Big Data

Introduction

Vast amount of data is being generated

- Society has become more networked and online, generating and sharing data
- Many devices, small, large, embedded have become networked and generating data (sensors, mobile devices, home, business appliances, cars etc.)

Importance of Big Data

Analyzing Big Data is seen as the source of creation of big **competitive advantage** in almost all areas

- National development
- Scientific research
- Understanding today, predicting future

Big Data

Definition

- Different perceptions about Big Data and different definitions
- About storing, processing, analyzing big collections of data
- Data from various sources with formatted or unformatted
- Interdisciplinary work, computer science, statistics, mathematics etc.

Definition (Author's Adopted)

Big Data refers to the massive volume of data that is difficult to collect, store, manage, analyze with traditional technologies

Big Data

Characteristics - Dimensions of Big Data

Five V's of Big Data

1. Volume
2. Variety
3. Velocity
4. Veracity
5. Value

Big Data

Characteristics - Dimensions of Big Data

Volume: Magnitude of data to store, to process

- Expected to be difficult to manage with current infrastructure
- Depends on the scalability of existing infrastructure

Velocity: Generation rate of the data

- Data may be received in batch, near or real time
- The rate of data may change overtime.
- Sometime rate of change regarded as another dimension as Variability [1, p. 139]

Big Data

Characteristics - Dimensions of Big Data

Variety: Data comes from variety of sources

- Sources may be sensors, mobile devices, online transactions, social networks, social media etc.

Variety: Data may be in various formats

Structured: Data which has well defined schema as in RDBMS. Structured data constitutes only 5%[1] of the Big Data.

Unstructured: Data whose structure is not known or the format is not clear or unknown

Hybrid: Data may have some forms of model or schema however data partially conforms, generally graph based or hierarchical.

Mixed: Data mixed with structured or unstructured parts

Big Data

Characteristics - Dimensions of Big Data

Veracity: Reliability of the data

- Data may be received from unknown or unverifiable sources.
- Origin, thrustwordiness, accountability, fidelity of data
- Very important since the decisions will be based on the information extracted from data

Veracity and Variety far more challenging[2]

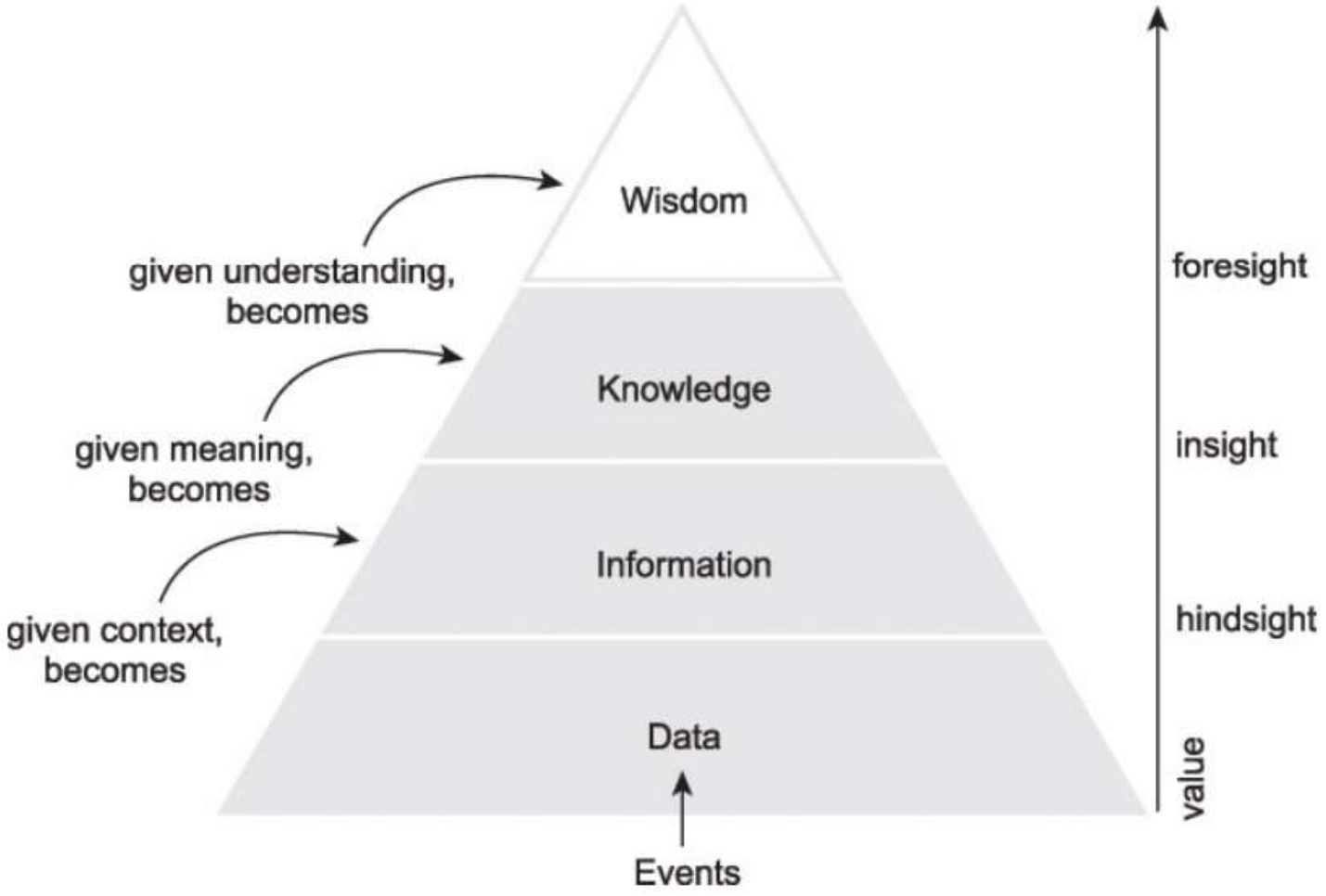
- Lesser studied areas.
- Very difficult to establish benchmark to compare against.

