Cartification

turning similarities into itemset frequencies

or how to turn every database into a supermarket

Bart Goethals

in collaboration with Emin Aksehirli and Jilles Vreeken

Universiteit Antwerpen

Historical note

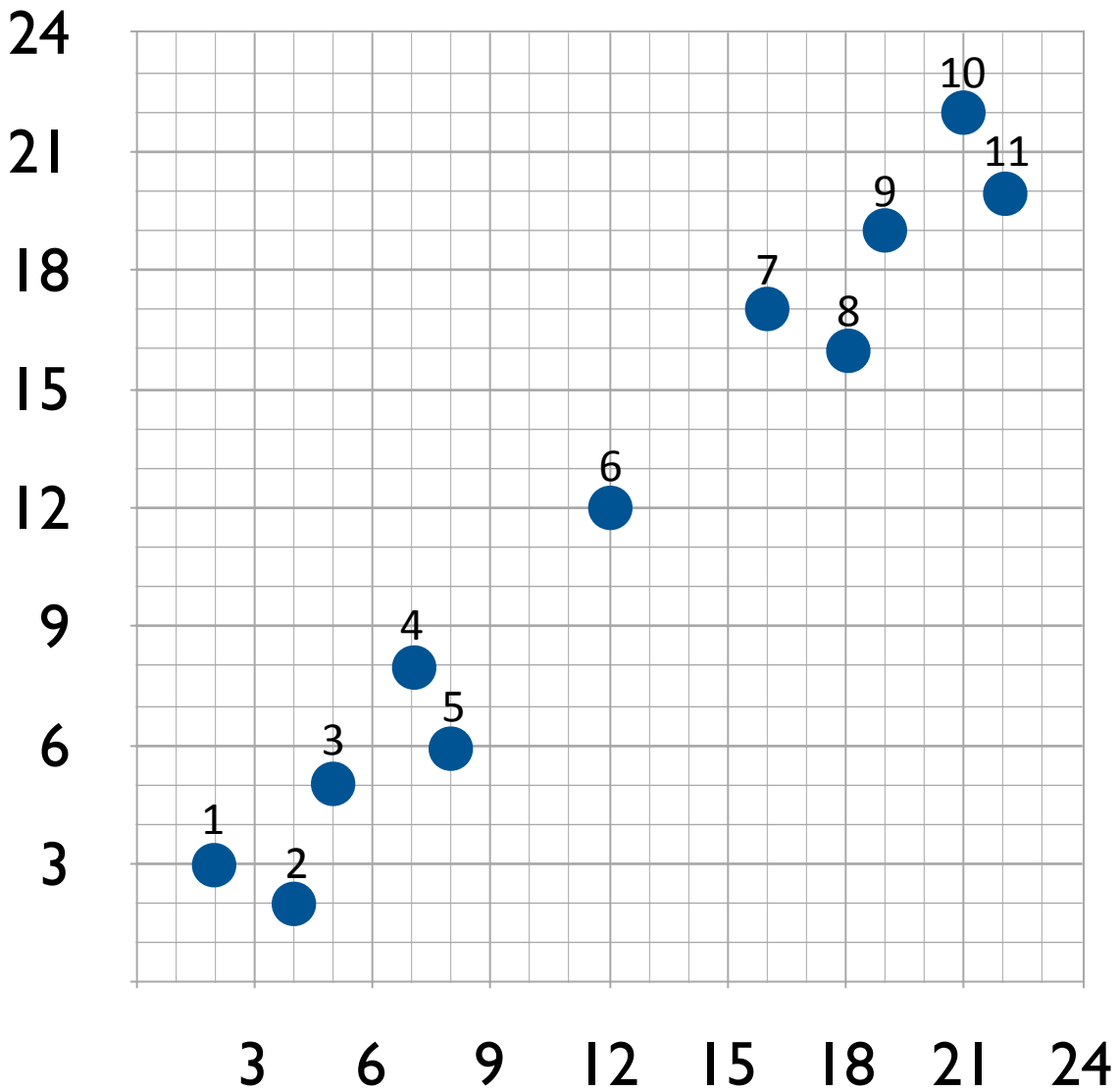
- Frequent itemset mining has a long history with 1000s of publications
- efficiency improvements
- constraints
- different interestingness measures
- a quest for search space pruning properties

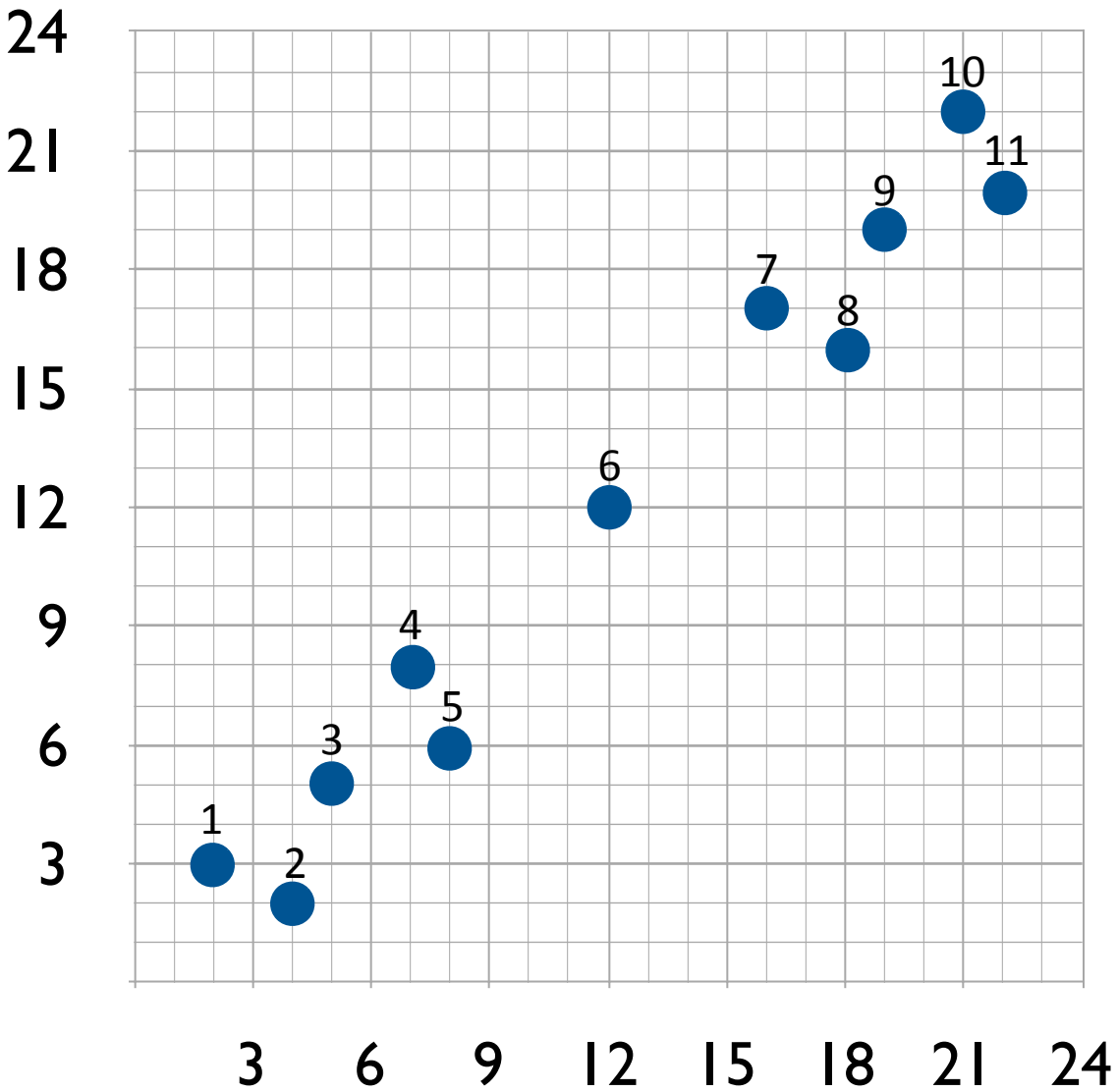
Clustering

- it's also about mining sets
- similarity measures are interestingness measures too
- what similarity measures have pruning properties?
- Sum of Squared Errors is monotone w.r.t. set size

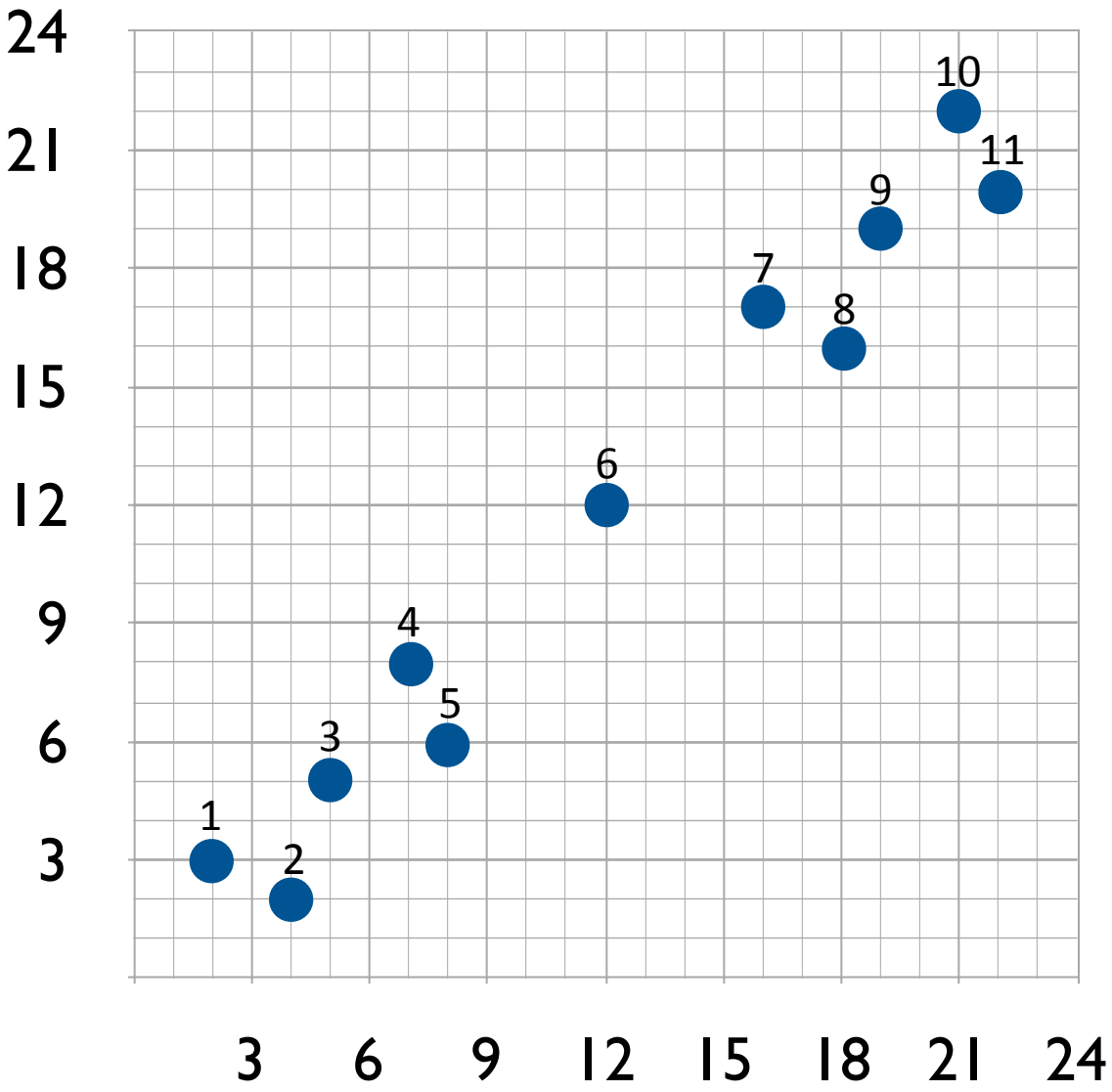
Cartification

- Idea!
- What if we can change the database into a supermarket?
-

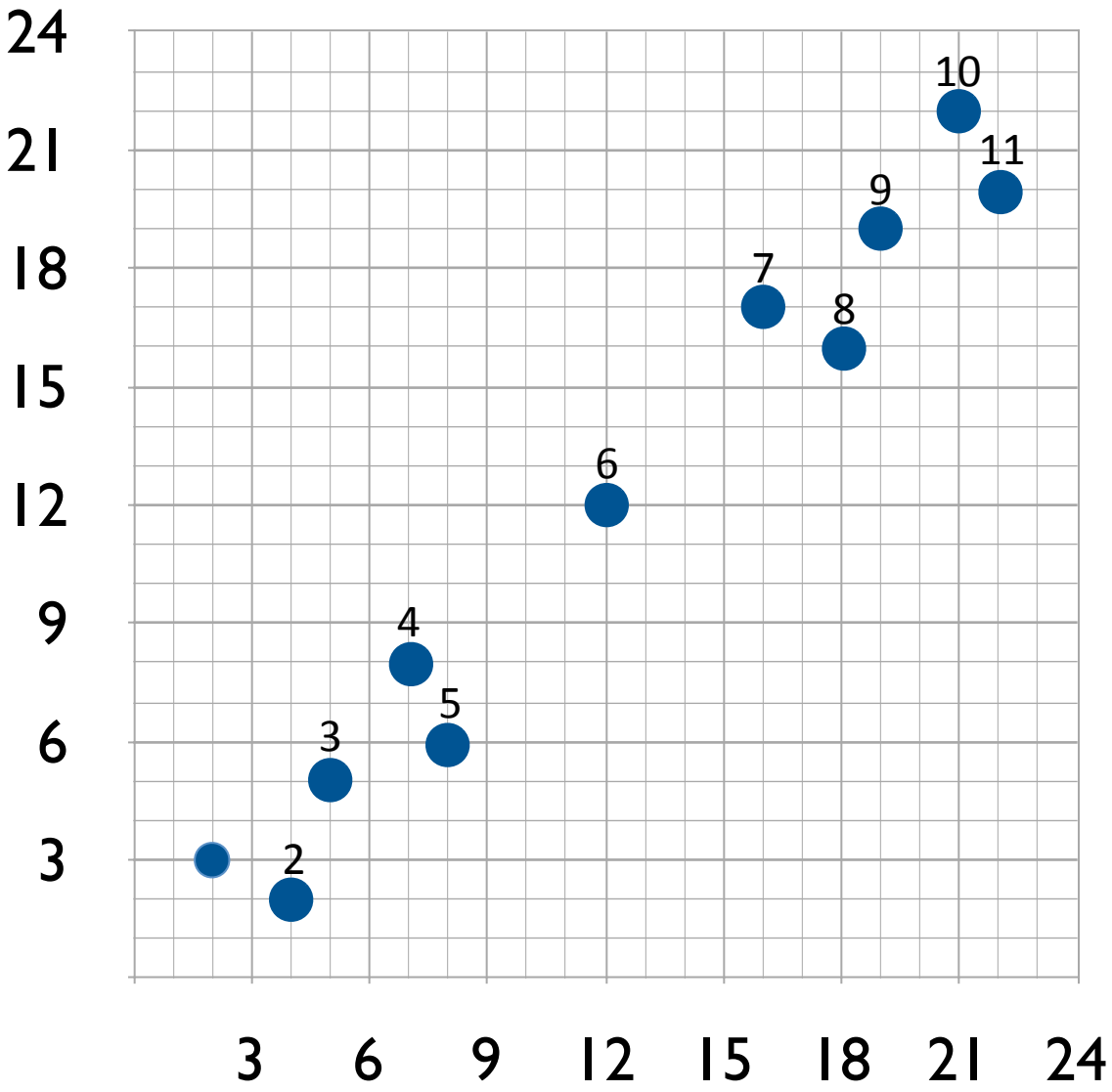

24
21
18
15
12
9
6
3
1
2
3
4
5
6
7
8
9
10
11
3
6
9
12
15
18
21
24



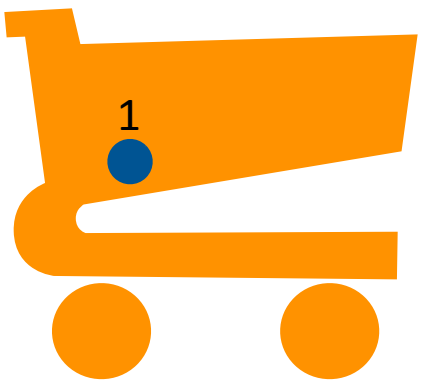
Select the k nearest neighbors of every point and add them to a cart

