DATA ENGINEERING

Prof. Lamia Ben Hiba
lamia.benhiba@um5s.net.ma
January 2026

Centrale
CASABLANCA

@lammiia

What is Data Engineering for you?

COURSE OBJECTIVES

Learning outcomes

- Demonstrate, through a case study, the impact of good data engineering on an enterprise's digital transformation
- Explain the foundations of data engineering
- Outline the definitions and concepts enabling gleaning value from Enterprise data
- Provide comprehensive and in-depth knowledge of the components of the data engineering lifecycle
- Develop technical competencies to implement data pipelines

COURSE OBJECTIVES

About this course

18

Total hours for this course

>50%

Sessions are dedicated to hands-on-labs

Unlimited

Number of hours to explore other aspects

Agenda

Overview of upcoming sessions

Data-Driven Enterprises

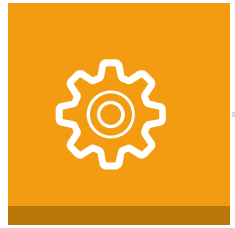
From Raw data to Decisions

Data Engineering

Data Engineering Lifecycle

Data Architecture

Data Architecture Design Patterns



Labs

Hands-on labs for data pipelines design and implementation

Closing

Data Quality, Ethical Concerns: Security, Privacy, etc.

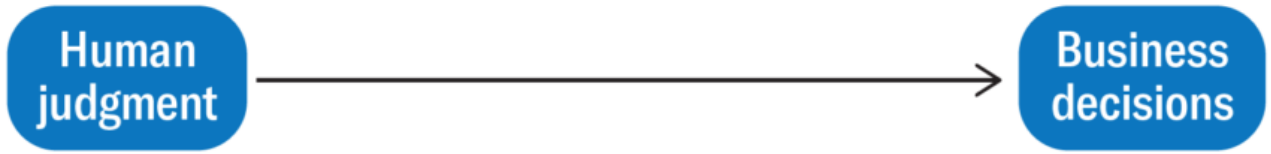
DATA-DRIVEN ENTERPRISE

Moneyball 2011

DATA DRIVEN ENTERPRISE

Decision Making

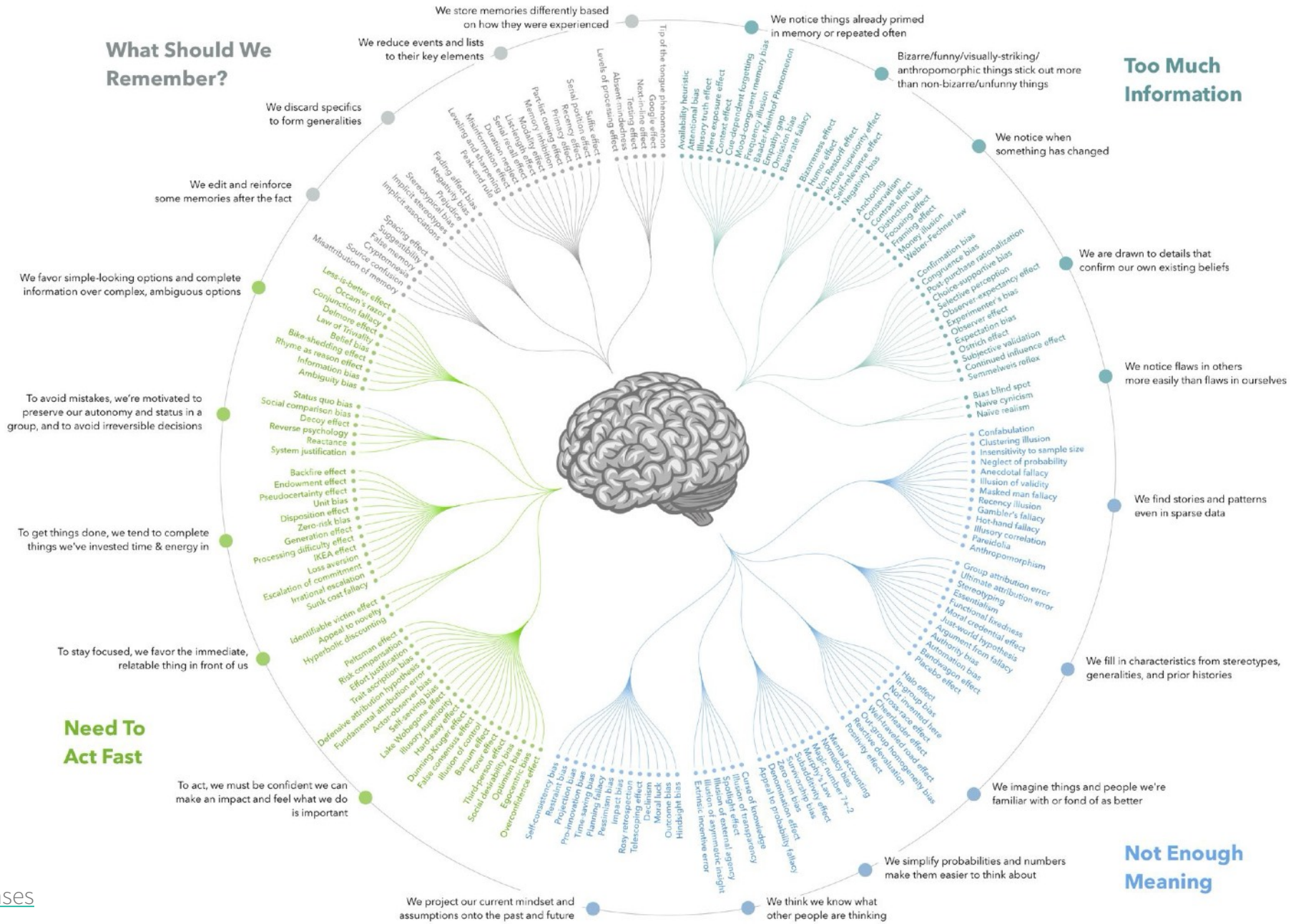
- 50-70 years ago, decision making was solely based on human intuition, developed from years of experience
- Relying exclusively on intuition is ineffective, capricious, faillible and limits an organization's capabilities