Value: the gain received by analyzing Big Data

- Primary reason for Big Data concept
- Solving a certain problem, support for a decision
- Gaining a competitive advantage etc.
- Requires human expertise, tooling, infrastructure, process
- Relevant to institution's ability to extract the value

Big Data

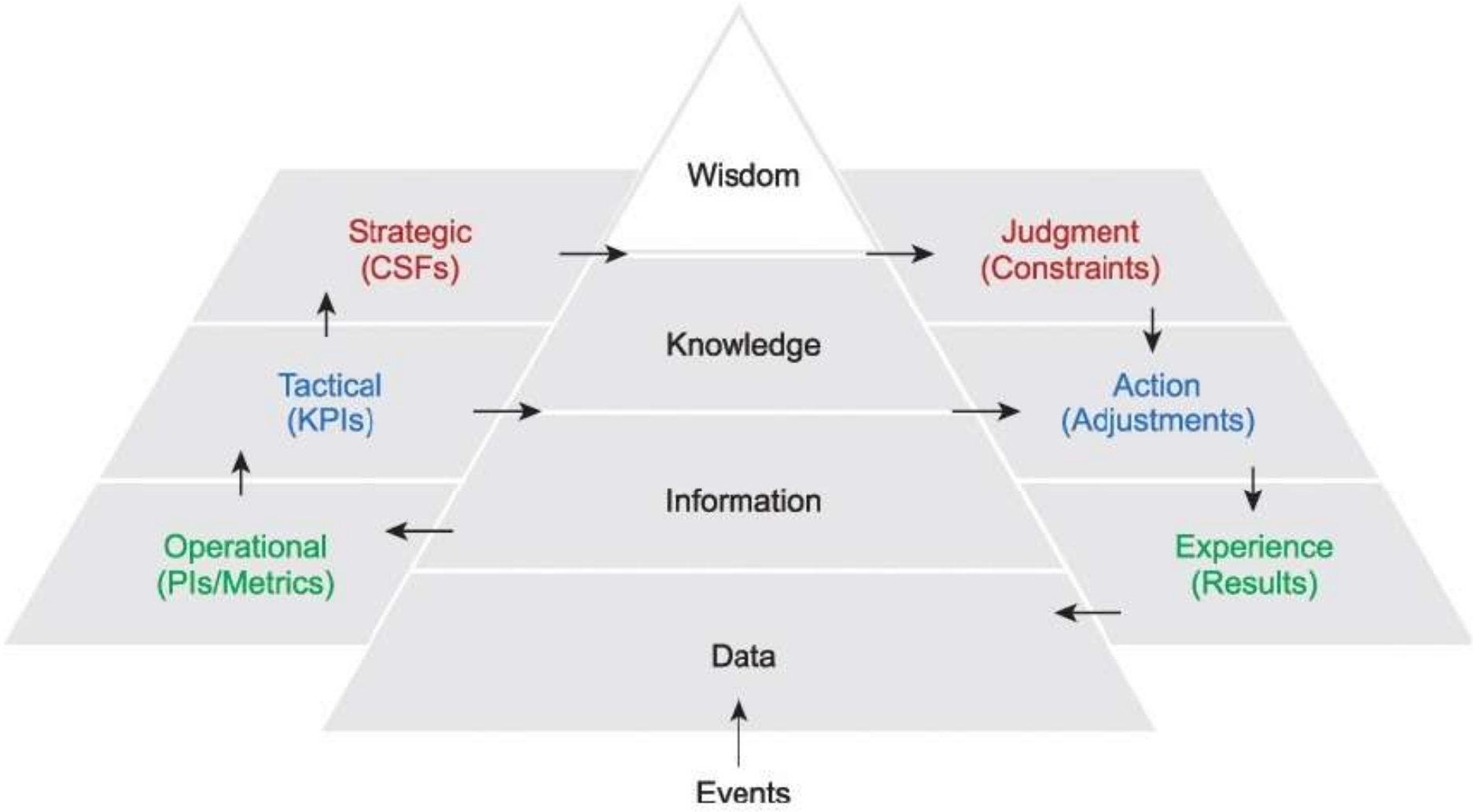
Characteristics - Dimensions of Big Data



