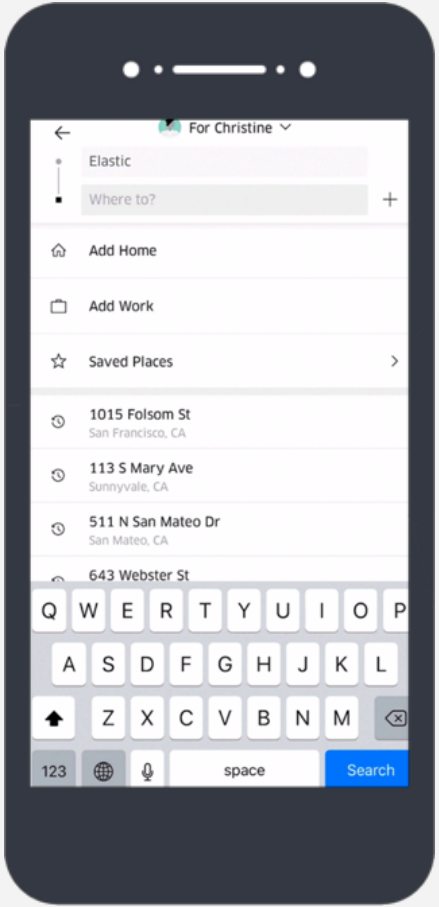
Machine Learning using Elasticsearch on Azure

