Data Analytics

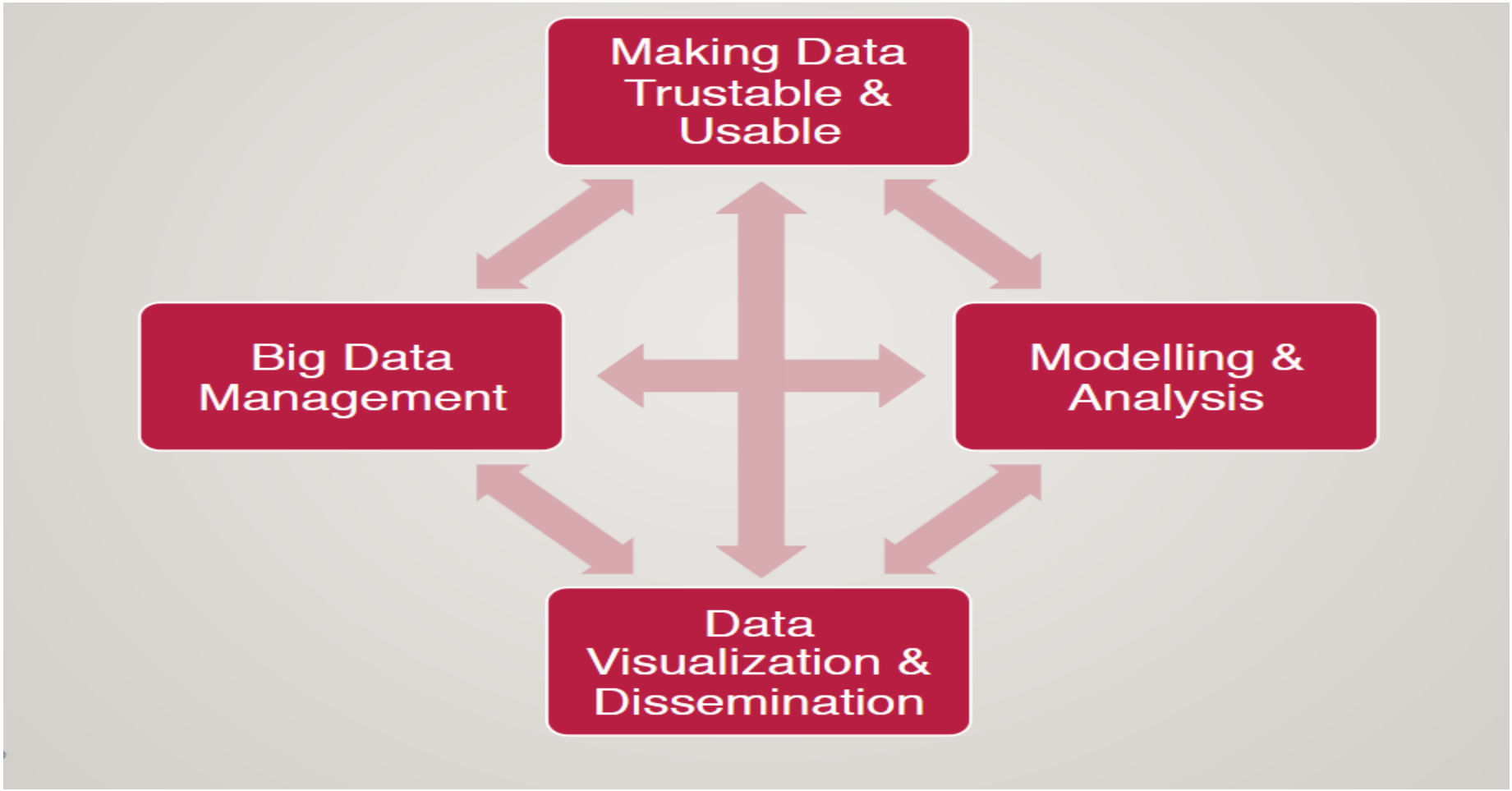

DETECTION
SOCIAL MEDIA
WWW
SERVICES
PROJECTS
MULTIMEDIA
NETWORK
PREDICTIVE
PROGRAM
ANALYTICS
BIG DATA
COMPUTING
INFORMATION
TECHNOLOGY
PROMOTION
CONTENT
PROCESSING
CONSUMER
ORGANIZATION
PLANNING
EVENTS
PROGRAMMING
CODING
SOFTWARE
MODELS
E-MARKETING
COMMUNICATION
BRANDING
COMPUTER
CONSUMER DEMAND MARKETS
WEB MARKETING
DATA MINING
DATA SCIENCE
INFORMATION
SOLUTIONS
MACHINE LEARNING
WEB SERVICES
MATHS PATTERN
ENGINEERING
PLANNING
MEDIA
STATISTICS
PROJECTS
BIG
DATA
MOBILE
TARGET
SERVICE
VISION
ENGINEERING
RESEARCH
PROBABILITY
WEB DEV
KDD
STRATEGY
WORLDWIDE
COMPUTING
VISUALIZATION
BIG
DATA
PRO
PRICING
CODING
SEGMENTATION
INFORMATION
DIGITAL
SOCIAL NETWORKS
SOCIAL
NETWORK

