AI4ALL

November 19, 2019

BOOSTING HUMAN EXPERTISE WITH SUBGROUP DISCOVERY

Speakers. Adnene Belfodil & Aimene Belfodil

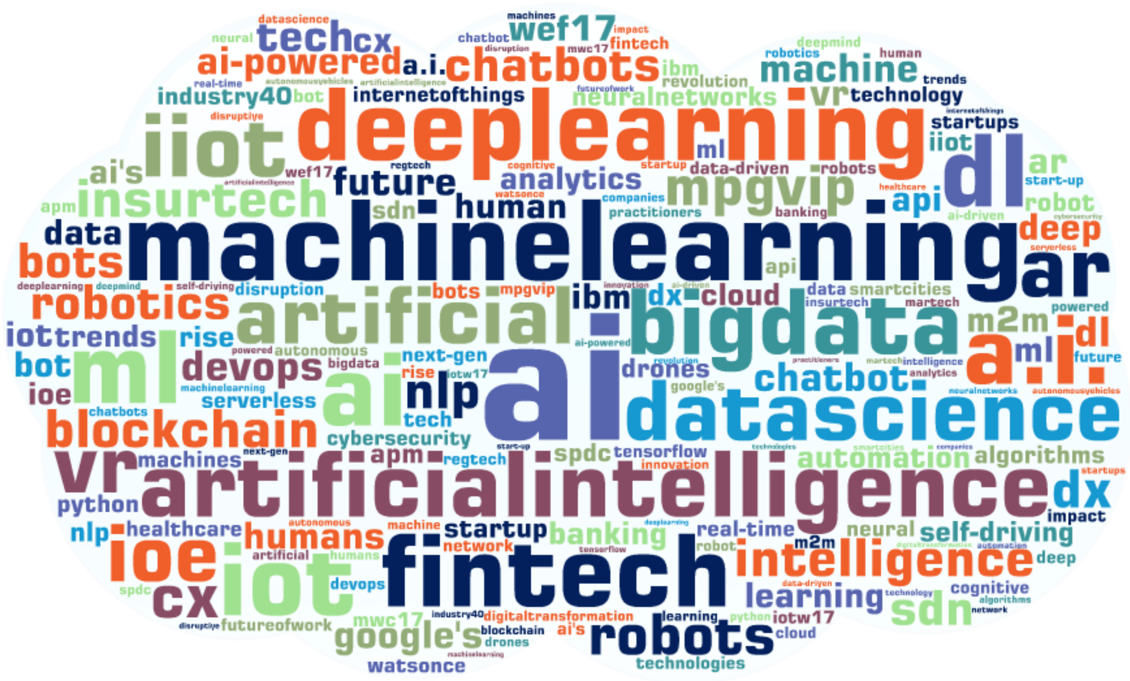
INTRODUCTION

ARTIFICIAL INTELLIGENCE AND KNOWLEDGE DISCOVERY

Artificial Intelligence

"AI [ed. Artificial Intelligence] currently encompasses a huge variety of subfields, ranging from the general (learning and perception) to the specific, such as playing chess, proving mathematical theorems, writing poetry, driving a car on a crowded street, and diagnosing diseases. AI is relevant to any intellectual task; it is truly a universal field."

Stuart J. Russell and Peter Norvig. 2010
[Artificial Intelligence A Modern Approach]

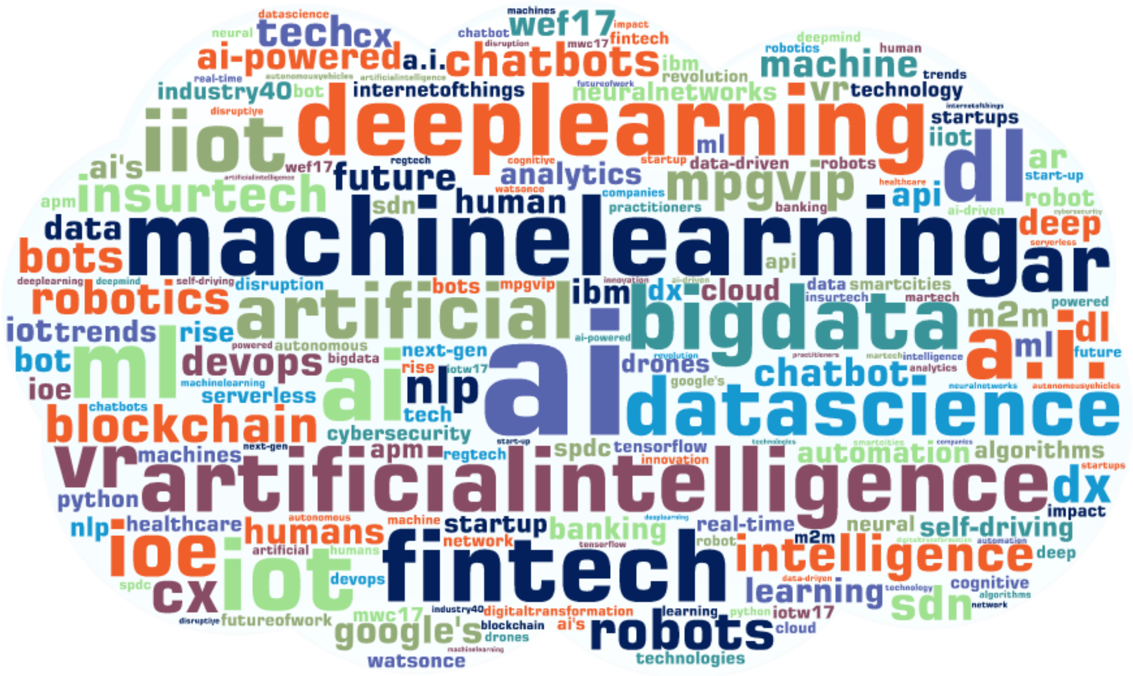


https://iiot-world.com/wp-content/uploads/2017/05/word-cloud.png

Artificial Intelligence

"AI [ed. Artificial Intelligence] currently encompasses a huge variety of subfields, ranging from the general (learning and perception) to the specific, such as playing chess, proving mathematical theorems, writing poetry, driving a car on a crowded street, and diagnosing diseases. AI is relevant to any intellectual task; it is truly a universal field."

Stuart J. Russell and Peter Norvig. 2010
[Artificial Intelligence A Modern Approach]

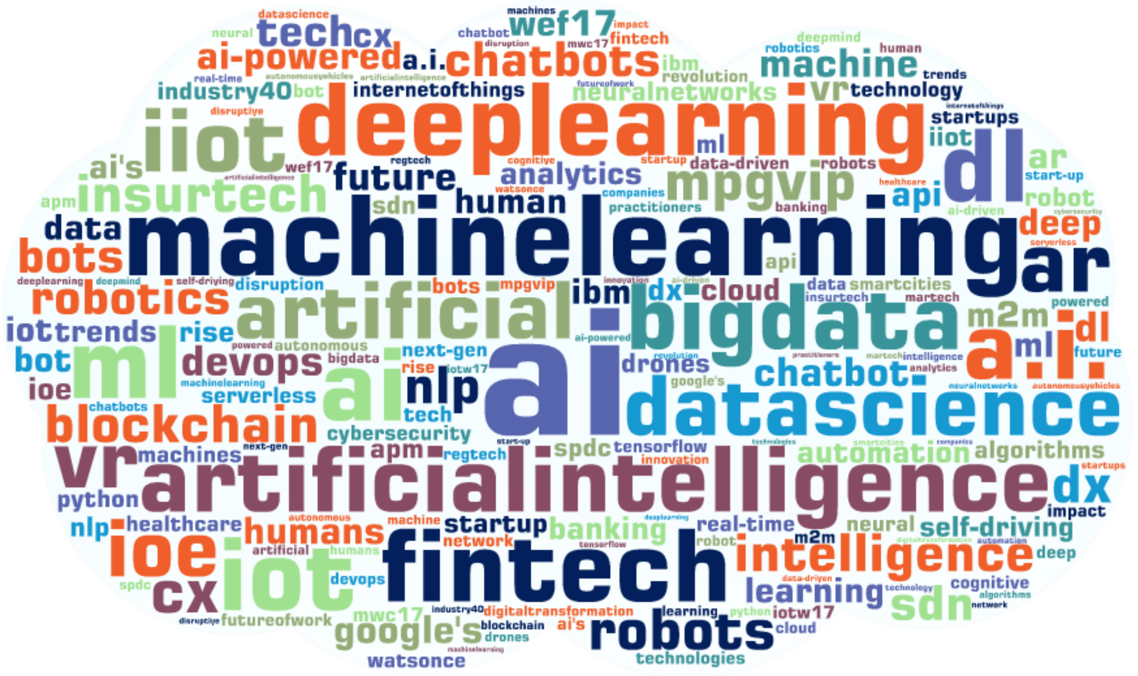


https://iiot-world.com/wp-content/uploads/2017/05/word-cloud.png

Artificial Intelligence

Artificial Intelligence (AI) is an umbrella term that simply means making computers act intelligently. It refers to computer software that can reason and adapt based on sets of rules. AI technologies include, but are not limited to,

- NLP,
- automated reasoning,
- machine learning,
- expert systems,
- robotics,
- computer vision,
- multi-agent systems,
- knowledge discovery.

