0010010110101111001111110101010101000
1010100010101110111000111010110011011
0010010110101111001111110101010101000

Data Mining

Luiz Celso Gomes Jr - André Santanchè
MC536 2013/2

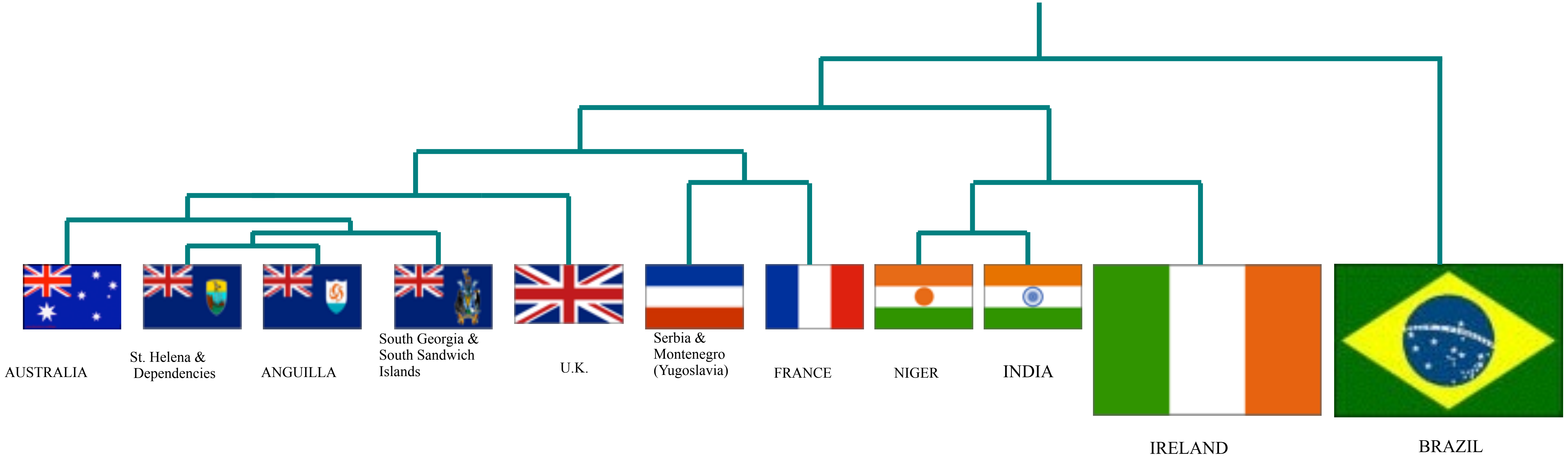
Outline

- Definition
- Architecture
- DM tasks
    - Classification
    - Clustering
    - Association Rules

A Gentle Introduction to Machine Learning and Data Mining for the Database Community

Dr Eamonn Keogh

University of California - Riverside
eamonn@cs.ucr.edu


AUSTRALIA
St. Helena & Dependencies
ANGUILLA
South Georgia & South Sandwich Islands
U.K.
Serbia & Montenegro (Yugoslavia)
FRANCE
NIGER
INDIA
IRELAND
BRAZIL

Importância dos Dados

- Empresas investem pesado em SGDBs
- A maioria das grandes empresas não sobreviveria sem seus dados
- Exemplo: Mastercard

Importância da Informação/Conhecimento

- Permite o gerenciamento coerente de dados operacionais em SGDBs
  - Cliente A já estourou seu limite?
  - Gerar boleto do Cliente B

Importância da Informação/Conhecimento

- Permite derivação de informações estratégicas em Sistemas de Suporte a Tomada de Decisão
  - Qual o perfil dos clientes que gastam mais?
  - Quais períodos do ano apresentam baixa no faturamento?
  - Há relação entre escolaridade e inadimplência?
  - Categorizar clientes entre compradores de ofertas, compradores pragmáticos, e compradores de itens de luxo
  - Outros?
- Processo de Data Mining (Mineração de Dados)

Database vs. Data Mining

- Query
  - Well defined
  - SQL
- Output
  - Subset of database
- Field
  - Mature

- Query
  - Poorly defined
  - No precise query language
- Output
  - Not a subset of database
- Field
  - Maturing, becoming very important

More Query Examples

Database

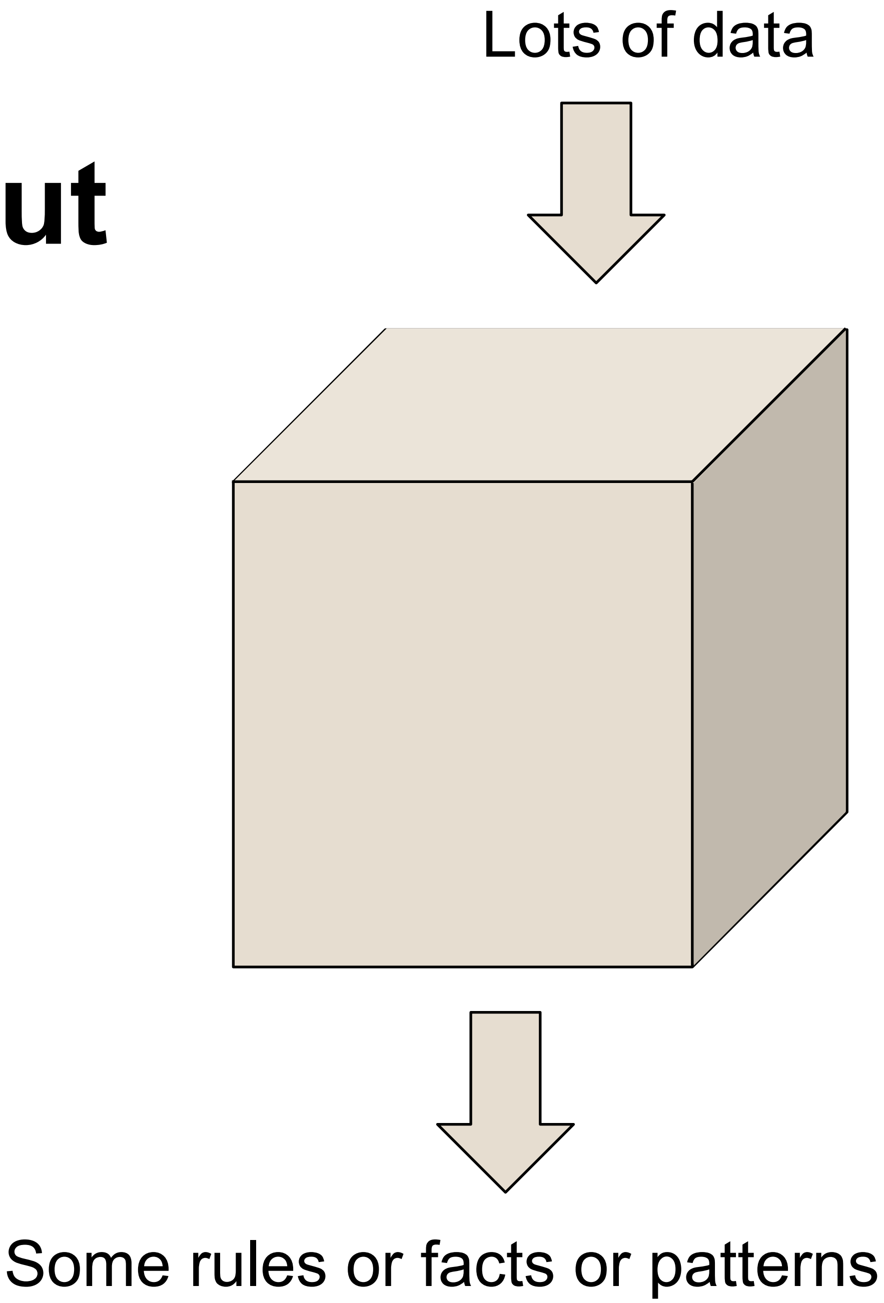
- Find all customers that live in Boa Vista
- Find all customers that use Mastercard
- Find all customers that missed one payment

Data mining

- Find all customers that are likely to miss one payment (**Classification**)
- Group all customers with simpler buying habits (**Clustering**)
- List all items that are frequently purchased with bicycles (**Association rules**)
- Find any “unusual” customers (**Outlier detection, anomaly discovery**)

All these examples show…

- Lots of raw data **in**
- *Some data mining*
- Facts, rules, patterns **out**



Definição

Mineração de Dados = Mineração de Conhecimento a partir de Dados.

Han 2006

Mineração de Dados se refere à análise de **grande quantidades de dados** com o objetivo de descobrir padrões que sejam **válidos**, **novos** (anteriormente desconhecidos), potencialmente **úteis**, e eventualmente **compreensíveis**.

Ramakrishnan and Gehrke 2003

Definição

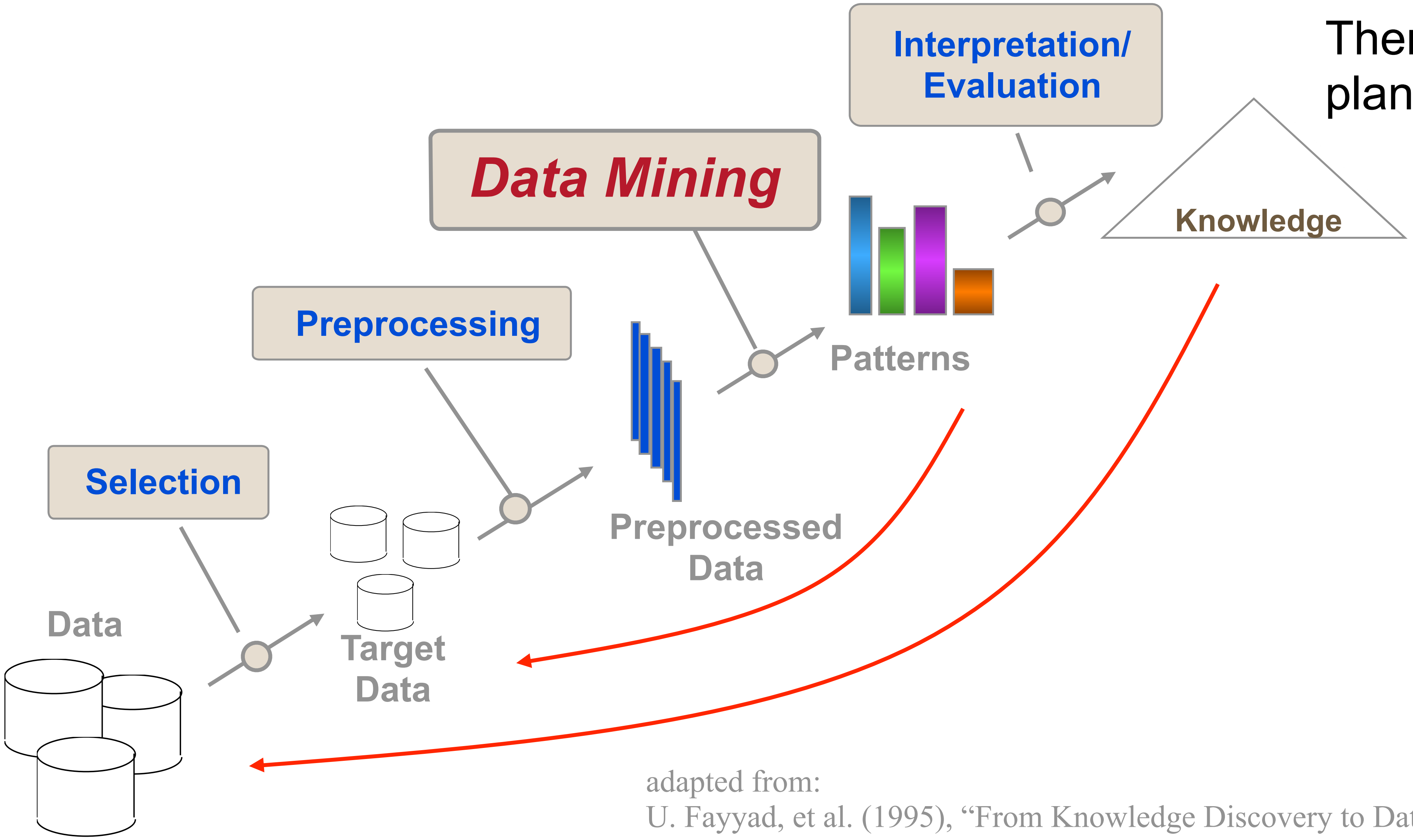
Mineração de Dados se refere à análise de grande quantidades de dados com o objetivo de descobrir padrões que sejam **válidos**, **novos**, **úteis**, e **compreensíveis**.

- Validade: os padrões são em geral válidos (generalizáveis)
- Novidade: o padrão não era conhecido antes
- Utilidade: é possível tomar certa atitude com base nos padrões
- Compreensibilidade: nós podemos interpretar e compreender os padrões

DM Alternative Names

- Knowledge discovery (mining) in databases (KDD)
- Knowledge extraction
- Data/pattern analysis
- Data archeology
- Data dredging
- Information harvesting
- Business intelligence ...

