


# Разработка на системи, базирани на големи данни

---

Ива Кръстева

Институт GATE



# Съдържание

- Големи данни – характеристики
  - Управление на големи данни
  - Анализ на големи данни
  - Системи, базирани на големи данни
  - Предизвикателство
-

# Предизвикателство



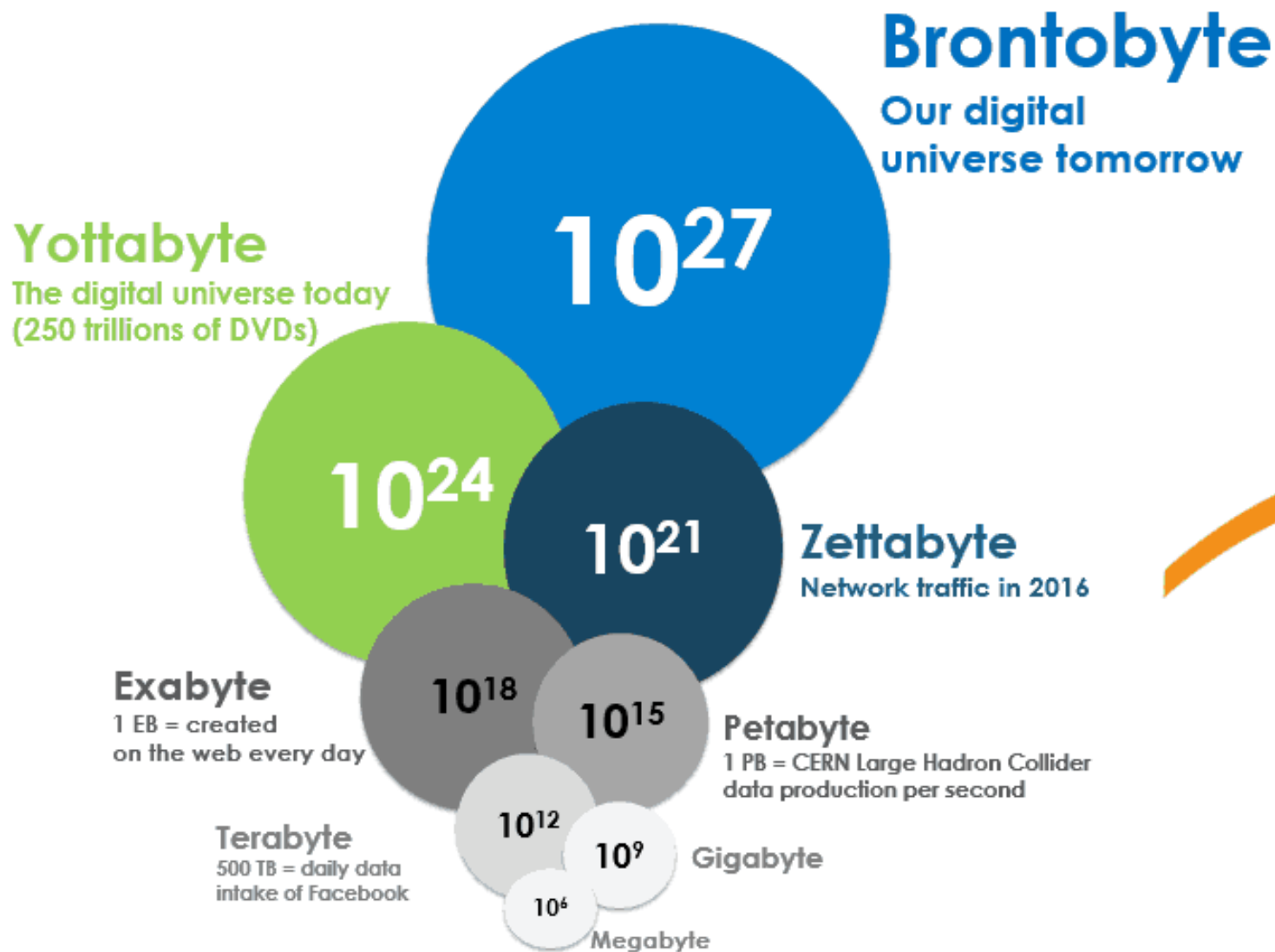
- Кои са основните характеристики на големите данни?
- Какви източници на поточни данни съществуват?
- Какви типове големи данни разпознаваме?
- Какви видове анализи върху данни извършваме?
- Какви примери за прогностичен анализ можем да дадем?

