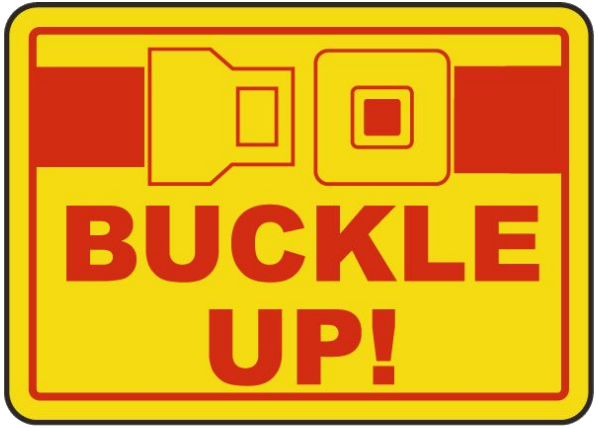
Introduction To Data Science Using Apache Spark



Chris Nicholls
Maritime Devcon 2017

A bit about me

Ontariario

LIES

STANFORD

STANFORD

STANFORD

salesforce

Some Words To Describe Me

Some Words To Describe Me

- Temperamentally Intense

Some Words To Describe Me

- Temperamentally Intense
- Honest

Some Words To Describe Me

- Temperamentally Intense
- Honest
- Expressive

Some Words To Describe Me

- Temperamentally Intense
- Honest
- Expressive
- Passionately Curious

Some Words To Describe Me

- Temperamentally Intense
- Honest
- Expressive
- Passionately Curious
- Zestful

Some Words To Describe Me

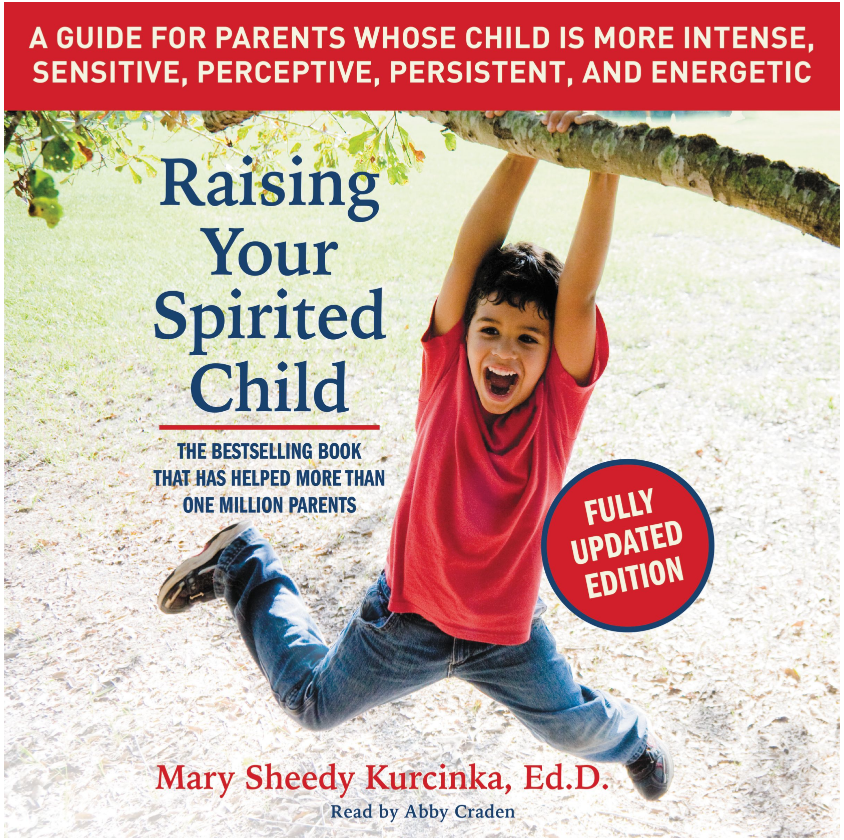
- Temperamentally Intense
- Honest
- Expressive
- Passionately Curious
- Zestful
- Selective

Some Words To Describe Me

- Temperamentally Intense
- Honest
- Expressive
- Passionately Curious
- Zestful
- Selective
- Committed To Goals

Some Words To Describe Me

- Temperamentally Intense
- Honest
- Expressive
- Passionately Curious
- Zestful
- Selective
- Committed To Goals

