August 29, 2017
Lunch Seminar @ NII

Introduction to Machine Learning

Mahito Sugiyama（杉山麿人）

Learning from Data
(from mlss.tuebingen.mpg.de/2013/schoelkopf_whatisML_slides.pdf)

- 1, 2, 4, 7, . . .
  - What are succeeding numbers?

Learning from Data

(from mlss.tuebingen.mpg.de/2013/schoelkopf_whatisML_slides.pdf)

- $1, 2, 4, 7, \ldots$
  - What are succeeding numbers?

$$1, 2, 4, 7, 11, 16, \ldots \qquad (a_n = a_{n-1} + n - 1)$$

Learning from Data

(from mlss.tuebingen.mpg.de/2013/schoelkopf_whatisML_slides.pdf)

- $1, 2, 4, 7, \ldots$
  - What are succeeding numbers?

$$1, 2, 4, 7, 11, 16, \ldots \quad (a_n = a_{n-1} + n - 1)$$

$$1, 2, 4, 7, 12, 20, \ldots \quad (a_n = a_{n-1} + a_{n-2} + 1)$$

Learning from Data

(from mlss.tuebingen.mpg.de/2013/schoelkopf_whatisML_slides.pdf)

- $1, 2, 4, 7, \ldots$
  - What are succeeding numbers?

$$1, 2, 4, 7, 11, 16, \ldots \quad (a_n = a_{n-1} + n - 1)$$

$$1, 2, 4, 7, 12, 20, \ldots \quad (a_n = a_{n-1} + a_{n-2} + 1)$$

$$1, 2, 4, 7, 13, 24, \ldots \quad (a_n = a_{n-1} + a_{n-2} + a_{n-3})$$

Learning from Data

(from mlss.tuebingen.mpg.de/2013/schoelkopf_whatisML_slides.pdf)

- $1, 2, 4, 7, \ldots$
  - What are succeeding numbers?

$1, 2, 4, 7, 11, 16, \ldots \quad (a_n = a_{n-1} + n - 1)$

$1, 2, 4, 7, 12, 20, \ldots \quad (a_n = a_{n-1} + a_{n-2} + 1)$

$1, 2, 4, 7, 13, 24, \ldots \quad (a_n = a_{n-1} + a_{n-2} + a_{n-3})$

$1, 2, 4, 7, 14, 28$ (divisors of 28)

Learning from Data

(from mlss.tuebingen.mpg.de/2013/schoelkopf_whatisML_slides.pdf)

- $1, 2, 4, 7, \ldots$
  - What are succeeding numbers?

$1, 2, 4, 7, 11, 16, \ldots$	$(a_n = a_{n-1} + n - 1)$
$1, 2, 4, 7, 12, 20, \ldots$	$(a_n = a_{n-1} + a_{n-2} + 1)$
$1, 2, 4, 7, 13, 24, \ldots$	$(a_n = a_{n-1} + a_{n-2} + a_{n-3})$
$1, 2, 4, 7, 14, 28$	(divisors of 28)
$1, 2, 4, 7, 1, 1, 5, \ldots$	(decimals of $\pi = 3.1415\ldots$, $e = 2.718\ldots$)

Learning from Data

(from mlss.tuebingen.mpg.de/2013/schoelkopf_whatisML_slides.pdf)

- $1, 2, 4, 7, \ldots$
  - What are succeeding numbers?

$1, 2, 4, 7, 11, 16, \ldots$	$(a_n = a_{n-1} + n - 1)$
$1, 2, 4, 7, 12, 20, \ldots$	$(a_n = a_{n-1} + a_{n-2} + 1)$
$1, 2, 4, 7, 13, 24, \ldots$	$(a_n = a_{n-1} + a_{n-2} + a_{n-3})$
$1, 2, 4, 7, 14, 28$	(divisors of 28)
$1, 2, 4, 7, 1, 1, 5, \ldots$	(decimals of $\pi = 3.1415\ldots$, $e = 2.718\ldots$)

- 1107 results (!) in the on-line encyclopedia (`https://oeis.org/`)

Remarkable Features of Machine Learning

- Which is the correct answer (or generalization) for succeeding numbers of $1, 2, 4, 7, \ldots$?
  - Any answer is possible!
  - There is no universal solution

Remarkable Features of Machine Learning

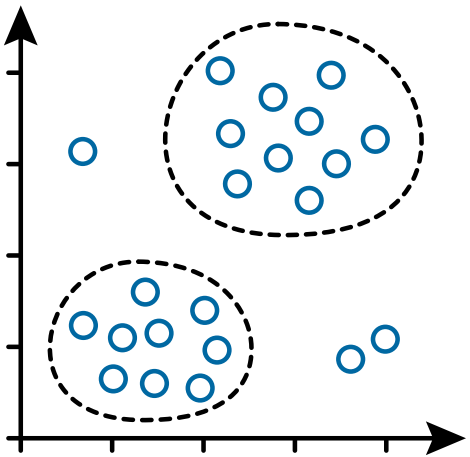
- Which is the correct answer (or generalization) for succeeding numbers of $1, 2, 4, 7, \ldots$?
  - Any answer is possible!
  - There is no universal solution

- Two features:
  (i) There are two agents (teacher and learner) in learning, which is different from "computation"
    - Results can be biased by a teacher (human being)