# 5 V's of big data



# Характеристики на големите данни: Обем

---



# Характеристики на големите данни: Скорост

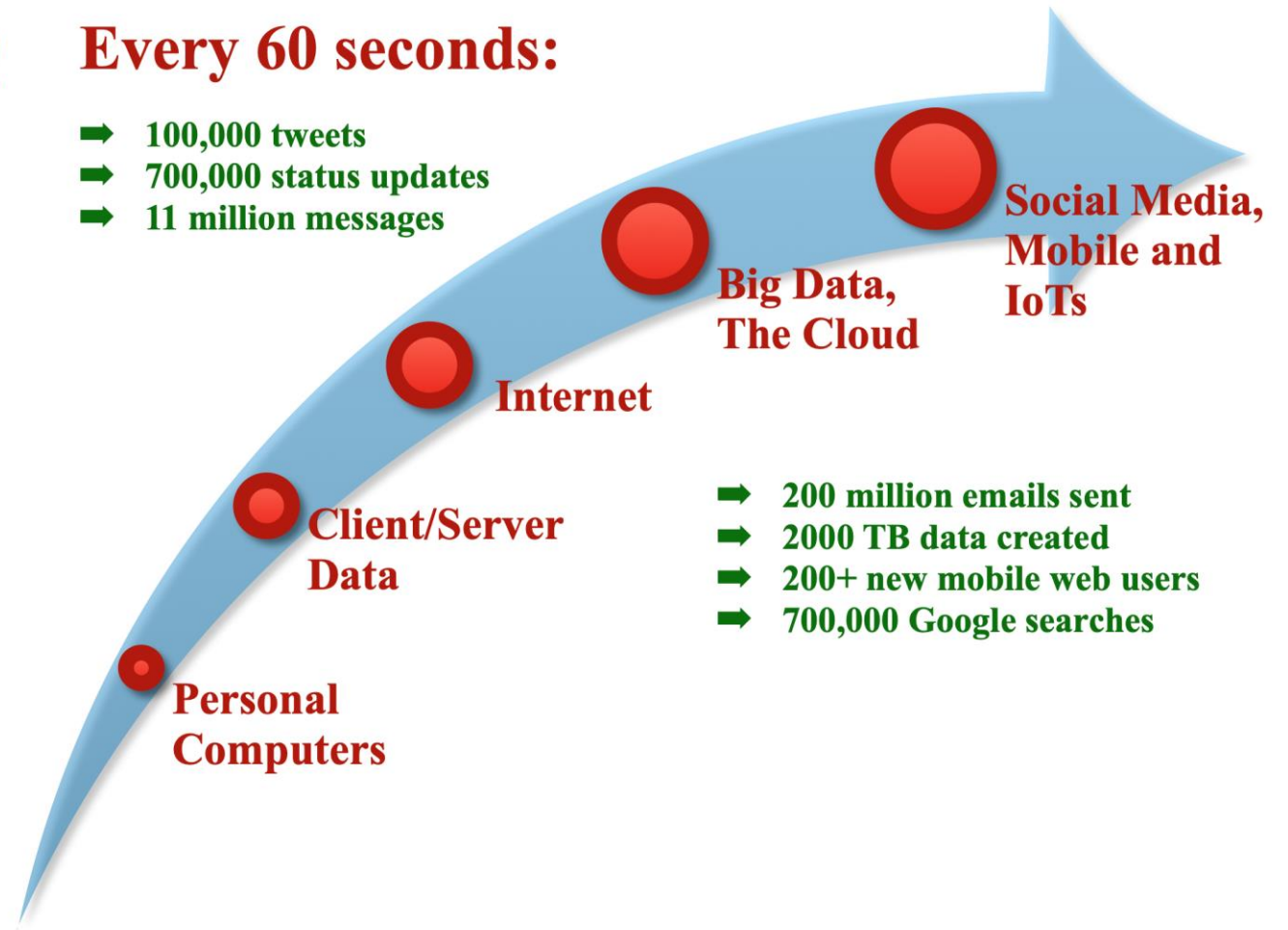
---

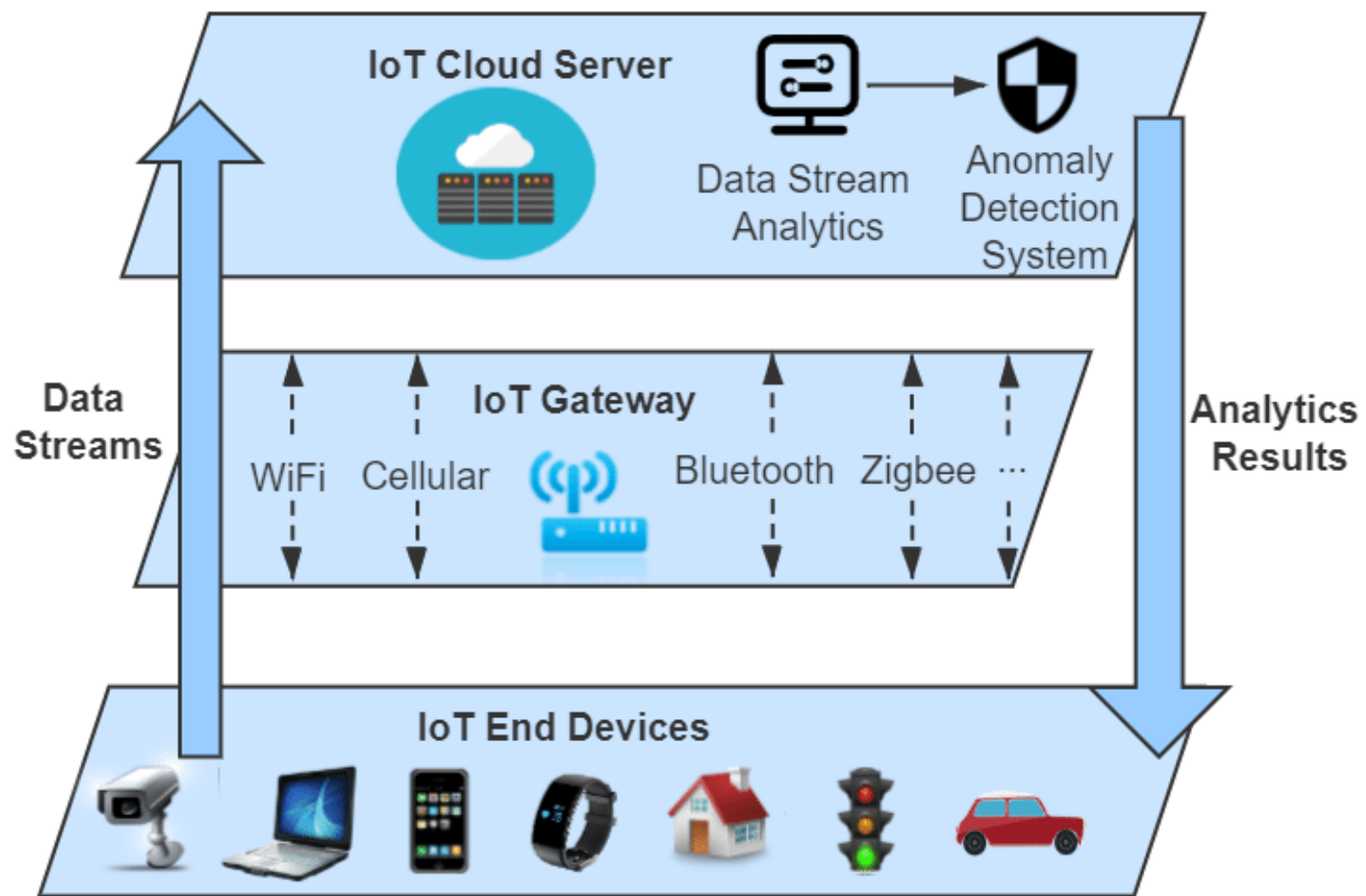
- Скорост (Velocity)
  - Streaming data
  - Batch data



