

# EXPLORATION OF HIGH-DIMENSIONAL DATA

EXPLAINABLE STRUCTURING

DISCOVERY OF RELEVANT CASES

---

**Joris Falip**, Frédéric Blanchard, Michel Herbin

[joris.falip@univ-reims.fr](mailto:joris.falip@univ-reims.fr)

[www.datacrunch.sh](http://www.datacrunch.sh)

March 20 2019

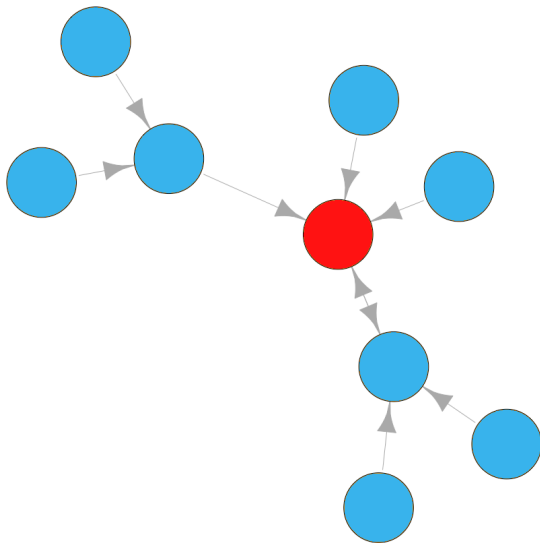
CRéSTIC, University of Reims Champagne-Ardenne, France

Help domain experts to visualize, explore and manipulate data

## Approach

- Leverage analogical reasoning
- Structure data according to their similarity
- Visually interpretable results

# OVERVIEW



## Definition

"Classification decisions are based on the similarity of stimuli to the stored exemplars." - R. M. Nosofsky

- Rely on past experiences
- Allows rapid decision making
- Can be applied to complex objects

# CONSTRAINTS

1. **Explainable** algorithm and **interpretable** results
2. Efficient with high-dimensionality

1. **Explainable** algorithm and **interpretable** results
2. Efficient with high-dimensionality

## Curse of dimensionality

"Having more dimensions usually means data sets tend to be sparse and all distances between data points tend to become harder to distinguish." - N. Tomasev et al.

# CONSTRAINTS

1. **Explainable** algorithm and **interpretable** results
2. Efficient with high-dimensionality

## Curse of dimensionality

"Having more dimensions usually means data sets tend to be sparse and all distances between data points tend to become harder to distinguish." - N. Tomasev et al.

No a priori knowledge → no dimensionality reduction

## Hubness

"High dimensionality causes some data points to be the nearest neighbor of many others points, thus becoming 'hubs'." - totally made up definition

Hubs exhibit interesting properties and can be used to improve traditional data mining approaches

- Compute on each dimension then aggregate
- Use ranks to avoid sparsity and outliers' exclusion

## Steps

1. Ranking individuals on each dimension
2. Aggregate the results into a score : the Degree of Representativeness
3. Link individuals to their neighbor with the highest DoR

# 1. RANKING INDIVIDUALS ON EACH DIMENSION

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |

# 1. RANKING INDIVIDUALS ON EACH DIMENSION

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |



| Feature 2 - ranking |      |      |     |
|---------------------|------|------|-----|
| Feature 1 - ranking |      |      |     |
|                     | Ind1 | Ind2 | ... |
|                     |      |      |     |
|                     |      |      |     |
|                     |      |      |     |
|                     |      |      |     |

# 1. RANKING INDIVIDUALS ON EACH DIMENSION

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |



| Feature 2 - ranking |      |      |     |
|---------------------|------|------|-----|
| Feature 1 - ranking |      |      |     |
|                     | Ind1 | Ind2 | ... |
| Ind1                | 1st  | 4th  | ... |
|                     |      |      |     |
|                     |      |      |     |

# 1. RANKING INDIVIDUALS ON EACH DIMENSION

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |



| Feature 2 - ranking |      |      |     |
|---------------------|------|------|-----|
| Feature 1 - ranking |      |      |     |
|                     | Ind1 | Ind2 | ... |
| Ind1                | 1st  | 4th  | ... |
| Ind2                | 2nd  | 1st  | ... |
| ...                 | ...  | ...  | ... |

