The shortest path to data science excellence

Louis Vainqueur
l.vainqueur@ucl.ac.uk

University of Reading, October 2018

Agenda

1. Overview of data science and AI

2. Creating data scientists

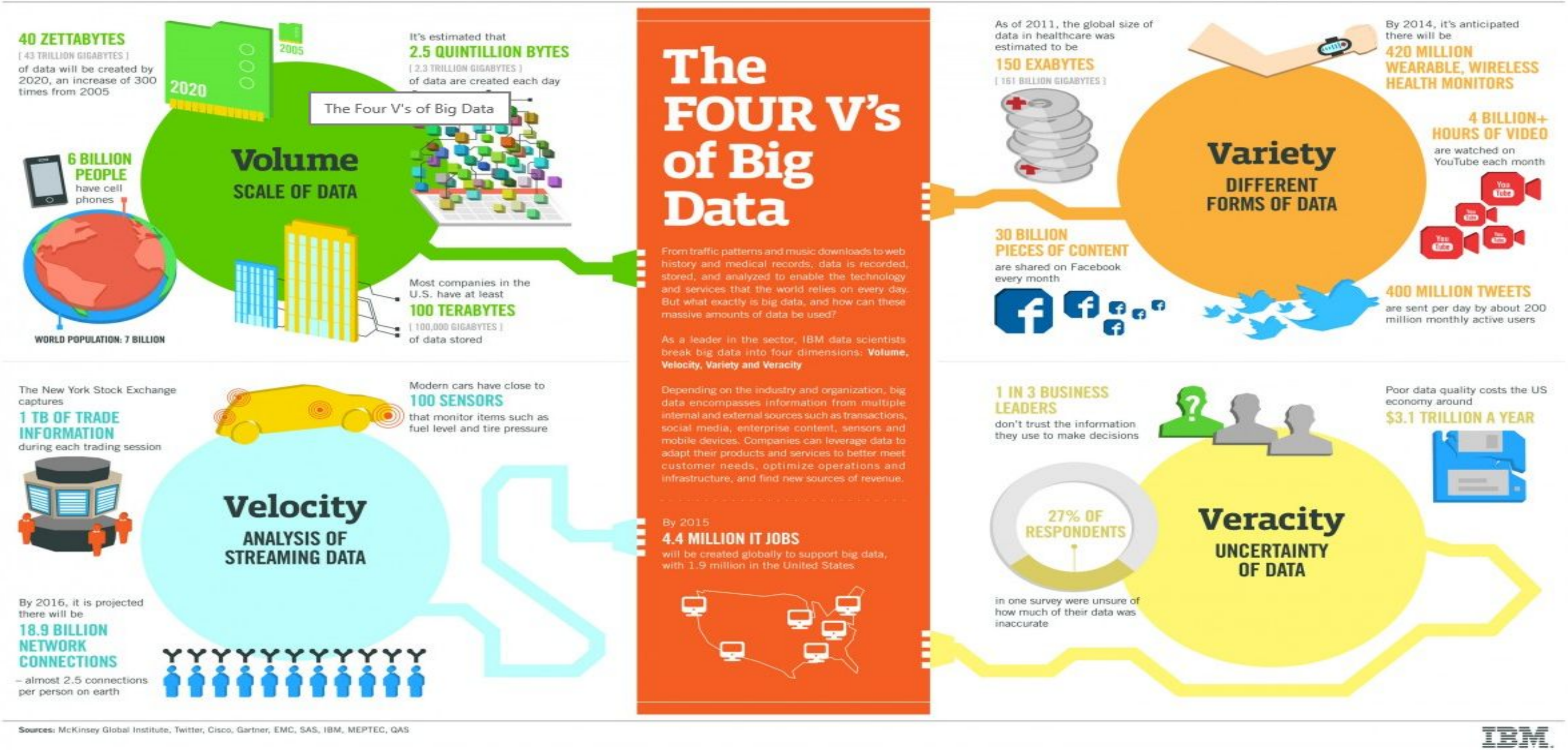
3. Recruiting data scientists

4. Aiming for data science success stories

A brief history of data science & AI

- **2002:** Doug Cutting and Mike Cafarella started an Apache project called to build an open source search engine
- **June 2003:** Successful 100-million-page demonstration of Nutch
- **Oct 2003:** Google File System paper released
- **Dec 2004:** Google MapReduce paper released
- **2005:** Cutting and Cafarella built a file system and processing framework based on concepts from Google's papers and ported Nutch on top to create the Hadoop core
- **2006:** Cutting joined Yahoo. Yahoo and Cutting spun out the storage and processing parts into an Apache project called