Knowledge Discovery in Databases: Process



Selection

Preprocessing

Data Mining

Interpretation/ Evaluation

Data

Target Data

Preprocessed Data

Patterns

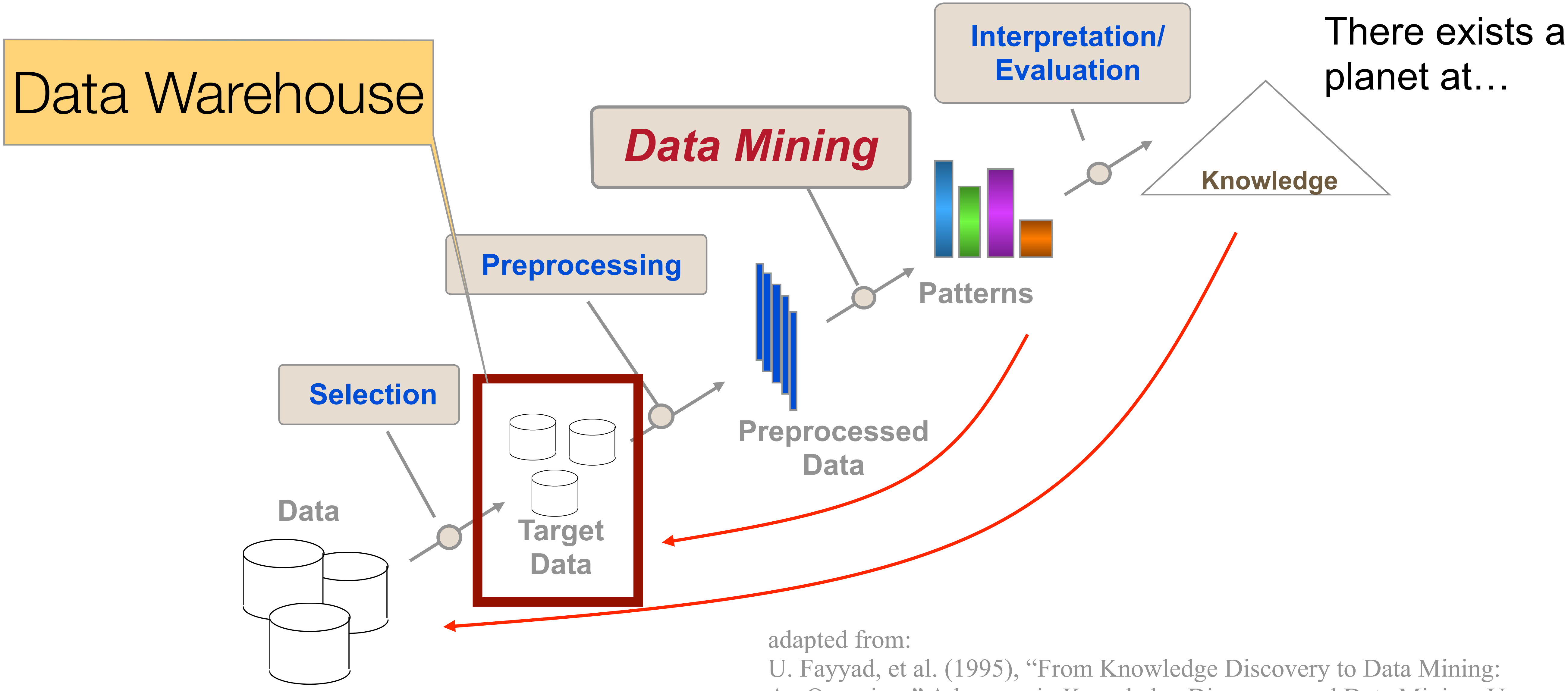
Knowledge

There exists a planet at…

adapted from:
U. Fayyad, et al. (1995), “From Knowledge Discovery to Data Mining: An Overview,” Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

12.2 3434.00232
11.2 3454.64555
23.6 4324.53435

Knowledge Discovery in Databases: Process

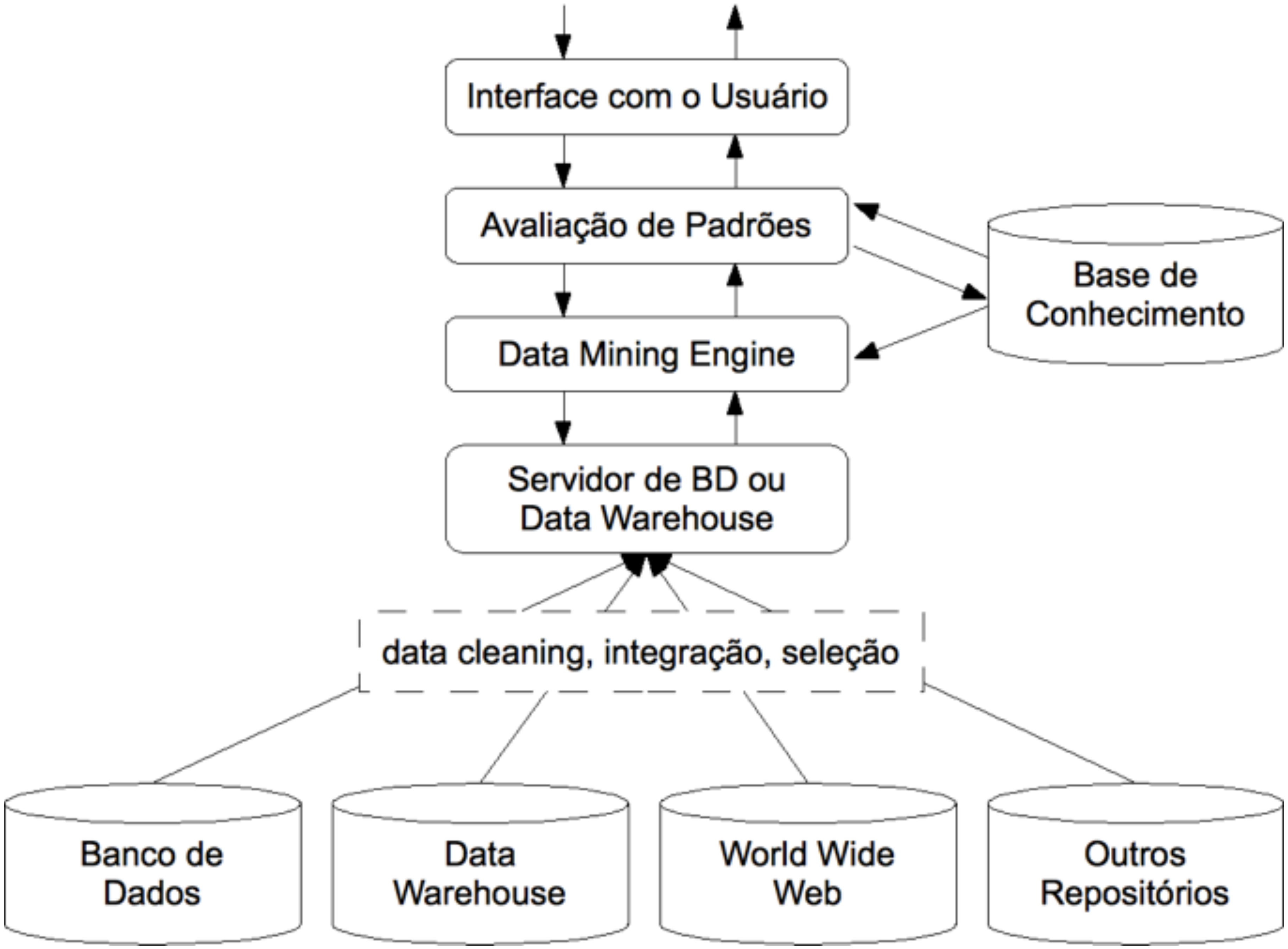


There exists a planet at…

adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

12.2 3434.00232
11.2 3454.64555
23.6 4324.53435

Arquitetura Típica



Han 2006

Multidisciplinaridade

Aprendizado de Máquina

Redes Sociais

Geoprocessamento

Computação de Alto Desempenho

Visualização de Informação

Bancos de Dados

Recuperação de Informação

Redes Neurais

Grafos

Estatística

Processamento de Linguagem Natural

Algoritmos Genéticos

Extração de Informação

Descoberta de Conhecimento em DM

- Regras de associação: quem compra leite tende a comprar também pão
- Hierarquias de classificação: classificar clientes em “compradores pragmáticos” ou “compradores compulsivos”
- Padrões sequenciais: quem compra câmera fotográfica tende a comprar acessórios em dois meses

Elmasri and Navathe 2005/ Han 2006

Descoberta de Conhecimento em DM

- Padrões em séries temporais: vende-se mais calçados esportivos na primavera que em qualquer outra estação
- Clusterização: agrupar clientes de acordo com os horários em que costumam comprar
- Outliers: identificar compras anormais, suspeitas de fraude

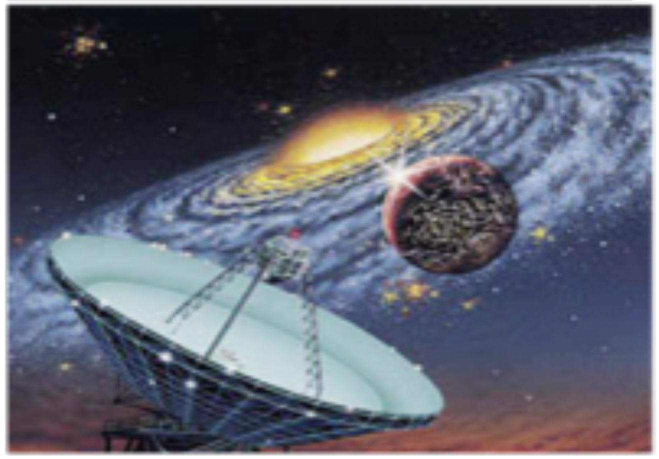
Elmasri and Navathe 2005/ Han 2006

Why is Data Mining Hard?

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis
- Over fitting
- Privacy issues

Scale of Data

Organization	Scale of Data
Walmart	~ 20 million transactions/day
Google	~ 8.2 billion Web pages
Yahoo	~10 GB Web data/hr
NASA satellites	~ 1.2 TB/day
NCBI GenBank	~ 22 million genetic sequences
France Telecom	29.2 TB
UK Land Registry	18.3 TB
AT&T Corp	26.2 TB



"The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don't have a clue as to what any of that data actually means"

(S. Cass, IEEE Spectrum, Jan 2004)

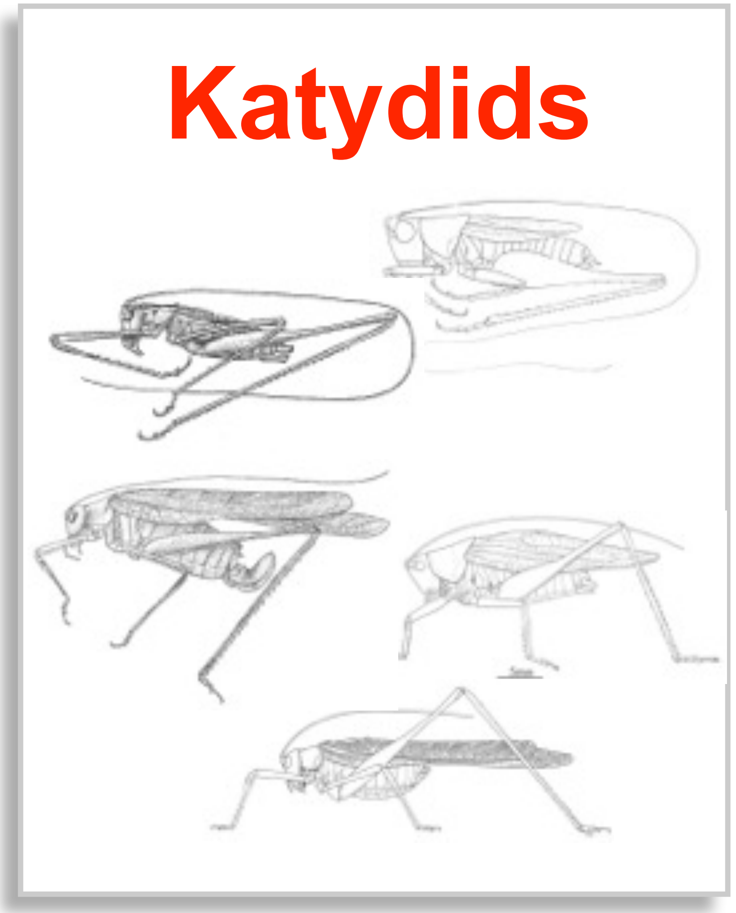
The Major Data Mining Tasks

- Classification
- Clustering
- Associations

Most of the other tasks (for example, outlier discovery or anomaly detection) make heavy use of one or more of the above.

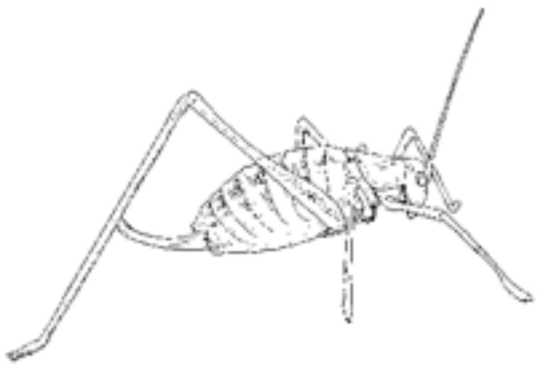
The Classification Problem
(informal definition)

Given a collection of annotated data. In this case 5 instances **Katydids** of and five of **Grasshoppers**, decide what type of insect the unlabeled example is.



Katydids

Grasshoppers

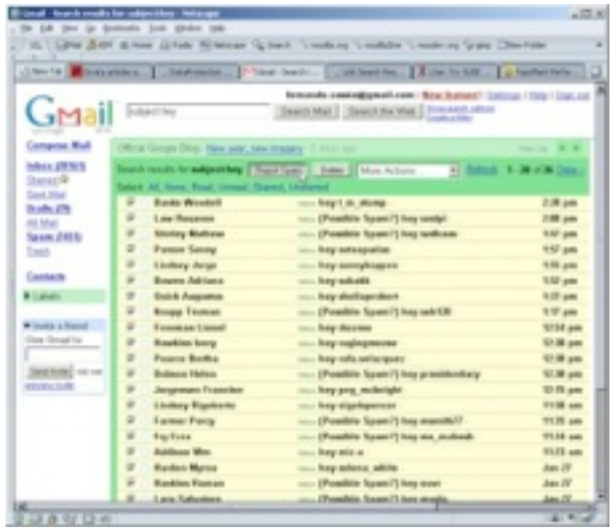


Katydid or **Grasshopper**?

