Arquitectura Big Data

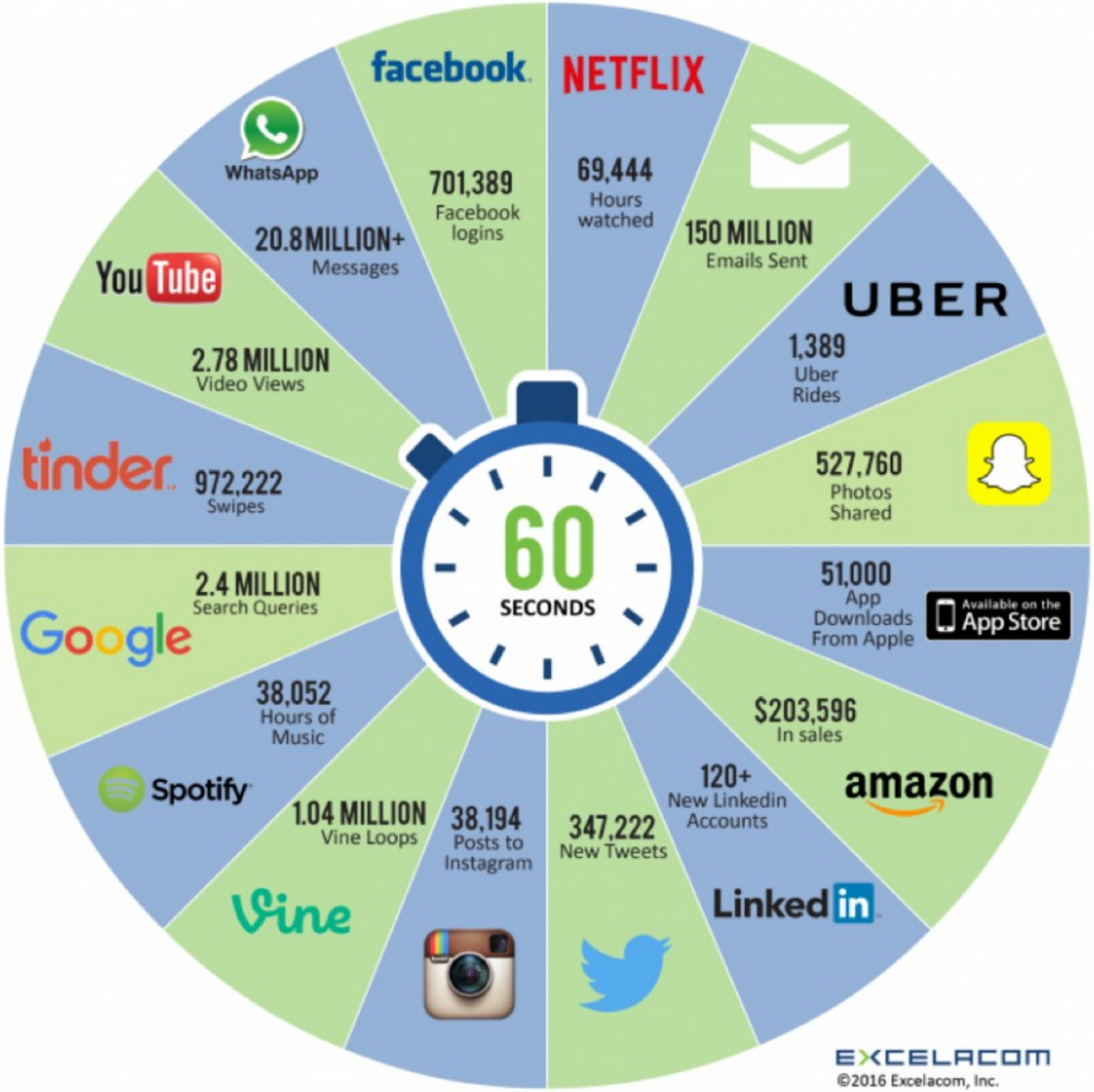
Angela Devia

ing.angela.devia@gmail.com

Científicas de Datos

www.cientificasdedatos.com

2016 What happens in an INTERNET MINUTE?



facebook
701,389 Facebook logins

NETFLIX
69,444 Hours watched

150 MILLION Emails Sent

UBER
1,389 Uber Rides

527,760 Photos Shared

51,000 App Downloads From Apple
Available on the App Store

$203,596 In sales
amazon

120+ New Linkedin Accounts
LinkedIn

347,222 New Tweets

38,194 Posts to Instagram

1.04 MILLION Vine Loops
Vine

38,052 Hours of Music
Spotify

2.4 MILLION Search Queries
Google

tinder 972,222 Swipes

2.78 MILLION Video Views
YouTube

WhatsApp
20.8 MILLION+ Messages

60 SECONDS

EXCELACOM

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

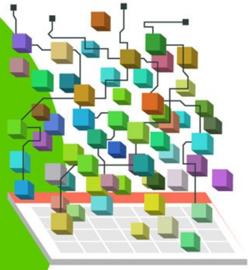
Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005



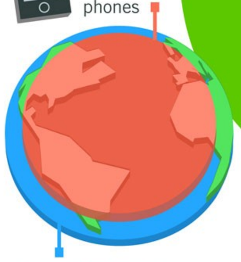
It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

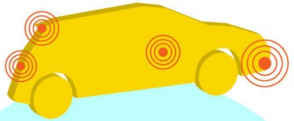
Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

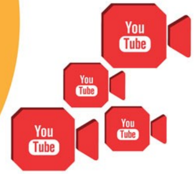
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month



