Dense Neighborhood Pattern Sampling in Numerical Data

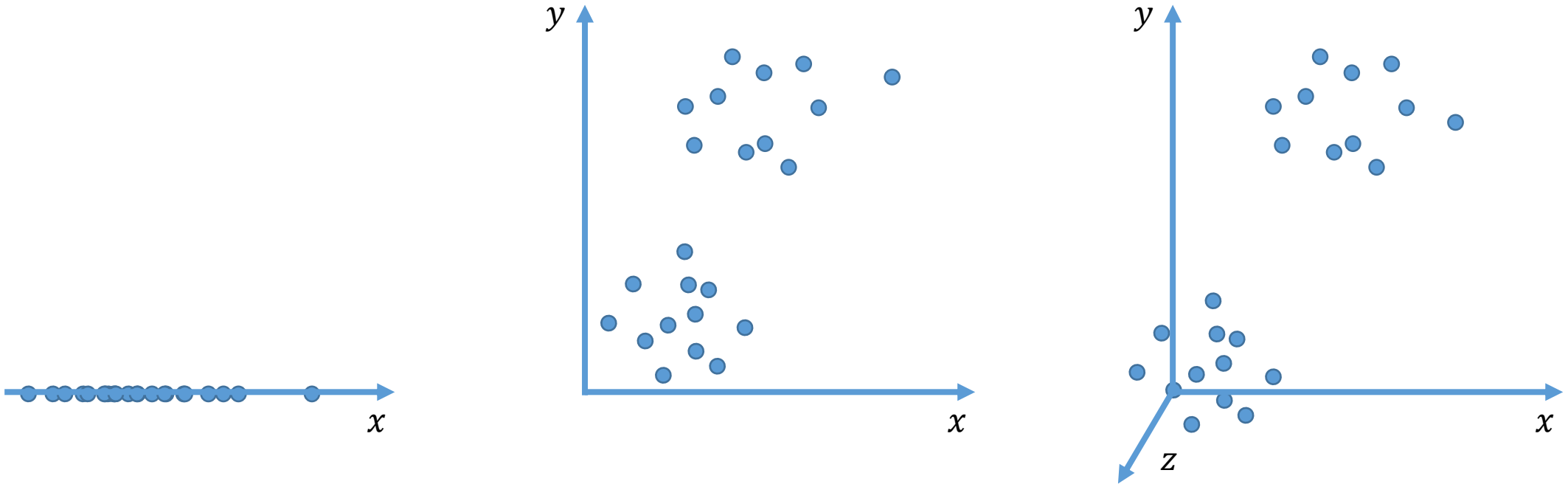
Arnaud Giacometti and Arnaud Soulet

University of Tours, France

Outlines

- Pattern mining in numerical data
- Neighborhood pattern sampling problem
- Three-step random procedure
- Experimental study on UCI benchmarks

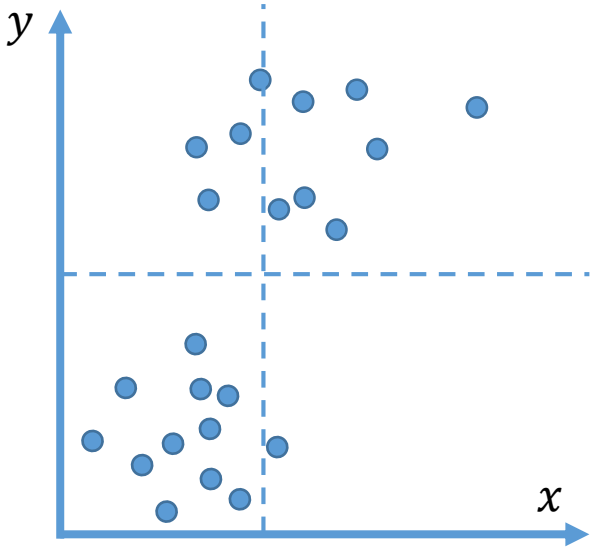
Pattern mining in numerical data?



Finding interesting patterns in all subspaces

(pattern = an anomalously high local density of data points [Hand, 2002])

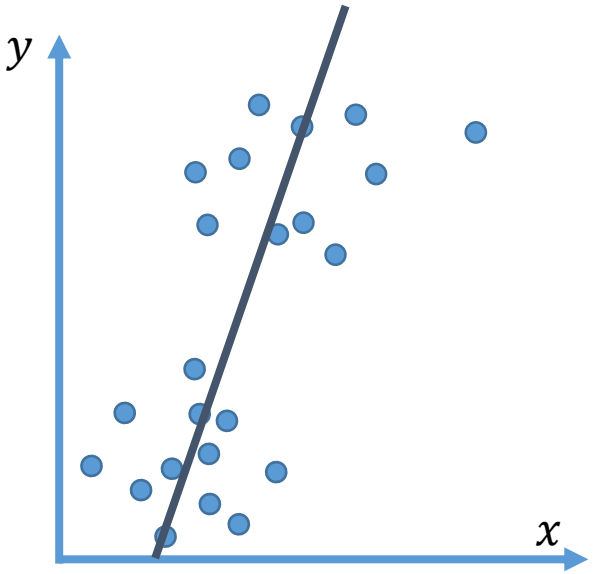
Pattern mining in numerical data?



Discretized patterns

[Srikant et al., 1996]

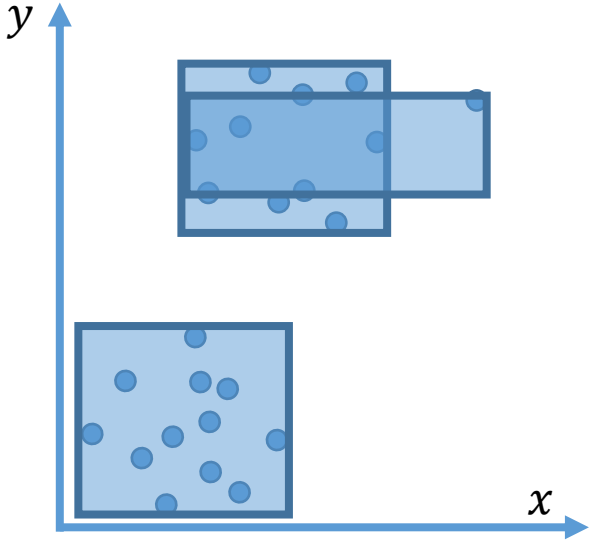
[Fukuda et al., 2001]



Subspace patterns

[Jaroszewicz et al., SDM07]

[Tatti, ICDM13]

