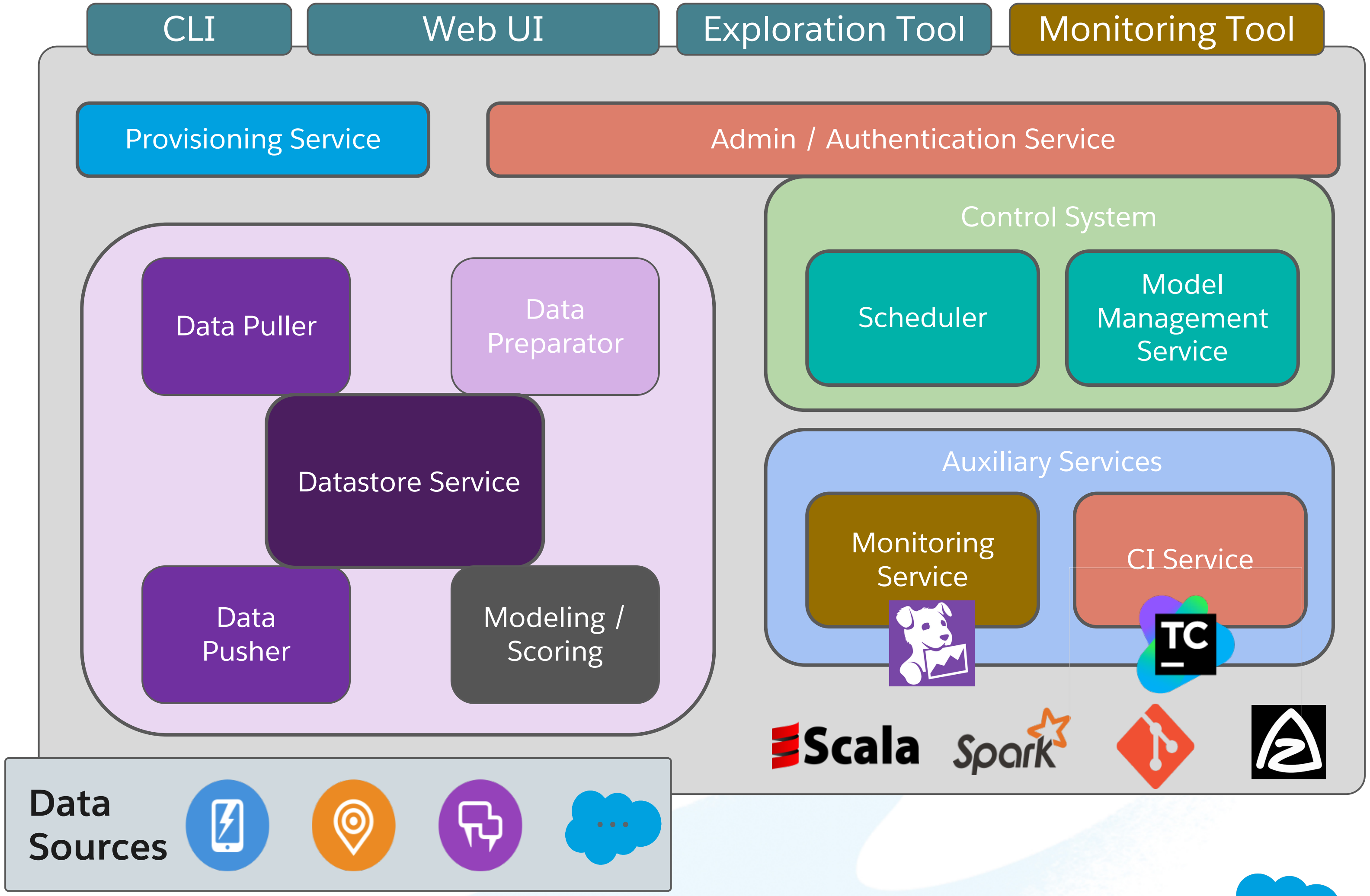
Why Data Services are Critical



How the Salesforce Einstein Platform Enables Data Scientists

Deploy, monitor and iterate on models in one location





How the Salesforce Einstein Platform Enables Data Scientists