30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

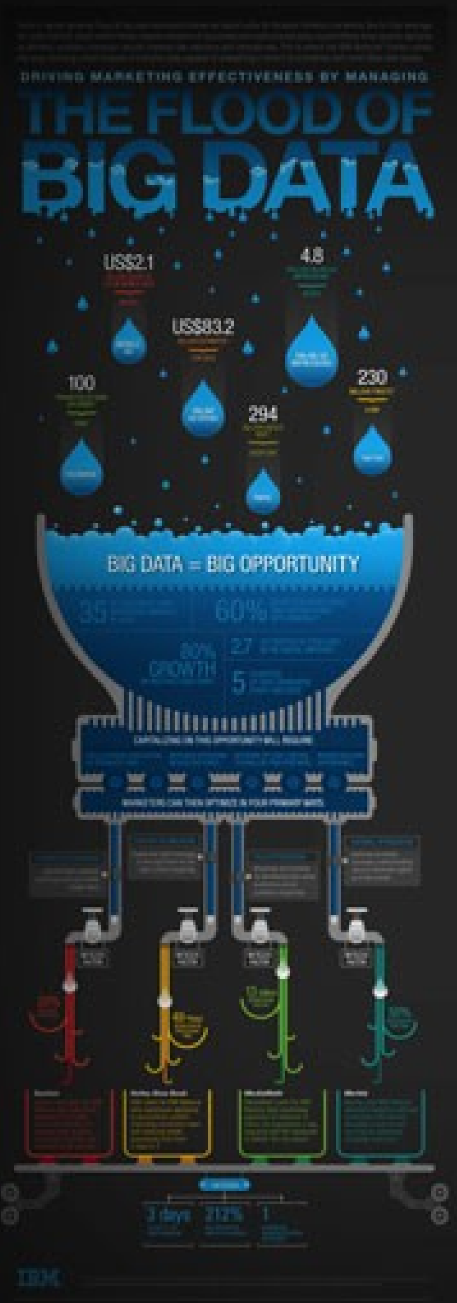
$3.1 TRILLION A YEAR

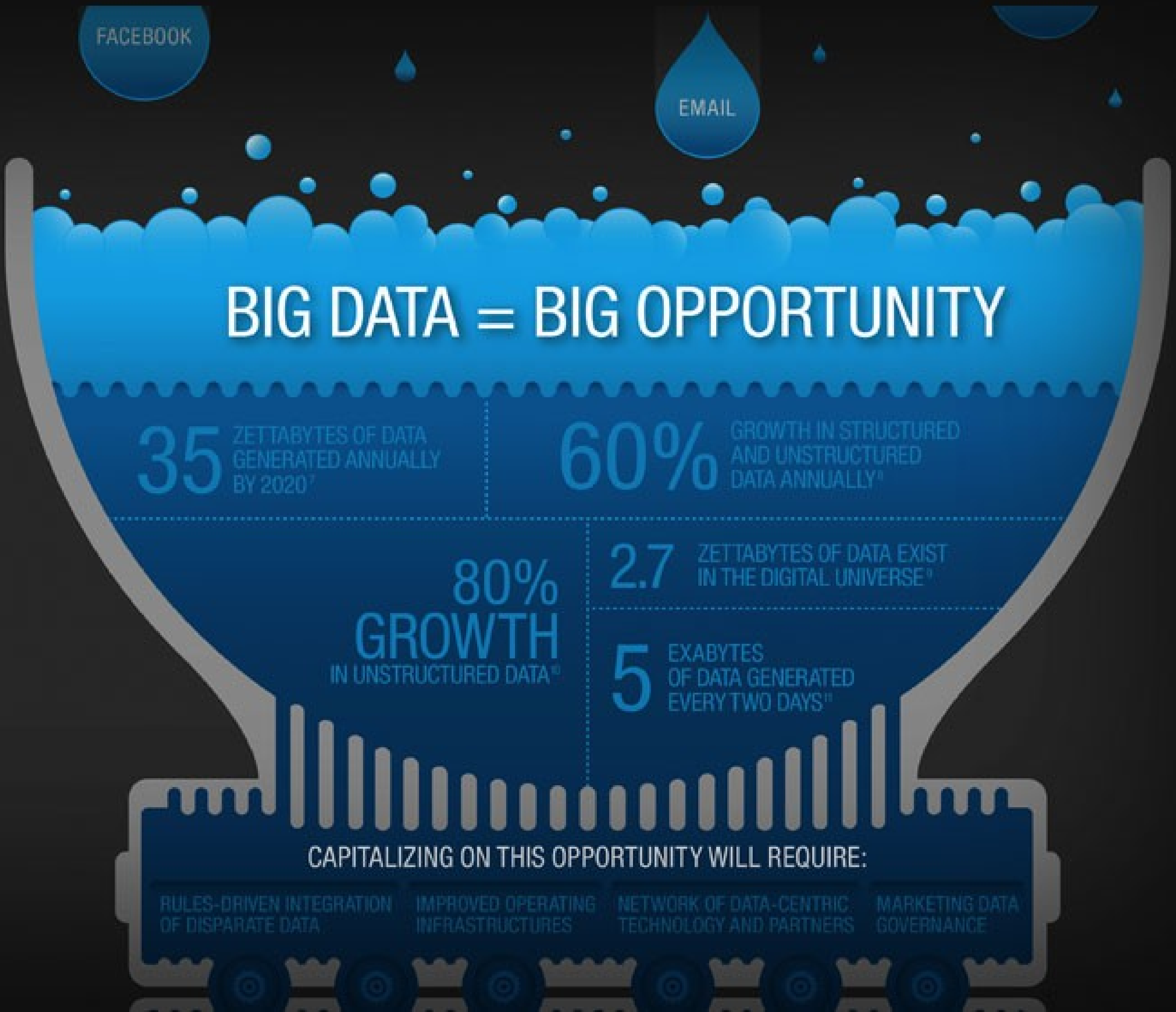
27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM





BIG DATA = BIG OPPORTUNITY

35 ZETTABYTES OF DATA GENERATED ANNUALLY BY 2020[7]

60% GROWTH IN STRUCTURED AND UNSTRUCTURED DATA ANNUALLY[8]

80% GROWTH IN UNSTRUCTURED DATA[10]

2.7 ZETTABYTES OF DATA EXIST IN THE DIGITAL UNIVERSE[9]

5 EXABYTES OF DATA GENERATED EVERY TWO DAYS[11]

CAPITALIZING ON THIS OPPORTUNITY WILL REQUIRE:

RULES-DRIVEN INTEGRATION OF DISPARATE DATA

IMPROVED OPERATING INFRASTRUCTURES

NETWORK OF DATA-CENTRIC TECHNOLOGY AND PARTNERS

MARKETING DATA GOVERNANCE

BIG DATA LANDSCAPE 2017

INFRASTRUCTURE

HADOOP ON-PREMISE
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

HADOOP IN THE CLOUD
amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere BigInsights, TREASURE DATA, Qubole, altiscale, CAZENA, CenturyLink

STREAMING
amazon web services, databricks, confluent, striim, GridGain, METAMARKETS, dataArtisans, DATATORRENT

NOSQL DATABASES
Google Cloud Platform, ORACLE, amazon DynamoDB, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, redislabs, influxdata

NEWSQL DATABASES
SAP Hana, Clustrix, Pivotal, nuodb, Cockroach LABS, memsql, splice, MariaDB, VOLTDB, citusdata, Trafodion, deepdb, paradigm4

GRAPH DBS
neo4j, ORACLE, OrientDB, InfiniteGraph, Objectivity

MPP DBS
TERADATA, VERTICA, NETEZZA, actian, kognitio, EXASOL, dremio

CLOUD EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, Infoworks

DATA TRANSFORMATION
talend, pentaho, alteryx, TRIFACTA, tamr, Paxata, StreamSets

DATA INTEGRATION
informatica, MuleSoft, snapLogic, TEALIUM, Segment, enigma, podium data, alooma, xplenty, ZALONI, import.io, Stitch

DATA GOVERNANCE
informatica, IBM, skyhigh, collibra, Alation

MGMT / MONITORING
amazon web services, New Relic, APPDYNAMICS, actifio, datadog, WAVEFRONT, unravel, splunk, Pepperdata, Anodot, Rocana, Numerify

STORAGE
amazon web services, Google Cloud Platform, Microsoft Azure, ALLUXIO, nimblestorage, Qumulo, COHO, panasas

CLUSTER SERVICES
amazon web services, kubernetes, docker, MESOSPHERE, CoreOS, pepperdata

APP DEV
Lightbend, rainforest, CASK

CROWDSOURCING
amazon mechanical turk, Upwork, WorkFusion, CrowdFlower

HARDWARE
Google TPU, ARM, nervana, Graphcore, MYTHIC, Cerebras, NVIDIA, Movidius, SCORTEX

ANALYTICS

DATA ANALYST PLATFORMS
Microsoft, pentaho, alteryx, Digital Reasoning, guavus, AYASDI, ATTIVIO, Datameer, Quid, ClearStory, OrigamiLogic, inter|ana, Bottlenose, ARIMO, ENDOR, MODE

DATA SCIENCE PLATFORMS
IBM, KNIME, dataiku, DOMINO, yhat, rapidminer, CONTINUUM ANALYTICS, Alpine, ALGORITHMIA, Angoss

BI PLATFORMS
Microsoft, amazon web services, DOMO, salesforce Wave Analytics, ARCADIA DATA, birst, atscale, GoodData, Sisense

VISUALIZATION
tableau, SAP, Google Cloud Platform, Qlik, looker, CELONIS, Periscope, MAPD, CHARTIO, ZEPL, ZOOMDATA, plotly

VERTICAL ANALYTICS
PREDIX, C3 IoT, CAPE, UPTAKE, TACHYUS, Alluvium, datarama

STATISTICAL COMPUTING
sas, SPSS, MATLAB

DATA SERVICES
Palantir, Mu Sigma, OPERA, DATA SCIENCE, kaggle, EXL, DataKind, FF

MACHINE LEARNING
Azure Machine Learning, amazon web services, Google Cloud Platform, H2O.ai, DataRobot, context relevant, VISENZE, SKYTREE, bonsai, DATARPM, deepsense.io, nutonian

HORIZONTAL AI
IBM Watson, Cortana, Face++, sentient, Voyager Labs, clarifai, Affectiva, CognitiveScale, SenseTime, Numenta, Gronocom, PETUUM, nara, OSARO, CURIOUS AI, BLUE VISION

SPEECH & NLP
Google Cloud Platform, twilio, amazon alexa, Narrative Science, semantic machines, MindMeld, Yseop, NUANCE, ARRIA, IDIBON, TalkIQ, snips, Gridspace, synergy

SEARCH
Attivio, elastic, exalead, ORACLE ENDECA, ThoughtSpot, Algolia, Lucidworks, swiftype, MAANA, alphasense, Searchlink, SINEQUA

LOG ANALYTICS
splunk, sumologic, loggly, kibana, logz.io

SOCIAL ANALYTICS
Hootsuite, NETBASE, DATASIFT, tracx, synthesio, simplereach, bitly, predata

WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics, mixpanel, AMPLITUDE, sumAll, Airtable, retention science, SIGOPT, granify, custora

CROSS-INFRASTRUCTURE/ANALYTICS
amazon web services, Google Cloud Platform, Microsoft, IBM, SAP, Hewlett Packard Enterprise, sas, 1010 data, vmware, TIBCO, TERADATA, ORACLE, NetApp

APPLICATIONS – ENTERPRISE

SALES
Aviso, CHORUS, INSIDESALES.COM, conversica, clari, AVISO, TACT, fuse machines, TROOPS

MARKETING - B2B
RADIUS, App Annie, EVERSTRING, Lattice, Infer, MINTIGO, Reflektion, 6sense, tubular, DataFox, ENGAGIO

MARKETING - B2C
Zeta, bloomreach, blue yonder, persado, ACTIONIQ, kahuna, BLUECORE, SAILTHRU, QUANTIFIND, mparticle, Amplero

CUSTOMER SERVICE
MEDALLIA, zendesk, CLARABRIDGE, Gainsight, CLICKFOX, NGDATA, DigitalGenius, aspect, automat, agent.ai, msg.ai, frame.ai

HUMAN CAPITAL
HireVue, entelo, hiQ, GIGSTER, textio, Wade&Wendy, Stella, pymetrics

