Big Data

School of Applied Science and Civil Engineering
Department of Data Science
H.B.Sun
2018.02.17

BIG
DATA
ANALYTICS
STORAGE
TECHNOLOGIES
INFORMATION
INTERNET
SETS
MANAGEMENT
PETABYTES
SYSTEMS
SOFTWARE
RESEARCH
SIZE
COMPLEX
SOCIAL
RELATIONAL
DATABASES
NETWORKS
RECORDS
HUNDREDS
CAPACITY
BIOLOGICAL
PROCESSING
SEARCH
PARALLEL
MASSIVELY
GROW
DIFFICULTY
GARTNER
SAN
INCLUDE
TOLERABLE
TERABYTES
PERFORMANCE
LOGS
NEEDED
TIME
ONE
LARGE
STORE
PROCESS
USING
MANAGE
ARCHIVES
MOVING
WITHIN
MAY
CURRENT
SENSOR
DEFINITION
SHARED
APPLIED
DISK
TARGET
TOOLS
ALSO
EVERY
EXAMPLES
PRACTITIONERS
CAPTURE
BUSINESS
DISTRIBUTED
GENOMICS
ABILITY
CASE
ELAPSED
WORKING
LARGER
CONTINUES
CITATION
TENS
SET
USE

我们从新获得的数据中获取社会和经济价值的能力受到缺乏专业知识的限制，处理这些数据需要独特的新技能和新工具……

数据科学领域正在社会科学与统计学，信息与计算机科学以及设计领域交汇…..

Big Data Landscape 2016 (Version 3.0)

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, CAZENA, TREASURE DATA, altiscale, Qubole

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon web services, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4, NUODB, memsql, splice MACHINE, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

Graph Databases
neo4j, GIRAPH, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, NETEZZA, Actian, kognitio, EXASOL, dremio

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

Data Transformation
alteryx, talend, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica, MuleSoft, snapLogic, BedrockData, xplenty

Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk>, DATADOG, rocana, DRIVEN, Anodot

Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

App Dev
apigee, CASK, Keen IO, Typesafe, DRIVEN

Crowd-sourcing
amazon mechanical turk, CrowdFlower, WorkFusion

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
Microsoft, GUAVUS, Datameer, Bottlenose, inter|ana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine DATA LABS, ARIMO, MODE, plotly, dataiku, tonian, DOMINO, sense, yhat, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Qlik, looker, Roambi, SISENSE, GOOMDATA, datorama, CHARTIO

BI Platforms
Power BI, amazon web services, DOMO, salesforce Wave Analytics, birst, GoodData, kyvos insights, platfora, atscale, ARCADIA, SISENSE

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk>, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
Hootsuite, NETBASE, DATASIFT, tracx, bitly, synthesio, simplereach

Real-Time
amazon web services, METAMARKETS, striim, confluent, DATATORRENT, dataArtisans

Machine Learning
Azure Machine Learning, amazon web services, H2O.ai, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, VISENZE, PredictionIO, glowfi.sh

Speech & NLP
NarrativeScience, NUANCE, WolframAlpha, semanticmachines, ARRIA, Gridspace, api.ai, cortical.io, maluuba, MindMeld, IDIBON, yseop

Horizontal AI
IBM Watson, Cortana, sentient technologies, VIV, nervana SYSTEMS, vicarious, nara, Numenta, HyperScience, SCALED INFERENCE, Descartes Labs, clarifai, DEXTRO, Geometric Intelligence, MetaMind

Search
hp, Autonomy, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

Data Services
uo, OPERA SOLUTIONS, Mu Sigma DO THE MATH, EXL analytics, DATASCIENCE, BINGHAMTON UNIVERSITY DATA SCIENCE, kaggle, datascope, DataKind

For Business Analysts
OrigamiLogic, ClearStory DATA, CIRRO, import.io

Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumall, Airtable, retention SCIENCE, custora

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, sense, QUANTIFIND, ACTIONIQ, fuse|machines, ENGAGIO

Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wise.io

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Ad Optimization
AppNexus, criteo, MediaMath, OpenX, rocketfuel, Integral, theTradeDesk, Adgorithms, dstillery, LiveIntent, TAPAD, DataXu, Appier, MOAT

Security
CYLANCE, CounterTack, cybereason, AREA 1 SECURITY, ThreatMetrix, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, siftscience, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
facebook, Clara, KASISTO, lumiata

Publisher Tools
outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
Affirm, LendingClub, OnDeck, Kreditech, zest finance, LendUp, Kabbage, tidemark, Payoff, INSIKT, uora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient technologies

Education/ Learning
KNEWTON, Clever, declara, PANORAMA EDUCATION, knowre

Life Sciences
23andMe, PATHWAY GENOMICS, Counsyl, deep genomics, Recombine, KYRUUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia health, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

Cross-Infrastructure/Analytics

amazon web services, Google, Microsoft, IBM, SAP, sas, 1010data, hp, Autonomy, VERTICA An HP Company, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

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

Query / Data Flow
SLAMDATA, HIVE, APACHE DRILL, Google Cloud Dataflow

Data Access
cassandra, APACHE HBASE, accumulo, mongoDB, kafka, CouchDB, riak, SciDB, OPENTSDB, nifi

Coordination
talend, ZOOKEEPER, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
R, ScalaLab, NumPy, SciPy

Machine Learning
mllib, mahout, Aerosolve, Apache SINGA, MADlib, Caffe, torch, CNTK, TensorFlow, learn, FeatureFu, VELES, WEKA, DIMSUM, jupyter, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
Zeppelin

Data Sources & APIs

Health
JAWBONE, GARMIN, practice fusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote

Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, S&P CAPITAL IQ, YODLEE, PREMISE, quandl, xignite, CB INSIGHTS, mattermark, StockTwits, estimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location / People / Entities
acxiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTODB, factual, PlaceIQ, CIRCULATE, placemeter, BASIS TECHNOLOGY, Sense 360

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, PLURALSIGHT, DataCamp, INSIGHT, DataElite, The Data Incubator, METIS

Last Updated 3/23/2016

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

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 / IN-MEMORY
amazon web services, databricks, confluent, striim, GridGain, METAMARKETS, DATATORRENT, dataArtisans, ORACLE Coherence, hazelcast, TERRACOTTA

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 machine, MariaDB, VOLTDB, citusdata, Trafodion, deepdb, paradigm4

GRAPH DBS
neo4j, IBM, ORACLE, GIRAPH, 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, UNIFI

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

DATA GOVERNANCE
informatica, IBM, WAVEFRONT, skyhigh, collibra, Alation, Waterline Data

MGMT / MONITORING
amazon web services, New Relic, APPDYNAMICS, actifio, DATADOG, unravel, splunk>, Anodot, rocana, Numerify, pagerduty

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, Keen IO, rainforest, CASK

CROWDSOURCING
amazon mechanical turk, Upwork, WorkFusion, CrowdFlower

HARDWARE
Google TPU, ARM, nervana SYSTEMS, Graphcore, MYTHIC, cerebras, NVIDIA, Movidius, SCORTEX

ANALYTICS

DATA ANALYST PLATFORMS
Microsoft, pentaho, alteryx, Digital Reasoning, GUAVUS, AYASDI, MATTIVIO, Dataweer, Quid, ClearStory, OrigamiLogic, inter|ana, Bottlenose, ARIMO, ENDOR, MODE

DATA SCIENCE PLATFORMS
IBM, KNIME, data iku, DOMINO, yhat, rapidminer, CONTINUUM ANALYTICS, Alpine DATA LABS, ALGORITHMIA, Angoss

BI PLATFORMS
Microsoft, DOMO, looker, amazon web services, Wave Analytics, atscale, ARCADIA DATA, GoodData, SISENSE

VISUALIZATION
tableau, SAP Lumira, Google Cloud Platform, Qlik, MAPD, celonis, Periscope DATA, ZEPL, CHARTIO, plotly, GOODDATA

VERTICAL ANALYTICS
PREDIX, C3 IoT, CAPE, UPTAKE, Orbital Insight, TACHYUS, Alluvium, datorama

STATISTICAL COMPUTING
sas, SPSS, MATLAB

DATA SERVICES
Palantin, Mu Sigma, DATASCIENCE, OPERA, BIG DATA DATA SCIENCE, kaggle, Fractal, EXL, DataKind, FF

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

HORIZONTAL AI
IBM Watson, Cortana, Face++, 旷视, sentient technologies, Voyager Labs, vicarious, clarifai, :) Affectiva, SENSETIME, CognitiveScale, Numenta, Cronocam, PETUUM, nara, OSARO, SI SCALED INTELLIGENCE, THE CURIOUS AI COMPANY, BLUE VISION LABORATORIES

SPEECH & NLP
Google Cloud Platform, twilio, amazon alexa, NarrativeScience, semantic machines, WolframAlpha, MindMeld, 出门问问, NUANCE, ARRIA, iDIBON, TalkIQ, cortical.io, maluuba, snips, vseop, Gridspace, SoundHound Inc.

SEARCH
Autonomy, elastic, EXALEAD, ORACLE ENDECA, ThoughtSpot, Algolia, Lucidworks, swiftype, MAANA, alphasense, Searchlink, SINEQUA

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

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

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

APPLICATIONS – ENTERPRISE

SALES
einstein, CHORUS, INSIDESALES.COM, conversica, Gong, clari, AVISO, TACT, fuse|machines, TROOPS

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

MARKETING - B2C
Zeta, bloomreach, blueyonder, PERSADO, kahuna, ACTIONIQ, BLUECORE, SAILTHRU, QUANTIFIND, mparticle, Amplero

CUSTOMER SERVICE
MEDALLIA, zendesk, CLARABRIDGE, Gainsight, ClickFox, NGDATA, DigitalGenius, appuri, AUTOMAT, frame.ai, msg.ai, INTERCOM

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

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

FINANCE
anaplan, uora, tidemark, S/4 HANA, TRADESHIFT

ENTERPRISE PRODUCTIVITY
slack, X, facebook, ORACLE, lumiata, diffbot, 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, sqrl, SOCURE, AREA 1 SECURITY, FORTSCALE, Keybase, sparkcognition

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus, MediaMath, criteo, xAd, rocketfuel, Integral Ad Science, theTradeDesk, OpenX, MOAT, dstillery, Adgorithms, drawbridge, LiveIntent, TAPAD, DataXu, gumgum, databerries, Appier, DYNAMIC YIELD, Yieldmo

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

GOVERNMENT
Socrata, OPENGOV, mark43, FiscalNote, OpenDataSoft

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

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

REAL ESTATE
Opendoor, VTS, CREDIFI, reonomy, COMPSTAK

INSURANCE
Metromile, Lemonade, CYENCE, Shift Technology, Tractable

HEALTHCARE
FLATIRON, KYRUUS, HealthTap, METABIOTA, Ginger.io, Glow, PANDYON, COTA, zebra, ovia, AiCure, enlitic, Qventus, imagia, ARTERYS, BAYLABS, CLOUD MEDX, IMAGEN, Kang Health, freenome, DIAGNOSIS PROJECT

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

TRANSPORTATION
UBER, ZOOX, TESLA, nuTonomy, CLEARPATH, nauto, drive.ai, pilot.ai, AMOTIVE, OTTO, OPTIMUS RIDE, nexar, comma.ai, netradyne, Civil Maps, NIO

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

COMMERCE
instacart, STITCH FIX, RetailNext, HowGood

OTHER
eHarmony, stem, rethink robotics, AMPER MUSIC, hopper, SIGHT MACHINE, BOXEVER, celect, VERDIGRIS, duetto, Unbabel, Jukedeck, Second Spectrum

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

OPEN SOURCE

FRAMEWORK
hadoop, Apache Kafka, hadoop MapReduce, YARN, Flink, TEZ, MESOS, Spark, CDAP

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

DATA ACCESS
nifi, mongoDB, cassandra, opentsdb, SciDB, CouchDB, OPENTSDB, riak, HBASE, Spanner, accumulo

COORDINATION
talend, Apache Zookeeper, Apache Ambari

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

STAT TOOLS
python, R, ScalaLab, NumPy, SciPy

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

SEARCH
elasticsearch, Solr, Lucene

LOG ANALYSIS
elasticsearch, kibana, logstash

VISUALIZATION
BEAKER, Rodeo

COLLABORATION
jupyter, Zeppelin, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry

DATA SOURCES & APIS

HEALTH
JAWBONE, 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, YODLEE, PREMISE, estimize, SECOND MEASURE, SYSTEM, Eagle Alpha, StockTwits, PLAID, mattermark, Thinknum

AIR / SPACE / SEA
PLANET LABS, Airware, spire, SKYCATCH, AIROBOTICS, Descartes Labs, UNDERSTORY, TELLUSLABS, WINDWARD, DroneDeploy, MarineTraffic

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

LOCATION INTELLIGENCE
FOURSQUARE, MapAnything, Sense 360, PlaceIQ, esri, factual, CARTO, Mapillary, STREETLINE

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

DATA RESOURCES

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

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

V2 – Last updated 5/3/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)

mattturck.com/bigdata2017

FIRSTMARK EARLY STAGE VENTURE CAPITAL

Agenda

- a Bird View of Big Data
- What and How Should We Do
- Big Data Technology Ecosphere
- Think about Statistics