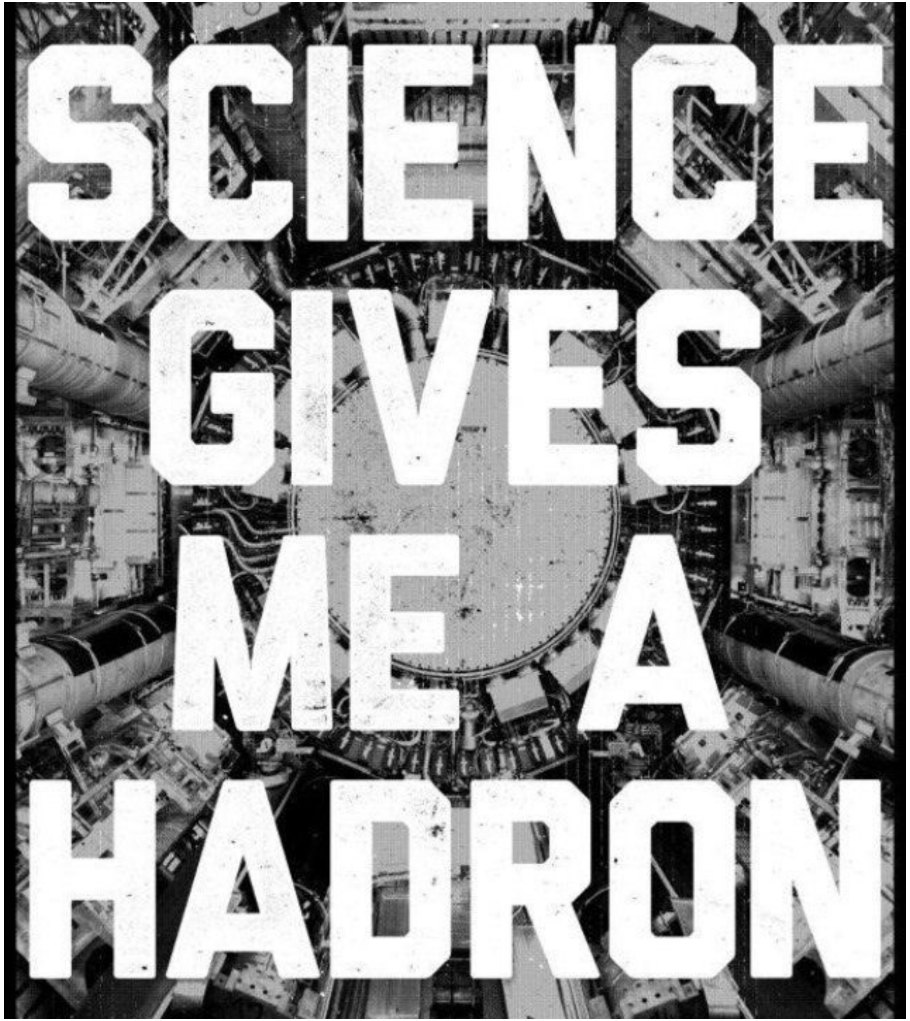

A GUIDE FOR PARENTS WHOSE CHILD IS MORE INTENSE, SENSITIVE, PERCEPTIVE, PERSISTENT, AND ENERGETIC
Raising Your Spirited Child
THE BESTSELLING BOOK THAT HAS HELPED MORE THAN ONE MILLION PARENTS
FULLY UPDATED EDITION
Mary Sheedy Kurcinka, Ed.D.
Read by Abby Craden

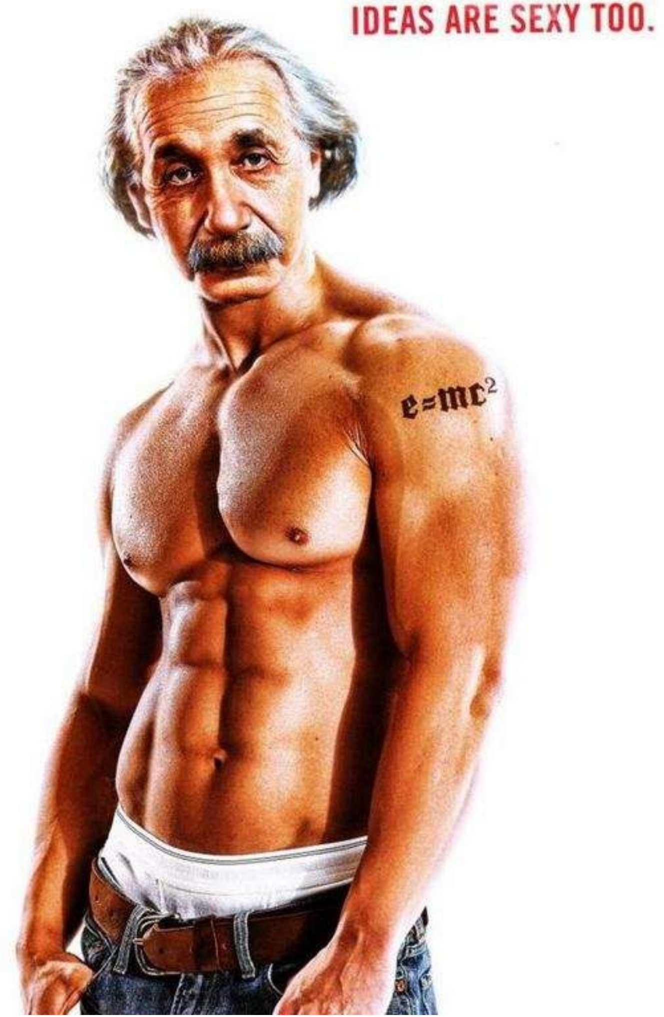
What Is Data Science?