# 1. RANKING INDIVIDUALS ON EACH DIMENSION

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |



| Feature 2 - ranking |      |      |     |
|---------------------|------|------|-----|
| Feature 1 - ranking |      |      |     |
|                     | Ind1 | Ind2 | ... |
| Ind1                | 1st  | 4th  | ... |
| Ind2                | 2nd  | 1st  | ... |
| ...                 | ...  | ...  | ... |

## 2. AGGREGATE THE RESULTS INTO A SCORE

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |



| Feature 2 - ranking<br>Feature 1 - ranking |      |      |     |
|--------------------------------------------|------|------|-----|
|                                            | Ind1 | Ind2 | ... |
| Ind1                                       | 1st  | 4th  | ... |
| Ind2                                       | 2nd  | 1st  | ... |
| ...                                        | ...  | ...  | ... |



| Feature 2 - scores<br>Feature 1 - scores |      |      |     |
|------------------------------------------|------|------|-----|
|                                          | Ind1 | Ind2 | ... |
| Ind1                                     |      |      | ... |
| Ind2                                     |      |      | ... |
| ...                                      | ...  | ...  | ... |

## 2. AGGREGATE THE RESULTS INTO A SCORE

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |



| Feature 2 - ranking |      |      |     |
|---------------------|------|------|-----|
| Feature 1 - ranking |      |      |     |
|                     | Ind1 | Ind2 | ... |
| Ind1                | 1st  | 4th  | ... |
| Ind2                | 2nd  | 1st  | ... |
| ...                 | ...  | ...  | ... |



| Feature 2 - scores |      |      |     |
|--------------------|------|------|-----|
| Feature 1 - scores |      |      |     |
|                    | Ind1 | Ind2 | ... |
| Ind1               | 100  |      | ... |
| Ind2               |      | 100  | ... |
| ...                | ...  | ...  | ... |

## 2. AGGREGATE THE RESULTS INTO A SCORE

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |



| Feature 2 - ranking<br>Feature 1 - ranking |      |      |     |
|--------------------------------------------|------|------|-----|
|                                            | Ind1 | Ind2 | ... |
| Ind1                                       | 1st  | 4th  | ... |
| Ind2                                       | 2nd  | 1st  | ... |
| ...                                        | ...  | ...  | ... |



| Feature 2 - scores<br>Feature 1 - scores |      |      |     |
|------------------------------------------|------|------|-----|
|                                          | Ind1 | Ind2 | ... |
| Ind1                                     | 100  |      | ... |
| Ind2                                     | 90   | 100  | ... |
| ...                                      | ...  | ...  | ... |

## 2. AGGREGATE THE RESULTS INTO A SCORE

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |



| Feature 2 - ranking<br>Feature 1 - ranking |      |      |     |
|--------------------------------------------|------|------|-----|
|                                            | Ind1 | Ind2 | ... |
| Ind1                                       | 1st  | 4th  | ... |
| Ind2                                       | 2nd  | 1st  | ... |
| ...                                        | ...  | ...  | ... |



| Feature 2 - scores<br>Feature 1 - scores |      |      |     |
|------------------------------------------|------|------|-----|
|                                          | Ind1 | Ind2 | ... |
| Ind1                                     | 100  | 75   | ... |
| Ind2                                     | 90   | 100  | ... |
| ...                                      | ...  | ...  | ... |

## 2. AGGREGATE THE RESULTS INTO A SCORE

| Dataset |         |         |     |
|---------|---------|---------|-----|
|         | Feat. 1 | Feat. 2 | ... |
| Ind1    | 67      | 173     | ... |
| Ind2    | 53      | 161     | ... |
| ...     | ...     | ...     | ... |

