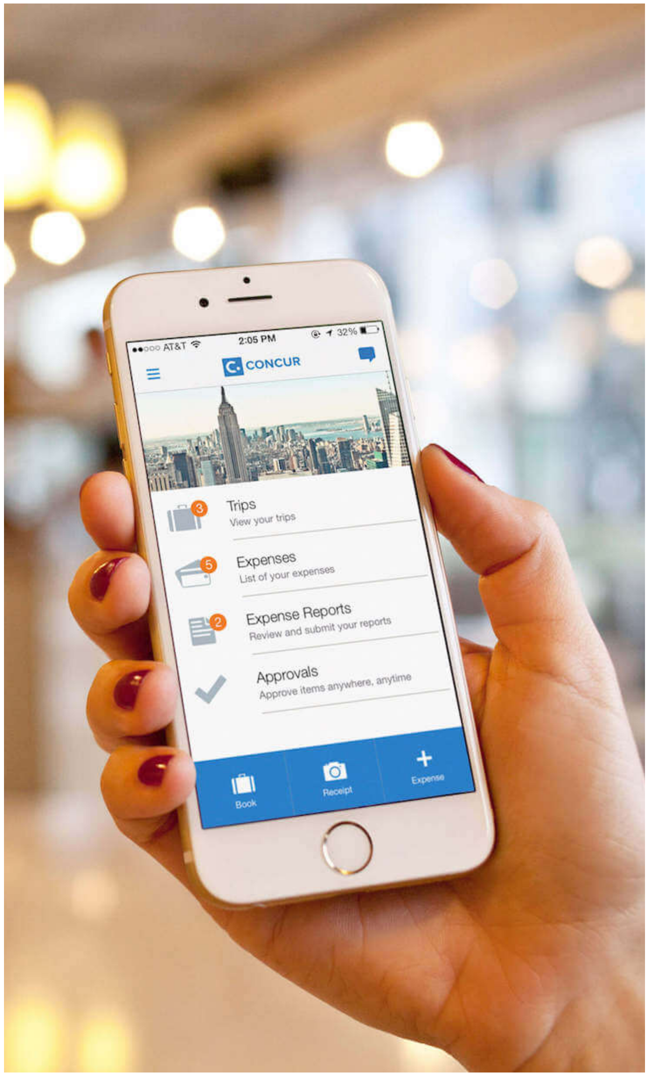
Machine Learning and Data Science for Performance and Quality Engineering

Gopal Brugalette
Principal Software Engineer
SAP Concur



SAP Concur

A busy day @ SAP Concur

183,000 trips booked

409,000 expense reports

1 million mobile logins

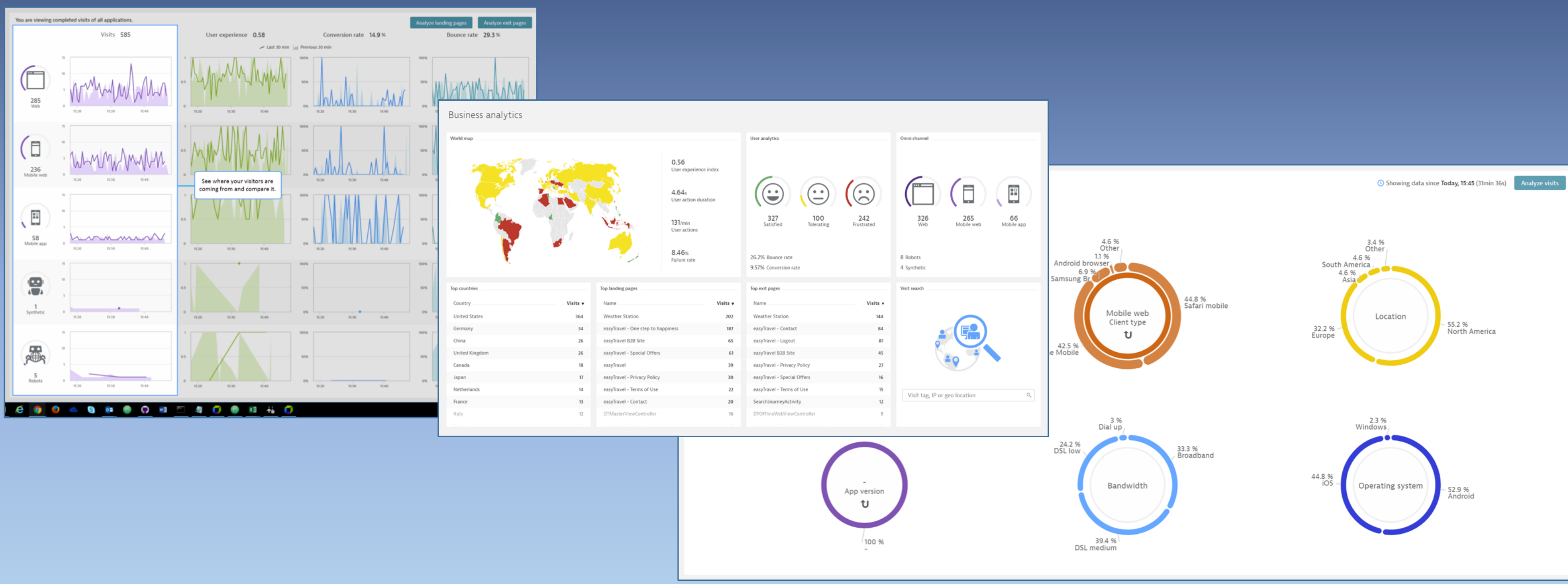
760,000 mobile receipts uploaded

32,000 clients, 100 countries

Gopal Brugalette

Principal Engineer, Performance

Performance Engineering is a Data Science



Data → Information → Insight → Action

What is Machine Learning?

Math enabling computers to do a what a human can-

Derive insights from data in a specific situation

1. $\nabla \cdot \mathbf{D} = \rho_V$
2. $\nabla \cdot \mathbf{B} = 0$
3. $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
4. $\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}$

Understand the problem, pick the algorithm

- What is the question?
- Machine learning algorithms
  - Supervised
    - Build a model using past data to make future predictions
  - Unsupervised
    - Understand the structure of the data, with no past data to compare

