

FORRESTER®

**Blazing Fast Machine Learning With
Apache Spark**

Perform Analytics Where Data Gravity Is Strongest

November 30, 2016 New York, New York

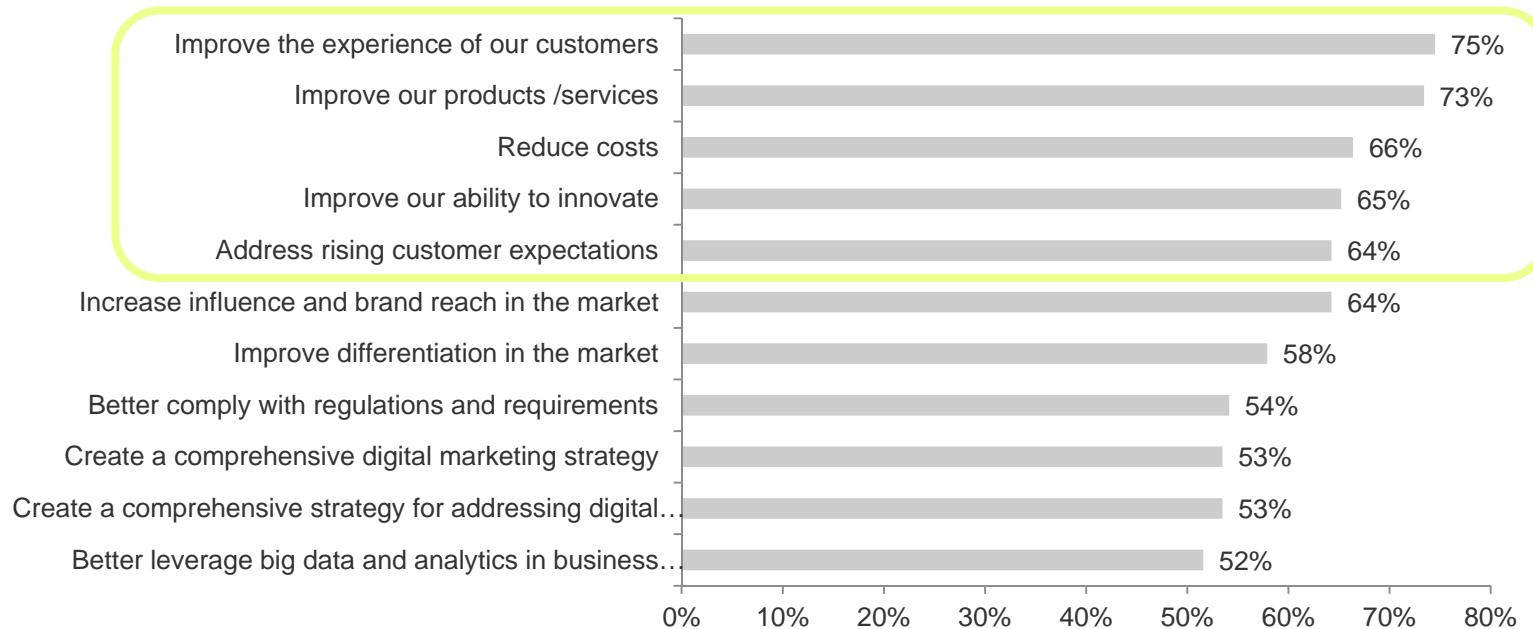
Mike Gualtieri, VP & Principal Analyst

Twitter: @mgualtieri

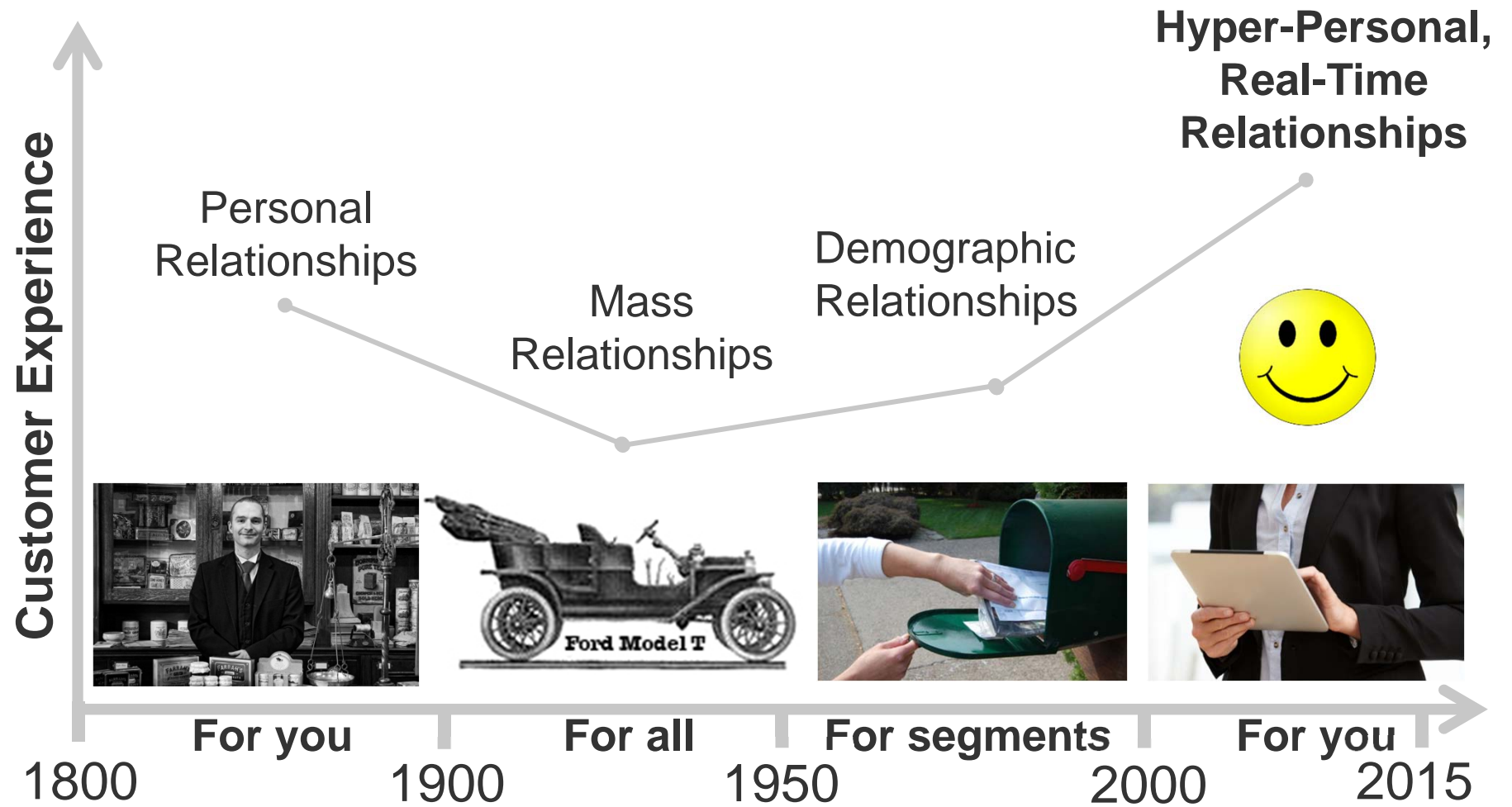


#Priority

Customer experience is a top priority for business leaders



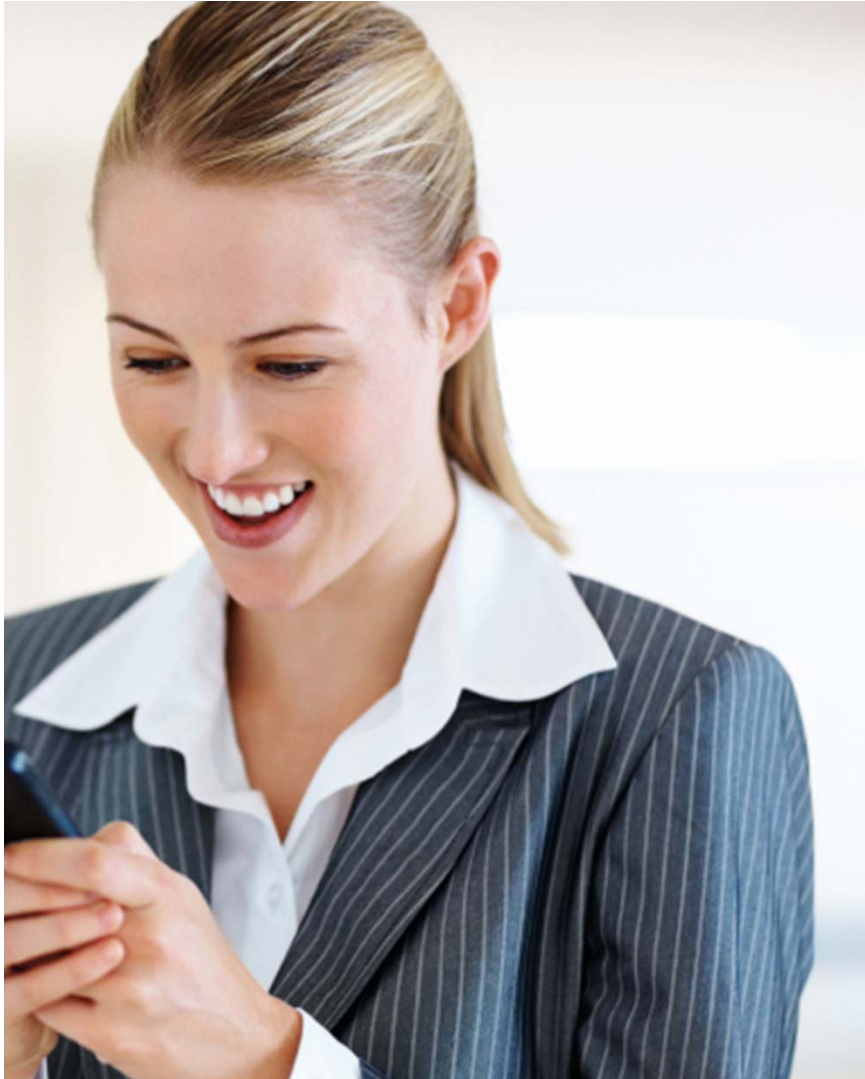
- > Base: 3,005 global data and analytics decision-makers
- > Source: Global Business Technographics Data And Analytics Online Survey, 2015



#Celebrity



Customers want and increasingly expect to be treated like celebrities.