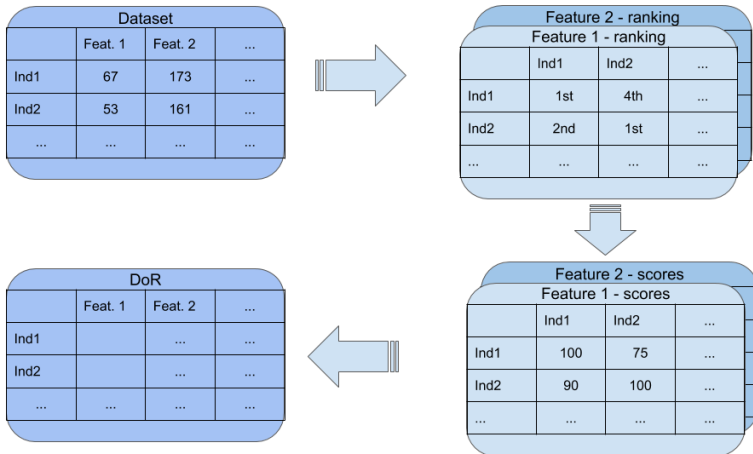


| Feature 2 - ranking<br>Feature 1 - ranking |      |      |     |
|--------------------------------------------|------|------|-----|
|                                            | Ind1 | Ind2 | ... |
| Ind1                                       | 1st  | 4th  | ... |
| Ind2                                       | 2nd  | 1st  | ... |
| ...                                        | ...  | ...  | ... |

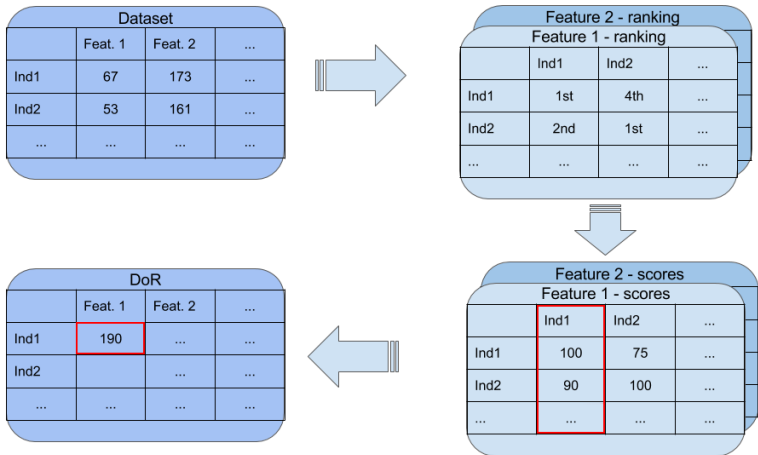


| Feature 2 - scores<br>Feature 1 - scores |      |      |     |
|------------------------------------------|------|------|-----|
|                                          | Ind1 | Ind2 | ... |
| Ind1                                     | 100  | 75   | ... |
| Ind2                                     | 90   | 100  | ... |
| ...                                      | ...  | ...  | ... |

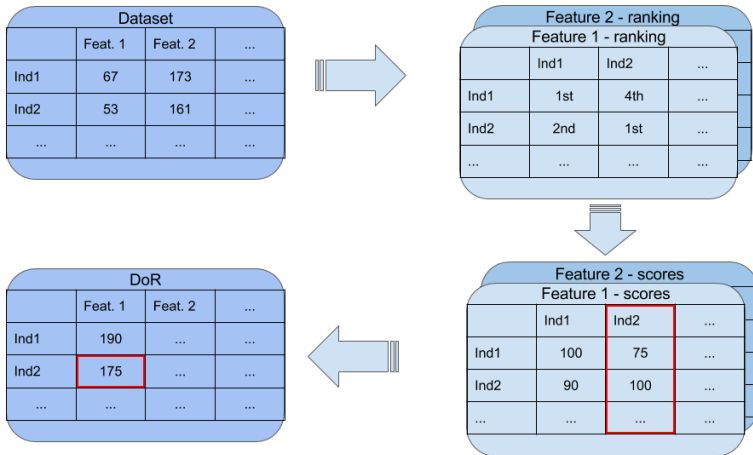
## 2. AGGREGATE THE RESULTS INTO A SCORE



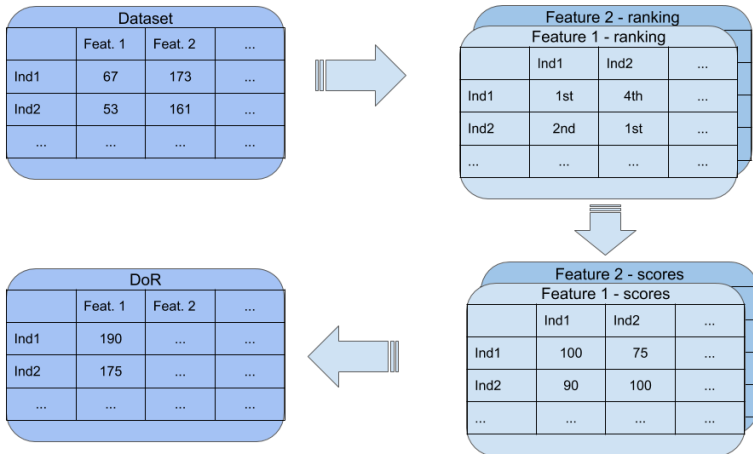
## 2. AGGREGATE THE RESULTS INTO A SCORE



