



**DATA  
SCIENCE**

# **GIỚI THIỆU NGÀNH DATA SCIENCE**





# 1 Giới thiệu ngành Data Science

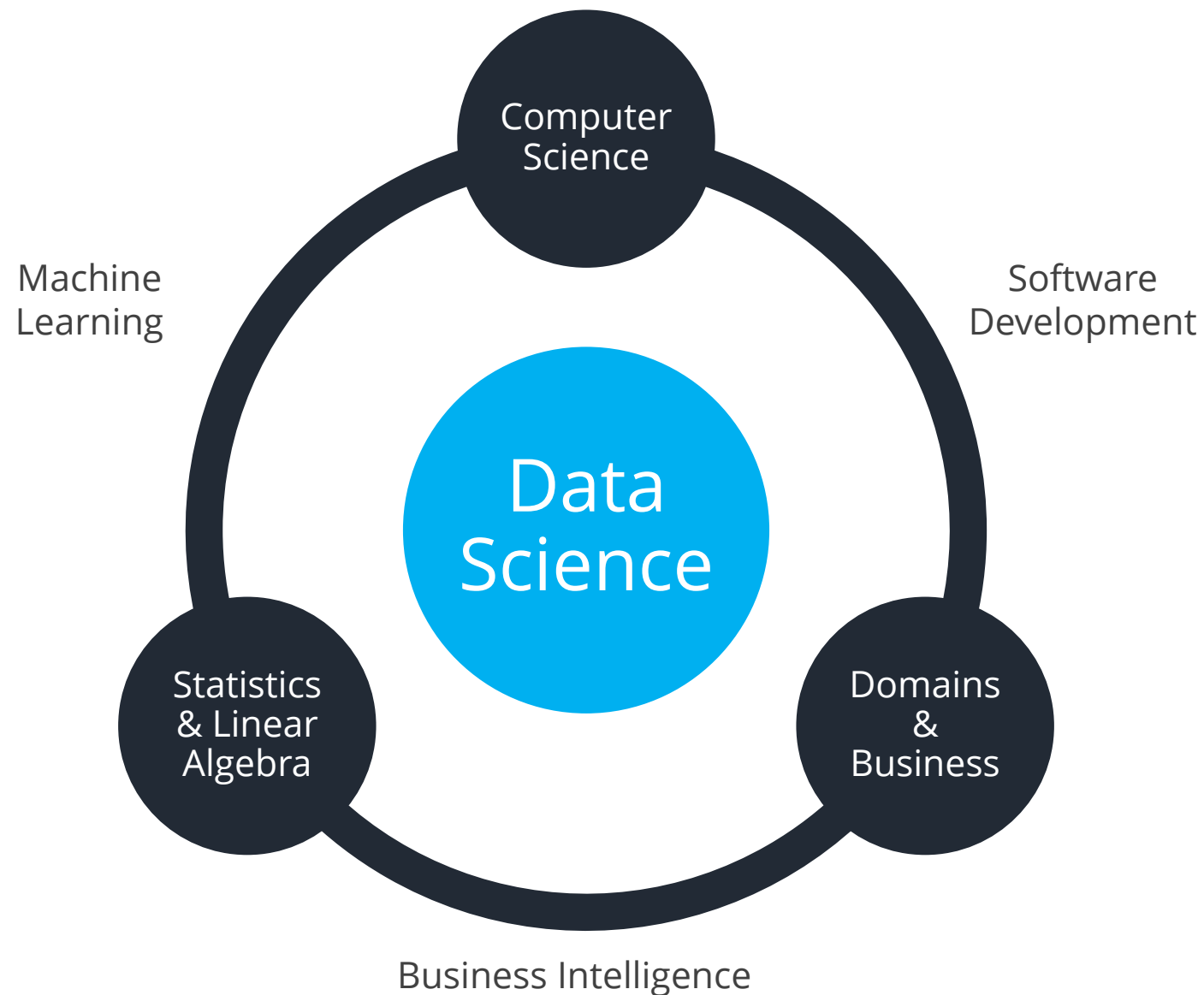




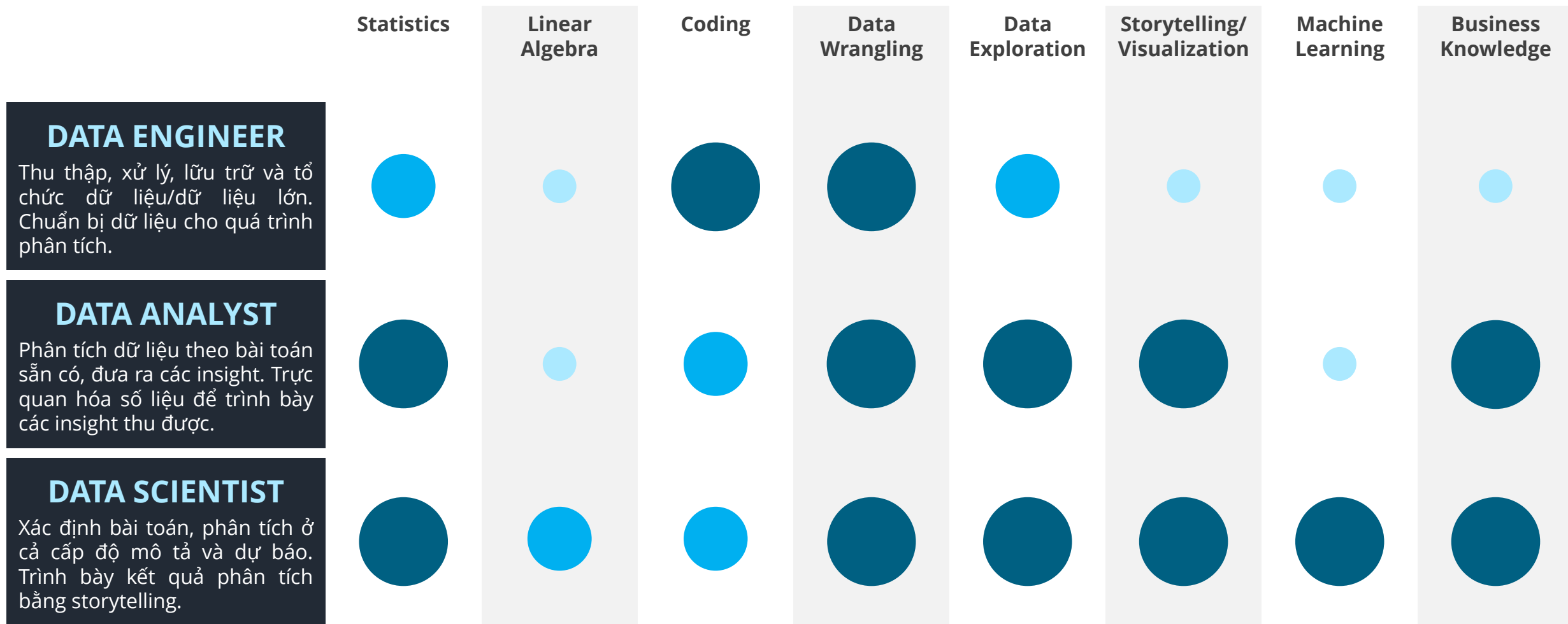
## Data Science là gì?

**“Khoa học Dữ liệu (Data Science) là lĩnh vực nghiên cứu dữ liệu để tìm ra các hiểu biết và tri thức có ý nghĩa.”**

# Nền tảng của Data Science



# Kỹ năng cần thiết cho Data Science



# Một số nghiệp vụ quan trọng

## Problem defining

- (1)** Xác định bài toán kinh doanh
- (2)** Thu thập các dữ liệu cần dùng để giải bài toán đó

## Data wrangling

- (1)** Phối hợp cùng Data Engineer đưa ra phương án tổ chức dữ liệu
- (2)** Biến đổi, làm sạch dữ liệu theo các kịch bản phân tích khác nhau