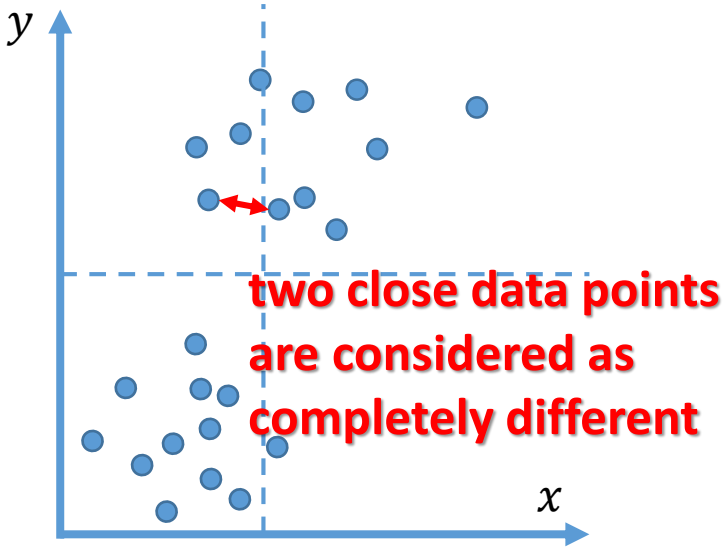


Interval patterns

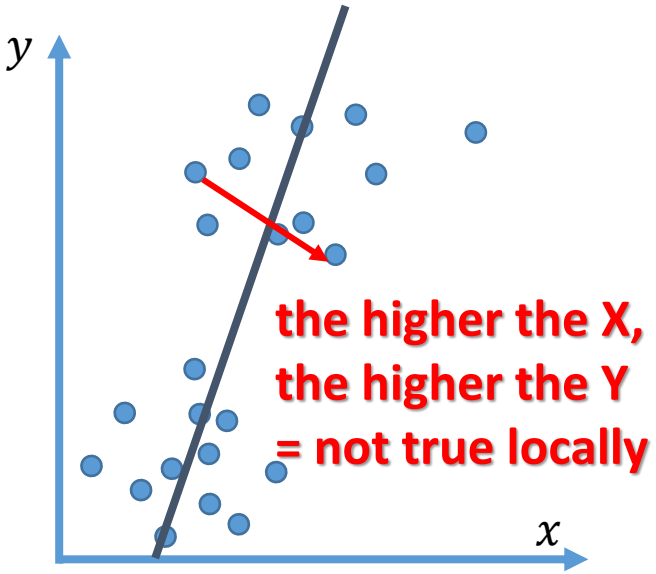
[Kaytoue et al., IJCAI11]

[Belfodil et al., IJCAI17]

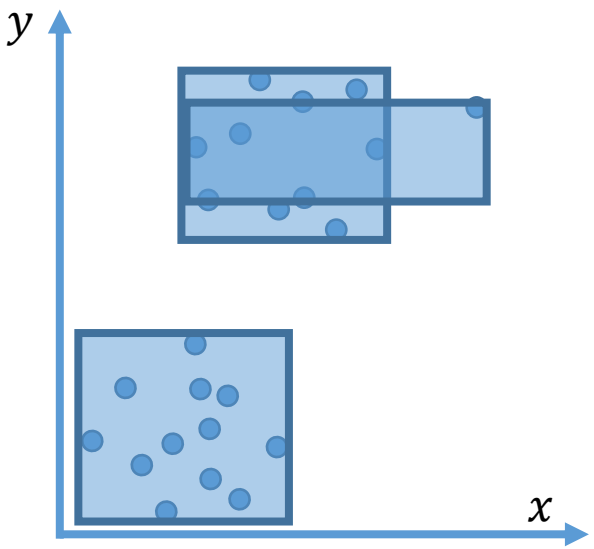
Pattern mining in numerical data?



Discretized patterns



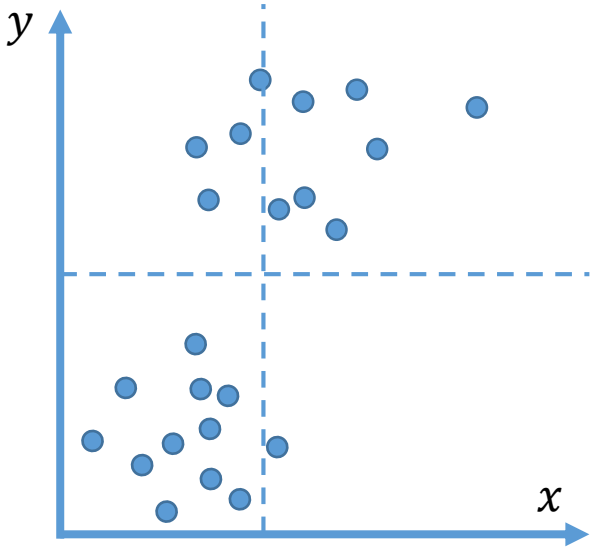
Subspace patterns



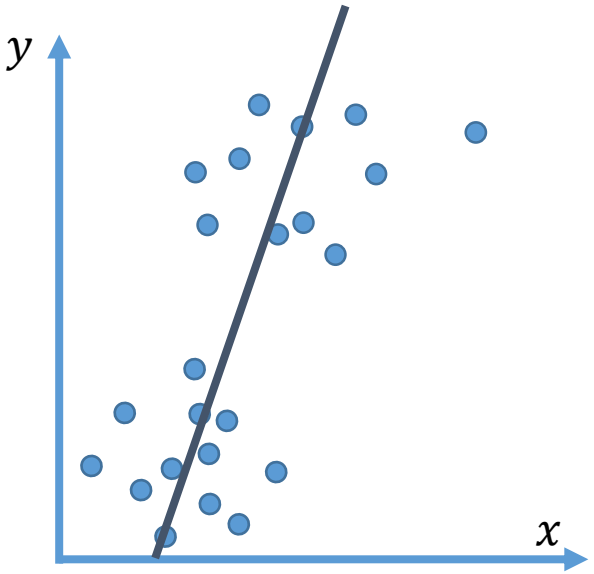
Interval patterns

loss of information

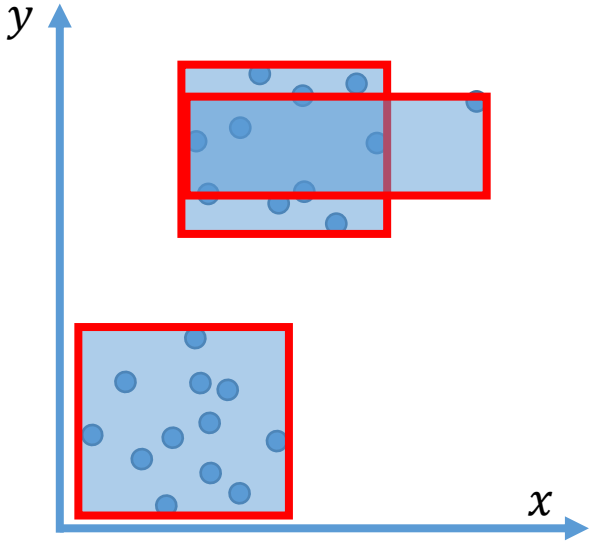
Pattern mining in numerical data?