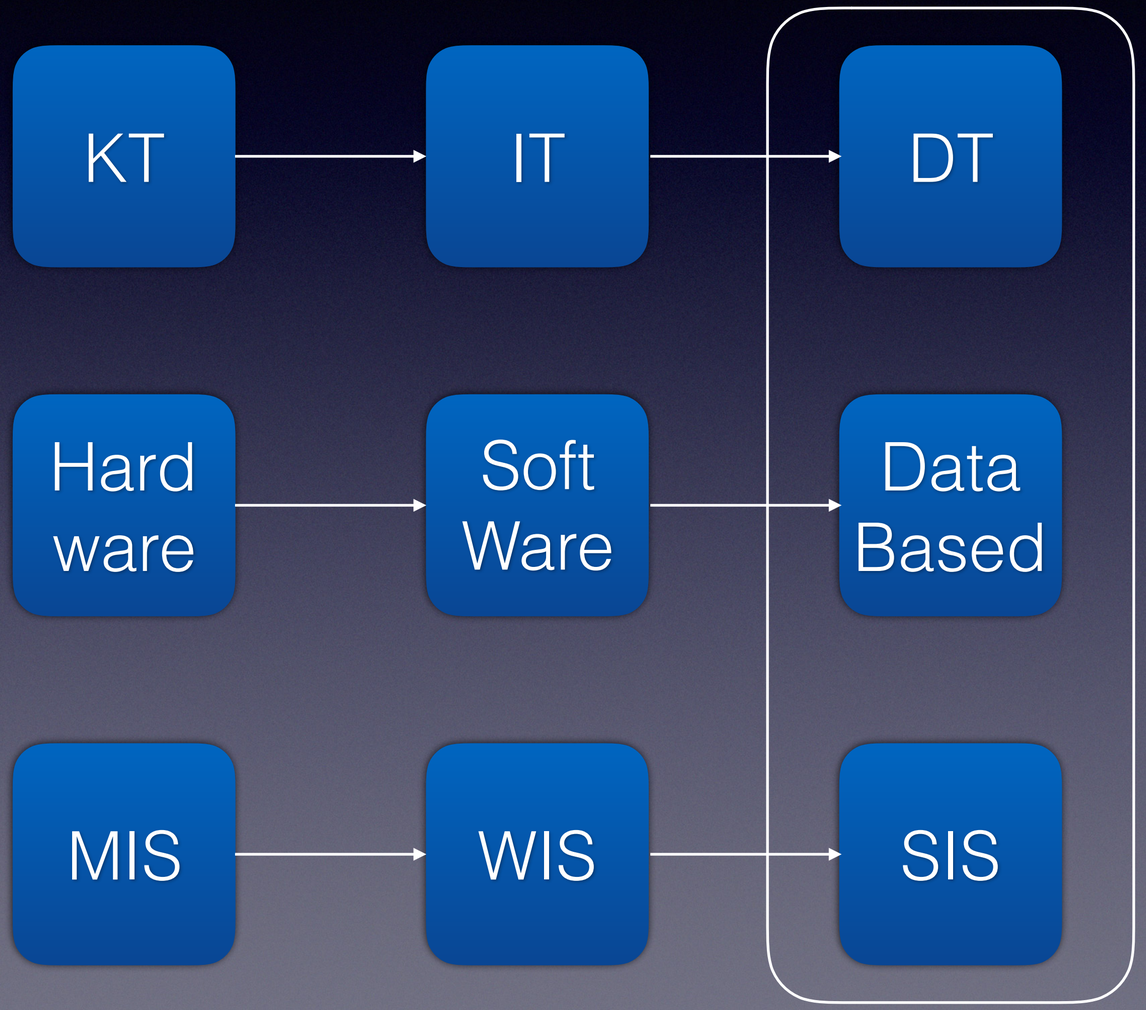
a Bird View of Big Data

- ★ The age of Big Data

a Bird View of Big Data

大数据引领时代变迁

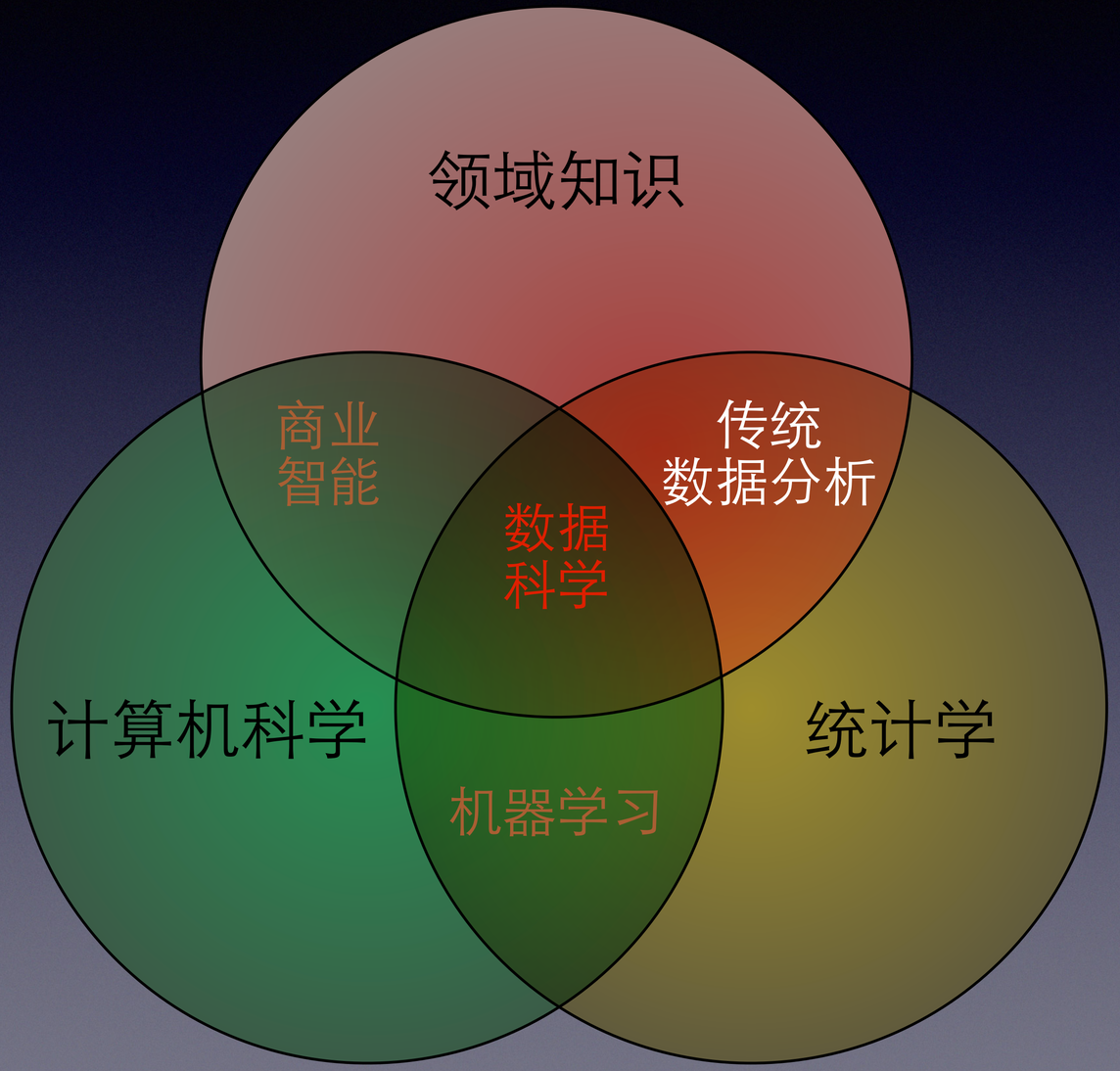
Data Driven
All in Data（AID）



a Bird View of Big Data

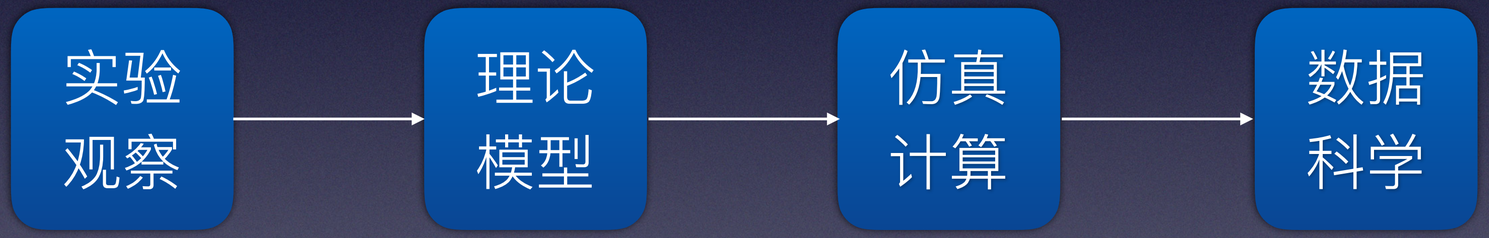
大数据引领知识重构

Crossing
Mixing
Uniting
Come together……



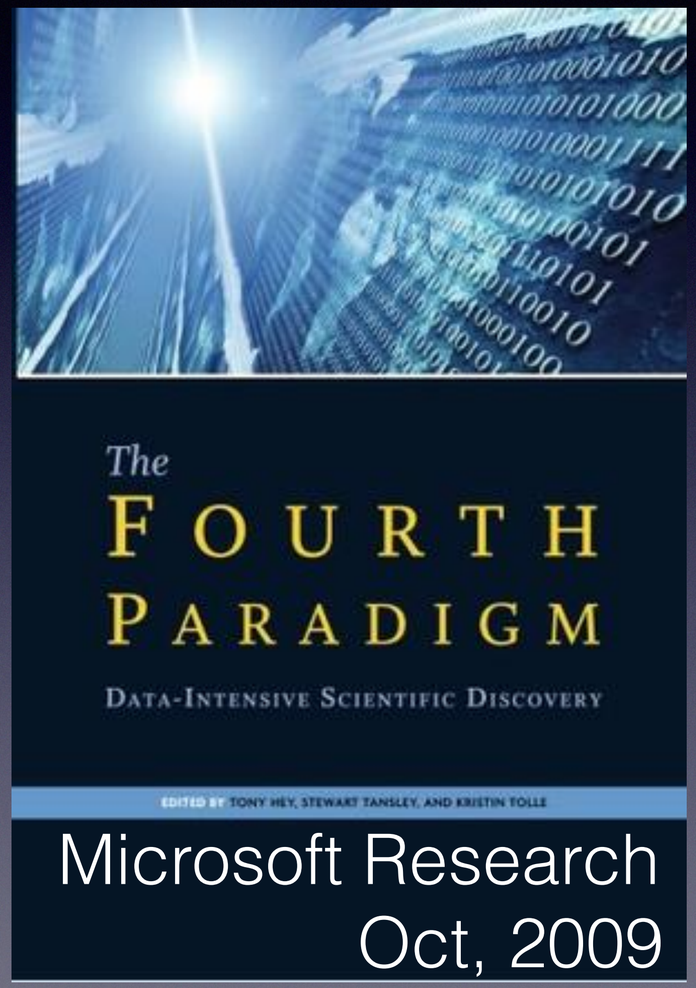
a Bird View of Big Data

- 大数据是一种思维方式
- 大数据是一种研究范式

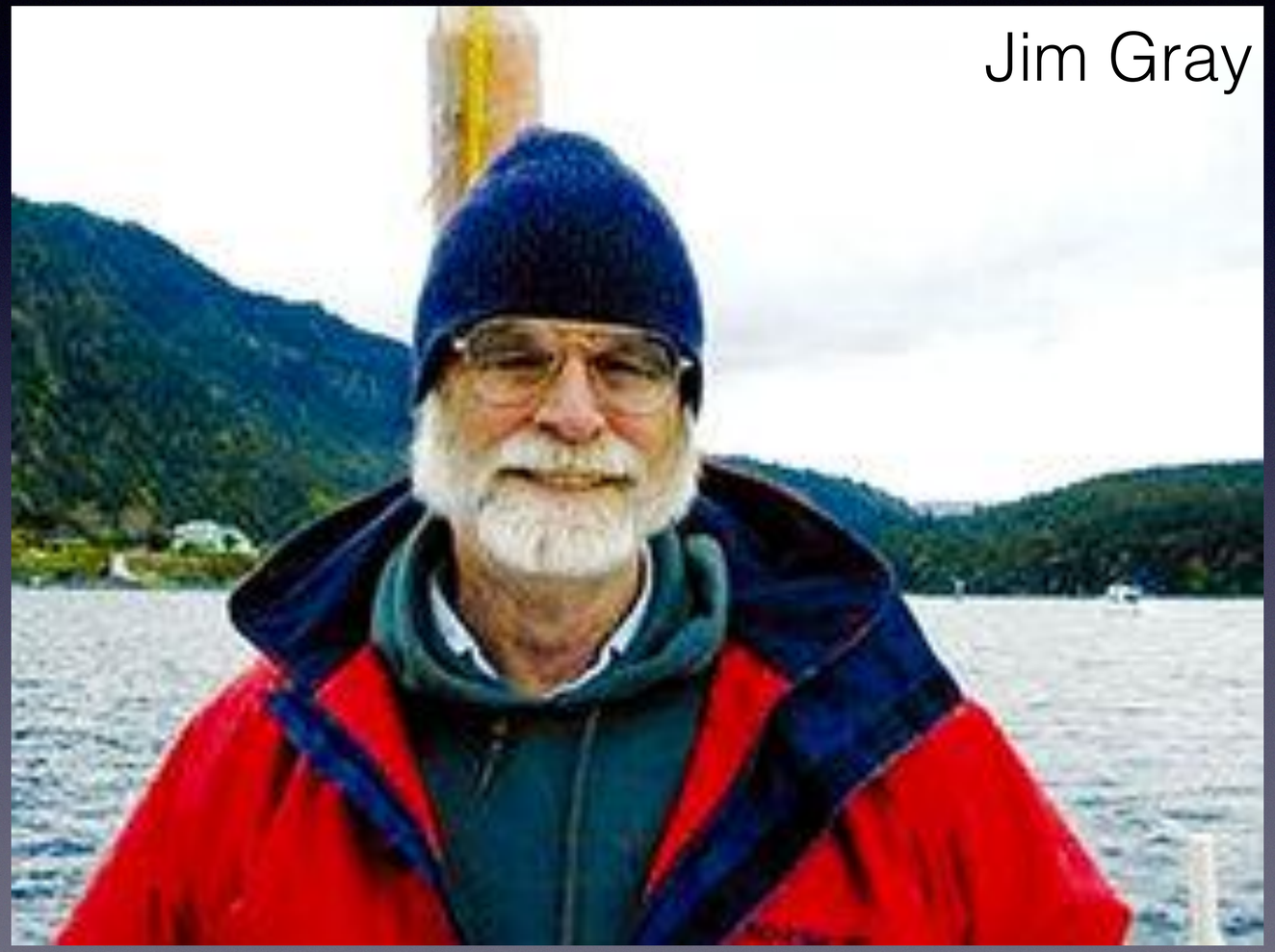


- 大数据是一种技术手段
- 大数据是一种战略资源

a Bird View of Big Data



Microsoft Research
Oct, 2009

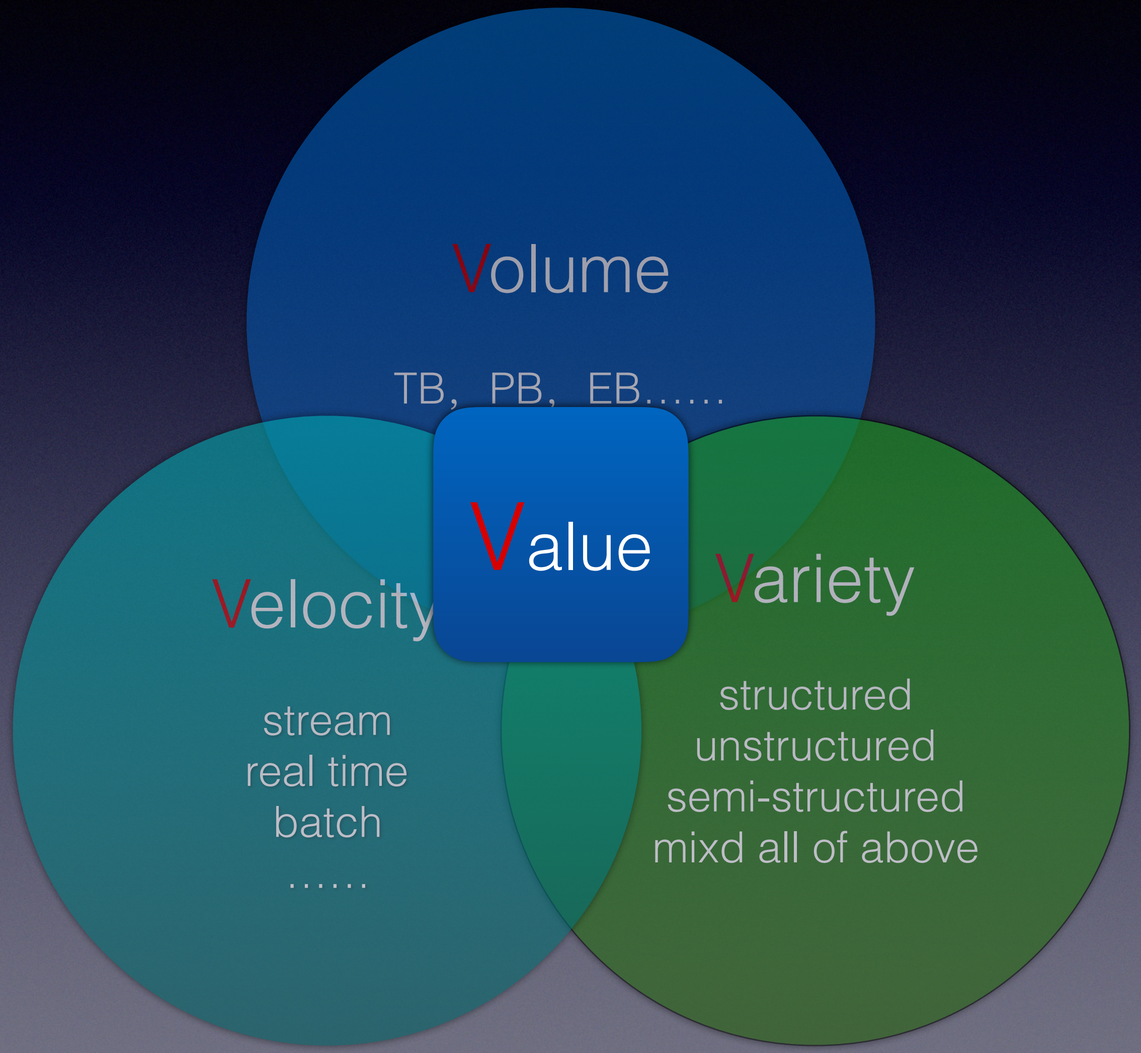


a Bird View of Big Data

★ Big Data is ……? —— Not only big!