Source: Eric Colson

HBR

COGNITIVE BIAS CODEX, 2016



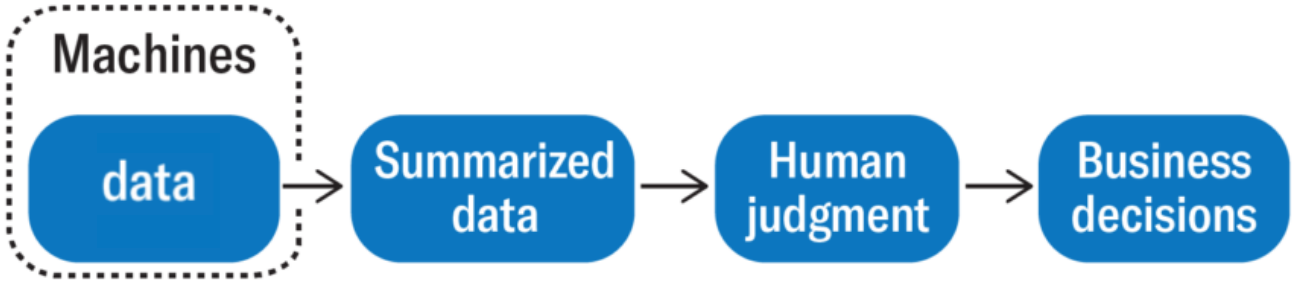
List of Cognitive Biases

Source: Better Humans

DATA DRIVEN ENTERPRISE

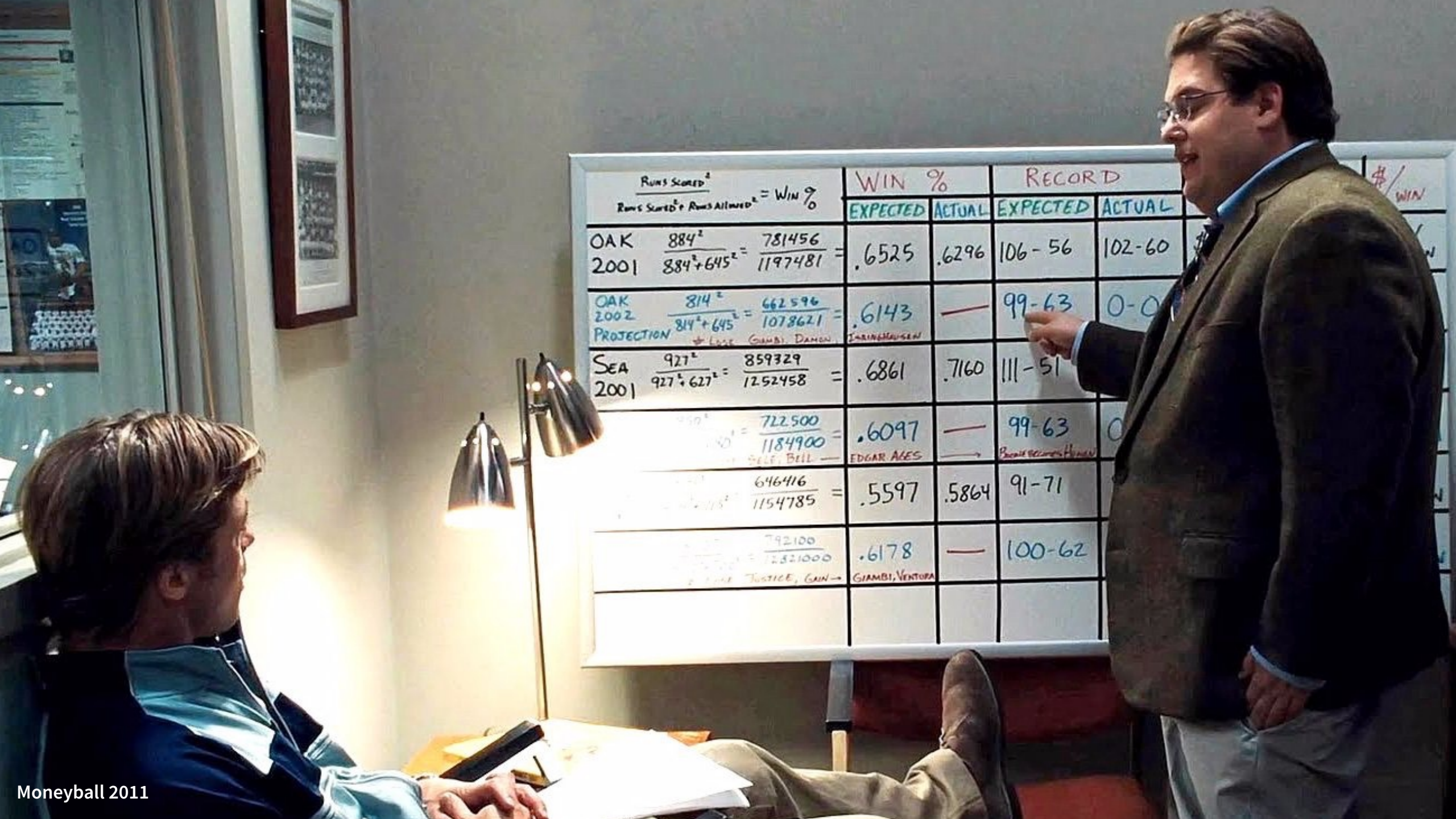
Decision Making

- Data-driven Decision Making is a process based on data captured from the organization's activities (transactions, customer behaviors etc.)
- Need to Aggregate and summarize unmanageable volumes of data, available in databases, system files etc.