Ravi Ramnani | Solutions Architect

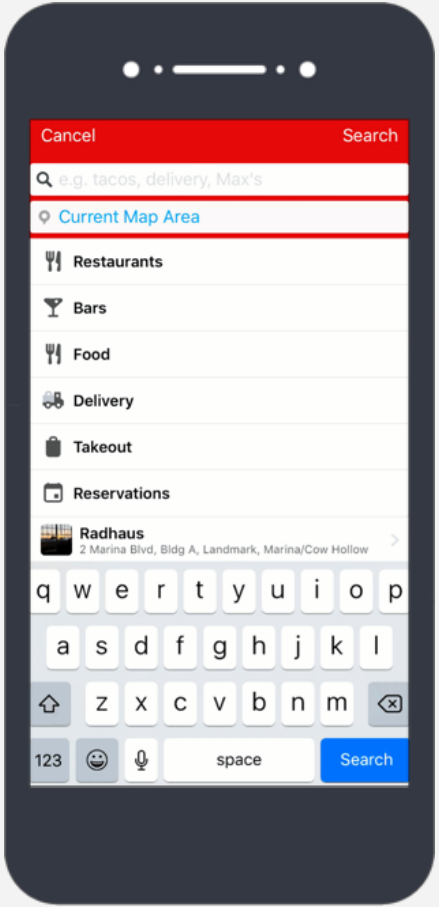
15th Dec, 2018

Elastic is a **search company**



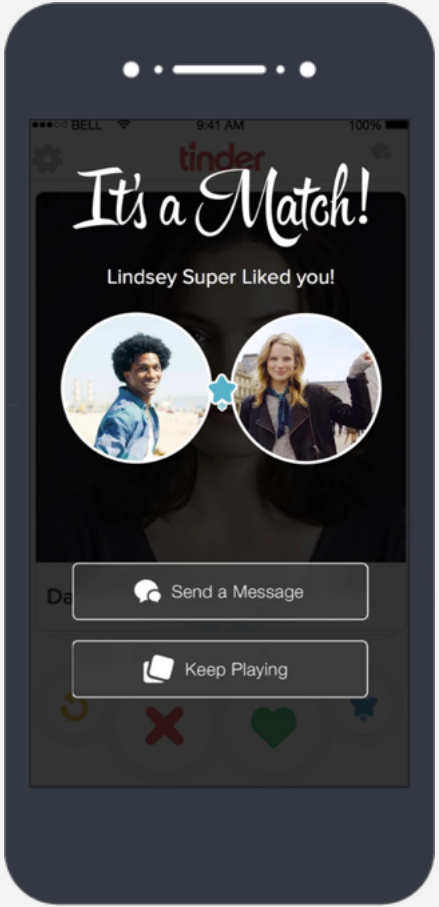
Uber

elastic



yelp®

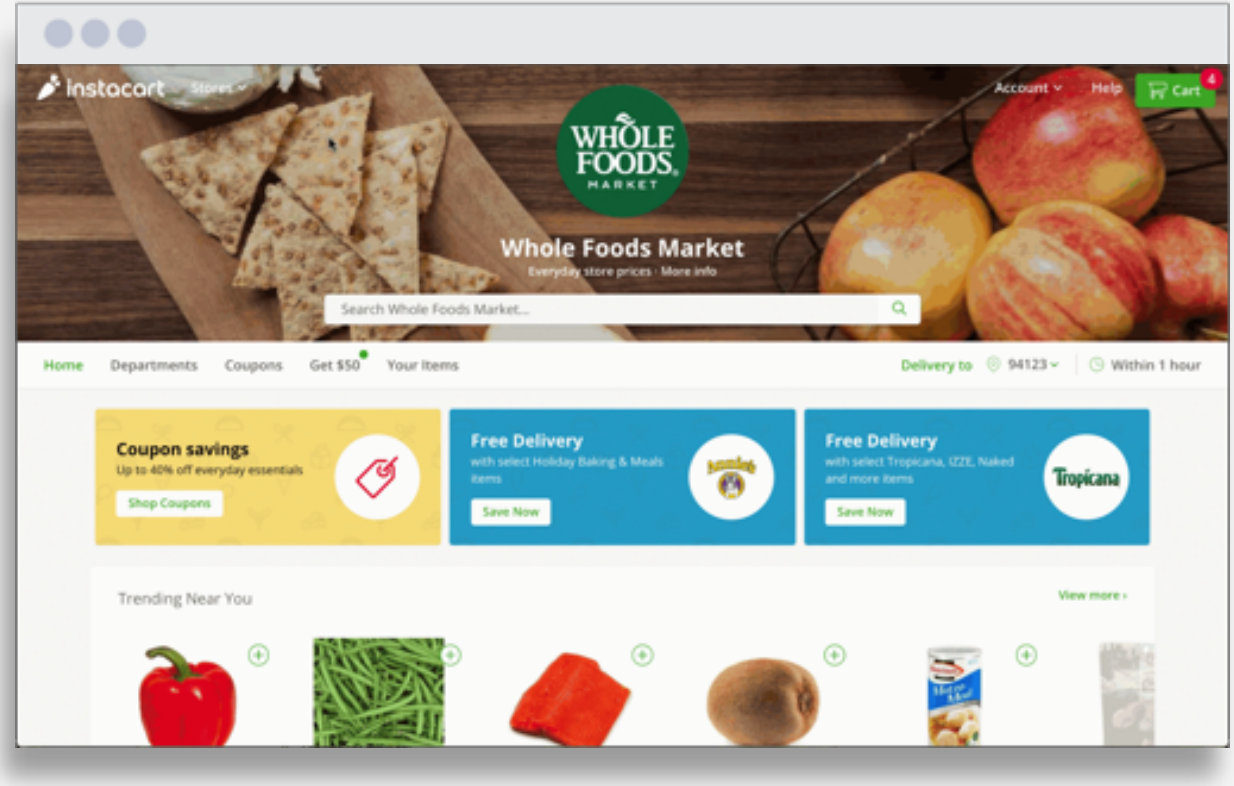
elastic

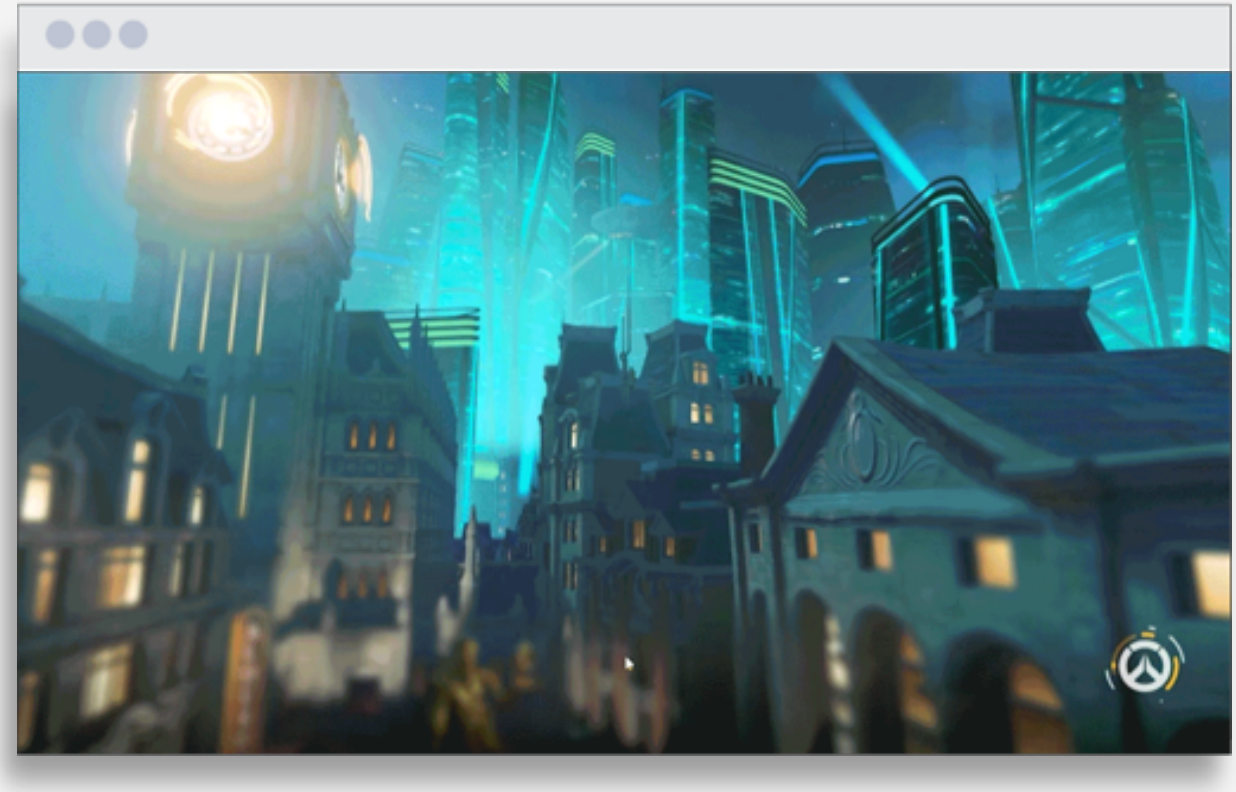


tinder™

elastic

instacart

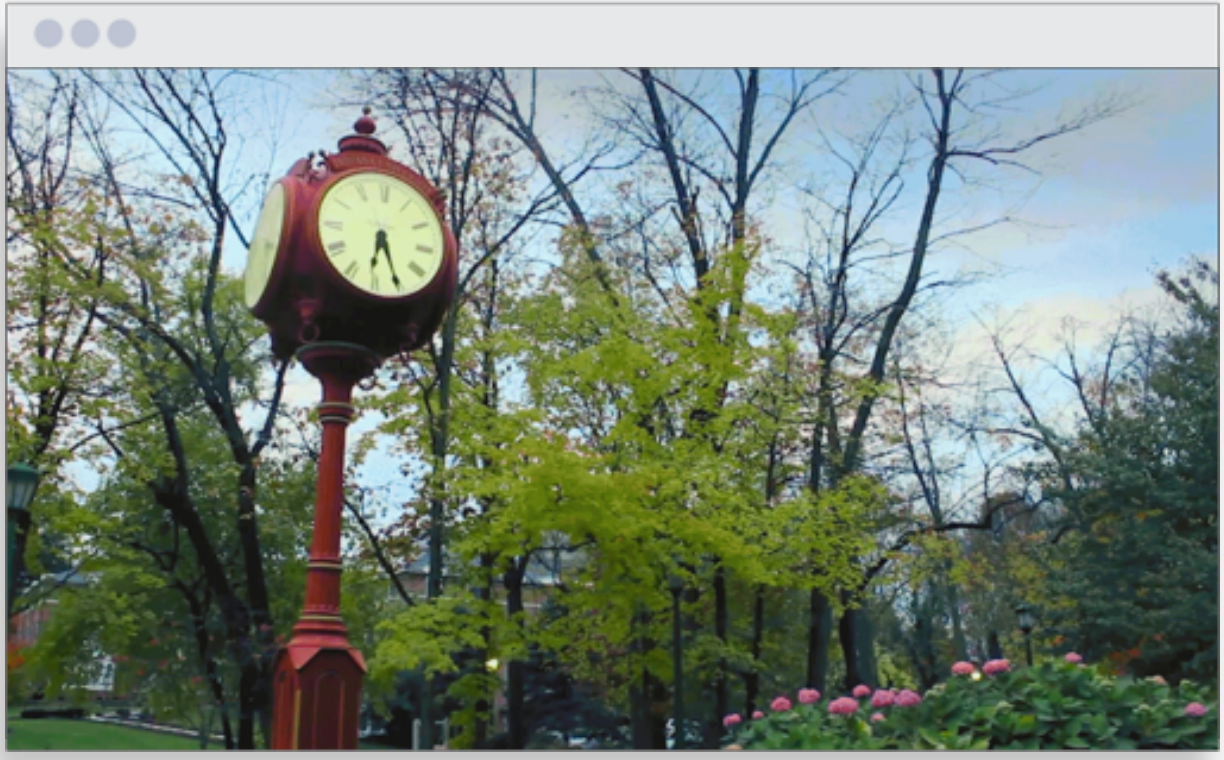

Instacart
Stores
Account
Help
Cart
4
WHOLE FOODS MARKET
Whole Foods Market
Everyday store prices · More info
Search Whole Foods Market...
Home
Departments
Coupons
Get $50
Your Items
Delivery to
94123
Within 1 hour
Coupon savings
Up to 40% off everyday essentials
Shop Coupons
Free Delivery
with select Holiday Baking & Meals items
Save Now
Free Delivery
with select Tropicana, IZZE, Naked and more items
Save Now
Tropicana
Trending Near You
View more



ACTIVISION BLIZZARD

elastic

INDIANA UNIVERSITY

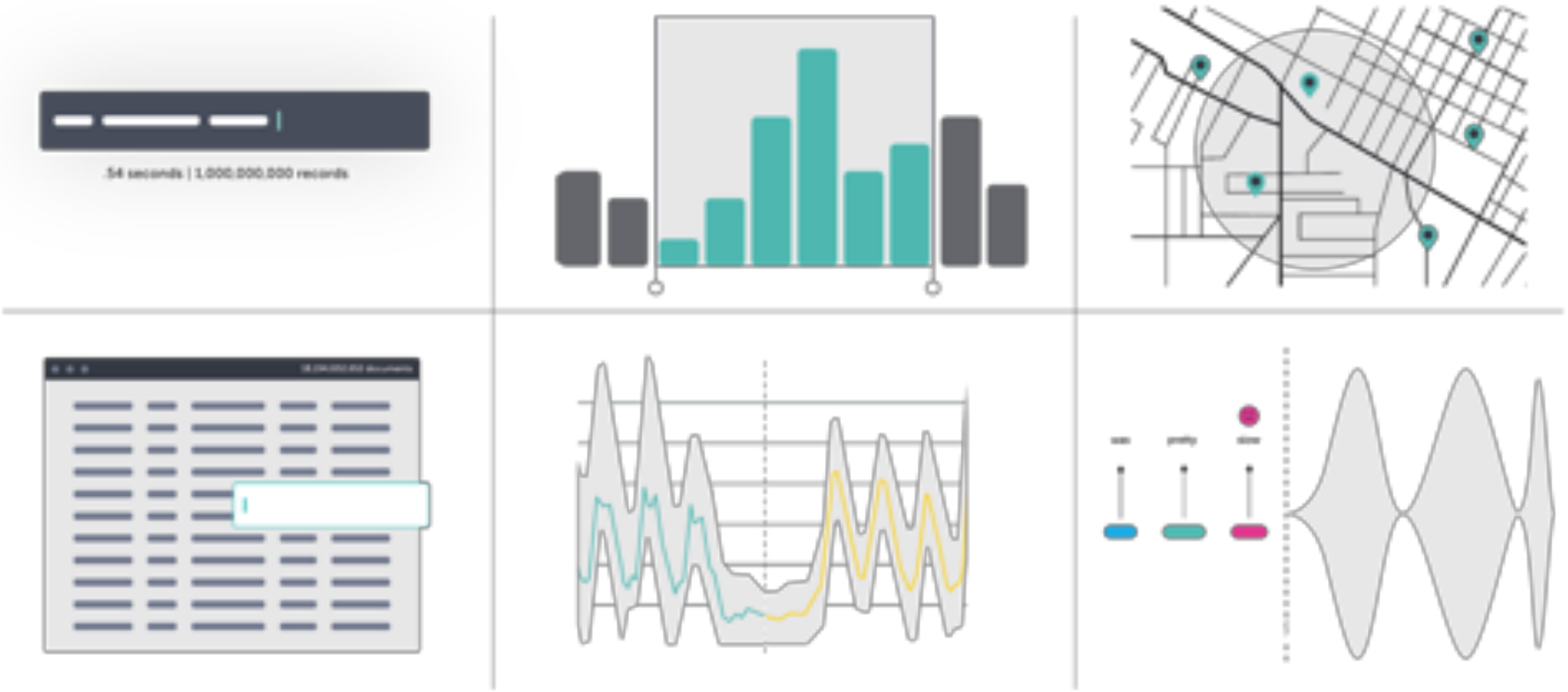


Customers across various **industries, segments, and geographies**

TECHNOLOGY	FINANCE	TELCO	CONSUMER	HEALTHCARE	PUBLIC SECTOR	AUTOMOTIVE / TRANSPORTATION	RETAIL
Adobe	BARCLAYS	orange	Uber	UnitedHealth Group	IU	RENAULT	Walmart
cisco	DISCOVER	COX	instacart	UCLA Health	United States Census Bureau	AIRBUS	The Home Depot
	USAA	SoftBank	tinder	Yale NewHaven Health	GSA	FUJITSU	ebay
Microsoft	collector bank	verizon	Activision Blizzard	Mayo Clinic	JPL Jet Propulsion Laboratory	SAMSUNG	Kroger
IBM	ING	Sprint	lyft	MERCK	WILSON NORTH CAROLINA		Walgreens

