So You Want To Be A Data Scientist?

Michael Lowe
IBM UK Limited

November 2018
Session IF

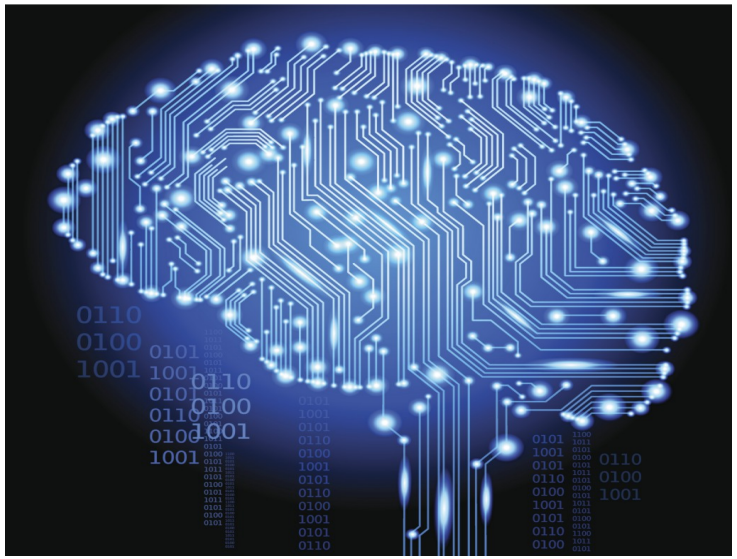
Notices and disclaimers

Legal Disclaimer

- © IBM Corporation 2017. All Rights Reserved.
- The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.
- References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.
- If the text contains performance statistics or references to benchmarks, insert the following language; otherwise delete:
  Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.
- If the text includes any customer examples, please confirm we have prior written approval from such customer and insert the following language; otherwise delete:
  All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.
- Please review text for proper trademark attribution of IBM products. At first use, each product name must be the full name and include appropriate trademark symbols (e.g., IBM Lotus® Sametime® Unyte™). Subsequent references can drop "IBM" but should include the proper branding (e.g., Lotus Sametime Gateway, or WebSphere Application Server). Please refer to http://www.ibm.com/legal/copytrade.shtml for guidance on which trademarks require the ® or ™ symbol. Do not use abbreviations for IBM product names in your presentation. All product names must be used as adjectives rather than nouns. Please list all of the trademarks that you use in your presentation as follows; delete any not included in your presentation. IBM, the IBM logo, Lotus, Lotus Notes, Notes, Domino, Quickr, Sametime, WebSphere, UC2, PartnerWorld and Lotusphere are trademarks of International Business Machines Corporation in the United States, other countries, or both. Unyte is a trademark of WebDialogs, Inc., in the United States, other countries, or both.
- If you reference Adobe® in the text, please mark the first use and include the following; otherwise delete:
  Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
- If you reference Java™ in the text, please mark the first use and include the following; otherwise delete:
  Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
- If you reference Microsoft® and/or Windows® in the text, please mark the first use and include the following, as applicable; otherwise delete:
  Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.
- If you reference Intel® and/or any of the following Intel products in the text, please mark the first use and include those that you use as follows; otherwise delete:
  Intel, Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- If you reference UNIX® in the text, please mark the first use and include the following; otherwise delete:
  UNIX is a registered trademark of The Open Group in the United States and other countries.
- If you reference Linux® in your presentation, please mark the first use and include the following; otherwise delete:
  Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both. Other company, product, or service names may be trademarks or service marks of others.
- If the text/graphics include screenshots, no actual IBM employee names may be used (even your own), if your screenshots include fictitious company names (e.g., Renovations, Zeta Bank, Acme) please update and insert the following; otherwise delete: All references to [insert fictitious company name] refer to a fictitious company and are used for illustration purposes only.

Agenda

- What is Data Science?
- What skills are needed?
- A peek at Machine Learning
- Data handling
- In-transaction model scoring
- A Word or Two about Spark
- Q & A



What is Data Science?

What is Data Science?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining. (Wikipedia)

Data Science: The sexiest job of the 21st century
HBR October 2012

"I think data-scientist is a sexed up term for a statistician"

"Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn't berate the term statistician."

Nate Silver August 2013. A journalist who correctly predicted the winner of all 50 states in the 2012 Presidential election.

What is Data Science?

THE FUTURE OF PAML* IS THE THOUSAND-MODEL VISION

The importance of data science cannot be understated. It is the electricity of artificial intelligence, the butterfly effect of the insights-driven business, and the chemical reaction of scalable intelligence across the enterprise.

The Forrester Wave™: Multimodal Predictive Analytics And Machine Learning Solutions, Q3 2018

*PAML = Predictive Analytics & Machine Learning

Data Science Jobs

September 6 – within 100 miles of London

Salary Estimate	
£35,000	(2283)
£40,000	(2005)
£50,000	(1428)
£55,000	(1138)
£65,000	(629)

Job Type	
Full-time	(1122)
Permanent	(808)
Contract	(168)
Temporary	(58)
Part-time	(30)

Title	
Senior Data Scientist	(87)
Data Engineer	(43)
Lead Data Scientist	(39)
Machine Learning Engineer	(31)
Quantitative Analyst	(21)
Software Engineer	(13)
Data Analyst	(13)
Process Development Tec…	(12)
Senior Scientist	(11)
Senior Data Engineer	(11)
Research Scientist	(11)
Python	(10)

Source: www.indeed.co.uk – 7,483 jobs, mean salary £44,675

What skills are needed?