a Bird View of Big Data

大数据的特征——4V



a Bird View of Big Data

大数据分析的目的——最大限度洞见（Insight）价值

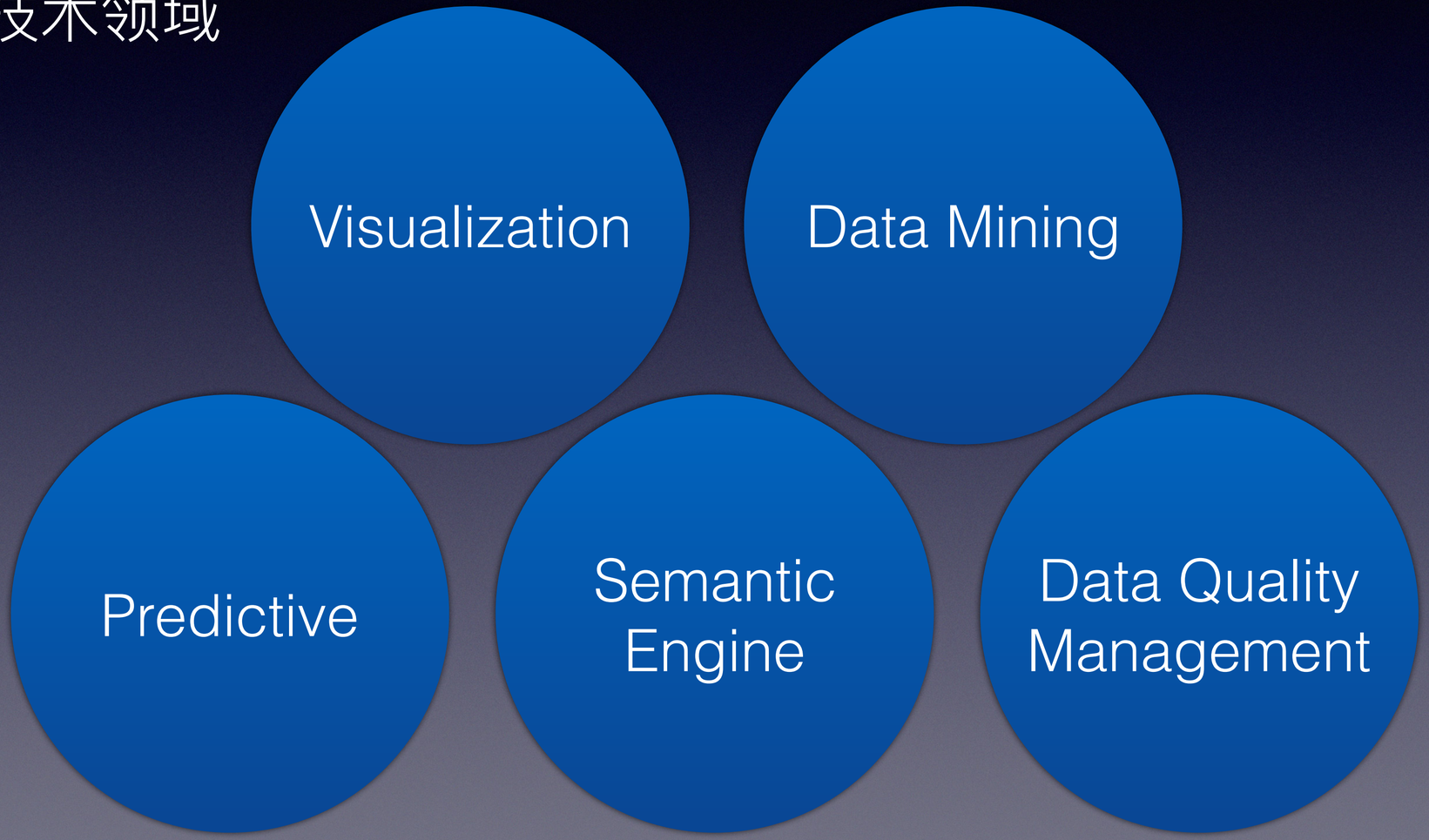
海量、不同来源、不同形式、包含不同信息的数据被整合、分析，原本孤立的数据变得互相联通。这使得人们通过数据分析，能发现小数据时代很难发现的新知识，创造新的价值。

a Bird View of Big Data

大数据分析的主要技术路线

a Bird View of Big Data

大数据分析的主要技术领域



a Bird View of Big Data

大数据分析的主要技术基础

数据

算法

计算

a Bird View of Big Data

大数据分析的主要技术挑战

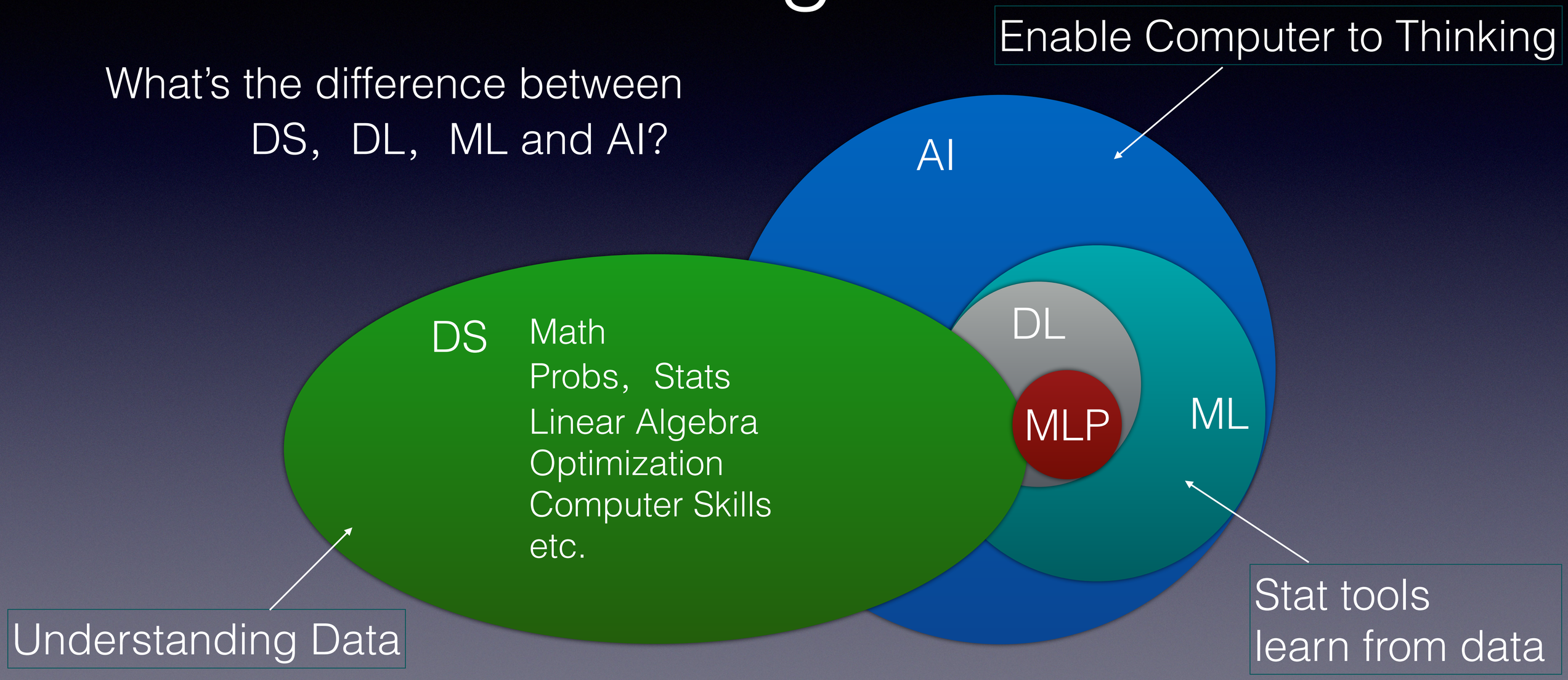
存储容量

存取速度

计算速度

a Bird View of Big Data

What's the difference between
DS, DL, ML and AI?



引擎——神经网络

燃料——大 数 据

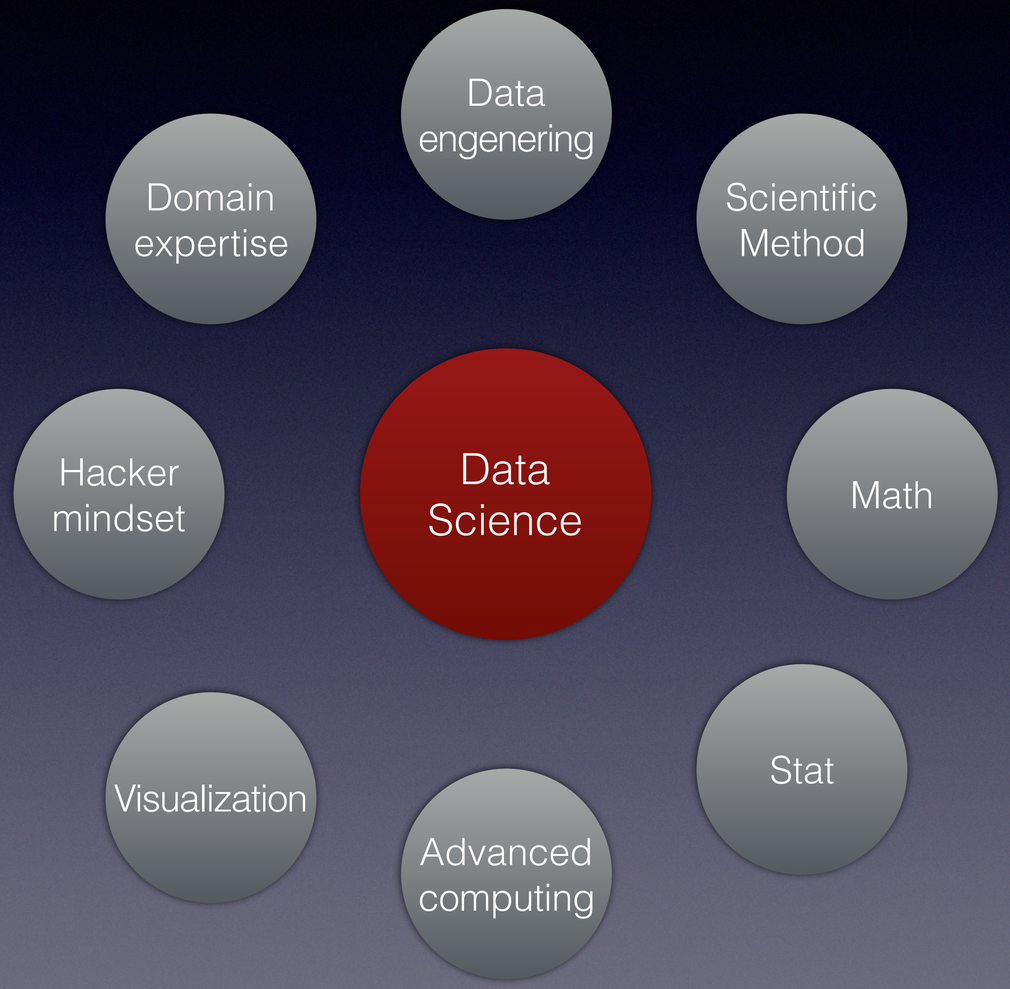
I belive this is our best shot at progress towards real AI.

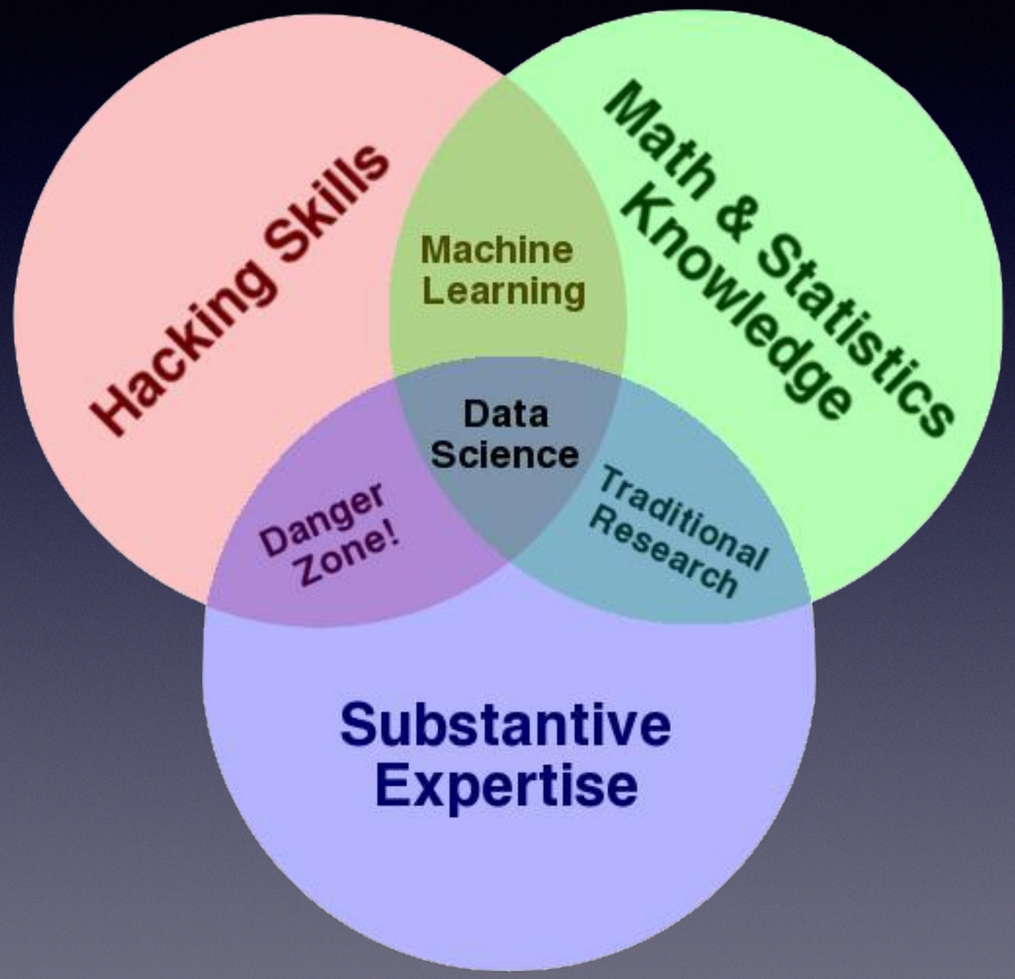
—Andrew Ng

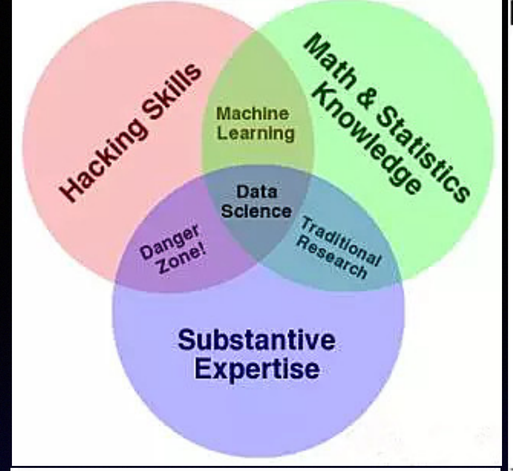
a Bird View of Big Data

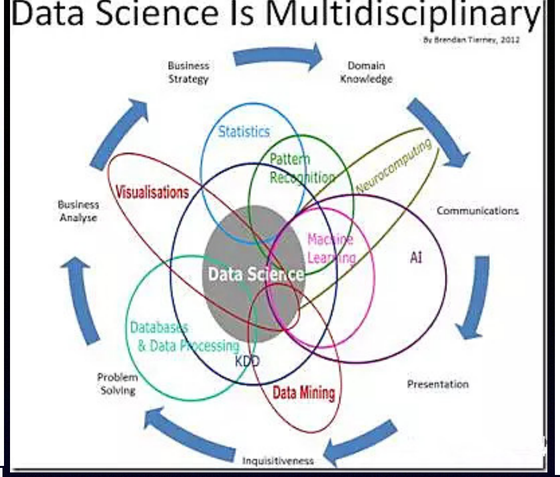
★ Data Science is…… ?