?

"The Sexiest Job of the 21st Century"

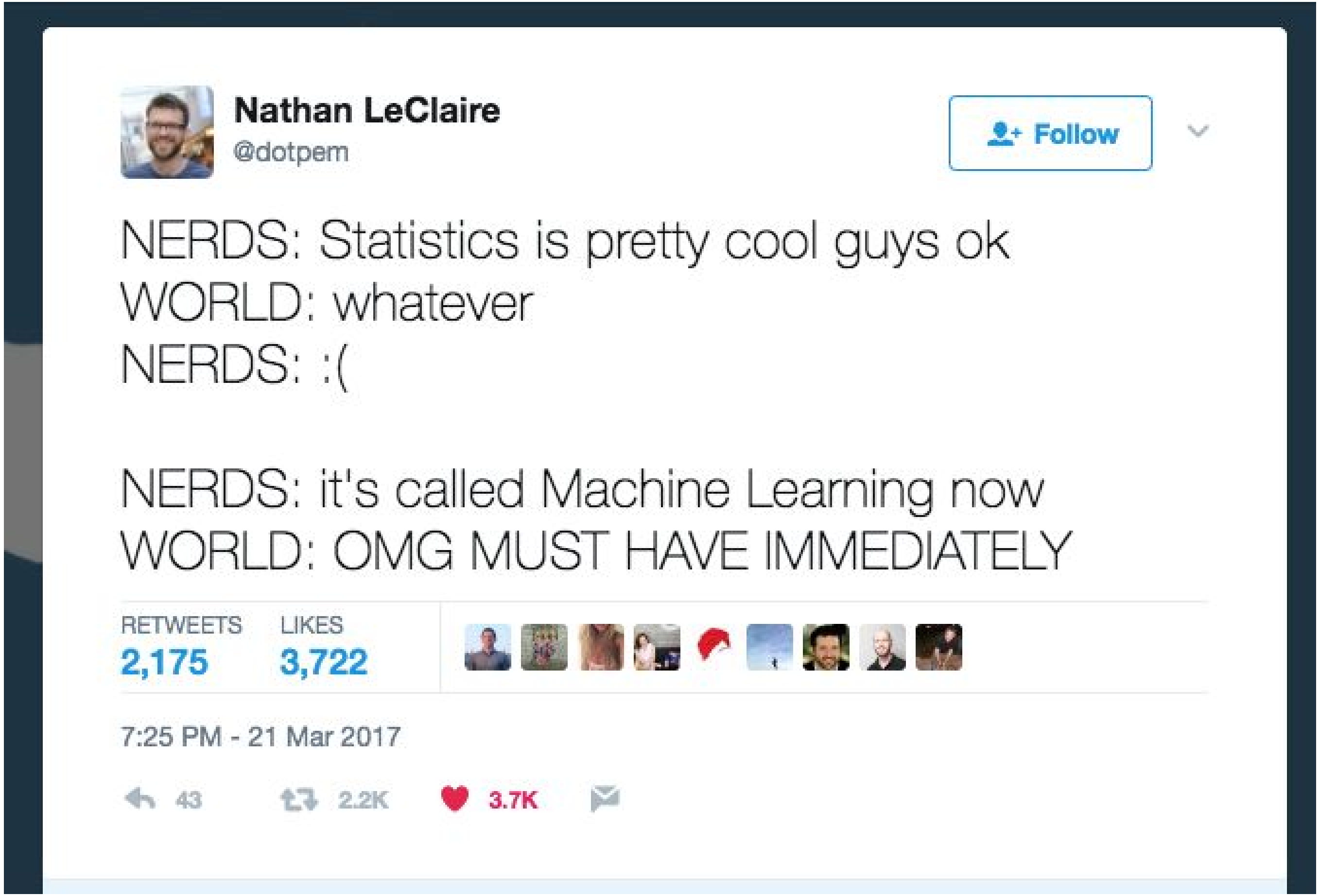
-Harvard Business Review


SCIENCE
GIVES
ME A
HADRON


IDEAS ARE SEXY TOO.
e=mc²

"The Sexiest Job of the 21st Century"