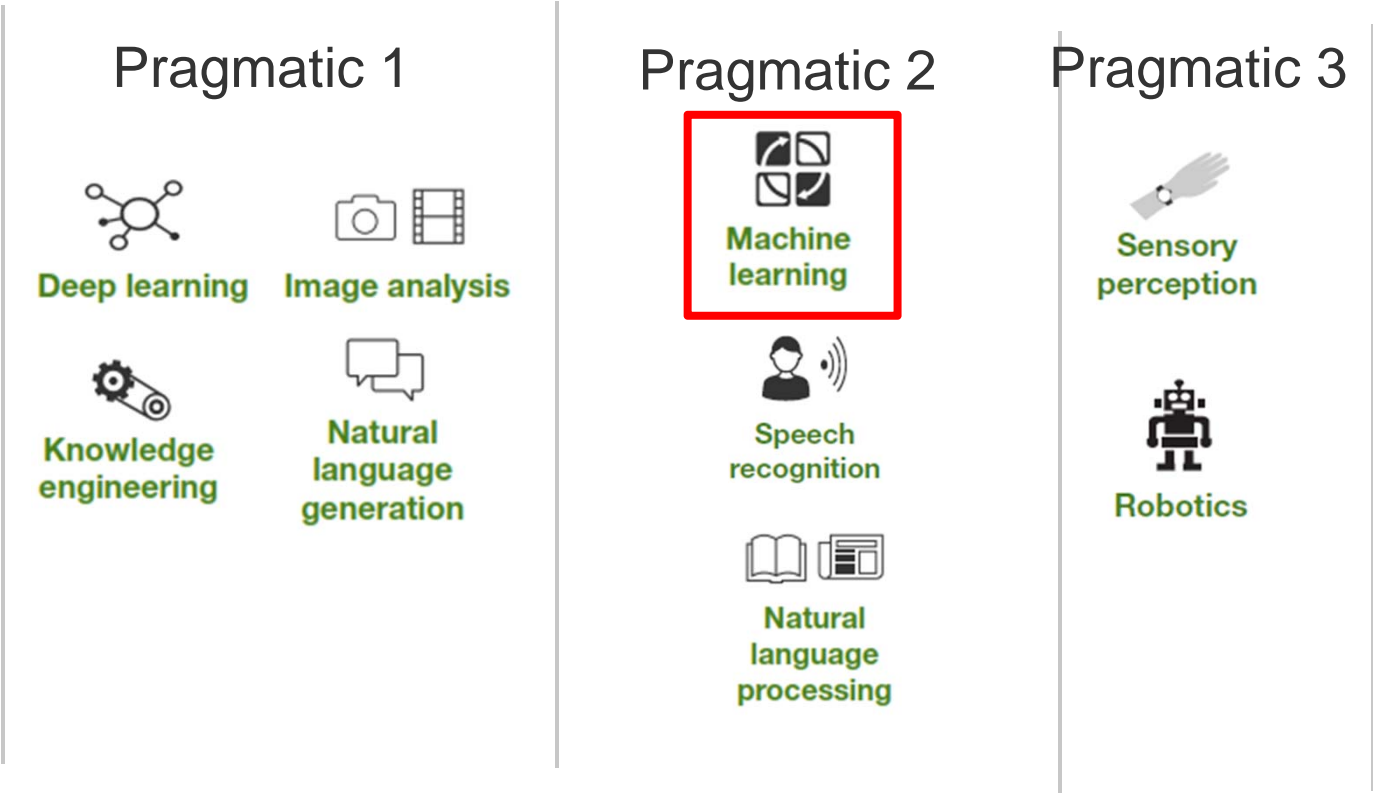


Celebrity experiences must:

- Learn individual customer characteristics and behaviors
- Detect customer needs and desires in **real-time**
- Make accurate decisions in **real-time**
- Adapt applications to serve an individual customer in **real-time**

#Cognitive

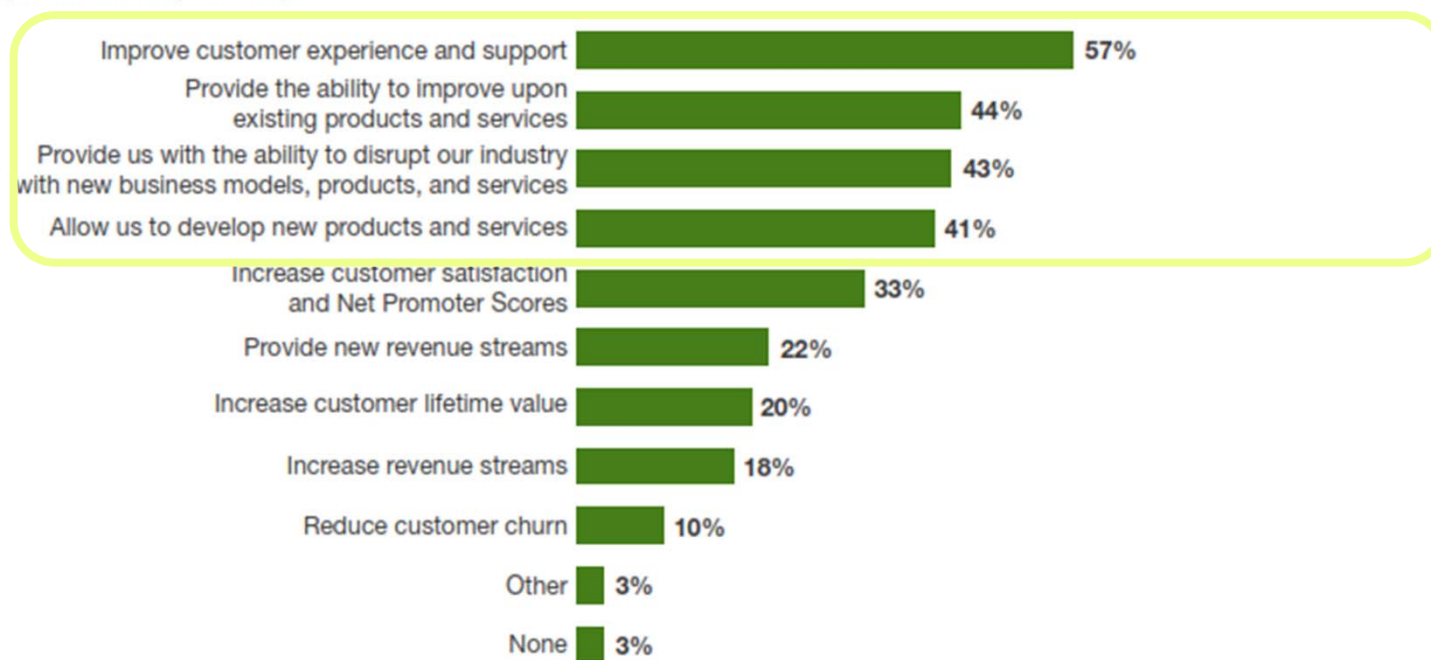
Cognitive building block technologies are very data, compute, and network intensive!



Enterprises believe the benefits of cognitive will be to improve customer experience, products, and business models

“What are the biggest strategic/growth benefits AI will contribute to your organization?”

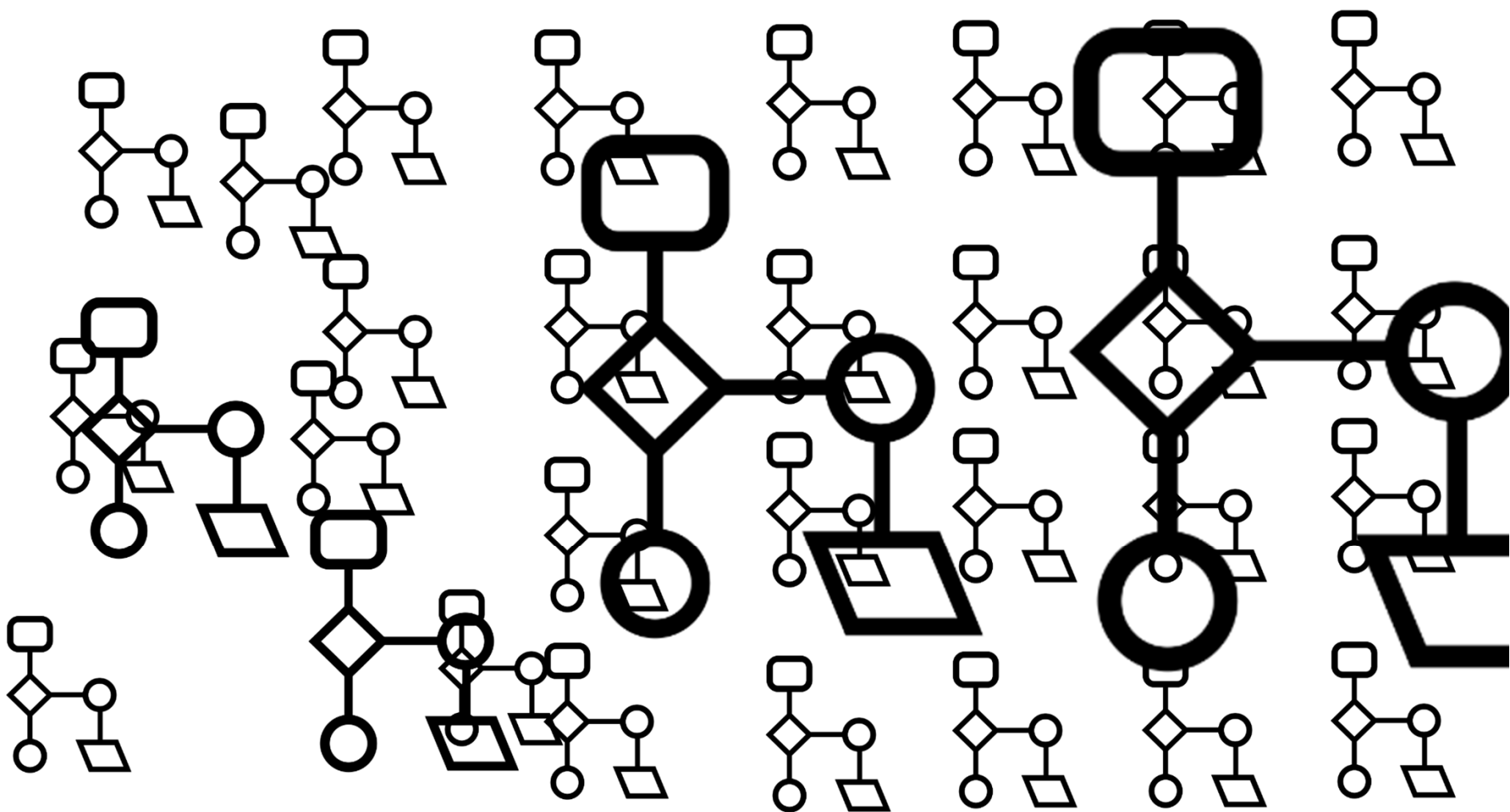
(Please select up to three)

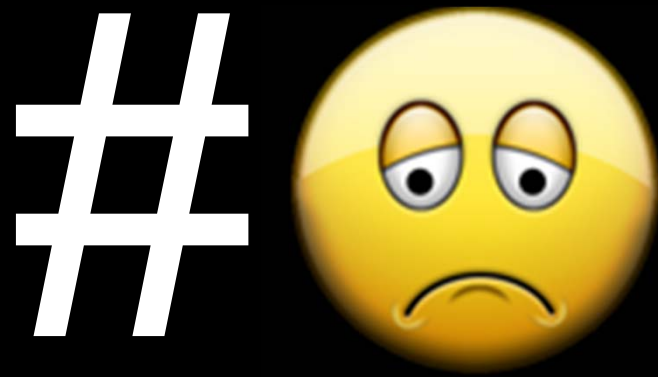


Base: 598 business and technology professionals

Source: Forrester's Q2 2016 Global State Of Artificial Intelligence Online Survey

#Models





Cognitive Models

10 characteristics + 10 behaviors + 10 needs =
30 cognitive models per customer

1 million customers x 30 models =
30 million cognitive models

#Data

A black and white photograph showing a close-up of a dark, textured surface covered with numerous water droplets of various sizes. The droplets are scattered across the frame, some appearing as bright highlights against the darker background. At the bottom of the image, there is a dark, semi-transparent rectangular box containing white text.

Data is like a drop of rain.




**It forms instantaneously in a
cloud...**



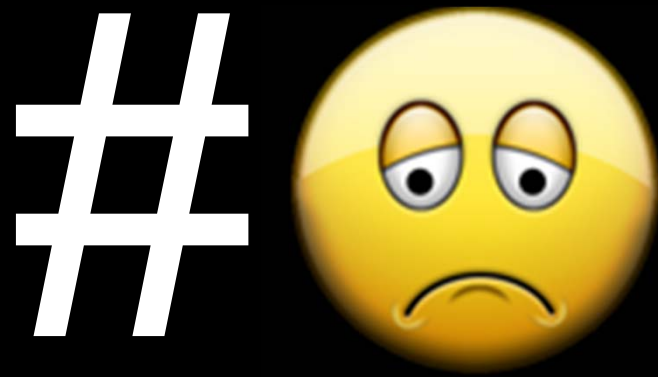
**...and travels far before it makes a
ripple.**



All data originates in real-time...