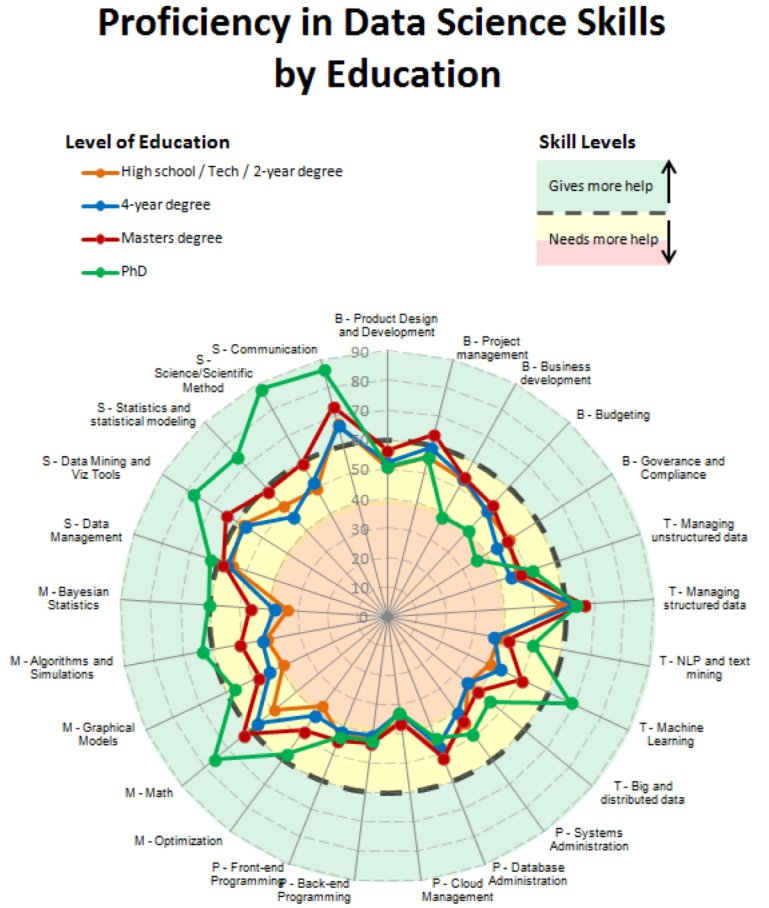
Academic Background

Data scientists are highly educated – 88% have at least a Master's degree and 46% have PhDs.

To become a data scientist, you could earn a Bachelor's degree in Computer science, Social sciences, Physical sciences, and Statistics. The most common fields of study are Mathematics and Statistics (32%), followed by Computer Science (19%) and Engineering (16%).

After your degree programme, you are not done yet. The truth is, most data scientists have a Master's degree or Ph.D and they also undertake online training to learn a special skill like how to use Hadoop or Big Data querying. Therefore, you can enrol for a master's degree program in the field of Data science, Mathematics, Astrophysics or any other related field. The skills you have learned during your degree programme will enable you to easily transition to data science.

https://www.kdnuggets.com/2018/05/simplilearn-9-must-have-skills-data-scientist.html



Tools and Languages

Top Analytics/Data Science Tools

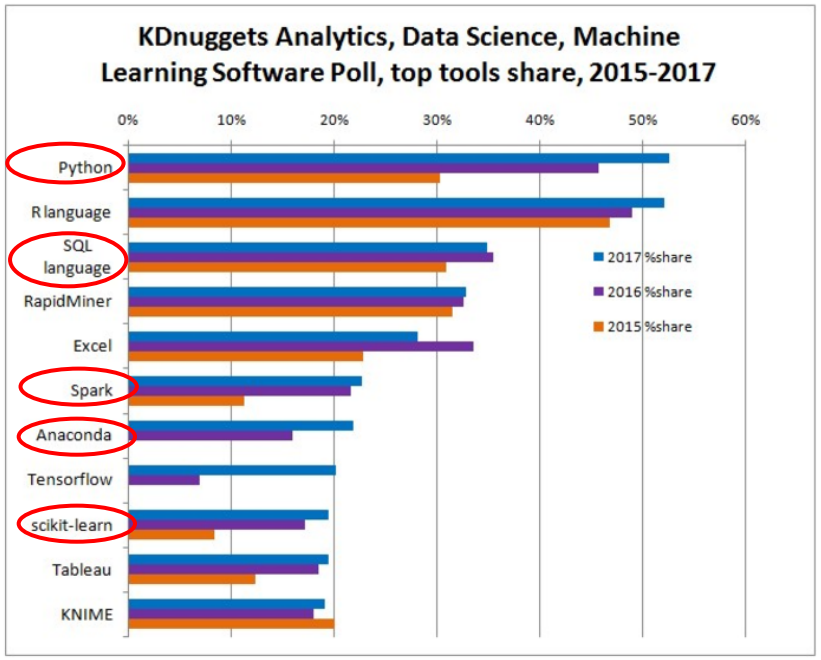


Fig 1: KDnuggets Analytics/Data Science 2017 Software Poll: top tools in 2017, and their usage in the 2015-6 polls

Table 1: Top Analytics/Data Science Tools in 2017 KDnuggets Poll

Tool	2017 % Usage	% change 2017 vs 2016	% alone
Python	52.6%	15%	0.2%
R language	52.1%	6.4%	3.3%
SQL language	34.9%	-1.8%	0%
RapidMiner	32.8%	0.7%	13.6%
Excel	28.1%	-16%	0.1%
Spark	22.7%	5.3%	0.2%
Anaconda	21.8%	37%	0.8%
Tensorflow	20.2%	195%	0%
scikit-learn	19.5%	13%	0%
Tableau	19.4%	5.0%	0.4%
KNIME	19.1%	6.3%	2.4%

Key Trends

- Apache Spark continued traction
- Python, Anaconda significant growth
- SQL style interaction is valuable

Source: http://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

Unexpected singularities in the Hessian matrix in NOMREG (Multinomial Logistic Regression)

This warning will be produced when there is a category of the dependent variable for which one of the predictors is constant. If this is the case, you can diagnose the problem by examining the regression coefficients resulting from the last iteration, which are shown in the Parameter Estimates table. Look for a set of coefficients where the magnitude of the intercept is very large, and one of the predictor coefficients is also large, in the opposite direction.

Covers @ FirstCovers.com

"Here I am, brain the size of a planet and they ask me to take you to the bridge. Call that job satisfaction? 'cause I don't."

A peek at Machine Learning

DemystifAIed

Artificial Intelligence

The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision making, and translation between languages

Machine Learning

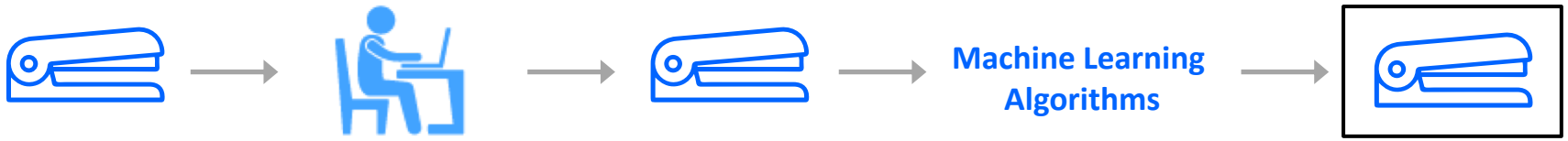
The capacity of a computer to learn from experience, i.e. to modify its processing on the basis of newly acquired information

Deep Learning

The study of artificial neural networks and related Machine Learning Algorithms that include more than one hidden layer.

