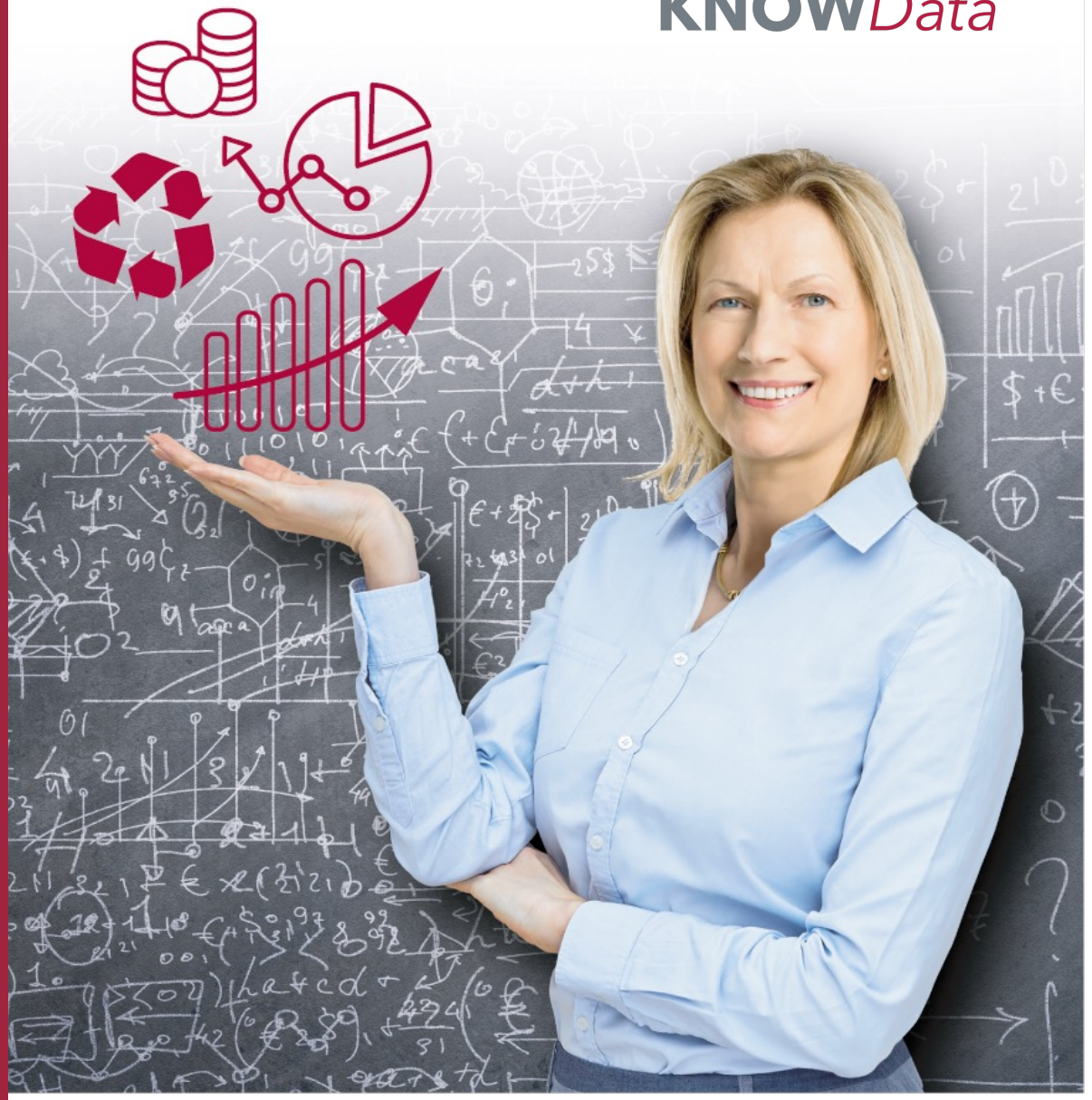
Big Data

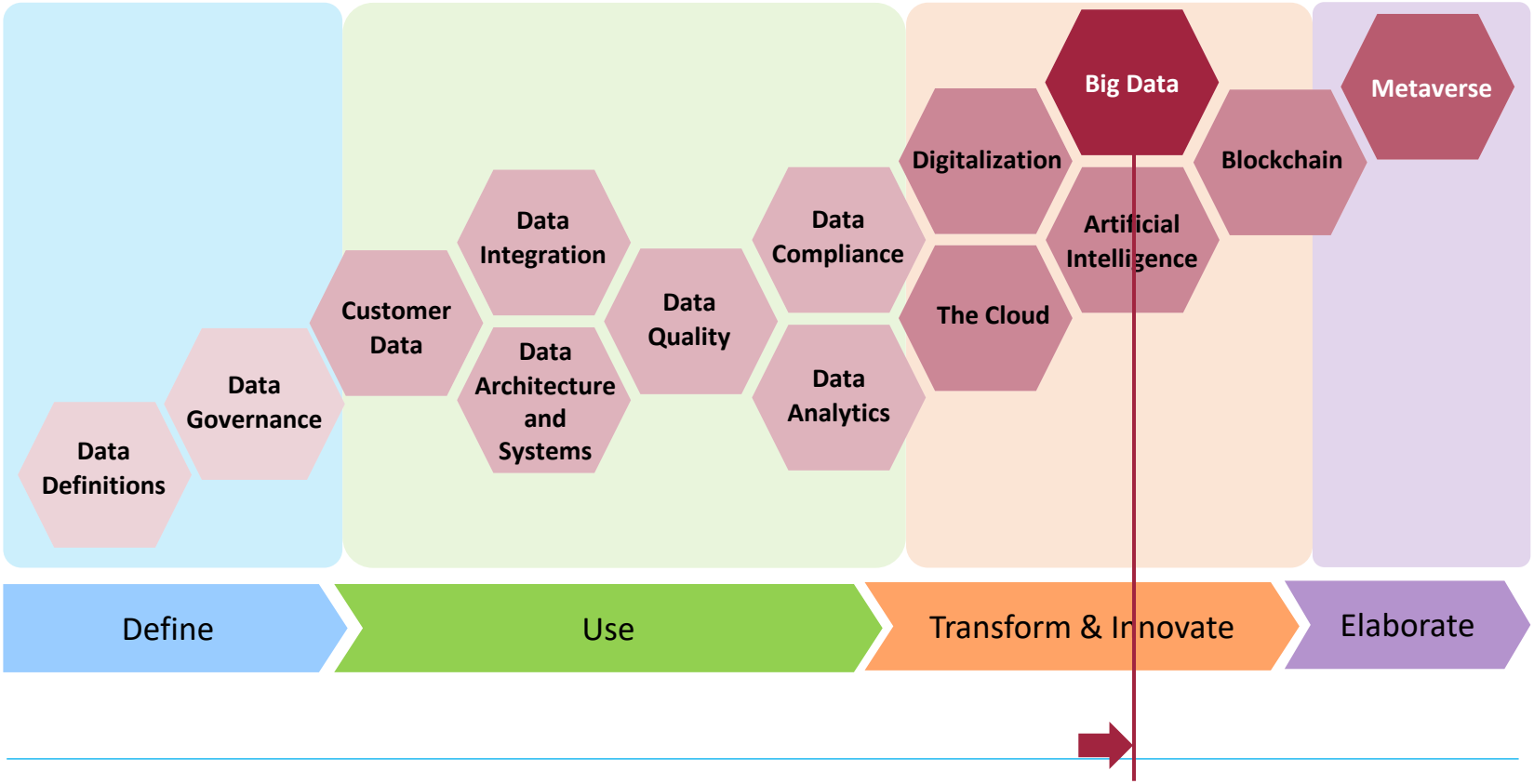


KNOWData

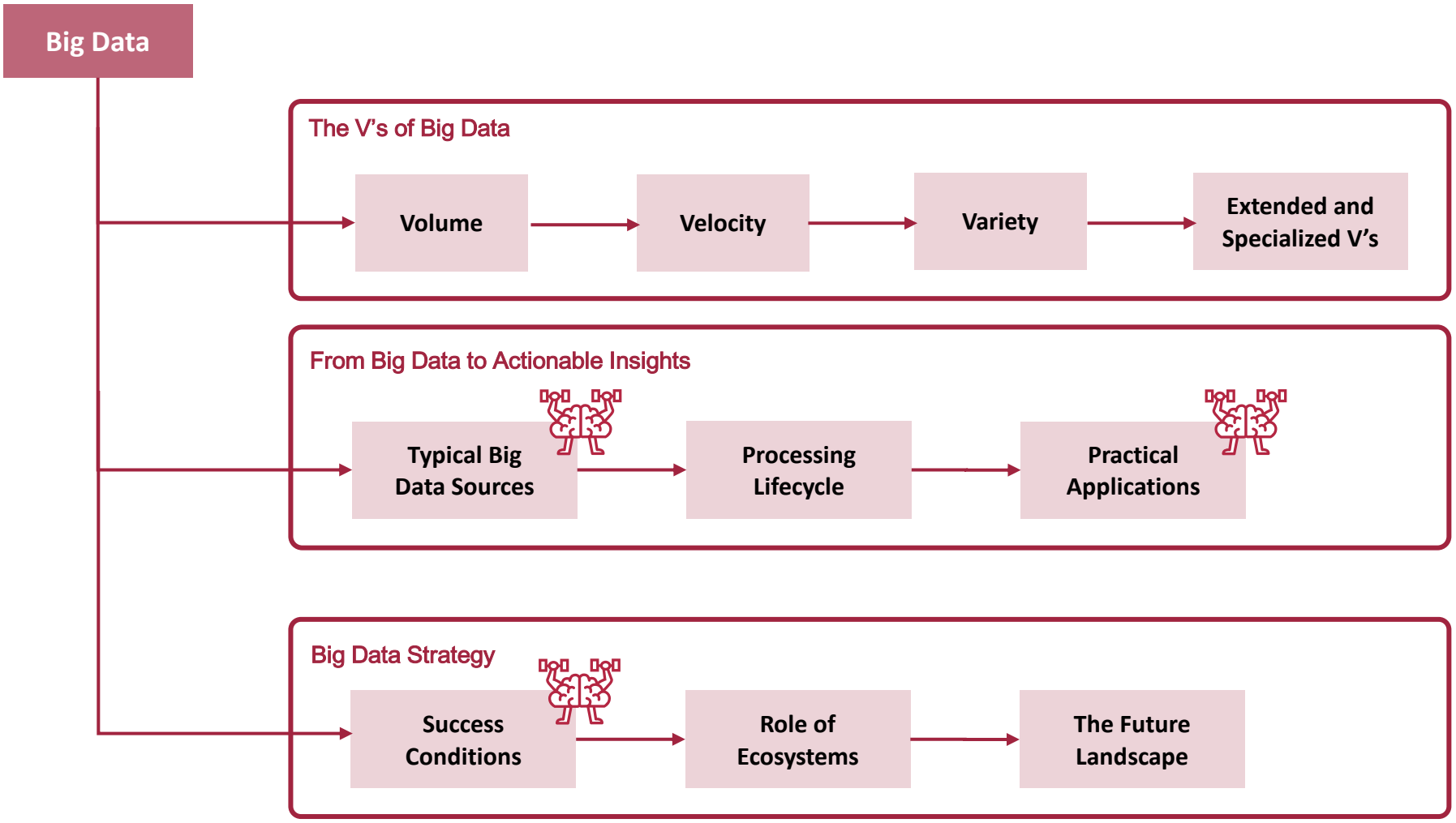


November 11, 2026

Lionel Pilorget


Data Definitions
Data Governance
Customer Data
Data Integration
Data Architecture and Systems
Data Quality
Data Compliance
Data Analytics
Digitalization
The Cloud
Big Data
Artificial Intelligence
Blockchain
Metaverse
Define
Use
Transform & Innovate
Elaborate

Structure of Lesson 11



Data Volume

Massive amount of data generated, collected, stored, and processed by organizations every day
In Big Data, this refers to datasets that are too large to be processed by traditional databases or software.