A wide-angle photograph of a large reservoir or dam. The water is a deep blue-green color. The surrounding landscape is rugged and mountainous, with dark, rocky slopes. The sky is a clear, bright blue with a few wispy white clouds. A thin red line, possibly a rope or cable, stretches across the water in the foreground. The overall scene is serene and majestic.

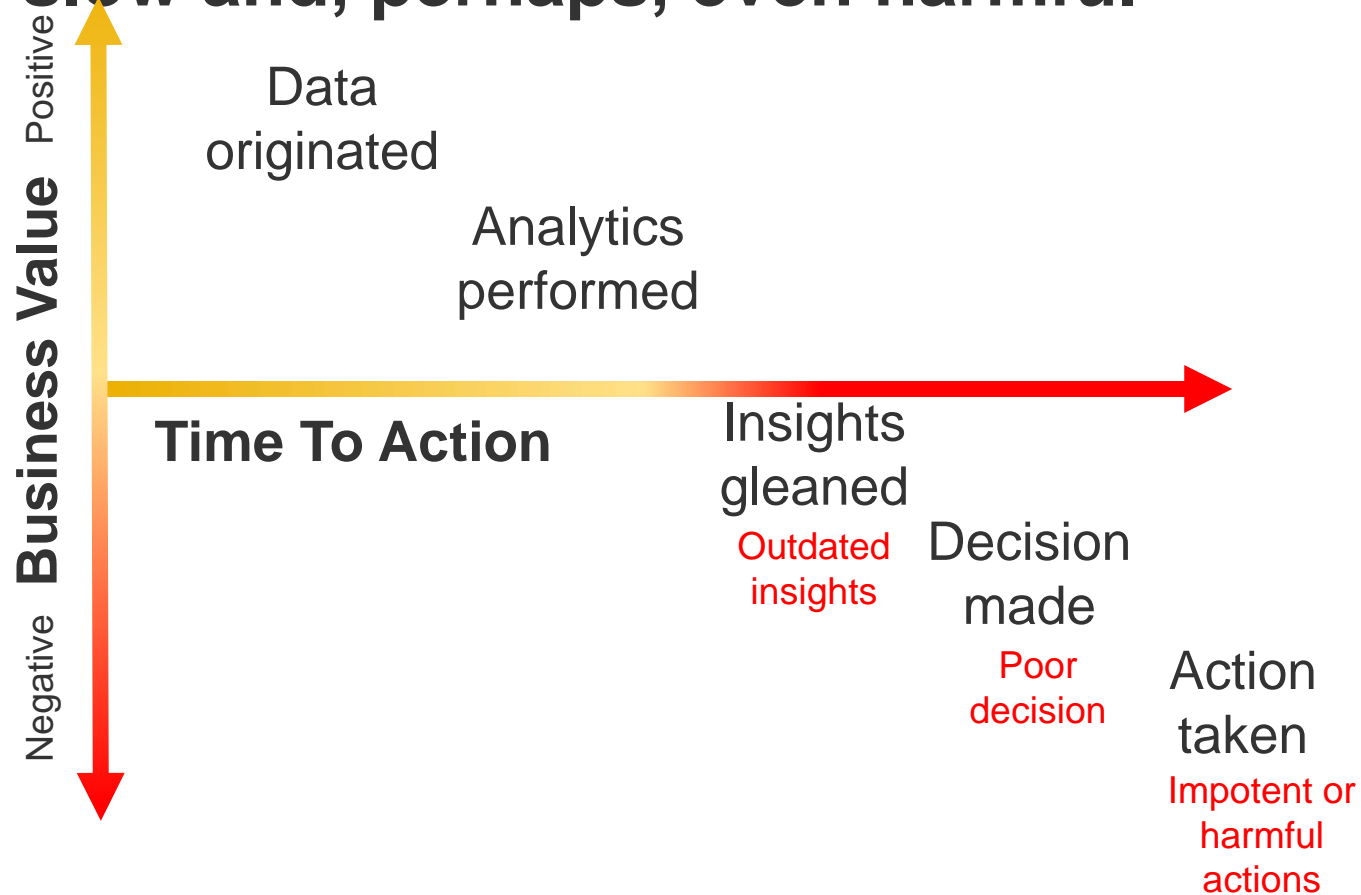
But, analytics to gain insights is usually done much, much later.



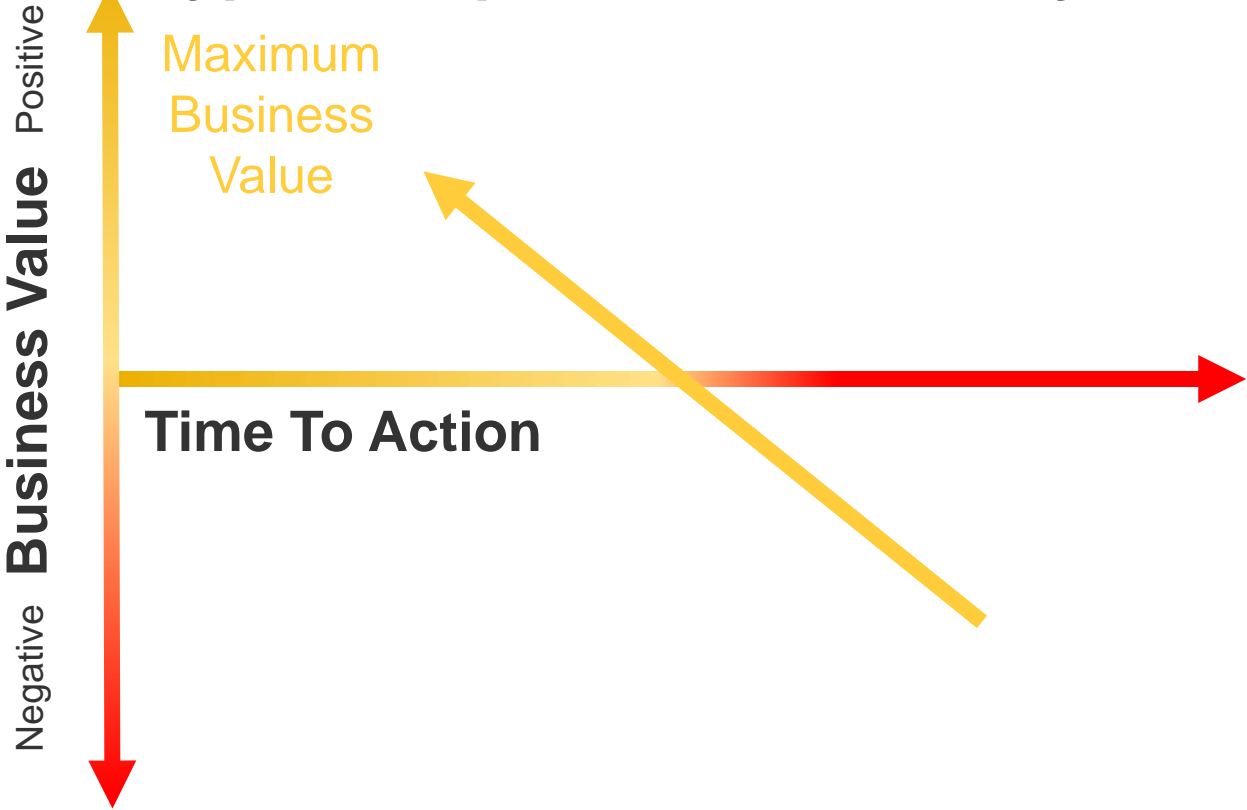


Insights are perishable.

Traditional analytics infrastructure is too slow and, perhaps, even harmful



Cognitive analytics require the power to hyper-compress the data analysis lifecycle





How can you know if you should you make an offer or send a gentle nudge **right now**?



Is this customer thinking about moving to a rival firm **right now**?



Crowborough CFRs @CrowboroughCFR · Feb 8

For those interested from my last post. [.flashaholics.co.uk/olight/olight-...](http://flashaholics.co.uk/olight/olight-...)
great head torch. **#olight**

CRB-03

Expand

← Reply ↻ Retweet ★ Favorite ⋮ More



ITOURLIGHT @ITOURLIGHT · Jan 30

Small but powerful flashlight **#Olight#itourlight**
itourlight.com/OLight-S10-Bat...

Expand

← Reply ↻ Retweet ★ Favorite ⋮ More



Kristen Williams @TheGunChick · Jan 23

Awesome video! Now that's some extreme testing! **#olight** fb.me/1cryjaS6

▶ View media

← Reply ↻ Retweet ★ Favorite ⋮ More



SOTG @studentofthegun · Jan 19

#View of **#LasVegas #Strip** for the **#Palazzo**. Thanks to Uncle Dick for the
invitation to the **#OLight...** instagram.com/p/jXn_7ZsF4o/

Expand

← Reply ↻ Retweet ★ Favorite ⋮ More



Followed by Manlyn Terrell

Battery Junction.com @BatteryJunction · Jan 8

Unique...
a rechargeable...

Expand

← Reply ↻ Retweet ★ Favorite ⋮ More



Hand Of Glory @HandOfGlory · Dec 21

Genuine night time music, up there with D. Wilson youtube.com/watch?

What are movers and shakers saying about equities that we cover **right now?**

#ML

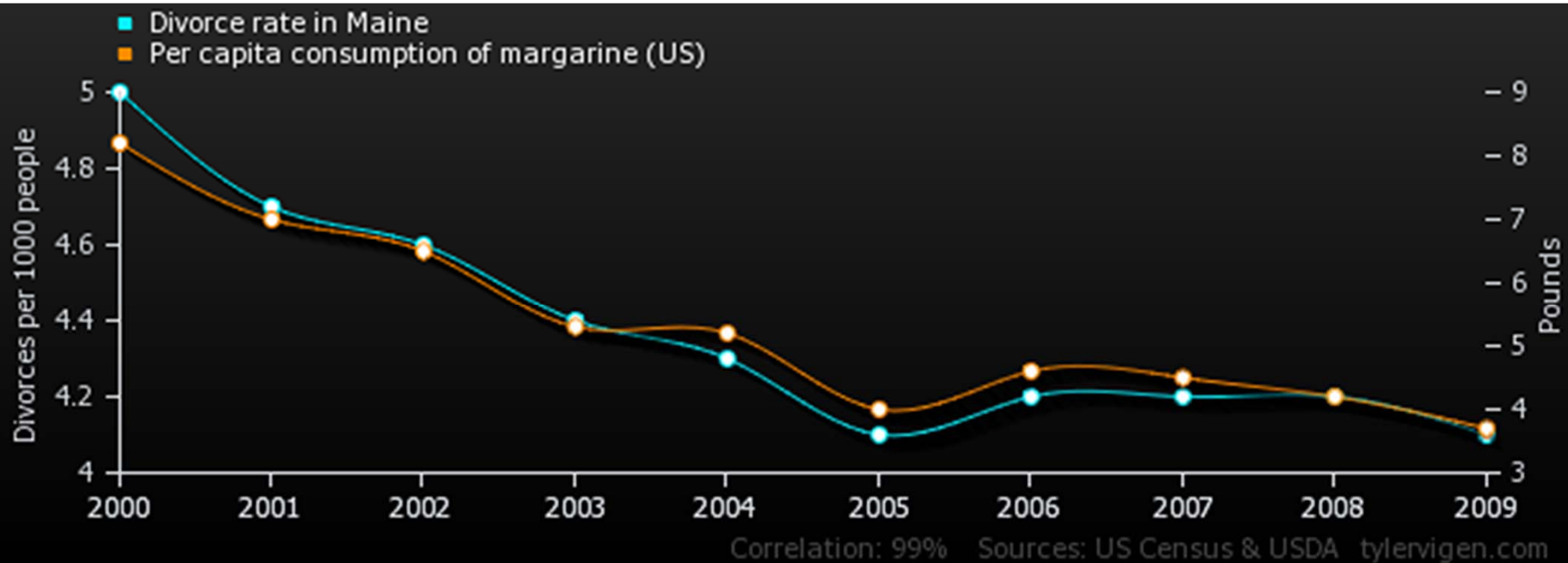
MACHINE

Algorithms that analyze data to find models –
models that can predict outcomes or
understand context with significant accuracy
and improve as more data is available.