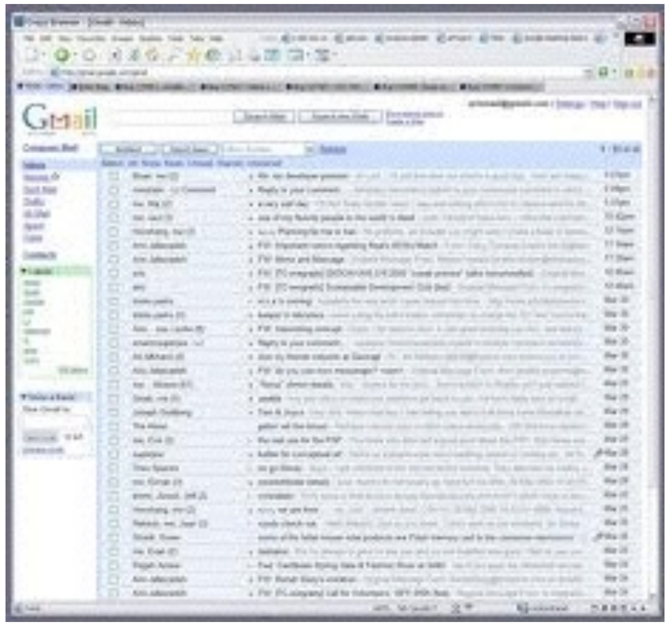
The Classification Problem

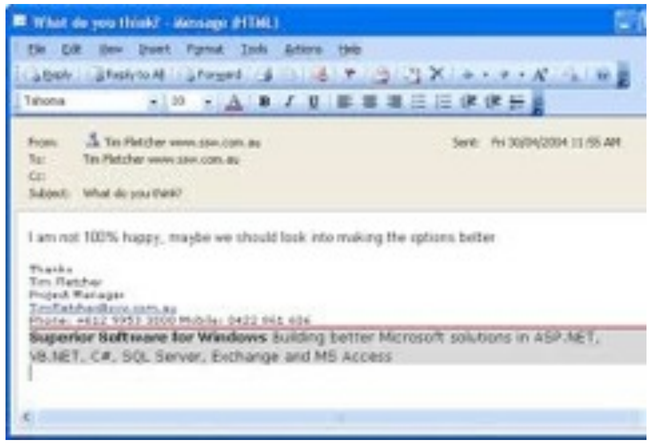
Given a collection of annotated data…

spam



email




What do you think? - Message (HTML)
File Edit View Insert Format Tools Actions Help
Reply Reply to All Forward
Tahoma 10

From: Tim Fletcher www.ssw.com.au Sent: Fri 30/04/2004 11:55 AM
To: Tim Fletcher www.ssw.com.au
Cc:
Subject: What do you think?

I am not 100% happy, maybe we should look into making the options better

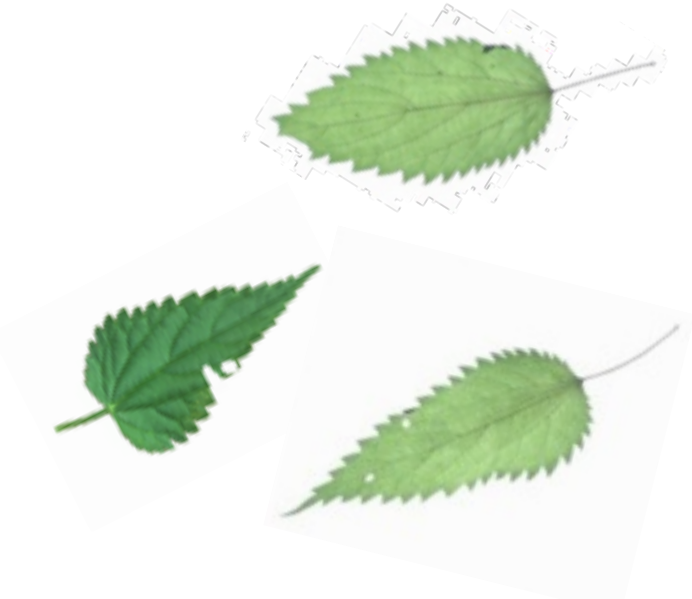
Thanks
Tim Fletcher
Project Manager
TimFletcher@ssw.com.au
Phone: +612 9953 3000 Mobile: 0422 941 436
Superior Software for Windows Building better Microsoft solutions in ASP.NET, VB.NET, C#, SQL Server, Exchange and MS Access

Spam or email?

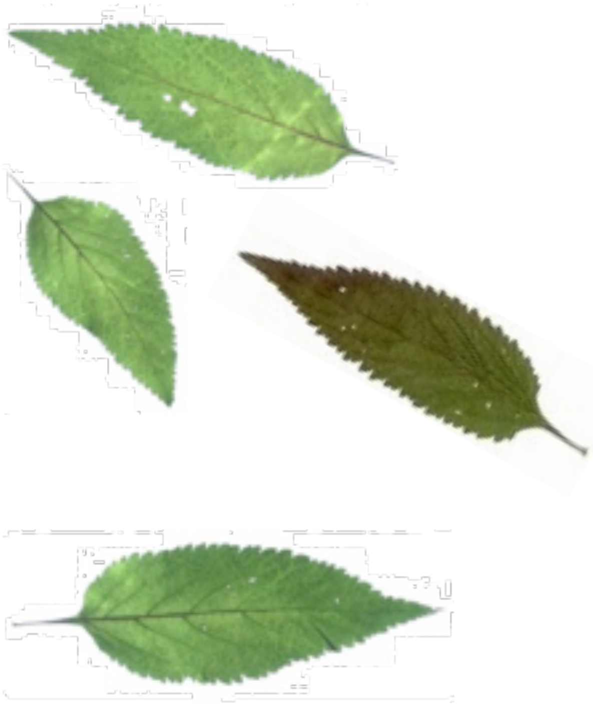
The Classification Problem

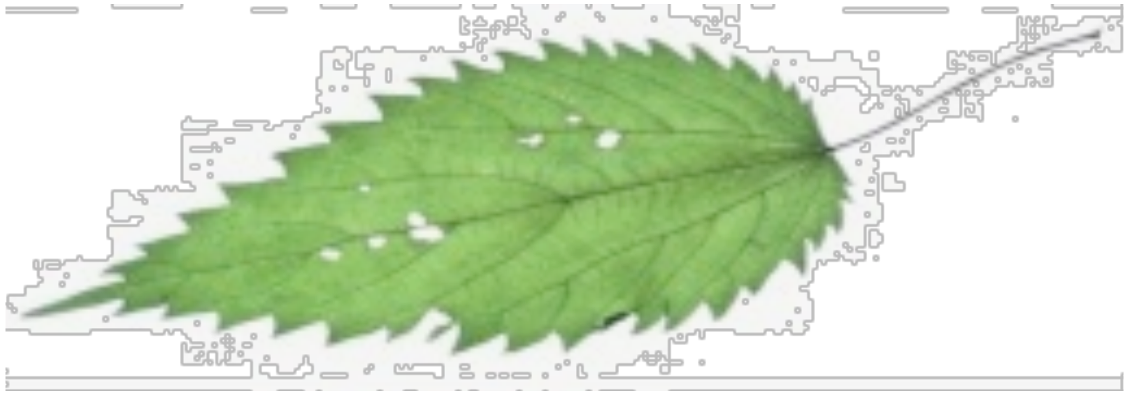
Given a collection of annotated data…

Stinging Nettle



False Nettle





Stinging Nettle or **False Nettle**?

The Classification Problem

Given a collection of annotated data…

Tsotras

Greek or Irish?

Greek

Gunopulos
Papadopoulos
Kollios
Dardanos

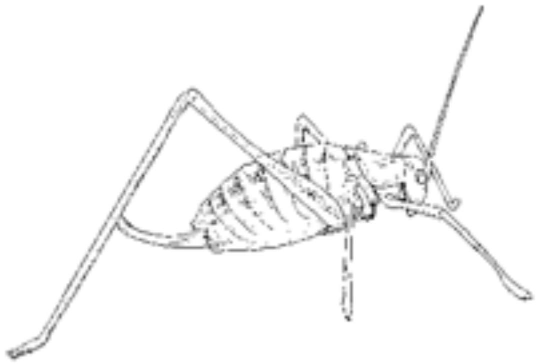
Irish

Keogh
Gough
Greenhaugh
Hadleigh

The Classification Problem

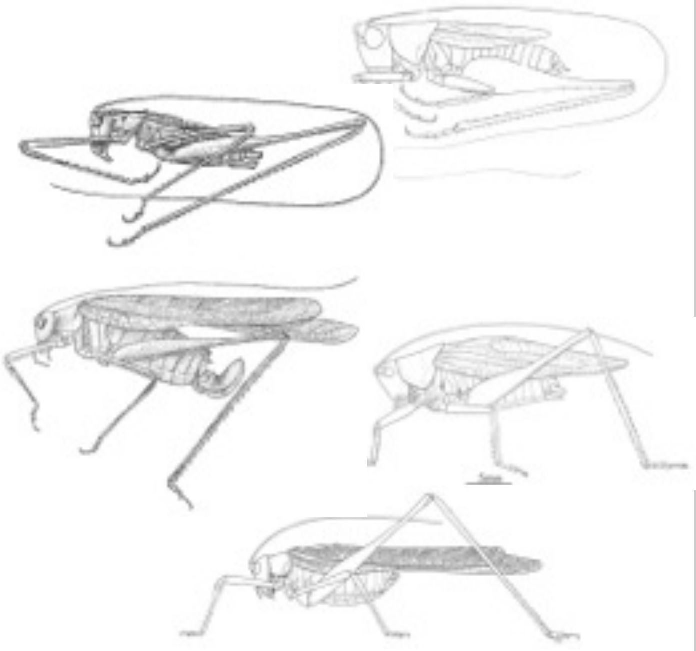
(informal definition)

Given a collection of annotated data. In this case 5 instances **Katydids** of and five of **Grasshoppers**, decide what type of insect the unlabeled example is.

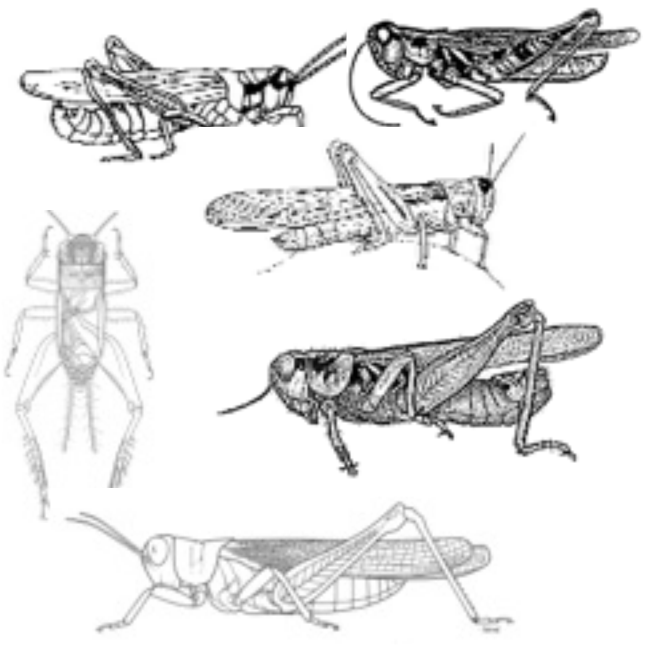


Katydid or **Grasshopper**?

Katydids



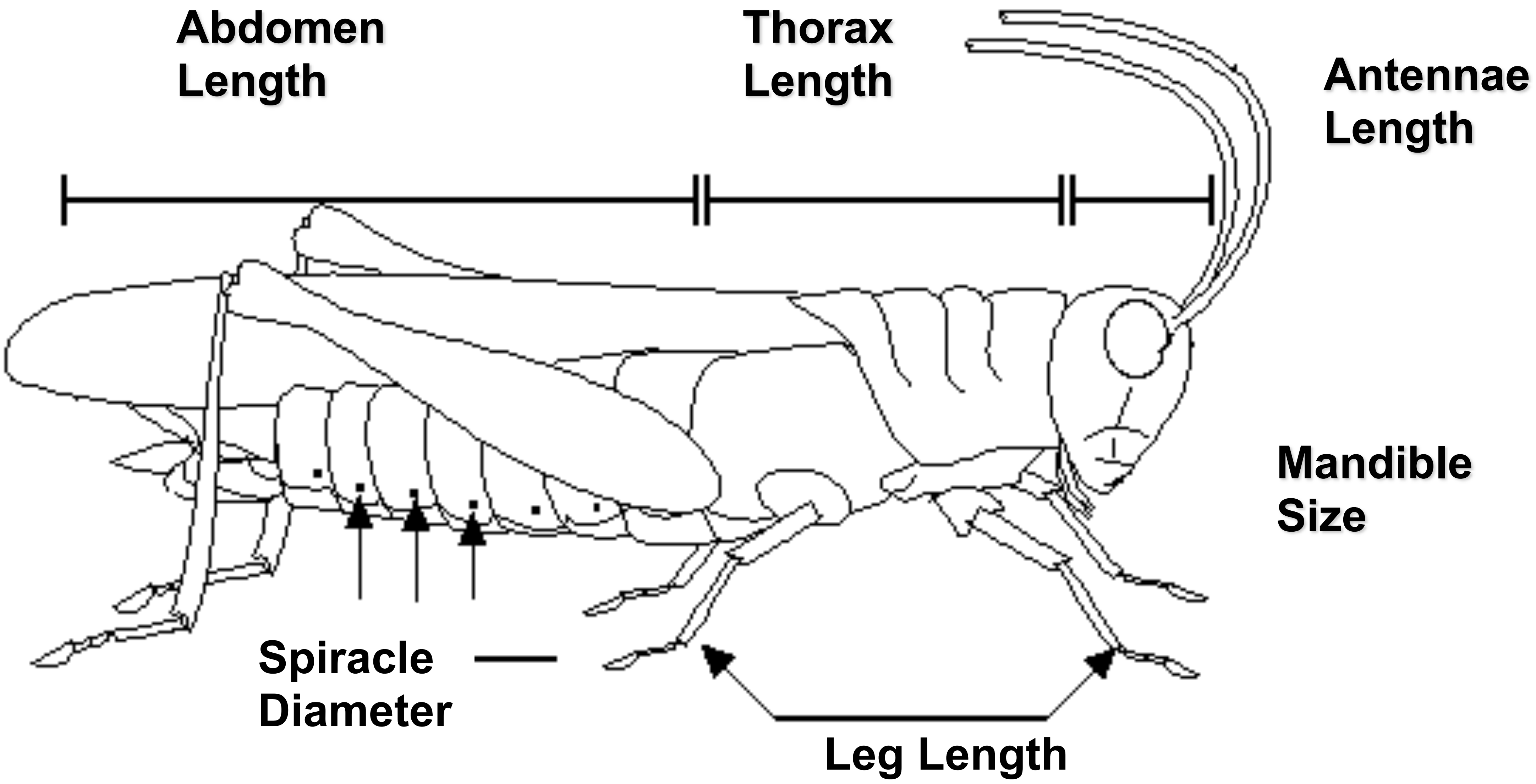
Grasshoppers



For any domain of interest, we can measure *features*

Color {Green, Brown, Gray, Other}

Has Wings?