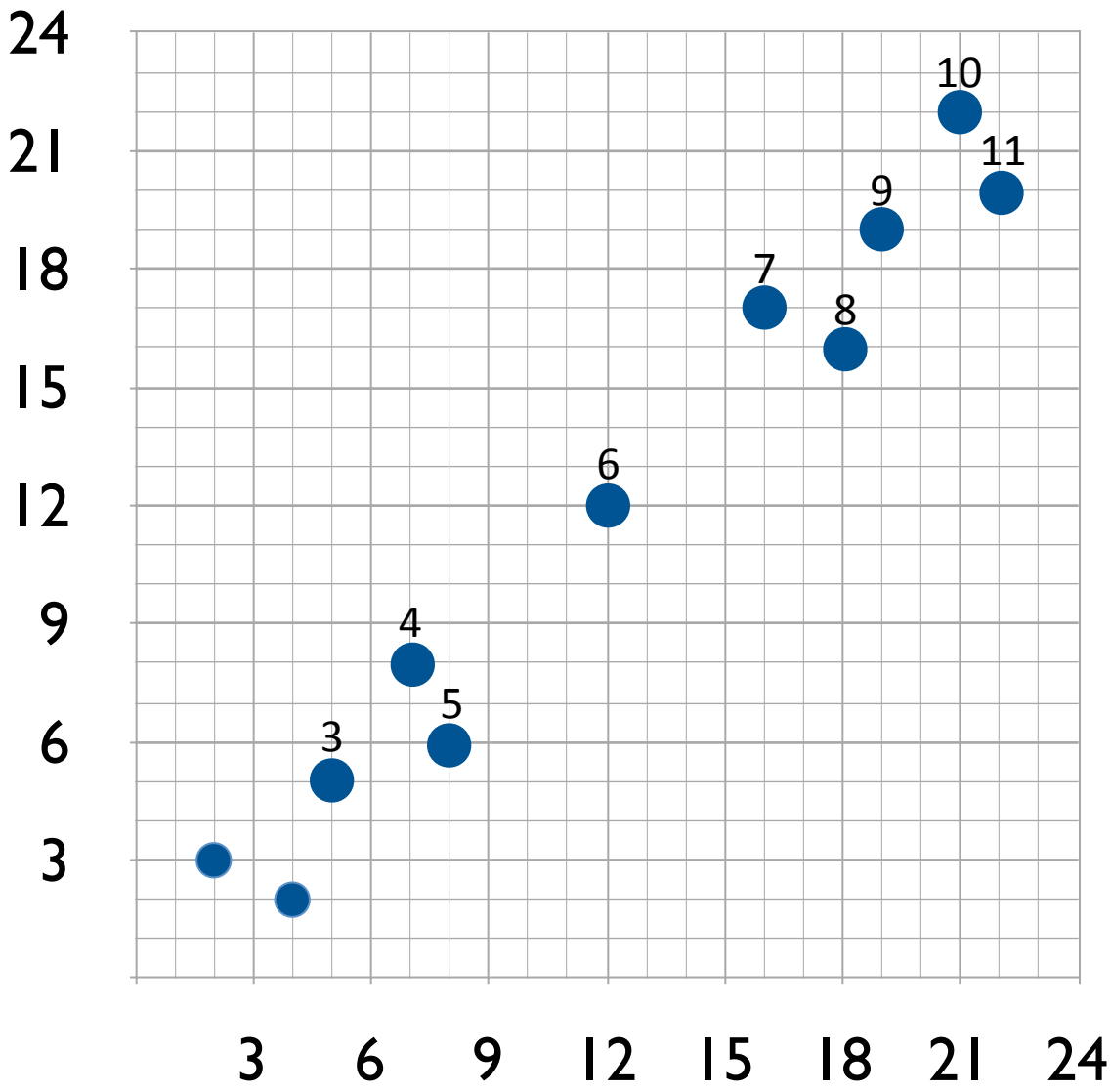


Select the k nearest neighbors of every point and add them to a cart

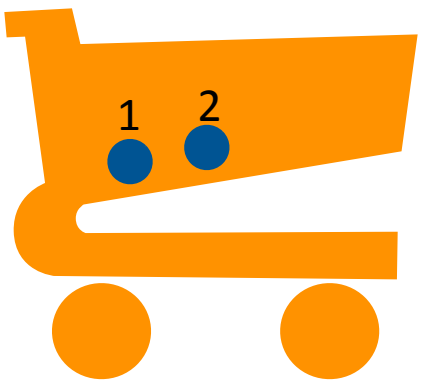


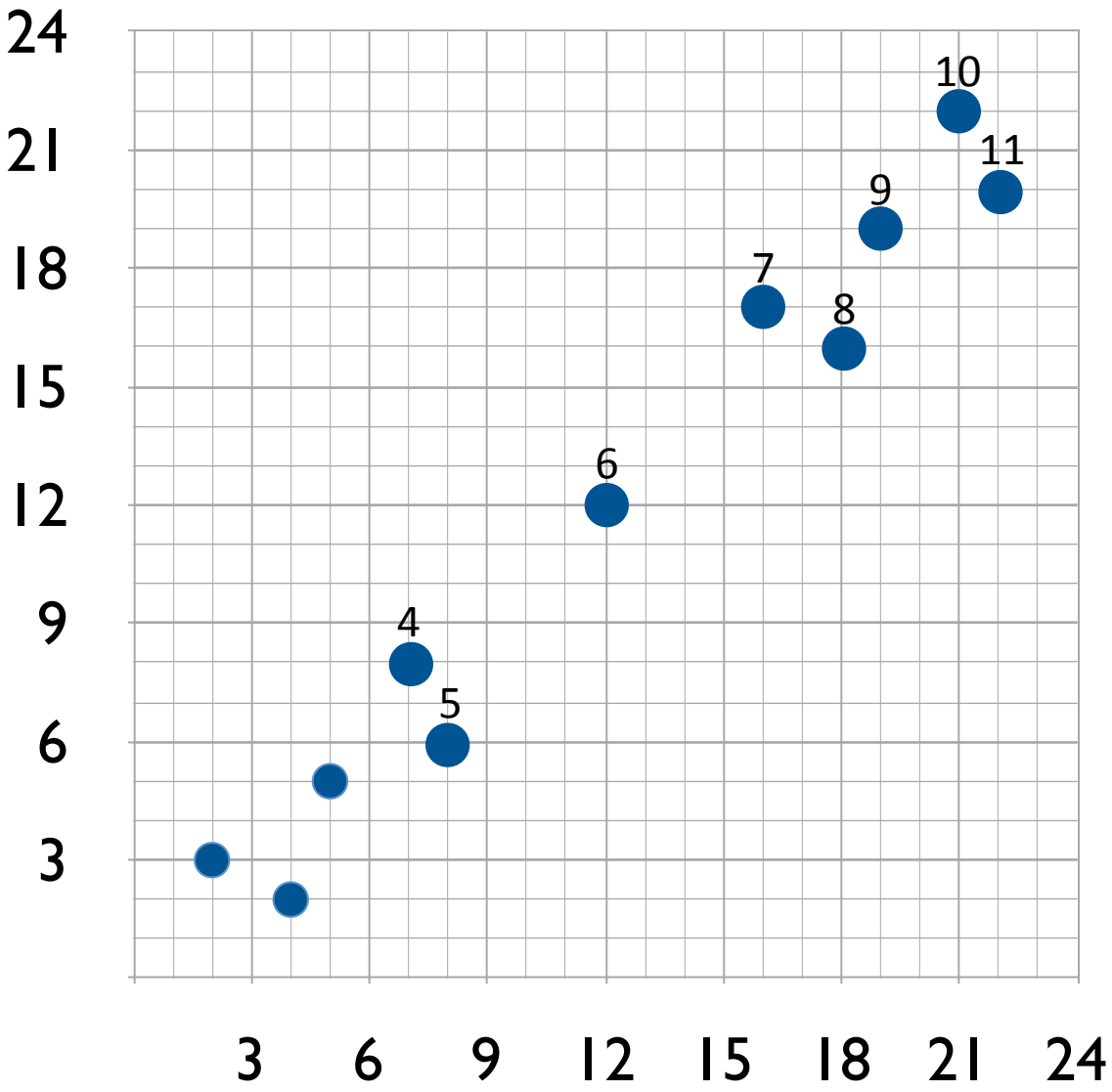
Select the k nearest neighbors of every point and add them to a cart



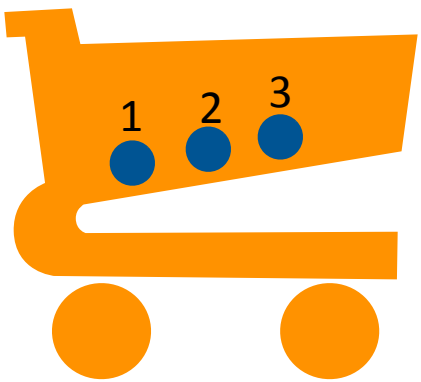


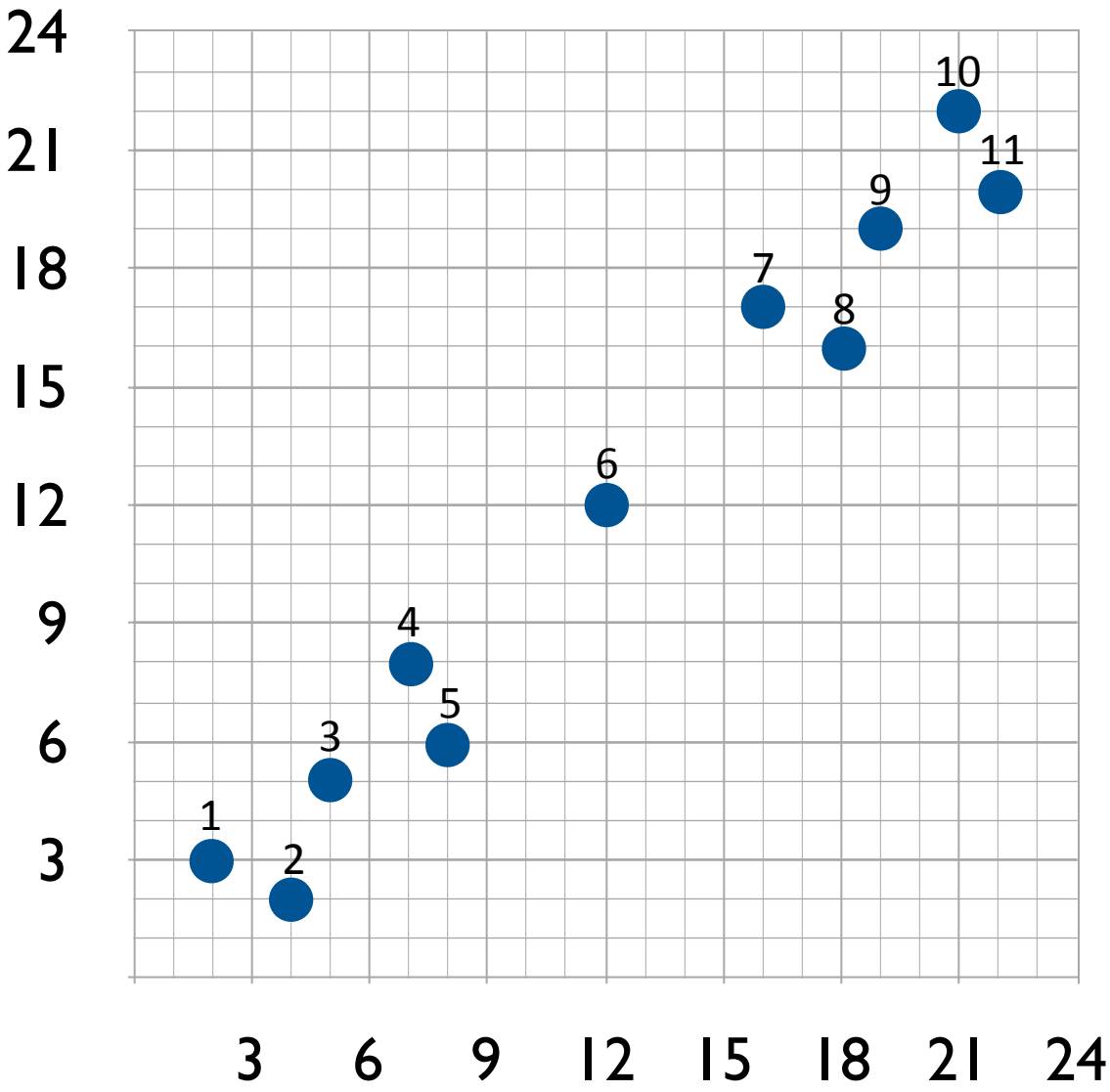
Select the k nearest neighbors of every point and add them to a cart

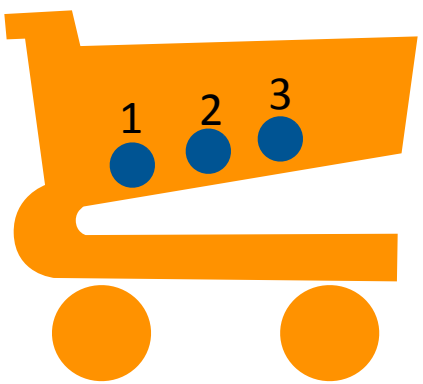


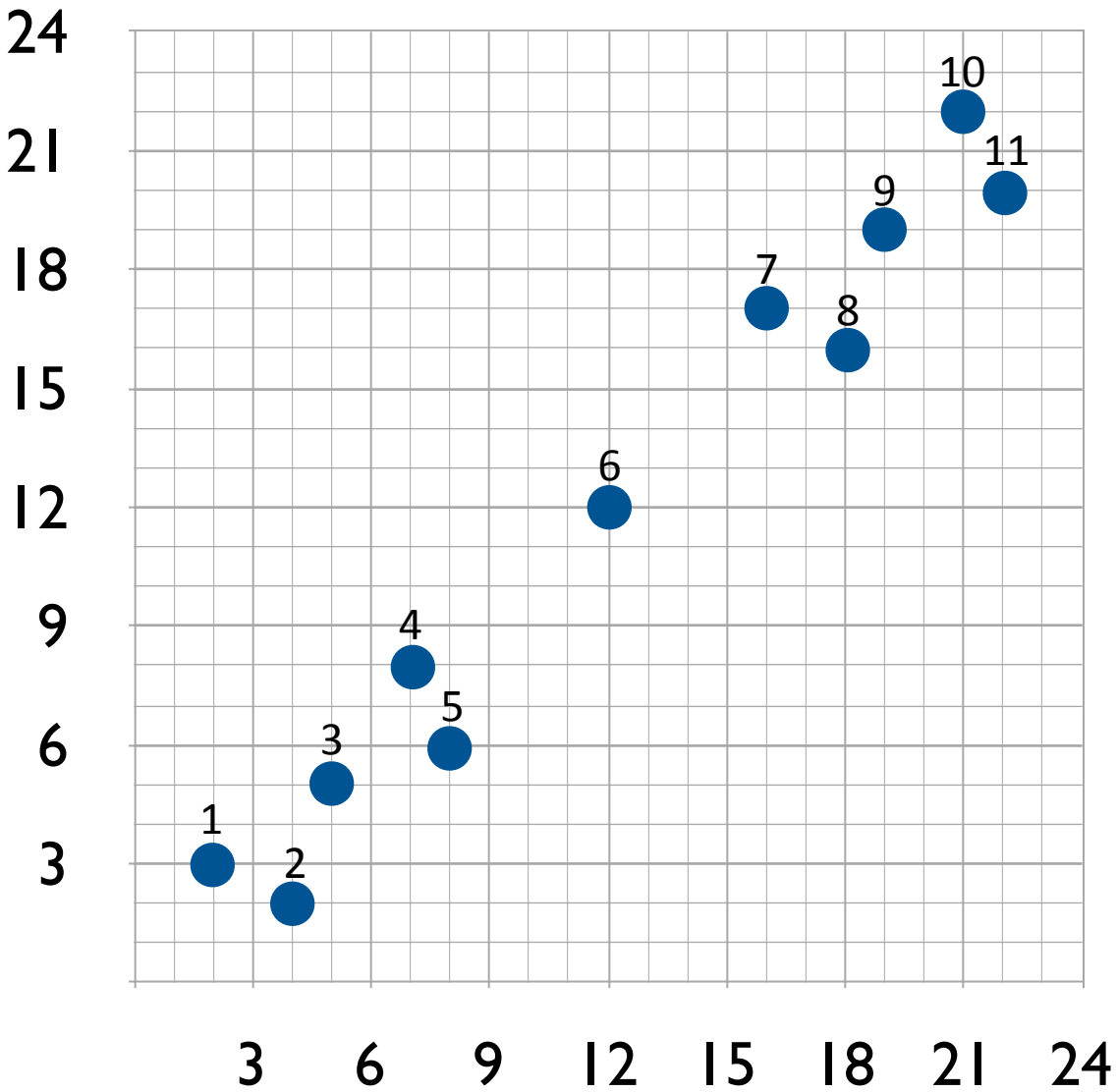


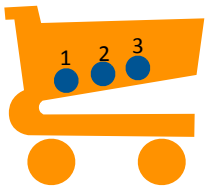
Select the k nearest neighbors of every point and add them to a cart

