HUMAN-IN-THE-LOOP OR HUMAN-IN-THE-WAY

Pitfalls of interactive data mining or : how to help the user

Albrecht Zimmermann

CODAG, GREYC, University of Caen

GREYC
Electronics and Computer Science Laboratory

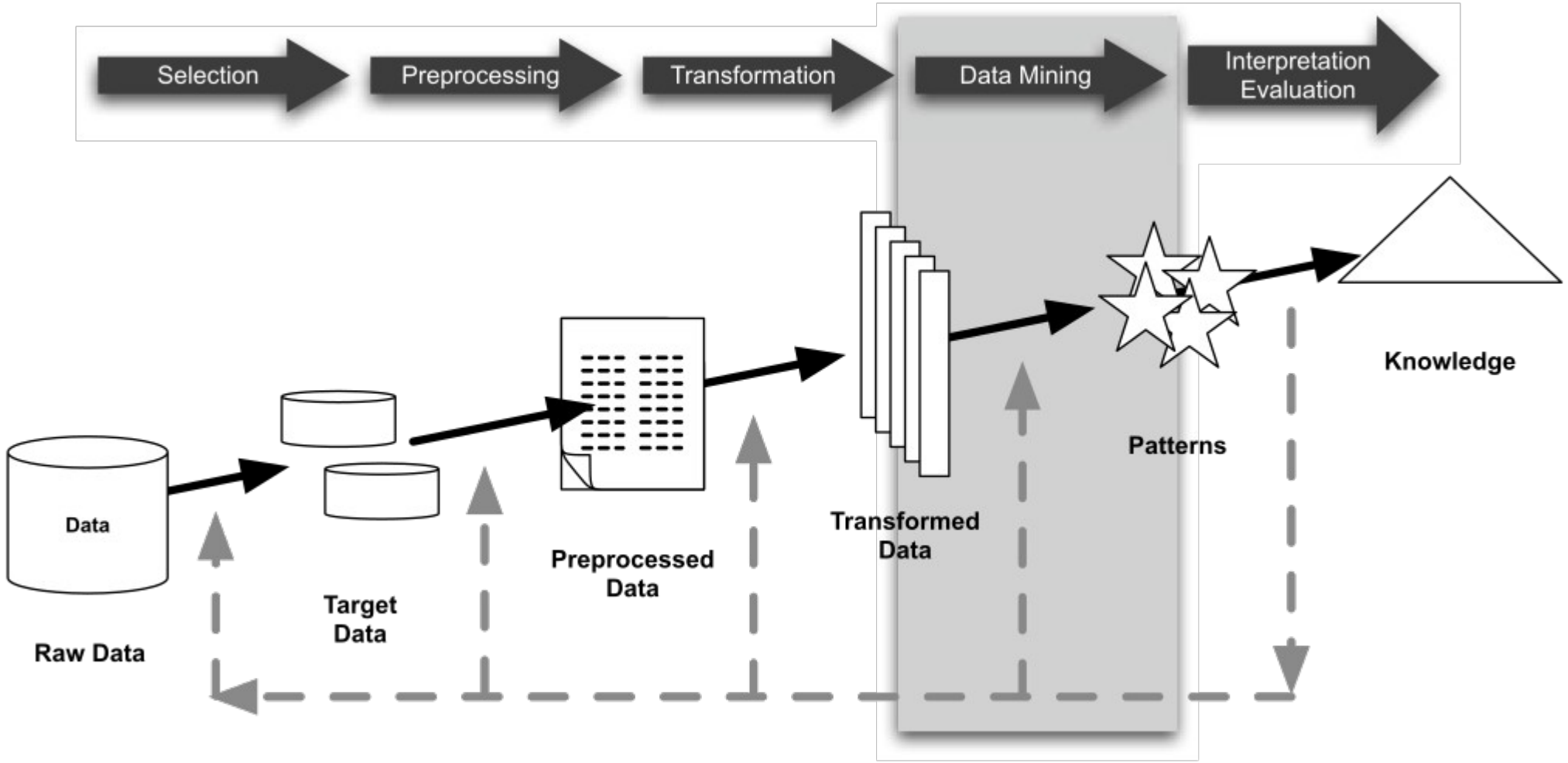
Normandie Université

UNICAEN

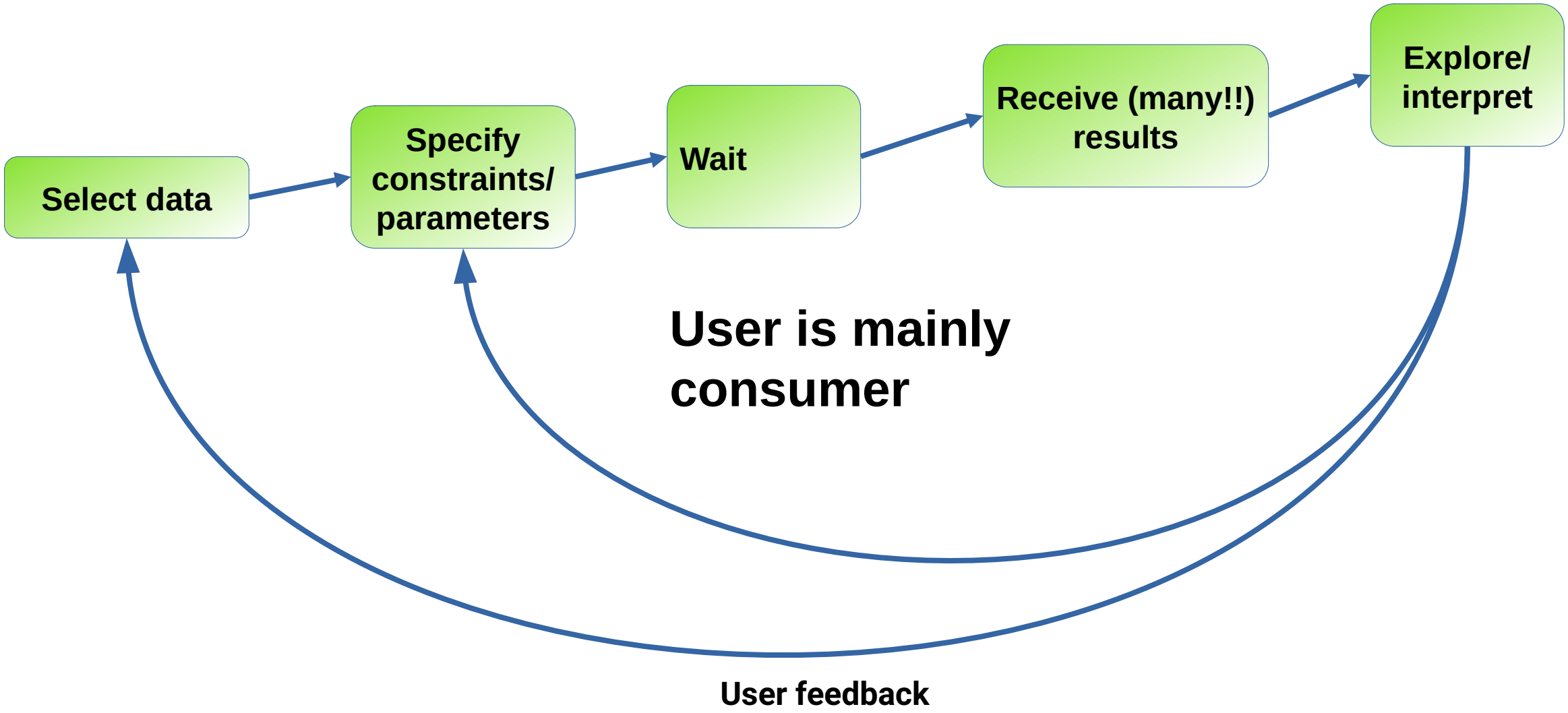
ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

