The Life Blood of Analytics

Data

Doug Gray

Reference texts: *Big Data @Work* and *The AI Advantage*
Dr. Thomas Davenport
https://www2.deloitte.com/insights/us/en/topics/analytics/data-management-techniques-approaches-tools.html

Data

- Data is… capital, the new oil, the new gold, more valuable than money.
- Data is generated by activity.
- Data generates more data.
- Data is proprietary and nonproprietary.
- Data is the lifeblood flowing through an organization.

Five Key Questions for Data

1. Relevance
2. Sourcing
3. Quantity
4. Quality
5. Governance (master data management, metadata)

Source: ***Competing on Analytics: The New Science of Winning***, T. Davenport, J. Harris p. 225

Data Governance

- Data governance is 80% about communication with data communities
- Data governance tools are important and necessary, but not sufficient
- Governance is more about governing people's *behavior*, e.g., locating, interpreting, using the data
- Establishing roles like data stewards or data owners
- https://www.forbes.com/sites/charlestowersclark/2019/01/23/the-ethics-of-data-governance-data-comes-with-benefits-and-liabilities/#12c03464215a

Value of Data Increases If It Is

- Correct (accurate)
- Complete
- Current (timely)
- Consistent [one version of the truth; in general (fact)]
- Context [one version of the truth; in context (semantic)]
- Controlled (integrity)
- Analyzed! (data alone is useless)

Source: ***Competing on Analytics: The New Science of Winning***, T. Davenport, J. Harris p. 231

Data and Competition

- Data, used properly, is a means to competitive advantage
- Data is the *enabler* of digital age businesses, e.g., Amazon, Netflix
- Data can *transform* legacy (analog) businesses, e.g., Walmart, GE

Data Sources

- Legacy systems and applications
- Clickstream data
- Third party data
- Digital sensors cost $0.40 each
- Digital video and still camera sales are increasing exponentially
- More mobile devices active now than there are ***people on the planet***
  - ***5.5 million new mobile devices connecting to the Internet every day***

DataScience@SMU

What Is “Big Data?”

Doug Gray


BIG DATA
MOBILE
RESEARCH
SOFTWARE
SOCIAL
NETWOKS
SOLUTIONS
MULTIMEDIA
NETWORK
BUSINESS
BRANDS
SOLUTIONS
WWW
ADVERTISING
SOCIALMEDIA
SERVICES
BIGDATA
INTERNET
VISION
ENGINEERING
B2B TARGET
BRAND
COM
BEHAVIOUR
STRATEGY
WORLDWIDE
WEB SERVICES
BIG
DATA
PROJECTS
SERVICE
BUZZ
INFORMATION
PLANNING
ORGANIZATION
WEB DEV
MEDIA
PROPRICING
DIGITAL
PROMOTION
CODING
E-MARKETING
SEGMENTATION
PLAN
COMMUNICATION
SOCIAL NETWORKS
TARGET
CONSUMER
CONTENT
ORGANZATION
PLANNING
EVENTS
PROGRAMMING
CODING
SOFTWARE
BRANDING
COMPUTER
CONSUMER DEMAND
MARKETS
WEB MARKETING
SOCIAL
STATISTICS
APPS

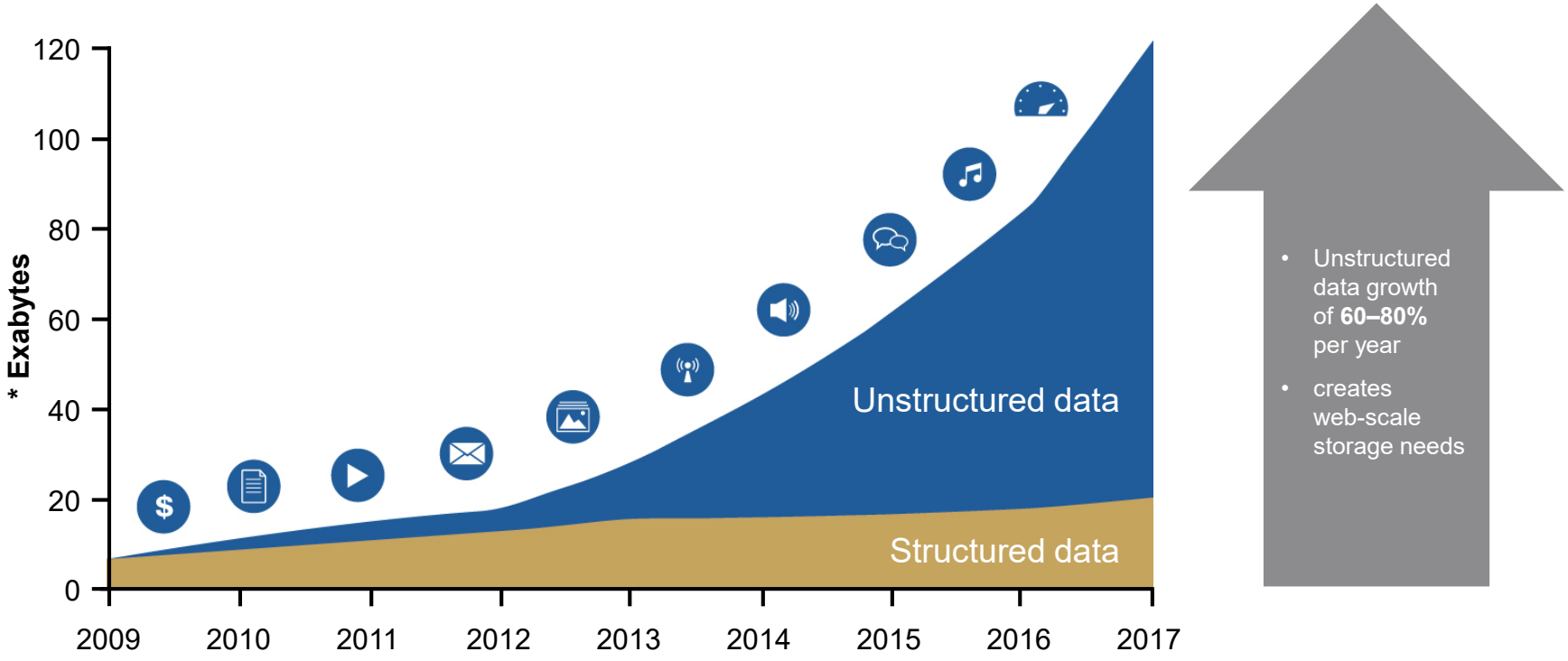
What Is Big Data?