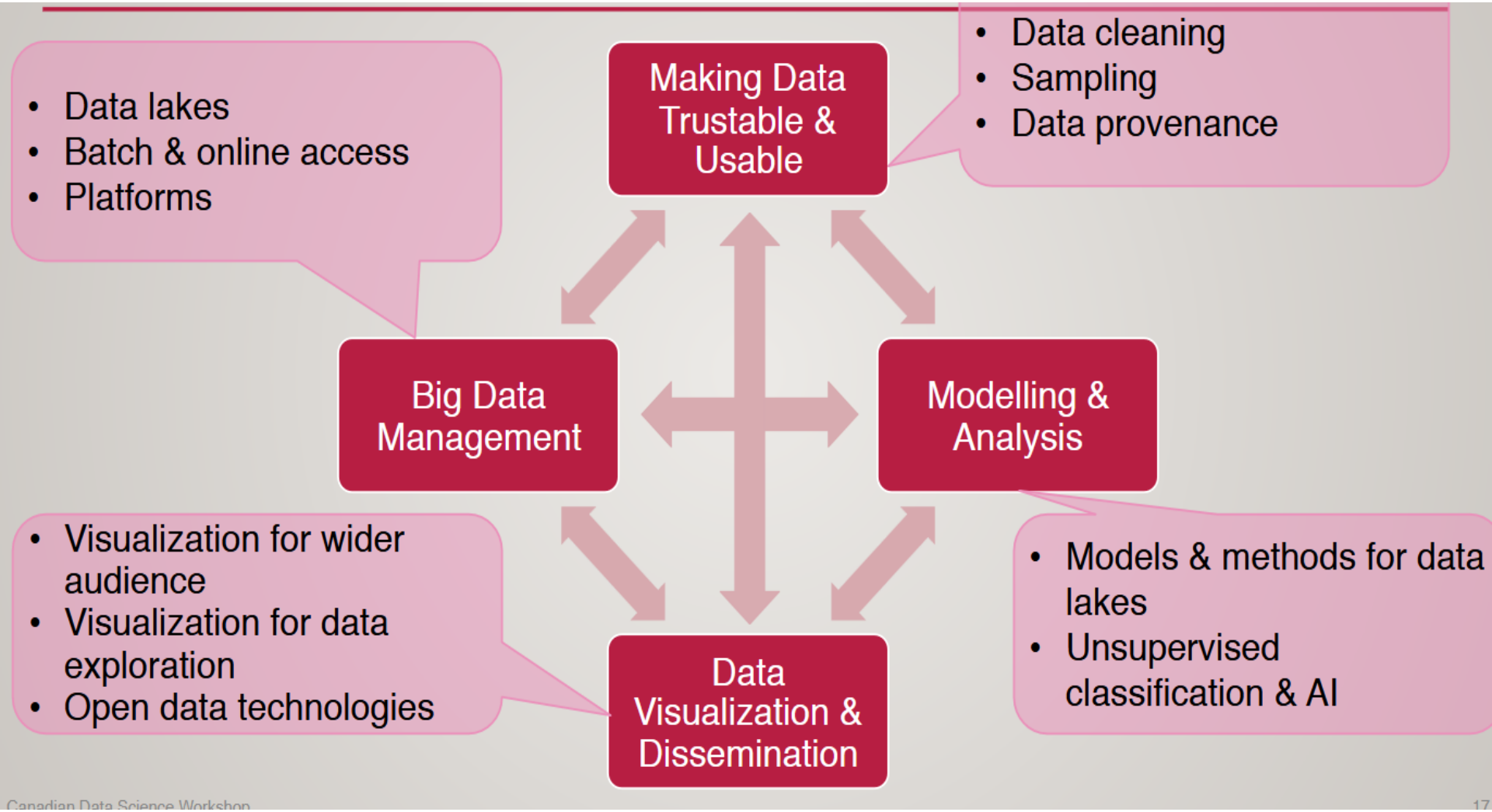
WHAT IS DATA SCIENCE?