Deploy, monitor and iterate on models in one location

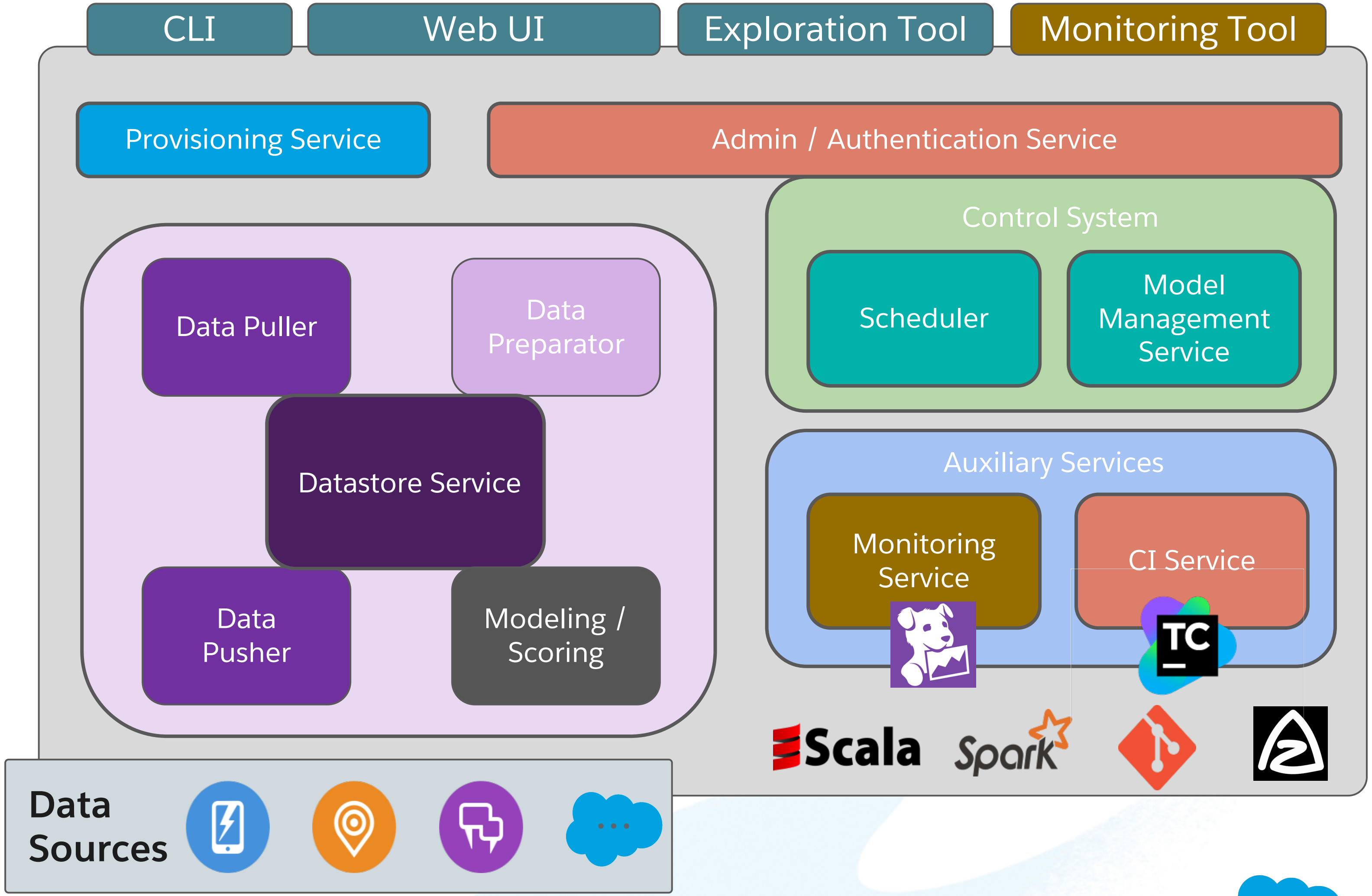
Microservice architecture

Customizable model-evaluation & monitoring dashboards

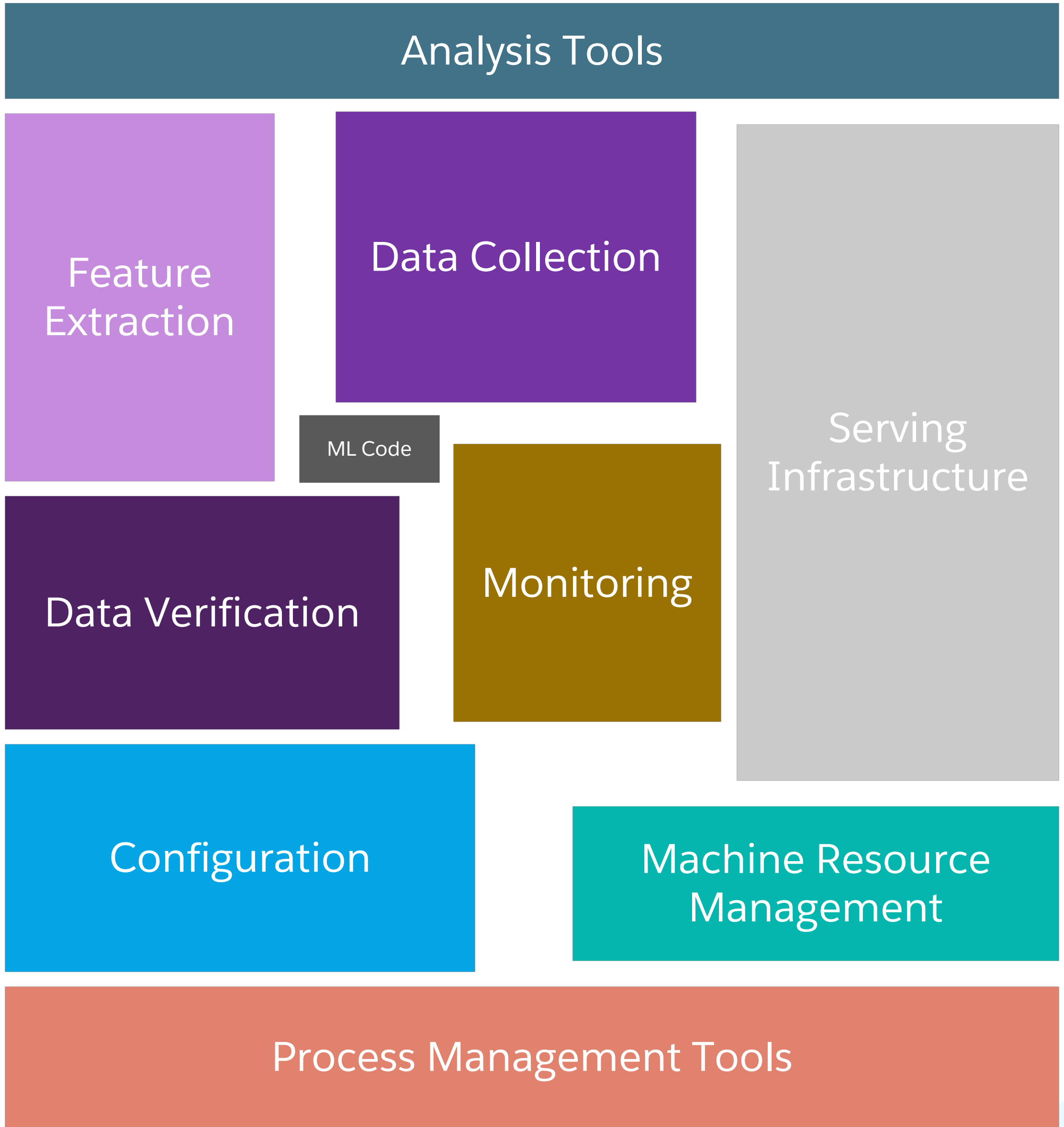
Scheduling and workflow management

In-platform secured experimentation and exploration

