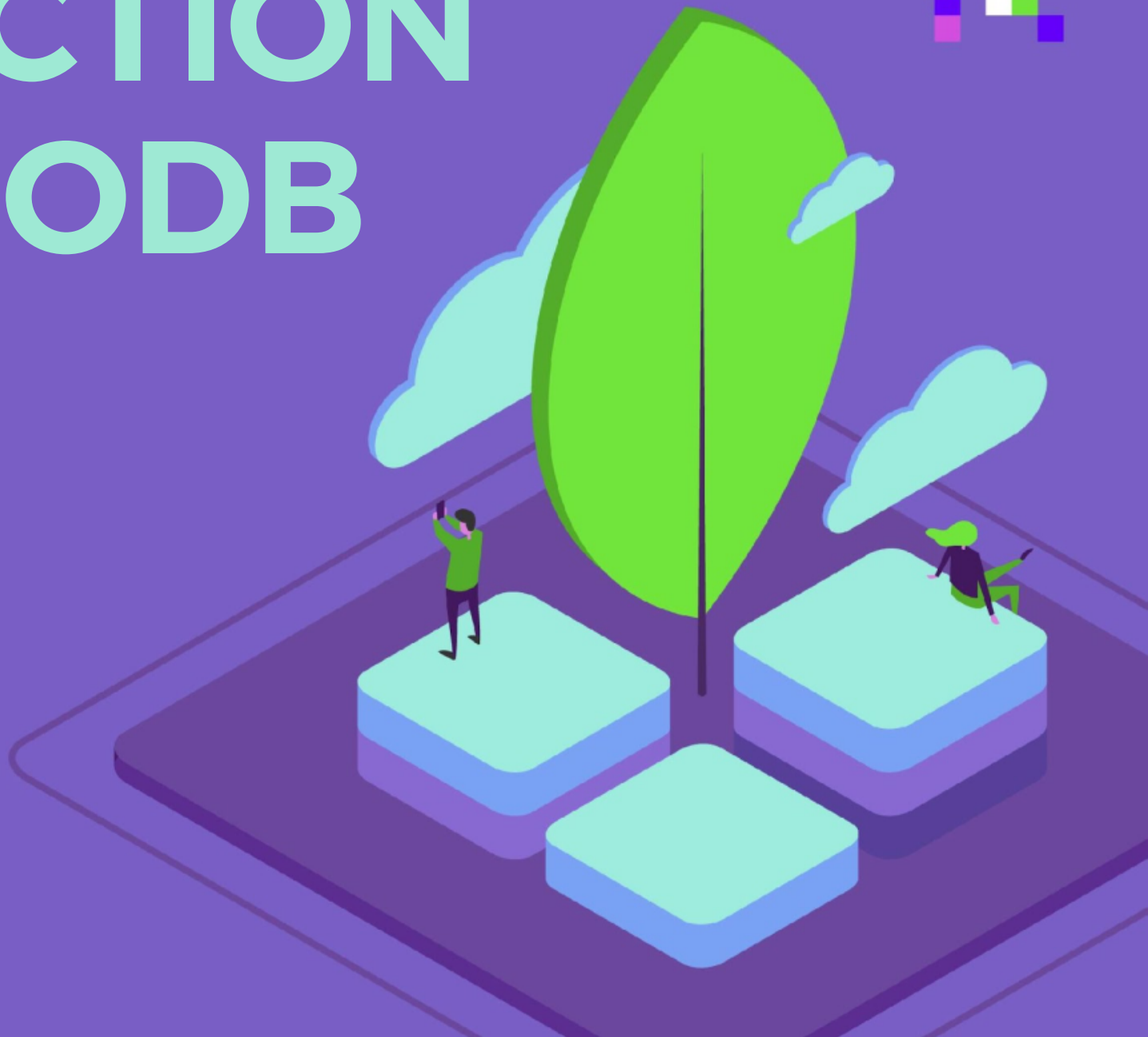
