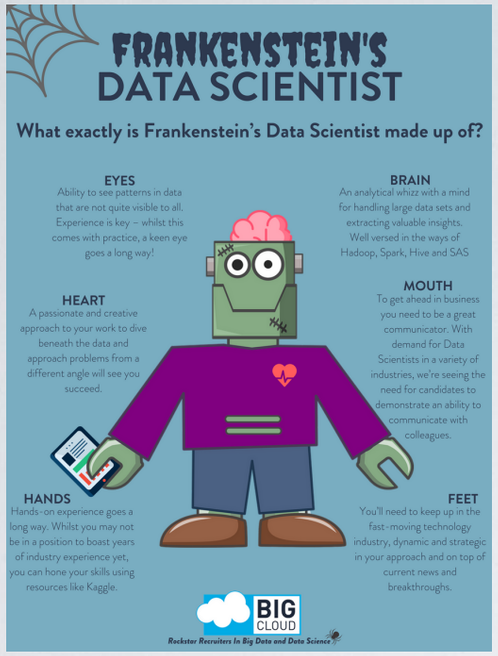
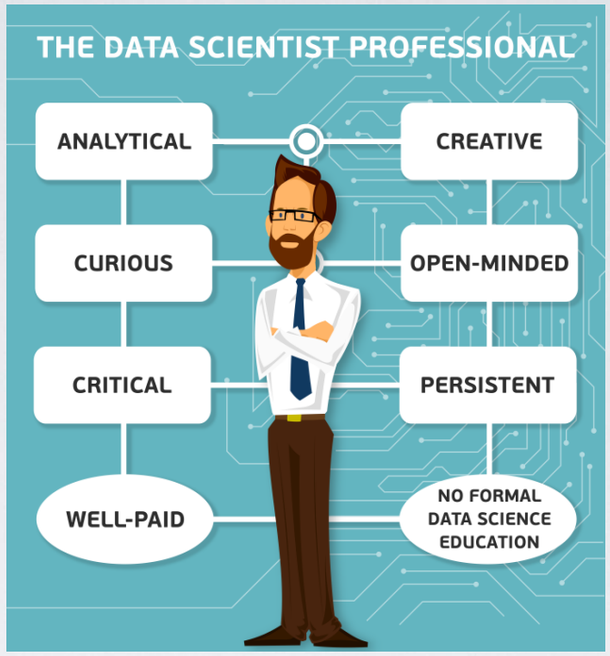
EFFECTIVE DATA SCIENCE

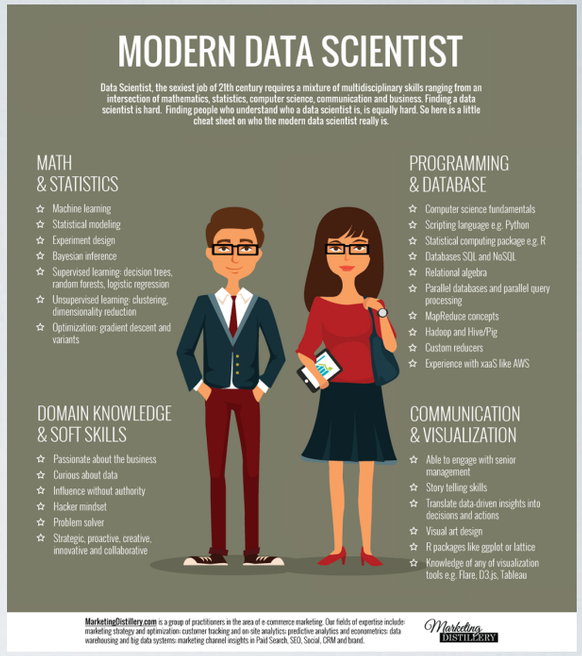
Jan 16, 2019 - Columbus Data Science

EXPEED SOFTWARE

WHAT IS(N'T) DATA SCIENCE







Who are Data Scientists?

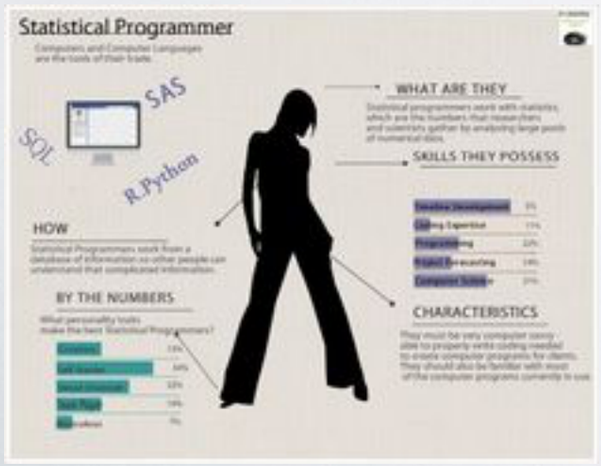
(As defined by Infographics on the first page of a google search)

