Big Data and Analytics
Use Cases

Dr. Abzetdin Adamov
School of Information Technology and Engineering
ADA University
www.site.ada.edu.az/~aadamov

Objectives

After completing this lesson, you should be able to:

- Technological Landscape
- Define Data Science and what a Data Scientist does
- List Big Data and Data Science Use Cases
- Define Standard Parameters for Use Cases

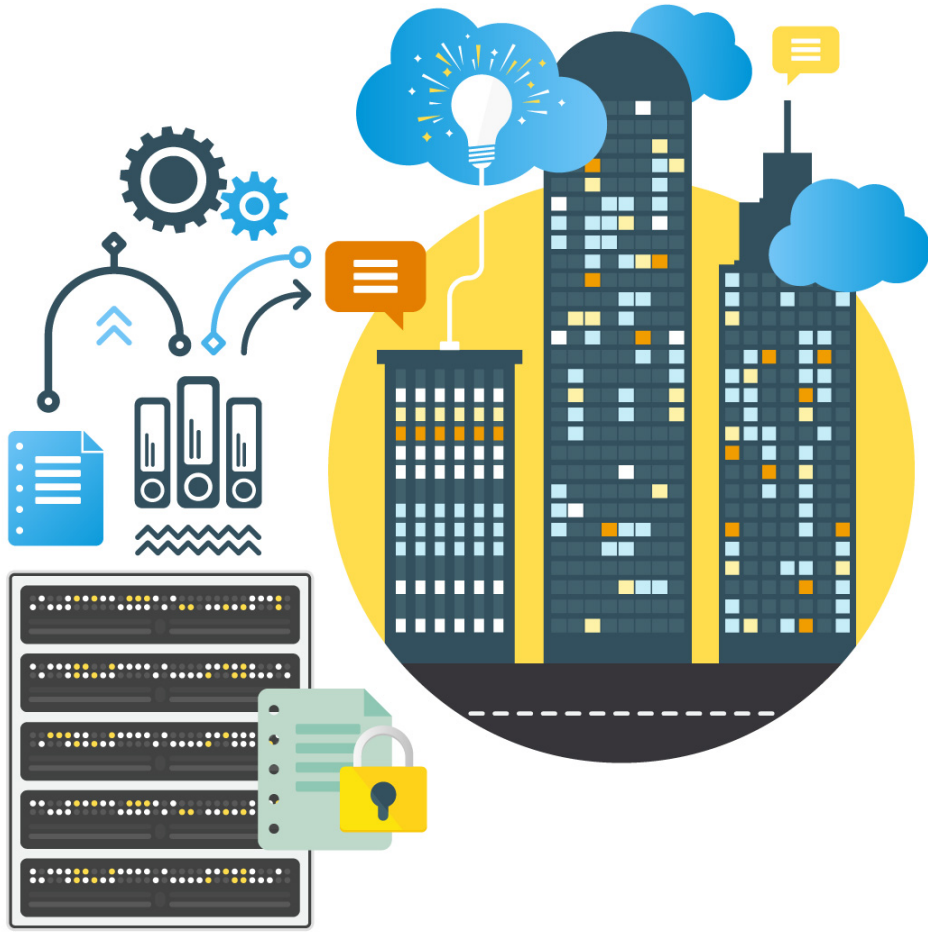
Most Requested Uses of Big Data

- Log Analytics & Storage
- Fusion of multi-INT (multi-formats)
- RFID Tracking & Analytics
- Fraud Detection & Modeling
- Risk Modeling & Management
- 360°View of a Person, Place, or Thing
- Warehouse Extension (case patterns)
- Email / Call Center Transcript Analysis
- Call Detail Record Analysis
- IBM Watson

Do we have Big Data?

real-time Big Data use cases

- Ad Technology
- Digital Marketing
- Fraud Detection
- Internet of Things
  - Cyberthreat Security
  - Network Monitoring
  - Personalized Medicine



Big Data Everywhere!

BIG DATA

Data that is TOO LARGE & TOO COMPLEX for conventional data tools to capture, store and analyze.

Shares traded on US Stock Markets each day:

7 Billion

Data generated in one flight from NY to London:

10 Terabytes

Number of tweets per day on Twitter:

400 Million

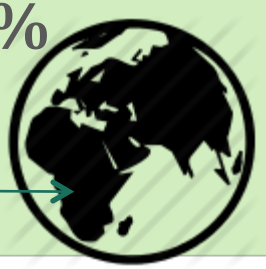
Number of 'Likes' each day on Facebook:

3 Billion

The 3V's of Big Data

VOLUME VARIETY VELOCITY

90%



OF THE WORLD'S DATA WAS GENERATED IN THE **LAST TWO YEARS**

What is Analytics?

Data on its own is useless unless you can make sense of it!

WHAT IS ANALYTICS?

The scientific process of transforming data into insight for making better decisions, offering new opportunities for a competitive advantage

The Case for Business Analytics

BUSINESS NEED

- The Business environment today is more complex than ever before.
- Businesses are expected to be diligently responsive to the increasing demands of customers, various stakeholders and even regulators.

SOLUTION

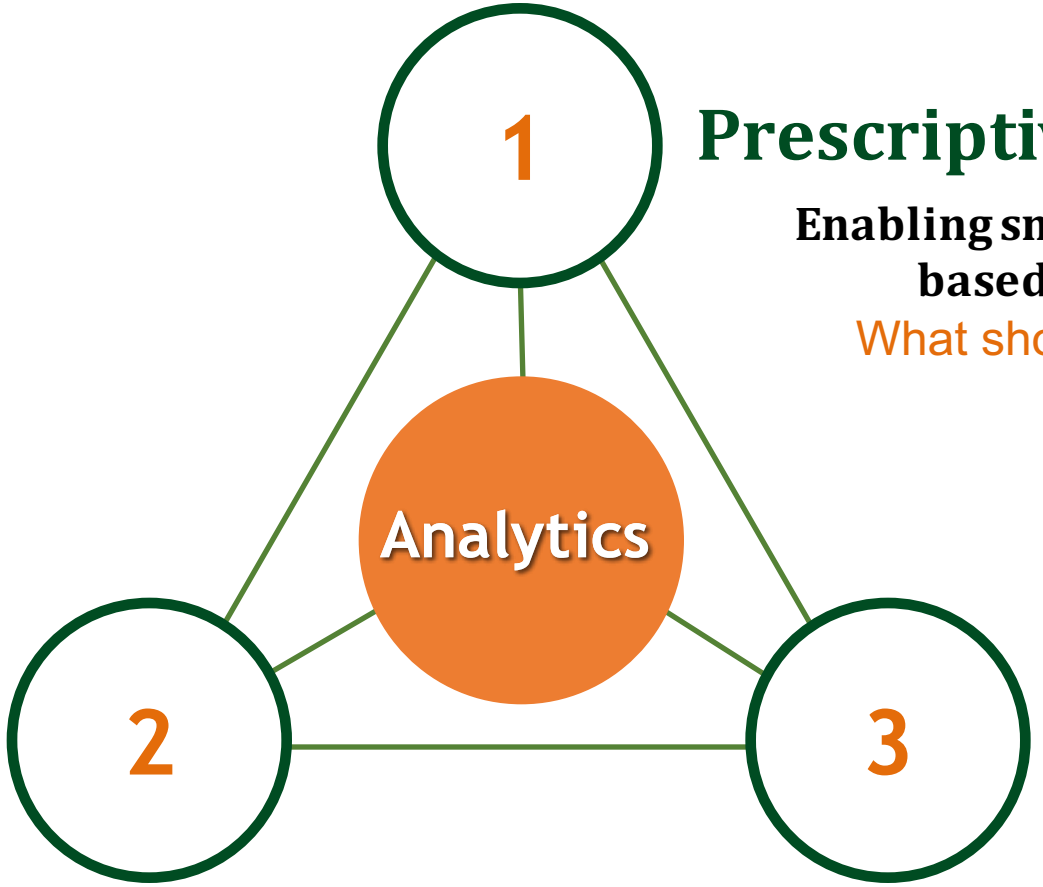
- Organizations have been turning to the use of analytics.
- More than 83% of Global CIOs surveyed by IBM in 2010 singled out Business Intelligence and Analytics as one of their visionary plans for enhancing competitiveness.

GOAL

In most cases the primary objective of an organization that seeks to turn to analytics is:

- Revenue/Profit growth
- Optimize expenditure

Types of Analytics



Prescriptive Analytics

Enabling smart decisions based on data

What should we do?

Predictive analytics

Predicting the future based on historical patterns

What could happen?

Descriptive analytics

Mining data to provide business insights

What has happened?

Disrupt or Be **Disrupted**

41% of executives say digital disruption increases risk of being put out of business