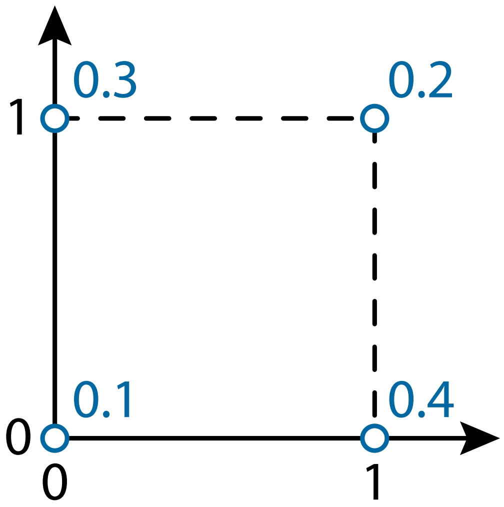
  (ii) Learning is an infinite process

Overview

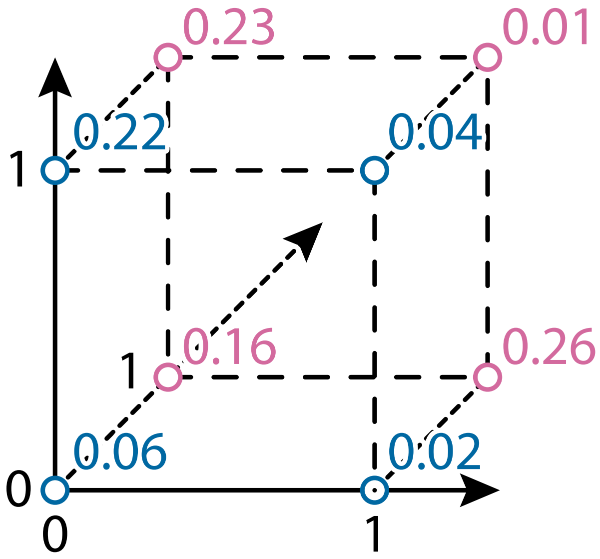
Clustering



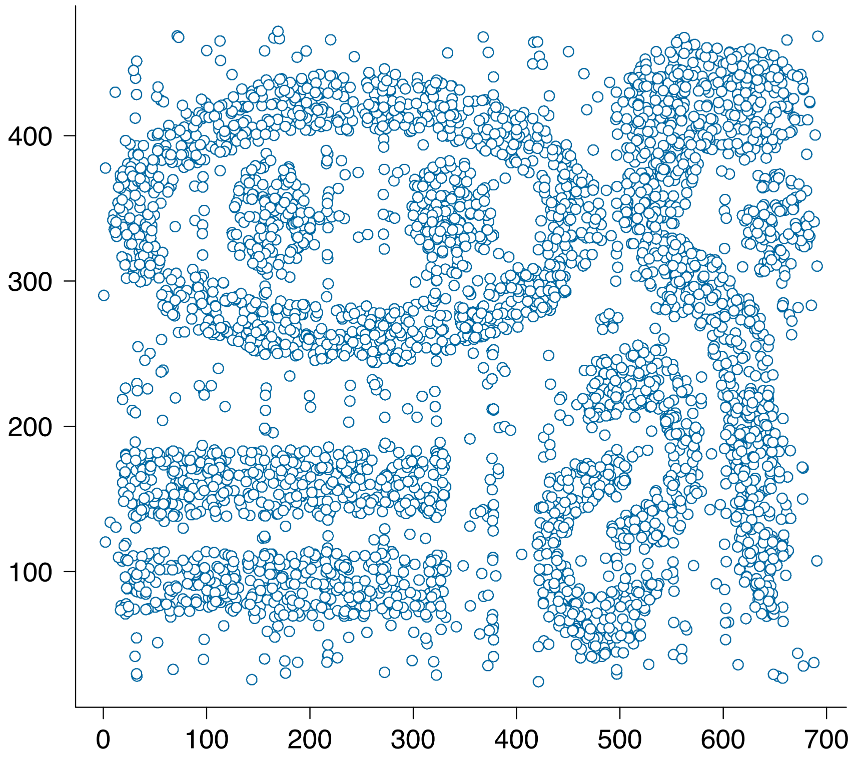
Pattern mining



Deep learning



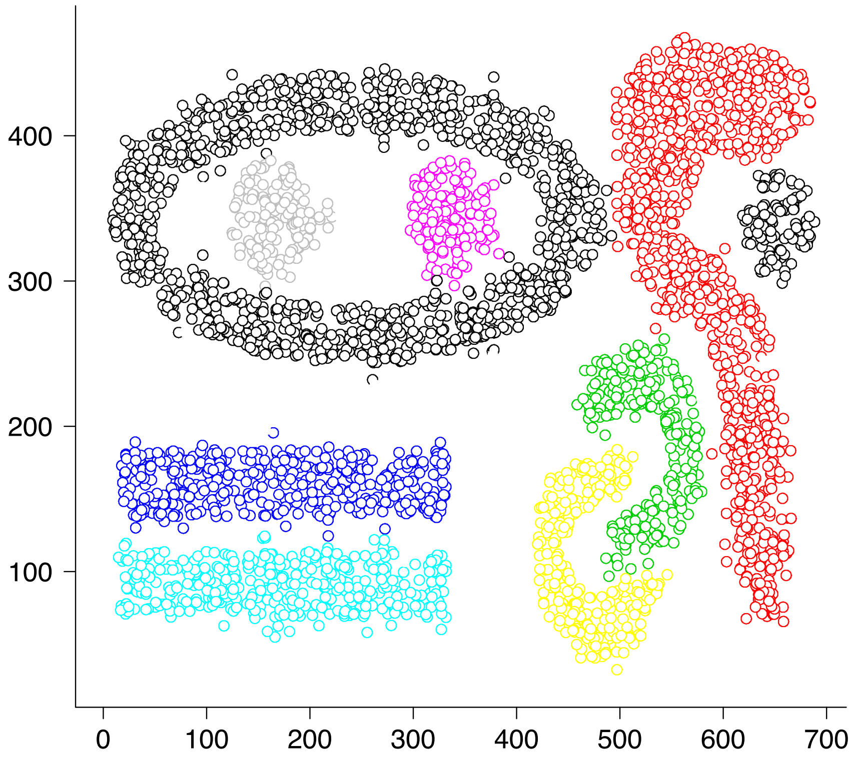
Clustering



- Data:

1	355.60	270.21
2	549.28	351.71
3	520.08	215.48
4	575.15	166.68
5	288.83	141.49
6	50.61	75.77
7	262.57	428.45
⋮	⋮	⋮
4000	309.395	365.09

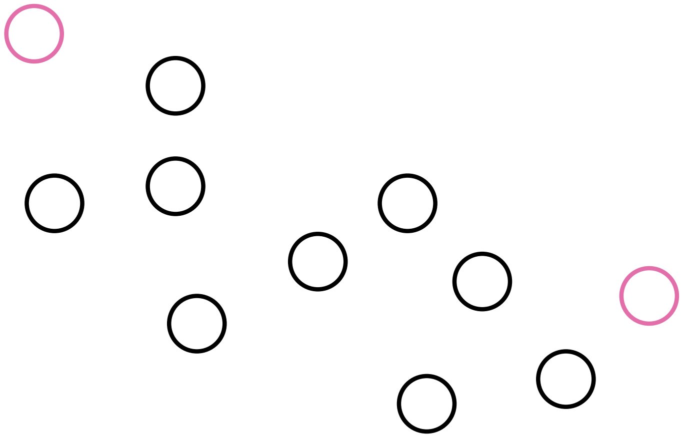
Goal of Clustering



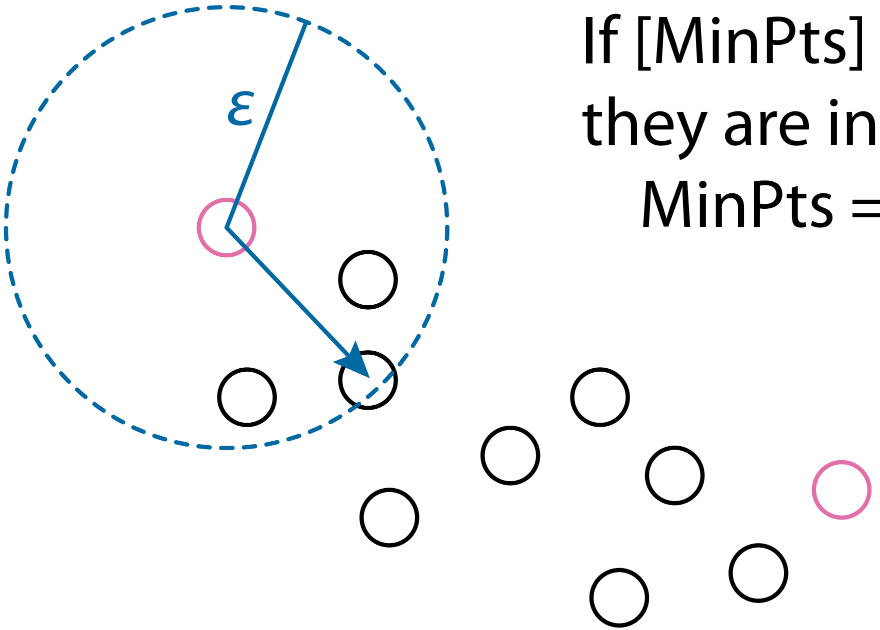
- Data:

1	355.60	270.21
2	549.28	351.71
3	520.08	215.48
4	575.15	166.68
5	288.83	141.49
6	50.61	75.77
7	262.57	428.45
⋮	⋮	⋮
4000	309.395	365.09

DBSCAN [Ester et al., 1996]



DBSCAN [Ester et al., 1996]



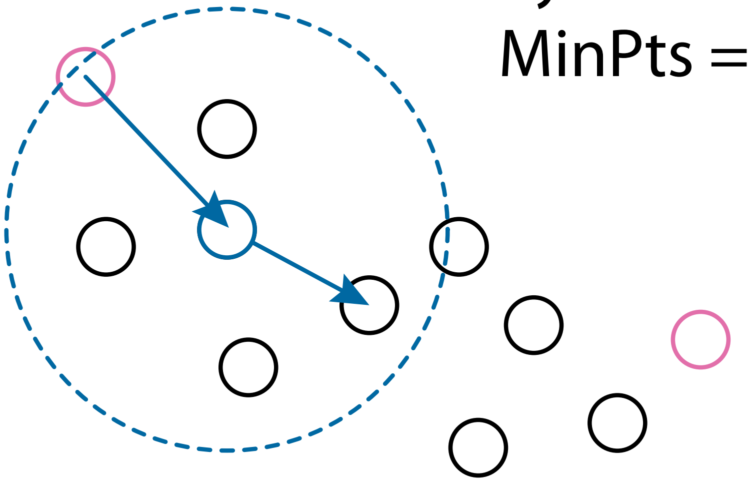
If [MinPts] points are in the same circle, they are in the same cluster

MinPts = 3 in this case

DBSCAN [Ester et al., 1996]

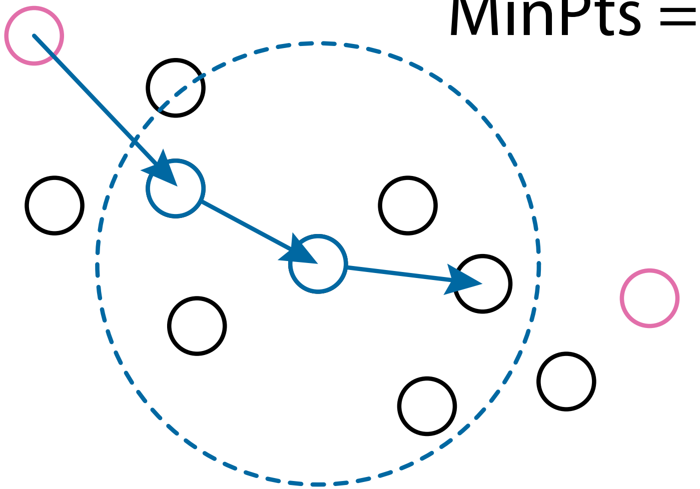
If [MinPts] points are in the same circle, they are in the same cluster

MinPts = 3 in this case



DBSCAN [Ester et al., 1996]

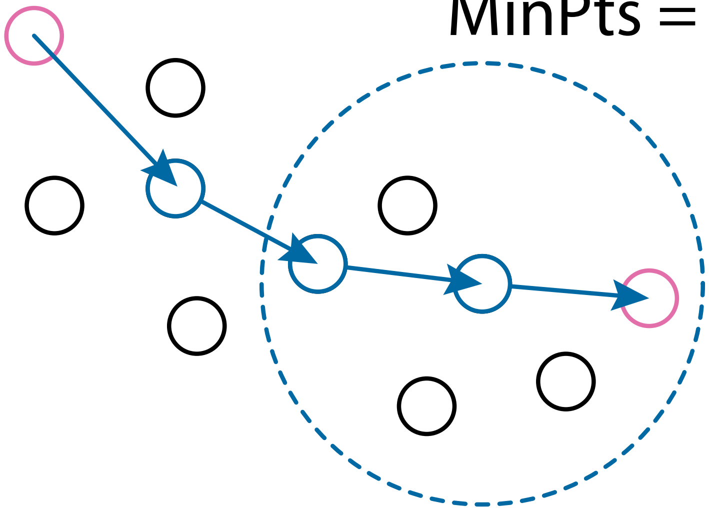
If [MinPts] points are in the same circle,
they are in the same cluster
MinPts = 3 in this case



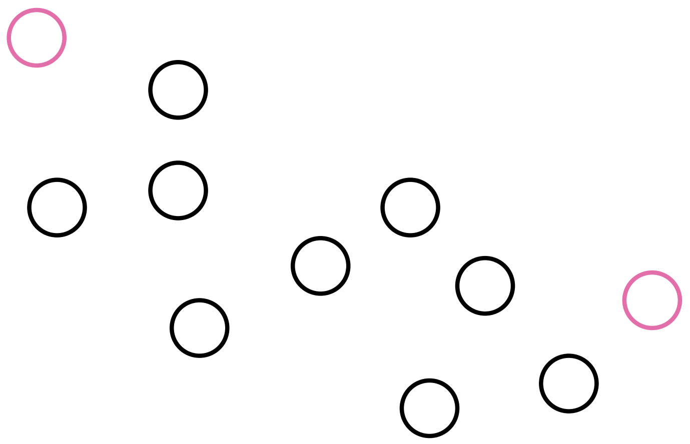
DBSCAN [Ester et al., 1996]