- **2007 to 2008:** Yahoo invests heavily in building out Hadoop
- **2008:** Cloudera, the first commercial Hadoop support company, is formed
- **2009:** Facebook develops Hive, an SQL-like framework for Hadoop
- **2011:** Yahoo spun out HortonWorks. Yahoo's Hadoop cluster has 42,000 nodes and hundreds of PT
- **2012:** Apache Hadoop v1.0 released. YARN introduced.
- **2013:** Support for running Hadoop on Windows introduced in v2.2
- **2014: Apache Spark** wins the Terrasort contest
- **2015**: Deep learning is everywhere
- **2016** : Bots, Spark, Deep learning
- **2017**: AI takes over : alpha Go
- **2018**: AlphaZero and Kubbernetes

A brief history of data science & AI



The Four V's of Big Data

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

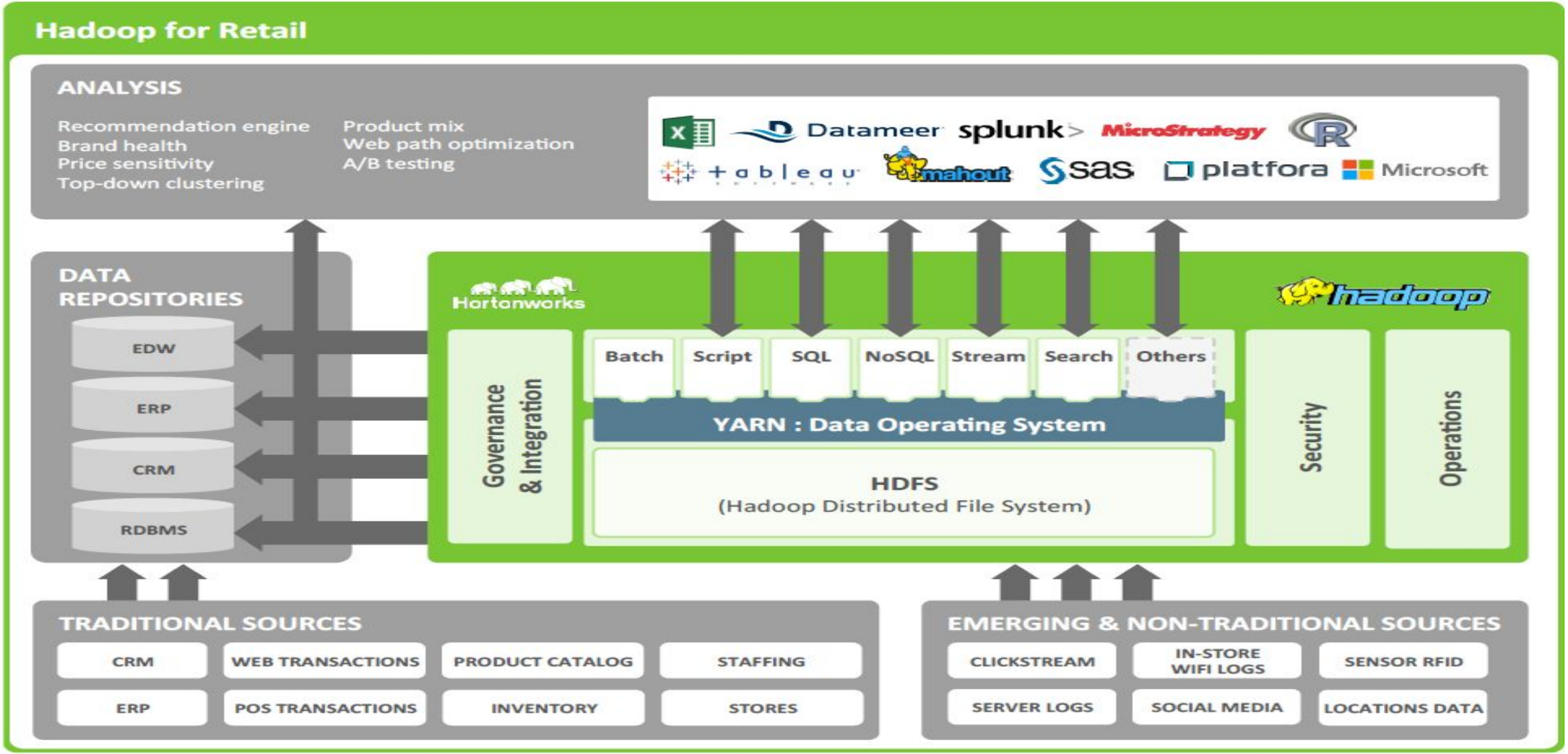
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