## Every 60 seconds:

- ➔ 100,000 tweets
- ➔ 700,000 status updates
- ➔ 11 million messages



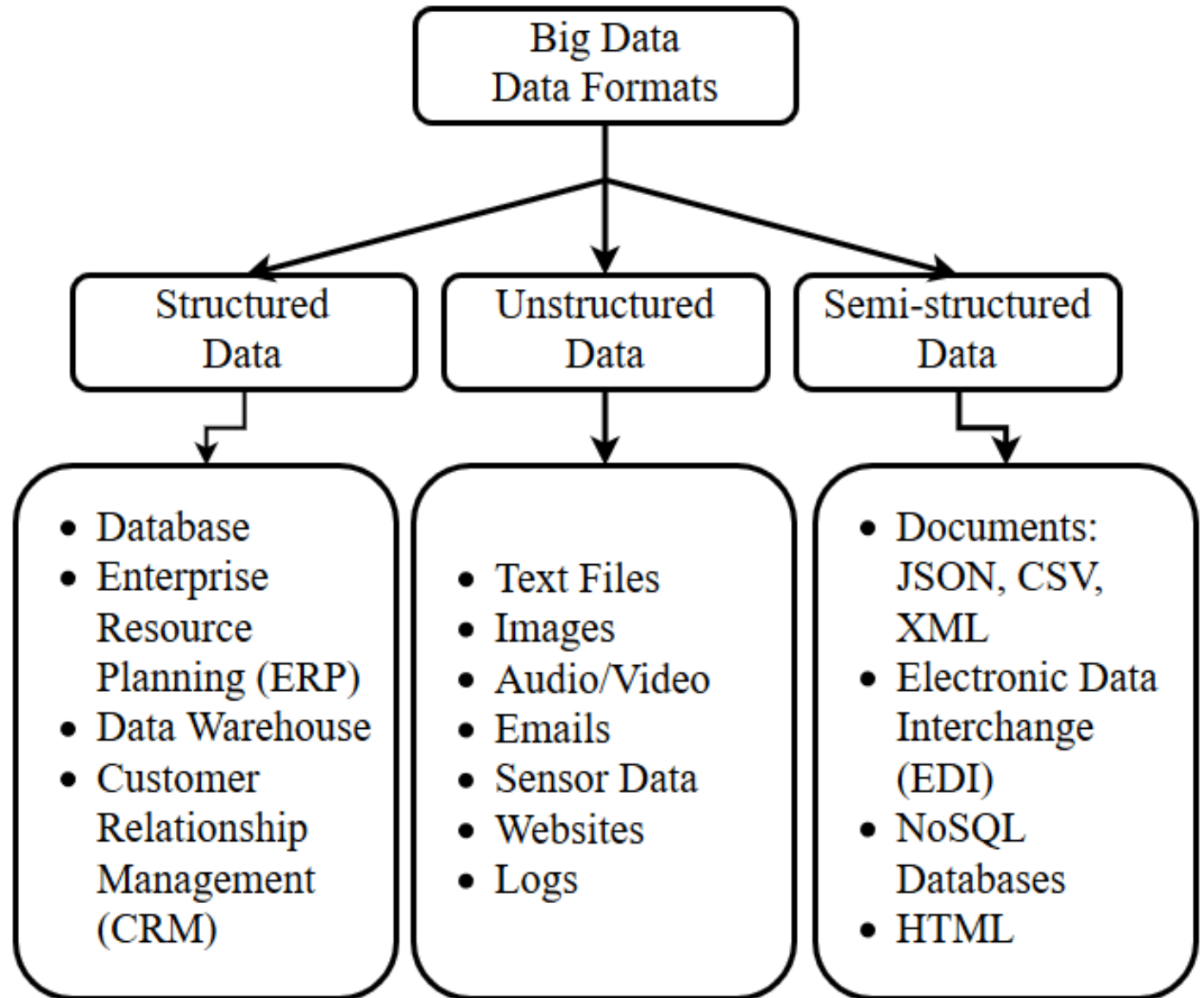


IoT  
Поточни  
Данни

## Характеристики на големите данни: Разнообразие

---

- Разнообразие (Variety)
  - Структурирани данни
  - Полу-структурирани данни
  - Неструктурирани данни





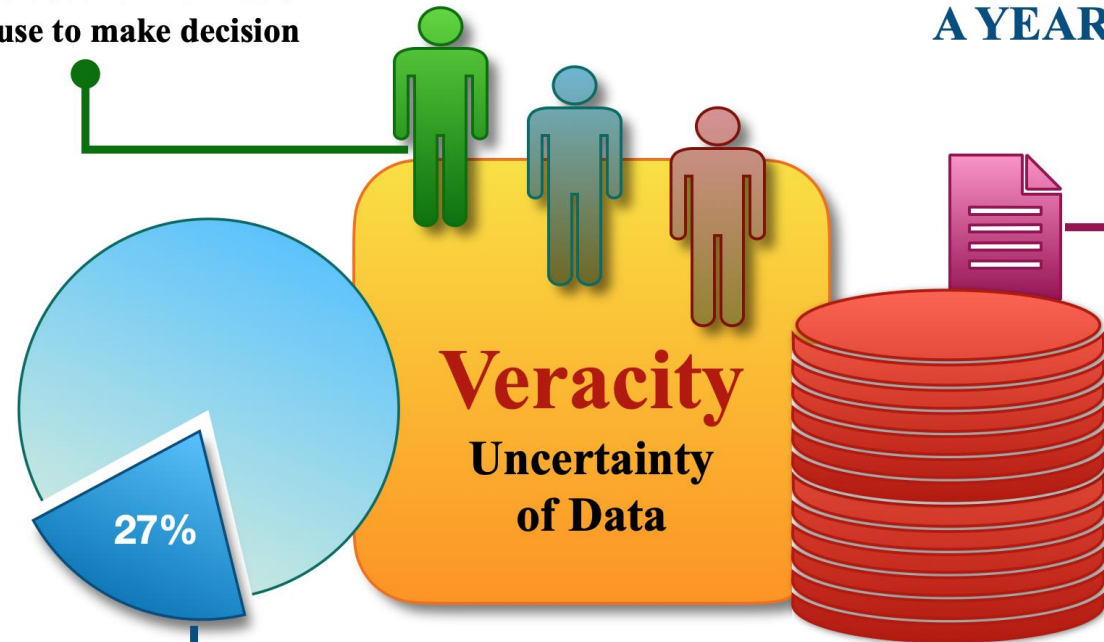
# Характеристики на големите данни: Достоверност

---

- Достоверност (Veracity)
  - Произход
  - Наличие
  - Качество
  - Точност
- Качество на данните  
-> качество на  
моделите



**1 IN 3 BUSINESS  
LEADERS**  
Don't trust the information  
they use to make decision



**OF RESPONDENTS**  
in one survey were unsure of how much  
of their data was inaccurate

Poor data quality cost the  
U.S. economy around  
**\$3.1  
TRILLION  
A YEAR**

1

## Big data impact on savings and profits

Source: TechJury, Tractica, Entrepreneur, Grazziti



**\$1** TRILLION

Savings by businesses through IoT by 2020.



**\$1** BILLION

Saved by Netflix using big data to improve customer retention.



**8-10** %

Increased profits by businesses that use big data.



**\$119** BILLION

Big data global revenue by 2025.