Regression analysis for prediction

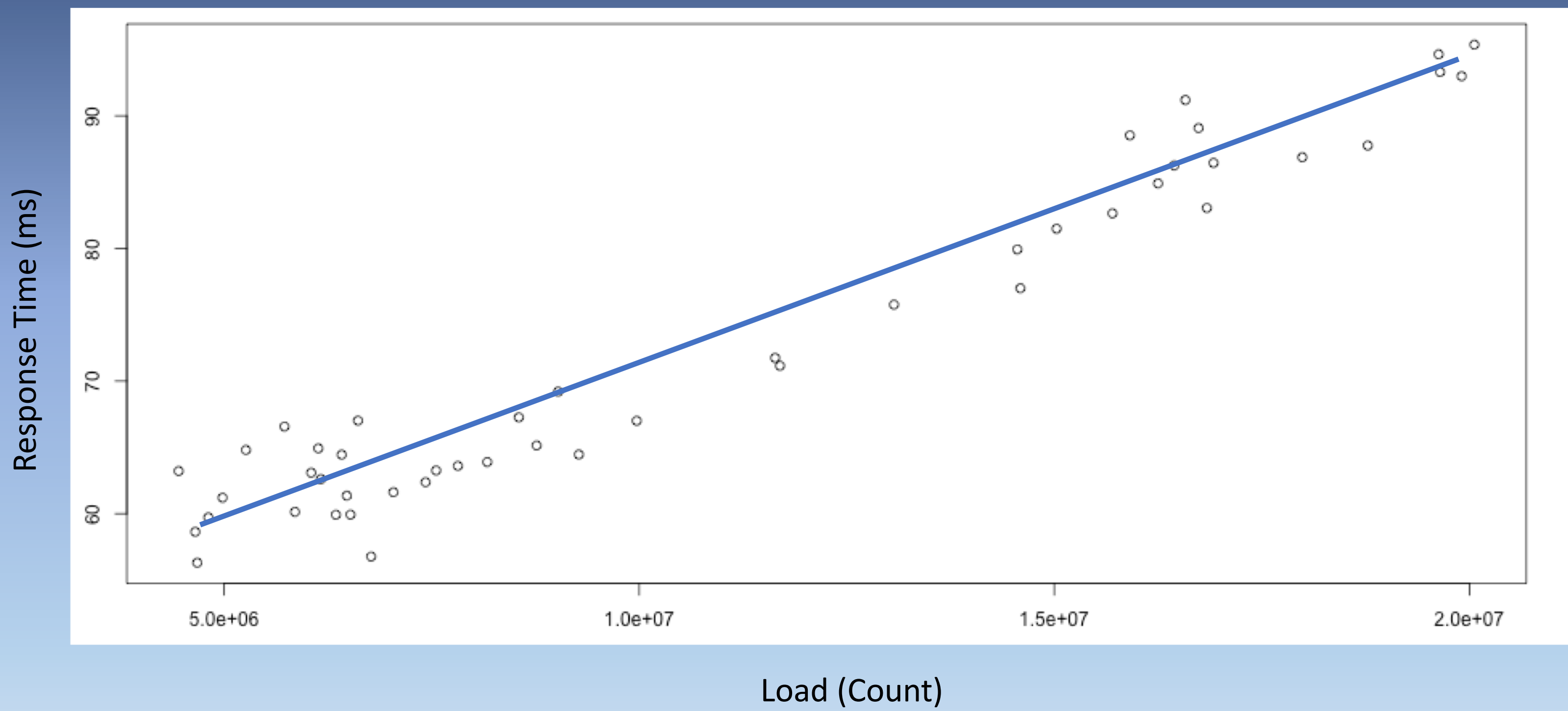
Goal: Build a Model [y=f(x)] to understand customer experience

Feature Engineering:

1. Understand your data
2. Look for dimensions or features
3. How are they related?

1	timestamp	Count	CPU	Memory	p25
2	1485156599	44759	58	67	214.0686
3	1485158399	49323	85	80	217.2732
4	1485160199	51611	61	58	219.4307
5	1485161999	53694	62	62	230.5242
6	1485163799	53590	67	57	224.6855
7	1485165599	48087	53	70	227.9708
8	1485167399	47291	53	94	234.5585
9	1485169199	44979	57	81	233.384
10	1485170999	47599	61	94	220.8943
11	1485172799	52629	57	56	215.9142
12	1485174599	66170	70	73	223.6415
13	1485176399	87112	53	74	243.2343
14	1485178199	112592	65	66	265.945
15	1485179999	135154	68	85	298.7804
16	1485181799	151021	72	51	336.3092
17	1485183599	162062	99	64	284.7538
18	1485185399	170519	96	97	278.3604
19	1485187199	171152	71	78	270.8226
20	1485188999	163063	63	61	263.2569
21	1485190799	145117	88	53	248.1356
22	1485192599	136043	51	64	238.2139
23	1485194399	130291	65	54	237.8973

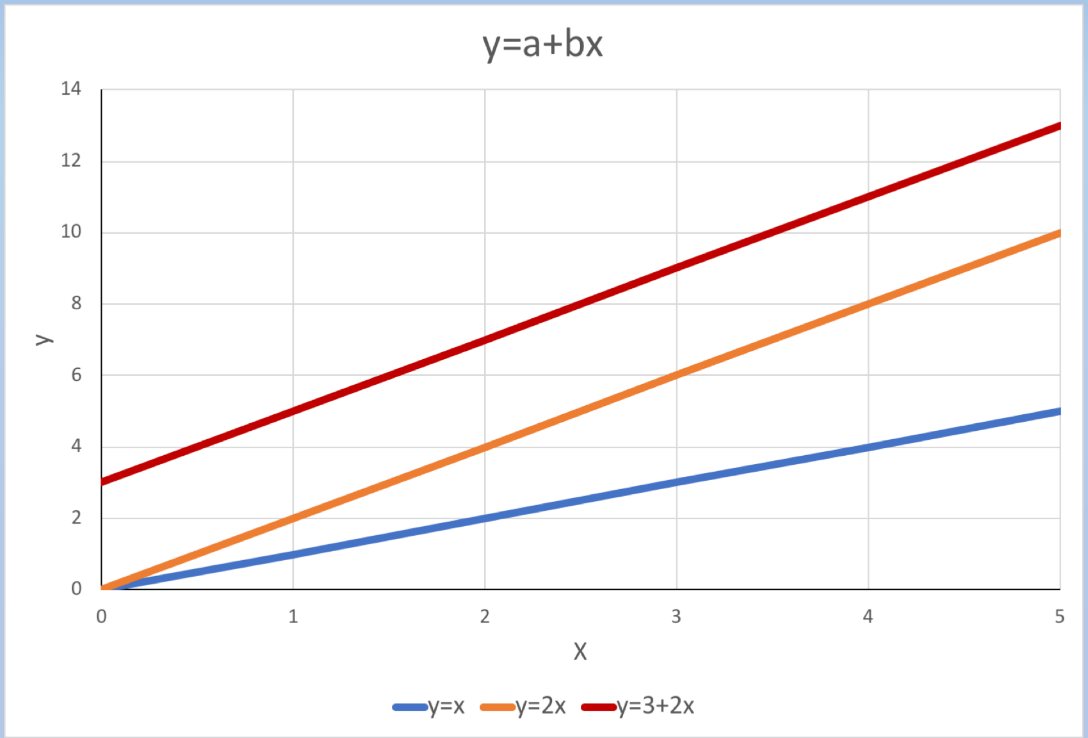
Predict Response Time based on Load



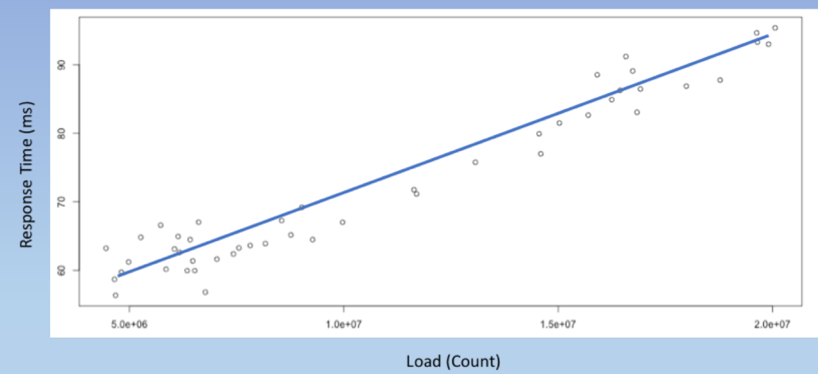
A little math

- *Response time = f(Load)*
- A linear model fits the data
  - $y = a + bx$
  - *Examples*

x	y=x	y=2x	y=3+2x
0	0	0	3
1	1	2	5
2	2	4	7
3	3	6	9
4	4	8	11
5	5	10	13



A little more math



- *Response time = f(Load)*
- A linear model fits the data
  - *y = a + bx*
- *Count is x, Response time (pnn) is y*
- Solve for a and b
  - $a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x \sum y)^2}$
  - $b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$
- *Response time = constant + factor * Load*

