॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥
IIT Jodhpur

Excellence above all
SME

Technology • Innovation • Leadership

The (Small/Relational) Database Approach

Dr. Deepak Saxena, SME IIT Jodhpur

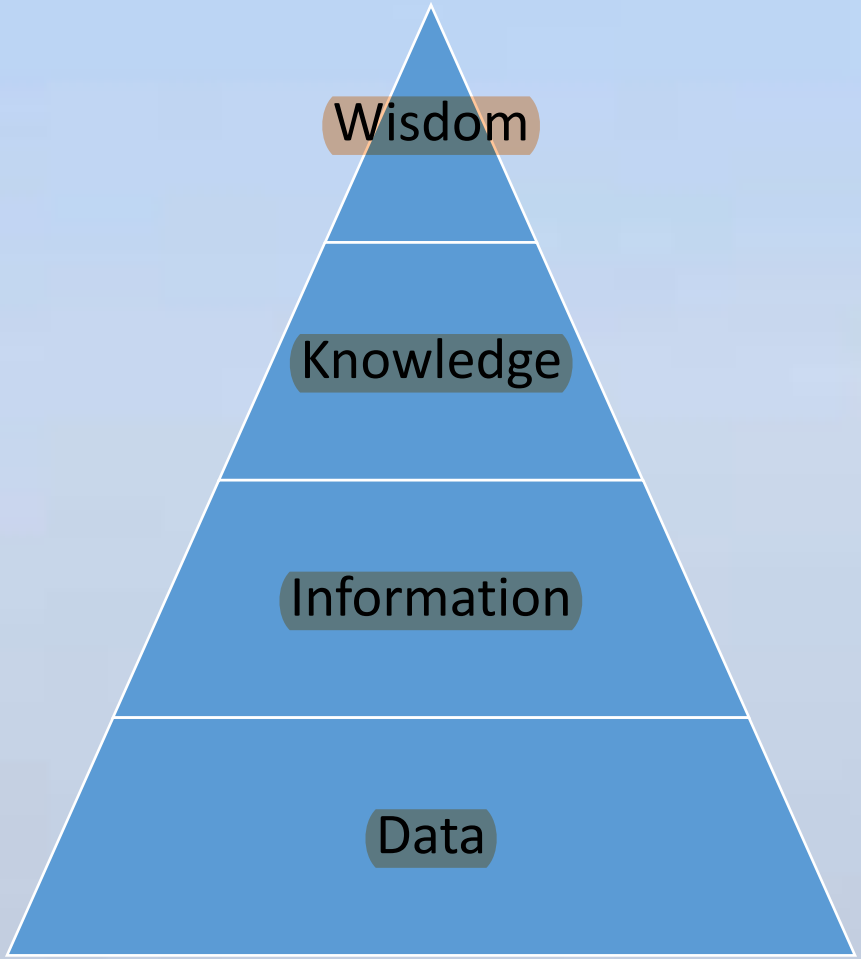

DATA
IS THE NEW OIL

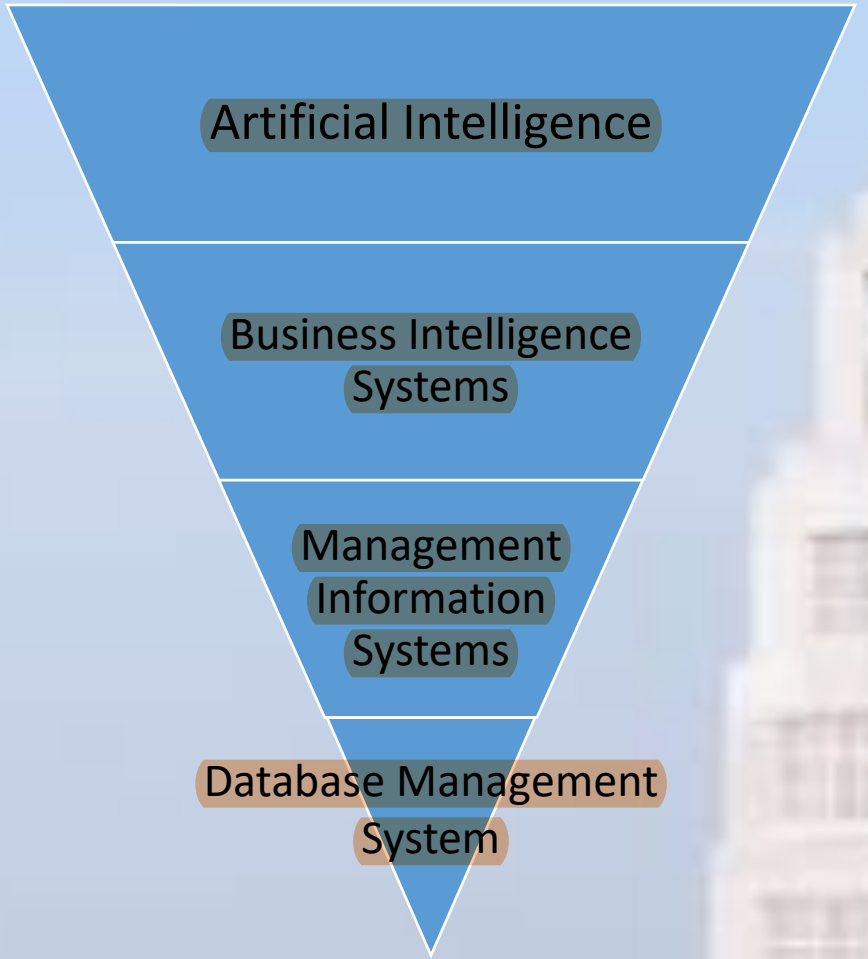

"Guide us, Oh Database Manager!"

Data, Information, and Metadata

- **Data**: Stored representations of objects and events that have meaning and importance in the users' environment.
  - *Objects*: Customer, Items
  - *Events*: Booking, Payment, Delivery
- **Information**: Data that have been processed (e.g. averages, trends etc.) in such as way as to increase the knowledge of the person who uses the data.
- **Metadata**: Data that describes the properties of the end-user data and the context of that data.
- **Database**: an organized collection of logically-related data

Information Systems for DIKW Model





Traditional File Processing Systems

- Program-file dependence
- Duplication of data
- Limited data sharing
- Lengthy development times
- Excessive program maintenance

Customer	Meal1	Date1	Cost1	Meal2	Date2	Cost2
J. Smith	Steak	2/1/2013	$ 20.00	Lobster	2/3/2013	$ 25.00
Jan						
R. Doyle	Veal	3/1/2013	$ 30.00	Shrimp	5/10/2013	$ 20.00
E. Pengler	Steak	2/5/2013	$ 20.00	Steak	7/8/2013	$ 20.00

Relational Data Model

- The relational data model represents data in the form of tables.
- Consists of the following three components:
  1. ***Data structure*** Data are organized in the form of tables, with rows and columns.
  2. ***Data manipulation*** Powerful operations (typically implemented using SQL) are used to manipulate data stored in the relations.
  3. ***Data integrity*** The model includes mechanisms to specify business rules that maintain the integrity of data when they are manipulated.

Relation

- A named, two-dimensional table of data.
- Each relation (or table) consists of a set of named columns and an arbitrary number of unnamed rows.
- An attribute is a named column of a relation.
- Each row of a relation corresponds to a record that contains data (attribute) values for a single entity.
- Example: EMPLOYEE1(EmpID, Name, DeptName, Salary)

Properties of Relations

1. Each relation (or table) in a database has a unique name.
2. An entry at the intersection of each row and column is atomic (or single valued).There can be only one value associated with each attribute on a specific row of a table; no multivalued attributes are allowed in a relation.
3. Each row is unique; no two rows in a relation can be identical.
4. Each attribute (or column) within a table has a unique name.
5. The sequence of columns (left to right) is insignificant. The order of the columns in a relation can be changed without changing the meaning or use of the relation.
6. The sequence of rows (top to bottom) is insignificant. As with columns, the order of the rows of a relation may be changed or stored in any sequence.

No multivalued attributes in a relation

(a) Table with repeating groups

EmpID	Name	DeptName	Salary	CourseTitle	DateCompleted
100	Margaret Simpson	Marketing	48,000	SPSS	6/19/2015
				Surveys	10/7/2015
140	Alan Beeton	Accounting	52,000	Tax Acc	12/8/2015
110	Chris Lucero	Info Systems	43,000	Visual Basic	1/12/2015
				C++	4/22/2015
190	Lorenzo Davis	Finance	55,000		
150	Susan Martin	Marketing	42,000	SPSS	6/16/2015
				Java	8/12/2015

(b) EMPLOYEE2 relation

EMPLOYEE2

EmpID	Name	DeptName	Salary	CourseTitle	DateCompleted
100	Margaret Simpson	Marketing	48,000	SPSS	6/19/2015
100	Margaret Simpson	Marketing	48,000	Surveys	10/7/2015
140	Alan Beeton	Accounting	52,000	Tax Acc	12/8/2015
110	Chris Lucero	Info Systems	43,000	Visual Basic	1/12/2015
110	Chris Lucero	Info Systems	43,000	C++	4/22/2015
190	Lorenzo Davis	Finance	55,000		
150	Susan Martin	Marketing	42,000	SPSS	6/19/2015
150	Susan Martin	Marketing	42,000	Java	8/12/2015

Anomalies

- An error or inconsistency that may result when a user attempts to update a table that contains redundant data.
- Indication that your table design is not proper.
- Three types:
  - Insertion anomalies
  - Deletion anomaly
  - Modification anomaly

EMPLOYEE2

EmpID	Name	DeptName	Salary	CourseTitle	DateCompleted
100	Margaret Simpson	Marketing	48,000	SPSS	6/19/2015
100	Margaret Simpson	Marketing	48,000	Surveys	10/7/2015
140	Alan Beeton	Accounting	52,000	Tax Acc	12/8/2015
110	Chris Lucero	Info Systems	43,000	Visual Basic	1/12/2015
110	Chris Lucero	Info Systems	43,000	C++	4/22/2015
190	Lorenzo Davis	Finance	55,000		
150	Susan Martin	Marketing	42,000	SPSS	6/19/2015
150	Susan Martin	Marketing	42,000	Java	8/12/2015

Solution: Normalization

EmpID	CourseTitle	DateCompleted
100	SPSS	6/19/2015
100	Surveys	10/7/2015
140	Tax Acc	12/8/2015
110	Visual Basic	1/12/2015
110	C++	4/22/2015
150	SPSS	6/19/2015
150	Java	8/12/2015

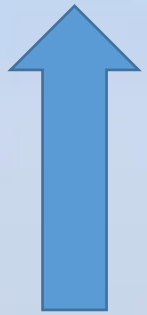
Relational Keys

- **Primary key:** An attribute or a combination of attributes that uniquely identifies each row in a relation.
- Example: EMPLOYEE1(EmpID, Name, DeptName, Salary)

- **Composite key:** A primary key that consists of more than one attribute.
- Example: DEPENDENT1 (EmpID, DependentName, Relationship)

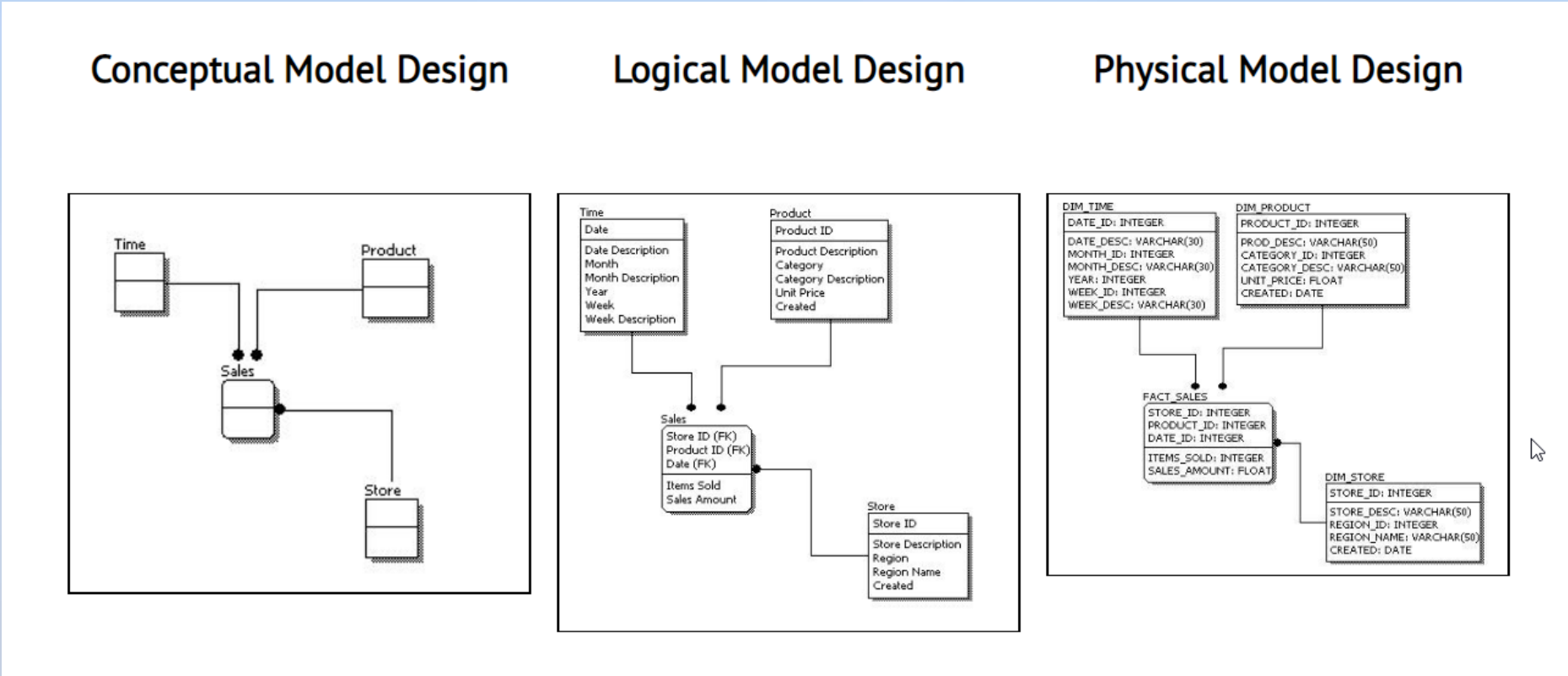
Relational Keys

- **Foreign key:** An attribute in a relation that serves as the primary key of another relation in the same database.
- DEPARTMENT(DeptID, DeptName, Location, Fax)
- EMPLOYEE1(EmpID, Name, DeptID, Salary)

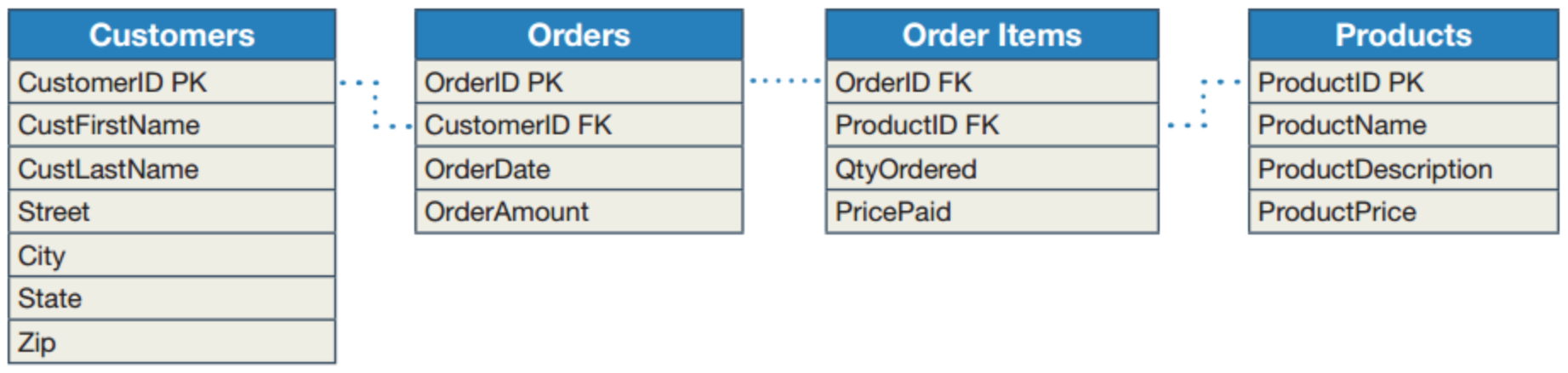


Foreign Key

- The database approach: Data models



- The database approach: Relational database



Integrity Constraints

- Domain Constraints
  - All of the values that appear in a column of a relation must be from the same domain.
  - A domain is the set of values that may be assigned to an attribute.
- Entity Integrity
  - No primary key attribute (or component of a primary key attribute) may be null.
- Referential Integrity
  - Either each foreign key value must match a primary key value in another relation or the foreign key value must be null.

Domain Constraints

- A domain definition usually consists of the following components: domain name, meaning, data type, size (or length), and allowable values or allowable range (if applicable).

TABLE 4-1 Domain Definitions for INVOICE Attributes

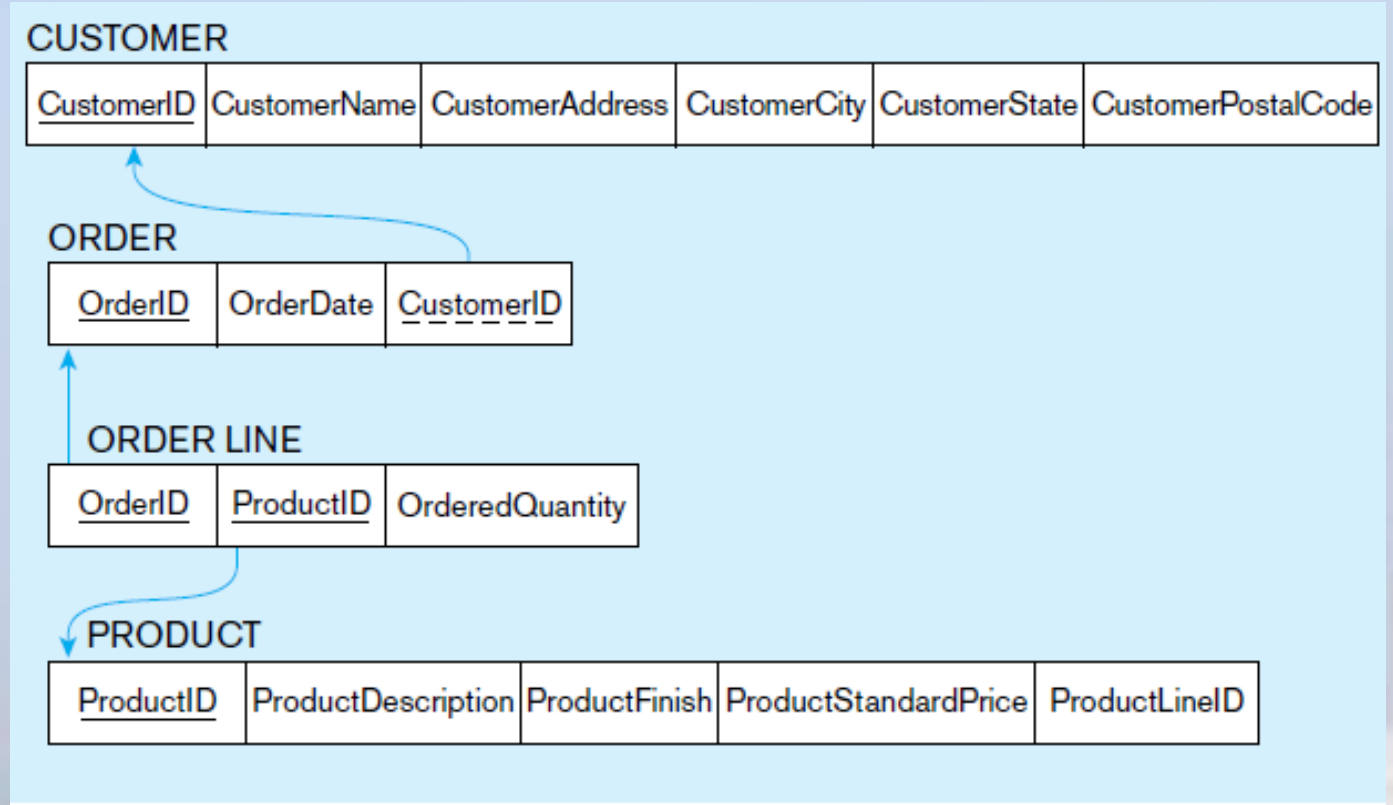
Attribute	Domain Name	Description	Domain
CustomerID	Customer IDs	Set of all possible customer IDs	character: size 5
CustomerName	Customer Names	Set of all possible customer names	character: size 25
CustomerAddress	Customer Addresses	Set of all possible customer addresses	character: size 30
CustomerCity	Cities	Set of all possible cities	character: size 20
CustomerState	States	Set of all possible states	character: size 2
CustomerPostalCode	Postal Codes	Set of all possible postal zip codes	character: size 10
OrderID	Order IDs	Set of all possible order IDs	character: size 5
OrderDate	Order Dates	Set of all possible order dates	date: format mm/dd/yy
ProductID	Product IDs	Set of all possible product IDs	character: size 5
ProductDescription	Product Descriptions	Set of all possible product descriptions	character: size 25
ProductFinish	Product Finishes	Set of all possible product finishes	character: size 15
ProductStandardPrice	Unit Prices	Set of all possible unit prices	monetary: 6 digits
ProductLineID	Product Line IDs	Set of all possible product line IDs	integer: 3 digits
OrderedQuantity	Quantities	Set of all possible ordered quantities	integer: 3 digits