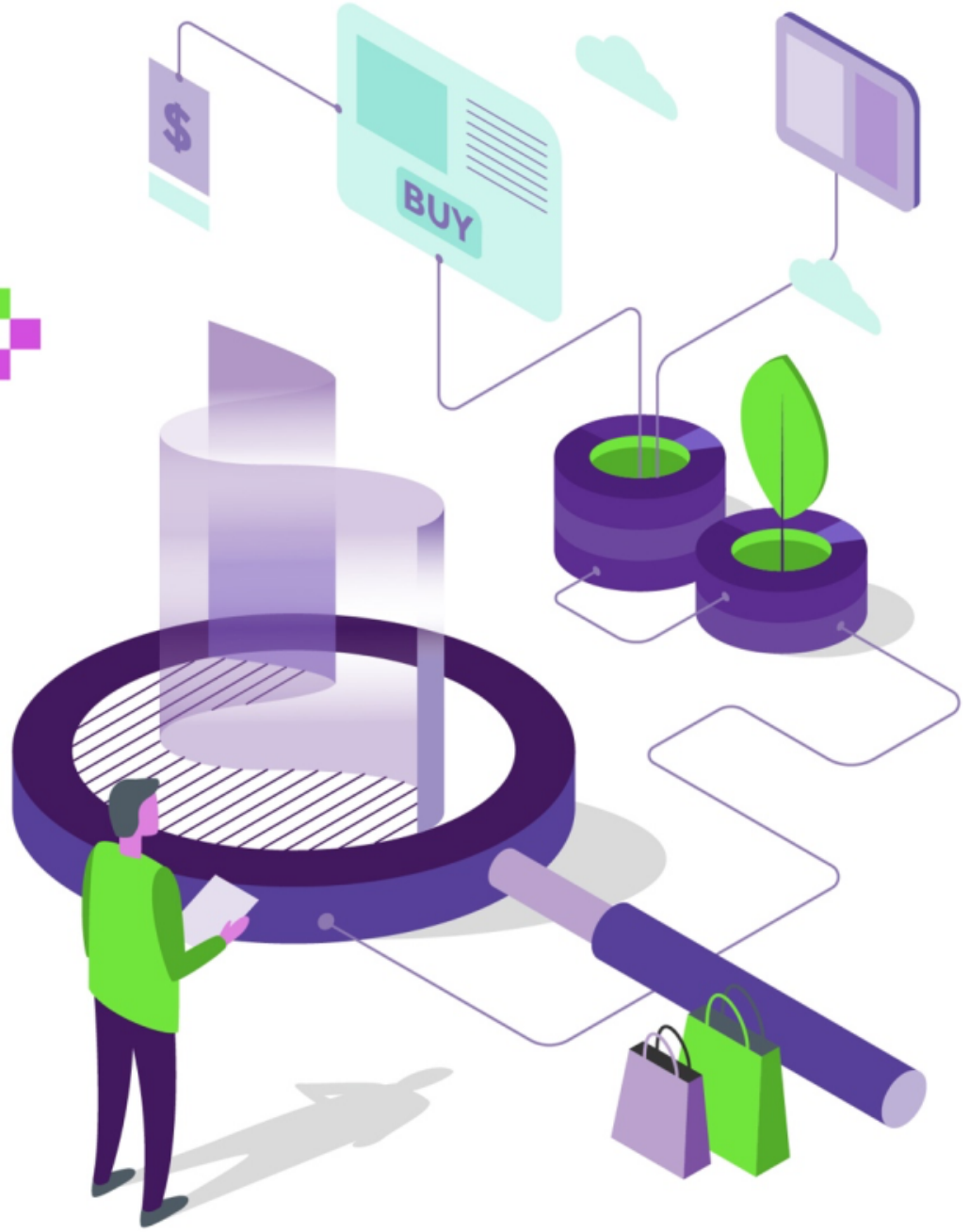
INTRODUCTION TO MONGODB