- Consumption of synthesized data using Reporting/Monitoring and data visualization tools (Business Intelligence)



Source: Eric Colson

HBR

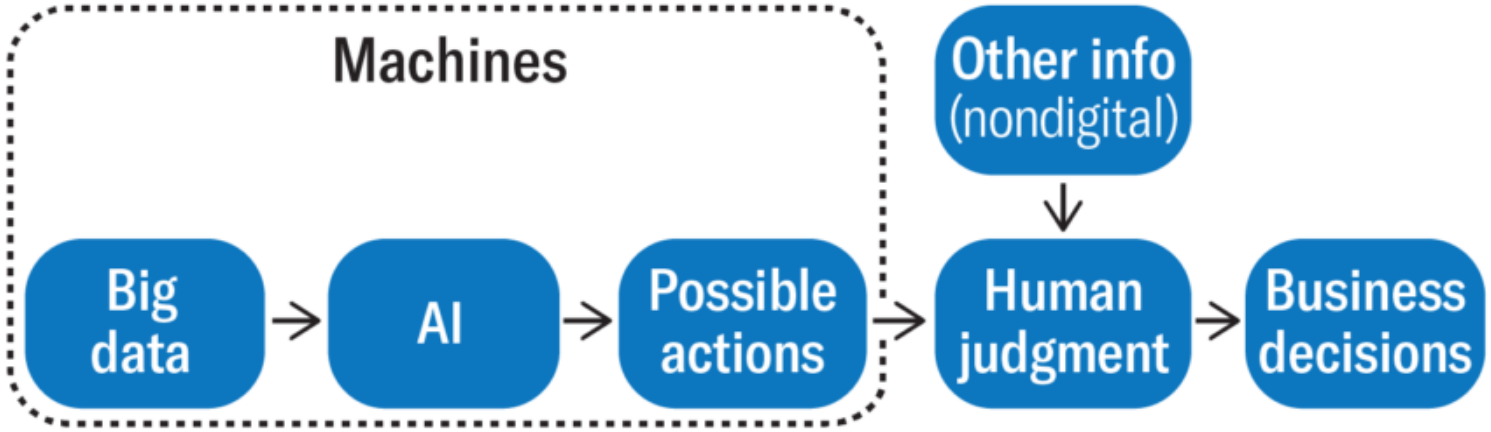


Moneyball 2011

DATA DRIVEN ENTERPRISE

Decision Making

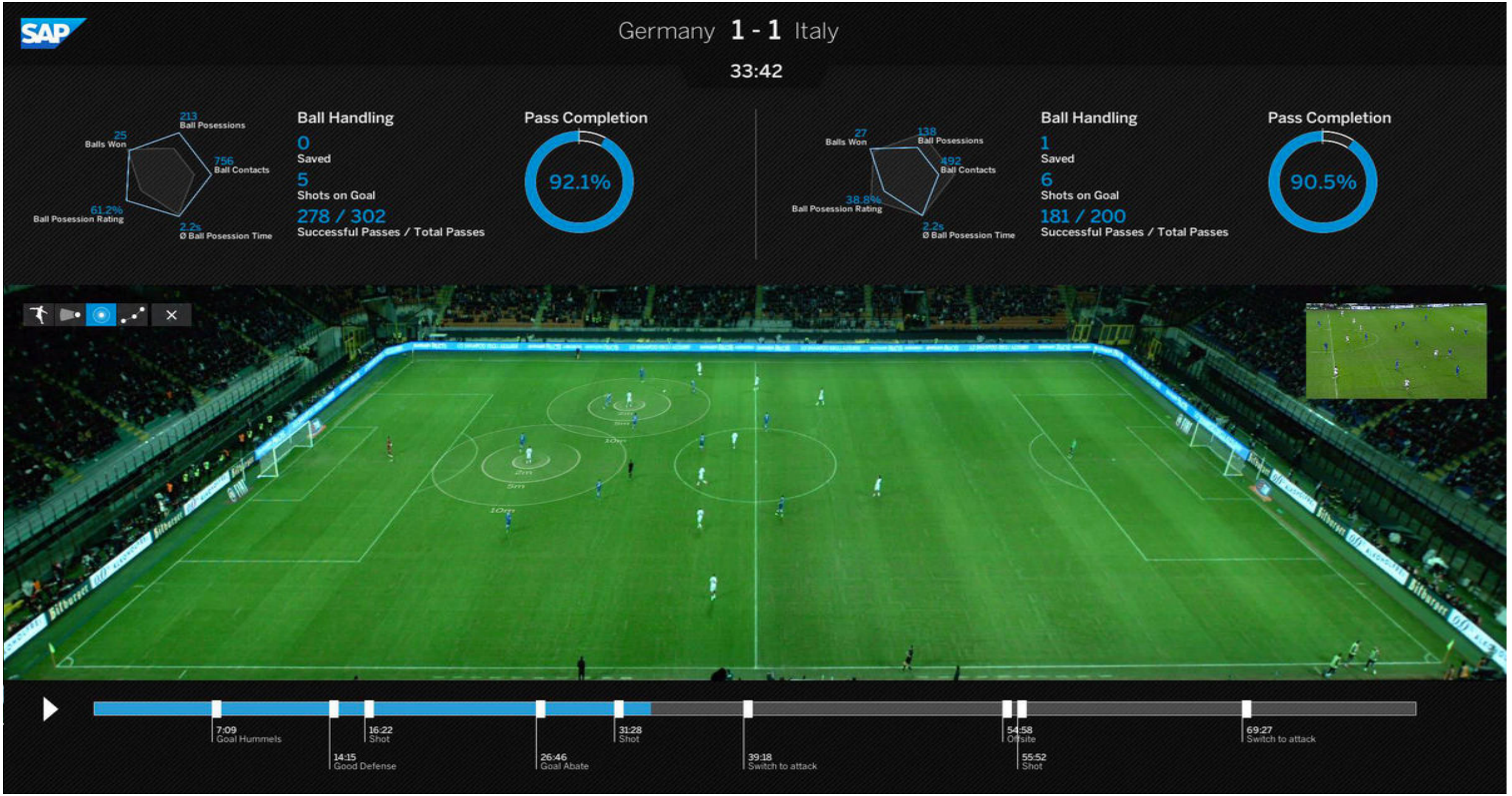
- Today, everyone is talking about “AI-driven Decision Making”
- Workflow benefits from the power of AI (ability to treat massive data, make sense of non linear relationships etc.) and human judgement that aligns with the organization’s culture, strategy, values and vision