Entity Integrity

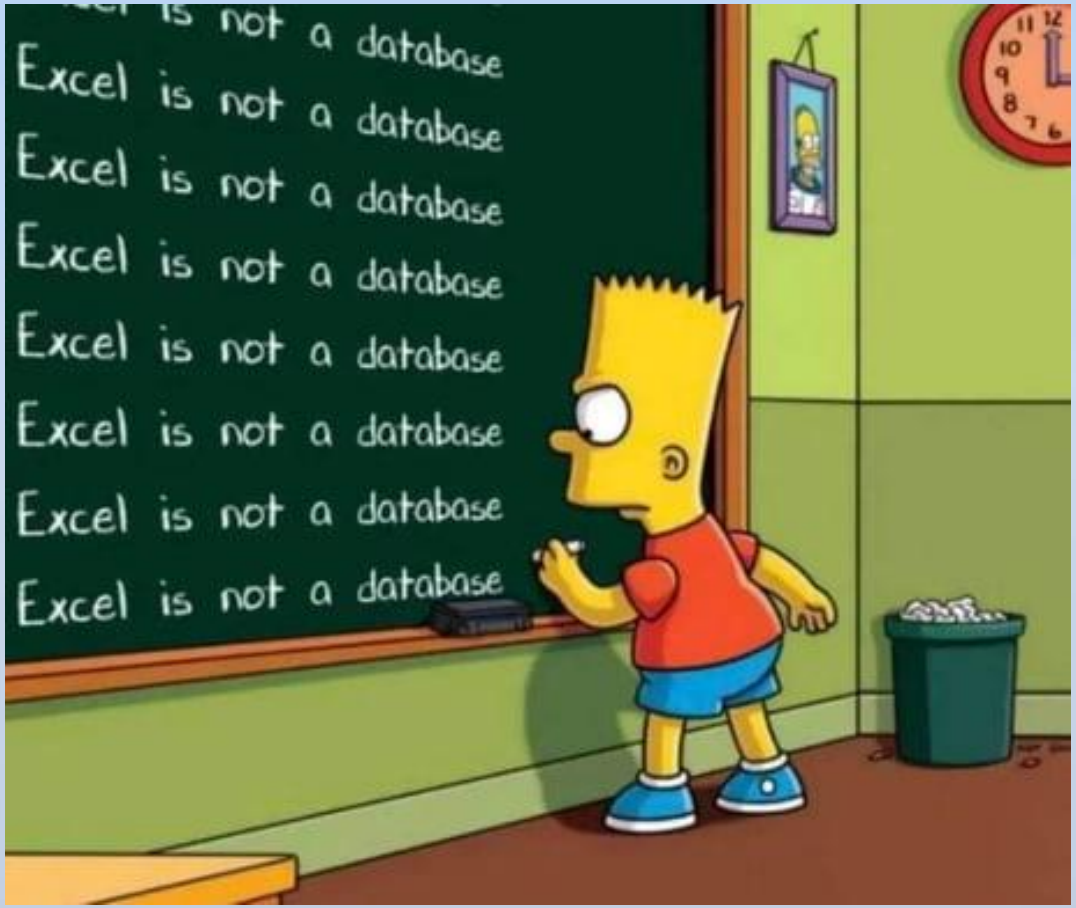
- The entity integrity rule is designed to ensure that every relation has a primary key and that the data values for that primary key are all valid.
- In particular, it guarantees that every primary key attribute is non-null.
- Null Values
  - A null is a value that may be assigned to an attribute when no other value applies or when the applicable value is unknown.
  - In reality, a null is not a value, but rather it indicates the absence of a value.
  - The inclusion of nulls in the relational model is somewhat controversial, because it sometimes leads to anomalous results.
  - However, Codd, the inventor of the relational model, advocates the use of nulls for missing values.

Referential Integrity

- Referential integrity constraint is a rule that maintains consistency among the rows of two relations.
- If there is a foreign key in one relation, either each foreign key value must match a primary key value in another relation or the foreign key value must be null.



Hence…


is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥
IIT Jodhpur

Excellence above all
SME

Technology • Innovation • Leadership

Big Data Introduction

Dr. Deepak Saxena, SME IIT Jodhpur

The size of small/traditional databases

Type of Database / Application	Typical Number of Users	Typical Size of Database
Personal	1	Megabytes
Multitier Client/Server	100–1000	Gigabytes
Enterprise resource planning	>100	Gigabytes–terabytes
Data warehousing	>100	Terabytes–petabytes

Enter Big Data

Why Big Data?

- Proliferation of devices that generate digital data
- Content generation and self-publishing
- Consumer Activity
- Machine data and Internet of Things
- Advances in natural science
- Plummeting cost of storage and processing power
- Open-source platforms
- Cloud computing

3 (or 5) (or 7) Vs of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value
- Variability
- Visualization

Original Vs

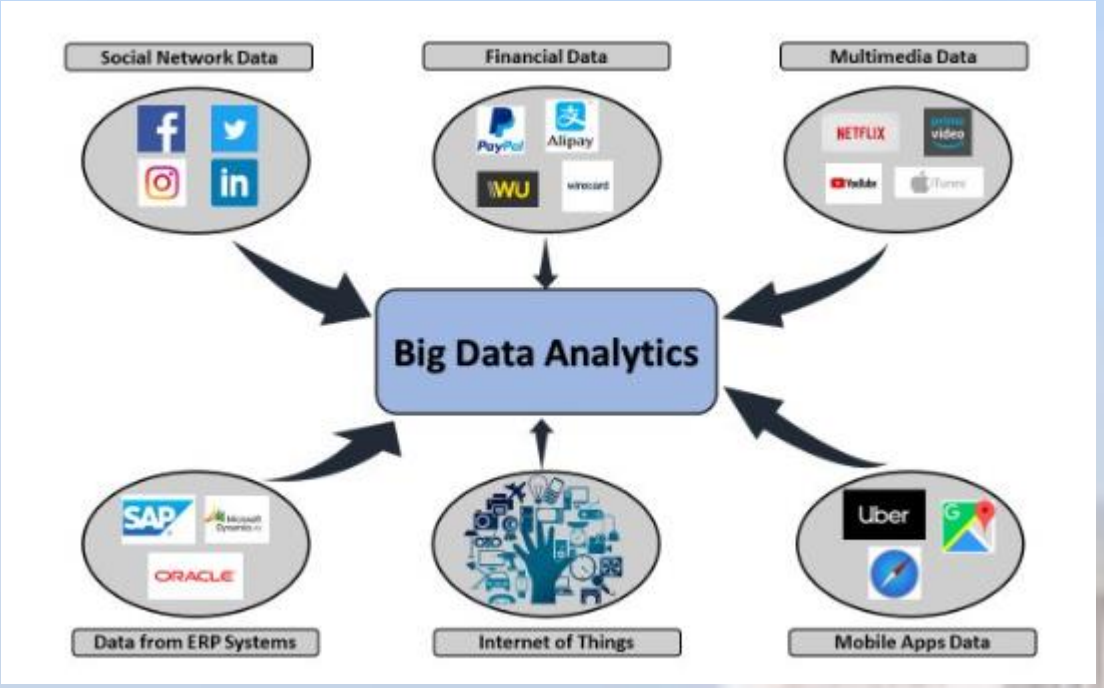
Volume

NAME	SYMBOL	VALUE	EQUAL VALUE
byte	b	8 bits	1 byte
kilobyte	Kb	1024 bytes	1 024 bytes
megabyte	MB	1024 KB	1 048 576 bytes
gigabyte	GB	1024 MB	1 073 741 824 bytes
terabyte	TB	1024 GB	1 099 511 627 776 bytes
Petabyte	PB	1024 TB	1 125 899 906 842 624 bytes
Exabyte	EB	1024 PB	1 152 921 504 606 846 976 bytes
Zetabyte	ZB	1024 EB	1 180 591 620 717 411 303 424 bytes
Yottabyte	YB	1024 ZB	1 208 925 819 614 629 174 706 176 bytes
Brontobyte	BB	1024 YB	1 237 940 039 285 380 274 899 124 224 bytes
Geopbyte	GB	1024 BB	1 267 650 600 228 229 401 496 703 205 376 bytes

- Refers to the generation of large chunk of data.
- It is estimated that the volume of digital data would be around 123 Zettabytes (1 Zettabyte equals 10^{21} bytes) in 2023, reaching a staggering 181 Zettabytes by 2025.
- Data Centres
  - The average data center occupies approximately 100,000 square feet of space.
  - The International Energy Agency estimates that 1% of all global electricity is used by data centers and that by 2025, data centers will consume 1/5 of the world's power supply.

Variety

- Not just relational structured data.
- Various forms of data in disparate formats from disparate sources are combined to generate a holistic picture.
- For instance, Google collects and combines data from various sources (Android operating system, Chrome browser, Gmail, Maps, Search history, Voice, and YouTube activity to name a few) to personalize its offerings to the users
- Social media is a big contributor of Big Data.



Velocity

- Refers to the speed with which data is generated.
- Correct and rapid processing of data in the real time has become extremely crucial for companies.
- Chinese retailer Alibaba processed around 544,000 orders per second on Singles day 2019.
- Amazon on Prime Day 2023
  - An incremental 163 petabytes of EBS storage capacity allocated – generating a peak of 15.35 trillion requests and 764 petabytes of data transfer per day.
  - 5,835 database instances running the PostgreSQL-compatible and MySQL-compatible editions of Amazon Aurora processed 318 billion transactions, stored 2,140 terabytes of data, and transferred 836 terabytes of data.
  - Amazon CloudFront handled a peak load of over 500 million HTTP requests per minute, for a total of over 1 trillion HTTP requests during Prime Day.
  - For more details: https://aws.amazon.com/blogs/aws/prime-day-2023-powered-by-aws-all-the-numbers

Veracity and Value

Veracity

- Refers to the authenticity and truthfulness of the data.
- This is more relevant in case of unstructured data, for instance making sure that service reviews are coming from authentic users and not from automated bots.

Value

- Refers to the outcomes, for instance efficiency or reputational gains, made possible by Big data analytics.
- Beyond operational gains, however, Big data capabilities may act as an enabler of digital enterprise transformation.

Variability and Visualization

Variability

- refers to the variability in the meaning when processing natural language data.
- For example, the term ‘great service’ would differ in meaning depending on being preceded by ‘got the reply within two hours’ or ‘still waiting for reply since last 15 days’.
- It may also refer to inconsistent speed of the data

Visualization

- Primarily relates to the appropriation of data as opposed to its inherent nature.
- Since Big data is unstructured and comes from a variety of sources, visualization of the trends helps in making sense of the vast tapestry of data.

Data Lake


@timoelliott
"What's a data lake for?
So you can drown in more data even faster!"

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥
IIT Jodhpur

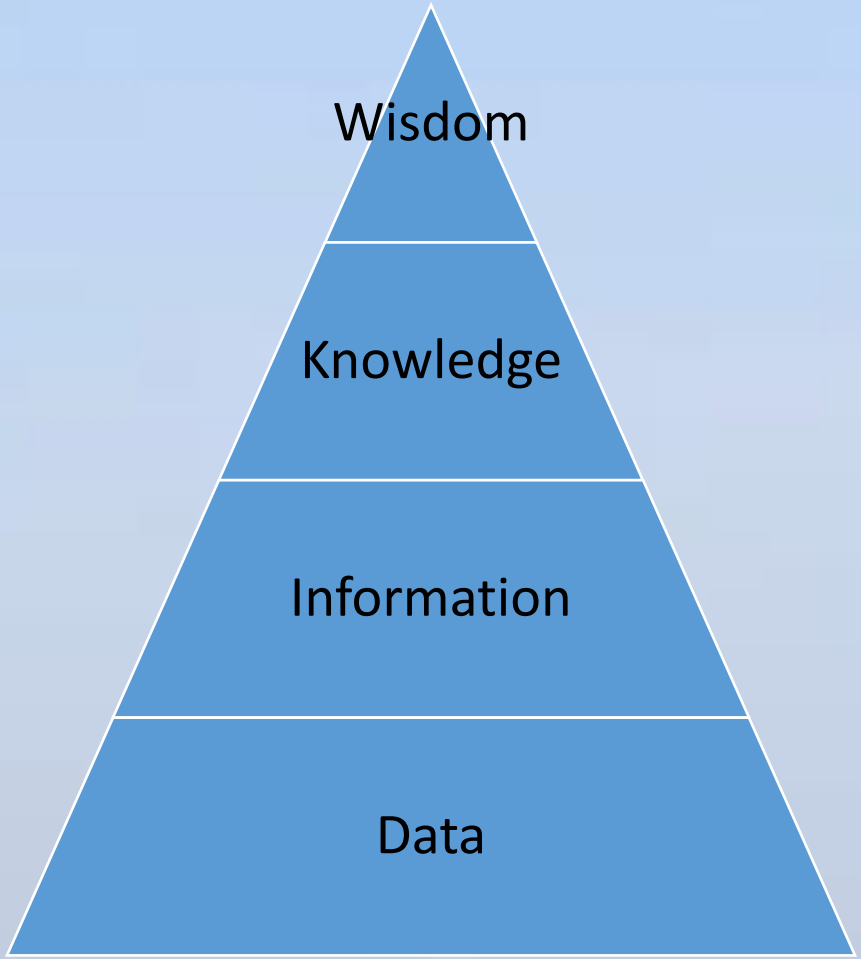
Excellence above all
SME

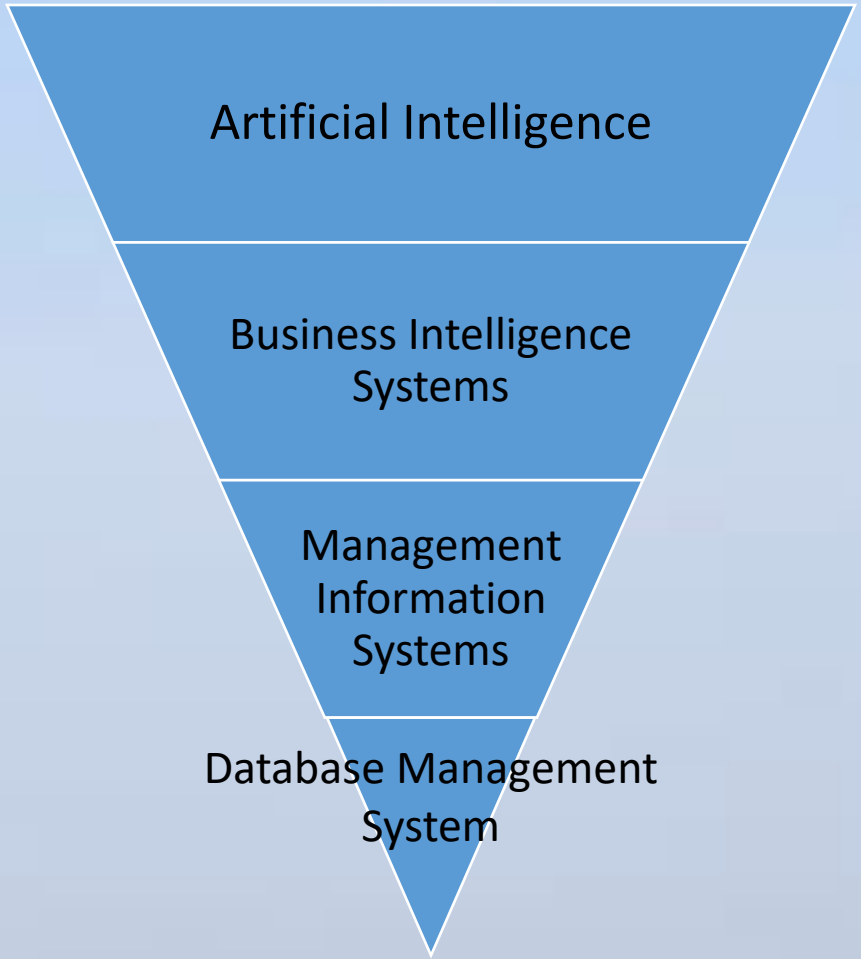
Technology • Innovation • Leadership

Data Science

Dr. Deepak Saxena, SME IIT Jodhpur

Remember this?





Big Data and Cloud Computing in AI

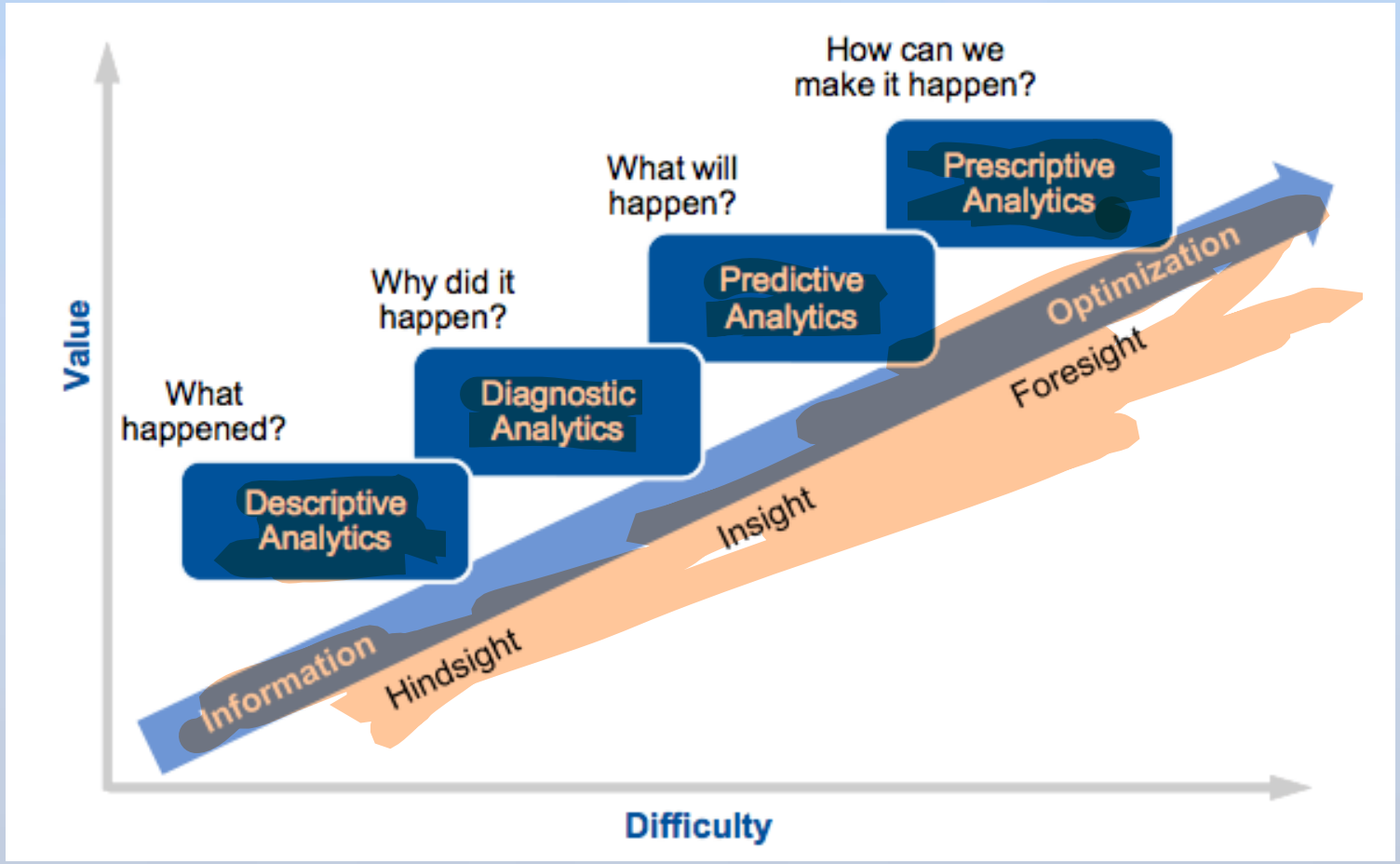
Big Data

- Training: We started amassing huge amounts of data that could be used for machine learning.

Cloud Computing

- We created software that would allow normal computers to work together with the power of a super-computer.