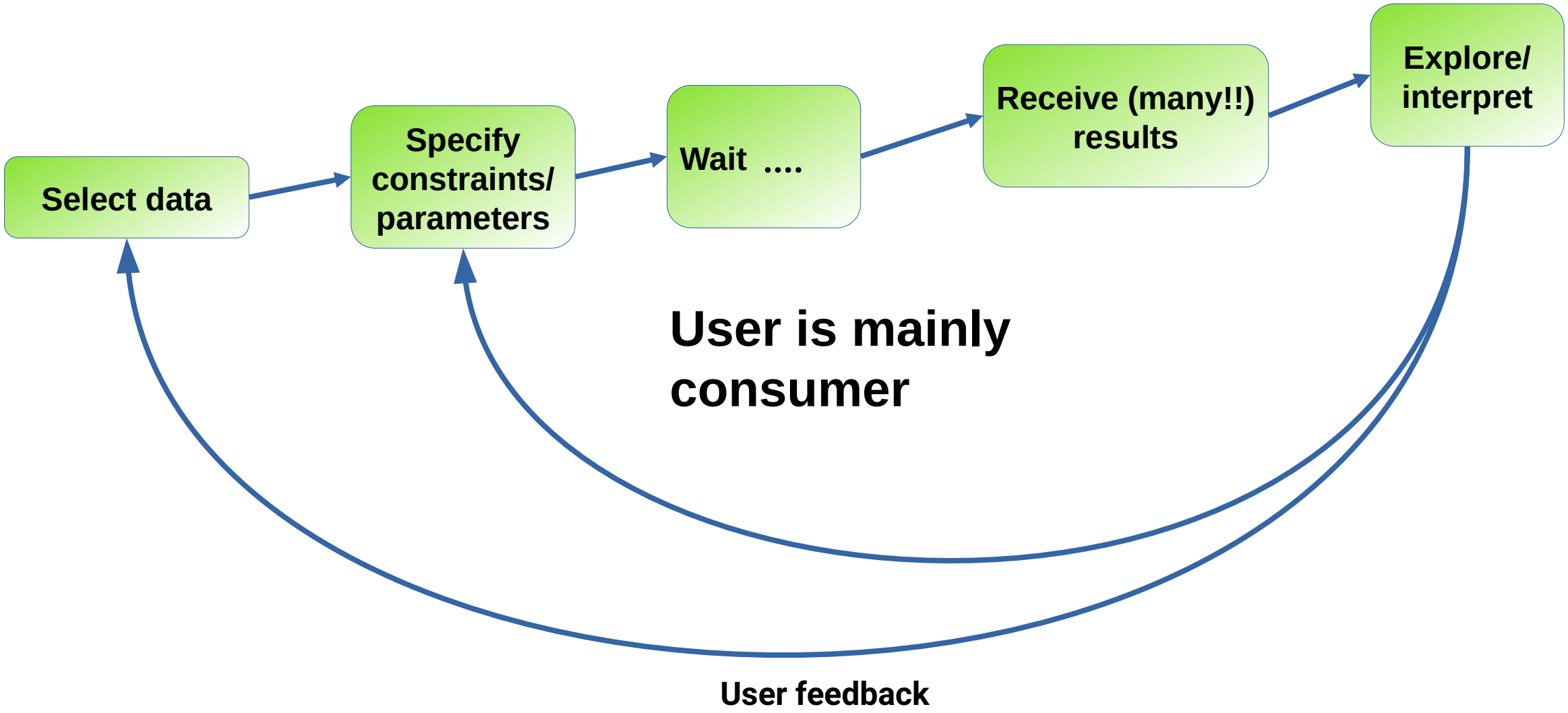
The classical knowledge discovery process