DIKW Pyramid[3]

Big Data

Characteristics - Dimensions of Big Data



Virtuous Cycle[3]

Big Data in Cloud

Cloud provides three important assets for Big Data[3]

1. External data sets
2. Scalable computing or processing capabilities
3. Scalable storage

- Used in the same context and ***cojoined***[4].
- Cloud computing provides ***the necessary hardware resources and the infrastructure*** to tackle the some of the challenges of Big Data.

Big Data Analytics

Data analysis vs. analytics

Data Analysis: finding patterns, relations, trends or hidden values in the data.

Data Analytics: general term which includes analysis. Includes collecting, storing, organizing, cleaning etc. in addition to analysis. Data analytic describes overall processes.

Big Data Analytic Categories

According to Purpose

Descriptive Analytic

Tries to understand or evaluate the past

- Uses historical data to model past behavior and patterns.
- Uses well-defined structured data
- Relatively easy to perform
- Does not require high skill set
- Contribution to institutions is not generally much
- Reports created for management automatically

Big Data Analytic Categories

According to Purpose

Predictive Analytic

Tries to predict the future

- Uses current data in addition to historical.
- Tries to answer what-if questions and predict the outcome of a scenario.
- Requires more skill set and infrastructure.
- Provides more value to the institution.

Big Data Analytic Categories

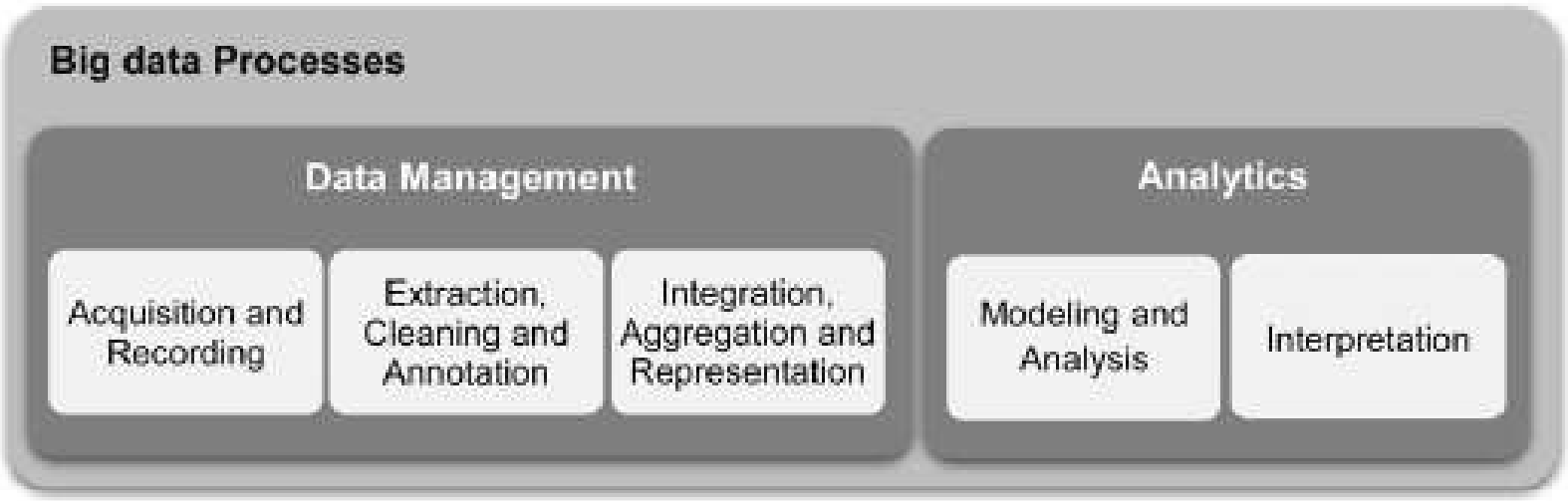
According to Purpose

Prescriptive Analytics

Tries to predict the future

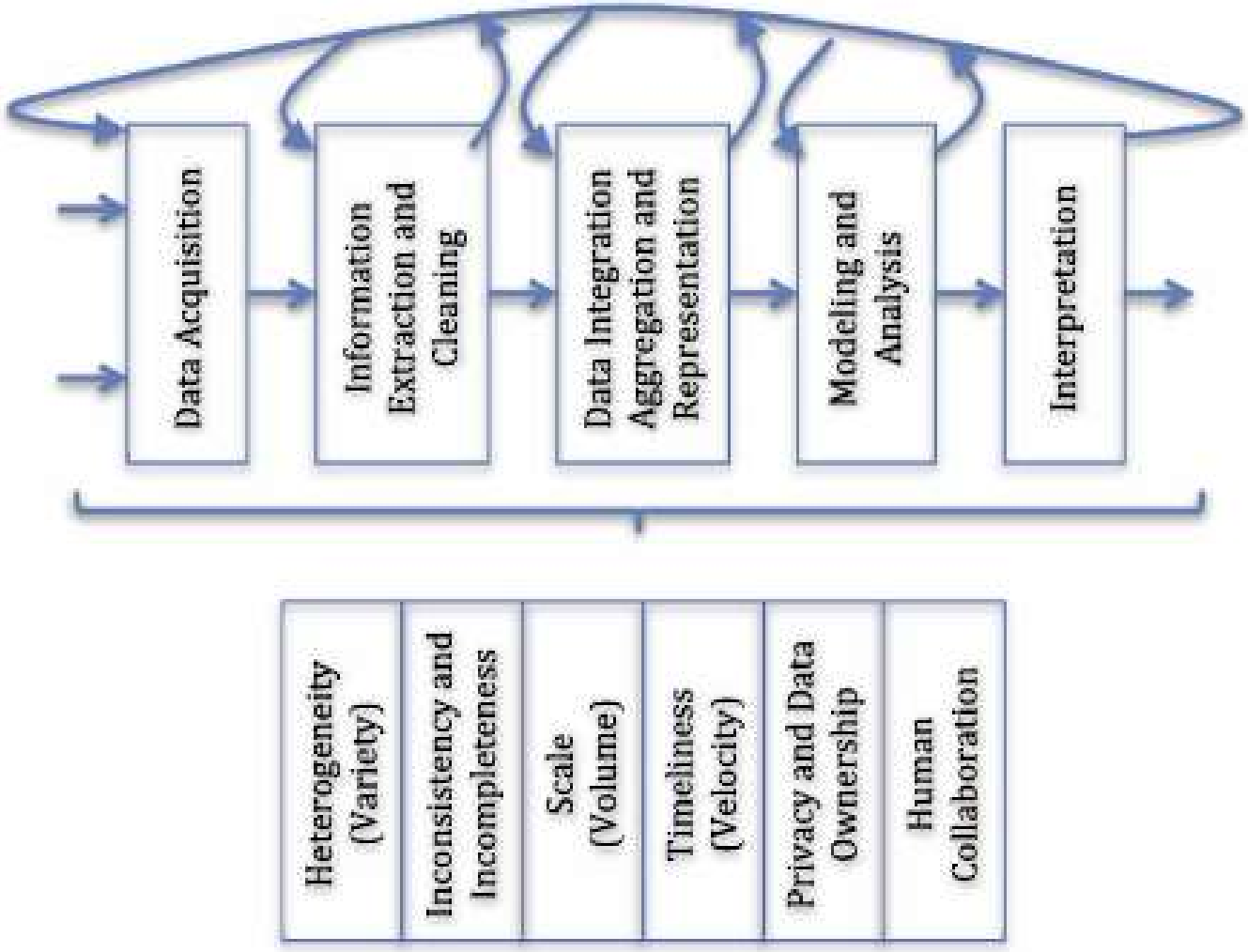
- By determining the best actions under business constrains
- Assess outcome of multiple actions or scenarios
- Based on predictive analysis but enhance it with situational understanding.
- Provides the highest value
- Requires the highest skillset, most sophisticated infrastructure, internal and external data