Gartner's Analytics Maturity Model



Descriptive Analytics

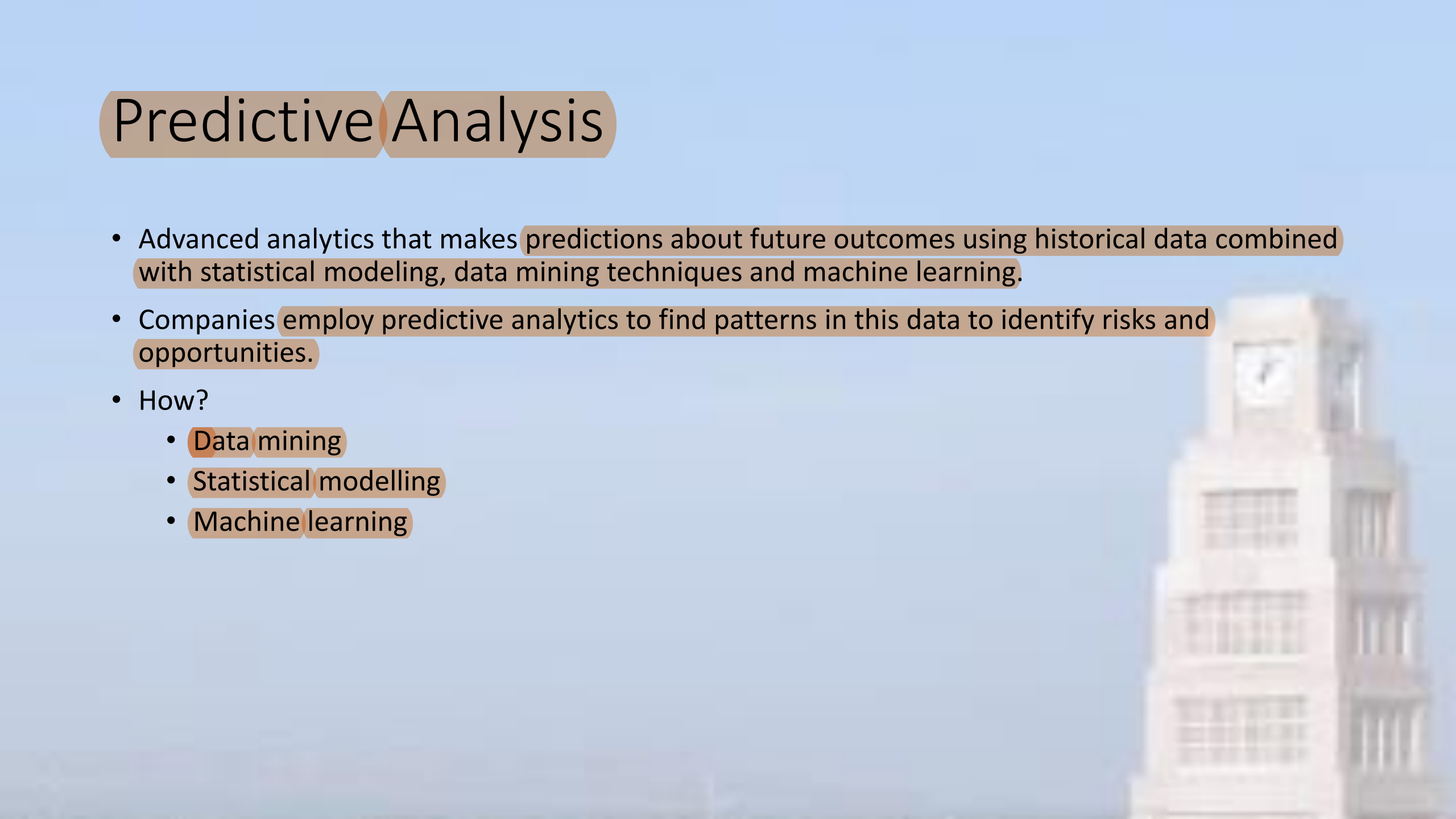
- Why?
  - Operational requirements
  - Data visibility
- How?
  - Excel would do
  - Relational databases (SQL)
  - Data Warehouse

Diagnostic Analytics

- Digs deeper for a root cause analysis.
- Characterized by statistical techniques such as correlation, regression, and/or hypotheses testing.
- How?
  - Excel would do
  - Tableau, PowerBI
  - Data mining

Predictive Analysis

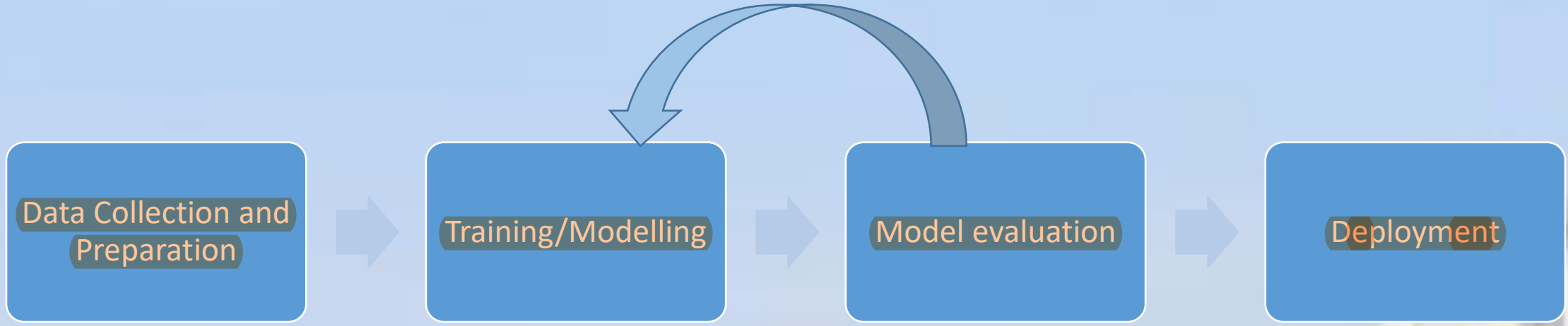
- Advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.
- Companies employ predictive analytics to find patterns in this data to identify risks and opportunities.
- How?
  - Data mining
  - Statistical modelling
  - Machine learning



Prescriptive analytics

- Prescriptive analytics is a process that analyzes data and provides instant recommendations on how to optimize business practices to suit multiple predicted outcomes.
- In essence, prescriptive analytics takes the "what we know" (data), comprehensively understands that data to predict what could happen, and suggests the best steps forward based on informed simulations.
- How?
  - Machine Learning
  - Deep Learning

Analytic Modelling Process



- Correct and complete data
- Feature engineering (ML)

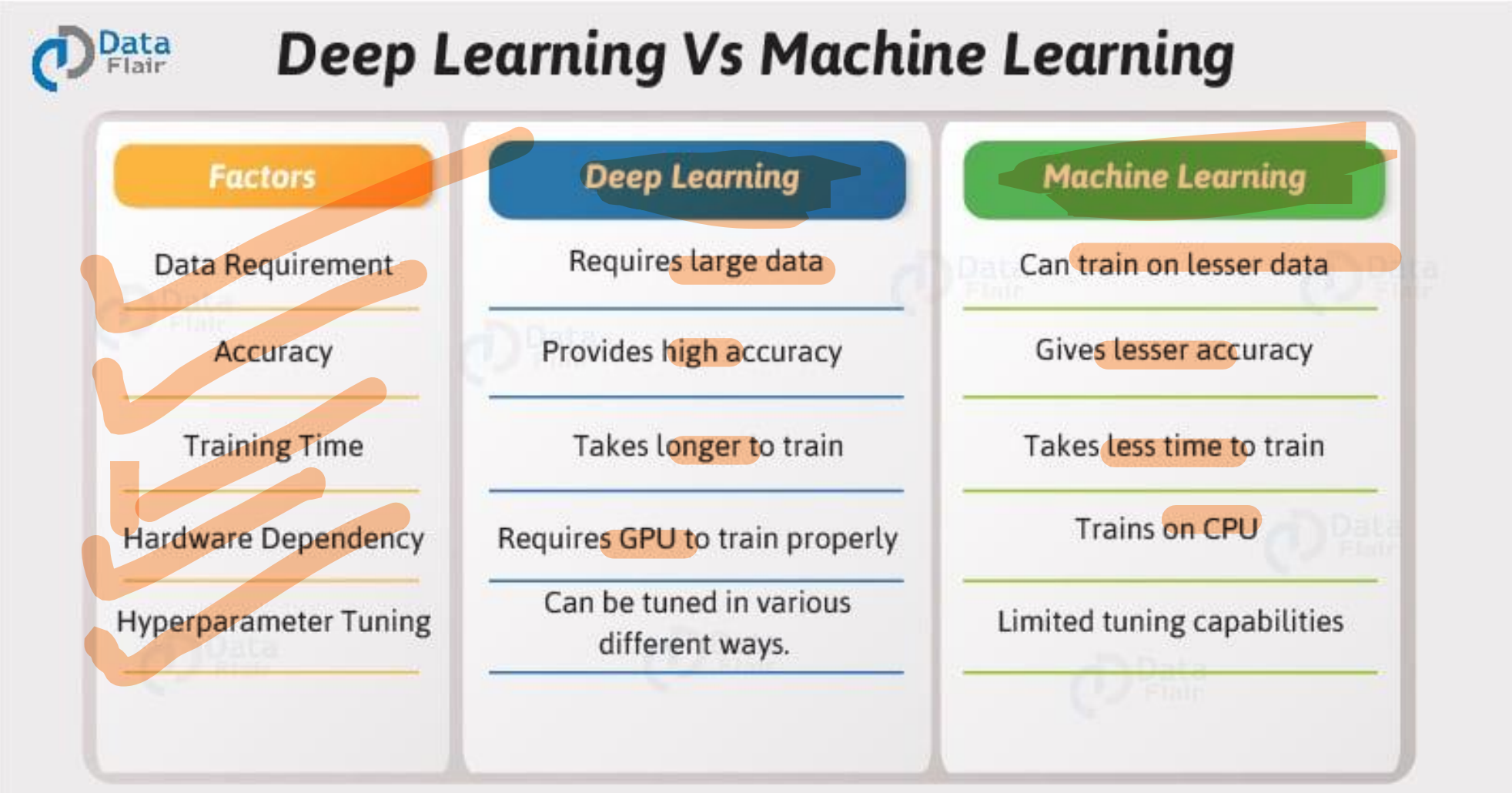
- Selecting Training data
- Applying algorithms for model building

- Model accuracy
- Competing Models

- Usage in business decision-making

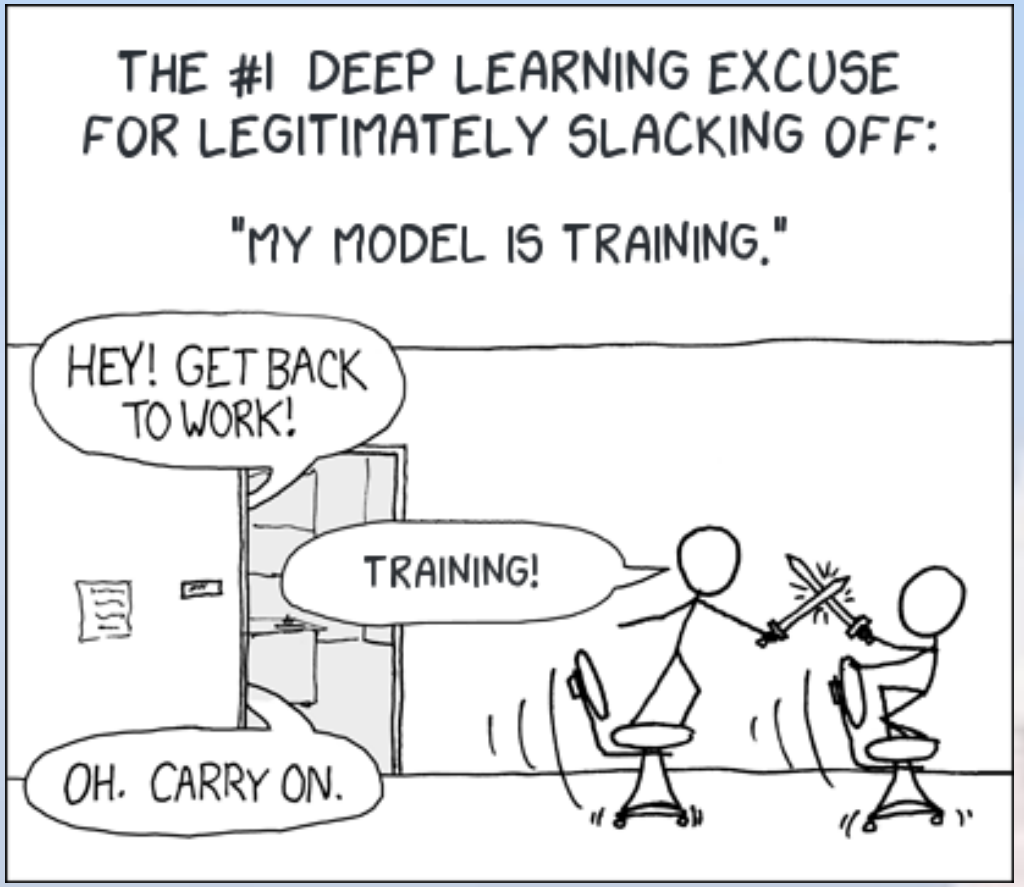
Model-fitting

ML vs DL?

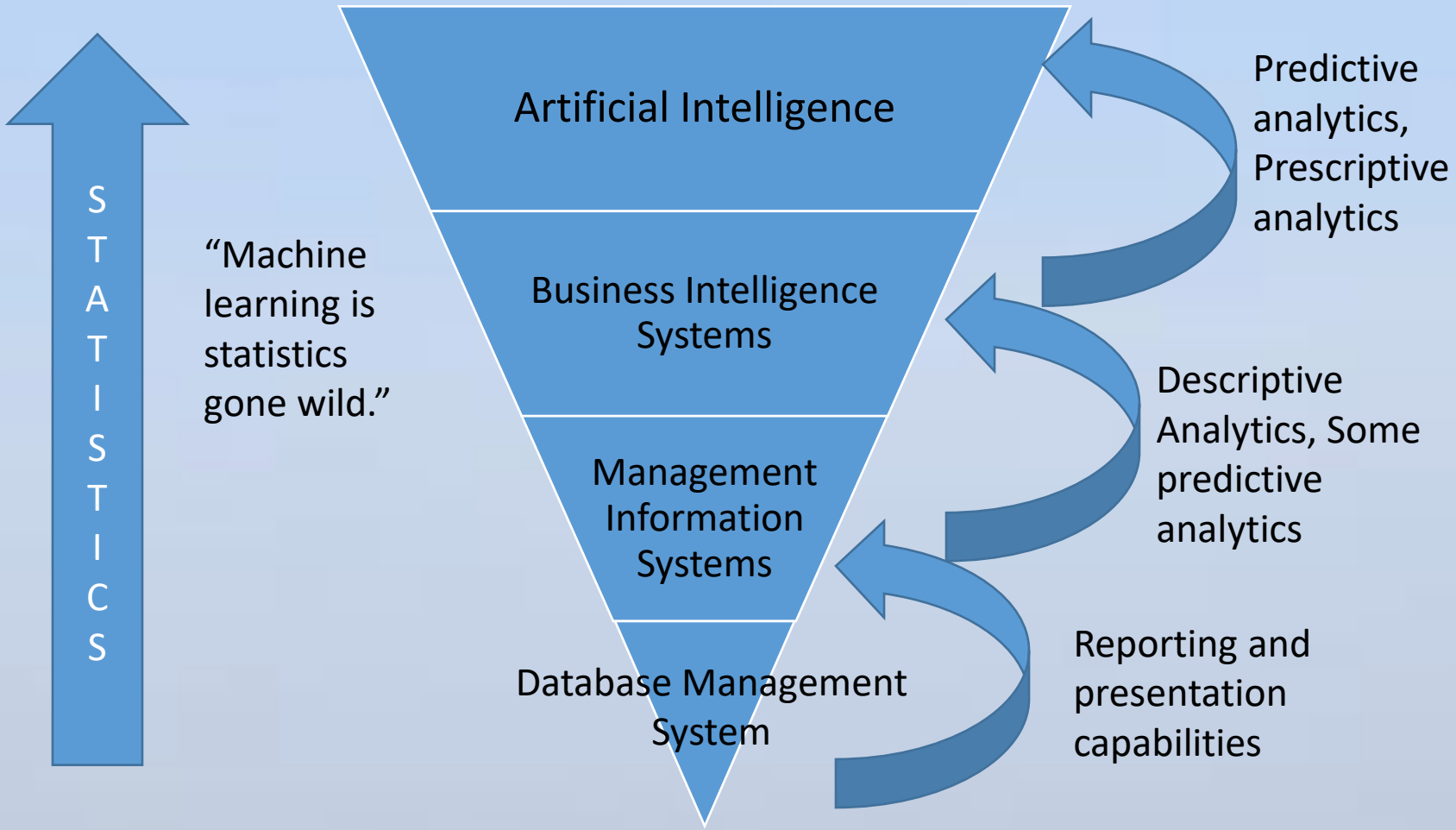

Data Flair
Deep Learning Vs Machine Learning
Factors
Deep Learning
Machine Learning
Data Requirement
Requires large data
Can train on lesser data
Accuracy
Provides high accuracy
Gives lesser accuracy
Training Time
Takes longer to train
Takes less time to train
Hardware Dependency
Requires GPU to train properly
Trains on CPU
Hyperparameter Tuning
Can be tuned in various different ways.
Limited tuning capabilities

ML vs DL?


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS


THE #1 DEEP LEARNING EXCUSE FOR LEGITIMATELY SLACKING OFF:
"MY MODEL IS TRAINING."
HEY! GET BACK TO WORK!
TRAINING!
OH. CARRY ON.

Making Sense



"Machine learning is statistics gone wild."

It all may be applicable together

- Descriptive analytics would flag a revenue shortfall;
- Diagnostic analytics might reveal that it was caused by a shortage of key inventory;
- Predictive analytics could forecast future supply and demand; and
- Prescriptive analytics could optimize pricing, based on the balance of supply and demand, as well as on the price elasticity of the customer base.

Points to ponder

1. You start at the bottom, advancing through the levels in sequence
2. Each higher level brings more value than the lower level before it
3. The way you manage these capabilities lie on the same spectrum

- There is no need to ‘complete’ building out descriptive analytics before moving on to advanced analytics.
- There is no certainty that higher levels of analytics bring more value.
- The different types of work described thrive under starkly different management methods.

How some people see it?

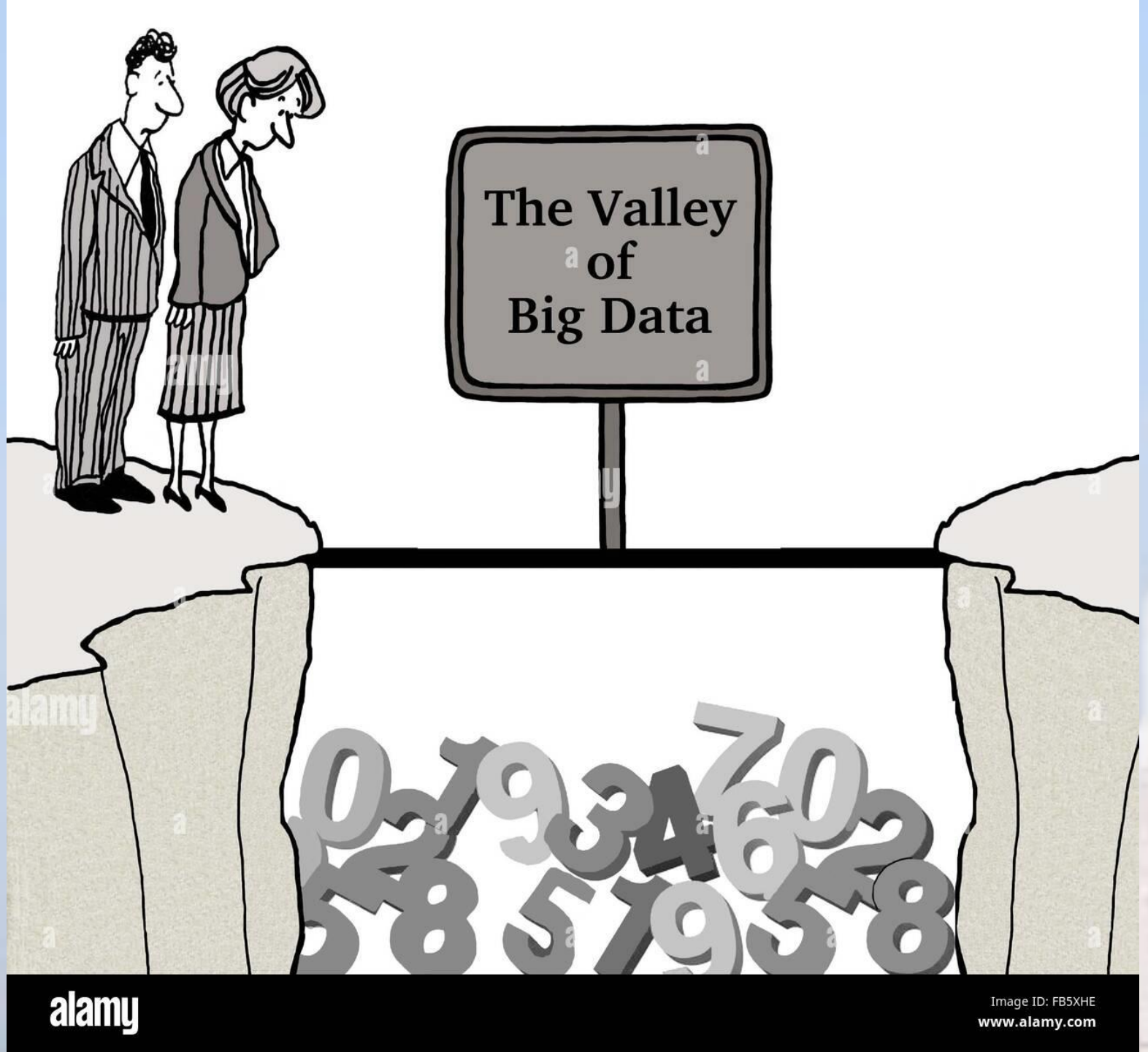
॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥
IIT Jodhpur

Excellence above all
SME

Technology • Innovation • Leadership

Big Data Use Cases

Dr. Deepak Saxena, SME IIT Jodhpur



Operations

- Predictive and preventive maintenance
  - Potential issues can be discovered by analyzing both structured data (equipment year, make, and model) and multi-structured data (log entries, sensor data, error messages, engine temperature, and other factors).
- Operational Efficiency
  - Analyze and assess production processes, proactively respond to customer feedback, and anticipate future demands.