- Three Vs of big data: ***volume, variety, velocity***
- Unstructured data: text, audio, video, images, alphanumeric

https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#54be10b360ba

Data Growth

Problem: traditional and legacy storage designed for transactional, not unstructured data



*1 exabyte = 1,000 petabytes = 1 million terabytes = 1 billion gigabytes **Source**iDC

What Is Big Data?

- Three Vs of big data: ***volume, variety, velocity***
- Unstructured data: text, audio, video, images, alphanumeric
- Large amounts of data measured in 100 terabytes, petabytes
- Constant flow of data
- Analyzed using automated machine learning, e.g., H2O.ai (ML at scale)
- Primarily used for data-based products

https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#54be10b360ba

Examples of Big Data?

- Social media data (natural language)
- Clickstream data (web activity and transaction data)
- Sensor data (engines, machines)
- Medical records data (test results, medical terminology)
- Financial, banking records
- Video, audio, images, computer system log data, phone records

Applications of Big Data?

- Social media data (natural language)
  - Customer sentiment analysis, e.g., 737 Max 8
- Clickstream data (web activity and transaction data)
  - Customer buying behavior ,e.g., Adobe Omniture, Google Analytics
- Sensor data (engines, machines)
  - Predictive maintenance, e.g., GE Predix
- Medical records data (test results, medical terminology)
  - Diagnosis, evidence-based medicine, e.g., Qure.ai
- Financial, banking records
  - Fraud detection, money laundering, e.g., London Whale
- Video, audio, images, computer system log data, phone records
  - Predicting terrorist attacks or hacking/malware behaviors, e.g., any three letter agency: NSA, CIA, FBI…

DataScience@SMU

Architecture with “Big” Data

Doug Gray

Big Data IT Architecture

The cloud and data lake and the enterprise data warehouse

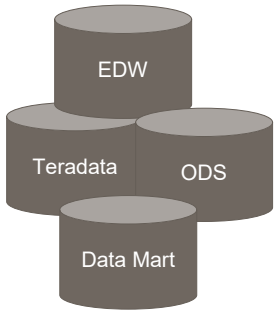
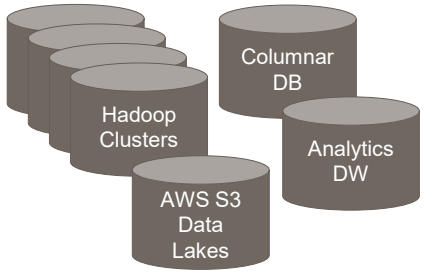
Big Data Analytics Architecture

- Data management
- ETL (extract, transform, load) tools and processes or now ELT (Ab Initio, Informatica)
- Repositories (Teradata, AWS, Tibco, Collibra)
  - Data warehouse, data lake, data mart, MDM Catalog, metadata library
- Analytical tools and applications (Alteryx, R, Xpress)
- Data visualization tools (Tableau, Qlik, Microstrategy); mobile-enabled
- Deployment processes (SDLC, ADLC, Agile, SAFe)

Source: ***Competing on Analytics: The New Science of Winning***, T. Davenport, J. Harris, pp. 46-47, DELTTAA Model Analytical Capabilities by Stage, p. 222

Big Data Technologies

- AWS S3 (simple storage service), formerly Hadoop, MapReduce, and Spark
  - Massively parallel processing, cloud-based commodity servers
  - Cloudera, HortonWorks, Pivotal
- Scripting languages
  - Python, Hive (SQL-like interface to HDP)
- Apache Projects
  - Mahout-Samsara
  - Pig

Classic Data Platforms	Emerging Data Platforms
	

Classic Analytics	Emerging Analytics
Statistics	Map reduce
Stochastics	Apache tools
Forecasting	Data mining
Simulation	Machine learning
Math programming	Text analytics
Econometrics	Pattern recognition

Industry trend is for classic and emerging data platforms, and analytics methods and technologies to co-exist and evolve side-by-side in an integrated manner without one necessarily replacing or outmoding the others.