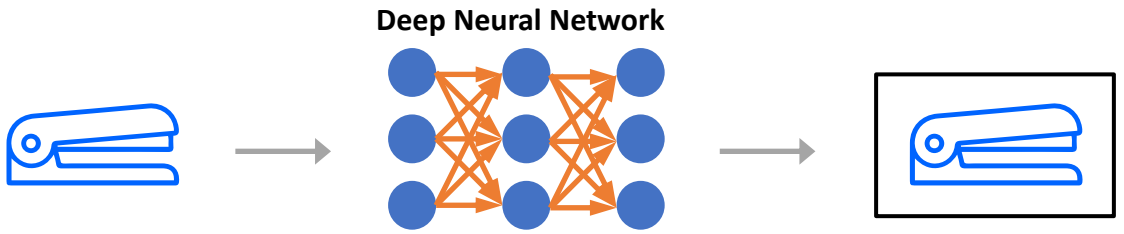
Machine Learning v Deep Learning

Machine Learning



Input → Feature Engineering → Features → Classification → Output

Deep Learning

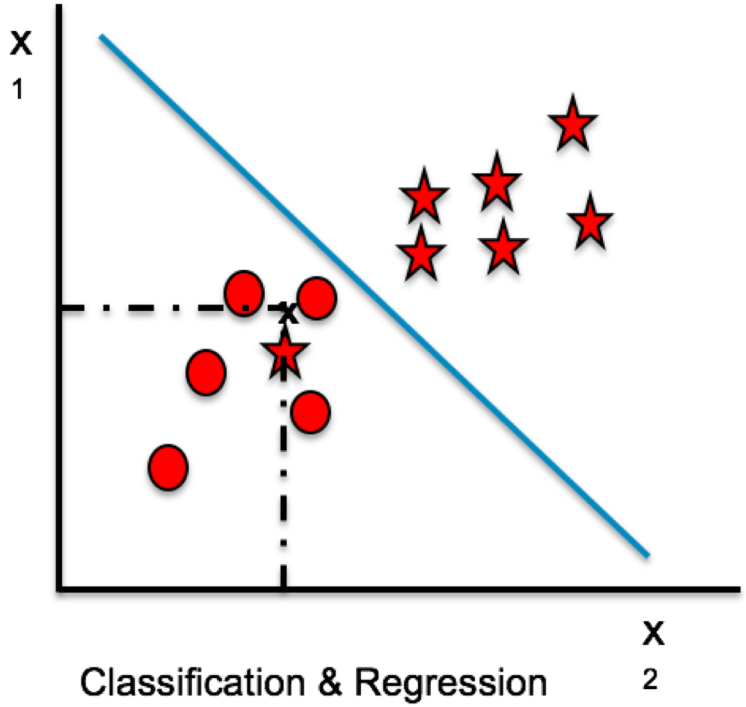


Input → Feature Extraction & Classification → Output

Types of Machine Learning

Supervised Learning

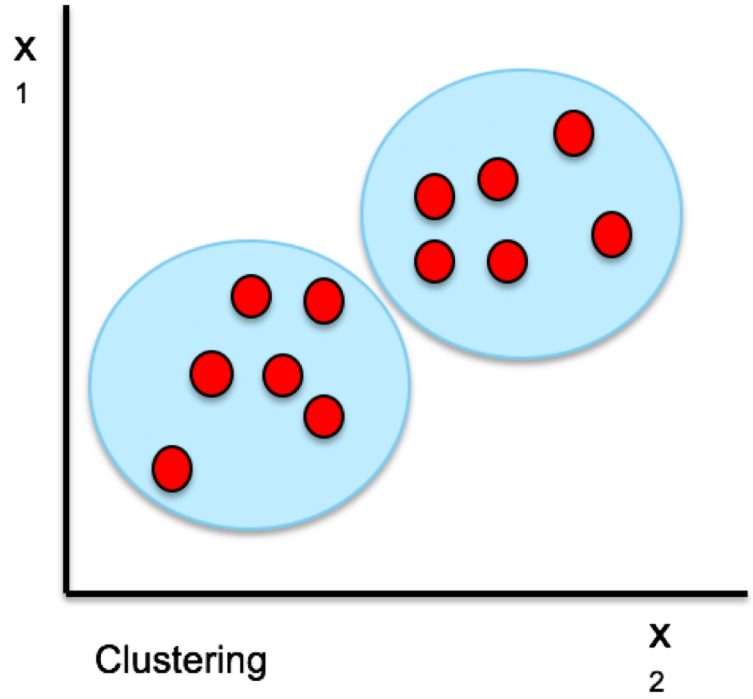