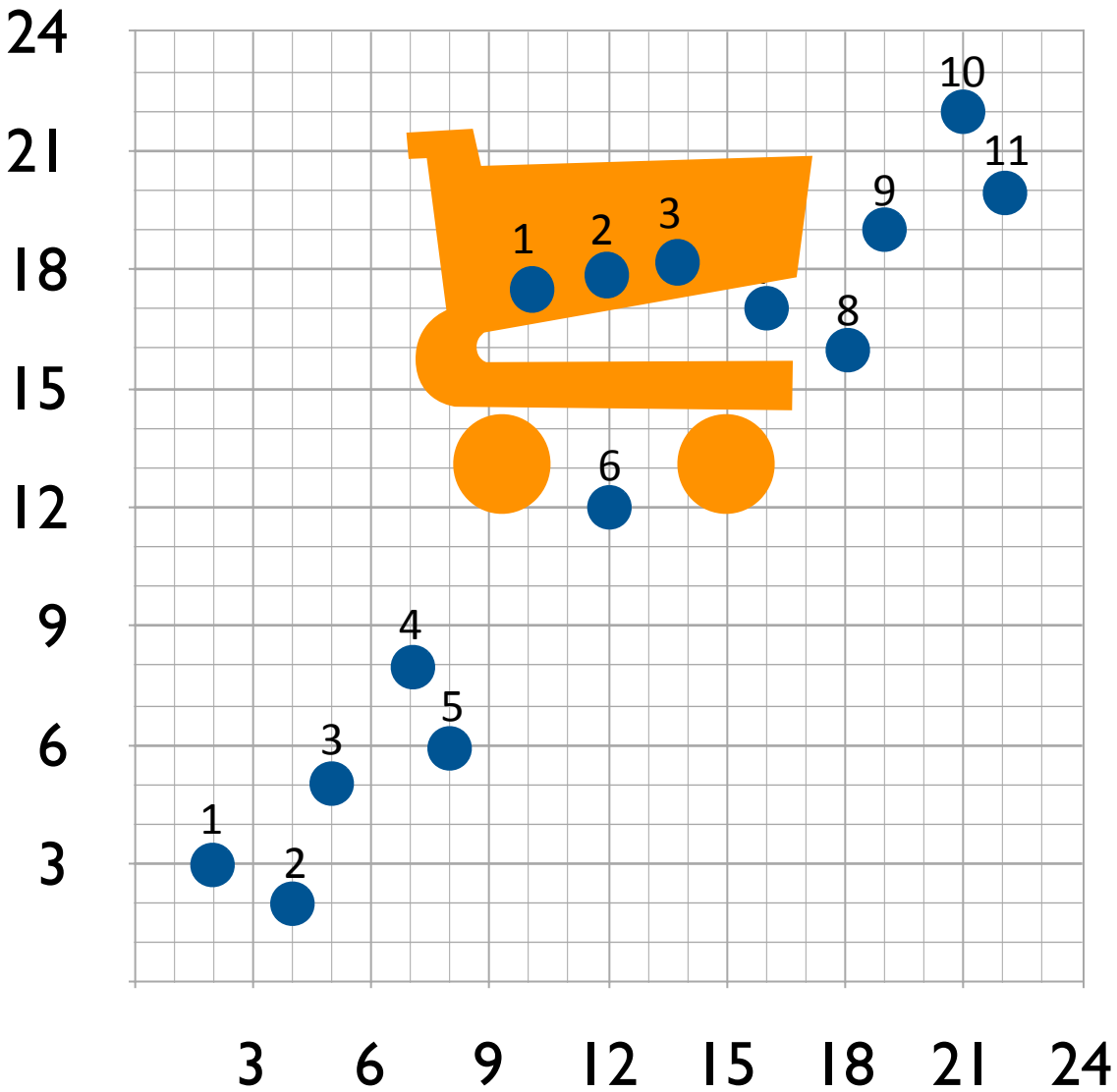


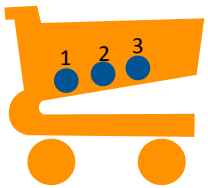

24
21
18
15
12
9
6
3
3
6
9
12
15
18
21
24
1
2
3
4
5
6
7
8
9
10
11

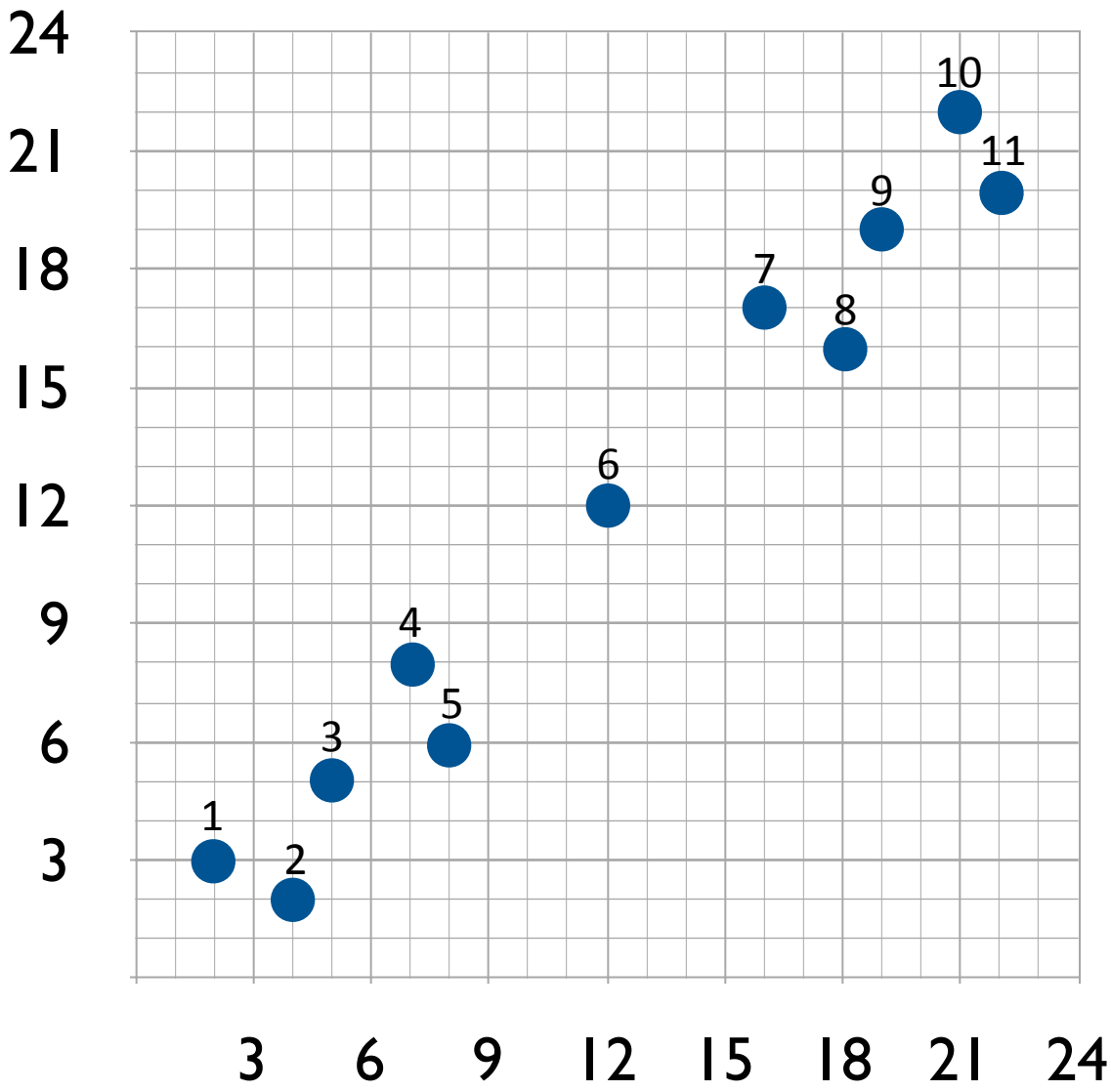

1
2
3

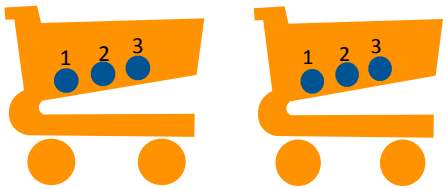

24
21
18
15
12
9
6
3
1
2
3
4
5
6
7
8
9
10
11
3
6
9
12
15
18
21
24

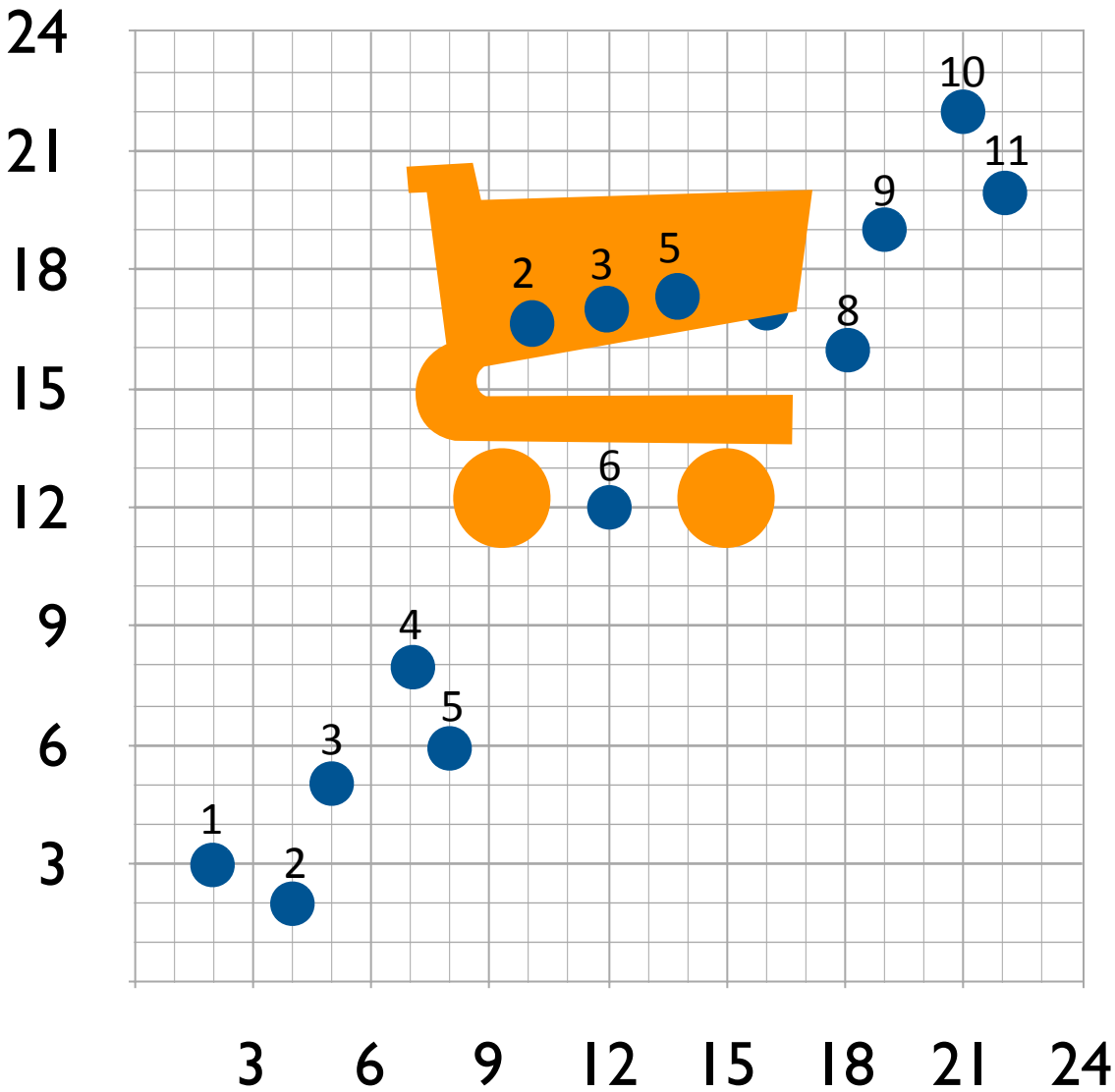

1
2
3

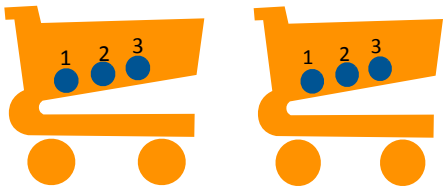

24
21
18
15
12
9
6
3
3
6
9
12
15
18
21
24
1
2
3
4
5
6
8
9
10
11

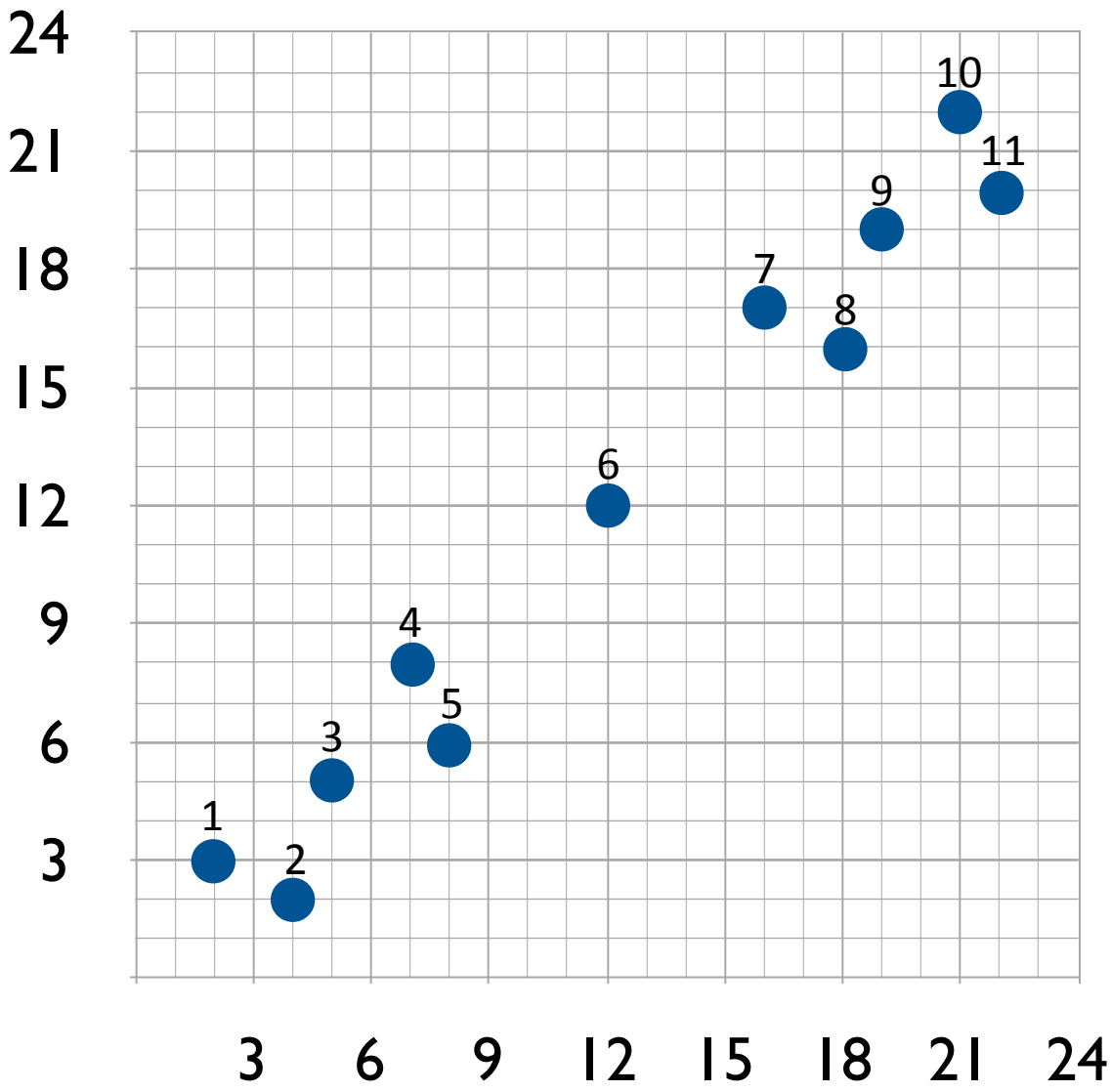

1
2
3

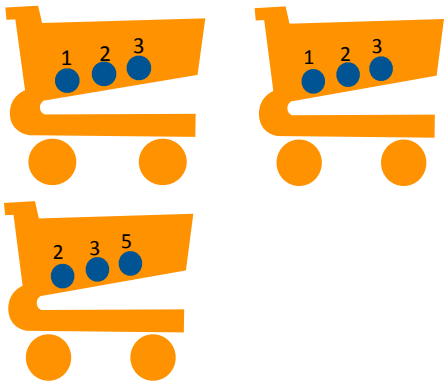

24
21
18
15
12
9
6
3
3
6
9
12
15
18
21
24
1
2
3
4
5
6
7
8
9
10
11

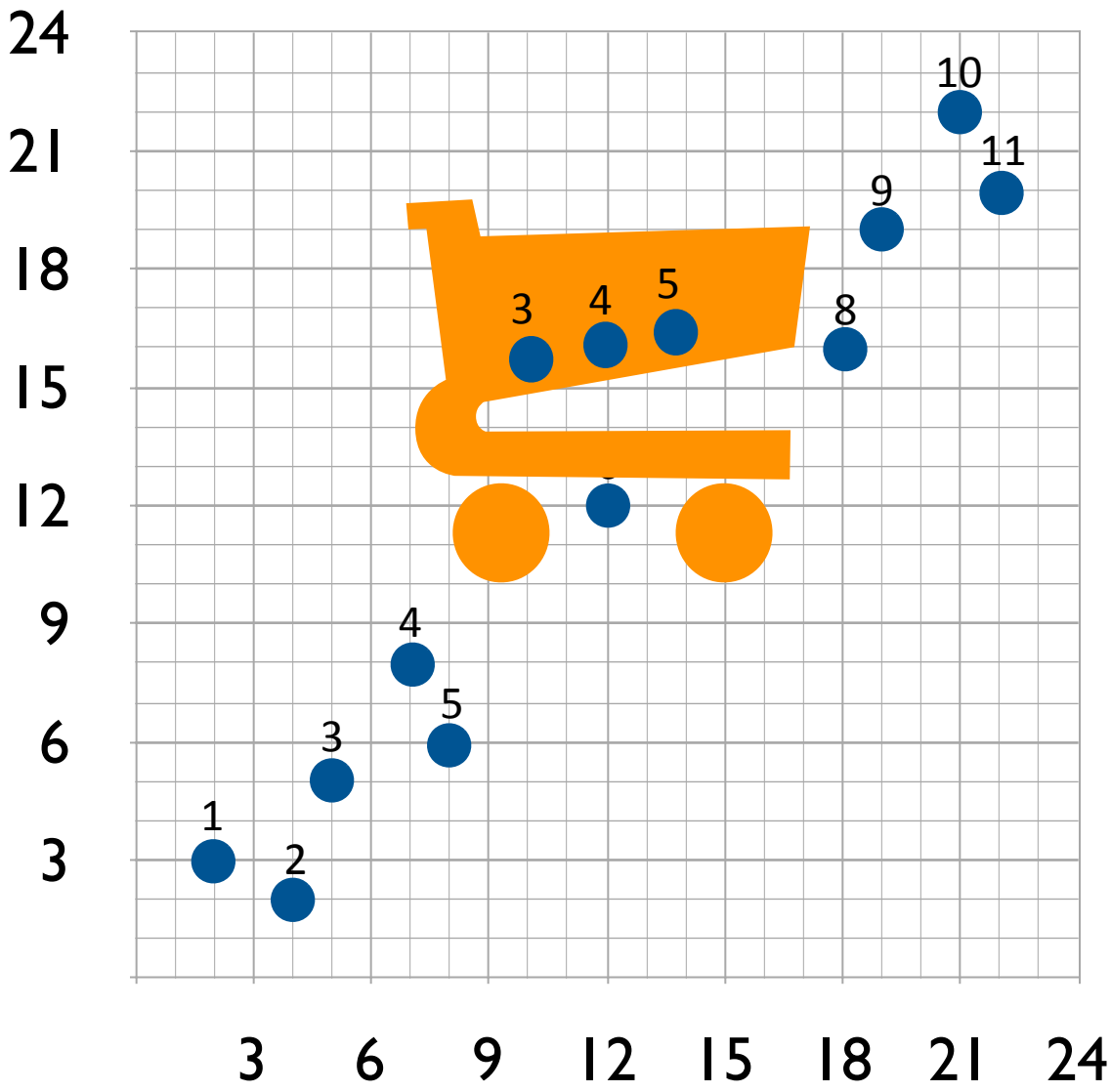

1 2 3
1 2 3

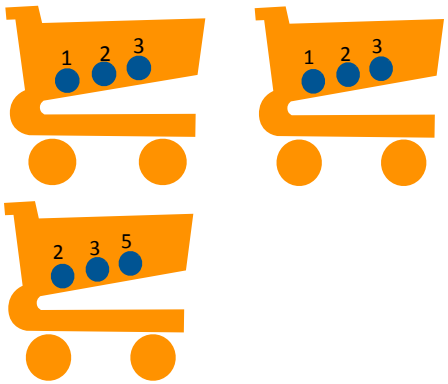

24
21
18
15
12
9
6
3
3
6
9
12
15
18
21
24
1
2
3
4
5
6
2
3
5
8
9
10
11

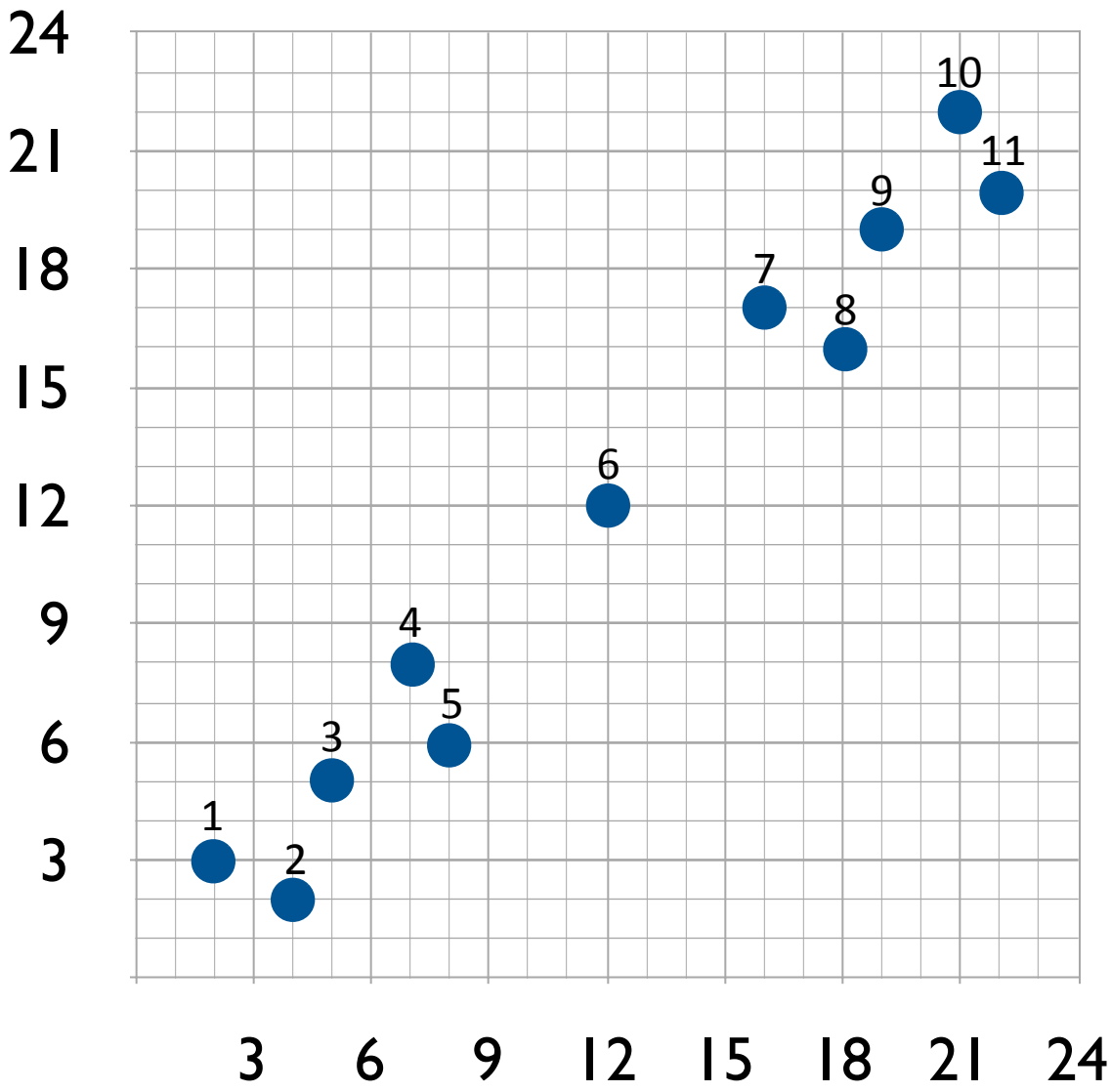

1
2
3
1
2
3

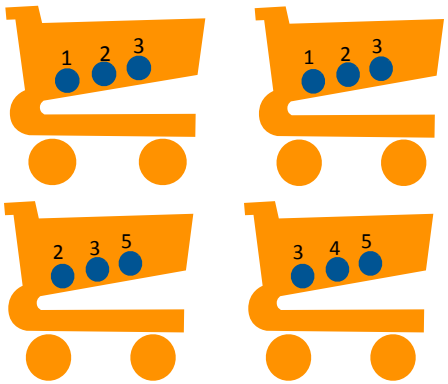

24
21
18
15
12
9
6
3
1
2
3
4
5
6
7
8
9
10
11
3
6
9
12
15
18
21
24

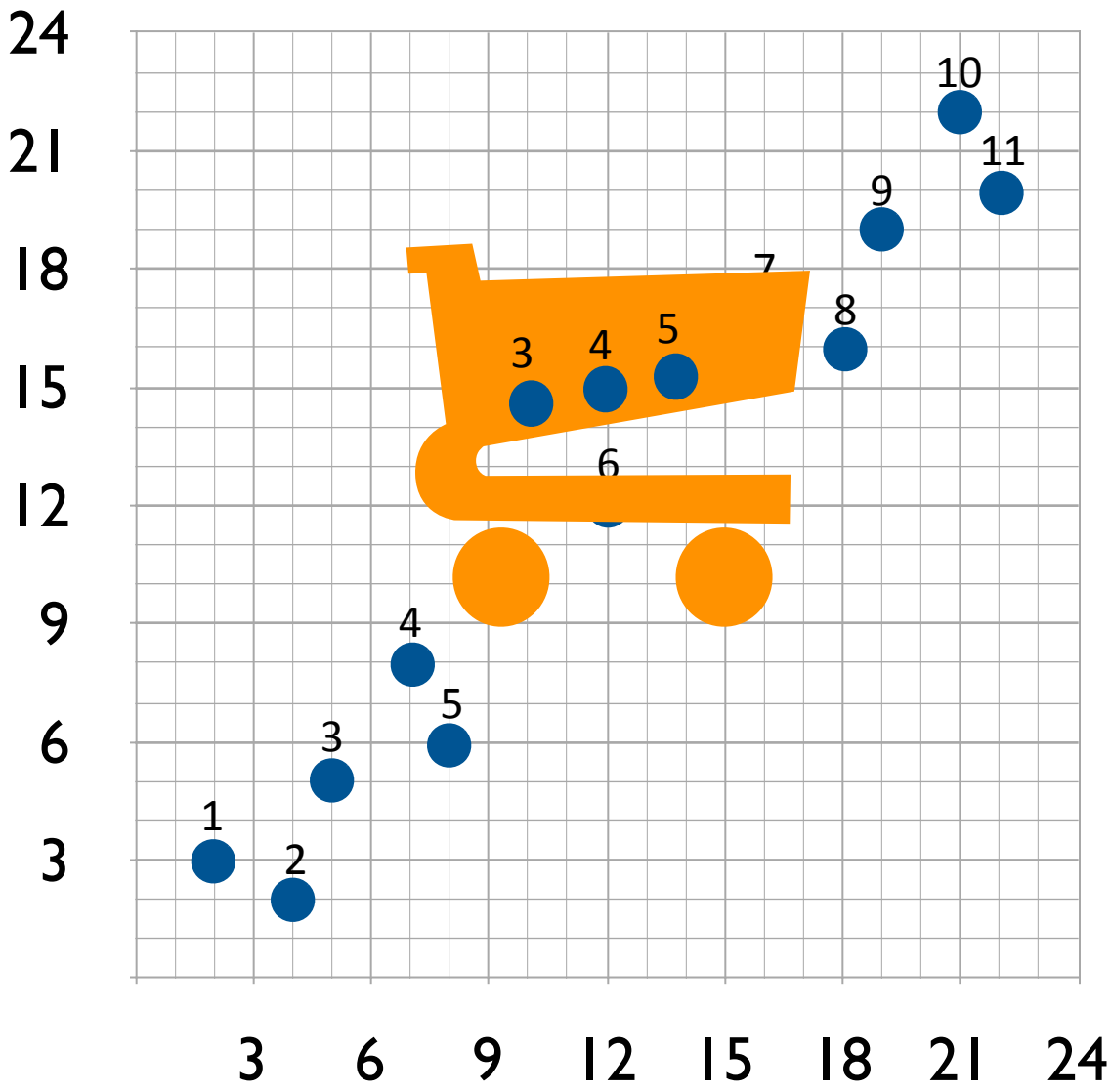

1 2 3