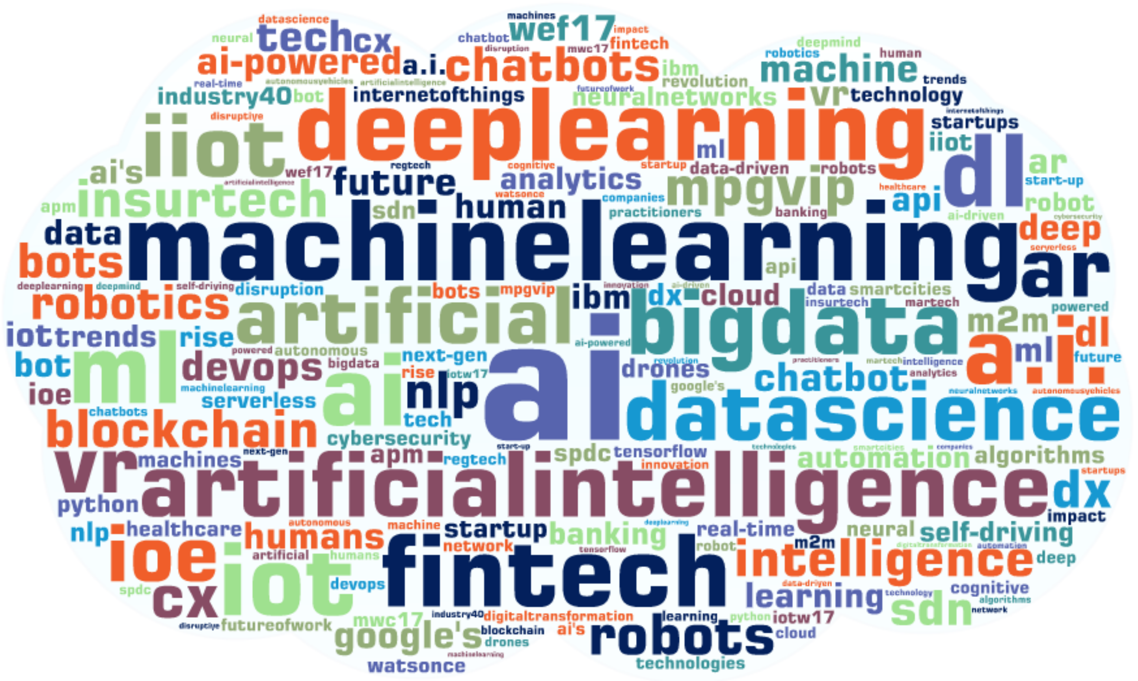


https://iiot-world.com/wp-content/uploads/2017/05/word-cloud.png

Artificial Intelligence

Artificial Intelligence (AI) is an umbrella term that simply means making computers act intelligently. It refers to computer software that can reason and adapt based on sets of rules. AI technologies include, but are not limited to,

- NLP,
- automated reasoning,
- machine learning,
- expert systems,
- robotics,
- computer vision,
- multi-agent systems,
- knowledge discovery.



https://iiot-world.com/wp-content/uploads/2017/05/word-cloud.png

Data Analysis vs Machine Learning

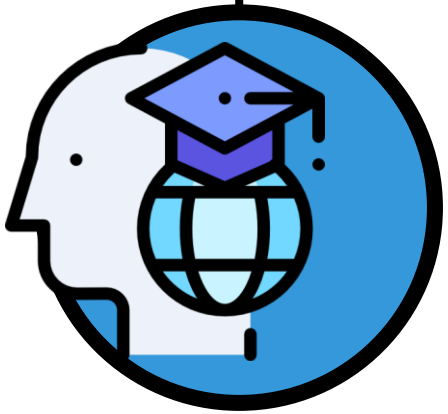
The ultimate purpose of the Machine Learning is automatization. i.e., to use a machine to substitute human cognition.

e.g.: for autonomously driving a car based on sensor inputs.

V.S.

In contrast, the purpose of (exploratory) data analysis is to use machine discoveries to synergistically boost human expertise.

e.g.: for understanding relationships between physicochemical properties of odorant molecules and their odors.



ed. Boley M. [realKD Blog, 2017]

Data Analysis vs Machine Learning

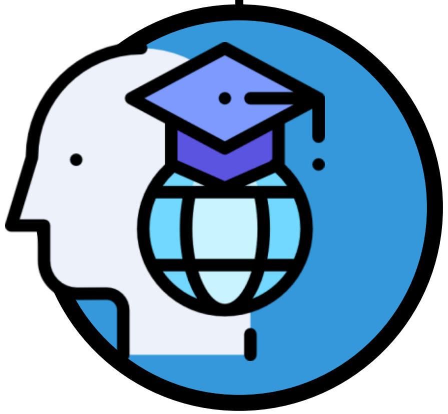
The ultimate purpose of the Machine Learning is automatization. i.e., to use a machine to substitute human cognition.

e.g.: for autonomously driving a car based on sensor inputs.

V.S.

In contrast, the purpose of (exploratory) data analysis is to use machine discoveries to synergistically boost human expertise.

e.g.: for understanding relationships between physicochemical properties of odorant molecules and their odors.

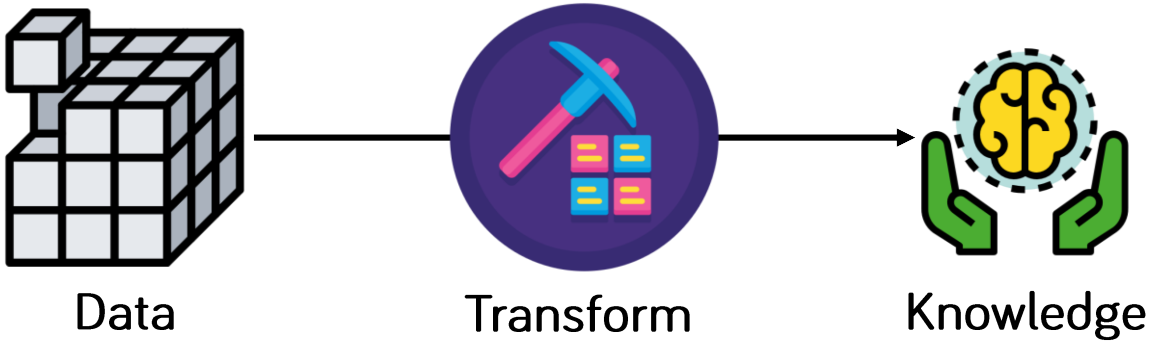


ed. Boley M. [realKD Blog, 2017]

EDA and KDD: Knowledge Discovery in Databases

Data Analysis: “Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”

Tukey, J.W. 1961

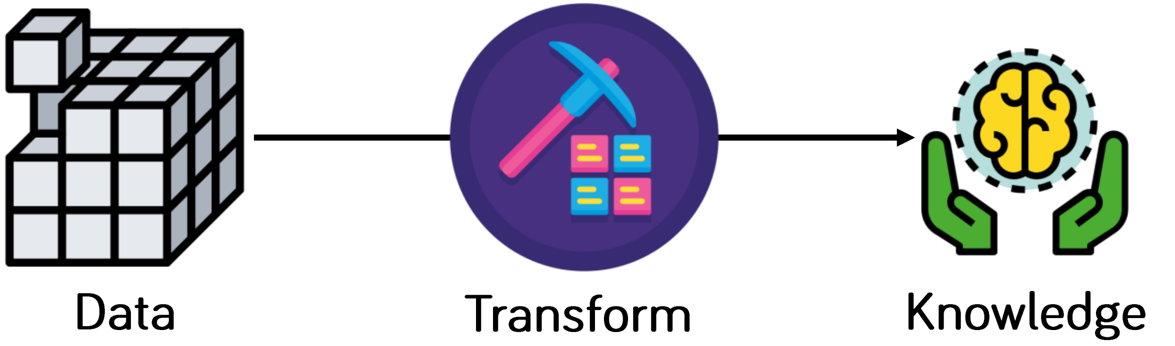


Knowledge Discovery in Databases: “The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.”

Fayyad U., Piatetsky-Shapiro G., Smyth P. 1996

Data Analysis: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

Tukey, J.W. 1961

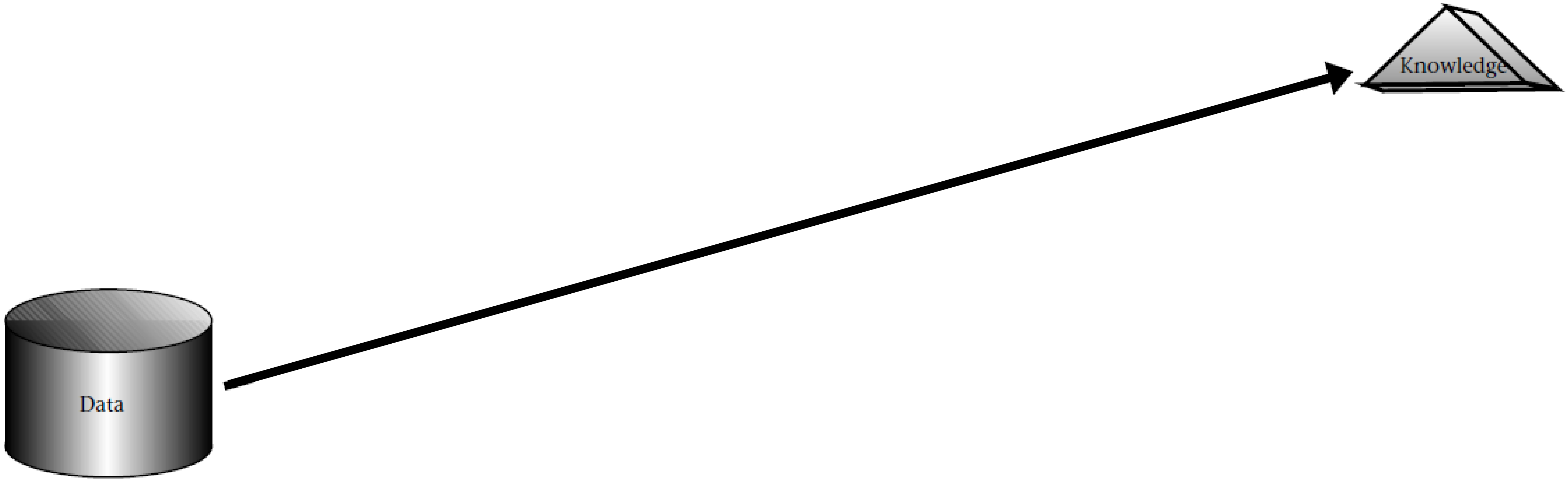


Knowledge Discovery in Databases: "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