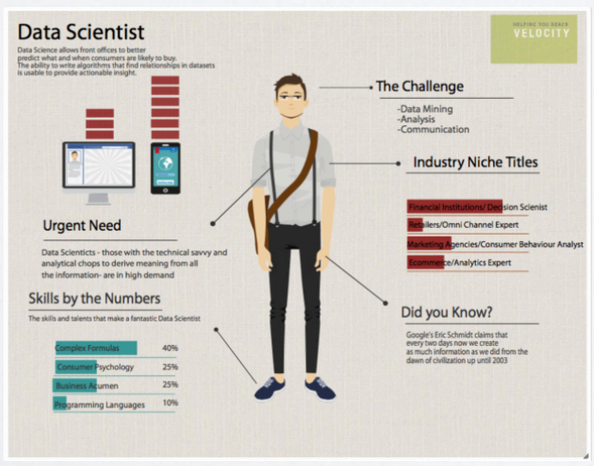
6/9 - Male

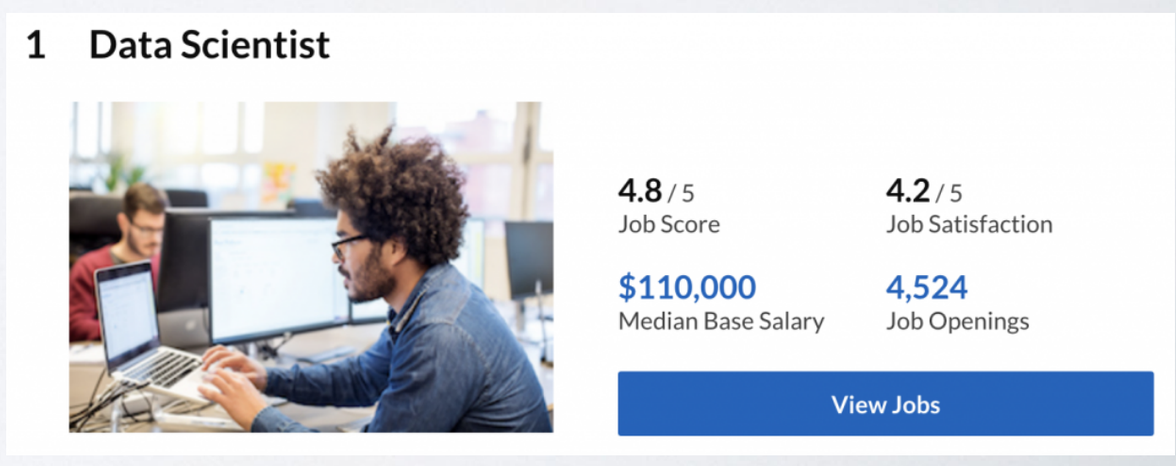
4/9 - Bearded

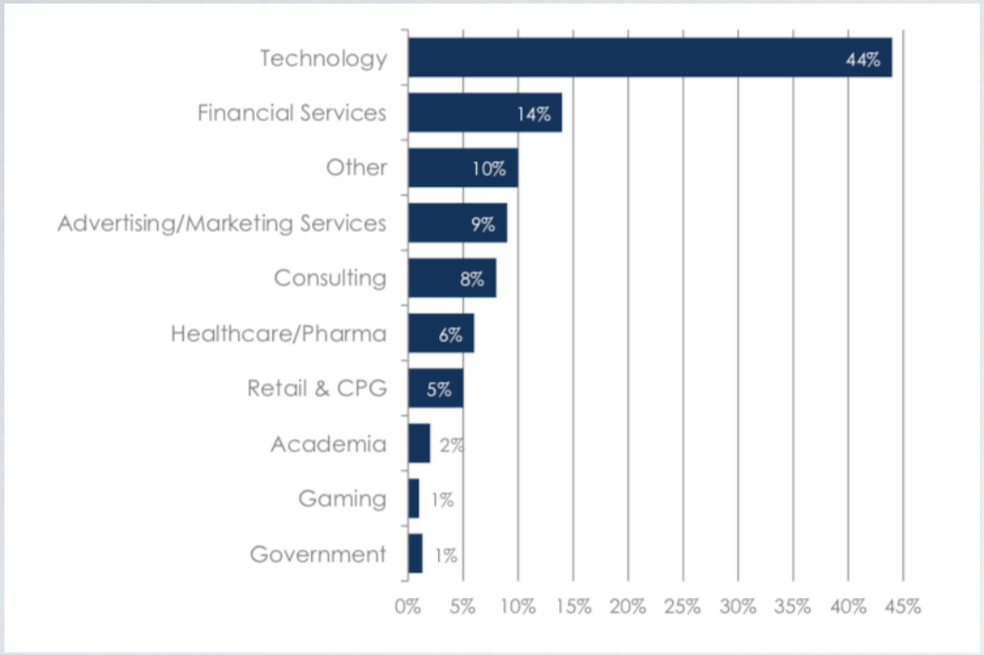
8/9 - Human

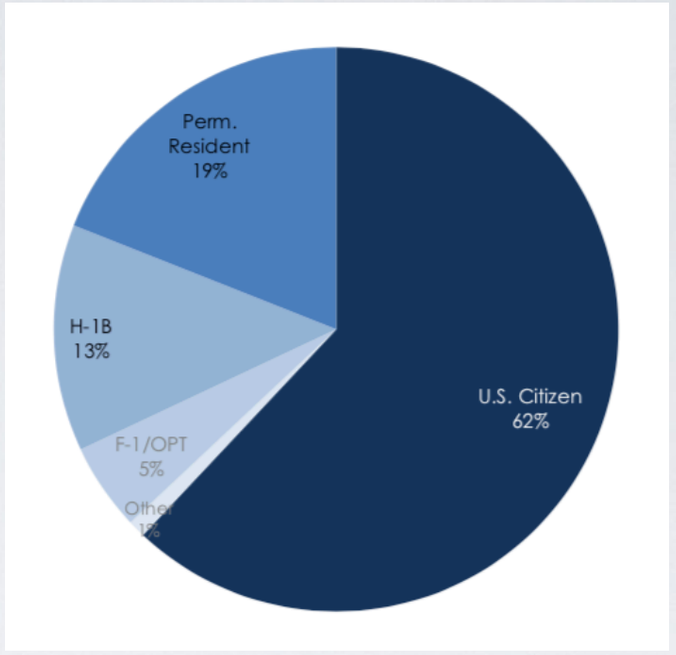
2/9 - Wear ties

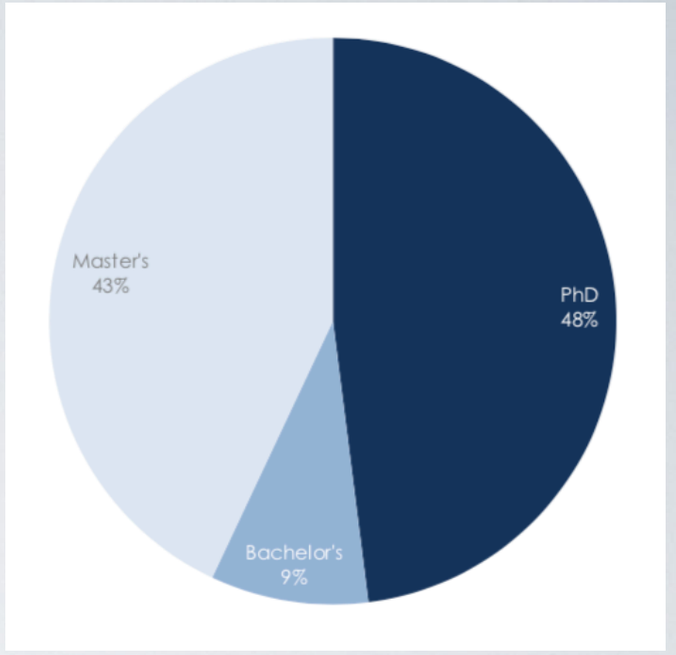












Who are Data Scientists?