Synonyms

Data size, Data scale, Data magnitude, Data quantity, Data mass

Core aspects

 Exponential Growth : Data is increasing rapidly (social media, IoT, sensors, transactions)

 Storage Requirements: Requires scalable storage solutions (cloud, distributed systems)

 Processing Power: Large datasets demand high computational capacity

 Structured & Unstructured Data: Includes databases, text, images, videos, logs, etc.

 Measured in Big Units: Terabytes (TB), Petabytes (PB), Exabytes (EB), Zettabytes (ZB)

Usage Examples of Data Volume

🌍 Social Media
- Facebook generates 4+ petabytes of data per day
- Millions of photos, videos, comments uploaded every hour

🎥 Video Streaming
- Netflix streams billions of hours of video monthly
- One HD movie ≈ 3–6 GB of data

🛒 E-Commerce
- Amazon processes millions of transactions daily
- Customer clicks, searches, reviews, and purchase history stored

🏥 Healthcare
- One MRI scan ≈ 100–200 MB
- Hospitals store petabytes of patient records and imaging data

🌐 Internet of Things (IoT)
- Smart factories generate terabytes of sensor data daily
- Connected cars produce gigabytes of data per hour

Data Velocity

The speed at which data is generated, produced, created, or refreshed In Big Data, this refers to the rate of change and the need for real-time or near-real-time processing.

Synonyms

Data flow rate, Stream frequency, Real-time throughput, Data speed

Core aspects

⏱️ Real-Time Generation: Continuous data from sensors, apps, financial markets, social media

🔄 Streaming Data: Data processed while it is being generated

⚡ Low Latency Requirements: Immediate response needed (e.g., fraud detection)

🌊 Continuous Data Flow: Not stored first, often processed instantly

🤖 Automated Processing: AI/algorithms analyse data instantly

Usage Examples of Data Velocity



💳 Fraud Detection in Banking

- Transactions analysed within milliseconds
- Suspicious activity blocked instantly

🎮 Online Gaming

- Player actions processed in milliseconds
- Real-time multiplayer synchronization
- Instant updates for scores, movements, and events

📱 Social Media Feeds

- Posts, likes, and comments update instantly
- Trending topics detected in real time

🚗 Autonomous Vehicles

- Sensors generate data continuously
- Immediate processing required for braking or steering

Data Variety

Different types and formats of data that are generated, collected, and analysed
In Big Data, this refers to the diversity of data sources, formats, and structures - from structured numerical data to unstructured text, images, and video.

Synonyms

Data diversity, Data heterogeneity, Multi-format data, Source multiplicity, Mixed data sources

Core aspects

- 📦 Structured Data: Tables, databases, spreadsheets
- 📝 Semi-Structured Data: JSON, XML, log files
- 🎥 Unstructured Data: Text, images, videos, audio, social media posts
- 🌐 Multiple Sources: Social media, sensors, transactions, emails, devices
- 🔄 Integration Challenges: Different formats require transformation and preprocessing

Usage Examples of Data Variety

🏎️ Formula 1 Car Telemetry

- Structured data (speed, RPM, tire pressure)
- Sensor streams (semi-structured logs)
- Onboard cameras (unstructured video) + team radio (unstructured audio)

🏠 Smart Home

- Voice recordings (audio - unstructured)
- Device status data (structured)
- Interaction logs (semi-structured)

🚗 GPS Coordinates

- Location data (structured coordinates)
- Route history (time-series data)

⌚ Wearables

- Heart rate & step count (structured sensor data)
- Sleep tracking logs (semi-structured)
- Health reports & activity summaries (mixed formats)

The Extended V's as Pillars of Maturity

Volume, Velocity & Variety define the problem.
Veracity, Value & Variability define the solution.

VERACITY ("Not all data is created equal")

- Definition: The quality, trustworthiness, and provenance of the data
- The Challenge: 80% of data science work is cleaning messy data. How do we quantify uncertainty?
- Key Question: Can we trust this signal, or is it noise?
- Example: Social media sentiment analysis – are these reviews real users or bots?

VALUE ("Data is worthless until it serves a purpose")

- Definition: The ability to transform data into actionable and measurable benefit
- The Challenge: How do we measure ROI on data? Avoiding "data rich, insight poor" syndrome
- Key Question: So what? Why are we collecting this?
- Example: Predictive maintenance – data isn't valuable until it prevents a machine from breaking

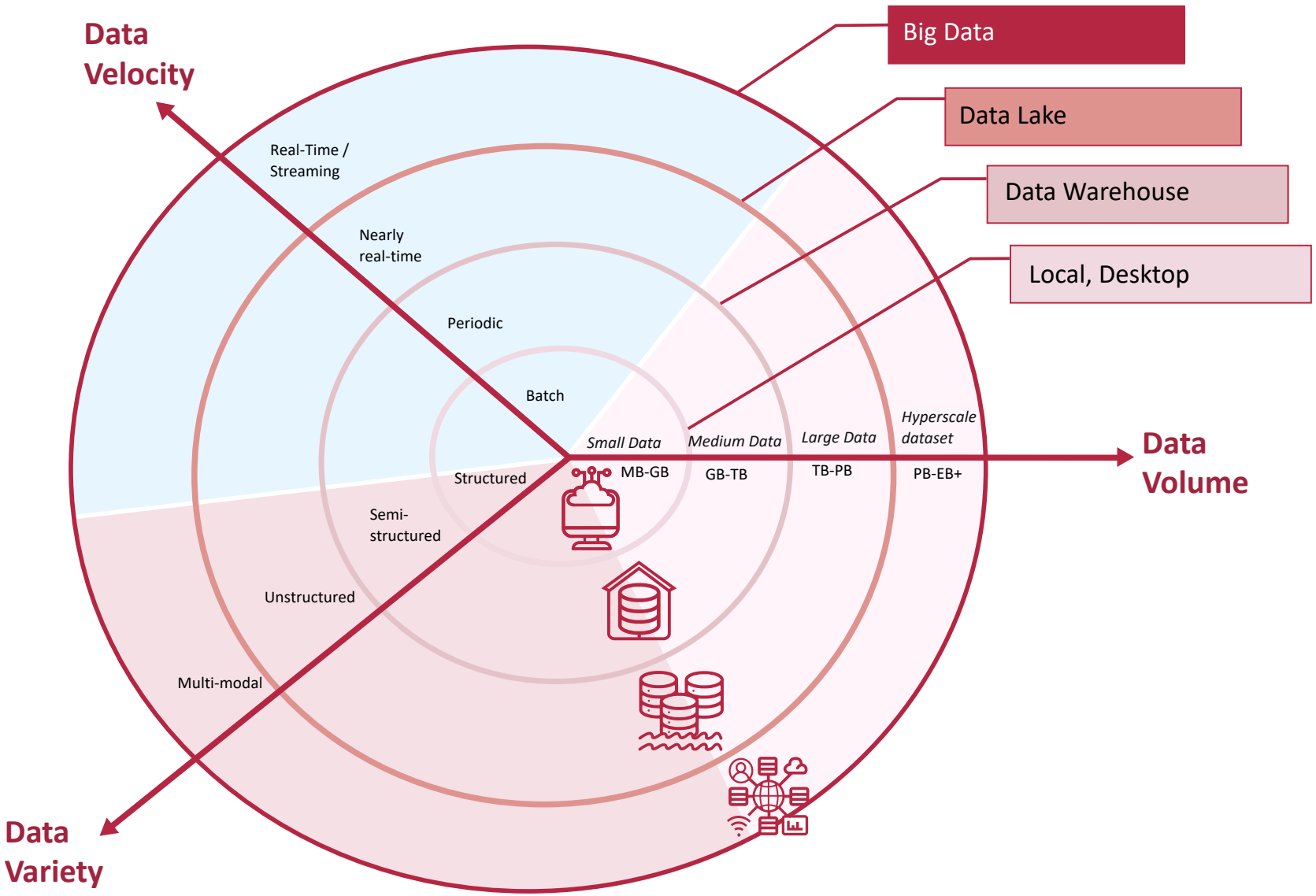
VARIABILITY ("Meaning is not static")

- Definition: The change in the meaning and distribution change over time/context
- The Challenge: Models trained on yesterday's data may be wrong today
- Key Question: Is this trend real, or is the context shifting?
- Example: Consumer behaviour during COVID-19 – historical shopping data became immediately obsolete

Specialized V's of Big Data

Specialized V	Context	Definition	Examples
Validity	Statistics, ML evaluation, scientific research	Degree to which data and models correctly measure what they intend to measure	Biased sampling, flawed survey design, overfitted ML model
Venue	Distributed Systems, platform studies, IoT	The heterogeneity of data sources and the jurisdictional/physical location where data resides, impacting latency, governance, and access	Social media data vs. industrial sensors vs. ERP systems
Vagueness	NLP, human-centered data, Cognitive Science	Ambiguity or imprecision inherent in data, human language and concepts where boundaries are not sharply defined	Sentiment labels ("positive"), subjective annotations, vague categories
Vocabulary	Semantic web, interoperability, data integration	The terminology, ontology, and metadata standards used to describe data	Internal Classification of Diseases, Resource Description Framework, schema mismatches
Volatility	Storage architecture, compliance, real-time systems	Lifespan and temporal stability of data relevance, the data retention policy	Streaming data decay, GDPR retention limits, log file rotation

Big Data Zone



From Big Data to Actionable Insights



Generic Sources of Big Data

Almost all Big Data can be traced back to four generic origins: Humans, Machines, Transactions, and Biology

Humans
Data generated directly by human communication, expression, or behaviour
Examples: Emails, messages, photos, videos, speech recordings, social media content

Machines
Data automatically generated by technical systems and sensors
Examples: IoT sensor data, server logs, GPS signals, industrial telemetry, network traffic

Transactions
Data created through economic, administrative, or digital exchanges
Examples: Purchases, banking operations, ERP / CRM entries, online orders

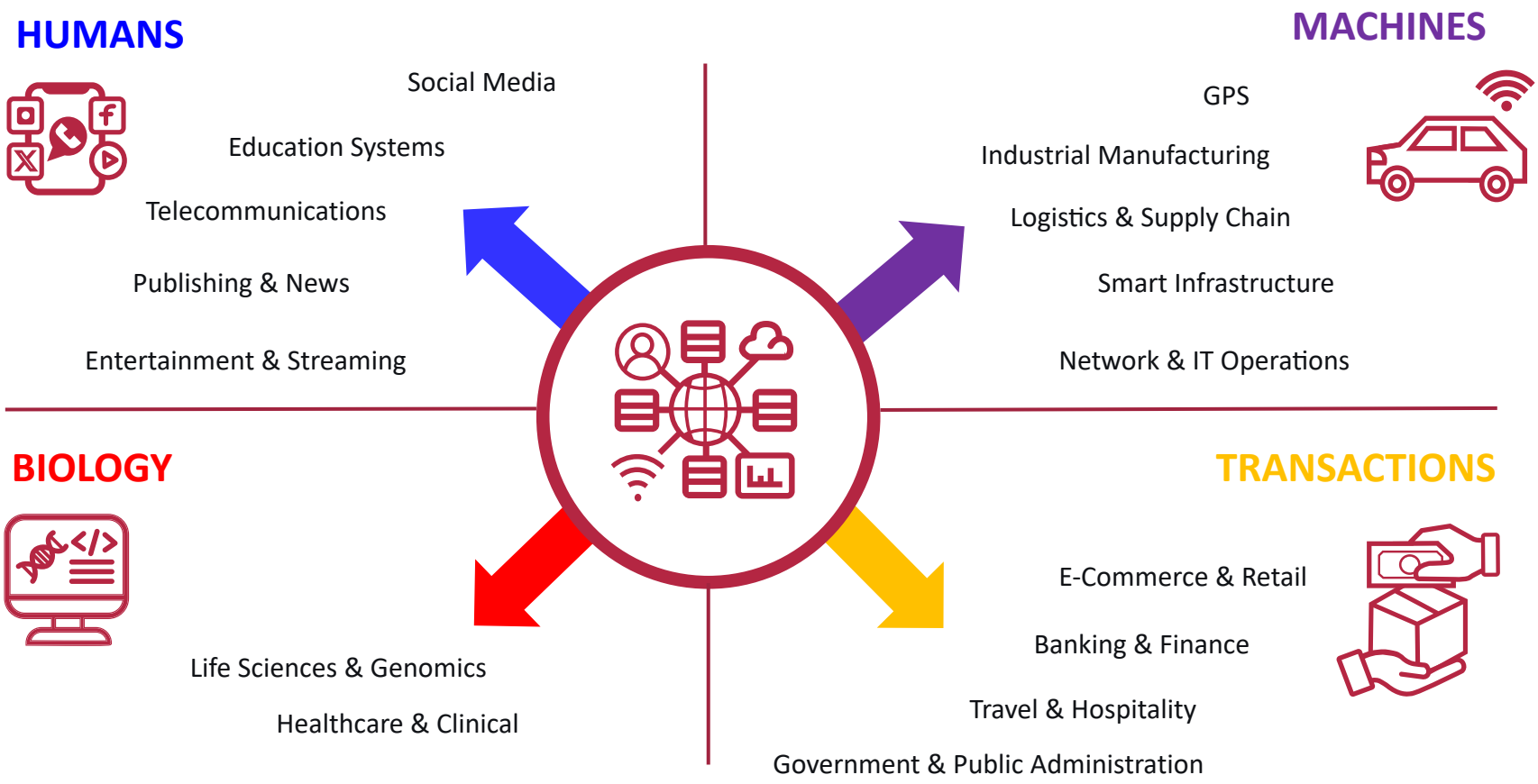
Biology
Data generated by living systems and natural processes
Examples: Genomics, Medical imaging, Proteomics, Environmental bio-data