Fayyad U., Piatetsky-Shapiro G., Smyth P. 1996

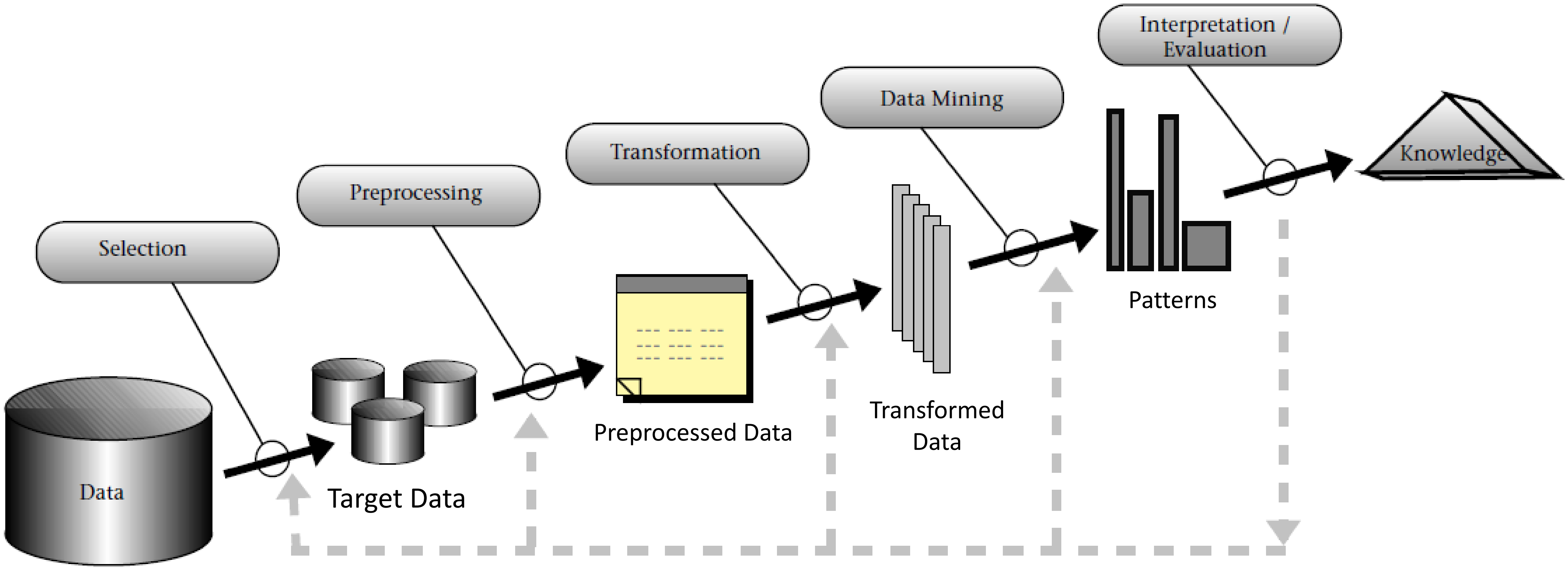
Knowledge Discovery in Databases

Fayyad U., Piatetsky-Shapiro G., Smyth P. [AAAI, 1996]



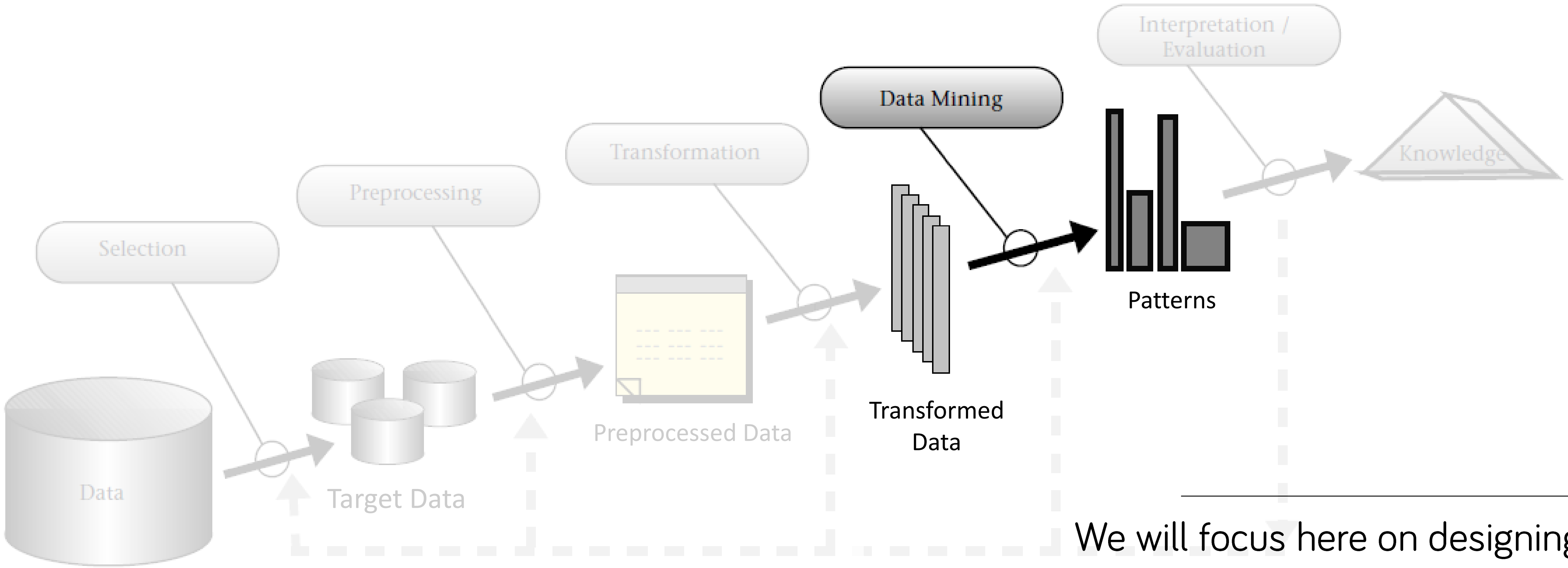
Knowledge Discovery in Databases – A Process

Fayyad U., Piatetsky-Shapiro G., Smyth P. [AAAI, 1996]



Knowledge Discovery in Databases – A Process

Fayyad U., Piatetsky-Shapiro G., Smyth P. [AAAI, 1996]



We will focus here on designing pattern discovery algorithms in prepared data.

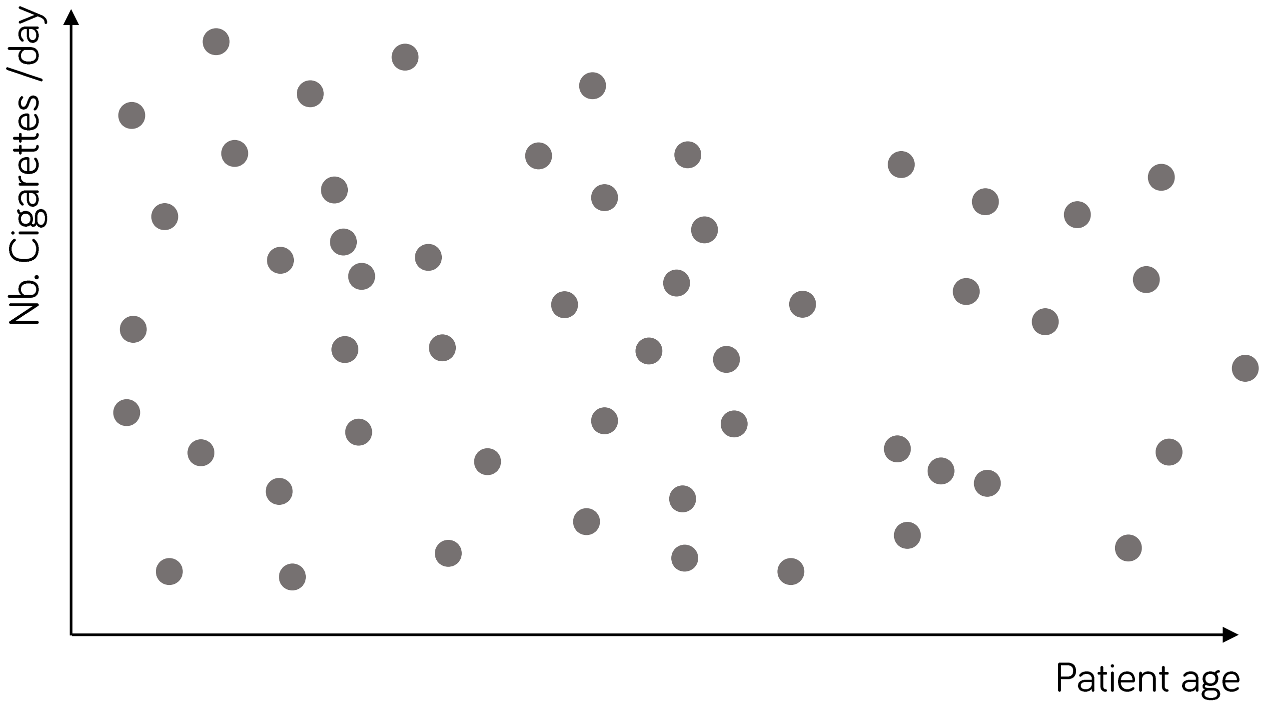
SUBGROUP DISCOVERY

WHAT IS IT ABOUT?

Subgroup Discovery : Toy Example 1

Subgroup Discovery*

The problem of finding statistically unusual subgroups in a given database



* S. Wrobel. An Algorithm for Multi-relational Discovery of Subgroups. In PKDD 1997.

Subgroup Discovery : Toy Example 1

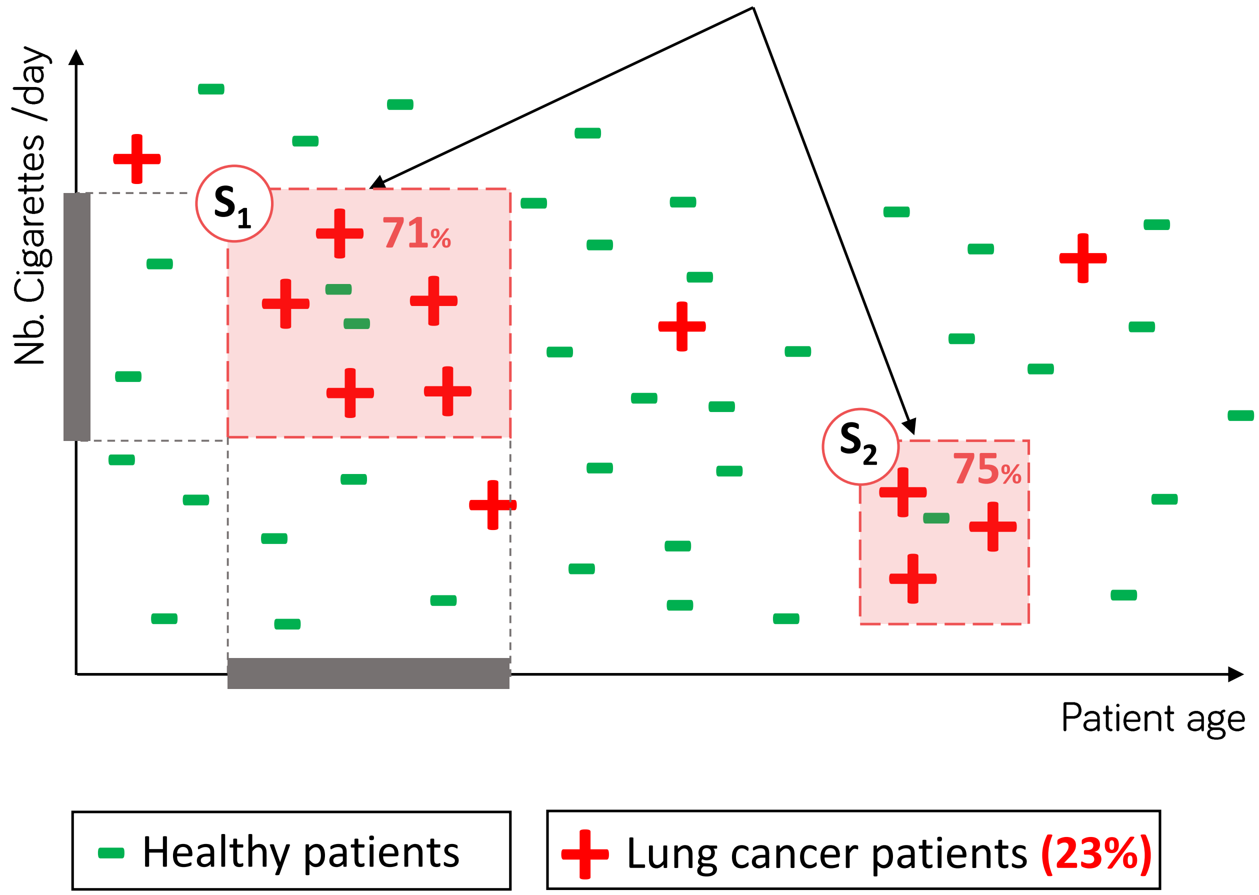
Subgroup Discovery*

The problem of finding statistically unusual subgroups in a given database

- Subgroups are subsets of individuals
- Subgroups are interesting
- Subgroups are described using individuals attributes