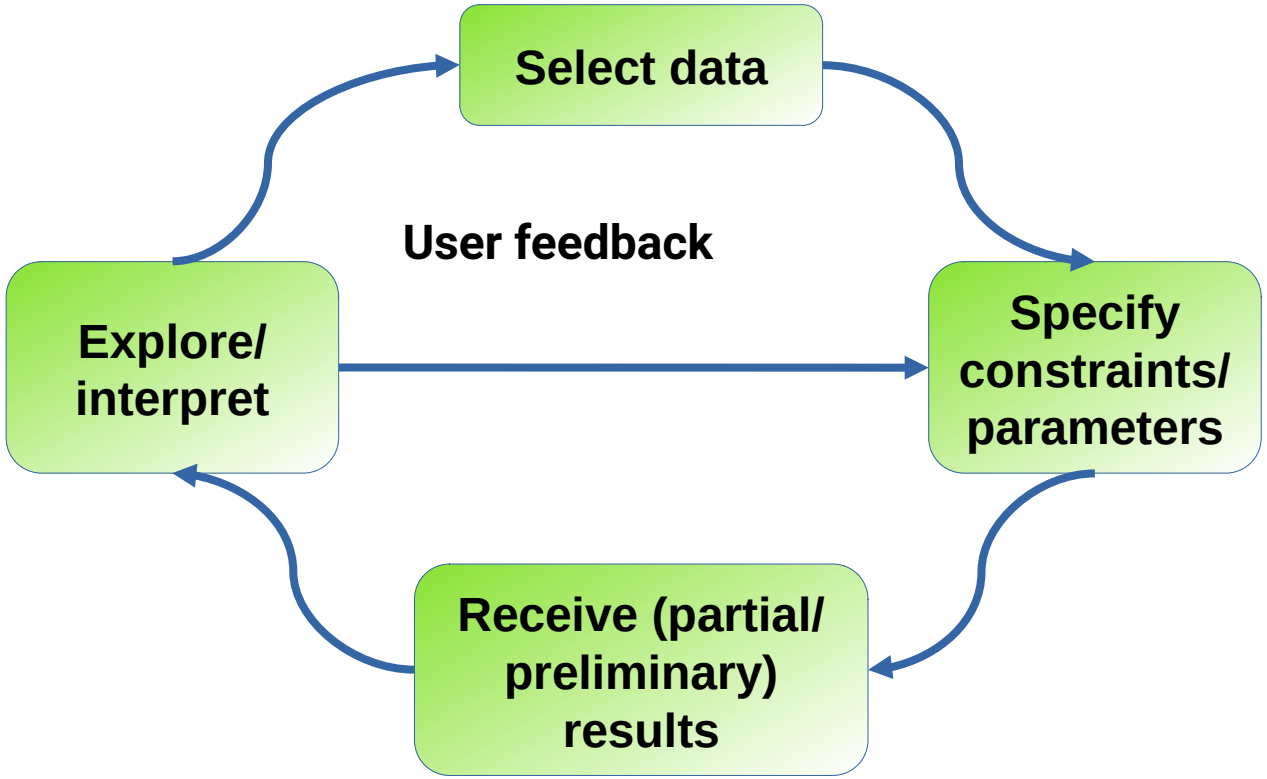
The classical « loop » (2)



The classical « loop » (2)