Key Sources of Big Data

The key sources of Big Data are massive and continuous streams of data generated by:

- **Social media**
  unstructured + graph data, high velocity
- **Web/clickstream**
  behavioural event streams
- **IoT/sensors**
  high-frequency time-series
- **Enterprise transactions**
  structured relational events
- **Mobile devices**
  contextual + telemetry data
- **Financial markets**
  ultra-low latency streams
- **Healthcare & bio**
  high-dimensional, privacy-sensitive
- **IT logs**
  operational monitoring streams
- **Multimedia**
  large unstructured content repositories

The Four Engines of Big Data



Example 01: Social Media Platforms

Massive volumes of real-time, heterogeneous data from Billions of users interacting are generated continuously across text, images, video, and networks



Fields of Application

- **Sentiment & Opinion Analysis:** Extraction of attitudes, emotions, and public opinion from posts and comments
- **Social Network & Graph Analysis:** Analysis of relationships, communities, influence structures, and information diffusion
- **Behavioural Analytics & Personalization:** Tracking user behaviour to optimize content, recommendations, and engagement
- **Real-Time Trend & Event Detection:** Identification of emerging topics, crises, or viral phenomena

Example 02: Internet of Things Devices

IoT devices generate massive volumes of machine-produced, time-stamped sensor data from physical systems and environments

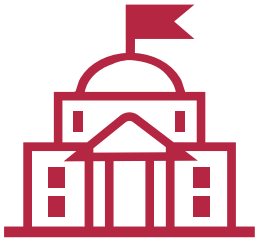


Fields of Application

- **Digital Twins**: Real-time data feeds virtual replicas of physical systems to simulate, monitor, and optimize performance
- **Health & Wearable Monitoring**: Wearable and medical IoT devices collect biometric signals for continuous assessment
- **Smart Cities & Infrastructure**: Sensors monitor traffic, air quality, utilities, and public infrastructure
- **Predictive Maintenance**: Continuous sensor monitoring detects anomalies and predicts equipment failure

Example 03: Government and Public Records

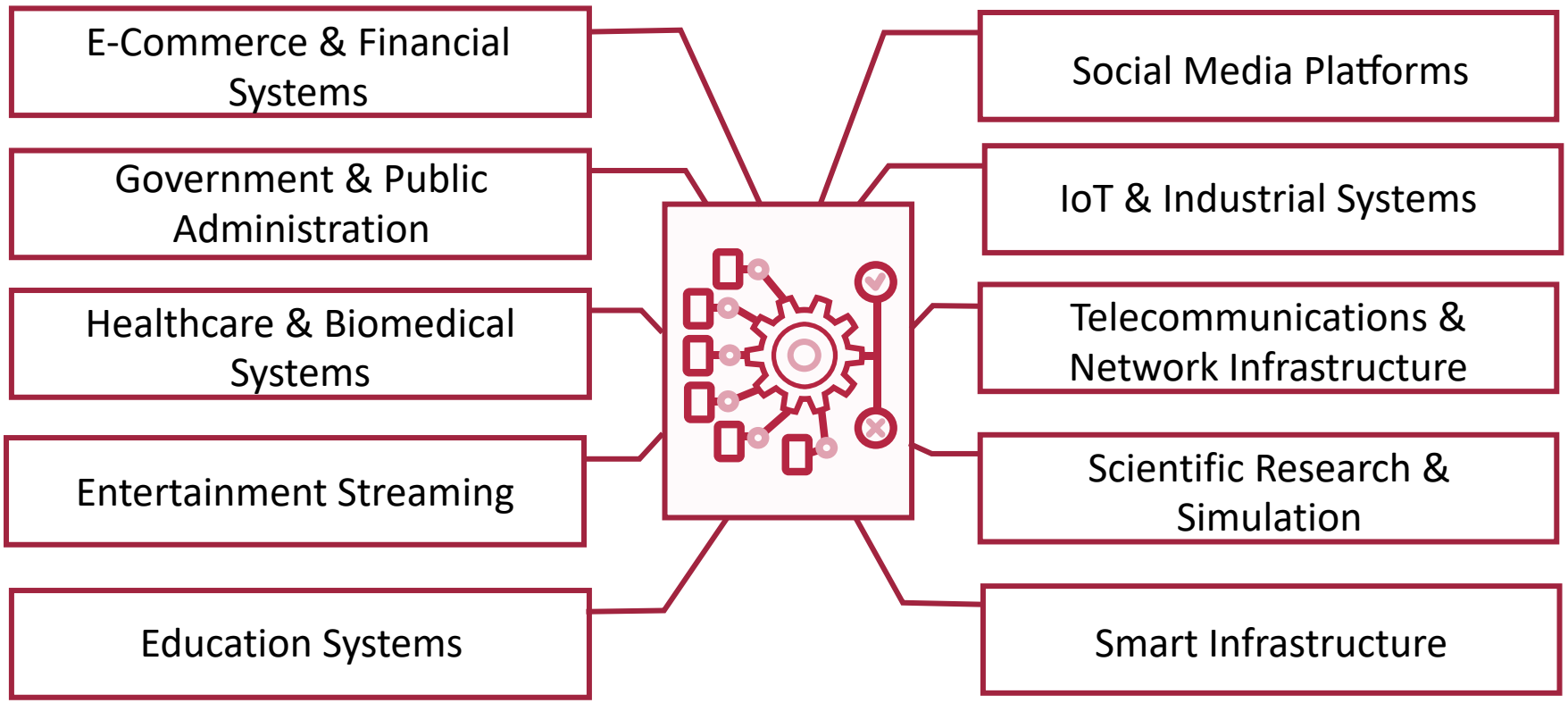
Datasets originated from administrative processes, legal obligations, and public service delivery, covering entire populations and long-time horizons



Fields of Application

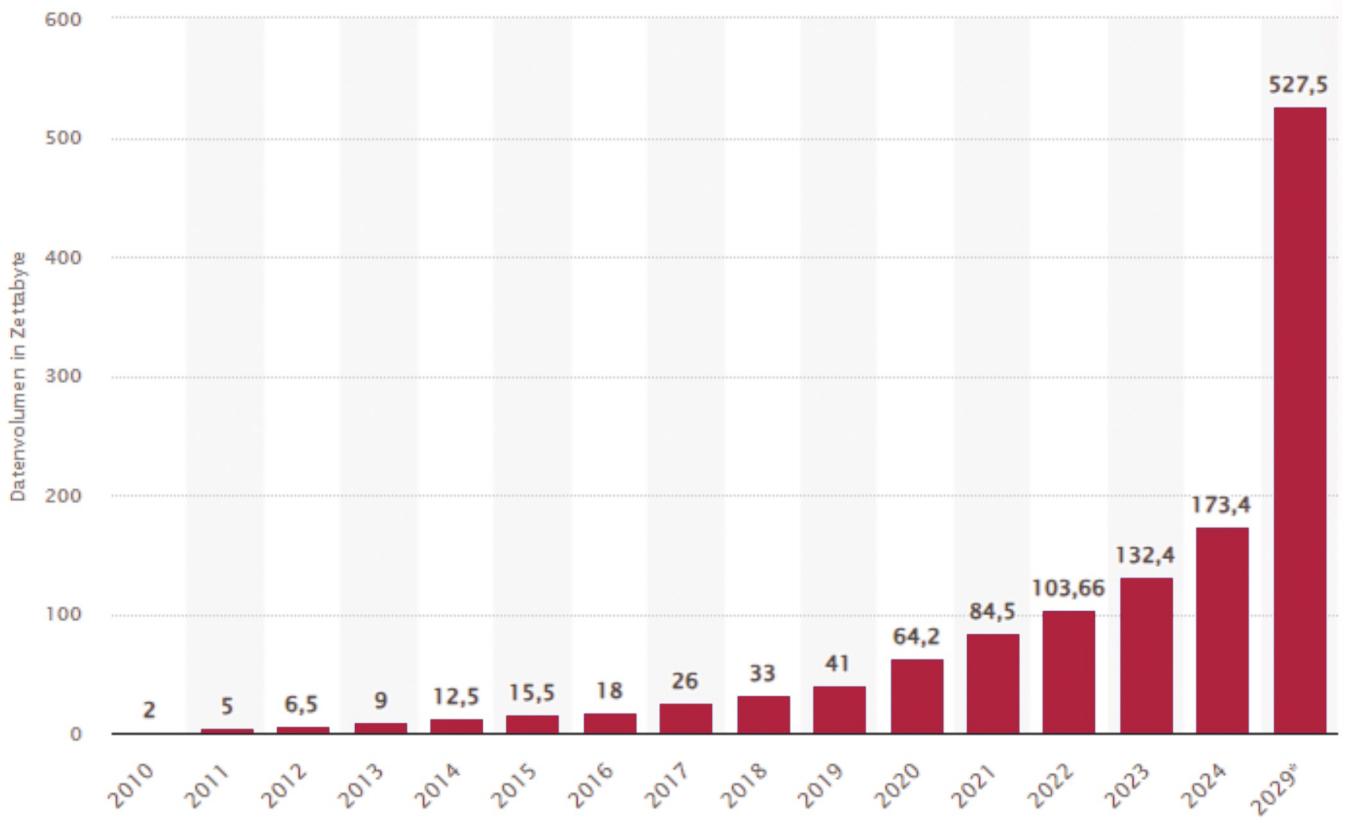
- **Public Service Optimization:** Analysis of service usage improves efficiency and allocation of resources
- **Open Data & Transparency:** Public datasets are released for research, innovation, and civic engagement
- **Scale & Historical Depth:** massive longitudinal datasets that track populations, economies, and land use over time
- ***Urban Mobility & Traffic Management:*** Real-time analysis of traffic flow, public transport usage (subway smart cards, bus GPS), and pedestrian movement patterns

Large Generators of Big Data



Some Statistics

Volume of digital data generated/replicated annually worldwide from 2010 to 2024 and forecast for 2029

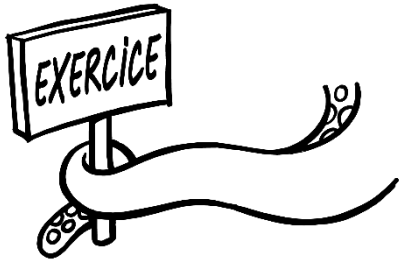


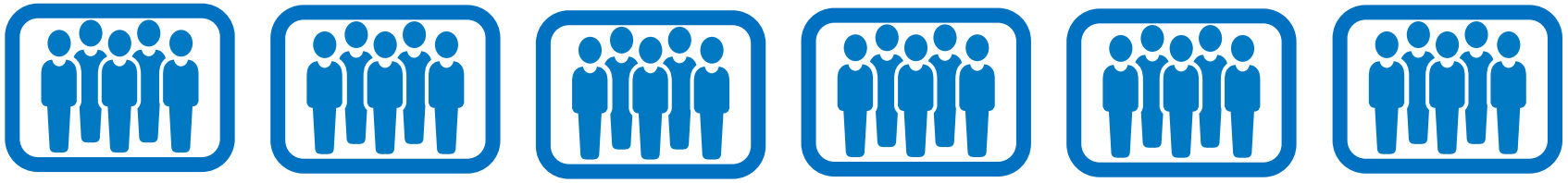
Top 10 Free & Open Big Data Repositories on the Web