Source: Eric Colson

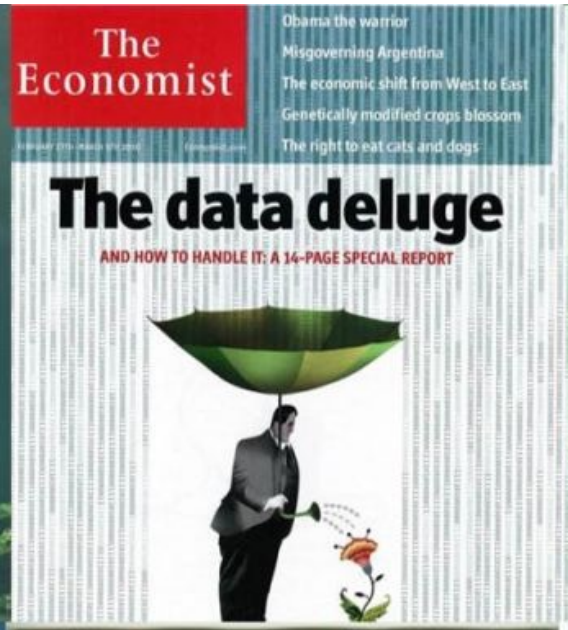
HBR

Image source: The Atlantic


SAP
Germany 1 - 1 Italy
33:42
213
Ball Posessions
25
Balls Won
756
Ball Contacts
61.2%
Ball Posession Rating
2.2s
Ø Ball Posession Time
Ball Handling
0
Saved
5
Shots on Goal
278 / 302
Successful Passes / Total Passes
Pass Completion
92.1%
27
Balls Won
138
Ball Posessions
492
Ball Contacts
38.8%
Ball Posession Rating
2.2s
Ø Ball Posession Time
Ball Handling
1
Saved
6
Shots on Goal
181 / 200
Successful Passes / Total Passes
Pass Completion
90.5%
7:09
Goal Hummels
14:15
Good Defense
16:22
Shot
26:46
Goal Abate
31:28
Shot
39:18
Switch to attack
54:58
Offsite
55:52
Shot
69:27
Switch to attack

Source: SAP


significance
statistics making sense
BIG DATA
SPECIAL ISSUE
DATA AND THE CITY
WHO ARE YOUR FACEBOOK FRIENDS?
MEET THE DATA FAMILY


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT


nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings


Science
AAAS
SPECIAL ISSUE
TECHNOLOGIES TRANSFORMING BIOLOGY


POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
OFFICER ALGORITHM
NEW WAYS OF SEEING
PLUS
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE


Harvard Business Review
The True Measure Of Success
10 Rules for Managing Global Innovation
What Ever Happened To Accountability?
GETTING CONTROL OF BIG DATA


Government Executive
MARCH/APRIL 2013
Does Government Ever Really Change?
Five Ways Women Get Ahead
BIG DATA
HOW IT'S MAKING AGENCIES SMARTER


TIME
A tale of two Libyas
The GOP's misinformation campaign
Could your baby be depressed?
Word up: A dictionary of slang
YOUR DATA FOR SALE
Everything about you is being tracked—get over it
BY JOEL STEIN
Owns a laptop
Age: 38-39
Likes: online news
Likes: Asian cuisine
Likes: green living
Sister is a lawyer
Spent $180 on intimate apparel & undergarments on Oct. 10, 2010
Dislikes: autos & vehicles
Married
Smart-phone user
Has had LASIK surgery
Likes: restaurants

DATA DRIVEN ENTERPRISE

Data Deluge

4%

Can exploit the totality of their data

36%

Do not have the necessary competencies

43%

Derive little tangible benefits from their data

Source: Seizing the information advantage - pwc.co.uk

CASE STUDY

- References:
  - Anderson-Lehman, Ron, et al. "Continental Airlines Flies High with Real-time Business Intelligence."
  - Wixom, Barbara H., et al. "Continental airlines continues to soar with business intelligence."

