Data Operations:

Or How I learned to stop data wrangling and love machine learning


Gentlemen, you can’t be data wrangling in here.
This is a machine learning company.


Nexla solves the challenges of data operations, to enable
greater focus on Machine Learning & Analytics

DATA PARTNERS

IN-HOUSE DATA

PUBLIC DATA SOURCES

Format:
JSON, XML, CSV, SQL, ZIP, AVRO, PARQUET

- FTP
- API
- S3
- HTTP
- Database
- HDFS
- ?

READY TO USE

DATAOPS: PILLARS

DATA OPERATIONS

Monitor & Manage

Dependable data flow from hundreds of sources. Data dictionary management. Alerts.

Connect & Move

Inter-company, multi-cloud and hybrid cloud

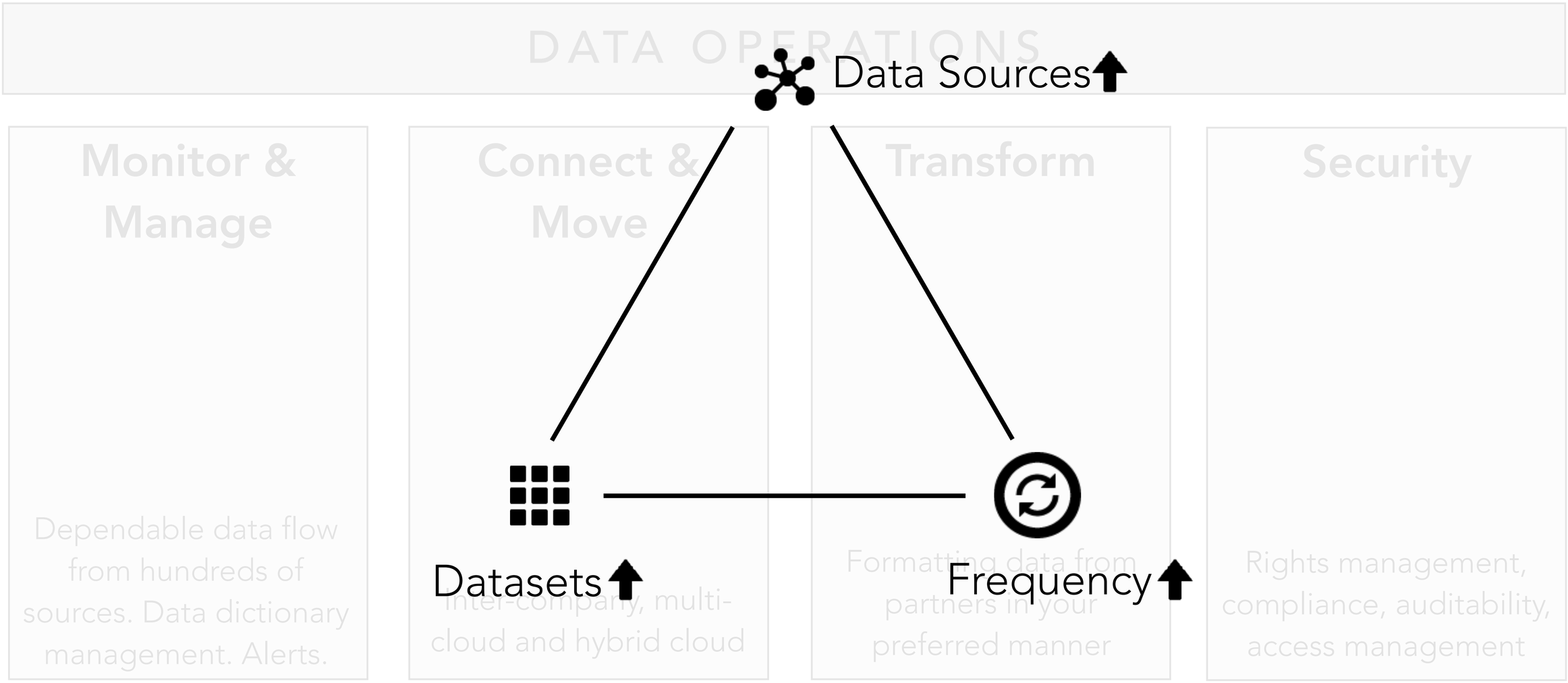
Transform

Serverless compute to format data from partners in your preferred manner

Security

Rights management, compliance, auditability, access management

DATAOPS: COMPLEXITY DRIVERS



DATAOPS: MACHINE LEARNING

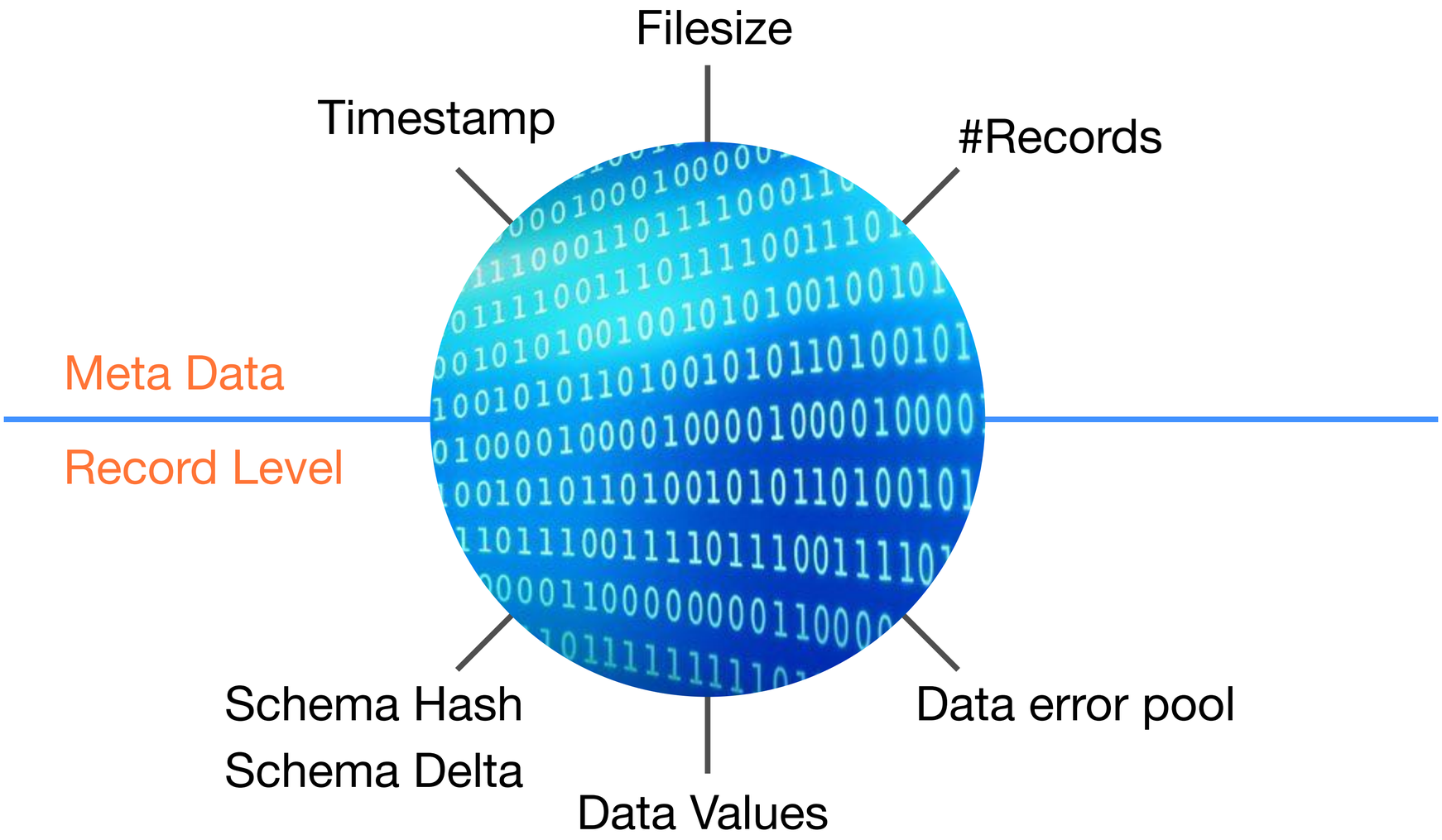
DATA OPERATIONS

Monitor & Manage

Dependable data flow from hundreds of sources. Data dictionary management. Alerts.