(As defined by a Burtch Works Survey)

91% - Have an advanced degree

40% on West Coast

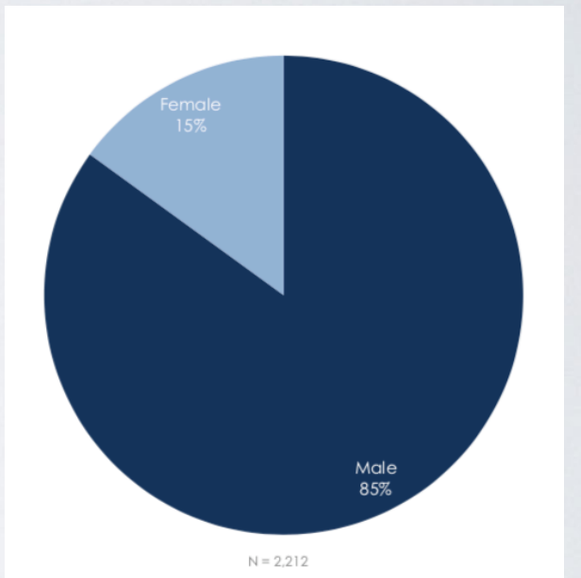
26% in Northeast

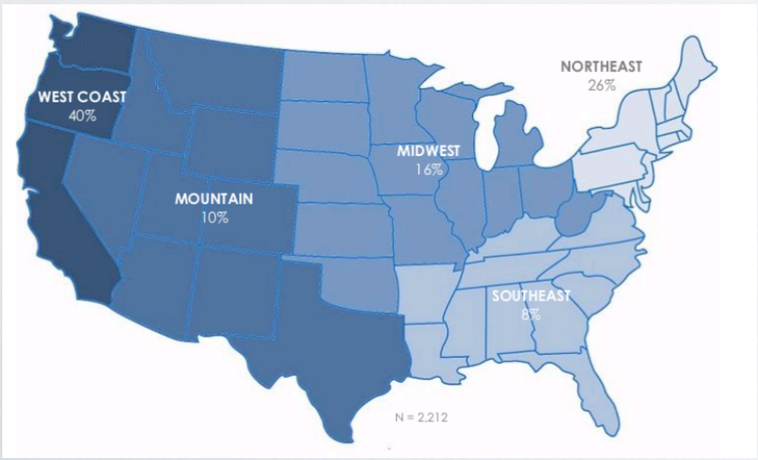
44% in Tech industry

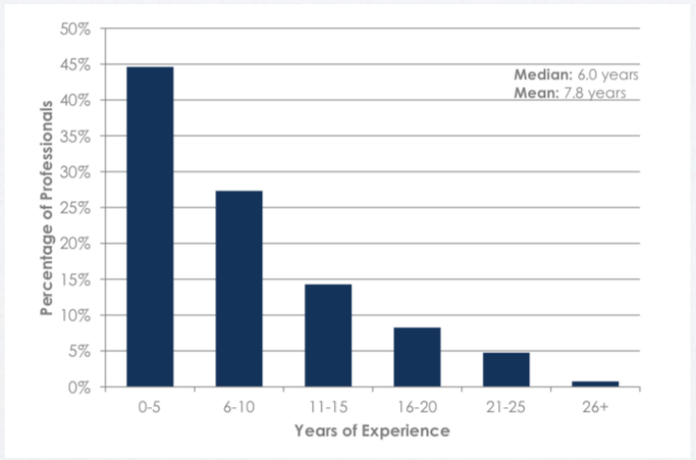
85% are male

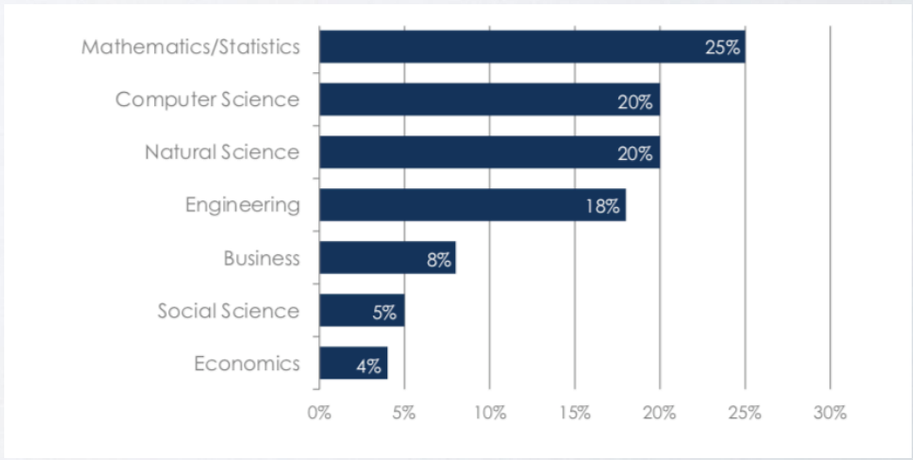
Median years of experience - 6

Note: 4/9 faces in Butch Works slide deck has a beard









WHAT IS DATA SCIENCE?

CCD DATA: THE *GOOD*, THE *BAD*, AND THE *UGLY*

PHILIP MASSEY and GEORGE H. JACOBY
Kitt Peak National Observatory, NOAO, P.O. Box 26732, Tucson, AZ 85726-6732

"Some days the magic works. Some days it doesn't."
Little Big Man's grandfather.

INTRODUCTION

We will describe three kinds of data: *good data*, *bad data*, and *ugly data* (also known as "truth"). The first of these, *good data*, actually doesn't exist, but rather is an ideal to strive for in terms of what you would like your CCD data to be like. *Bad data* consists of data that you are better off throwing out—something went wrong; you either didn't notice or couldn't do anything about it, but just throw it out. *Ugly data* is what you get from *real* CCDs, and is just fine—just a little bit, well, ugly. It's our plan to help you distinguish between the last two cases.

WHAT IS DATA SCIENCE?