elastic

Search is a **constant/foundation**



Technology **differentiation**

SCALE

Distributed by design

SPEED

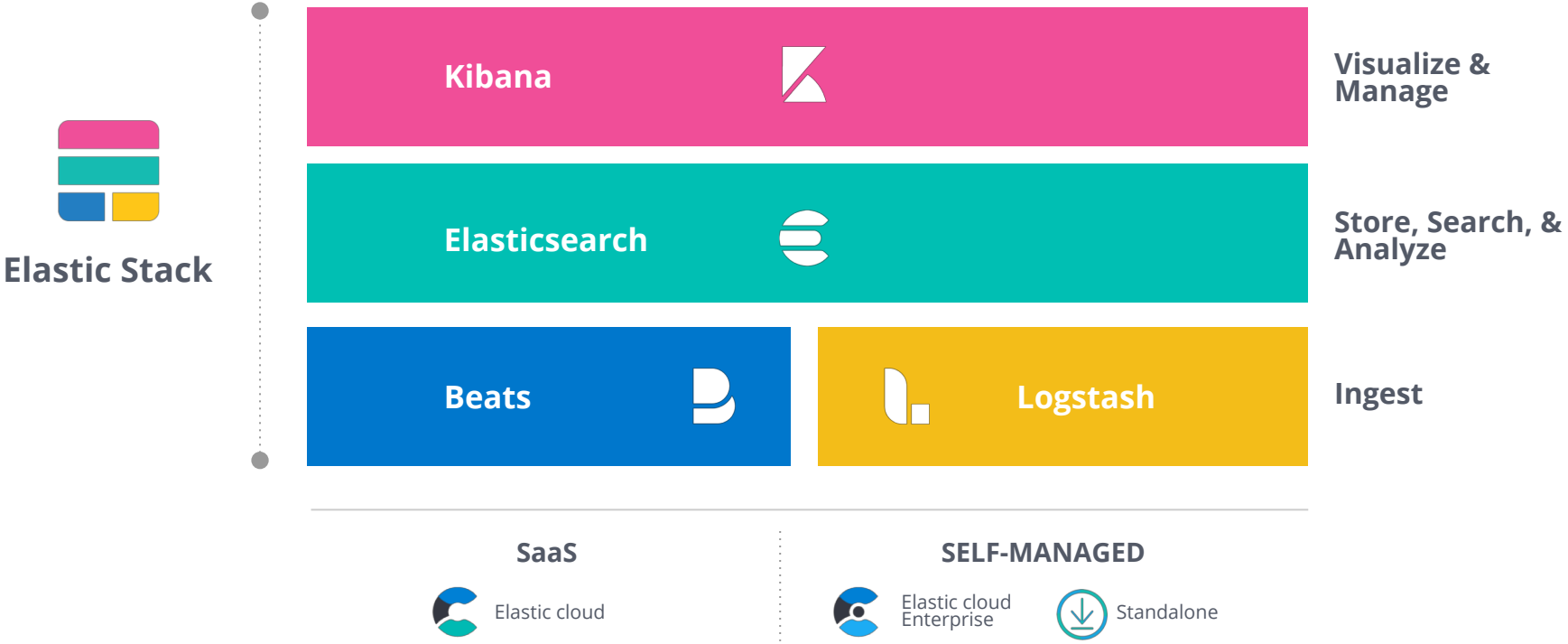
Find matches in milliseconds

RELEVANCE

Get highly relevant results

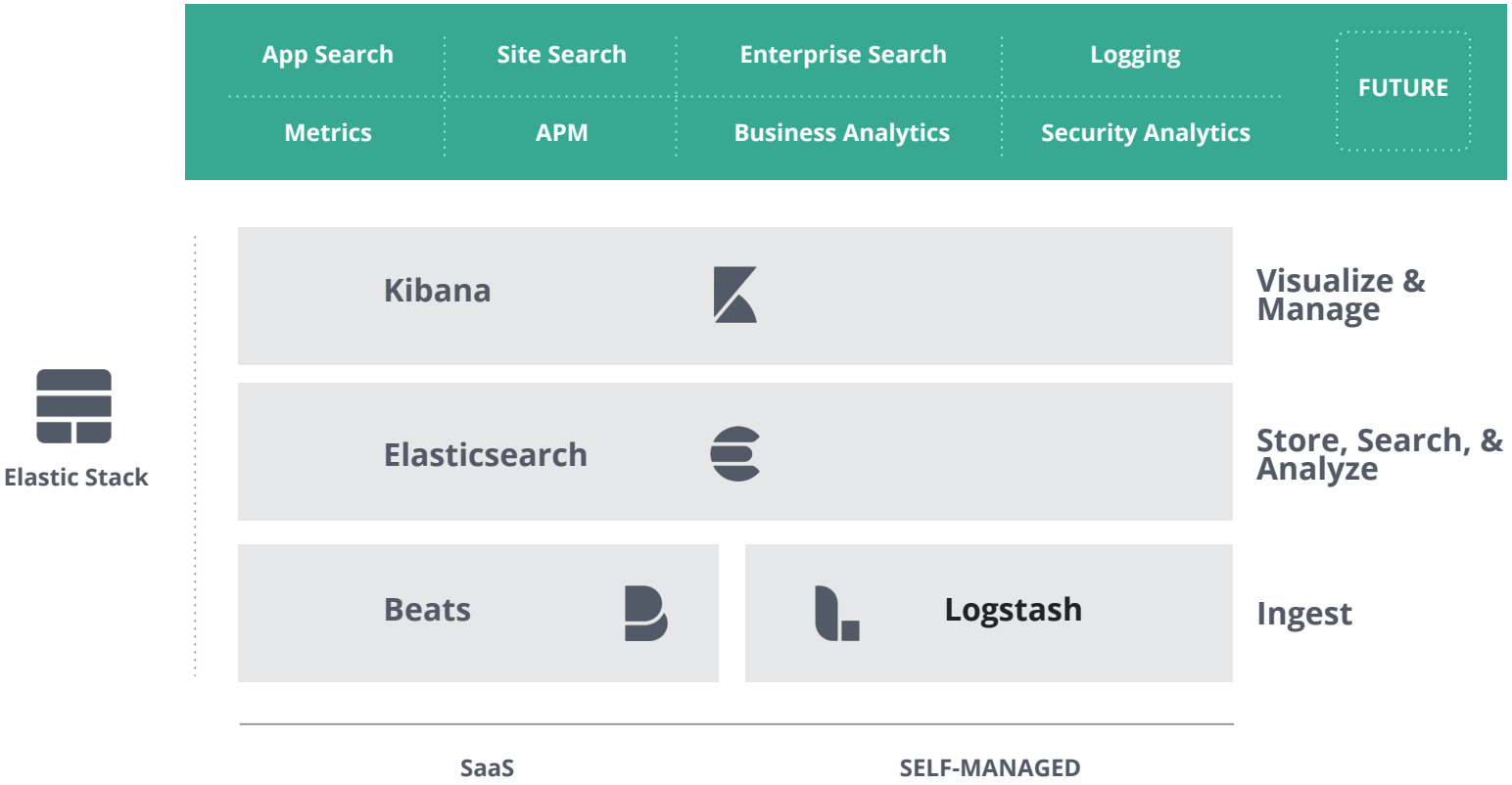
Elastic Products

Elastic **Stack**



elastic

Solutions



elastic

AI and ML

elastic

AI and ML

Algorithmic Approaches

Supervised	Unsupervised	Reinforcement
• Labelled Data • Driven by objective • NNs, SVMs, Decision Trees	• No Label • No objective, only Data • K-means, Apriori	• No Label • Driven by objective • Q-Learning, SARSA, DQL

Anomaly Detection Concepts

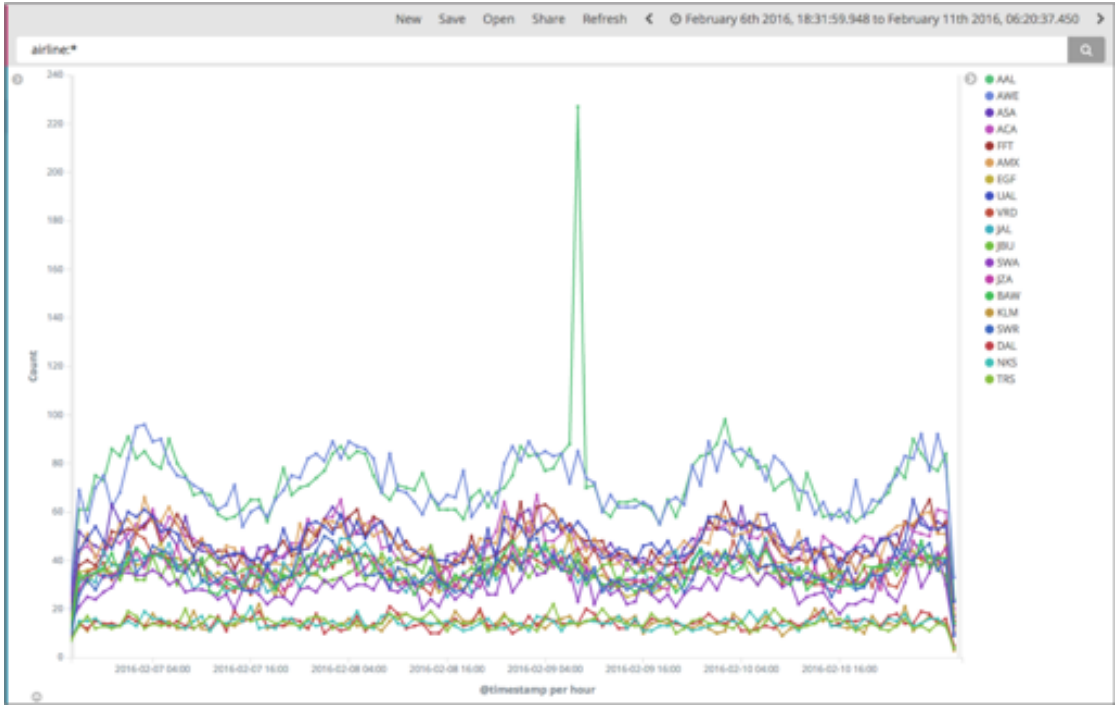
Terminology

- ***Machine Learning***
  - Broad term, but X-Pack Machine Learning is automated anomaly detection for time-series data (for now).
- ***Anomaly Detection***
  - Discovery of what's "weird" or "different", not what's "bad"
- ***Unsupervised Learning***
  - Learning without human-labeled examples (without being "taught"). Rely only on the data
- ***Bayesian***
  - An approach based on probability in which prior results are used to calculate probabilities of certain present or future events