Характеристики  
на големите  
данни: Стойност

## Top benefits of data analytics

Sources: Chicago Analytics Group

25%



faster innovation  
cycles

17%



improved business  
efficiencies

13%



more effective  
R&D

12%



better  
product/service

# Обработка на големи данни - предизвикателства

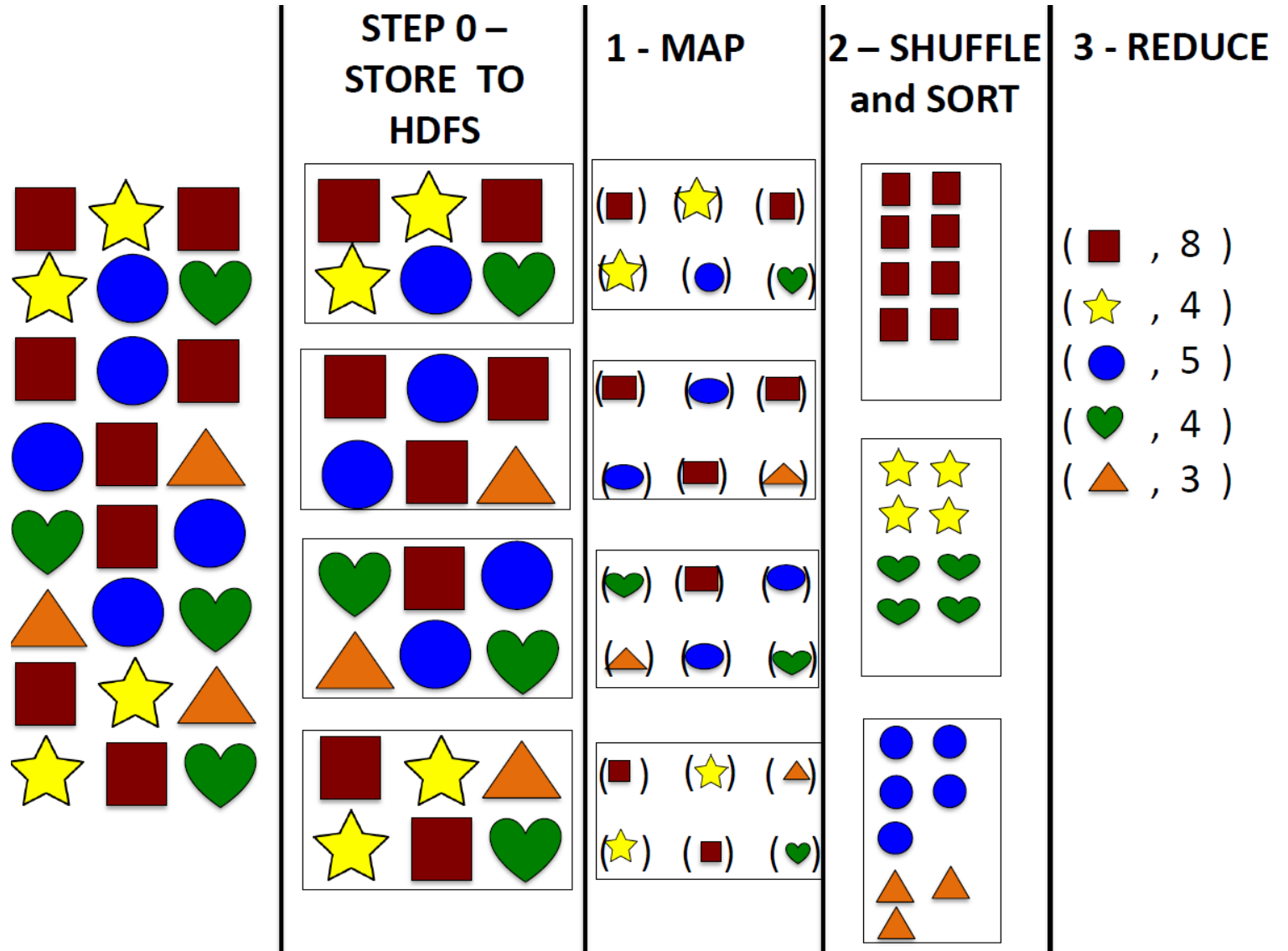


# Програмни модели за големи данни

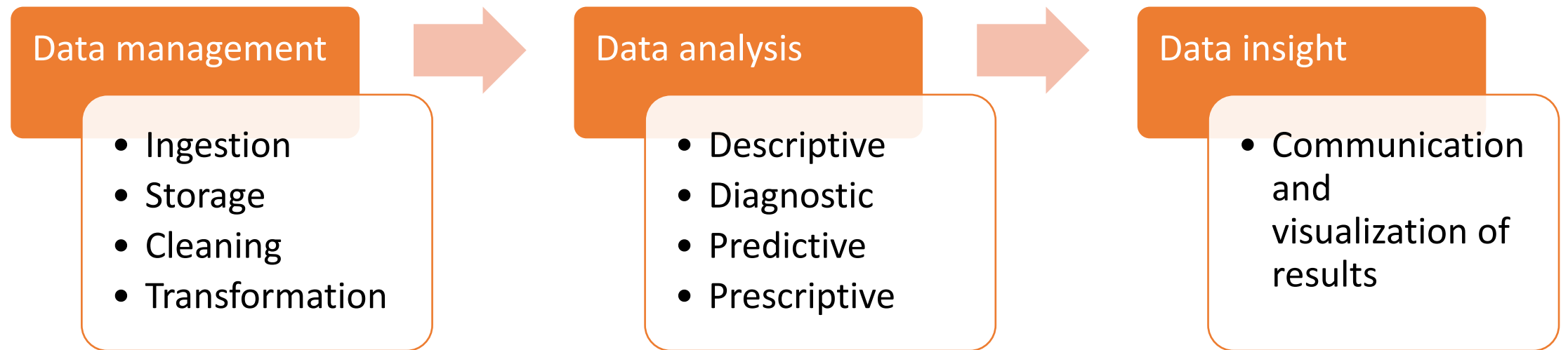
---

- Библиотеки + езици за обработка на големи данни
- Изисквания
  - Поддръжка на операции върху големи данни – бърз достъп до данните, разпределена обработка, работа с разпределени системи
  - Поддръжка на възстановяване от срыв и репликация
  - Възможност за скалиране на архитектурата
  - Оптимизирани за определени типове – документи, таблици, граф-бази, поточни данни...

# MapReduce



# Big Data Value Chain



# Управление на големи данни

Data ingestion – транспортиране на данните от източника до получателя

- batch data ingestion – на определен интервал или при промяна
- streaming data ingestion – постоянно транспортиране

Data exploration – изследване на параметрите и основни характеристики на данните

- Големина на пакета, брой колони, редове
- Типове данни, разпределение, диапазон
- Визуализация на данните
- Изследване на зависимости

# Управление на големи данни

## Data cleaning – откриване и премахване на аномалии

- Липсващи стойности
- Синтактично некоректни данни
- Неконсистентни данни
- Отклонения (outliers)

## Data transformation - трансформиране на данните в подходящ формат за обработка

- Нормализация и стандартизация на данните



# Управление на големи данни

## Data integration – интегриране на данни от различни източници

- Конфликти в семантиката на данните
- Разлики в описанието/представянето на еднакви данни
- Разлика в структурата на данните

## Data validation

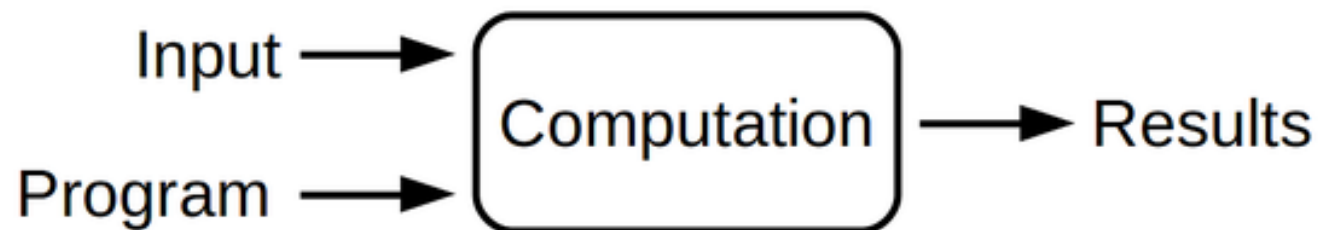
- Accuracy
- Completeness
- Uniqueness
- Consistency
- Timeliness

# Анализ на големи данни

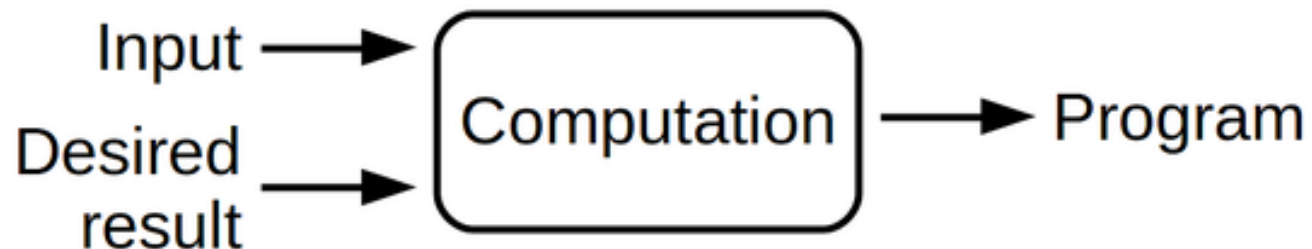
---

- Използване на анализа на данни
  - Подпомага взимането на бизнес решение
  - Компонент на система
- Машинно обучение <-> традиционно програмиране