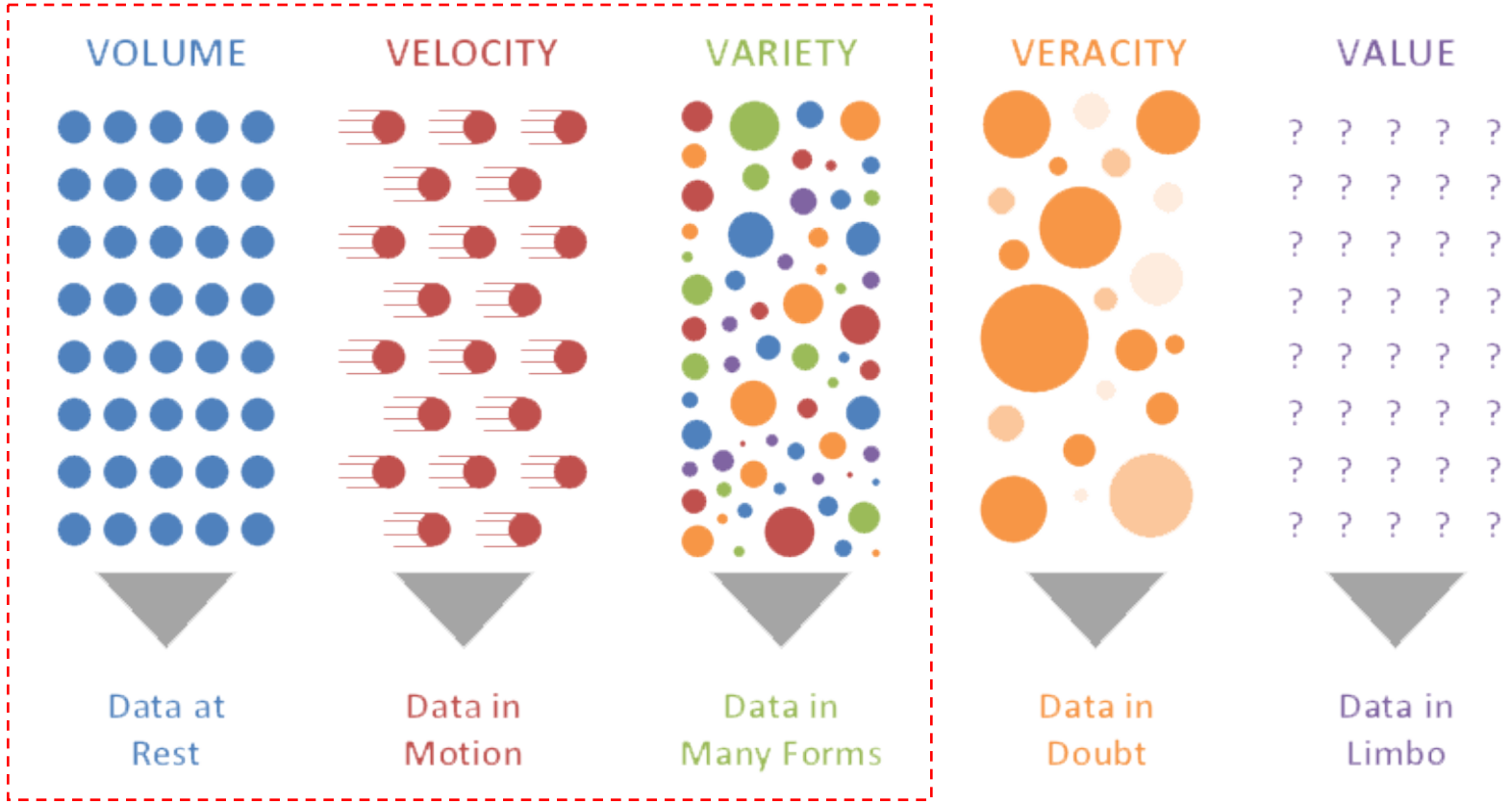
Digital Vortex, 2015

"New technologies combine to create a business innovation platform, not just a technology platform, helping transform every industry on the planet."

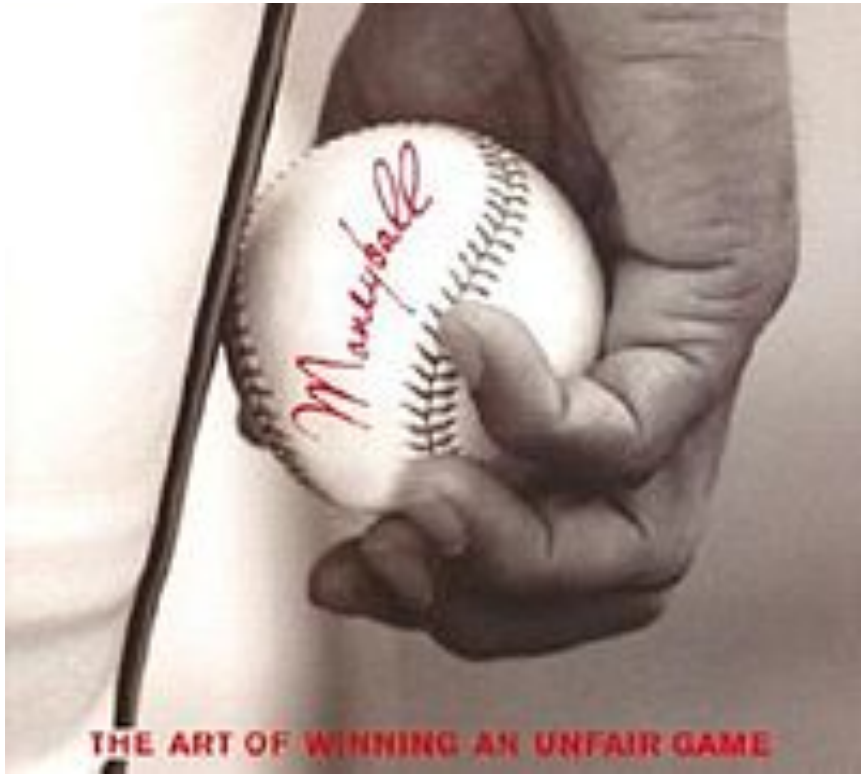
IDC (Dec. 2014)

Challenging Environment – 5Vs

5 Vs that best describe the nature of Big Data Problem

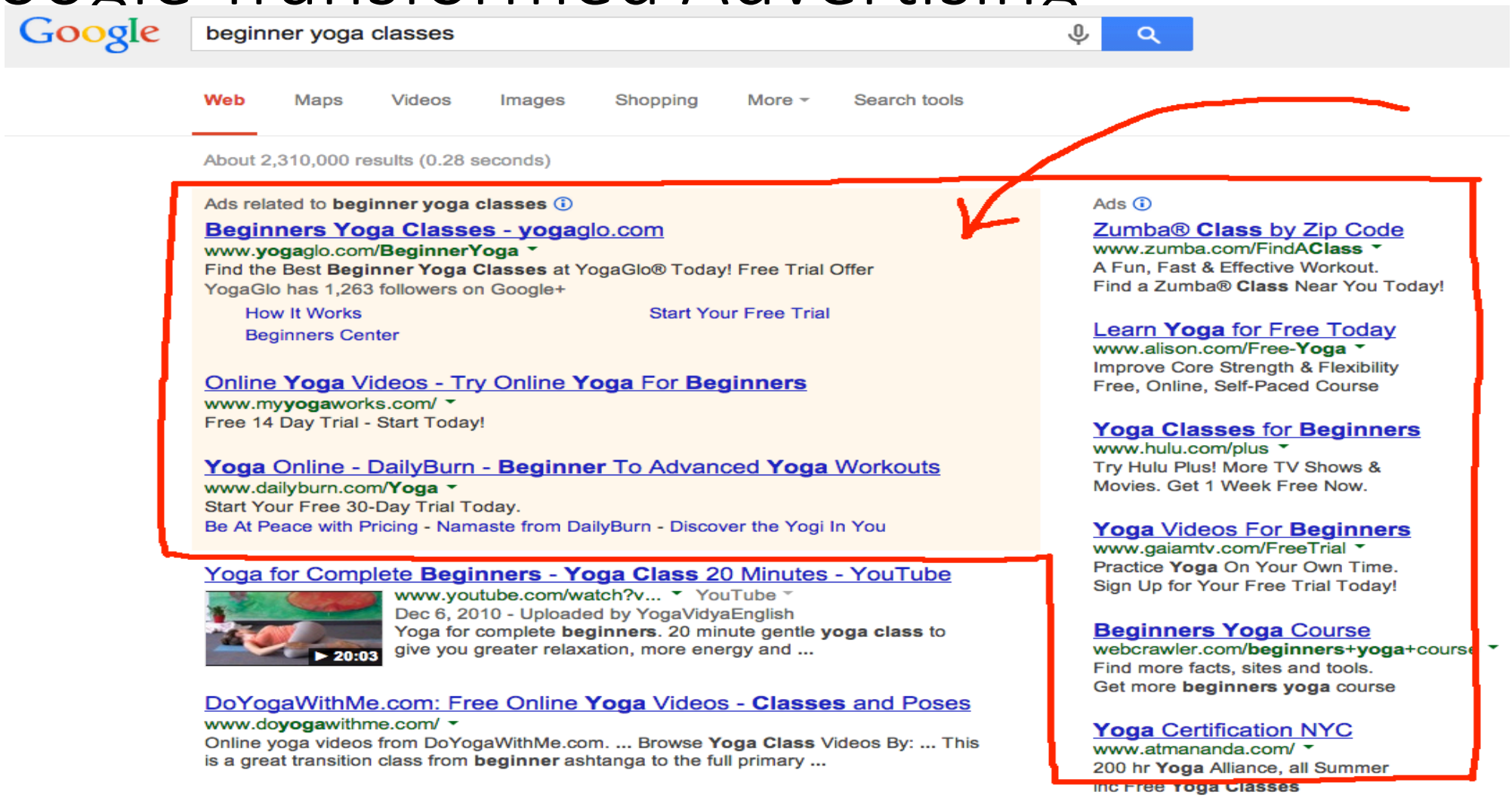


Billy Bean Transformed Baseball Scouting

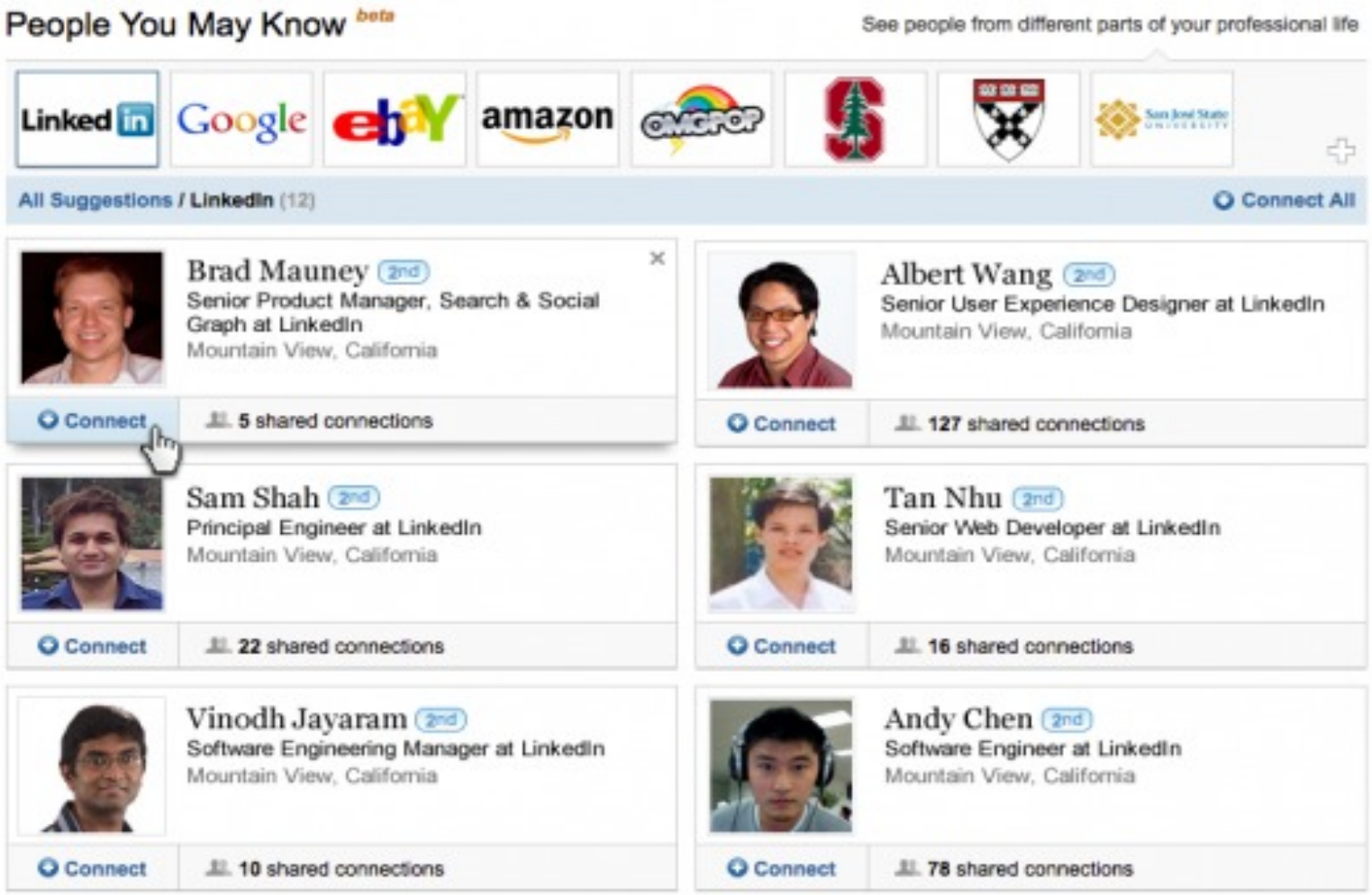


The use of data analytics has changed the way many major league front offices do business – emphasizing data over traditional player metrics.