Count	p25	p50	p75	p95	p99
44759	214.0686	301.0763	435.0048	912.0915	1732.1838
49323	217.2732	308.0325	450.0636	947.6567	1736.1739
51611	219.4307	311.6029	446.6124	927.6483	1663.055
53694	230.5242	327.7388	477.4424	1009.7936	2065.0375
53590	224.6855	317.2768	453.6071	928.3359	1819.799
48087	227.9708	318.1334	455.3174	934.8043	1721.309
47291	234.5585	341.2053	538.0538	1151.8259	2165.4265
44979	233.384	335.0738	513.9694	1127.9793	2276.5243
47599	220.8943	313.1973	470.1585	1006.3628	1849.6871
52629	215.9142	299.0448	432.6091	933.8607	1847.893
66170	223.6415	318.3739	511.4039	1151.6963	2277.8392
87112	243.2343	355.4192	612.5432	1502.2516	3062.444
112592	265.945	403.0888	692.3006	1765.0567	3348.6161
135154	298.7804	487.5251	895.0957	2446.268	4311.8653
151021	336.3092	598.2396	1094.7428	2913.9519	5072.793
162062	284.7538	382.7318	592.8598	1466.7603	3462.5878
170519	278.3604	363.9555	521.2324	1088.573	2261.4422
171152	270.8226	351.8303	497.7192	1023.7396	2072.0174
163063	263.2569	343.124	488.4148	1025.0027	2052.6454
145117	248.1356	319.2189	448.8661	938.2041	1832.6257
136043	238.2139	306.4644	431.255	924.5088	1926.0287
130291	237.8973	306.6233	430.9984	912.6513	1862.1043
131844	239.291	308.657	430.7724	901.0013	1853.0945
129200	249.9657	326.0272	467.8967	1102.1019	3103.9653
132239	238.8009	308.9355	438.5559	971.7552	2112.7195
125707	232.5378	299.0806	420.0991	900.5856	1947.956
124926	230.4473	295.8031	416.9463	915.8902	1898.809

A little code

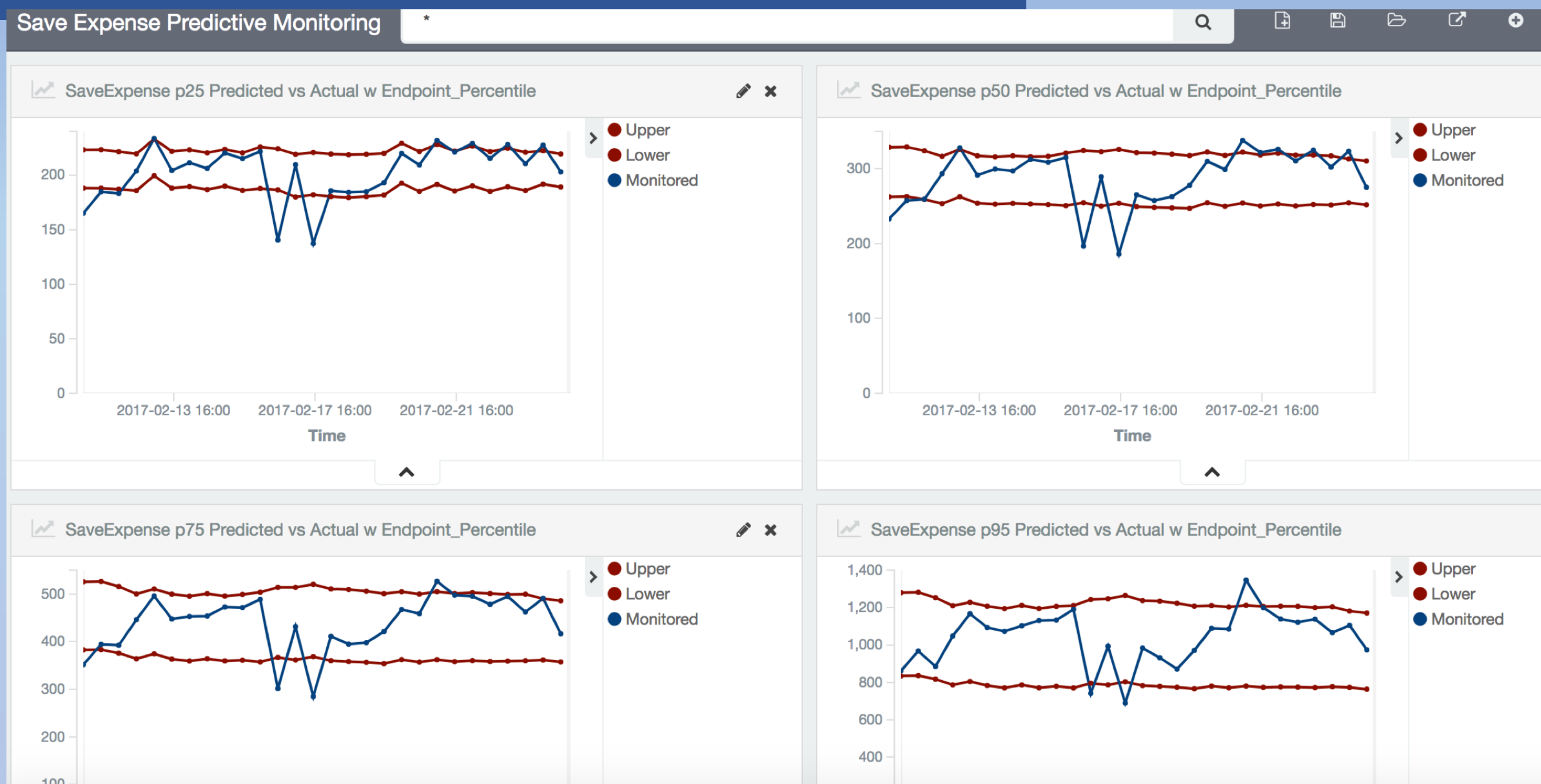
```r
for(model in rt_model_names) {
  f <- paste(model, "~", "Count")
  modelpnn <- paste(endpoint, model, sep = '')
modelset[[modelpnn]] <- lm(f, data=model_numbers)
}
```

Train the models

Make a prediction

```r
 lastmonitorprediction <- t(data.frame(predict(modelset[[e_m]],
        data.frame(Count=monitor_numbers$Count), interval="prediction")[2:3]))
```

Predictive Modeling of Response Times



Model predicts response times (red) based on measured load and compares it to monitored [actual] response times (blue)

Regression Model Use Cases

- Evaluate Changes in Production before peak load
- Less than optimal performance
- Normalize Performance for Load
- Detecting Outages