1 2 3
2 3 5

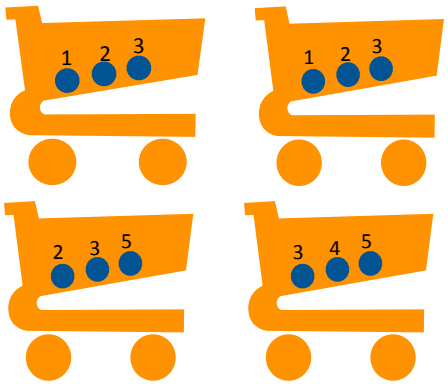

24
21
18
15
12
9
6
3
1
2
3
4
5
3
4
5
8
9
10
11
3
6
9
12
15
18
21
24

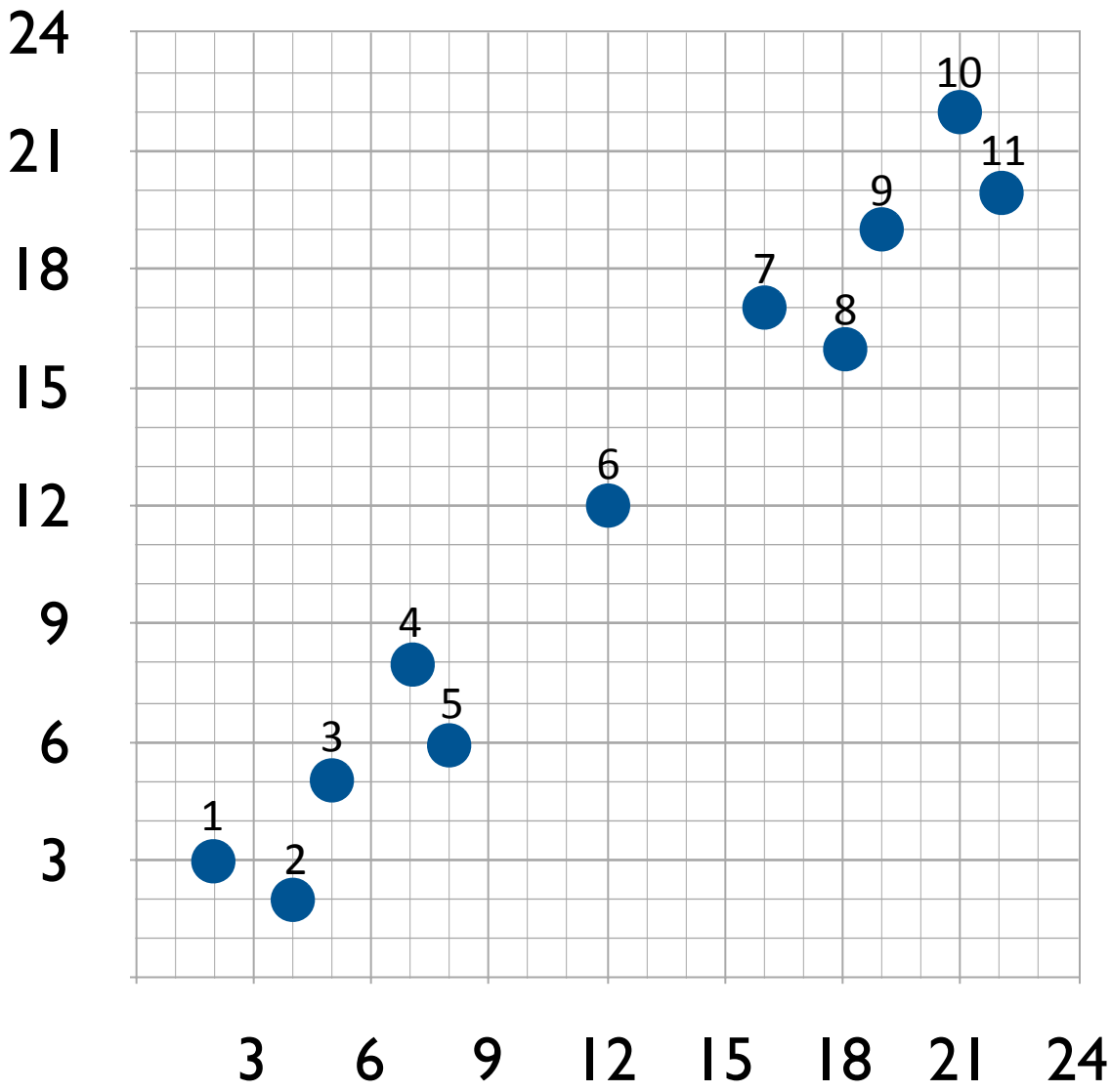

1 2 3
1 2 3
2 3 5

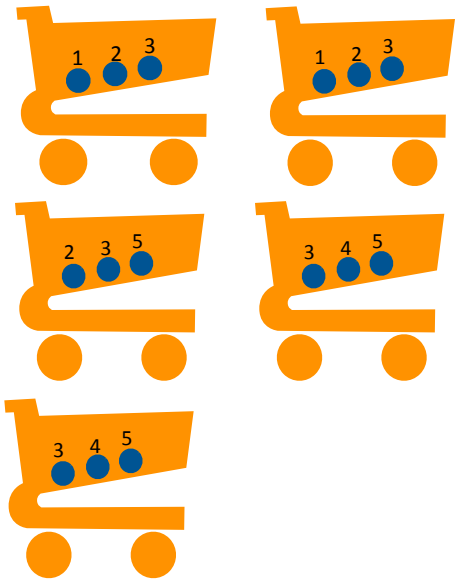

24
21
18
15
12
9
6
3
3
6
9
12
15
18
21
24
1
2
3
4
5
6
7
8
9
10
11

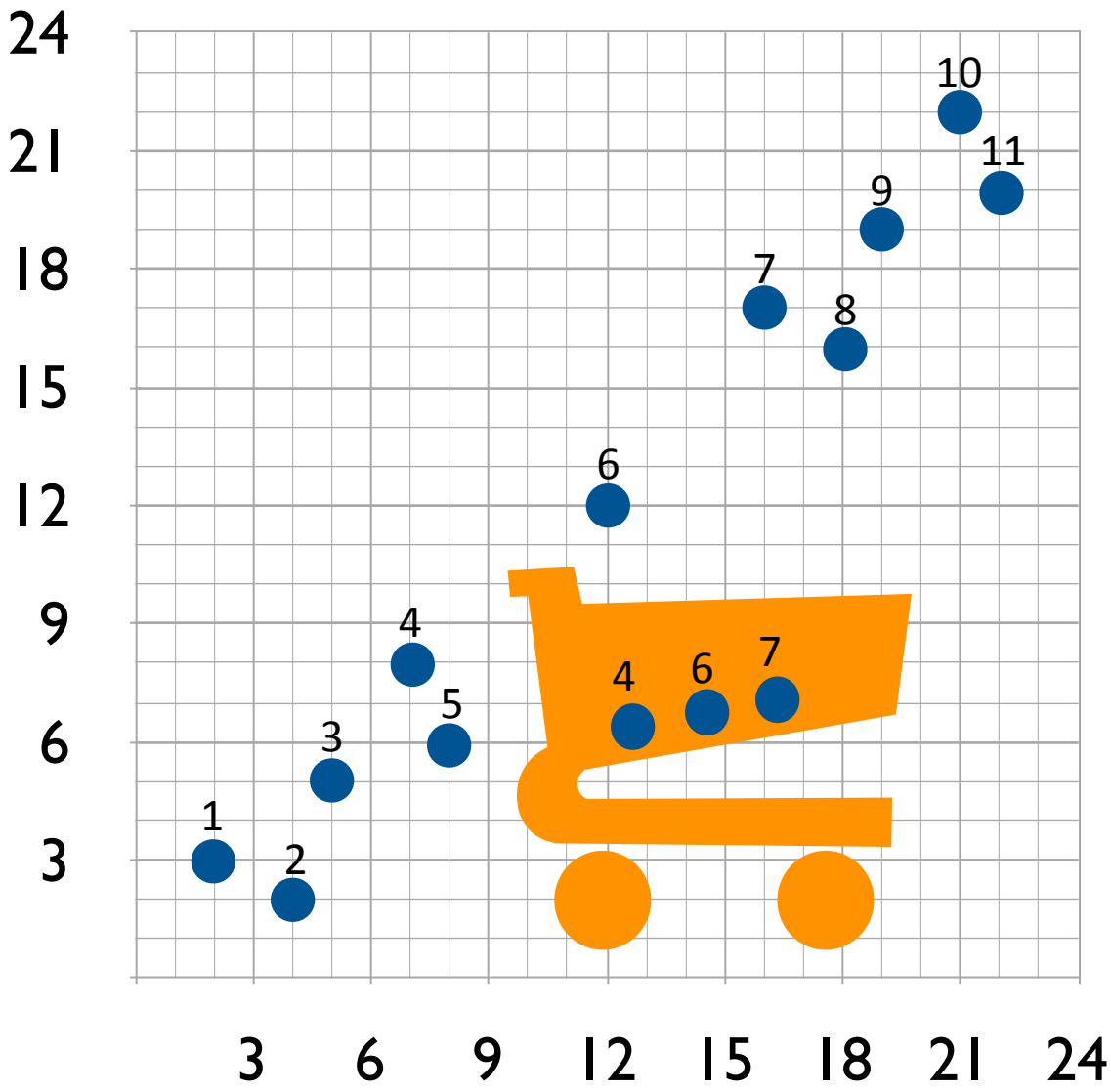

1 2 3
1 2 3
2 3 5
3 4 5

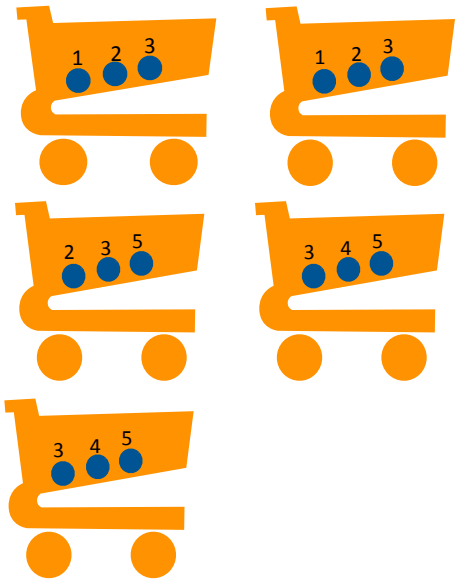

24
21
18
15
12
9
6
3
3
6
9
12
15
18
21
24
1
2
3
4
5
6
7
8
9
10
11

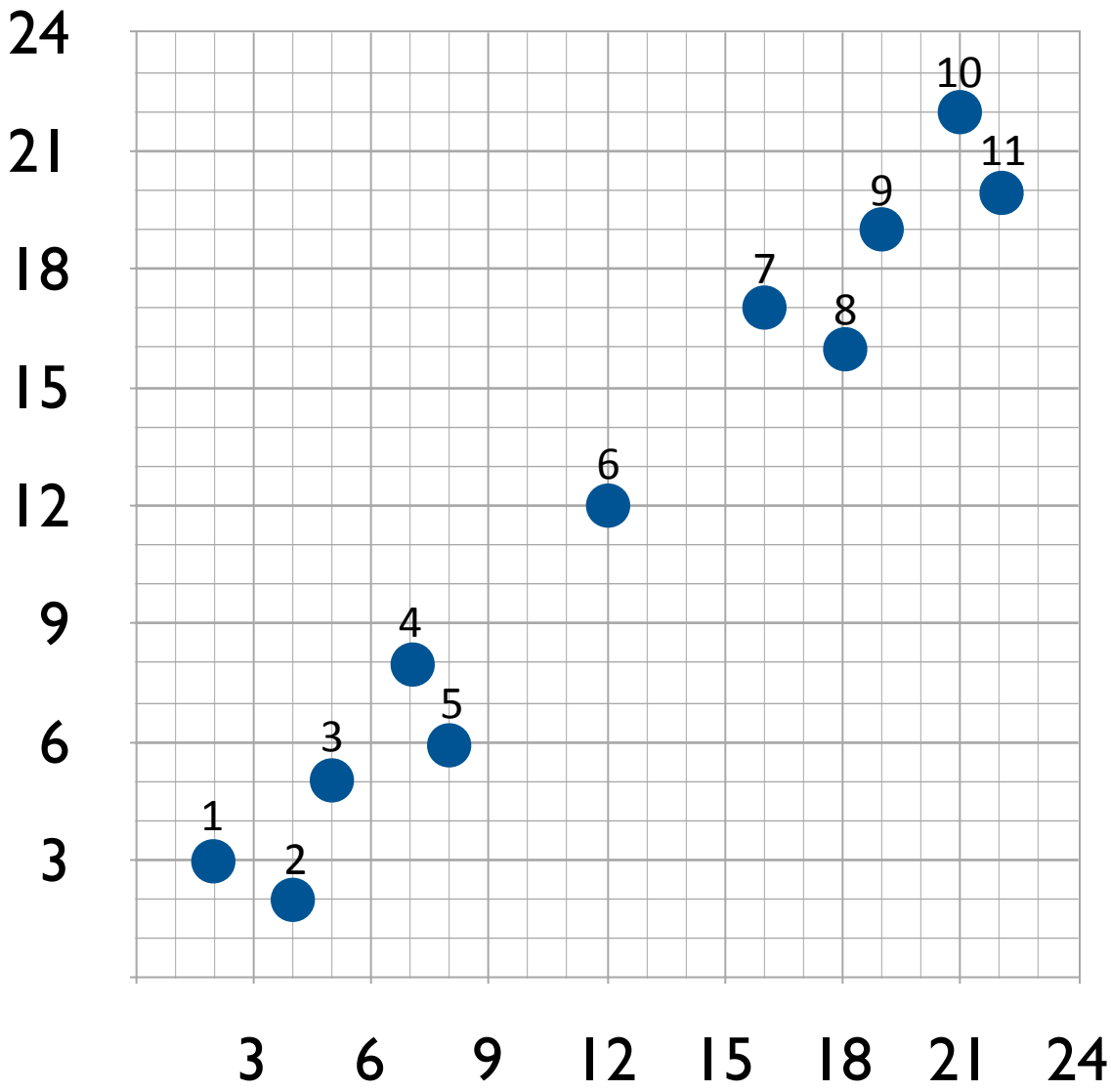

1 2 3
1 2 3
2 3 5
3 4 5

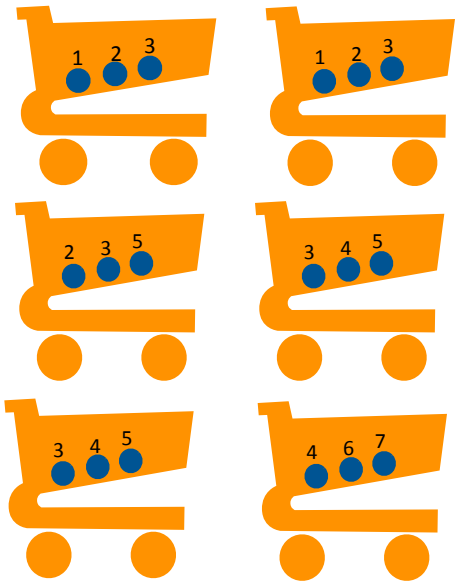

24
21
18
15
12
9
6
3
1
2
3
4
5
6
7
8
9
10
11
3
6
9
12
15
18
21
24

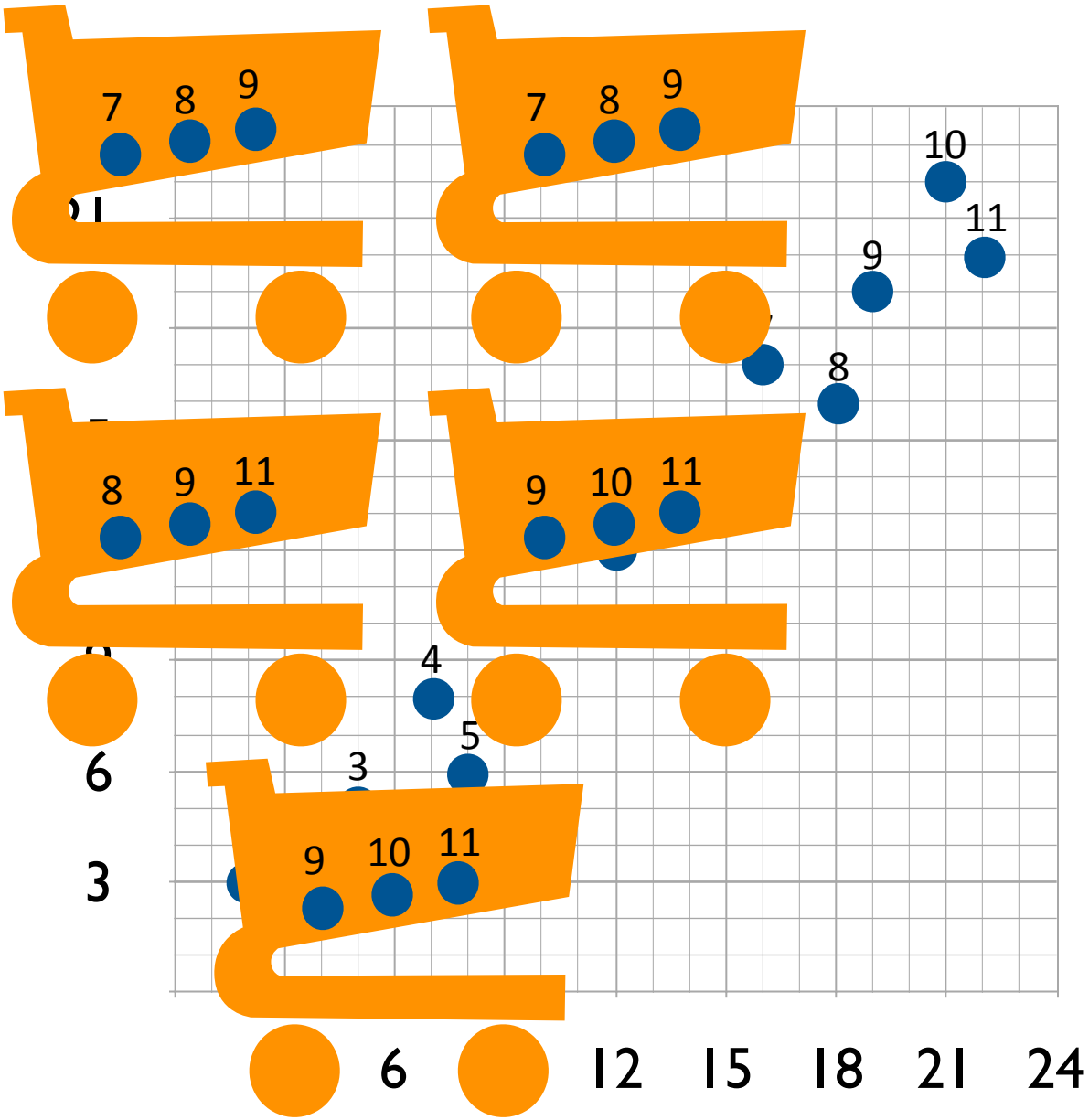

1 2 3
1 2 3
2 3 5
3 4 5
3 4 5

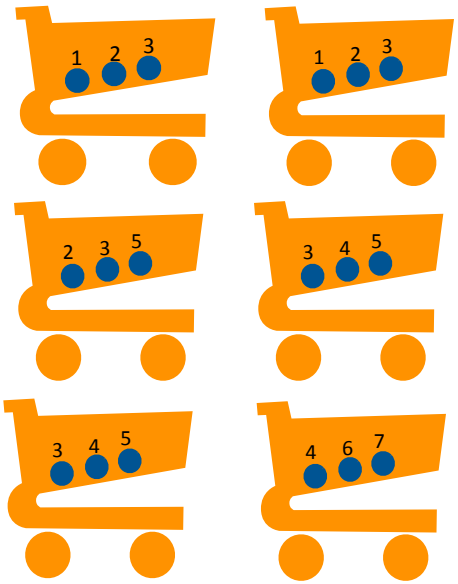

24
21
18
15
12
9
6
3
3
6
9
12
15
18
21
24
1
2
3
4
5
6
7
8
9
10
11
4
6
7

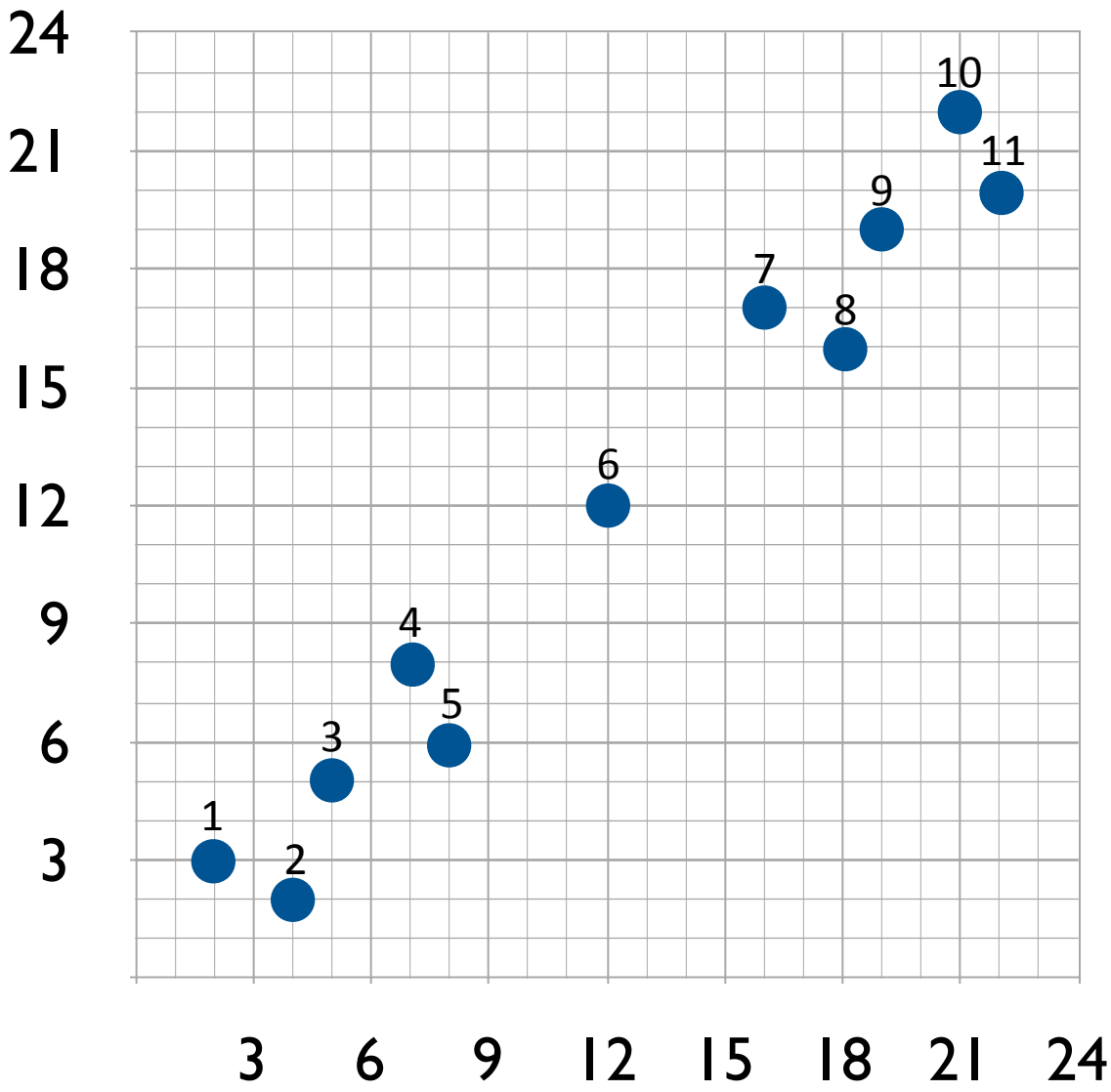

1 2 3
1 2 3
2 3 5
3 4 5
3 4 5

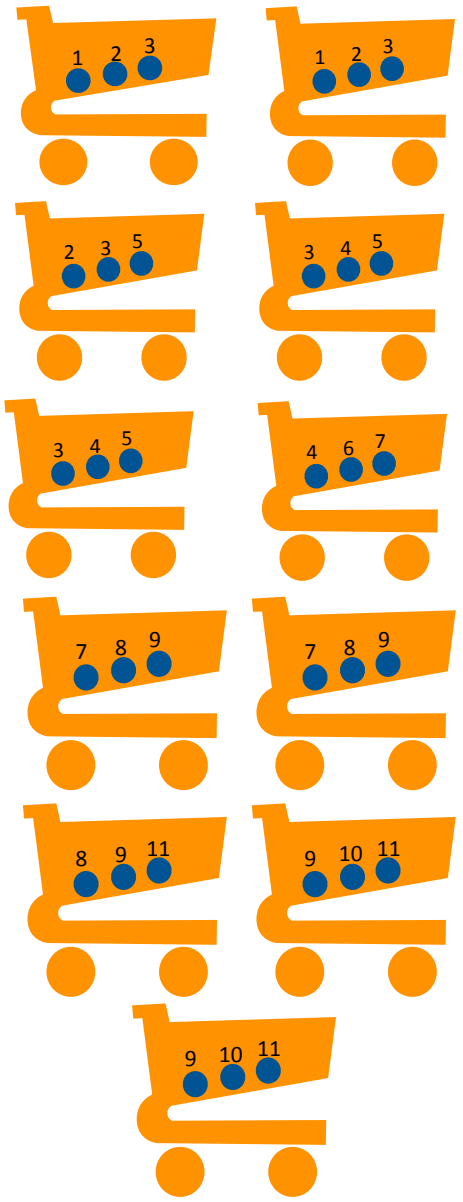

24
21
18
15
12
9
6
3
1
2
3
4
5
6
7
8
9
10
11
3
6
9
12
15