## Traditional programming

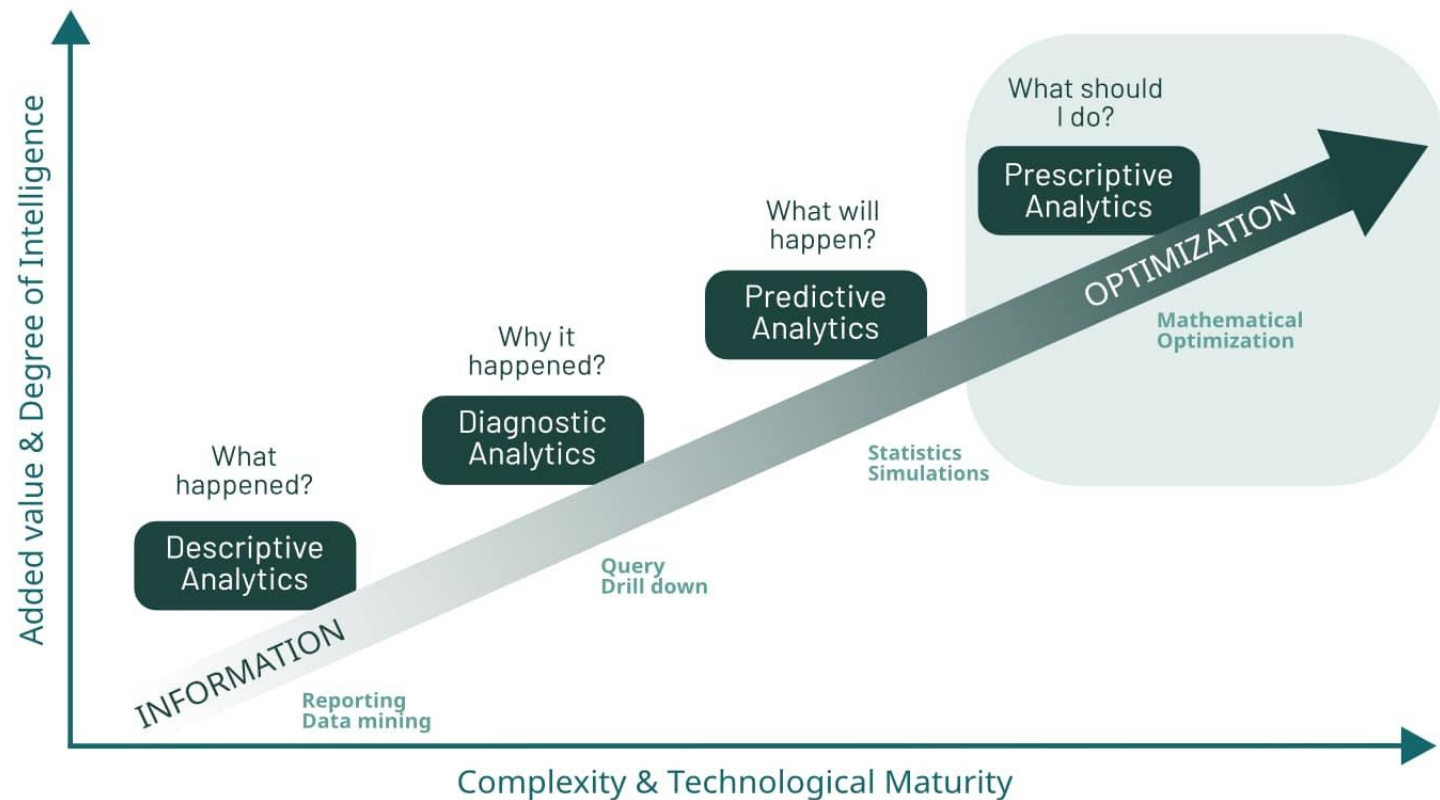


## Machine learning

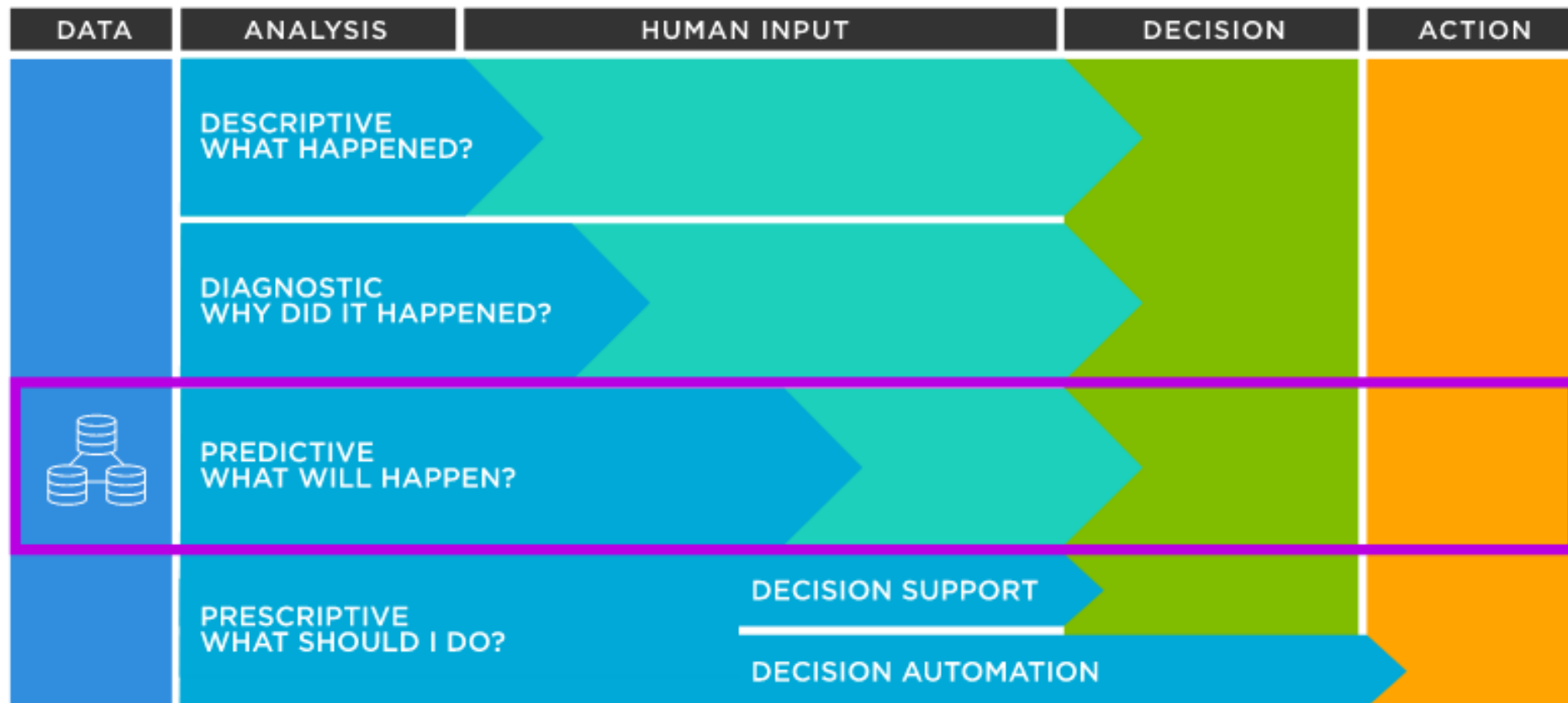


## Видове анализи върху данни

- Описателен анализ
- Диагностичен анализ
- Прогностичен анализ (Прогнозиране)
- Предписващ анализ