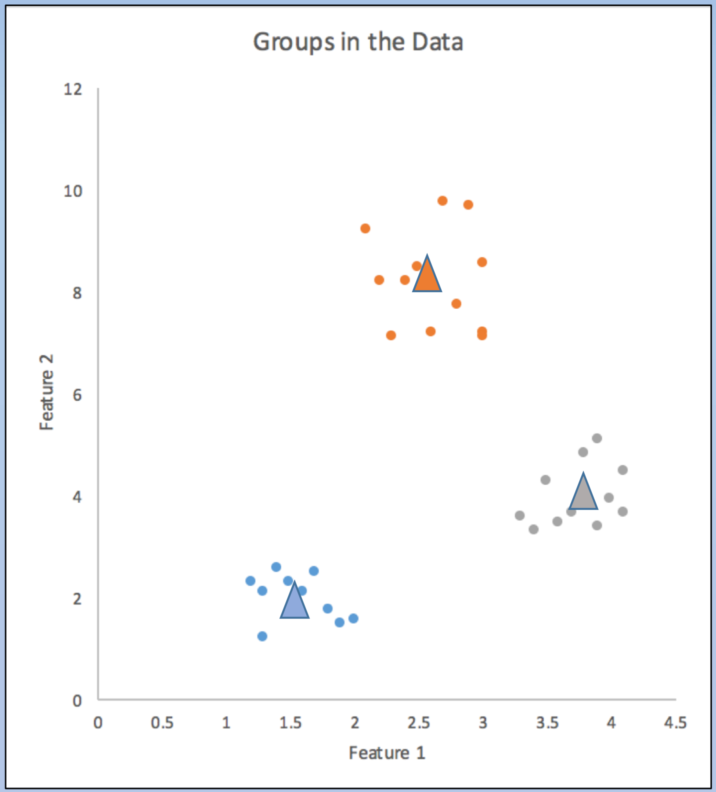
The hard parts

- Pick an algorithm & model
- Feature engineering
- Data wrangling
- Training set
- Productionizing

Clustering

- Kmeans Clustering
  - Unsupervised
  - Segments data by similarity of features into *K* number of groups

- Algorithm
  - Select center for each of K groups (centroid)
  - Assign each point to nearest centroid
  - Calculate new center as mean of points in the centroid
  - Iterate

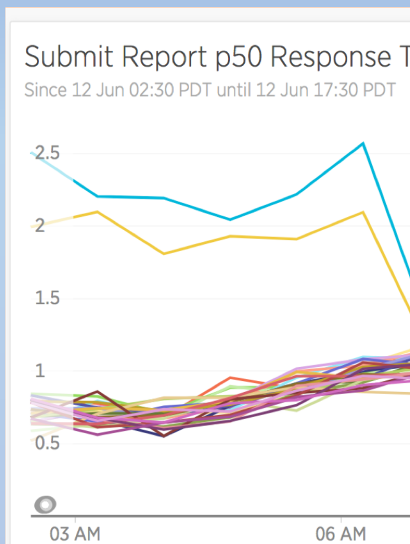


A little code

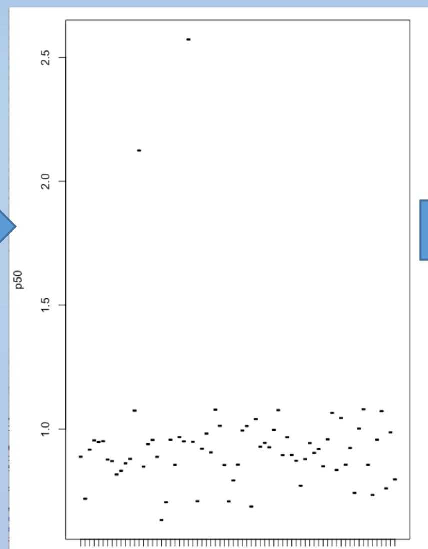
```
17 serverinfo <- read.csv("submitreport.csv")
18 serverinfo <- serverinfo[,-3]
19 server.cluster <- kmeans(serverinfo[,2], 2, nstart=20)
20 server.cluster
```

- 17 Read the data
- 18 Clean it up
- 19 Execute Kmeans for *K*=2
- 20 Print it out

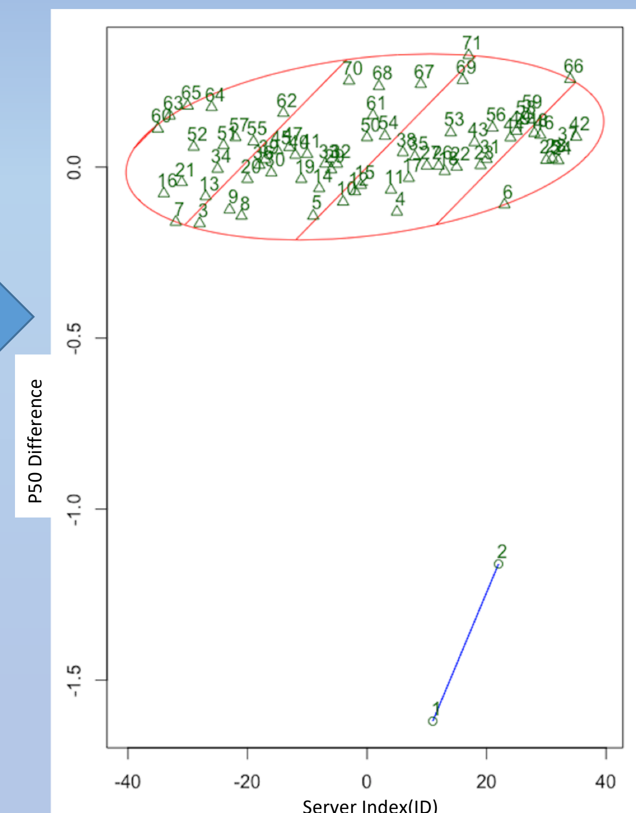
KMeans Clustering of server performance



Lines represent a server

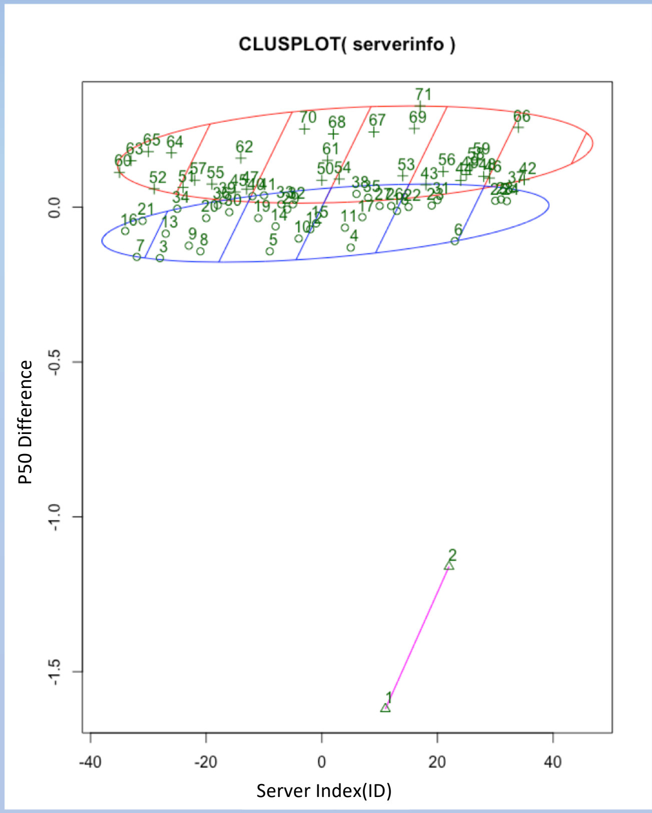


Server Index(ID)