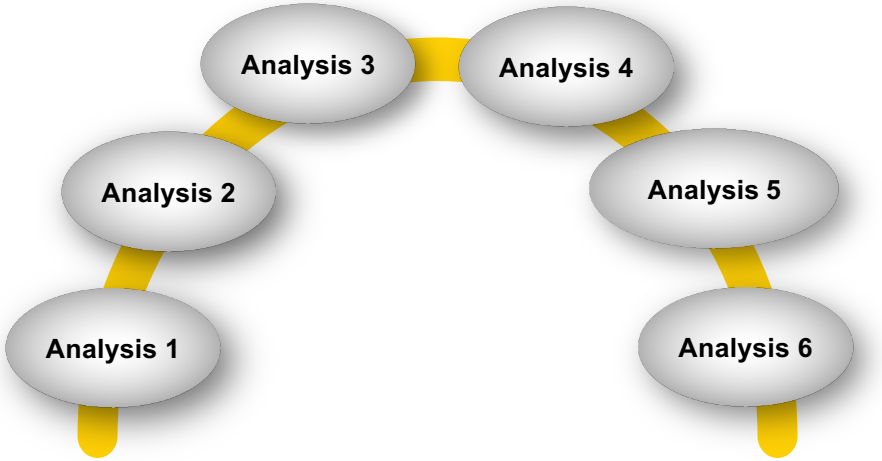
US Government's open data portal with thousands of datasets across many domains (health, energy, agriculture, finance, climate)	https://www.data.gov
Central access point for public datasets from European institutions and member countries	https://data.europa.eu/en
Comprehensive global socio-economic, demographic, and environmental datasets from UN agencies	https://data.un.org
Extensive global development indicators including economic, health, and social data	https://data.worldbank.org
Massive repository of datasets uploaded by the data science community, covering many fields and ready for ML projects	https://www.kaggle.com/datasets
A specialized search engine for datasets across the web, aggregating from multiple sources	https://datasetsearch.research.google.com
Extensive weather, climate, and environmental data from the U.S. National Oceanic and Atmospheric Administration	https://www.ncdc.noaa.gov/cdo-web
Clean and visualizable socio-economic and health statistics spanning decades for almost all countries	https://www.gapminder.org/data
Authoritative country-level data on demographics, economics, geography, and military	https://www.cia.gov/the-world-factbook
Large collection of public datasets hosted on AWS infrastructure, spanning genomics, geospatial, satellite imagery, and more	https://registry.opendata.aws

From Repository to Insight: An Open Data Exploration

Select <u>one</u> of the ten open big data repositories. Investigate a specific dataset within it and define a valuable insight that could be extracted for a real-world company, organization, or research project







The Big Data Pipeline

From Raw Data to Action: The Four Essential Steps

Ingest	Store	Process	Serve
Acquiring raw data from diverse sources (social media, IoT sensors, transactions, logs)	Persisting data in scalable, fault-tolerant storage systems	Cleaning, validating, aggregating, analysing, and modelling data	Presenting insights to users, applications, or decision systems
Collect	***Persist***	***Transform***	***Deliver***
			
(Apache Kafka)	*(HDFS)*	*(Apache Spark)*	*(Tableau, Power BI)*

The Complete Big Data Lifecycle

Step	Aim	Technical Aspects	Example (Smart City Traffic)
1. Data Ingestion	Collect data from multiple sources	Batch vs streaming, APIs, Kafka, Flume	Traffic sensors, ride-sharing apps, social media feeds
2. Data Storage & Management	Store and organize large volumes of data	HDFS, NoSQL, Data Lakes, schema-on-read/write	Data lake storing historical traffic, ride, and sensor data
3. Data Processing & Analysis	Clean, transform, and extract insights	ETL/ELT, Spark, Hadoop, ML models, stream/batch processing	Filtering sensor anomalies, predicting congestion hotspots, optimizing traffic lights
4. Visualization & Decision Making	Present insights for action	Dashboards, BI tools, alerts, reporting	Live congestion maps, predictive alerts, recommendations for infrastructure changes

HDFS = Hadoop Distributed File System / ETL = Extract, Transform, Load / ELT = Extract, Load, Transform / ML = Machine Learning / BI = Business Intelligence

Big Data Tool Landscape

Infrastructure	• **HDFS (Hadoop Distributed File System):** Distributed file system for storing massive datasets across clusters; supports parallel processing and fault tolerance *Storage of large-scale datasets for batch analytics, supports parallel processing* • **Apache Kafka**: Distributed event streaming platform for high-throughput, real-time data ingestion and stream processing *IoT streams, log aggregation, real-time data pipelines* • **Apache Spark**: Distributed processing engine for large-scale batch and streaming data Big data ETL, real-time analytics, ML, graph analytics
Cloud & Warehouse Services	• **Amazon EMR**: Managed cloud service for running Hadoop, Spark, and related big data frameworks at scale *Cloud-based batch and streaming processing, ETL, large-scale analytics* • **Google BigQuery**: Fully-managed, serverless cloud data warehouse for interactive SQL analytics on very large datasets *Large-scale analytical queries, dashboards, business intelligence* • **Snowflake**: Cloud-native data warehouse with separated storage and compute *Cloud analytics, data sharing, BI, modern ELT pipelines*
Enterprise Platforms	• **Databricks**: Unified analytics platform built around Spark for big data ETL, ML, and streaming *Big data pipelines, collaborative data science, ML engineering* • **Cloudera**: Enterprise data platform integrating Hadoop, Spark, Kafka, and governance tools for large-scale data lake and hybrid cloud deployment *Enterprise data lakes, batch and streaming analytics, governance, ML pipelines*