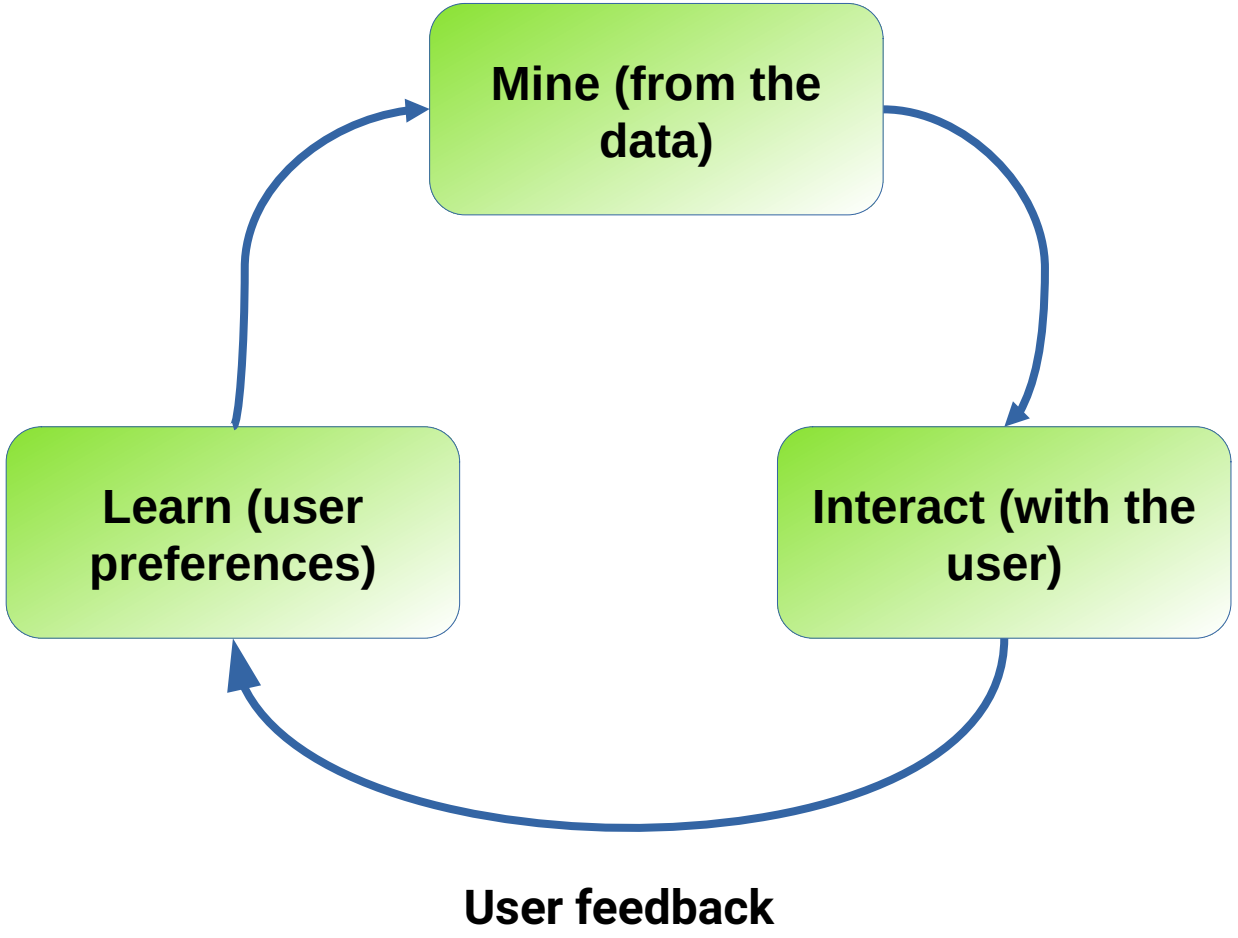


The iterative mining loop



User becomes a much more important part of the machinery

The implicit iterative mining loop



User becomes an integral part of the machinery
→ becomes implicit constraint/parameter adjustment component
→ becomes implicit data selection module

Literature blue print

1. Mining/sampling module
2. Patterns represented by descriptors (binary, e.g. item presence, and/or numerical, e.g. frequency)
3. Users feedback translated into vector labels
4. Numerical (preference) function learned

Requirements to put the human in the loop usefully

1. Result representation
2. Feedback option
3. Translation of feedback into internal model
4. Appropriate approximation of preference function
5. Correct internal model of the user

Result representation

- (Ordered) lists of patterns
  - Ordered by what ?
  - Requires user to relate them to each other
- Augmented with (some) statistics
  - Requires to keep background stats in mind
  - Not too many as to not overwhelm – not too few/which ones ?

Result representation

- (Ordered) lists of patterns
  - Ordered by what ?
  - Requires user to relate them to each other
- Augmented with (some) statistics
  - Requires to keep background stats in mind
  - Not too many as to not overwhelm – not too few/which ones ?

Order risks biasing user

Result representation

- (Ordered) lists of patterns
  - Ordered by what ?
  - Requires user to relate them to each other
- Augmented with (some) statistics
  - Requires to keep background stats in mind
  - Not too many as to not overwhelm – not too few/which ones ?

Order risks biasing user

Larger values risk introducing bias

Result representation

- + Data
  - Whole data set ?
  - Linked to pattern ?
    - Rest of data hidden ?
  - Dimensionality problem
    - Original presentation ?
    - PCA or similar ?
- Lots of literature on interactive visual clustering analysis...not for now

Result representation

- + Data
  - Whole data set ?
  - Linked to pattern ?
    - Rest of data hidden ?
  - Dimensionality problem
    - Original presentation ?
    - PCA or similar ?
- Lots of literature on interactive visual clustering analysis...not for now

Uncovered data might be even more interesting

Result representation

- + Data
  - Whole data set ?
  - Linked to pattern ?
    - Rest of data hidden ?
  - Dimensionality problem
    - Original presentation ?
    - PCA or similar ?
- Lots of literature on interactive visual clustering analysis...not for now

Uncovered data might be even more interesting

Change presentation based on pattern ?