Automatically finds different groups

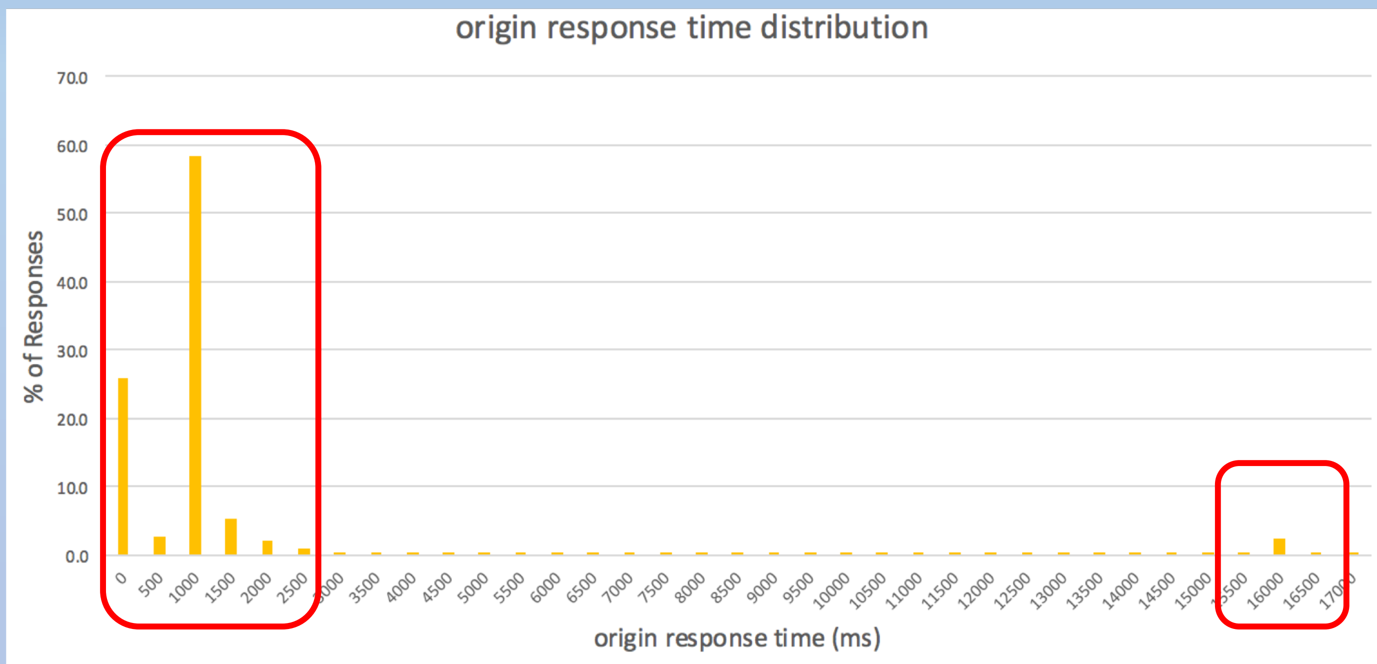
K= 3



K =3 identified unique clusters

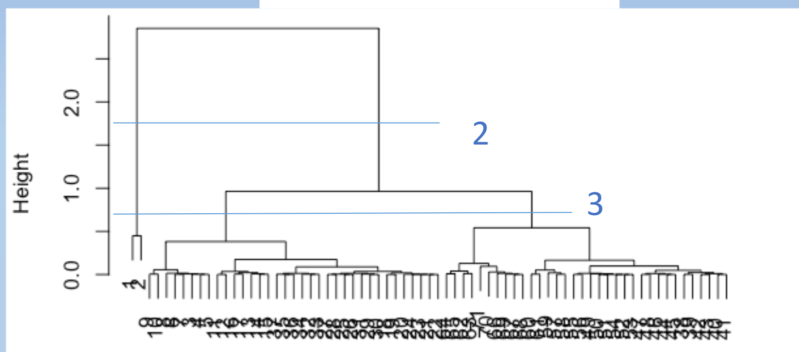
- Red Cluster is Data Center Server Group A
- Blue Cluster is Data Center Server Group B
- Purple is Servers with Power Saving On

Clustering to look for multi-modal distributions



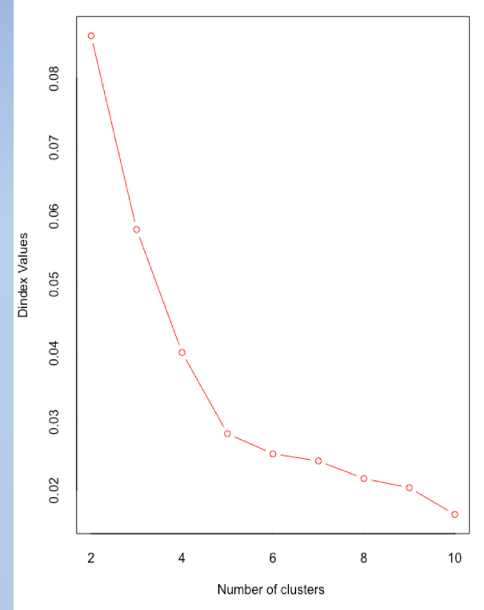
How many clusters?

hierarchical clustering




```r
c.dist <- dist(serverinfo[,2], method = "euclidean")
h.fit <- hclust(c.dist, method = "ward.D2")
plot(h.fit)
```

Elbow Method



Various Calculated Methods

	KL	CH	Hartigan	CCC	Scott	Marriot
Number_clusters	2.0000	5.0000	3.0000	2.0000	3.0000	5.0000

Libraries do it for you

```r
servers.nb <- NbClust(serverinfo$p50, distance = "euclidean", min.nc = 2,
                      max.nc = 8, method = "ward.D2", index ="all")
num.clusters <- length(unique(servers.nb$Best.partition))
```

Big Data & Data Science

- Large data sets needed
- Visualization needed
- Where does it all go?
- Get it to people?
- More data is better?

The Use Case

Question

- How can we understand the Customer Experience (performance) over time?

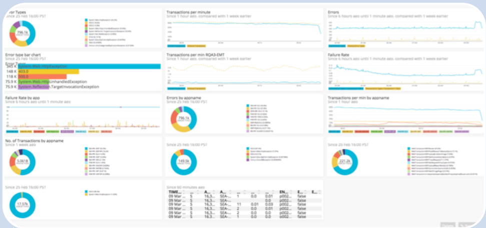
Requirements

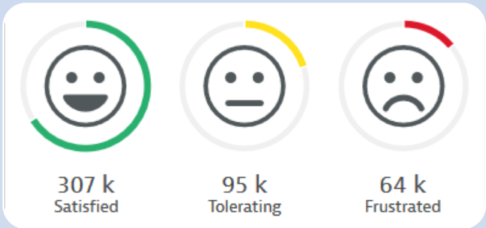
- Leadership reporting
- Long-term trending
- Agile/Dev-Op team accountability