What is “Abnormal”?

What’s abnormal here?

Why?

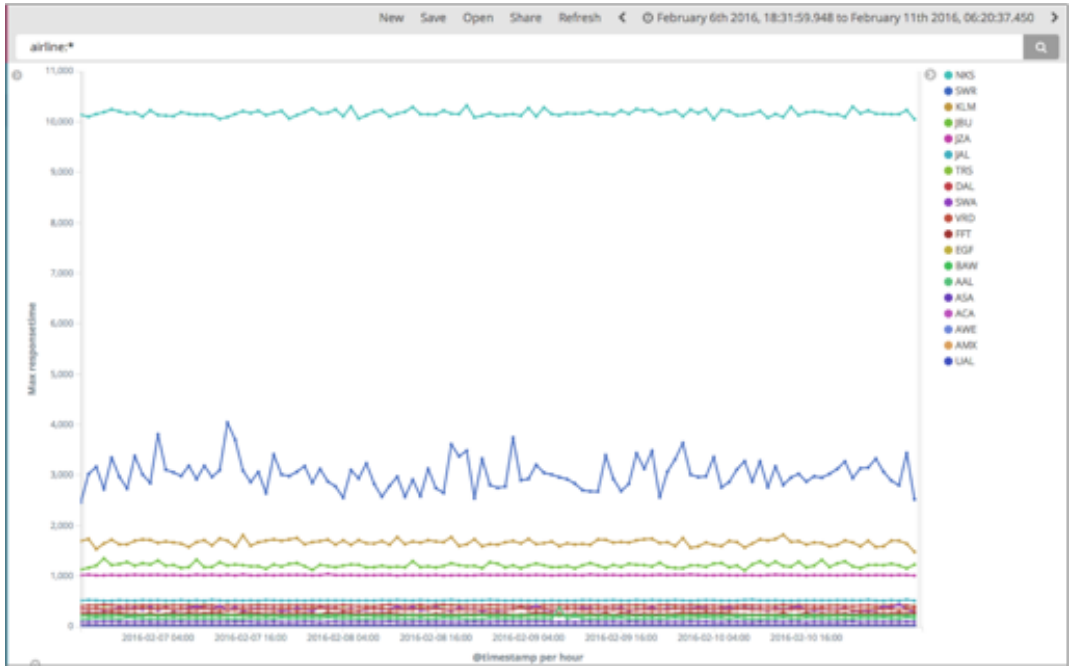


elastic

What is “Abnormal”?

What’s abnormal here?

Why?



elastic

What is “Abnormal”?

What’s abnormal here?

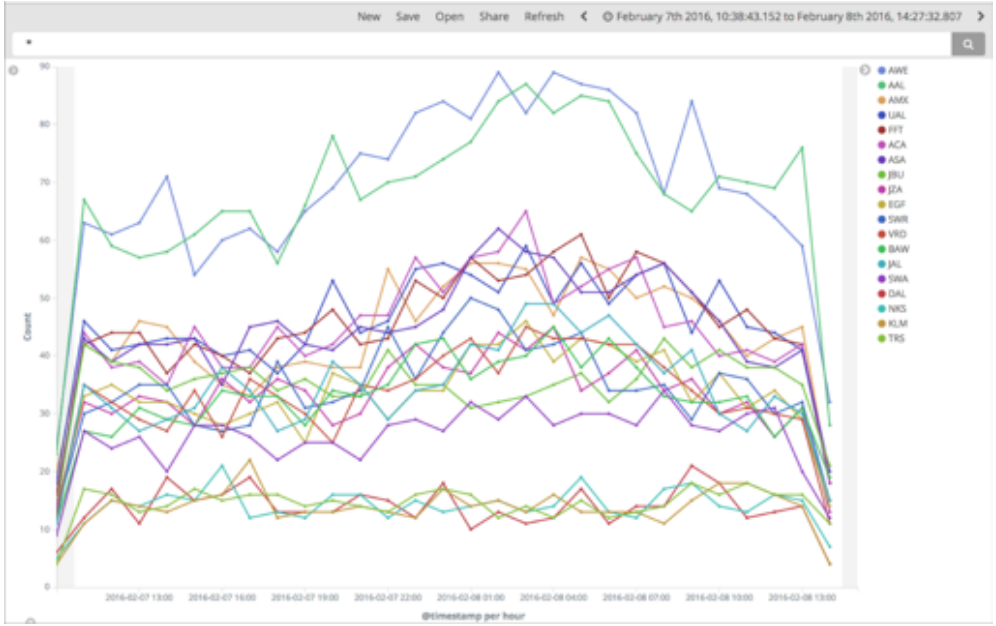
Why?

What is “Normal”?

In general, this question can be answered in two ways:

1) Something behaves in a consistent way with respect to itself, over time

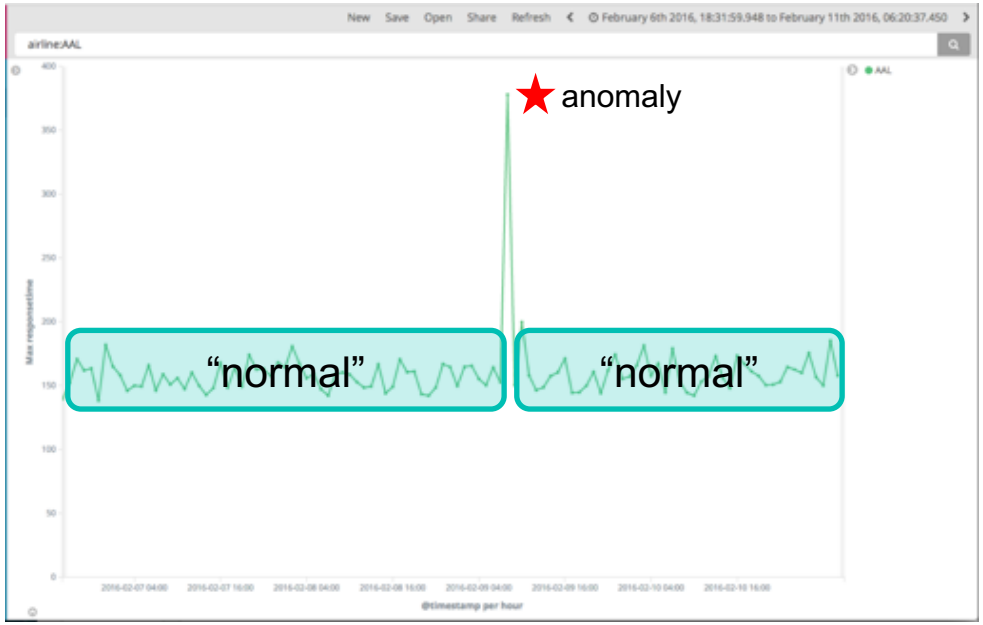
2) Something behaves in a consistent way compared against similar entities



elastic

What is “Abnormal”?