Source: big data @work, T. Davenport, 2014.

Data Lake Overview

Increase speed at which information is curated, added to the platform, and access is provided!

- **Data storage**
  - Collects everything for longer periods of Time
  - Increases the durability of your data
- **Data catalog**
  - Lets you search and dive in anywhere
- **Access controls**
  - Flexible access
- **Charge storage costs to owner**
  - Enhanced data ownership
- **Streaming and real-time analysis**
  - Can be the target for a streaming data platform

Data Lake Overview

Increase speed at which information is curated, added to the platform, and access is provided!

- **Democratize**
  - Data access to accelerate more insights
- **Collecting and store**
  - Any data at scale and at low costs
- **Securing and protecting**
  - All of data stored in the central repository
- **Search**
  - Quickly search and find the relevant data
- **Easily** perform new types of data analysis
  - Using the right tool for the right job
- **Query the data**
  - Defining the data's structure at the time of use

Comparison of a Data Lake to an Enterprise Data Warehouse

Data lake

- Complementary to EDW (not replacement)
- Schema on read (no predefined schemas)
- Structured/semi-structured/unstructured data
- Fast ingestion of new data/content

Data warehouse

- Data lake can be source for EDW
- Schema on write (predefined schemas)
- Structured data only
- Time-consuming to introduce new content

Comparison of a Data Lake to an Enterprise Data Warehouse

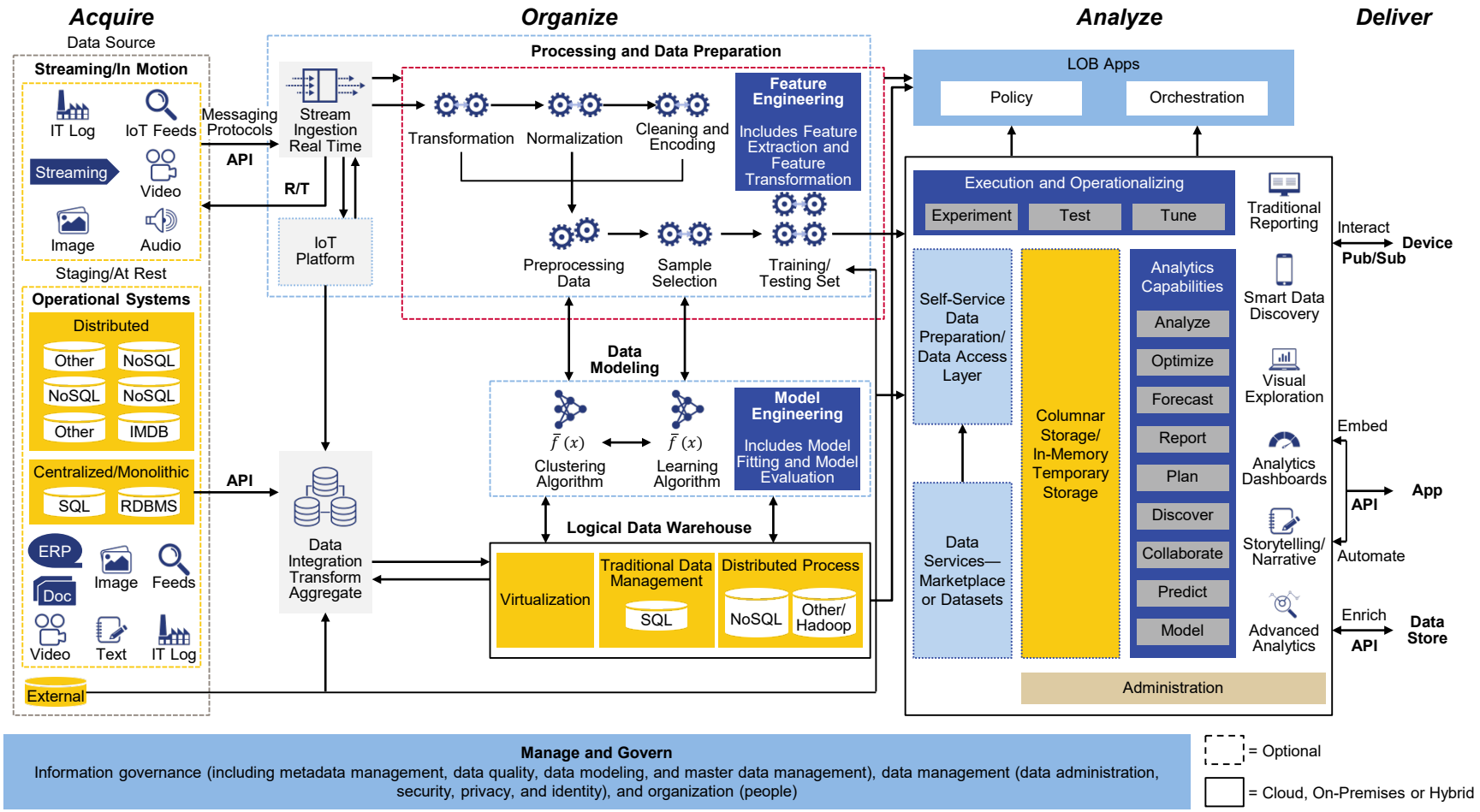
Data lake

- Data science and prediction/advanced analytics and BI use cases
- Data at low level of detail/granularity
- Loosely defined SLAs
- Flexibility in tools (open source/tools for advanced analytics)