Data Scientists focus their efforts on modeling and evaluating results



Why Stop at Microservices for Supporting Your ML Code?



Why stop here?

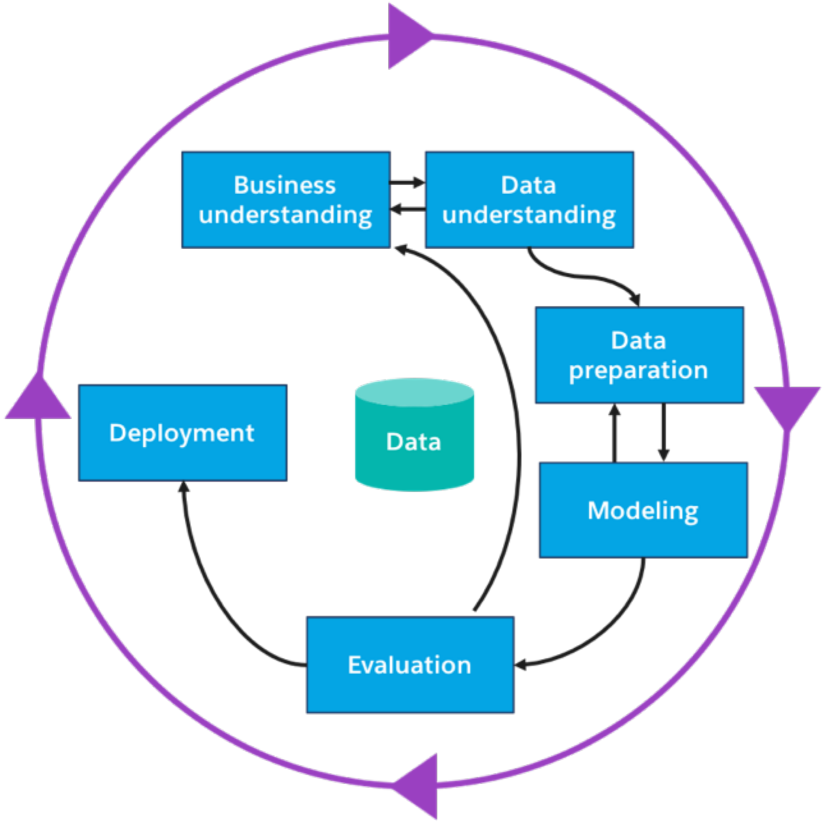
Your ML code can also be just a collection of microservices!