- Source uptime
- Data Frequency
- Data Volume
- Schema Changes
- Data Changes
- Auditability

DATAOPS MONITORING: SIGNALS



DATA PARTNERS

IN-HOUSE DATA

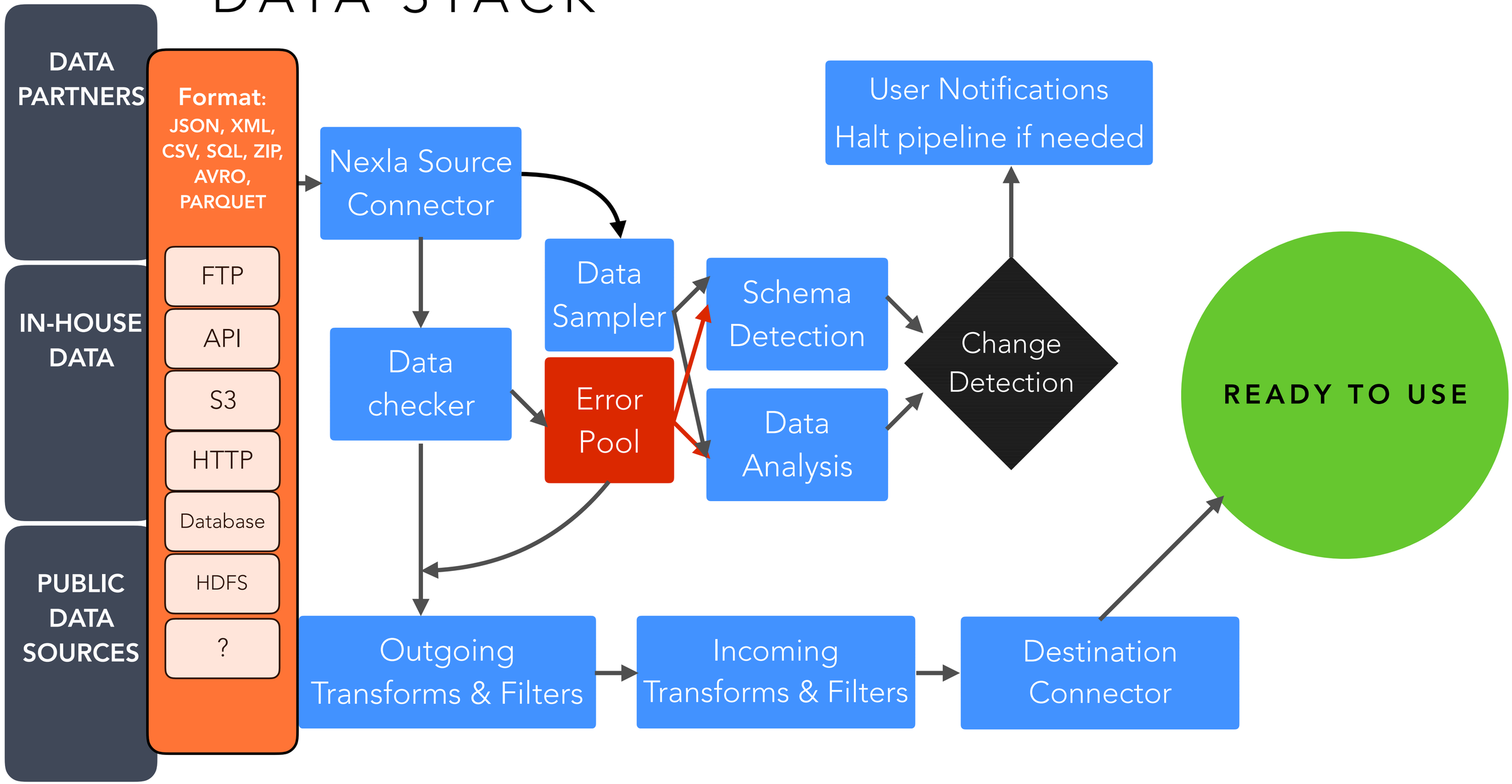
PUBLIC DATA SOURCES

Format:
JSON, XML, CSV, SQL, ZIP, AVRO, PARQUET

- FTP
- API
- S3
- HTTP
- Database
- HDFS
- ?