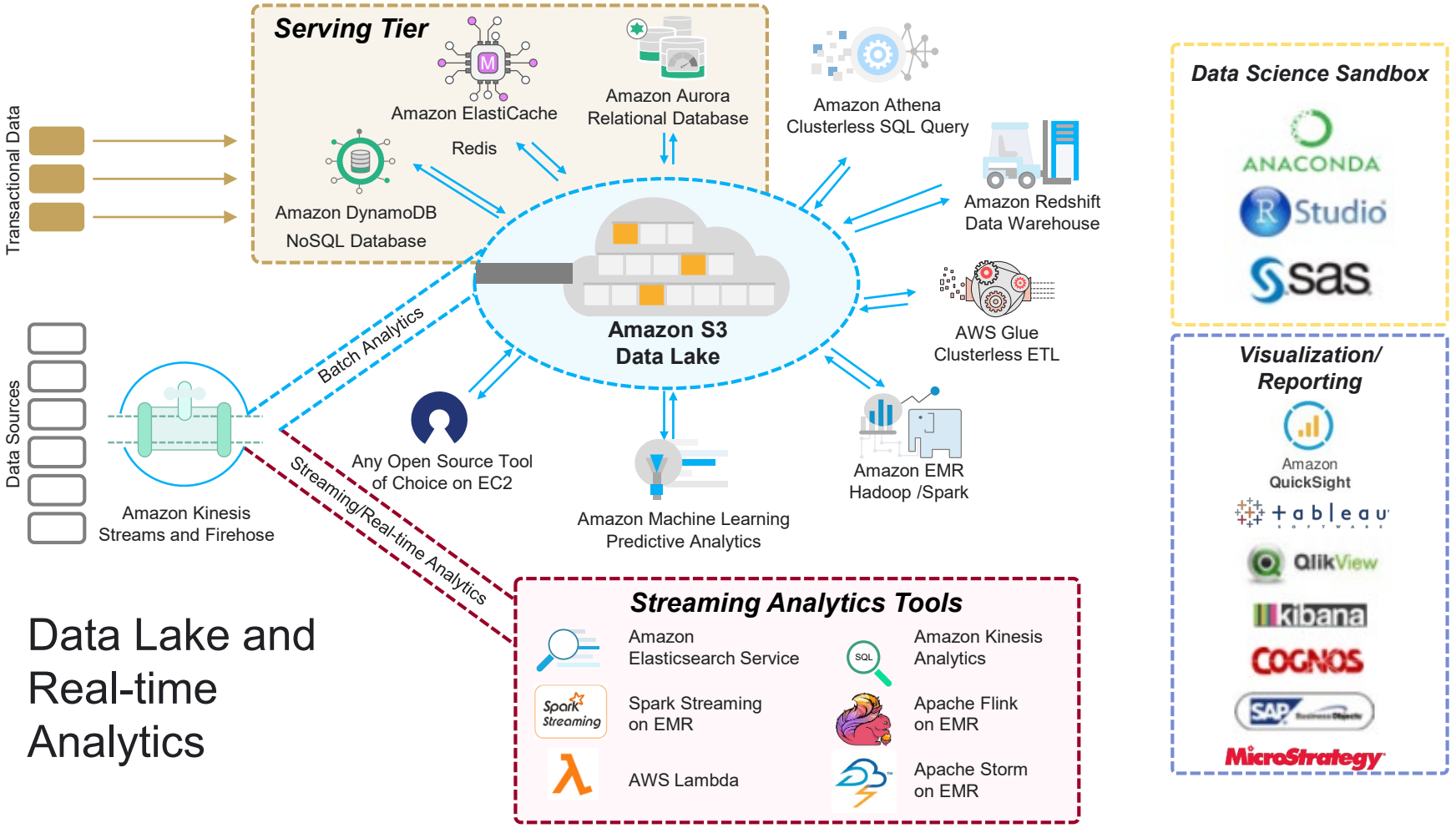
Data warehouse

- BI use cases only (no prediction/advanced analytics)
- Data at summary/aggregated level of detail
- Tight SLAs (production schedules)
- Limited flexibility in tools (SQL only)