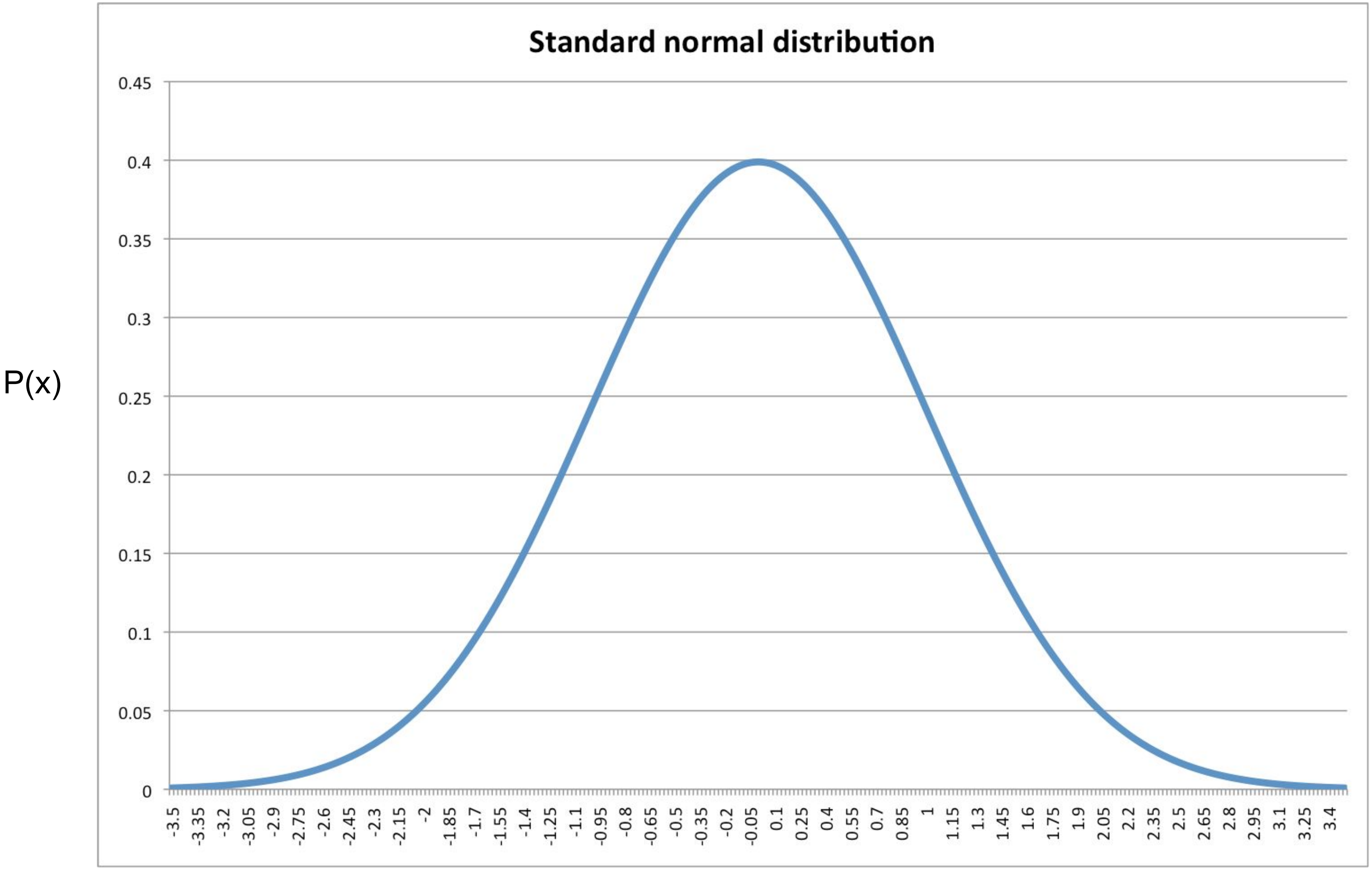
-Harvard Business Review



NERDS: Statistics is pretty cool guys ok
WORLD: whatever
NERDS: :(

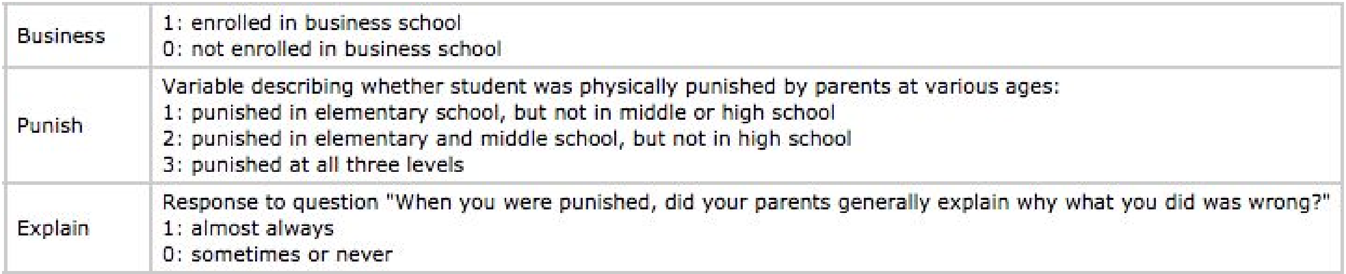
NERDS: it's called Machine Learning now
WORLD: OMG MUST HAVE IMMEDIATELY

Let’s dig into this for a minute…


Standard normal distribution
P(x)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-3.5
-3.35
-3.2
-3.05
-2.9
-2.75
-2.6
-2.45
-2.3
-2.15
-2
-1.85
-1.7
-1.55
-1.4
-1.25
-1.1
-0.95
-0.8
-0.65
-0.5
-0.35
-0.2
-0.05
0.1
0.25
0.4
0.55
0.7
0.85
1
1.15
1.3
1.45
1.6
1.75
1.9
2.05
2.2
2.35
2.5
2.65
2.8
2.95
3.1
3.25
3.4

“If you found a wallet on the street, would you: return it, keep the money, or keep both?”

Outcome	Pronoun	Business	Punish	Explain
Keep $	He	0	2	0
Keep $	He	1	2	1
Return	He	0	1	1
Return	She	0	2	0
Keep Both	She	0	1	1
...				



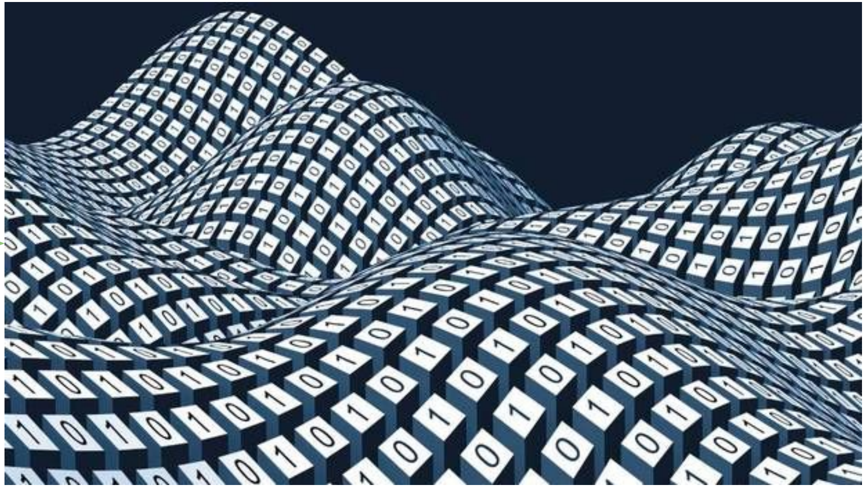
Business	1: enrolled in business school 0: not enrolled in business school
Punish	Variable describing whether student was physically punished by parents at various ages: 1: punished in elementary school, but not in middle or high school 2: punished in elementary and middle school, but not in high school 3: punished at all three levels
Explain	Response to question "When you were punished, did your parents generally explain why what you did was wrong?" 1: almost always 0: sometimes or never

“The Wallet Problem” - Allison, Paul David. 1999. *Logistic Regression Using the SAS System: Theory and Application*. SAS Institute

What’s changed since 1999?

¡BIG
DATA
Revolution!

BIG DATA
Revolution!



Outcome	Pronoun	Business	Punish	Explain
Keep $	He	0	2	0
Keep $	He	1	2	1
Return	He	0	1	1
Return	She	0	2	0
Keep Both	She	0	1	1
...				

Outcome	Pronoun	Business	Punish	Explain	Age	Birthplace	Mother’s age	Father’s age	...	...
Keep $	He	0	2	0	...	...	...	...	...	...
Keep $	He	1	2	1	...	...	...	...	...	...
Return	He	0	1	1	...	...	...	...	...	...
Return	She	0	2	0	...	...	...	...	...	...
Keep Both	She	0	1	1	...	...	...	...	...	...
...										