If [MinPts] points are in the same circle, they are in the same cluster

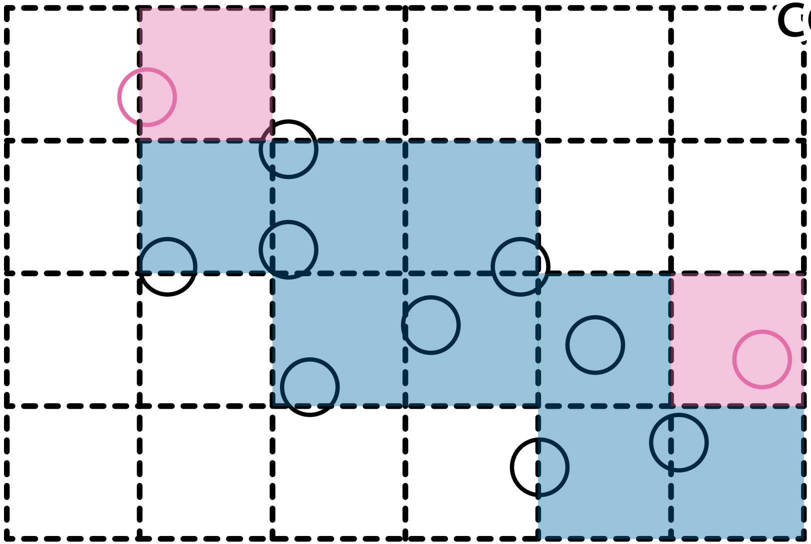
MinPts = 3 in this case



BOOL [Sugiyama and Yamamoto, 2011]



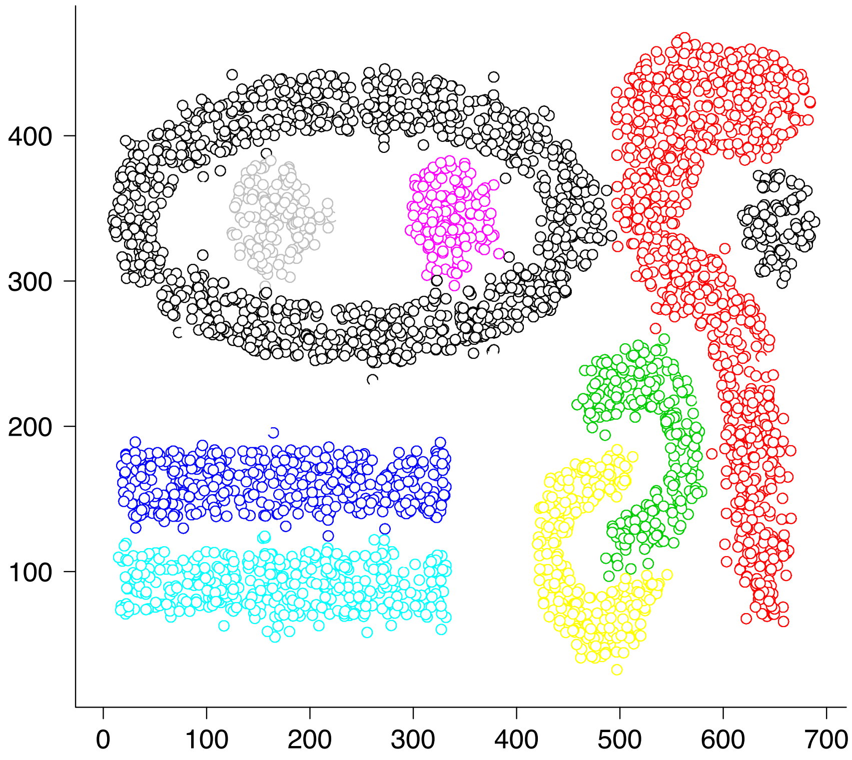
BOOL [Sugiyama and Yamamoto, 2011]



Discretize data and connect them if contiguous

$O(n^2) \Rightarrow O(n)$ by radix sort

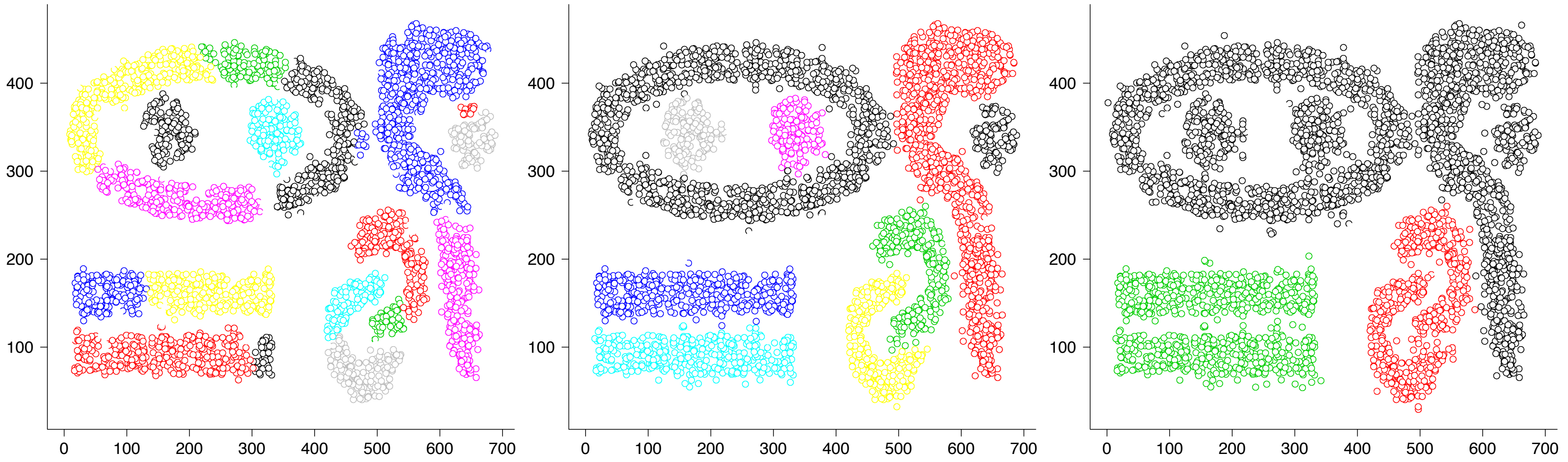
Result of DBSCAN with $\varepsilon = 14$ and MinPts = 10



- Data:

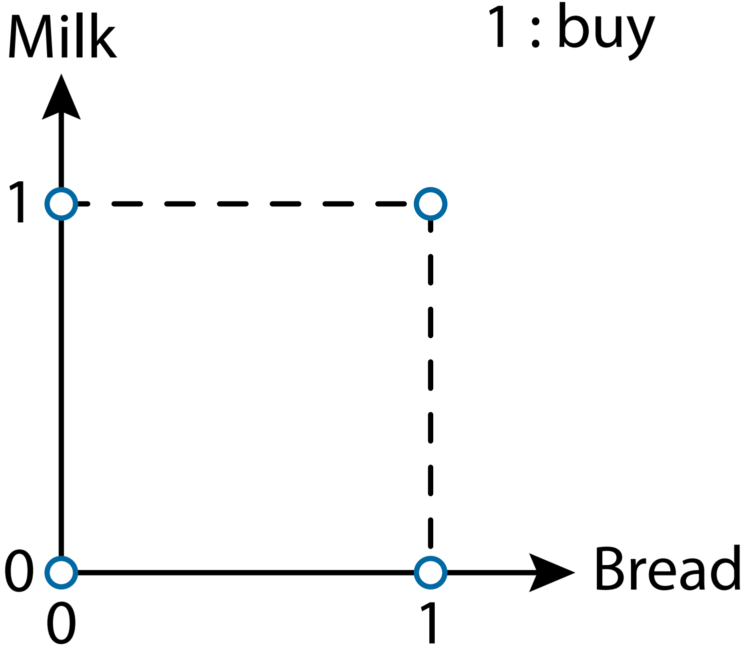
1	355.60	270.21
2	549.28	351.71
3	520.08	215.48
4	575.15	166.68
5	288.83	141.49
6	50.61	75.77
7	262.57	428.45
⋮	⋮	⋮
4000	309.395	365.09

Clustering Results Are Arbitrary



- ε = 12, 14, 16 (from left to right), MinPts = 10
  - Bias is introduced through a parameter ε and MinPts

Binary Data → Pattern Mining

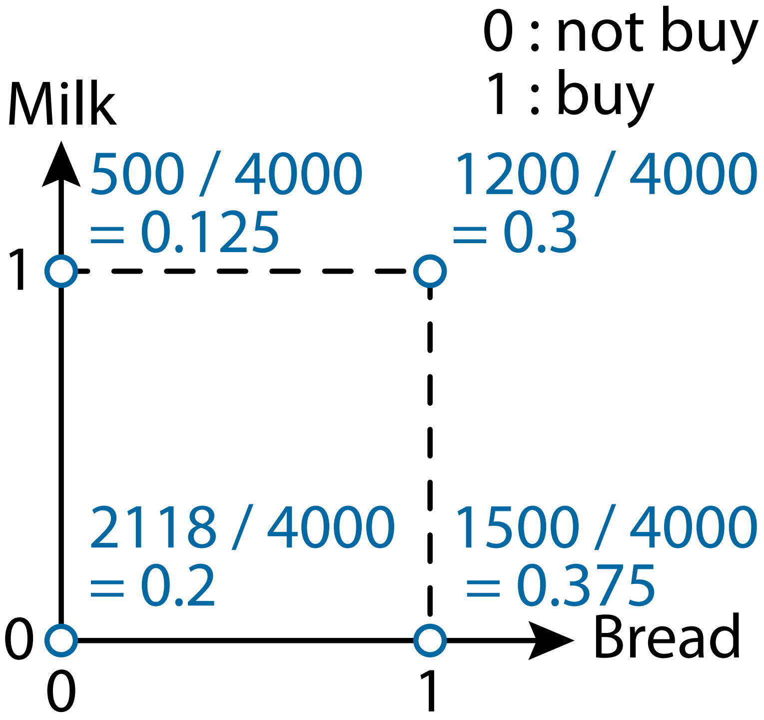


0 : not buy
1 : buy

- Data:

	Bread	Milk
1	0	1
2	1	1
3	1	1
4	1	0
5	0	0
6	1	0
⋮	⋮	⋮
4000	1	0

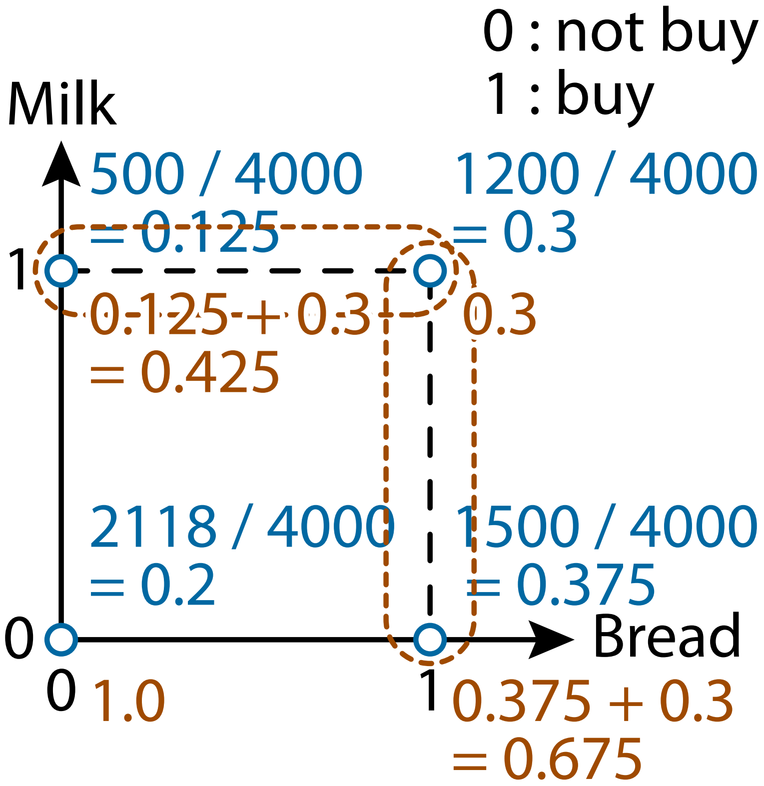
Find Frequent Patterns



- Data:

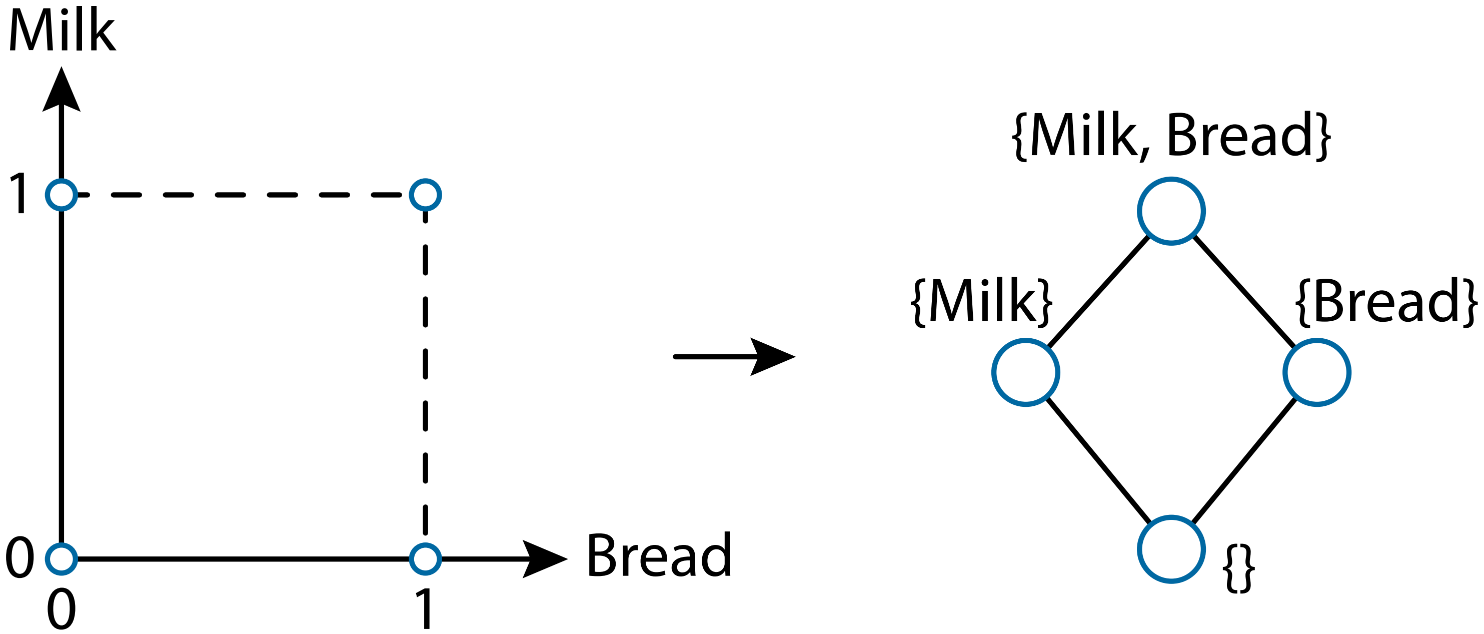
	Bread	Milk
1	0	1
2	1	1
3	1	1
4	1	0
5	0	0
6	1	0
⋮	⋮	⋮
4000	1	0

Find Frequent Patterns

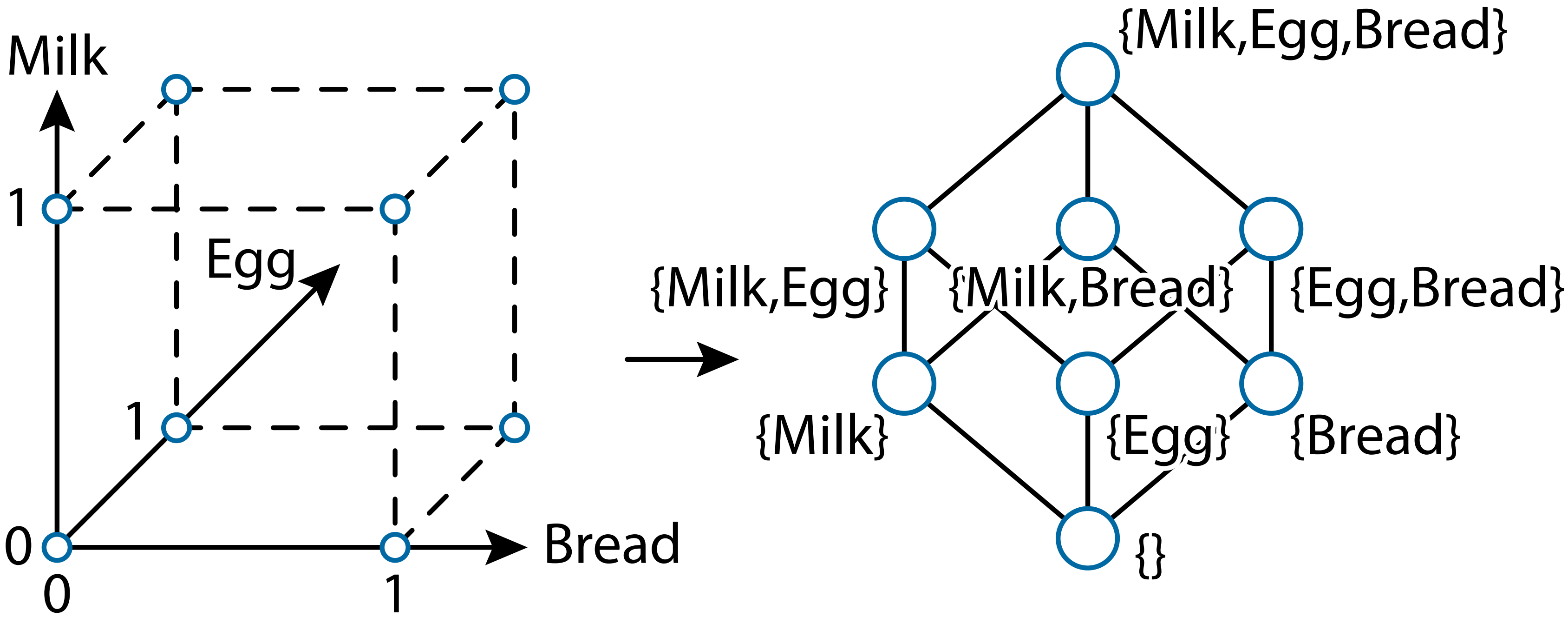


- 67.5% of customers bought {bread}
- 42.5% of customers bought {milk}
- 30% of customers bought {bread, milk}
- Combinations of items are patterns

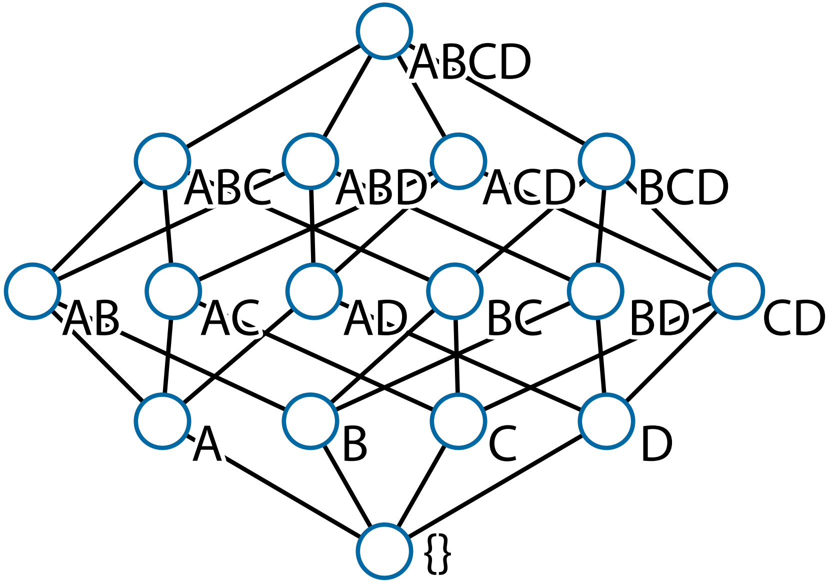
Lattice Representation (2D)