LEGAL
RAVEL, Seal, Everlaw, JUDICATA, Brevia, ROSS

FINANCE
anaplan, zuora, tidemark, SAP S/4HANA

ENTERPRISE PRODUCTIVITY
slack, x.ai, facebook, ORACLE, Clara, talla, butter.ai, KASISTO

BACK OFFICE AUTOMATION
HyperScience, captricity, AppZen

SECURITY
TANIUM, CYLANCE, StackPath, DARKTRACE, illumio, CODE42, ThreatMetrix, DataGravity, VECTRA, CipherCloud, Guardian Analytics, Cybereason, ANOMALI, siftscience, SIGNIFYD, SentinelOne, SecurityScorecard, BlueTalon, Recorded Future, feedzai, Sqrrl, SOCURE, PERIMETER X, Fraugster

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus, MediaMath, criteo, xAd, integral ad science, rocketfuel, theTradeDesk, OpenX, MOAT, distillery, Adgorithms, drawbridge, LiveIntent, TAPAD, DataXu, gumgum, Oppizi

EDUCATION
Knewton, Clever, declara, kidaptive, PANORAMA, knowre, gradescope

GOVERNMENT
Socrata, OPENGOV, mark43, enigma, FiscalNote, OpenDataSoft

FINANCE - LENDING
OnDeck, affirm, Kreditech, AVANT, Zestfinance, INSIKT, TALA, MoneyLion, TrueAccord, trooly, cignifi, aire, Active.Ai

FINANCE - INVESTING
Dataminr, KENSHO, Quantopian, NUMERAI, iSENTIUM, claritymoney, ALGORIZ

REAL ESTATE
Opendoor, VTS, CREDIFI, reonomy, COMPSTAK

INSURANCE
metromile, Lemonade, CYENCE, Shift Technology, Tractable

HEALTHCARE
FLATIRON, KYRUUS, Healthtap, METABIOTA, Gingerio, Glow, Babylon, COTA, zebra, ovia, AiCure, enlitic, Qventus, imagia, DAVIAS, IMADEN, Kang Health

LIFE SCIENCES
23andMe, color, zymergen, BenevolentAI, ZEPHYR HEALTH, 3Scan, Clear Labs, Citrine, twoXAR, Phosphorus, Atomwise, deep genomics

TRANSPORATION
UBER, ZOOX, TESLA, nuTonomy, CLEARPATH, nauto, drive.ai, pilot.ai, OPTIMUS RIDE, OTTO, nexar, comma.ai, CivilMaps

AGRICULTURE
FARMERS BUSINESS NETWORK, FarmersEdge, FarmLogs, BLUERIVER, mavrx, prospera

COMMERCE
Instacart, STITCH FIX, RetailNext, HowGood

OTHER
eHarmony, stem, rethink robotics, hopper, BOXEVER, celect, VERDIGRIS, duetto, Second Spectrum

OPEN SOURCE

FRAMEWORK
hadoop, YARN, TEZ, Flink, MESOS, Spark, CDAP

QUERY / DATA FLOW
Spark SQL, presto, SLAMDATA, APACHE DRILL, Google Cloud Dataflow

DATA ACCESS
cassandra, nifi, mongoDB, RethinkDB, CouchDB, OPENTSDB, riak, HBASE, Spanner, ACCUMULO

COORDINATION
talend, Apache Zookeeper, Apache Ambari

STREAMING
Spark, APEX, Flink, beam, kafka, druid, STORM

STAT TOOLS
R, ScalaLab, NumPy, SciPy

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, theano, Caffe, torch, MXNet, CNTK, DMTK, Keras, SINGA, OpenAI, neon, FeatureFu, DSSTNE, mllib, Chainer, VELES, WEKA, DIMSUM, DL4J, Aerosolve

SEARCH
elasticsearch, Solr, Lucene

LOG ANALYSIS
elasticsearch, kibana, logstash

VISUALIZATION
BEAKER, Rodeo

COLLABORATION
jupyter, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry

DATA SOURCES & APIS

HEALTH
JAWBONE, VALIDIC, practice fusion, fitbit, GARMIN, Human API, kinsa

IOT
GE Digital, UPTAKE, ThingWorx, helium, samsara

FINANCIAL & ECONOMIC DATA
Bloomberg, THOMSON REUTERS, DOW JONES, S&P CAPITAL IQ, CB Insights, xignite, quandl, YODLEE, PREMISE, Estimize, SECOND MEASURE, Eagle Alpha, StockTwits, PLAID, mattermark

AIR / SPACE / SEA
PLANET LABS, Airware, spire, SKYCATCH, AIRBOTICS, DroneDeploy

PEOPLE / ENTITIES
acxiom, Experian, EPSILON, InsideView, Crimson Hexagon, SAFEGRAPH, BASIS

LOCATION INTELLIGENCE
foursquare, Sense360, PlaceIQ, esri, factual, CARTO, Mapillary, STREETLINE

OTHER
qualtrics, DATA.GOV, data.world, panjiva

DATA RESOURCES

INCUBATORS & SCHOOLS
PLURALSIGHT, galvanize, DataCamp, DataElite, INSIGHT, The Data Incubator, METIS