Have the user learn the representation ?

- Enforces certain thought patterns (like a language)
- Out of sight, out of mind
- Clashes w/promise of democratic DM

Have the user learn the representation ?

- Enforces certain thought patterns (like a language)
- Out of sight, out of mind
- Clashes w/promise of democratic DM

Think of pattern list *vs* pattern graph

Feedback options

- Like/dislike – what's the meaning ?
  - Right vs wrong ?
  - Known vs unknown ?
  - Doesn't look interesting ?
  - Don't understand ?
- (Pairwise) ranking
  - But the list's already sorted
  - Meaning of low-ranked patterns ?

Feedback options

- Like/dislike – what's the meaning ?
  - Right vs wrong ?
  - Known vs unknown ?
  - Doesn't look interesting ?
  - Don't understand ?
- (Pairwise) ranking
  - But the list's already sorted
  - Meaning of low-ranked patterns ?

Telling the user ≠ being understood this way

Feedback options

- Like/dislike – what's the meaning ?
  - Right vs wrong ?
  - Known vs unknown ?
  - Doesn't look interesting ?
  - Don't understand ?
- (Pairwise) ranking
  - But the list's already sorted
  - Meaning of low-ranked patterns ?

Telling the user ≠ being understood this way

Harder to move things down

Feedback options

- Delete/filter patterns
  - Stronger than dislike but semantic problem stays
  - Affects data ?
- Tag for keeping
  - Should be taken into account in the future ?
- Tag for extending
- Create new descriptors
  - For the data ?
  - For patterns ?

Feedback options

- Delete/filter patterns
  - Stronger than dislike but semantic problem stays
  - Affects data ?
- Tag for keeping
  - Should be taken into account in the future ?
- Tag for extending
- Create new descriptors
  - For the data ?
  - For patterns ?

E.g. delete data containing patterns ?
Or involved attributes ?

Feedback options

- Delete/filter patterns
  - Stronger than dislike but semantic problem stays
  - Affects data ?
- Tag for keeping
  - Should be taken into account in the future ?
- Tag for extending
- Create new descriptors
  - For the data ?
  - For patterns ?

E.g. delete data containing patterns ?
Or involved attributes ?

Interaction with presentation ?

Feedback options

- Select data
  - Give more weight to these data ?
  - Work *only* on this data ?
  - Effect on prior patterns ?
- Explicit constraint adjustment
  - Kind of what we wanted to avoid

Feedback options

- Select clusters
  - → Like ? Equal to selecting data ?
- Customizing/splitting/merging clusters (Geono-Cluster, Das et al. ‘20)
  - Do they make algorithmic sense ?
- Change feature weights ?
  - Effects on data ?
  - On patterns ?
  - On presentation ?

Feedback options

- Select clusters
  - → Like ? Equal to selecting data ?
- Customizing/splitting/merging clusters (Geono-Cluster, Das et al. ‘20)
  - Do they make algorithmic sense ? **How to deal w/it if not ?**
- Change feature weights ?
  - Effects on data ?
  - On patterns ?
  - On presentation ?

Feedback options

- Select clusters
  - → Like ? Equal to selecting data ?
- Customizing/splitting/merging clusters (Geono-Cluster, Das et al. ‘20)
  - Do they make algorithmic sense ?
- Change feature weights ?
  - Effects on data ?
  - On patterns ?
  - On presentation ?

How to deal w/it if not ?

Revisit old patterns ?

Feedback options

- Undo ?
  - How far back ?
  - What’s it mean ?

Feedback options

- Undo ?
  - How far back ?
  - What’s it mean ?

Do we roll back the learned preference function ?

Have the user learn the feedback options ?

1. What if none is wanted ?
   - Learn work-arounds ?
2. Additional predefined options ?
3. Limits ways of thinking about pattern interestingness
4. How about « demonstration-based interaction » ?