Gartner Reference Architecture



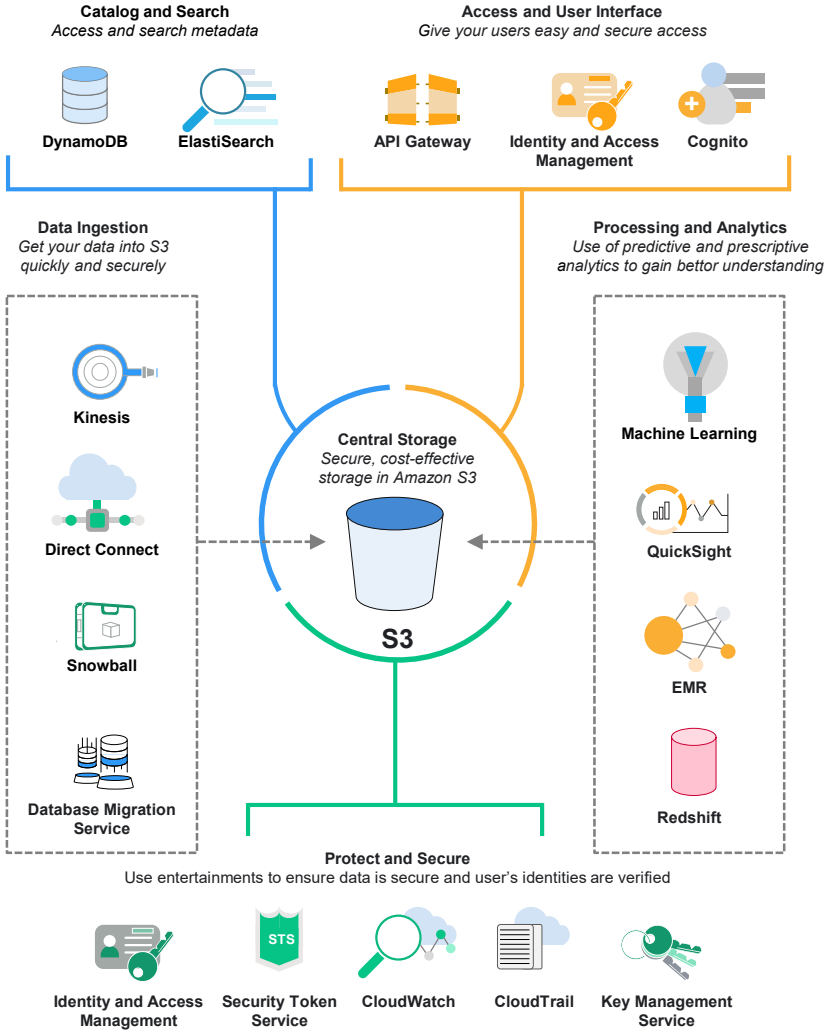
AWS: Reference Architecture



Data Lake and Real-time Analytics

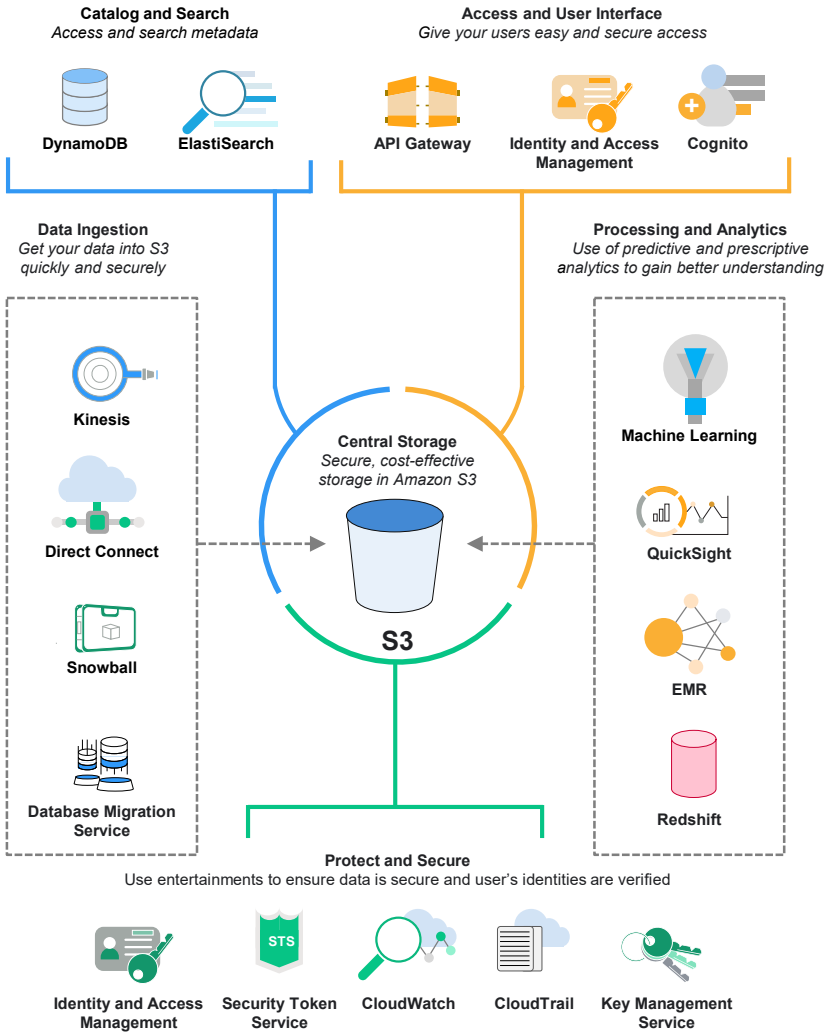
AWS Professional Services Offerings

- Data ingestion
  - IoT workshop, assessment, and accelerator
  - Ingestion accelerator
- Catalog, search, and storage
  - Data lake accelerator
  - Data warehousing accelerator
  - RDBMS migration accelerator
- Protect and secure access
  - Big data security assessment