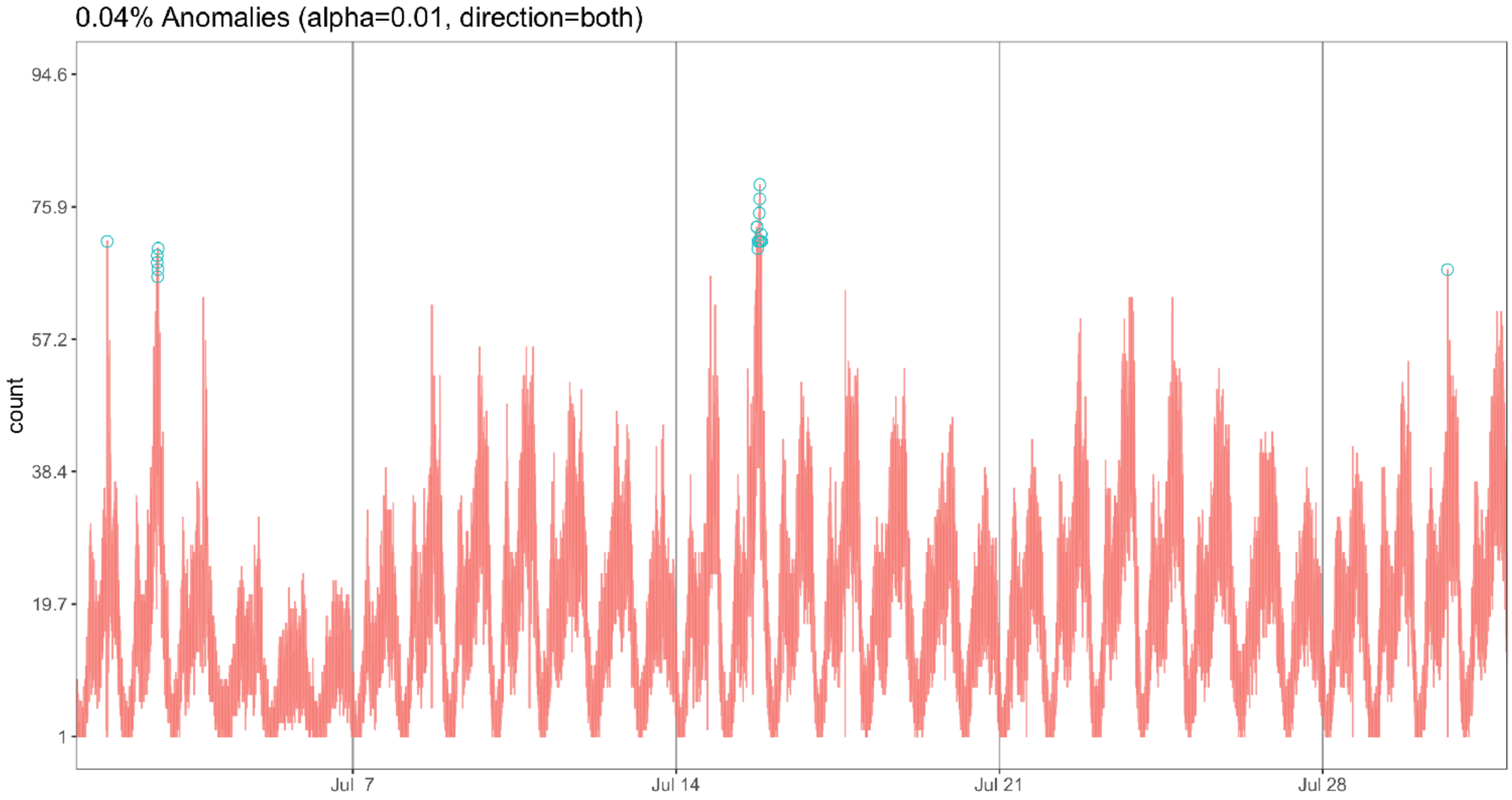
READY TO USE

DATA STACK


DATA PARTNERS
IN-HOUSE DATA
PUBLIC DATA SOURCES
Format: JSON, XML, CSV, SQL, ZIP, AVRO, PARQUET
FTP
API
S3
HTTP
Database
HDFS
?
Nexla Source Connector
Data checker
Data Sampler
Error Pool
Schema Detection
Data Analysis
Change Detection
User Notifications
Halt pipeline if needed
READY TO USE
Outgoing Transforms & Filters
Incoming Transforms & Filters
Destination Connector

Timestamp	Filename	# of rows	Schema Hash
2016-11-17 06:00:00	events_2016_11_17_6	5653	4b2c83bf78746a086dde0144f91d1ac8
2016-11-17 07:00:00	events_2016_11_17_7	5103	4b2c83bf78746a086dde0144f91d1ac8
2016-11-17 08:00:00	events_2016_11_17_8	5123	4b2c83bf78746a086dde0144f91d1ac8
2016-11-17 09:00:00	events_2016_11_17_9	9506	4b2c83bf78746a086dde0144f91d1ac8
2016-11-17 10:00:00	events_2016_11_17_10	5975	4b2c83bf78746a086dde0144f91d1ac8
2016-11-17 11:00:00	events_2016_11_17_11	5998	4b2c83bf78746a086dde0144f91d1ac8
2016-11-17 12:00:00	events_2016_11_17_12	6284	4b2c83bf78746a086dde0144f91d1ac8
2016-11-17 12:00:00	events_2016_11_17_12	3716	622adc13abd0fccd00e27a0659f8f1d4
2016-11-17 13:00:00	events_2016_11_17_13	8290	622adc13abd0fccd00e27a0659f8f1d4
2016-11-17 14:00:00	events_2016_11_17_14	9717	622adc13abd0fccd00e27a0659f8f1d4
2016-11-17 15:00:00	events_2016_11_17_15	8465	622adc13abd0fccd00e27a0659f8f1d4

```
res = AnomalyDetectionTs(data, direction='both',
      plot=TRUE,max_anoms=0.02,alpha=0.01)
```



SUPERVISED CLASSIFICATION

- Build a classification model for normal and outliers based on a training data
  - Customers upload a training set with labeled data
  - Typically training set has > 99.99% valid data and < 0.01 invalid data
  - Classification algorithms don’t work very well with imbalanced classes
- To solve this, we create a new dataset as follows:
  - Only select invalid data
  - Resample the original dataset to add (3-4x valid data items)
  - The new dataset is more balanced and is used as a training set
- Pros-Cons
  - Very accurate when training sets are available. Not usually the case.
  - Doesn’t work very well for frequently evolving data (stock market prices)

K-MEANS CLUSTERING

- Unsupervised detection without a training set
- Use K means algorithm to identify clusters
- Identify cluster boundaries
    - Using percentiles (95th percentile from all distances between data points and their respective cluster centers)
- Outliers are further away from cluster centres
- For each new data point,
    - Compute the distance of the data point from each cluster center
    - For Closest cluster, if the distance >boundary, the data point is an outlier

LOOKING FURTHER...

- We will continue to explore other techniques
  - Markov Chains; Hierarchical Temporal Memory Networks; Half space trees; Combination of algorithms


We are Hiring s Running closed Betas s Launching in March 2017