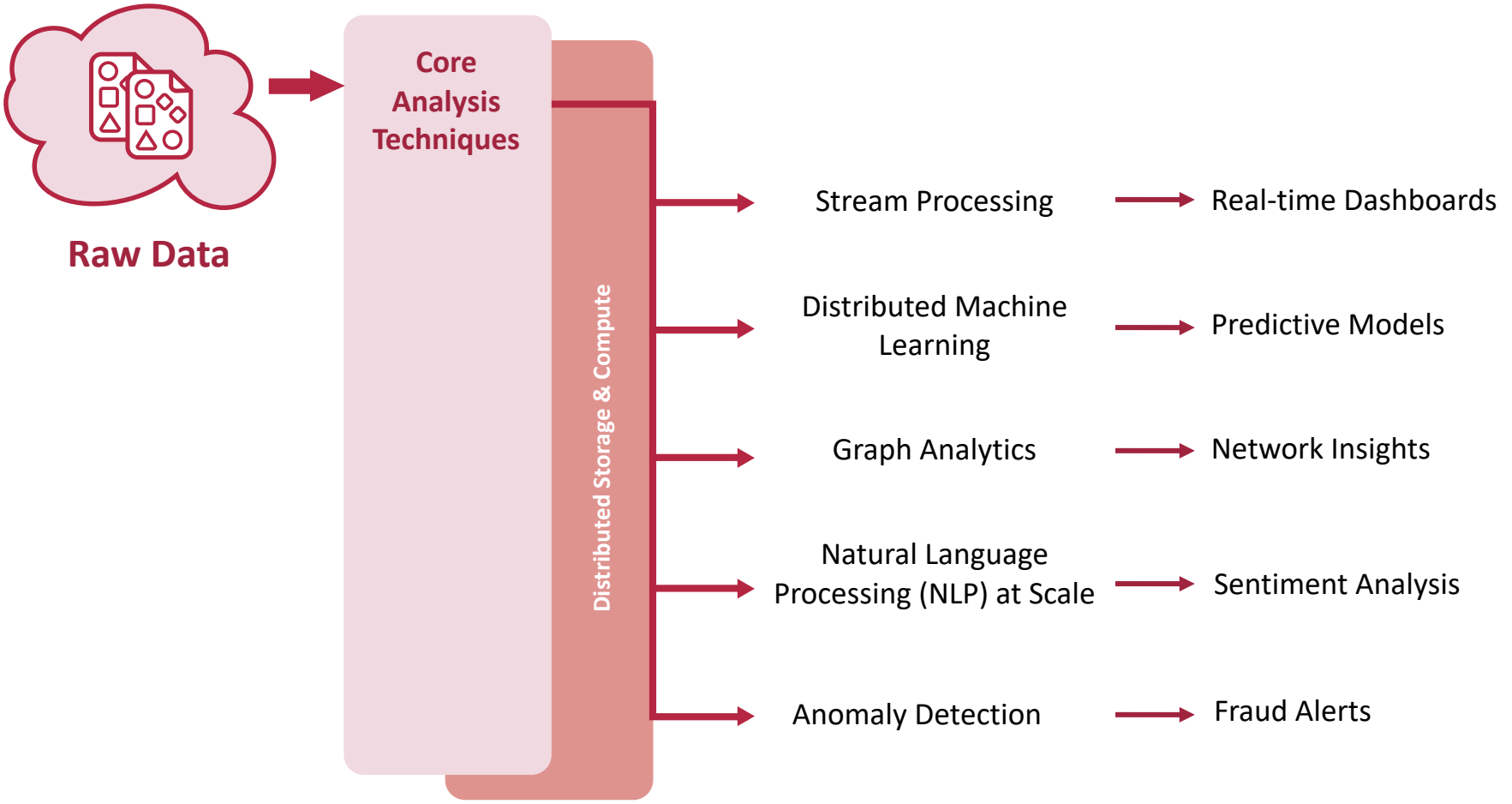
The Three Pillars of Big Data Storage



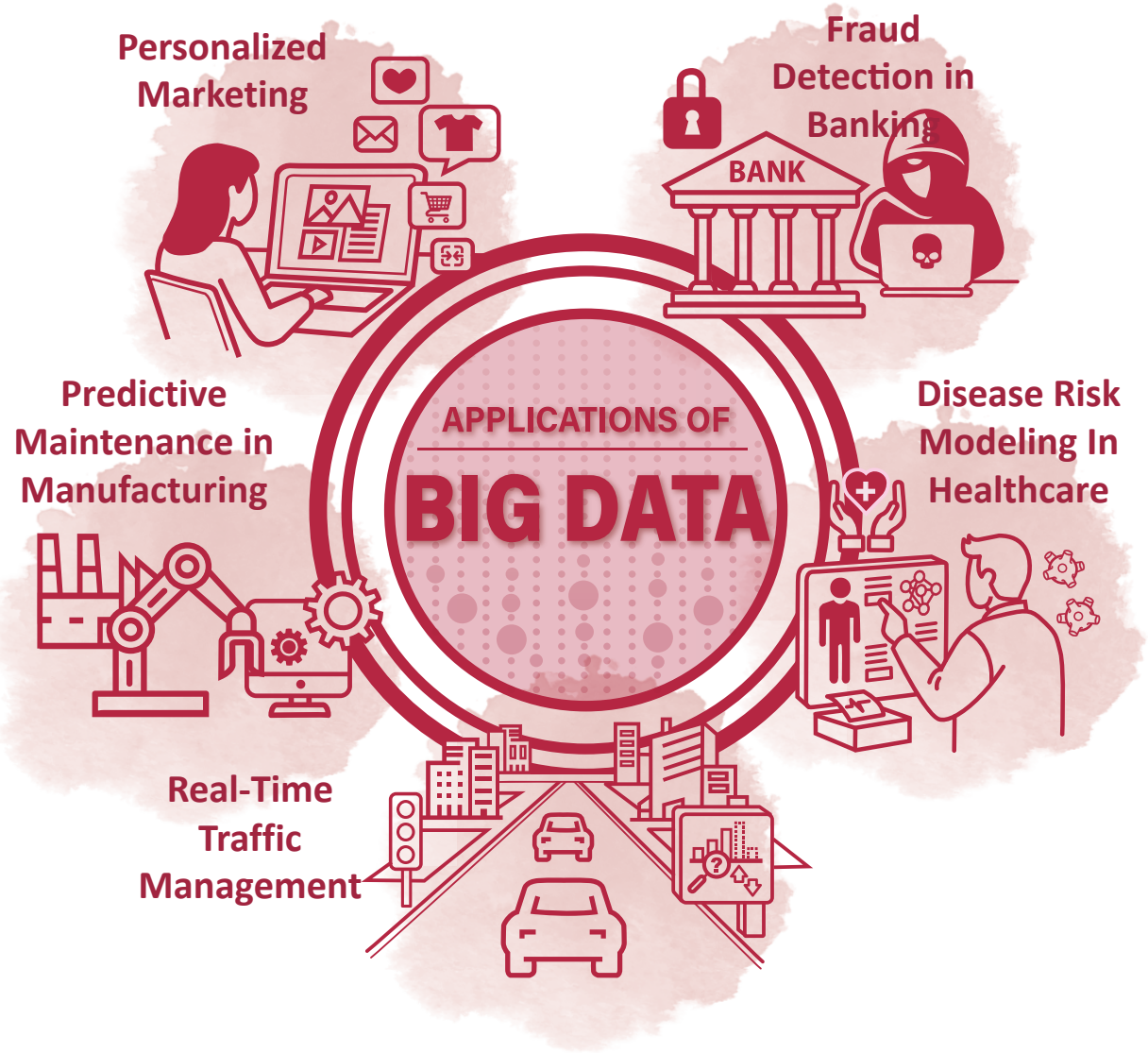
Pillar	Description	Use Cases	Technical Features	Solution
Distributed File Systems	Store files across multiple machines, presenting a single unified namespace	• Raw data storage • Data lakes • Batch processing foundation	Block-based storage, replication, fault tolerance, horizontal scaling, optimized for sequential batch reads	HDFS
NoSQL Databases	Non-relational databases optimized for scalability and flexible data models	• High-velocity writes • Semi-structured data • Low-latency access • Time-series data	Horizontal sharding, schema flexibility, high write throughput, low-latency access	Cassandra, MongoDB
Cloud / Analytical Data Warehouse	Systems optimized for large-scale analytical queries on structured data and semi-structured data	• Business intelligence • Complex SQL analytics • Reporting and dashboards • Enterprise performance management	Columnar storage, MPP (massively parallel processing), SQL optimization, separation of storage & compute	Snowflake, BigQuery

Big Data Value Pipeline

Extracting Value from Big Data



Applications of Big Data



Key Fields of Big Data Applications

Big Data is transforming industries by enabling data-driven decision-making, optimizing operations, and enhancing customer experiences



Patient monitoring, predictive diagnostics, genomics



Fraud detection, risk modelling, algorithmic trading, customer analytics



Personalized recommendations, inventory optimization, dynamic pricing



Predictive maintenance, process optimization, IoT monitoring



Streaming analytics, content recommendation, audience analysis



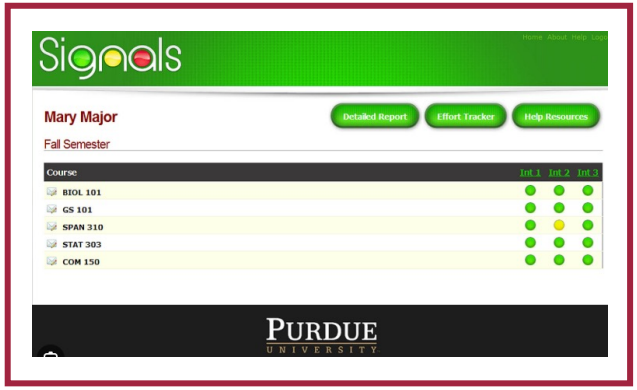
Smart cities, open data initiatives, public service optimization

Big Data Adoption: A Cross-Sector Maturity Assessment

Sector	Primary Data Sources	Key Analytics Use	Maturity Level	Business Outcome
Banking	Transactions, credit bureaus, web logs, ATM sensors	**Real-time fraud detection**	Advanced	50% fraud reduction, $10B+ saved annually
Insurance	Claims, telematics, IoT, weather, health records	**Usage-based pricing**	Advanced	20% lower claims, personalized premiums
Retail	POS transactions, loyalty cards, web clicks, inventory	**Personalized recommendations**	Advanced	10-30% revenue lift, 20% inventory reduction
Healthcare	EHR, medical imaging, wearables, genomics	**Predictive diagnostics**	Maturing	30% earlier disease detection, personalized treatment
Manufacturing	IoT sensors, production logs, quality control	**Predictive maintenance**	Maturing	40% downtime reduction, 25% maintenance savings
Logistics	GPS, RFID, weather, traffic, fuel sensors	**Dynamic route optimization**	Maturing	15% fuel savings, 20% faster deliveries
Telecom	Call records, network logs, social media	**Churn prediction**	Advanced	20% churn reduction, $1B+ annual savings
Energy	Smart meters, weather, grid sensors, satellite	**Load forecasting**	Maturing	15% energy savings, grid stability
Media	Viewing history, ratings, social media	**Content recommendation**	Advanced	50%+ engagement from recommendations
Agriculture	Soil sensors, satellite, weather, equipment telemetry	**Precision farming**	Maturing	25% higher yields, 30% water savings
Government	Census, traffic cameras, tax records, permits	**Smart city operations**	Emerging	20% congestion reduction, faster services
Education	LMS logs, video views, assessments, surveys	**Personalized learning paths**	Emerging	15% better student outcomes, reduced dropouts

Example 01: Reducing Drop-out Rates

Purdue University – Course Signals



- **Use of Big Data**

  Course Signals analyses multiple educational data sources in real time, including:

  - Learning Management System interaction logs (e.g., readings, assignments, visits)
  - Performance data (grades, quiz results)
  - Historical student records
  - Engagement and effort indicators
- **Analysis**: It processes these using predictive models to estimate *risk levels* for each student and presents them via a **traffic-light system** (green/yellow/red) for both students and instructors
- **Results**: Students in Courses using Signals have **higher retention and success rates** compared to peers not using the system

Source: studyres.com/doc/15603353/the-role-of-learning-analytics-in-future-education-models

Example 02: Wearable Devices

Wearable devices like smartwatches, fitness trackers, smart rings, and biosensors

- **Use of Big Data**

  Wearables generate **massive volumes of structured and semi-structured data,** Including:
  - Heart rate, blood pressure, oxygen saturation
  - Sleep patterns, step counts, activity levels
  - GPS/location tracking for mobility patterns
  - Exercise intensity and calories burned
  - Sometimes social/interaction data (connected apps)



- **Analysis**
  - Predictive Modelling: Identify risk of heart disease or sleep disorders
  - Anomaly Detection: Detect irregular heart rhythms or falls
  - Trend Analysis: Personal fitness trends over weeks/months

- **Results**
  - Real-time personalized feedback for healthier behavior
  - Integration with apps, cloud platforms, and AI provides continuous learning and insight refinement

Example 03: Disease Tracking

Real-Time Epidemic Monitoring and Public Health



- **Data Sources**
  - Electronic Health Records (EHRs) from hospitals and clinics
  - Mobile and GPS location data for population movement
  - Social media posts mentioning symptoms or outbreaks
  - Pharmacy and lab testing data
  - Environmental sensors (air quality, climate conditions)
  - Sometimes social/interaction data (connected apps)
- **Analysis**
  - Predictive Modelling: Forecast disease spread and infection rates
  - Anomaly Detection: Identify unusual clusters of symptoms
  - Geospatial Analysis: Map outbreak hotspots and mobility patterns
  - Network Analysis / Graph Analytics: Model transmission chains and contact networks
- **Results**
  - Early warning systems for epidemics and pandemics (e.g., influenza, COVID-19)
  - Supports policy decisions such as vaccination campaigns or quarantine measures
  - Improves resource allocation in healthcare systems (beds, vaccines, testing kits)
  - Enables research in epidemiology using large, heterogeneous datasets

Example 04: Precision Farming



Agriculture faces challenges like resource scarcity (water, fertilizer), crop disease, unpredictable weather, and pressure to maximize yield

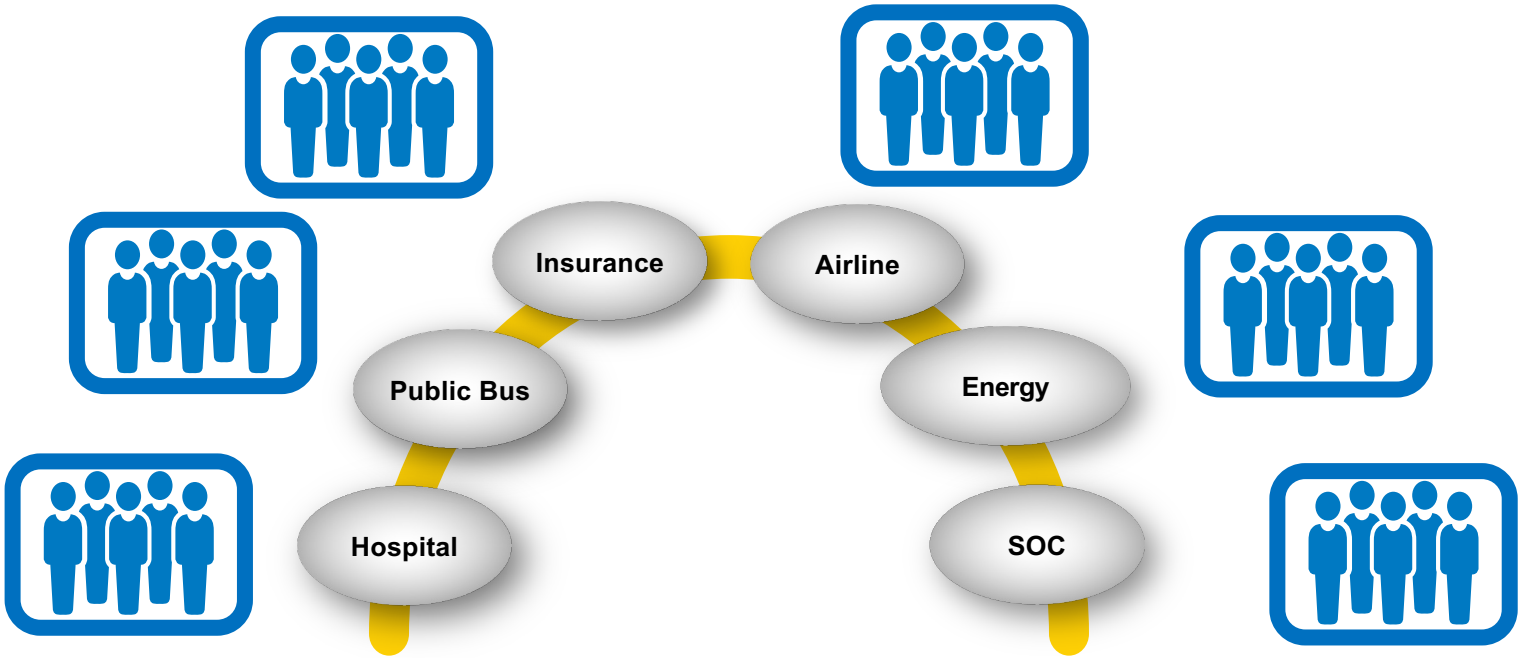
- **Use of Big Data**
  - IoT sensors in soil, crops, and irrigation systems
  - Drone and satellite imagery (NDVI, crop health monitoring)
  - Weather and climate data (temperature, rainfall, humidity)
  - Machinery telemetry and GPS tracking
  - Market prices and supply chain data
  - Sometimes social/interaction data (connected apps)
- **Analysis**
  - Predictive Modeling: Forecast crop yield and disease outbreaks
  - Spatial / Geospatial Analysis: Optimize planting and irrigation by field zones
  - Time-Series Analysis: Monitor soil moisture, growth cycles, and weather impacts
  - Anomaly Detection: Identify pest infestations or crop stress early
- **Results**
  - Optimized **irrigation and fertilizer usage**, reducing costs and environmental impact
  - Enables **real-time farm management decisions** based on sensor feedback

How Can Big Data Generate Actionable Insights?