- Go Forward business plan : Real time BI and Data warehousing

CASE STUDY

Organization: Continental Airlines

- Founded in 1934, Houston, Texas.
- 5th biggest airline company in USA, 7th internationally
- 50 Millions of passengers per year, 2300 daily flights to more than 227 destinations

CASE STUDY

Before the Go-Forward Plan

- On the 10 American airline compagnies, Continentale was **ranked 10th** in performance (delays, complaints, overbooking etc.)
- Facing financial problems, the company was in danger of bankruptcy
- Information was not available. **No consolidation** of data scattered across the company
- Lack of visibility on multi-transit flights: market and consumer behavior studies were impossible, no optimisation undertaken
- Externalized IS, limited periodic reports, no support of ad hoc queries
- Each department had its own **approach** for data management and reporting

CASE STUDY

The Go-Forward Plan

- 4 interconnected parts, executed simultaneously:
  - *Fly to win*: Identify products that customers like and are prepared to buy
  - *Fund the future*: Adjust costs and cash flows
  - *Make Reliability a Reality*: Guarantee reliability to clients (delay, luggages, security…)
  - *Working together*: Create an enterprise culture in favor of the plan's success

CASE STUDY

The Go-Forward Plan

- Implementation of a Datawarehouse for decision support mainly for the management of **revenue and pricing**: 6 months of development, multitude of data sources: flight, customers, stocks data etc.
- Centralisation and consolidation of data (flights, customers, finance, security etc.)
- Applications integrating these data such as **“Demand-driven Dispatch”** ($5 millions / year of additional revenue), **“Goodwill Letters”** ($6 millions in revenue), **“Group Snoop”** ($2 millions savings)