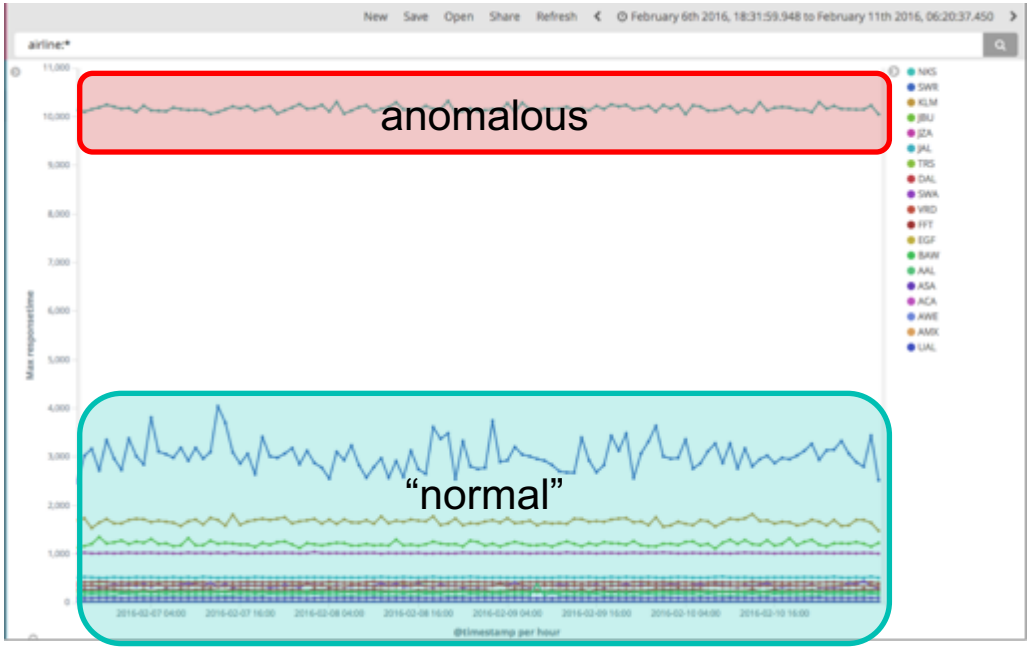
1) If something changes its behavior, compared to its own history – that change is ***anomalous***.



elastic

What is “Abnormal”?

2) If something is drastically different than others within a population, then that entity is ***anomalous***.

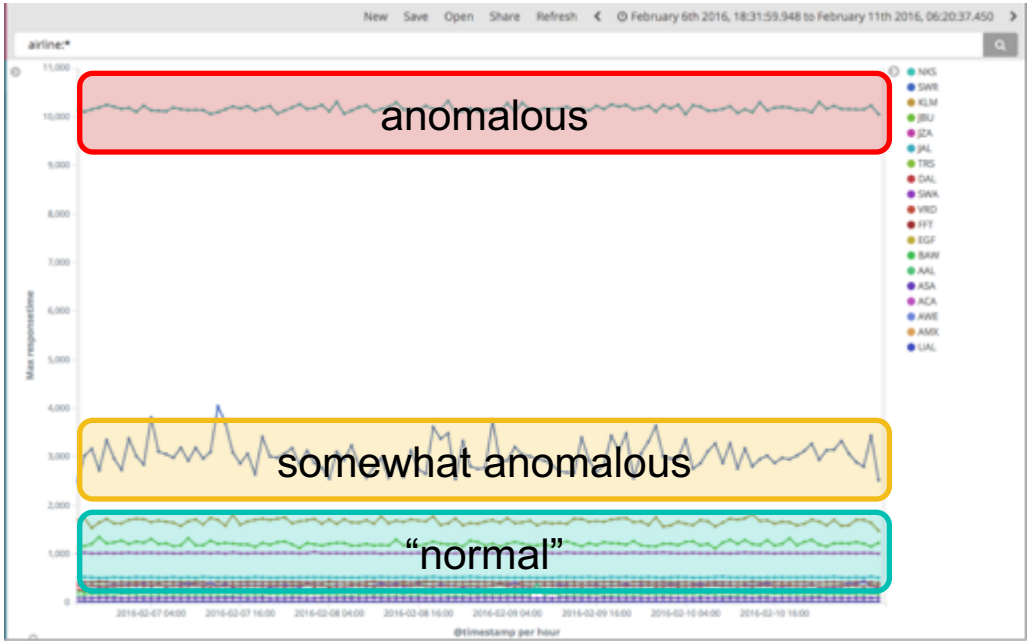


elastic

What is “Abnormal”?

2) If something is drastically different that others within a population, then that entity is ***anomalous***.

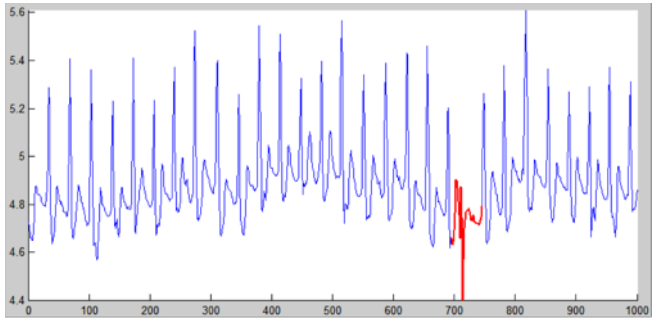
There’s also the concept of being “somewhat anomalous”



elastic

In Summary, Anomalousness is:

1) When an entities' ***behavior changes*** significantly and suddenly



2) When an entity is drastically ***different than others*** within a population

elastic

How to Learn “Normal”

elastic

An Analogy

- How could I learn how much Postal-mail you get daily and how could I use that information to predict how much you might get tomorrow?

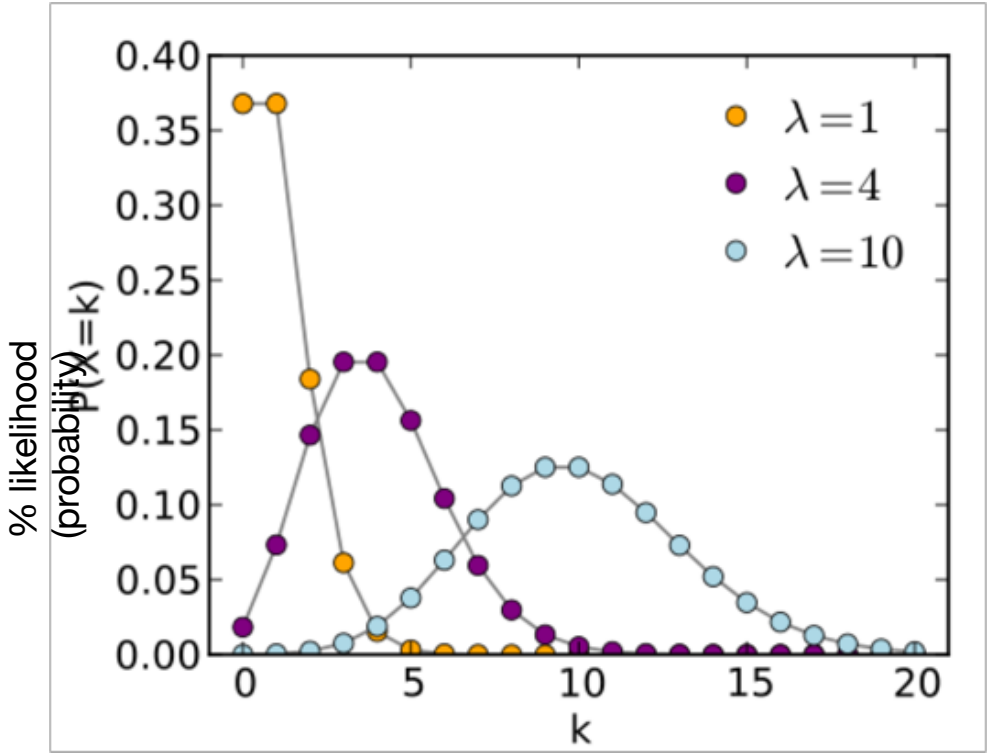
I could stand on your front porch and…

- Watch your mail delivery volume for a while, but how long?
  - 1 day?
  - 1 week?
  - 1 month?

- Notice, that you intuitively feel like you'll gain accuracy in your predictions with more data that you see.

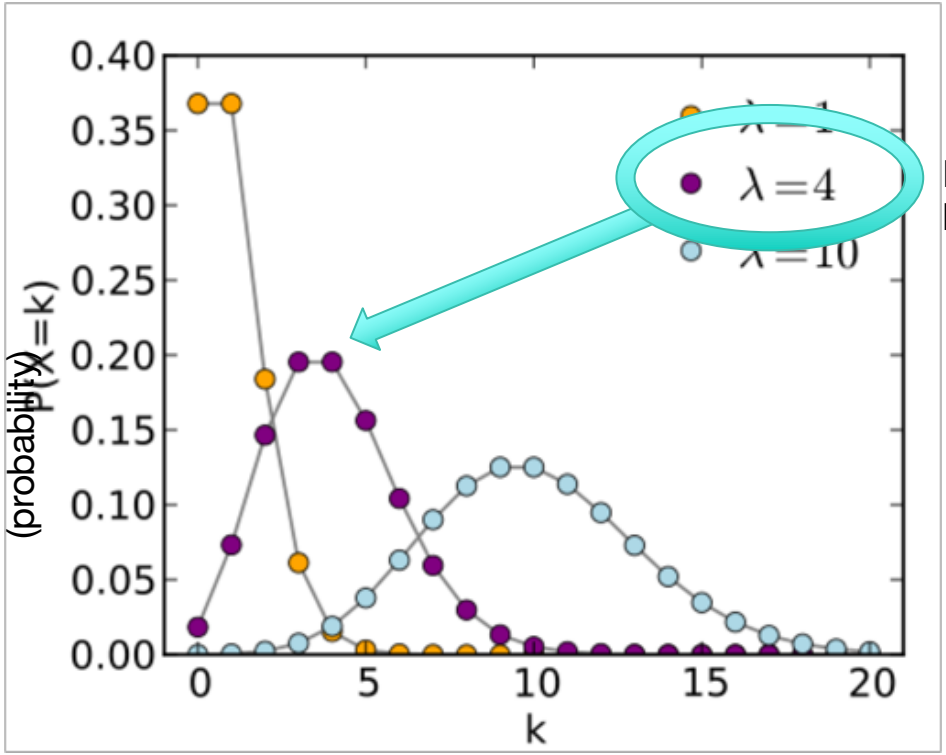
- Use those observations to create a model…