Marketing and Retail

- Product development
  - By classifying key attributes of past and current products and then modeling the relationship between those attributes and the commercial success of the offerings, you can build predictive models for new products and services.
- Customer Experience
  - By gathering data from social media, web visits, call logs and other company interactions, and other data sources, companies can improve customer interactions and maximize the value delivered.
  - Big data analytics can be used to deliver personalized offers, reduce customer churn, and proactively handle issues.
- Reduce Customer Churn
  - By analyzing the data about service quality, convenience, and other factors, companies can predict overall customer satisfaction. They can set up alerts when customers are at risk of churning—and take action with retention campaigns and proactive offers.

Marketing and Retail

- Customer Lifetime Value
  - Big data provides you with insights on customer behavior and spending patterns so you can identify your best customers.
  - Such customers need to be treated with special care.
- In-store shopping experience
  - Many retailers are starting to analyze data from mobile apps, in-store purchases, and geolocations to optimize merchandizing encourage customers to complete purchases.
- Pricing Analytics
  - Price optimization and discounting.

Big data sources for customer data

- Visits to your digital platforms: websites, apps and kiosks'.
- Interactions with customer support: phone, email, online chat, etc.
- Social media, including direct messaging, tweets, and posts on accounts you own, or they own.
- Records of physical movement, including store videos and movement logs
- Additional sensor data from sensors, RFID tags, personal fitness trackers, etc., which may provide data such as bio-medical readings, accelerometer data, external temperature, etc.

Healthcare

- Genomic Research
  - Researchers can identify disease genes and biomarkers to help patients pinpoint health issues they may face in the future.
  - The results can even allow healthcare organizations to design personalized treatments.
- Patient Experience and Outcomes
  - With big data, healthcare organizations can create a 360-degree view of patient care as the patient moves through various treatments and departments.
- Pandemic Management
  - detection of existing cases, prediction of future outbreak, anticipation of potential preventive and therapeutic agents, and assistance in informed decision-making
- Claim Frauds
  - For every healthcare claim, there can be hundreds of associated reports in a variety of different formats.
  - Big data helps healthcare organizations detect potential fraud by flagging certain behaviors for further examination.

Financial Services

- Fraud and compliance
  - Using big data, companies can identify patterns that indicate fraud and aggregate large volumes of information to streamline regulatory reporting.
- Risk modelling
  - Financial services companies can bring together a large volume of data, create advanced risk models, and do this quickly without adversely affecting other projects.

At least so far, it is still BIG Data

Big Data Use Cases – Public Good

Dr. Deepak Saxena, SME IIT Jodhpur

Tax Administration

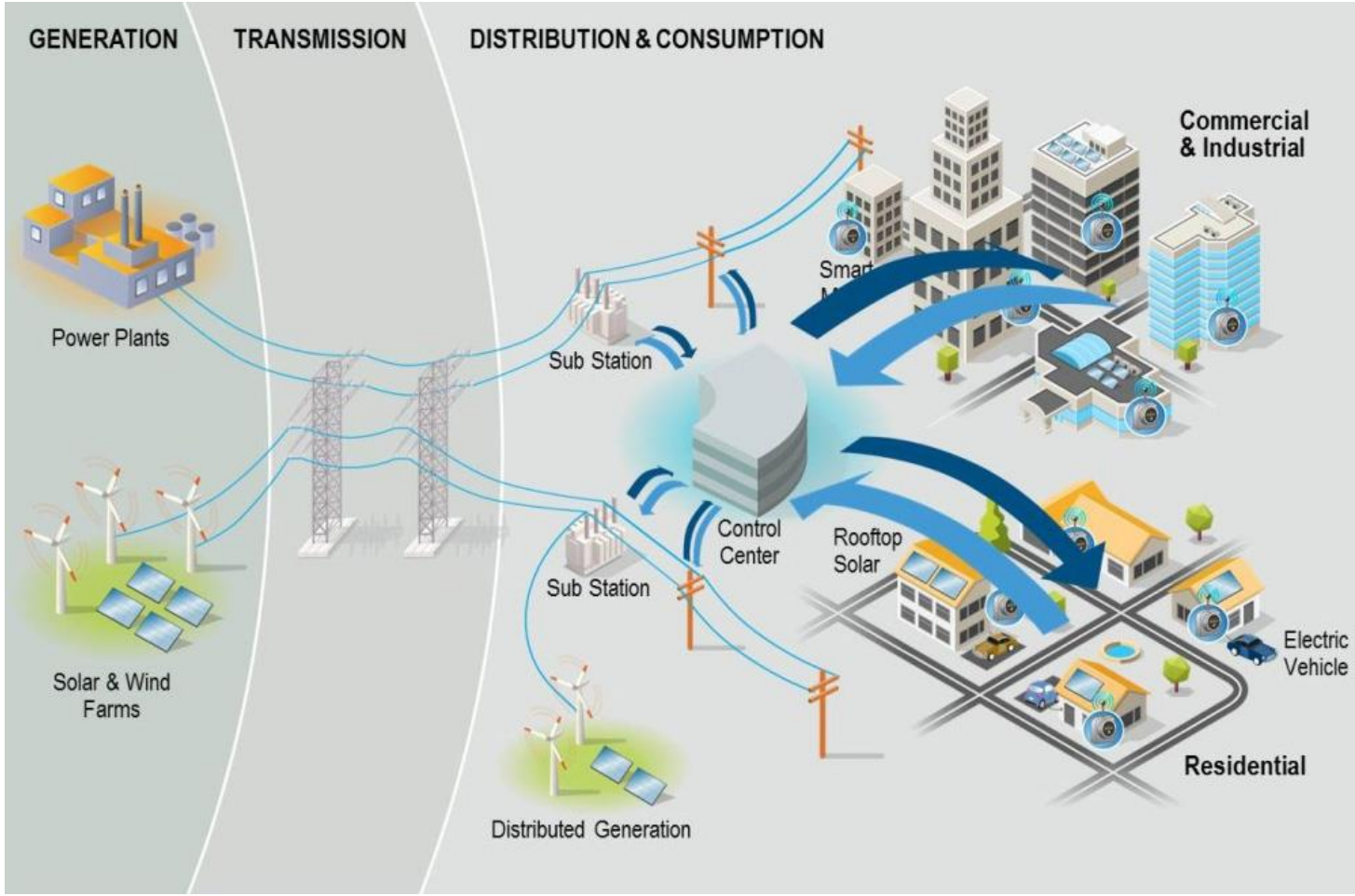
- Supporting users/businesses in filing correct taxes
- Identifying tax frauds
- Identification of high risk entities

Transportation

- Route planning
- Real-time congestion handling
- Ensuring reliability of the services
- Self-driving cars

Energy Management

- Production planning
- Fault detection
- Smart meters/homes



Public Health

- Electronic health records
- Patient-centered care
- Predictive care
- Infectious disease planning and control

Disaster Management

- Disaster prediction
- Real-time disaster assistance

Challenges

- Data quality
- Infrastructure cost
- Adequate Human Resource
- Privacy and Security

New Trends

- Shared services
- Cloud computing
- Open data
- Regulatory compliance

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥
IIT Jodhpur

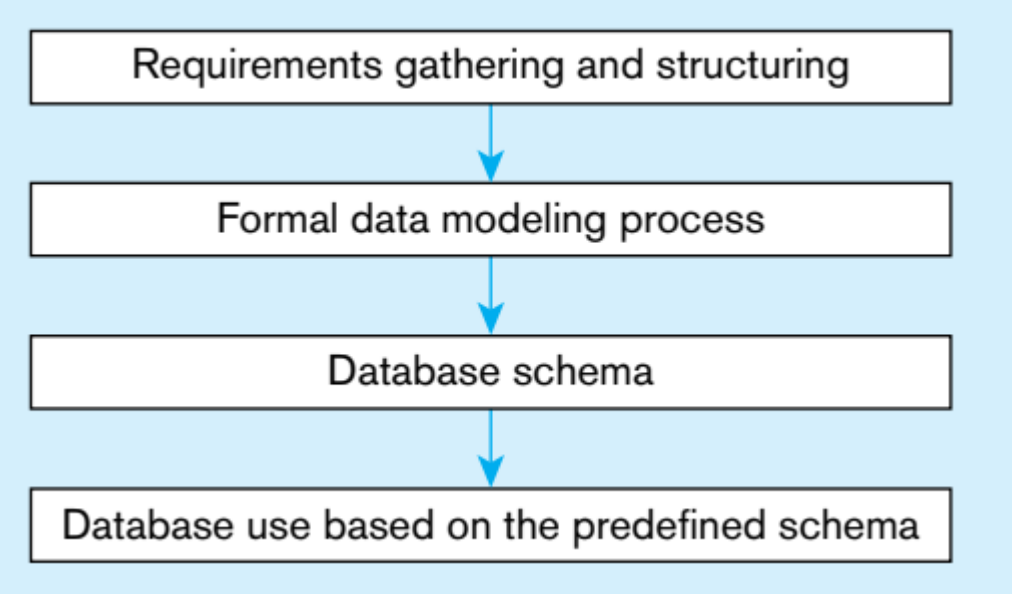
Excellence above all
SME

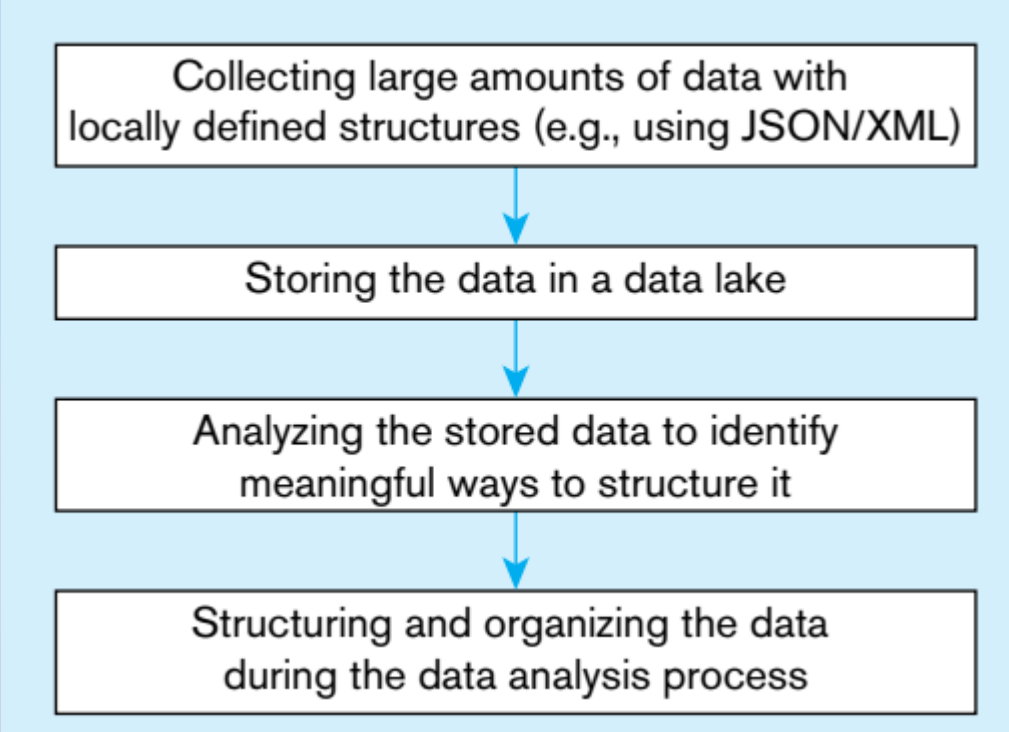
Technology • Innovation • Leadership

Big Data Ecosystem - Database

Dr. Deepak Saxena, SME IIT Jodhpur

Relational vs Big Data Approach



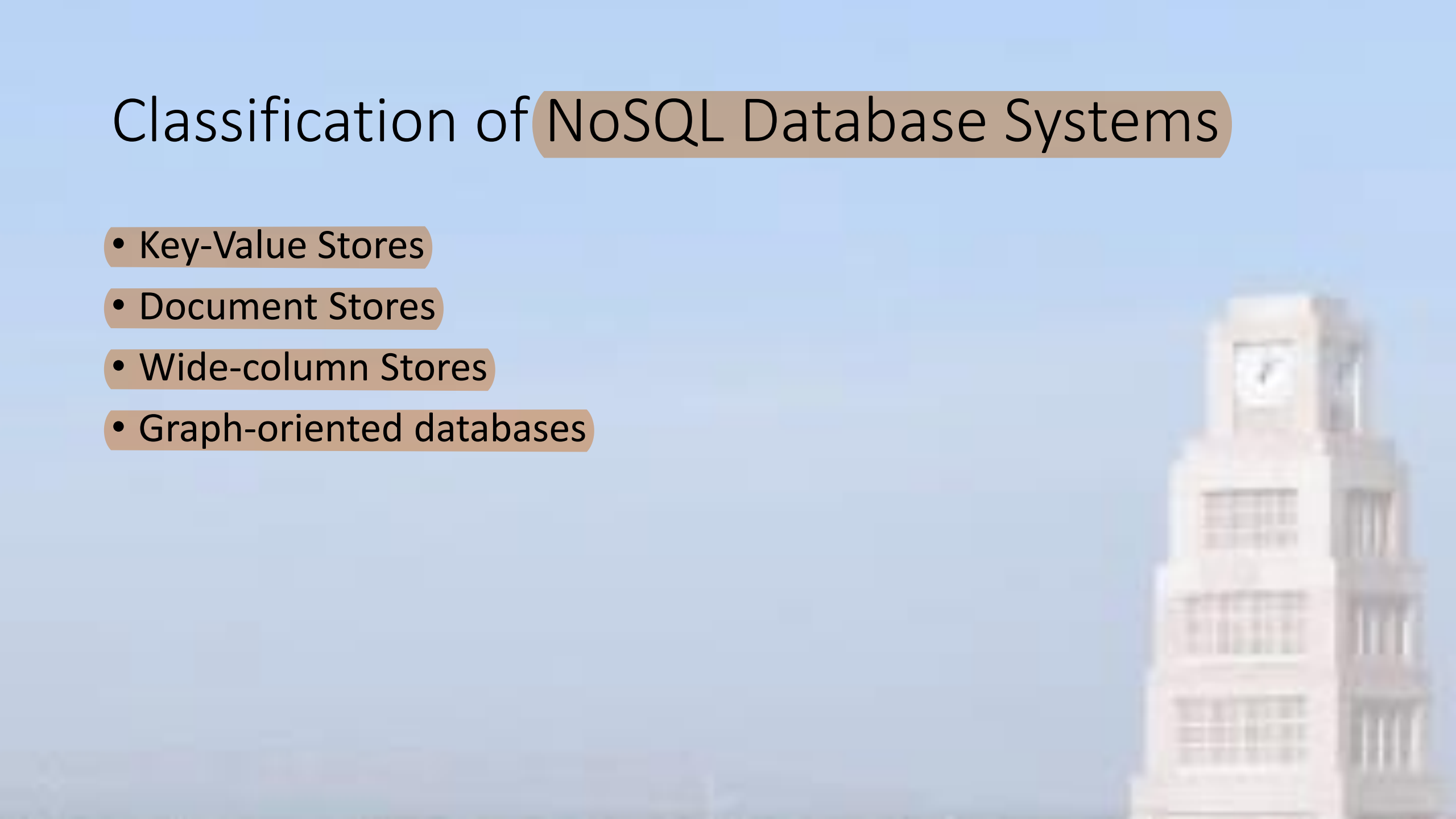


Schema on Write
SQL (Structured Query Language)
Suitable for Data Warehouse

Schema on Read
NoSQL (Not Only SQL)
Suitable for Data Lake

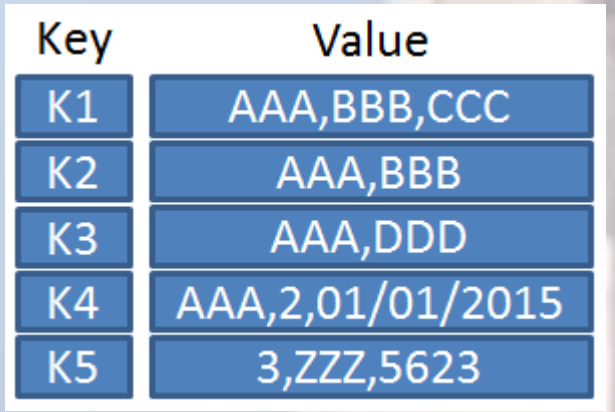
Classification of NoSQL Database Systems

- Key-Value Stores
- Document Stores
- Wide-column Stores
- Graph-oriented databases



Key-Value Stores

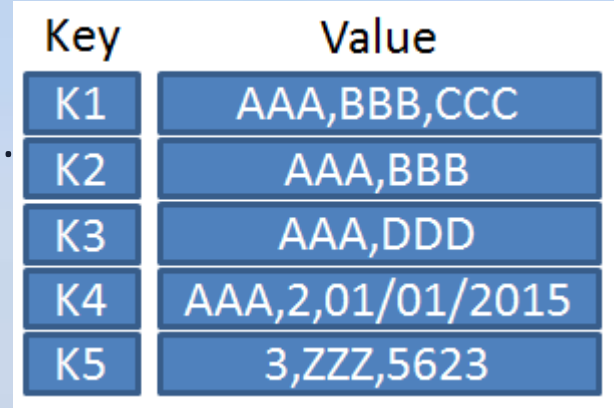
- A key-value store database maintains a structure that allows it to store and access “values” based on a “key”.
- The “key” is typically a string, with or without specific meaning.
- If some part of the “value” needs to be changed, the entire collection will need to be updated.
- Software solutions: REDIS, Amazon DynamoDB
- When to use key value database?
  - Handling Large Volume of Small and Continuous Reads and Writes
  - Storing Basic Information
  - Applications with Infrequent Updates and Simple Queries
  - Key-Value Databases for Volatile Data



Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

Key-Value Stores

- Use Cases
  - Session management on a large scale.
  - Using cache (in-memory database) to accelerate application responses.
  - Storing personal data on specific users.
  - Product recommendations, storing personalized lists.
  - Managing each player's session in massive multiplayer online games.



Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

Key Value Databases

Advantages

- Simplicity
- Speed
- Scalability
- Easy to move

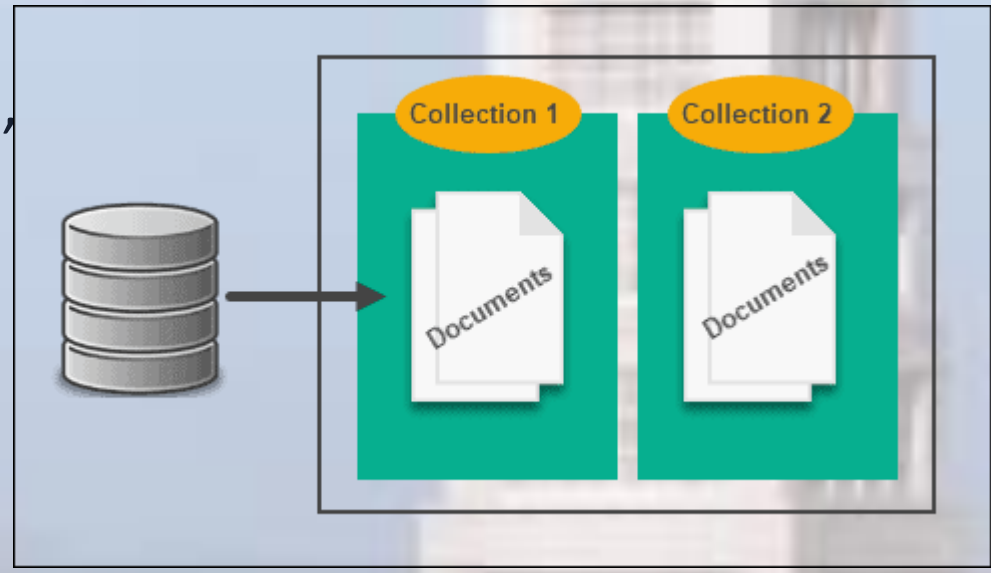
Disadvantages

- Simplicity
- No query language
- Values can't be filtered

Document Stores

- A document in this context is a structured set of data formatted using a standard such as JSON, BSON, or XML.
- The key difference between key-value stores and document stores is that a document store has the capability of accessing and modifying the contents of a specific document based on its structure.
- The “documents” may have a hierarchical structure, and they *do no*t typically reference each other.
- Software Solutions: MongoDB, Amazon DocumentDB

Key	Document
1001	{ "CustomerID": 99, "OrderItems": [{ "ProductID": 2010, "Quantity": 2, "Cost": 520 }, { "ProductID": 4365, "Quantity": 1, "Cost": 18 }], "OrderDate": "04/01/2017" }
1002	{ "CustomerID": 220, "OrderItems": [{ "ProductID": 1285, "Quantity": 1, "Cost": 120 }], "OrderDate": "05/08/2017" }



Document Stores: Use Cases

- Customer data management and personalization
- Internet of Things (IoT) and time-series data
- Product catalogs and content management
- Payment processing
- Mobile apps
- Operational analytics
- Real-time analytics

Document Stores

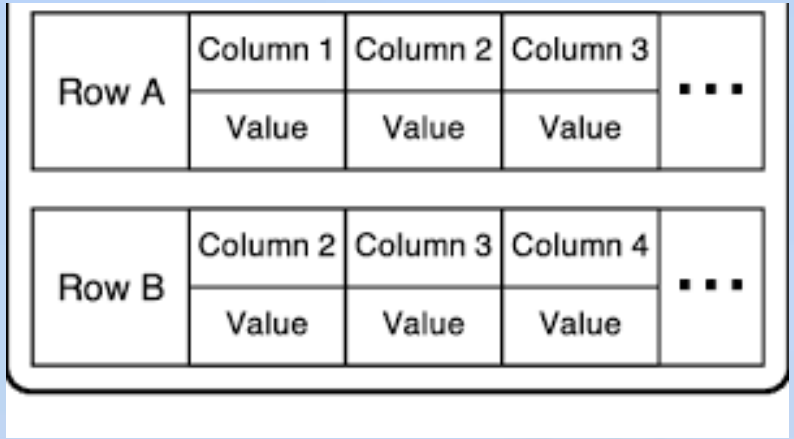
Advantages

- Schema-less
- Faster creation and maintenance
- No foreign keys
- Open formats
- Built-in versioning

Disadvantages

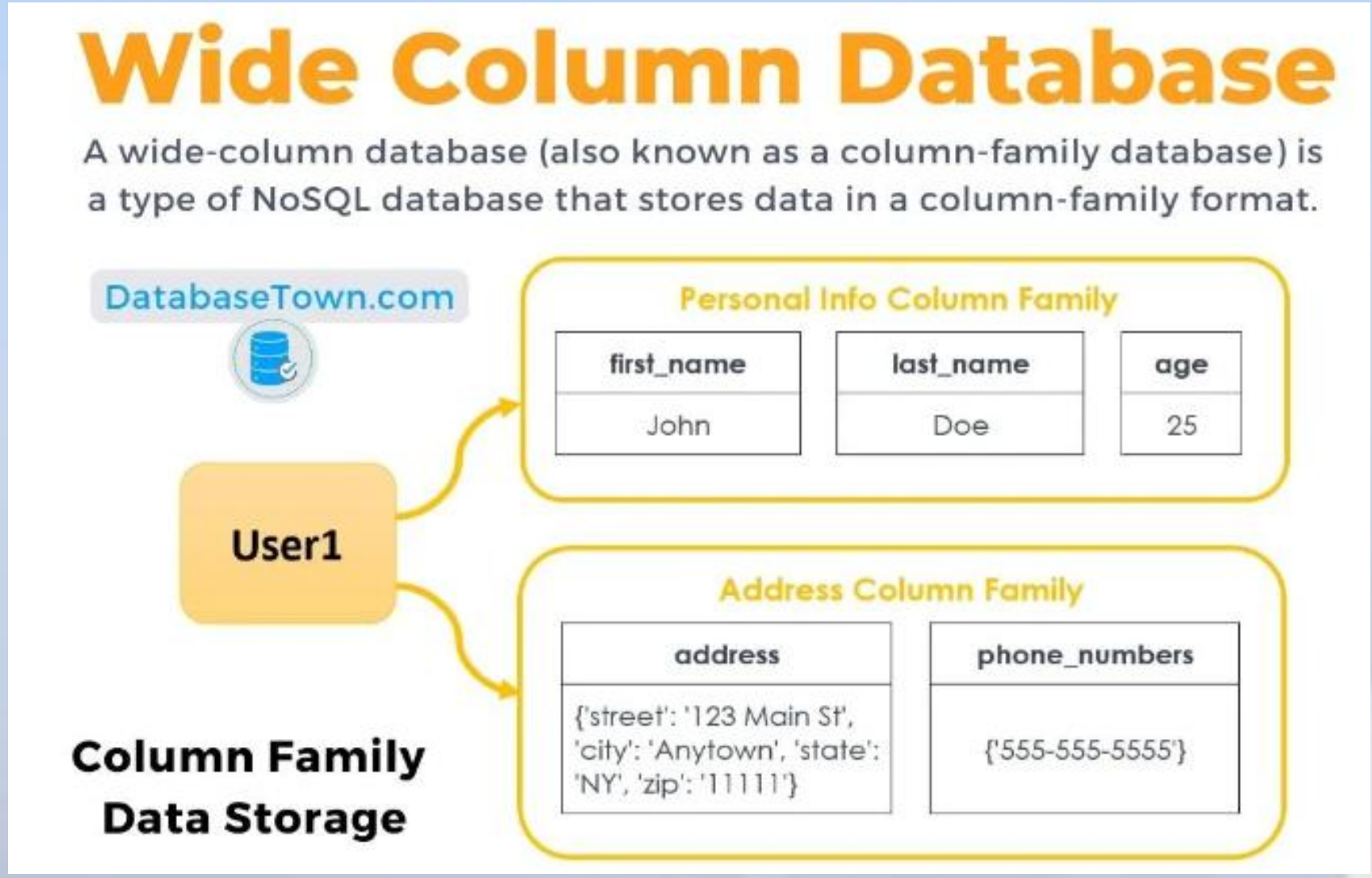
- Consistency-Check Limitations
- Security

Wide-column Stores

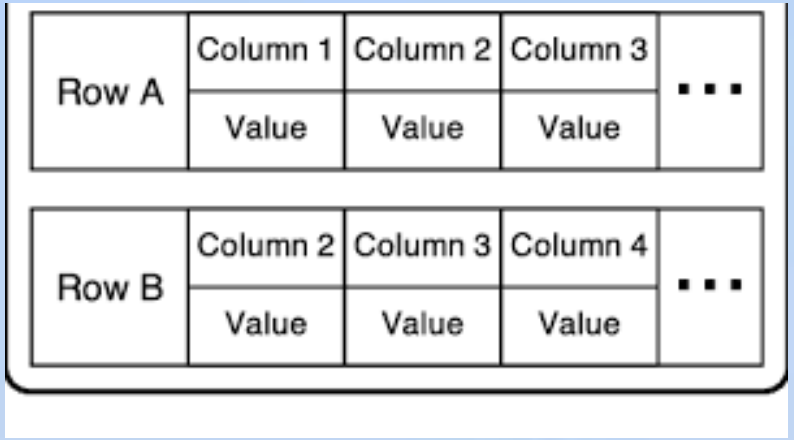


- Consist of rows and columns, and their characteristic feature is the distribution of data based on both key values (records) and columns, using “column groups” or “column families” to indicate which columns are best to be stored together.
- They allow each row to have a different column structure (there are no constraints defined by shared schema), and the length of the rows varies.
- A column is only written if there is a data element for it.
- Software solutions: Apache Cassandra, BigTable, ScyllaDB

Wide-column Stores



Wide-column Stores: Use Cases



- Log data
- IoT (Internet of Things) sensor data
- Time-series data, such as temperature monitoring or financial trading data
- Attribute-based data, such as user preferences or equipment features
- Real-time analytics
- High throughput data such as gaming or e-commerce

Wide-column Stores

Advantages

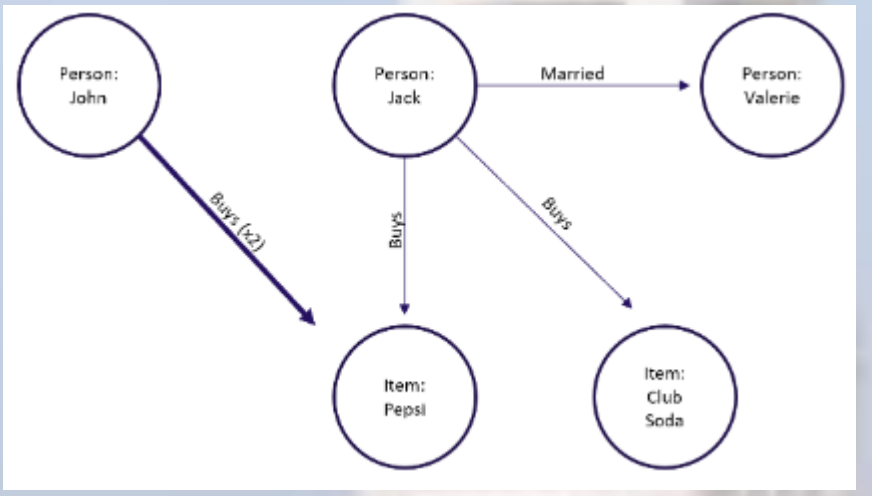
- High performance
- Flexible and efficient data model
- Scalability
- Distributed Systems
- Handling high write throughput

Disadvantages

- Limited querying capabilities
- Limited data modelling
- Limited support for advanced features
- Data Migration

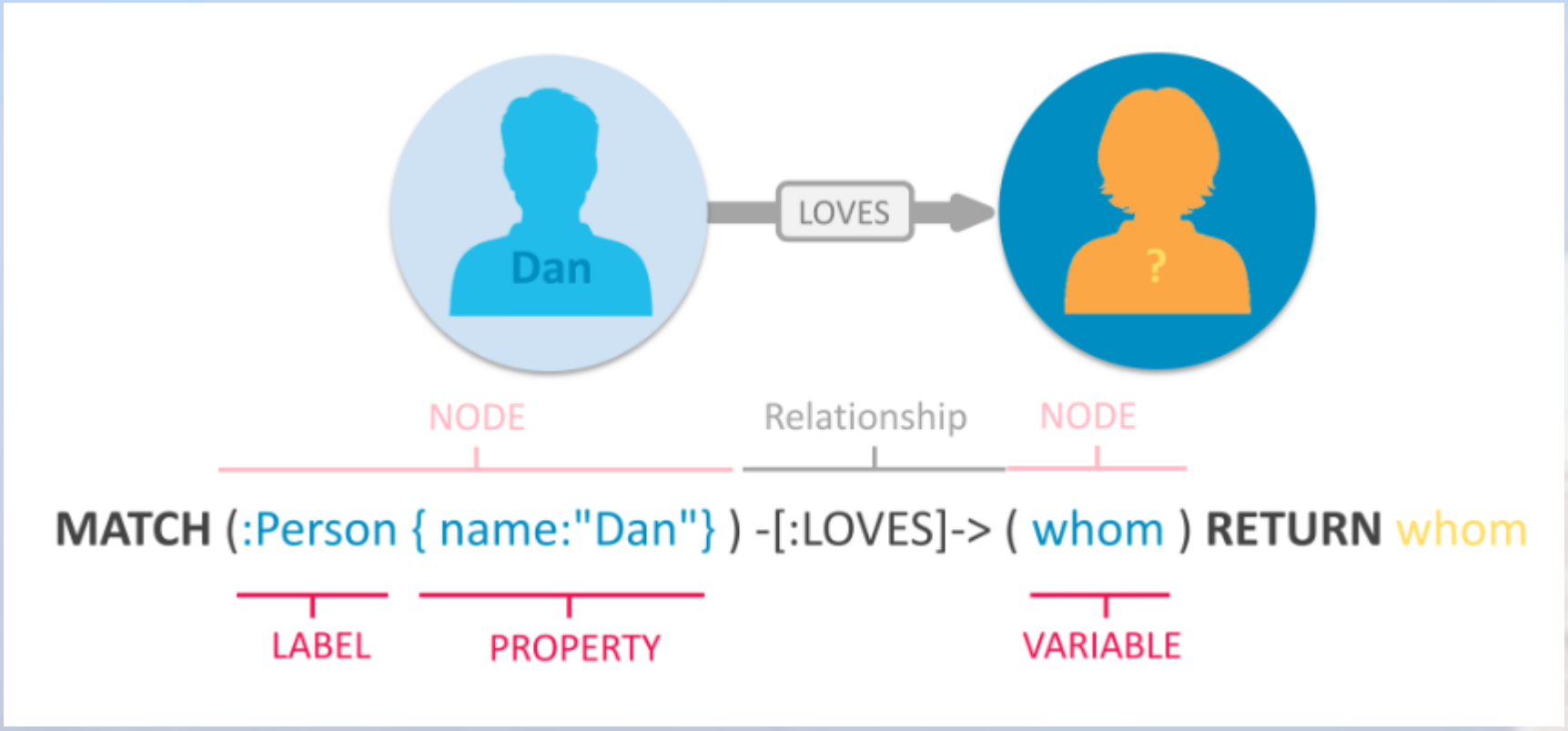
Graph-oriented databases

- Data in a graph-oriented database is stored in nodes with properties (named attribute values), and the connections between the nodes represent relationships between the real-world instances.
- The collections of attributes associated with each node may vary.
- Relationships may also have attributes associated with them.
- Software solutions: Neo4J, Oracle's Graph Database



The property graph model in Neo4J

In Neo4j, information is organized as nodes, relationships, and properties.



The property graph model in Neo4J

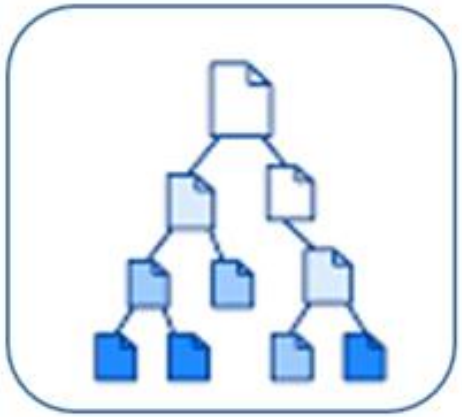
- ***Nodes*** are the entities in the graph.
  - Nodes can be tagged with labels, representing their different roles in your domain. (For example, Person).
  - Nodes can hold any number of key-value pairs, or properties. (For example, name)
  - Node labels may also attach metadata (such as index or constraint information) to certain nodes
- ***Relationships*** provide directed, named, connections between two node entities (e.g., Person LOVES Person).
  - Relationships always have a direction, a type, a start node, and an end node, and they can have properties, just like nodes.
  - Nodes can have any number or type of relationships without sacrificing performance.
  - Although relationships are always directed, they can be navigated efficiently in any direction.

Graph-oriented databases

- Graph databases allow:
  - Navigate deep hierarchies,
  - Find hidden connections between distant items, and
  - Discover inter-relationships between items.
- Use Cases
  - Fraud detection
  - Detection of money laundering
  - Bill of materials
  - Cybersecurity
  - Contact tracing
  - Product recommendations
  - AI (Feature Engineering, Neural Networks)