Each group identifies what Big Data reveals in a specific sector and defines one valuable insight for a real organization



1. Hospital
2. City Bus Company
3. Insurance Company
4. Airline Company
5. Energy Utility Company
6. Cybersecurity Operations Centre (SOC)



Big Data Strategy



What is a Big Data Strategy?

A comprehensive plan that defines how your organization will leverage large-scale data assets to achieve its business goals



What are the Goals of Big Data Strategy?

- **Real-time and near-real-time decision intelligence**
  Support instant analytics on streaming data (e.g., fraud detection, dynamic pricing, predictive maintenance)
- **Cross-domain data integration**
  Combine siloed enterprise data with external and high-variety sources to uncover hidden patterns not visible in isolated datasets
- **Hyper-personalization at population scale**
  Deliver individualized recommendations or experiences across millions of users simultaneously
- **Drive real-time operational optimization**
  Adjust business processes and logistics instantaneously based on live data streams
- **Data monetization at ecosystem scale**
  Create new revenue streams by packaging, sharing, or commercializing large-scale data assets

Strategic Goals of Big Data: From Firms to Economies

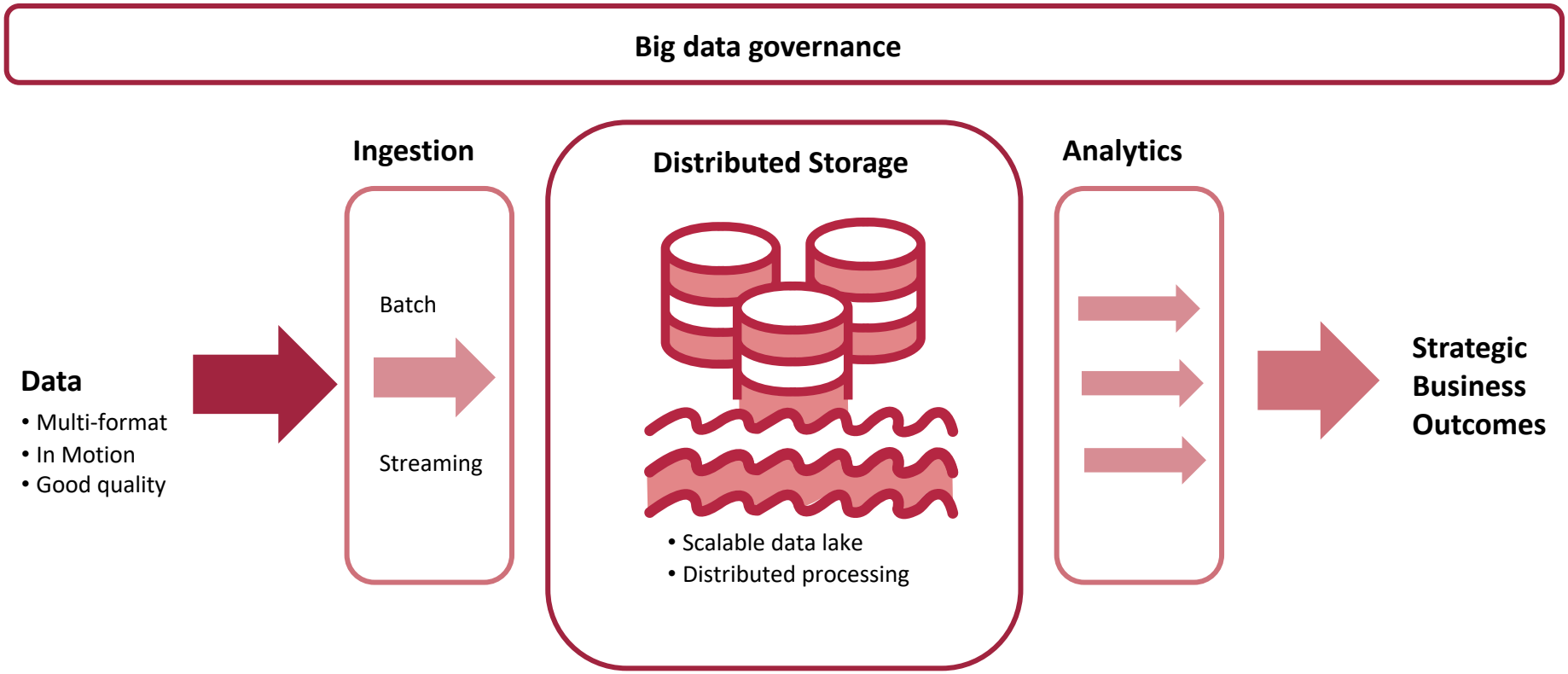
"Generating value at the different stages of the data value chain will be at the centre of the future knowledge economy. Good use of data can also bring opportunities to more traditional sectors such as transport, health or manufacturing. Improved analytics and processing of data, especially of big data, will make it possible to:

- *transform Europe's service industries by generating a wide range of innovative information products and services;*
- *increase the productivity of all sectors of the economy through improved business intelligence;*
- *address more efficiently many of the challenges that face our societies;*
- *improve research and speed up innovation;*
- *achieve cost reductions through more personalised services;*
- *increase efficiency in the public sector."*

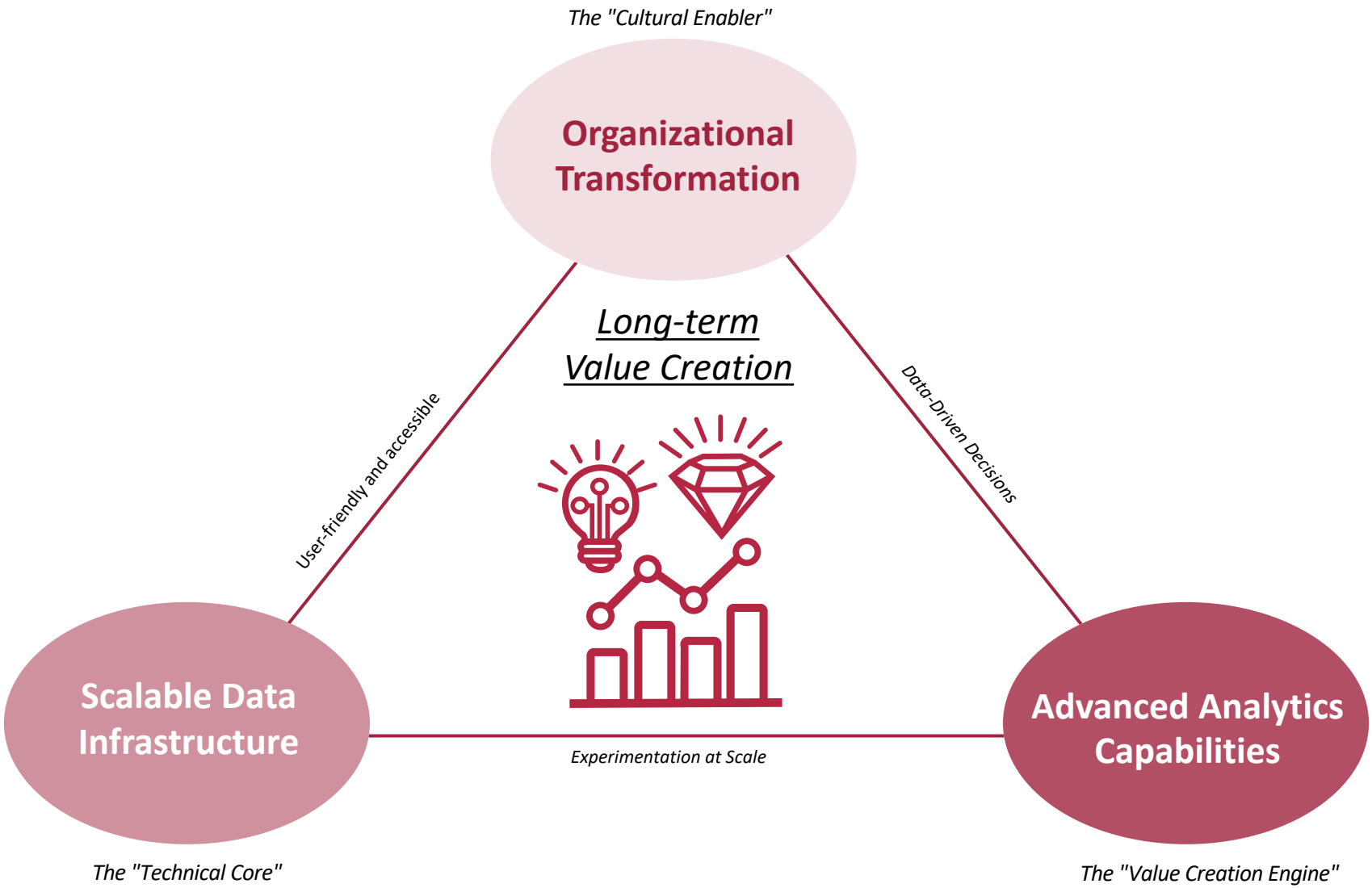
European Commission

Source: digital-strategy.ec.europa.eu/en/policies/big-data

Building Blocks of a Big Data Strategy



Strategic Planning for Big Data: Three Pillars of Long-Term Value Creation



Big Data Readiness Assessment: Evaluating the Three

Maturity Check
Scalable Data Infrastructure

- Degree of cloud adoption / distributed architecture
- Ability to handle high volume, velocity, variety
- Existence of data lake / Lakehouse
- Streaming and real-time processing capabilities
- Data integration across silos
- Scalability and cost elasticity
- Security and privacy at scale

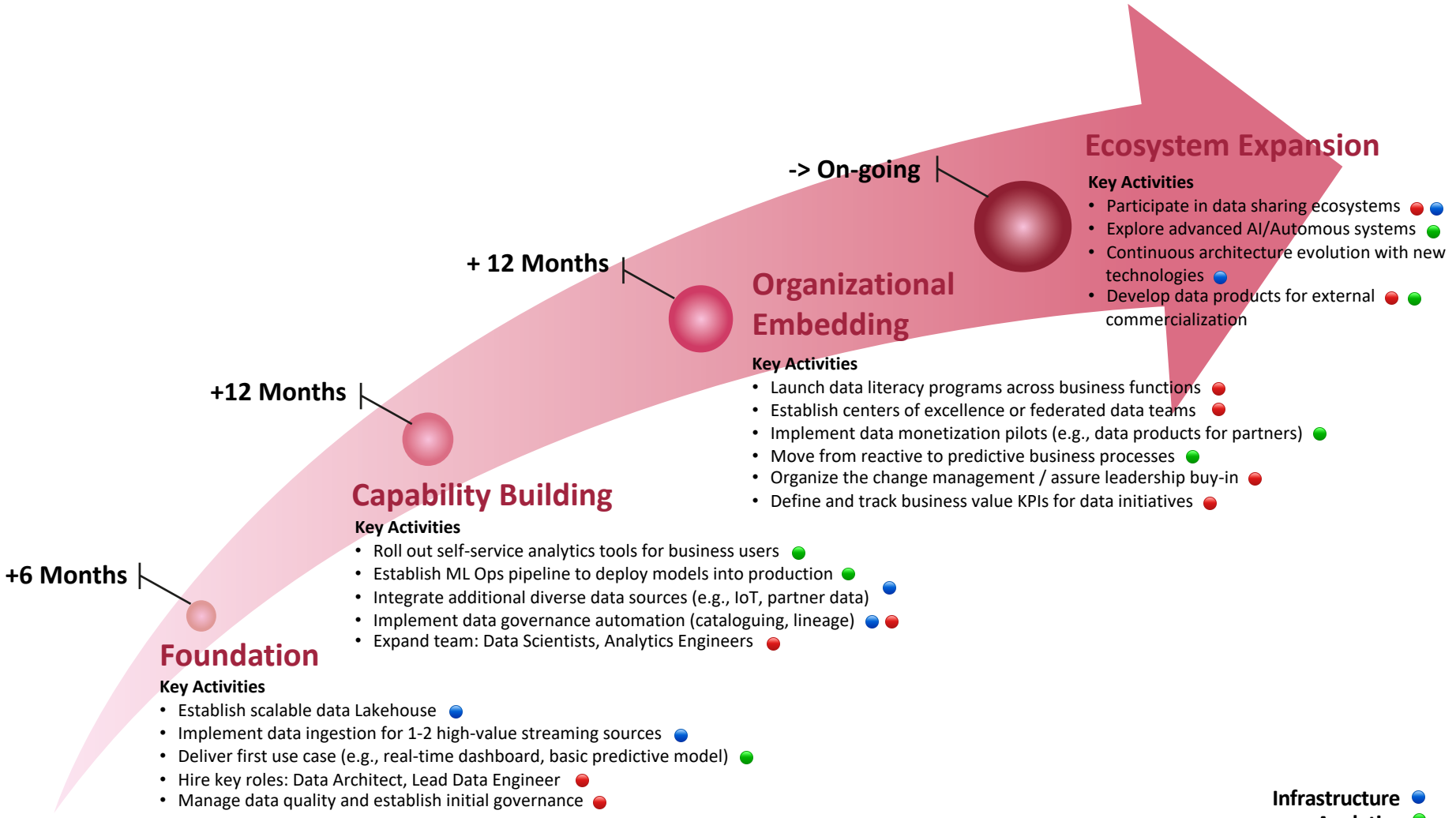
Capability Check
Advanced Analytics Capabilities