AWS Professional Services Offerings

- Processing analytics
  - EMR accelerator
  - ETL accelerator
  - Data science workshop
  - Agile analytics accelerator
- Automation, operations integration
  - Security accelerator
  - DevOps and automation workshop and accelerator
  - Operations accelerator



Financial Implications of Big Data Tech

Metric	Big data Hadoop MPP	Data warehouse
Cumulative 3-year CF	$152 million	$53 million
NPV	$138 million	$46 million
IRR	524%	74%
Breakeven	4 months	26 months

DataScience@SMU

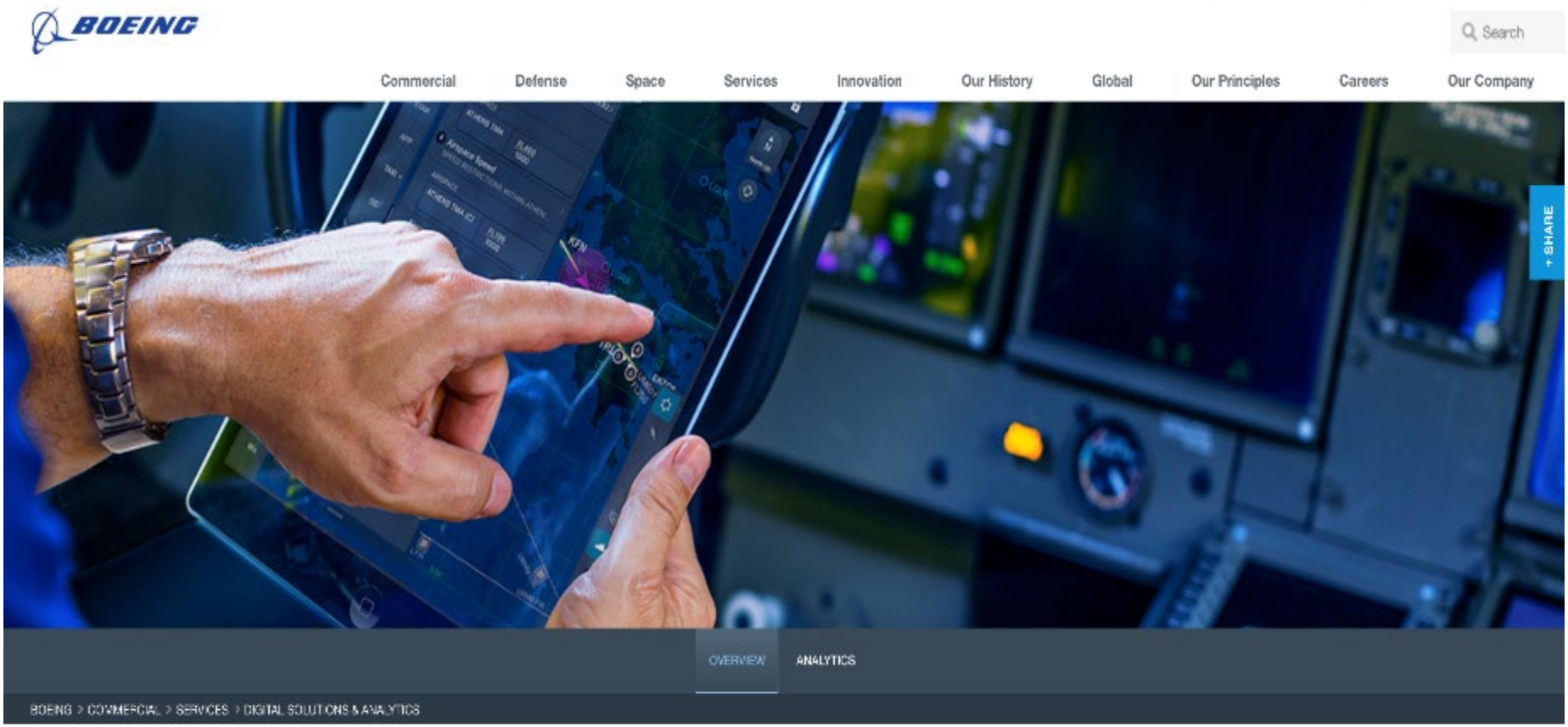
Mini Case Studies

Doug Gray

Adobe Web Site Clickstream Data

Web Site A/B Testing Using a Statistical Test

Boeing Digital GE Aviation
Digital Aircraft Sensor Data



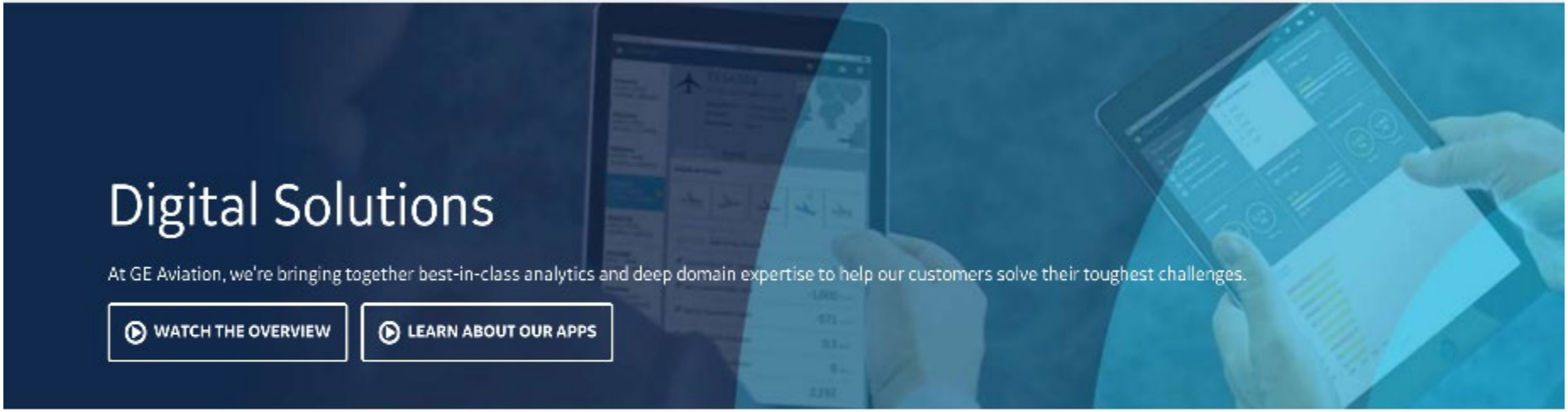