CASE STUDY

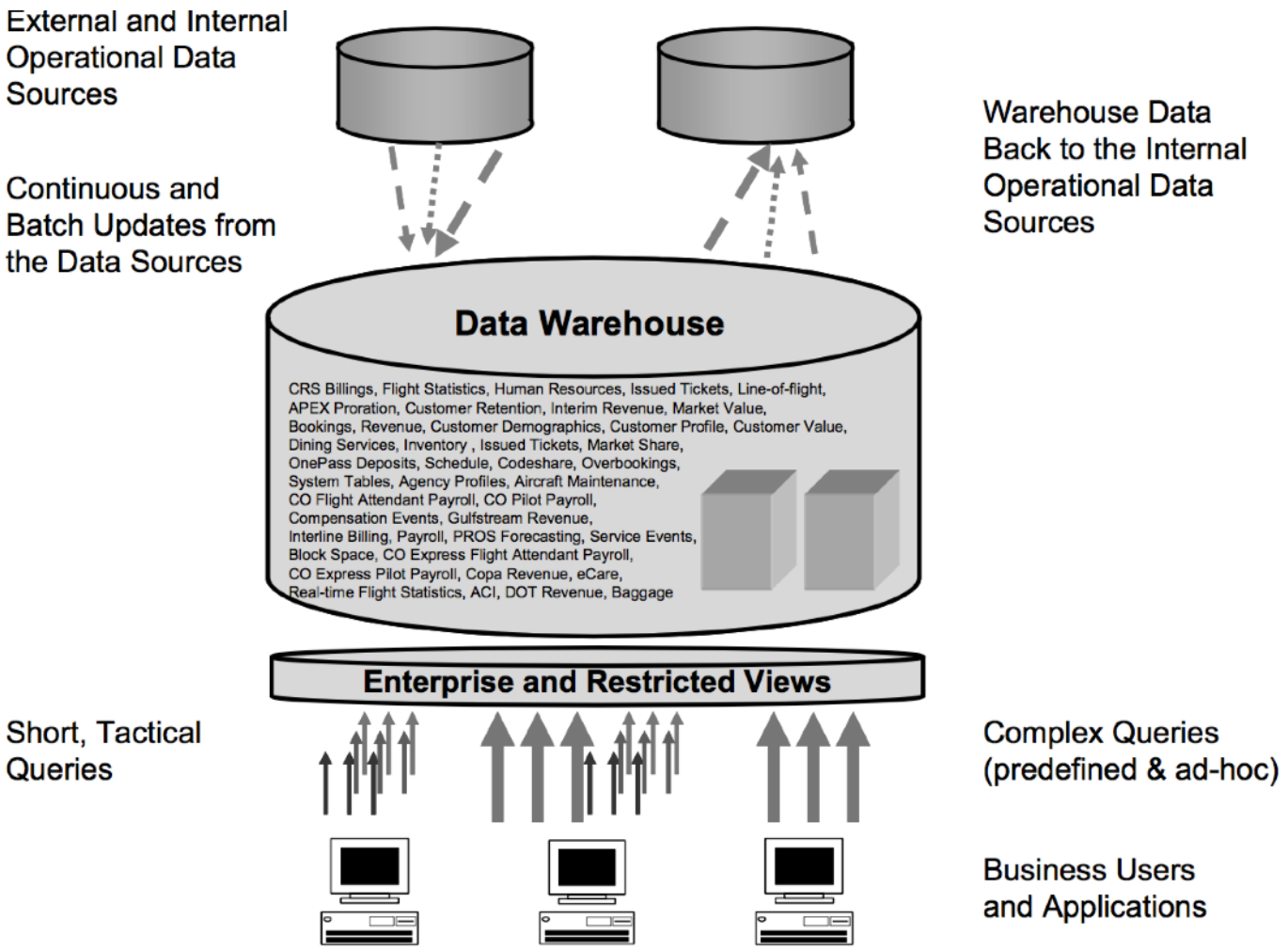
The Go-Forward Plan

- Transition to a Datawarehouse providing information in **real-time**, actionable, that support tactical decision-making process across the organization and business processes
- Implementation of Datawarehouse-based real-time applications (Fare design, tracking flight reservation, customer segmentation , market studies….)
- Implementation of a **flight management dashboard** to help identify flight issues quickly and consequently manage reservations in ways to guarantee customer satisfaction and airline profitability

CASE STUDY

The Go-Forward Plan

- Datawarehouse with 42 themes, 35 datamarts, 29 applications and 1292 users
- Critical information from analyses are fed back into operational systems
- Real time >> Right Time (seconds or batch mode in hours)



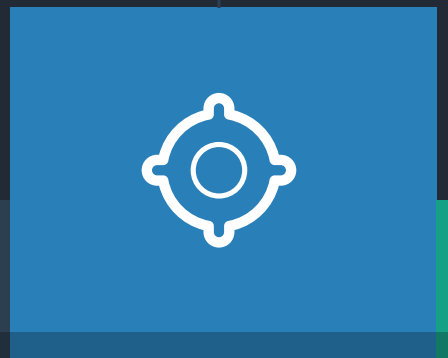
Anderson-Lehman, Ron, et al. "Continental Airlines Flies High with Real-time Business Intelligence."

CASE STUDY

Post Go-Forward Plan

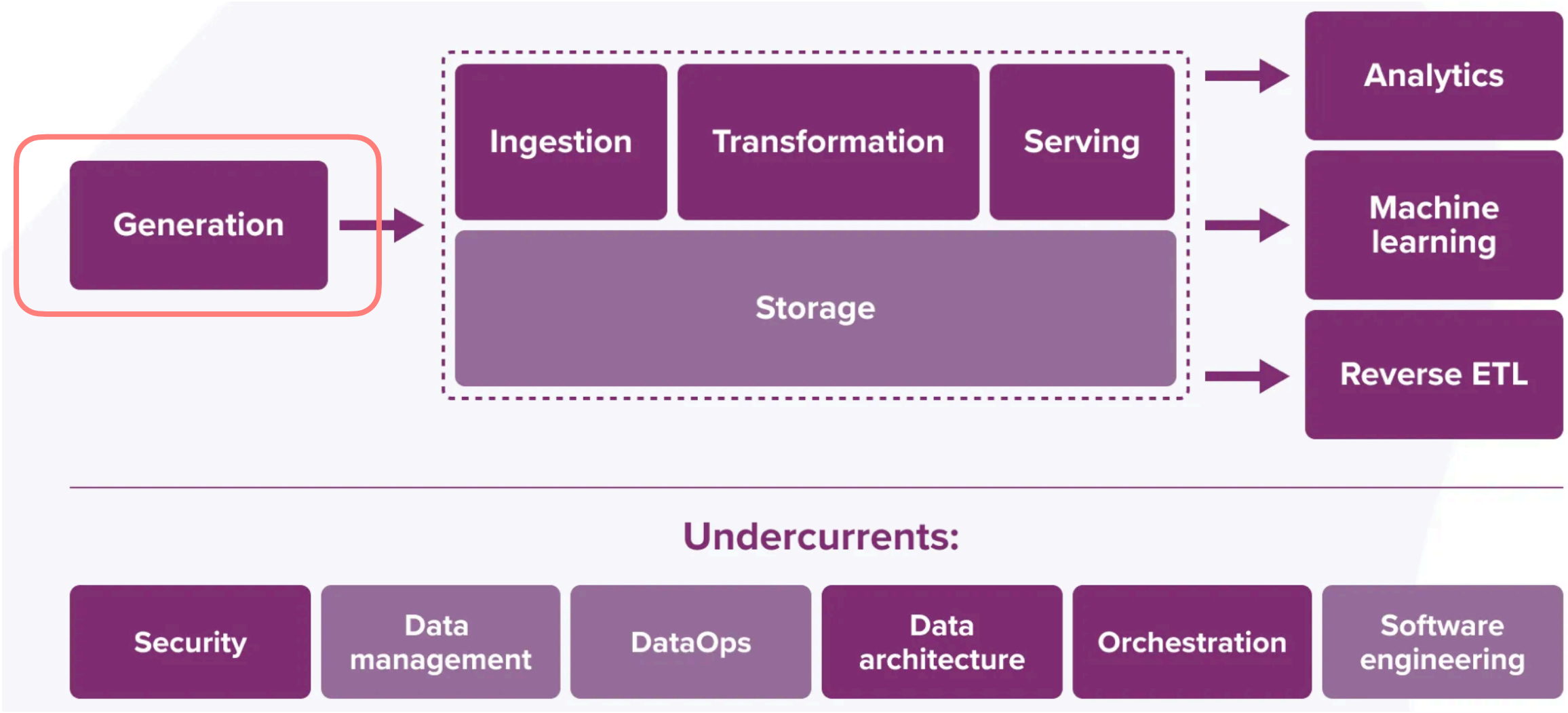
- On a **30M$ investment on 6 years**, ROI of more than 1000%: **More than 500M$ in minimized costs and generated revenue**
- Optimized fares, better fraud detection
- Increasing rate of customer satisfaction
- **Ranked 1st American Airline Company (2004)**
- From **First** to **Favorite**