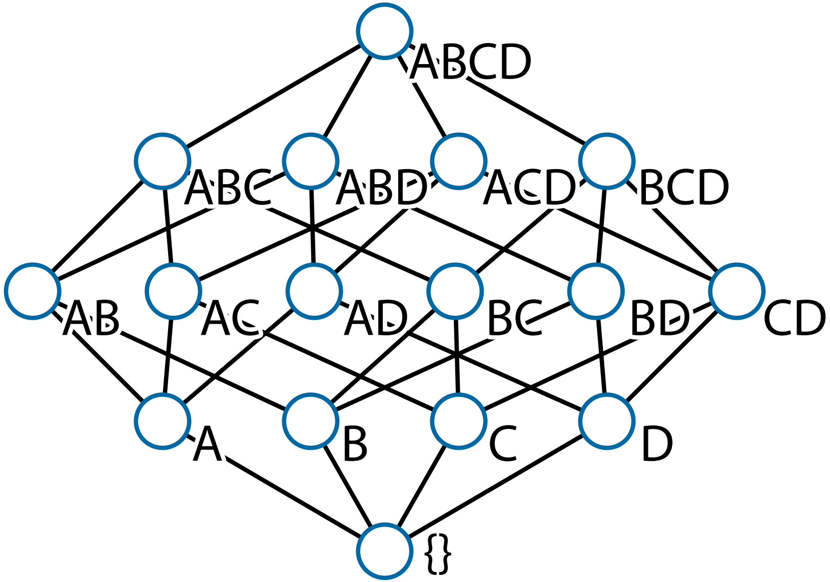
Lattice Representation (3D)



Lattice Representation (4D)



Combinatorial Explosion



The number of patterns for *n* items

$2^2 = 4$

$2^3 = 8$

$2^4 = 16$

...

$2^{10} = 1024$

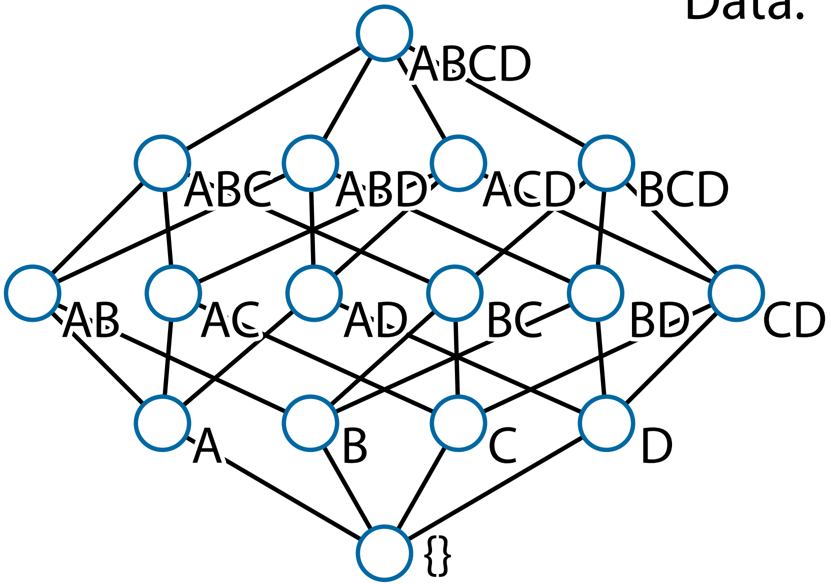
$2^{20} = 1048576$

...

$2^{100} > 1000000000000000000000000000000$

Combinatorial explosion!

Frequency = Upper Set

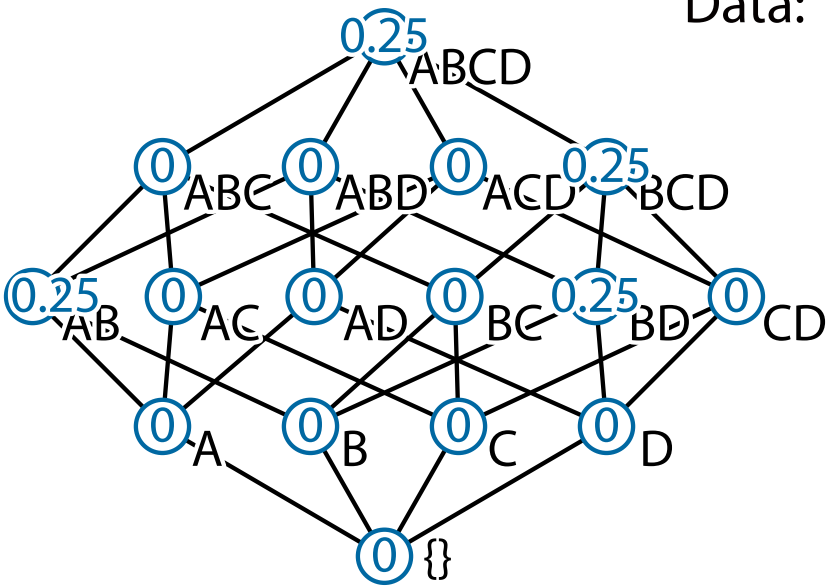


Data:

ID	A	B	C	D
1	1	1	1	1
2	1	1	0	0
3	0	1	0	1
4	0	1	1	1

Task:
Find all patterns (sets of features) whose frequency ≥ 2/4

Frequency = Upper Set



Data:

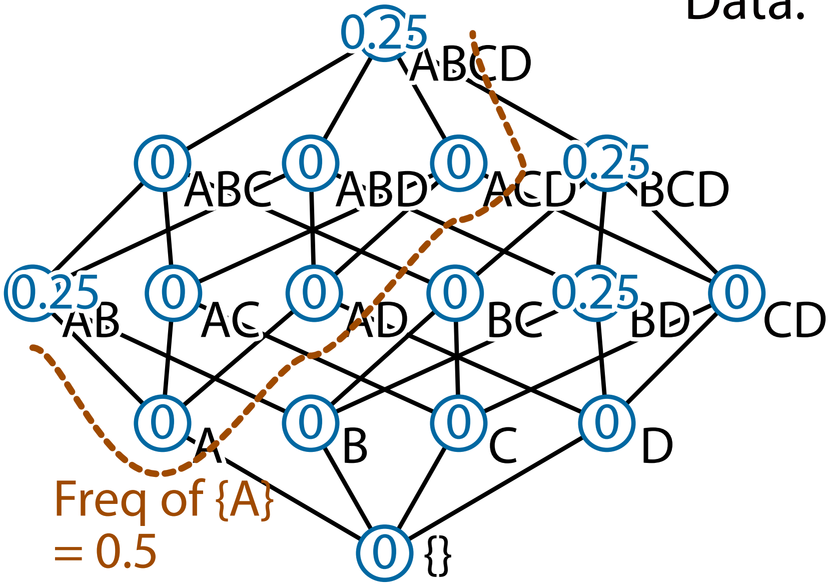
ID	A	B	C	D
1	1	1	1	1
2	1	1	0	0
3	0	1	0	1
4	0	1	1	1

Task:

Find all patterns (sets of features) whose frequency ≥ 2/4

■ : Probability ■ : Frequency

Frequency = Upper Set



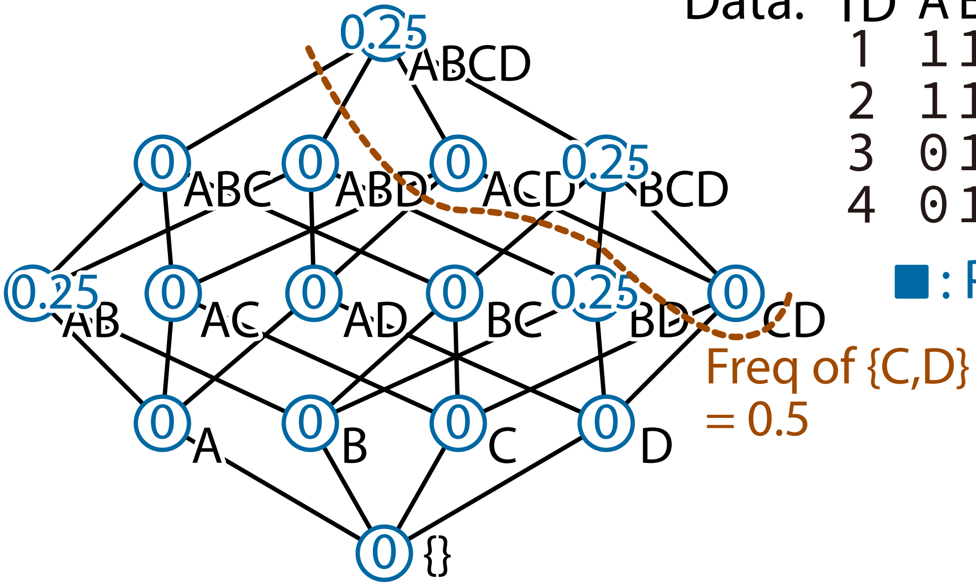
Data:

ID	A	B	C	D
1	1	1	1	1
2	1	1	0	0
3	0	1	0	1
4	0	1	1	1

Task:
Find all patterns
(sets of features)
whose frequency ≥ 2/4

■ : Probability ■ : Frequency

Frequency = Upper Set



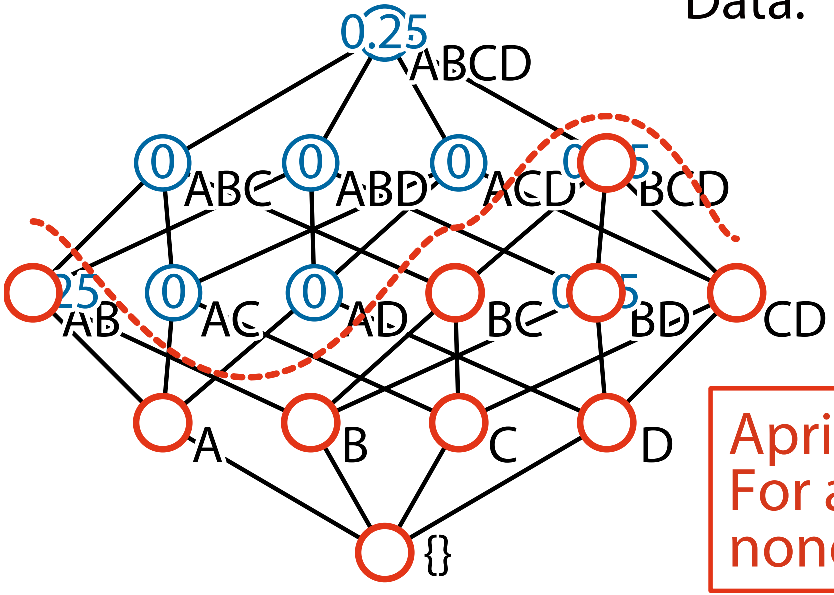
Data:

ID	A	B	C	D
1	1	1	1	1
2	1	1	0	0
3	0	1	0	1
4	0	1	1	1

Task:
Find all patterns (sets of features) whose frequency ≥ 2/4

■ : Probability ■ : Frequency

Apriori Strategy For Pattern Mining



Data:

ID	A	B	C	D
1	1	1	1	1
2	1	1	0	0
3	0	1	0	1
4	0	1	1	1

Task:
Find all patterns (sets of features) whose frequency ≥ 2/4

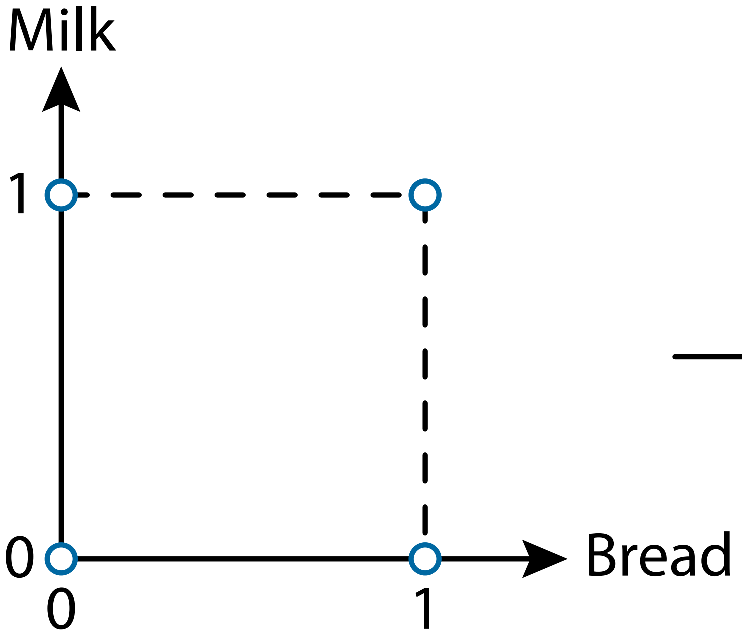
■ : Probability ■ : Frequency

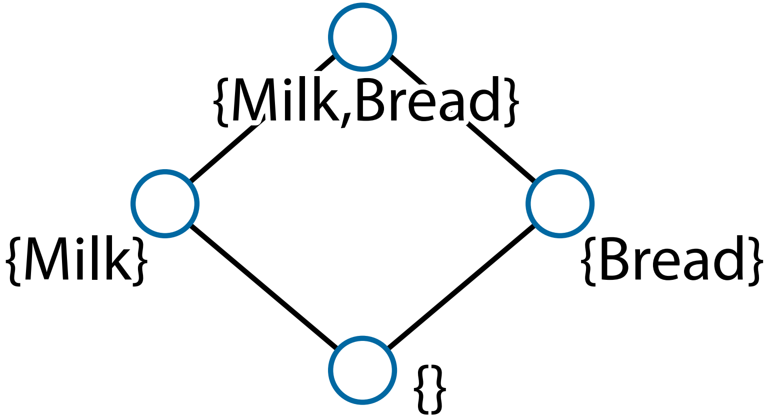
Apriori principle [Agrawal and Srikant, 1994]:
For a pattern with frequency η,
none of its superset's frequency $\geq \eta$

Recent Advances

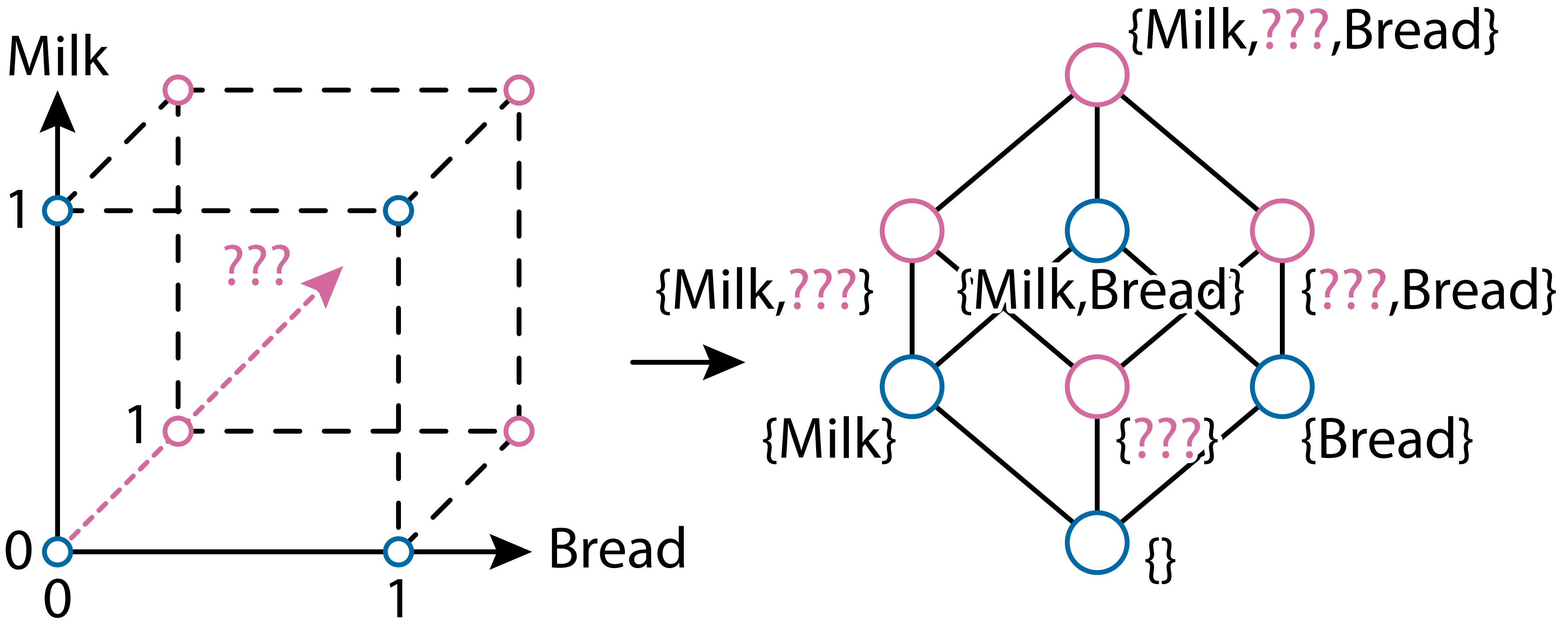
- The fastest algorithm: LCM [Uno et al., 2004]
- Introduce statistical assessment of frequency to compute *p*-value
  - LAMP [Terada et al., 2013]
  - Graph mining [Sugiyama et al., 2015]
  - Westfall-Young light [Llinares-López et al., 2015]
  - A review paper [Sugiyama 2017] (in Japanese)