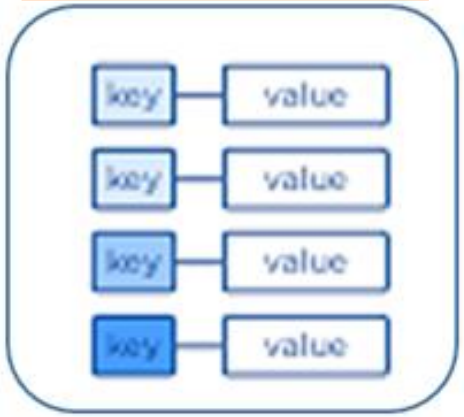
To summarize

MongoDB, Amazon DocumentDB



Document Store

REDIS, Amazon DynamoDB

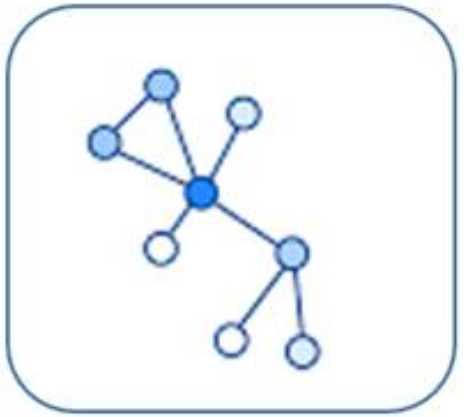


Key-Value Store

Apache Cassandra, BigTable, ScyllaDB

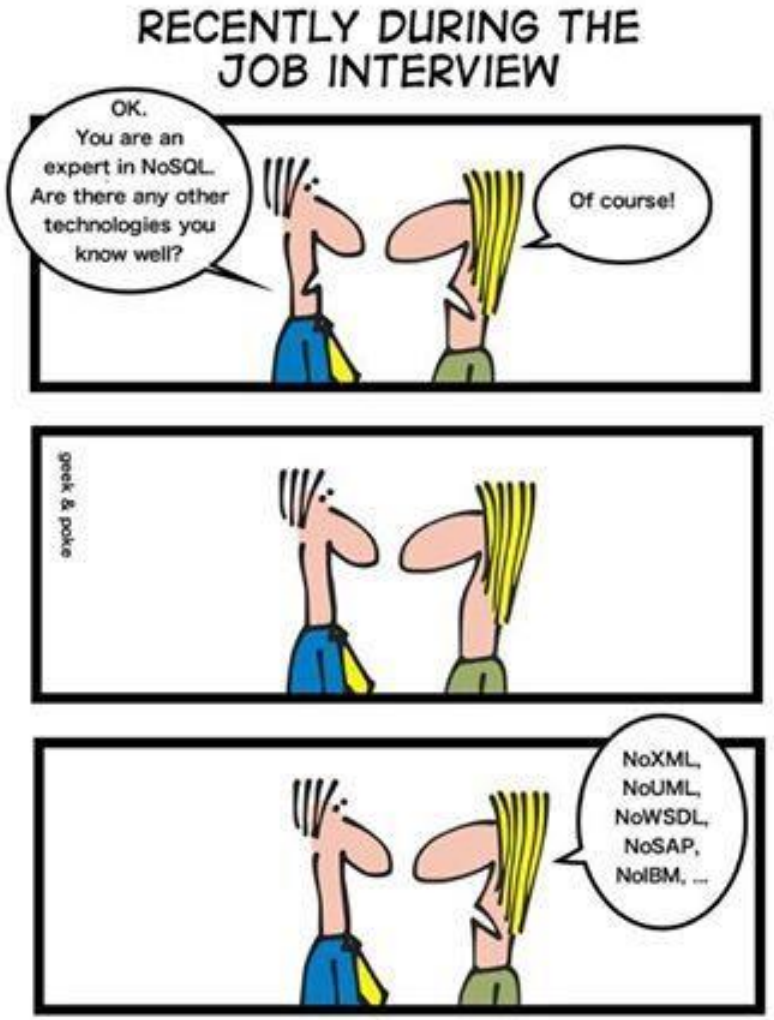
Wide-Column Store

Neo4J, Oracle's Graph Database



Graph Store

What not to claim in your job interview…



॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥
IIT Jodhpur

Excellence above all
SME

Technology • Innovation • Leadership

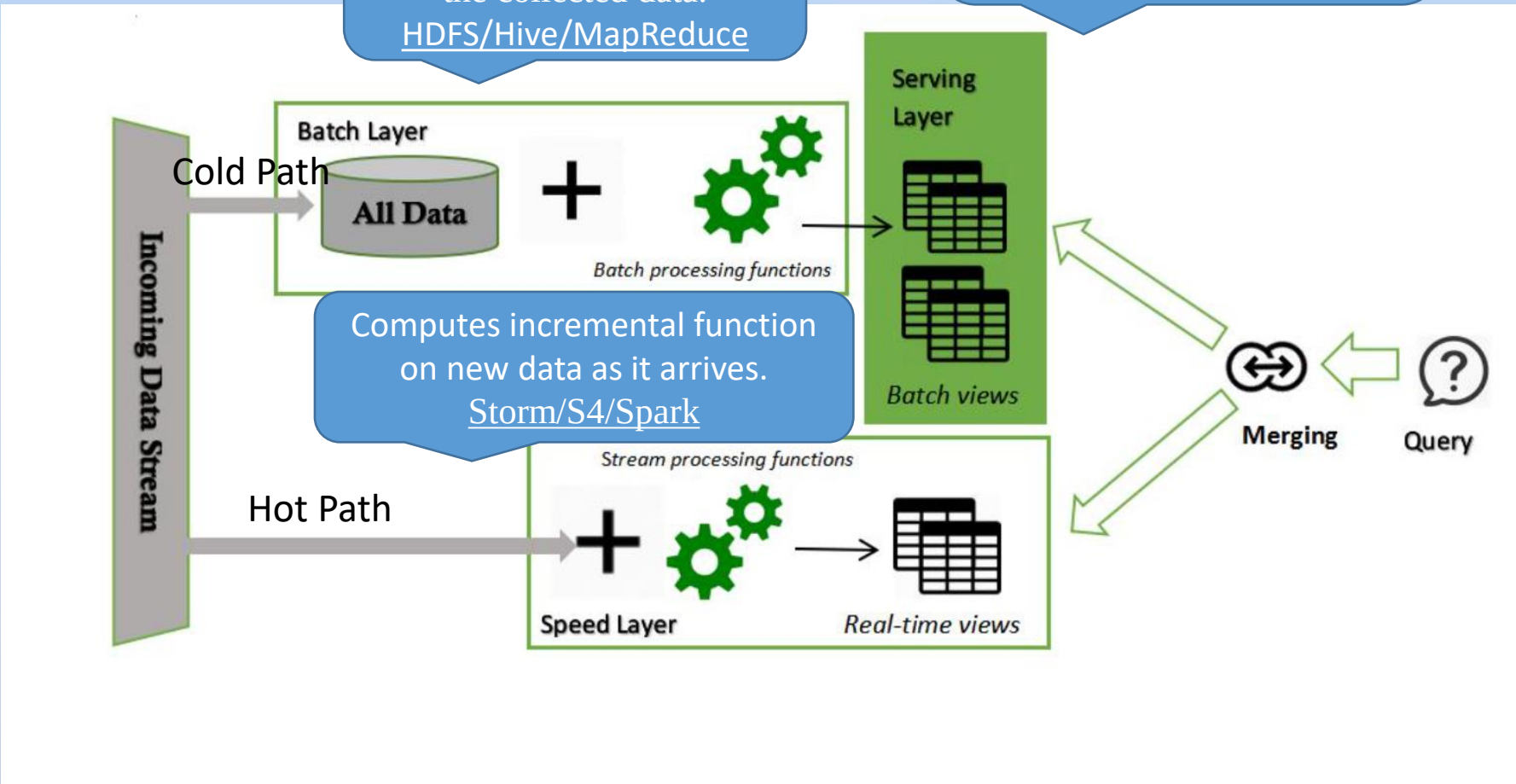
Big Data Ecosystem - Architecture

Dr. Deepak Saxena, SME IIT Jodhpur

Big Data Architecture

- Lambda
- Kappa
- Microservice
- Zeta
- IoT

Lambda Architecture



Lambda Architecture

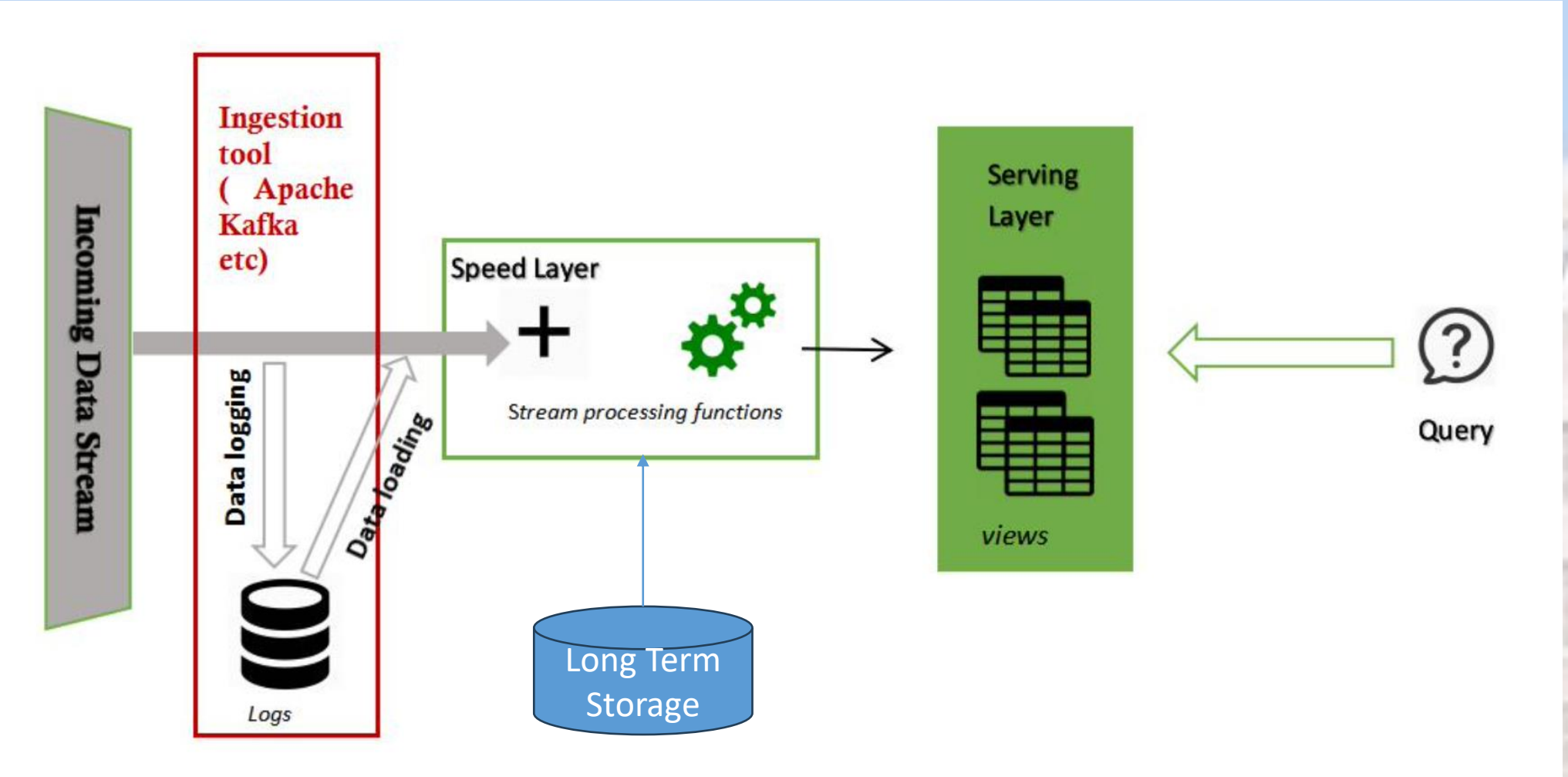
Advantages

- Better accuracy
- Higher throughput
- Lower latency
- Resilient
- Fault tolerant

Challenges

- Synchronization of batch and speed layer
- Need to maintain two separate codebase

Kappa Architecture



Kappa Architecture

Advantages

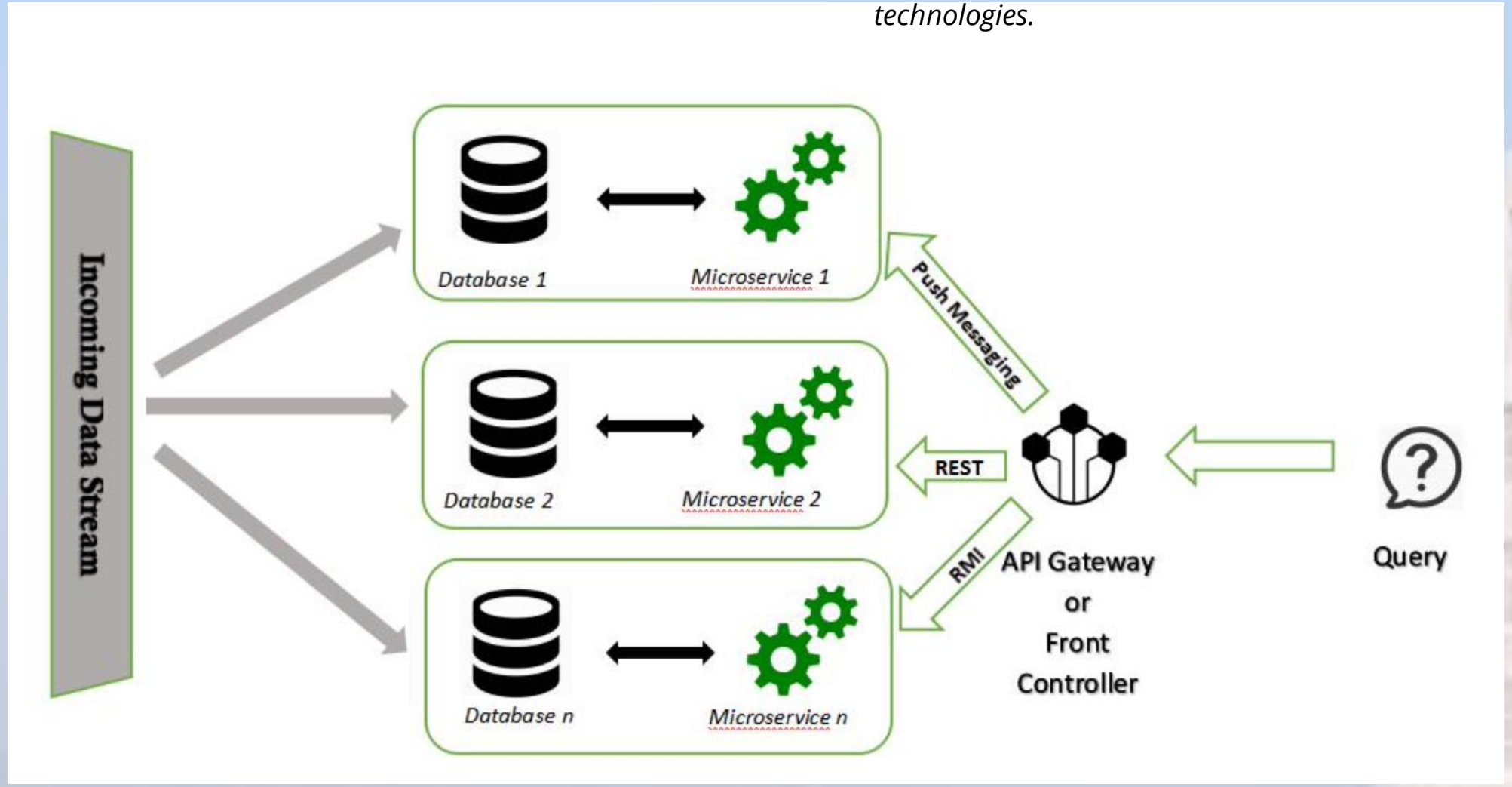
- Simplified architecture
- Single codebase

Challenges

- Only analytical operations are possible not transactional ones.
- It is important to know that the data is not conserved for a long term; data is kept for a limited predefined period after which it is discarded.

Microservice Architecture

The microservice architectural style is an approach to developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API. These services are built around business capabilities and independently deployable by fully automated deployment machinery. There is a bare minimum of centralized management of these services, which may be written in different programming languages and use different data storage technologies.



REST: Representational State Transfer

RMI: Remote Method Invocation

Microservice Architecture

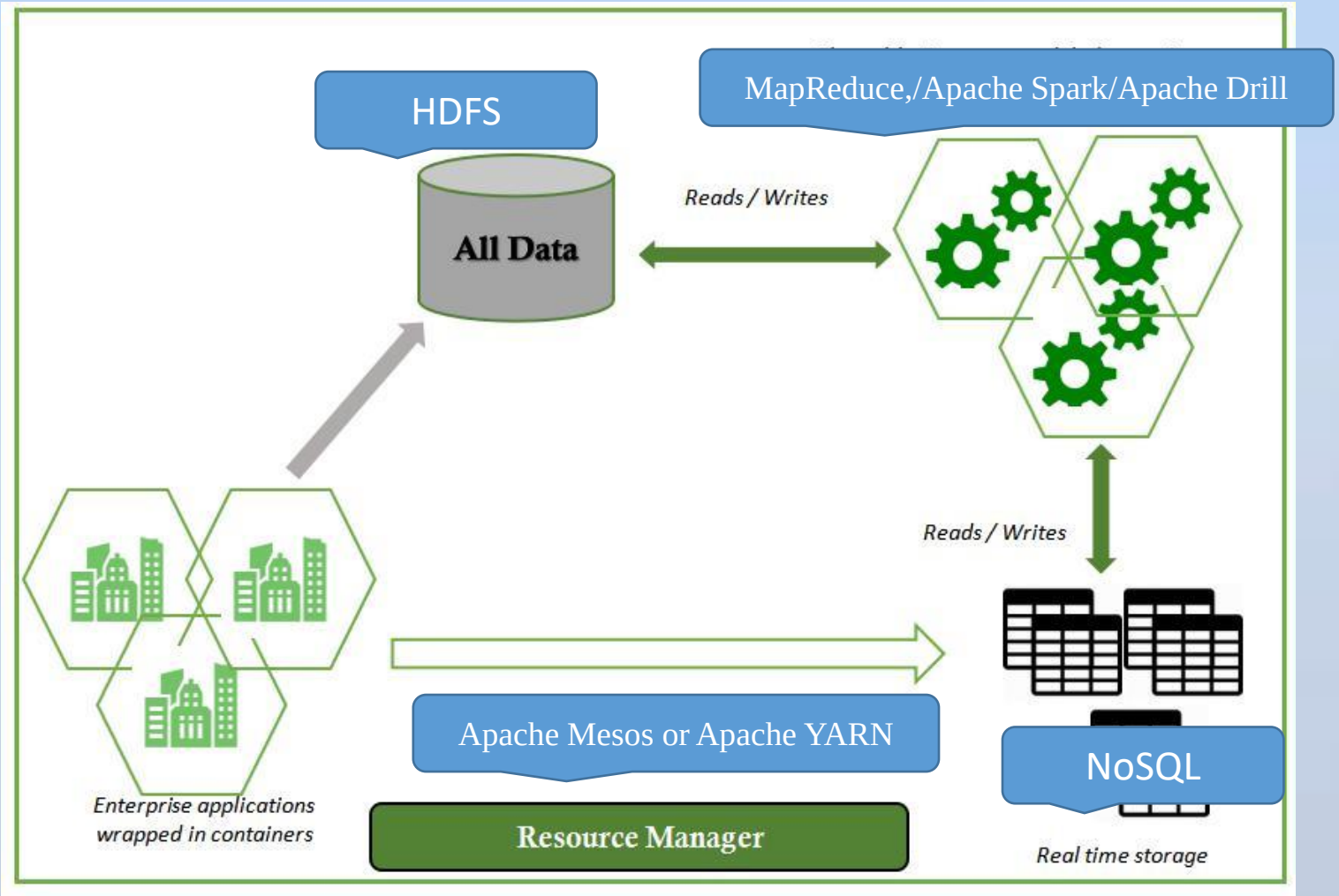
Advantages

- Faster development, testing, and deployment
- Fault tolerant
- Limited vendor/technology lock-in
- Easy onboarding and maintenance

Challenges

- An inter-service communication mechanism is required, and its development is quite complex.
- Though the deployments are faster, they are more complex to setup

Zeta Architecture



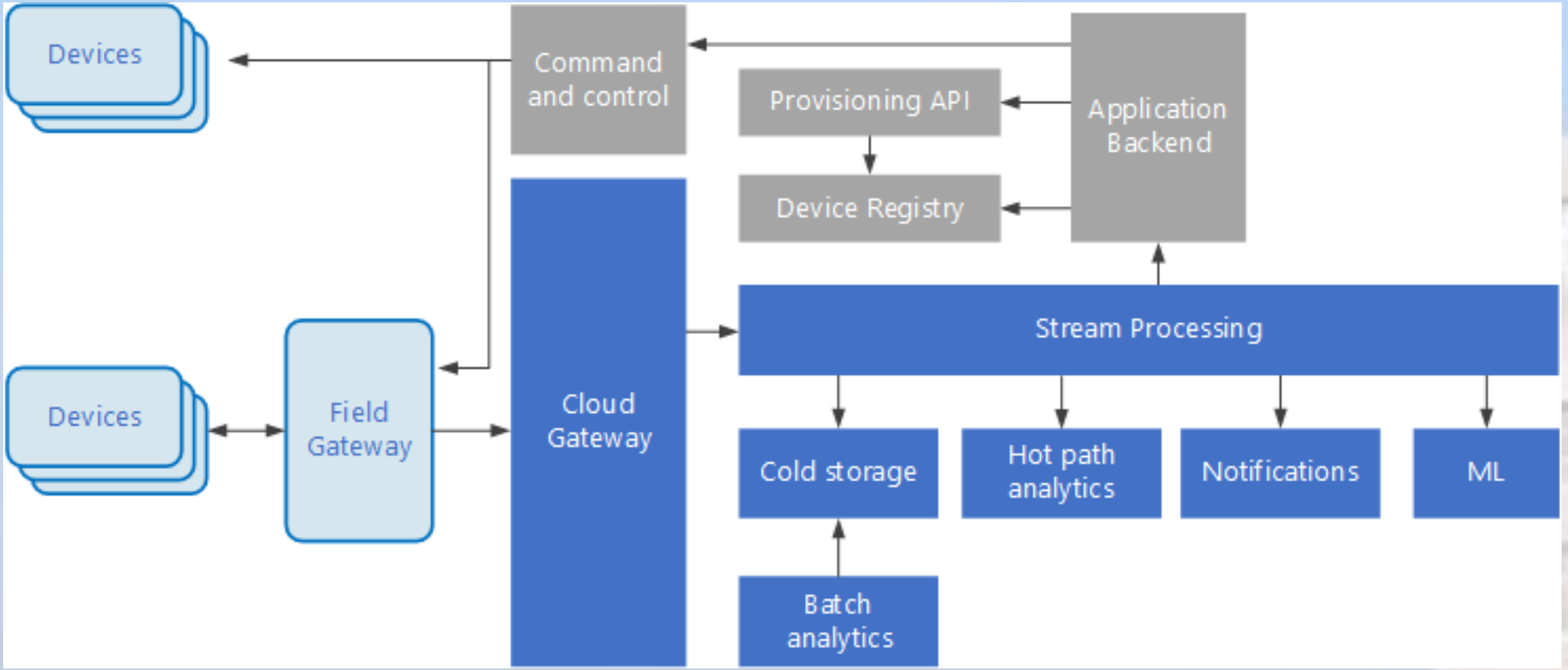
The goal of the container is to guarantee a single, standardized method of deployment. It also implies that deployed resources are isolated containers that don't concern about any environment changing, i.e. they deploy in the same manner in local environment as in prod environment. Thanks to this isolation, the containers can be freely moved between machines with the guarantee of repeatability (results on local server will be the same as on prod's one). A famous example of isolated containers is Docker, but can also be used Kubernetes or Mesos.

Zeta Architecture

Advantages

- Better utilization of hardware
- Near real-time backup

IoT Architecture



Architectures / Features	Lambda	Kappa	Iot-a	Microservices	Zeta
Analysis type	Batch/Real-time	Real-time	Batch / Real-time	Batch/ Real-time	Batch/ Real-time
Processing methodology	Query and reporting	Query and reporting	Query and reporting/ Analytical/ Predictive analysis	Query and reporting/ Analytical	Query and reporting
Data frequency	Real-time feeds	Continuous feeds	On-demand feeds	On-demand feeds	On- demand feeds
Data type	Master data	Transactional	Master data	Transactional data	Transactional data
Content format	Structured, Semi-structured & Unstructured	Structured, Semi-structured & Unstructured	Structured, Semi-structured & Unstructured	Structured, Semi-structured & Unstructured	Structured,Semi-structured & Unstructured
Data sources	Human & Machine generated, web or social media	Machine & Human generated, Web or social media	Machine generated	Internal data sources, machine generated	Web and social media, Internal Data sources
Data consumers	Human	Human	Human/ Other data repositories	Business process	Enterprise applications

Big Data Project Implementation

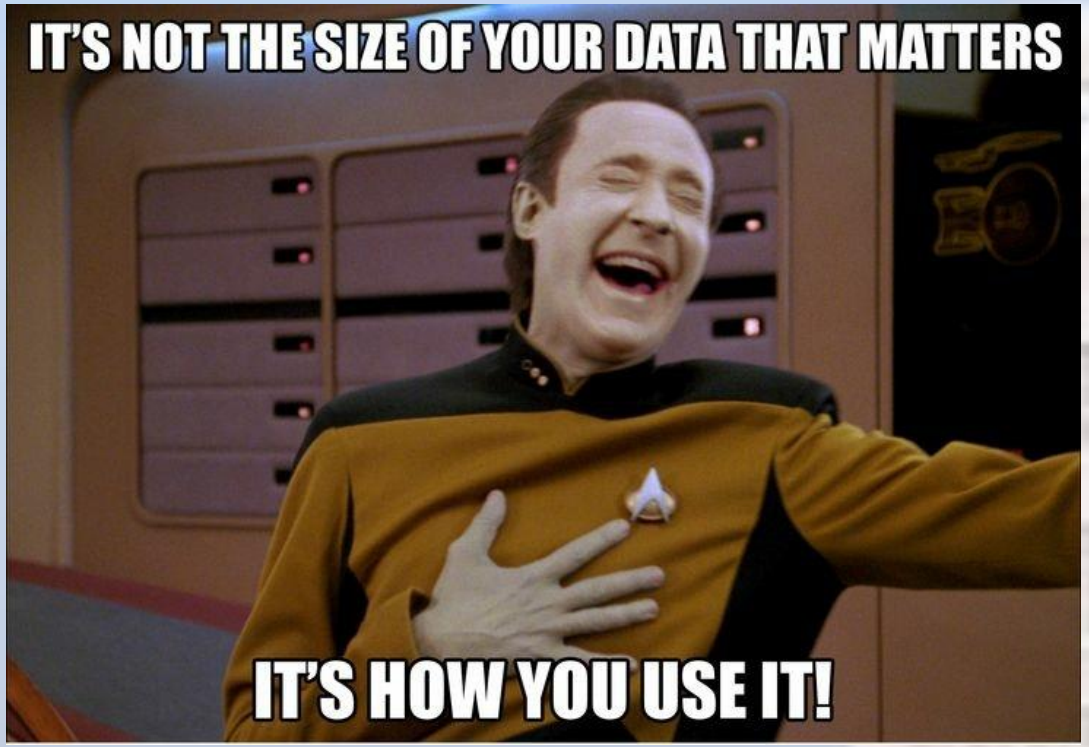
Dr. Deepak Saxena, SME IIT Jodhpur

Key Aspects

- Strategic
- Business
- Analytics
- Technical

Strategic Aspects

- Strategic vision dictate the purpose and principles that underpin how you use data.
  - Customer intimacy
  - Product leadership
  - Operational excellence