Abdomen Length

Thorax Length

Antennae Length

Mandible Size

Spiracle Diameter

Leg Length

We can store features in a database.

The classification problem can now be expressed as:

- Given a training database (My_Collection), predict the class label of a previously unseen instance

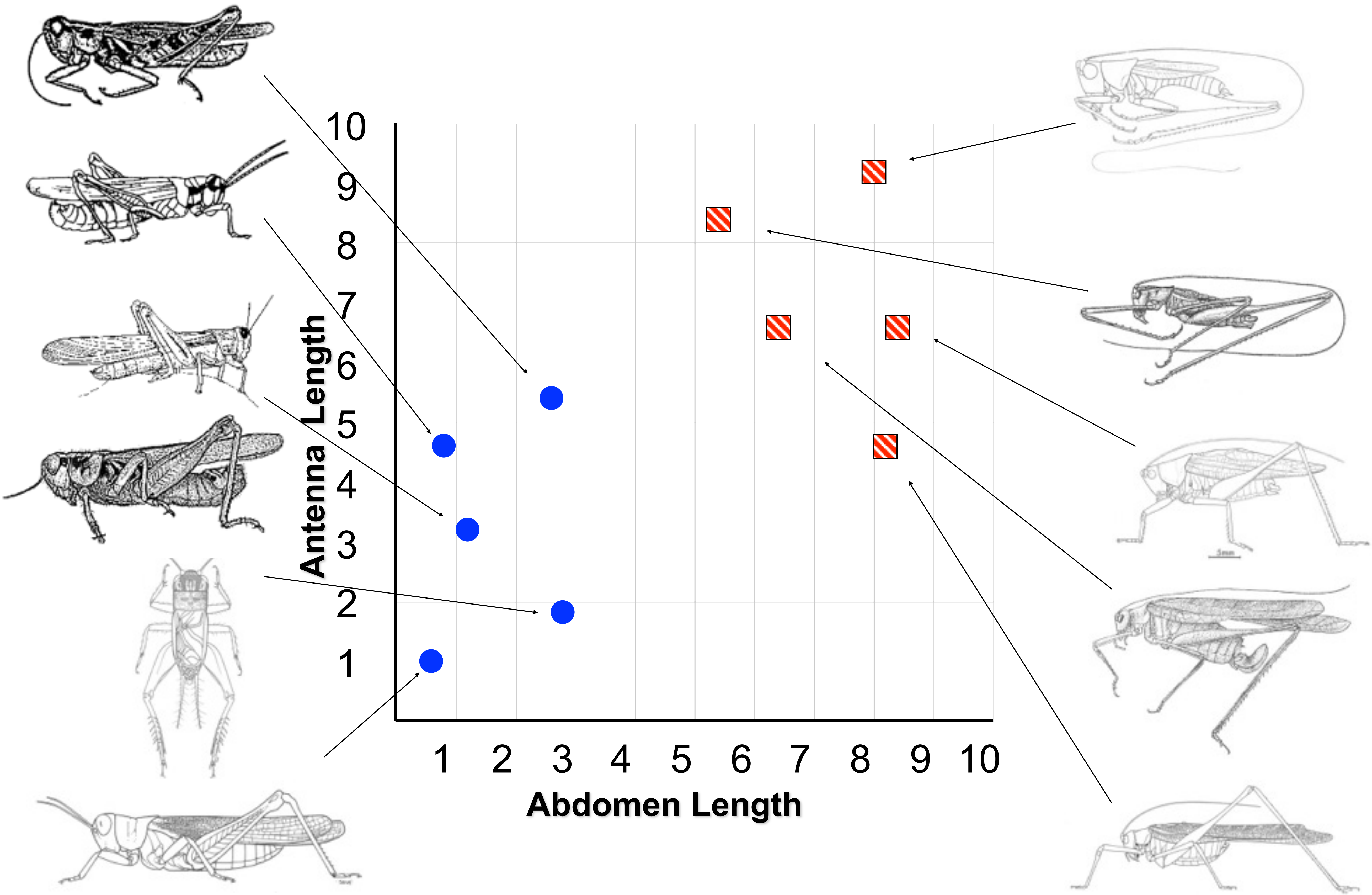
My_Collection

Insect ID	Abdomen Length	Antennae Length	Insect Class
1	2.7	5.5	Grasshopper
2	8.0	9.1	Katydid
3	0.9	4.7	Grasshopper
4	1.1	3.1	Grasshopper
5	5.4	8.5	Katydid
6	2.9	1.9	Grasshopper
7	6.1	6.6	Katydid
8	0.5	1.0	Grasshopper
9	8.3	6.6	Katydid
10	8.1	4.7	Katydids

previously unseen instance =

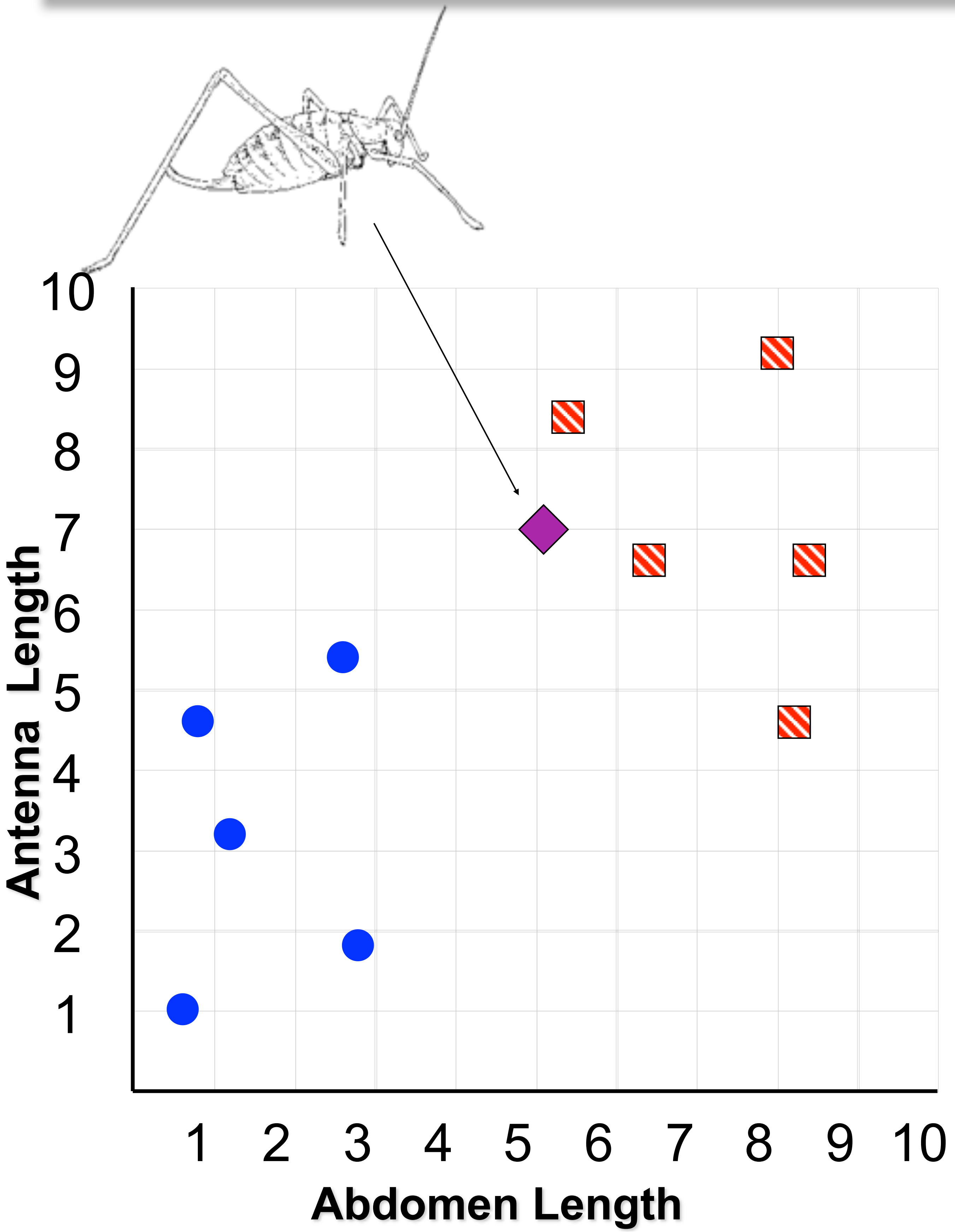
11	5.1	7.0	???????

Grasshoppers **Katydids**



previously unseen instance =

11	**5.1**	**7.0**	???????



- We can “project” the previously unseen instance into the same space as the database.

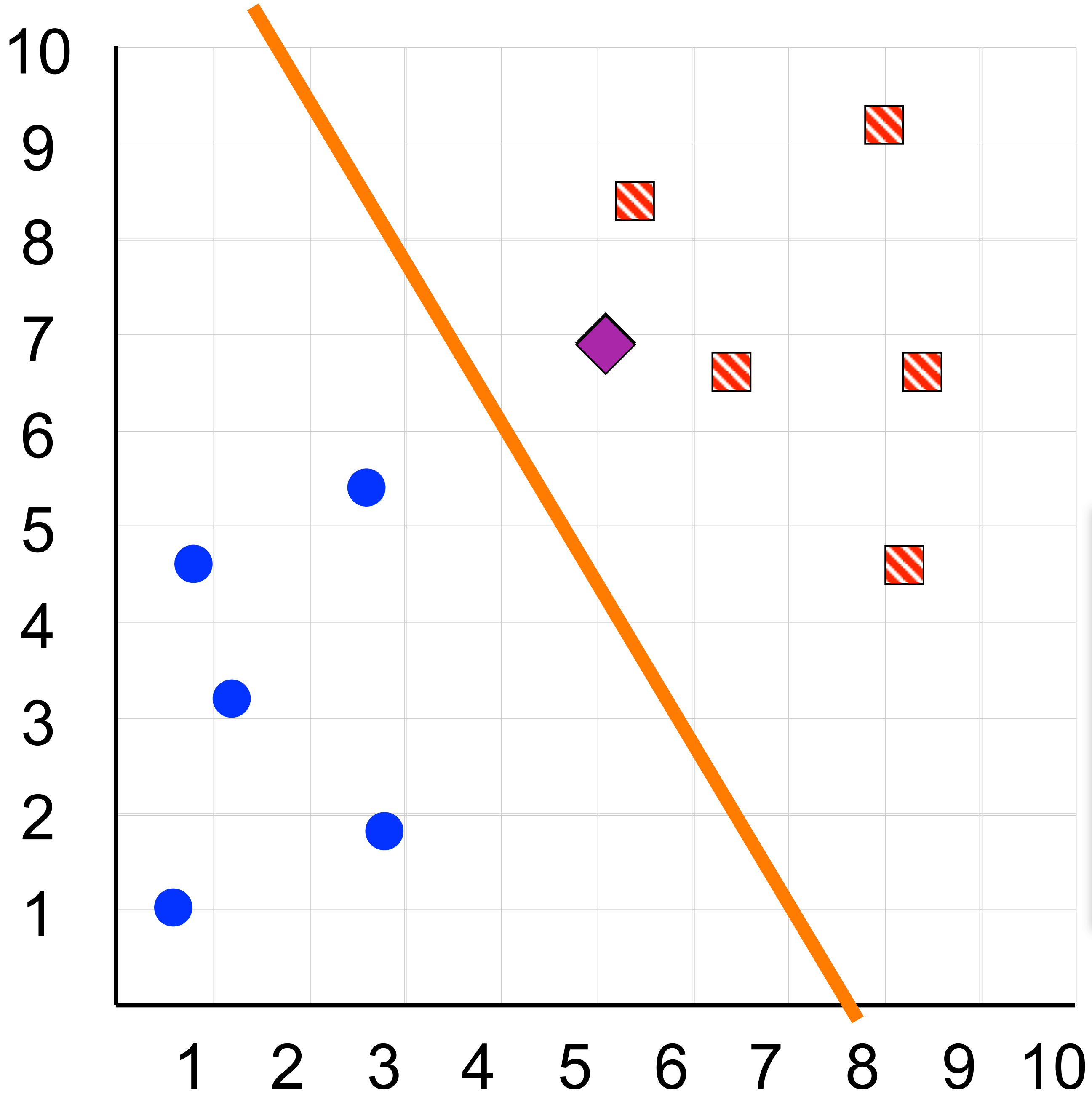
- We have now abstracted away the details of our particular problem. It will be much easier to talk about points in space.