Probability Distribution Function



elastic

Probability Distribution Function

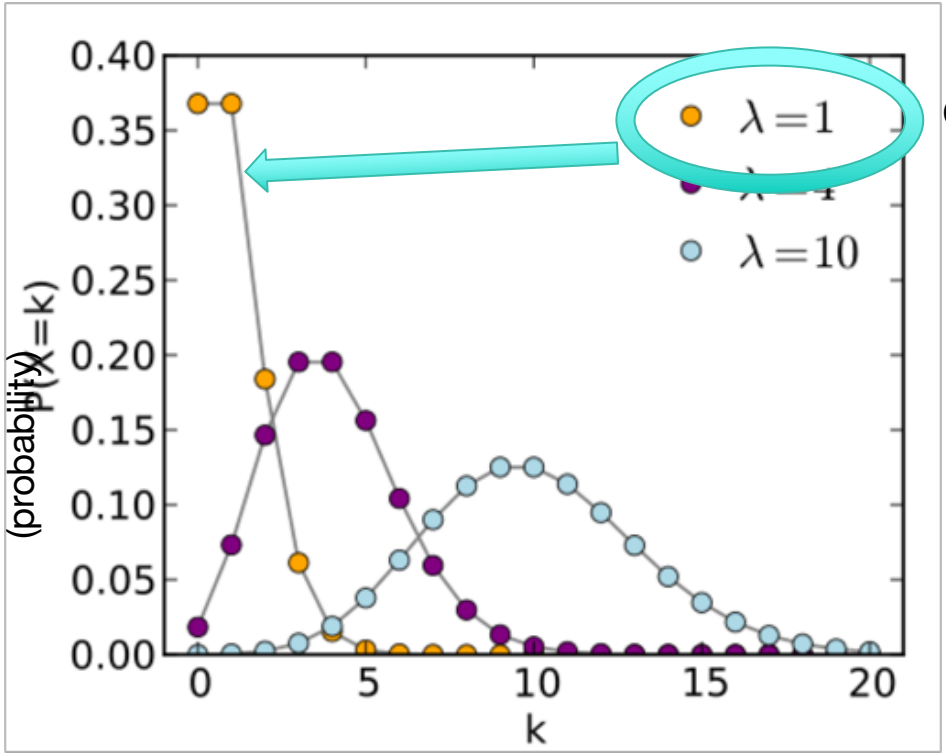


Best for my house

pieces of mail per day

elastic

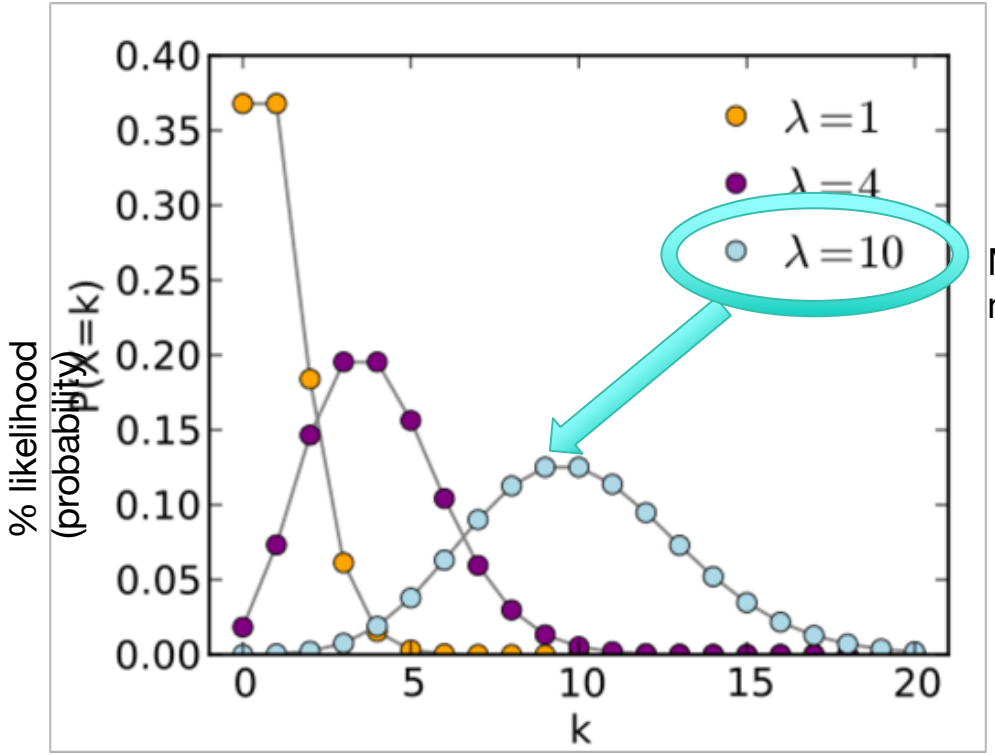
Probability Distribution Function



College Student?

elastic

Probability Distribution Function

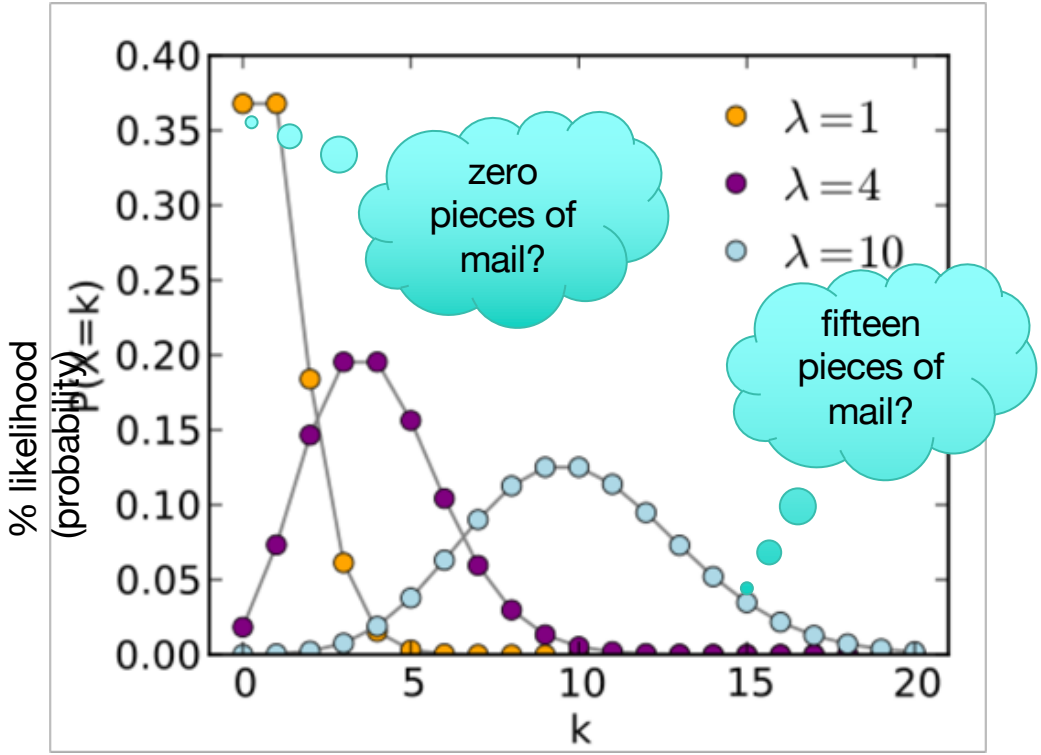


My mom

pieces of mail per day

elastic

Using the Model to Find What is Unexpected



elastic

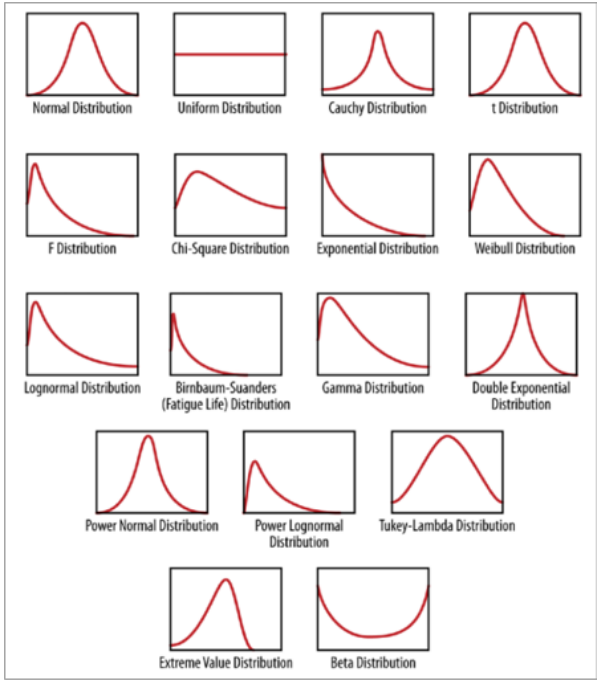
Relate back to IT/Security data

- # Pieces of mail = # events of a certain type
  - number of failed logins
  - number of errors in a log
  - number of events with certain status codes
- Or, metrics
  - response time
  - utilization
  - orders per minute

=> Every kind of data will need its own unique “model” (probability distribution function)

How does one pick a model?