Introduce Unknown Variable → Deep Learning

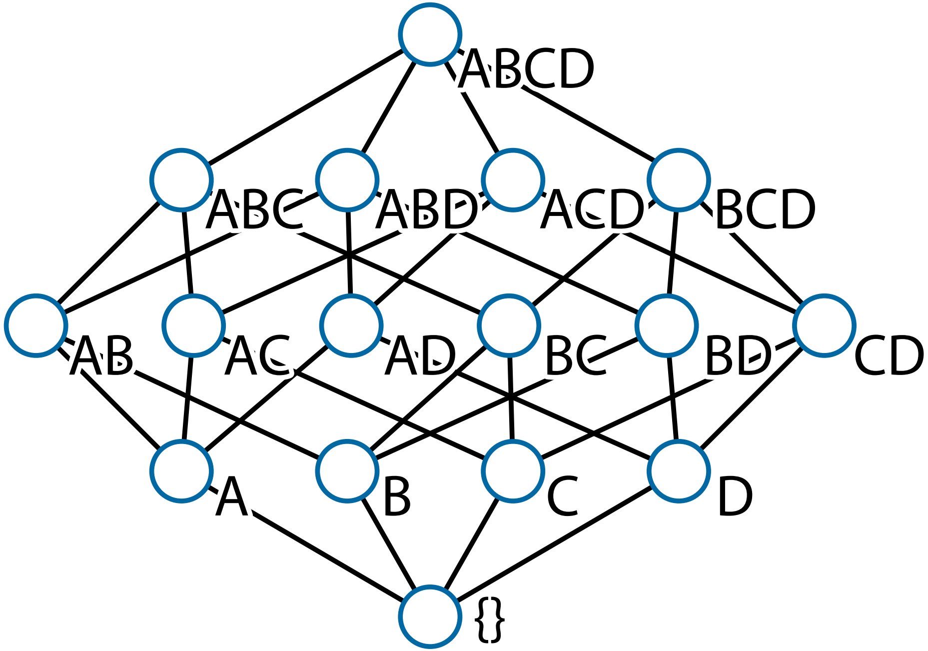




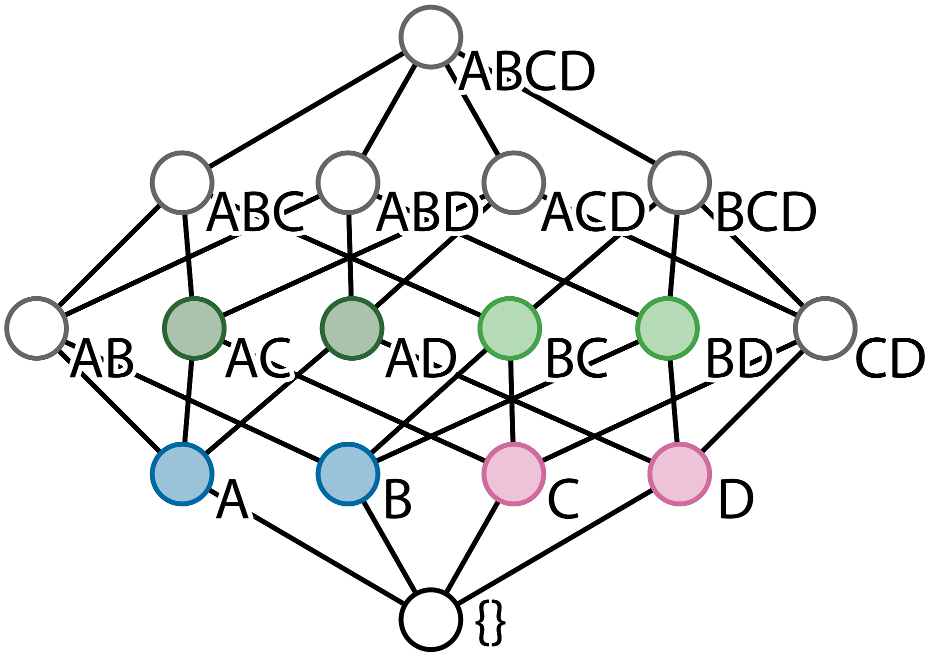
Introduce Unknown Variable → Deep Learning