▨ Katydids

● Grasshoppers

Simple Linear Classifier

R.A. Fisher
1890-1962



If previously unseen instance above the line
then
class is **Katydid**
else
class is **Grasshopper**

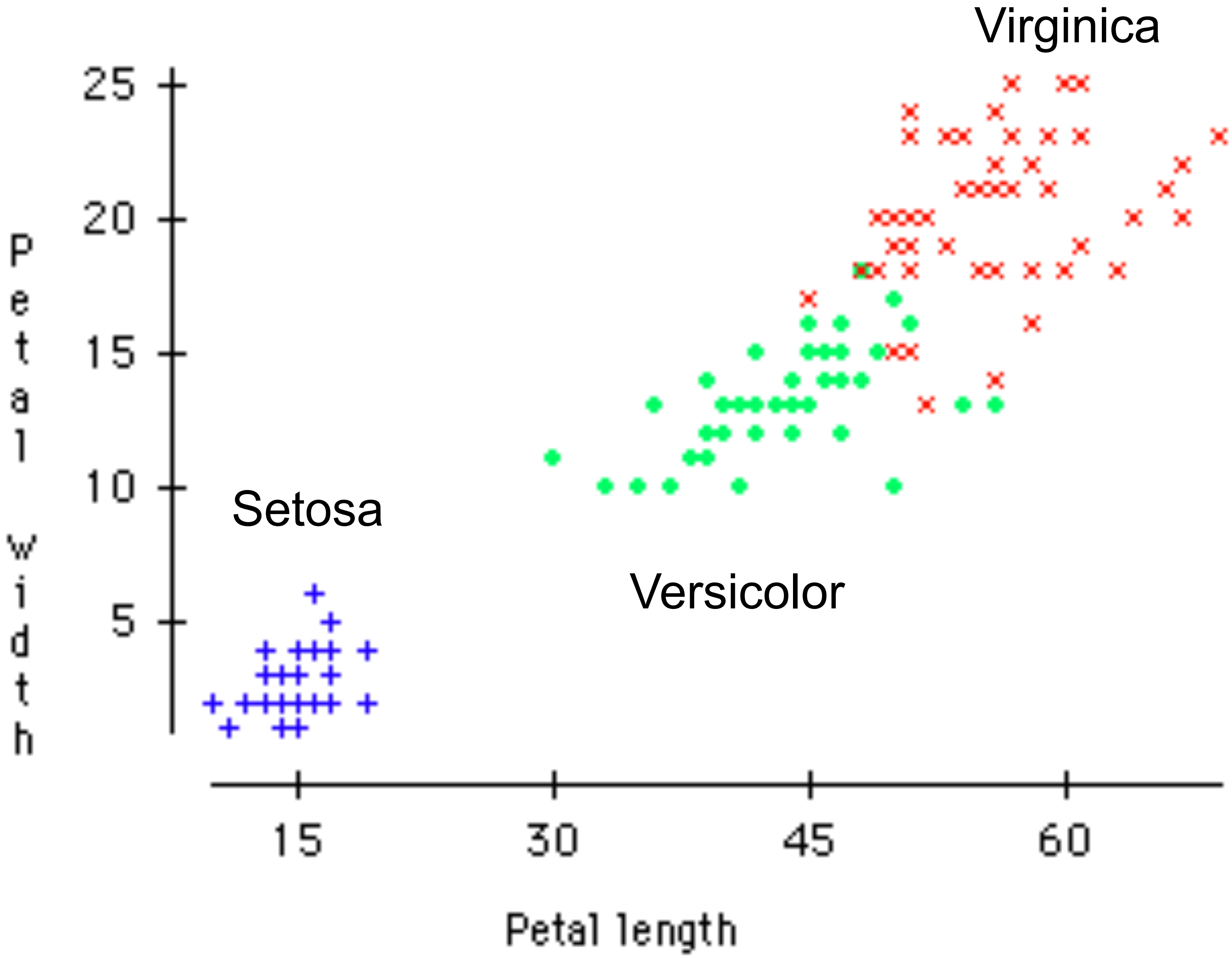
▨ **Katydids**
● **Grasshoppers**

A Famous Problem

R. A. Fisher's Iris Dataset.

- 3 classes
- 50 of each class

The task is to classify Iris plants into one of 3 varieties using the Petal Length and Petal Width.

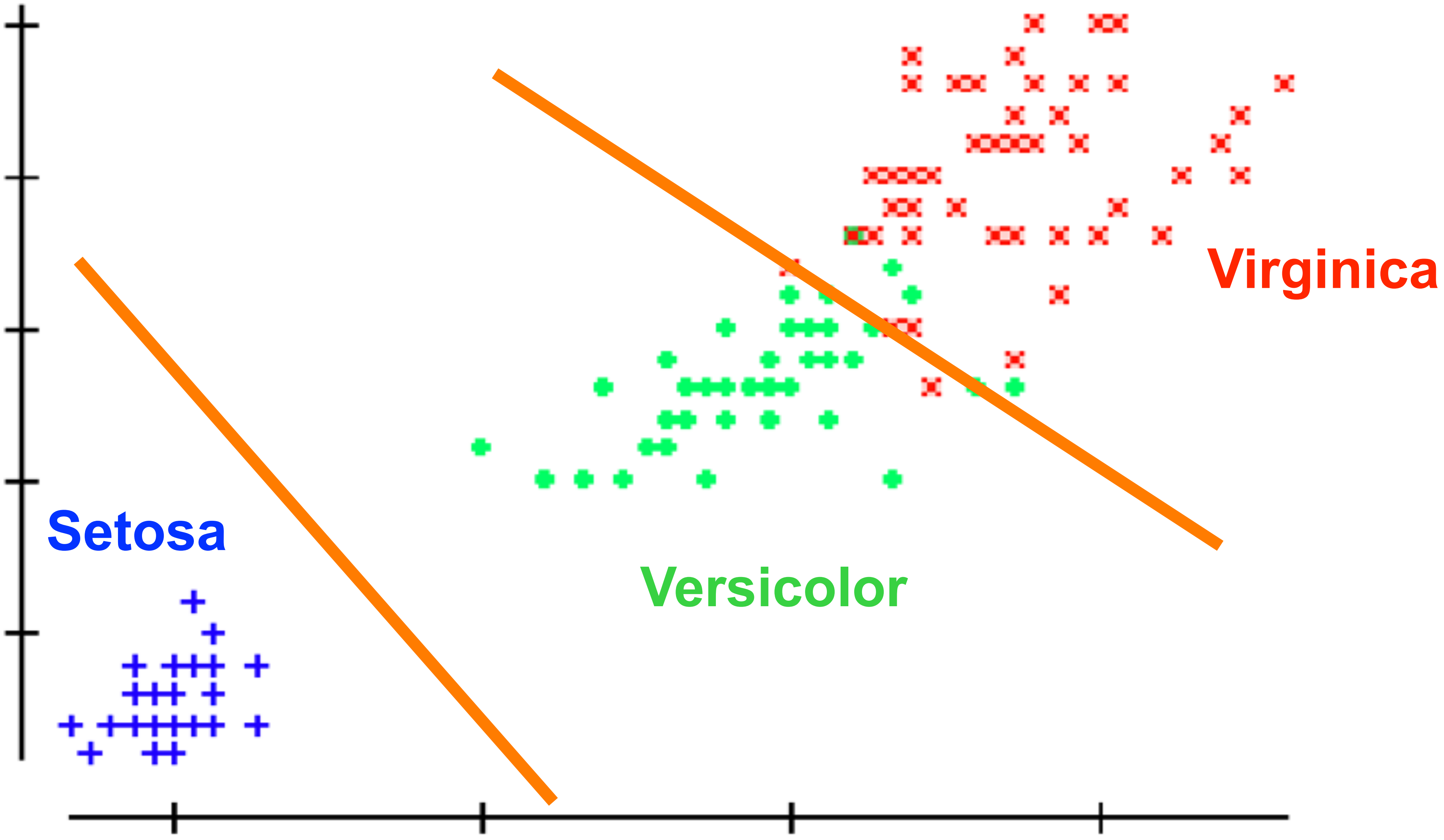


Iris Setosa

Iris Versicolor

Iris Virginica

We can generalize the piecewise linear classifier to N classes, by fitting N-1 lines. In this case we first learned the line to (perfectly) discriminate between **Setosa** and **Virginica**/ **Versicolor**, then we learned to approximately discriminate between **Virginica** and **Versicolor**.

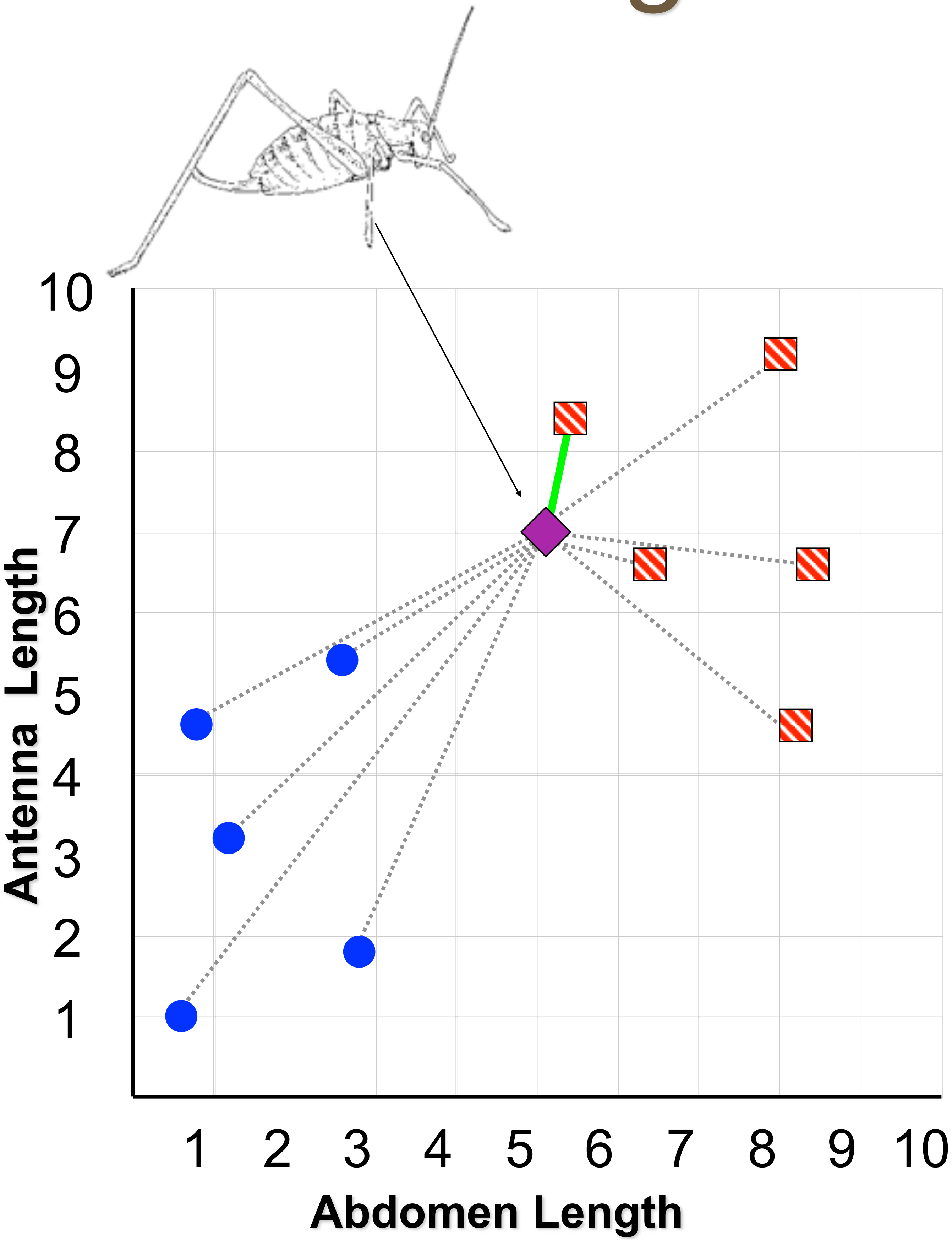


If petal width > 3.272 – (0.325 * petal length) **then** class = **Virginica**
Elseif petal width…

We have now seen one classification algorithm, and we are about to see more. How should we compare them?

- Predictive accuracy
- Speed and scalability
  - time to construct the model
  - time to use the model
  - efficiency in disk-resident databases
- Robustness
  - handling noise, missing values and irrelevant features, streaming data
- Interpretability:
  - understanding and insight provided by the model