Think

Pair

Share

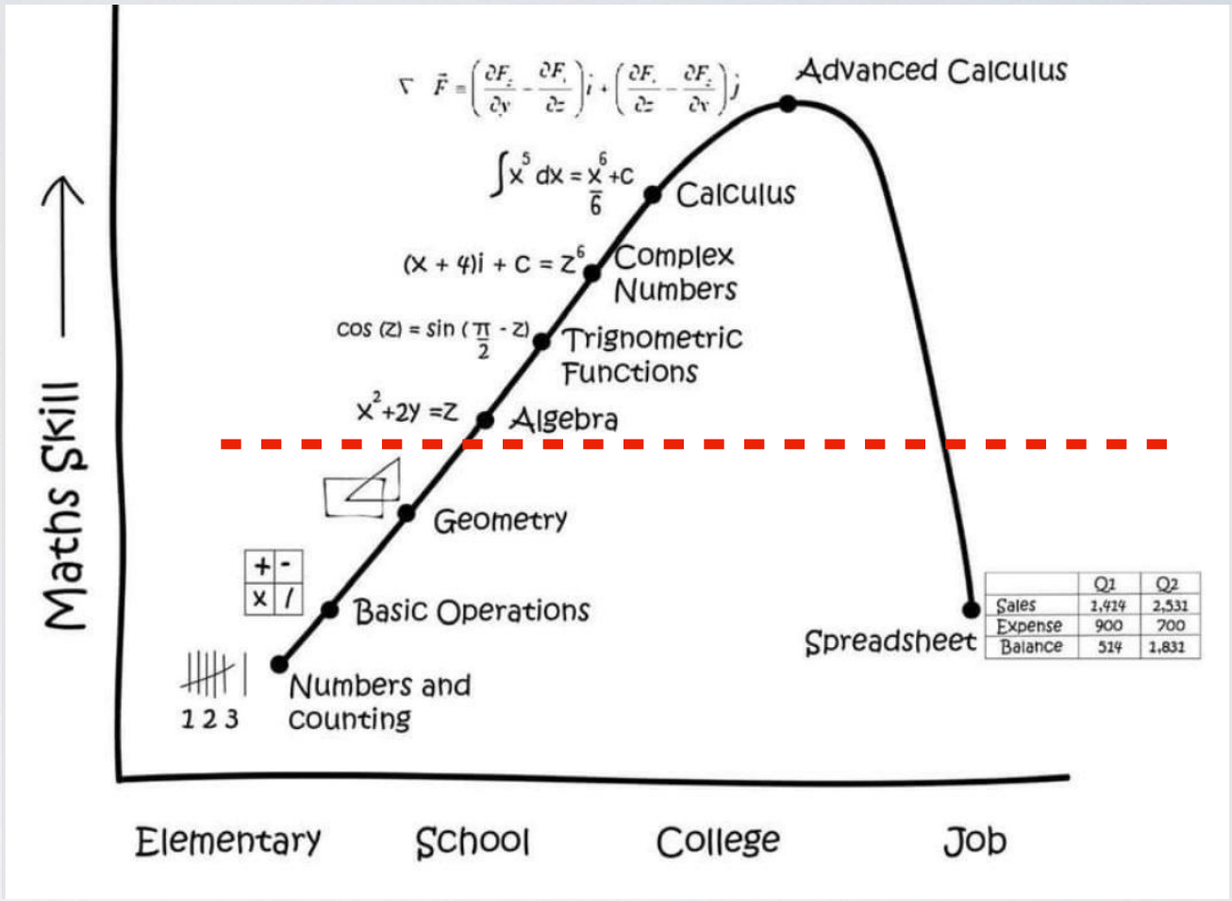
Source: IBM® BigInsights™

WHAT IS DATA SCIENCE?

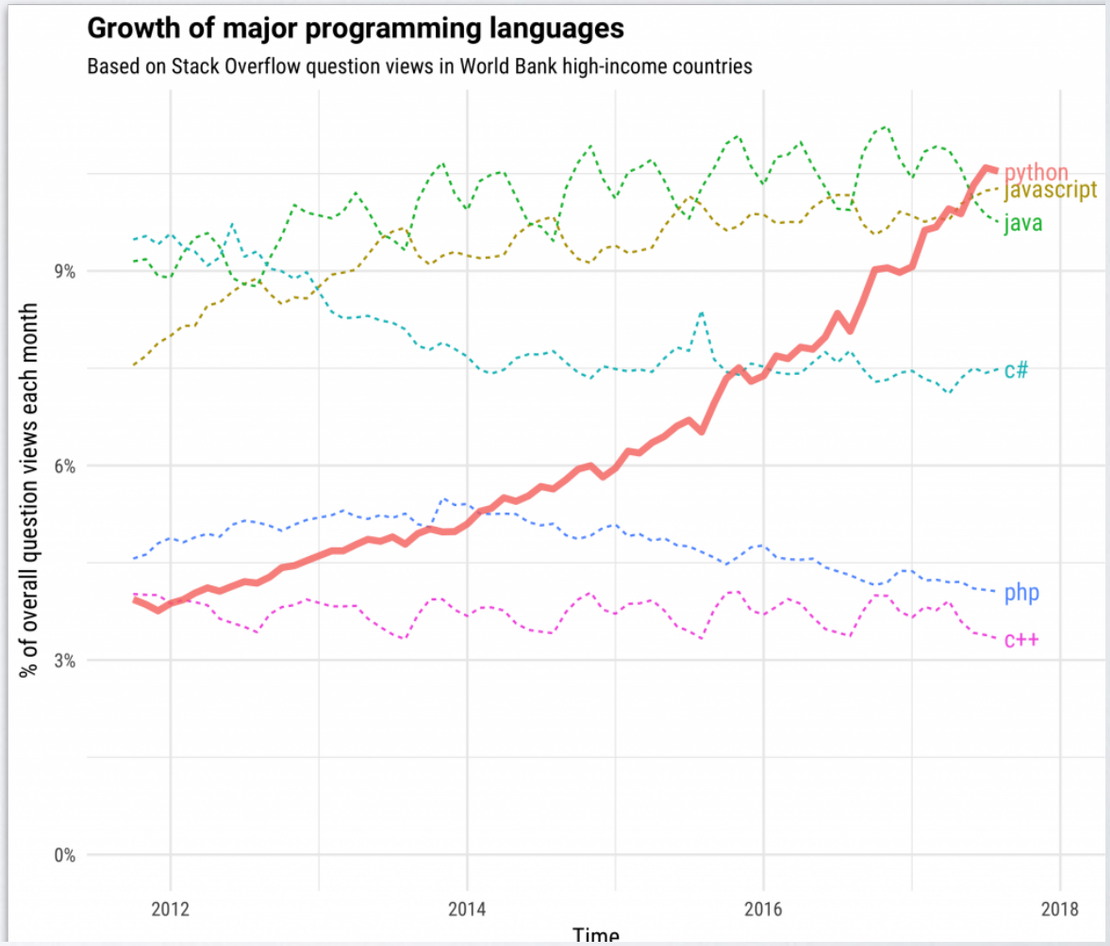
- Critical Thinking
- Problem Solving
- Computer Programing
- Analytics
- Machine Learning
- Stuff and stuff

Source: IBM® BigInsights™


What is
not data
science?