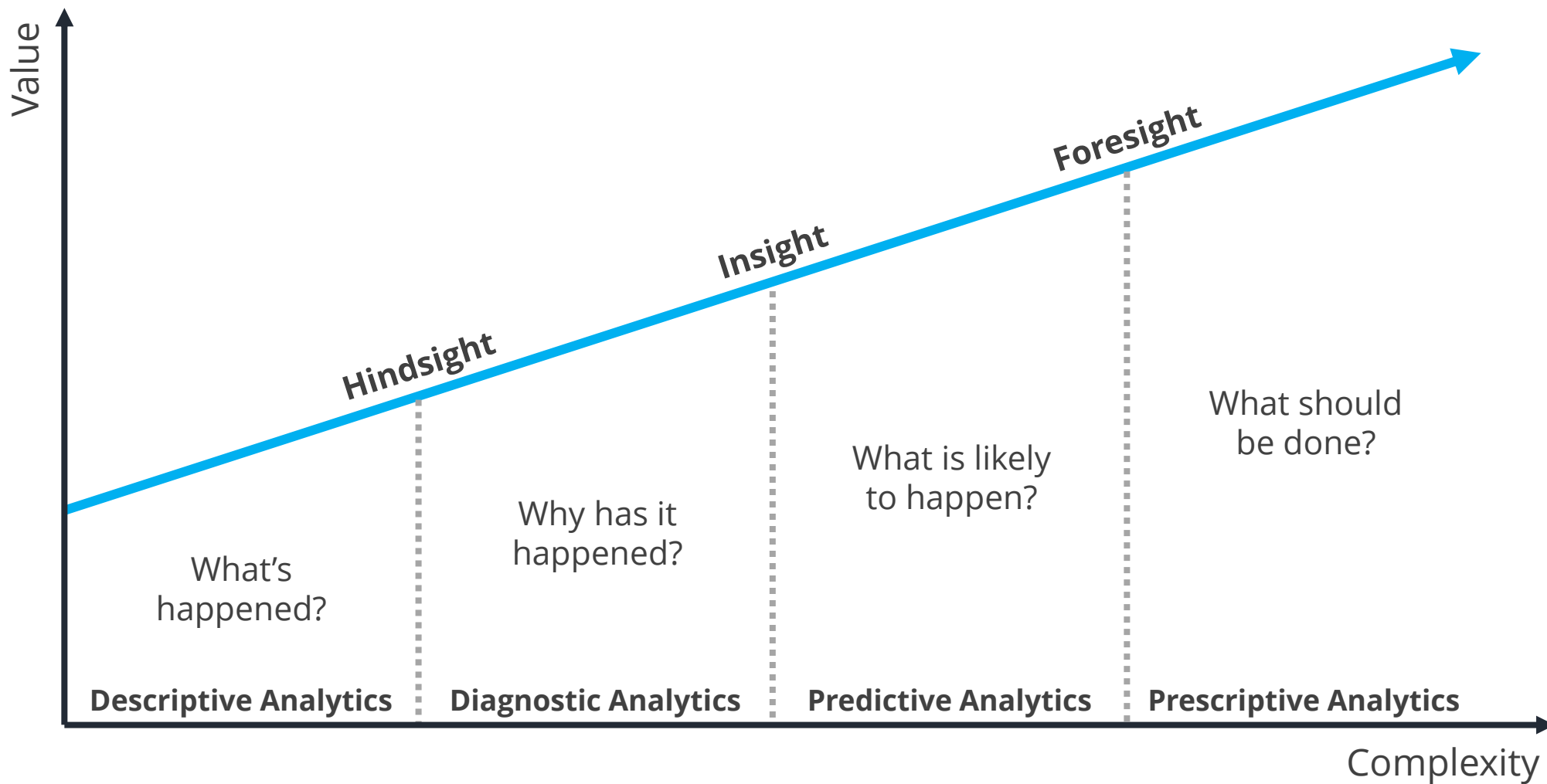
## Forecasting

- (1)** Sử dụng các mô hình thống kê và Machine Learning để dự báo giá, lượng,...
- (2)** Giải thích dự báo bằng ngôn ngữ end-user

## Visualization

- (1)** Trực quan hóa để nhìn ra quy luật của số liệu
- (2)** Trình bày tri thức rút ra từ số liệu dưới dạng dashboard và storytelling

## 4 cấp độ phân tích dữ liệu



# Data Science và Artificial Intelligence

## Data Science (Khoa học Dữ liệu)

- Sử dụng Machine Learning
- Gắn liền với kinh doanh
- Đầu ra là thông tin, quyết định