Nearest Neighbor Classifier



Evelyn Fix
1904-1965

Joe Hodges
1922-2000

If the **nearest** instance to the previously unseen instance **is a Katydid**

    class is **Katydid**

else

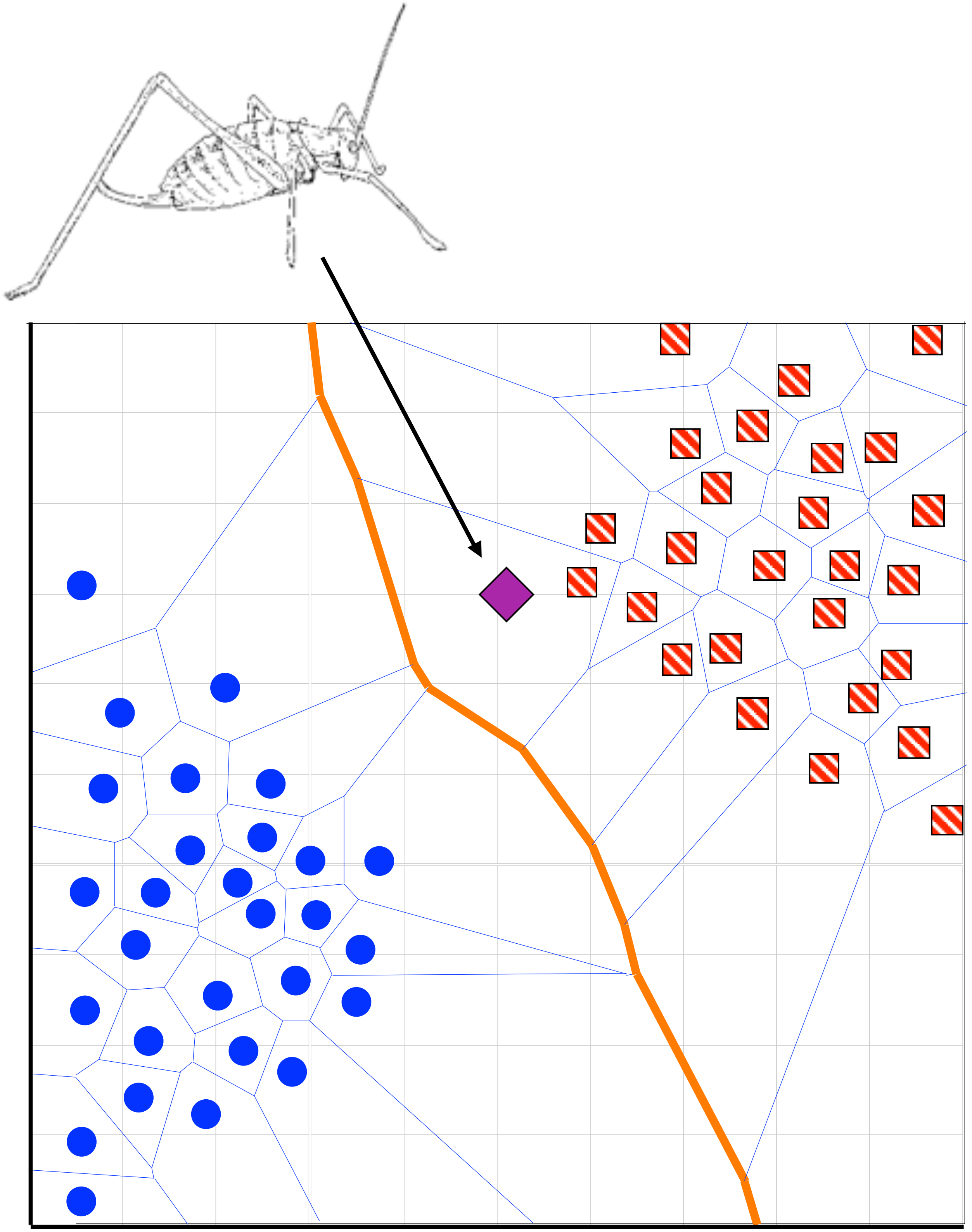
    class is **Grasshopper**

▨ **Katydids**

● **Grasshoppers**

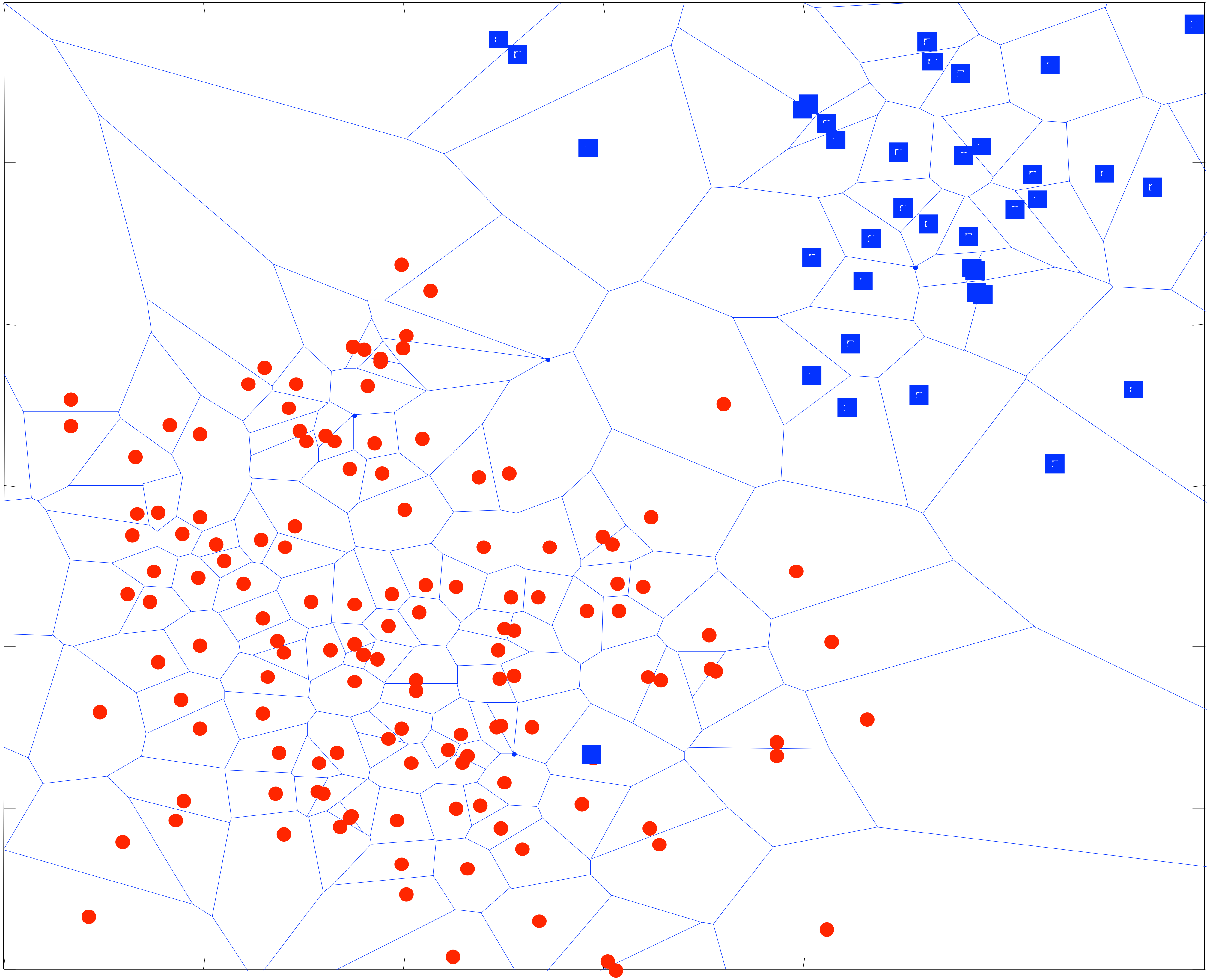
We can visualize the nearest neighbor algorithm in terms of a decision surface…

Note the we don’t actually have to construct these surfaces, they are simply the implicit boundaries that divide the space into regions “belonging” to each instance.



This division of space is called Dirichlet Tessellation (or Voronoi diagram, or Theissen regions).

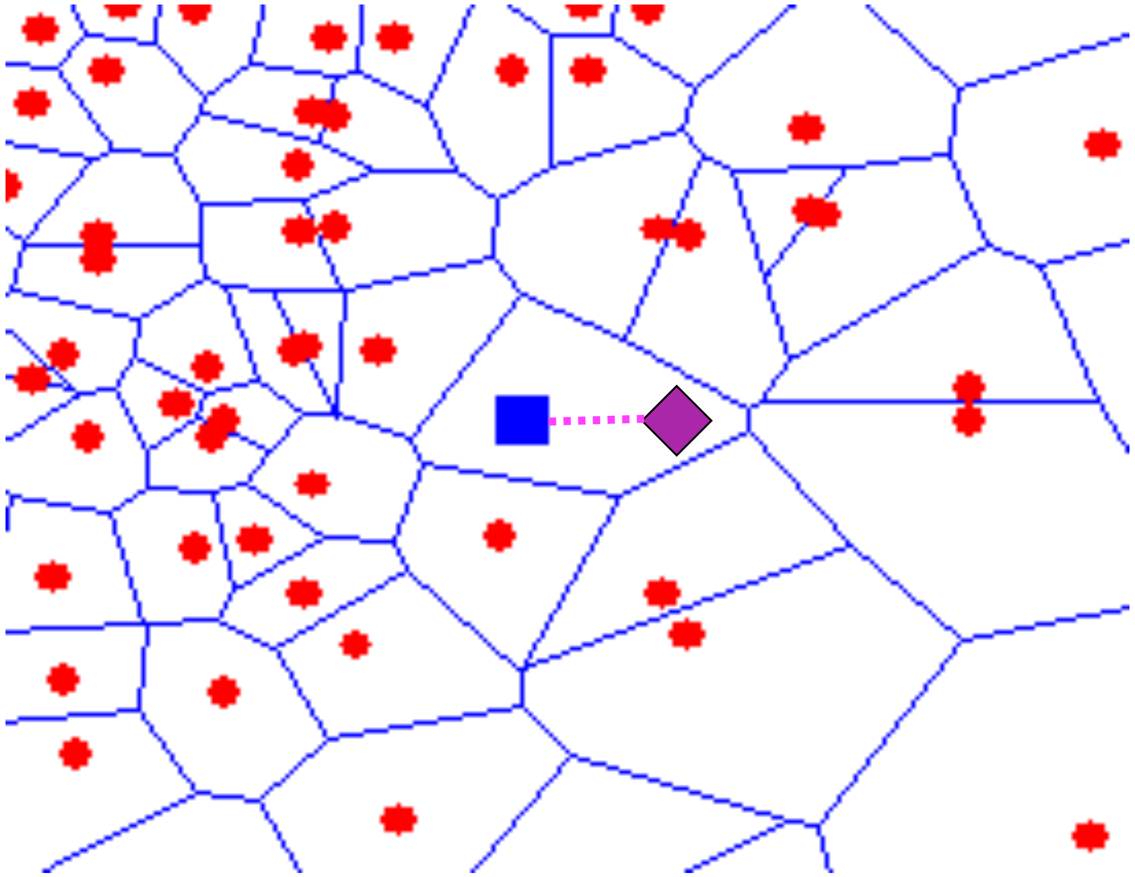
The nearest neighbor algorithm is sensitive to outliers…



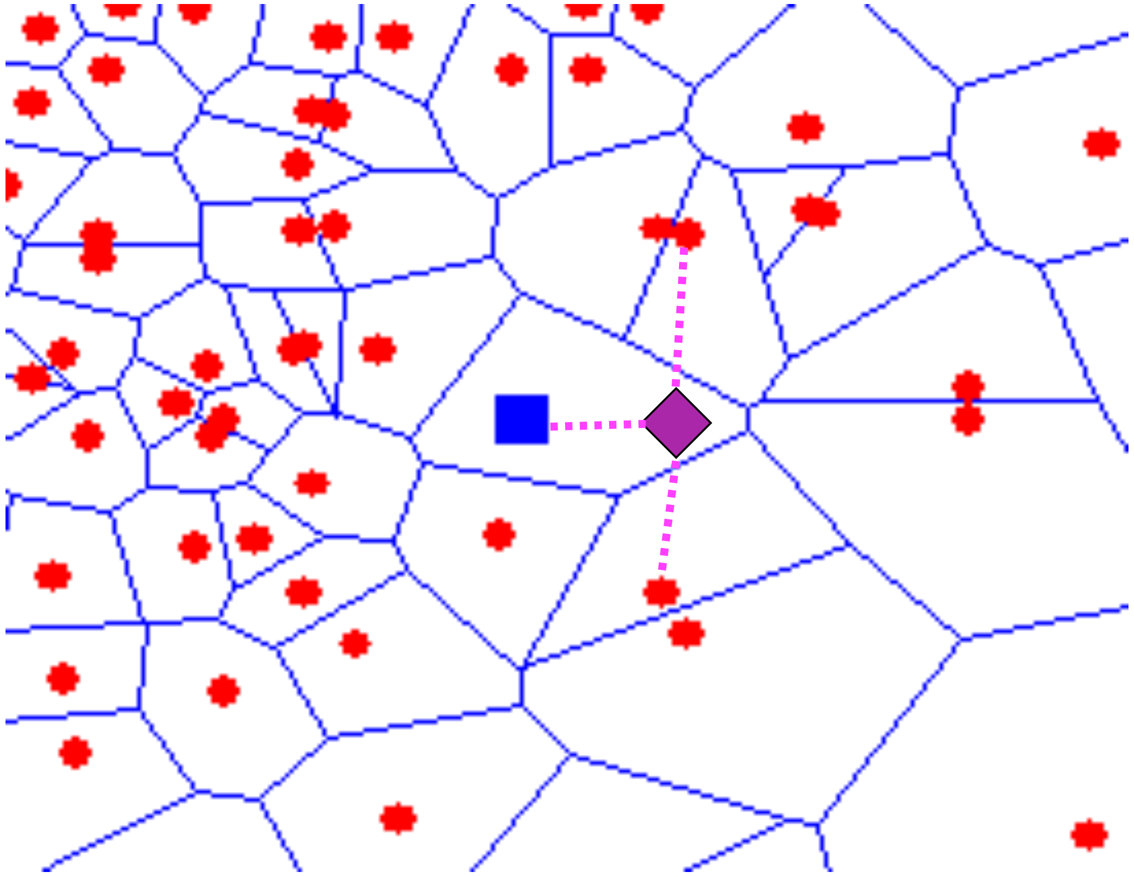
The solution is to…

We can generalize the nearest neighbor algorithm to the K- nearest neighbor (KNN) algorithm.

We measure the distance to the nearest K instances, and let them vote. K is typically chosen to be an odd number.



K = 1



K = 3

…In fact, we can use the nearest neighbor algorithm with any distance/similarity function

For example, is “*Faloutsos*” Greek or Irish? We could compare the name “*Faloutsos*” to a database of names using string edit distance…

edit_distance(*Faloutsos*, *Keogh*) = 8
edit_distance(*Faloutsos*, *Gunopulos*) = 6

Hopefully, the similarity of the name (particularly the suffix) to other Greek names would mean the nearest nearest neighbor is also a Greek name.