## 2. AGGREGATE THE RESULTS INTO A SCORE



## 2. AGGREGATE THE RESULTS INTO A SCORE



### 3. LINK INDIVIDUALS TO THEIR NEIGHBOR WITH THE HIGHEST DOR

|      | Feat. 1 | Feat. 2 | Feat. 3 | Sum |
|------|---------|---------|---------|-----|
| Ind1 | 18      | 48      | 30      | 96  |
| Ind2 | 12      | 17      | 5       | 34  |
| Ind3 | 49      | 36      | 80      | 165 |
| Ind4 | 59      | 79      | 30      | 168 |

### 3. LINK INDIVIDUALS TO THEIR NEIGHBOR WITH THE HIGHEST DOR

|      | Feat. 1 | Feat. 2 | Feat. 3 | Sum |
|------|---------|---------|---------|-----|
| Ind1 | 18      | 48      | 30      | 96  |
| Ind2 | 12      | 17      | 5       | 34  |
| Ind3 | 49      | 36      | 80      | 165 |
| Ind4 | 59      | 79      | 30      | 168 |

### 3. LINK INDIVIDUALS TO THEIR NEIGHBOR WITH THE HIGHEST DOR

|      | Feat. 1 | Feat. 2 | Feat. 3 | Sum |
|------|---------|---------|---------|-----|
| Ind1 | 18      | 48      | 30      | 96  |
| Ind2 |         | 17      | 5       | 22  |
| Ind3 | 49      | 36      | 80      | 165 |
| Ind4 | 59      | 79      | 30      | 168 |

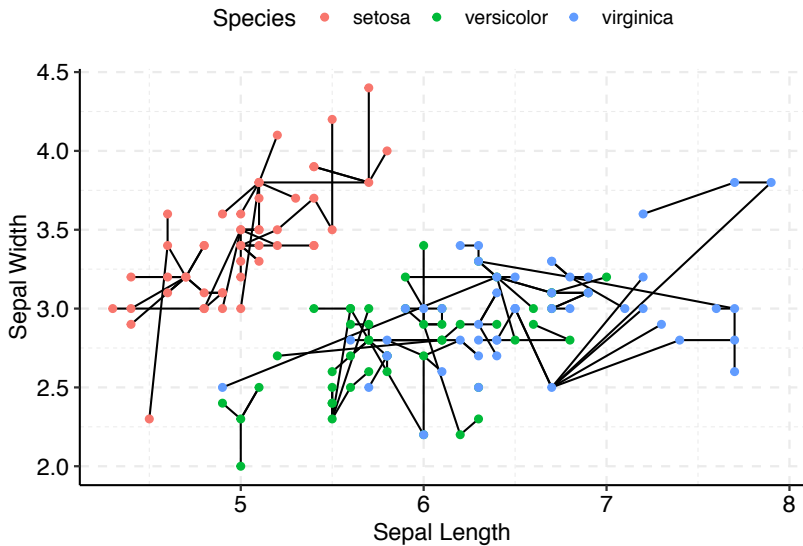
### 3. LINK INDIVIDUALS TO THEIR NEIGHBOR WITH THE HIGHEST DOR

|      | Feat. 1 | Feat. 2 | Feat. 3 | Sum |
|------|---------|---------|---------|-----|
| Ind1 | 18      | 48      | 30      | 96  |
| Ind2 |         |         | 5       | 5   |
| Ind3 | 49      | 36      | 80      | 165 |
| Ind4 | 59      | 79      | 30      | 168 |