18
21
24


1 2 3
1 2 3
2 3 5
3 4 5
3 4 5
4 6 7


7 8 9
7 8 9
10
11
9
8
8 9 11
9 10 11
4
5
3
9 10 11
6
3
6 12 15 18 21 24


1 2 3
1 2 3
2 3 5
3 4 5
3 4 5
4 6 7


24
21
18
15
12
9
6
3
3
6
9
12
15
18
21
24
1
2
3
4
5
6
7
8
9
10
11

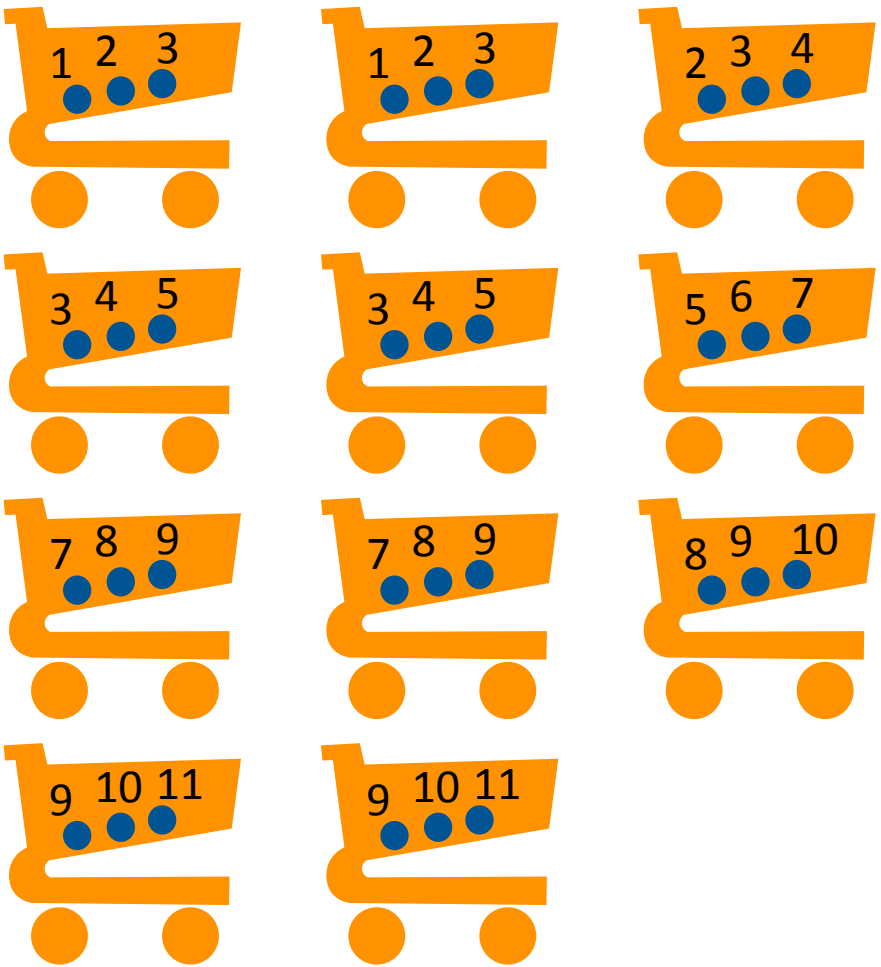

1 2 3
1 2 3
2 3 5
3 4 5
3 4 5
4 6 7
7 8 9
7 8 9
8 9 11
9 10 11
9 10 11

k-nearest neighbors on Y-axes

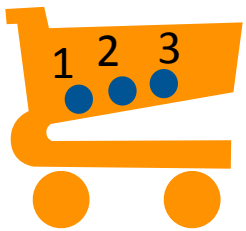
k-nearest neighbors on X-axes

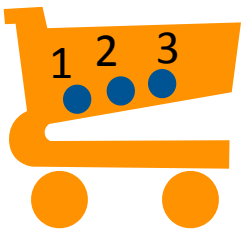
k-nearest neighbors on Y-axes

k-nearest neighbors on X-axes

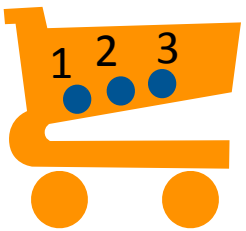


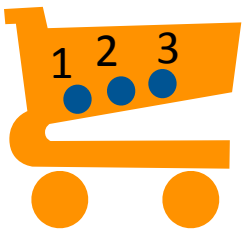
k-nearest neighbors on Y-axes





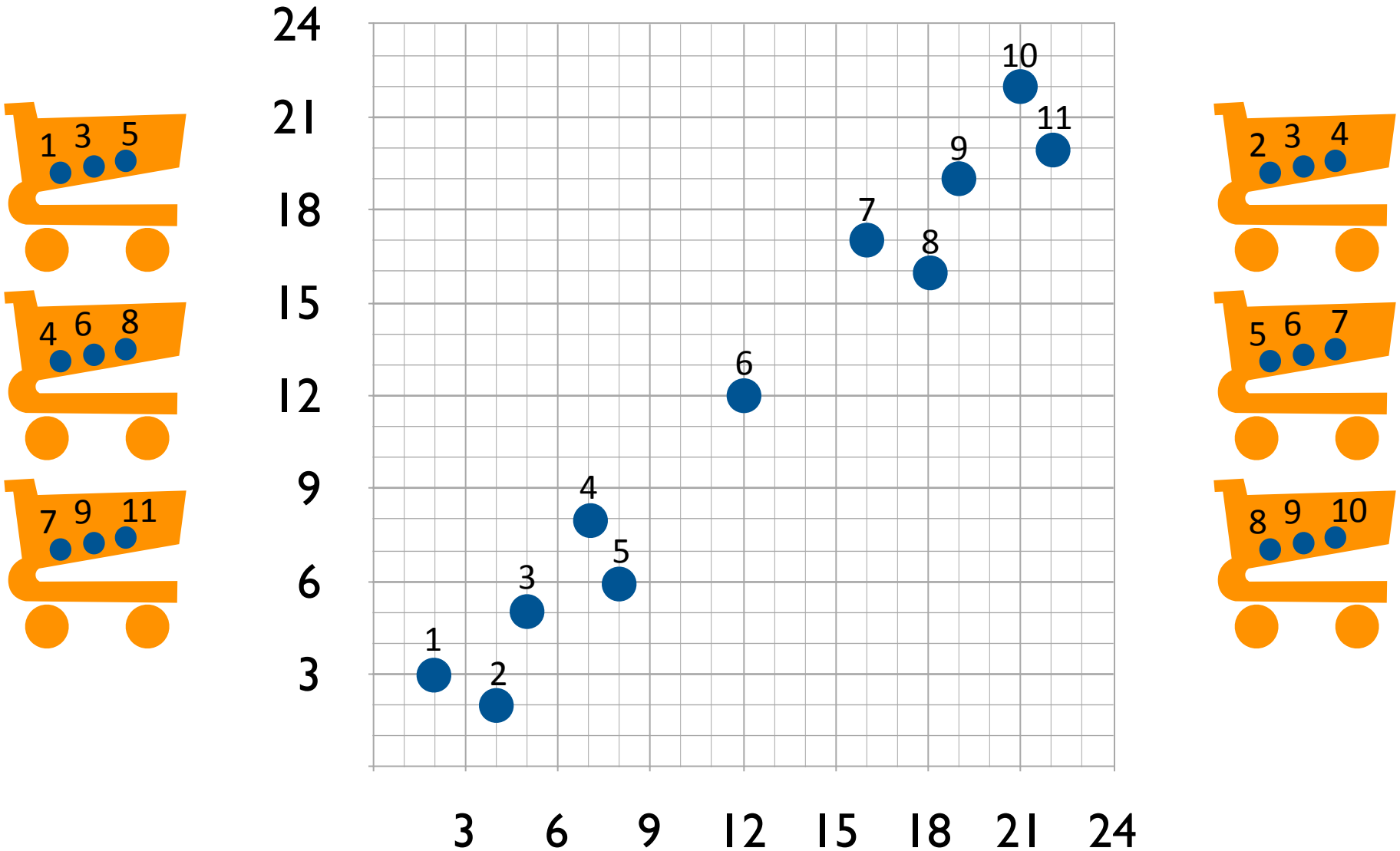
k-nearest neighbors on X-axes



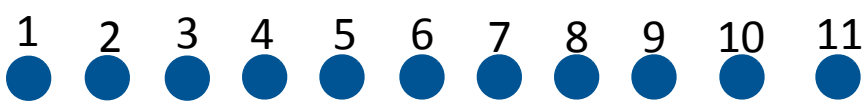


k-nearest neighbors on Y-axes

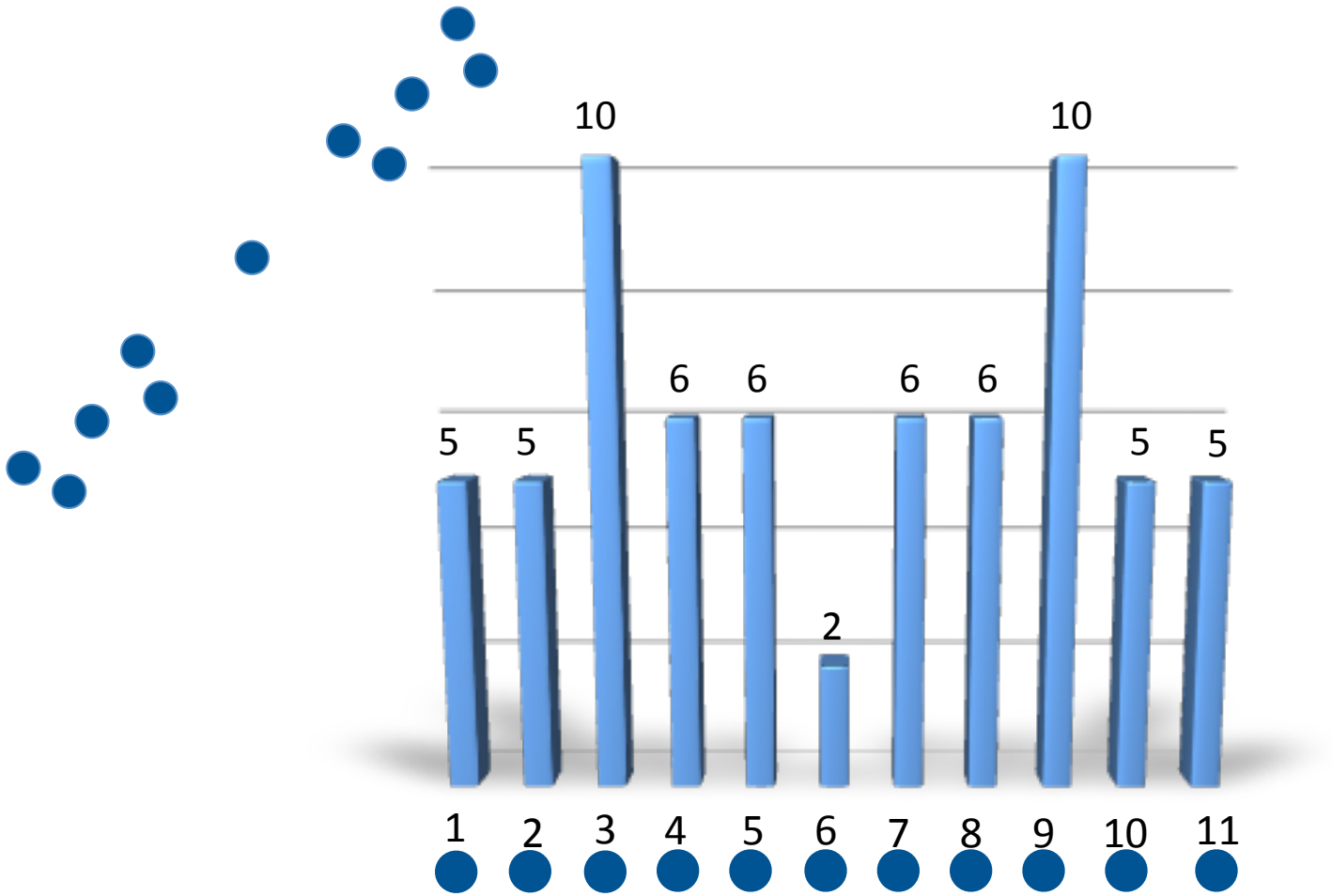
k-nearest neighbors on X-axes



Item Frequencies for k=3

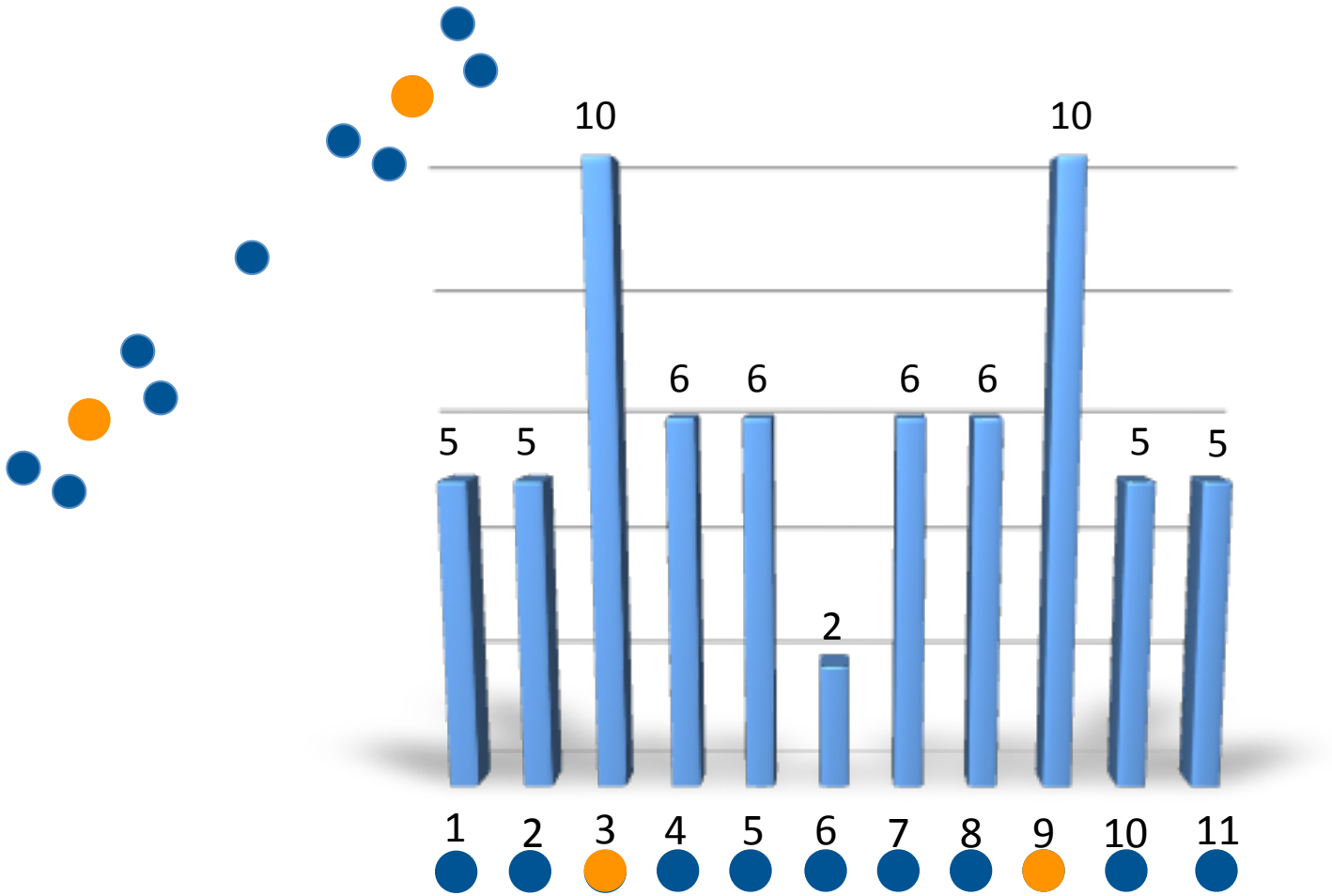


Item Frequencies for k=3



Item Frequencies for k=3