- Models trained from labeled data



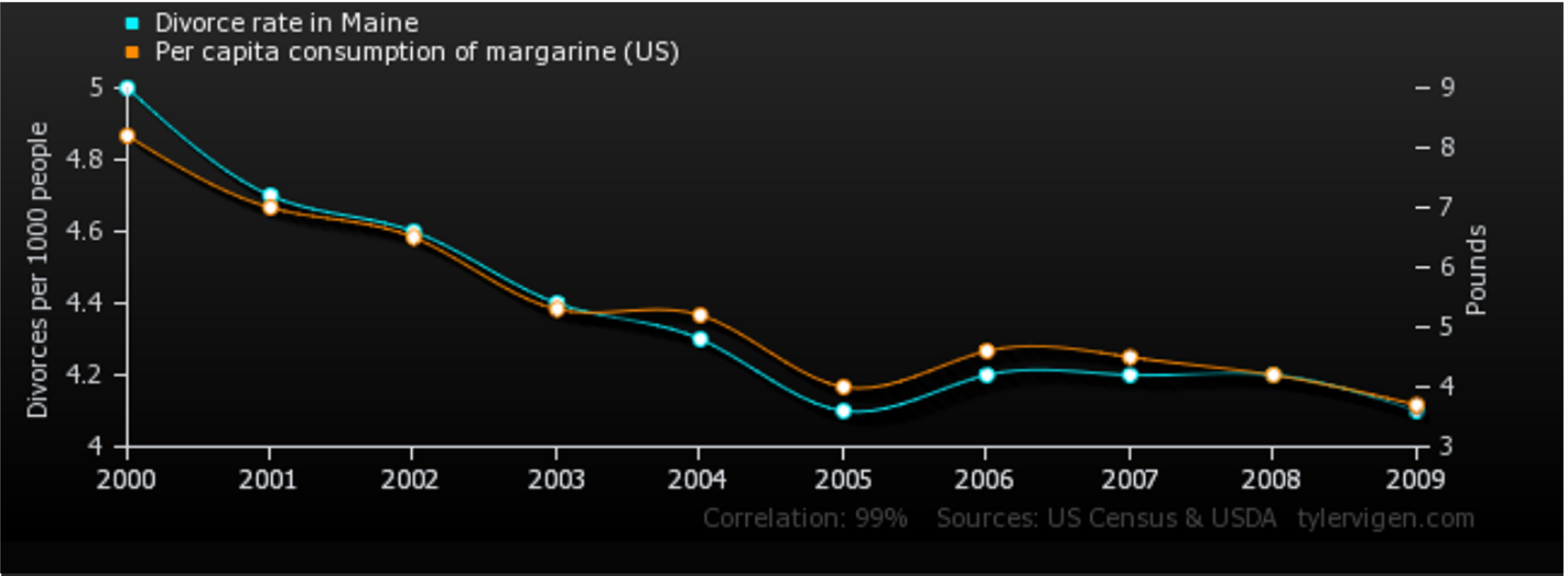
Classification & Regression

Unsupervised Learning

- Models trained from unlabeled data

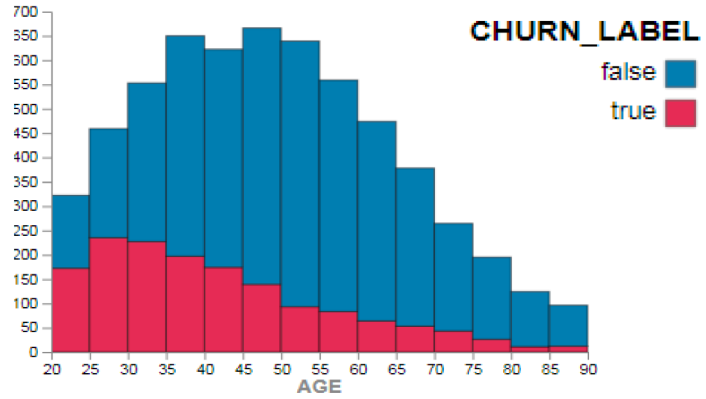


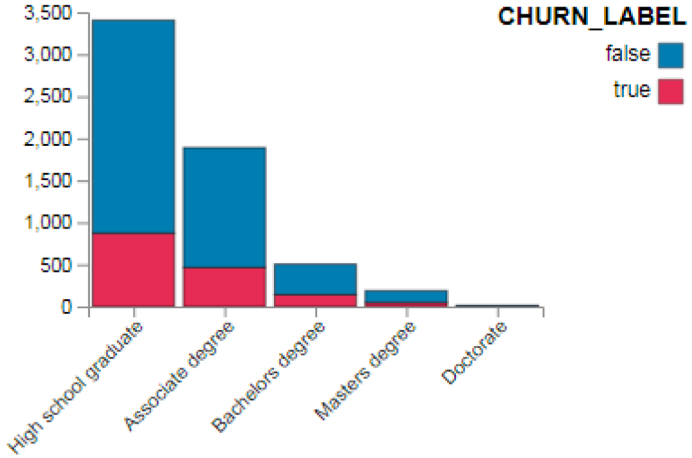
Clustering

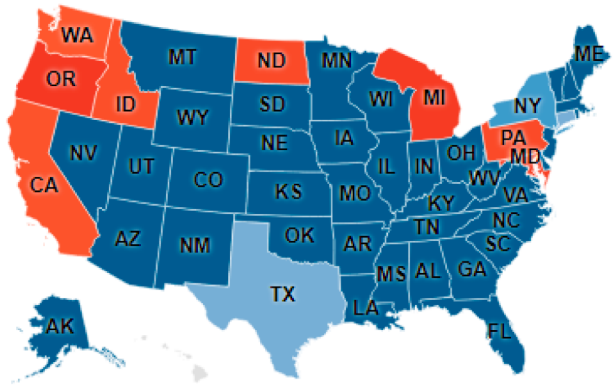


Correlation does not imply causation.