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

A brief history of data science & AI



Hadoop for Retail

ANALYSIS

Recommendation engine
Brand health
Price sensitivity
Top-down clustering

Product mix
Web path optimization
A/B testing

DATA REPOSITORIES

EDW
ERP
CRM
RDBMS

Governance & Integration

Batch
Script
SQL
NoSQL
Stream
Search
Others

YARN : Data Operating System

HDFS
(Hadoop Distributed File System)

Security

Operations

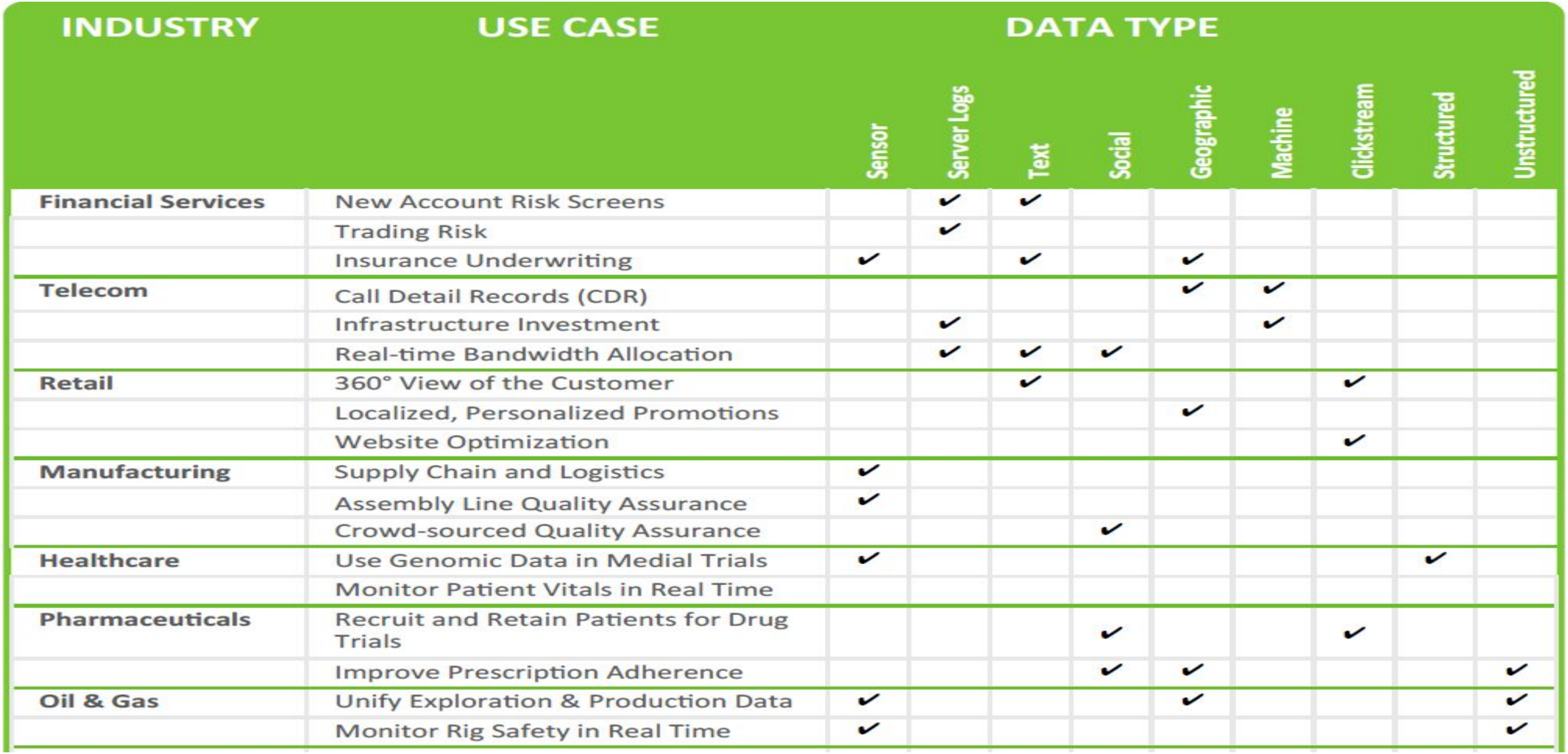
TRADITIONAL SOURCES

CRM
WEB TRANSACTIONS
PRODUCT CATALOG
STAFFING
ERP
POS TRANSACTIONS
INVENTORY
STORES

EMERGING & NON-TRADITIONAL SOURCES

CLICKSTREAM
IN-STORE WIFI LOGS
SENSOR RFID
SERVER LOGS
SOCIAL MEDIA
LOCATIONS DATA

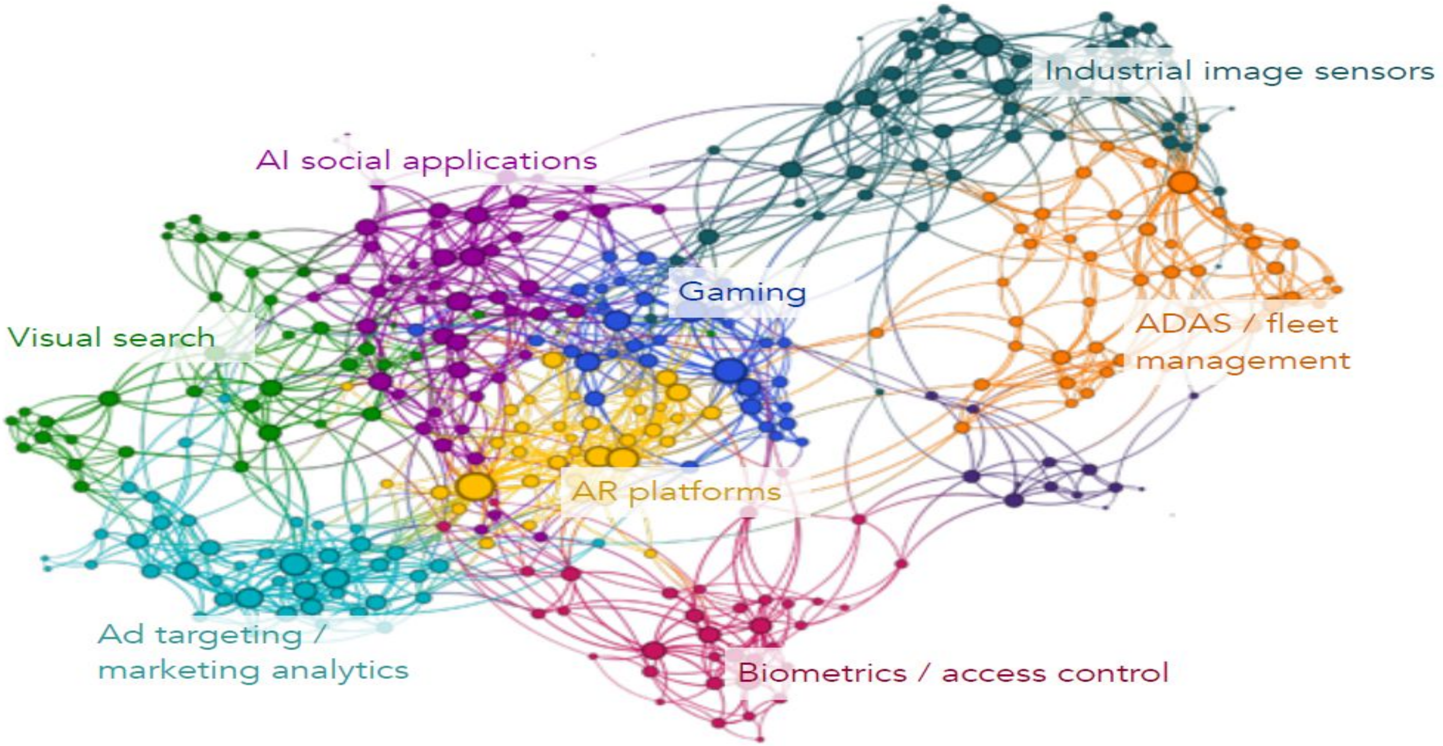
A brief history of data science & AI



INDUSTRY	USE CASE	DATA TYPE								
		Sensor	Server Logs	Text	Social	Geographic	Machine	Clickstream	Structured	Unstructured
Financial Services	New Account Risk Screens		✓	✓						
	Trading Risk		✓							
	Insurance Underwriting	✓		✓		✓				
Telecom	Call Detail Records (CDR)					✓	✓			
	Infrastructure Investment		✓				✓			
	Real-time Bandwidth Allocation		✓	✓	✓					
Retail	360° View of the Customer			✓				✓		
	Localized, Personalized Promotions					✓				
	Website Optimization							✓		
Manufacturing	Supply Chain and Logistics	✓								
	Assembly Line Quality Assurance	✓								
	Crowd-sourced Quality Assurance				✓					
Healthcare	Use Genomic Data in Medial Trials	✓							✓	
	Monitor Patient Vitals in Real Time									
Pharmaceuticals	Recruit and Retain Patients for Drug Trials				✓			✓		
	Improve Prescription Adherence				✓	✓				✓
Oil & Gas	Unify Exploration & Production Data	✓				✓				✓
	Monitor Rig Safety in Real Time	✓								✓