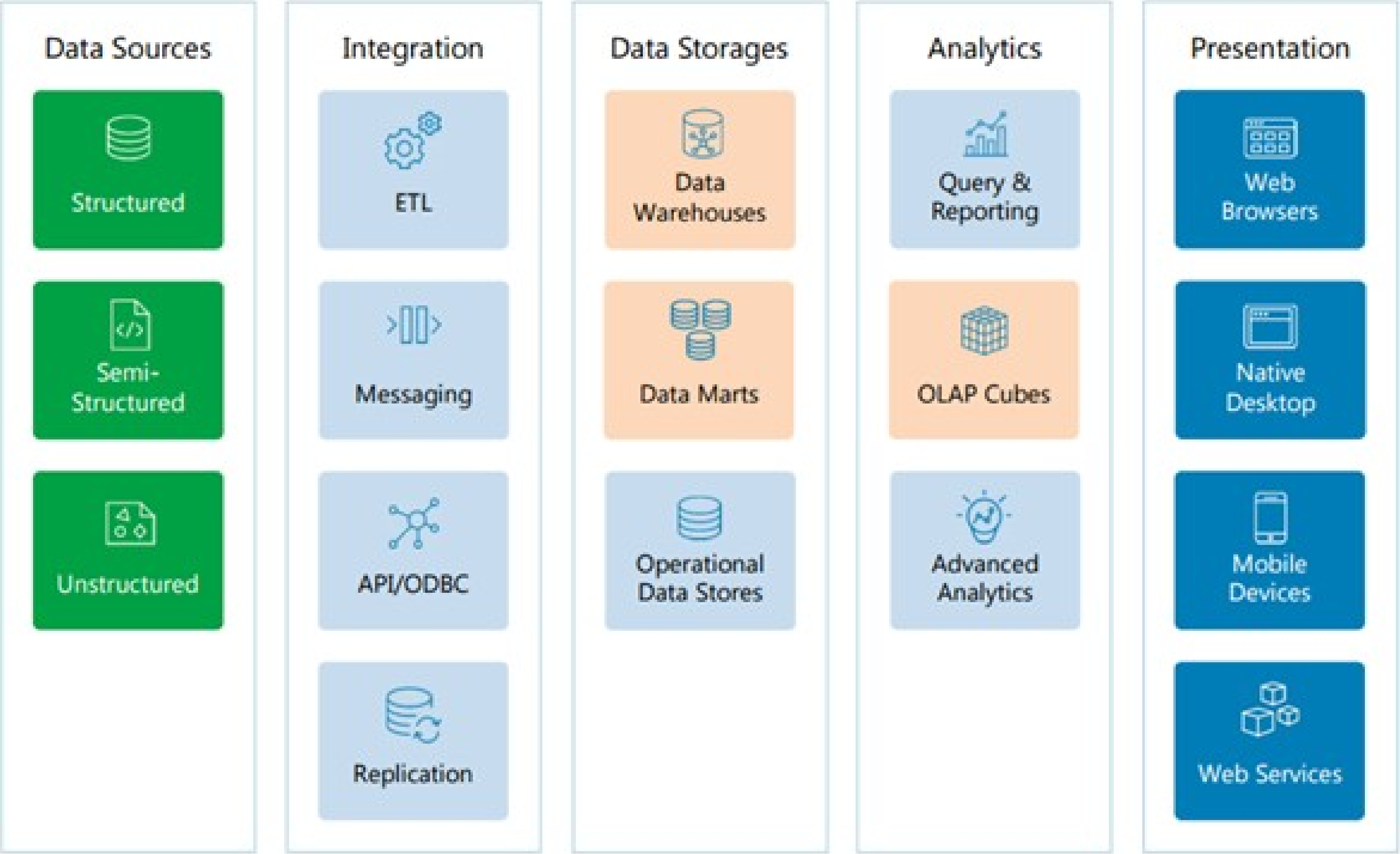
RESEARCH
facebook research, OpenAI, MIRI, CSAIL, DFKI, ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE, AI2

Last updated 4/5/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)   mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Extended Relational Reference Architecture



Data Sources

- Structured
- Semi-Structured
- Unstructured

Integration

- ETL
- Messaging
- API/ODBC
- Replication

Data Storages

- Data Warehouses
- Data Marts
- Operational Data Stores

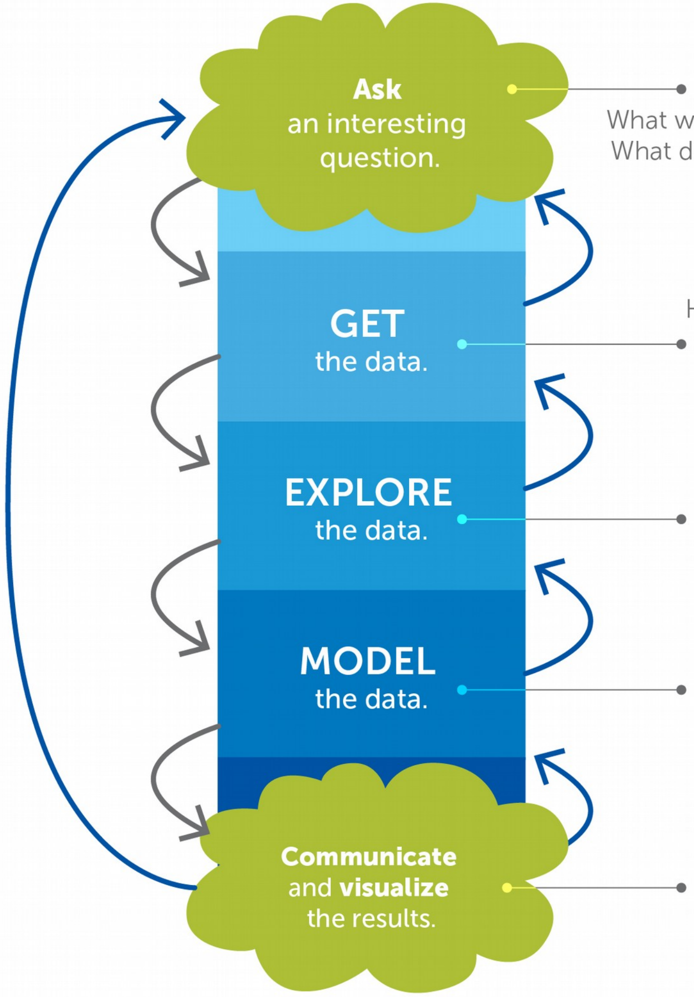
Analytics

- Query & Reporting
- OLAP Cubes
- Advanced Analytics

Presentation

- Web Browsers
- Native Desktop
- Mobile Devices
- Web Services

The
Data Science Process



Ask an interesting question.

What is the scientific **goal?**
What would you do if you had all the **data?**
What do you want to **predict** or **estimate?**

GET the data.

How were the data **sampled?**
Which data are **relevant?**
Are there **privacy** issues?

EXPLORE the data.

Plot the data.
Are there **anomalies?**
Are there **patterns?**

MODEL the data.