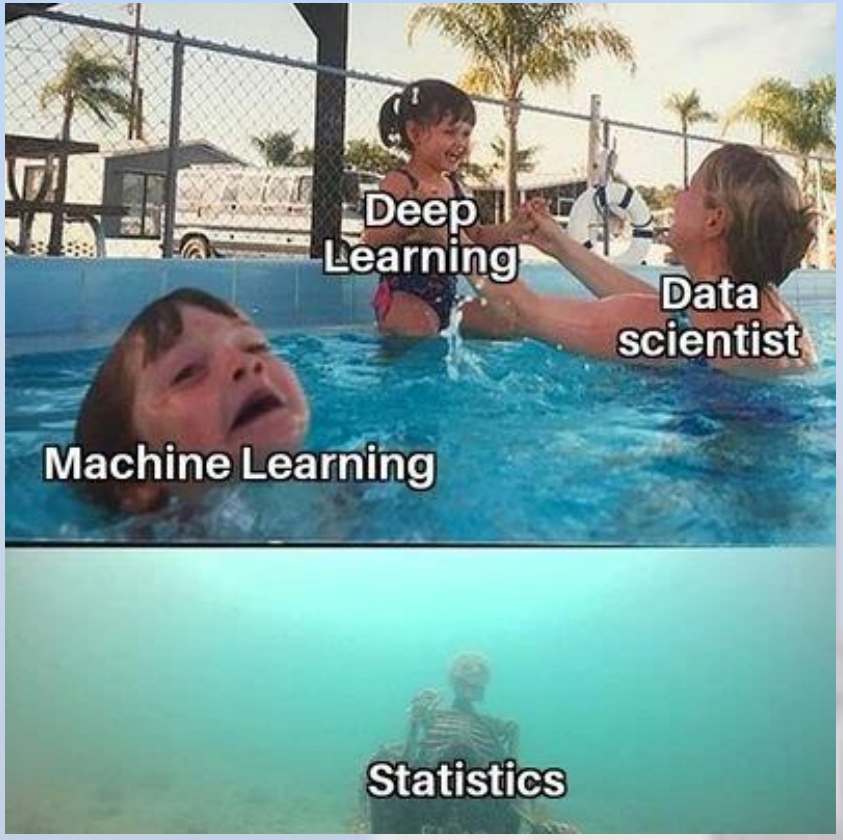


Business Aspect

- Harvest the intuition of colleagues with the deepest understanding of the customer, the product and the market.
- Get them involved at the start of your analytics programme and keep them closely involved at each stage of development.
- Business Involvement
  - To explain the basics to the analytics team
  - To give initial insights/hunches
  - To provide continuous feedback during data analysis and modelling

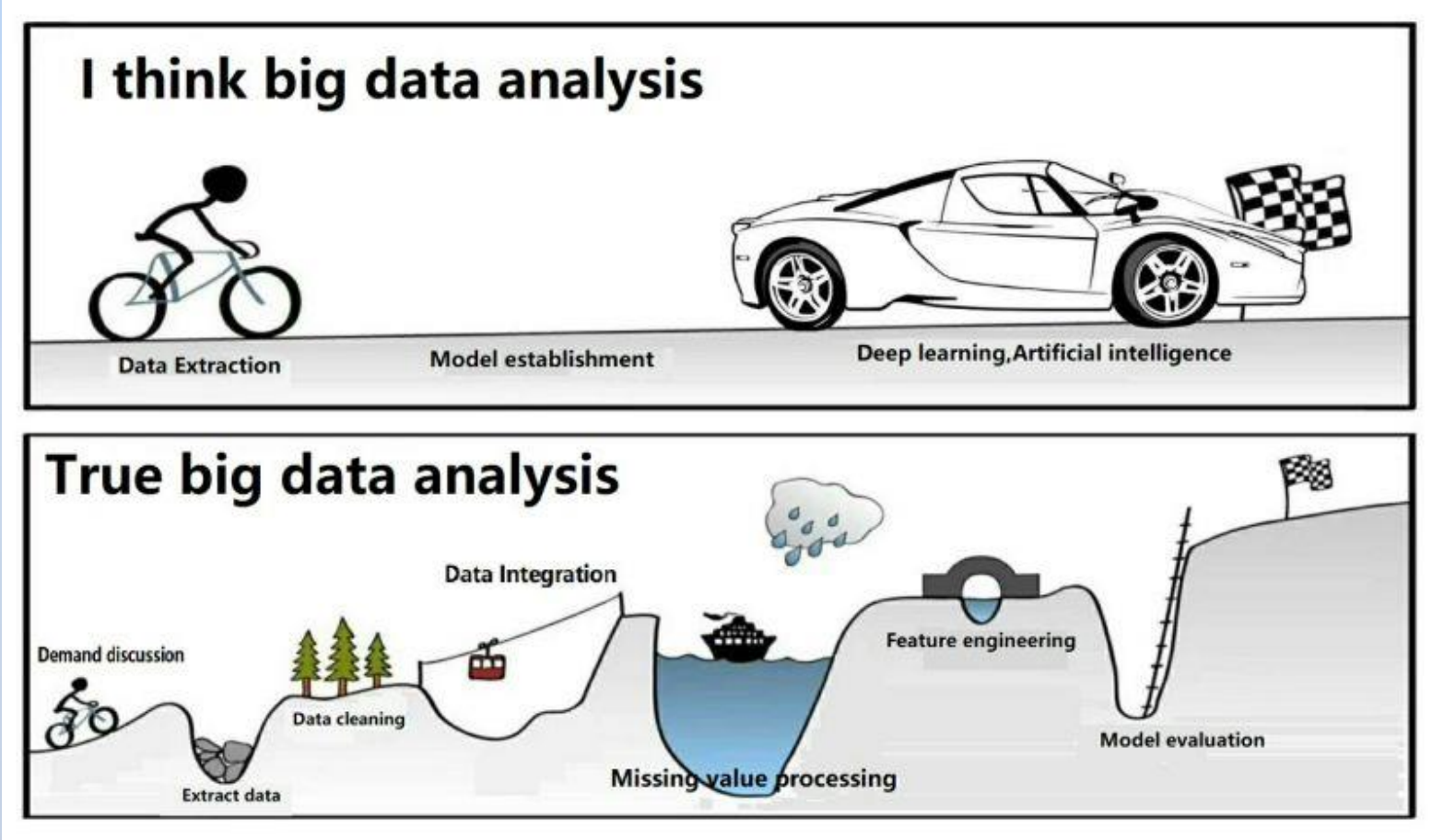
Analytics Aspects

- Bring in someone with a strong background in developing and deploying analytic models.
- Identify high-potential analytic applications
- Choose from a broad selection of analytic methods and models
- Analytical Tools to be used
  - Execution speed;
  - Ease of development
  - The ability of the language to easily interface with relevant technologies
  - The breadth of the user / support base
- Build vs Buy / Open-Source vs Proprietary



Technical Aspects

- The extent and accuracy of the various data stores within the organization.
- Determining new Tables
- Computational Infrastructure / Cloud Service Configuration
- Acceptable choice of development language, frameworks and operating system;
- Requirements for version control and documentation; and
- Requirements and resources for testing (QA) and deployment to production.



Strategic Input

- Review the strategic goals of the company
- Distinguishing between the long- and short-term strategic goals.

Business Input

- Review the KPIs used within the organization
- Identify pain points within the organization.

Analytics Input

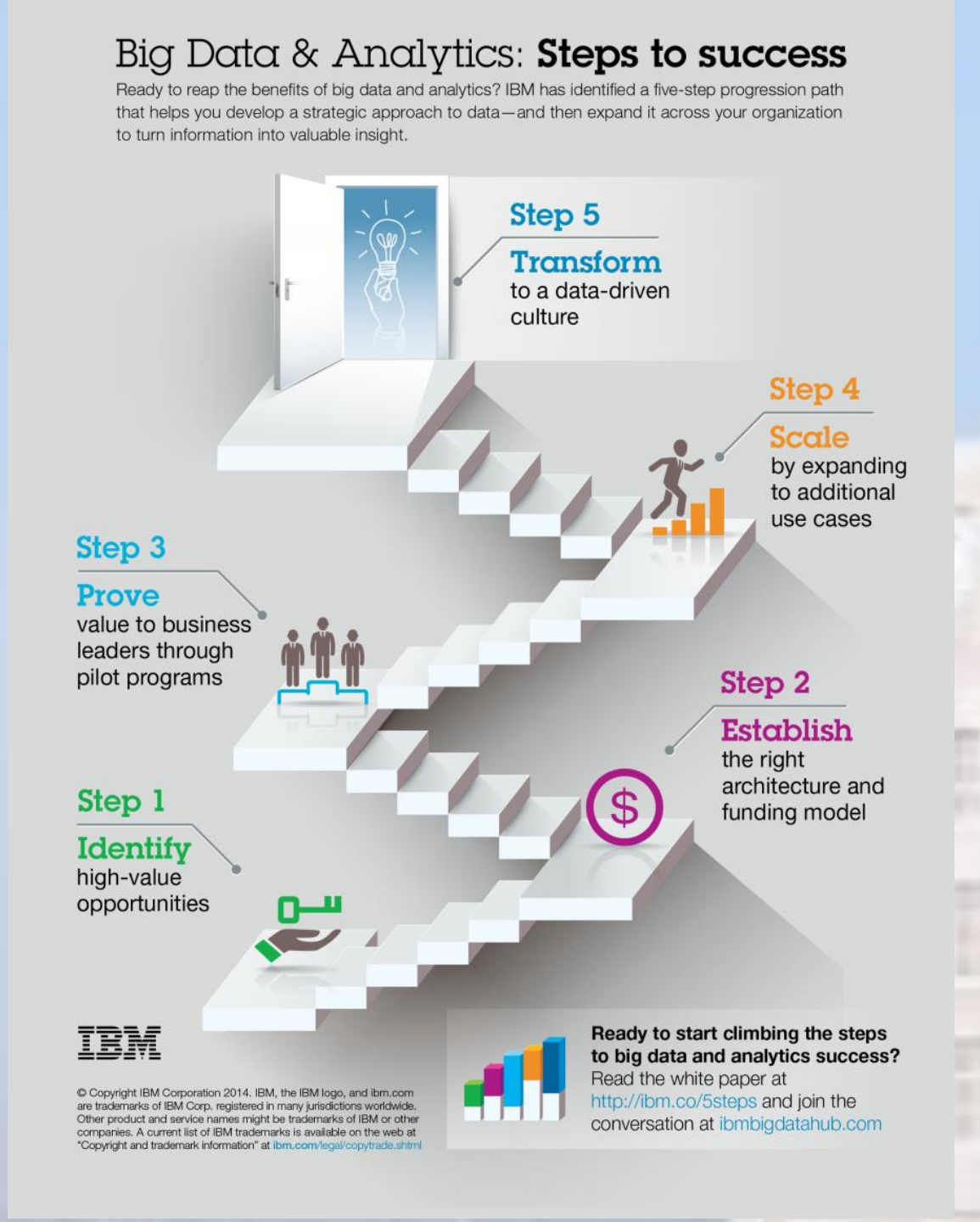
Identify which of those business objectives can be matched to standard analytic tools or models that may bring business value in:

- relieving a pain point,
- raising a KPI, or
- providing an innovative improvement

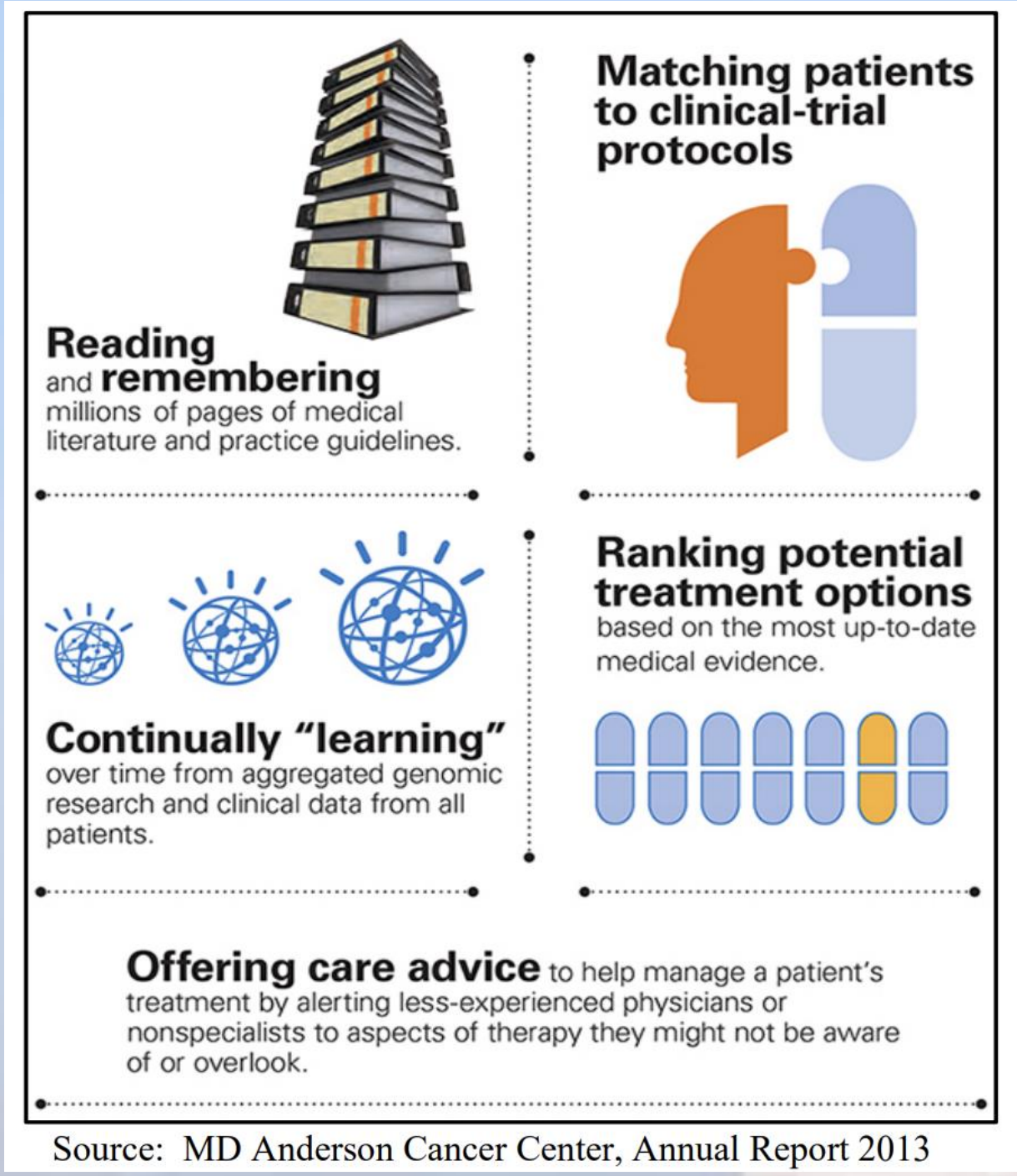
Technical Input

- Highlighting technical limitations and opportunities
- Providing the initial data input
- Taking responsibility for eventual deployment of analytics solutions

What IBM Says



IBM Watson



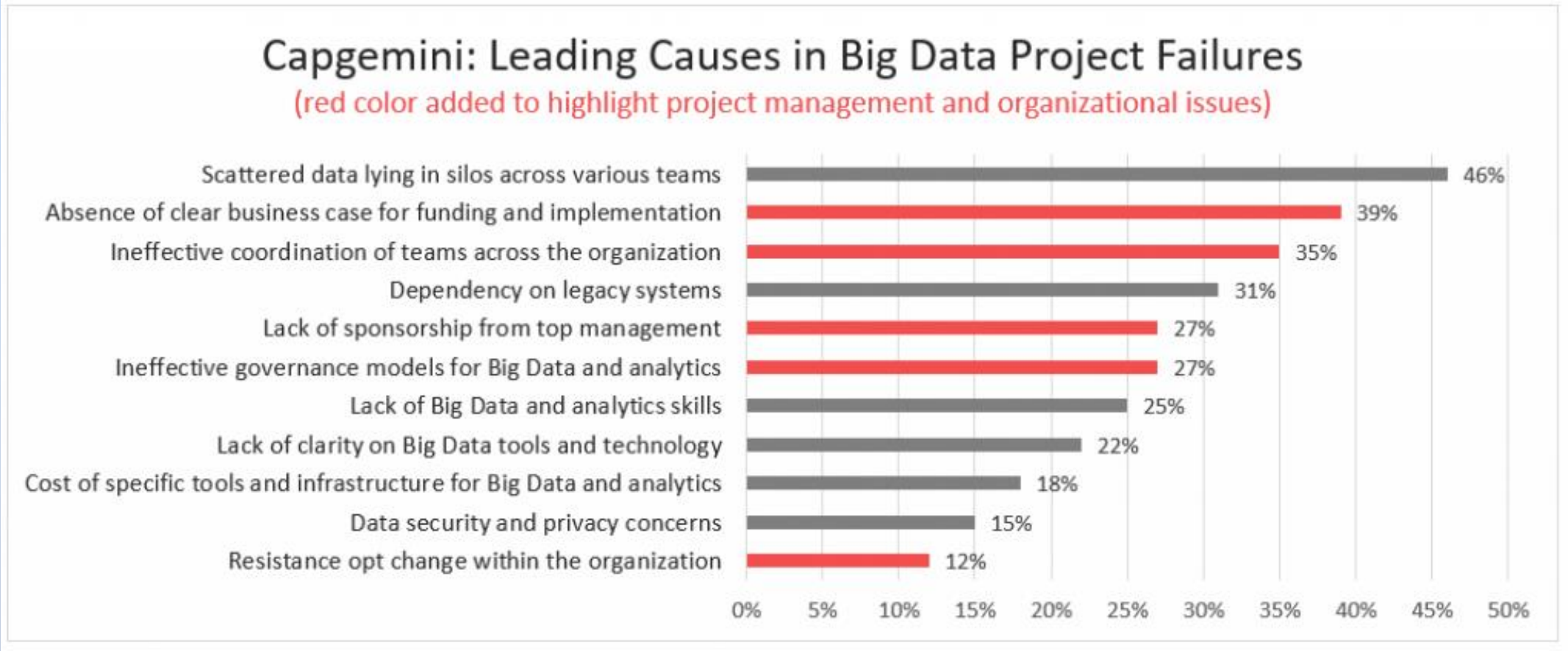
Source: MD Anderson Cancer Center, Annual Report 2013

However, IBM Watson Failed because it…

- Did not use proper contracting and procurement procedures
- Failed to follow IT Governance processes for project approval
- Did not effectively monitor vendor contract delivery
- During the collaboration, the hospital switched to a new electronic health record system and Watson could not tap patient data.
- Watson struggled to decipher doctors' notes and patient histories.