Factors driving customer churn

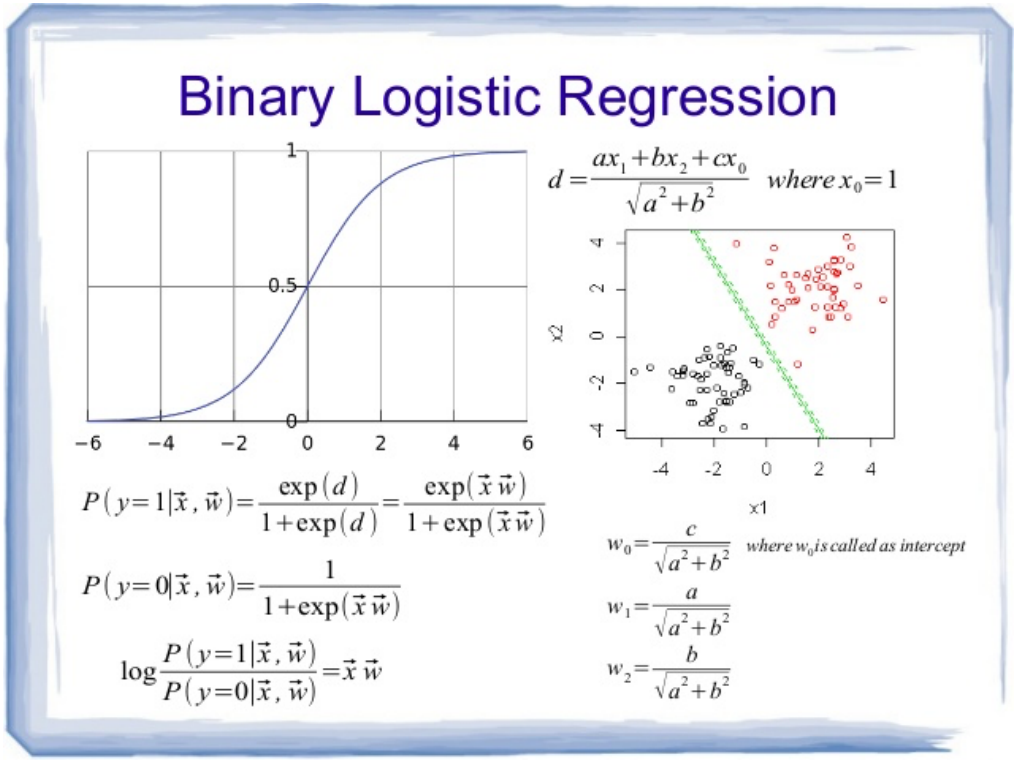






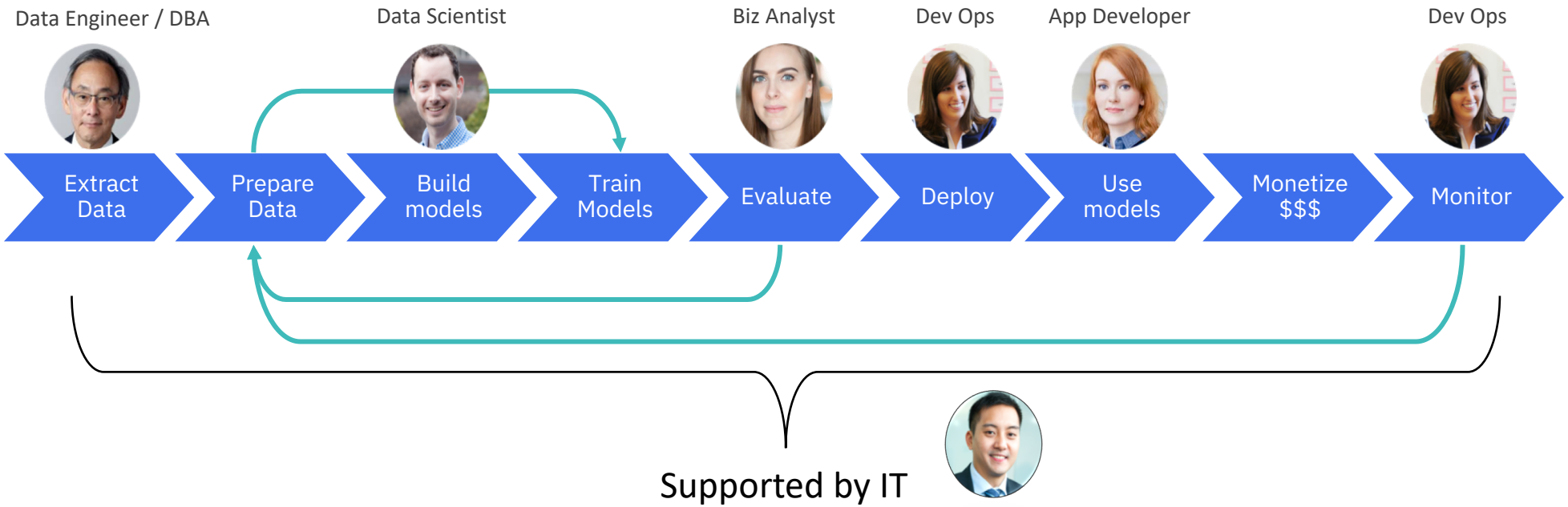
Business Rules v Machine Learning

	< 35	35-45	46-55	> 55	
AGE	100	80	60	20	
	High School Grad	Assoc Degree	Bachelors Degree	Masters Degree	Doctorate
EDUC LEVEL	100	80	50	20	10
	OR	CA	WA	TX	NY
STATE	100	80	70	20	10



Can the model outperform the knowledge captured in the business rules?

Data Science is a Team Sport
with Iterative Development



The team needs a connected infrastructure for data, development, and iteration.

IT needs to service and operationalize the process.

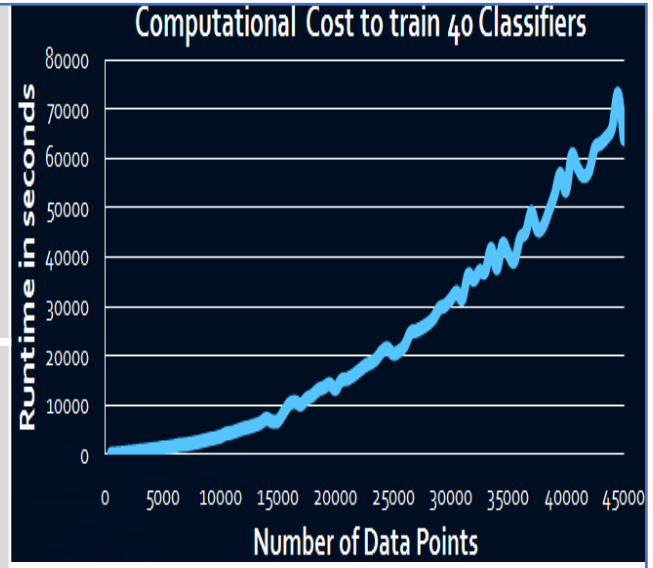
The challenges of the Data Scientist

More than 50 different algorithms

SVM, Neural Net, Decision Trees/Forests, Naïve Bayes, Regression, SMO, K-nearest Neighbor Clustering, Rules, …

Explosive # of parameter choices per algorithm

Kernel type, pruning strategy, number of trees in a forest, learning rate, …



Wide variation in performance across different algorithm implementations/user defined algorithms

SPSS vs. SAS vs. Python vs. WEKA vs SPARK …

Trying new combinations and parameters is time intensive

Computational cost for training a single SVM can exceed 24 hours

Data Handling

Points to Ponder for DBAs

- Missing values (e.g. nulls)
- Transforming nominal variables
- Model training and scoring
- Being lazy with Spark

Transforming Nominal Variables

Python Example (scikit-learn)

```
from sklearn.preprocessing import LabelEncoder

sex_encoder = LabelEncoder()
sex_encoder.fit(d2['SEX'])
d2['sex_code'] = sex_encoder.transform(d2['SEX'])
state_encoder = LabelEncoder()
state_encoder.fit(d2['STATE'])
d2['state_code'] = state_encoder.transform(d2['STATE'])

features = ['AGE', 'ACTIVITY', 'EDUCATION', 'NEGTWEETS', 'INCOME', 'sex_code',
'state_code']
```

← generates a value of 0 or 1

← generates values 0 - 49

features is now an array of **numbers**

Model Training

```
from sklearn.model_selection import train_test_split

X, y = d2.loc[:, features], d2.loc[:, 'CHURN']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=0)
print("The number of training data is ", X_train.shape[0])
print("The number of test data is ", X_test.shape[0])

# Train a logistic regression
from sklearn.linear_model import LogisticRegressionCV
from sklearn import metrics, cross_validation
logreg_cv = LogisticRegressionCV(Cs=10)
logreg_cv.fit(X_train, y_train)
```