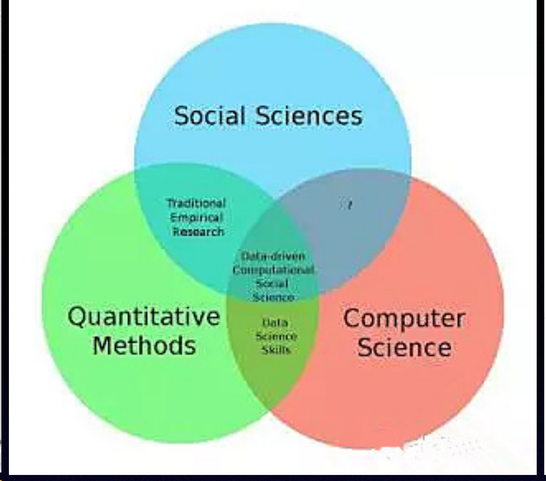
关于数据科学的定义

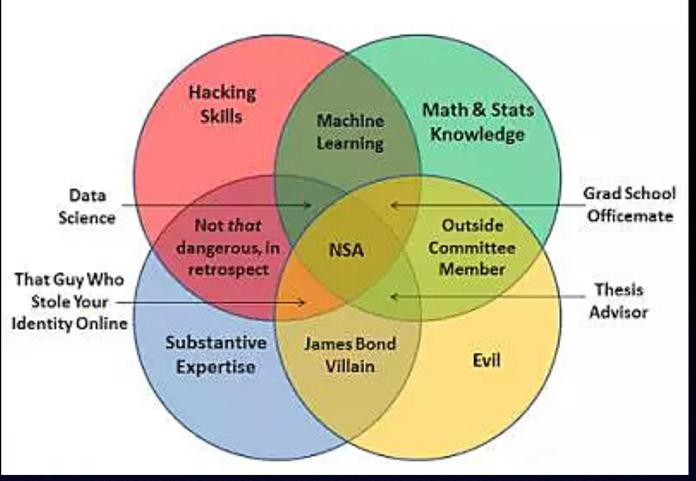

Data engenering
Domain expertise
Scientific Method
Hacker mindset
Data Science
Math
Visualization
Advanced computing
Stat

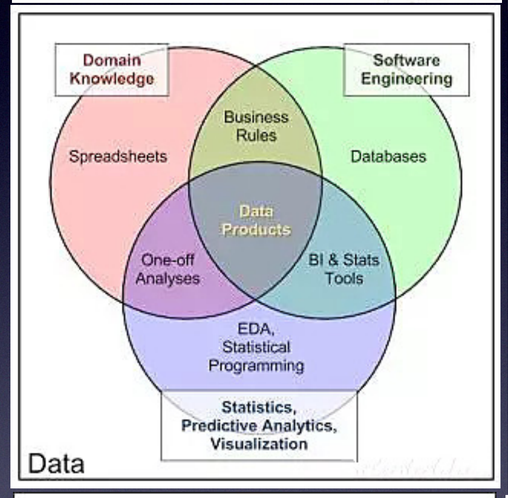

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

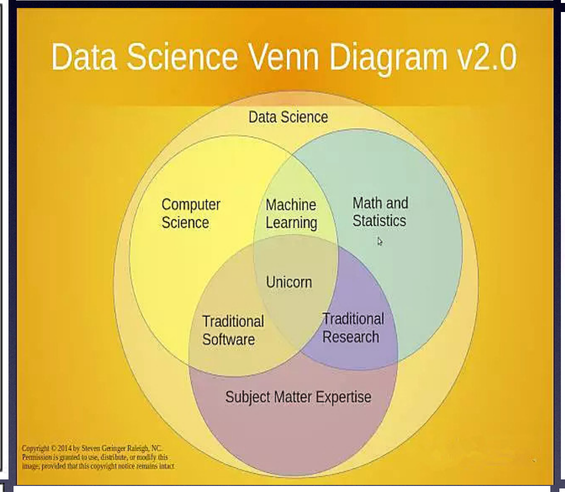

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

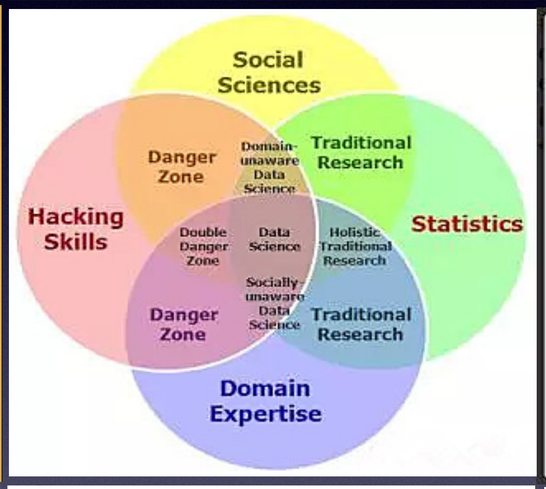

Data Science Is Multidisciplinary
By Brendan Tierney, 2012
Business Strategy
Domain Knowledge
Statistics
Pattern Recognition
Neurocomputing
Visualisations
Business Analyse
Communications
Machine Learning
AI
Data Science
Databases & Data Processing
KDD
Problem Solving
Data Mining
Presentation
Inquisitiveness

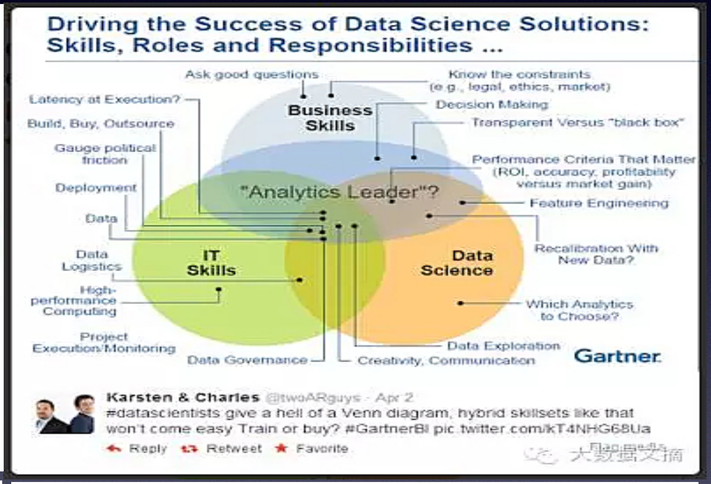

Social Sciences
Traditional Empirical Research
Data-driven Computational Social Science
Quantitative Methods
Data Science Skills
Computer Science

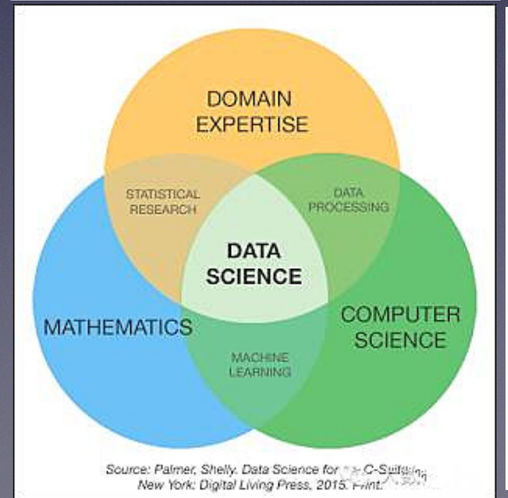

Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Not that dangerous, in retrospect
NSA
Outside Committee Member
Grad School Officemate
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil

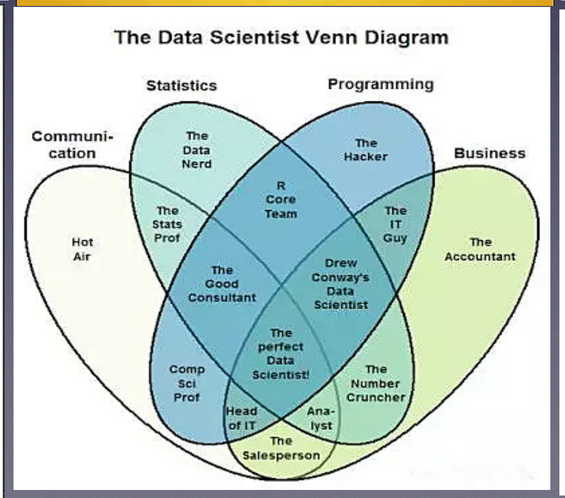

Domain Knowledge
Software Engineering
Business Rules
Spreadsheets
Databases
Data Products
One-off Analyses
BI & Stats Tools
EDA, Statistical Programming
Statistics, Predictive Analytics, Visualization
Data

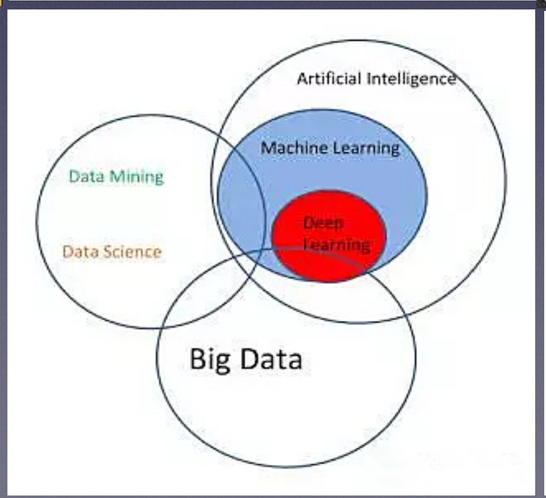

Data Science Venn Diagram v2.0
Data Science
Computer Science
Machine Learning
Math and Statistics
Unicorn
Traditional Software
Traditional Research
Subject Matter Expertise

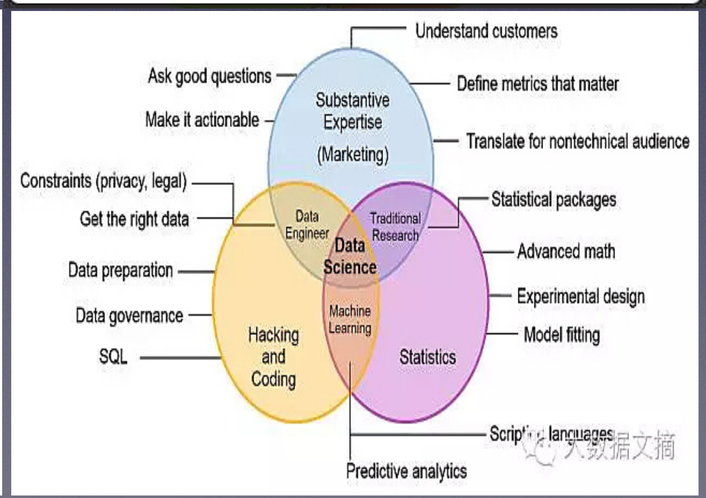

Social Sciences
Danger Zone
Domain-unaware Data Science
Traditional Research
Hacking Skills
Statistics
Double Danger Zone
Data Science
Holistic Traditional Research
Socially unaware Data Science
Danger Zone
Traditional Research
Domain Expertise

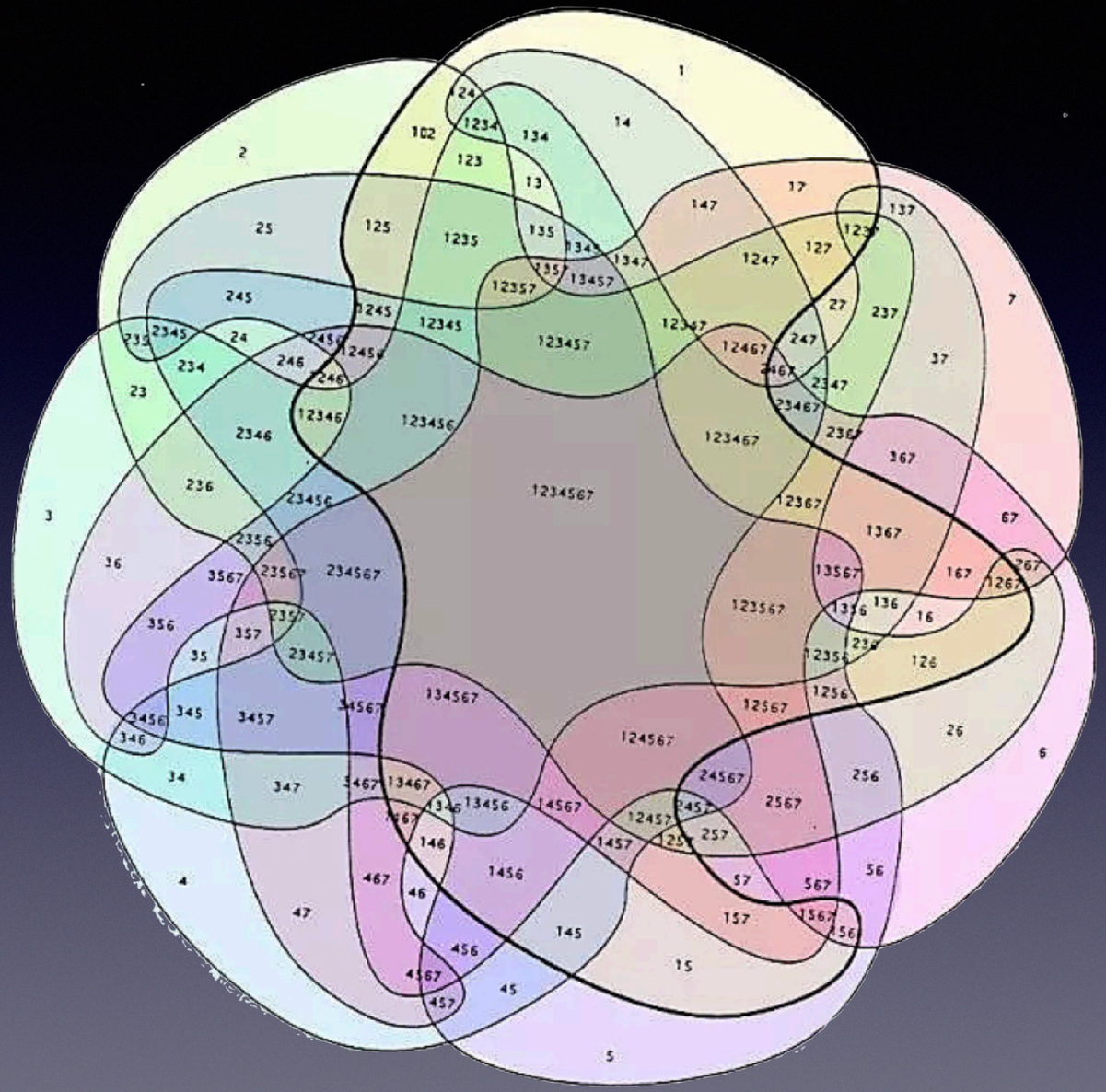

Driving the Success of Data Science Solutions: Skills, Roles and Responsibilities ...
Ask good questions
Know the constraints (e.g., legal, ethics, market)
Latency at Execution?
Business Skills
Decision Making
Build, Buy, Outsource
Transparent Versus "black box"
Gauge political friction
Performance Criteria That Matter (ROI, accuracy, profitability versus market gain)
Deployment
"Analytics Leader"?
Feature Engineering
Data
Recalibration With New Data?
Data Logistics
IT Skills
Data Science
High-performance Computing
Which Analytics to Choose?
Project Execution/Monitoring
Data Exploration
Data Governance
Creativity, Communication
Gartner
Karsten & Charles @twoARguys · Apr 2
#datascientists give a hell of a Venn diagram, hybrid skillsets like that won't come easy. Train or buy? #GartnerBI pic.twitter.com/kT4NHG68Ua
Reply Retweet Favorite


DOMAIN EXPERTISE
STATISTICAL RESEARCH
DATA PROCESSING
DATA SCIENCE
MATHEMATICS
COMPUTER SCIENCE
MACHINE LEARNING
Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.


The Data Scientist Venn Diagram
Statistics
Programming
Communi-cation
Business
The Data Nerd
The Hacker
R Core Team
The Stats Prof
The IT Guy
Hot Air
The Accountant
The Good Consultant
Drew Conway's Data Scientist
The perfect Data Scientist!
Comp Sci Prof
The Number Cruncher
Head of IT
Ana-lyst
The Salesperson


Artificial Intelligence
Machine Learning
Data Mining
Deep Learning
Data Science
Big Data


Understand customers
Ask good questions
Define metrics that matter
Make it actionable
Substantive Expertise (Marketing)
Translate for nontechnical audience
Constraints (privacy, legal)
Statistical packages
Get the right data
Data Engineer
Traditional Research
Data Science
Advanced math
Data preparation
Experimental design
Data governance
Machine Learning
Model fitting
SQL
Hacking and Coding
Statistics
Scripting languages
Predictive analytics


1
2
3
4
5
6
7
12
13
14
15
16
17
23
24
25
26
27
34
35
36
37
45
46
47
56
57
67
123
124
125
126
127
134
135
136
137
145
146
147
156
157
167
234
235
236
237
245
246
247
256
257
267
345
346
347
356
357
367
456
457
467
567
1234
1235
1236
1237
1245
1246
1247
1256
1257
1267
1345
1346
1347
1356
1357
1367
1456
1457
1467
1567
2345
2346
2347
2356
2357
2367
2456
2457
2467
2567
3456
3457
3467
3567
4567
12345
12346
12347
12356
12357
12367
12456
12457
12467
12567
13456
13457
13467
13567
14567
23456
23457
23467
23567
24567
34567
123456
123457
123467
123567
124567
134567
234567
1234567

About the UC Berkeley School of Information
About MIDS Faculty
What is Data Science?
Which Degree Program is Right for You?
FAQ
In the News
2U, Our Program Partner

? REQUEST INFO

What is Data Science?