# Видове анализи върху данни



# Елементи на системите, базирани на големи данни



# Системи, базирани на големи данни - примери

Препоръчване на  
продукти

Планиране на  
складови  
наличности

Adaptive cruise  
control in a car

Групиране на  
играчи по умения

Модни тенденции  
на база на постове  
в социалните  
медии

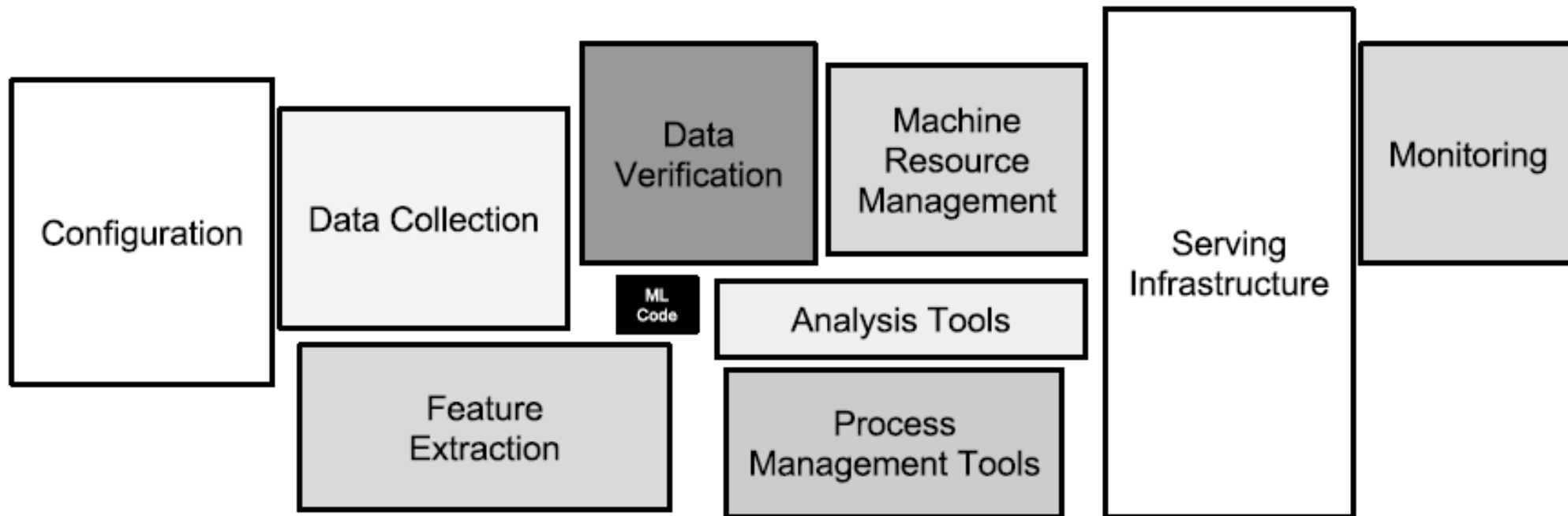
Проследяване и  
предсказване на  
нива на  
заболяване

Предписания за  
отстраняване на  
мрежови  
проблеми

...

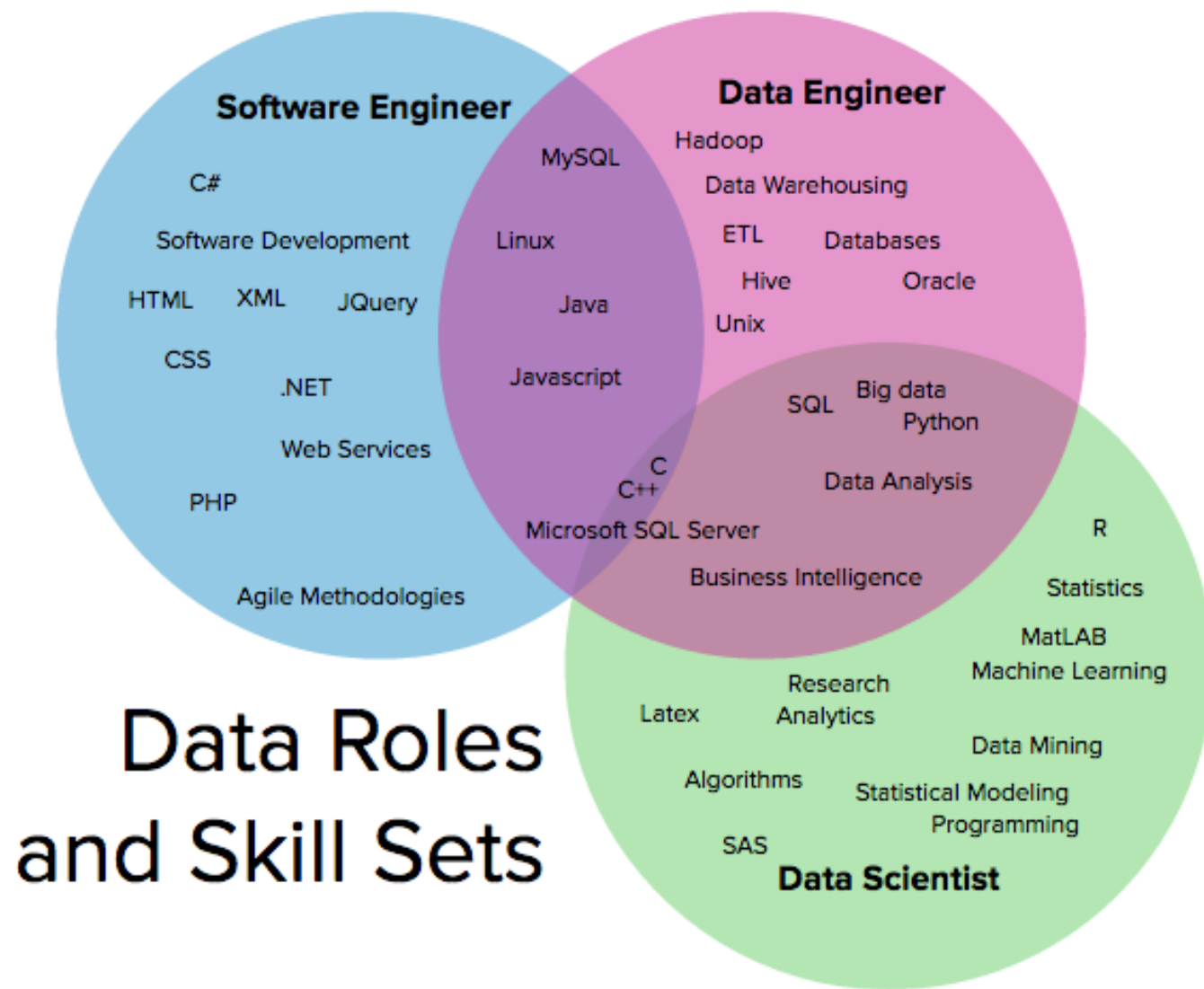
# Системи, базирани на големи данни - КОМПОНЕНТИ

---



Системи,  
базирани на  
големи  
данни—  
умения

---





# Анализ на изискванията - предизвикателства

## Цели за анализ <-> функционални изисквания

- Свойства на моделите
- Оценката на модел зависи от качеството и достоверността на данните