Discretized patterns



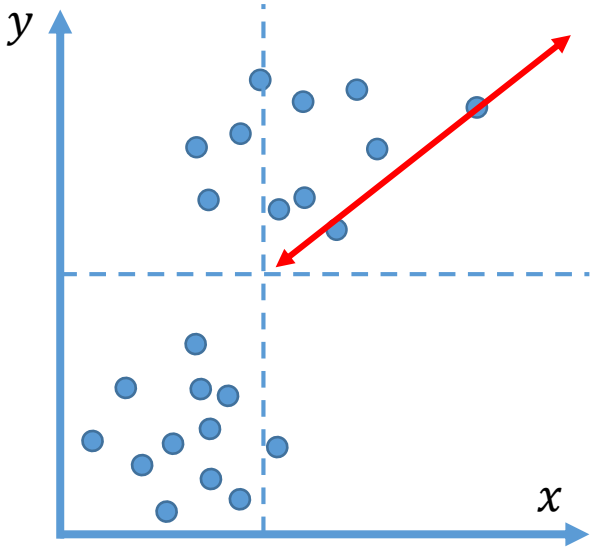
Subspace patterns



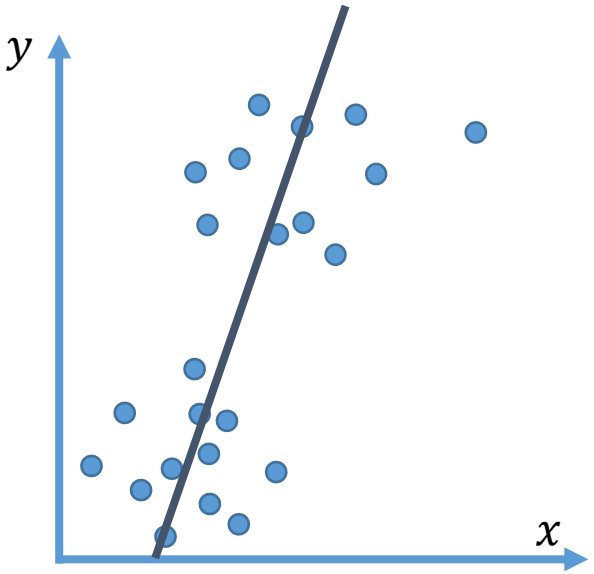
Interval patterns

loss of information + combinatorial explosion

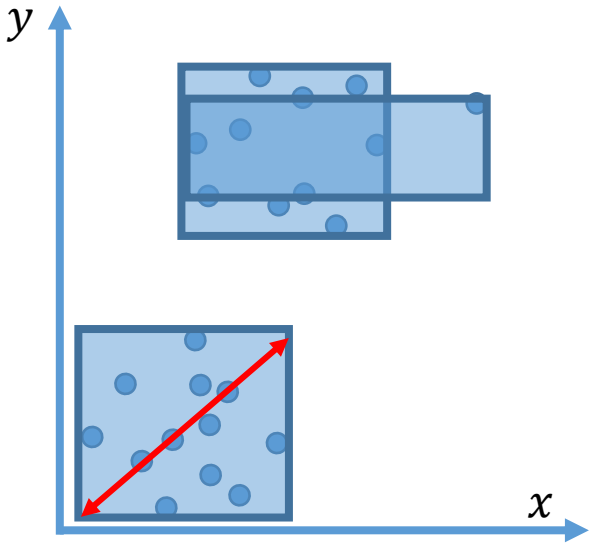
Pattern mining in numerical data?



Discretized patterns



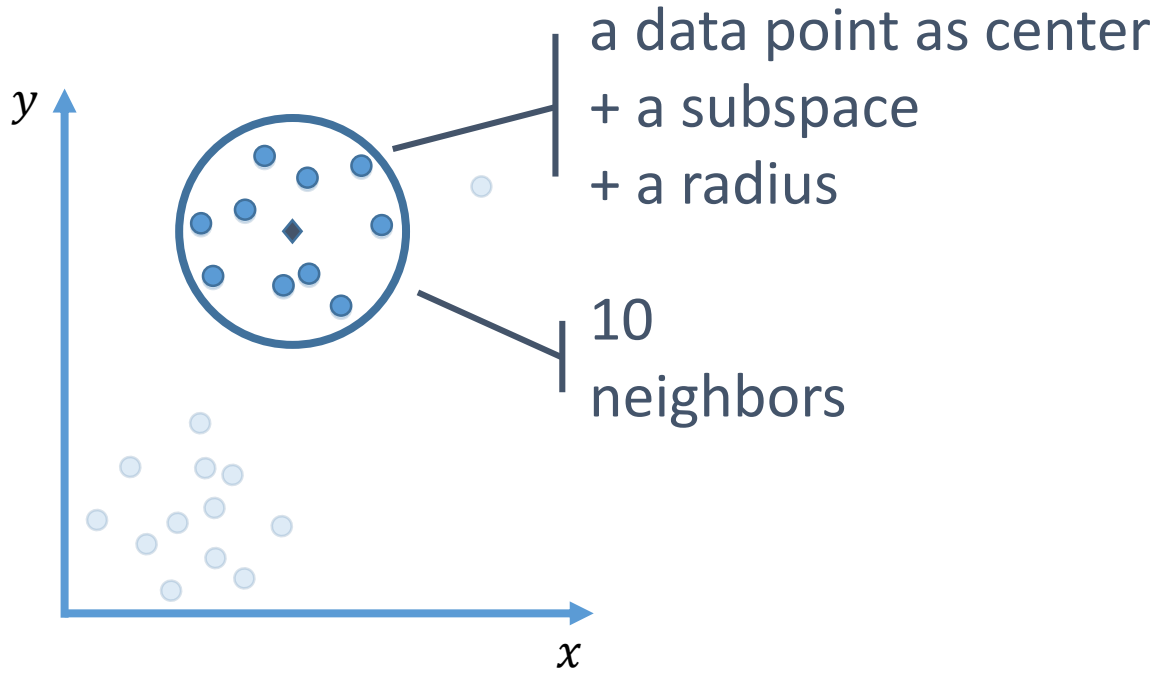
Subspace patterns



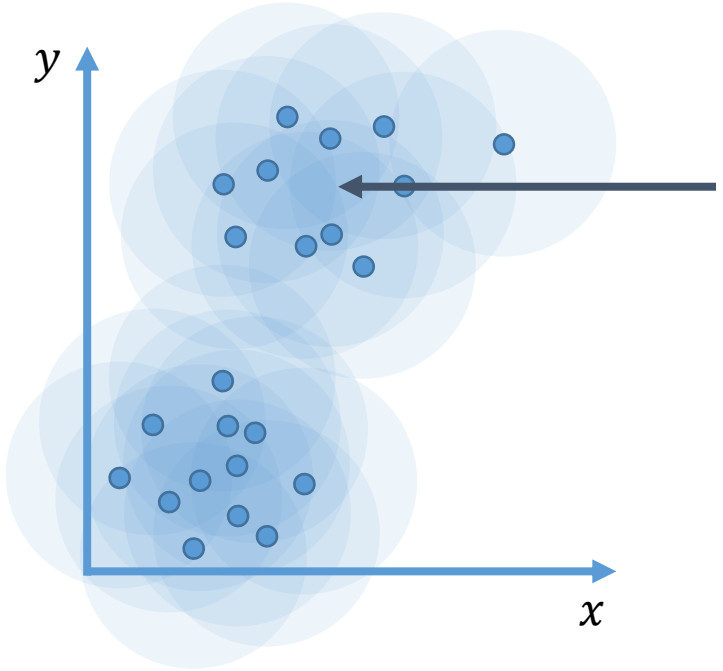
Interval patterns

loss of information + combinatorial explosion + curse of dimensionality

Neighborhood patterns (for contrast mining)
[Konijn et al., PAKDD13]

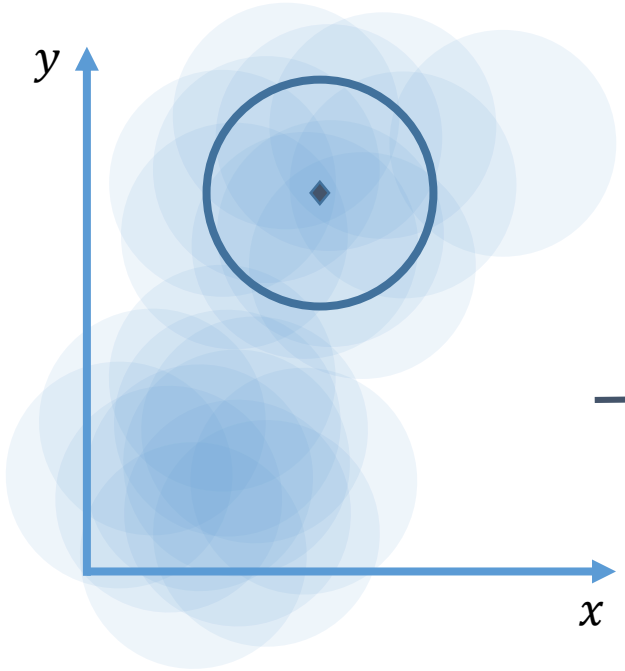


Limit of neighborhood patterns
[Konijn et al., PAKDD13]



very good place
for a center…
but no data point!