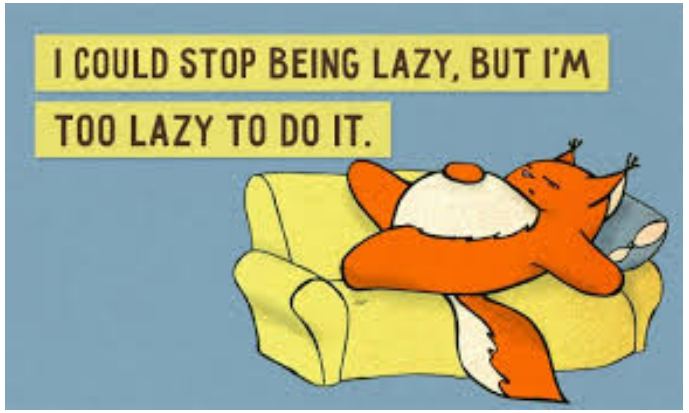
Model Scoring

```
import pandas as pd
cust = pd.DataFrame({'AGE': 41,
                     'ACTIVITY': 1,
                     'EDUCATION': 3,
                     'NEGTWEETS': 12,
                     'INCOME': 20000,
                     'sex_code': sex_encoder.transform(['M']),
                     'state_code': state_encoder.transform(['TX'])},
                    columns = ['AGE', 'ACTIVITY', 'EDUCATION', 'NEGTWEETS', 'INCOME', 'sex_code', 'state_code'])

print(cust)
print()
print("Predicted probability: ", logreg_cv.predict_proba(cust))
print("Predicted churn label: ", logreg_cv.predict(cust))
```

```
   AGE  ACTIVITY  EDUCATION  NEGTWEETS  INCOME  sex_code  state_code
0   41         1          3         12   20000         1          42

Predicted probability:  [[ 0.00151332  0.99848668]]
Predicted churn label:  [1]
```

Being Lazy with Spark

- Spark is designed to process data in memory
- How to avoid memory being swamped with large tables?
- Spark uses ”lazy” evaluation
  - not all statements result in an action
  - encourage use of filters and temporary tables
  - predicate pushdown is enabled



Lazy Example – Scala & Spark

```
val sc=new SparkContext("local[*]",”custsum")
val sqlContext = new SQLContext(sc)
val optionscustSum = scala.collection.mutable.Map[String, String]();
optionscustSum.put("driver", "com.ibm.db2.jcc.DB2Driver");
val myurl="jdbc:db2://10.7.1.139:4750/MOPDBC0:user=sudb101;password=xxxxxxxx;"
optionscustSum.put("url", myurl);
optionscustSum.put("dbtable", "SUDB101.CUST_SUM");

val custSumDF = sqlContext.read.format("jdbc").options(optionscustSum).load();
val highincome = custSumDF.filter("INCOME > 200000");
val count = highincome.count();

println ("Count of high earners is " +  count);  // first action here

highincome.registerTempTable("highearners"); // enables SQL against data frame

val texashigh = sqlContext.sql("SELECT * FROM highearners WHERE STATE = 'TX’”)

println ("Count of high earners in Texas is " +  texashigh.count());
```

```
Count of high earners is 69
Count of high earners in Texas is 16
```

In-Transaction Scoring

Storing a model in Db2 for z/OS

```
CREATE TABLE MODEL_REPOSITORY
   (ARTIFACT_ID        INTEGER NOT NULL,
    VERSION            INTEGER NOT NULL,
    NAME               VARCHAR(500) FOR MIXED DATA NOT NULL,
    DESCRIPTION        VARCHAR(1000) FOR MIXED DATA
     WITH DEFAULT NULL,
    MODEL_OWNER     VARCHAR(100) FOR MIXED DATA
     WITH DEFAULT NULL,
    INPUT_SCHEMA       CLOB(32000) FOR MIXED DATA
     WITH DEFAULT NULL
      INLINE LENGTH 0,
    PROJECT            VARCHAR(200) FOR MIXED DATA
     WITH DEFAULT NULL,
    STATUS             VARCHAR(100) FOR MIXED DATA
     WITH DEFAULT NULL,
    MODEL_DATA         CLOB(1 G) FOR MIXED DATA WITH
      DEFAULT NULL
      INLINE LENGTH 0,
    CREATED_BY         VARCHAR(100) FOR MIXED DATA
     WITH DEFAULT NULL,
    CREATED_DATE       TIMESTAMP (6) WITHOUT TIME ZONE
     WITH DEFAULT)

-- ETC ETC
```

CICS / COBOL Example

Code the input and output parameters used by the model

```
*****************************************************************
* DATA STRUCTURE FOR MODEL INPUT                                *
*****************************************************************
 01  CHURNIN.
        06 EDUCATION                      COMP-2 SYNC.
        06 AGE                            COMP-2 SYNC.
        06 SEX-length                     PIC S9999 COMP-5 SYNC.
        06 SEX                            PIC X(255).
        06 NEGTWEETS                      COMP-2 SYNC.
        06 INCOME                         COMP-2 SYNC.
        06 ACTIVITY                       COMP-2 SYNC.
        06 STATE-length                   PIC S9999 COMP-5 SYNC.
        06 STATE                          PIC X(255).

*****************************************************************
* DATA STRUCTURE FOR MODEL OUTPUT                               *
*****************************************************************
 01  CHURNOUT.
         06 prediction                     COMP-2 SYNC.
         06 probability OCCURS 2           COMP-2 SYNC.

 01 I PIC 9(2) VALUE 1.
```

CICS / COBOL Example

Pass parameters using a CICS container and call the scoring program.