* S. Wrobel. An Algorithm for Multi-relational Discovery of Subgroups. In PKDD 1997.

Subgroups of patients with high lung cancer prevalence

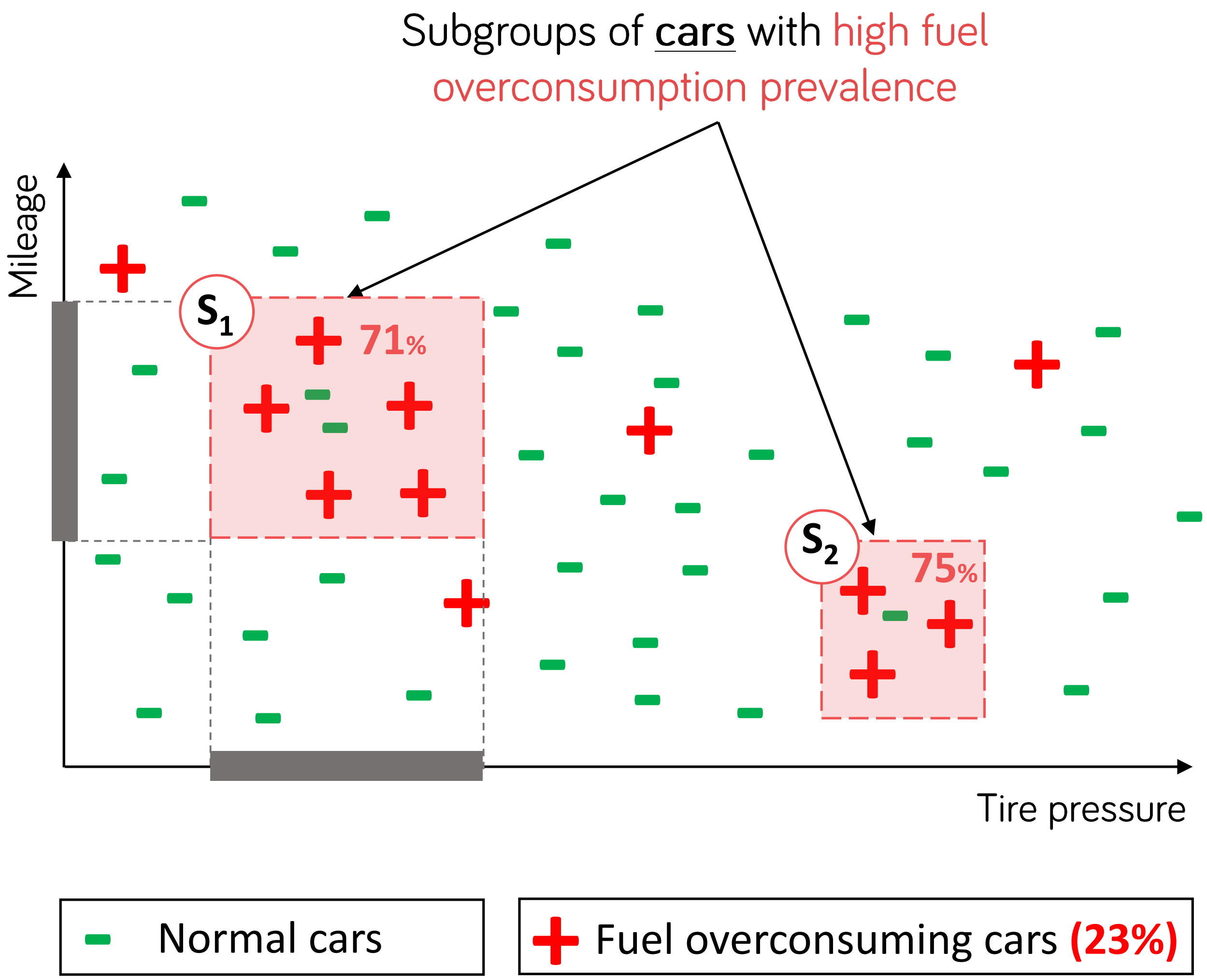


Subgroup Discovery : Toy Example 2

Subgroup Discovery*

The problem of finding statistically unusual subgroups in a given database

- Subgroups are subsets of individuals
- Subgroups are interesting
- Subgroups are described using individuals attributes



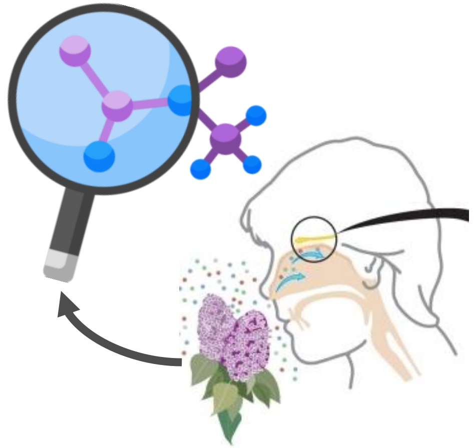
* S. Wrobel. An Algorithm for Multi-relational Discovery of Subgroups. In PKDD 1997.

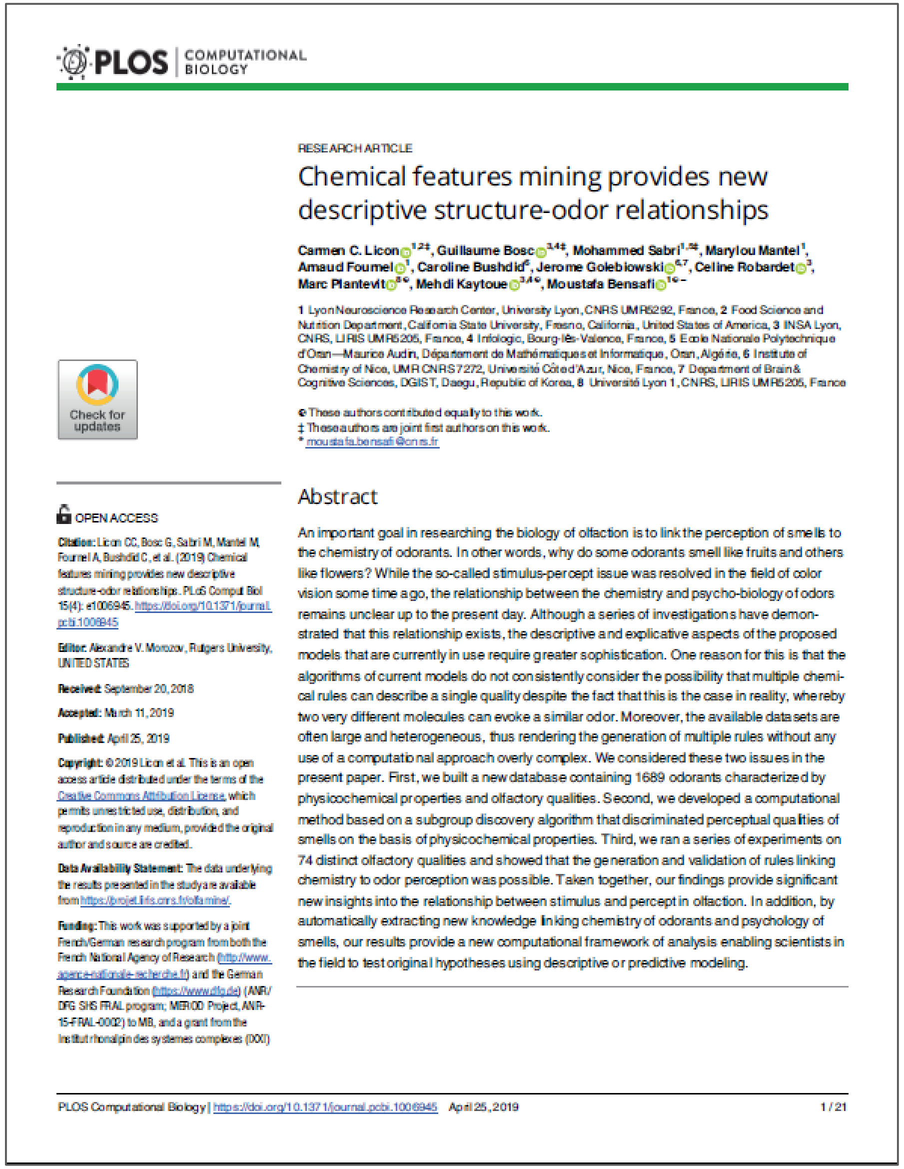
Subgroup Discovery : A Real Use Case

C. C. Licon, G. Bosc, M. Sabri, M. Mantel, A. Fournel, C. Bushdid, J. Golebiowski, C. Robardet, M. Plantevit, M. Kaytoue, M. Bensafi. Chemical features mining provides new descriptive structure-odor relationships. PLOS Computational Biology 2019

Collaboration between Computer Scientists, Neuroscientists and Chemists

Find relationships between the physicochemical properties of odorant molecules and their olfactory qualities





PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Chemical features mining provides new descriptive structure-odor relationships