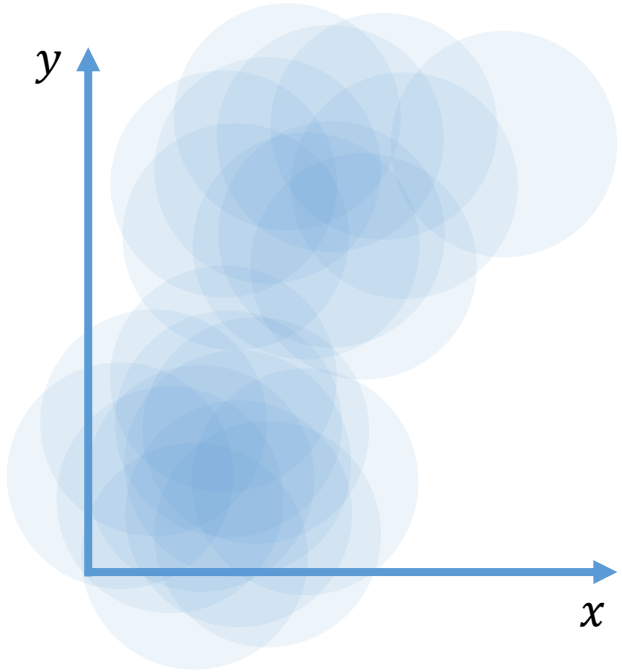
Challenge of neighborhood patterns = infinite search space



we have to consider the fullspace as language

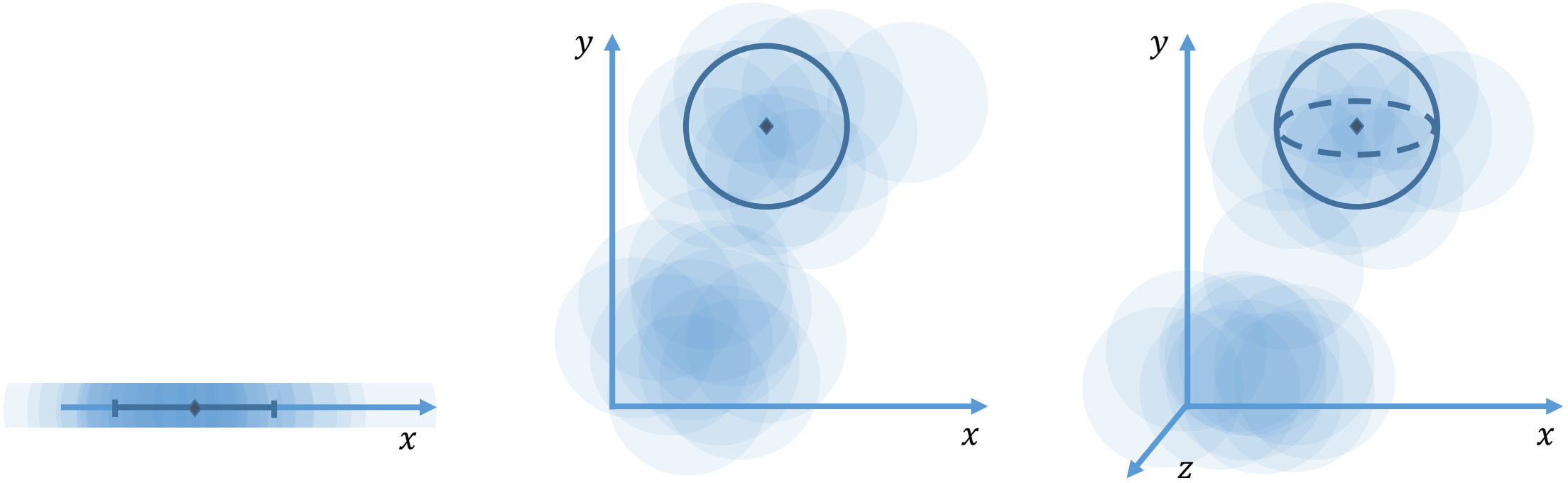
Neighborhood pattern sampling problem

Return a random neighborhood pattern with a probability proportional to its number of neighbors (given a radius)



Neighborhood pattern sampling problem

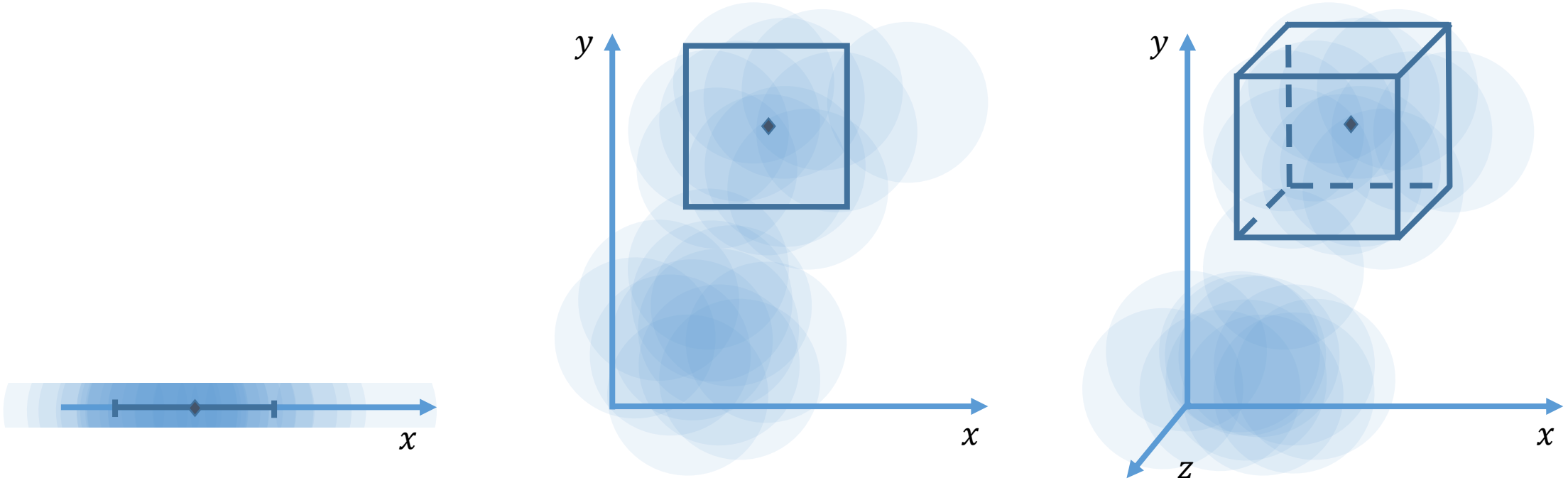
Return a random neighborhood pattern with a probability proportional to its ~~number of neighbors~~ density (given a radius)