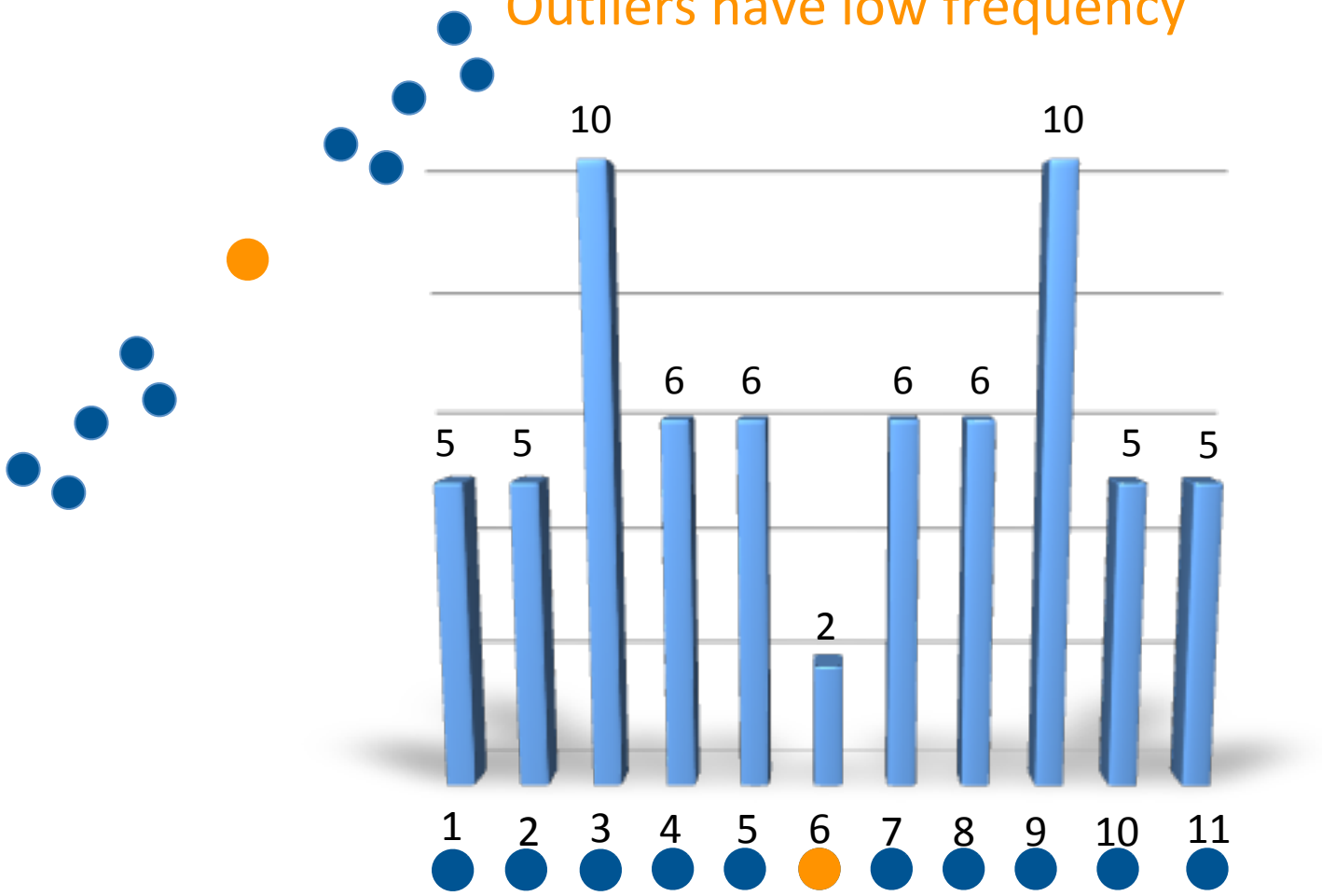
Central points have high frequency



Item Frequencies for k=3

Central points have high frequency

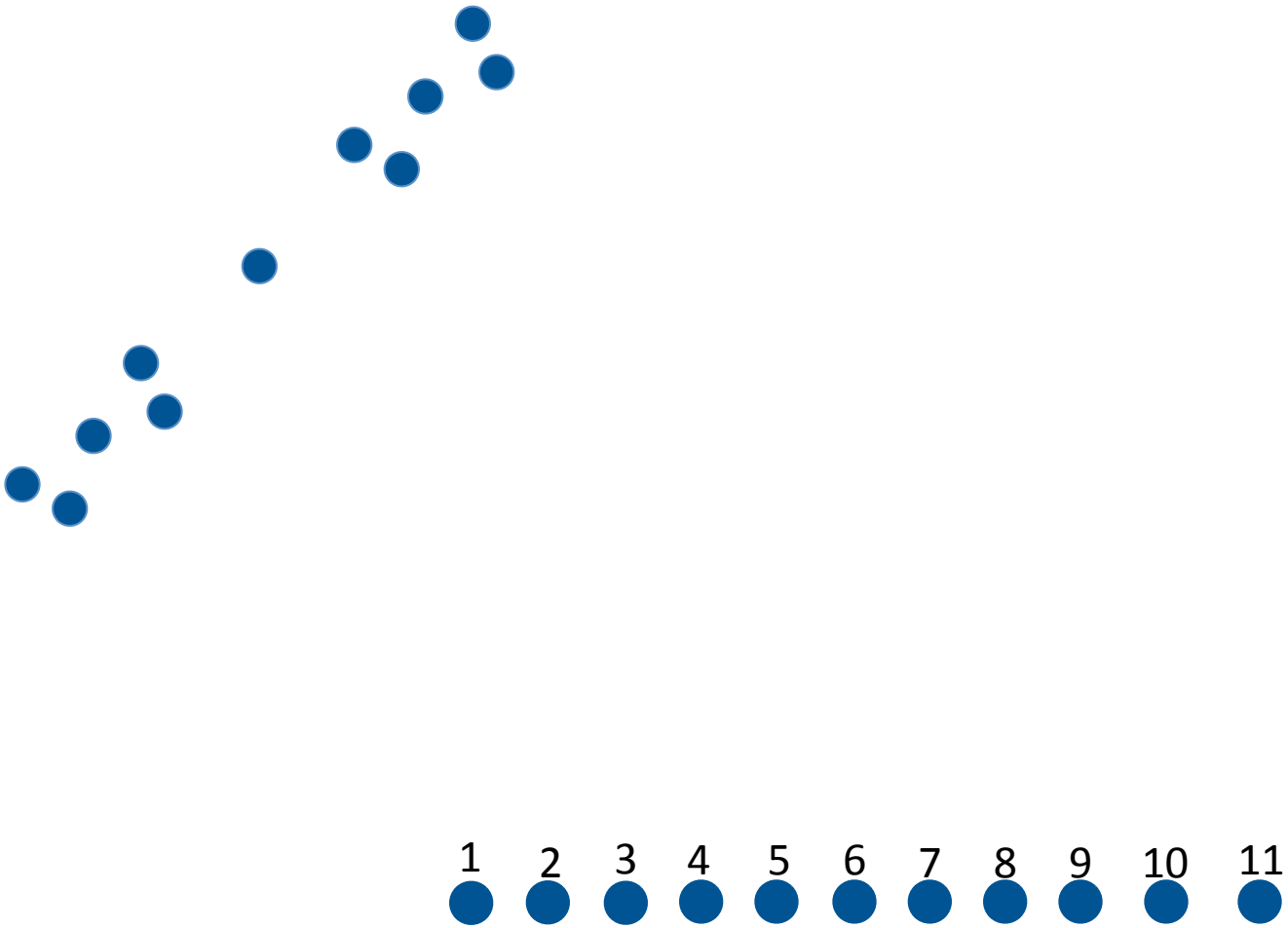
Outliers have low frequency



Item Frequencies for k=6

Increasing k corresponds to zooming out

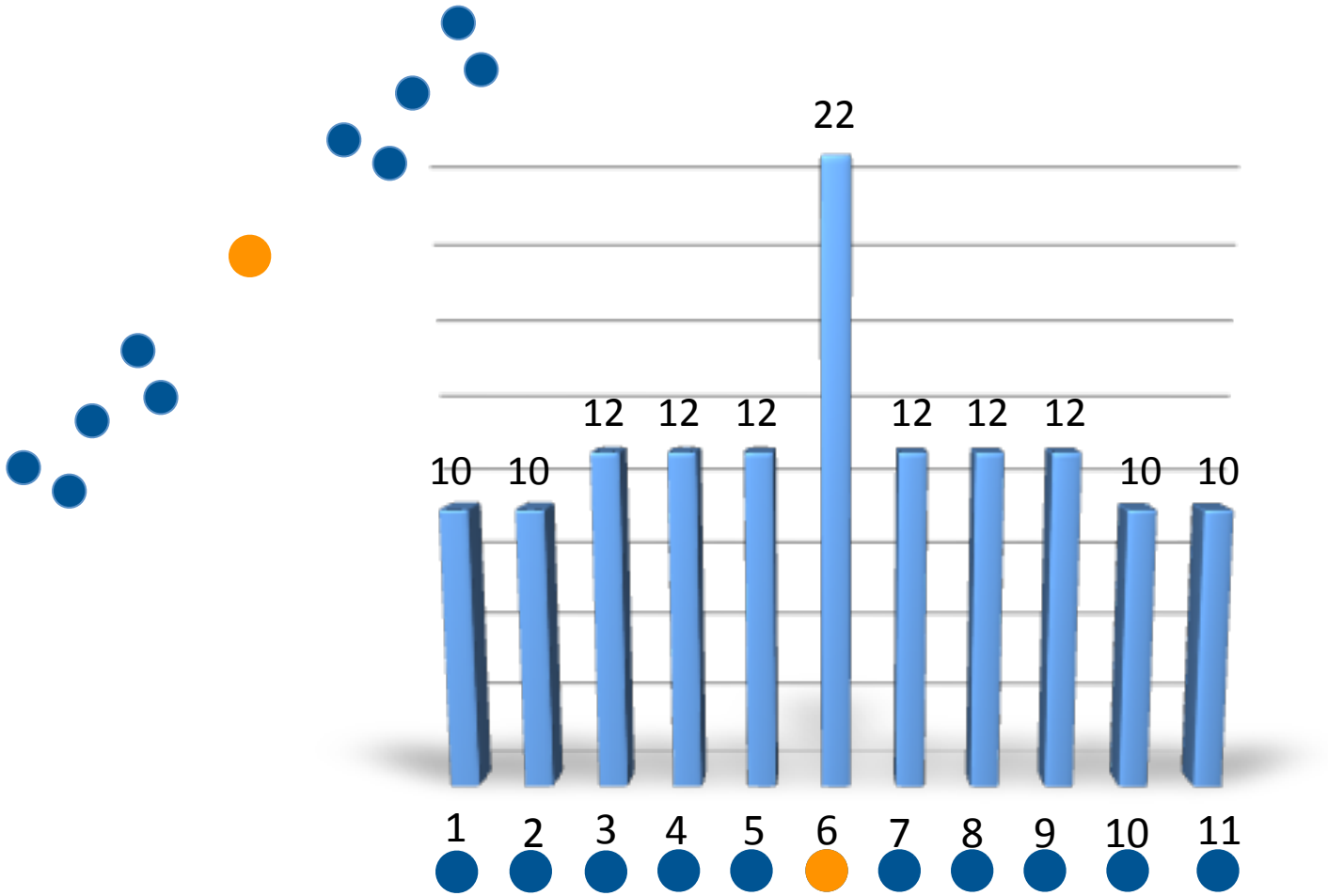
Central points have high frequency



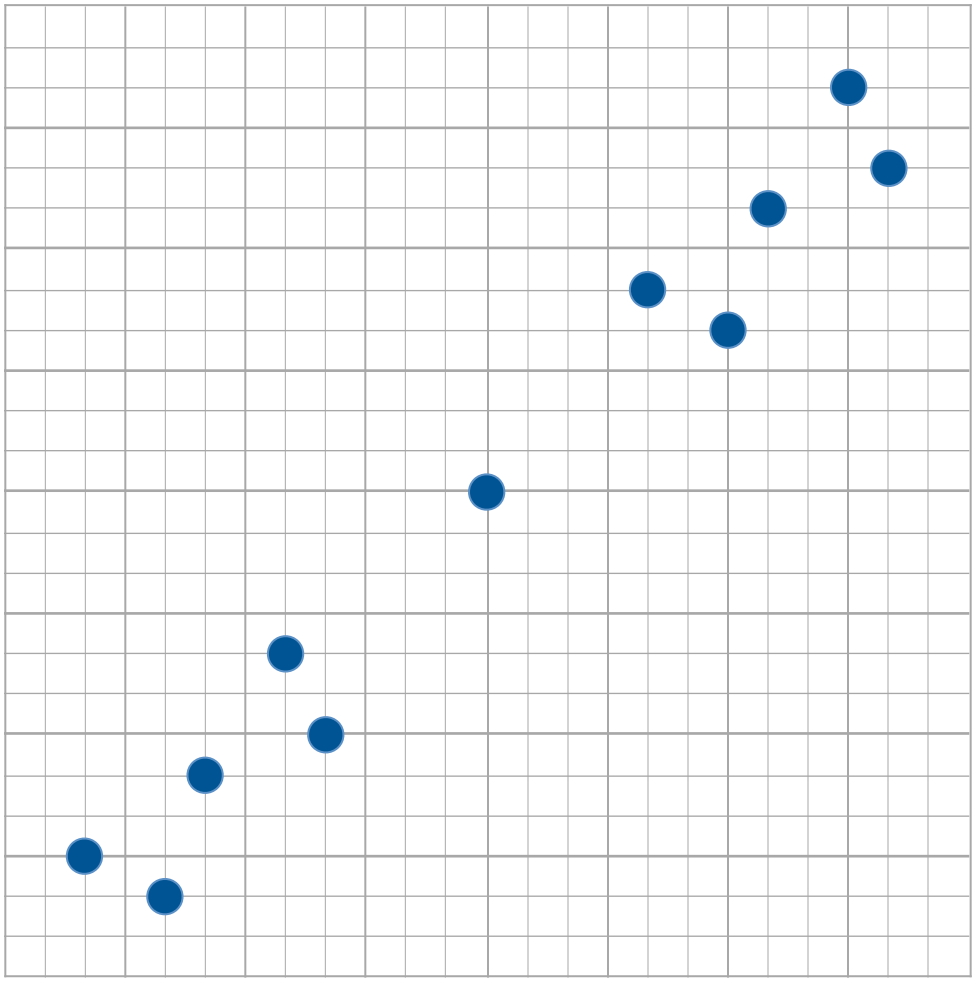
Item Frequencies for k=6

Increasing k corresponds to zooming out

Central points have high frequency

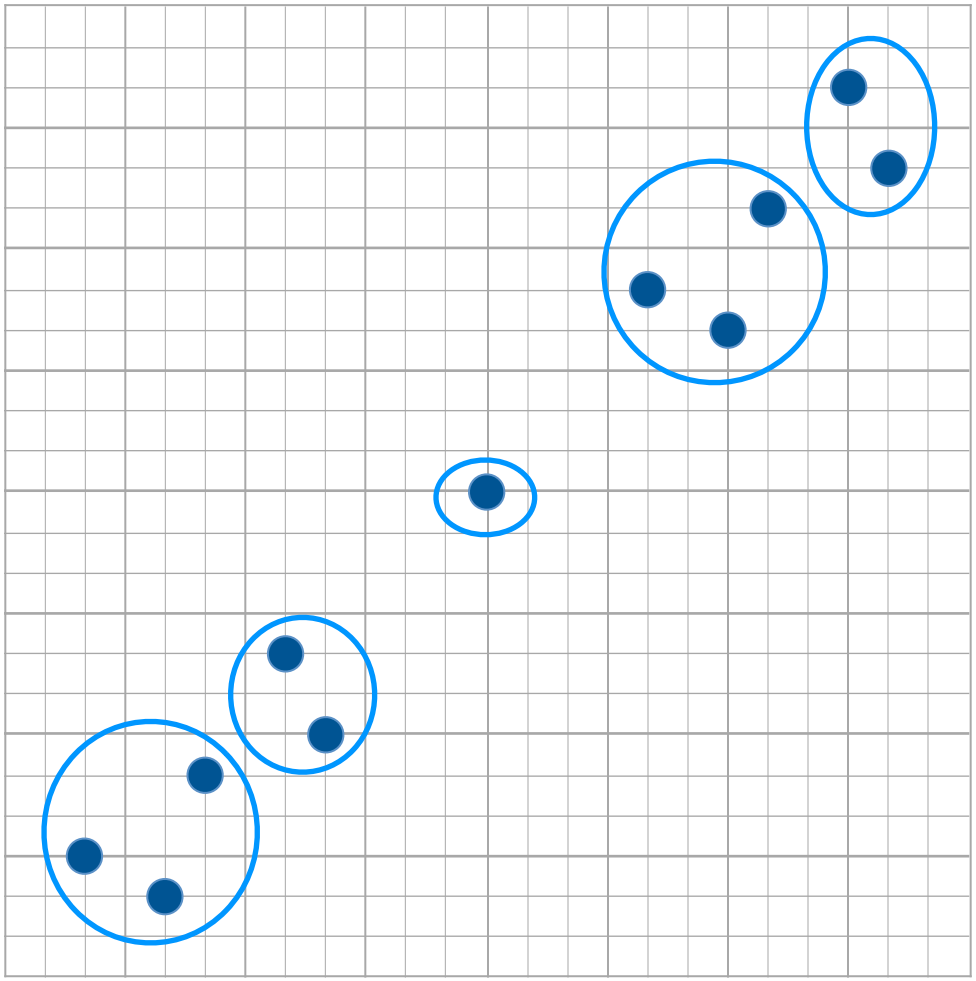


Frequent Sets/Clusters



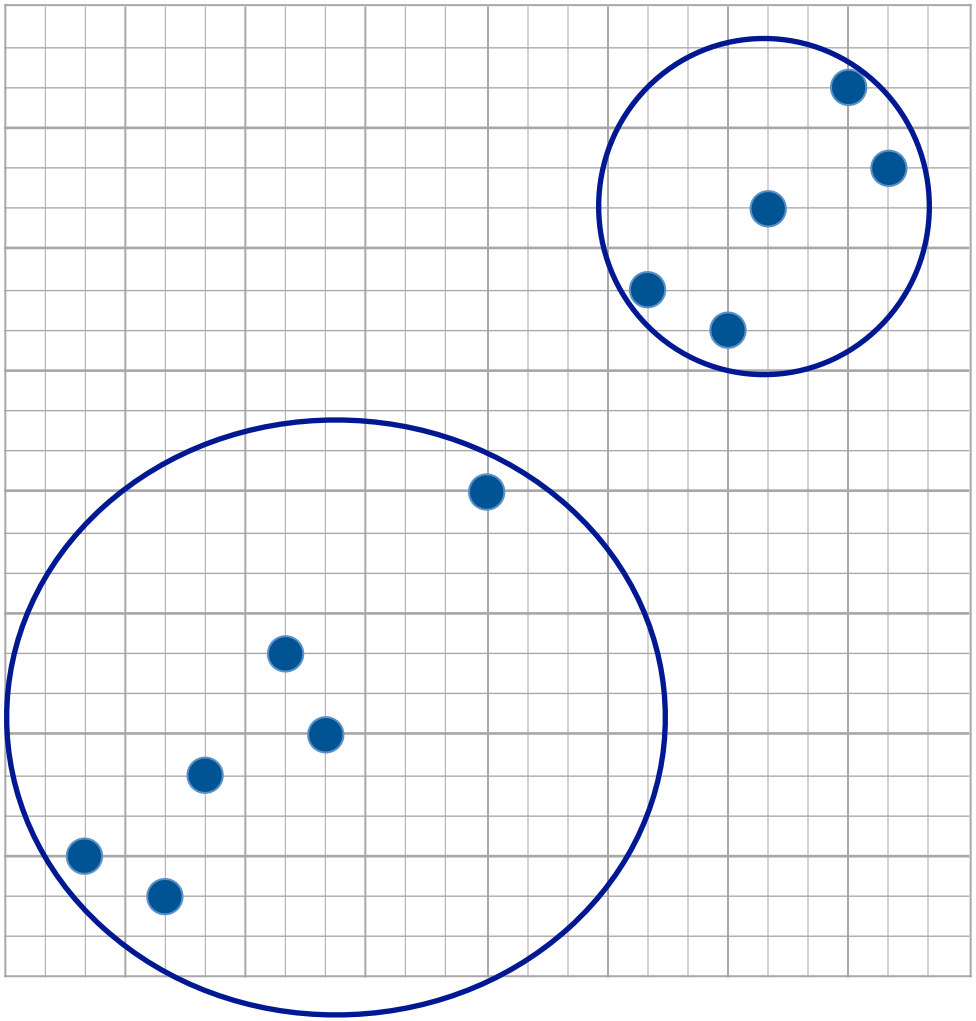
Frequent Sets/Clusters

- for k=3



Frequent Sets/Clusters

- for k=3
- for k=6



Subspace Clustering

- Support is computed for every dimension separately
- For every cluster, non-supporting dimensions can be disregarded
  (no curse of dimensionality!)

Ongoing research

- fixed k vs. auto k
- validation
- combining subspaces
- number of (subspace) clusters