Carmen C. Licon[1,2‡], Guillaume Bosc[3,4‡], Mohammed Sabri[1,5‡], Marylou Mantel[1], Arnaud Fournel[1], Caroline Bushdid[6], Jerome Golebiowski[6,7], Celine Robardet[3], Marc Plantevit[8☯], Mehdi Kaytoue[3,4☯], Moustafa Bensafi[1☯*]

1 Lyon Neuroscience Research Center, University Lyon, CNRS UMR5292, France, 2 Food Science and Nutrition Department, California State University, Fresno, California, United States of America, 3 INSA Lyon, CNRS, LIRIS UMR5205, France, 4 Infologic, Bourg-lès-Valence, France, 5 Ecole Nationale Polytechnique d'Oran—Maurice Audin, Département de Mathématiques et Informatique, Oran, Algérie, 6 Institute of Chemistry of Nice, UMR CNRS 7272, Université Côte d'Azur, Nice, France, 7 Department of Brain & Cognitive Sciences, DGIST, Daegu, Republic of Korea, 8 Université Lyon 1, CNRS, LIRIS UMR5205, France

☯ These authors contributed equally to this work.
‡ These authors are joint first authors on this work.
* moustafa.bensafi@cnrs.fr

Check for updates

OPEN ACCESS

Citation: Licon CC, Bosc G, Sabri M, Mantel M, Fournel A, Bushdid C, et al. (2019) Chemical features mining provides new descriptive structure-odor relationships. PLoS Comput Biol 15(4): e1006945. https://doi.org/10.1371/journal.pcbi.1006945

Editor: Alexandre V. Morozov, Rutgers University, UNITED STATES

Received: September 20, 2018

Accepted: March 11, 2019

Published: April 25, 2019

Data Availability Statement: The data underlying the results presented in the study are available from https://projet.liris.cnrs.fr/olfamine/.

Funding: This work was supported by a joint French/German research program from both the French National Agency of Research (http://www.agence-nationale-recherche.fr) and the German Research Foundation (https://www.dfg.de) (ANR/DFG SHS FRAL program; MEROD Project, ANR-15-FRAL-0002) to MB, and a grant from the Institut rhonalpin des systemes complexes (IXXI)

Abstract

An important goal in researching the biology of olfaction is to link the perception of smells to the chemistry of odorants. In other words, why do some odorants smell like fruits and others like flowers? While the so-called stimulus-percept issue was resolved in the field of color vision some time ago, the relationship between the chemistry and psycho-biology of odors remains unclear up to the present day. Although a series of investigations have demonstrated that this relationship exists, the descriptive and explicative aspects of the proposed models that are currently in use require greater sophistication. One reason for this is that the algorithms of current models do not consistently consider the possibility that multiple chemical rules can describe a single quality despite the fact that this is the case in reality, whereby two very different molecules can evoke a similar odor. Moreover, the available datasets are often large and heterogeneous, thus rendering the generation of multiple rules without any use of a computational approach overly complex. We considered these two issues in the present paper. First, we built a new database containing 1689 odorants characterized by physicochemical properties and olfactory qualities. Second, we developed a computational method based on a subgroup discovery algorithm that discriminated perceptual qualities of smells on the basis of physicochemical properties. Third, we ran a series of experiments on 74 distinct olfactory qualities and showed that the generation and validation of rules linking chemistry to odor perception was possible. Taken together, our findings provide significant new insights into the relationship between stimulus and percept in olfaction. In addition, by automatically extracting new knowledge linking chemistry of odorants and psychology of smells, our results provide a new computational framework of analysis enabling scientists in the field to test original hypotheses using descriptive or predictive modeling.

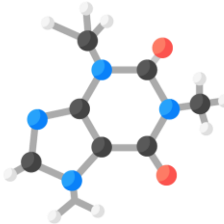
Subgroup Discovery : A Real Use Case

X	MW	nAT	nC	...	Vanillin	Woody	...
1	150	21	11		×	×	...
2	128	24	9				...
3	136	29	10	...		×	...
4	152	23	11	...		×	...
5	151	27	12	...	×		
6	154	27	10	...	×		...
...	...	...	...	...	...	...	...

Data: 1689 Molecules, 82 Properties, 74 Odors

Algorithm

Subgroup

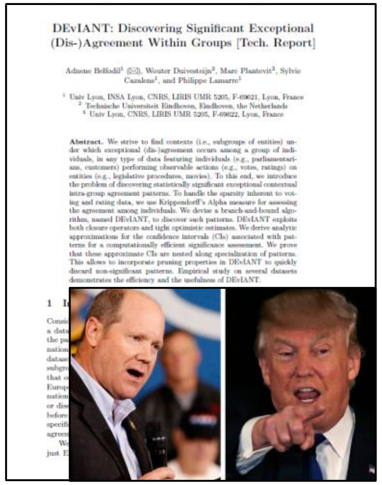


$$MW(X) \in [128, 151] \wedge$$
$$nAT(X) \in [20, 27] \wedge$$
$$nC(X) \in [9,12] \rightarrow Vanilin$$

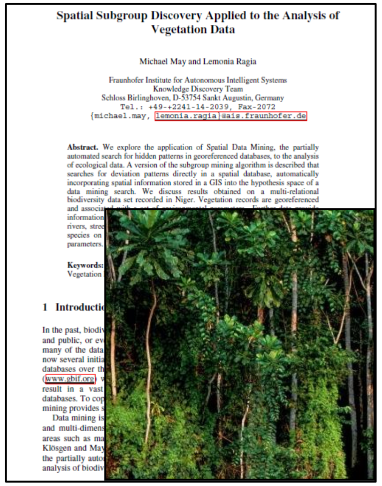
*Molecules with such physicochemical attributes are likely to smell Vanillin

The "Likely" is quantified by some interestingness measure

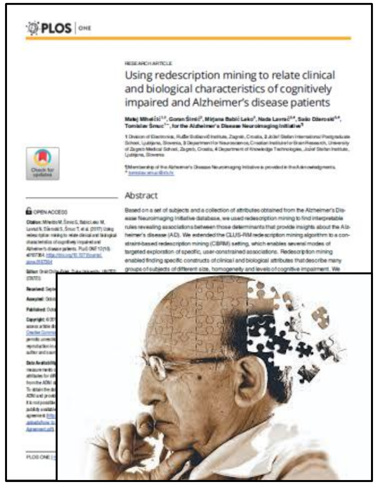
Subgroup Discovery : More Use Cases



Find exceptional disagreements within groups of individuals.



Find dependencies of a plant species on other plant species and environmental parameters.



Find relationship between biological and clinical attributes of cognitively impaired and Alzheimer's disease patients.

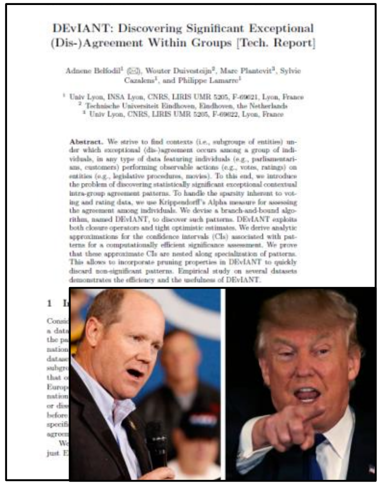
A. Belfodil, W. Duivesteijn et al. DEvIANT: Discovering Significant Exceptional (Dis-) Agreement Within Groups. In ECML/PKDD 2019.

M. May, L. Ragia. Spatial subgroup discovery applied to the analysis of vegetation data. In PAKM 2002.

M. Mihelčić, G. Šimić et al. Using redescription mining to relate clinical and biological characteristics of cognitively impaired and Alzheimer's disease patients. In PLOS One 2017.

● ● ●

Subgroup Discovery : More Use Cases



DEvIANT: Discovering Significant Exceptional (Dis-)Agreement Within Groups [Tech. Report]

Adnene Belfodil[1], Wouter Duivesteijn[2], Marc Plantevit[3], Sylvie Cazalens[1], and Philippe Lamarre[1]

[1] Univ Lyon, INSA Lyon, CNRS, LIRIS UMR 5205, F-69621, Lyon, France
[2] Technische Universiteit Eindhoven, Eindhoven, the Netherlands
[3] Univ Lyon, CNRS, LIRIS UMR 5205, F-69622, Lyon, France

Abstract. We strive to find contexts (i.e., subgroups of entities) under which exceptional (dis-)agreement occurs among a group of individuals, in any type of data featuring individuals (e.g., parliamentarians, customers) performing observable actions (e.g., votes, ratings) on entities (e.g., legislative procedures, movies). To this end, we introduce the problem of discovering statistically significant exceptional contextual intra-group agreement patterns. To handle the sparsity inherent to voting and rating data, we use Krippendorff's Alpha measure for assessing the agreement among individuals. We devise a branch-and-bound algorithm, named DEvIANT, to discover such patterns. DEvIANT exploits both closure operators and tight optimistic estimates. We derive analytic approximations for the confidence intervals (CIs) associated with patterns for a computationally efficient significance assessment. We prove that these approximate CIs are nested along specialization of patterns. This allows to incorporate pruning properties in DEvIANT to quickly discard non-significant patterns. Empirical study on several datasets demonstrates the efficiency and the usefulness of DEvIANT.

Find exceptional disagreements within groups of individuals.

Spatial Subgroup Discovery Applied to the Analysis of Vegetation Data

Michael May and Lemonia Ragia

Find dependencies of a plant species on other plant species and environmental parameters.