LEARNING

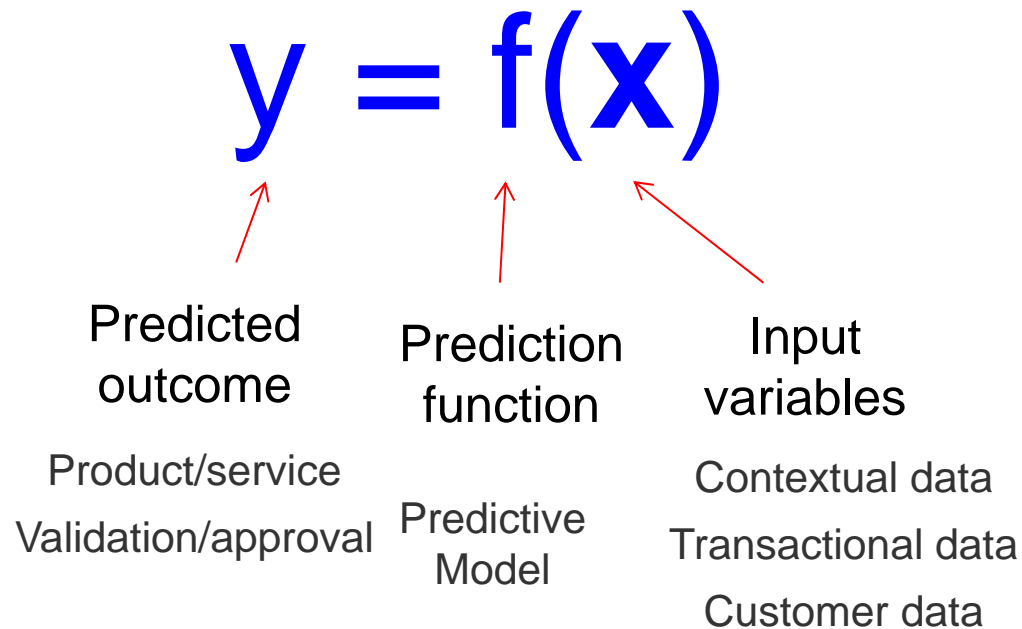
Models can be very powerful and profitable, but understand that:

- › Models are about probabilities, **NOT** absolutes
 - E.g. 78% chance you will like *Westworld*
- › Accurate models may **NOT** exist for every question
 - E.g. Elections, economic indicators, fashion, etc...
- › Machine learning models are based on correlation and probably **NOT** causative



Correlation does not imply causation.

Machine learning generated logic (models) are functions that takes input variables, apply a formula and/or rules to predict an outcome.



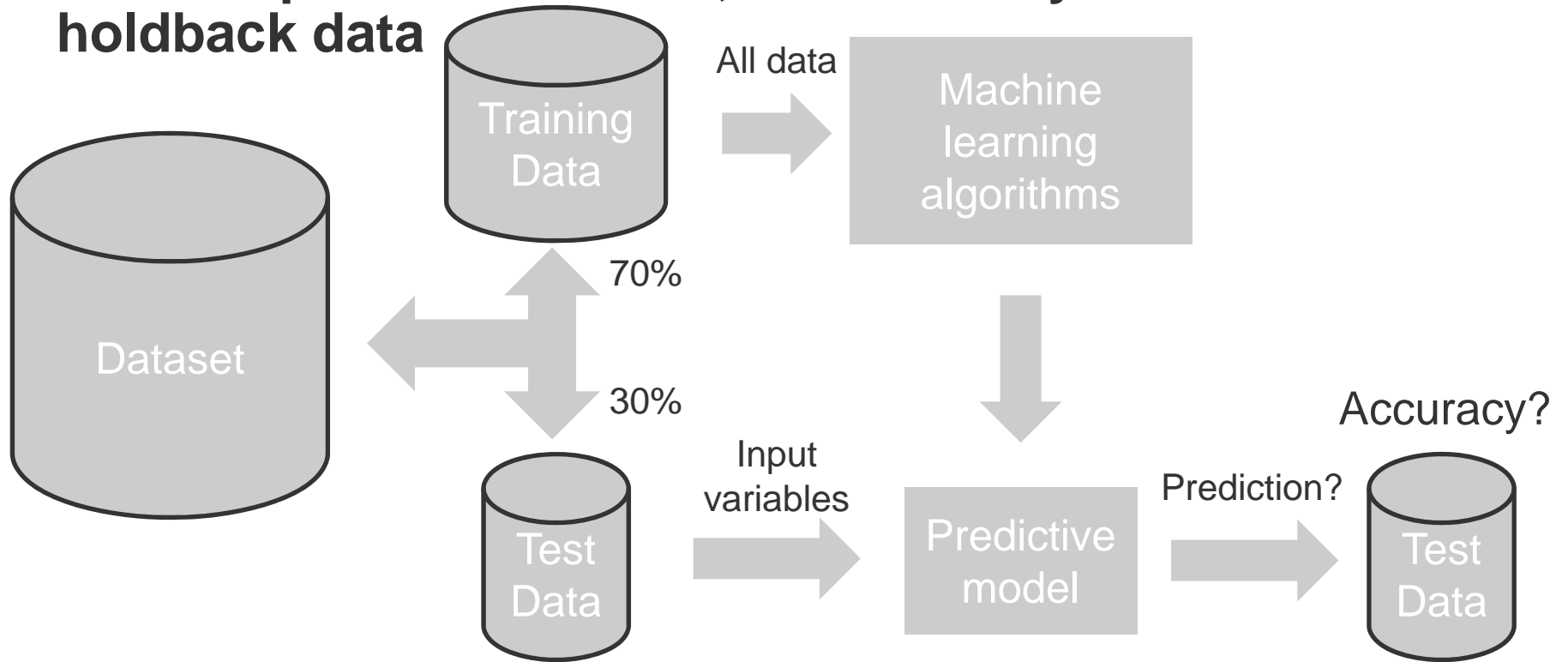


Should we play tennis?

Training data for “Play tennis?”

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

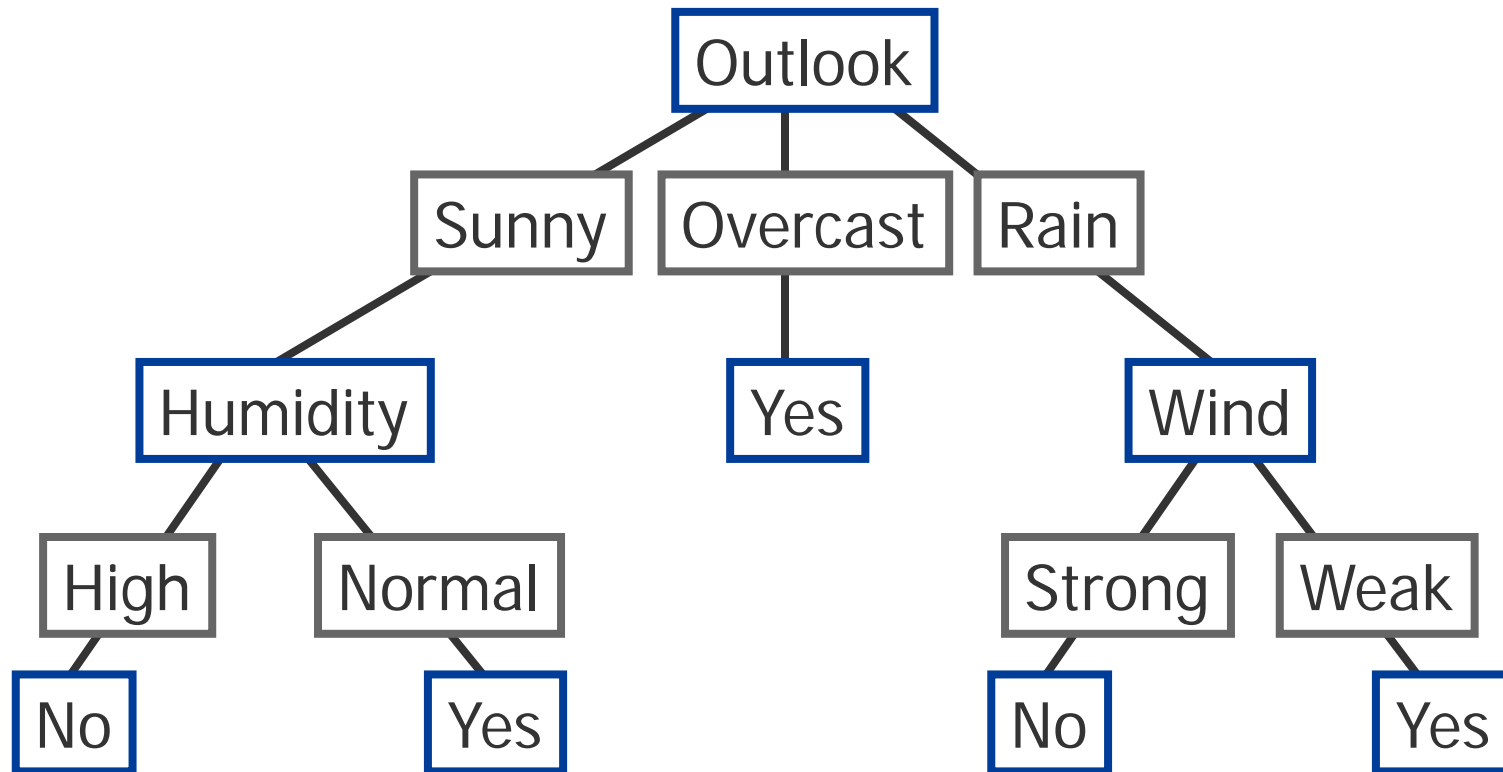
Machine learning algorithms use training data to create a predictive model; it's accuracy is tested on holdback data



Training data for “Play tennis?”

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Learned decision logic for “Play tennis?” is created automatically by machine learning



Unique for You

Michael, we think you might enjoy:

Die Hard

Our best guess for you 4.7



Customer average rating 4.1



Add DVD



Not Interested

YOUR RECENT ACTIVITY

05/18 We shipped The Wire: Season 1: Disc 1

SUGGESTIONS FOR YOU

You have [new suggestions](#) in Movies You'll ♥

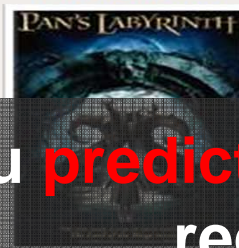
Critically-acclaimed Violent Sci-Fi & Fantasy

Your taste preferences created this row.

Critically-acclaimed
Sci-Fi & Fantasy

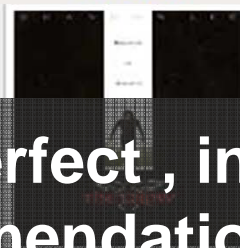
Violent

Pan's Labyrinth



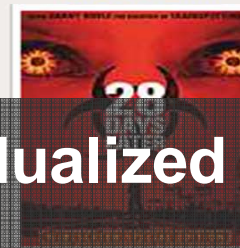
Not Interested

The Crow



Not Interested

28 Days Later



Not Interested

Alien: Collector's Edition



Not Interested

How can you **predict** a perfect, individualized product recommendation?