## Artificial Intelligence (Trí tuệ Nhân tạo)

- Sử dụng Machine Learning
- Gắn liền với công nghệ
- Đầu ra là công cụ, phần mềm



# Business Analytics và Business Analysis

## Business Analytics (Phân tích Kinh doanh)

Là một biến thể của Data Analytics

## Business Analysis (Phân tích Nghiệp vụ)

Là cầu nối giữa bộ phận IT và khách hàng

# Ngôn ngữ lập trình cho Data Science

## SQL

- Là ngôn ngữ truy vấn Database
- Không hỗ trợ Visualization, hỗ trợ hạn chế Data Wrangling
- Mức độ phổ biến cao

## R

- Là ngôn ngữ lập trình chuyên sâu về Statistics
- Hỗ trợ tốt: Data Wrangling, Visualization
- Khó tiếp cận, mức độ phổ biến không cao (thường được sử dụng cá nhân)

## Python

- Là ngôn ngữ lập trình đa mục đích
- Hỗ trợ tốt: Data Wrangling, Visualization, Machine Learning
- Dễ đọc, dễ học, mức độ phổ biến cao
- Không chuyên sâu về Statistics



## 2 Đối tượng làm việc của Data Analyst



## Dữ liệu bảng (Tabular data)

	Category	Datetime	Value	

Mỗi cột là một thuộc tính/chiều thông tin



# Biểu đồ (Chart)



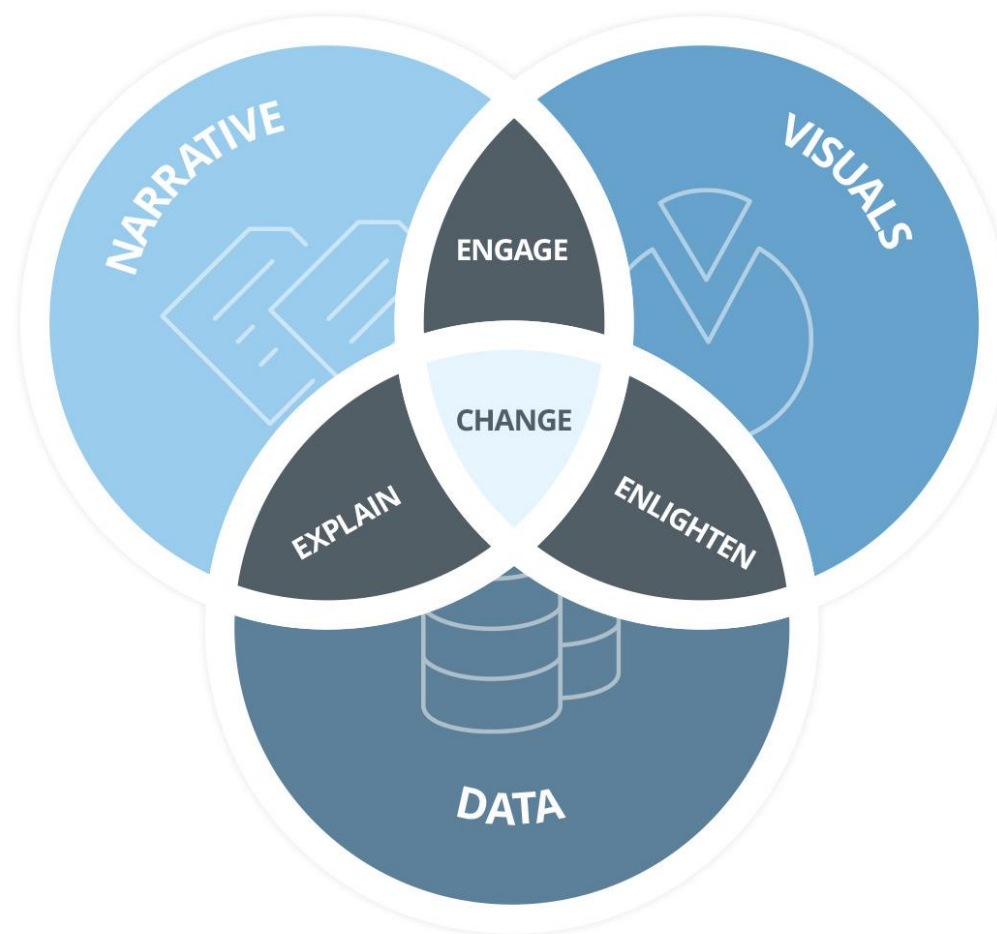
Mỗi biểu đồ biểu diễn 2 đến 4 chiều thông tin