Maths Skill
Elementary
School
College
Job
Numbers and counting
Basic Operations
Geometry
Algebra
Trignometric Functions
Complex Numbers
Calculus
Advanced Calculus
Spreadsheet
1 2 3
+ -
x /
$\nabla \vec{F} = \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right)i + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right)j$
$\int x^5 dx = \frac{x^6}{6} + c$
$(x + 4)i + c = z^6$
$\cos(z) = \sin\left(\frac{\pi}{2} - z\right)$
$x^2 + 2y = z$
Q1 Q2
Sales 1,414 2,531
Expense 900 700
Balance 514 1,831


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS
xkcd


Growth of major programming languages
Based on Stack Overflow question views in World Bank high-income countries
% of overall question views each month
9%
6%
3%
0%
2012
2014
2016
2018
Time
python
javascript
java
c#
php
c++

HOW DOES DS BENEFIT FROM COMMUNITY

BIG DATA & AI LANDSCAPE 2018

INFRASTRUCTURE

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

HADOOP IN THE CLOUD
aws, Microsoft Azure, Google Cloud, IBM InfoSphere BigInsights, TREASURE DATA, Qubole, altiscale, CAZENA, CenturyLink

STREAMING / IN-MEMORY
aws, databricks, striim, confluent, GridGain, DATATORRENT, dataArtisans, ORACLE Coherence, hazelcast, TERRACOTTA

NoSQL DATABASES
Google Cloud, aws, ORACLE, Microsoft Azure, mongoDB, MarkLogic, AEROSPIKE, DATASTAX, ArangoDB, Couchbase, redislabs, SCYLLA.

NewSQL DATABASES
SAP HANA, Clustrix, Pivotal, NUODB, Cockroach LABS, Cloud Spanner, MEMSQL, ORACLE, influxdata, MariaDB, TIMESCALE, VOLTDB, citusdata, splice MACHINE, paradigm4, Apache Trafodion

GRAPH DBs
neo4j, Amazon Neptune, IBM, ORACLE, OrientDB, GIRAPH, InfiniteGraph, Objectivity

MPP DBs
TERADATA, VERTICA, IBM Data Warehouse Systems, actian, Kognitio, Exasol, dremio

CLOUD EDW
aws, Google Cloud, Microsoft Azure, Pivotal, snowflake, Infoworks

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

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

DATA GOVERNANCE
Informatica, IBM, SailPoint, McAfee Skyhigh Security Cloud, collibra, Alation, Waterline Data, OKERA

MGMT / MONITORING
aws, New Relic, actifio, rubrik, APPDYNAMICS, DATADOG, dynatrace, WAVEFRONT by vmware, splunk>, SignalFx, druva, Moogsoft, unravel, pagerduty, Numerify, Anodot

STORAGE
aws, Google Cloud, Microsoft Azure, PURESTORAGE, ALLUXIO, nimblestorage, Qumulo, panasas, COHESITY

CLUSTER SVCS
aws, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata

APP DEV
Lightbend, Keen IO, rainforest, CASK

CROWD-SOURCING
amazon mechanical turk, Upwork, Figure Eight, scale, HIVE

HARDWARE
Google TPU, arm, intel AI, GRAPHCORE, MYTHIC, NVIDIA, cerebras, Movidius, WAVE COMPUTING, HAILO

GPU DBs
kinetica, MAPD, SQREAM TECHNOLOGIES, blazegraph, BLAZINGDB, brytlyt, PG-Strom

ANALYTICS

DATA ANALYST PLATFORMS
Microsoft, pentaho, alteryx, Digital Reasoning, GUAVUS, AYASDI, ATTIVIO, Datameer, Quid, incorta, interana, ClearStory, Origami logic, ASCEND.IO, ENDOR, MODE, Bottlenose

DATA SCIENCE PLATFORMS
IBM, KNIME, data iku, DOMINO, rapidminer, CONTINUUM ANALYTICS, ALGORITHMIA, DATAWATCH ANGOSS

BI PLATFORMS
Microsoft, aws, DOMO, Wave Analytics, looker, ATSCALE, ARCADIA DATA, SISENSE, GoodData, birst

VISUALIZATION
tableau, SAP Lumira, Google Cloud, CELONIS, Qlik, Periscope Data, ZEPL, GOOMDATA, plotly, CHARTIO

MACHINE LEARNING
Azure Machine Learning, aws, Google Cloud, H2O.ai, DataRobot, gamalon, ELEMENT AI, VERSIVE, ViSENZE, deepsense.io, bonsai

COMPUTER VISION
Microsoft Azure, Amazon Rekognition, Cloud Vision API, clarifai, 商汤 sensetime, EVER AI, deepomatic, twenty bn

HORIZONTAL AI
IBM Watson, Cortana, Face++ 旷视, sentient, Voyager Labs, vicarious, Affectiva, PROPHESEE, CognitiveScale, Numenta, PETUUM, SCALED INFERENCE, narrologics, CURIOUS AI, OSARO, BLUE VISION

SPEECH & NLP
Google Cloud, twilio, amazon alexa, narrative science, Wolfram Alpha, semantic machines, Mobvoi, NUANCE, SoundHound Inc., voicera, MindMeld, cortical.io, snips, maluuba, Gridspace, yseop

SEARCH
elasticsearch, ORACLE ENDECA, THOUGHTSPOT, EXALEAD, Lucidworks, coveo, swiftype, ATTIVIO, algolia, alphasense, MAANA, omni:us, SINEQUA

LOG ANALYTICS
splunk>, sumologic, solarwinds LOGGLY, TIMBER, kibana, logz.io

SOCIAL ANALYTICS
Hootsuite, sprinklr, NETBASE, synthesio, tracx, simplereach, bitly, predata, SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics, mixpanel, AMPLITUDE, sumall, Airtable, RESCI, SIGOPT, granify, custora

APPLICATIONS – ENTERPRISE

SALES
einstein, CHORUS, INSIDESALES.COM, conversica, GONG, clari, aviso, tact.ai, fuse|machines, TROOPS

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

MARKETING - B2C
zeta, bloomreach, SendGrid, BlueYonder, [PERSADO], kahuna, ACTIONIQ, SAILTHRU, BLUECORE, QUANTIFIND, mparticle, Amplero, amperity

CUSTOMER SERVICE
MEDALLIA, zendesk, CLARABRIDGE, Gainsight, NG DATA, DigitalGenius, afiniti, AUTOMAT, frame.ai, msg.ai, INTERCOM

HUMAN CAPITAL
HireVue, entelo, hiQ, GIGSTER, textio, RESTLESS BANDIT, Wade&Wendy, Stella, Clustree, pymetrics, mya, uncommon