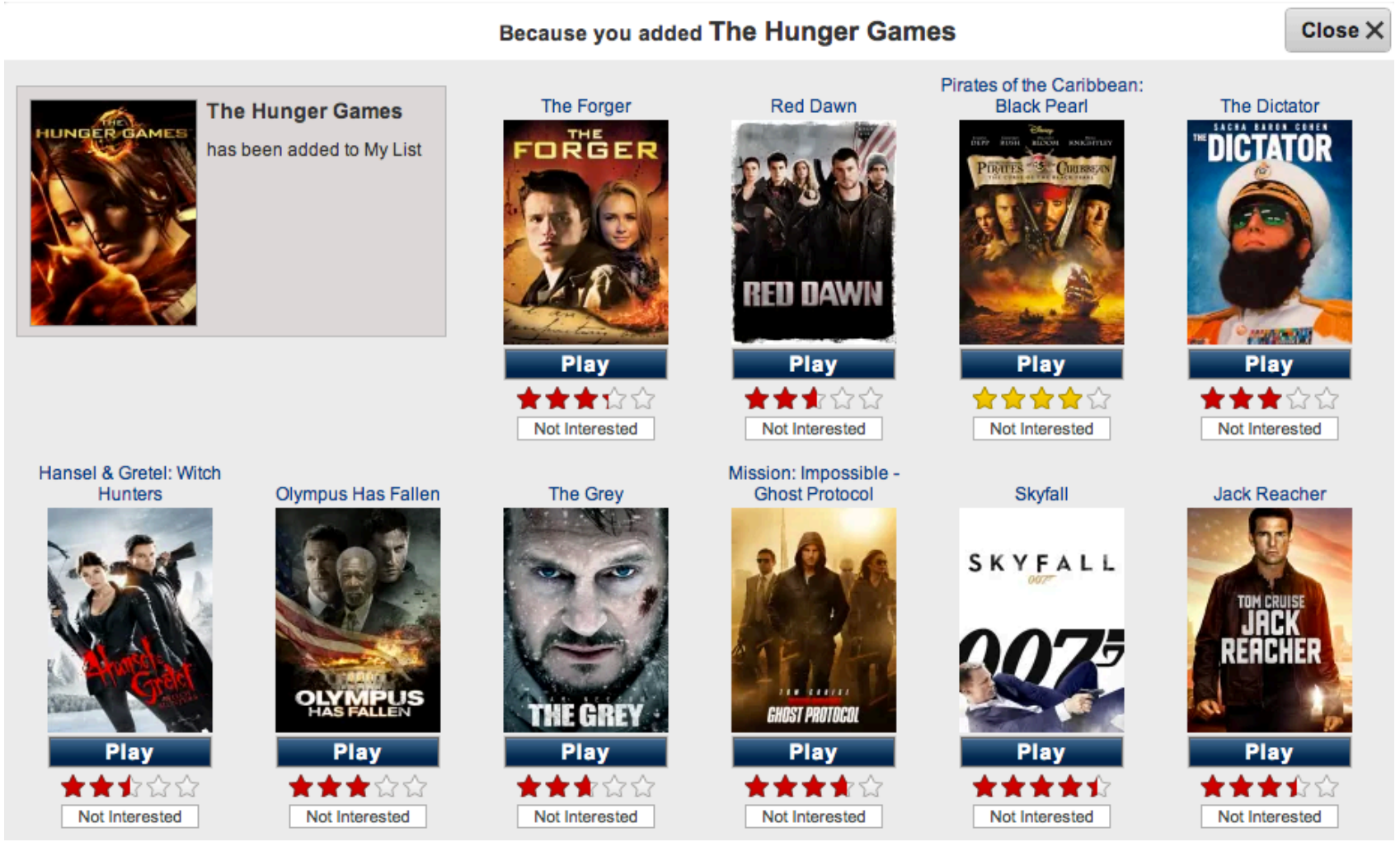
Google Transformed Advertising



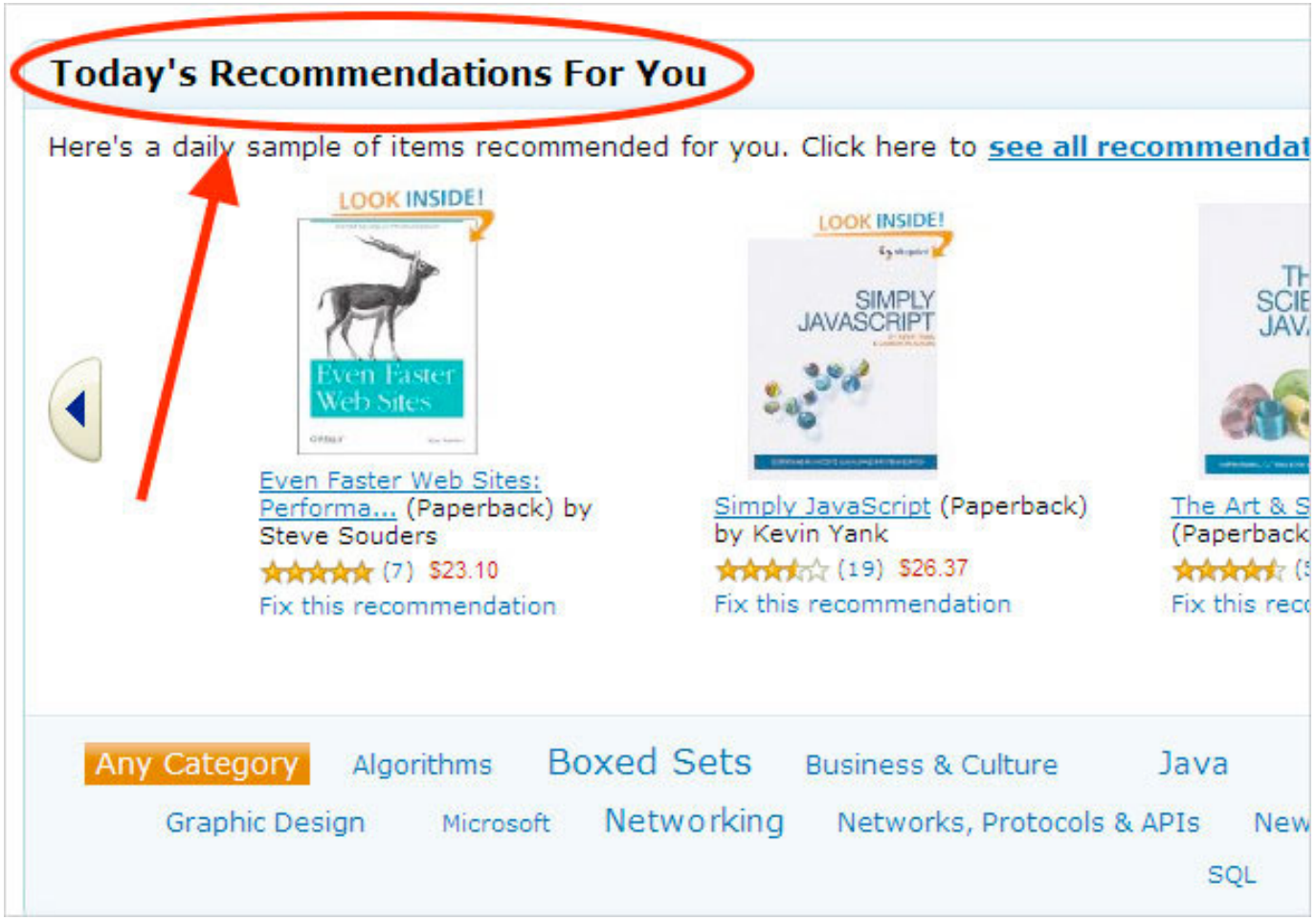
LinkedIn Transformed Network Growth



Netflix Transformed Streaming Video



Amazon Transformed Product Recommendations



Retail Transformed Market Basket Analysis

Question

Which products are very frequently purchased together?

Solution Approach

Determine frequent item sets from transaction logs
Use to design physical store layout accordingly
Use to design special offers and coupon strategy

Example

The urban legend of Diapers & Beer

Financial Use Cases Transformed Analytics

Customer Profiling

Financial firms use parameters about customers to determine risk

- Likelihood of customer repaying a loan
  - Collect data (FICO, net worth, etc.)
  - Build a predictive model
  - Use model to price loans

Fraud Detection

Credit card companies look at transaction factors to detect fraud

- Likelihood any given transaction is fraudulent
  - Track spending habits
  - Build a spending model
  - Alert when transaction falls outside model

Demand for Data Scientists

The Data Science Skillset Continuum

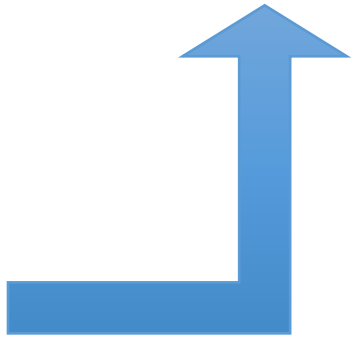
Software engineer — Data Engineer — Data Scientist — Applied Scientist — Research Scientist

Role	Data Engineer	Applied Scientist
Function	Build production-grade data products	Find signal/meaning in the data Applies statistical/ML models and tunes the algorithm
Good at….	Data and Systems architecture Hadoop, PIG/HIVE, MapReduce, ops/admin Java, Python, Perl, SQL, C++, NoSQL (HBase, Cassandra, Mongo)	Statistics, Machine learning Text processing, NLP R, Python, Matlab, SAS, SQL Scripting Visualization / telling the story

What is Data Science?