# Dashboard



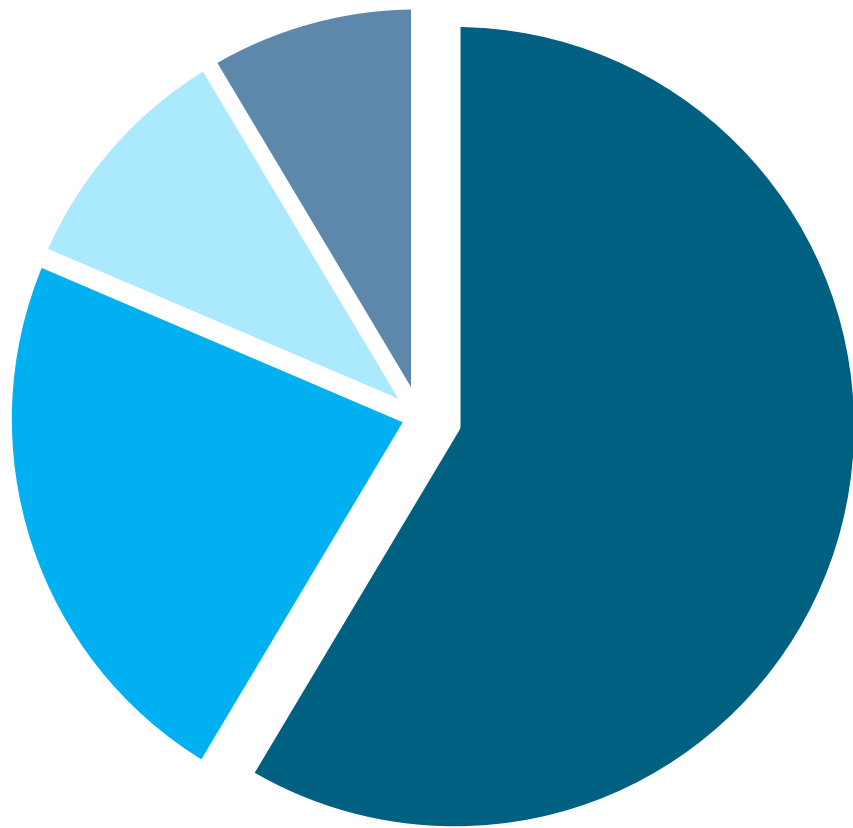
Báo cáo trực quan gồm nhiều biểu đồ và chỉ số