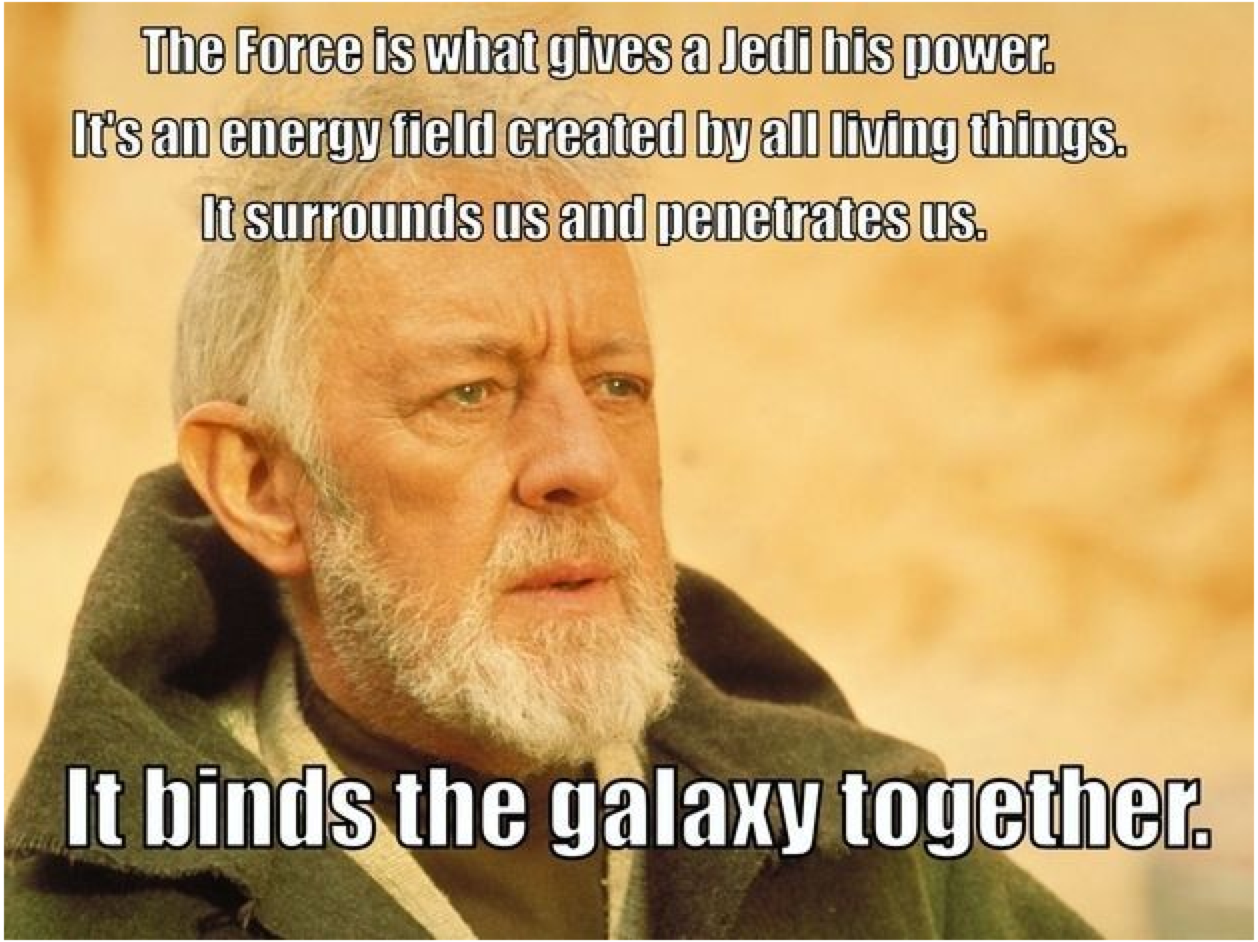
What Is Data Science?

It’s just the evolution of analysis

- Image Classification and Identification
- Relevancy
- Recommenders
- Deep Learning
- AI

- Identifying Marketing Opportunities
- Fraud Detection
- Targetting

- Making data-driven decisions


The Force is what gives a Jedi his power.
It's an energy field created by all living things.
It surrounds us and penetrates us.
It binds the galaxy together.