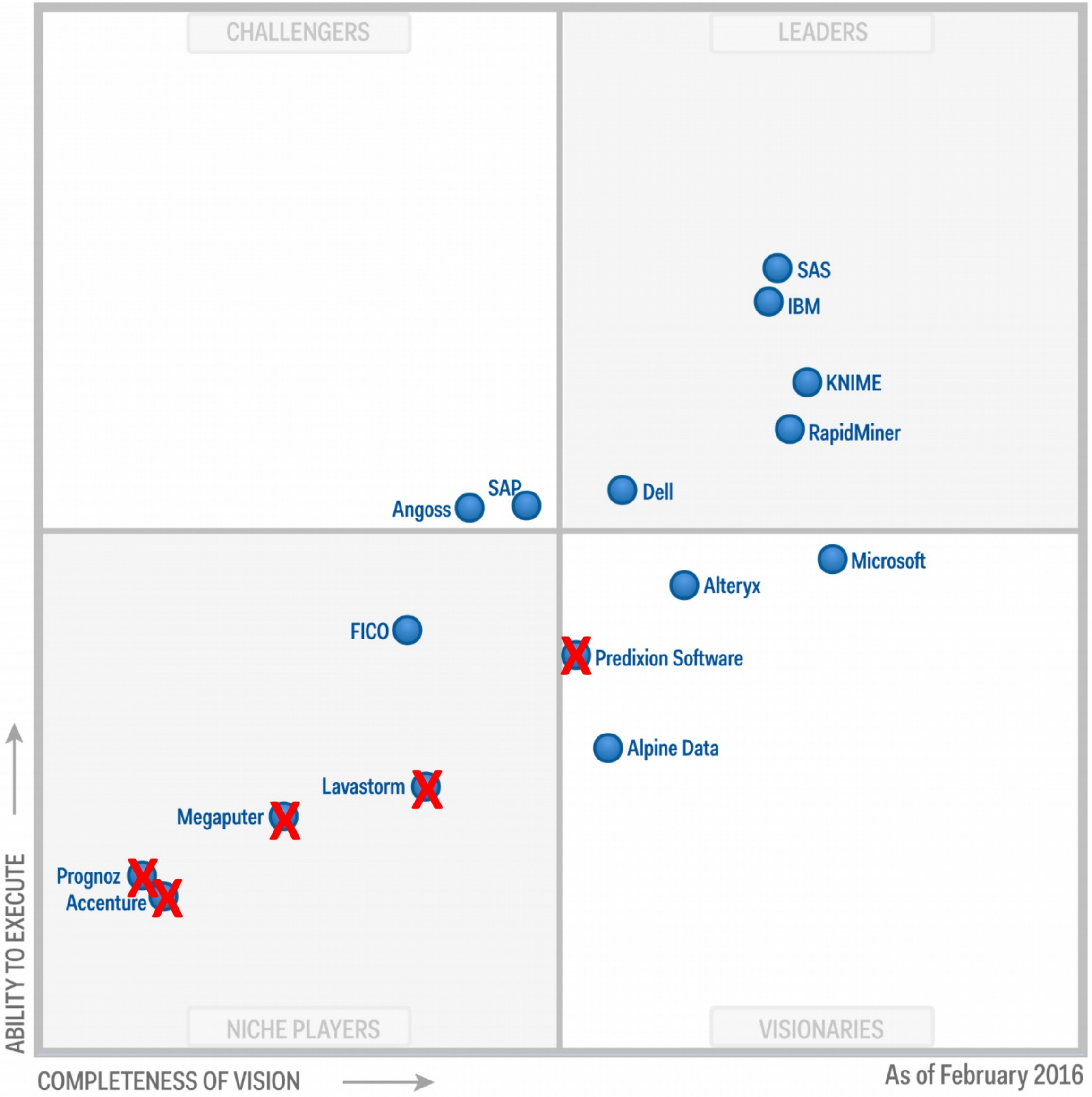
Build a model.
Fit the model.
Validate the model.

Communicate and **visualize** the results.

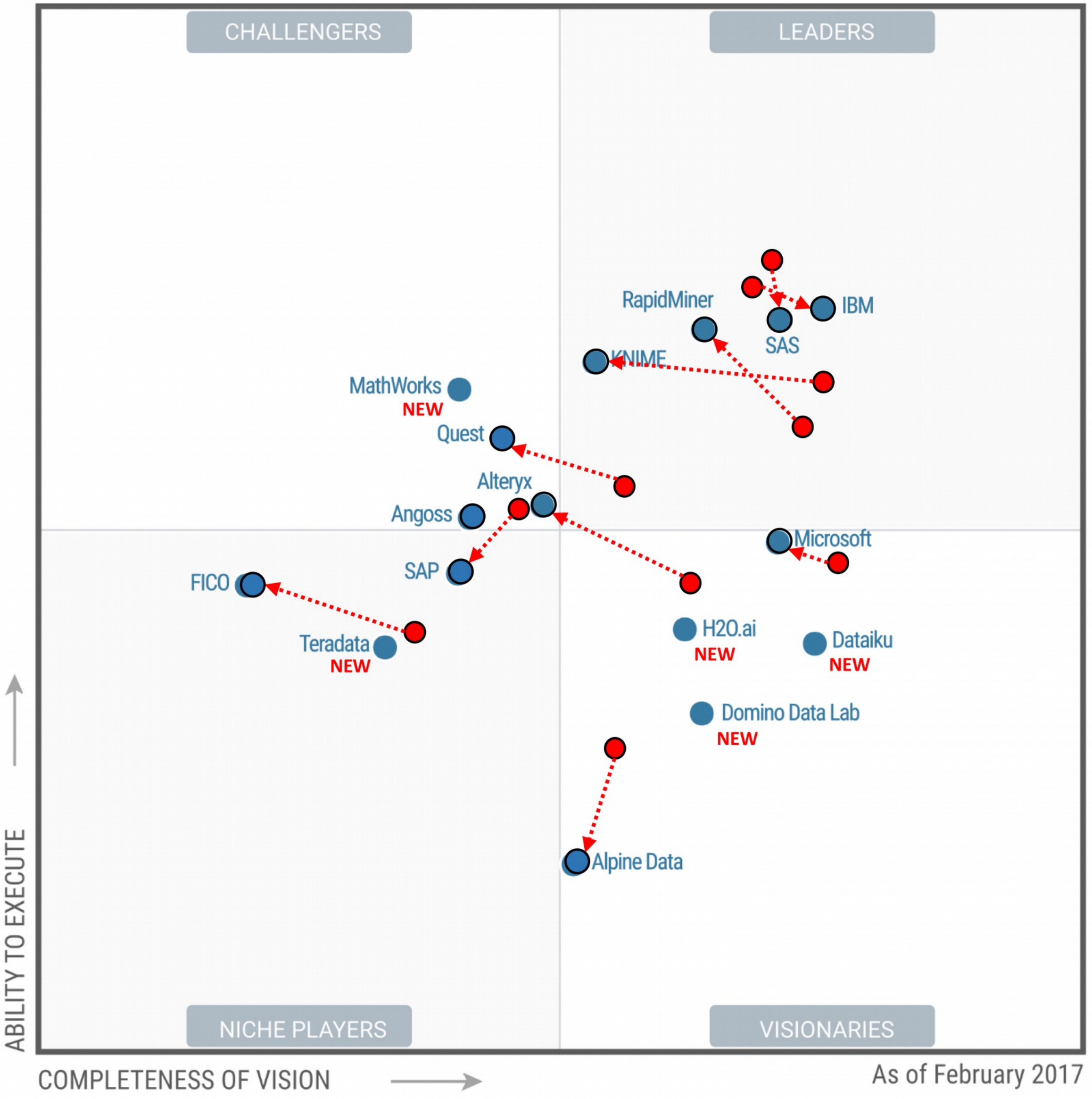
What did we **learn?**
Do the results make **sense?**
Can we tell a **story?**

Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course http://cs109.org/.

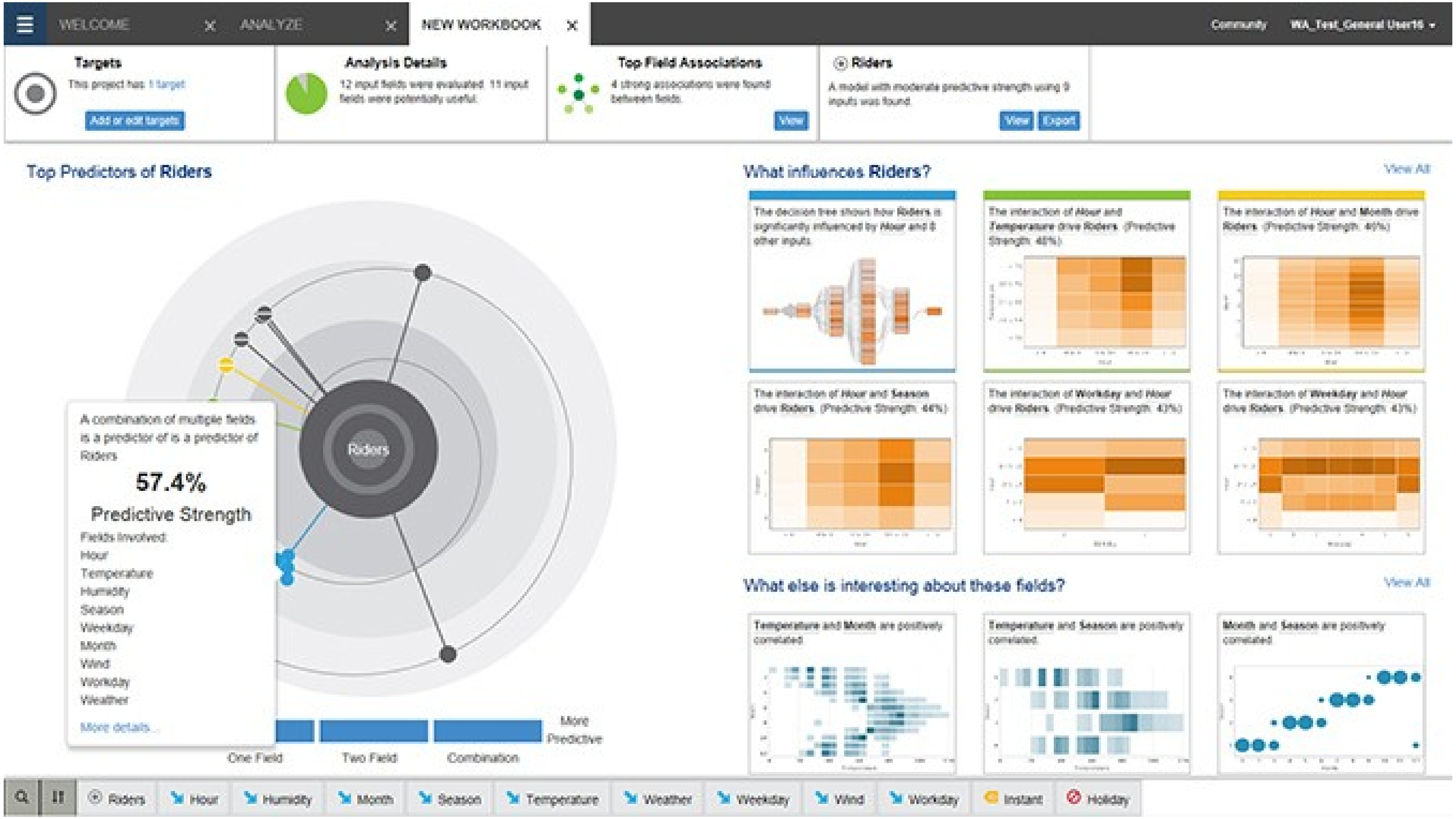
2016 Magic Quadrant for Advanced Analytics Platforms



2017 Magic Quadrant for Data Science Platforms



IBM Watson



Targets

This project has 1 target

Analysis Details

12 input fields were evaluated. 11 input fields were potentially useful.

Top Field Associations

4 strong associations were found between fields.