# Storytelling



Kể chuyện với thông tin, hình ảnh và dữ liệu

## Một số biểu đồ cơ bản



### Pie Chart

**Biểu diễn tỉ trọng thành phần**

**Chiều thông tin:**

- Kích thước pie - Value
- Màu sắc - Category



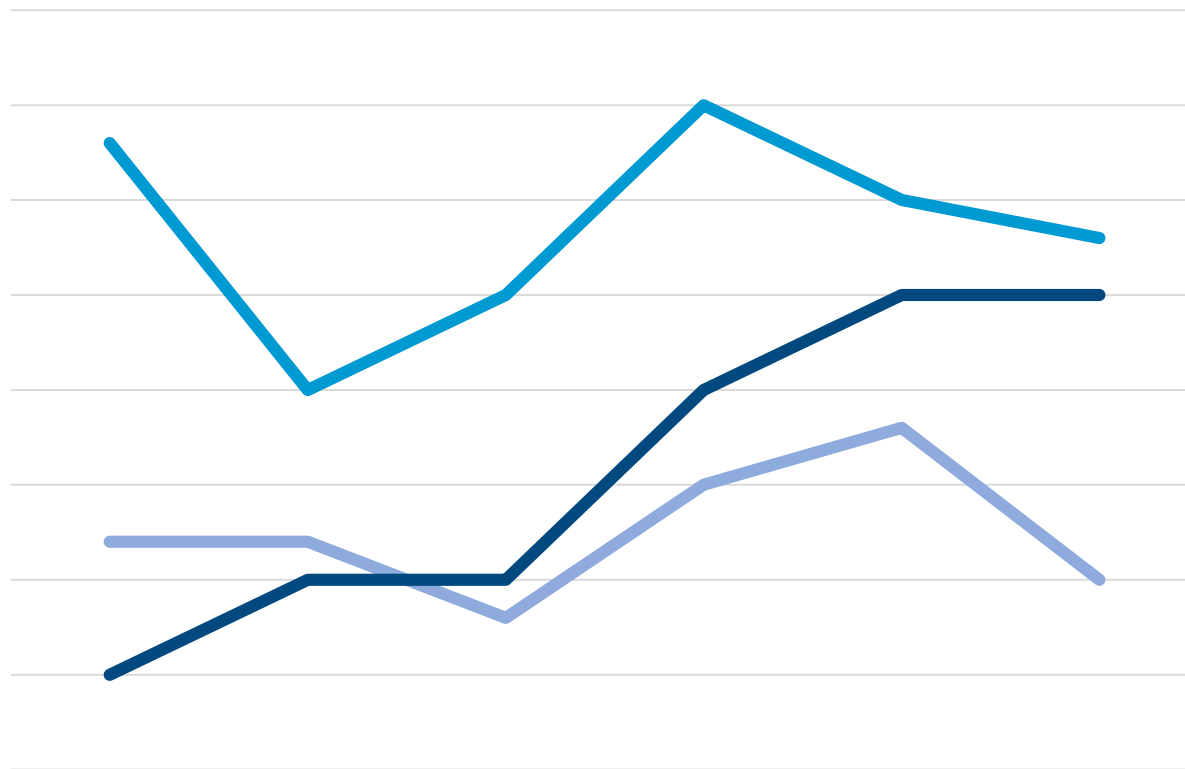
# Một số biểu đồ cơ bản

## Line Chart

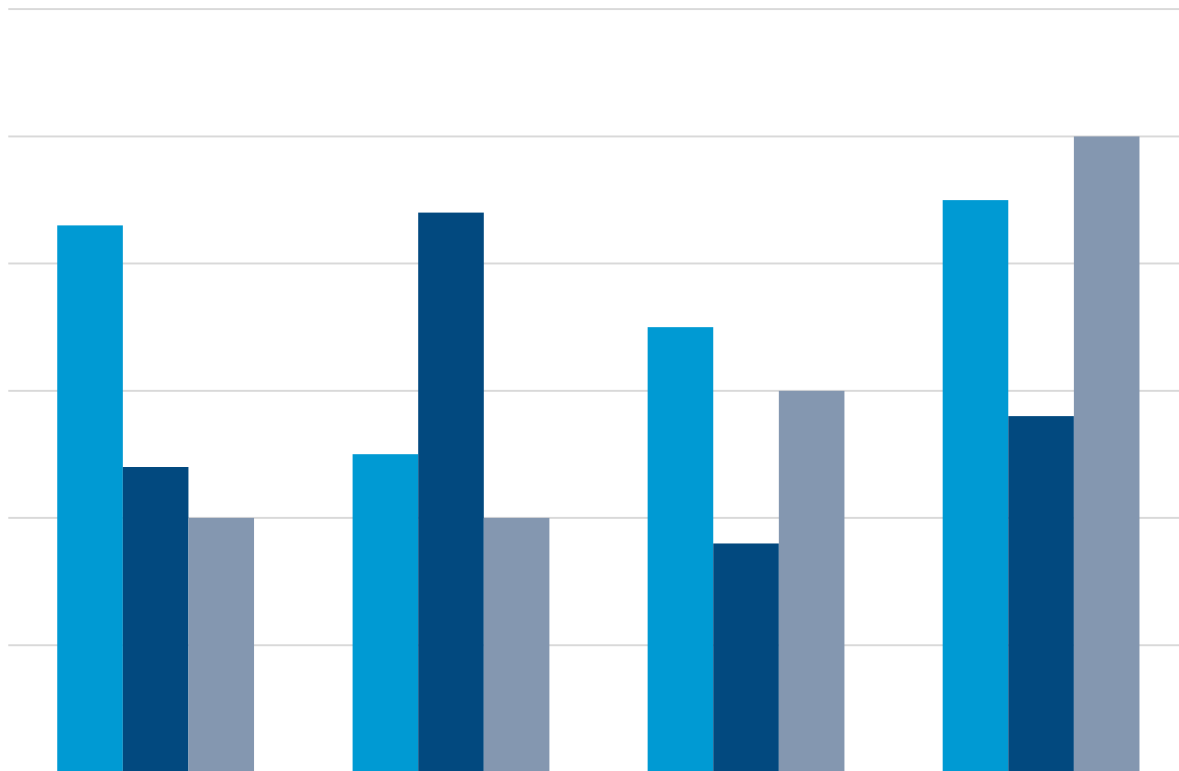
Biểu diễn biến động theo thời gian

Chiều thông tin:

- Trục x - Datetime
- Trục y - Value
- Màu sắc - Category



## Một số biểu đồ cơ bản



### Group Bar Chart

So sánh thành phần theo thời gian

Chiều thông tin:

- Trục x - Datetime hoặc Category
- Trục y - Value
- Màu sắc - Category

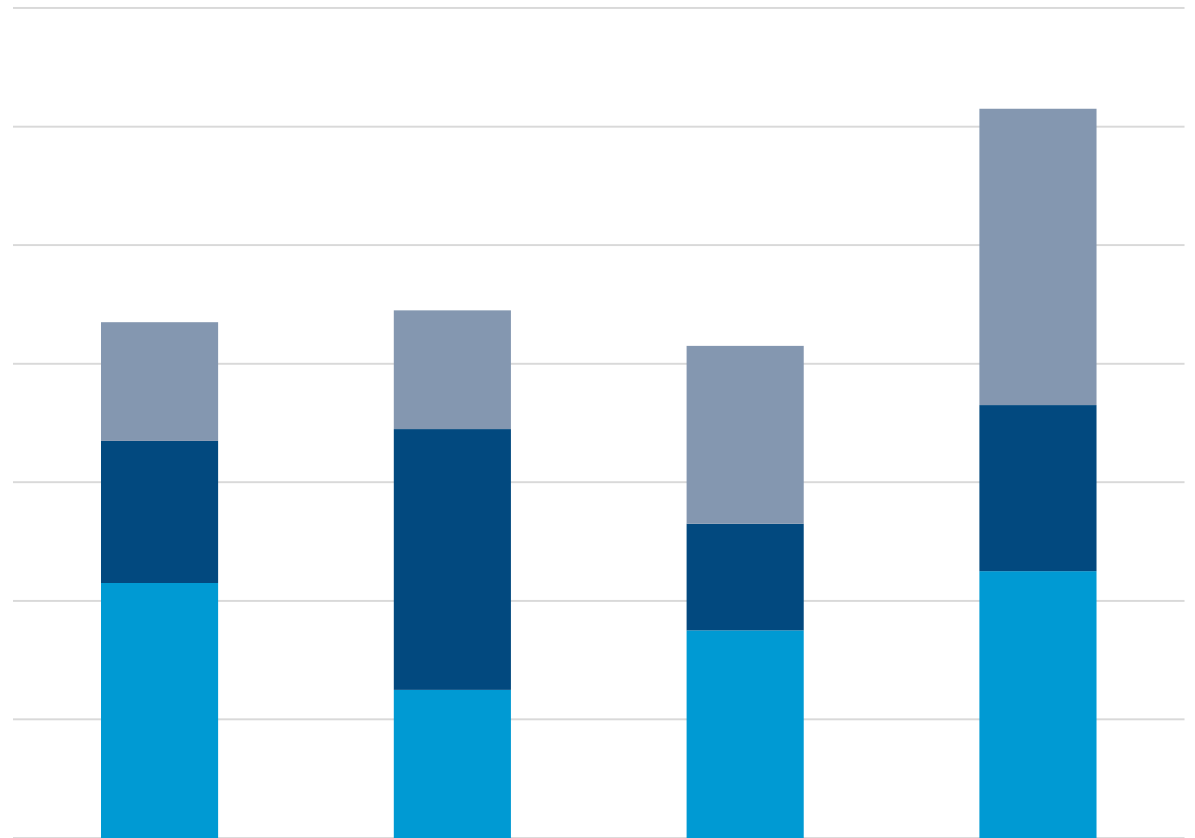
# Một số biểu đồ cơ bản

## Stacked Bar Chart

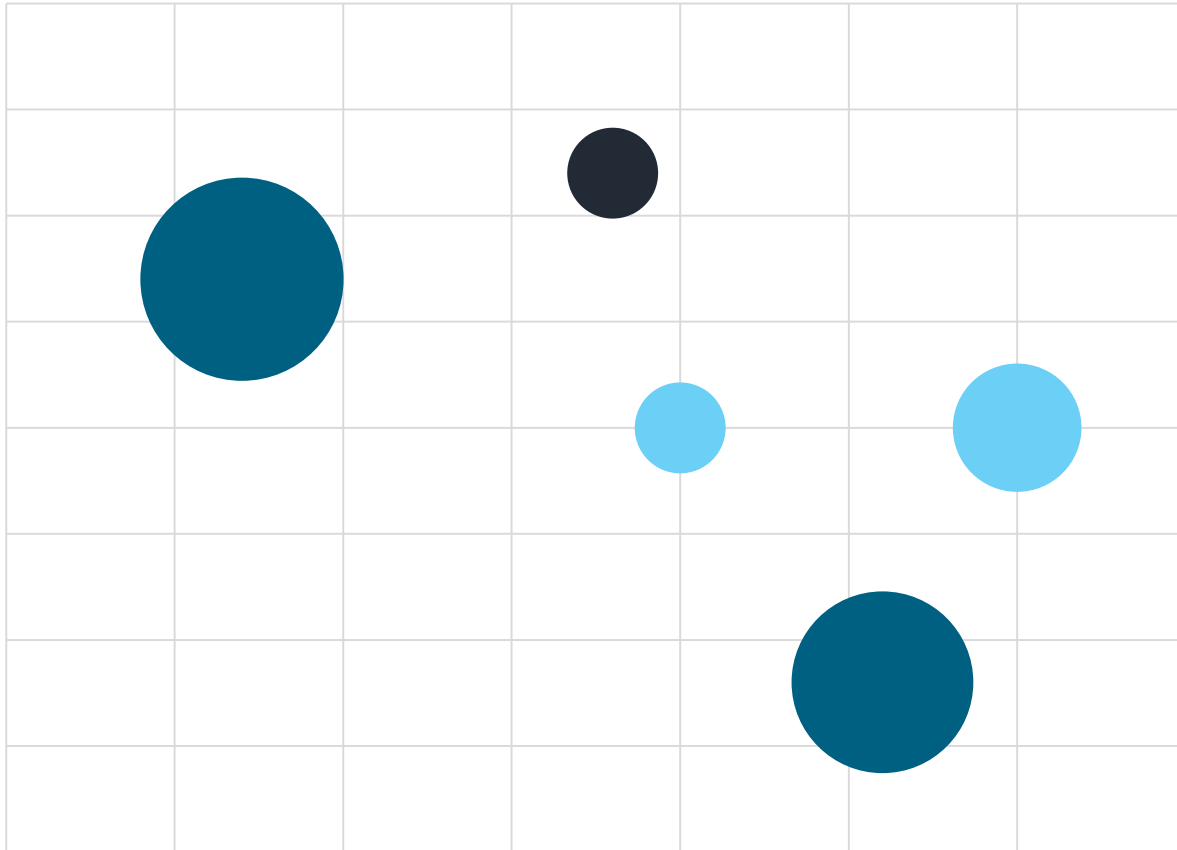
So sánh tổng số theo thời gian

Chiều thông tin:

- Trục x - Datetime hoặc Category
- Trục y - Value
- Màu sắc - Category



## Một số biểu đồ cơ bản



### Scatter Plot

**Biểu diễn mật độ phân bố**

**Thông tin:**

- Trục x - Value hoặc Category
- Trục y - Value hoặc Category
- Kích thước điểm - Value
- Màu sắc - Category





**DISCUSSION**  
**Thank you for your attention!**

---