Normalizing by the volume for dealing with different **subspaces**

Neighborhood pattern sampling problem

Return a random neighborhood pattern with a probability proportional to its ~~number of neighbors~~ density (given a radius)



Normalizing by the volume for dealing with different **p-norms**

Methods for pattern sampling

Stochastic procedure

- Output Space Sampling for Graph Patterns [Hasan et Zaki, VLDB09]
- Fast Query Execution for Retrieval Models Based on Path-Constrained Random Walks [Lao et Cohen, KDD10]

Two-step random procedure

- Direct local pattern sampling by efficient two-step random procedures [Boley et al., KDD11]
- Instant Exceptional Model Mining Using Weighted Controlled Pattern Sampling [Moens et Boley, IDA14]

SAT procedure

- Flexible constrained sampling with guarantees for pattern mining [Dzyuba et al., DAMI17]

Two-step random procedure
[Boley et al., KDD11]

TId	Items
1	A B C
2	A B
3	B C
4	C

freq(AB) = 2
freq(AC) = 1

Frequent itemset sampling: return a random itemset with a probability proportional to its frequency

(AB has twice more chance to be drawn than AC.)

Two-step random procedure
[Boley et al., KDD11]

TId	Items
1	A B C
2	A B
3	B C
4	C

Weight ω
$7 = 2^3-1$
$3 = 2^2-1$
$3 = 2^2-1$
$1 = 2^1-1$

❶

TId	Itemsets
1	A, B, C, AB, AC, BC, ABC
2	A, B, AB
3	B, C, BC
4	C

❶ Calculate the number of itemsets per transaction

Two-step random procedure
[Boley et al., KDD11]

TId	Items
1	A B C
2	**A B**
3	B C
4	C

Weight ω
$7 = 2^3-1$
$\mathbf{3 = 2^2-1}$
$3 = 2^2-1$
$1 = 2^1-1$

❶

TId	Itemsets
1	A, B, C, AB, AC, BC, ABC
2	**A, B, AB**
3	B, C, BC
4	C

❶ Draw a transaction $t \in \mathcal{D}$ with a probability proportional to the number of itemsets contained in t

Two-step random procedure
[Boley et al., KDD11]

TId	Items
1	A B C
2	**A B**
3	B C
4	C

Weight ω
$7 = 2^3-1$
$\mathbf{3 = 2^2-1}$
$3 = 2^2-1$
$1 = 2^1-1$

❶

TId	Itemsets
1	A, B, C, AB, AC, BC, ABC
2	**A, B, AB**
3	B, C, BC
4	C

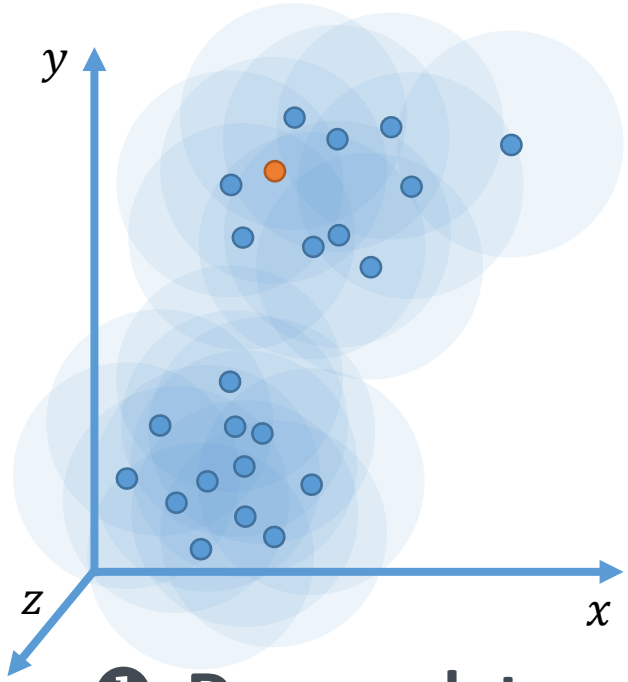
❷

Sample
A, A
B, B, B
C, C, C
AB, AB
AC
BC, BC
ABC

❶ **Draw a transaction $t \in \mathcal{D}$ with a probability proportional to the number of itemsets contained in t**

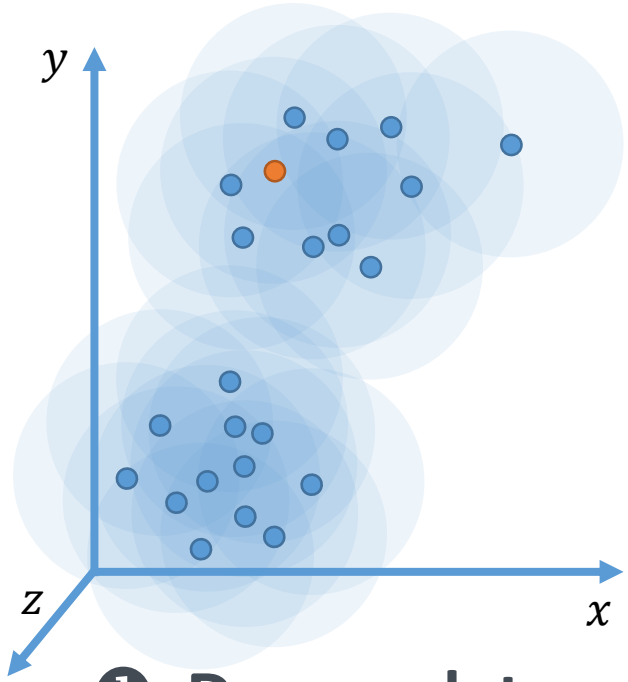
❷ **Draw uniformly an itemset from t**

Three-step random procedure (1)

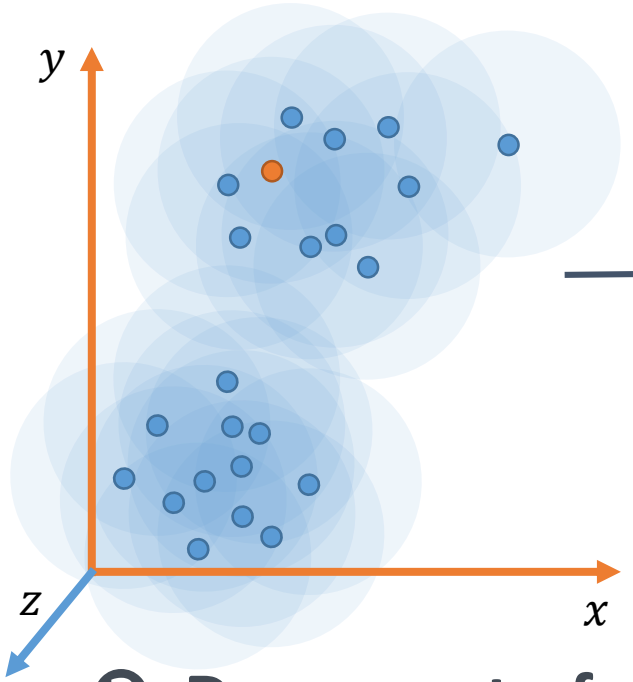


❶ **Draw a data point p**

Three-step random procedure (2)