Which one best fits your data?

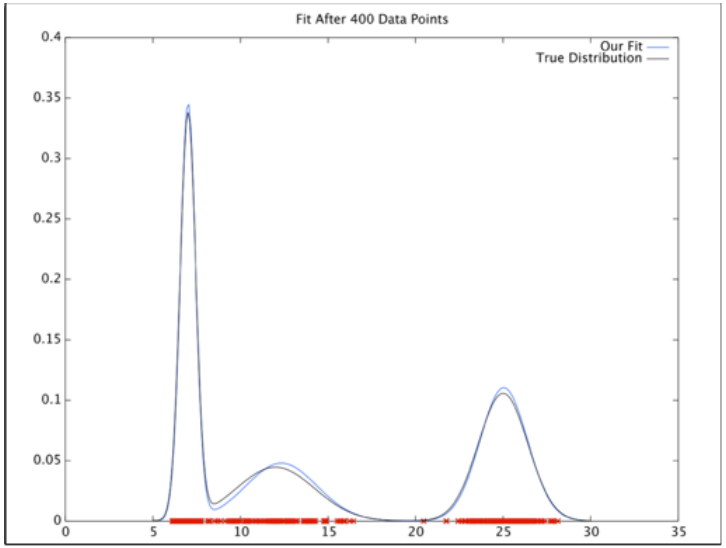


source: “Doing Data Science”
O’Neil & Schutt

elastic

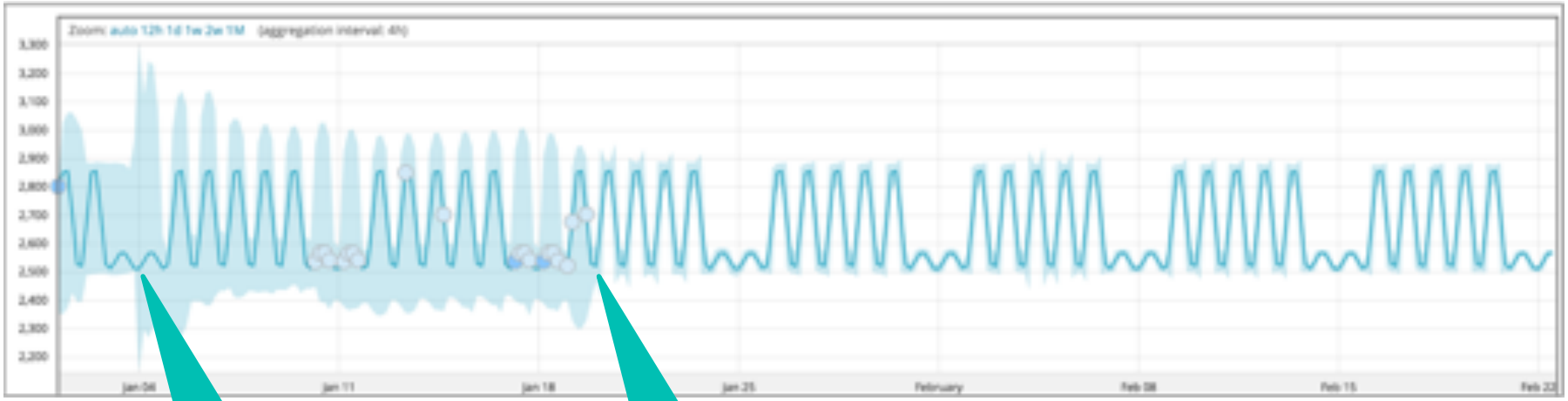
Machine Learning picks it for you

- ML uses sophisticated machine-learning techniques to best-fit the right statistical model for your data.
- Better models = better outlier detection = less false alarms
- Anomalies occur when observation is in low probability area



elastic

The Model's Evolution in Time



Elastic Machine Learning

- Online Unsupervised Learning
- Index Visualizer
- Anomaly Detection
- Forecasting

elastic

Anomaly Detection

- Single or Multi-metric time series
- Outliers in population
- Rare events categorization

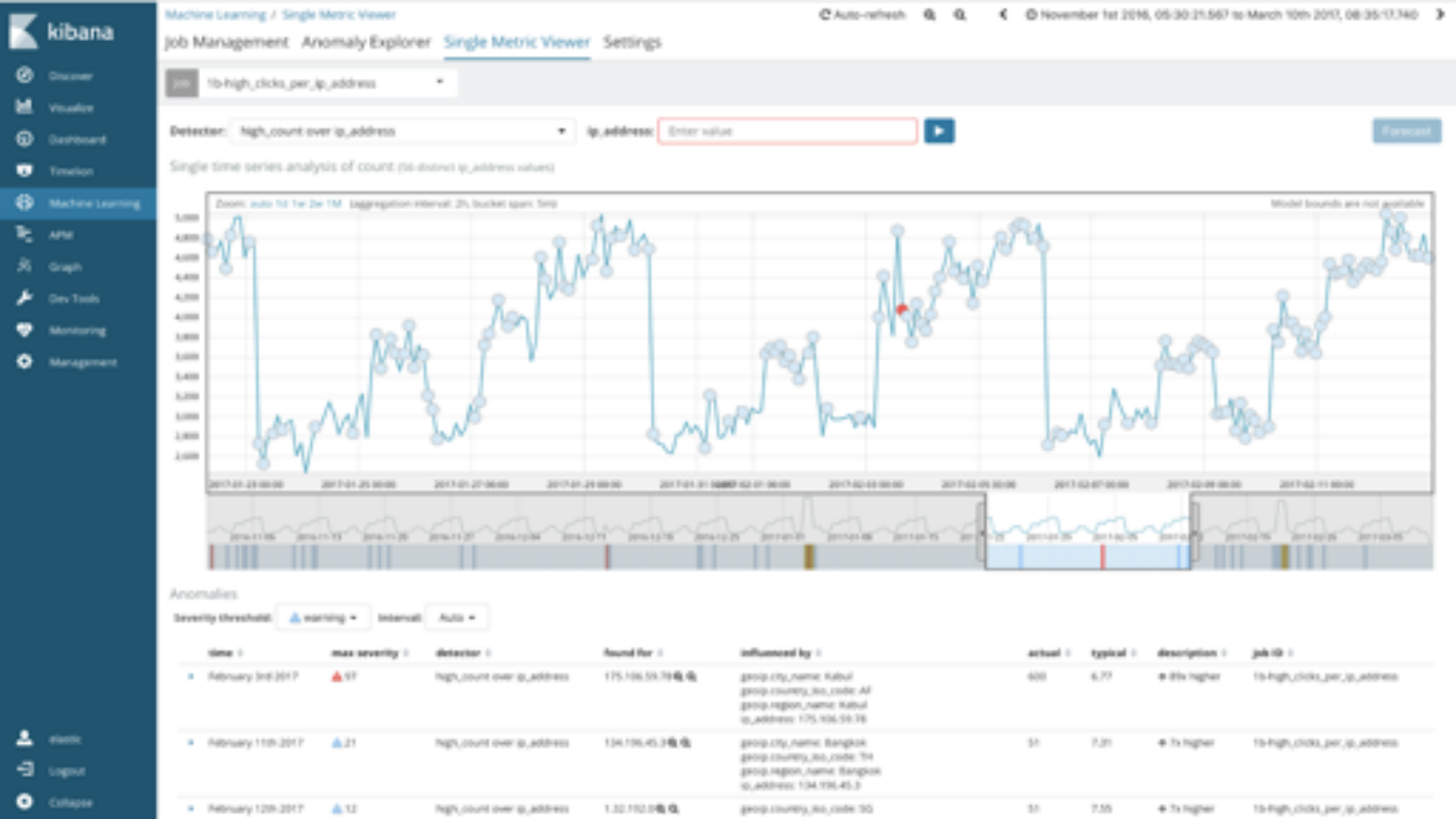
What you can do?