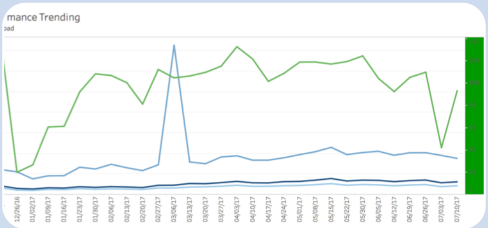
Approach

- KPI’s and derived metrics
- Long Term Storage
- Easy Visualization

Approach Iterations







Dashboard Dump

- Too much data
- Too many questions

Apdex Overload

- What does it mean
- Where's the insight

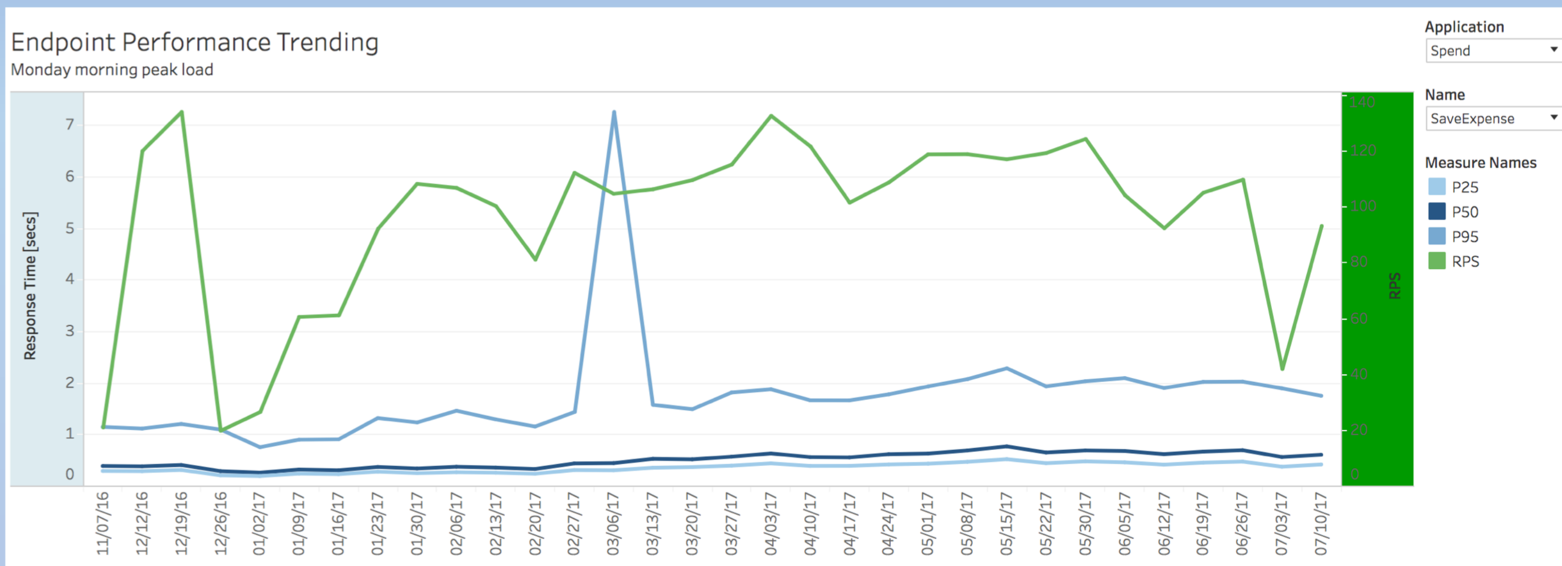
A simple metric, trended over time

- Peak hour performance, week over week
- Easy to get
- Easy to understand

Data → Information → Insight → Action

Performance Metric Trending

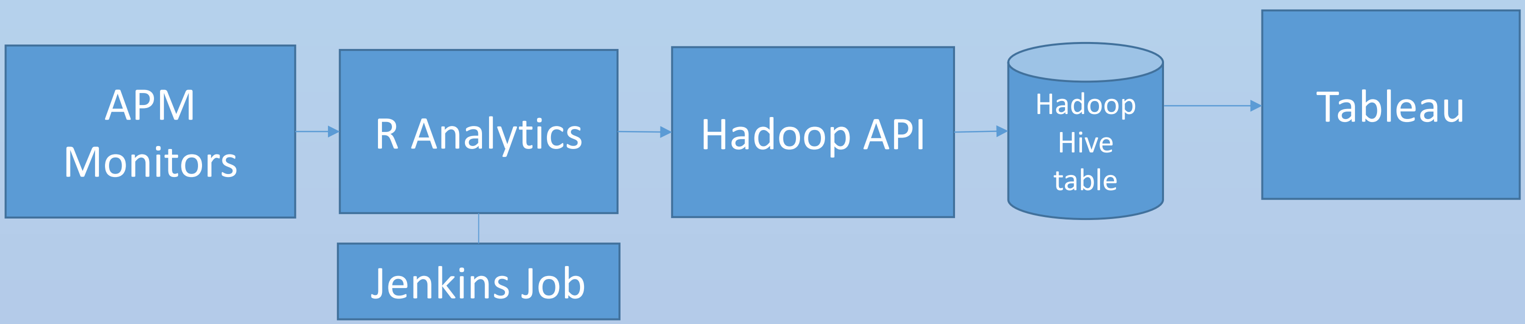
A Tableau Dashboard



Long term trending of customer experience through key endpoint performance

- Response time Distributions (25%, 50%, 95%)
- Peak Hour – Monday Morning 7-8 AM PST

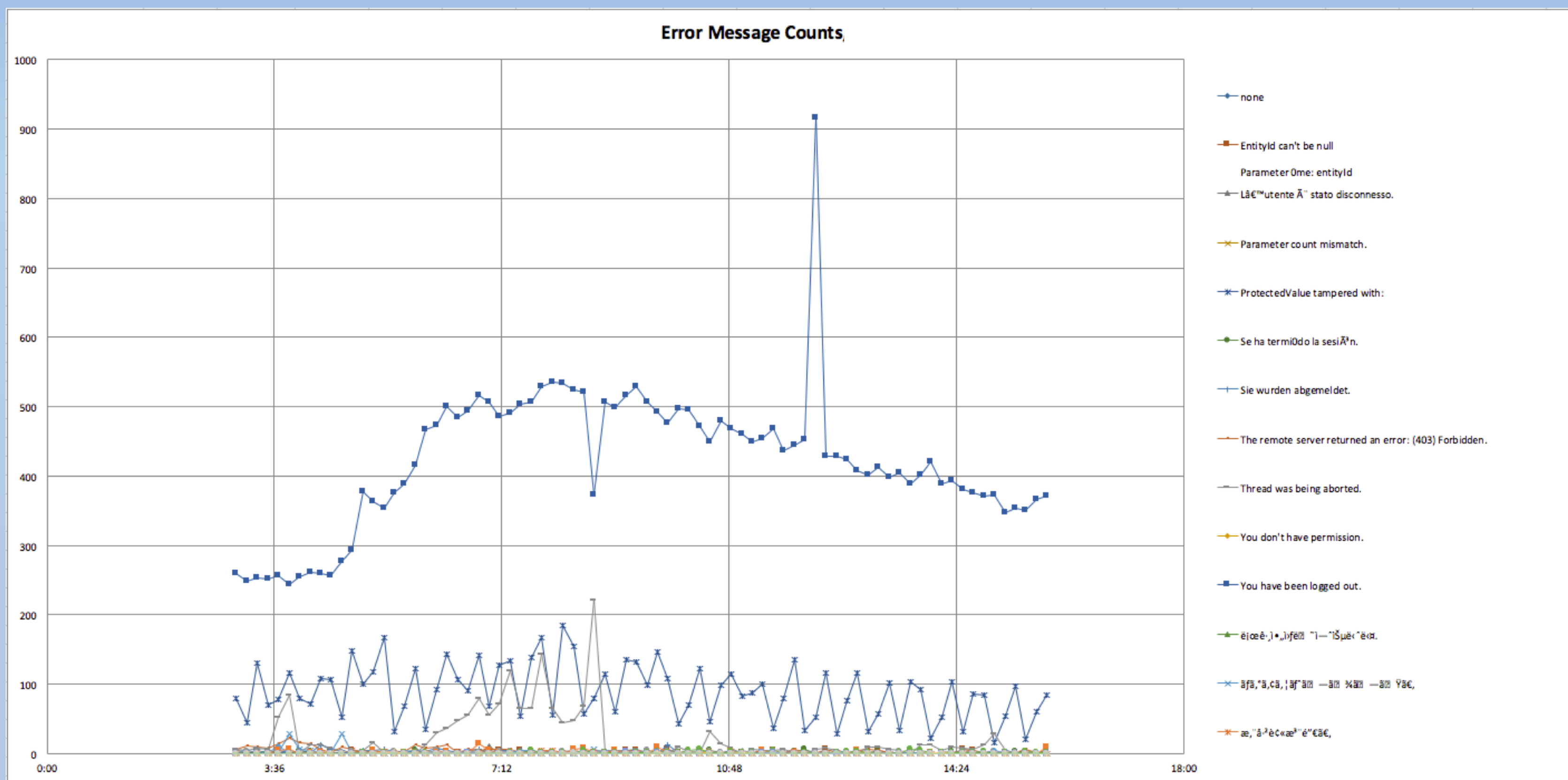
Solution Architecture



Machine Learning for Outage RCA

- Question
  - Can we use ML for Root Cause Analysis and Prediction of major system outages?
- Premise
  - Application error logs contain sufficient information to detect an issue
  - Application error logs contain sufficient details to identify and distinguish between system failure modes
- But is this true?