- Data engineering maturity
- Real-time analytics capability
- Experimentation capability (A/B at scale)
- Availability of large-scale model training infrastructure
- Use of ML beyond pilots
- MLOps maturity (deployment, monitoring, retraining)

Readiness Check
Organizational Transformation

- Leadership commitment to data-driven strategy
- Data governance adapted to distributed environments
- Clear ownership of data products
- Data-driven decision culture
- Availability of specialized roles (data engineers, ML engineers)
- Cross-functional collaboration

The Big Data Implementation Journey: A Phased Roadmap to Maturity

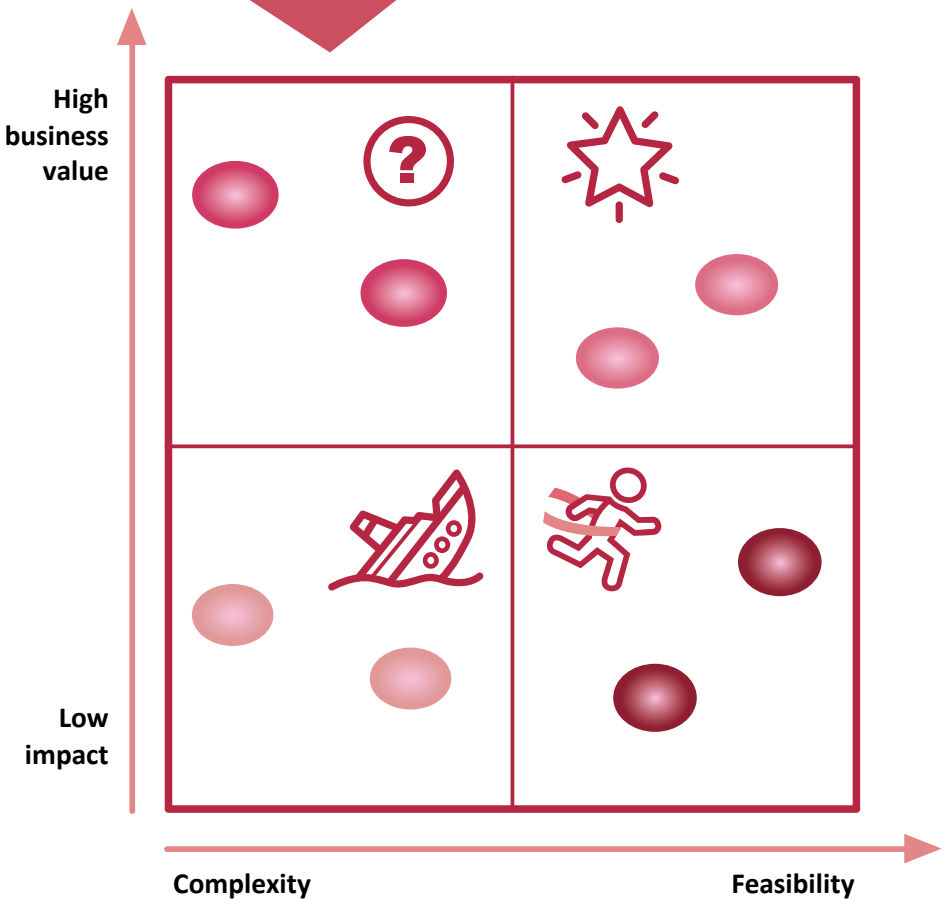


Identifying and Prioritizing Big Data Use Cases

Checklist

Dimension	Description
Data Availability & Quality	Do we already have the data? Is it complete, clean, and accessible?
Technology & Skills Gap	Do we currently possess the required technology stack and the in-house talent to execute this, or must we invest heavily to acquire them?
Implementation Velocity	Can this be delivered in a short, agile timeline (weeks/months), or is it a long-term project (years) that will tie up resources?
Integration Complexity	How difficult is it to integrate the output into existing business processes, operational systems, and user workflows?
Revenue Generation	Does the use case directly create new revenue, enable upselling/cross-selling, or form the basis of a new data product for partners?
Cost Optimization & Efficiency	Will it significantly reduce operational costs, automate manual processes, or optimize supply chains to save money?
Strategic Alignment	Does this use case directly support a top-tier corporate goal
Decision Quality & Agility	Does it improve the speed and accuracy of decisions? Does it move the business from reactive to predictive ?

Rating



4 Challenges of Big Data Strategy Implementation

Insight

Impact

Expectation Gap



The difficulty in translating complex analytics into tangible, operational business

Data Accumulation

Data Usability

Data Swamp



The failure of governance and integration, turning data assets into unmanageable liabilities

Technical Expertise

Business Integration

Talent Chasm



The scarcity of hybrid professionals who can bridge the gap between statistical methodologies and real-world business strategy

Innovation Freedom

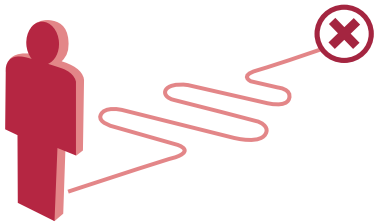
Regulatory Constraint

Governance Paradox



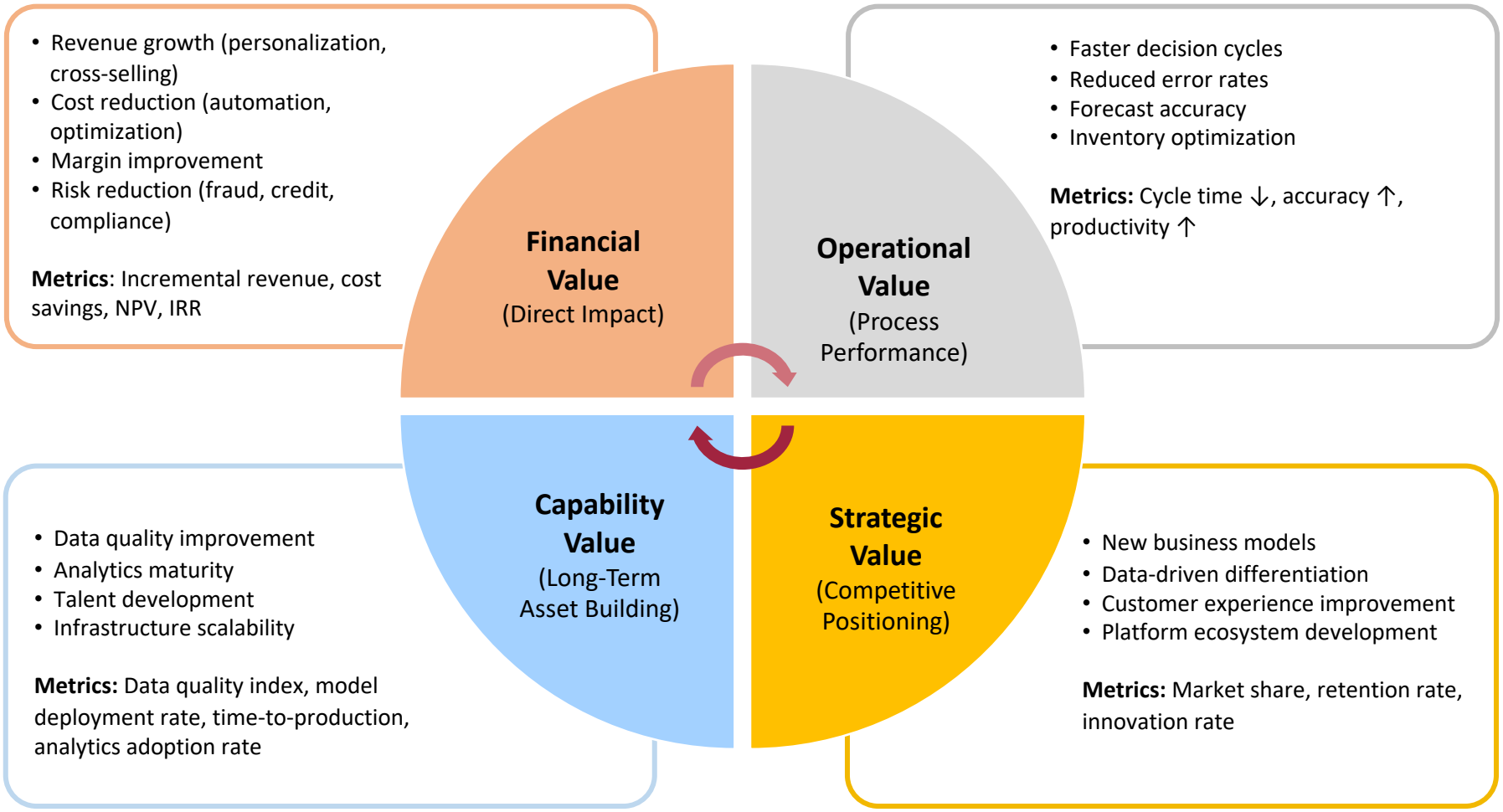
The tension between aggressive data collection for innovation and the strict requirements of privacy, ethics, and compliance

Four Layers of Bias in Big Data Strategy



Cultural Biases	Analytical Biases	Data & Infrastructure Biases	Sampling and Population Bias
• *Confirmation Bias:* Seeking evidence to support pre-existing beliefs • *Silo Bias:* Hoarding data within departments, preventing enterprise value • *Short-termism:* Prioritizing quick wins over sustainable data architecture • *Automation Bias:* Blindly trusting model outputs over human intuition and over-delegating judgment to algorithms	• *Correlation Fallacy:* Mistaking correlation for causation • *Overfitting (The Nostalgia Bias):* Building models that work perfectly on past data but fail in the present	• *Data Abundance Bias:* Equating high volume with high quality/insight • *Technology Solutionism:* Believing that buying software/tools alone solves business problems	• *Sampling Bias:* Non-representative data • *Selection Bias:* Systematic exclusion of relevant groups and removing "outliers" that are relevant & meaningful • *Survivorship Bias:* Learning only from successes • *Historical Bias:* Embedding past discrimination

Success Metrics for Big Data



Advising the CEO on Big Data Initiatives

You are an external consultant hired to fix Big Data implementation failures



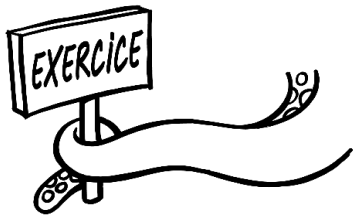
1. Data Science
2. Sales VP
3. CFO
4. Marketing
5. Customer Support
6. Research & Development