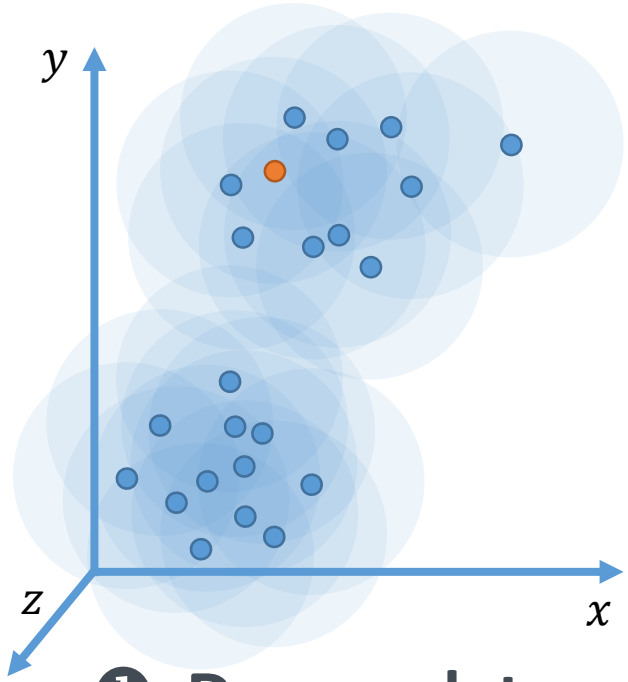
❶ **Draw a data point p**



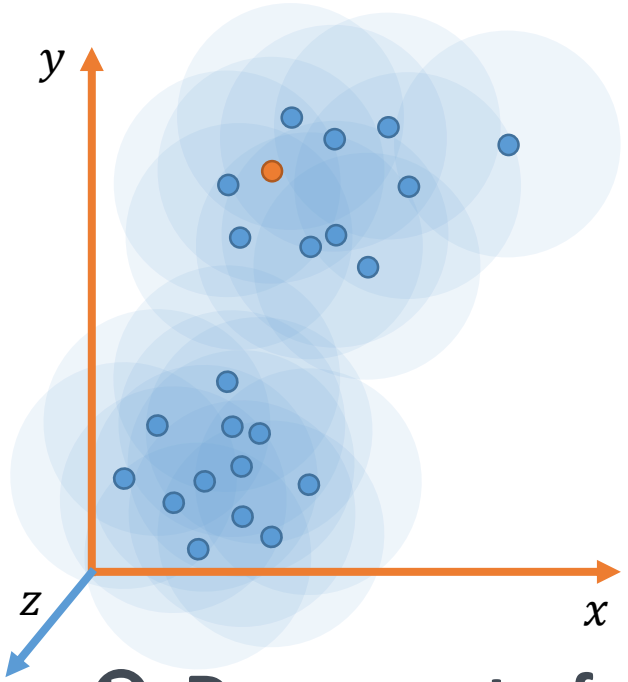
❷ **Draw a set of dimensions D**

we have to consider the fullspace as language!

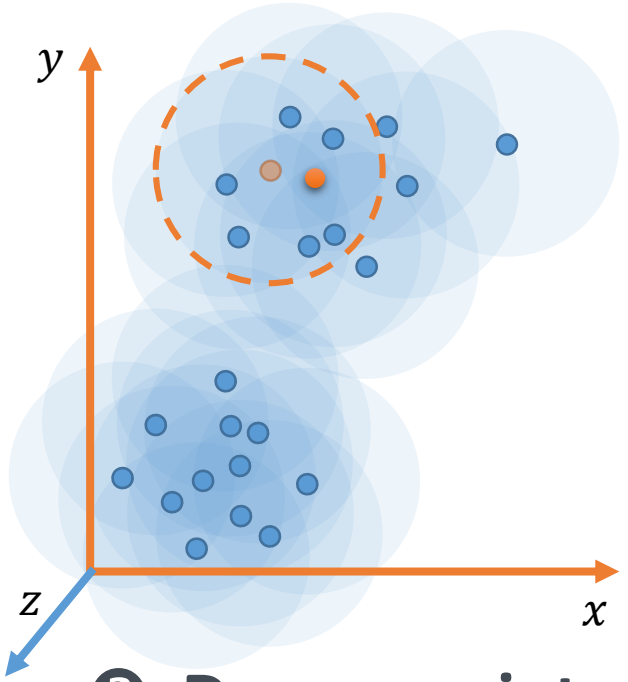
Three-step random procedure (3)



❶ Draw a data point p

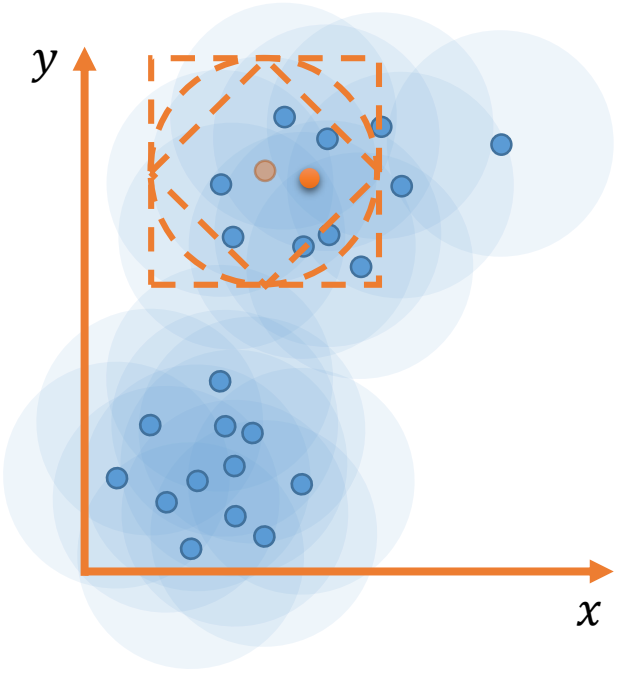


❷ Draw a set of dimensions D



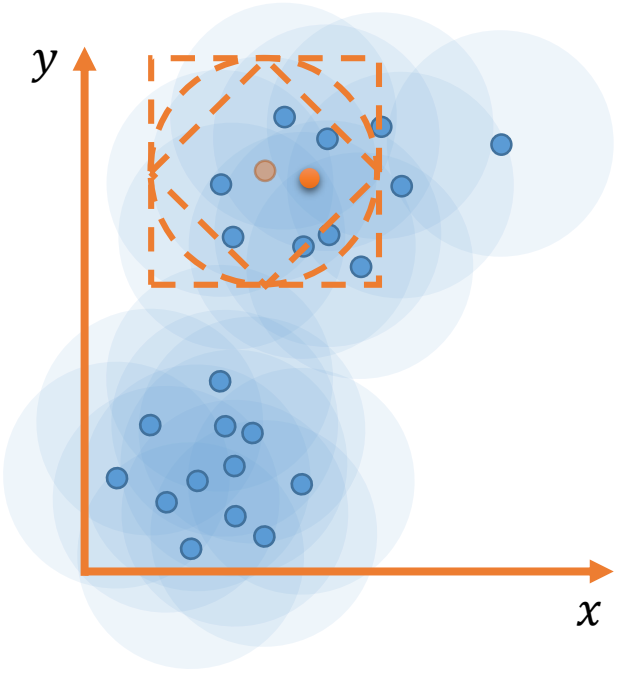
❸ Draw a point q in the neighborhood of $p[D]$

Third step: How?