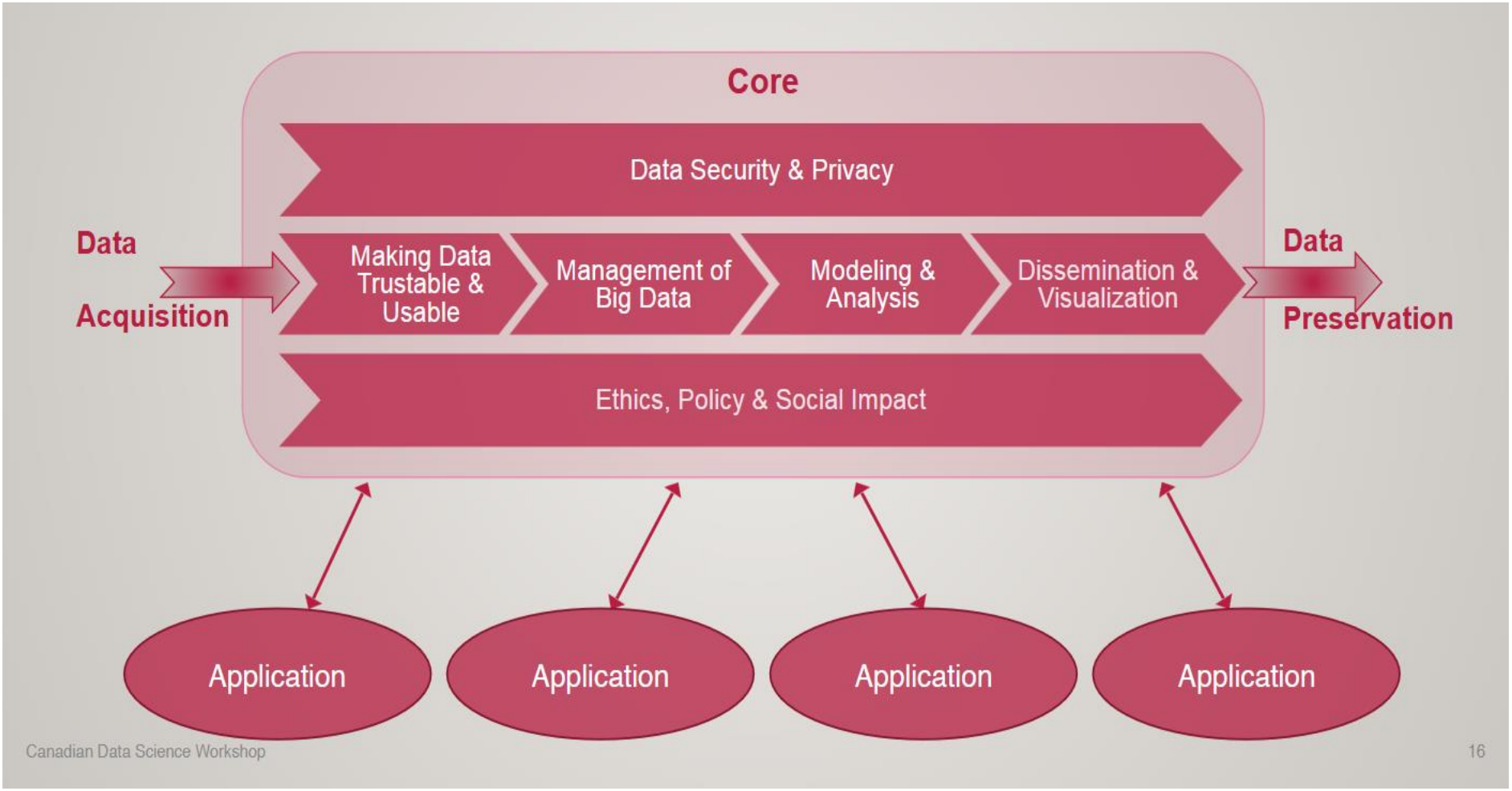
"Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."

- “Data science intends to analyze and understand actual phenomena with ‘data’. In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human, and social phenomena with data from a different point of view from the established or traditional theory and method.”