Have the user learn the feedback options ?

1. What if none is wanted ?
   - Learn work-arounds ?
2. Additional predefined options ?
3. Limits ways of thinking about pattern interestingness
4. How about « demonstration-based interaction » ?

Facebook has 7 options and most of the time I STILL don't know which one applies

Have the user learn the feedback options ?

1. What if none is wanted ?
   - Learn work-arounds ?
2. Additional predefined options ?
3. Limits ways of thinking about pattern interestingness
4. How about « demonstration-based interaction » ?

Facebook has 7 options and most of the time I STILL don't know which one applies

Learned feedback options from user ?

Translation

1. Cannot-/must-link constraints
   - Explicit : fine but limiting feedback
   - Implicit : truly what the user meant ?
2. Weights for elements/features/descriptive statistics
   - Equal weight = equal importance ?
3. Classification examples
   - Depends on meaning of feedback labels
4. Ranking examples

Preference/quality approximation ?

1. Regression/classification function
   - Linear ?
   - Multiplicative ?
   - Cannot model complex relationships
2. A single decider enough ?
3. « Don’t know » needed ?
4. Set of instance level constraints ?
   - Encodes all the possible information ?

Ensembles ?

Preference/quality approximation ?

1. Regression/classification function
   - Linear ?
   - Multiplicative ?
   - Cannot model complex relationships

Why not choose more powerful learners, e.g. trees ?

2. A single decider enough ?

Ensembles ?

3. « Don’t know » needed ?
4. Set of instance level constraints ?
   - Encodes all the possible information ?

Correct internal model ?

1. Does user know what he/she's looking for ?
   - Something frequent ?
   - Something unexpected ?
   - Something counter-intuitive ?
2. Are they too locked into what they look for ?
   - → exploration/exploitation dilemma
3. Can they tell random noise from structure ?
   - Calibrate to user ?

Which leads me to….

Can we put the human in the loop ?

Which leads me to….

Can we put the human in the loop ?

In my opinion, right now : NO !

Which leads me to….

Can we put the human in the loop ?

In my opinion, right now : NO !

Solving it’s gonna be hard because the problems are not DM-expertise...

Meta-questions (h/t Bruno Crémilleux)

1) User feedback is expensive

   1) Can we define/develop oracles to simulate the user ?

   2) Could it be easier to have the user react more generally to a « results », e.g. a cluster, and propagate this to patterns ?

2) How to best/convingly evaluate iterative mining systems ?

3) How can pattern sampling best exploited for iterative mining ? Only to speed up things ?

4) How could we combine implicit user preferences (like the learned approximations), preferences already encoded in the data (like ranked labels), and subjective interestingness in the form of background knowledge about the data as given by users ?

Literature list

- D. Xin et al., Discovering Interesting Patterns Through User's Interactive Feedback. 2006
- S. Rueping, Ranking Interesting Subgroups. 2009
- E. Galbrun & P. Miettinen, A Case of Visual and Interactive Data Analysis: Geospatial Redescription Mining. 2012
- V. Dzyuba & M. van Leeuwen. Interactive Discovery of Interesting Subgroup Sets. 2013
- V. Dzyuba et al., Interactive Learning of Pattern Rankings. 2014
- M. van Leeuwen, Interactive Data Exploration using Pattern Mining. 2014
- M. Boley et al., One Click Mining - Interactive Local Pattern Discovery through Implicit Preference and Performance. 2016

Literature list (cont.)

- M. Bhuyian & M. Al Hasan, Interactive Knowledge Discovery from Hidden Data through Sampling of Frequent Patterns. 2016
- M. Bhuyian & M. Al Hasan, PRIIME: A Generic Framework for Interactive Personalized Interesting Pattern Discovery. 2016
- V. Dzyuba & M. van Leeuwen, Learning what matters - Sampling interesting patterns. 2017
- A. Giacometti & A. Soulet, Interactive Pattern Sampling for Characterizing Unlabeled Data. 2017
- K. Puolamäki et al., Interactive visual data exploration with subjective feedback: an information-theoretic approach. 2020
- S. Das et al., Geono-cluster: Interactive visual cluster analysis for biologists. 2020