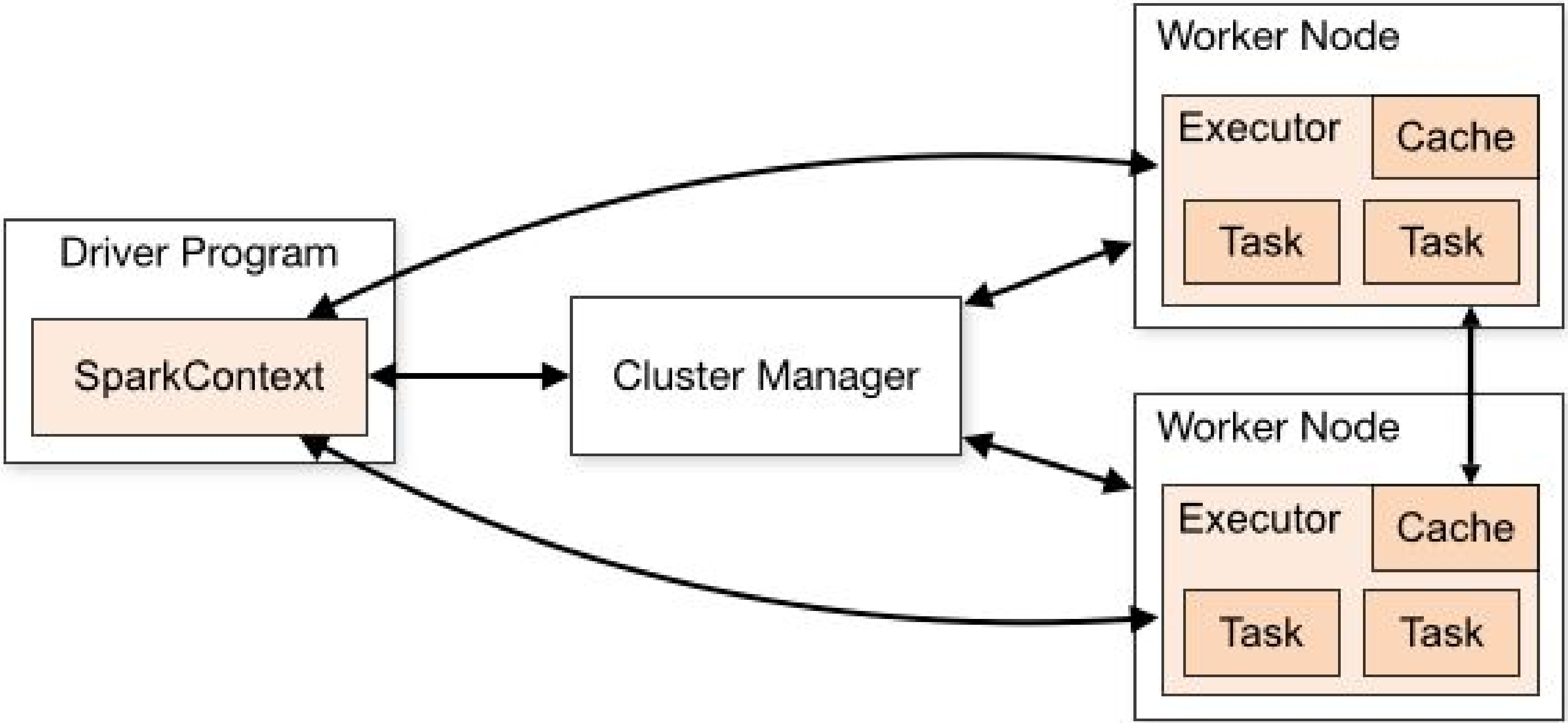
APACHE
Spark
TM

Apache Spark™ is a fast and general engine for large-scale data processing.

- Apache Spark is an open-source cluster-computing framework for large-scale data processing
- Built to be compatible with hadoop (specifically HDFS)
- Scala, Java, Python, R
- Access many data sources including HDFS, Cassandra, HBase, and S3
- Large community
  - Over 200 contributors
- Used in production by over 1000 organizations

Timeline:

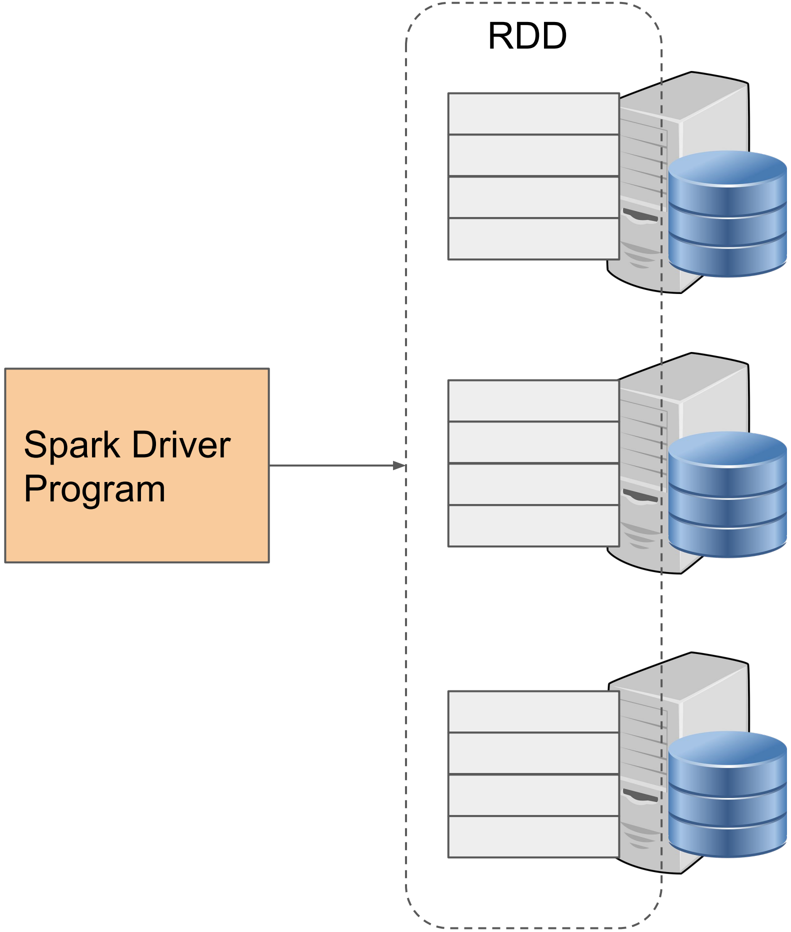
- Created by Matei Zaharia in 2009 at UC Berkeley AMPLab
- Open-sourced in 2010
- Donated to Apache Software Foundation in 2013
- Initial release:
  - May 30, 2014
- Current Stable Version:
  - v2.1.1 / May 2, 2017