CORE RESEARCH ISSUES & INTERACTIONS



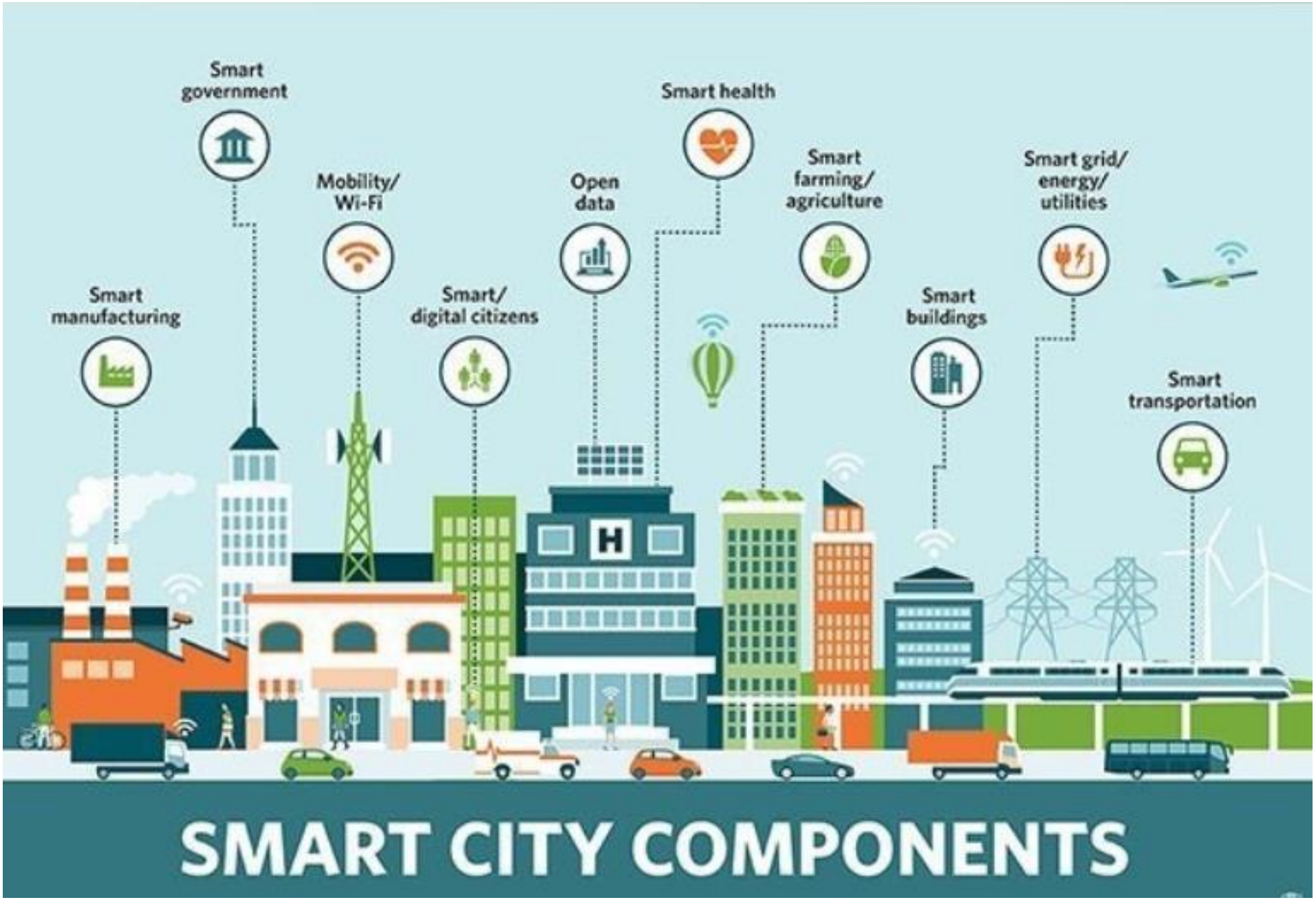

Making Data Trustable & Usable
• Data cleaning
• Sampling
• Data provenance
• Data lakes
• Batch & online access
• Platforms
Big Data Management
Modelling & Analysis
• Visualization for wider audience
• Visualization for data exploration
• Open data technologies
Data Visualization & Dissemination
• Models & methods for data lakes
• Unsupervised classification & AI


Core
Data Security & Privacy
Data
Acquisition
Making Data Trustable & Usable
Management of Big Data
Modeling & Analysis
Dissemination & Visualization
Data
Preservation
Ethics, Policy & Social Impact
Application
Application
Application
Application

- DM support
- Data preparation for big data management
- Cleaning for data analysis
- DM for ML
- ML for DM
- Visual analytics

DATA SCIENCE APPLICATION

- Fraud detection
- Recommender systems
- Predicting acts
- Smart cities
- Optimization – ex. IPL/Moneyball


Smart government
Smart health
Mobility/ Wi-Fi
Open data
Smart farming/ agriculture
Smart grid/ energy/ utilities
Smart manufacturing
Smart/ digital citizens
Smart buildings
Smart transportation
H
SMART CITY COMPONENTS

What is Data Analytics

- The intent of Data Analytics is to transform raw data into valuable information.
- Data analytics is used in today's business world by examining the data to generate models for predictions of patterns and trends.
- When used effectively, data analytics gives us the ability to search through large and unstructured data to identify unknown patterns or relationships, which when organized, is used to provide useful information.

- **Data Science:**

A comprehensive discipline combining statistics, machine learning, programming, and domain expertise to extract actionable insights from structured and unstructured data, often including predictive modeling and complex analysis to forecast future outcomes.

- **Data Analytics:**

The process of examining data to draw conclusions and inform business decisions by analyzing trends and patterns within existing data, typically using simpler statistical methods and visualization techniques.

- **Data Mining:**

A subset of data science that specifically focuses on discovering hidden patterns and relationships within large datasets using advanced algorithms and techniques to uncover previously unknown insights.

Example Use Cases:

- **Data Science:**

  Developing a predictive model to identify customers likely to churn, analyzing complex customer behavior patterns to personalize marketing campaigns.

- **Data Analytics:**

  Analyzing website traffic data to identify peak usage times, generating customer segmentation reports based on purchase history.

- **Data Mining:**

  Identifying fraudulent transactions in a large financial dataset, discovering customer preference trends in a large social media dataset.

What is the value of Big Data and Data Analytics

- 85% of CEOs put a high value on Data Analytics.
- 80% of CEOs place data mining and analysis as the second-most important strategic technology.
- Business analytics tops CEO's list of priorities.
- Data Analytics could generate up to $3 trillion revenue per year.

The Power of Data Analytics

- With a wealth of data on their hands, companies are empowered by using data analytics to discover various patterns, investigate anomalies, forecast future behavior, and so forth.
- Patterns discovered from historical data enable businesses to identify future opportunities and risks.
- In addition to producing more value externally, studies show that data analytics affects internal processes, improving productivity, utilization, and growth.

Benefits and Costs of Data Analytics

ETL: Extract - Transform - Load

- Reformatting, cleansing, and consolidating large volumes of data from multiple sources and platforms can be especially time consuming.

- Data analytics professionals estimate that they spend between 50 percent and 90 percent of their time cleaning data for analysis.

- The cost to scrub the data includes the salaries of the data analytics scientists and the cost of the technology to prepare and analyze the data.

- As with other information, there is a cost to produce these data.

The impact of Data Analytics

- Many companies address the likely possibility that the data their organizations hold influence their market value.
- Facebook, for example, has a large amount of its market value driven by the number of users on the platform and the amount of data those users contribute which is sold to third parties.