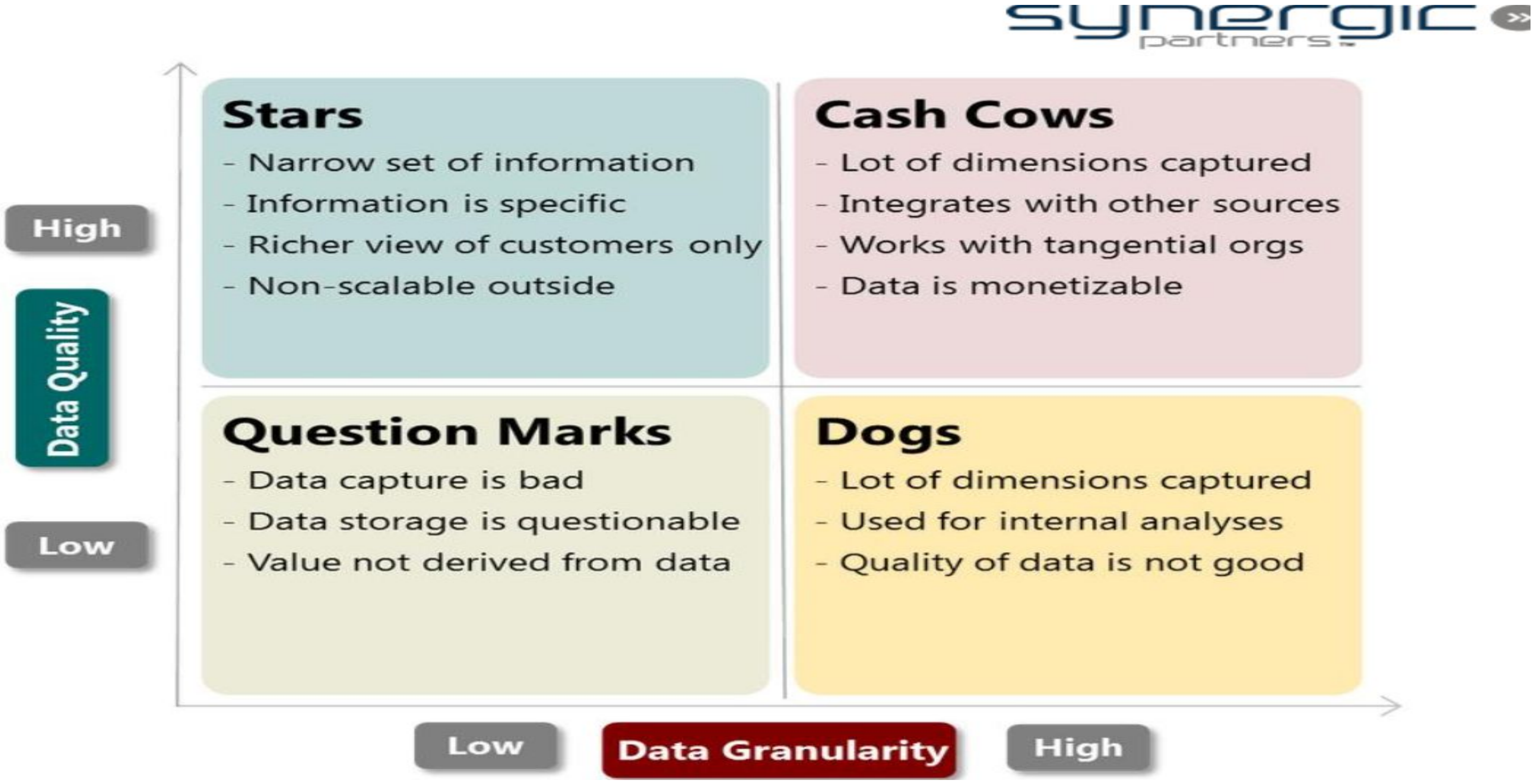
A brief history of data science & AI

Computer Vision startup ecosystem



Quid

A brief history of data science & AI



Stars

- Narrow set of information
- Information is specific
- Richer view of customers only
- Non-scalable outside

Cash Cows

- Lot of dimensions captured
- Integrates with other sources
- Works with tangential orgs
- Data is monetizable

Question Marks

- Data capture is bad
- Data storage is questionable
- Value not derived from data

Dogs

- Lot of dimensions captured
- Used for internal analyses
- Quality of data is not good

Agenda

1. Overview of data science and AI
2. Creating data scientists
3. Recruiting data scientists
4. Aiming for data science success stories

Obstacles on the road to AI excellence

1. Legacy systems and architectures
    - The ‘all or nothing’ approach
    - The system compatibility issues

2. Skills gap and team resistance
    - Limited amount of skilled employees with the necessary combination of skills
    - Efficient data science teams should include people with Business, engineering & statistics backgrounds

3. Irrational expectations

Agenda

1. Overview of data science and AI
2. Creating data scientists
3. Recruiting data scientists
4. Aiming for data science success stories

Recruiting data scientists: the leader

1. Start your journey with very experienced hire (5-10+ years)
2. The first hire should be a **communicator**
3. The first hire must have data engineering and **infrastructure know how**
4. The first hire must have **algorithmic training**
5. The first hire must have **statistical analysis knowledge**