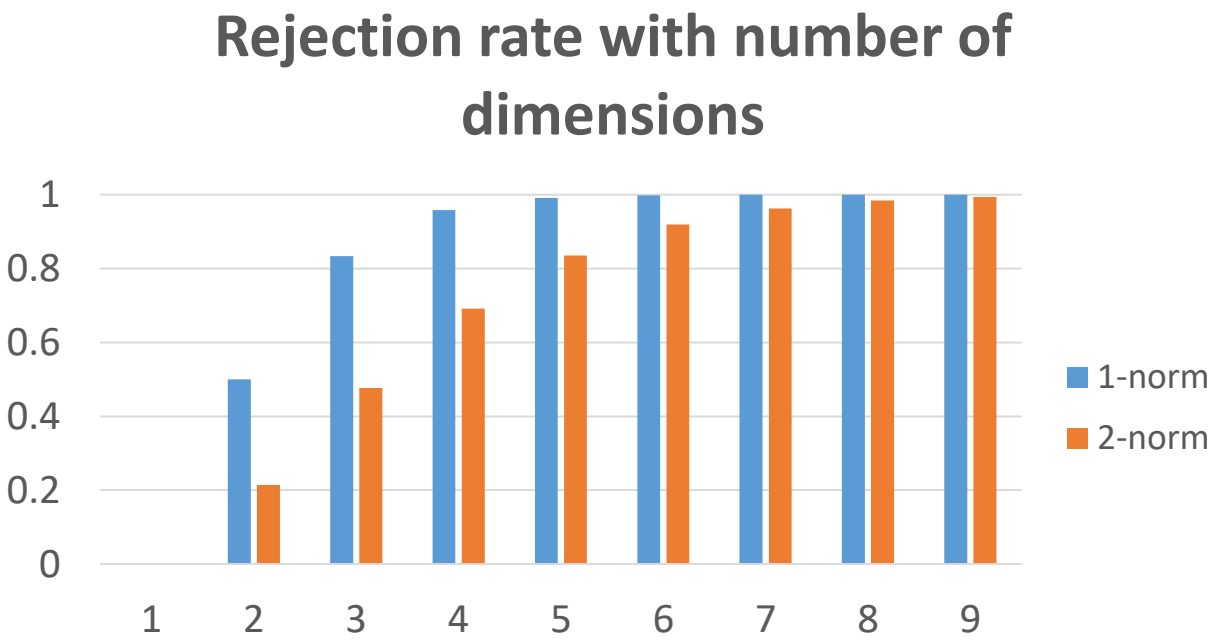


- ∞**-norm:** select uniformly a point in a hypercube
- **1-norm: ?**
- **2-norm: ?**

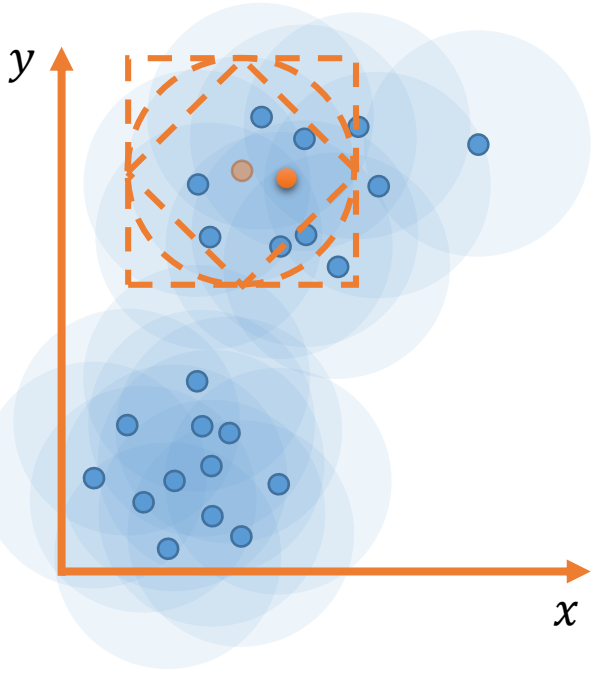
Third step: How?



- **∞-norm:** select uniformly a point in a hypercube
- **1-norm: ?**
- **2-norm: ?**



Third step: How?



- **∞-norm:** select uniformly a point in a hypercube
- **1-norm:** selecting a point from a unit simplex, uniformly at random [Smith and Tromble, 2004]
- **2-norm:** property of normal distribution where points are uniformly distributed on the unit k-sphere [Muller, 1959]

Contributions of three-step random procedure

- **Exact sampling method of neighborhood patterns with a probability proportional to its density** (given a radius and a p-norm)

- **Generalization of the two-step random procedure** (mixed data)

- **Third step is not costly:**
  - 1-norm: O(log(#data points) + #dimensions x log(#dimensions))
  - 2-norm and ∞-norm: O(log(#data points) + #dimensions)
  - In practice, average time per pattern: 70 ms

Experimental study on UCI benchmarks

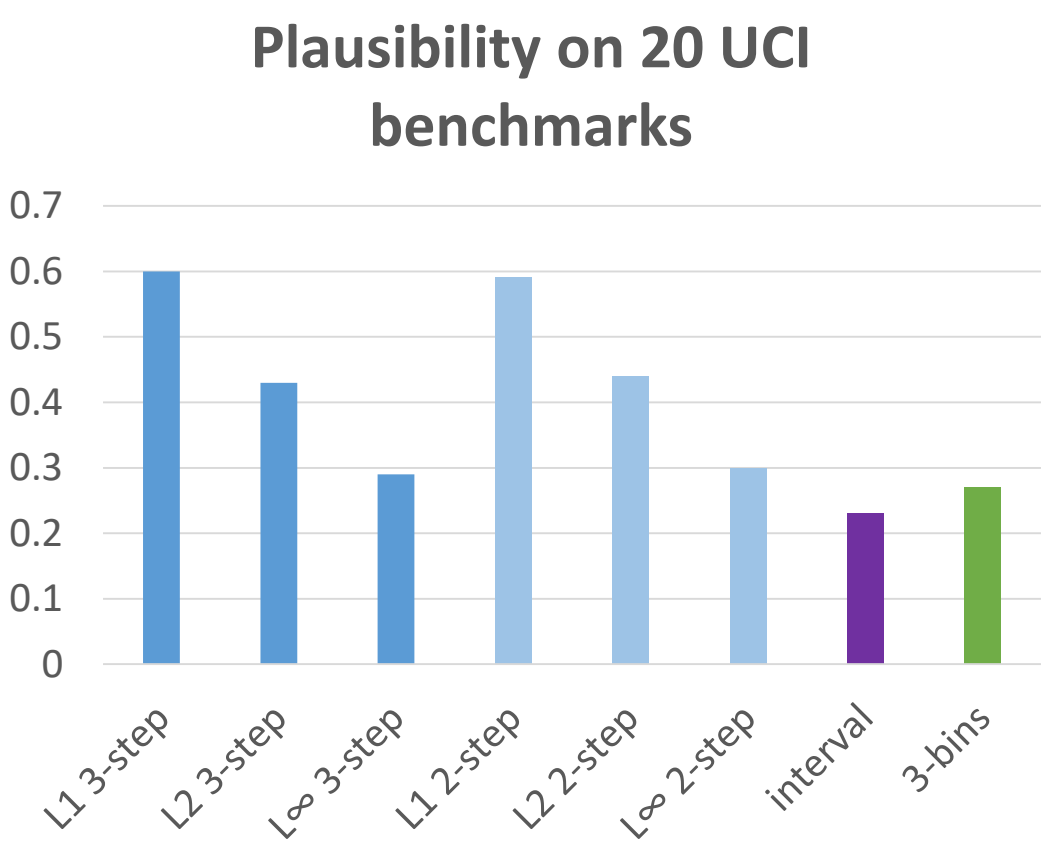
Protocol

- Z-score for numerical data
- Radius = 1

Questions

1. What is the proportion of patterns that we do not observe in a randomized dataset? (= plausibility)
2. What is the proportion of distinct patterns? (= diversity)
3. What is the accuracy of a sample-based associative classifier?