- association rule mining for numerical data
- ...


250 ml

European Mammals database

- presence records of 124 mammals
- avg. temperature, elevation, rainfall, temp. range
- For 2183 areas of 50x50km areas in Europe

Easily visualised

Well-known regions

Allowing us to validate clusters

Cartifying Mammals

Two similarity measures

Cartifying Mammals

Two similarity measures

- presence of mammals
  - $| t_1 \cap t_2 | / | t_1 \cup t_2 |$
  - over all binary attributes at once

Cartifying Mammals

Two similarity measures

- presence of mammals
  - $| t_1 \cap t_2 | / | t_1 \cup t_2 |$
  - over all binary attributes at once

- for the numeric attributes
  - Euclidean distance
  - over the four numeric attributes combined

Cartifying Mammals

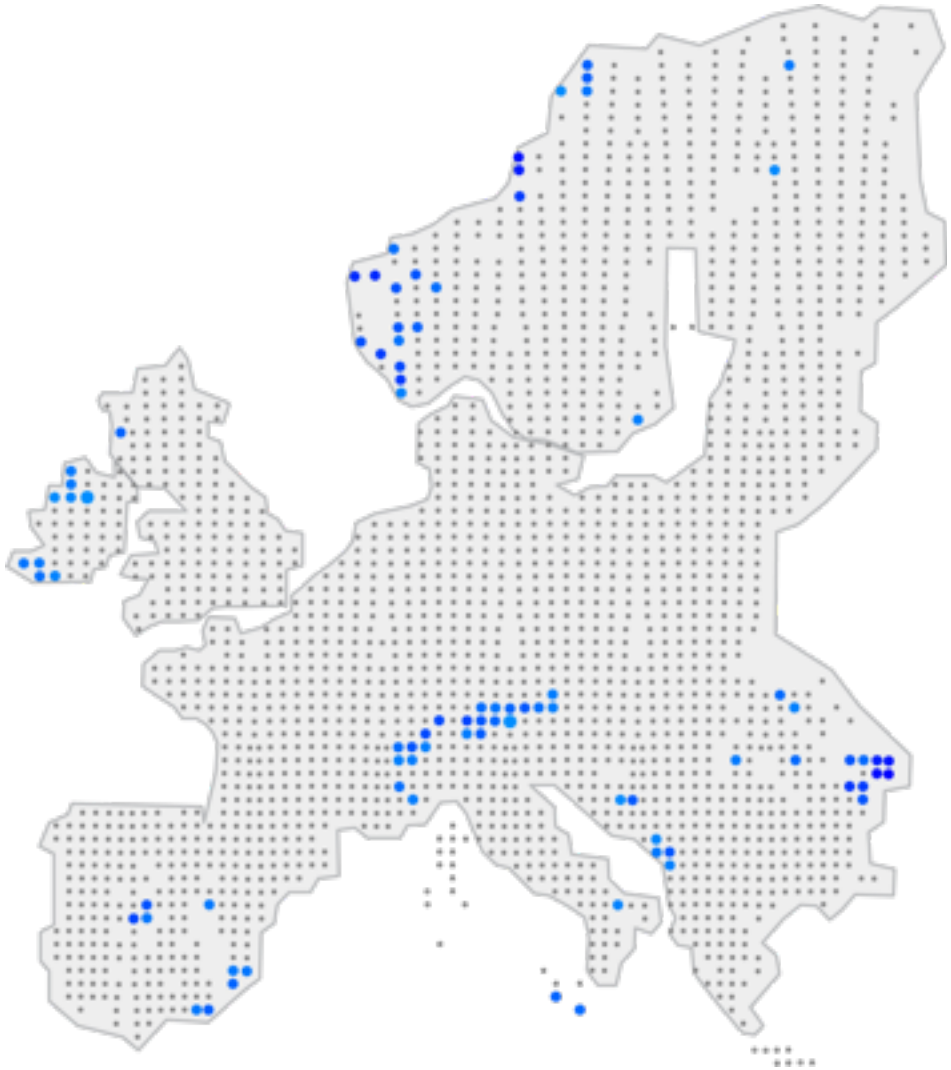
Two similarity measures

- presence of mammals
  - $| t_1 \cap t_2 | / | t_1 \cup t_2 |$
  - over all binary attributes at once

- for the numeric attributes
  - Euclidean distance
  - over the four numeric attributes combined