What offers should you make to your customer if they are eCommerce'ing **right now?**

Image source: iStockphoto



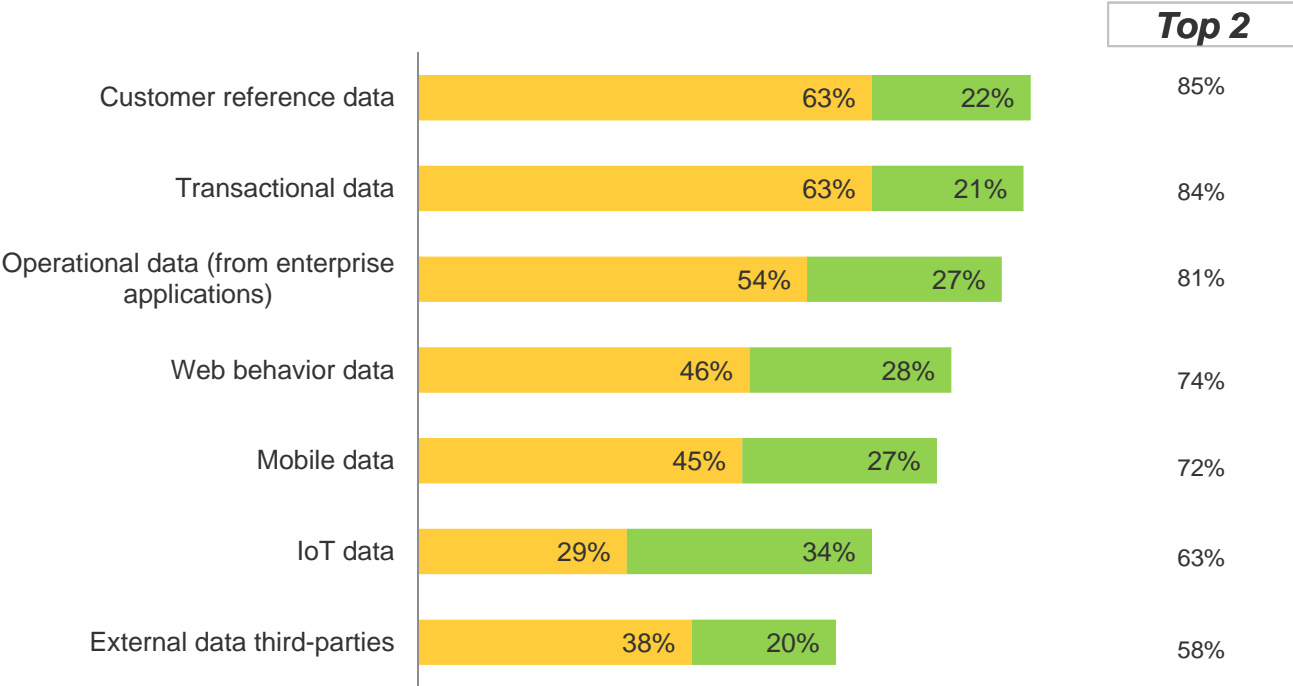
“Knowledge is **the power to predict in real-time.**”

Francis Bacon (1561–1626)

Founder of the modern scientific method to establish causation between phenomenon.

Data scientists recognize importance of transactional data in building predictive models

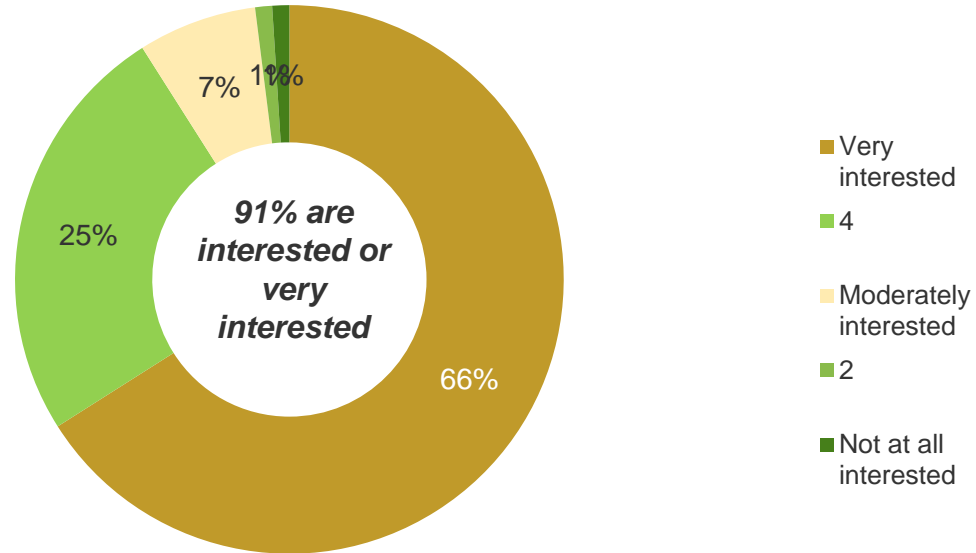
“Thinking specifically about building predictive models, which of the following best describes the importance of the data needed to build accurate models?”



Base: 100 data science and data analytics leaders at enterprises within the US
 Source: A commissioned study conducted by Forrester Consulting on behalf of IBM, April 2016

91% of data scientists express interest in real-time data use for modeling

“If there were no drawbacks (e.g. SLA concerns, resource consumption concerns) how interested would you be in having real-time data to use for modeling?”



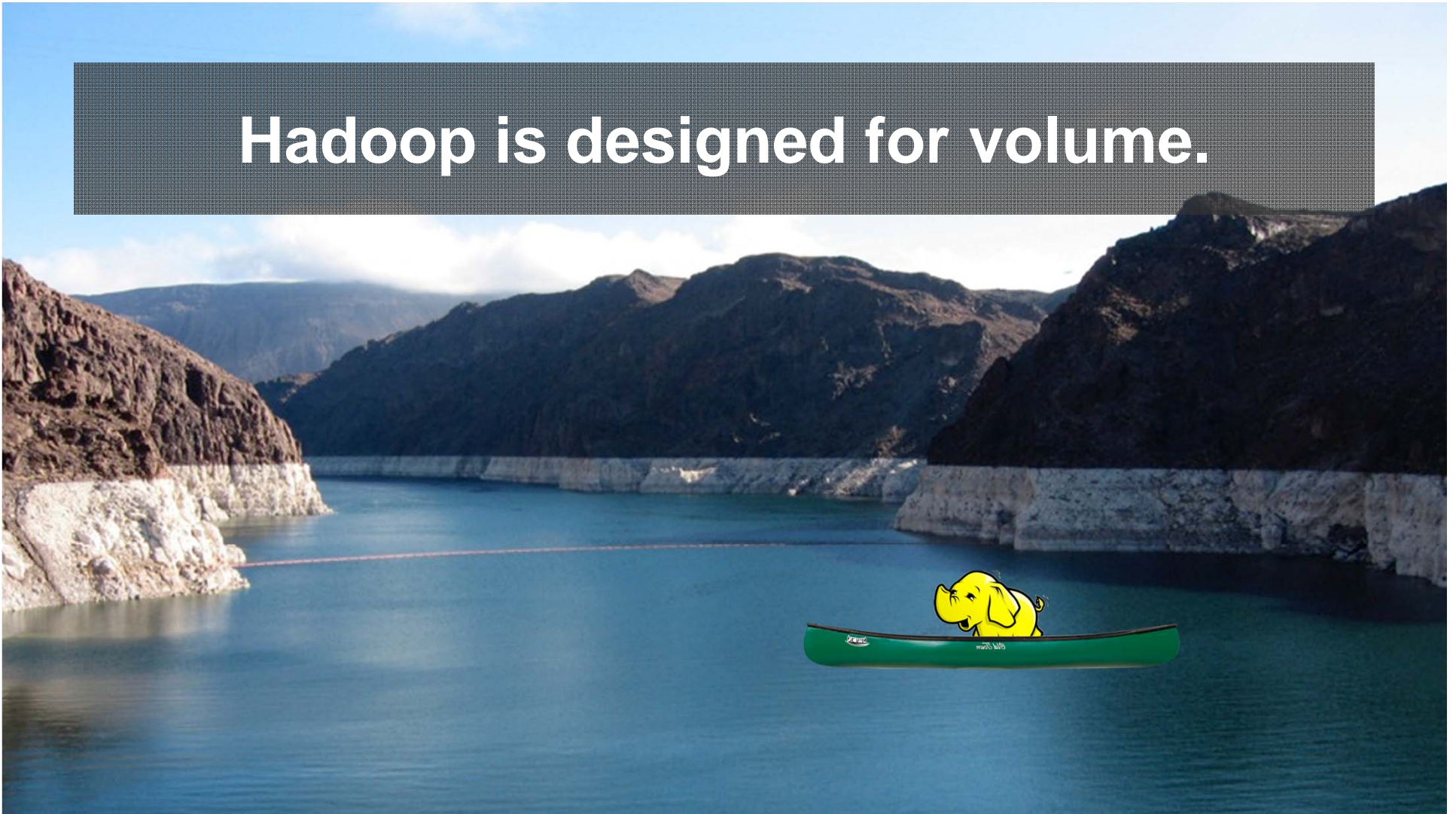
Base: 100 data science and data analytics leaders at enterprises within the US
Source: A commissioned study conducted by Forrester Consulting on behalf of IBM, April 2016