Riders

A model with moderate predictive strength using 9 inputs was found.

A combination of multiple fields is a predictor of is a predictor of Riders

57.4%

Predictive Strength

Fields Involved:
Hour
Temperature
Humidity
Season
Weekday
Month
Wind
Workday
Weather

More details...

What influences Riders?

The decision tree shows how Riders is significantly influenced by Hour and 8 other inputs.

The interaction of Hour and Temperature drive Riders. (Predictive Strength: 46%)

The interaction of Hour and Month drive Riders. (Predictive Strength: 45%)

The interaction of Hour and Season drive Riders. (Predictive Strength: 44%)

The interaction of Workday and Hour drive Riders. (Predictive Strength: 43%)

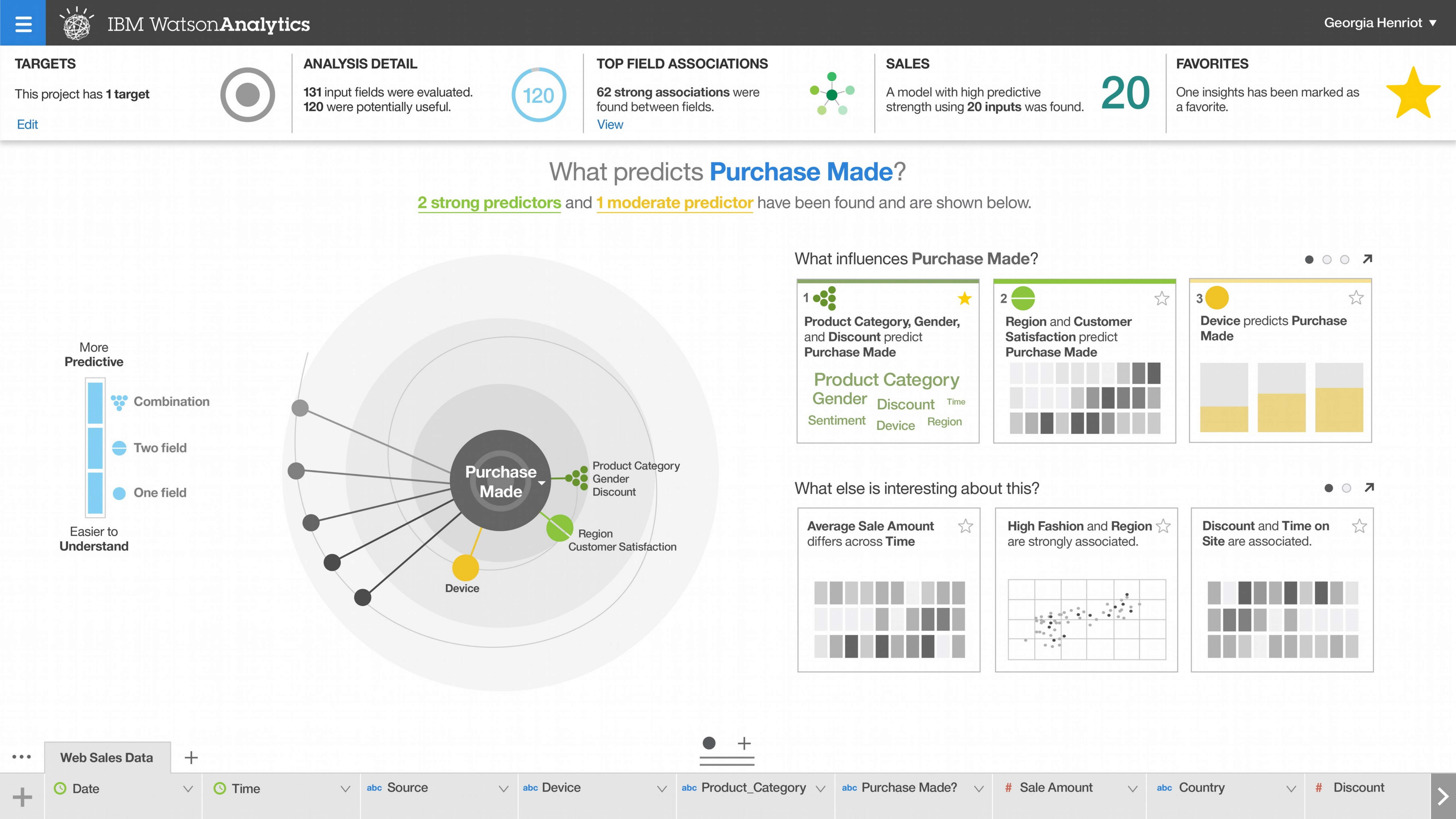
The interaction of Weekday and Hour drive Riders. (Predictive Strength: 43%)

What else is interesting about these fields?

Temperature and Month are positively correlated.

Temperature and Season are positively correlated.

Month and Season are positively correlated.



IBM WatsonAnalytics

Georgia Henriot ▾

TARGETS

This project has **1 target**

Edit

ANALYSIS DETAIL

131 input fields were evaluated.
120 were potentially useful.

TOP FIELD ASSOCIATIONS

62 strong associations were found between fields.

View

SALES

A model with high predictive strength using **20 inputs** was found.

FAVORITES

One insights has been marked as a favorite.

What predicts **Purchase Made**?

2 strong predictors and **1 moderate predictor** have been found and are shown below.

What influences **Purchase Made**?

1 **Product Category, Gender,** and **Discount** predict **Purchase Made**

2 **Region** and **Customer Satisfaction** predict **Purchase Made**

3 **Device** predicts **Purchase Made**

What else is interesting about this?

Average Sale Amount differs across **Time**

High Fashion and **Region** are strongly associated.

Discount and **Time on Site** are associated.

IBM resources

- https://www.ibm.com/analytics/us/en/industry/government/

- **IBM Certified Data Architect - Big Data**

http://www-03.ibm.com/certify/certs/50001701.shtml

- https://datascience.ibm.com/

Felicitaciones por tu interés en los datos y cómo usarlos □

Ángela Devia - ing.angela.devia@gmail.com

Gracias