Using redescription mining to relate clinical and biological characteristics of cognitively impaired and Alzheimer's disease patients

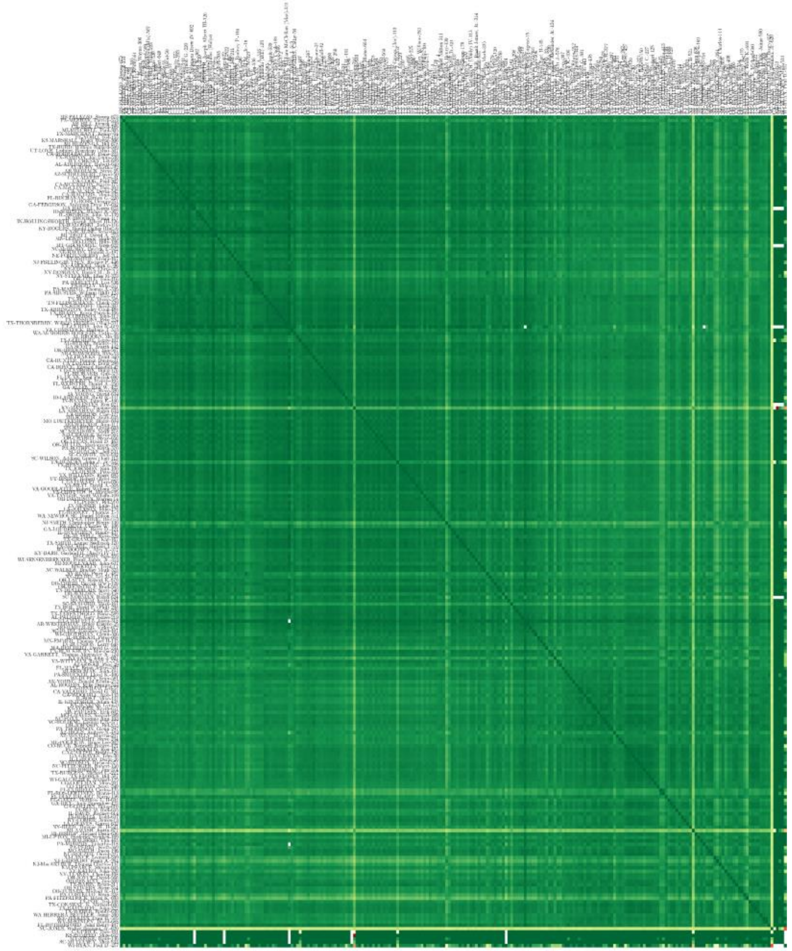
Find relationship between biological and clinical attributes of cognitively impaired and Alzheimer's disease patients.

A. Belfodil, W. Duivesteijn et al. DEvIANT: Discovering Significant Exceptional (Dis-) Agreement Within Groups. In ECML/PKDD 2019.

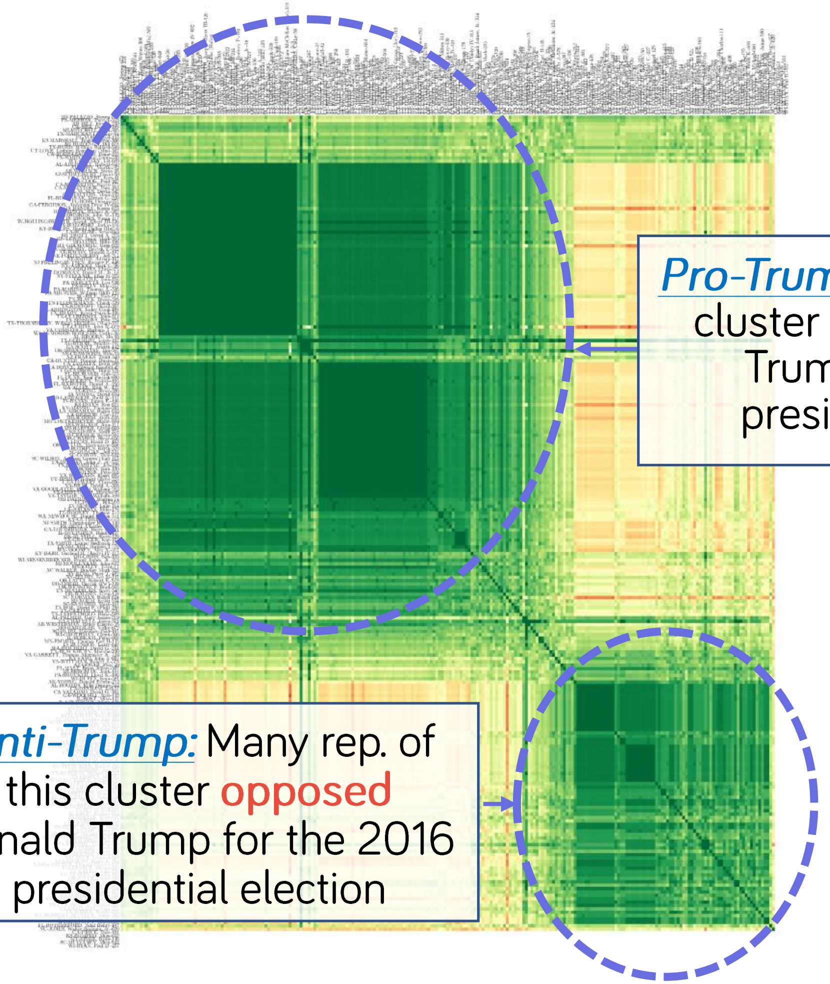
M. May, L. Ragia. Spatial subgroup discovery applied to the analysis of vegetation data. In PAKM 2002.

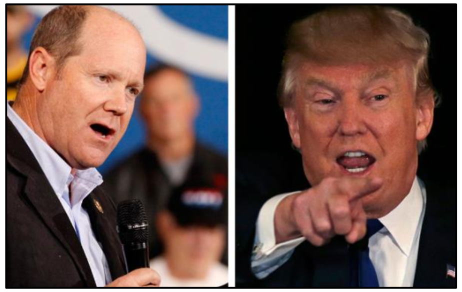
M. Mihelčić, G. Šimić et al. Using redescription mining to relate clinical and biological characteristics of cognitively impaired and Alzheimer's disease patients. In PLOS One 2017.

Subgroup Discovery : A Use Case in the US Congress



Overall intra-agreement between Republicans
Group





Contextual intra-agreement between Republicans in Government and Administrative Issues related voting sessions
Context

SUBGROUP DISCOVERY

HOW DOES IT WORK?

Subgroup Discovery: A generic framework

Data

What are the data we have?

Design **efficient algorithms** to find the **best set of hypotheses** from the **data**

Interestingness

What is the task we are targeting? How should we assess the quality of our discoveries?

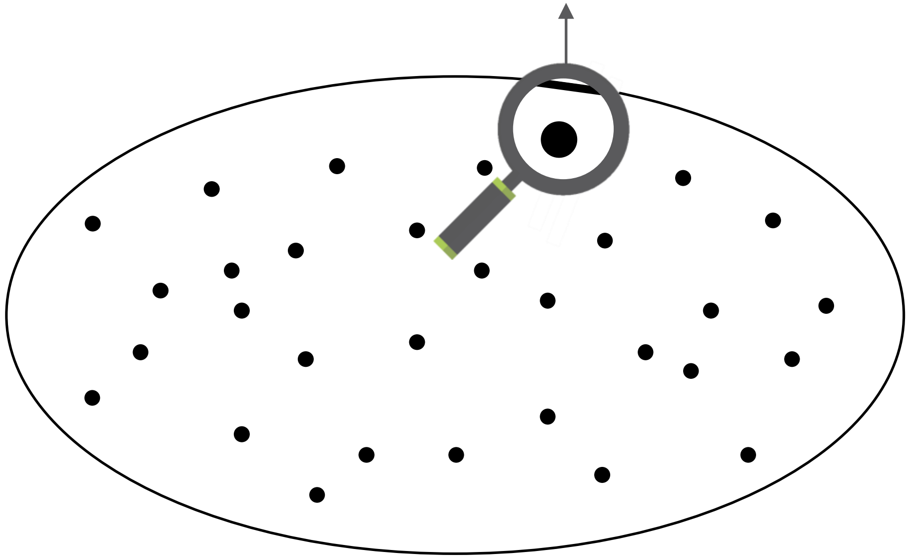
Hypothesis Language

What kind of hypotheses we want to investigate?

> *" … The problem of finding statistically unusual subgroups in a given database"*
>
> S. Wrobel [PKDD, 1997]

Hypotheses Spaces

Patient Age = Young and Number of Cigarettes per day < 15

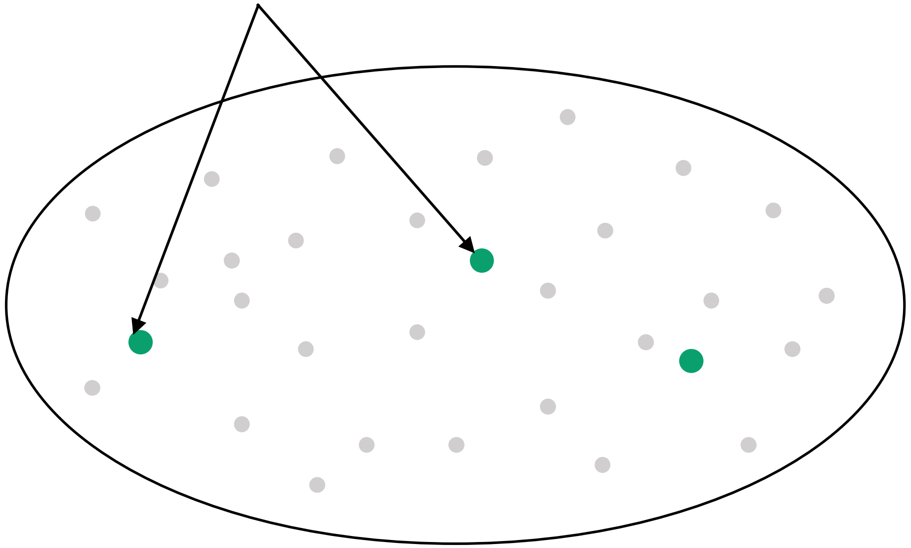


Hypotheses Search Space

Algorithm

Elicit the interesting hypotheses out from the data

Interesting Hypotheses

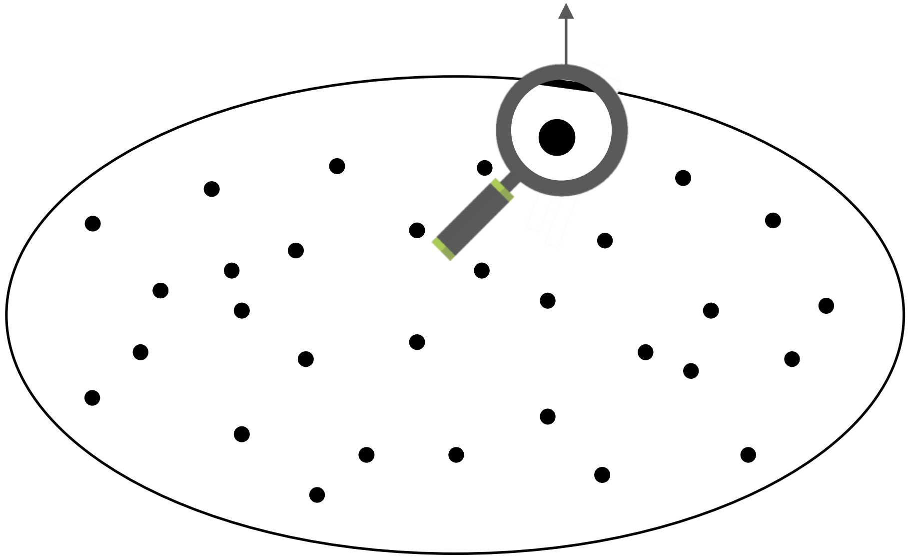


Collection of interesting hypotheses (subgroups)