Digital Solutions & Analytics

GE Aviation

CUSTOMER LOGIN

Search

COMMERCIAL MILITARY B&GA DIGITAL MARINE COMPANY CAREERS PRESS BLOG CUSTOMER SUPPORT



Digital Solutions

At GE Aviation, we're bringing together best-in-class analytics and deep domain expertise to help our customers solve their toughest challenges.

WATCH THE OVERVIEW

LEARN ABOUT OUR APPS

Explore the digital solutions from GE Aviation that are best suited for your specific needs.

Start by selecting the fields from industries, outcomes and/or products that apply to you to see a personalized list of digital offerings that may help solve your toughest challenges.

IBM Watson
Medical Records Diagnosis

Ford Motor Company

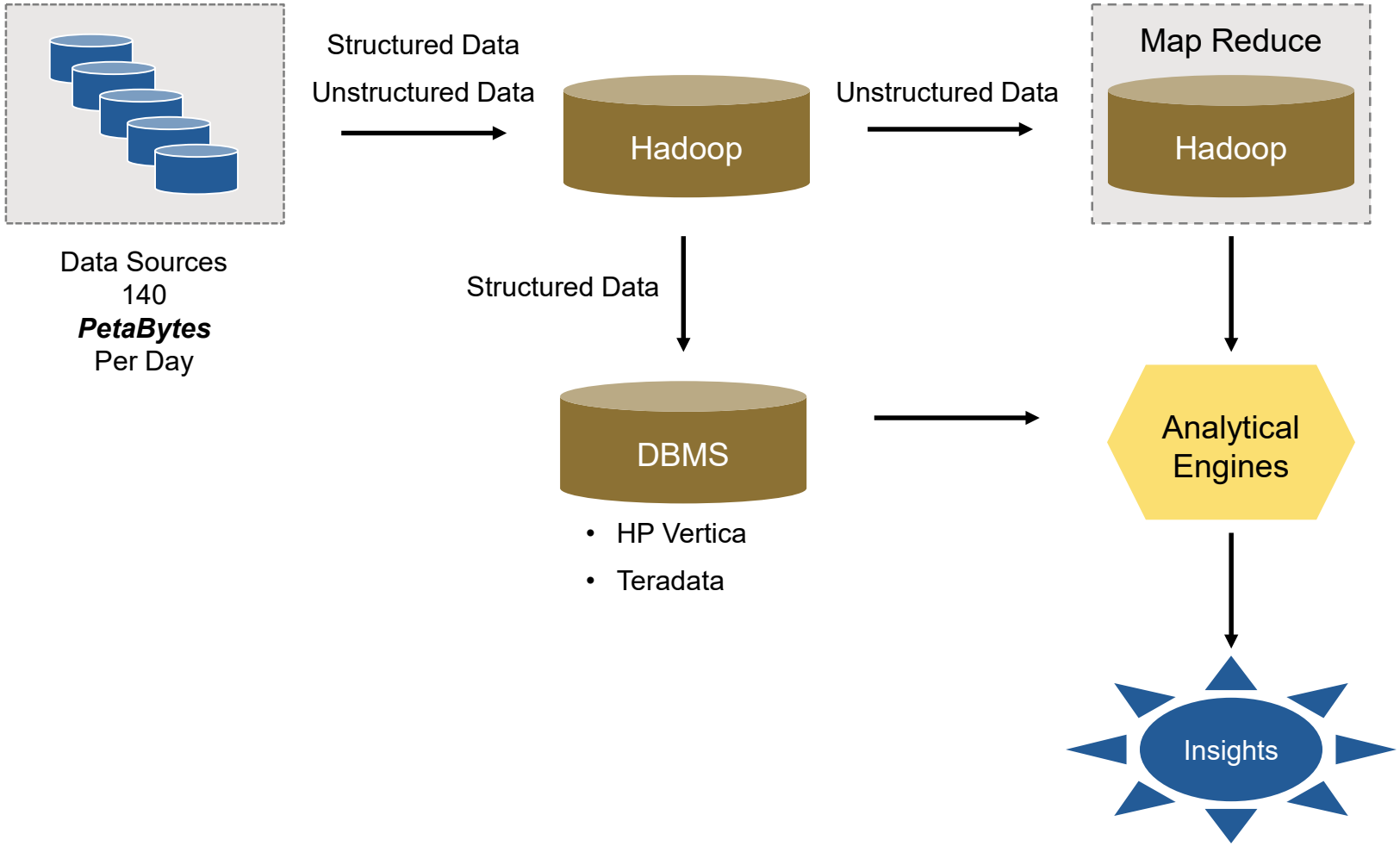
Palantir Technologies (“Gotham”) Terror Prevention