Auto Machine Learning

Building reusable ML code

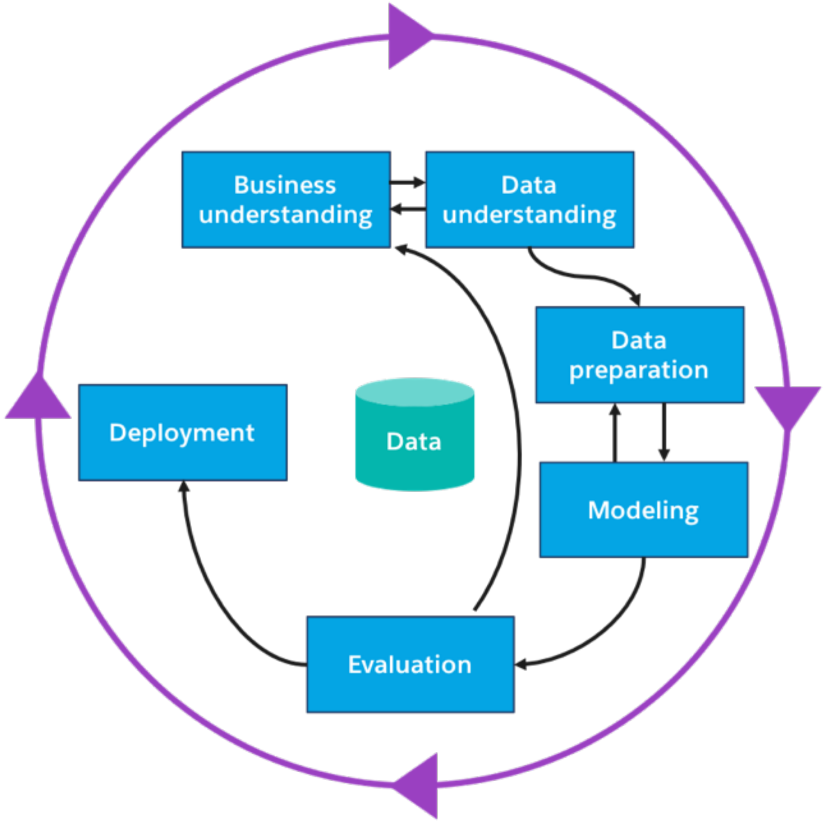
Leveraging Platform Services to Easily Deploy 1000s of Apps