- Drop in Orders?
- Unusual Traffic
- Early hints before things go wrong

- MITM detection
- Malware? Insider Threat?
- Maliciously running processes

- Latency in response times
- Low click through rate on ads

elastic


kibana
Discover
Visualize
Dashboard
Timelion
Machine Learning
APM
Graph
Dev Tools
Monitoring
Management
elastic
Logout
Collapse
Machine Learning / Single Metric Viewer
Auto-refresh
November 1st 2016, 05:30:21.567 to March 10th 2017, 08:35:17.740
Job Management
Anomaly Explorer
Single Metric Viewer
Settings
Job
1b-high_clicks_per_ip_address
Detector:
high_count over ip_address
ip_address:
Enter value
Forecast
Single time series analysis of count (56 distinct ip_address values)
Zoom: auto 1d 1w 2w 1M
Aggregation interval: 2h, bucket span: 5mb
Model bounds are not available
Anomalies
Severity threshold:
warning
Interval:
Auto
time
max severity
detector
found for
influenced by
actual
typical
description
job ID
February 3rd 2017
99
high_count over ip_address
175.106.59.78
geoip.city_name: Kabul
geoip.country_iso_code: AF
geoip.region_name: Kabul
ip_address: 175.106.59.78
600
6.77
89x higher
1b-high_clicks_per_ip_address
February 11th 2017
21
high_count over ip_address
134.196.45.3
geoip.city_name: Bangkok
geoip.country_iso_code: TH
geoip.region_name: Bangkok
ip_address: 134.196.45.3
51
7.31
7x higher
1b-high_clicks_per_ip_address
February 12th 2017
12
high_count over ip_address
1.32.152.0
geoip.country_iso_code: SG
51
7.55
7x higher
1b-high_clicks_per_ip_address

Job: 2a-high_purchase_quantity

Detector: high_mean(total_quantity)-(high_purchase_quantity)-(to

user: Enter value

Forecast

Single time series analysis of avg(total_quantity) (52 distinct user values)

show forecast

Anomalies

Severity threshold: warning

Interval: Auto

No matching results found

kibana

Discover
Visualize
Dashboard
Timelion
Machine Learning
APM
Graph
Dev Tools
Monitoring
Management

elastic
Logout
Collapse

security-analytics-winlogbeat-*

Use full security-analytics-winlogbeat-* data

Search... (e.g. status:200 AND extension:PHP)

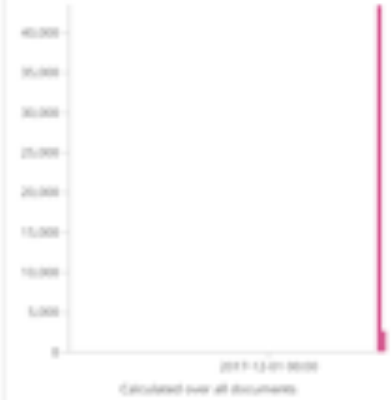
Sample 5000 documents per shard from a total of **46295** documents

Metrics

9 fields exist in documents (**29** in total) show empty fields

Filter

document count



Calculated over all documents

destination_port

43 documents (0.9%)
20 distinct values

min	median	max
0	47,001	60,942

top values

47,001	16.3%
13	14.0%
49,669	9.3%
17	7.0%
123	7.0%
49,664	4.7%
49,665	4.7%
49,678	4.7%
50,350	4.7%
52,191	4.7%

Calculated from sample of 5000 documents per shard

event_id

5000 documents (100.0%)
34 distinct values

min	median	max
1	4,674	53,504

top values

5,140	31.3%
4,634	20.8%
4,674	12.5%
4,624	11.4%
4,688	5.1%
4,103	4.8%
11	3.4%
5	2.8%
4,663	1.4%
5,157	0.9%

Calculated from sample of 5000 documents per shard

process_id

4991 documents (99.8%)
47 distinct values

min	median	max
4	4	8,344

top values

script_block_text_length

2 documents (0.0%)
2 distinct values

min	median	max
80	1,383.5	2,687

top values

source_port

43 documents (0.9%)
24 distinct values

min	median	max
0	47,464	60,546

top values

Create Job

Use the Advanced job wizard to create a job to find anomalies in this data:

Advanced

Use the full range of options to create a job for more advanced use cases.

Elastic Stack - Elasticsearch, Kibana and Logstash

Elastic

Microsoft preferred solution. Learn more

Elastic's addition to the Microsoft Azure Marketplace means developers can utilize preconfigured templates built by Elastic to more easily and quickly deploy an Elasticsearch cluster on Azure. Elastic is the company behind the Elastic Stack, a suite of products that include Elasticsearch, Logstash, Kibana and Beats, which are focused on scalability and ease-of-use to help you make sense of your data.

This is a Bring-Your-Own-License (BYOL) solution template. The BYOL model gives users the option to add critical commercial enhancements such as

- cluster and data security
- user authentication
- stack monitoring
- alerting and notifications
- graph
- machine learning capabilities

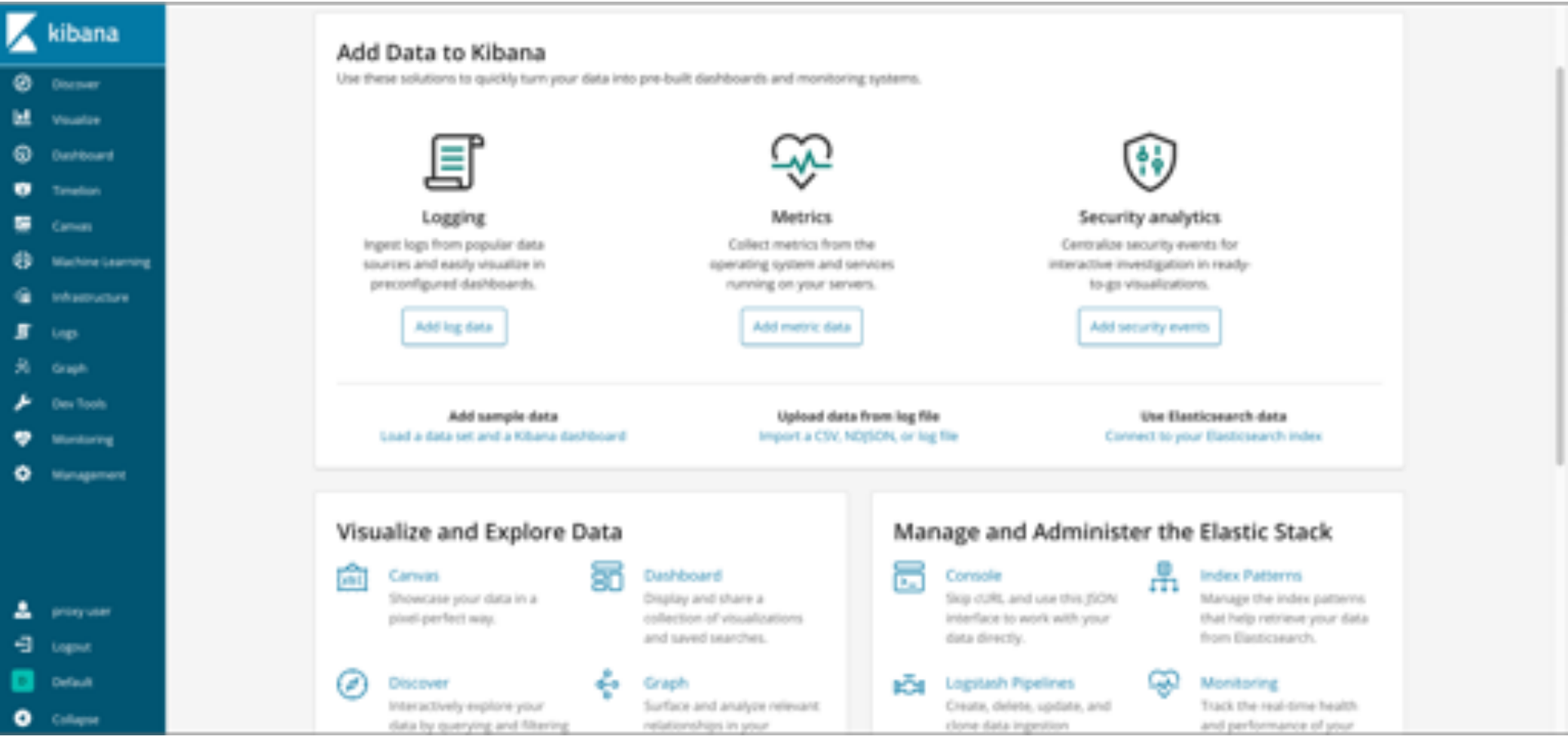
through an Elastic subscription purchased directly from Elastic. This subscription also provides global, Enterprise class support for Elasticsearch on Azure. You can learn more about Elastic Stack features at www.elastic.co.

The solution template installs with a 30 day trial license for Elastic platinum features. For information about an Elastic subscription, contact azuremarketplace@elastic.co. Be sure to also check out our documentation to learn more about this offering.

Elasticsearch cluster with up to 50 data nodes (assuming you have a sufficiently large VM quota), with optional Kibana and Logstash. Elastic Stack platinum features like monitoring, security, alerting, graph and machine learning are automatically activated as part of a 30 day trial license.

Save for later

Create


kibana
Discover
Visualize
Dashboard
Timelion
Canvas
Machine Learning
Infrastructure
Logs
Graph
Dev Tools
Monitoring
Management
proxy user
Logout
Default
Collapse
Add Data to Kibana
Use these solutions to quickly turn your data into pre-built dashboards and monitoring systems.
Logging
Ingest logs from popular data sources and easily visualize in preconfigured dashboards.
Add log data
Metrics
Collect metrics from the operating system and services running on your servers.
Add metric data
Security analytics
Centralize security events for interactive investigation in ready-to-go visualizations.
Add security events
Add sample data
Load a data set and a Kibana dashboard
Upload data from log file
Import a CSV, NDJSON, or log file
Use Elasticsearch data
Connect to your Elasticsearch index
Visualize and Explore Data
Canvas
Showcase your data in a pixel-perfect way.
Dashboard
Display and share a collection of visualizations and saved searches.
Discover
Interactively explore your data by querying and filtering
Graph
Surface and analyze relevant relationships in your
Manage and Administer the Elastic Stack
Console
Skip cURL and use this JSON interface to work with your data directly.
Index Patterns
Manage the index patterns that help retrieve your data from Elasticsearch.
Logstash Pipelines
Create, delete, update, and clone data ingestion
Monitoring
Track the real-time health and performance of your