- https://www.youtube.com/watch?v=26YRBjOtR2w
- https://www.youtube.com/watch?v=pbfGzFMxmHo

AT&T Big Data Foundry Plano, TX

AT&T will be a ***data-powered*** enterprise by chairman mandate by 2020.

- Their big data foundry pulls **140 Petabytes** (1 Petabyte = *1,000 Terabytes*) off of their wire line, wireless, and satellite (DirecTV) networks every single day!



DataScience@SMU

AI in Business

Doug Gray

Artificial Intelligence (AI)

- ***Amara's law*** (futurist Roy Amara)
- *"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."*

***The AI Advantage*, Tom Davenport**
https://hbr.org/2019/03/the-ai-roles-some-companies-forget-to-fill

Definitions

- Artificial intelligence (AI)
  - The theory and development of *computer systems* able to perform tasks that normally require *human intelligence*, such as visual perception, speech recognition, decision-making, and translation between languages
- Machine learning (ML)
  1. A type of artificial intelligence that allows software applications to become more accurate in predicting outcomes without being explicitly programmed
  2. A field of computer science that uses statistical techniques to give computer systems the ability to “learn” (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed

Artificial Intelligence (AI) Technologies

- Statistical machine learning (ML)
- Neural networks and deep learning
- Natural language processing (NLP): semantic (domain specific) vs. statistical
- Voice recognition
- Image recognition
- Rules-based expert systems
- Robots
- Robotic process automation (RPA)

Artificial Intelligence (AI) Applications

- Repetitive and/or dangerous task automation
- Fraud detection, money laundering using ML-based pattern recognition
- Cybersecurity intrusion and attack detection
- Personal assistant, e.g., Amazon Alexa, Echo, Apple Siri, chatbots, intelligent agents
- Decision-making
- Driverless cars
- Insurance claims handling

AI Supports Three Important Business Activities

1. Automating structured and repetitive work processes, often via robotics or RPA: robotic process automation
2. Gaining insight through extensive analysis of structured data, most often using ML: machine learning
3. Engaging with customers and employees using NLP: natural language processing chatbots, intelligent agents, and ML: machine learning

Companies and Industries Benefitting Most from AI

- Technology
  - Digital natives—**FANG: F**acebook, **A**mazon, **N**etflix, **G**oogle
  - Cisco: digital marketing
- Healthcare
  - Disease detection, treatment, surgical robotics
- Manufacturing: automotive
  - GE: predictive maintenance on aircraft engines, turbines, windmills
  - Toyota: robots
- Airlines
  - Southwest: real-time decision-making in irregular operations recovery

Companies and Industries Benefitting Most from AI

- Financial services
  - Vanguard: Personal advisor services (PAS) robotic investment advisor
  - Bank of America: Erica
- Life sciences
  - Pfizer: new drug designs and trials and patient drug treatment regimens
- Retail
  - Macy's: mobile and web site app shopping assistant
  - Levi's: virtual stylist
  - Lowe's: LoweBot robot aisle navigator
- Agriculture and farming
  - Monsanto: digital optimized planting
  - https://monsanto.com/innovations/modern-agriculture/articles/digital-farming-technology-around-world/

How Google Uses AI

- Understand images in Google Photos
- Enable Waymo cars to recognize and distinguish objects safely
- Significantly improve sound and camera quality in our hardware
- Understand and produce speech for Google Home
- Translate over 100 languages in Google Translate
- Caption over 1 billion YouTube videos in 10 languages
- Improve the efficiency of our data centers
- Suggest short replies to emails
- Help doctors diagnose diseases, such as diabetic retinopathy
  - See also Qure.ai: AI-based radiology
- Discover new planetary systems
- Create better neural networks (AutoML)
- …and much, much more!

https://www.cnbc.com/2018/04/27/alphabet-founders-letter-2017-sergey-brin-on-ai.html

Key Takeaways

- ***Augmentation, not automation***
  - Job elimination may be a side effect of AI, but is not usually the primary objective.
- ***Get rich slow***
  - Invest steadily in AI over time, avoid the hype and trough of disillusionment.
  - Match business problems that matter most economically to AI.
  - Take the long view.
  - Slow and steady will win the race on AI.

Key Takeaways

- ***Digital twins of analog world entities***
  - AI models embedded in processes and systems, as with analytics, generates the most value
  - AI fades into the background and you barely know it is there
- ***"This s*** is still hard (to do)"***
  - Data, cognitive technologies (software, hardware), systems integration, and qualified experts to build, deploy, operate solutions
  - Experiment, fail fast, learn, grow and move on; find reliable partners

DataScience@SMU