A New Field Emerges

There is significant and growing demand for data-savvy professionals in businesses, public agencies, and nonprofits. The supply of professionals who can work effectively with data at scale is limited, and is reflected by rapidly rising salaries for data engineers, data scientists, statisticians, and data analysts.

A recent study by the McKinsey Global Institute concludes, "a shortage of the analytical and managerial talent necessary to make the most of Big Data is a significant and pressing challenge (for the U.S.)." The report estimates that there will be four to five million jobs in the U.S. requiring data analysis skills by 2018, and that

a Bird View of Big Data

美国密歇根大学在2015年9月宣布了一个1亿美金的“数据科学项目（Data Science Initiative）”，计划在未来4年聘请35名新教授，支持与数据相关的跨学科研究。该项目的网站上有如下对数据科学的描述：

“数据科学是科学发现和实践的结合，其包括对大量类型各异的数据进行收集，管理，清理，分析，可视化和结果解释。其应用遍及各种科学，平移和交叉领域。”

a Bird View of Big Data

数据科学六个部分

1.数据探索和准备
2.数据表示和变换
3.数据编程与计算
4.数据建模与分析
5.数据可视化与展示
6.数据科学的科学

——“50 Years of Data Science.”Donoho, David. 2015.

a Bird View of Big Data

“没有系统的数据科学作为指导的大数据研究，就如同不利用工程科学的知识来建造桥梁，很多桥梁可能会坍塌，并带来严重的后果。”

——Michael Jordan, Berkeley UCLA

a Bird View of Big Data

统计学和计算机科学都不是数据科学的全部。

统计学的多数传统核心内容，尚未做好处理以十亿为单位的记录及非结构化数据的准备，统计理论只在数据科学中扮演部分角色；同样，计算机科学的核心领域虽然擅长高效获取并存储大量结构化和非结构化数据，但在通过建模、分类和可视化等形式将数据转换成信息的能力上存在短板。

统计学贡献了基础方法，计算机科学提供了实现和计算的工具。

无可否认，数据科学还有艺术的一面。

a Bird View of Big Data

大数据视野下的统计学与计算机科学

- 对数据定义的延伸，迫使统计学与计算机科学从各自的传统核心领域进入边缘领域
- 对数据分析的过程，迫使统计学与计算机科学在边缘领域进行的高度融合，奠定了数据科学的基础

按照严谨的学科概念，迄今为止，数据科学还没有形成一个学科，基本上还只是一个概念，其内涵和外延尚在探索之中……但有一点是公认的：

数据科学=从数据中得到问题答案的科学

Born to Appling

Computer Science

- Liebniz – Binary Logic.
- Babbage, Lovelace
- Boolean Algebra
- Punch cards.
- Turing machines
- Information Theory
- Weiner & Cybernetics
- Von Neumann Architecture.
- Sort & Search Algorithms – Dijkstra, Kruskal, Shell Sort, …
- Heuristics – Simulated Annealing, …
- Text/ string search
- 1974 Peter Naur "Concise Survey of Computer Methods", **Data Science, Datalogy**
- Knuth – Art of Computer Programming.
- Database Marketing
- Data Mining, Knowledge Discovery
- "Data science, classification, and related methods."

Data Technology

- Catrography
- Astronomical Charts.
- William Playfair
- Charles Minard
- Florence Nightingale.
- First IBM Computers
- DBMS.
- Graph Algorithms
- Multigrid methods
- Tree based methods.
- Removable Disk drives
- Relational DBMS.
- 1989 First KDD Workshop
- Gregory Piatetsky-Shapiro.
- Desktop, floppy
- SQL, OOP
- High level languages.
- William Cleveland: Data Science
- Leo Breimann: Statistical Modeling: 2 Cultures

Visualization

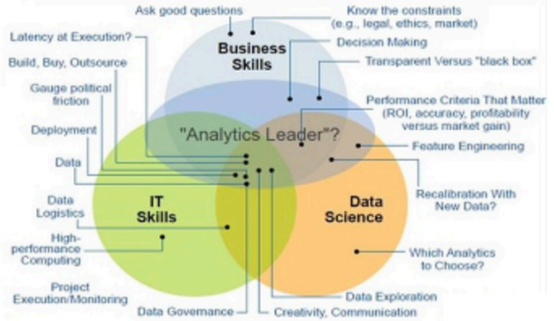
- John Tukey
- Jacques Bertin.
- Edward Tufte.
- Grammar of Graphics
- Word Cloud, Tag Cloud.

Mathematics/ OR

- Calculus
- Logarithms
- Newton-Raphson.
- Optimization Methods
- Fourier and other transforms
- Matrix & Generalizations
- Non-euclidean geometries.
- Applications to Military, manufacturing, Communications.
- Networks
- Assignment Problems
- Automation
- Scheduling

Statistics

- Probability
- Correlation
- Bayes Theorem.
- Regression, Least Squares
- Time Series.
- Theoretical Foundations of Modern Stats
- Hypothesis, DOE
- Mathematical Statistics.
- **1962** John W. Tukey, Future of Data Analysis
- Bayesian Methods
- Time Series Methods (Box Cox, Survival, etc.)
- Stochastic Methods.
- 1976 – SAS Institute
- 1977 The International Association for Statistical Computing (IASC).
- Simulation, Markov
- Computational Statistics.
- Decision Science
- Pattern recognition
- Machine learning.



Pre 1800s | 1800-1900 | 1900-1940 | 1940-1960 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010

a Bird View of Big Data

- ★ Data Scientist is……?

a Bird View of Big Data

数据科学家综合使用一整套科学工具与技术（数学，计算，视觉，分析，统计，试验，艺术，问题界定，模型建立与检验等），用于从数据收集中获得新发现、洞见（Insight）与价值。

a Bird View of Big Data

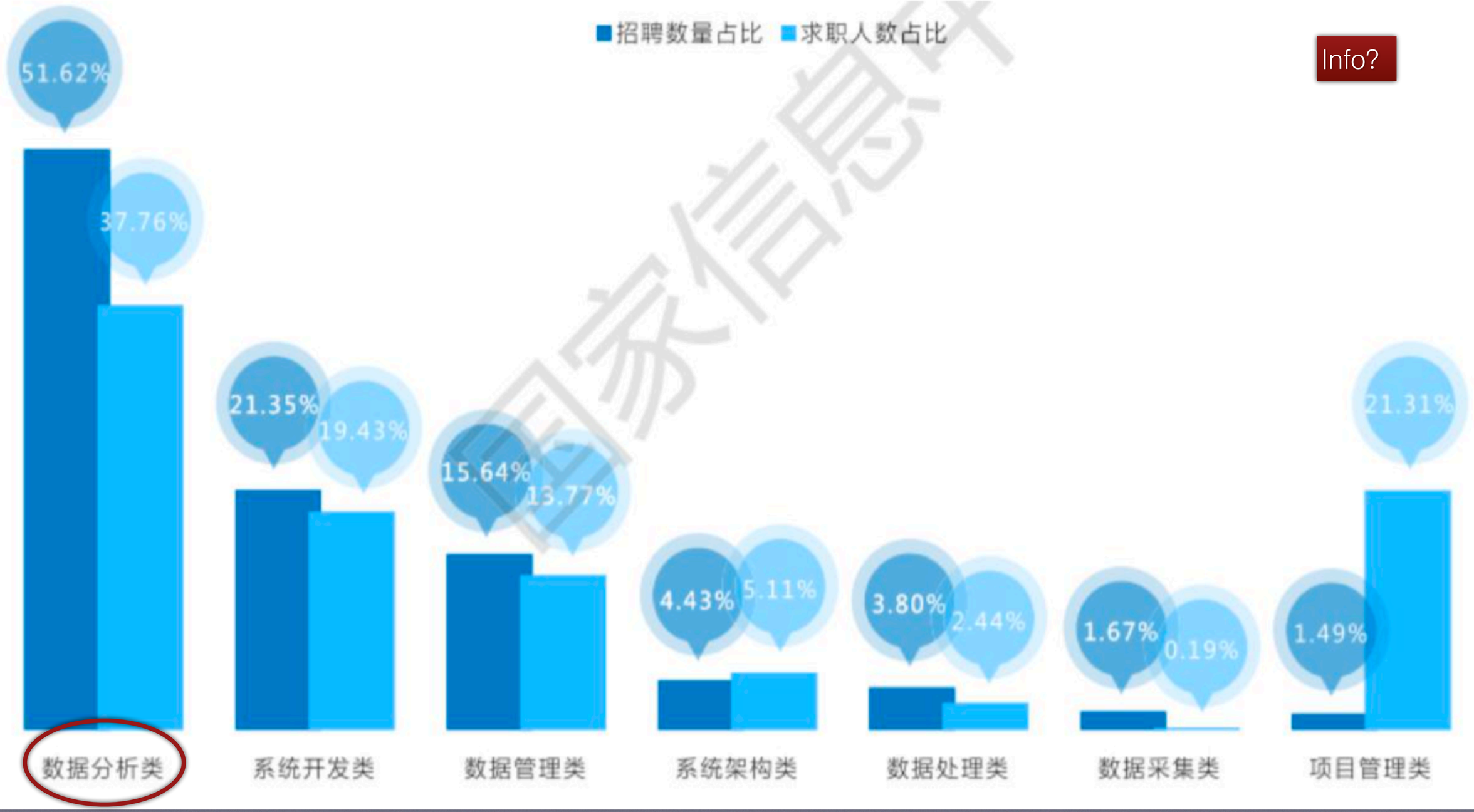
- I’m not a DBA, but I’m good at SQL.
- I’m not a programmer but I am good at programming.
- I’m not a statistician but I am good at applying statistical techniques.

—— 《Enterprise Data Analysis and Visualization: An Interview Study》， by Sean Kandel, Andreas P aepcke, Joseph M. Hellerstein, and Jeffrey Heer

What and How Should We Do?

What and How Should We Do?

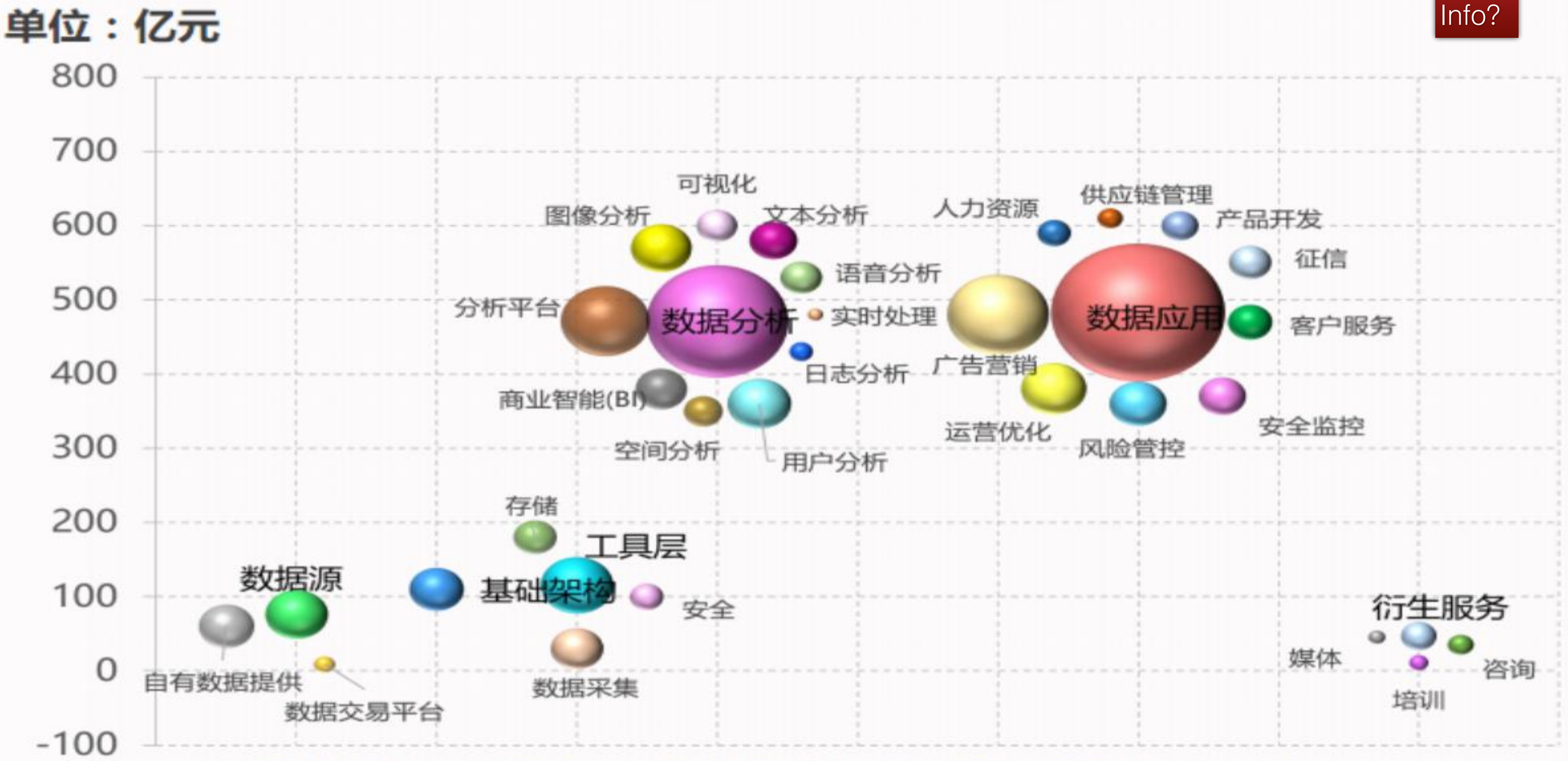
- What?


招聘数量占比
求职人数占比
51.62%
37.76%
21.35%
19.43%
15.64%
13.77%
4.43%
5.11%
3.80%
2.44%
1.67%
0.19%
1.49%
21.31%
数据分析类
系统开发类
数据管理类
系统架构类
数据处理类
数据采集类
项目管理类

Info?

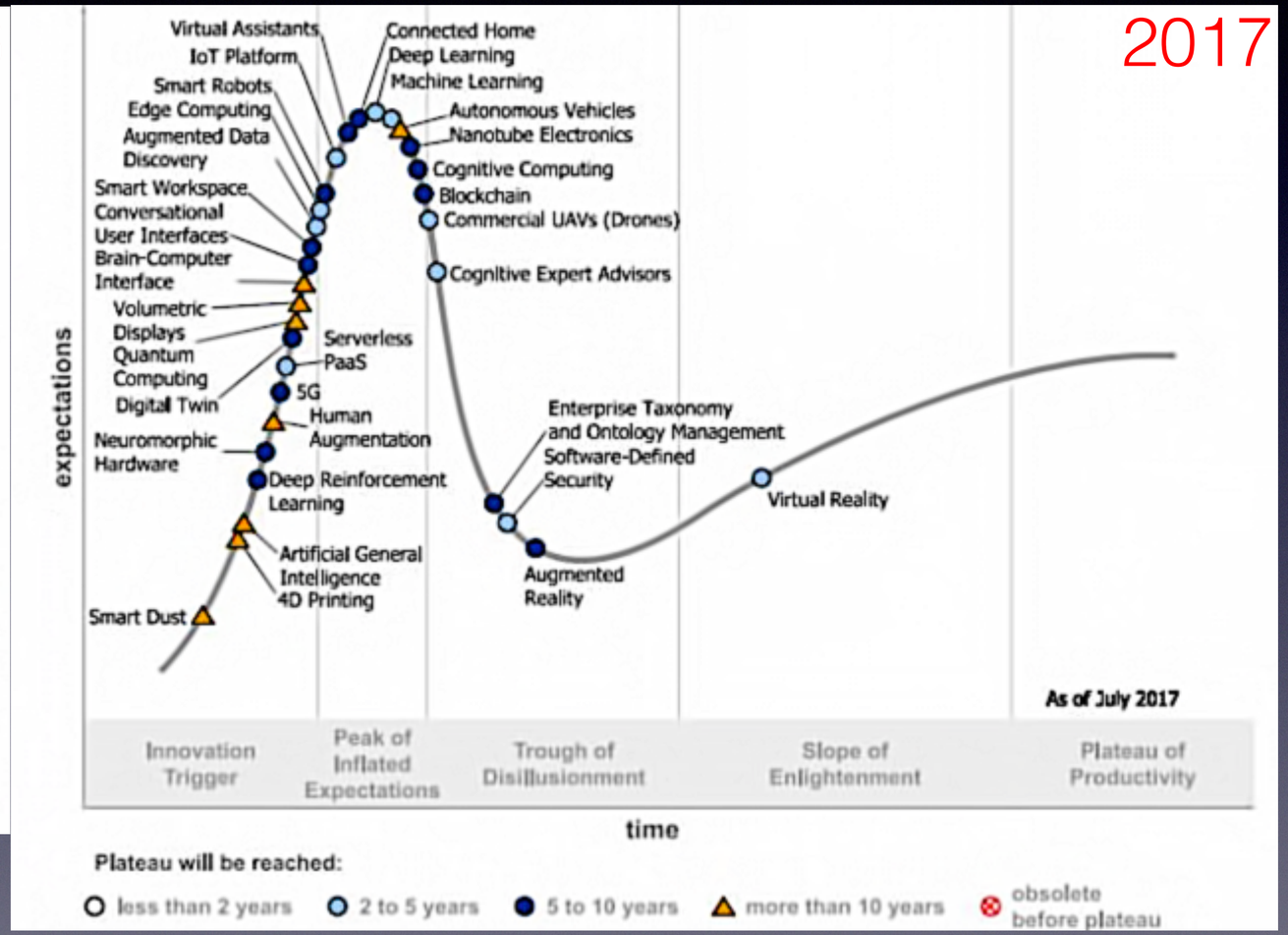
2012-2016 各产业项目融资情况

Info?