Retrieve output parameters and process the result

```
*****************************************************************
* PASS THE INPUT DATA RECORD TO SCORING VIA CICS CHANNEL AND    *
* CONTAINER ALN_INPUT_DATA.                                     *
*****************************************************************
     EXEC CICS PUT CONTAINER('ALN_INPUT_DATA') CHANNEL('CHAN')
                   FROM(CHURNIN) BIT END-EXEC.

*****************************************************************
* USE CICS LINK TO CALL THE SCORING PROGRAM ALNSCORE TO         *
* PERFORM PREDICTION AGAINST THE SPECIFIED INPUT RECORD.        *
*****************************************************************
     EXEC CICS LINK PROGRAM('ALNSCORE') CHANNEL('CHAN')
                   END-EXEC.

*****************************************************************
* GET THE SCORING RESULT BACK VIA CICS CHANNEL AND CONTAINER    *
* ALN_OUTPUT_DATA.                                              *
*****************************************************************
     EXEC CICS GET CONTAINER('ALN_OUTPUT_DATA') CHANNEL('CHAN')
                   INTO(CHURNOUT) END-EXEC.

     MOVE CHURNOUT to Commarea-data.

*****************************************************************
* PRINT OUT THE SCORING RESULT.                                 *
*****************************************************************
     DISPLAY 'PREDICTION     :' PREDICTION.
     PERFORM UNTIL I = 3
     DISPLAY 'PROBABILITY-' I
     DISPLAY PROBABILITY(I)
     ADD 1 TO I
     END-PERFORM.
```

Performance Test Result for CICS Online Scoring Service

Testing environment

- z13 / zOS 2.2 / 1 GCP / 4 zIIP
- CICS Integrated Scoring running in the same region as the transaction

Testing object

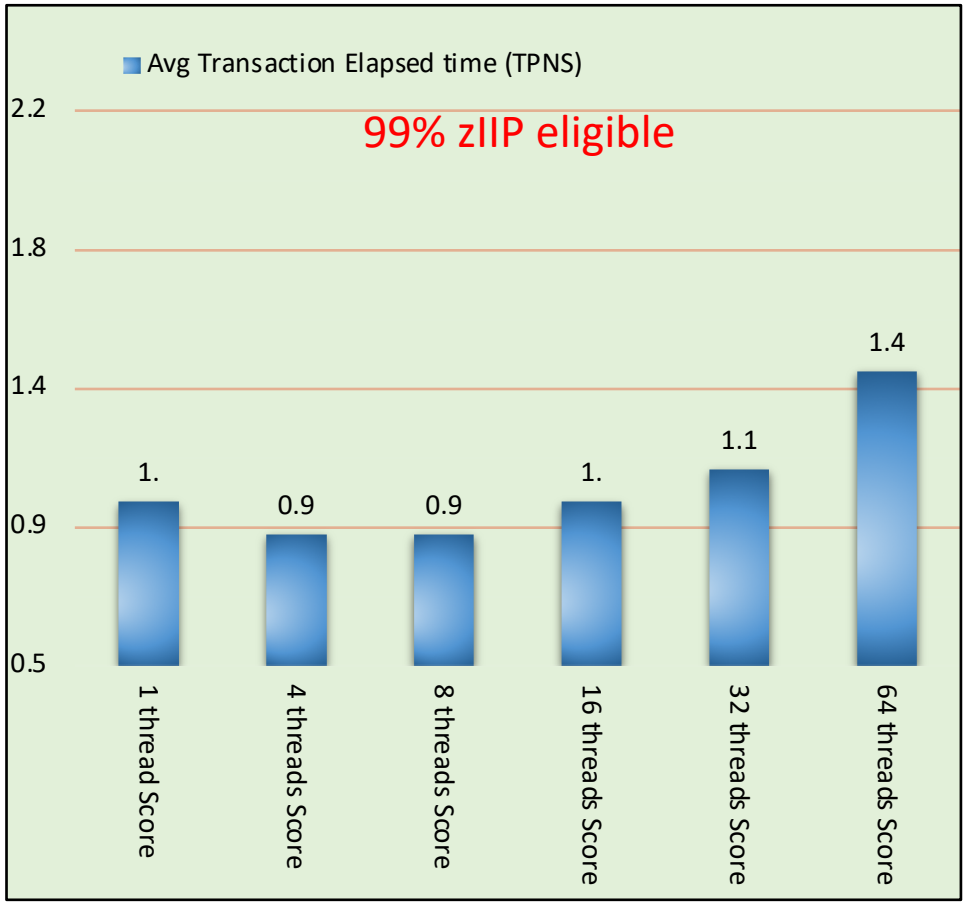
- Linear Regression model (45 stage, 105 input fields)

Testing approach

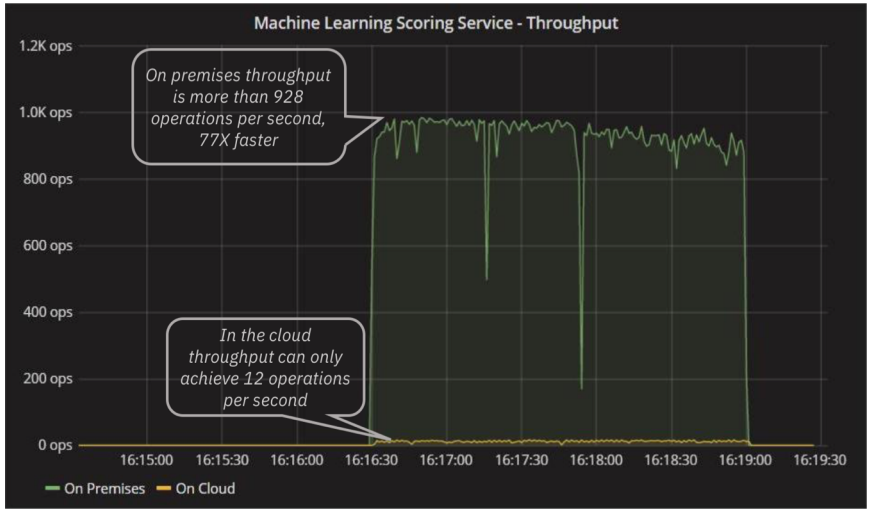
- TPNS drives the test from another LPAR to avoid extra CPU/memory overhead
- Score transaction is the same Baseline transaction adding a single scoring call after the query