Data Scientists on App #1

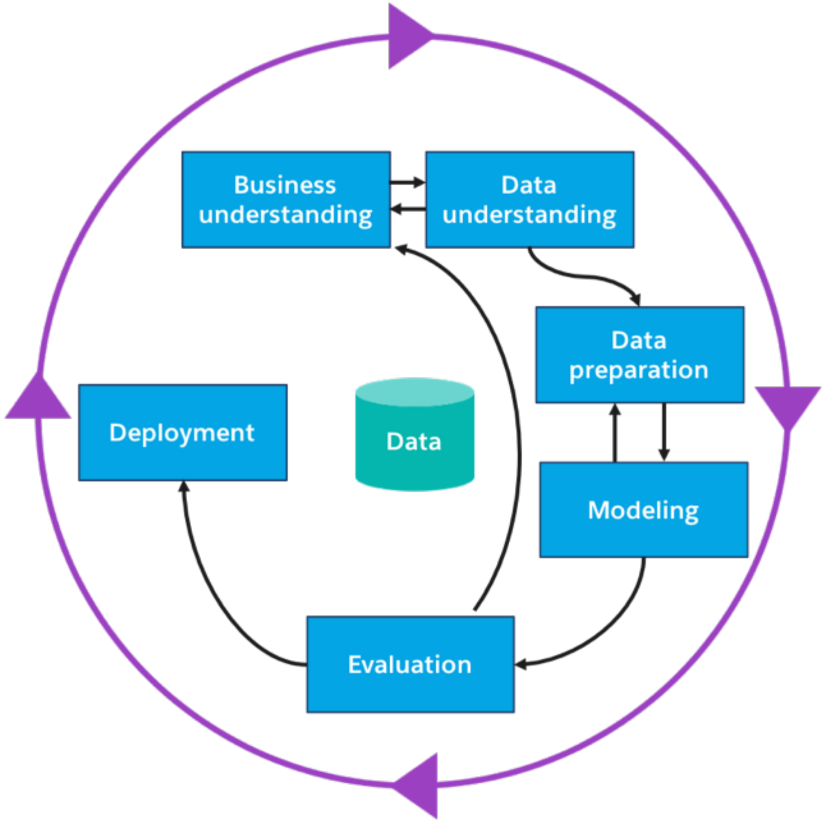


Leveraging Platform Services to Easily Deploy 1000s of Apps

Data Scientists on App #1

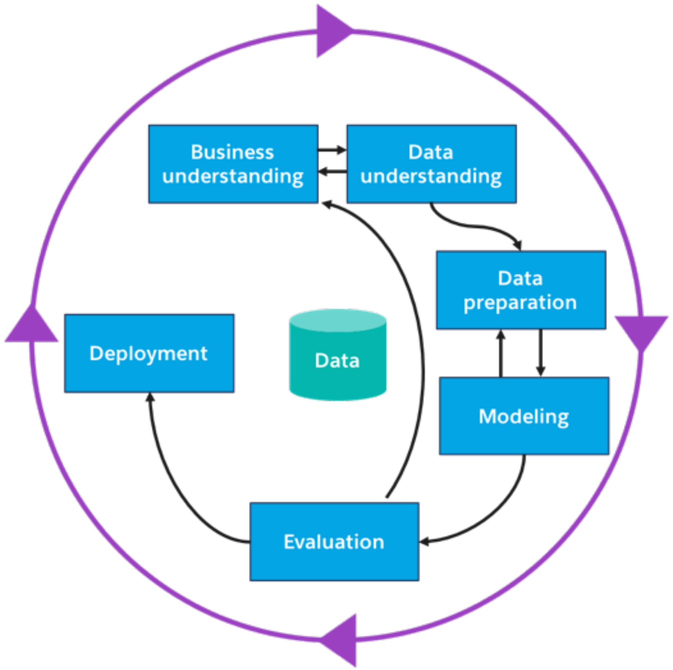


Data Scientists on App #2

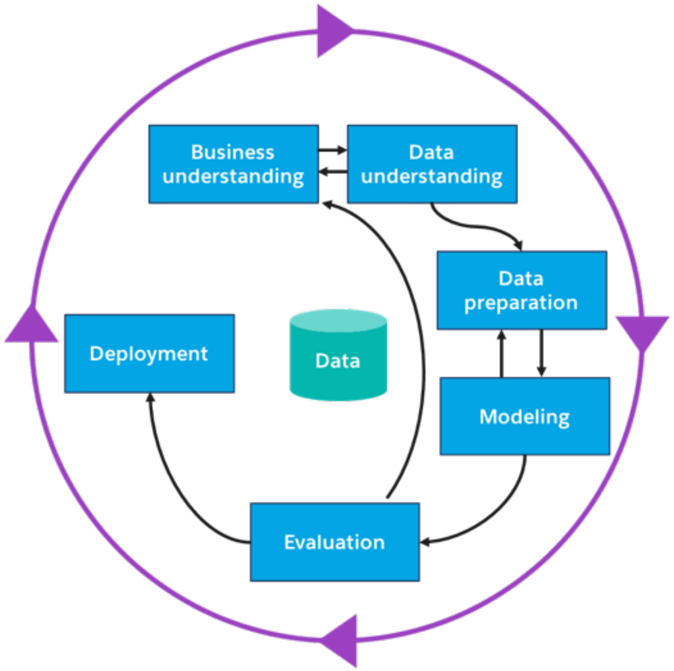


Let’s Add a Third App

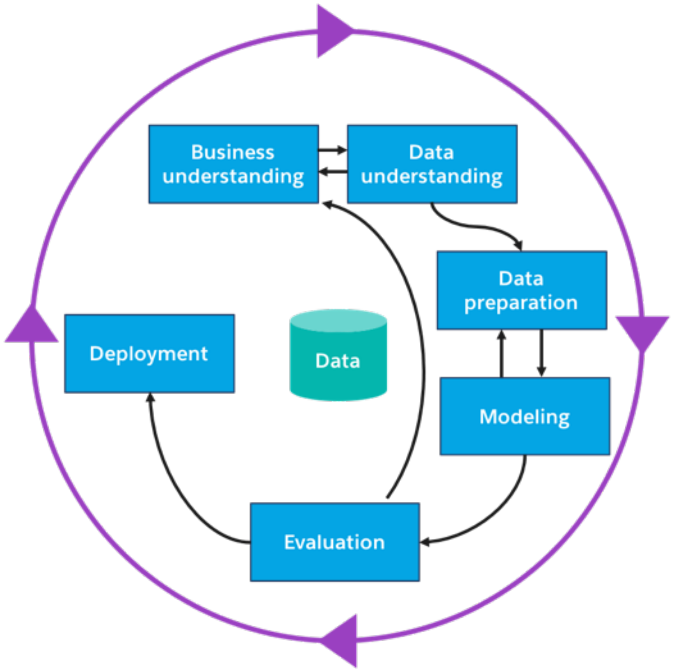
Data Scientists on App #1



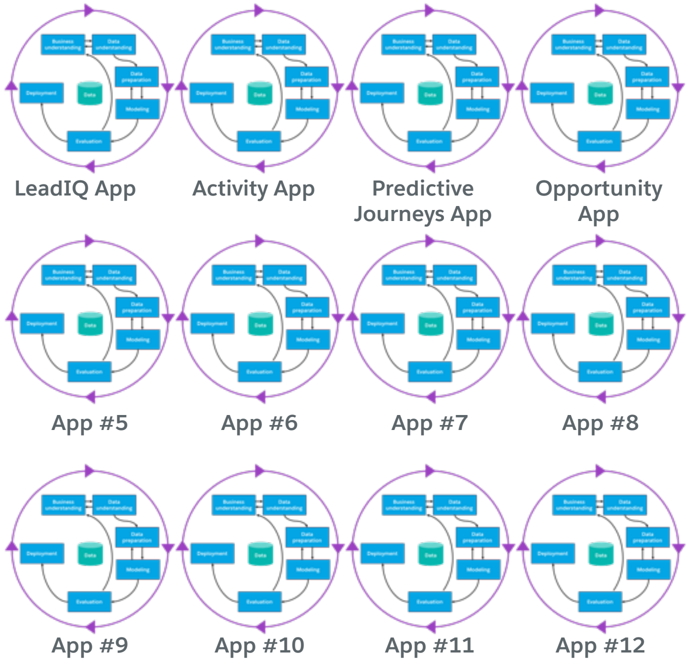
Data Scientists on App #2

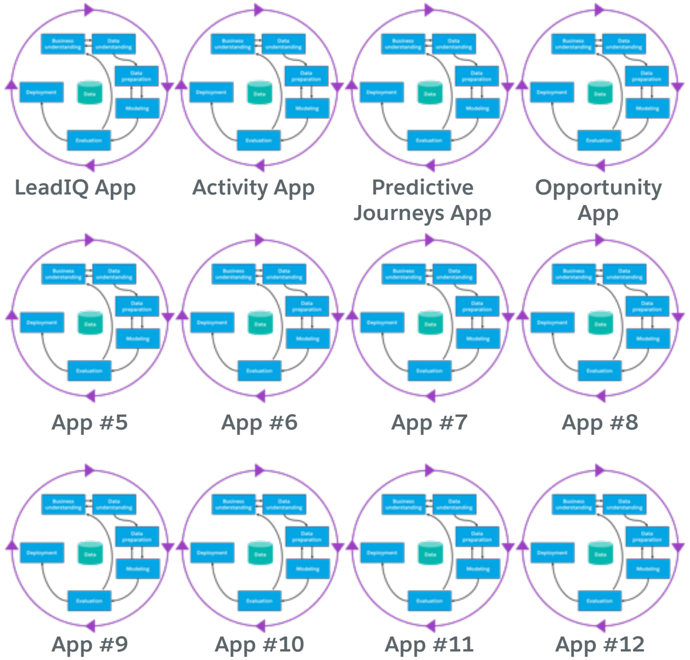


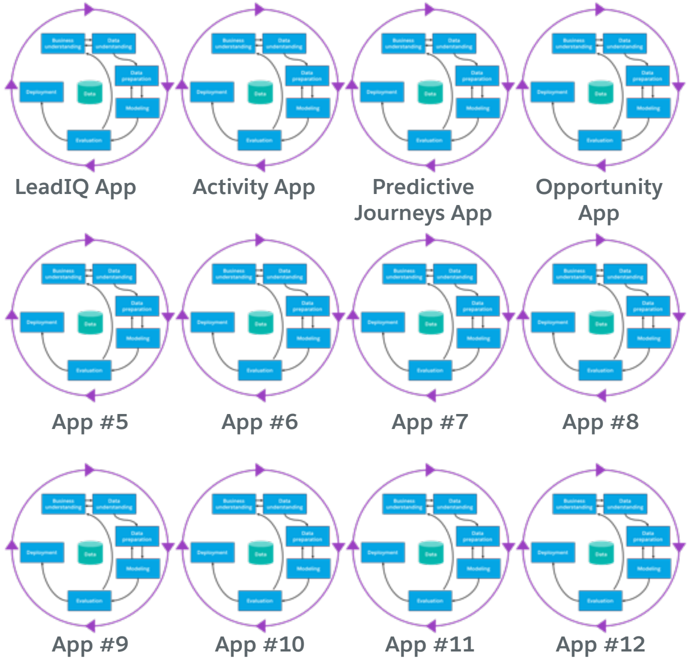
Data Scientists on App #3



How This Process Would Look in Salesforce







150,000 customers

Einstein's New Approach to AI

Democratizing AI for Everyone

Classical Approach → Data Sampling → Feature Selection → Model Selection → Score Calibration → Integrate to Application → Artificial Intelligence

Einstein Auto-ML → AI for CRM
Discover
Predict
Recommend
Automate

Data already prepped
Models automatically built
Predictions delivered in context

Repeatable Elements in Machine Learning Pipelines

AutoML for feature engineering

Categorical Variables

NAME	TITLE
Jim Steele	Senior VP
John Gardner	Senior VP
Andy Smith	Vice President
Test User	Vice President
Test User	CEO
Test User	Vice President
Test User	Chairperson
Test User	CEO

Text Fields

DESCRIPTION
A blessing in disguise
Time flies when you're having fun
Alles hat ein Ende, nur die Wurst hat zwei
um den heißen Brei herumreden
We'll cross that bridge when we come to it
You can say that again
Your guess is as good as mine

Numerical Buckets

number of employees
90
16
224
192
335
12
621
72
560
80
24
0
208

Repeatable Elements in Machine Learning Pipelines

AutoML for feature engineering

Categorical Variables

NAME	TITLE	Senior VP	CEO	Vice President
Jim Steele	Senior VP	1	0	0
John Gardner	Senior VP	1	0	0
Andy Smith	Vice President	0	0	1
Test User	Vice President	0	0	1
Test User	CEO	0	1	0
Test User	Vice President	0	0	1
Test User	Chairperson	0	0	0
Test User	CEO	0	1	0