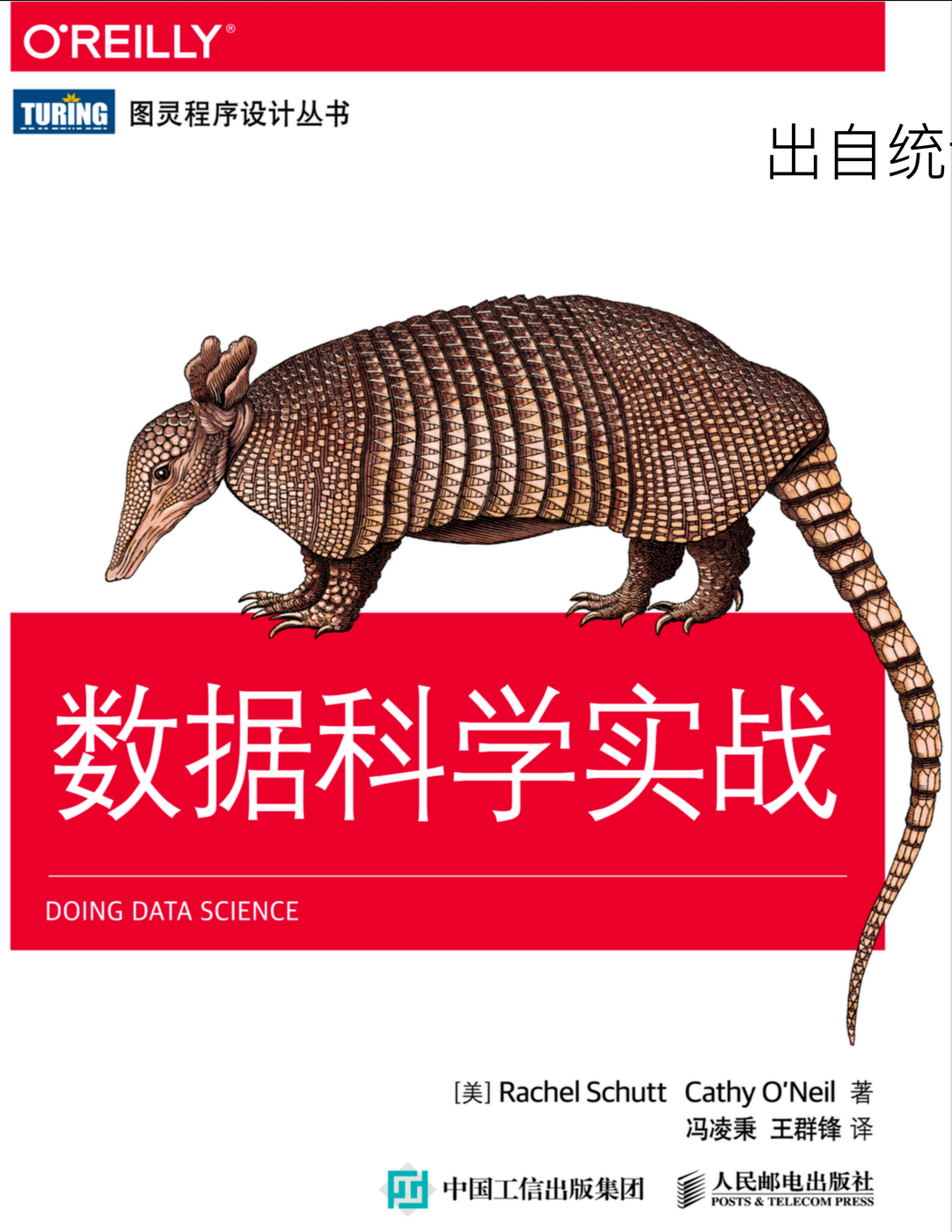
数据来源：中国大数据和人工智能产业分析平台（中国信通院、数据猿，2017.03）

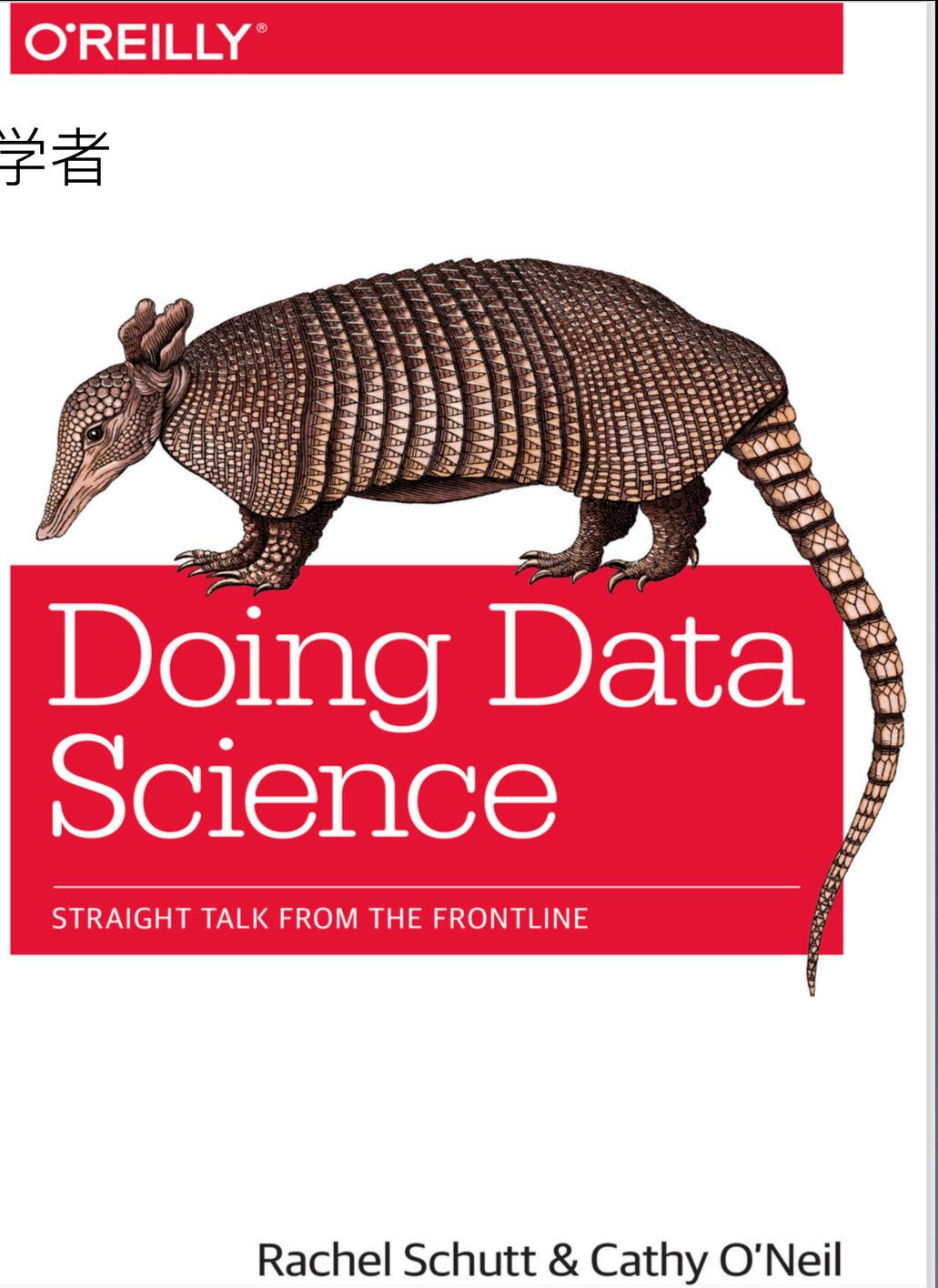
计算机技术与工具
数学、统计学
思维、意识
视野、精神
表达、沟通
得有一颗好奇、执着的❤️
法律、道德
行业、领域专门知识
图片来源：视觉中国 www.vcg.com



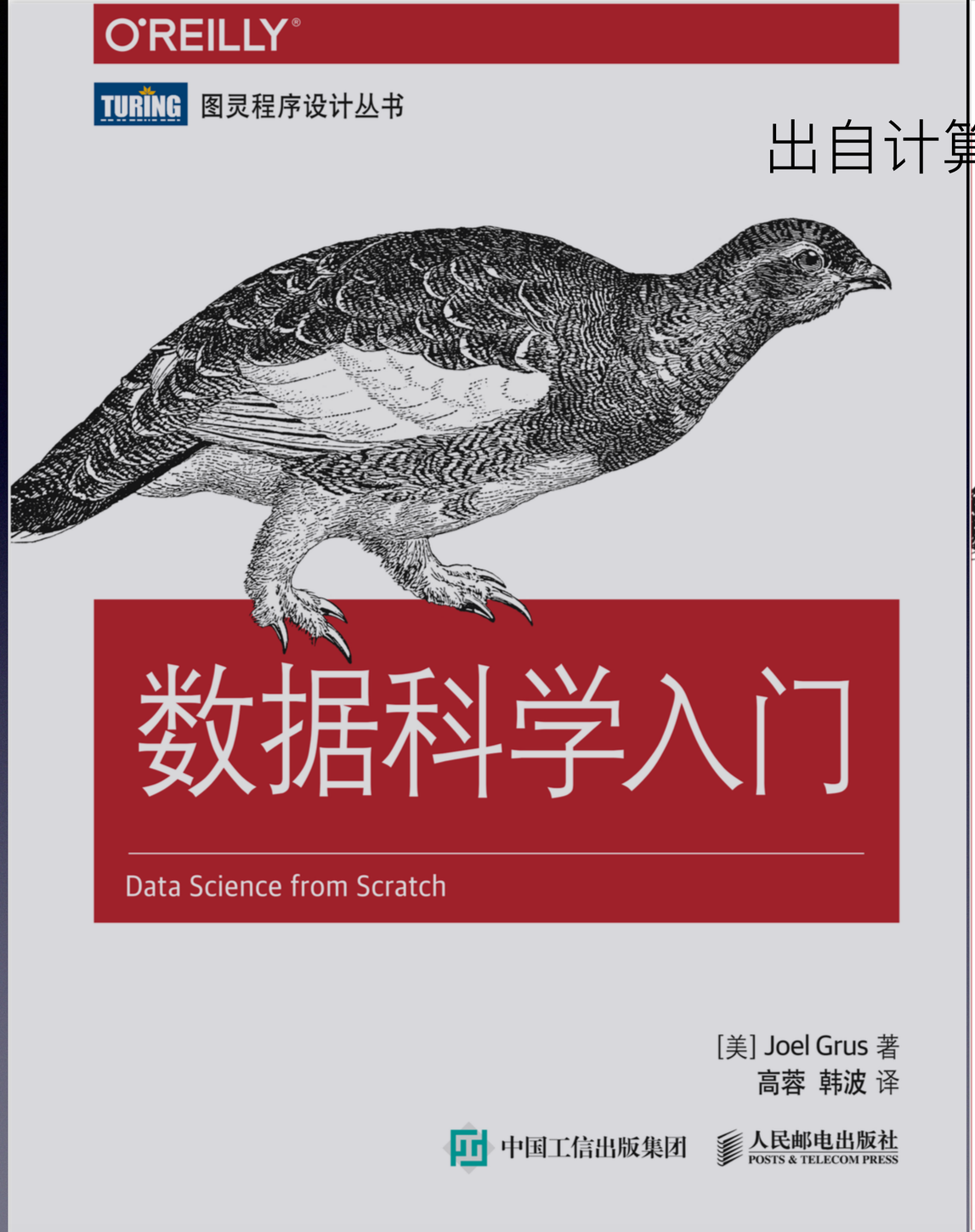
知名 IT 咨询公司 Gartner 7 月发布了2017年新兴技术成熟度曲线

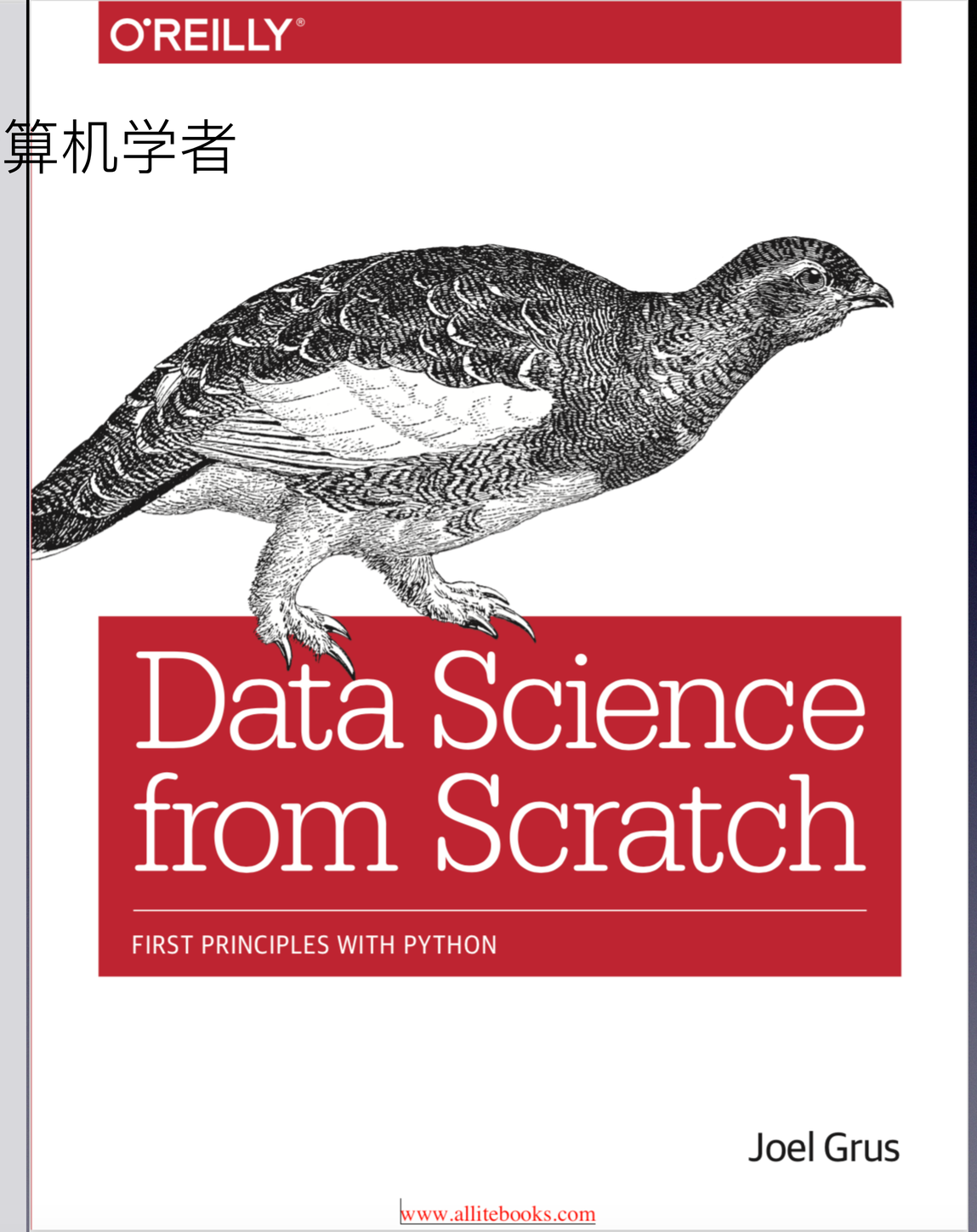
出自统计学者


O'REILLY
TURING 图灵程序设计丛书
数据科学实战
DOING DATA SCIENCE
[美] Rachel Schutt Cathy O'Neil 著
冯凌秉 王群锋 译
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS


O'REILLY
Doing Data Science
STRAIGHT TALK FROM THE FRONTLINE
Rachel Schutt & Cathy O'Neil

出自计算机学者





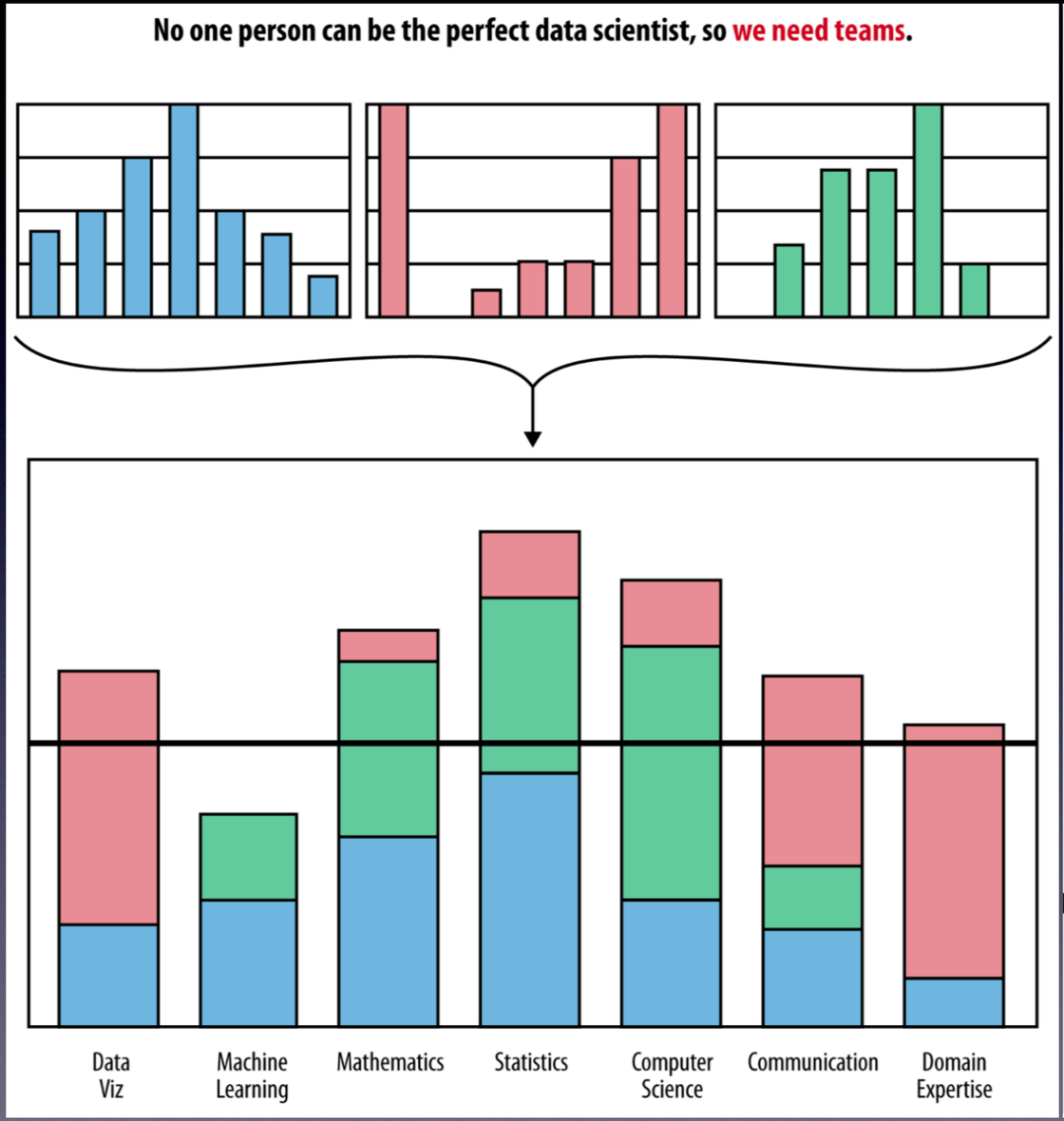
www.allitebooks.com

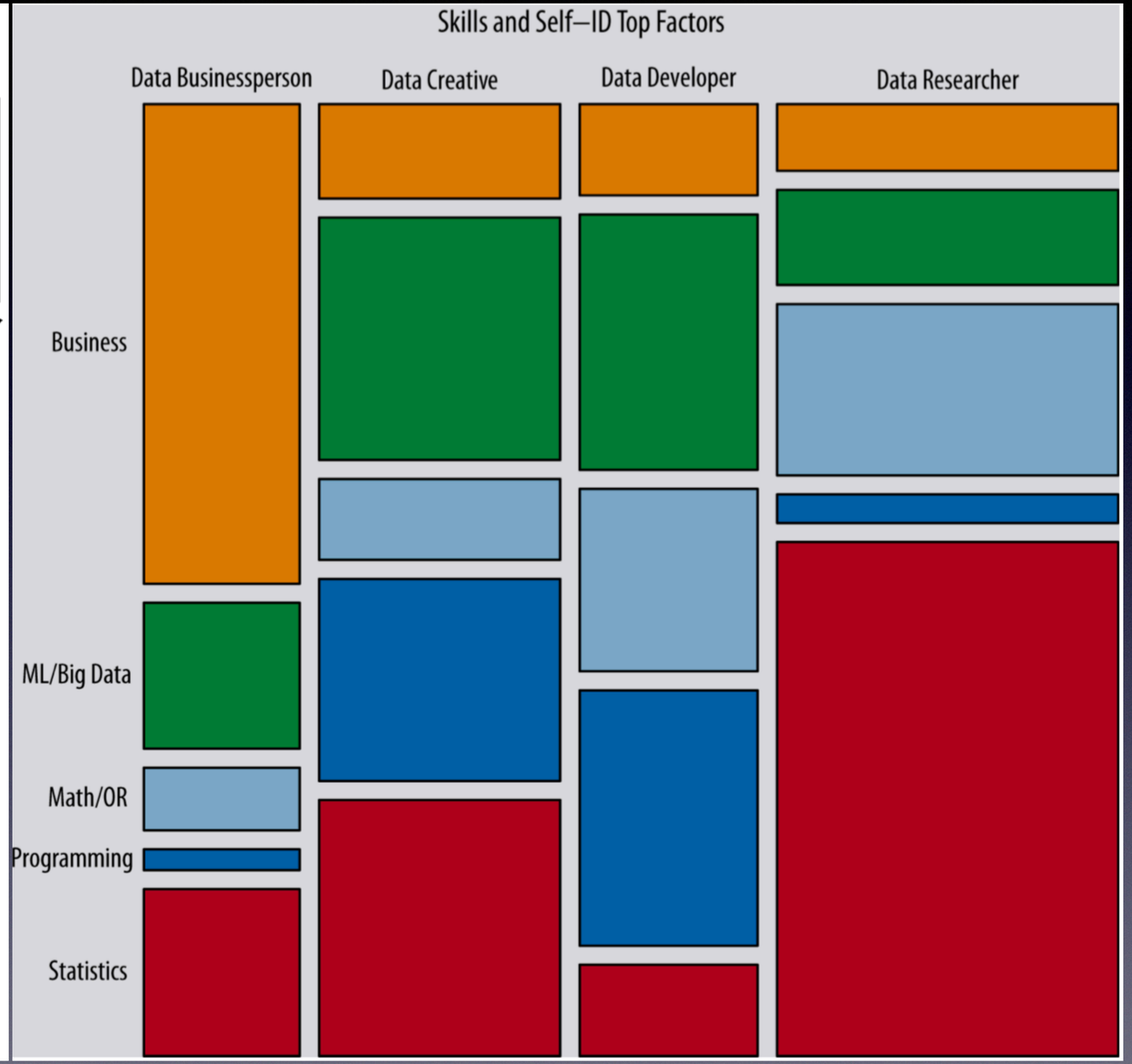
What and How Should We Do?

- How?

未来的智能工作空间

借喻大数据学者的工作模式


No one person can be the perfect data scientist, so we need teams.
Data Viz
Machine Learning
Mathematics
Statistics
Computer Science
Communication
Domain Expertise


Skills and Self–ID Top Factors
Data Businessperson
Data Creative
Data Developer
Data Researcher
Business
ML/Big Data
Math/OR
Programming
Statistics

What and How Should We Do?

买菜	数据
洗菜	数据清洗
切菜	数据整理、加工
找到对应的菜谱	逻辑分析、框架、算法
烹饪	使用相应的计算机语言和软件分析数据
盛到合适的器皿里	数据可视化
端到食客面前	数据报告

……

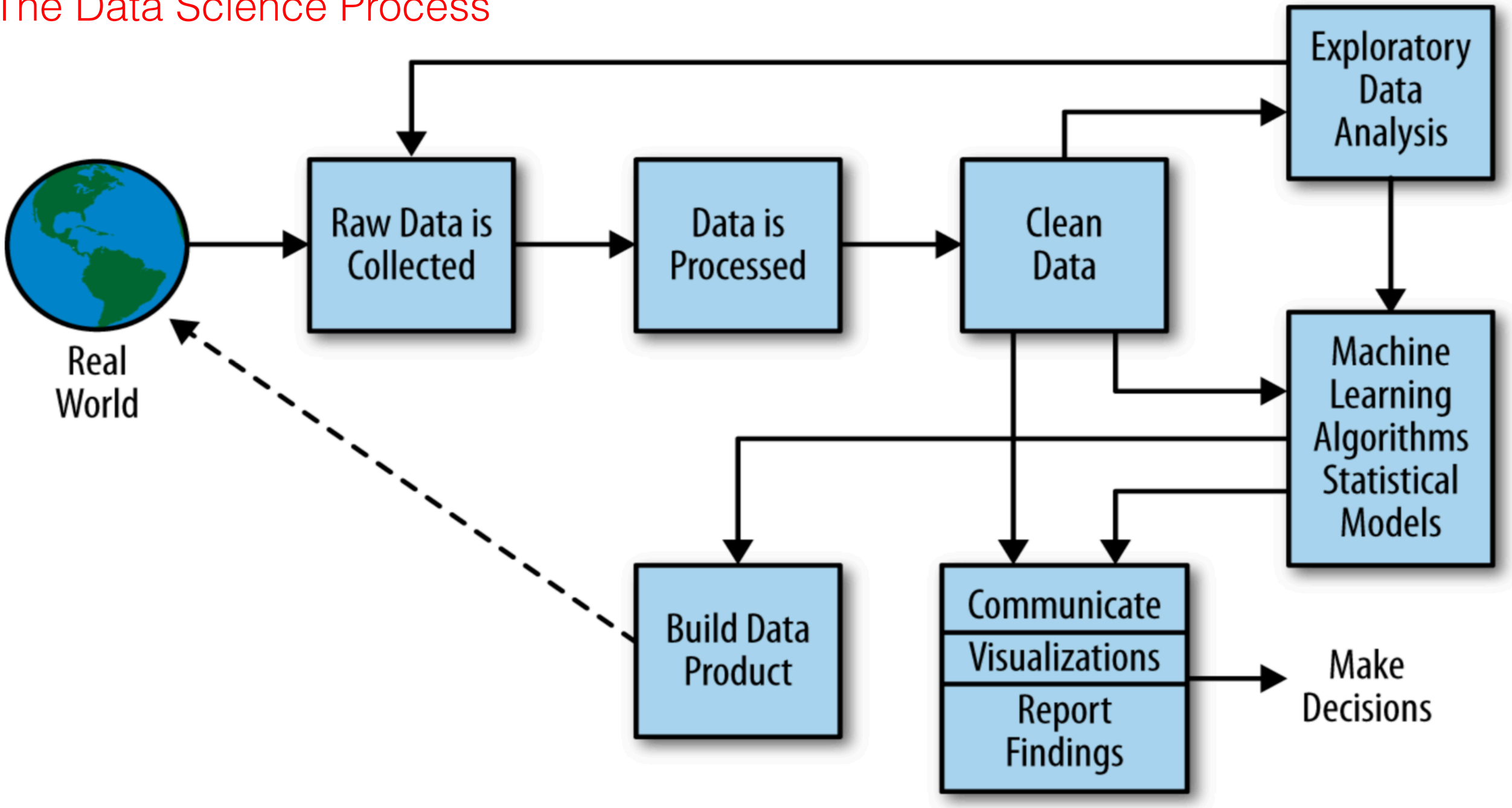
厨师给我们什么启示?

What and How Should We Do?

我们的问题

食材、炊具都有了，但却无法做出一道色香味俱佳的菜……

The Data Science Process



The Data Scientist'Role in Process

为什么？我要解决
什么样的研究课题?

问问题：
该记录和收集
什么样的数据?

形成假设

我希望它看起来
是什么样的?

收集原始数据

处理数据

清理数据

人类行为
生物学
金融
互联网
制药
社会学
奥运会

– 电子邮件
– 日志
– 药品记录
– 调查
– 抽血
– 奥运记录
–《纽约时报》网页（文章）

– 管道
– 网络爬虫
– 清理
– 整理
– 组合
– 处理

– 清理
– 奇异值
– 缺失值
– 调试
– 表

What and How Should We Do?

要学会编剧本，讲故事……

What and How Should We Do?

英文——英文——英文
百度——百度——百度
官网——官网——官网

重要的事情说三遍😀

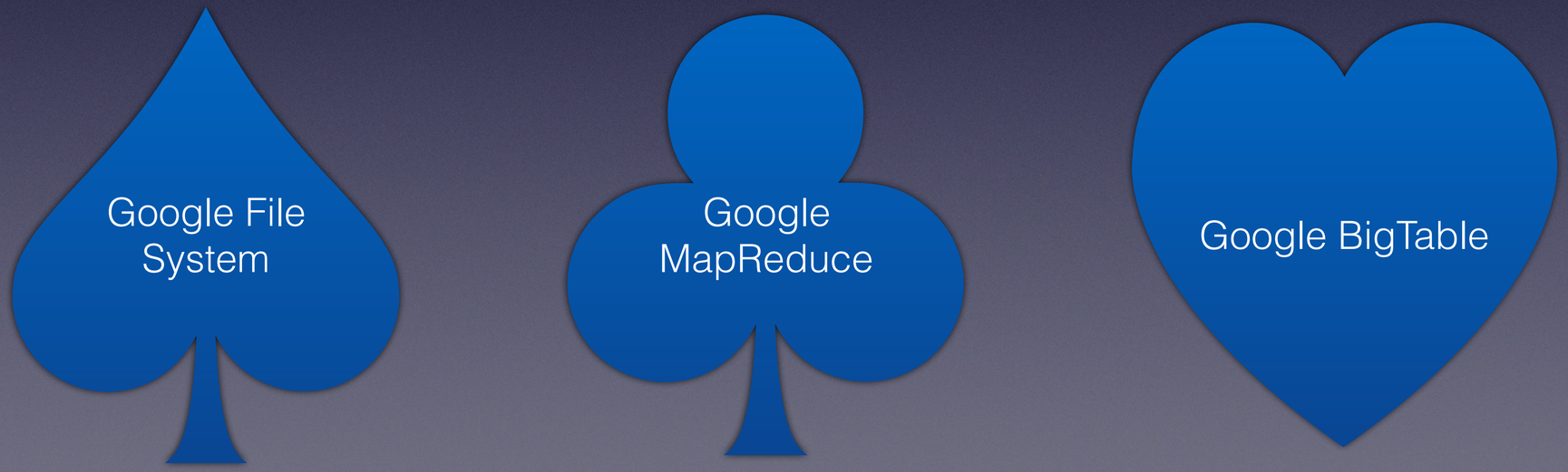
Big Data Technology Ecosphere

大数据技术生态其实是一个JiangHu江湖……

Big Data Technology Ecosphere

在一个夜黑风高的夜晚，江湖第一大帮会Google三本阵法修炼秘籍流出，大数据江湖从此纷争四起……

这三本秘籍分别为：



Big Data Technology Ecosphere

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

Google*

Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber

{fay,jeff,sanjay,wilsonh,kerr,m3b,tushar,fikes,gruber}@google.com

Google, Inc.