Exercise: Analyse and Propose Solutions

Department	Departmental View
Data Science	*“We just got access to the data warehouse.' It's a mess. The sales data is in Salesforce, the support data is in Excel files on managers' laptops, and the product usage data doesn't exist. We have PhDs from MIT ready to build models, but we have nothing to build them on. We are bored.”*
Sales VP	*"I don't need your algorithms. I've been selling software for 20 years. I know my customers. Last quarter, your model told us to stop calling a certain type of client, and I lost two deals because we followed the computer's advice. I'm going back to my gut instinct."*
CFO	*"I approved the budget for big data projects, but I need to justify it to the board. The IT is asking for $2M to build a new 'Data Lake' on AWS. I asked them what the Return on Investment is, and they said, 'It enables optionality.' I can't take 'optionality' to the bank. I need numbers."*
Marketing	*"I asked the data team to help me predict which new customers are likely to buy our premium package. They built a model that is 99% accurate on last year's data! But this year, we launched a new product and changed our pricing. The model is useless now. Why did they build it on old stuff?"*
Customer Support	*"We have millions of support tickets. We want to use AI to auto-respond to the simple ones. I gave the data team 1 million tickets from 2023. They built a great model. But 2023 was a weird year—we had a major bug that caused a specific type of complaint. Now the AI only knows how to respond to that old bug, not the new issues customers have today."*
R&D	*"We are building the next generation of our product, and we want to embed AI features directly into it. But Legal says we cannot use certain customer behavioral data due to GDPR concerns. IT says integrating real-time analytics will require redesigning the entire architecture. Finance says there is no clear revenue projection yet. If we wait for perfect compliance and ROI clarity, our competitors will launch first. Should we move fast and take the risk, or slow down and lose innovation speed?"*

1. Diagnose the root cause (bias / structural tension / capability gap)
2. Propose concrete actions to unblock the situation
3. Propose a metric to measure success in 6 months



Big Data Ecosystem (BDE)

Data Generators	Data Collectors	Data Aggregators	Data Consumers
• Physical World: IoT, Machines, Smart Cities • Human Interaction: Social Media, Clickstreams • Business Ops: Enterprise Systems (ERP, CRM), Transactions	• Real-time: Streaming (Kafka, Kinesis), Edge Gateways • Batch: Log Collectors, APIs, ETL tools	• Storage: Data Lake • Processing: Engines • Serving: Warehouse (Snowflake), NoSQL (Cassandra)	• Decision makers • Analytics: Business Analysts • Science: Data Scientists • Applications: AI Systems, Recommendation Engines
Volume	**Velocity**	**Variety**	**Value**









Key Components of the Big Data Ecosystem

Data Governance & Security

- Access control
- Privacy & compliance
- Metadata & lineage
- Data quality management
- Monitoring & auditing



Data Ingestion

- Streaming
- Batch
- APIs
- Log Collectors
- Edge Gateways



Data Factory (Processing & Consumption)

- Processing
- Query
- Business Intelligence & Analytics
- Data Science



Foundation (IT Infrastructure & Storage)

- Data Lakes
- Distributed File Systems (HDFS, S3)
- Cluster Managers (Kubernetes, YARN)
- Compute Resources (CPU, Memory, Network)



BDE in Action: Real-World Usage Examples

Healthcare: Predictive Patient Monitoring

- Monitors streaming vital signs
- Lab results and electronic health records
- Historical patient data

↓

- Early Warning Alerts: Notification via mobile app of at-risk patients
- Clinical Dashboard
- Reduction in mortality
- Shorter hospital stays

Smart City: Traffic Management

- Road loop detectors and cameras capturing vehicle counts
- Anonymous location data from connected vehicles and navigation apps
- Real-time weather conditions affecting traffic

↓

- Lights adjust automatically to reduce wait times
- Real-time route guidance to avoid congestion
- Lower emissions

Smart Manufacturing: Industry 4.0

- Vibration, temperature, and pressure data from assembly line equipment
- Quality control results

↓

- Alerts before equipment failure, reducing unplanned downtime
- Reduction in downtime
- Lower maintenance costs

Financial Services: Fraud Detection

- **Transaction:** Credit card swipes, online payments
- Customer profiles
- Geolocation
- Blacklists

↓

- Suspicious transactions blocked in under 100ms
- Fraudulent transactions blocked
- Reduction in false positives

New and Different Capabilities of the BDE

Smart-Data Aggregation and Custodian

- Ecosystem Orchestrator
- Governance Enforcer
- Value Creator
- Trusted sources
- Quality of data

Transformative Capabilities

Unified Customer 360° View

Breaking down silos to truly understand the customer

Real-time Decisions

From hindsight to insight to foresight—in milliseconds

Predictive Intelligence

Anticipating the future, not just reporting the past

Data as an Asset

Monetizing data directly or enabling new business models

Importance of a Successful BDE

A well-developed Big Data Ecosystem ensures:

- Seamless data integration
- Enhanced decision-making with real-time analytics
- Improved data availability for strategic planning
- Unlocks Business Value from Data
- Reliable and high-quality data (Data is accurate, complete, timely, and governed with clear standards, making it trustworthy for analytics and AI)
- Strong governance and security (Policies for privacy, access control, compliance, and ethical use are embedded into the ecosystem, reducing risk)
- A data literate culture (People understand how to use data responsibly and effectively, supported by training, documentation, and clear ownership)

Evolving BDE: Integration of Emerging Technologies

Modern Big Data Ecosystems are evolving to integrate **AI, IoT, and Blockchain**, enabling new capabilities:

Enable Smart Data Management

- IoT generates massive, continuous data streams
- AI automatically organizes, cleans, and enriches data
- Blockchain ensures trusted, tamper-proof data provenance

Enhance Automated Decision-making

- AI models analyse data and trigger actions autonomously
- Real-time insights drive intelligent automation
- Blockchain enables secure, multi-party data sharing

Improve Real-time Data Analytics

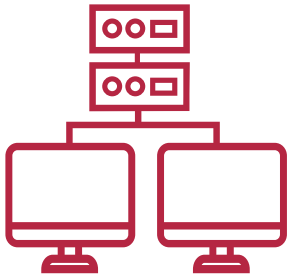
- IoT streaming data processed instantly by AI
- AI-driven predictive analytics for proactive decisions
- Blockchain ensures transparent and verifiable audit trails

Future of Big Data Technology

From Centralized Processing to Distributed Intelligence

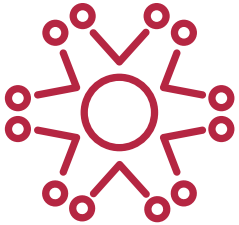
Centralized Infrastructure

- Data lakes
- Batch processing
- Central AI



Distributed Intelligence

- Edge computing
- Federated learning
- Real-time analytics

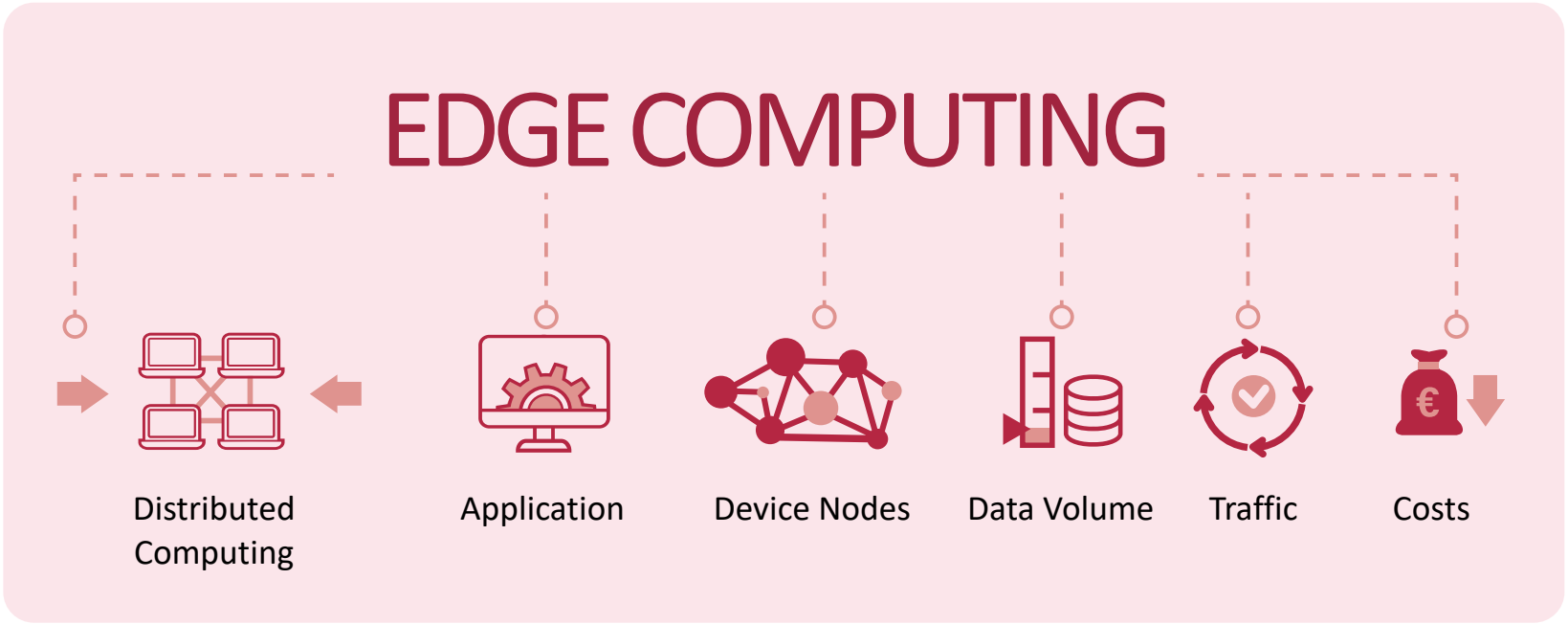


Autonomous Industry 4.0



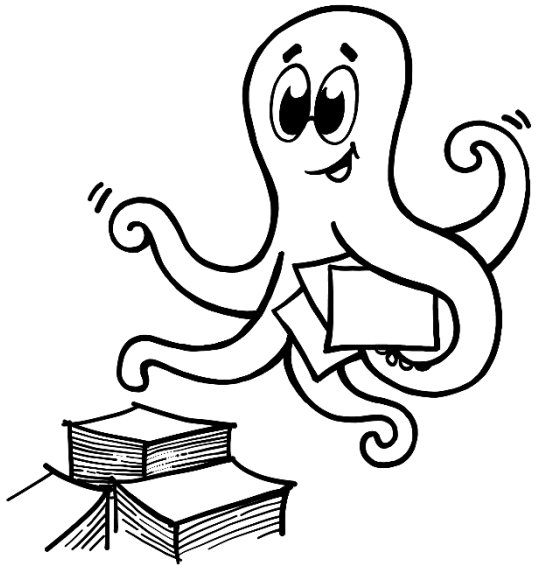
Edge Computing for Data Processing

Performing data processing close to the data source (e.g., IoT devices, sensors, or local gateways) instead of sending all data to a centralized cloud or data center to reduce latency, to achieve lowers bandwidth usage, and to enable real-time analytics and faster decision-making



Take away and must know!

- **Explain** the concept of Big Data and its key characteristics using the V-framework, and distinguish Big Data from traditional data management approaches
- **Describe** how Big Data is transformed into actionable insights
- **Assess** the strategic conditions under which Big Data initiatives create business value, including organizational readiness and the role of data ecosystems



Literature

- Almeida F., Calistru C. (2013) The Main Challenges and Issues of Big Data Management. International Journal of Research Studies in Computing
- Ebner K. et alii (2014) Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments. Proceedings of the 47th Hawaii International Conference on System Sciences

  ISBN 978-1-4799-2505-6
- Laney D. (2001) 3D Data Management: Controlling Data Volume, Velocity, and Variety. Stanford
- Nair, R., Narayanan, A. (2012) Getting Results from Big Data-a Capabilities-Driven Approach to the Strategic Use of Unstructured Information. Booz Co.



Link

https://www.bigdataframework.org/



KNOWLEDGE

www. know-ledge.ch