6. The first hire must have **business experience**

Recruiting data scientists: the set up

1. Initially, pick a couple of test projects with short development cycles to evidence success

2. Express strong managerial support by including him in strategic meetings and identify pockets of resistance

1. Ask him to devote 10% of his time for internal training

2. Find a way to let existing IT managers and analysts be part of the process

3. Ask the devops team to be fully available to him to set up the environment from day 1: it is going to much take longer than they both think

Recruiting data scientists: the team

1. Don’t let the leader recruit alone

2. Communication skills almost as important as technical skills

1. Mix Phds and Non Phds

1. Physicists, economists and biochemists often doing well , don’t reject psychology grads and hire as many linguists as possible

2. Have a portfolio strategy , there is no unicorn : mix optimization, statistics, nlp, graph analytics, data mining expert

3. Beware claims about deep learning mastery

Agenda

1. Overview of data science and AI

2. Creating data scientists

3. Recruiting data scientists

4. Aiming for data science success stories

Creating data scientists

1. Don’t let the leader recruit alone

2. Communication skills almost as important as technical skills

1. Mix Phds and Non Phds

1. Physicists, economists and biochemists often doing well , don’t reject psychology grads and hire as many linguists as possible

2. Have a portfolio strategy , there is no unicorn : mix optimization, statistics, nlp, graph analytics, data mining expert

3. Beware claims about deep learning mastery