Error Log Counts and Correlations

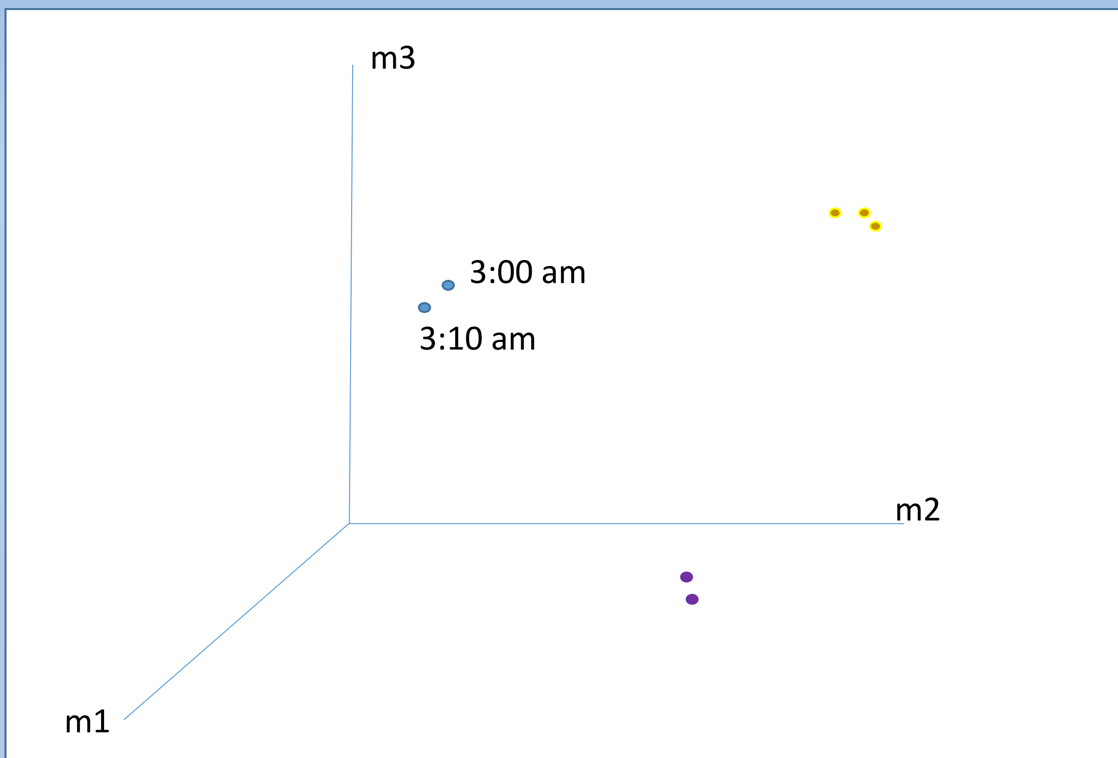


Form a Vector

- Count messages in small time slices
- Each time slice forms a message count vector
  - 3:00 am (0,2,2,1,3,6,6,1,260,...)
  - 3:10 am (0,0,1,0,45,3,5,11,5,0,249,...)
- Normalize for load

message	3:00:00 AM	3:10:00 AM	3
none	0	0	
	2	0	
Lâ€™utente Ã¨ stato disconnesso.	2	1	
Parameter count mismatch.	1	0	
ProtectedValue tampered with:	80	45	
Se ha termi0do la sesiÃ³n.	2	3	
Sie wurden abgemeldet.	3	5	
The remote server returned an error: (403) Forbidden.	6	11	
Thread was being aborted.	6	5	
You don't have permission.	1	0	
You have been logged out.	260	249	
ë¡œê·¸ì•„ì›ƒë�˜ì—ˆìŠµë‹ˆë‹¤.	1	0	
ãƒã‚°ã‚¢ã‚¦ãƒˆã�—ã�¾ã�—ã�Ÿã€‚	2	2	
æ‚¨å·²è¢«æ³¨é”€ã€‚	2	0	
La session est terminÃ©e.	0	3	
U bent afgemeld.	0	1	
Ð’Ñ‹Ð¿Ð¾Ð»Ð½ÐµÐ½ Ð²Ñ‹Ñ…Ð¾Ð´.	0	1	
Received empty ArHeader	0	0	
Error reading JObject from JsonReader. Path '', line 0, position 0.	0	0	
Object reference not set to an instance of an object.	0	0	
A sua sessÃ£o foi encerrada.	0	0	
	0	0	
Jste odhlÃ¡Å¡eni.	0	0	
Ha cerrado sesiÃ³n.	0	0	
OstÄ…piÅ‚o wylogowanie.	0	0	
æ‚¨å·²ç¶“ç™»å‡ºã€‚	0	0	
Du Ã¤r nu utloggad.	0	0	
Task or EntityCode must be defined	0	0	
You do not have appropriate role to Advance Request Workflow	0	0	
	0	0	
' ', hexadecimal value 0x16, is an invalid character.	0	0	
' ', hexadecimal value 0x0B, is an invalid character.	0	0	
' ', hexadecimal value 0x10, is an invalid character. Line 79, position 7.	0	0	
Input string was not in a correct format.	0	0	
Index was outside the bounds of the array.	0	0	
Bad JSON escape sequence: \o. Path '[0].data[5]', line 1, position 111.	0	0	

Message Space



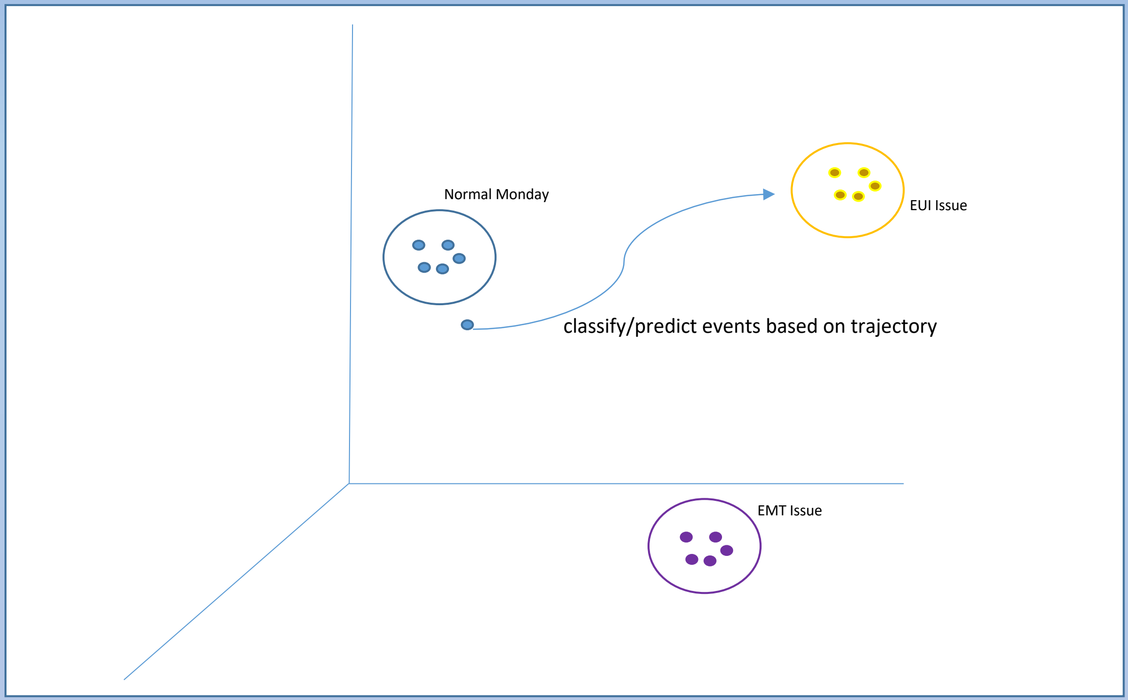
Message count vectors define events in a message space