DATA ENGINEERING LIFECYCLE



DATA ENGINEERING LIFECYCLE

Value chain



Source: Fundamentals of data engineering - Wiley

DATA ENGINEERING LIFECYCLE

Data Generation

- Data is generated by everything from cameras and traffic sensors to heart rate monitors, enabling richer insights into human and “thing” behavior
- In organizations, data can originate from external sources (third-party datasets), or be a by-product of a particular business process (sales, stock, patient care etc.) often managed in databases or file systems
- Data Generation marks the starting point of the data engineering lifecycle and encompasses the raw material of all downstream processes
- Understanding source systems includes defining what data exists, where it comes from, and how it enters the data ecosystem

DATA ENGINEERING LIFECYCLE

Data Generation

- Data can be generated from various **sources**:
  - Surveys
  - Web data
  - APIs
  - Operational Databases (OLTP systems: Relational or NoSQL)
  - OLAP systems (Data Warehouse…)
  - External data (from third parties)
  - Logs
  - Sensors (IoT)
  - ….

DATA ENGINEERING LIFECYCLE

Data Generation - Data types

- We distinguish different types of data:
    - Manual vs. Automated
    - Internal vs. External
    - Quantitative vs. Qualitative
    - Raw vs. derived
    - Structured, semi-structured, unstructured
    - Static vs. Streaming

DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Manual vs. Automated

- Manual data originate from human input (survey entries or clinical notes). It is prone to errors and can introduce subjectivity
- Automated data is generated by systems, sensors, or applications that capture events continuously with minimal human intervention

DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Internal vs. External

- Internal data is produced within the organization's own systems
- High control and governance
- External data can be sourced from partners, APIs, or open datasets
- It extends analytical scope but raises concerns about reliability, format heterogeneity, and access limitations

EQUIFAX

EPSILON

Datarade

AWS Data Exchange

DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Quantitative vs. Qualitative

- Quantitative data consist of numeric records (e.g. how many; how much; or how often)
- They can be expressed in:
  - Discrete: data that can only take exact values
  - Continuous: there is an infinity of intermediary values between any two values