LEGAL
RAVEL, Seal, Everlaw, JUDICATA, EBREVIA, PREMONITION, ROSS, casetext

FINANCE
Anaplan, zuora, SAP S/4HANA, TRADESHIFT

ENTERPRISE PRODUCTIVITY
slack, ORACLE, lumiata, DIFFBOT, clara, talla, butter.ai, Kasisto

BACK OFFICE AUTOMATION
UiPath, HyperScience, captricity, AppZen, WorkFusion

SECURITY
TANIUM, CYLANCE, zscaler, StackPath, illumio, CODE42, CipherCloud, DARKTRACE, ANOMALI, ThreatMetrix, VECTRA, cybereason, Guardian Analytics, DATAVISOR, siftscience, SICNIFYD, SentinelOne, SecurityScorecard, SOCURE, AREA 1 SECURITY, BlueTalon, Recorded Future, feedzai, CyberX, sparkcognition, IronNet Cybersecurity

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus, MediaMath, criteo, xAd, Integral Ad Science, ORACLE MOAT, OpenX, datorama, theTradeDesk, Adgorithms, dstillery, LiveIntent, TAPAD, dataxu, gumgum, Appier, DYNAMIC YIELD, teemo, yieldmo

EDUCATION
Liulishuo, KNEWTON, Clever, eclara, kidaptive, PANORAMA, knowre, gradescope

GOVERNMENT
OPENGOV, mark43, FiscalNote, OpenDataSoft

FINANCE - LENDING
ondeck, Affirm, 拍拍贷 ppdai.com, JIANPU.AI, Kreditech, AVANT, TALA, zestfinance, Upstart, INSIKT, upgrade, 100Credit, WeLab, Wecash闪现普惠, TrueAccord, MoneyLion, Active.Ai, aire, cignifi

FINANCE - INVESTING
Dataminr, Quantopian, ADDEPAR, NUMERAI, iSENTIUM, ALGORIZ, RavenPack, PAGAYA

REAL ESTATE
REDFIN, Opendoor, VTS, CREDIFI, reonomy, COMPSTAK, CAPE ANALYTICS

INSURANCE
Metromile, Lemonade, CYENCE, Shift Technology, TRACTABLE

HEALTHCARE
flatiron, Clover, KYRUUS, HealthTap, METABIOTA, Ginger.io, Glow, babylon, 3DMed, zebra, PathAI, ovia, TEMPUS, patientslikeme, AiCure, RECURSION, prognos, enlitic, imagia, Qventus, BAYLABS, ARTERYS, CLOUD MEDX, IMAGEN, Kang Health, PAIGE, DATAVANT, HUMAN DIAGNOSIS PROJECT, innoviccer, LeanTaaS

LIFE SCIENCES
23andMe, color, iCarbonX, BenevolentAI, verily, WuXiNextCODE, ZEPHYR HEALTH, Zymergen, INTUITIVE GENOMICS, SCAN, Clear Labs, freenome, NANOPORE, DNAnexus, Phosphorus, CITRINE, twoXAR, Atomwise, deep genomics, SOPHIA GENETICS, OWKIN

TRANSPORTATION
UBER, TESLA, ZOOX, CLEARPATH, nuTonomy, drive.ai, nauto, AIMOTIVE, PILOT.AI, NIO, OPTIMUS RIDE, moovit, nexar, comma.ai, netradyne, Civil Maps, German Autolabs

AGRICULTURE
FARMERS BUSINESS NETWORK, Granular, BLUE RIVER TECHNOLOGY, FarmersEdge, FarmLogs, mavrx, TerrAvion, prospera

COMMERCE
instacart, STITCH FIX, Dia & Co, RetailNext, HowGood

INDUSTRIAL
C3 IoT, PREDIX, UPTAKE, osisoft, NIGHT MACHINE, TACHYUS, SCORTEX, Alluvium

OTHER
eharmony, stem, rethink robotics, Amper, ByteDance, hopper, celect, BOXEVER, VERDIGRIS, duetto, Unbabel, Jukedeck, Second Spectrum, remesh, ASAPP

CROSS-INFRASTRUCTURE/ANALYTICS
aws, Google Cloud, Microsoft, IBM, SAP, Hewlett Packard Enterprise, SAS, 1010DATA, vmware, TIBCO, TERADATA, ORACLE, NetApp, syncsort

OPEN SOURCE

FRAMEWORK
hadoop HDFS, Apache Kylin, hadoop MapReduce, Flink, YARN, TEZ, MESOS, Spark, CDAP

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

DATA ACCESS
cassandra, nifi, mongoDB, Apache Parquet, CouchDB, OPENTSDB, riak, APACHE HBASE, Cloud Spanner, ACCUMULO

COORDINATION
talend, Apache Zookeeper, DOCKER, Apache Ambari

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

STAT TOOLS
python, ScalaLab, NumPy, SciPy

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, theano, torch, MADlib, scikit-learn, Caffe, Microsoft Cognitive Toolkit, OpenAI, DMTK, Keras, PaddlePaddle, Apache SINGA, FeatureFu, mxnet, Chainer, VELES, WEKA, neon™, DIMSUM, DSSTNE, mlib, DL4J, MAHOUT, Aerosolve

SEARCH
elasticsearch, Solr, Lucene

LOGGING & MONITORING
elasticsearch, kibana, SENTRY, logstash, Prometheus

VISUALIZATION
BeakerX, Rodeo

COLLABORATION
jupyter, Zeppelin, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry

DATA SOURCES & APIs

HEALTH
VALIDIC, practicefusion, fitbit, GARMIN, HUMAN API, kinsa

IOT
GE Digital, UPTAKE, thingworx, helium, samsara, AUGURY, estimote

FINANCIAL & ECONOMIC DATA
Bloomberg, THOMSON REUTERS, DOW JONES, S&P CAPITAL IQ, CB INSIGHTS, xignite, Quandl, ENVESTNET YODLEE, PREMISE, estimize, SECOND MEASURE, Eagle Alpha, StockTwits, PLAID, Thinknum

AIR / SPACE / SEA
Orbital Insight, planet., SKYCATCH, Airware, AIROBOTICS, spire, PRECISIONHAWK, UNDERSTORY, Descartes Labs, WINDWARD, tellusIabs, DroneDeploy, MarineTraffic

PEOPLE / ENTITIES
acxiom, experian, EPSILON, InsideView, Crimson Hexagon, BASIS TECHNOLOGY, Quantcast, SAFEGRAPH

LOCATION INTELLIGENCE
FOURSQUARE, MapAnything, sense360, PlaceIQ, esri, factual, CARTO, Mapillary, Streetline, cuebiq