  Ex. Jio-Meta deal
- Data analytics often also involves data management and business intelligence with knowledge of business functional areas.
- Today there is an increasing number of investments in data analytics and increasing demand for data analytics–related tasks

The impact of Data Analytics

- The real value of data comes from the use of data analytics.
- Companies are getting much smarter about using data analytics to discover various patterns, investigate anomalies, forecast future behavior, and so forth.
- For example, companies can use their data to do more directed marketing campaigns based on patterns observed in their data.
- That can give them a competitive advantage and it can also be used on historical data to enable businesses to identify future opportunities and risks.

Data Analysis

- Analysis: Turning raw **data**
- into useful **information**

- Purpose: To provide answers to questions being asked by a health program

- Even the greatest amount and best quality of data mean nothing if data are not properly analyzed—or analyzed at all.

Data Analysis

- Analysis does not mean using a computer software package.
- Analysis is looking at the data in light of the questions you need to answer:
- How would you analyze data to determine: "Is my program resulting its objectives?"

Answering Program Questions

- **Question:** Is my program reslting its objectives?
- **Analysis**: Compare program targets and actual program performance to learn how far you are from the targets
- **Interpretation**: Why have you achieved or not achieved a target, and what does this mean for your program?
- Answering may require more information.

Descriptive Analysis

- Describes the sample/target population (demographic and clinical characteristics)
- Does not define causality; tells you what, not why
- Example: Average number of clients seen per month

Basic Terminology and Concepts

- **Statistical terms**
  - Ratio
  - Proportion
  - Percentage
  - Rate
  - Mean
  - Median
  - Trend

Central Tendency

Measures of the location of the middle or the center of a distribution of data

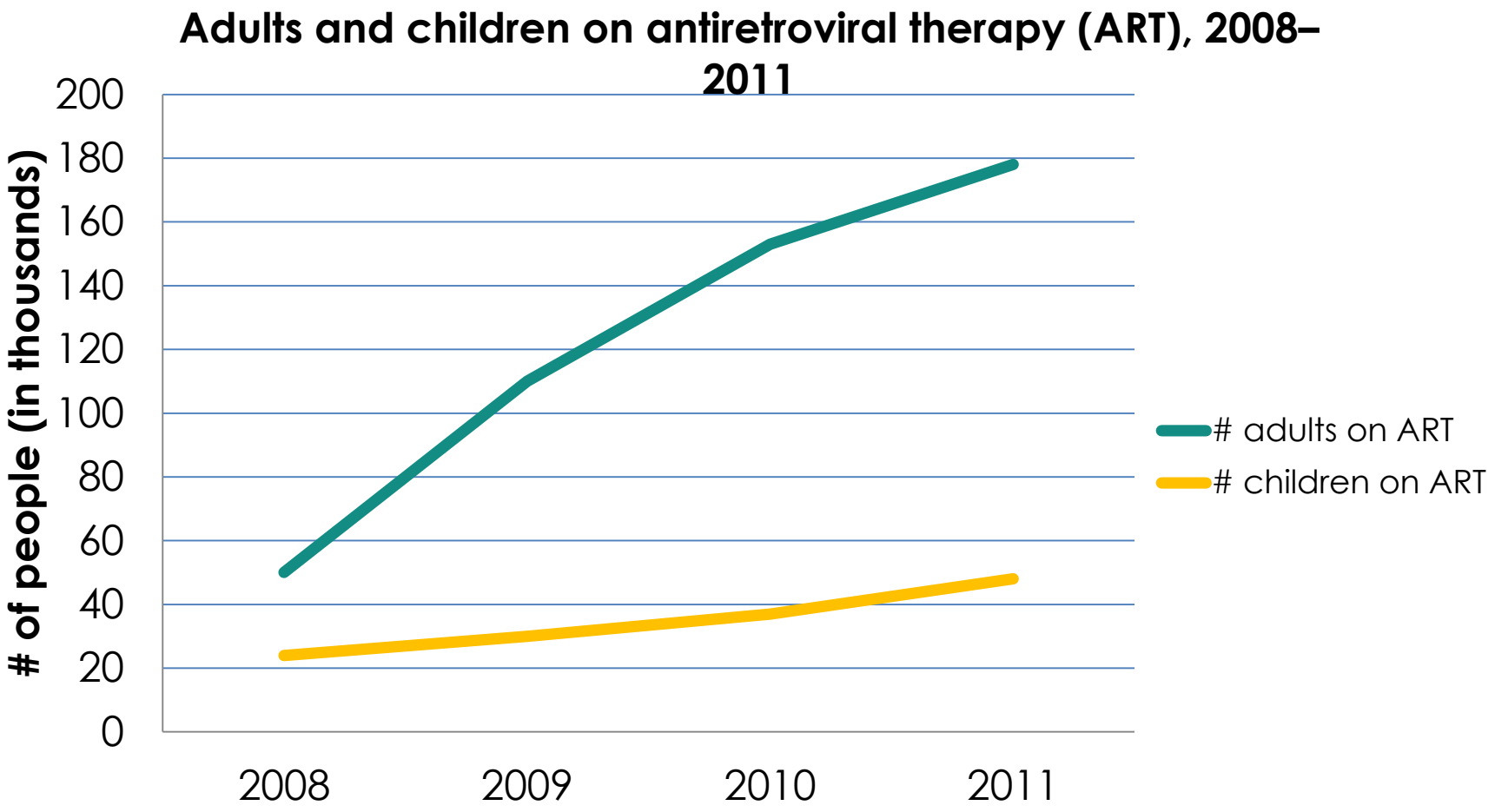
- Mean
- Median
- Mode
- 5 no summary

Trend

- A trend is a pattern of gradual change in a condition, output, or process, or an average or general tendency of a series of data points to move in a certain direction over time, represented by a line or curve on a graph.