DATA ENGINEERING LIFECYCLE

Data Generation - Data types

DATA ENGINEERING LIFECYCLE

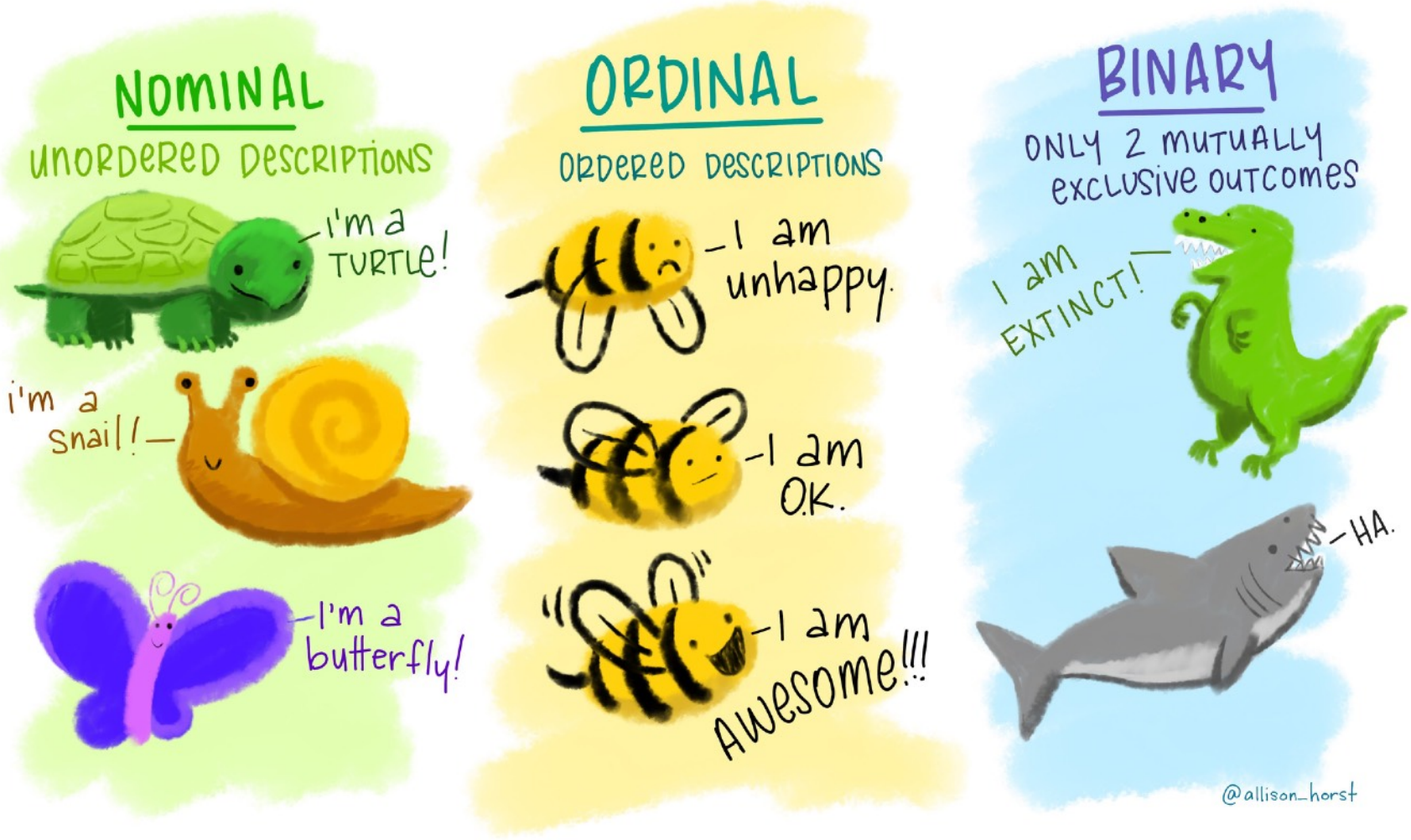
Data Generation - Data types

Quantitative vs. Qualitative

- Qualitative data are non numerical, categorical data (e.g. what type) that can be expressed as:
  - Nominal data: categories that are mutually exclusive without any numerical value
  - Binomial data: place things in one of two mutually exclusive categories
  - Ordinal data : ordered categories where certain observations are greater than others

DATA ENGINEERING LIFECYCLE

Data Generation - Data types



DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Raw vs. Derived

- Raw data do not necessitate any treatments before use (individual traffic counts through an intersection)
- Derived data are produced through additional processing or analysis of captured data (total number of counts or counts per hour)
- Raw data are the input of the data value chain, derived data are created in the process

DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Raw vs. Derived

- Primary data Data collected by the Enterprise itself for a specific purpose
- Secondary data are made available to others to reuse and analyse that are generated by someone else
- Tertiary data are a form of derived data, such as counts, categories, and statistical results
- Primary data are generally tailored to the specific needs and focus of the Enterprise

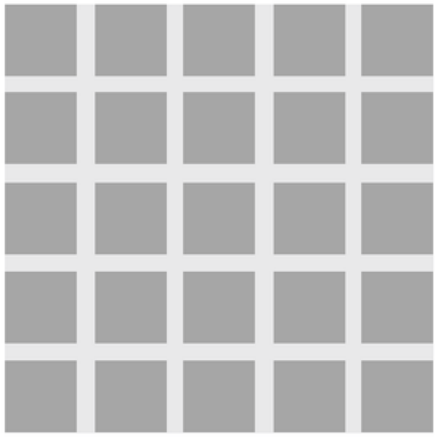
DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Structured vs. Unstructured

- Structured data can be easily organized, stored and transferred in a defined data model set out in a table or relational database that have a consistent format
- Structured data can be processed, searched, queried, combined, and analyzed relatively straightforwardly and easily processed by computers

Structured data



Database, CRM, ERP

Unstructured data

Text, audio, videos

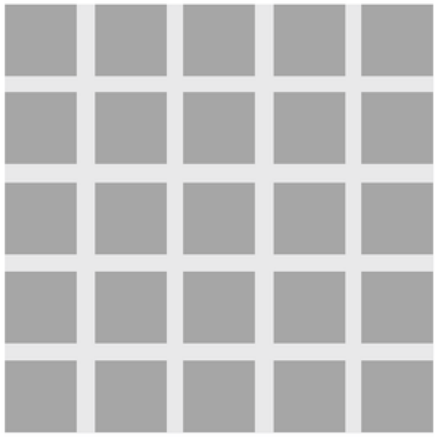
DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Structured vs. Unstructured

- Semi-Structured data are loosely structured data that have no predefined data model/schema and thus cannot be held in a relational database
- Their structure are irregular, implicit, flexible and often nested hierarchically, but they have a reasonably consistent set of fields and the data are tagged
- We can thus separate content semantically, define metadata, sort, order and structure the data to some extent

Structured data



Database, CRM, ERP

Unstructured data

Text, audio, videos

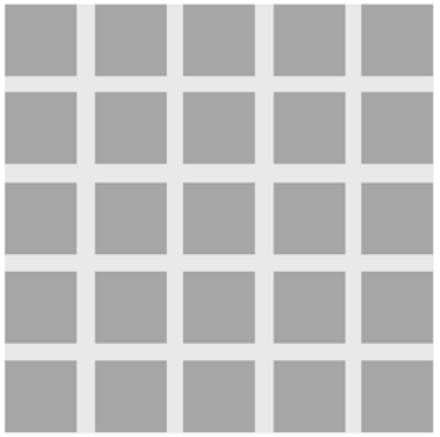
DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Structured vs. Unstructured

- Unstructured data do not have a defined data model or common identifiable structure
- Each individual element, such as narrative text or photo, may have a specific structure or format, but not all data within a dataset share the same structure
- while they can often be searched and queried, they are not easily combined or computationally analyzed

Structured data



Database, CRM, ERP

Unstructured data

Text, audio, videos

DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Static vs. Streaming

- Static data doesn’t change frequently and typically generated or updated at fixed intervals.
- Static data is usually captured as snapshots or batches
- Streaming (or real-time) data is continuously produced and transmitted as events occur
- Streaming data is often generated by systems, applications, or sensors that publish updates immediately

DATA ENGINEERING LIFECYCLE

Data Generation - Data types

Dimensions	Data Types		
Personal	Volunteered	Observed	Inferred
	Private		Public
	Identified		Pseudonymised
Non-Personal	Anonymous		Machine Data
Timeliness	Instant/Live		Historic

Source: GSMA