Big Data Analysis Process



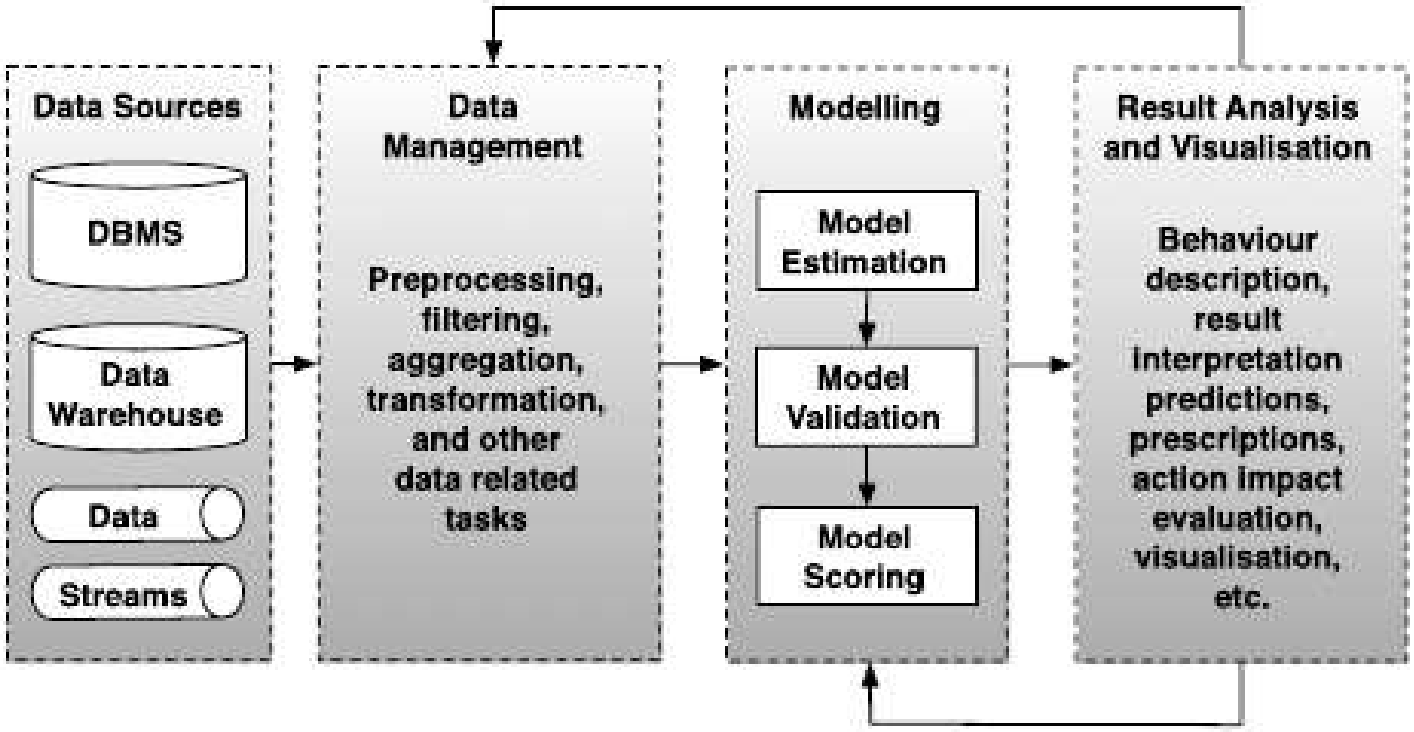
Big Data Analysis Process[1]