- To follow a trend you must not only be aware of what is currently happening but also be astute enough to predict what is going to happen in the future.

Calculating Trends



Key Messages

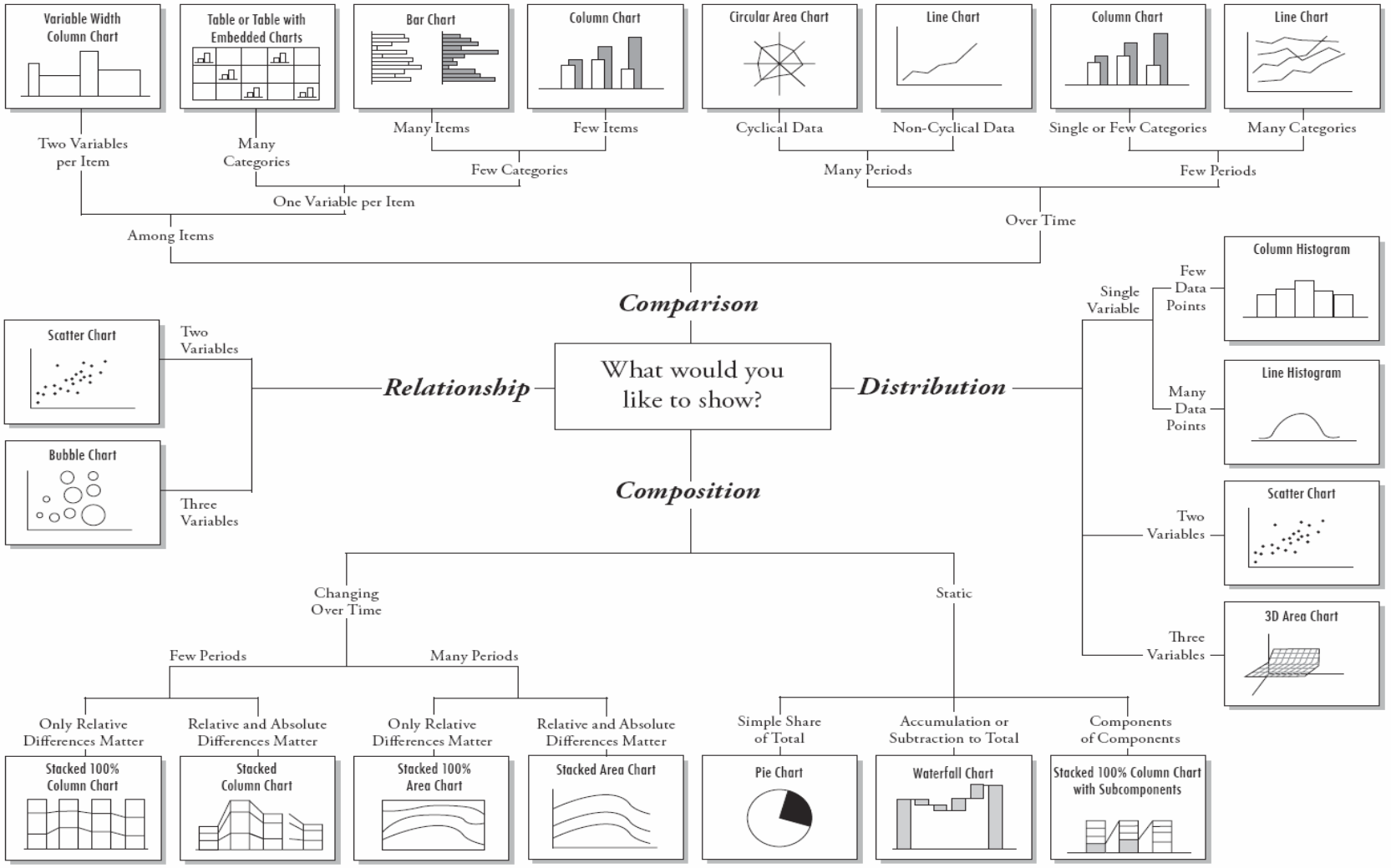
- Purpose of analysis: Provide answers to programmatic questions

– Descriptive analyses describe the sample or target population.

– Descriptive analyses do not define causality. That is, they tell you *what,* not *why*.

SELECT THE RIGHT CHART

Types of Charts



5 Questions to Ask Yourself When Choosing a Chart

1. Want to compare values?

Charts are perfect for comparing one or many value sets, and they can easily show the low and high values in the data sets.