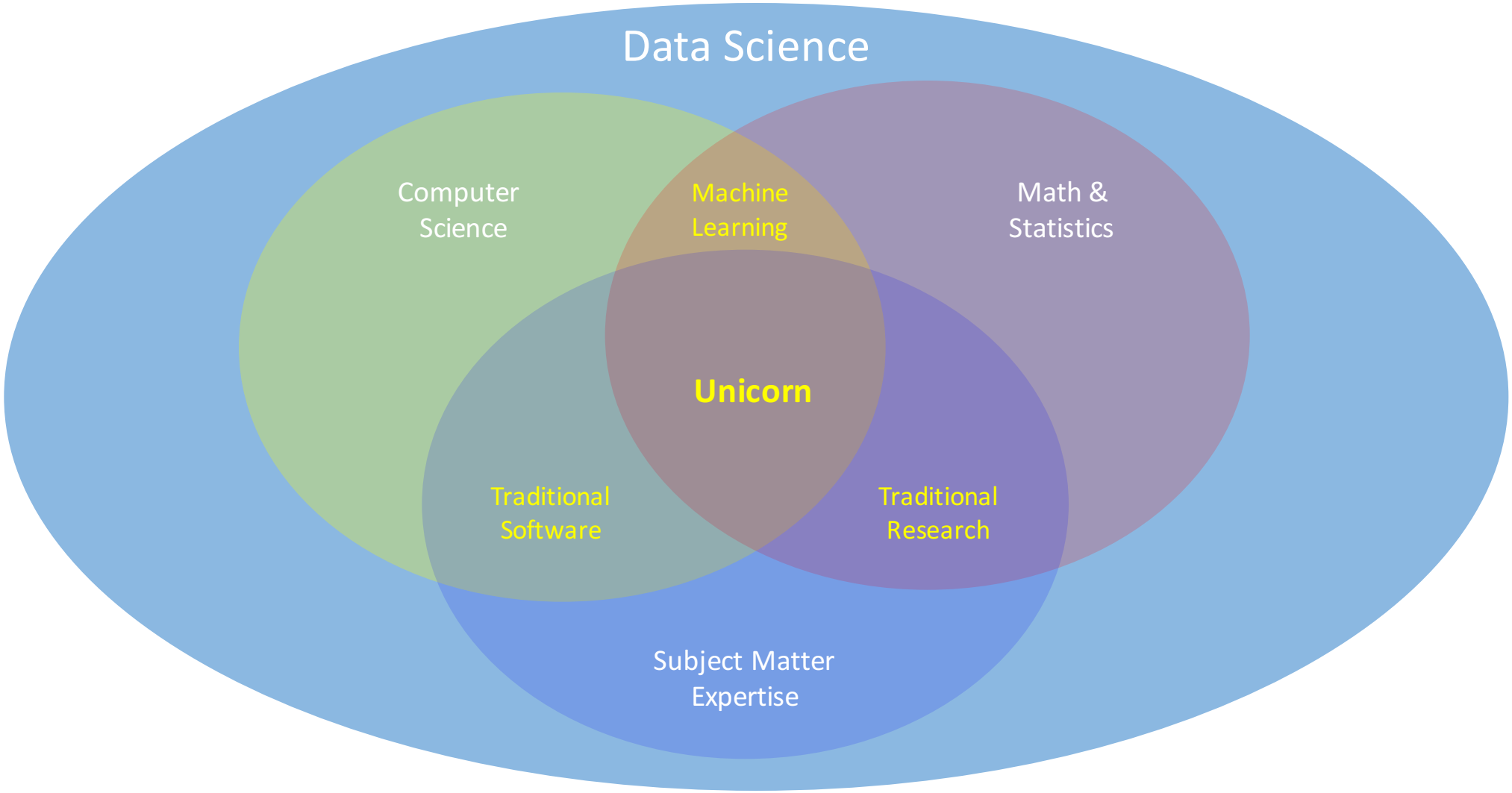
"Building software products (aka Data Products) whose core functionality relies on applying statistical or machine learning methods to data."

What is a data scientist?
A person who does **this**



Nice historical perspective:
http://whatsthebigdata.com/2012/04/26/a-very-short-history-of-data-science/

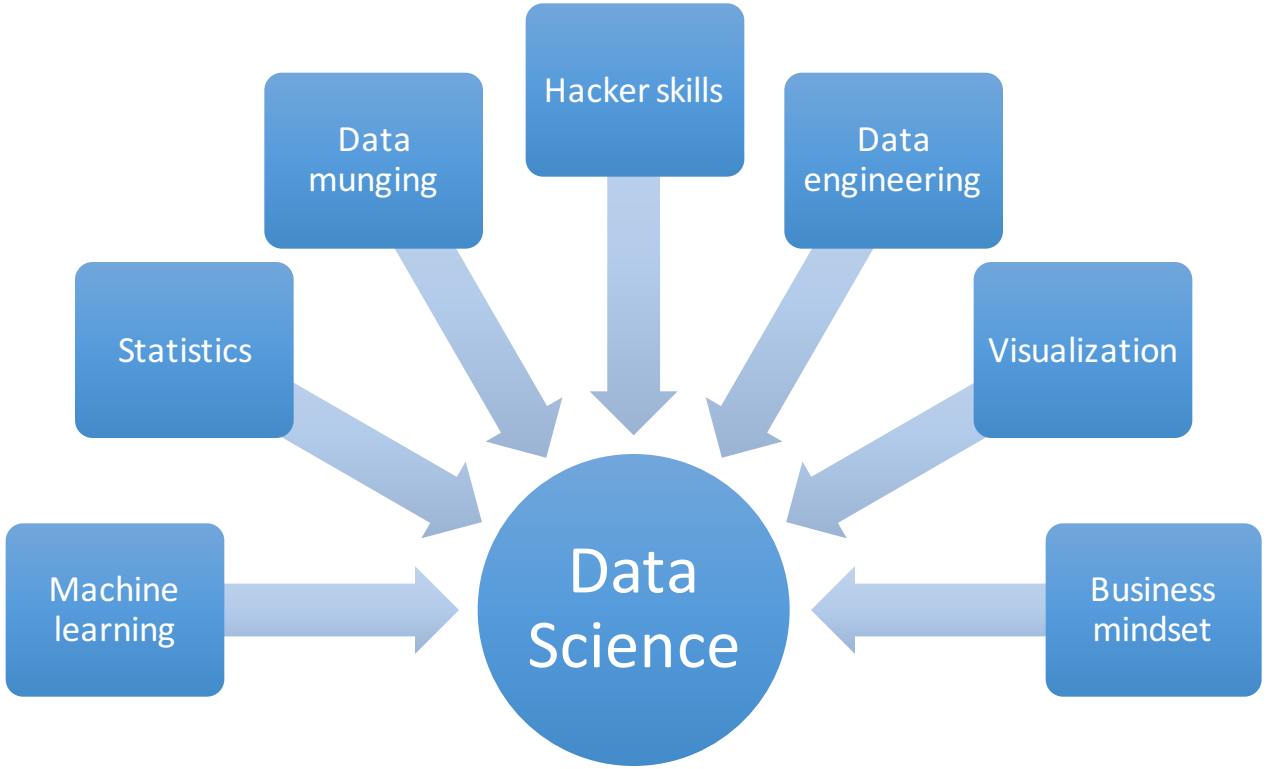
Data Science Skill Set



Data Scientist Skill Set

Our perspective:

- It's a hybrid role: data engineer + applied scientist
- Combines many disciplines:

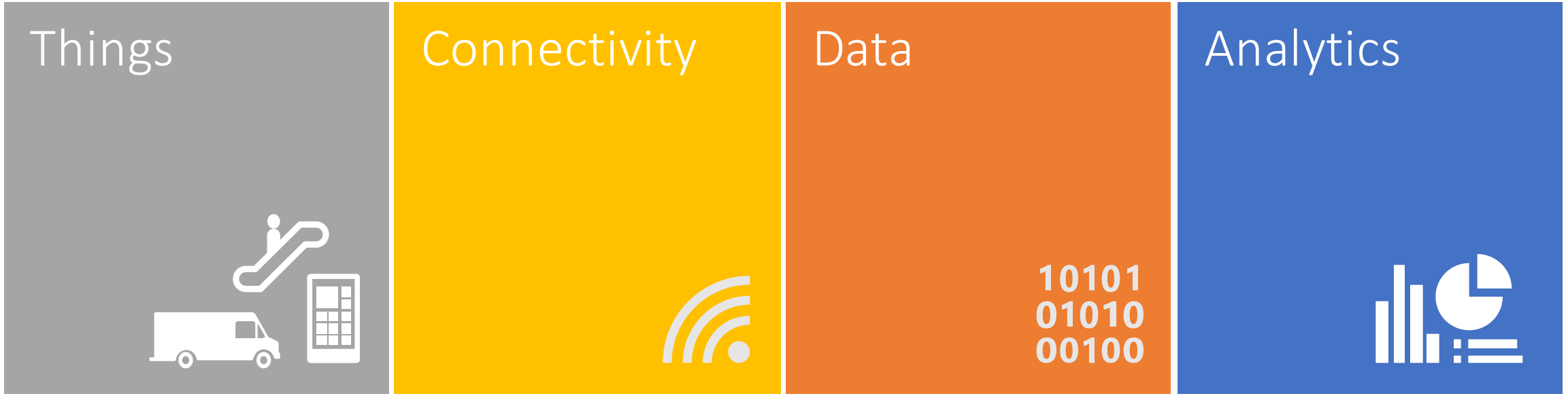


What is Big Data?