Big Data Analysis Pipeline



Big Data Analysis Pipeline [2, 5]

Big Data Analysis Workflow



Big Data Analysis Workflow[6]

What is an Anomaly?

Definition

What is an Anomaly?

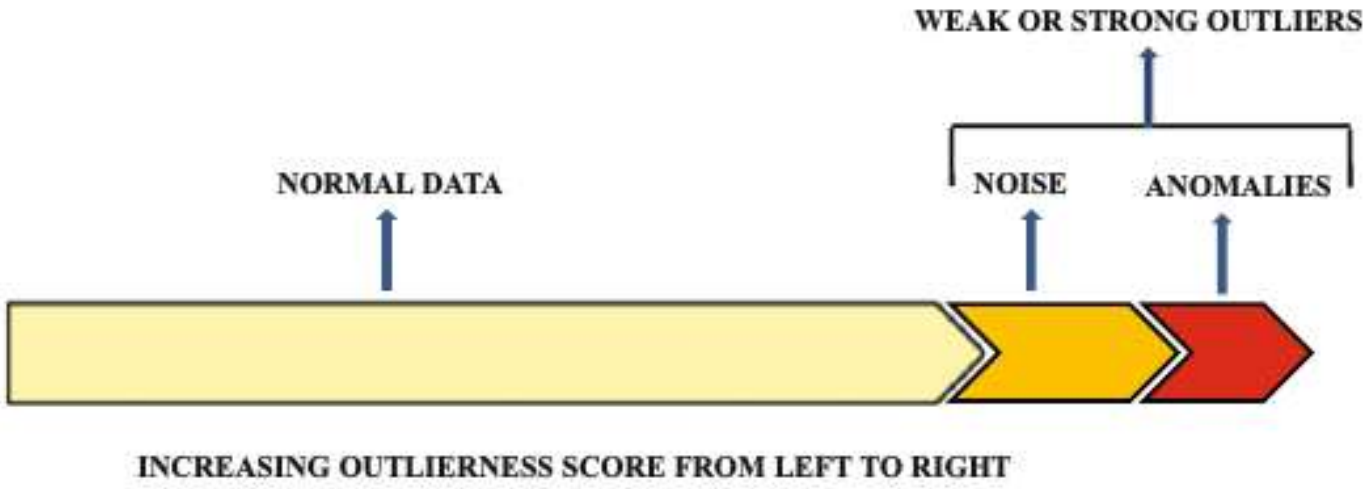
"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." [7]

Synonym to abnormalities, discordants, deviants, or **anomalies** in the data mining and statistics literature

Important application areas

- Intrusion detection systems
- Credit-card fraud
- Interesting sensor events
- Medical diagnosis
- Law enforcement
- Earth science

Anomaly Detection



Outlierness Spectrum[8]

Noise vs. Anomaly

- No clear distinction
- Application vs. Analyst specific
- Noise removal/identification is important
- Supervised and unsupervised scenarios

Anomaly Detection Algorithms

Models

- Model the *normal data*
- Calculate the anomaly score of each data

Output

- Anomaly Scores
- Binary Labels

General Classification of Models

- Probabilistic and Statistical
- Linear
- Proximity Based
- High Dimensional Models
- Outlier Ensembles
- Supervised
- Time-series, multidimensional, streaming
- Categorical, Text and Mixed Attribute Data
- Discreet Sequences
- Spacial
- Graphs and Networks

Problems of Anomaly Detection Algorithms

- The difficulty in identifying the boundaries between normal and abnormal cases
- The difficulty of collecting abnormal data to learn from (labeled normal and abnormal data)
- Most models are not suitable for anomaly detection for streaming data
- Most models can not cope with big data
- Most models are suitable for a very customized environments, not applicable in general
- Most models does not give an insight about the reasons of abnormality, just provides scores