- Use these charts to show comparisons:
    - Column/bar
    - Circular area
    - Line
    - Scatter plot
    - Bullet

5 Questions to Ask Yourself When Choosing a Chart

2. Want to show the composition of something?

- To show how individual parts make up the whole of something (such as the device used for mobile visitors to your website, or total sales broken down by sales rep)

  - Use these charts to show composition:
    - Pie
    - Stacked bar
    - Stacked column
    - Area

5 Questions to Ask Yourself When Choosing a Chart

3. Want to understand the distribution of your data?

- Distribution charts help you to understand outliers, the normal tendency, and the range of information in your values.

  - Use these charts to show distribution:
    - Scatter plot
    - Line
    - Column
    - Bar

5 Questions to Ask Yourself When Choosing a Chart

- **4. Interested in analyzing trends in your data set?**
- If you want more information about how a data set performed during a specific period, there are specific chart types that do this extremely well.

  – Use these charts to analyze trends:
    - • Line
    - • Dual-axis line
    - • Column

5 Questions to Ask Yourself When Choosing a Chart

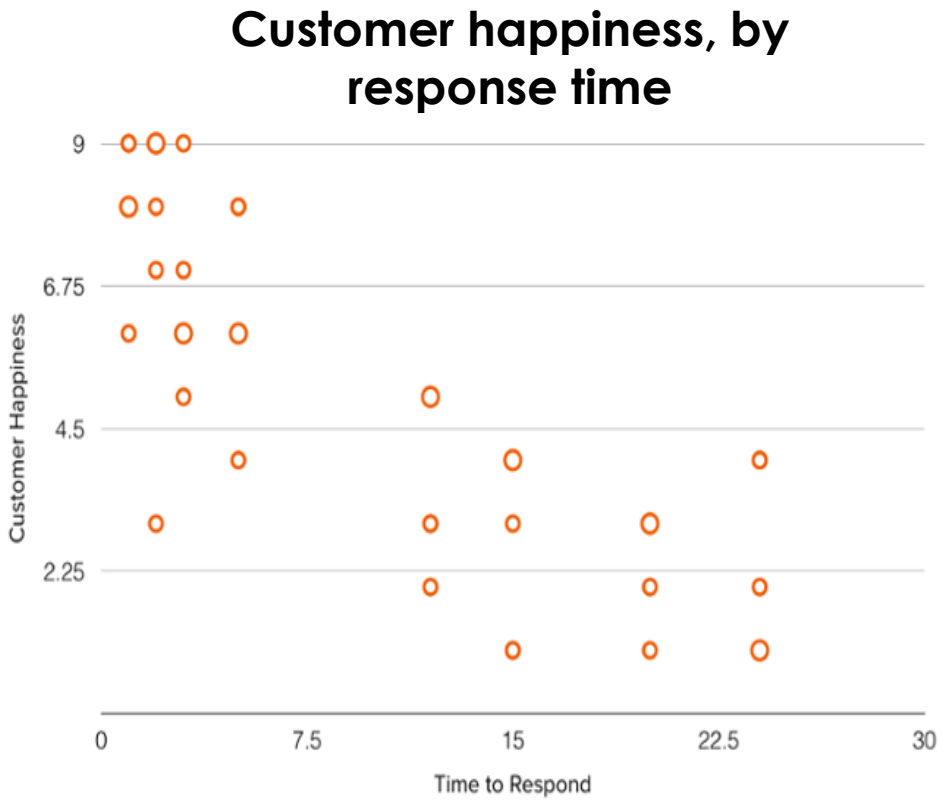
5. Want to better understand the relationships among value sets?

- Relationship charts are designed to show how one variable relates to one or many different variables. You could show how something positively affects (or has no effect, or negatively affects) another variable.

  - Use these charts to show relationships:
    - Scatter plot
    - Bubble
    - Line

Scatter plot

- Can show relationship between two variables, or reveal the distribution trends
- Should be used when there are many data points, and you want to highlight similarities in the data set
- Useful when you are looking for outliers or want to understand the distribution of your data



Types of Data Science Tools

Data science tools are categorized into several broad types, each with specific functions and common denominators.

- **Data Acquisition and Data Science Storage Tools**: These data science tools are primarily concerned with collecting, storing, and managing data—databases (MySQL, PostgreSQL, MongoDB), data warehouses (Redshift, Snowflake), and data lakes (Hadoop, S3).
- **Data Cleaning and Preparation Tools**: These data science tools—Pandas, NumPy, and OpenRefine — clean, transform, and prepare data for analysis.
- **Data Exploration and Visualization Tools**: These data science tools—Power BI, Tableau, Matplotlib, and Seaborn — aid in understanding and communicating data through visual representations.
- **Machine Learning and Modeling Tools**: These data science tools—Scikit-learn, TensorFlow, PyTorch, and Keras — provide algorithms and frameworks for building and training machine learning models.
- **Model Deployment and Management Tools**: These data science tools help deploy and manage machine learning models in production environments, such as MLflow and Kubeflow.
- **Big Data Tools**: These data science tools, Apache Spark and Apache Flink, are designed to handle massive datasets.
- **Natural Language Processing (NLP) Tools**: These data science tools are used for tasks involving human language, such as text classification, sentiment analysis, and machine translation – NLTK, spaCy, and Gensim.
- **Deep Learning Tools**: These data science tools—TensorFlow, PyTorch, and Keras — are specialized for building and training deep neural networks.