Harness the growing and changing nature of data

- Challenge is combining transactional data stored in relational databases with less structured data
- Big Data = All Data
- Get the right information to the right people at the right time in the right format

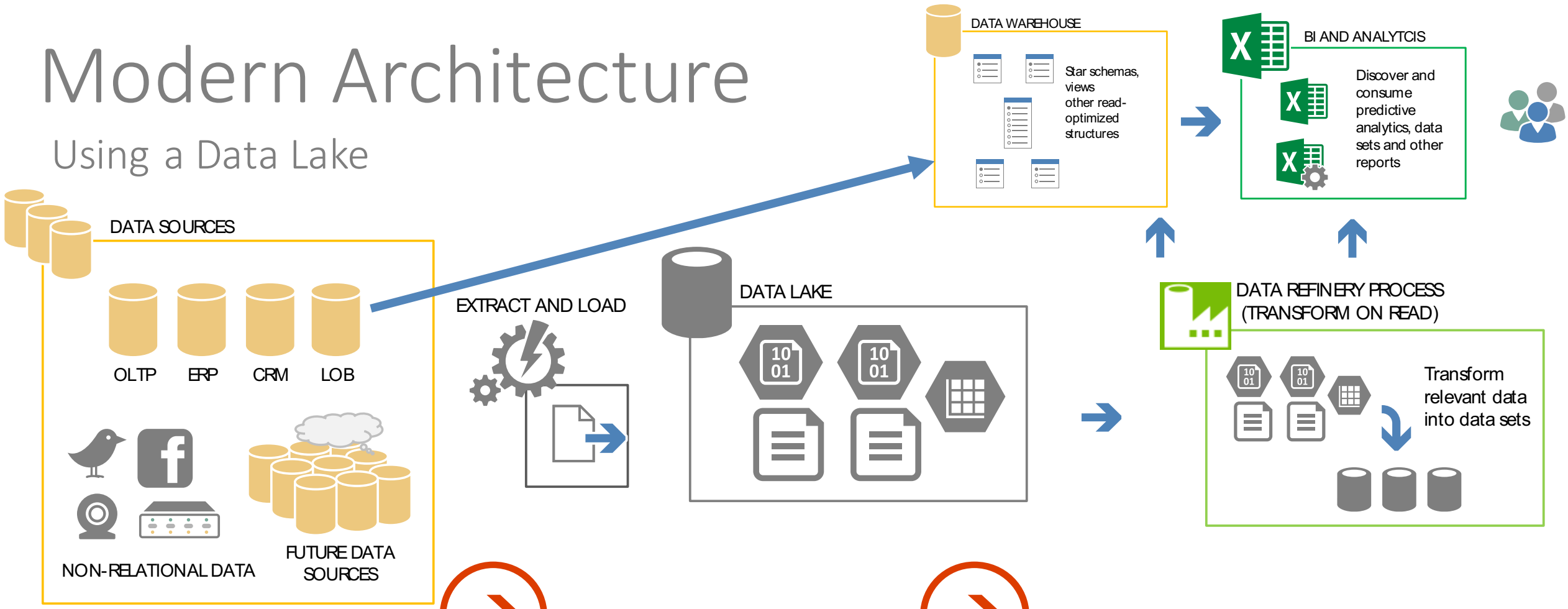
What is the Internet of Things?



IoT = sensor-acquired data

Modern Architecture

Using a Data Lake



All data sources are considered

Leverages the power of on-prem technologies and the cloud for storage and capture

Native formats, streaming data, big data

Extract and load, no/minimal transform

Storage of data in near-native format

Orchestration becomes possible

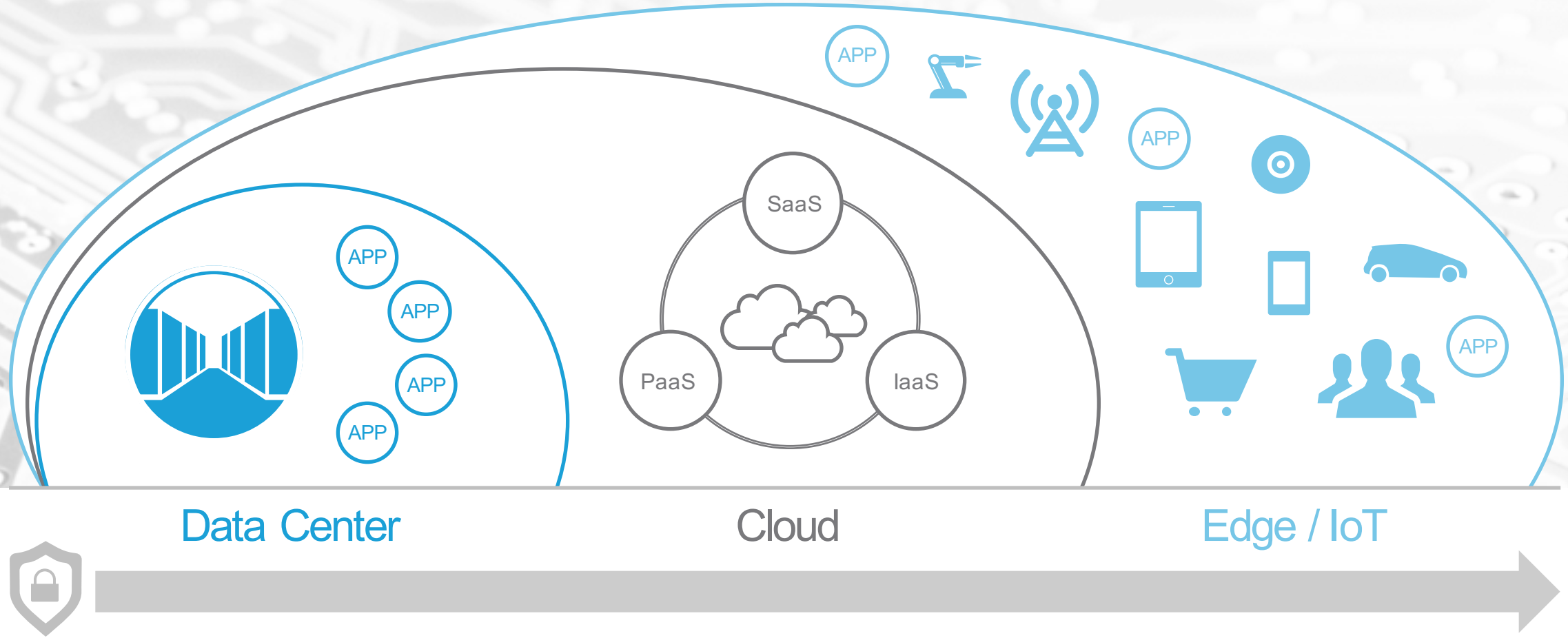
Streaming data accommodation becomes possible

Refineries transform data on read

Produce curated data sets to integrate with traditional warehouses

Users discover published data sets/services using familiar tools

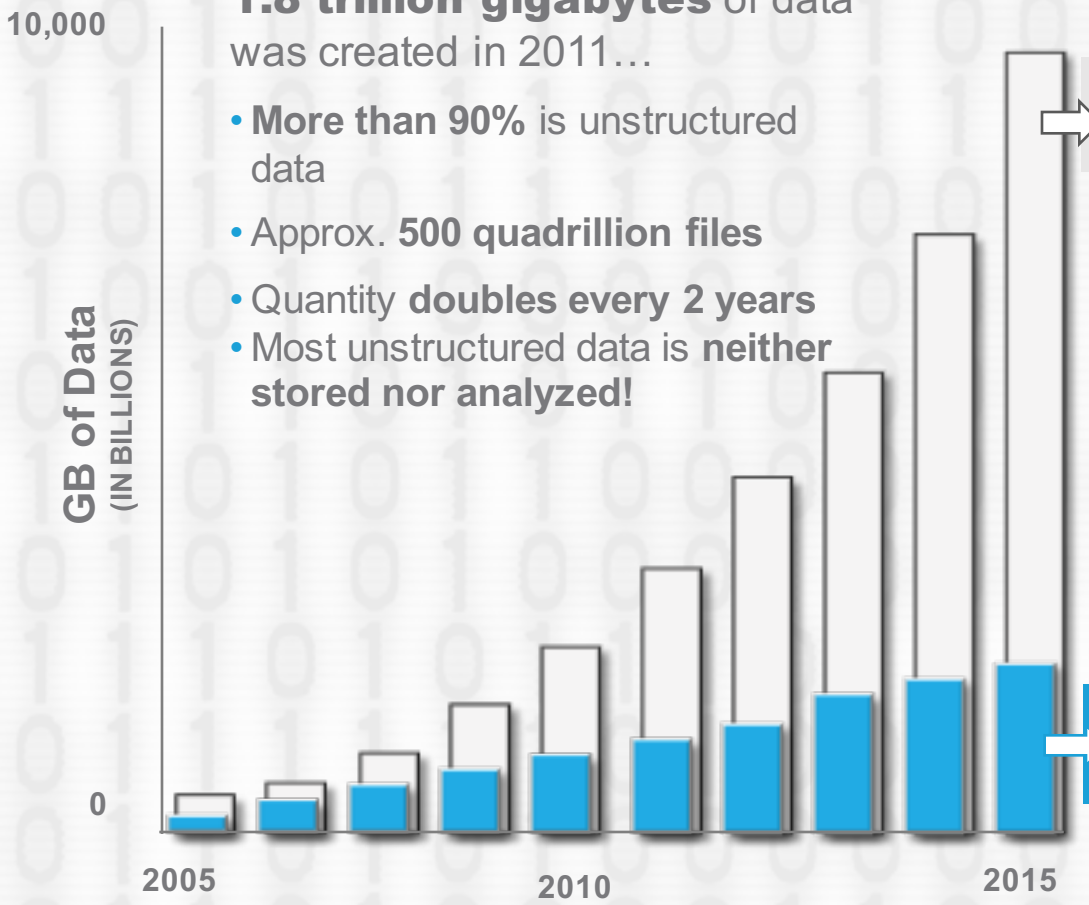
A Broad Perspective To Set The Scene

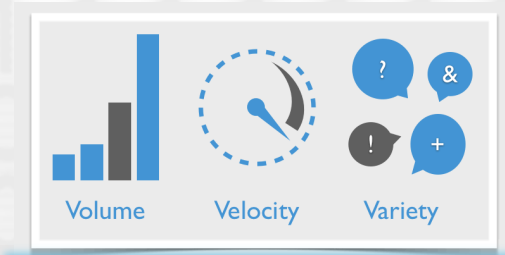


The Explosion of Unstructured Data

1.8 trillion gigabytes of data was created in 2011…

- **More than 90%** is unstructured data
- Approx. **500 quadrillion files**
- Quantity **doubles every 2 years**
- Most unstructured data is **neither stored nor analyzed!**