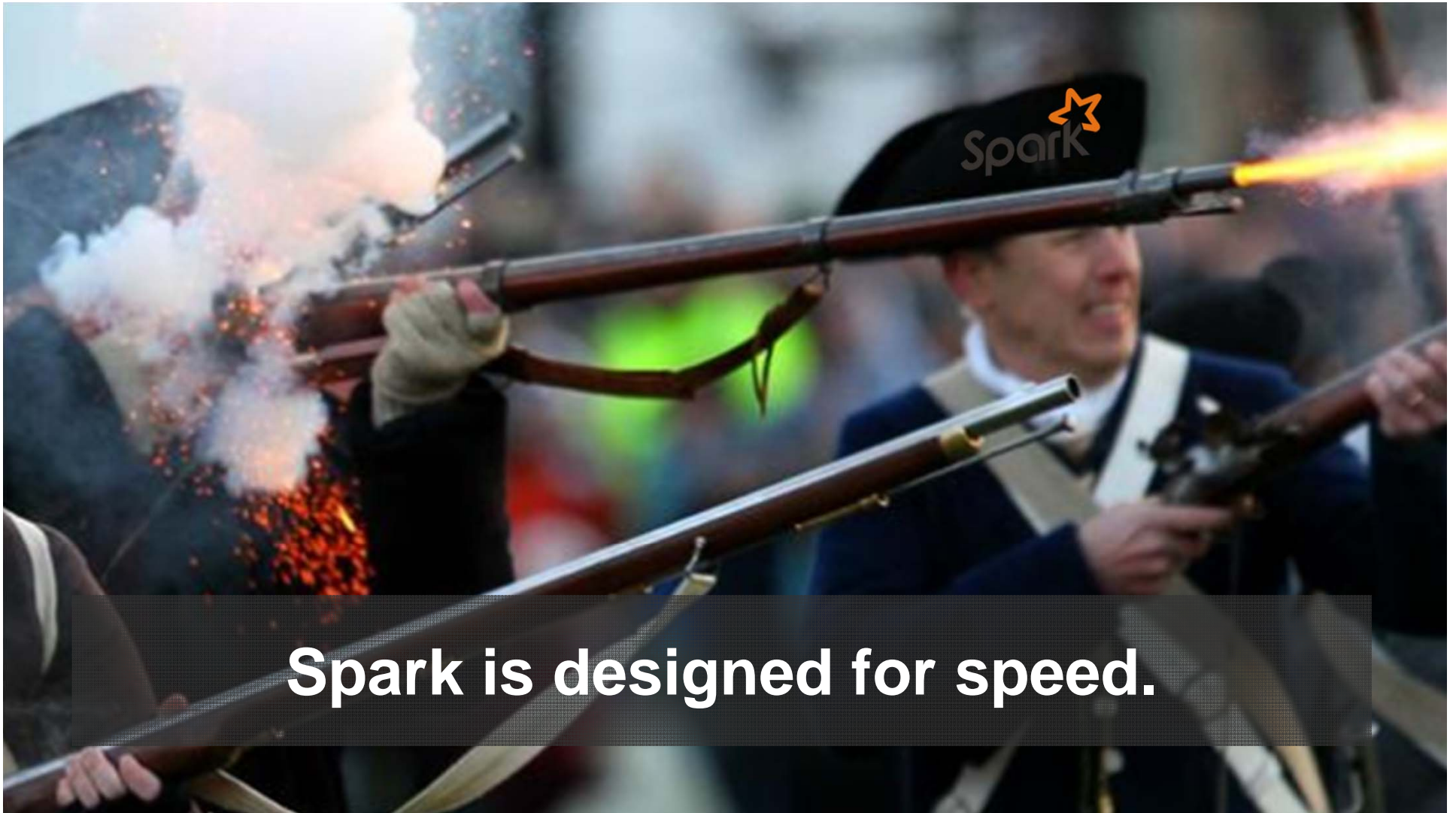
#



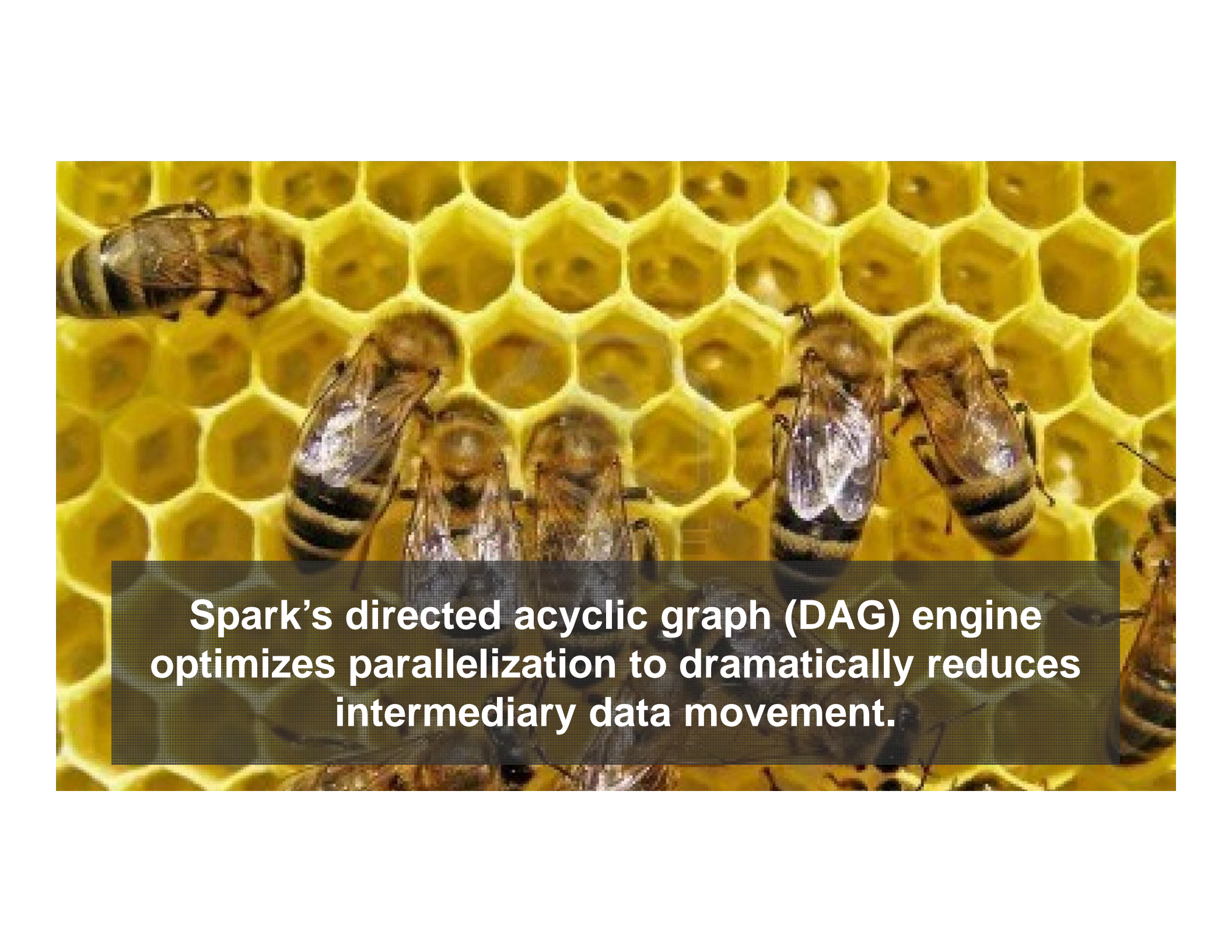
Spark 

Hadoop is designed for volume.





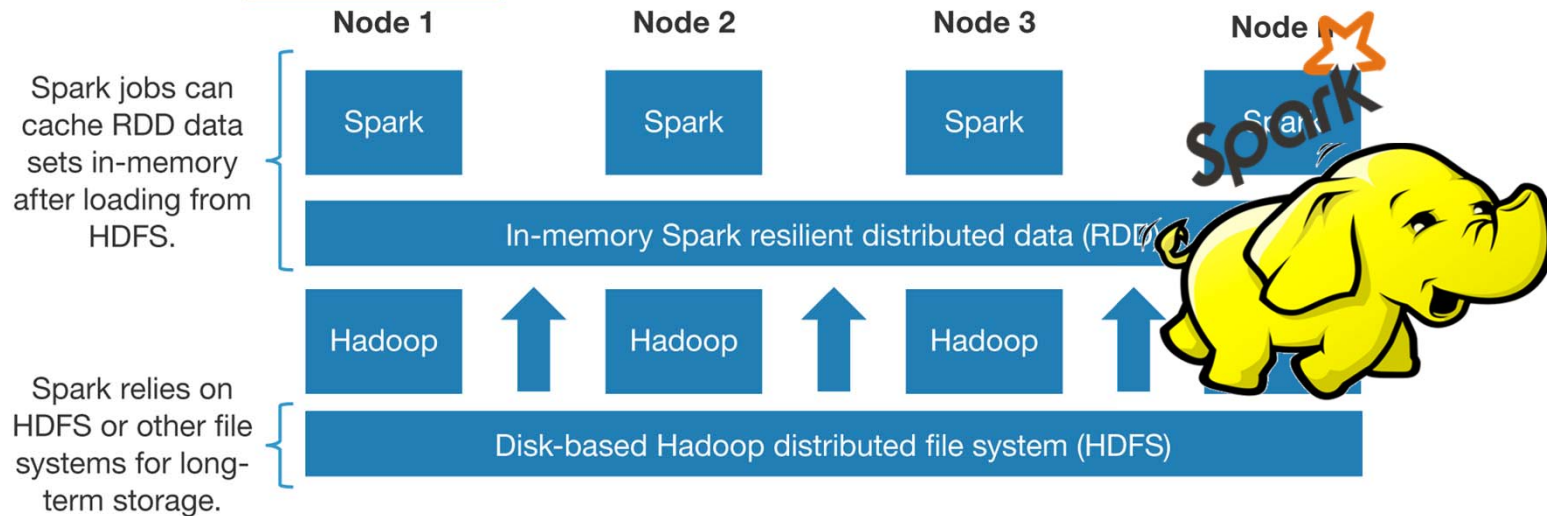
Spark is designed for speed.



Spark's directed acyclic graph (DAG) engine optimizes parallelization to dramatically reduce intermediary data movement.

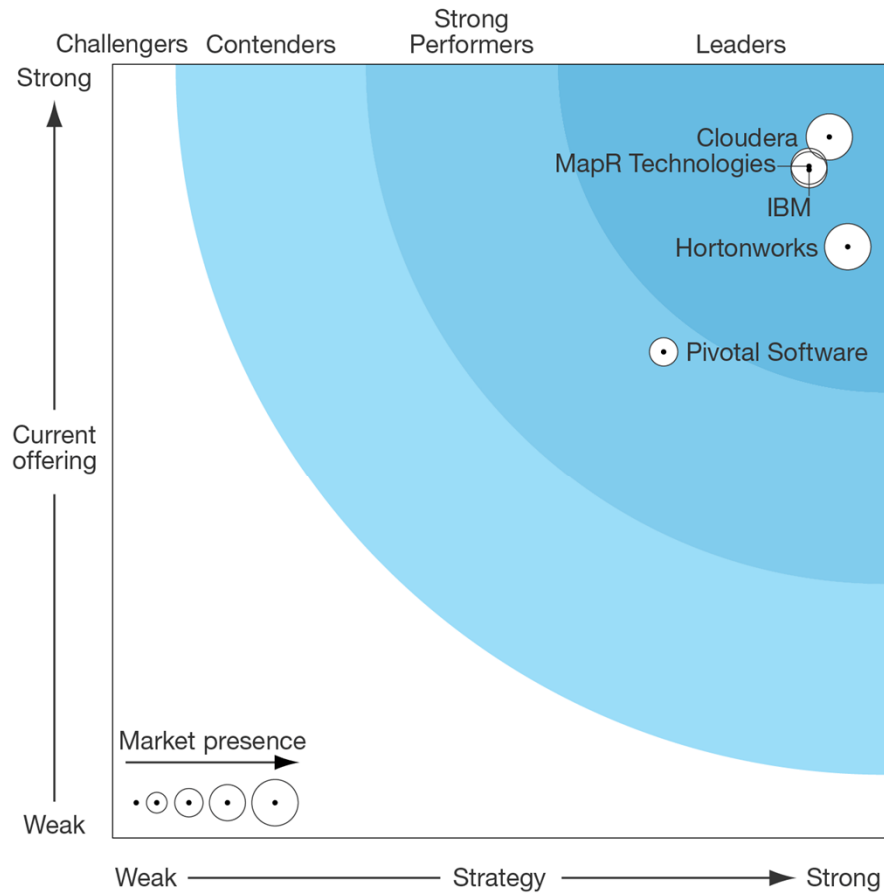
Spark and Hadoop can coexist in the same cluster.

Spark and Hadoop can coexist on the same nodes in a cluster.