More interesting details on:
https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html

This is not a one-off case



Some best practices

- Become Data Driven
  - Keep asking questions about your business
  - Create and monitor KPIs
  - Get new ideas from other sectors
  - Organise your data
- Get the right people onboard
  - Data scientists
  - Data-driven businesspeople
  - Try new vendors
- Break down silos
- Focus on business value
- Measure results
- Stay Agile

Or Maybe…


DATA
TimoElliott.com
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥
IIT Jodhpur

Excellence above all
SME

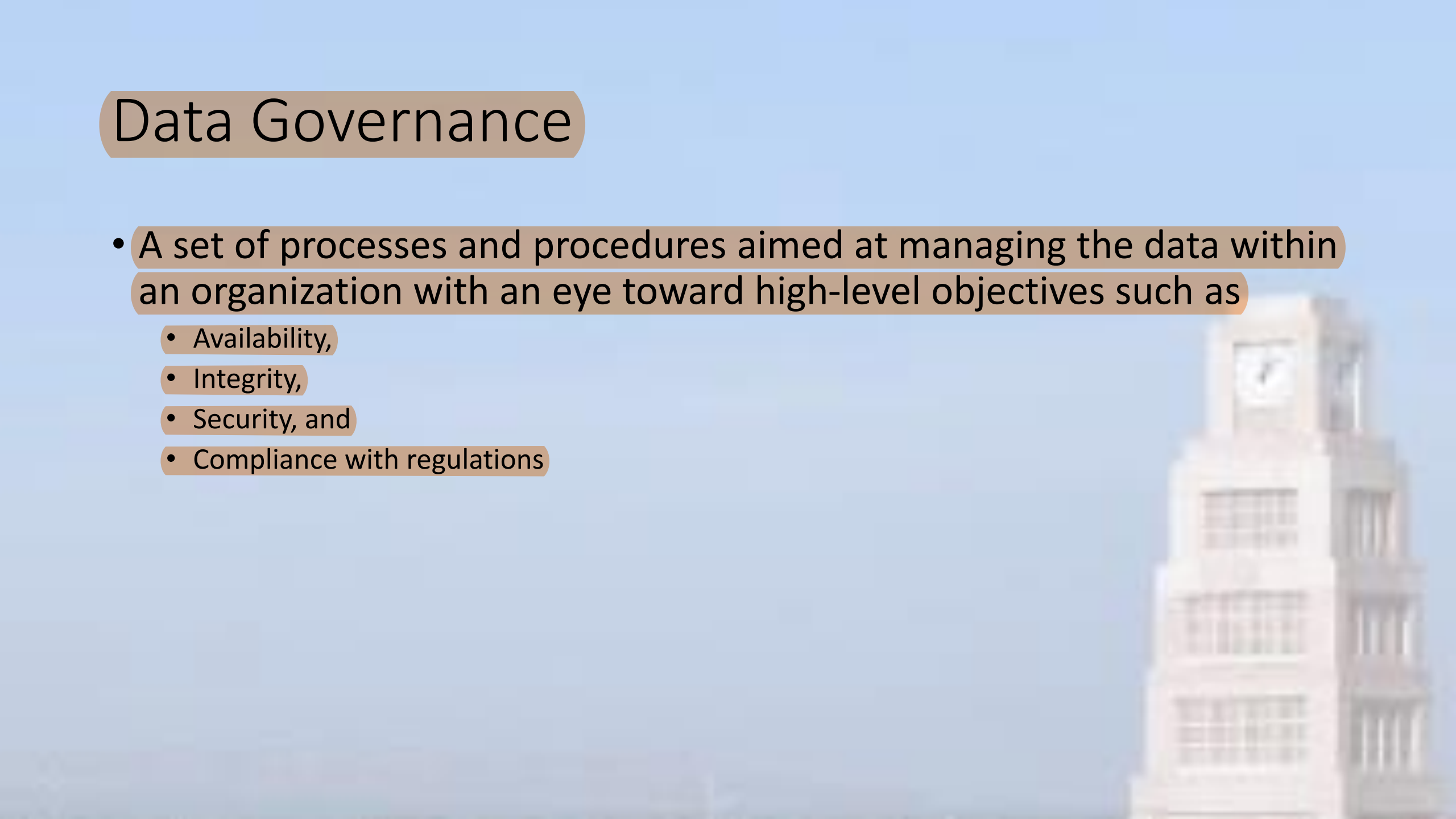
Technology • Innovation • Leadership

Data Governance and Legal Compliance

Dr. Deepak Saxena, SME IIT Jodhpur

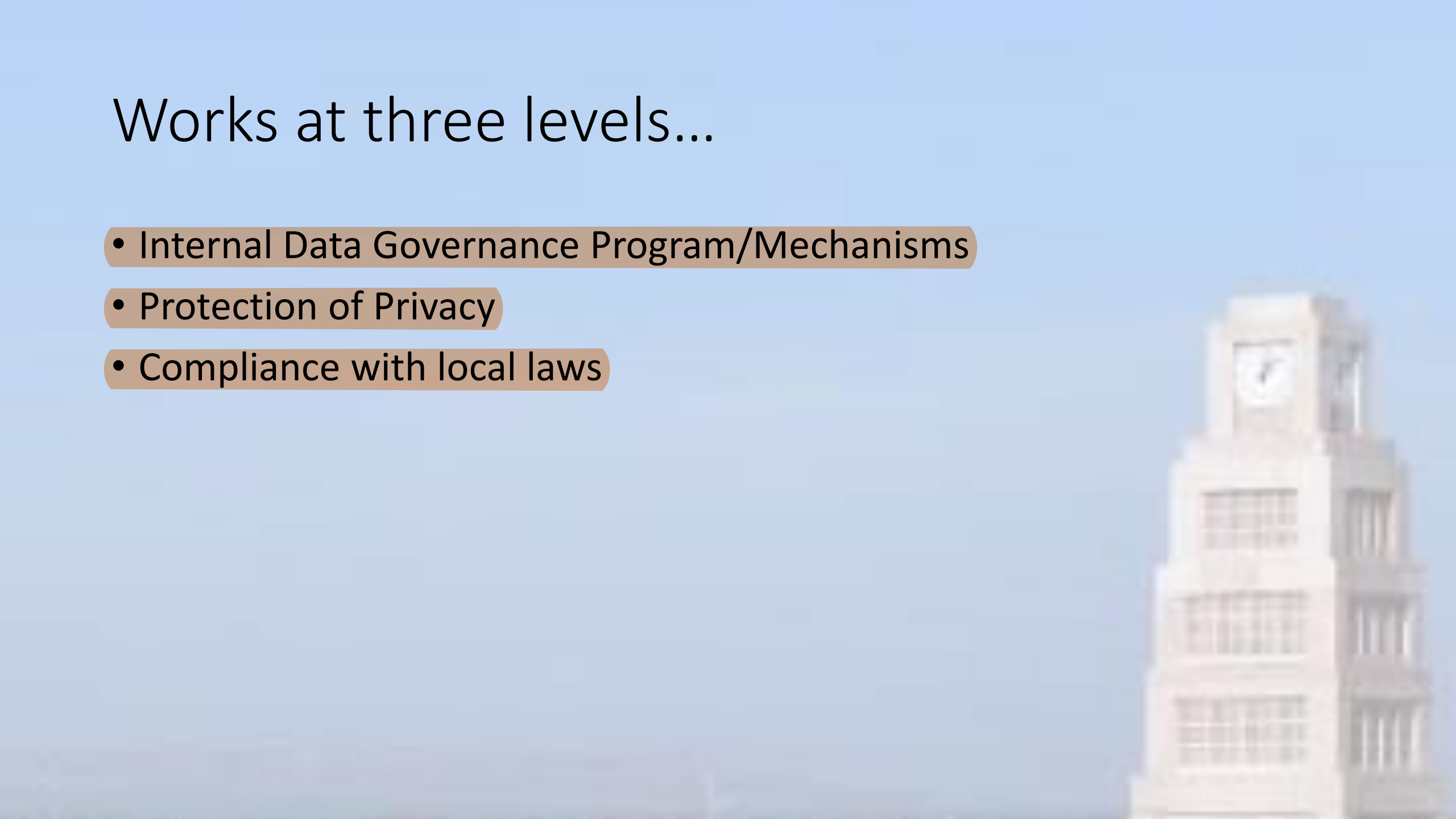
Data Governance

- A set of processes and procedures aimed at managing the data within an organization with an eye toward high-level objectives such as
  - Availability,
  - Integrity,
  - Security, and
  - Compliance with regulations



Works at three levels…

- Internal Data Governance Program/Mechanisms
- Protection of Privacy
- Compliance with local laws



Data Governance

Who's part of the program on data governance?

DATA GOVERNANCE TEAM

A data governance manager heads a program office that may also include data architects and governance specialists.

DATA GOVERNANCE COUNCIL OR COMMITTEE

Typically made up of executives from all business units, it sets data policies and standards and resolves issues.

DATA STEWARDS

Stewards oversee data sets and are in charge of implementing governance policies and monitoring compliance with them.

CHIEF DATA OFFICER

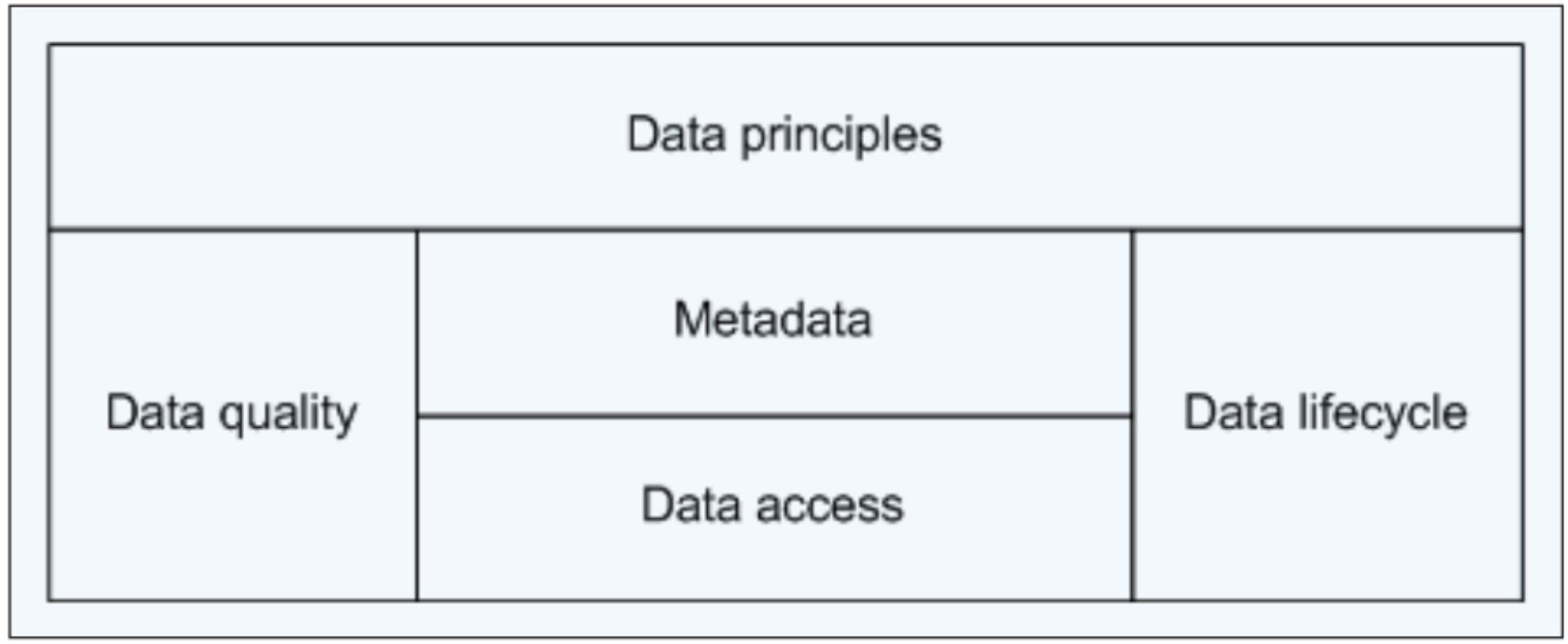
CDOs often have overall responsibility and accountability for their organization's data governance program.

DATA QUALITY ANALYSTS AND ENGINEERS

They work with the governance team and data stewards to fix data errors and track data quality metrics.

ILLUSTRATION: TARTILA/ADOBE STOCK

Decision Domains for Data Governance



Source: Khatri, V., & Brown, C. V. (2010). Designing data governance. Communications of the ACM, 53(1), 148-152.

Data Governance domains	Domain decisions
Data Principles *Clarifying the role of data as an asset*	• What are the uses of data for the business? • What are the mechanisms for communicating business uses of data on an ongoing basis? • What are the desirable behaviors for employing data as assets? • How are opportunities for sharing and reuse of data identified? • How does the regulatory environment influence the business uses of data?
Data Quality *Establishing the requirements of intended use of data*	• What are the standards for data quality with respect to accuracy, timeliness, completeness and credibility? • What is the program for establishing and communicating data quality? • How will data quality as well as the associated program be evaluated?
Metadata *Establishing the semantics or "content" of data so that it is interpretable by the users*	• What is the program for documenting the semantics of data? • How will data be consistently defined and modeled so that it is interpretable? • What is the plan to keep different types of metadata up-to-date?

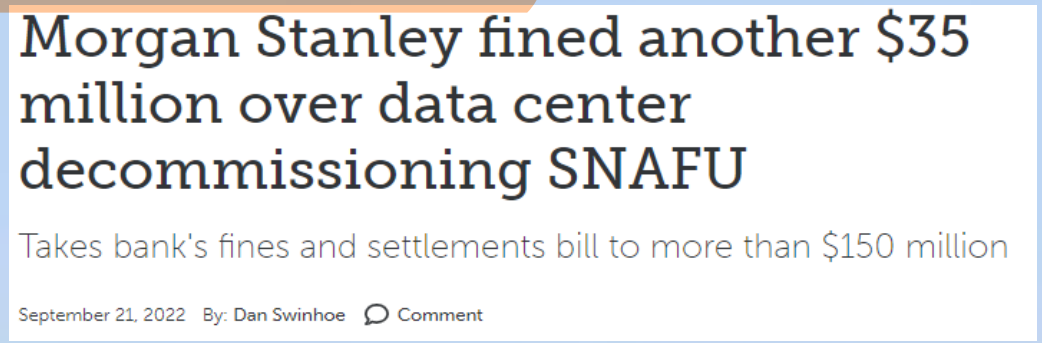
Source: Khatri, V., & Brown, C. V. (2010). Designing data governance. Communications of the ACM, 53(1), 148-152.

Data Governance domains	Domain decisions
Data Access *Specifying access requirements of data*	• What is the business value of data? • How will risk assessment be conducted on an ongoing basis? • How will assessment results be integrated with the overall compliance monitoring efforts? • What are data access standards and procedures? • What is the program for periodic monitoring and audit for compliance? • How is security awareness and education disseminated? • What is the program for backup and recovery?
Data Lifecycle *Determining the definition, production, retention and retirement of data*	• How is data inventoried? • What is the program for data definition, production, retention, and retirement for different types of data? • How do the compliance issues related to legislation affect data retention and archiving?

Source: Khatri, V., & Brown, C. V. (2010). Designing data governance. Communications of the ACM, 53(1), 148-152.

Data Privacy and Protection

- **Personally-identifiable information (PII)** is data that is unique to an individual
- **Data privacy** relates to what data you may collect, store and use, such as whether it is appropriate to use web cookies to track online browsing without user consent.
- **Data protection** relates to the safeguarding and redistribution of data you have legally collected and stored. It addresses questions such as whether you can store private data of residents in data centres outside of the country.



Morgan Stanley fined another $35 million over data center decommissioning SNAFU

Takes bank's fines and settlements bill to more than $150 million

September 21, 2022 By: Dan Swinhoe Comment

Google and Facebook fined $240 million for making cookies hard to refuse

Posted: January 7, 2022 by Pieter Arntz

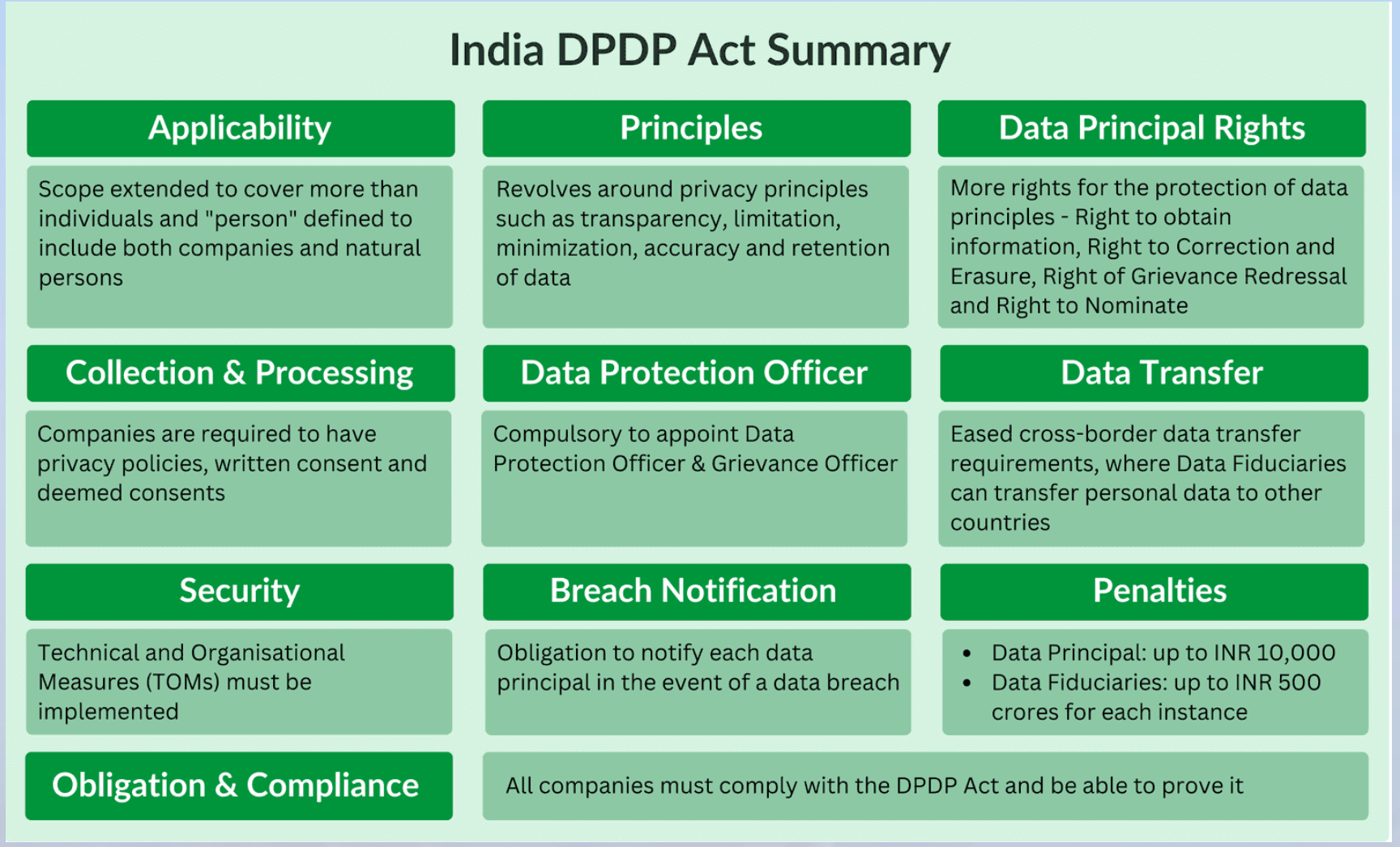
PIXELS • SOCIAL MEDIA

Irish data watchdog fines Instagram €405 million over children's privacy

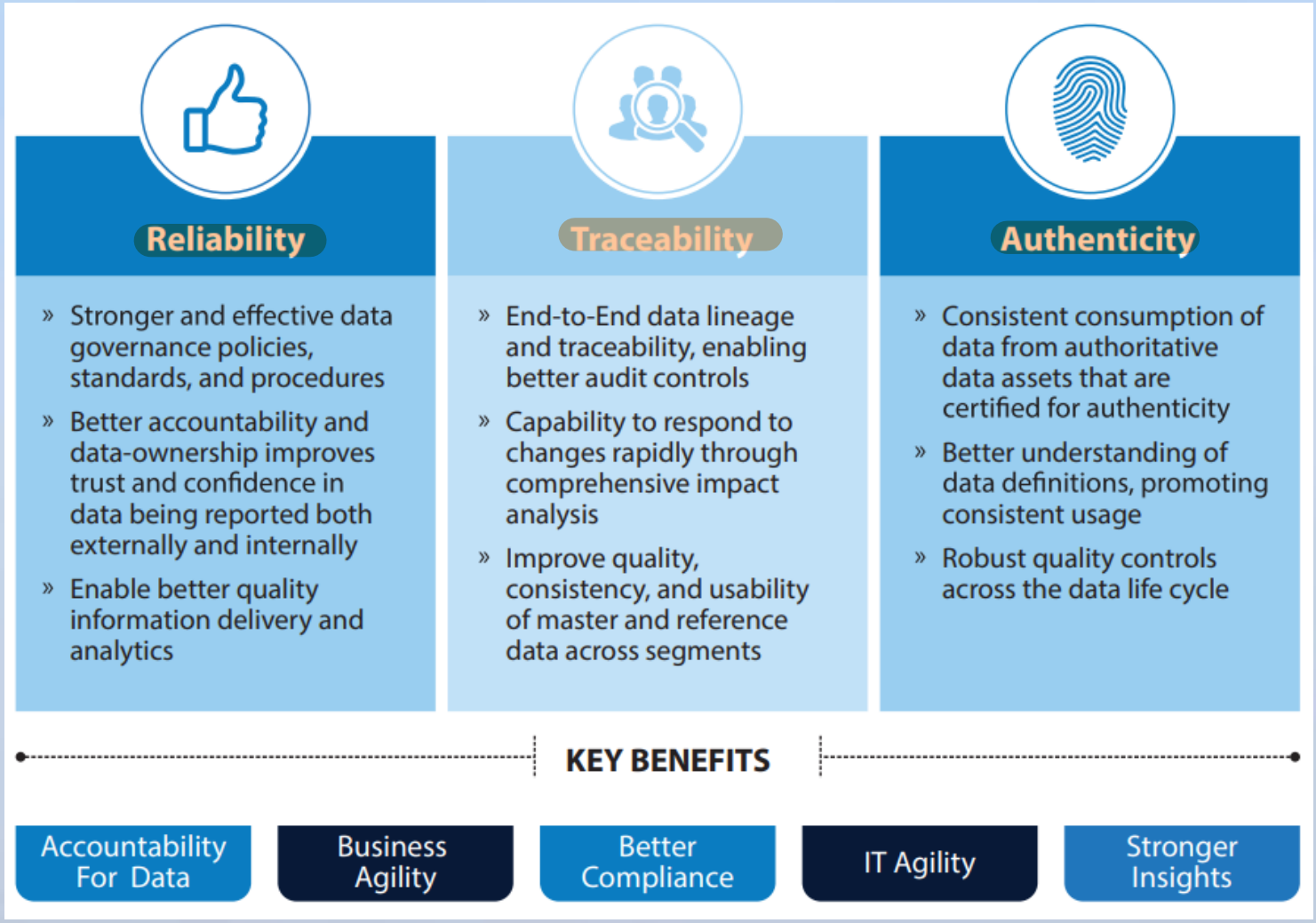
Ireland's Data Protection Commission on Monday said it had fined Instagram a record €405 million for breaching regulations on the handling of children's data.

Compliance with Local Laws – GDPR

Compliance with Local Laws - DPDP



Benefits of Seamless Data Governance



Source: Infosys