OTHER
qualtrics, DATA.GOV, data.world, enigma, mobilewalla

DATA RESOURCES

DATA SERVICES
Palantir, Mu Sigma, OPERA, SILICON VALLEY DATA SCIENCE, fractal, kaggle, EXL, DataKind

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

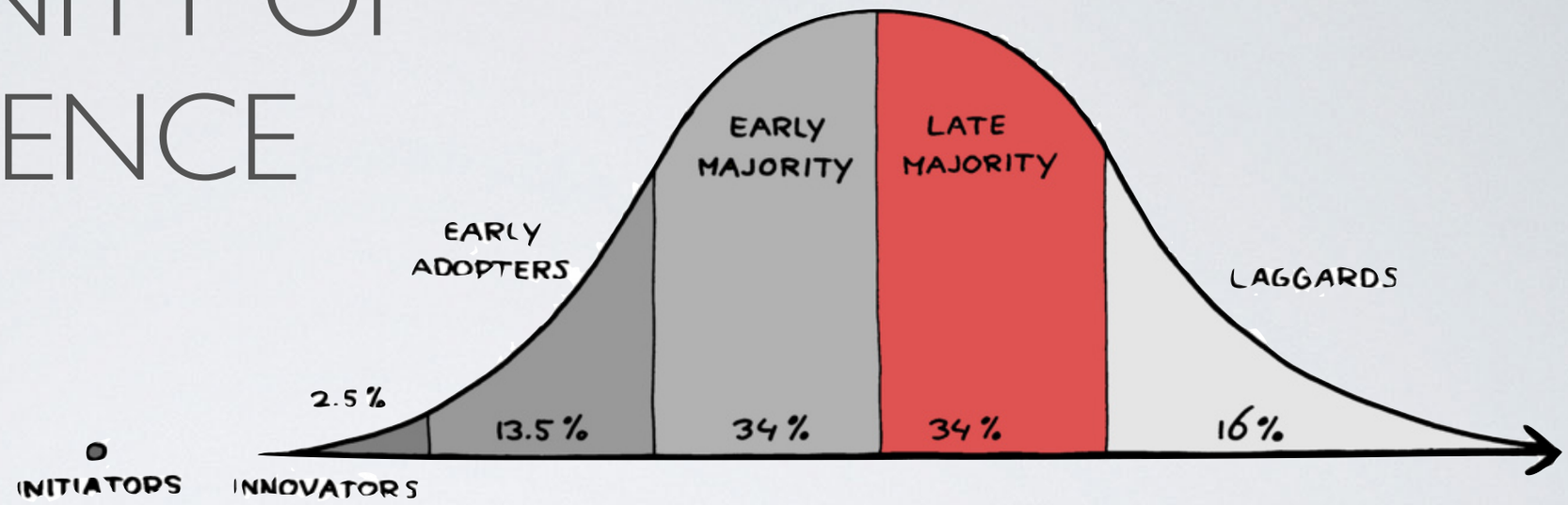
RESEARCH
OpenAI, facebook research, MILA, MIRI, VECTOR INSTITUTE, CSAIL, DFKI, AI2, ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

V1 – Last updated 6/19/2018

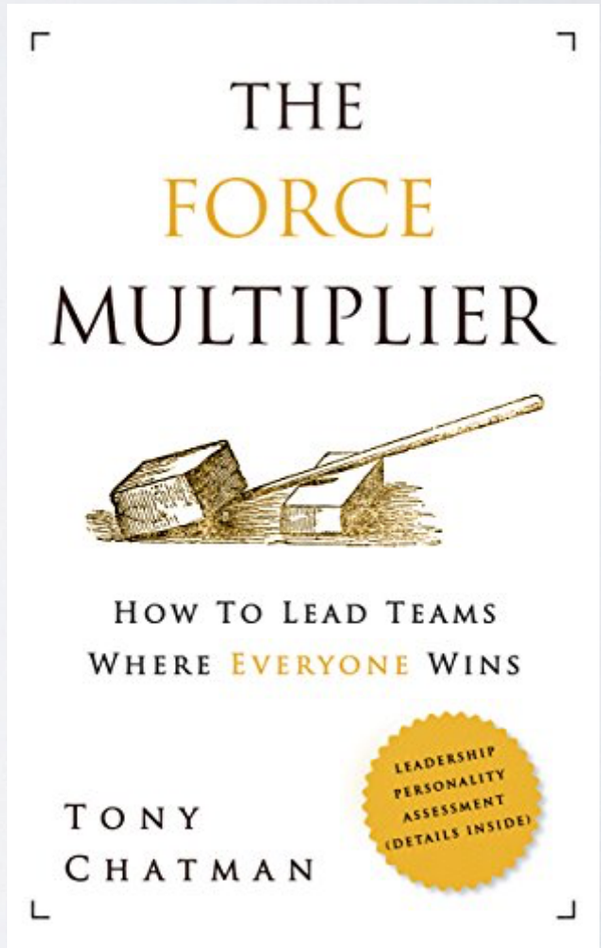
 mattturck.com/bigdata2018

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

A COMMUNITY OF DATA SCIENCE



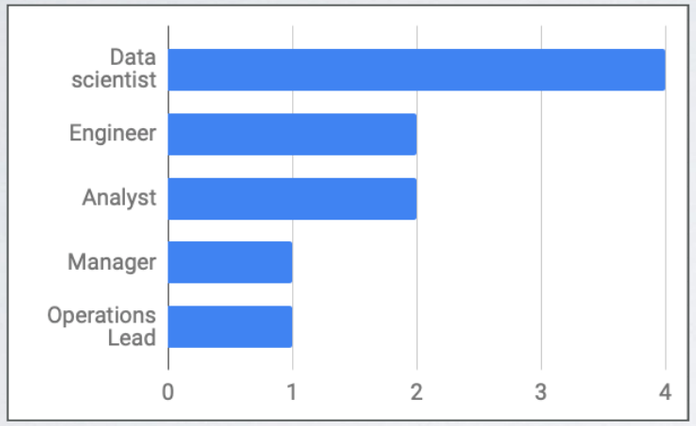
- Network
- Learn stuff
- Showcase your skills
- Get answers to questions
- Discuss new technologies
- Explain where you are stuck
- Brainstorm
- Force Multiplier