121127

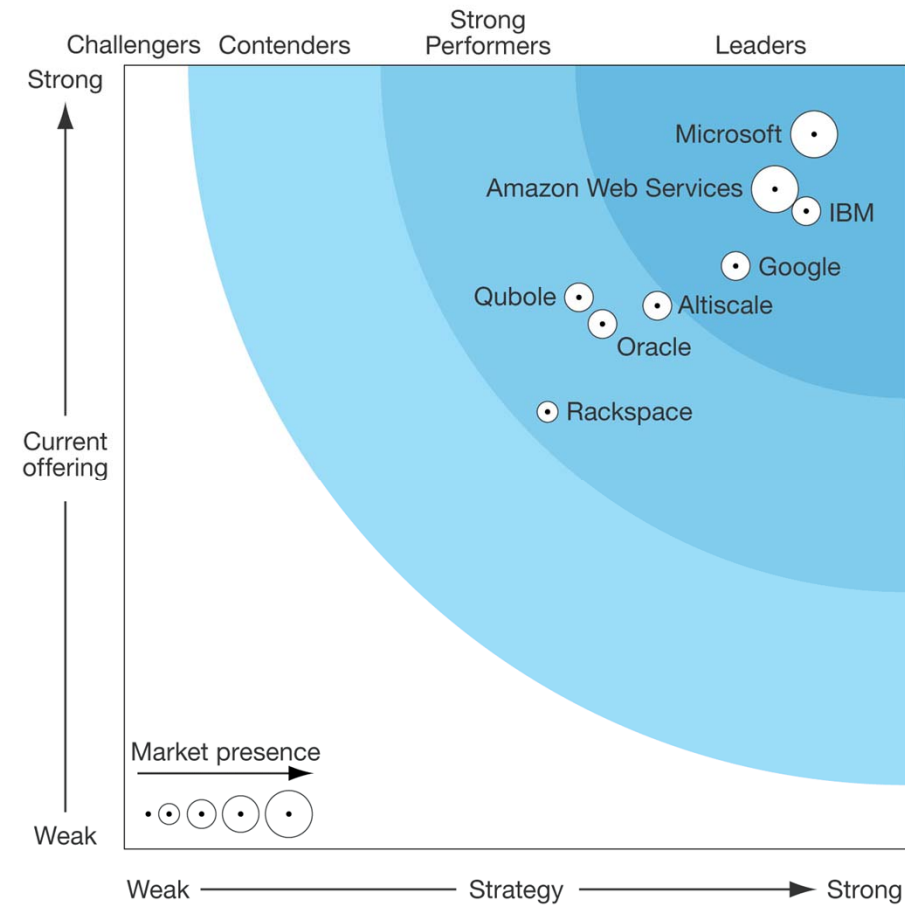
Source: Forrester Research, Inc. Unauthorized reproduction or distribution prohibited.



Source: Forrester Research, Inc. Unauthorized reproduction

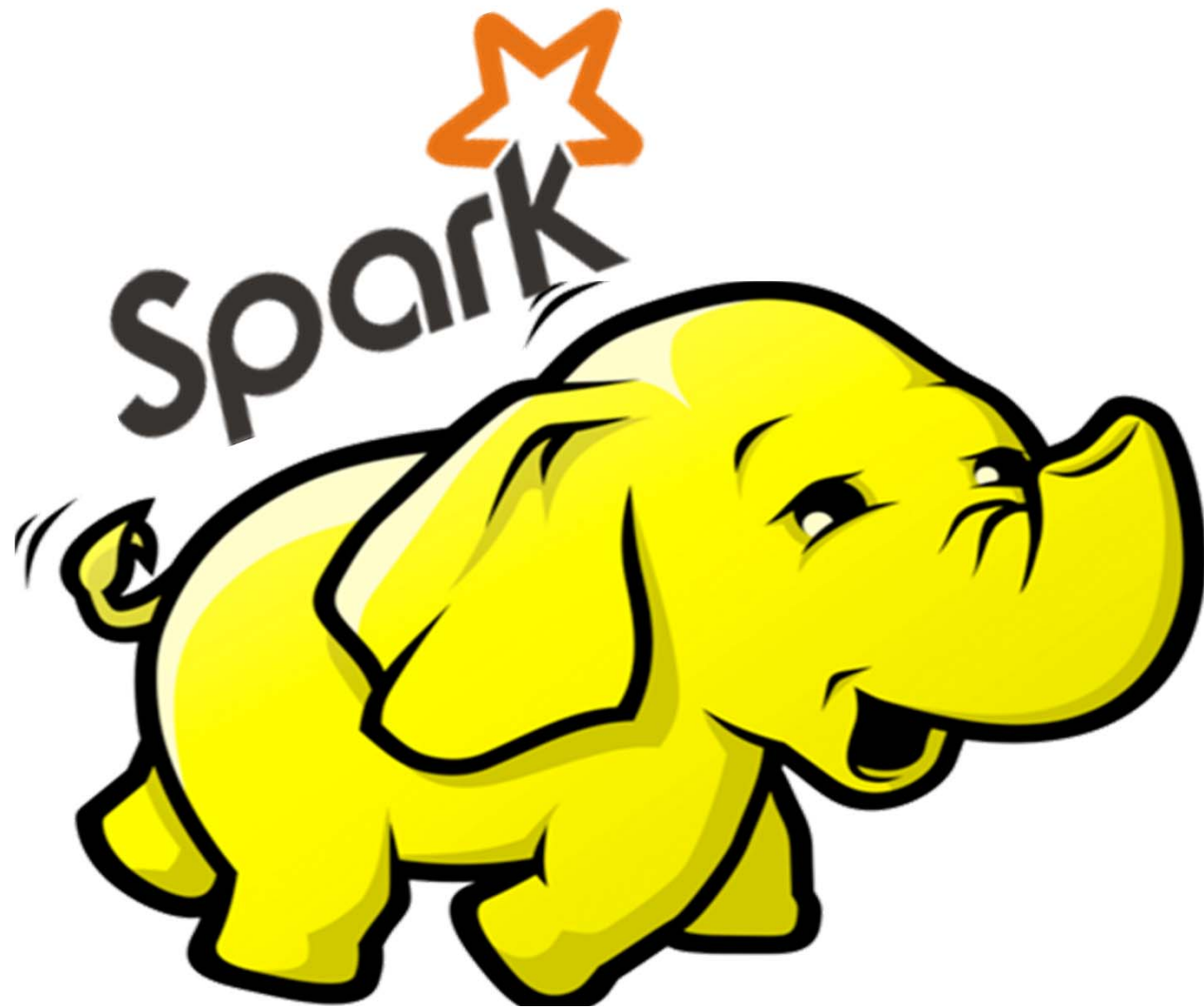
Source: Forrester Research

The Forrester Wave™: Big Data Hadoop Distributions, Q1 2016



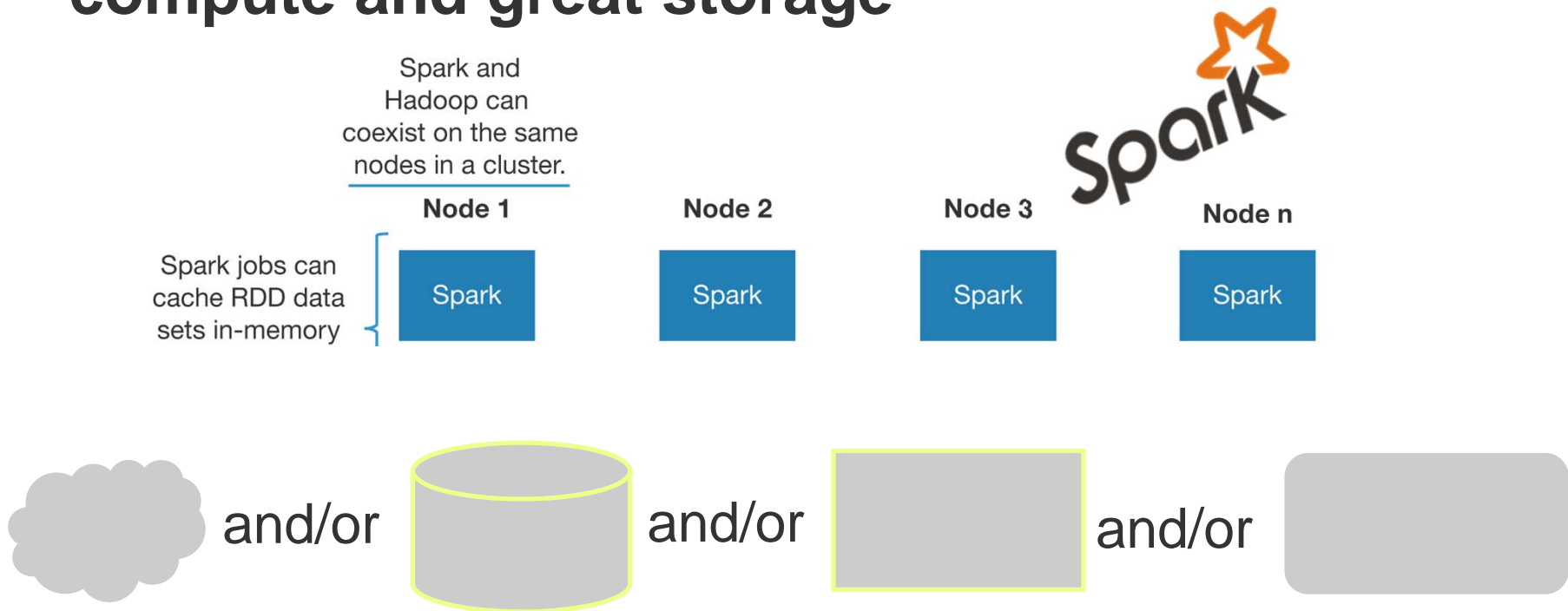
The Forrester Wave™: Big Data Hadoop Cloud, Q1 2016

Source: Forrester Research, Inc. Unauthorized reproduction

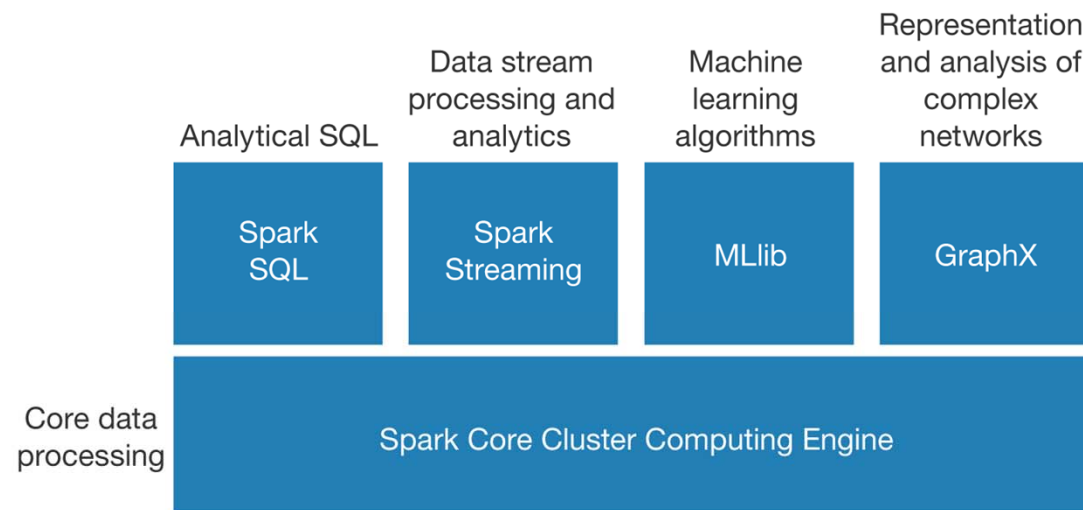




Spark doesn't need Hadoop; it just needs great compute and great storage



Spark also includes a growing number of specialized tools



Source: Apache Spark(<https://spark.apache.org/>)