## Explainability

- Обяснение на модела
- Обяснение на анализа

## Дискриминация

- Biased models

## Законови изисквания

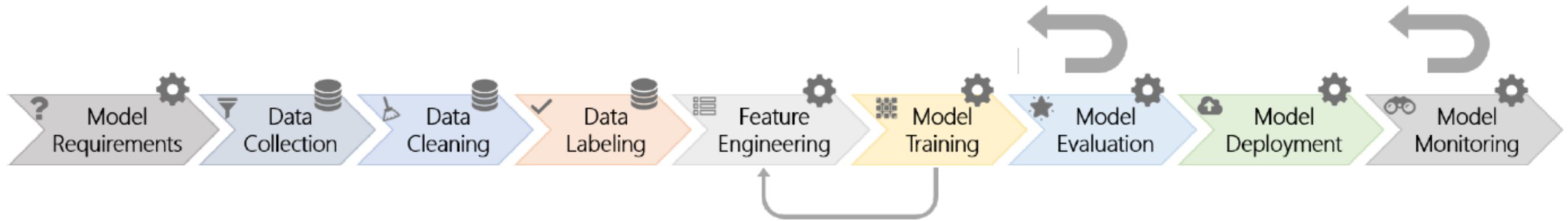
- Защита на данни

# Системи, базирани на големи данни - дизайн

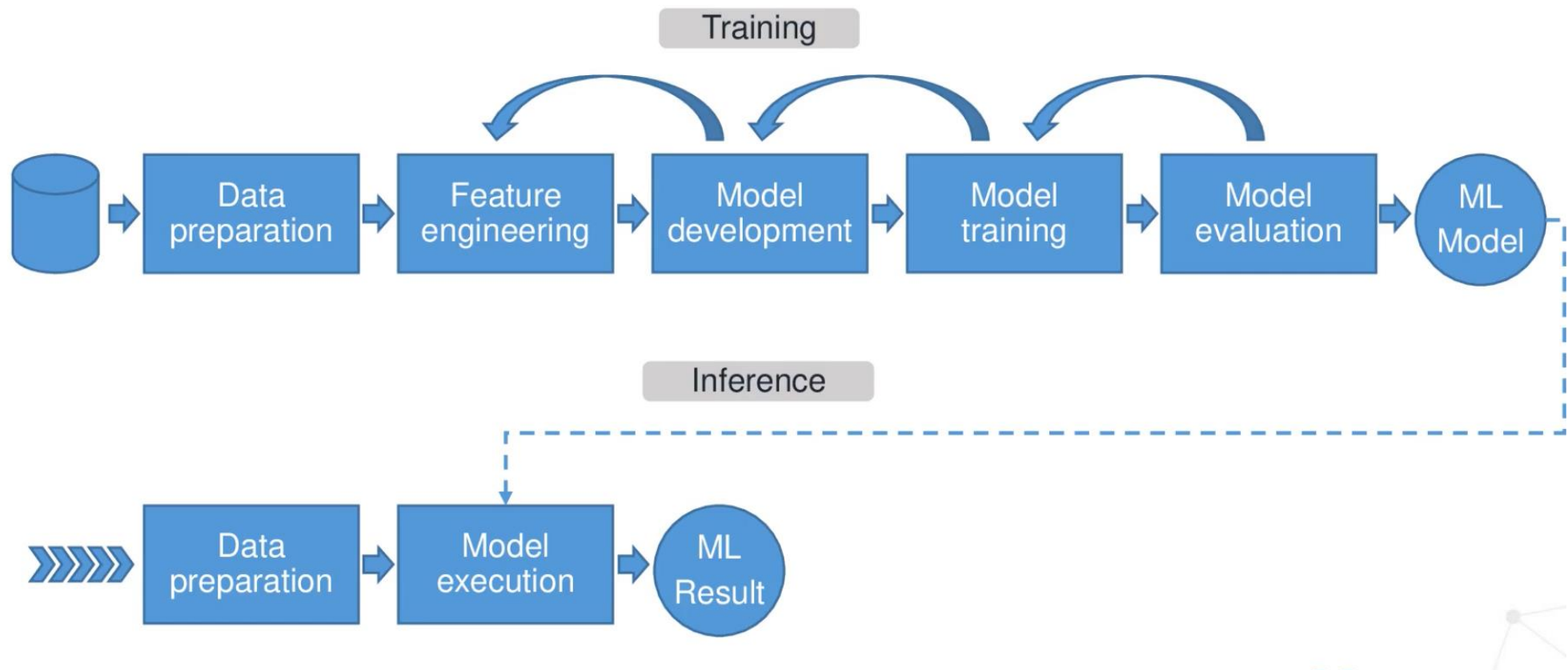
- Получаване на обратна връзка за работата на моделите
  - Колко често да се взаимодейства с потребителите
  - Каква е стойността на функционалността за потребителя
  - Каква е цената на неточно предсказване за потребителя
  - Колко често е неточно предсказването

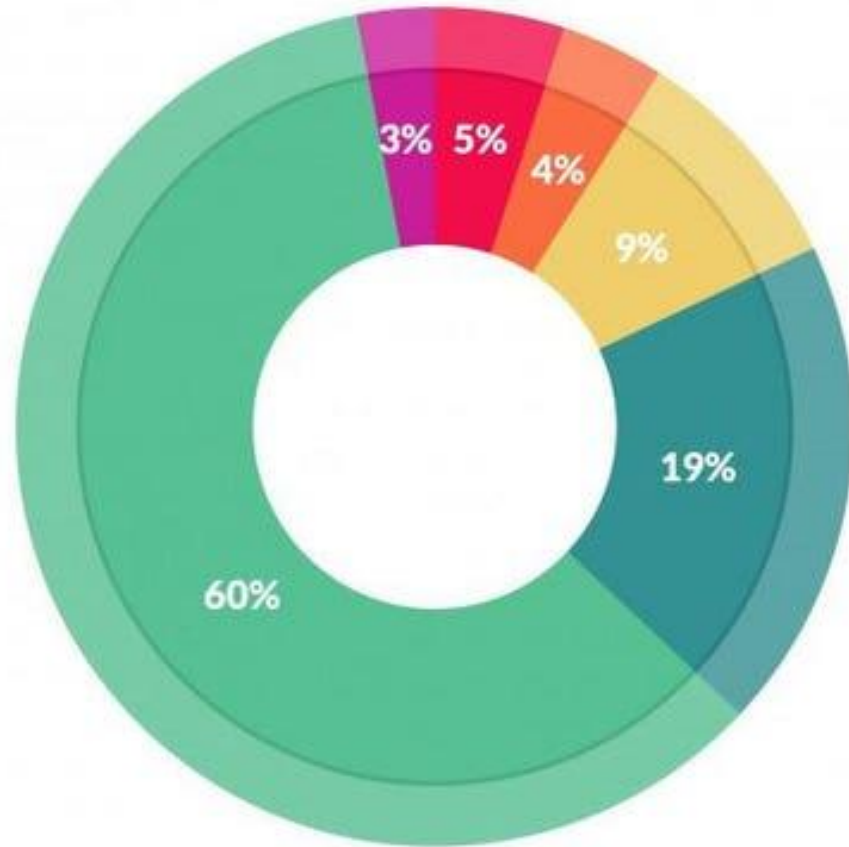
# Процес на създаване на модел за анализ на данни

---



# Трениране и изпълнение на модел





### What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

# Множества от данни



# Системи, базирани на големи данни – осигуряване на качеството



Качеството на моделите  $\leftrightarrow$  качеството на цялата система



"Good enough" quality



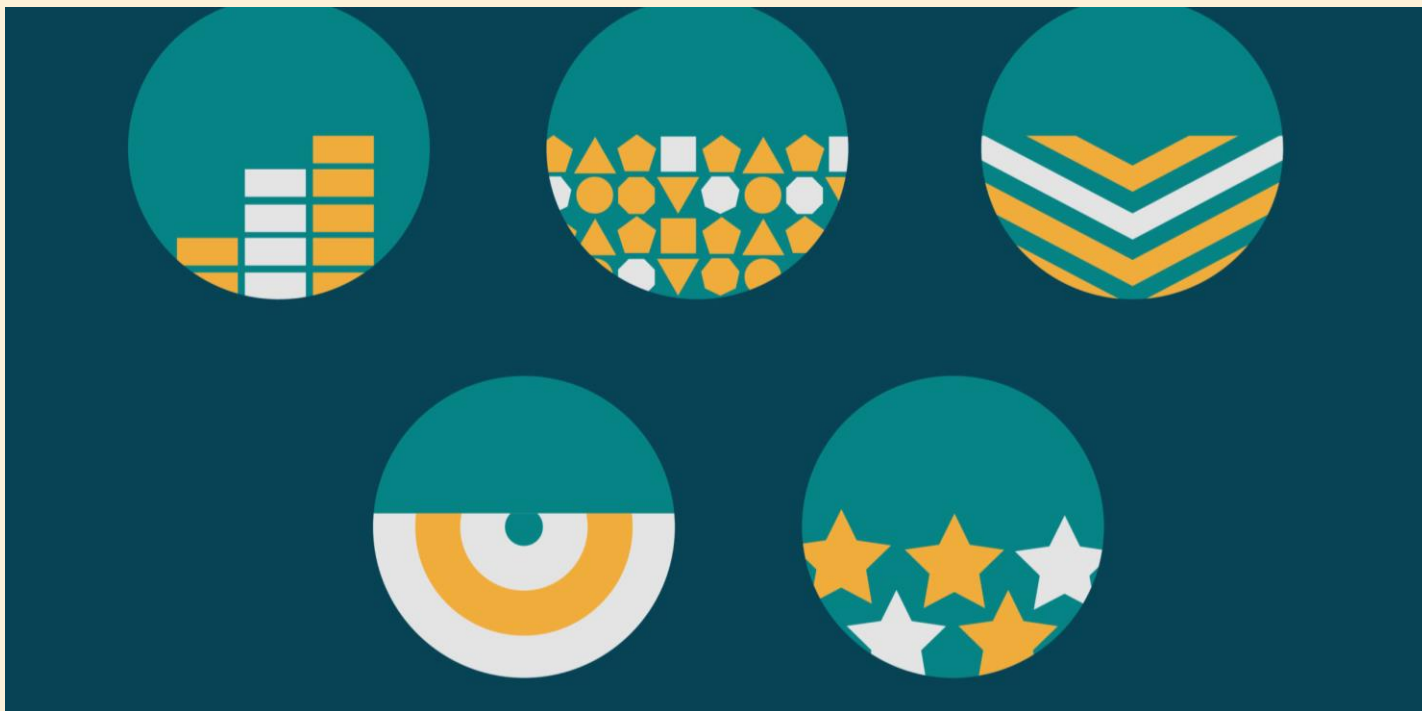
По-точни предсказания могат да имат много по-висока цена