ID	**Name**	**Class**
1	Gunopul**os**	**Greek**
2	Papadopoul**os**	**Greek**
3	Kolli**os**	**Greek**
4	Dardan**os**	**Greek**
5	Keo**gh**	**Irish**
6	Gou**gh**	**Irish**
7	Greenhau**gh**	**Irish**
8	Hadlei**gh**	**Irish**

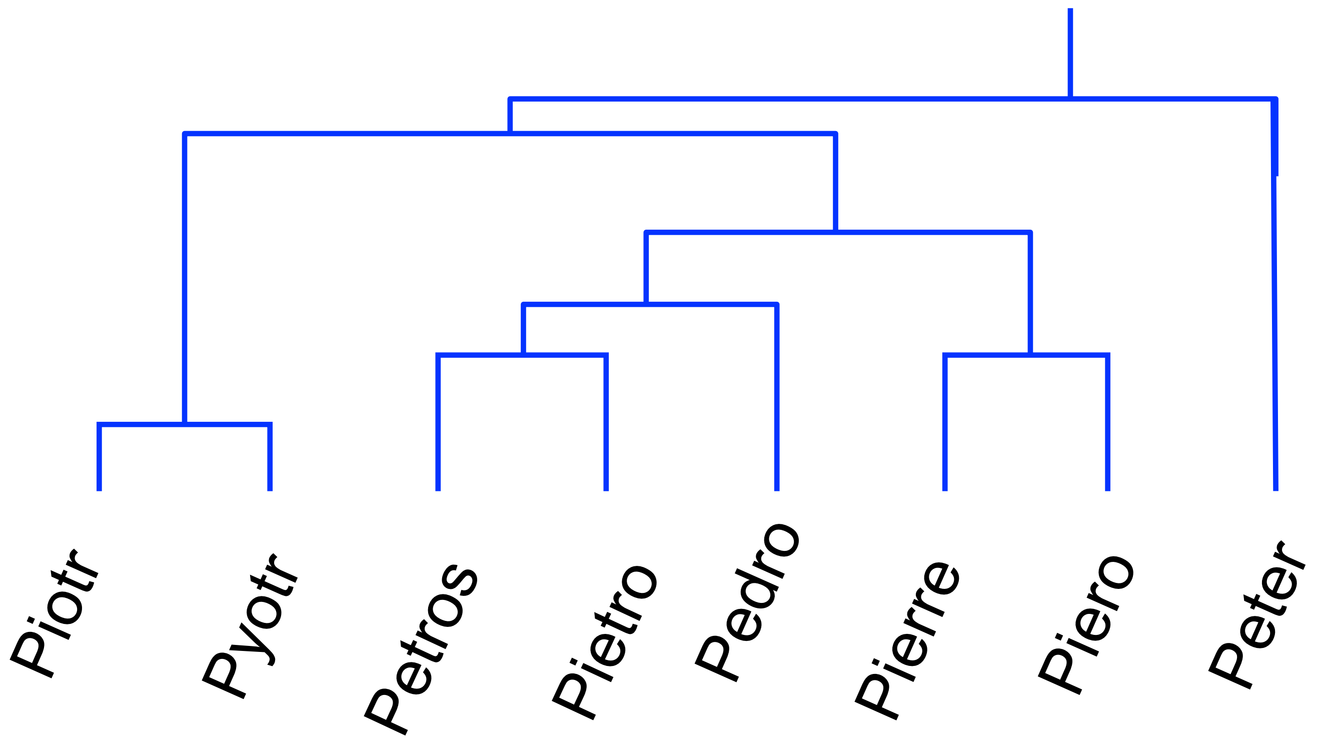
Specialized distance measures exist for DNA strings, time series, images, graphs, videos, sets, fingerprints etc…

Edit Distance Example

It is possible to transform any string *Q* into string *C*, using only *Substitution*, *Insertion* and *Deletion*.
Assume that each of these operators has a cost associated with it.

The similarity between two strings can be defined as the cost of the cheapest transformation from *Q* to *C.*

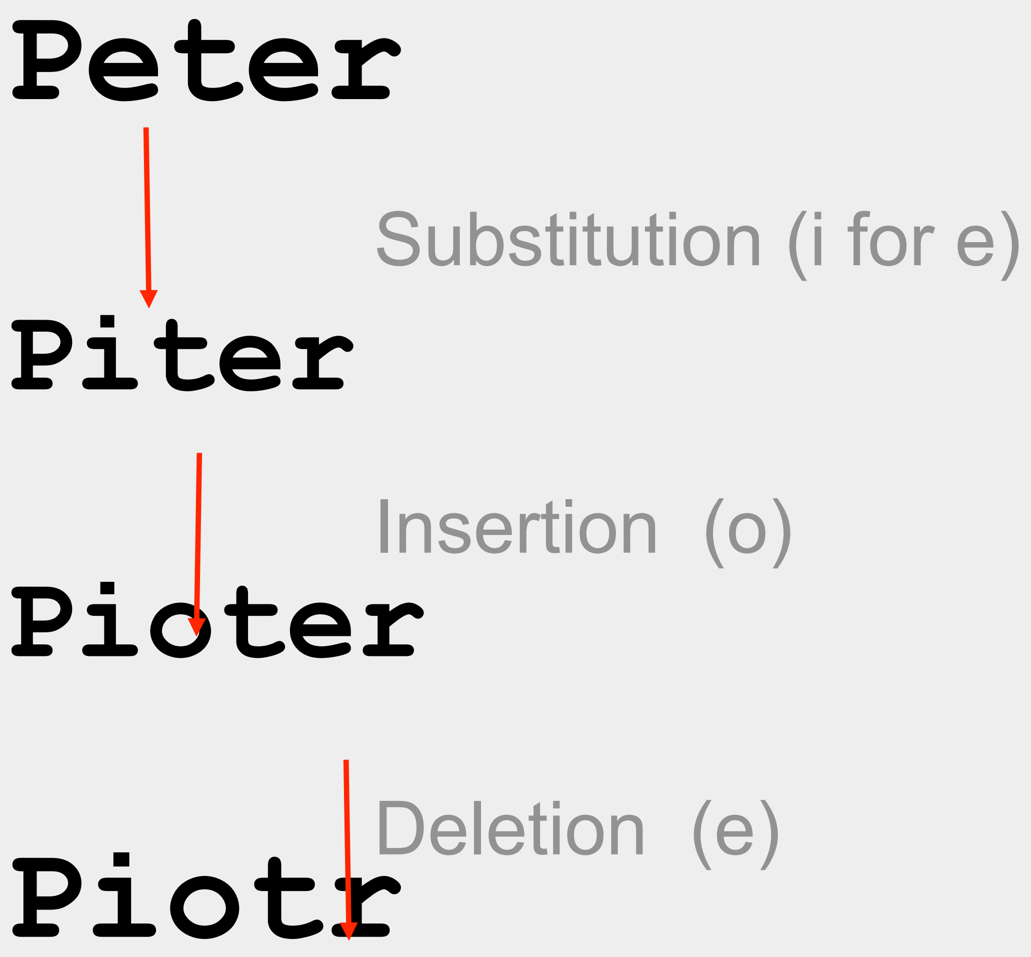
Note that for now we have ignored the issue of how we can find this cheapest transformation



How similar are the names "Peter" and "Piotr"?
Assume the following cost function

Substitution	1 Unit
Insertion	1 Unit
Deletion	1 Unit

D(**`Peter,Piotr`**) is 3



Peter

Substitution (i for e)

Piter

Insertion (o)

Pioter

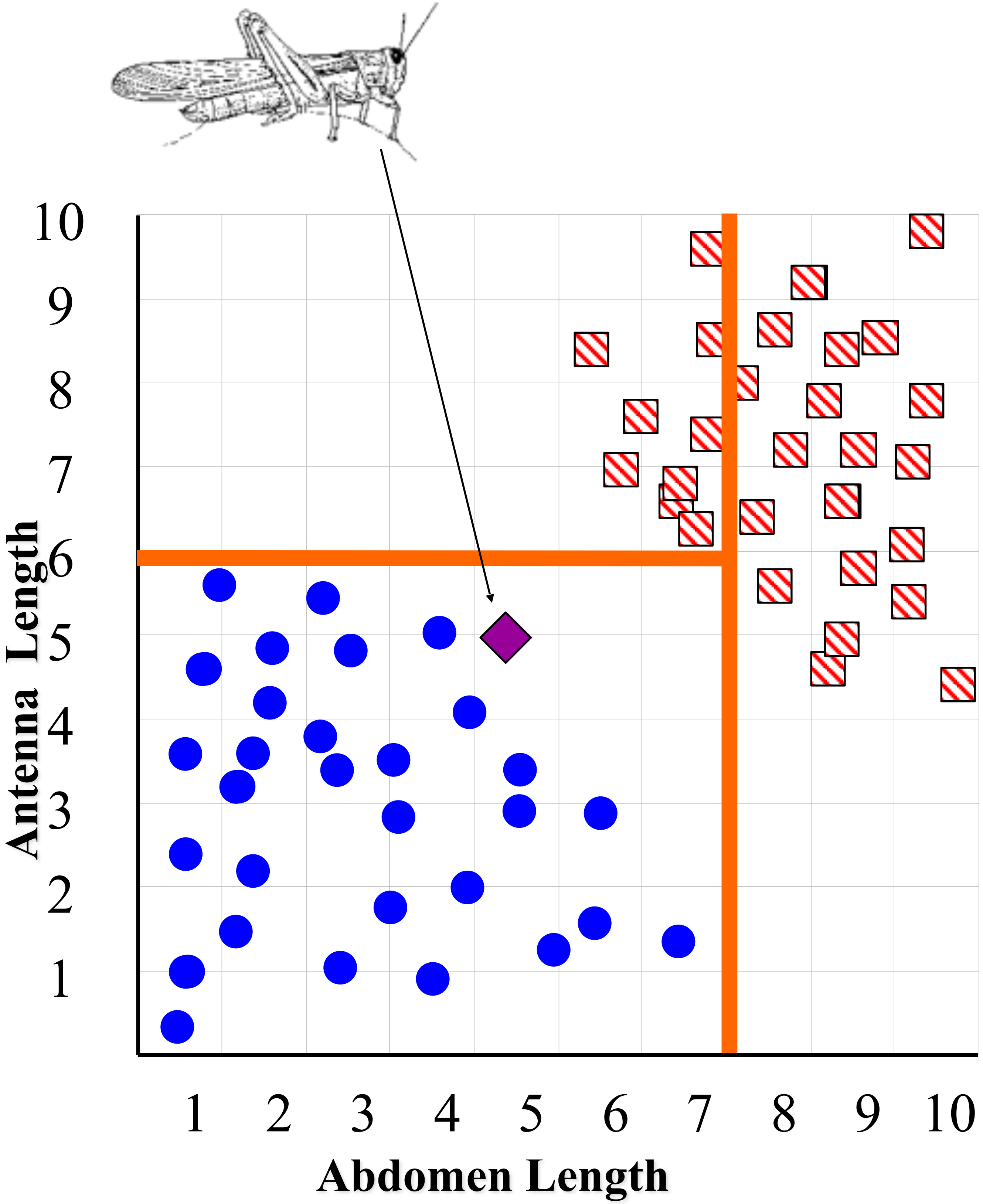
Deletion (e)

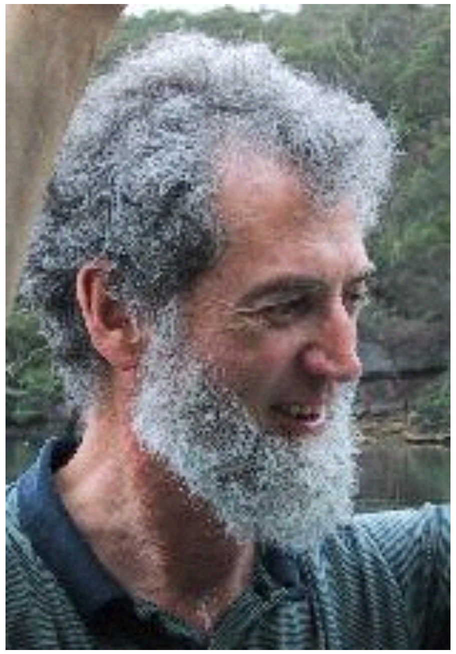
Piotr

Advantages/Disadvantages of Nearest Neighbor

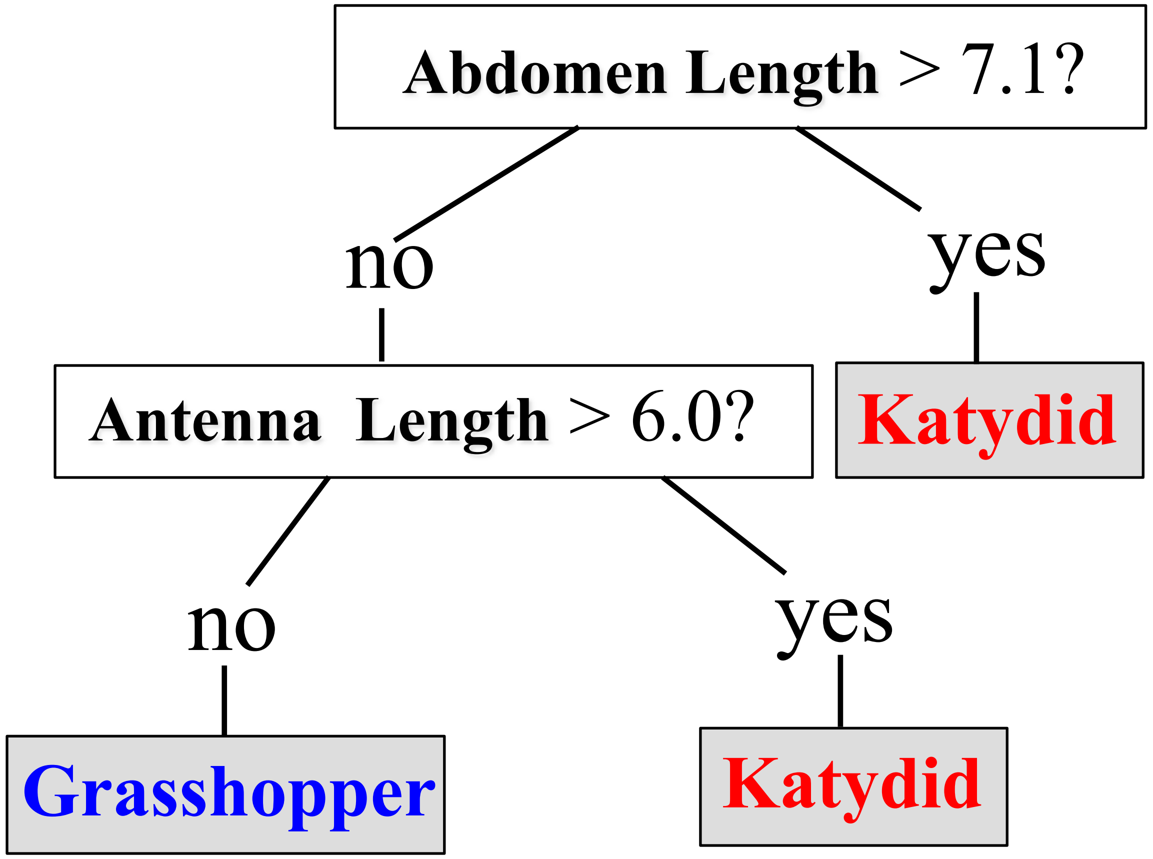
- Advantages:
  - Simple to implement
  - Handles correlated features (Arbitrary class shapes)
  - Defined for any distance measure
  - Handles streaming data trivially
- Disadvantages:
  - Very sensitive to irrelevant features.
  - Slow classification time for large datasets
  - Works best for real valued datasets

Decision Tree Classifier





Ross Quinlan



Abdomen Length > 7.1?

no

yes

Antenna Length > 6.0?