121127

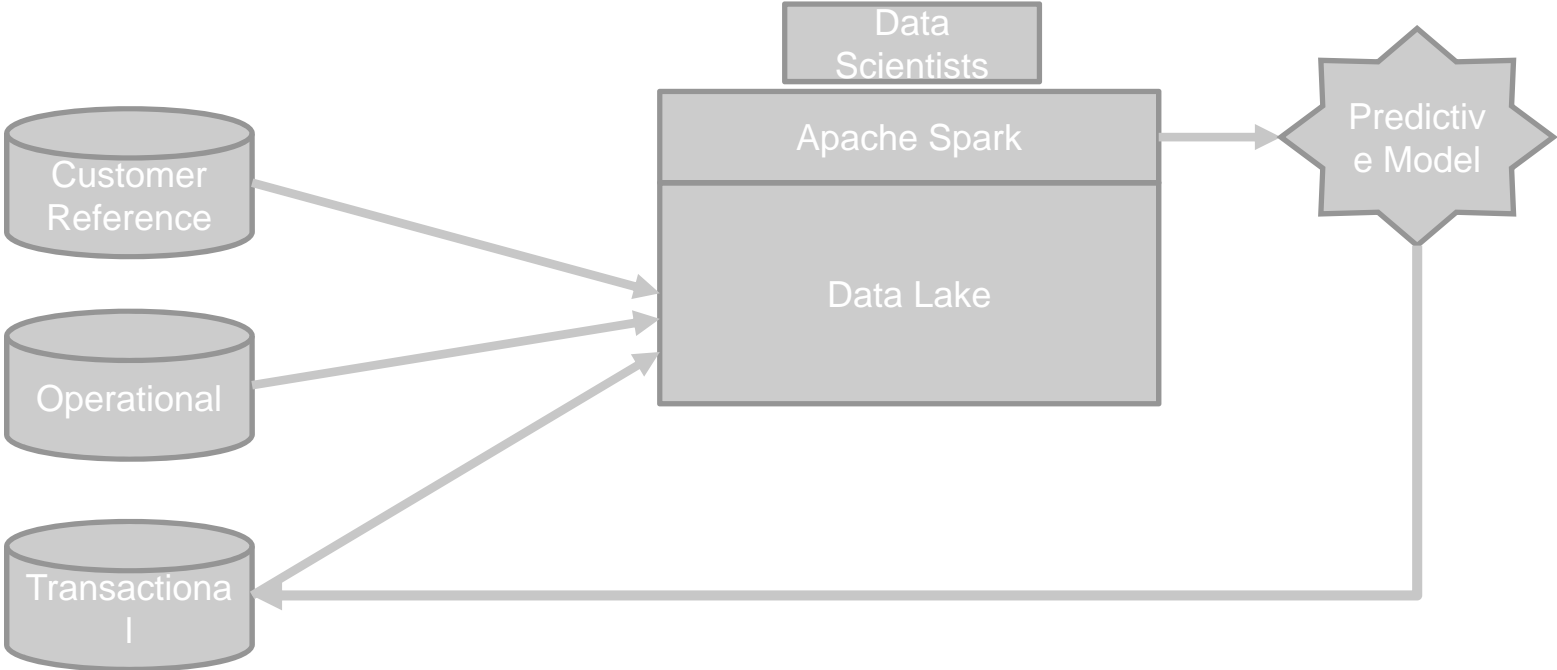
Source: Forrester Research, Inc. Unauthorized reproduction or distribution prohibited.

#DataGravity



Data lakes are repositories for data from multiple sources.

The data lake approach to analytics can require excessive movement of the data.



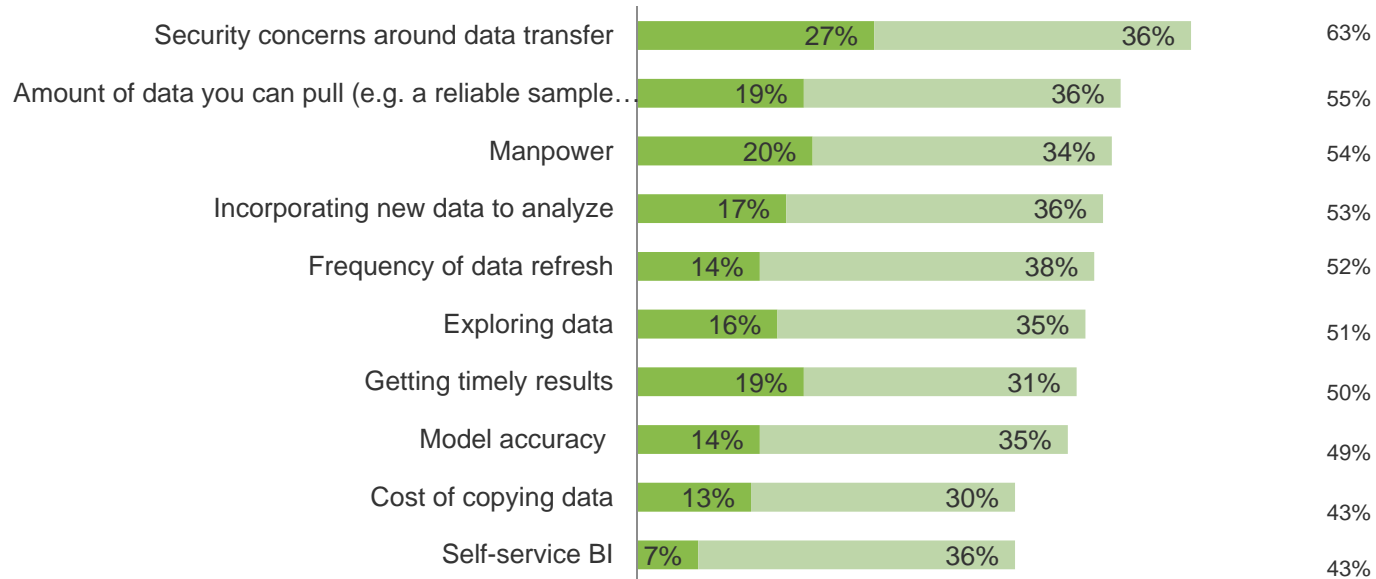
Moving transactional data in analytics models is challenging

“How challenging are each of the following as your organization tries to incorporate operational and transactional data into your analytics models?”

(Only including responses for very challenging and challenging)

■ Very challenging ■ 4

Top 2



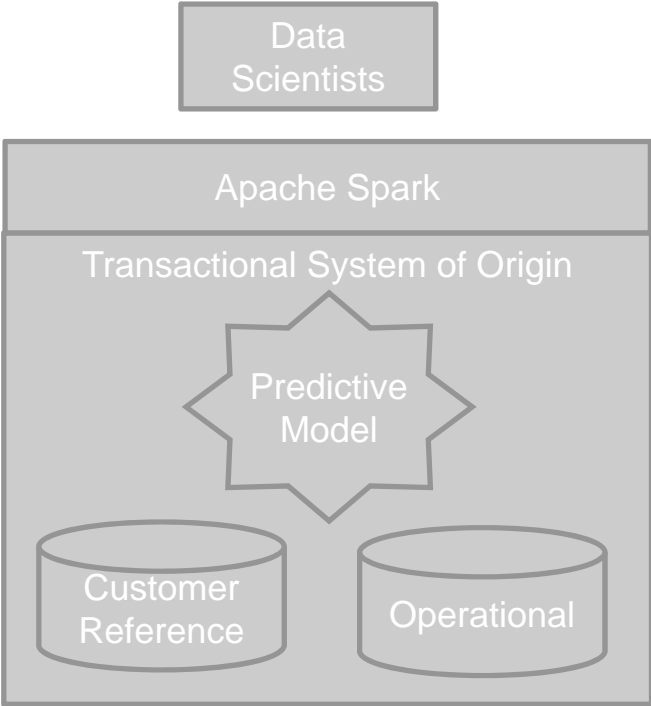
Base: 168 IT managers responsible for mainframe strategy at enterprises within the US, UK and Germany

Source: A commissioned study conducted by Forrester Consulting on behalf of IBM, April 2016



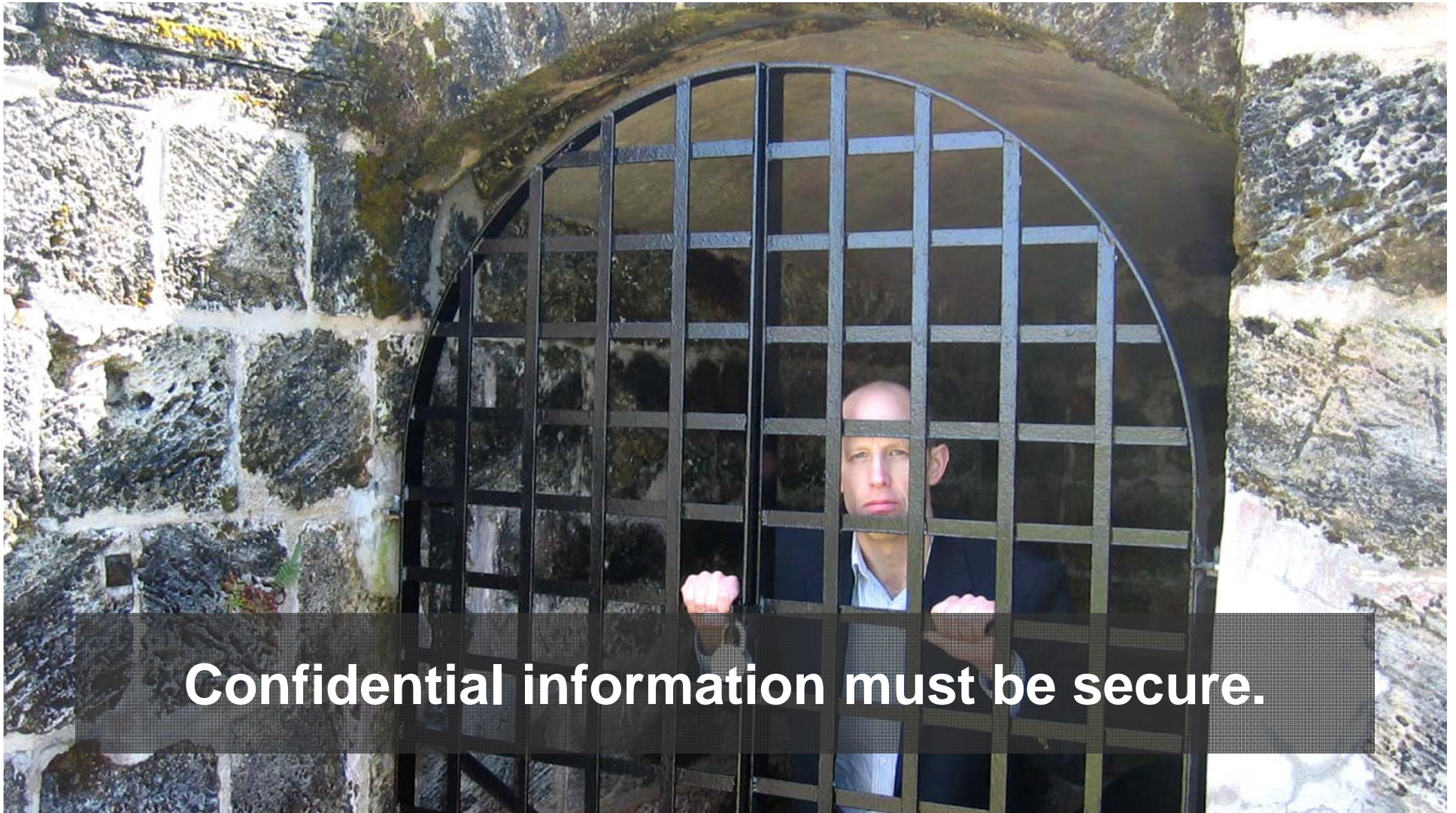
Data gravity approach performs analytics where the preponderance of the data originates.

The data gravity approach minimizes movement of data and reduces time to model



#





Confidential information must be secure.



Scale to handle any amount of data.



Analyze blazingly fast.



Fault-tolerance is non-negotiable for real-time applications.



Solutions and/or platforms must fit seamlessly into existing architectures.



Be open to the tools that data scientists need to use now and in the future.

A young woman with long, straight blonde hair is sitting on a white sofa. She is wearing a light blue t-shirt and is smiling broadly while looking down at a tablet computer she is holding in her hands. The background is a softly blurred indoor setting with light-colored walls and curtains.

Embed and act on real-time predictive models during transactions.



#Opportunity



Consider data gravity and inline predictive for advanced analytics.

Recommendations

- › **Measure data gravity for customer reference data, transactions, and operational data.**
- › **Deploy Apache Spark where data gravity is strongest.**
- › **Let data scientists build more accurate predictive models, faster.**
- › **Deploy predictive models directly within transactional systems.**

Thank you

Mike Gualtieri

mgualtieri@forrester.com

Twitter: @mgualtieri