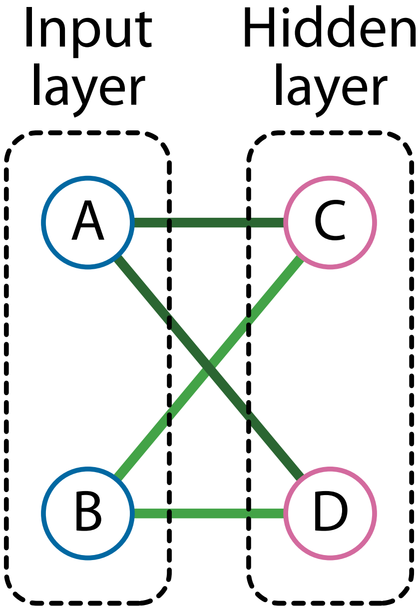


Lattice and (Deep) Boltzmann Machine

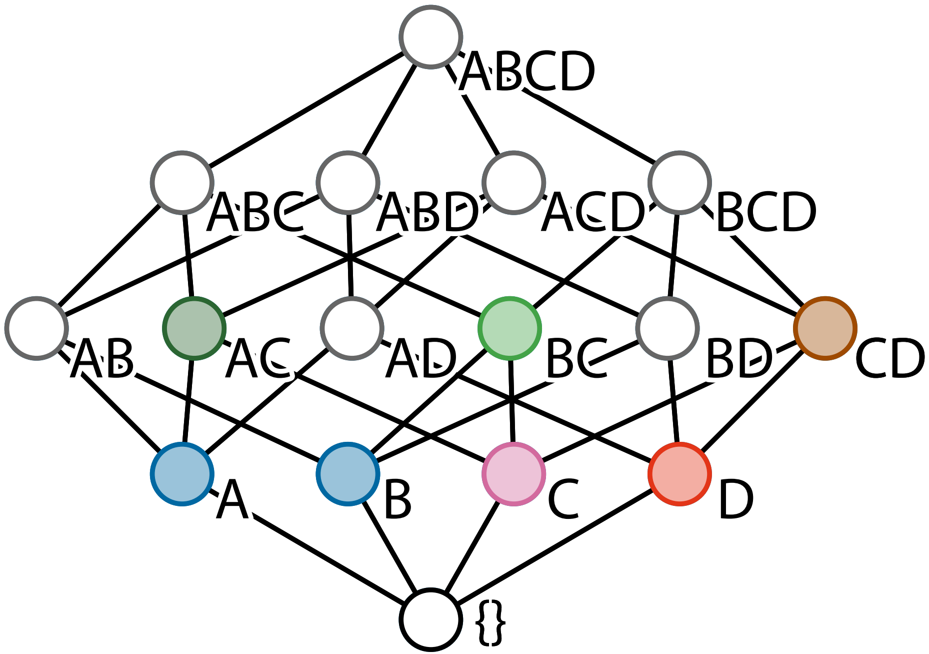


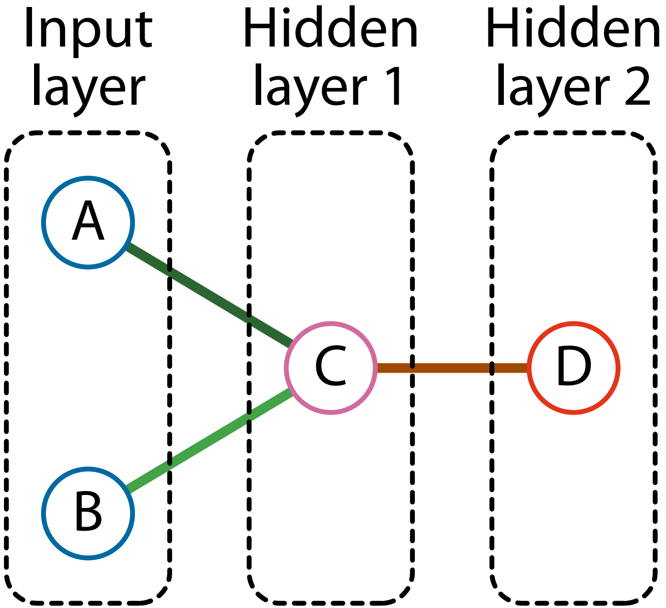
Lattice and (Deep) Boltzmann Machine



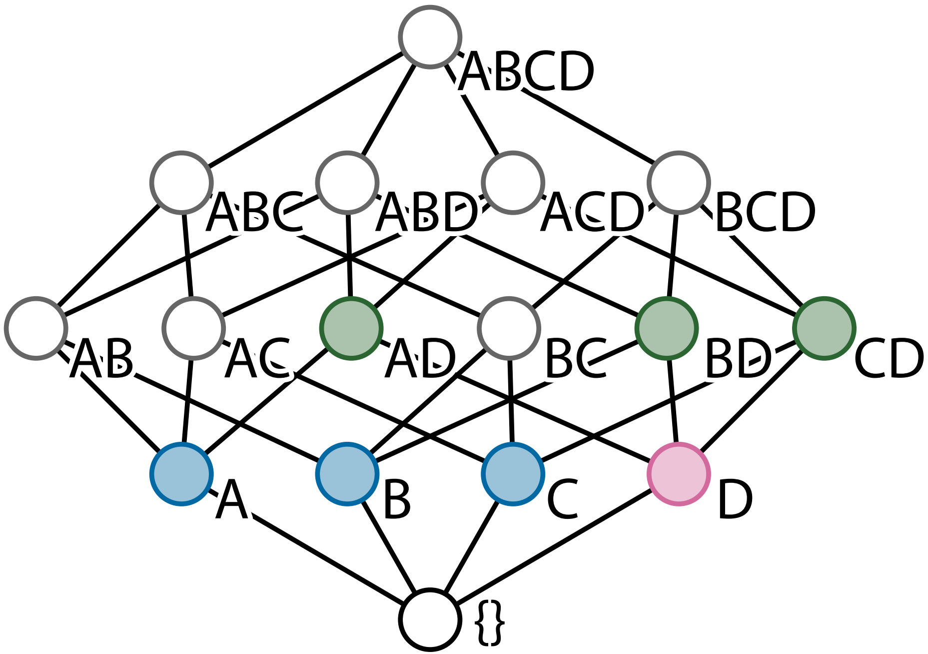


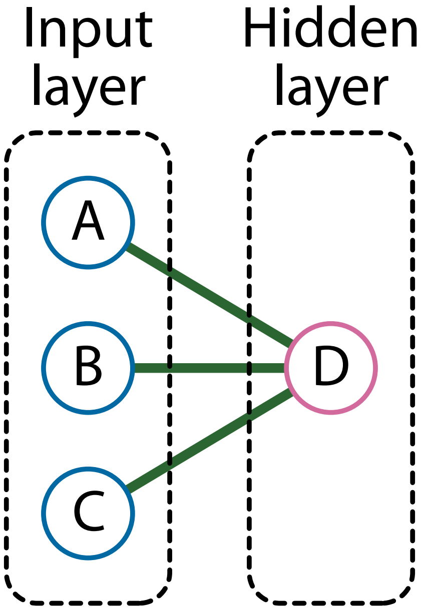
Lattice and (Deep) Boltzmann Machine



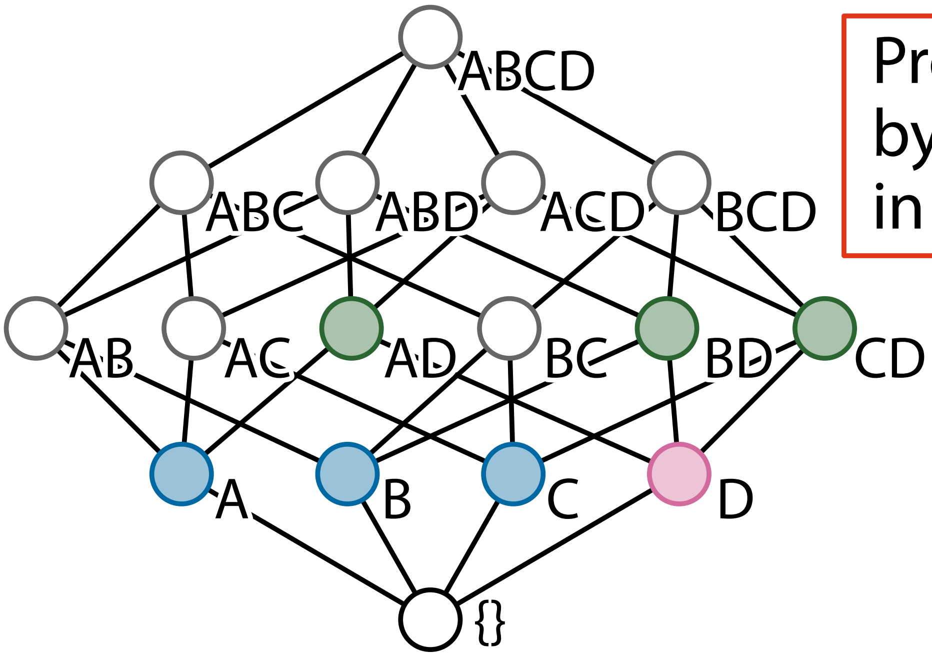


Lattice and (Deep) Boltzmann Machine



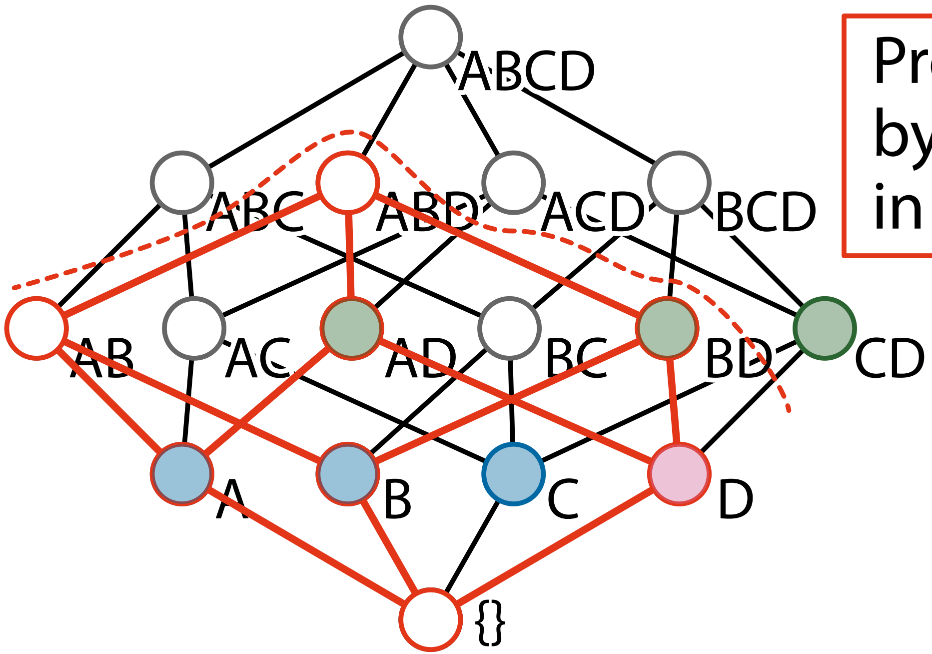


Model Distribution



Probabilities p are obtained by the sum of coefficients θ in the lower sets

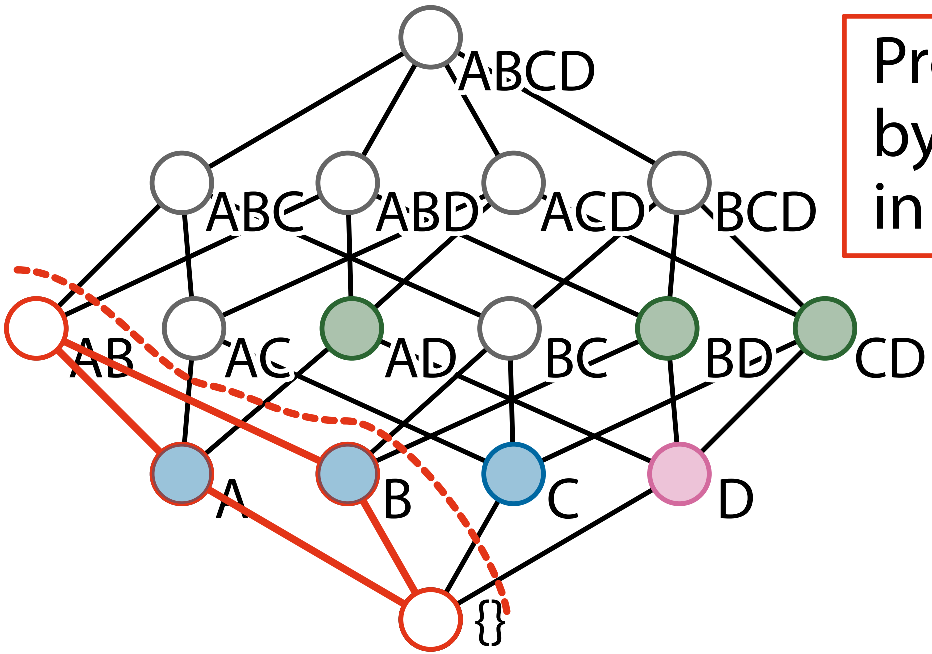
Model Distribution



Probabilities p are obtained by the sum of coefficients θ in the lower sets

$$\log p(\mathrm{ABD}) = \theta(\mathrm{AD}) + \theta(\mathrm{BD}) + \theta(\mathrm{A}) + \theta(\mathrm{B}) + \theta(\mathrm{D})$$

Model Distribution

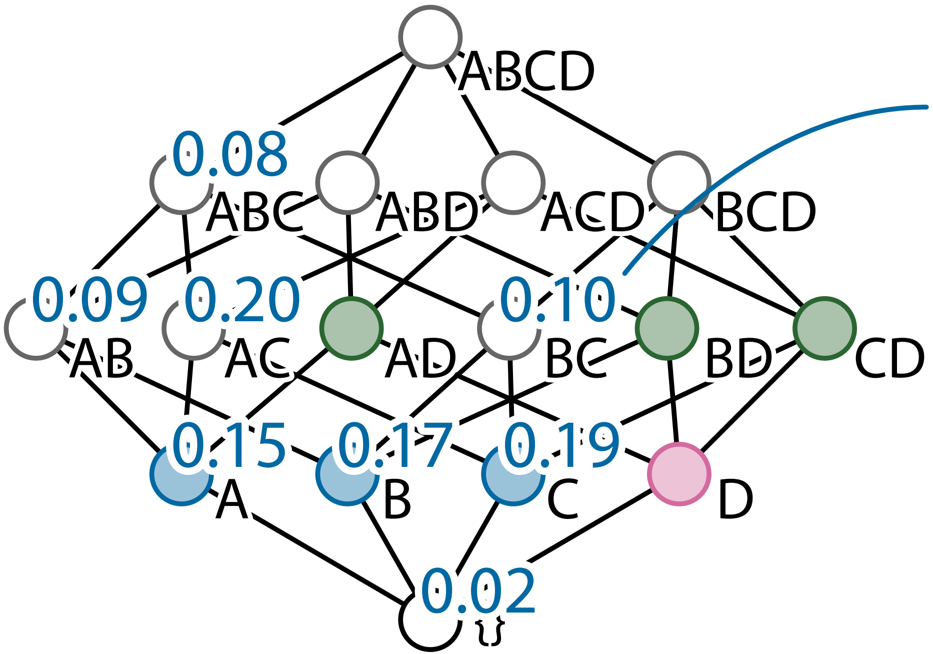


Probabilities p are obtained by the sum of coefficients θ in the lower sets

$$\log p(\mathrm{ABD}) = \theta(\mathrm{AD}) + \theta(\mathrm{BD}) + \theta(\mathrm{A}) + \theta(\mathrm{B}) + \theta(\mathrm{D})$$

$$\log p(\mathrm{AB}) = \theta(\mathrm{A}) + \theta(\mathrm{B})$$

Learn θ That Maximizes Log-Likelihood



q: Empirical dist. from data

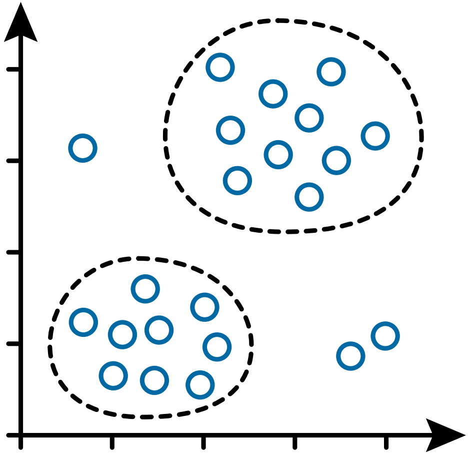
Log-likelihood
$= \sum q(x) \log p(x)$

Recent Study

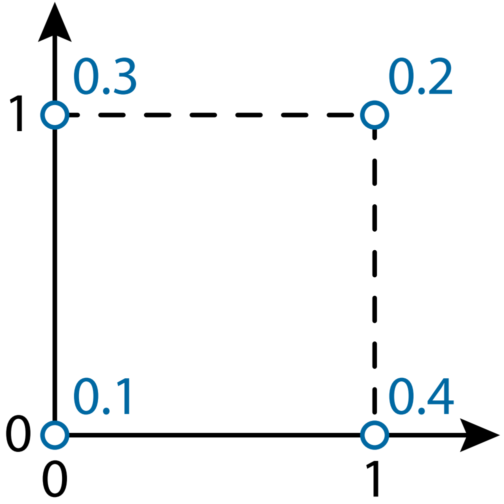
- Analyze distributions on lattices (posets) by information geometry
  - Information decomposition [Sugiyama et al., 2016]
  - Application to tensor balancing [Sugiyama et al., 2017]

Conclusion

Clustering



Pattern mining



Deep learning