Understanding Specific Needs for Data Science Tools

When picking data science tools, think about these things:

- What's your mission? What do you want to achieve with your data?
- Is your data a giant mountain or a small hill? Large or Small data set.
- How big and complex is it?
- What are the data science tools required to handle your specific data challenges?
- Do your people know their stuff? What are their skills and experience?
- Does it play well with others? Can the data science tools work with your existing systems? Compatibility
- Can it grow with you? Will it handle more data and bigger projects in the future?
- How much can you spend? Set a budget and stick to it.
- Is there a community behind it? Look for data science tools with lots of help and resources.
- What have others done? See how other businesses have used similar data science tools.

The Data Science Tools Main Functions

- Data scientists collect data from various sources, ensuring it's clean and ready for performance.
- Scrub away errors, inconsistencies, and missing notes.
- Explore the data, search for patterns and trends, and use visualization tools to paint a picture of the data, making complex information understandable and engaging.
- With a clear vision of the desired outcome, select the appropriate algorithms to build models.
- Train these models, feeding them data to learn and improve.
- Once the model is ready, evaluate its performance, ensuring it meets the desired standards.
- Then deploy it into the real world, where it can perform its magic.
- To handle data's increasing volume and complexity, scientists rely on data science automation tools to streamline repetitive tasks.

Choosing the Right Data Science Tools

When choosing data science tools, think about these things:

- Look for data science tools that handle data efficiently without sacrificing accuracy or speed.
- Choose data science tools that are intuitive and easy to learn, especially if you're new to data science.
- A strong community provides valuable resources, tutorials, and assistance when you encounter challenges.
- Consider your budget and choose data science tools that align with your financial resources.
- As data grows and projects become more complex, data science tools should be able to scale with your needs.
- Choose tools that increase workloads and data volumes without compromising performance.

Data Analytics Tools Benefits

- ***Improved Productivity***

Many data science tools automate repetitive tasks, freeing up data scientists to focus on more strategic work. These data science tools streamline workflows and provide shortcuts, reducing the time it takes to complete tasks. By standardizing processes and workflows, data science tools ensure consistency and reproducibility in data analysis.

- ***Scalability***

Data science tools are designed to control massive datasets efficiently, allowing organizations to analyze large-scale data without compromising performance. Many data science tools data science support parallel processing, enabling faster execution of computationally intensive tasks. Cloud-based data science platforms provide scalable infrastructure adjusted to meet changing demands.

- ***Reproducibility***

Tools often integrate with version control systems, allowing data scientists to track changes and reproduce experiments accurately. Well-documented code and workflows ensure that experiments are easily replicated and understood by others. Using standardized data science tools and techniques can improve reproducibility and team collaboration.

- ***Faster Experimentation***

Data science tools facilitate rapid prototyping, allowing data scientists to quickly test different models and approaches. Tools support iterative web development, enabling data scientists to refine models based on feedback and experimentation. Many data science tools provide features for tracking experiments, making it easier to compare results and identify the best-performing models.

Tailoring Data Science Tools to Specialized Needs

Specialized Needs in Data Science Tools

- **Industry**: A healthcare organization may require tools to store sensitive patient data and comply with regulations like HIPAA. A financial institution, on the other hand, may need tools to drive large datasets and perform complex financial calculations.
- **Data Type and Volume**: A business dealing with unstructured text data may need NLP tools, while a company working with large-scale numerical data may require tools that lift distributed computing.
- **Use Case**: A business aiming to build a predictive model for customer churn may require different data science tools than a business trying to optimize supply chain logistics.
- **Team Expertise**: The data science team's skills and experience also influence tool selection. If the team is proficient in a particular programming language or framework, it may be more efficient to choose tools that align with their expertise.
- **Integration with Existing Systems**: The data science tools must be able to integrate seamlessly with the organization's existing IT infrastructure, including databases or data warehouses.

Five Essential Data Science Tools

- **Python**: This versatile programming language handles everything from simple data manipulation to complex machine-learning models.
- **NumPy**: If you need to perform complex mathematical operations on large datasets, NumPy is your go-to data science tool.
- **Power BI**: This business intelligence tool transforms raw data into stunning visualizations that tell a compelling story.
- **Hadoop**: If you're dealing with massive datasets too large to fit on a single computer, Hadoop is your solution. It distributes the data across multiple machines and processes it efficiently.
- **Amazon Web Services (AWS)**: AWS offers a wide range of data science services, from computing power to storage and machine learning tools. It's a personal assistant who handles all your data needs.

Generally, data scientists care about the following

- **Data understanding**: the characteristics, quality, and structure of the data they're working with.
- **Feature engineering**: selecting and creating relevant features from the data to improve model performance.
- **Model selection**: Choosing the most appropriate algorithms and techniques for the given problem and data.
- **Model evaluation**: Assessing the performance of the models using appropriate metrics and validation techniques.
- **Interpretability**: Understanding how the model works and being able to explain its predictions to stakeholders.

KNIME *Analytics Platform*

- **Visual workflows for data science**
- **care about the method, not the code**
- **Data science demands collaboration**
- **Learning data science shouldn't require a coding expertise.**

www.knime.com

Defining Project Goals, Data Types, and Scope