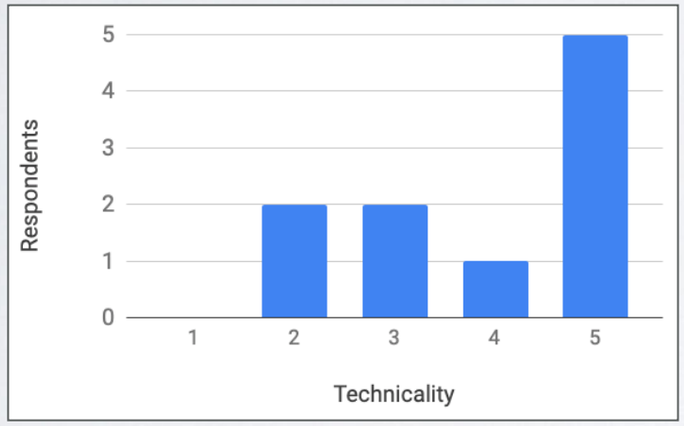


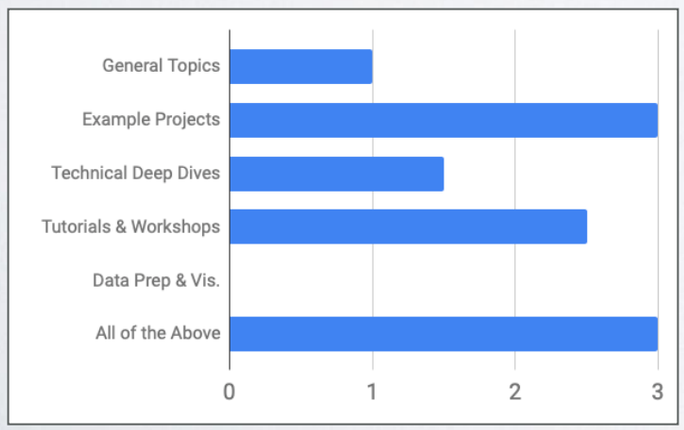
EXPEED SOFTWARE

WHAT DOES OUR COMMUNITY WANT/NEED

COLUMBUS DATA SCIENCE POLL







C-Bus Data Science Audience Survey

What's your current role?

- ○ Data scientist
- ○ Engineer
- ○ Analyst
- ○ Manager
- ○ Other: _____

How technical are you?

	1	2	3	4	5	
I'm interested and want to learn more	○	○	○	○	○	I code on the reg

What sort of talks are interesting to you?

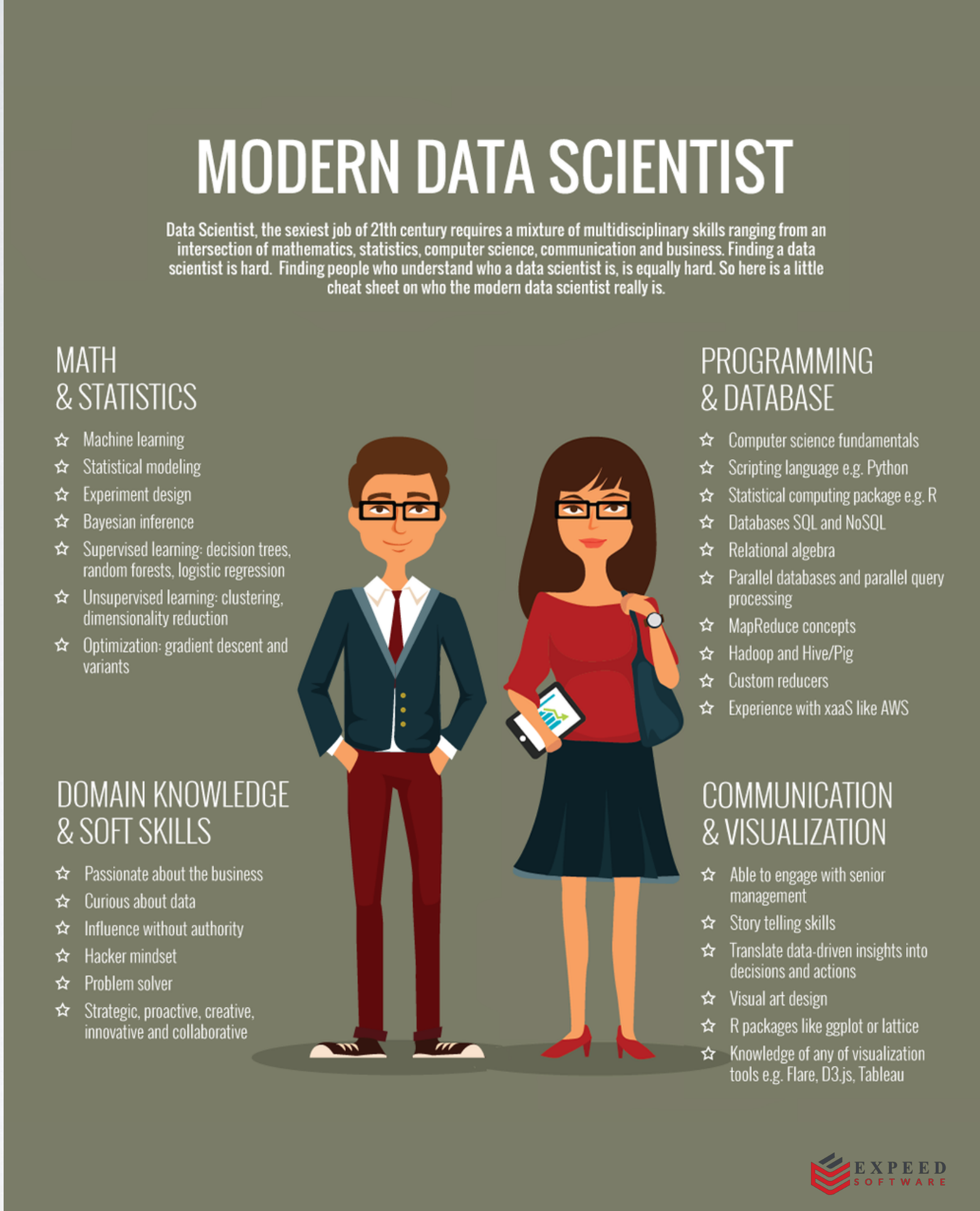
- ○ What is data science? / Building data science teams
- ○ Data science projects / a day in the life
- ○ Technical deep dives
- ○ Tutorials & workshops
- ○ Data exploration & visualization
- ○ Other: _____

Anything else we should know?

Your answer

A COMMUNITY OF DATA SCIENCE

- Network
- Share projects
- Share tools
- Showcase your skills
- Discuss new technologies
- Practice presenting
- Discuss new technologies
- Argue over the use of tabs or spaces



"This is our circus lets train those monkeys to dance!"

– Socrates