UNSTRUCTURED DATA

STRUCTURED DATA

Source: Cloudera

Application Centric Infrastructure

Simple

Scalable

Secure

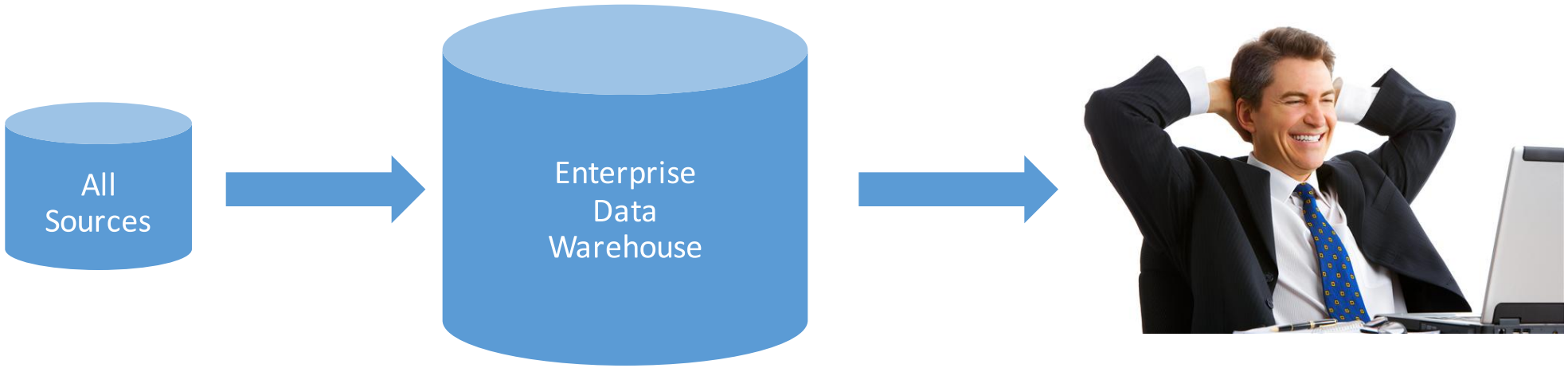
Open

Manageable
& Visible

Cost Effective

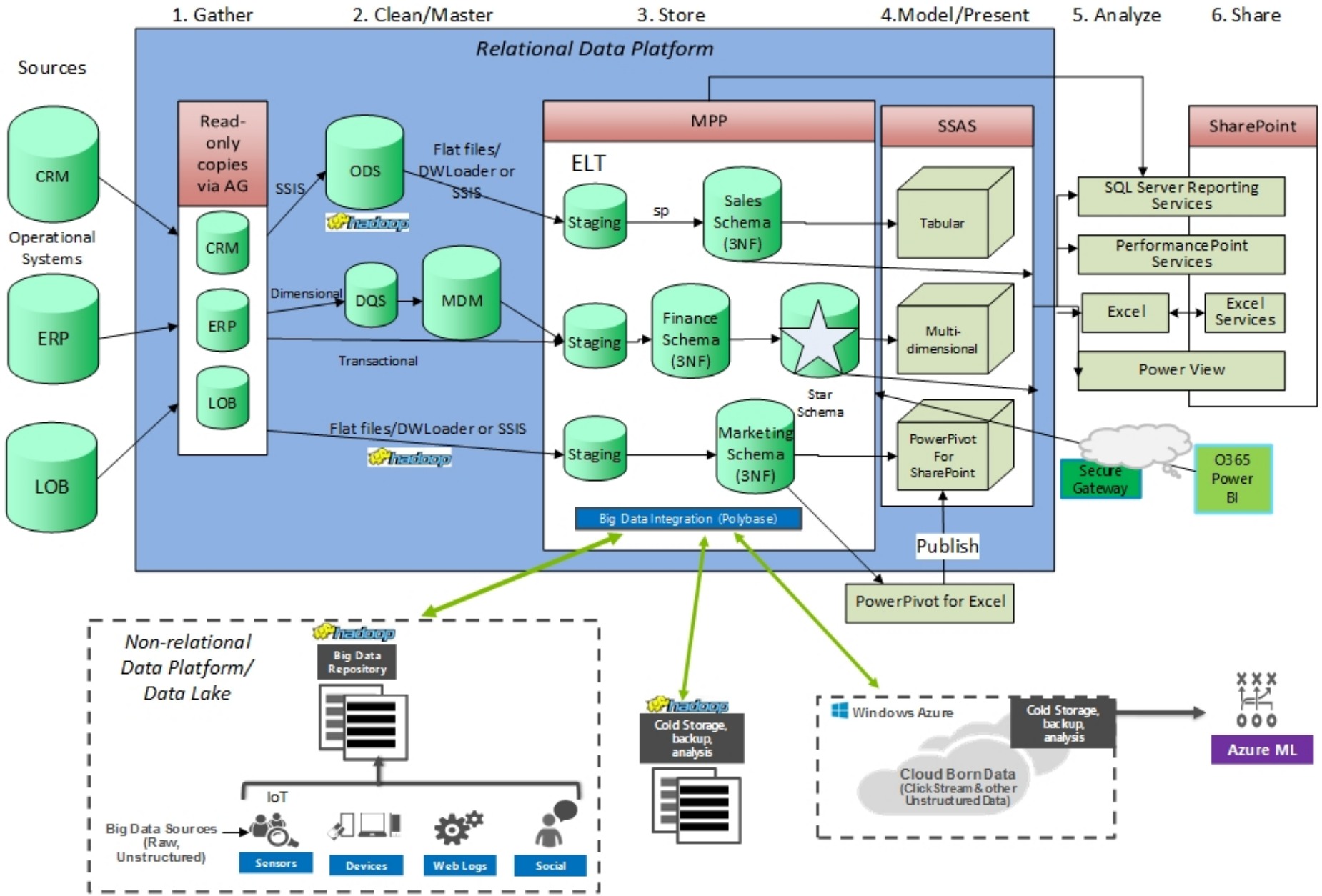
The Dream

Modern Data Warehouse



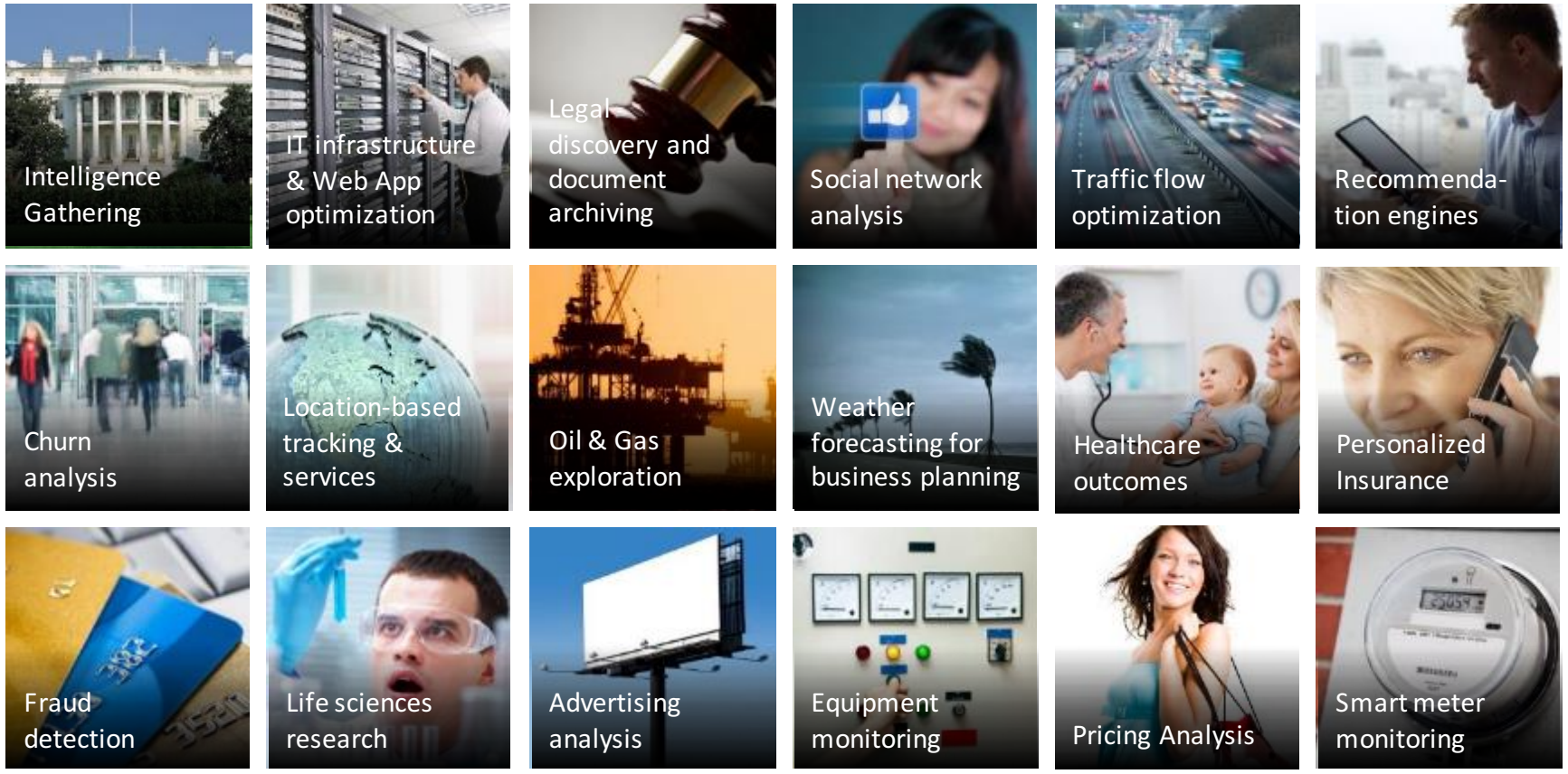
The Reality

Modern Data Warehouse



Let’s set ON light bulbs in your head

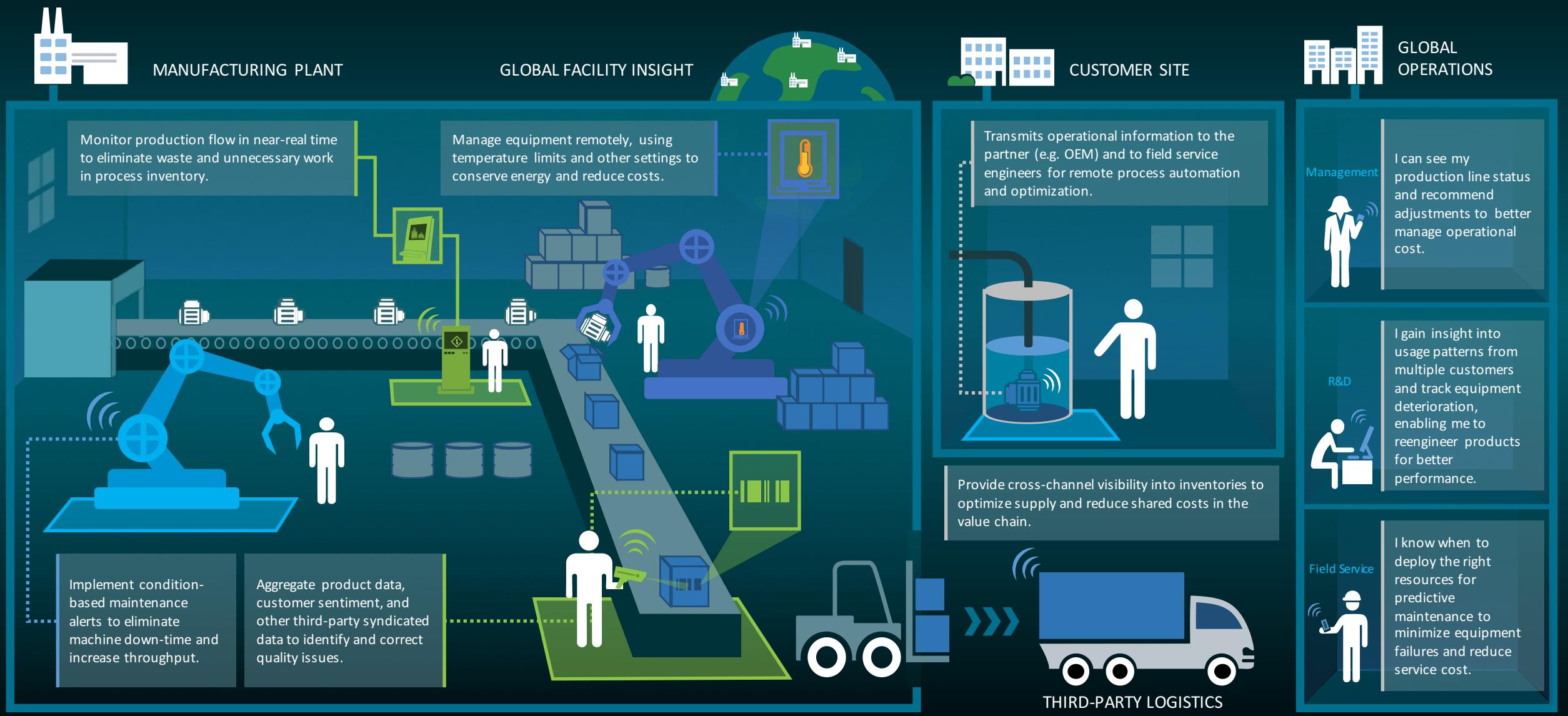
Data Analytics is needed everywhere



CROSS-INDUSTRIAL USE CASES

Microsoft’s Vision

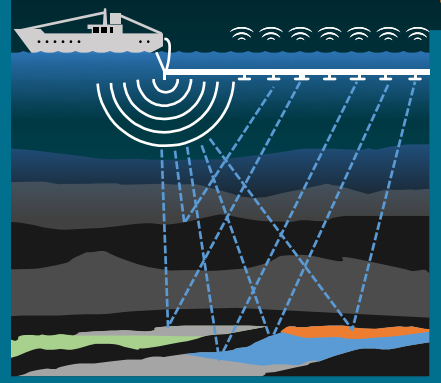
The Internet of Things – Manufacturing



The Internet of Things – Oil & Gas

1. Exploration

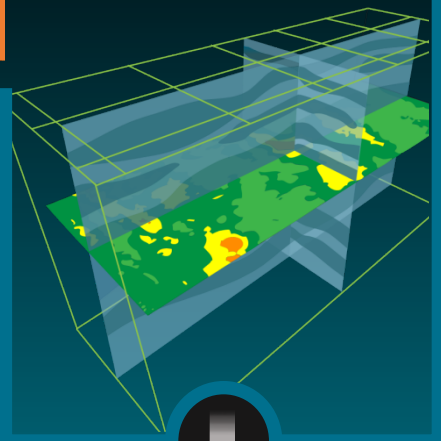
Find new hydrocarbon reservoirs quicker with seismic data uploaded to the cloud and prepared for analysis



Operations Control Center

Consolidate data from surveys, drill logs, and external sources to generate advanced reservoir models and production forecasts

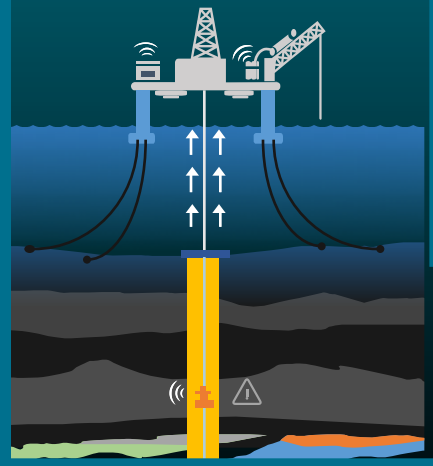
2. Development



3. Drilling

Combine near real-time drilling and seismic data to optimize drilling trajectories and recovery potential, while minimizing environmental risk

4. Production



Maximize recovery by monitoring near real-time production data and generating alerts for conditional maintenance needs

Production Manager