Data → Knowledge

Hypotheses Spaces

Patient Age = Young and Number of Cigarettes per day < 15

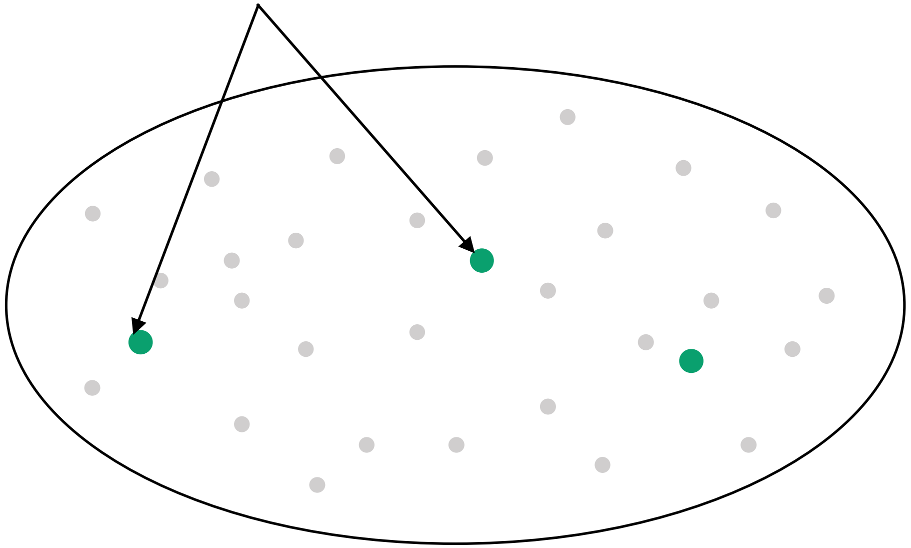


Hypotheses Search Space

Algorithm

Elicit the interesting hypotheses out from the data

Interesting Hypotheses



Collection of interesting hypotheses (subgroups)

Data → Knowledge

AI4ALL. TALK
Adnene Belfodil and Aimene Belfodil

Subgroup Discovery Algorithmic Challenges

Interactivity

Algorithms need to be fast for better interactivity

Efficiency

Algorithms need to avoid consuming too much resources

Subgroup Discovery: Our Research Interest

Understand the theory lying behind subgroup discovery

- A. Belfodil, M. Kaytoue, M. Kaytoue and A. Napoli. On Ordered Sets in Pattern Mining. An ECML/PKDD Tutorial in 2019.
- A. Belfodil, M. Kaytoue and S. O. Kuznetsov. On Pattern Setups and Pattern Multistructures. In Arxiv 2019.

Provide efficient and fast algorithms

- A. Belfodil, A. Belfodil and M. Kaytoue. Anytime Subgroup Discovery in Numerical Domains with Guarantees. In ECML/PKDD 2018.
- A. Belfodil, A. Belfodil et al. FSSD – A Fast and Efficient Algorithm for Subgroup Set Discovery. In DSAA 2019.

Provide practical tools to end users

- A. Belfodil, S. Cazalens et al. Identifying exceptional (dis) agreement between groups. In DMKD Journal 2019 (to appear).
- A. Belfodil, W. Duivesteijn et al. DEvIANT: Discovering Significant Exceptional (Dis-) Agreement Within Groups. In ECML/PKDD 2019.

Thesis Work

Aimene Belfodil

An Order-Theoretic Point of View on Subgroup Discovery (2019)

Adnene Belfodil

Exceptional Model Mining for Behavioral Data Analysis (2019)

Subgroup Discovery: Our Research Interest

Understand the theory lying behind subgroup discovery

- A. Belfodil, M. Kaytoue, M. Kaytoue and A. Napoli. On Ordered Sets in Pattern Mining. An ECML/PKDD Tutorial in 2019.
- A. Belfodil, M. Kaytoue and S. O. Kuznetsov. On Pattern Setups and Pattern Multistructures. In Arxiv 2019.

Provide efficient and fast algorithms

- A. Belfodil, A. Belfodil and M. Kaytoue. Anytime Subgroup Discovery in Numerical Domains with Guarantees. In ECML/PKDD 2018.
- A. Belfodil, A. Belfodil et al. FSSD – A Fast and Efficient Algorithm for Subgroup Set Discovery. In DSAA 2019.

Provide practical tools to end users

- A. Belfodil, S. Cazalens et al. Identifying exceptional (dis) agreement between groups. In DMKD Journal 2019 (to appear).
- A. Belfodil, W. Duivesteijn et al. DEvIANT: Discovering Significant Exceptional (Dis-) Agreement Within Groups. In ECML/PKDD 2019.

Thesis Work

Aimene Belfodil
An Order-Theoretic Point of View on Subgroup Discovery (2019)

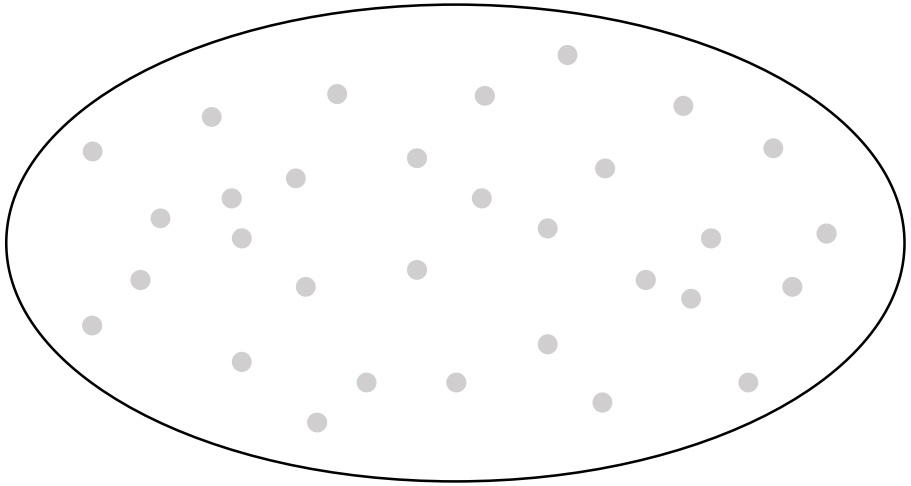
Adnene Belfodil
Exceptional Model Mining for Behavioral Data Analysis (2019)