Repeatable Elements in Machine Learning Pipelines

AutoML for feature engineering

Text Fields DESCRIPTION	Word Count	Word Count (no stop words)	Is English	Sentiment
A blessing in disguise	4	2	1	1
Time flies when you're having fun	6	3	1	1
Alles hat ein Ende, nur die Wurst hat zwei	9	4	0	0
um den heißen Brei herumreden	6	4	0	-1
We'll cross that bridge when we come to it	7	3	1	0
You can say that again	5	1	1	0
Your guess is as good as mine	7	3	1	0

Repeatable Elements in Machine Learning Pipelines

AutoML for feature engineering

Numerical Buckets		
number of employees	->	employee bucket
90	->	10-99
16	->	10-99
224	->	100-499
192	->	100-499
335	->	100-499
12	->	10-99
621	->	500-1000
72	->	10-99
560	->	500-1000
80	->	10-99
24	->	10-99
0	->	0-9
208	->	100-499

What Now? How autoML can choose your model

```
>>> from sklearn import svm
>>> from numpy import loadtxt as l, random as r
>>> clf = svm.SVC()
>>> pls = numpy.loadtxt("leadFeatures.data", delimiter=",")
>>> testSet = r.choice(len(pls), int(len(pls)*.7), replace=False)
>>> X,  y = pls[-testSet,:-1], pls[-testSet:,-1]
>>> clf.fit(X,y)
SVC(C=1.0, cache_size=200, class_weight=None,
        coef0=0.0,decision_function_shape=None, degree=3,
        gamma='auto', kernel='rbf', max_iter=-1,
        probability=False, random_state=None, shrinking=True,
        tol=0.001, verbose=False)
>>> clf.score(pls[testSet,:-1],pls[testSet,-1])
 0.8857142857142857142568
```

Should we try other model forms?
Features?
Kernels or hyperparameters?

Each use case will have its own model and features to use. We enable building separate models and features with 1 code base using OP

A tournament of models!

CUSTOMER A

Customer ID	1	2	3	4	5
Age	22	30	45	23	60
Age Group	A	A	B	A	C
Gender	M	F	F	M	M
Valid Address	Y	N	N	Y	Y

MODEL GENERATION

Model 1

Model 2

Model 3

Model 4

...

MODEL TESTING

Model 1
83% accuracy

Model 2
91% accuracy

Model 3
73% accuracy

Model 4
89% accuracy

...

A tournament of models!

CUSTOMER A

Customer ID	1	2	3	4	5
Age	22	30	45	23	60
Age Group	A	A	B	A	C
Gender	M	F	F	M	M
Valid Address	Y	N	N	Y	Y

CUSTOMER B

Customer ID	1	2	3	4	5
Age	34	22	66	58	41
Age Group	B	A	C	C	B
Gender	M	M	F	M	F
Valid Address	Y	Y	N	N	Y

MODEL GENERATION

Model 1

Model 2

Model 3

Model 4

...

MODEL TESTING

Model 1

Model 2
Customer A

Model 3

Model 4
Customer B

...

Deploy Monitors, Monitor, Repeat!

134	215	98.51%
Models in Production	Models Trained (curr.month)	Models with Above Chance Performance

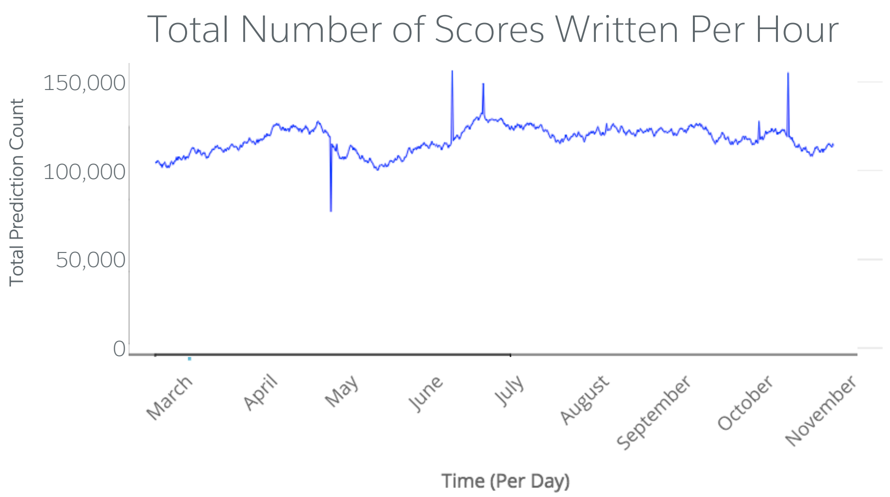
8	35,573,664
Experiments Run this Week	Predictions Written Per Day (7 day avg)