Integrate all upstream data onto a unified platform to facilitate analytics, information sharing, and organizational transition

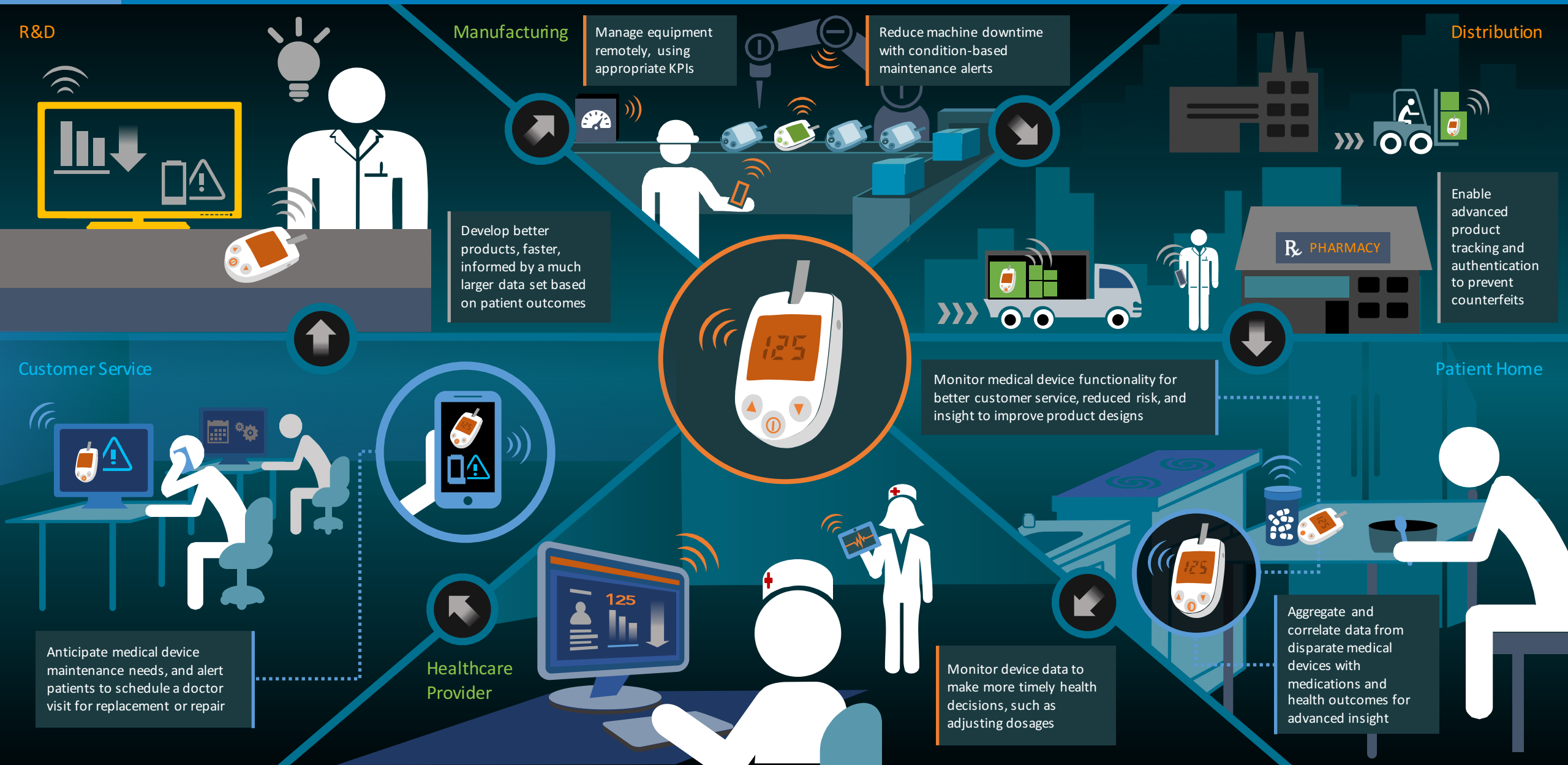
Geologist

Utilize advanced 3D and 4D visualizations based on analytic algorithms to model subsurface geology

Onsite personnel

Establish near real-time communication and automatically publish events and alarms to the field to guide and protect onsite personnel and assets

The Internet of Things – Pharma



The Internet of Things – Healthcare

PATIENT HOME

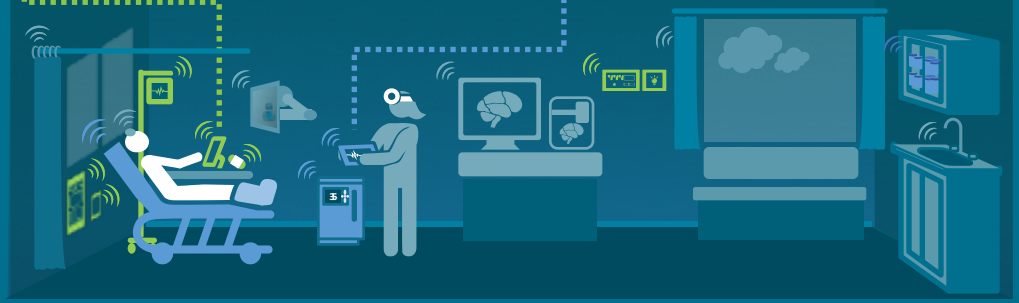
Monitor patient conditions with in-home medical devices that alert care team staff when a health event occurs.

Transform the vehicle into a smart environment that monitors health indicators.

HOSPITAL

Enable an interactive experience between patients and collaborative care teams, and reduce response times by providing remote access to the latest patient data.

Connect patient data to contextual data, so the latest patient data automatically displays on care provider devices based on their location and role.

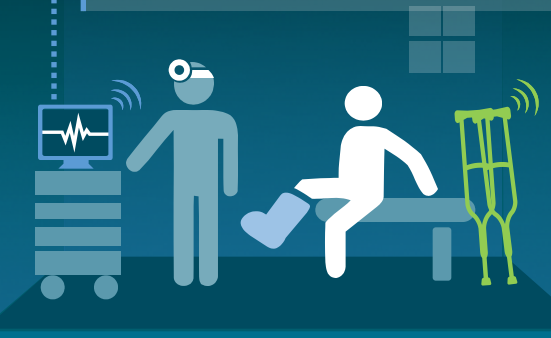


Combine data from various sources to uncover insights that enable an enhanced patient journey, improved operational efficiency, and better risk management.

Make patient data visible and actionable in near real-time, enabling improved outcomes through data-driven decision making, better coordination and error reduction.