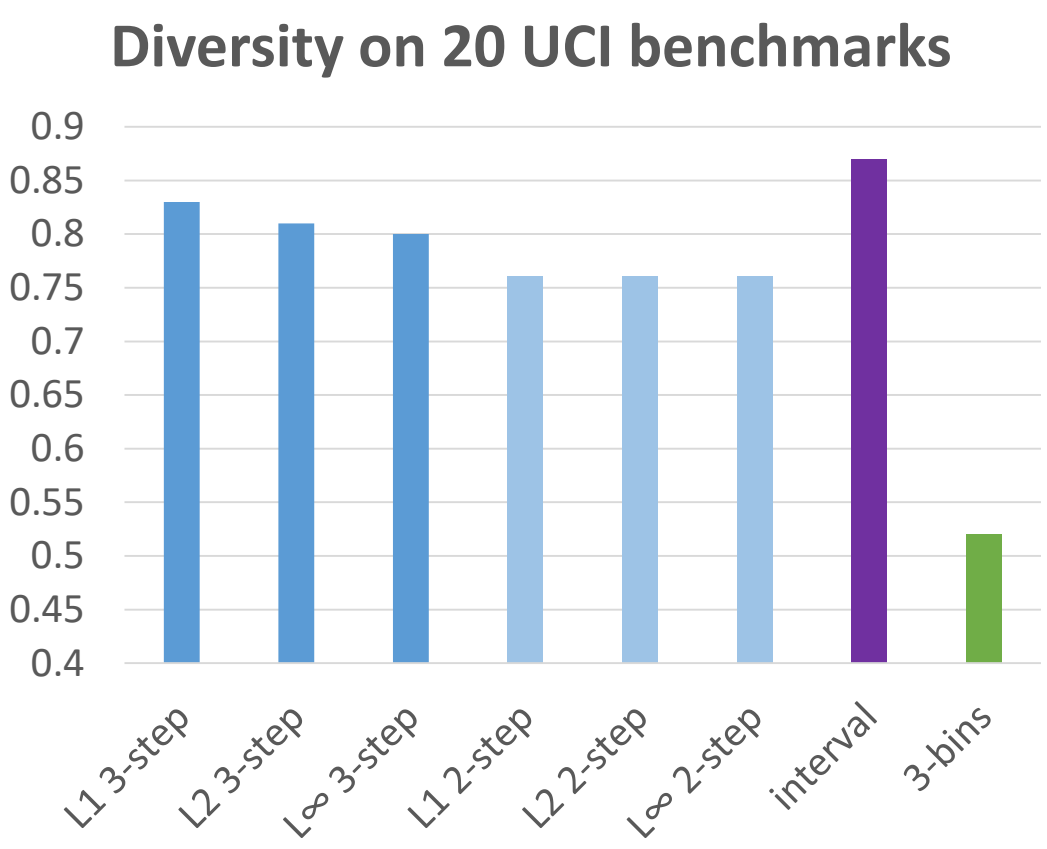
1. Plausibility of sampled patterns



Plausibility = proportion of patterns that we do not observe in a randomized dataset

- **L1 > L2 > L∞ > interval/3-bins**

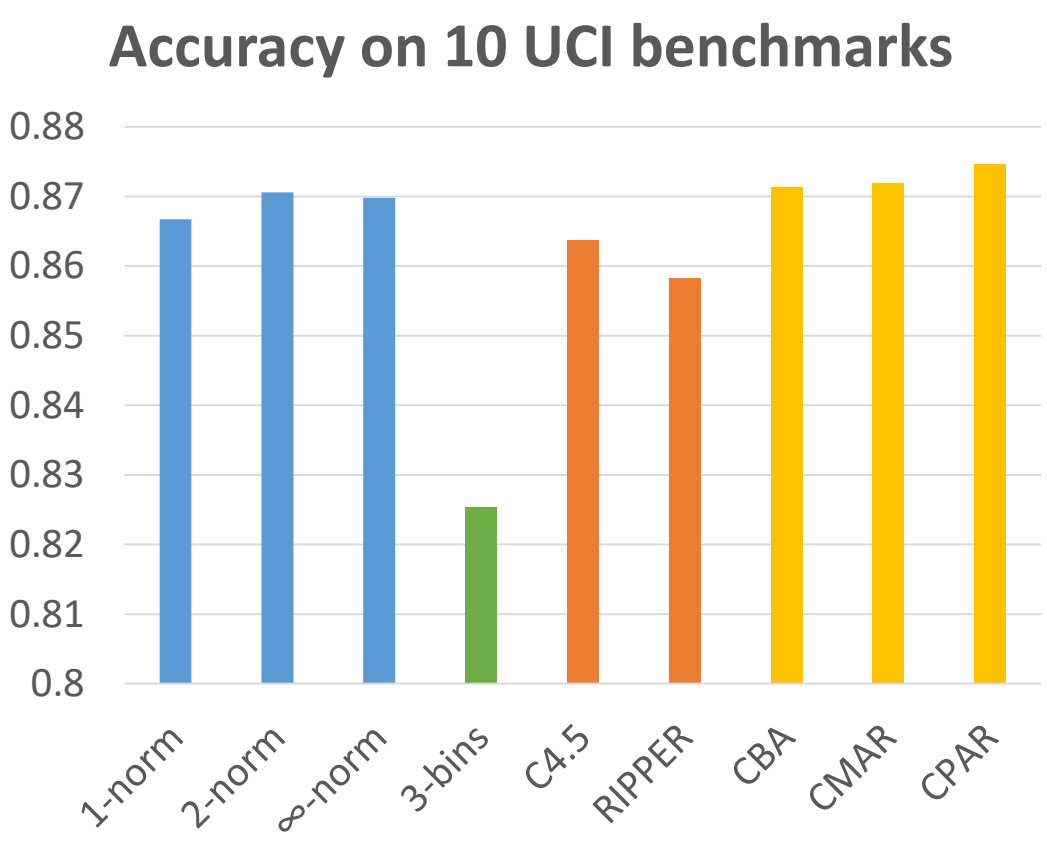
2. Diversity of sampled patterns



Diversity = proportion of distinct patterns

- **Diversity enhancement thanks to the third step**
- **Diversity as good as interval pattern mining**

3. Sampling-based associative classification



CBA-like method with a sample of 10k patterns

- **Accuracy as good as classic methods with a complete extraction**
- **Sampling-based classifier built in seconds**

Conclusion

- **Pattern sampling in numerical data:**
  - **Instant** discovery of patterns despite an infinite search space
  - **High quality** patterns
  - **Lossless** pattern mining

- **Next steps:**
  - Interactive pattern mining
  - Instant subspace clustering

Thank you!