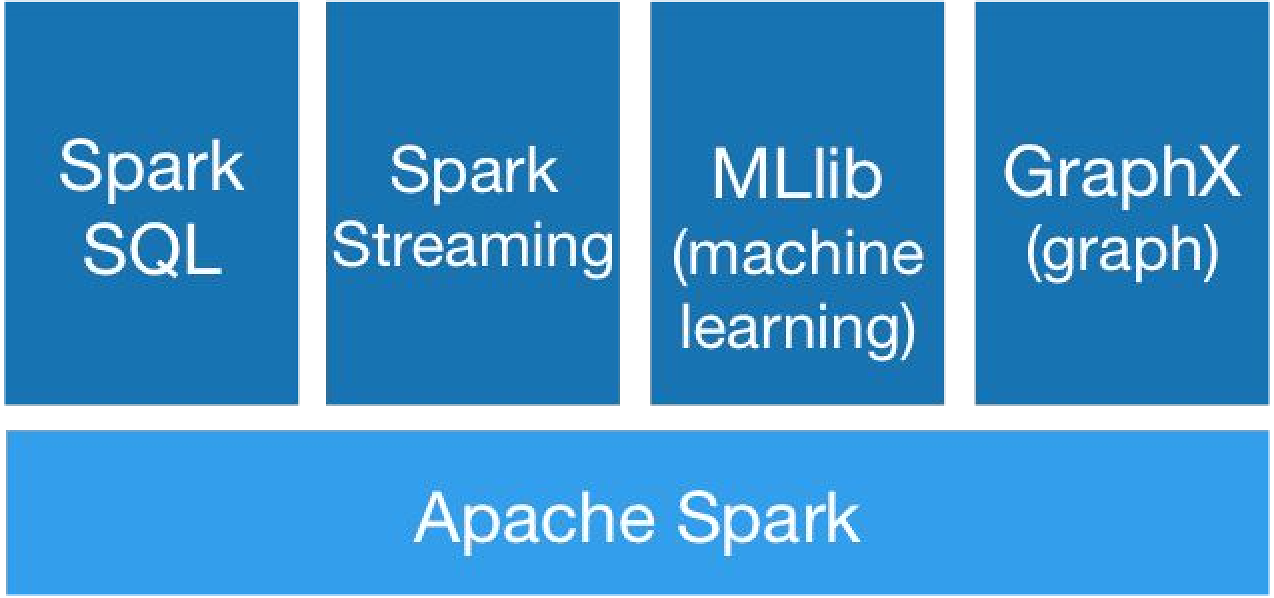

Driver Program
SparkContext
Cluster Manager
Worker Node
Executor
Cache
Task
Task
Worker Node
Executor
Cache
Task
Task

Demo Time - submit a job

```
$SPARK_HOME/bin/spark-submit --class
org.apache.spark.examples.SparkPi \
    --master yarn \
    --num-executors 10 \
    --executor-cores 2 \
    $SPARK_HOME/examples/jars/spark-examples*.jar \
    100
```

Resilient Distributed Dataset





Spark SQL

Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

Some other things

- Cluster Managers
  - Mesos, YARN, Stand-alone
- Hive
  - “Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.”
- Databricks
  - The SAAS offering from the original creators of Spark

Resources

- https://github.com/chrisnicholls/spark-yarn-hadoop-cluster-vagrant

In summary….

In summary….

In summary….

+

In summary....

+

=