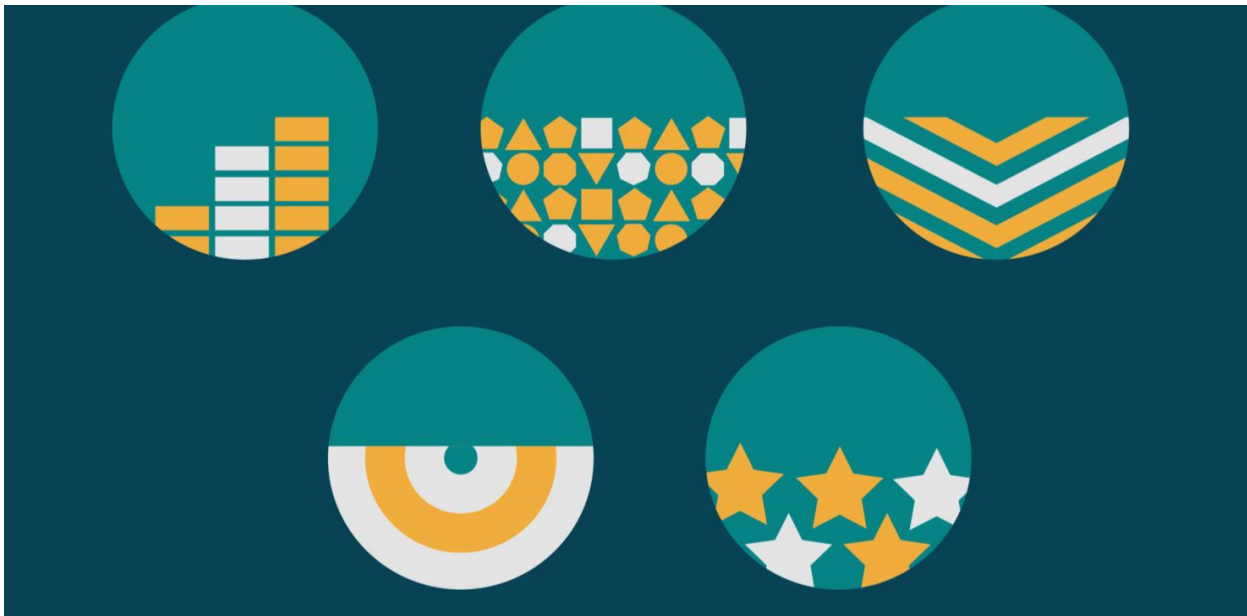
# Предизвикателство



- Кои са характеристиките на големите данни?
- Какви източници на поточни данни съществуват?

# Какви типове големи данни разпознаваме?

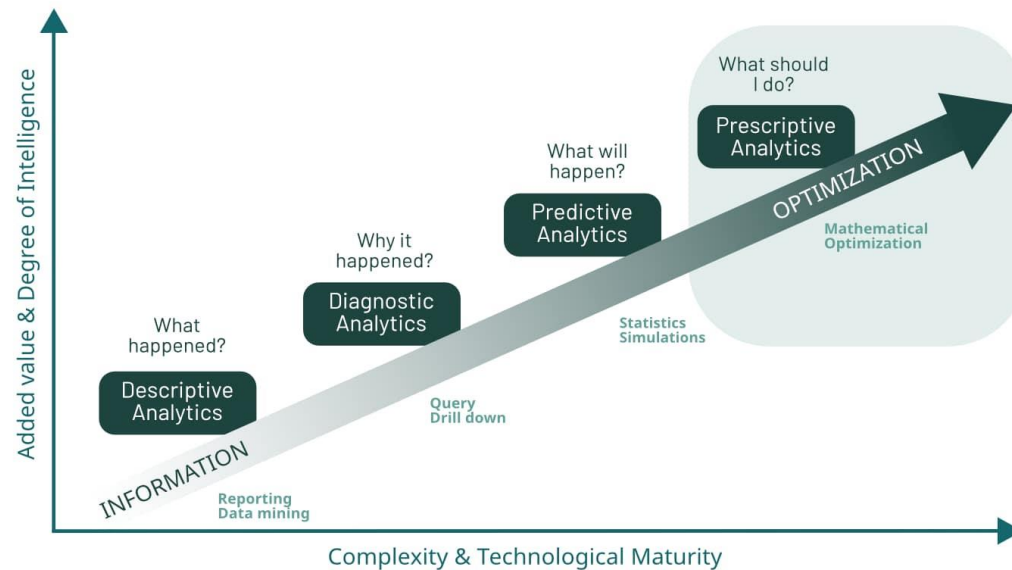
```
####<Oct 3, 2018 3:12:31,634 AM PDT> <Notice> <Server> <localhost> <AdminServer> <weblogic.socket.ServerListenThread>  
<20bd7f9b-0053-464a-8c9c-a8ce03b7e222-0000000d> <1538561551634> <[severity-value: 17] [rid: 12]  
<Channel "Default" is now listening on 192.0.2.254:7001 for protocols iiop, t3, ldap, snmp, http.>  
####<Oct 3, 2018 3:13:01,582 AM PDT> <Info> <Server> <localhost> <Application> <weblogic.socket.ApplicationEventHandler>  
<61cf3d4a-2158-379b-3e6f-b2cd46a6e432-0000000d> <1558561341536> <[severity-value: 38] [rid: 08]  
<Channel "Success" is now executing on 192.0.2.236:7002 for protocols http, t2, event handling from input.>
```



```
{  
  first_name: "John",  
  last_name: "Doe",  
  order_id: "769345",  
  order_total: "32.65"  
},  
{  
  first_name: "Mary",  
  last_name: "Moe",  
  order_id: "769458",  
  order_total: "58.43"  
}
```

First_Name	Last_Name	Order_Id	Order_Total
John	Doe	769345	32.65
Mary	Moe	769458	58.43

# Предизвикателство



- Какви видове анализи върху данни извършваме?
- Какви примери за прогностичен анализ можем да дадем?

The background of the image is a close-up, shallow depth-of-field shot of a large number of wooden question marks. The question marks are made of light-colored wood and are scattered across the frame, some in sharp focus and others blurred. A horizontal white bar is positioned across the lower third of the image, containing the text.

Въпроси