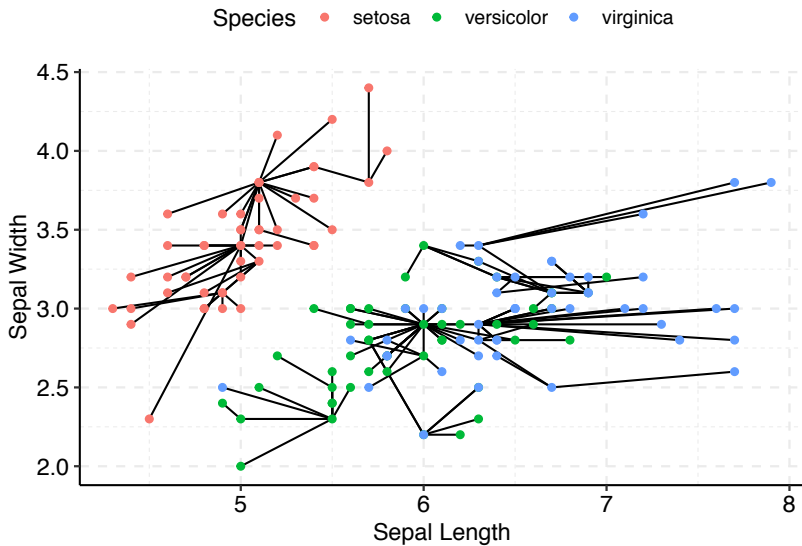
### 3. LINK INDIVIDUALS TO THEIR NEIGHBOR WITH THE HIGHEST DOR

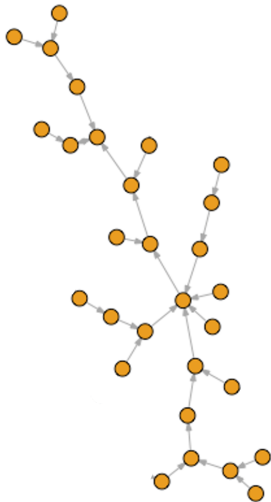
|      | Feat. 1 | Feat. 2 | Feat. 3 | Sum |
|------|---------|---------|---------|-----|
| Ind1 | 18      | 48      | 30      | 96  |
| Ind2 |         |         | 5       | 34  |
| Ind3 | 49      | 36      |         | 85  |
| Ind4 | 59      | 79      | 30      | 168 |

## IRIS - K = 25



## IRIS - K = 50

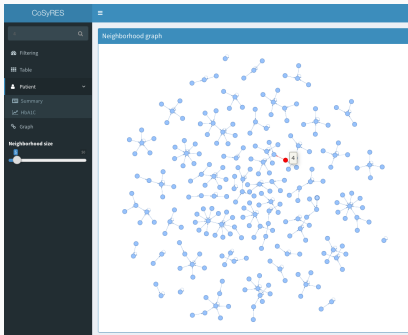




# EXPERIMENTS

| exemplar structuring |        |      |          | nearest neighbor |      |          |
|----------------------|--------|------|----------|------------------|------|----------|
| dataset              | number | size | diameter | number           | size | diameter |
| normal               | 18     | 16.7 | 5.2      | 22               | 13.6 | 2.8      |
| residential          | 64     | 5.8  | 2.7      | 105              | 3.5  | 2.1      |
| communities          | 256    | 8.7  | 3.2      | 400              | 5.5  | 2.5      |

# PROTOTYPE



## In a nutshell

- Structuring of high-dimensional data
- Interpretable results for domain experts
- Graph structure suitable for exploration

# CONCLUSION

## In a nutshell

- Structuring of high-dimensional data
- Interpretable results for domain experts
- Graph structure suitable for exploration

## Future work

- Automatically find best tradeoff for neighborhood size  $K$
- Use all values of  $K$  to determine the **DoR** score

- “Rapid Decision Making on the Fire Ground”  
2010 - G. Klein et al.
- “On the Surprising Behavior of Distance Metrics in High Dimensional Space”  
2001 - C. C. Aggarwal et al.
- “The Role of Hubness in Clustering High-Dimensional Data”  
2014 - N. Tomasev et al.

# EXPLORATION OF HIGH-DIMENSIONAL DATA

EXPLAINABLE STRUCTURING

DISCOVERY OF RELEVANT CASES

---

**Joris Falip**, Frédéric Blanchard, Michel Herbin

[joris.falip@univ-reims.fr](mailto:joris.falip@univ-reims.fr)

[www.datacrunch.sh](http://www.datacrunch.sh)

March 20 2019

CRéSTIC, University of Reims Champagne-Ardenne, France