DATA Analysis



Analysis Techniques

Projection:

```
_id: ObjectId("5c9fee8966d8c1392df10eaf")
post_id: 3
user_id: 42
body: "Nullam a viverra magna"
topic: "politics"
likes: 403
dislikes: 293
views: 790
date_created: 2018-08-22T11:06:40.000+00:00
```

```
_id: ObjectId("5c9fee8966d8c1392df10ead")
post_id: 1
user_id: 2
body: "Aliquam eget suscipit odio"
topic: "sports"
likes: 168
dislikes: 341
views: 598
date_created: 2018-05-07T17:10:12.000+00:00
```

Analysis Techniques

Projection:

```
_id: ObjectId("5c9fee8966d8c1392df10eaf")
topic: "politics"
likes: 403
dislikes: 293
views: 790
date_created: 2018-08-22T11:06:40.000+00:00
```

```
_id: ObjectId("5c9fee8966d8c1392df10ead")
topic: "sports"
likes: 168
dislikes: 341
views: 598
date_created: 2018-05-07T17:10:12.000+00:00
```

Analysis Techniques:

Exploration

- Goal: Learn about how our documents are distributed with respect to each attribute that we're focused on

- Different approaches for Categorical vs. Quantitative attributes

Analysis Techniques:

Exploration: Categorical Data

- Query for each distinct category
  - Find the number of documents in each category
  - Find the percentage of documents in each category

- Percentage:

$$\frac{\text{Amount in one category}}{\text{Total amount}} \times 100$$

Analysis Techniques:

Exploration: Categorical Data

Example:

- 2000 blog posts
- 379 have a topic of sports

- 18.95% of posts have a topic of sports

```
_id: ObjectId("5c9fee8966d8c1392df10ead")
topic: "sports"
likes: 168
dislikes: 341
views: 598
date_created: 2018-05-07T17:10:12.000+00:00
```

$$\frac{379}{2000} \times 100 = 18.95\%$$

Analysis Techniques:

Exploration: Quantitative Data

- For each quantitative attribute:
  - Use sorting to find the min and max values
  - Calculate the midpoint value
  - Query for documents above and below the midpoint
    - Find the number
    - Find the percentage

Midpoint:

$$\frac{\text{minimum} + \text{maximum}}{2}$$

Analysis Techniques:

Exploration: Quantitative Data

```
_id: ObjectId("5c9fee8966d8c1392df10eaf")
topic: "politics"
likes: 403
dislikes: 293
views: 790
date_created: 2018-08-22T11:06:40.000+00:00
```

Example:

- All blog posts have a “likes” field
- Minimum amount of likes on any post is 4
- Maximum amount is 528
- Midpoint is 266

$$\frac{4 + 528}{2} = 266$$

Analysis Techniques:

Exploration: Quantitative Data

- Using the midpoint as a basis for exploration can work well when:
  - There are enough documents above and below the midpoint to represent a significant portion of the data

- Does not work well when:
  - There are significant outliers in the dataset

Analysis Techniques:

Exploration: Quantitative Data

- Example:
  - 3,500 purchases in total
  - Minimum Unit Price of 0.5
  - Maximum Unit Price of 165
  - Midpoint is 82.75
  - 3,496 purchases below the midpoint (99.89%)
  - 4 purchases above the midpoint (0.11%)


```
_id: ObjectId("5ca2ecd50dadcc5e8fccfd21")
InvoiceNo: "536392"
StockCode: "22827"
Description: "RUSTIC  SEVENTEEN DRAWER SIDEBOARD"
Quantity: 1
InvoiceDate: 2010-01-12T10:29:00.000+00:00
UnitPrice: 165
CustomerID: "13705"
Country: "United Kingdom"
```


Reason for this:

- Purchase with the highest unit price is an outlier