SUBGROUP DISCOVERY

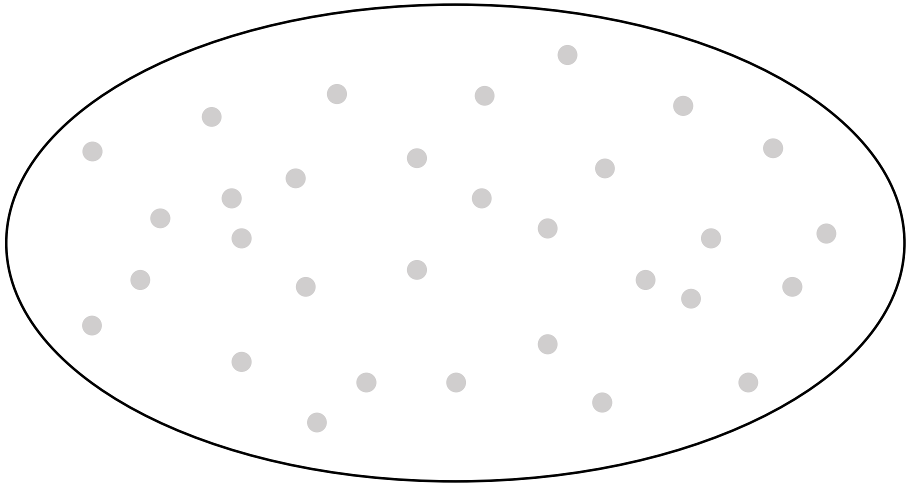
TOWARD FAST ALGORITHMS

Exhaustive VS Heuristic Search Paradigm

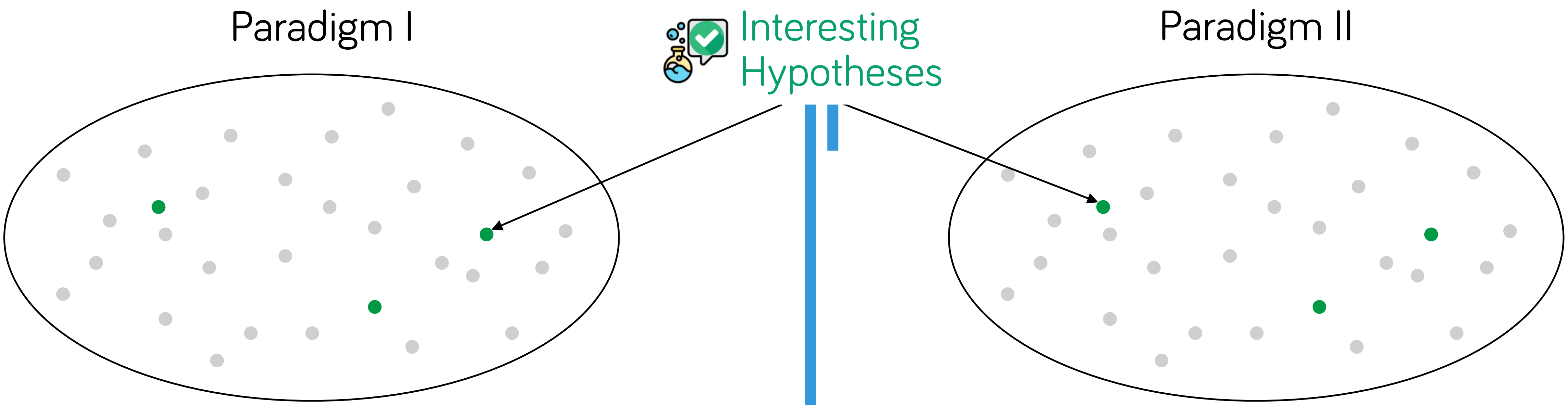
Paradigm I



Paradigm II



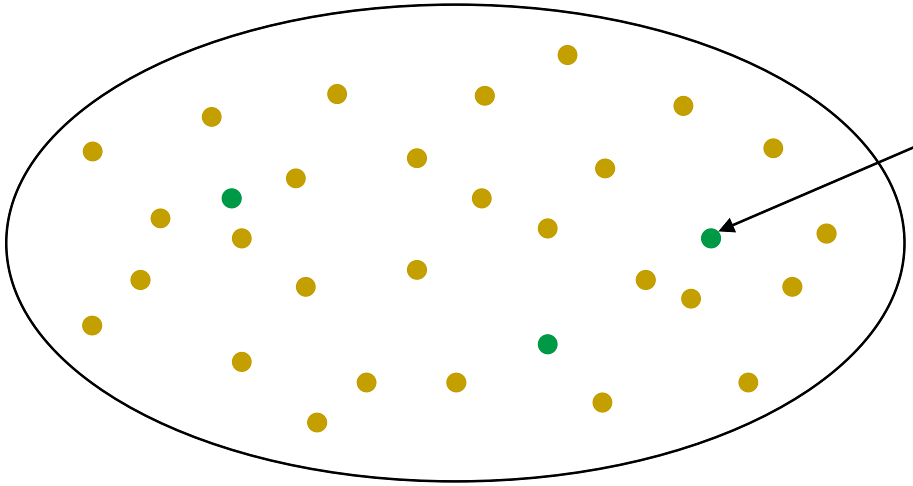
Exhaustive VS Heuristic Search Paradigm



Exhaustive VS Heuristic Search Paradigm

Interesting Hypotheses

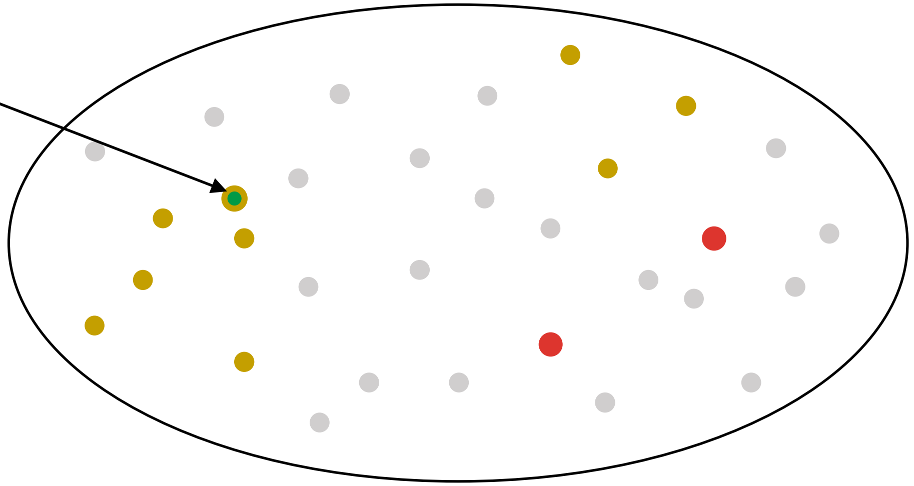
Exhaustive Search Paradigm



All subgroups are looked for and then post-processed to obtain the best subgroups

- *Guarantees to find the best subgroups*
- *Time and resources consuming*

Heuristic Search Paradigm

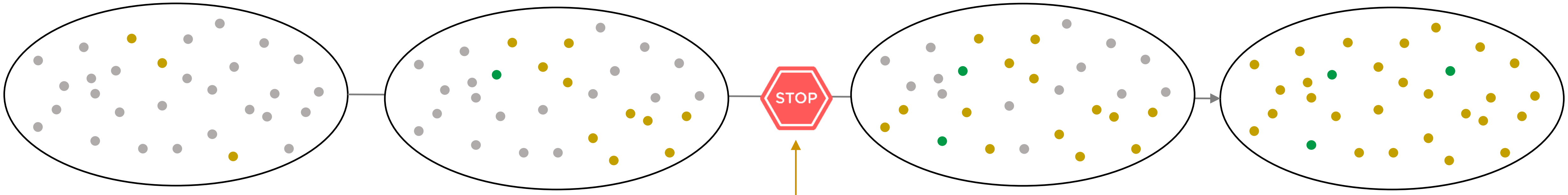


Only few subgroups are looked for and then post-processed to obtain the best subgroups

- *Efficient algorithms: fast and not resource greedy*
- *Lack of exhaustivity: some (to most) interesting subgroups can be ignored*

Refine and Mine : a Good Tradeoff

Good trade-off between the two paradigms



Few patterns after 5”

More patterns after 1’

Many patterns after 20’

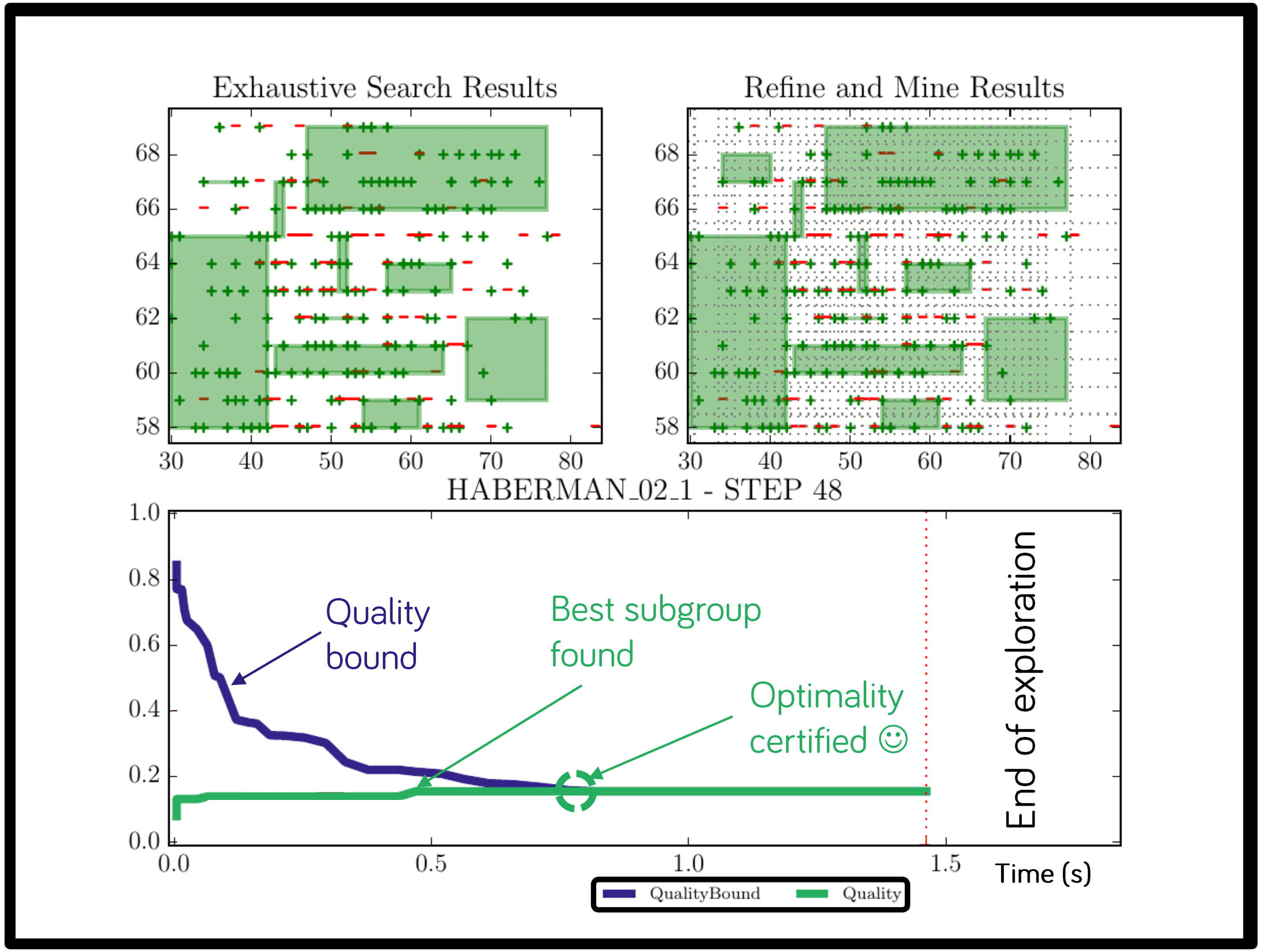
All patterns after the End

A. Belfodil, A. Belfodil and M. Kaytoue. Anytime Subgroup Discovery in Numerical Domains with Guarantees. In ECML/PKDD 2018.

Awarded as the best data mining paper in ECML/PKDD 2018

Propose Refine&Mine, an Anytime algorithm which yields progressively interesting subgroups whose quality improves over time.

Refine and Mine : a Good Tradeoff

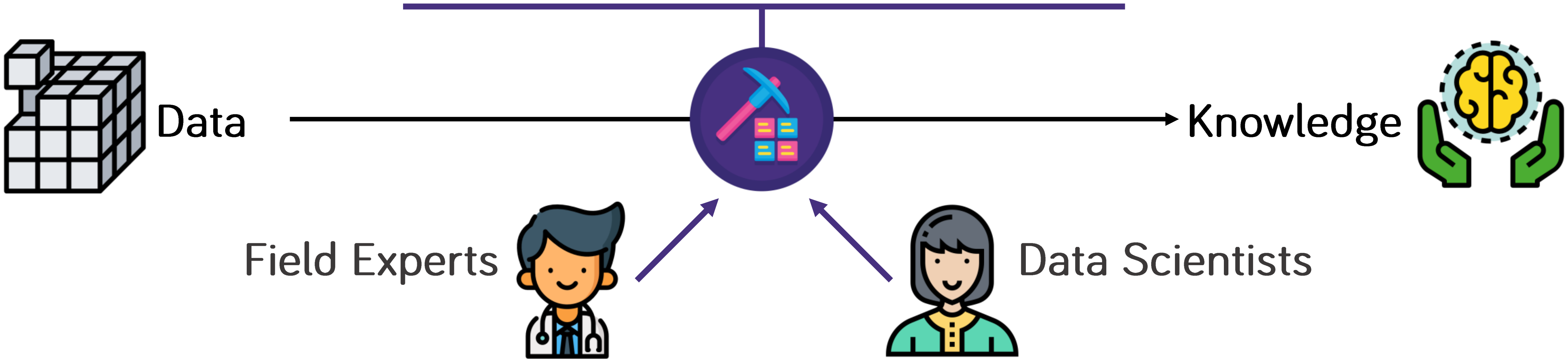


CONCLUDING REMARKS

WHAT SHOULD WE TAKE HOME?

Conclusion

Subgroup Discovery is a generic framework that can be used to enhance human expertise by exploiting the data.



How can we leverage subgroup discovery framework to facilitate the discovery of interesting hypotheses from clinical data?

THANKS
FOR YOUR TIME

? QUESTIONS

Contact : adnene.belfodil@gmail.com
aimene.belfodil@gmail.com