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

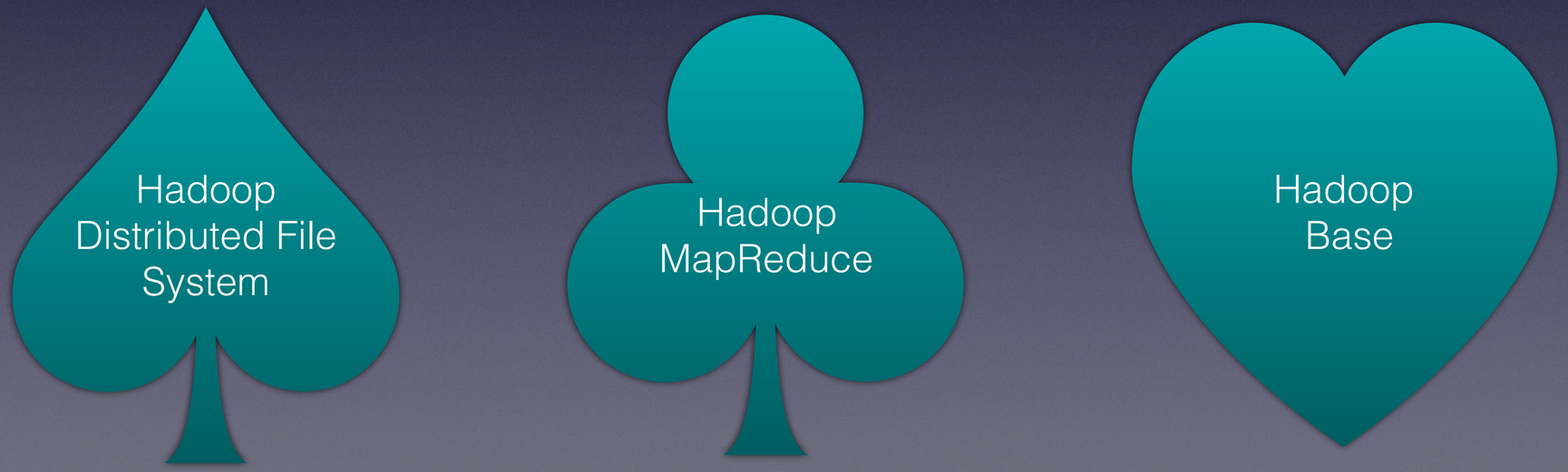
jeff@google.com, sanjay@google.com

Google, Inc.

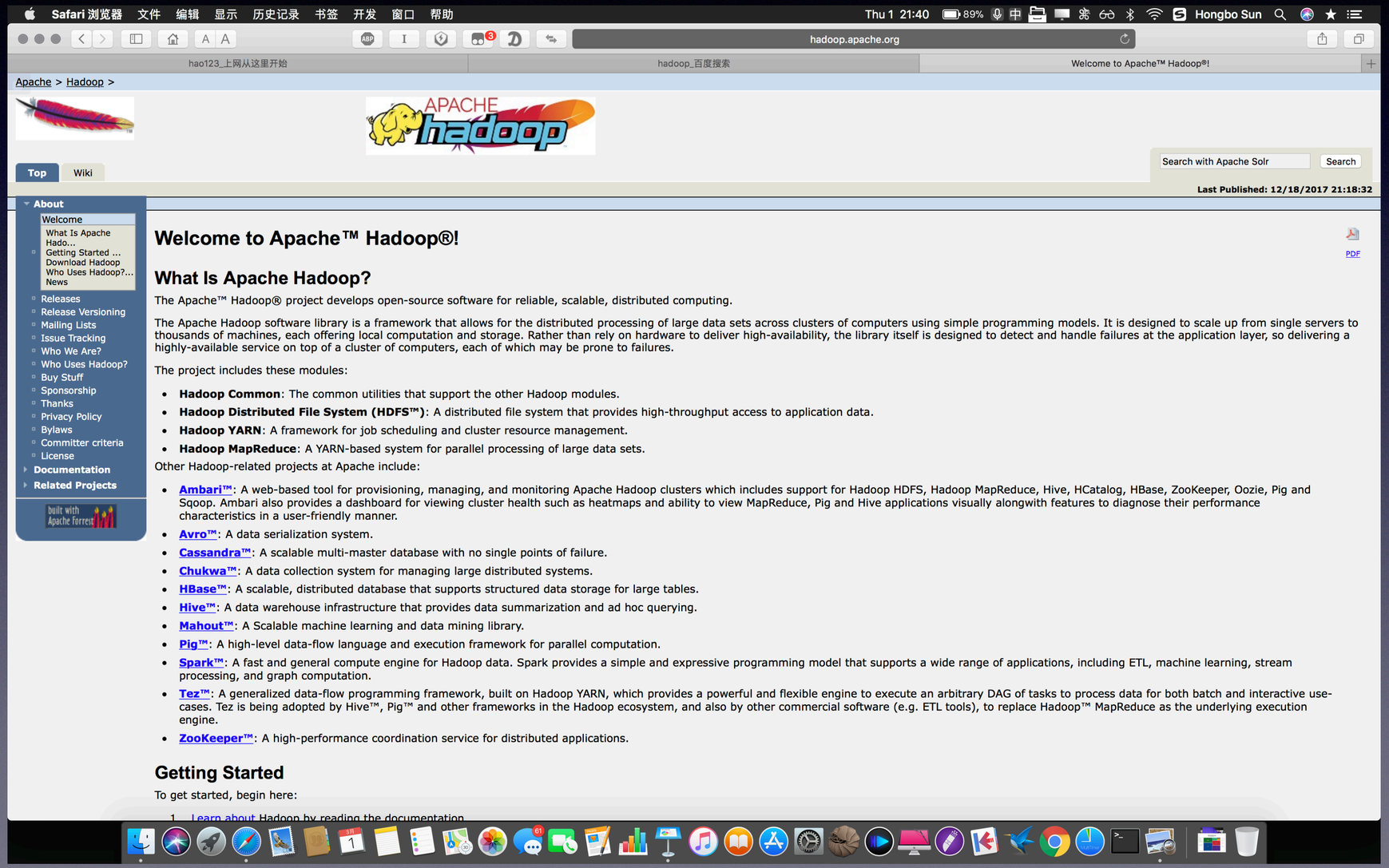
2003年到2004年间，Google发表了MapReduce、Google File System（GFS）和BigTable三篇技术论文，提出了一套全新的分布式计算理论 和技术框架，奠定了大数据的技术基础。

Big Data Technology Ecosphere

在Google三大秘籍流出之后，江湖上，致力于武学开放的Apache根据这三本秘籍分别研究出了对应的武学巨著《hadoop》，并开放给各大门派研习，Hadoop包括三大部分，分别是Hadoop Distributed File System（HDFS）、Hadoop MapReduce和Hadoop Base（HBase）：



Big Data Technology Ecosphere



Big Data Technology Ecosphere

之后，在各大门派的支持下，Hadoop不断衍生和进化各种拥有独门绝技的分支流派，其中独门绝技争斗最激烈的当属计算技术，其次是查询技术，存储技术基本无太多变化，HDFS无人能憾，雄霸天下。

接下来……

Big Data Technology Ecosphere

- 传统数据仓库派说你MapReduce修炼太复杂，老子不会编程，老子以前用SQL吃遍天下，为了将这拨人收入门下，并降低大数据修炼难度，遂出了HIVE，pig、impla等SQL on Hadoop的简易修炼秘籍；
- 伯克利派说你MapReduce只重招数，内力无法施展，且不同的场景需要修炼不同的技术，太过复杂，于是推出基于内力（内存）的Spark，意图解决所有大数据计算问题；
- 流式计算相关门派说你Hadoop只能憋大招（批量计算），太麻烦，于是出了SparkStreaming、Storm，S4等流式计算技术，能够实现数据一来就即时计算；
- Apache看各大门派纷争四起，推出flink，想一统流式计算和批量计算，大有一统江湖之势……

Big Data Basic Ecosystem

- 计算模式
  - 集中式
  - 分布式
  - 网格式
  - 云计算
- 主流计算框架 Google
  - MapReduce — 分布式计算框架
  - GFS — 分布式文件系统
  - BigTable — 基于 GFS 的数据存储系统
- 实现
- 主流计算系统 Apache
  - Hadoop MapReduce — 大规模离线并行存储计算
  - Hadoop DFS — 高可用、分布式文件系统
  - Hadoop YARN — 分布式资源调度系统
  - Hadoop Commom — 核心支持部件
- 改进
- 泛化生态系统
  - Spark — 基于Hadoop改良的分布式计算系统，离线内存计算。可用运行于Hadoop 系统，也可standalone
  - Storm — 基于 Hadoop改进的分布式计算系统，在线实时计算
  - Cassandra — 对 Hadoop DFS的改进
  - Hadoop Base — 基于 BigTable 的数据存储系统
- 数据存管系统
  - 传统 DB — MS-SQL，MySQL，Oracle，etc.
  - 新兴 DB
    - NoSQL — MongoDB，HBase，etc.
    - Cloud DB
      - BigTable，GoogleBase，etc.（Google）
      - RDS(Amazon)
      - Microsoft SQL Azure
      - HBase(Apache)
      - .....
- 云端集成服务系统(All in Cloud)
  - Google Cloud Plaform
  - Amazon Web Service
  - Ali Cloud，Tencent Cloud，baidu Cloud，etc.
- 操作系统
  - WinOS — 个人用户
  - MacOS — 个人用户
  - Lin OS — 企业用户
  - Virtualbox
  - Google Chrome
- 基础语言
  - Markdown — 动态报告、网页发布
  - Latex — 动态报告、论文写作
  - Python — 数据分析、一般性编程语言、极具粘合性
  - R — 专注数据分析的语言、粘合性越来越强
  - Java — Java是一门面向对象编程语言，吸收了C++的优点，摈弃其难点，简单易用，功能强大

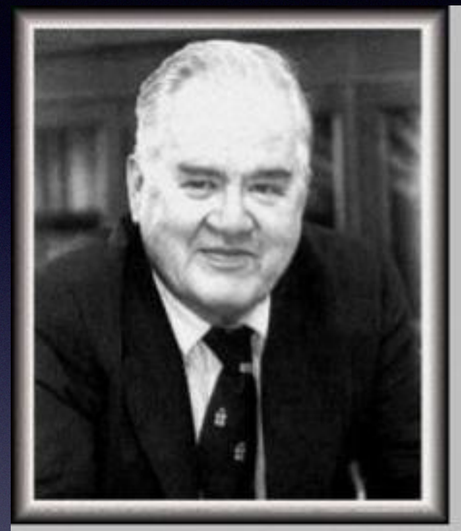
Think about Statistics

统计学怎么了?

Think about Statistics

统计学迷失了方向，与时代渐行渐远……

Think about Statistics



John Wilder Tukey

1915年6月16日-2000年7月26日

- *"An appropriate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."*

48

对正确问题的适当答案比对近似问题的确切答案更有价值.

Think about Statistics

统计学家必须摒弃 "证明真理的守护者" 的作用, 并抵制试图提供一劳永逸的解决办法和整洁的统一的问题。

——John Wilder Tukey

Think about Statistics

“很长一段时间我觉得自己是统计学家，对统计推断情有独钟，将从小样本上研究得到的结论推广到更大的群体。但随着数理统计的发展，我越发觉得这个路数不大对……总的来说，我觉得自己感兴趣的是数据分析，它包括：分析数据的过程，解释该过程得到结果的技术，合理计划收集数据的方案使得之后的分析过程更方便准确，以及所有分析中需要用到的仪器和数学理论。”

——“The Future of Data Analysis”，John Wilder Tukey，1962

概括就是：仅仅研究数学理论不是数据科学，数据科学的内容涵盖更广。

Think about Statistics

大数据时代的到来对统计学提出了全新挑战：

- 统计推断的核心失去了前提；
- 面对结构复杂、类型多样的海量数据，巨大计算量和存储量使其统计模型、方法难以承受；
- 比较注重理论研究，对应用重视不够。

Think about Statistics

大数据时代统计学的前途：

- 大数据分析的很多根本性问题和小数据分析并没有本质区别；
- 依然存在相当多的问题，并不属大数据问题，是大数据解决不了的；
- 统计学依然是数据分析的灵魂，但需要改进、改良和改变。

Think about Statistics

大数据时代统计学的机遇与挑战：

统计学用数学做工具，但不是数学

统计学受到计算机科学的强烈挑战

统计学受到计算框架和算法的挑战

大数据时代统计学的机遇与对策：

反思统计思维与方法的改进

加强分布式、并行统计计算方法的研究

加强计算机技术的使用

Think about Statistics

需要统计人思考的问题：

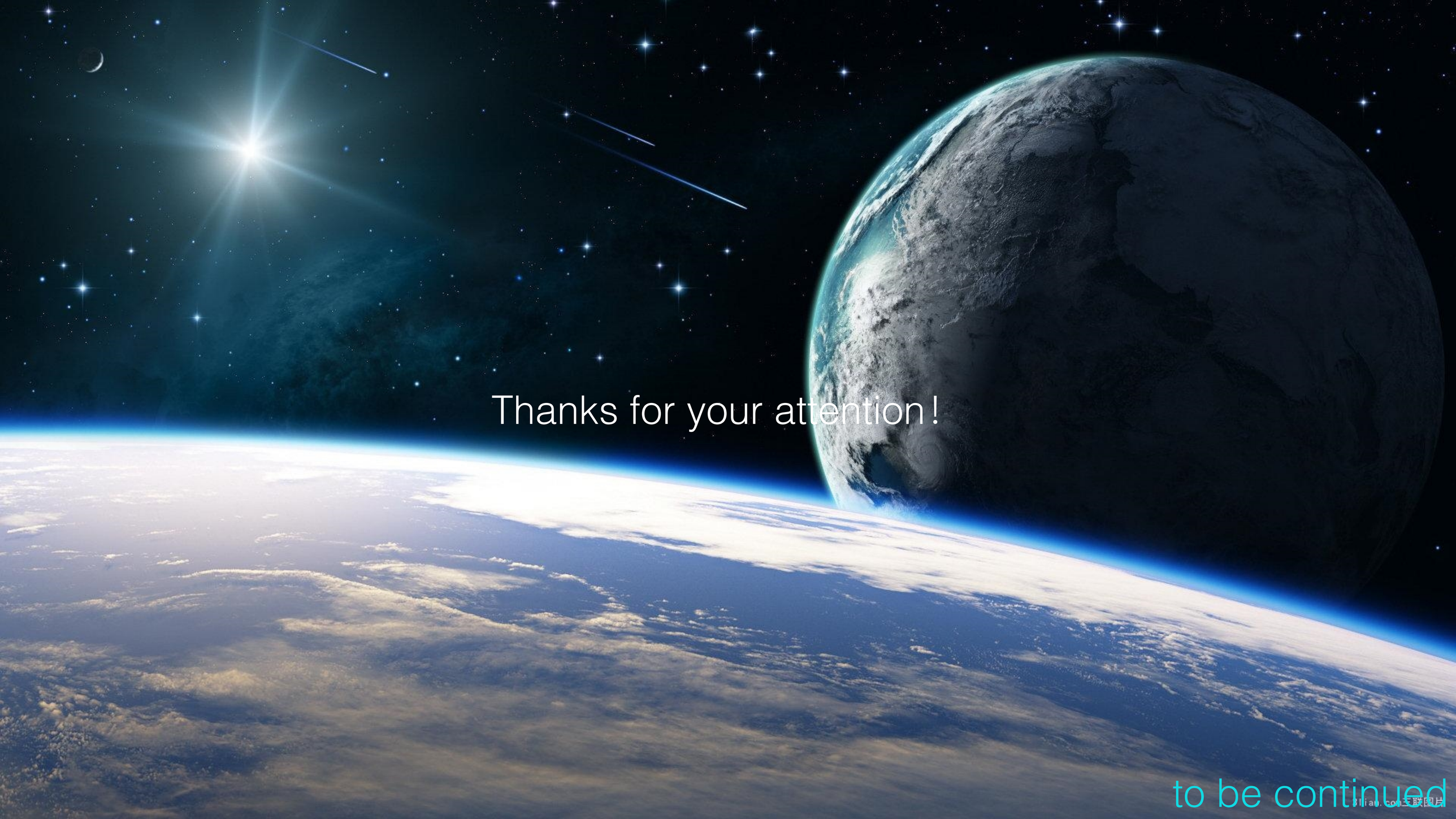
大数据时代统计学的地位、作用和角色是什么?

大数据时代统计学可能被替代吗?

Think about Statistics

需要统计人改变的问题：

让统计学回归数据分析的本质——应用驱动！


Thanks for your attention!
to be continued