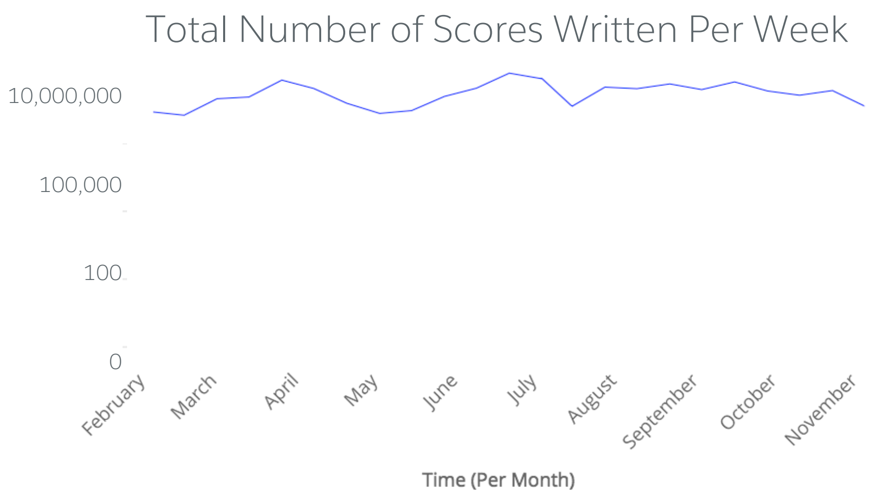
Sample Dashboard on Simulated Data

Deploy Monitors, Monitor, Repeat!

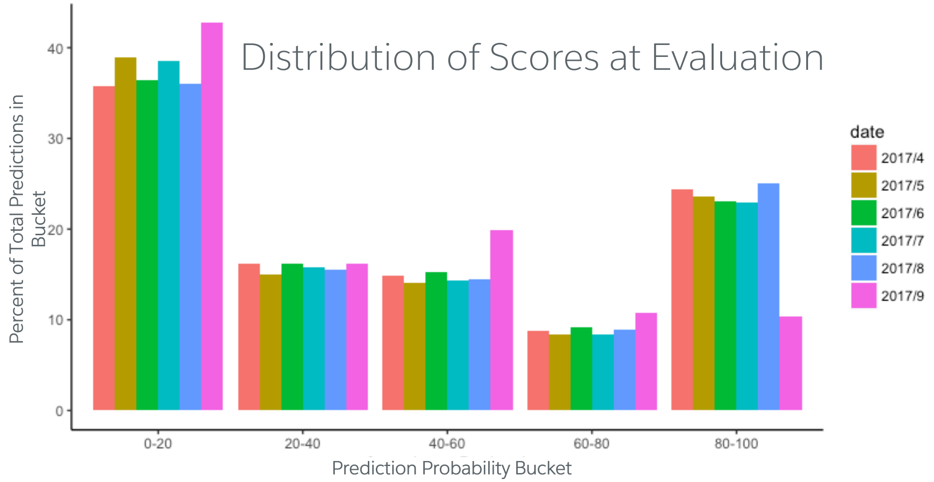
Pipelines, Model Performance, Scores – Invest your time where it is needed!

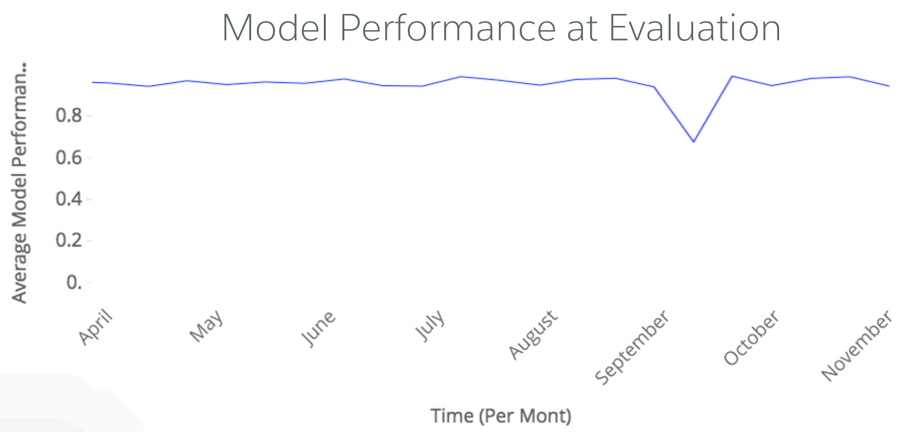
105,874
Scores Written Per Hour(1 day moving avg)





0.86
Evaluation auROC





Sample Dashboard on Simulated Data

Deploy Monitors, Monitor, Repeat!

134 Models in Production	216 Models Trained (curr.month)	99.25% Models with Above Chance Performance

12 Experiments Run this Week	35,573,664 Predictions Written Per Day (7 day avg)

Sample Dashboard on Simulated Data

Key Takeaways

- Deploying machine learning in production is hard
- Platforms are critical for enabling data scientist productivity
  - Plan for multiple apps… **always**
  - To ensure enabling rapid identification of areas of improvement and efficacy of new approaches provide
    - Monitoring services
    - Experimentation frameworks
- Identify opportunities for reusability in all aspects, even your machine learning pipelines
- **Help simplify the process of experimenting, deploying, and iterating**


Thank You