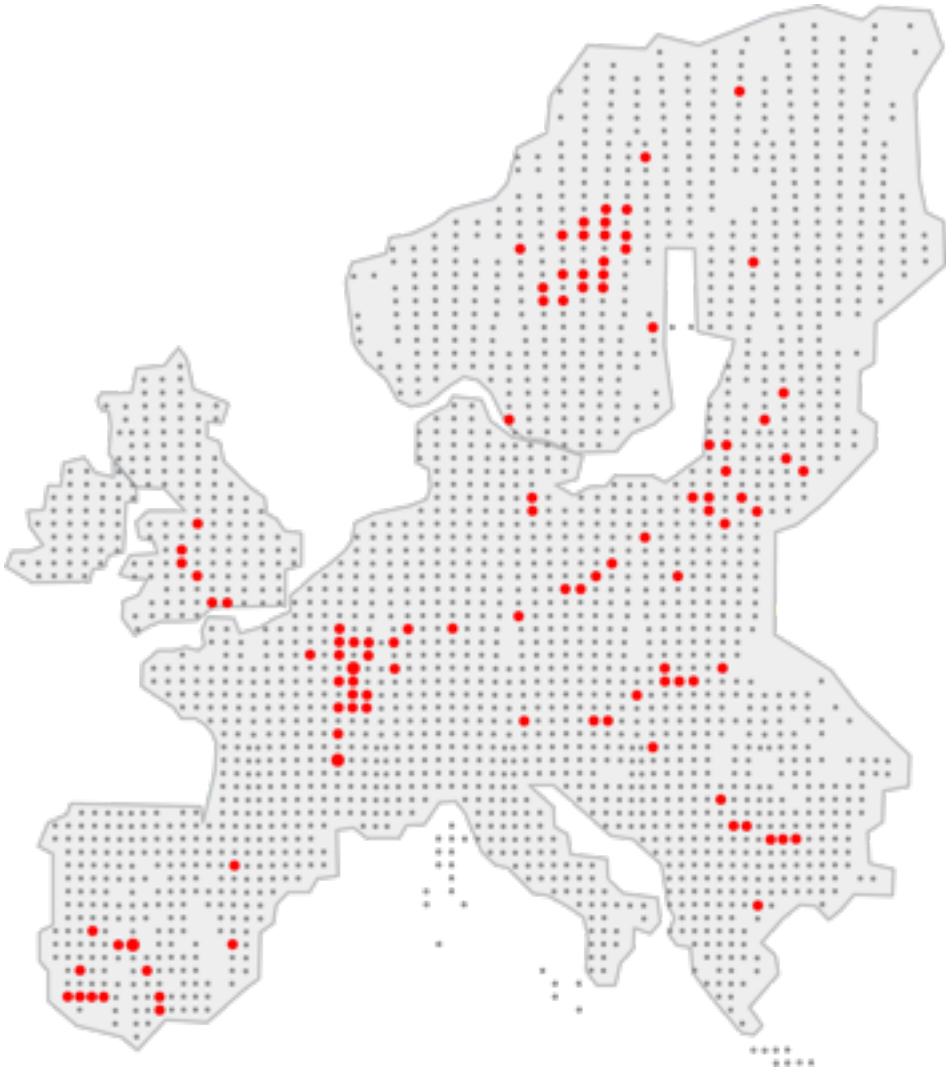
Providing two cartified databases, C_b and C_n, which we concatenate into one, C

Outliers



items with *supp* < 300 plotted, for data *k* = 400

Centers



items with *supp* > 120 plotted, for data *k* = 40

Clustering & Cartification

Clustering & Cartification

Many possibilities

Clustering & Cartification

Many possibilities

We use hierarchical agglomerative clustering

- support($X \cup Y$) as similarity

Clustering & Cartification

Many possibilities

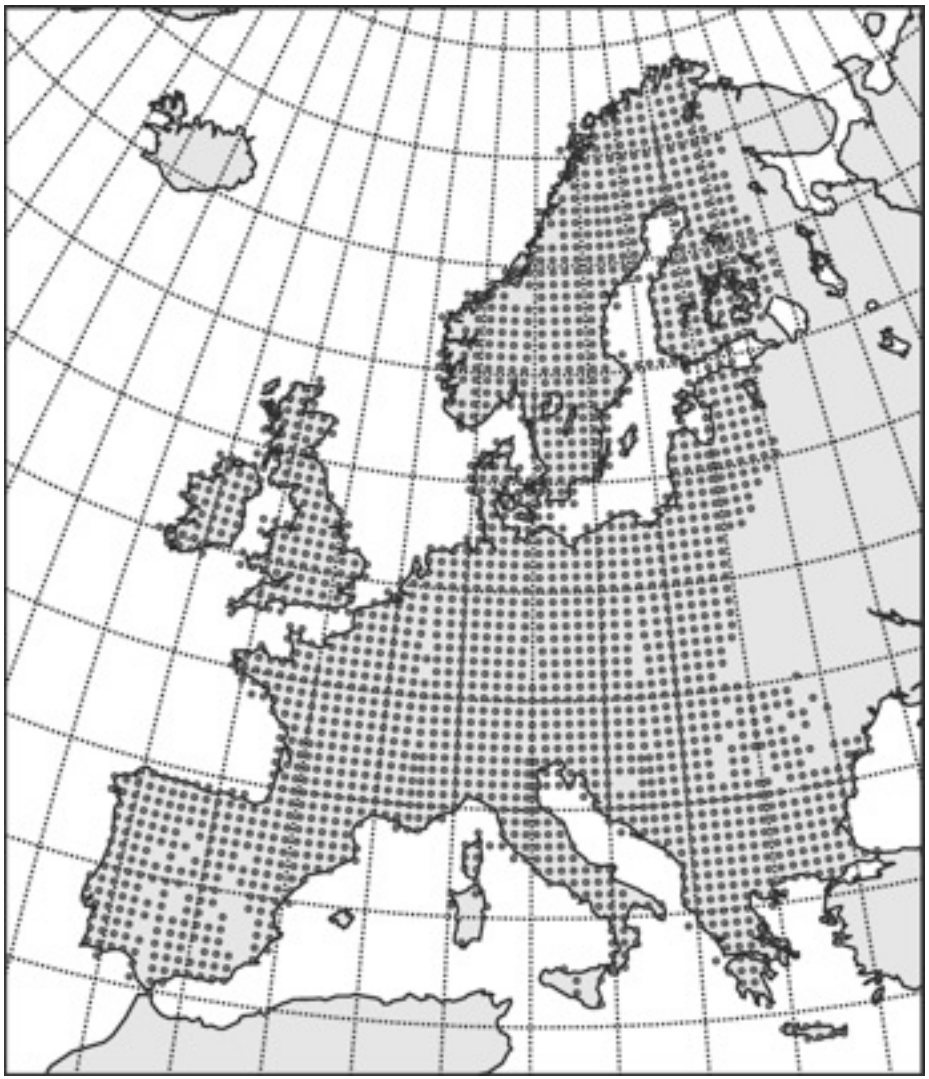
We use hierarchical agglomerative clustering

- support($X \cup Y$) as similarity

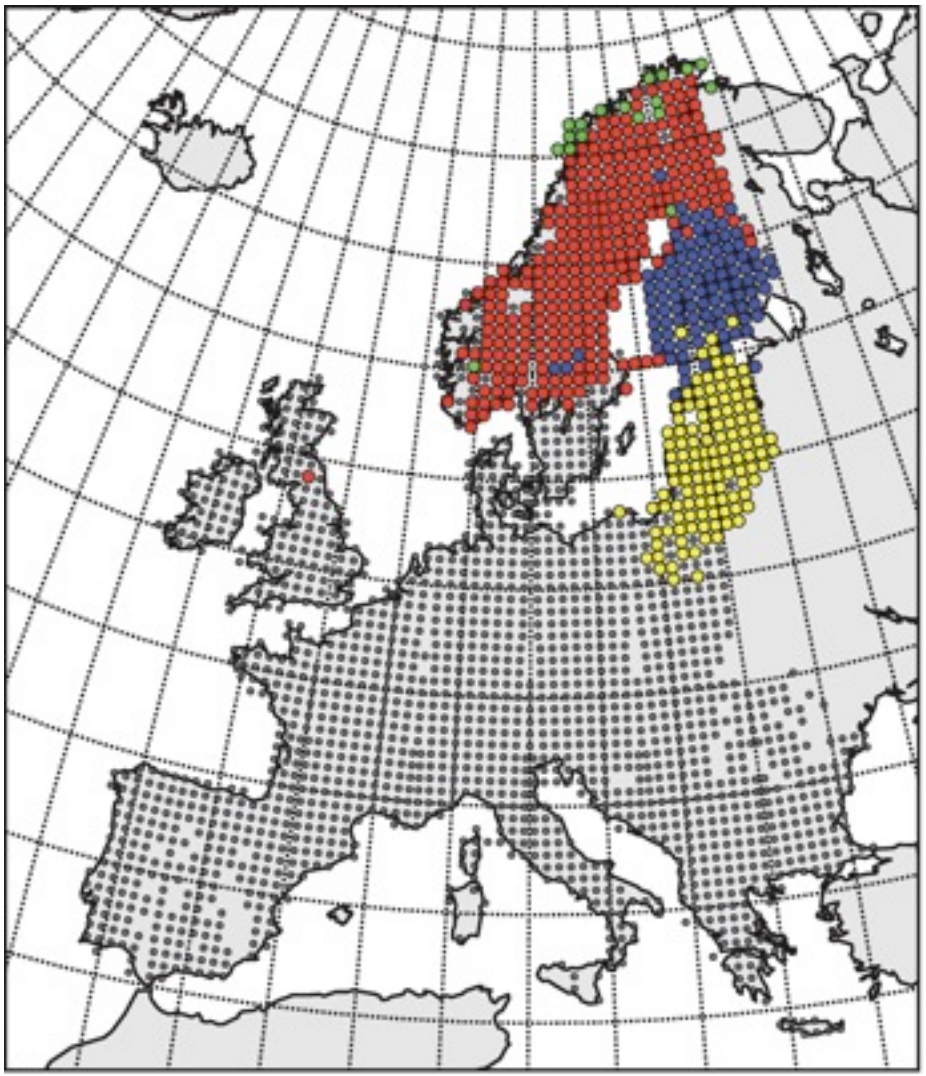
Simplest approach we could think of

+ it works!
- $X \cup Y$ needs to exist in database

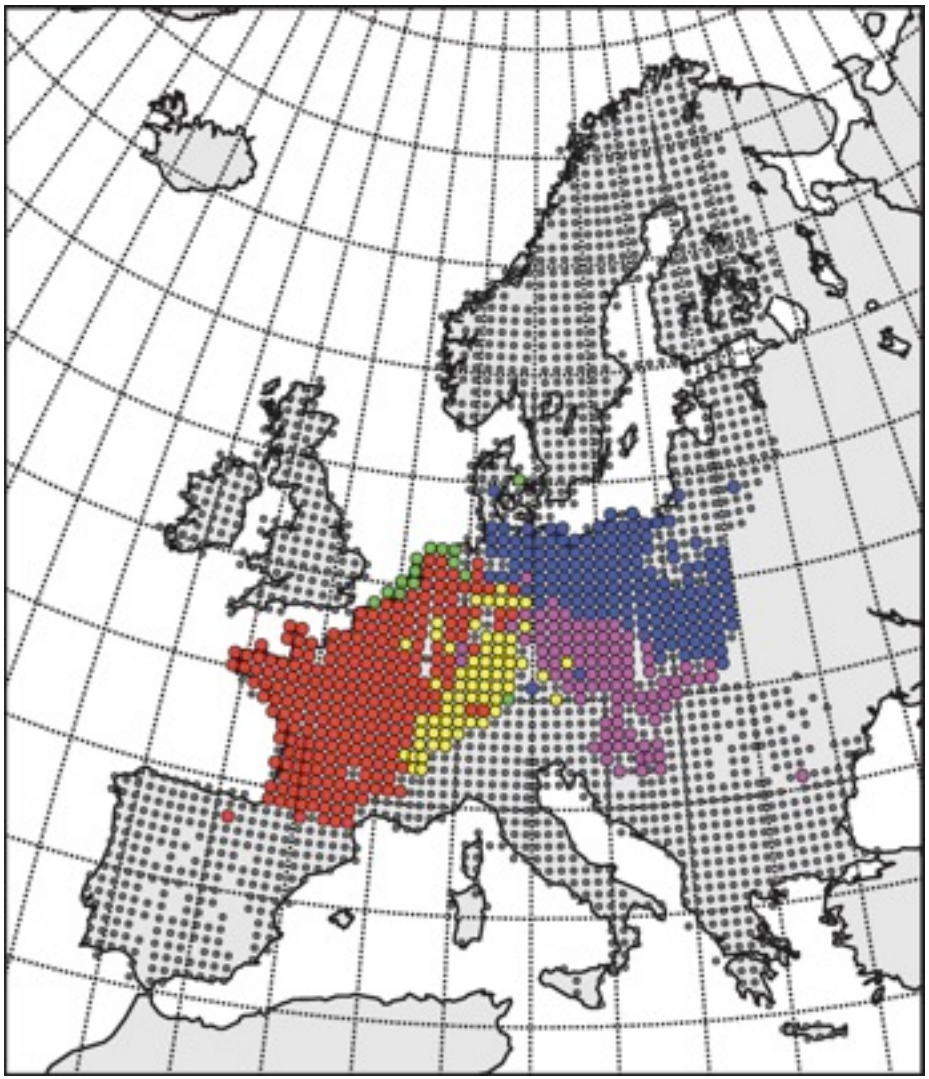
Clusters of Mammals



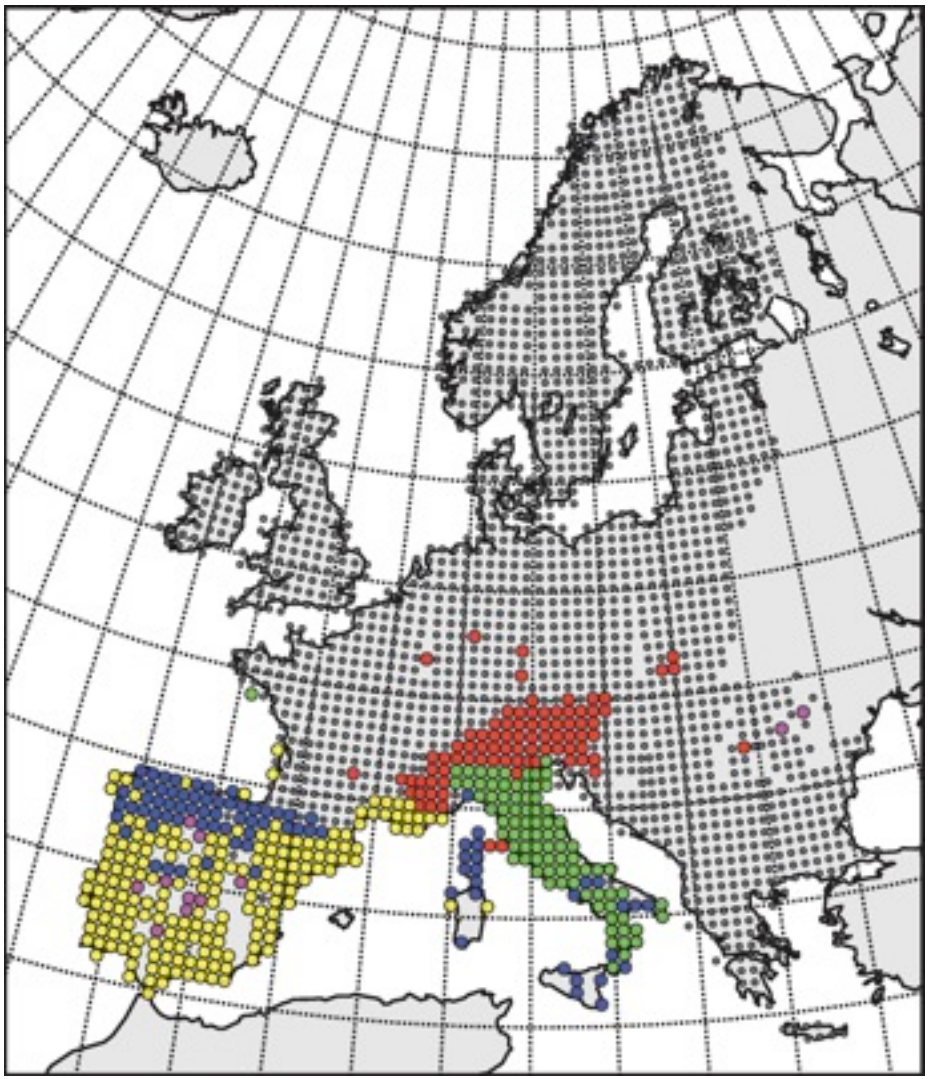
Clusters of Mammals



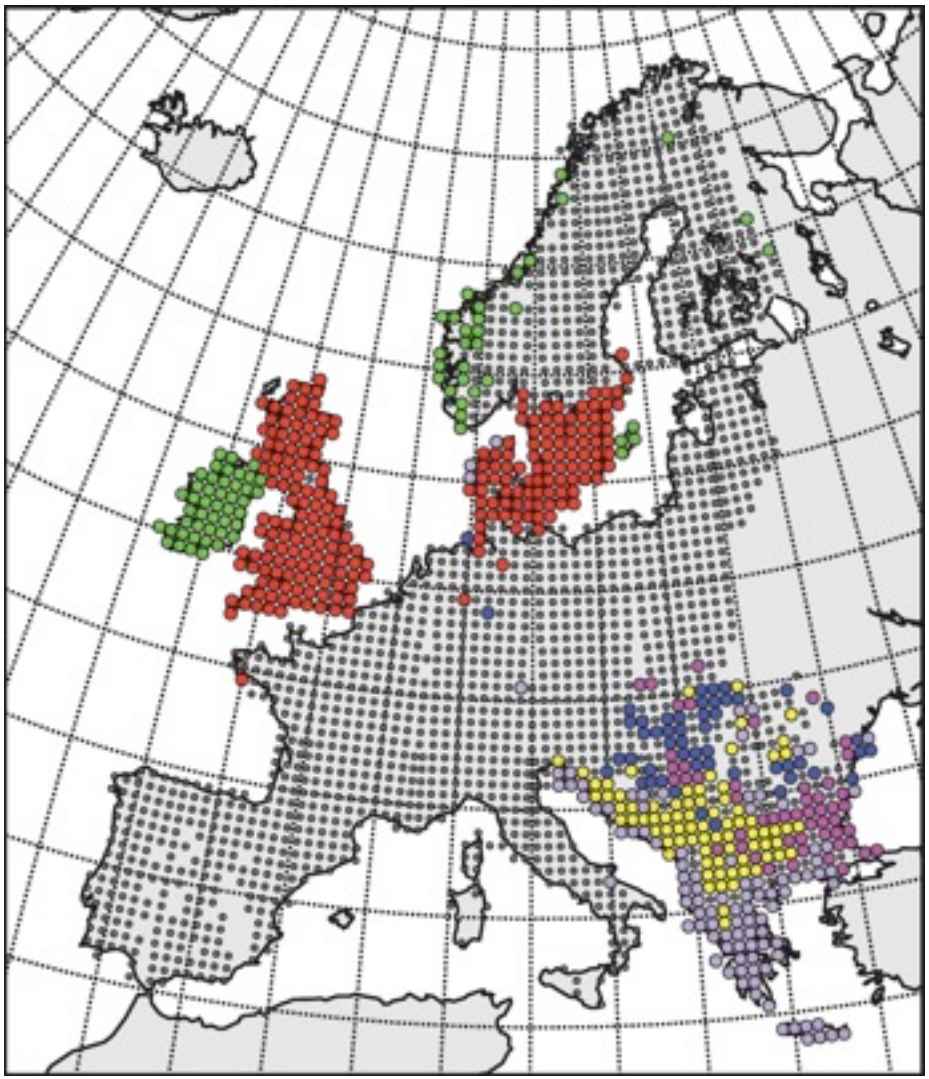
Clusters of Mammals



Clusters of Mammals



Clusters of Mammals



More Dimensions

More Dimensions

UCI's Breast-cancer and Ionosphere

- 9 and 34 numeric attributes, respectively
- We cartified these separately, then clustered
- Found clusters have high class purity