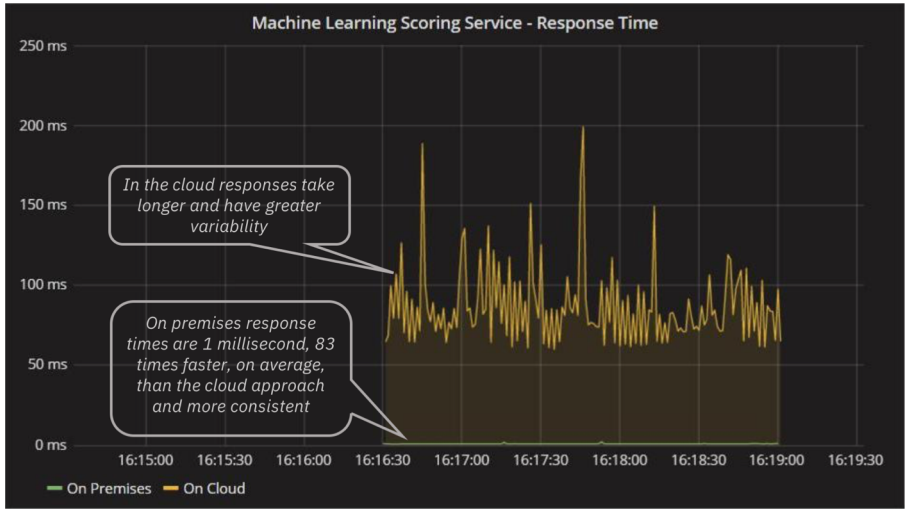
Test result summary

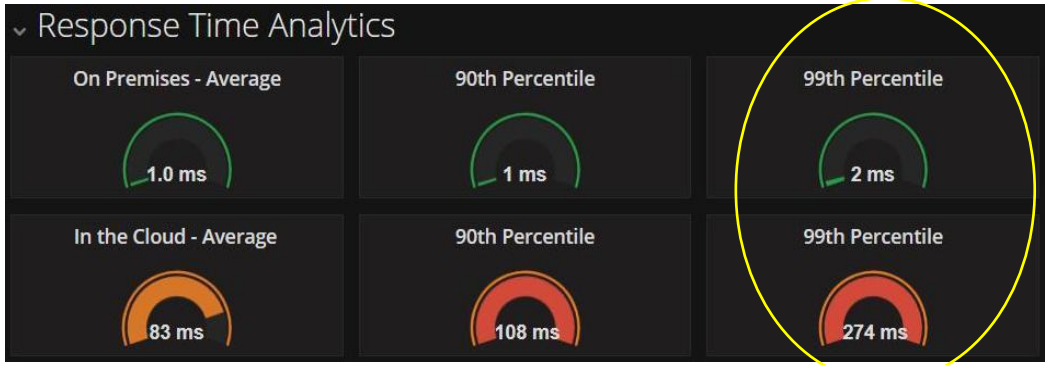
- Best Average Elapsed time is 0.9 ms (8 threads)
- Best TPS is 1000+ (64 threads)



On-Prem v Cloud Comparison

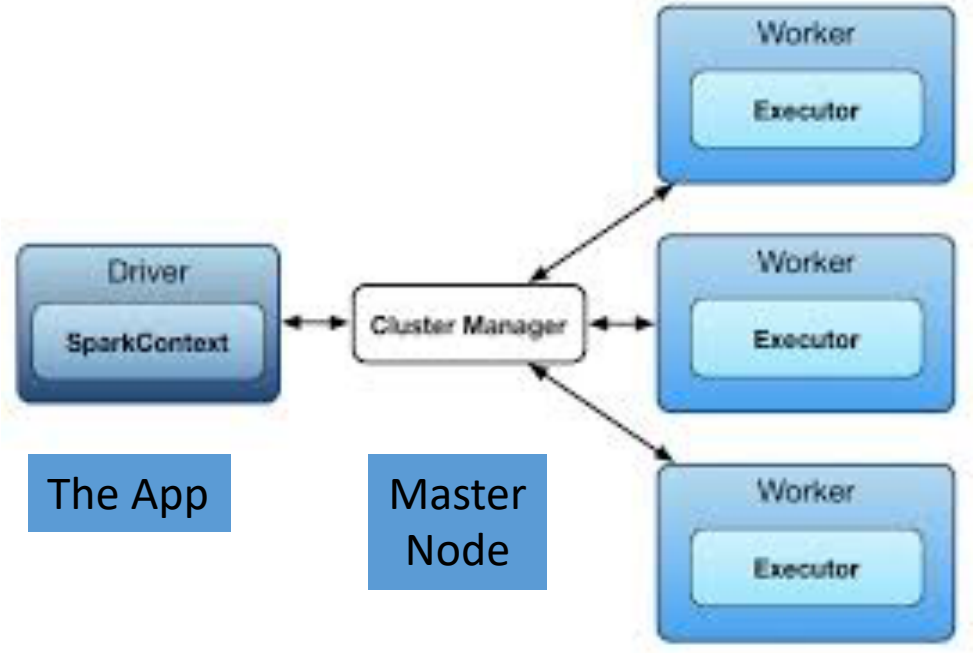






A Word or Two About Spark

Spark Likes Resources



Spark architecture on a distributed cluster

Spark Component	Memory Defaults
Master	1 GB
Worker (Slave)	1 GB
Each Executor	1 GB
Driver	1 GB

Spark on z/OS will spawn multiple executors under a single worker.

It will use as many zIIPs as are available, but can spill to CPs if needed.

IBM benchmarks showed 15% - 20% throughput benefit by enabling SMT on z13 zIIP.

Spark Web UI – Main Screen

Alive Workers: 1
Cores in use: 17 Total, 10 Used
Memory in use: 304.1 GB Total, 65.0 GB Used
Applications: 2 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-20161028162401-xx.xx.xx.xx-1055	xx.xx.xx.xxx:1055	ALIVE	17 (10 Used)	304.1 GB (65.0 GB Used)

Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
app-20161028162913-0002 (kill)	TestRunner: Sort	2	15.0 GB	2016/10/28 16:29:13	ROBIN	RUNNING	12 s
app-20161028162800-0001 (kill)	TestRunner: Aggregate	8	25.0 GB	2016/10/28 16:28:00	ROBIN	RUNNING	1.4 min

Spark Web UI – Drill Down on Worker Node

Running Executors (3)

ExecutorID	Cores	State	Memory	Job Details	Logs
0	4	LOADING	25.0 GB	**ID:** app-20161028162800-0001 **Name:** TestRunner: Aggregate **User:** SPARKID	stdout stderr
0	2	LOADING	15.0 GB	**ID:** app-20161028162913-0002 **Name:** TestRunner: Sort **User:** SPARKID	stdout stderr
1	4	LOADING	25.0 GB	**ID:** app-20161028162800-0001 **Name:** TestRunner: Aggregate **User:** SPARKID	stdout stderr

Spark Tuning is Imperative

zOS for Apache Spark Resource Tuning.pdf

http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102684

Indispensable advice for changing defaults and taming Spark!

Q & A

And thanks for listening ☺

We want your feedback!

- Please submit your feedback online at ….
  - http://conferences.gse.org.uk/2018/feedback/IF

- Paper feedback forms are also available from the Chair person

- This session is IF