OUTPATIENT FACILITY

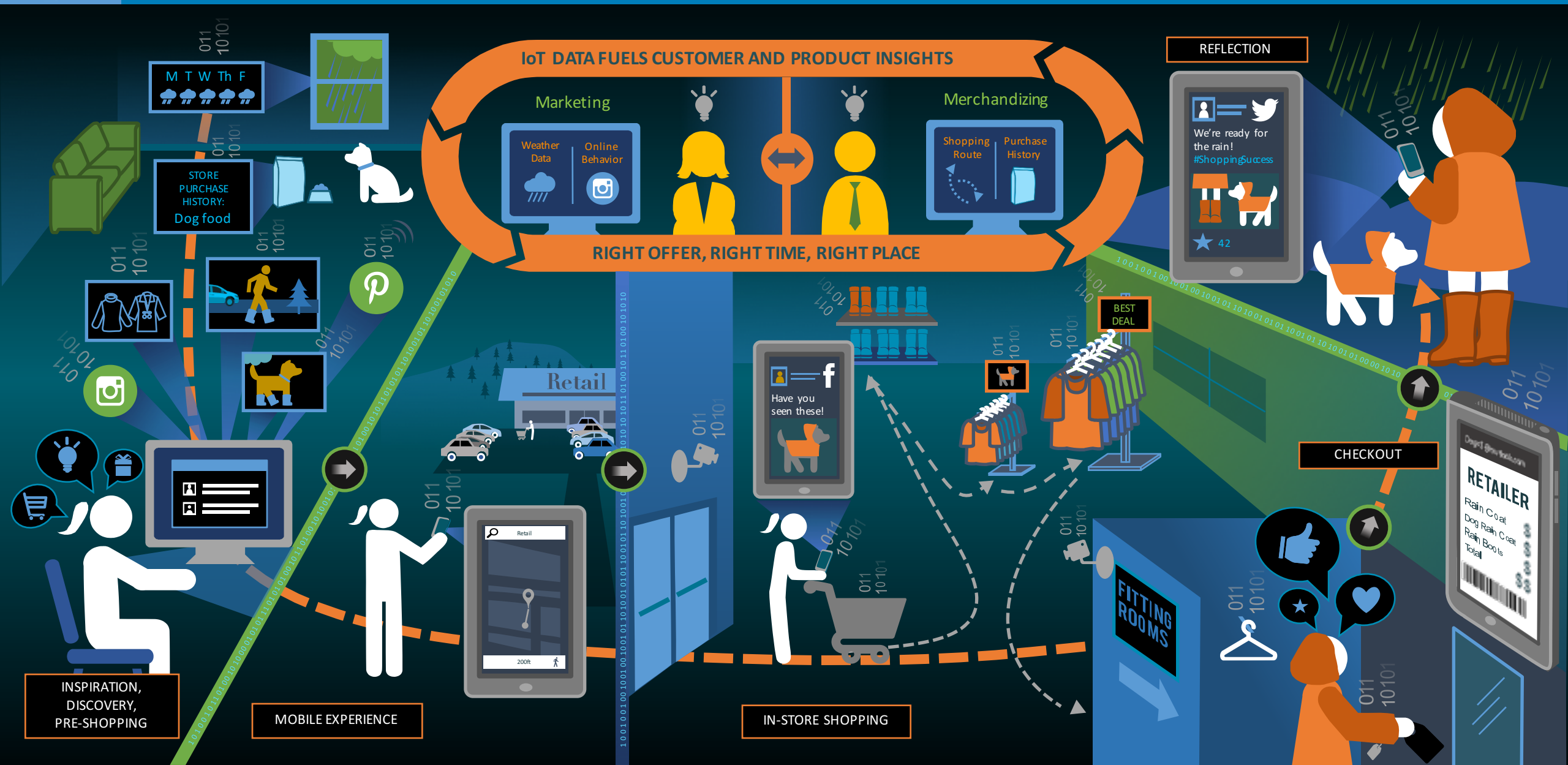
Make authorized patient data accessible from a unified point, enabling a holistic view of the patient's journey so providers can optimize each care interaction.



HEALTHCARE ECOSYSTEM

Integrate data from existing and non-traditional sources to drive Big Data analytics, enabling care process innovation and healthcare transformation.

The Internet of Things – Retail



CROSS-INDUSTRIAL USE CASES

based on Standard Parameters

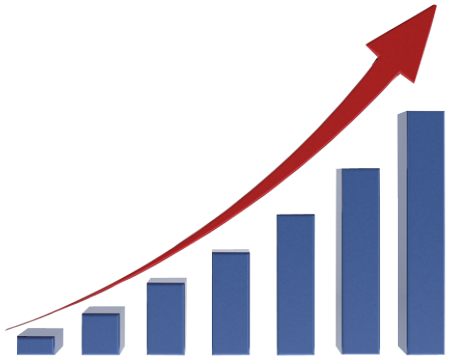
USE-CASE PARAMETERS

1. Company

2. Data Sources

3. Techniques

4. Business Value



EXAMPLE: DATA SOURCE

- Server Telemetry
- Monitoring Logs
- Memory Consumption
- Processing (CPU) Consumption
- Network Flow

EXAMPLE: TECHNIQUES

- Pattern Recognition
- Proactive Monitoring
- Early Alert Delivery

EXAMPLE: BUSINESS VALUE

PROFIT = REVENUE – COST

To increase this…

… increase this…

…or decrease this…

EXAMPLE: TELECOMMUNICATIONS

generic logo company

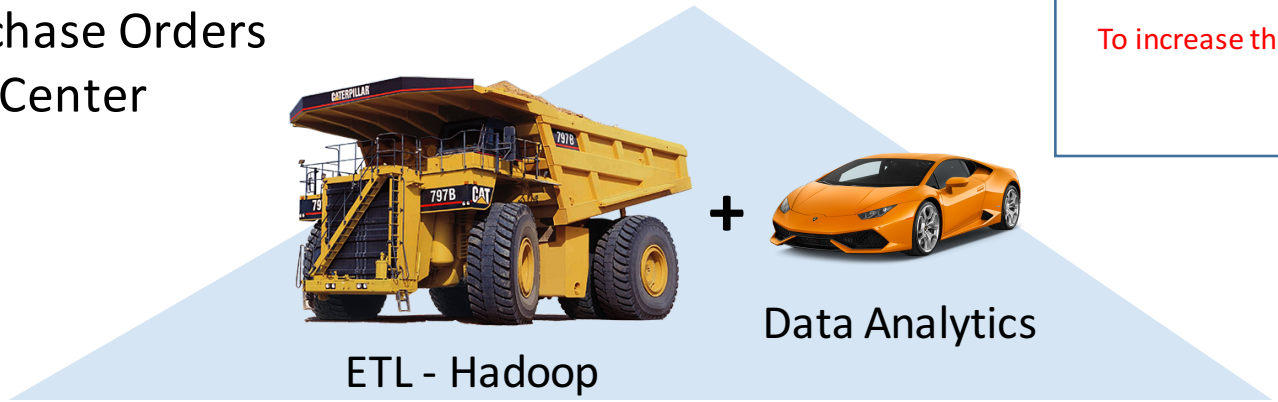
Problem: ETL Offload

Data Sources

- Customer Records
- Contract Data
- Purchase Orders
- Call Center

Techniques

- ETL
- Analytics

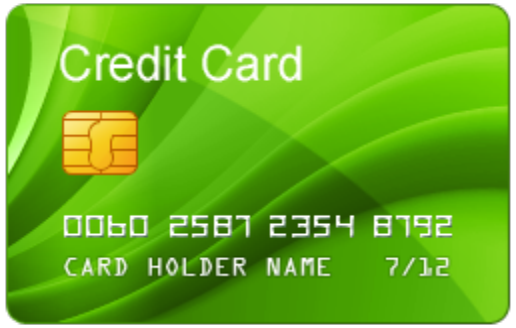


Business Value

PROFIT = REVENUE – COST

To increase this… … increase this… …or decrease this…

EXAMPLE: CARD ISSUER



Problem: Low Profit

Data Sources

- Customer Purchase History
- Merchant Designations
- Merchant Special Offers

Techniques

- ETL
- Machine Learning

Business Value

PROFIT = REVENUE – COST

To increase this…

… increase this…

…or decrease this…

EXAMPLE: WASTE & RECYCLING

Problem: Idle Alerts

Data Sources

- Truck Geolocation Data
  - 20 000 trucks
  - 5 sec interval
- Landfill Geo-boundaries

Techniques

- Realtime Streaming
- Batch Computation
- Graph Algorithms

Business Value

PROFIT = REVENUE – COST

To increase this… … increase this… …or decrease this…

+

APACHE STORM™

hadoop

hadoop Map Reduce

Short Path –
Graph Algorithms

+

- Immediate Alerts
- Tax Reduction
- Route Optimization

EXAMPLE: BANKING

Problem: Fraud Detection

Data Sources

- Anti-Money Laundering
- Consumer Transactions

Techniques

- Bayesian Learning
- Peer Group Analysis

Business Value

PROFIT = REVENUE – COST

To increase this… … increase this… …or decrease this…

+ hadoop +

= Suspicious Events

EXAMPLE: DNA ANALYSIS

ancestry.com®

Problem: Search Relevance, DNA Matching

Data Sources

- Birth, Death, Census,
- Military, Immigration Records
- Search Behavior Activity
- DNA SNP

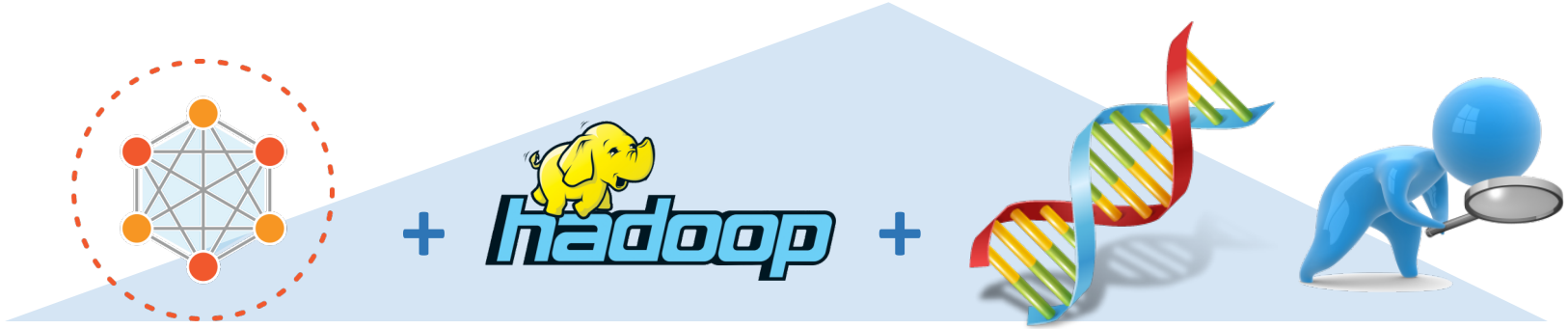
Techniques

- Record Linking
- Search Relevance
- Clickstream
- Security Forensics
- DNA Matching

Business Value

PROFIT = REVENUE – COST

To increase this… … increase this… …or decrease this…



Q & A

Abzetdin Adamov, Assoc Prof.
Email me at: aadamov@ada.edu.az
Follow me at: @
Link to me at: www.linkedin.com/in/adamov
Visit my blog at: aadamov.wordpress.com