Train the model

- Classify these points as different events
  - “Normal”, “DB Issue:, “App Server”, etc.
- Train the analysis engine to recognize these events
- Use K*nn* or other classifier to identify what type of event is occurring in real time
- Improve Root Cause Analysis

Classify and predict events

- Identify RCA
- Predict/Prevent Issues



Intelligent Clustering of Error Messages

- Can we group messages based on similarity?

- Method:
  - Clean messages
  - Create a DTM (document term matrix)
  - Kmeans- Clustering to group messages
- While this works, there is very little semantic similarity between messages.

Clustering them in this way was not valuable.

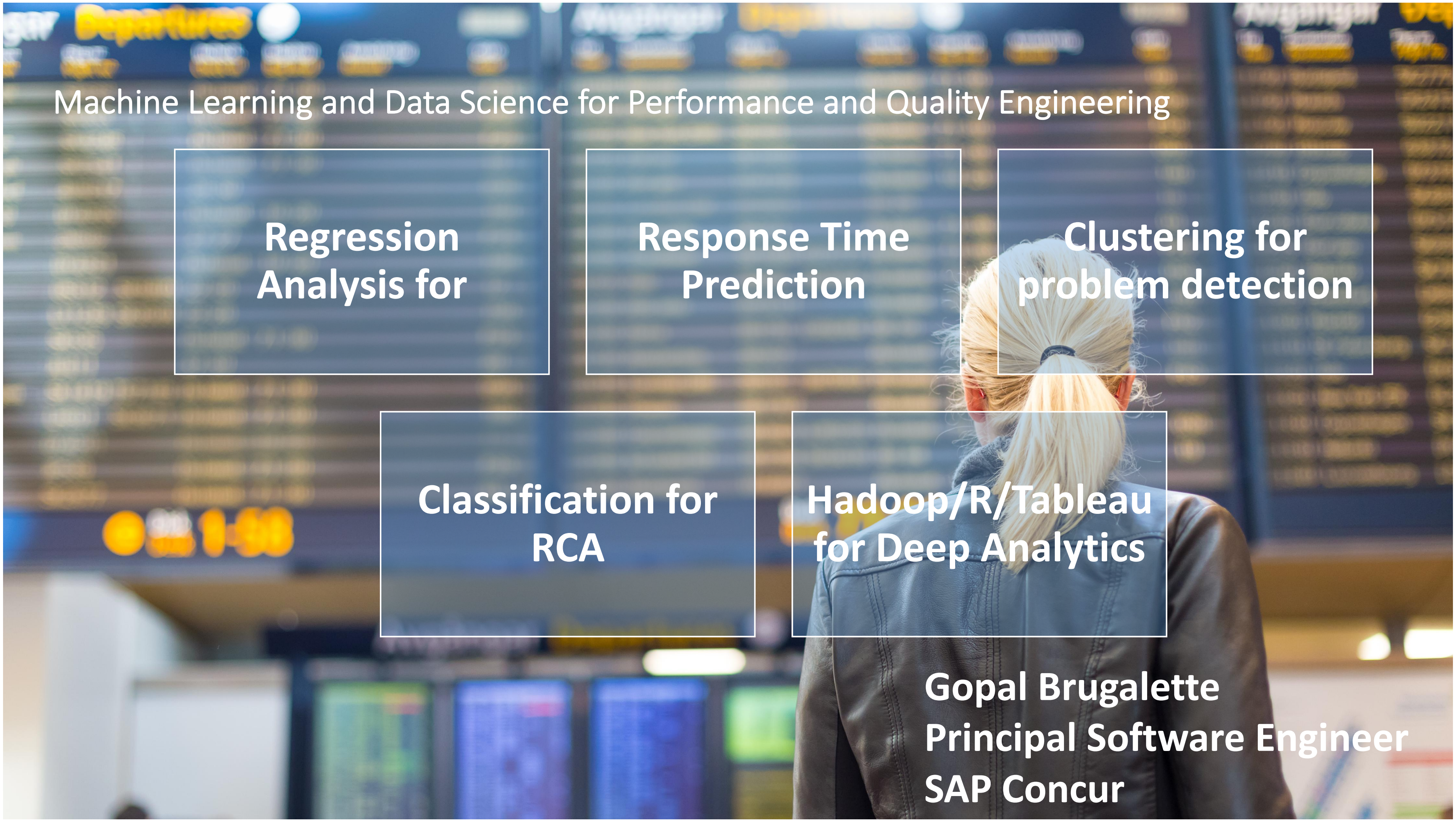
Message DTM

	abgemeldet	aborted	advance	afgemeld	appropriate	är	arheader	bad	bent	blevet	can	cerrado	character	count	disconnesso	don
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thread was being aborted.	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
You have been logged out.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Clustered Messages

Console ~/git/R-reports/ServerAnalysis/

```
ProtectedValue tampered with: gWiDX9t7nHbCBdQlHOsga4yOQ7ZiESQ[4]
5
ProtectedValue tampered with: gWjzpUl0NEVlRw7FOpLbXj6BJ6RCvygRfNA[4]
5
You have been logged out.
2
You have been logged out.
2
You have been logged out.
2
You have been logged out.
2
You have been logged out.
2
You have been logged out.
2
ProtectedValue tampered with: gWiOGOy$sRwInoxw9sjgAeZG9L2$suzm6Y[4]
5
You have been logged out.
2
You have been logged out.
2
You have been logged out.
2
ProtectedValue tampered with: gWkMTh$pJGPClNIXUhiK2YV$puvfXjFY9uzI0s[4]
5
ProtectedValue tampered with: gWkt0$so4hn3pp9RZQwGgUkKIQjokWmoQ[4]
5
ProtectedValue tampered with: gWkt0$so4hn3pp9RZQwGgUkKIQjokWmoQ[4]
5
ProtectedValue tampered with: gWkt0$so4hn3pp9RZQwGgUkKIQjokWmoQ[4]
5
ProtectedValue tampered with: gWmye1yF6dG7W3C$sEQbqQt$pNekoRcwxx8Sg[4]
1
EntityId can't be null\r\nParameter name: entityId
6
You have been logged out.
2
Thread was being aborted.
4
You have been logged out.
2
```


Machine Learning and Data Science for Performance and Quality Engineering
Regression Analysis for
Response Time Prediction
Clustering for problem detection
Classification for RCA
Hadoop/R/Tableau for Deep Analytics
Gopal Brugalette
Principal Software Engineer
SAP Concur