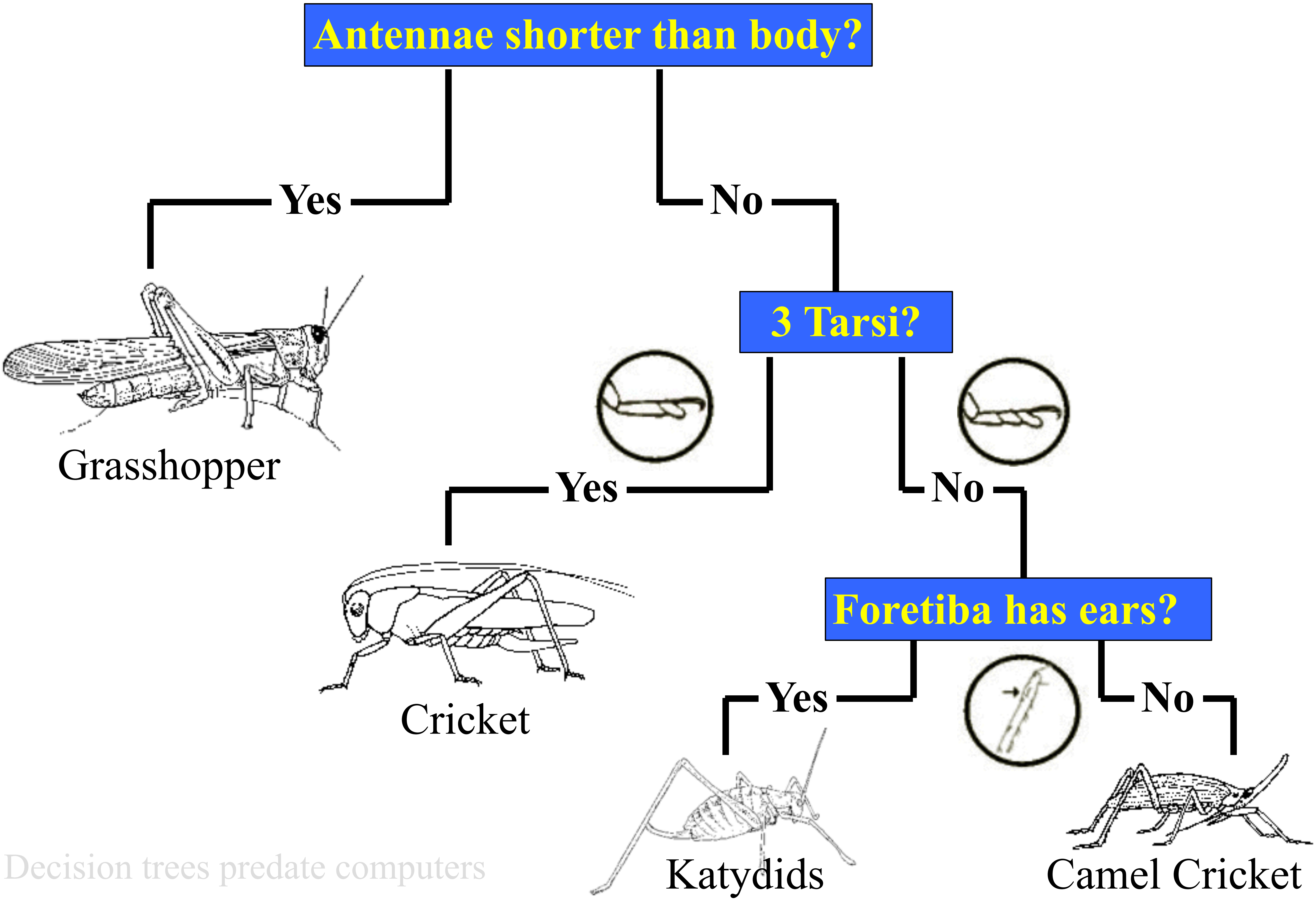
Katydid

no

yes

Grasshopper

Katydid



Antennae shorter than body?

Yes

Grasshopper

No

3 Tarsi?

Yes

Cricket

No

Foretiba has ears?

Yes

Katydids

No

Camel Cricket

Decision trees predate computers

Decision Tree Classification

- Decision tree
  - A flow-chart-like tree structure
  - Internal node denotes a test on an attribute
  - Branch represents an outcome of the test
  - Leaf nodes represent class labels or class distribution
- Decision tree generation consists of two phases
  - Tree construction
    - At start, all the training examples are at the root
    - Partition examples recursively based on selected attributes
  - Tree pruning
    - Identify and remove branches that reflect noise or outliers
- Use of decision tree: Classifying an unknown sample
  - Test the attribute values of the sample against the decision tree

How do we construct the decision tree?

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they can be discretized in advance)
  - Examples are partitioned recursively based on selected attributes.
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Advantages/Disadvantages of Decision Trees

1) Advantages:
   1) Easy to understand (Doctors love them!)
   2) Easy to generate rules
- Disadvantages:
  - May suffer from overfitting.
  - Classifies by rectangular partitioning (so does not handle correlated features very well).
  - Can be quite large – pruning is necessary.
  - Does not handle streaming data easily

Naïve Bayes Classifier

Thomas Bayes
1702 - 1761

Very useful and widely used. We won’t go into details.

Bayes Classifiers

- Bayesian classifiers use **Bayes theorem**, which says

$$p(c_j \mid d) = \frac{p(d \mid \mathrm{c}_j)\, p(c_j)}{p(d)}$$

- $p(c_j \mid d)$ = probability of instance *d* being in class c_j,

  This is what we are trying to compute

- $p(d \mid \mathrm{c}_j)$ = probability of generating instance *d* given class c_j,

  We can imagine that being in class c_j, causes you to have feature *d* with some probability

- $p(c_j)$ = probability of occurrence of class c_j,

  This is just how frequent the class c_j, is in our database

- $p(d)$ = probability of instance *d* occurring

  This can actually be ignored, since it is the same for all classes

Advantages/Disadvantages of Naïve Bayes

- Advantages:
  - Fast to train (single scan). Fast to classify
  - Not sensitive to irrelevant features
  - Handles real and discrete data
  - Handles streaming data well
- Disadvantages:
  - Assumes independence of features

Summary of Classification

We have seen 4 major classification techniques:

- Simple linear classifier, Nearest neighbor, Decision tree, Naïve Bayes.

There are other techniques:

- Neural Networks, Support Vector Machines, Genetic algorithms..

In general, there is no one best classifier for all problems. You have to consider what you hope to achieve, and the data itself…

Let us now move on to the other classic problem of data mining and machine learning, Clustering…

Exercício 1

- Considere o banco de dados da empresa Toyota contendo informações sobre funcionários, vendas, fornecedores, concessionárias, modelos, concorrência, etc.

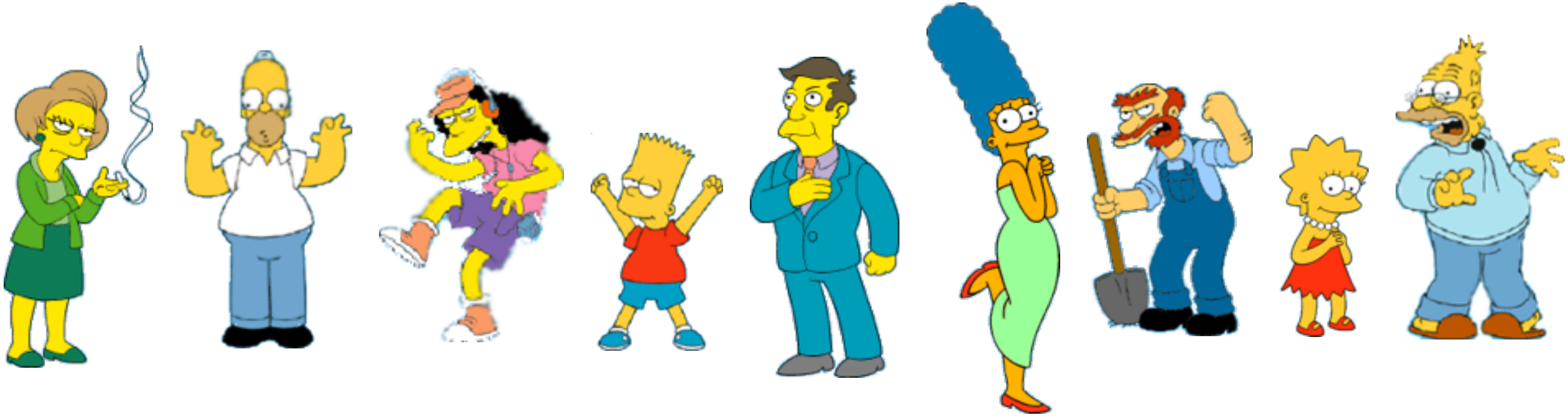
- Descreva três tarefas de classificação que poderiam ser usadas para guiar as estratégias de negócio da empresa. Use fontes de dados diversificadas.

Clustering

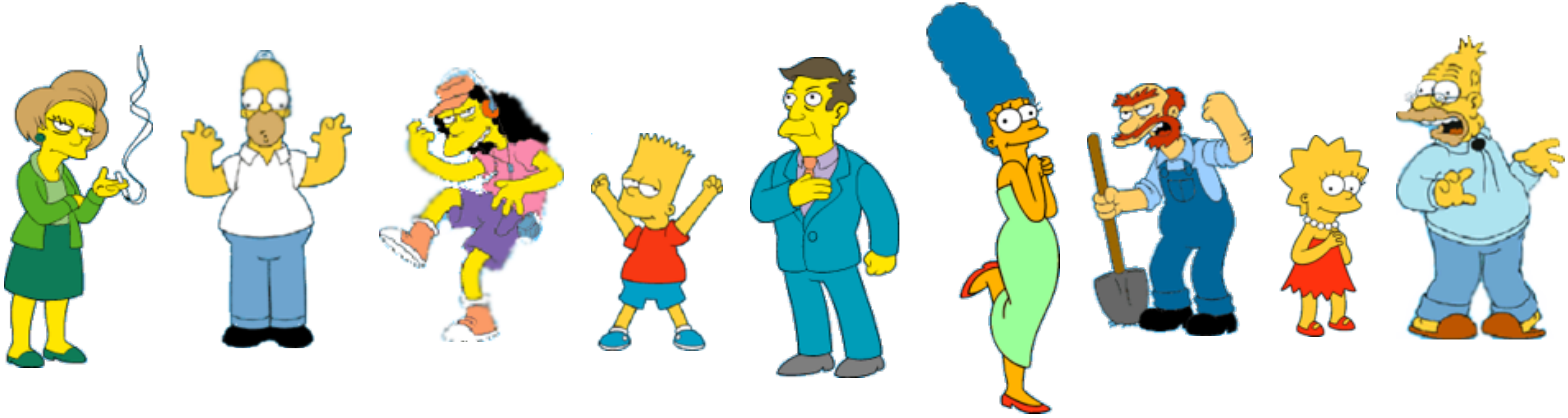
Also called *unsupervised learning*, sometimes called *classification* by statisticians and *sorting* by psychologists and *segmentation* by people in marketing

- Organizing data into classes such that there is
    - high intra-class similarity
    - low inter-class similarity
- Finding the class labels and the number of classes directly from the data (in contrast to classification).
- More informally, finding natural groupings among objects.

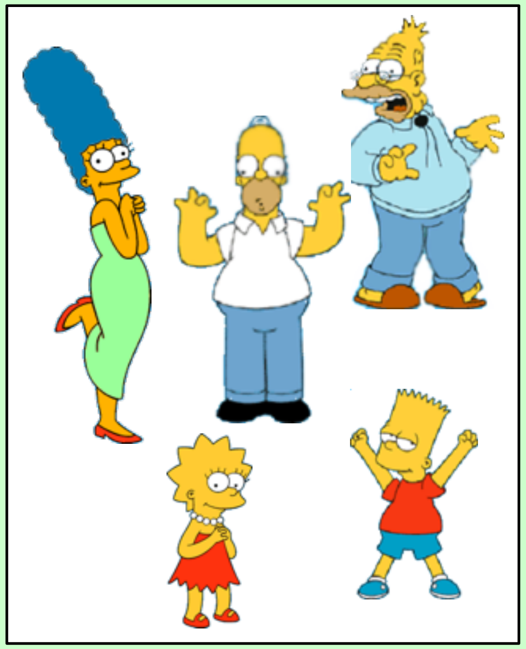
What is a natural grouping among these objects?



What is a natural grouping among these objects?

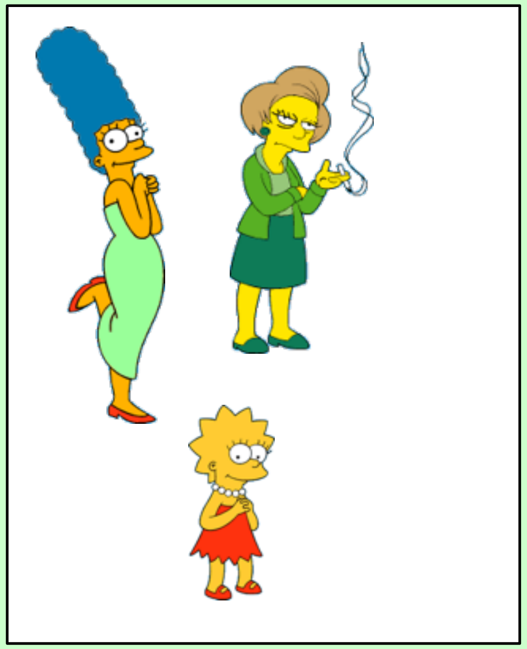


Clustering is subjective



Simpson's Family

School Employees

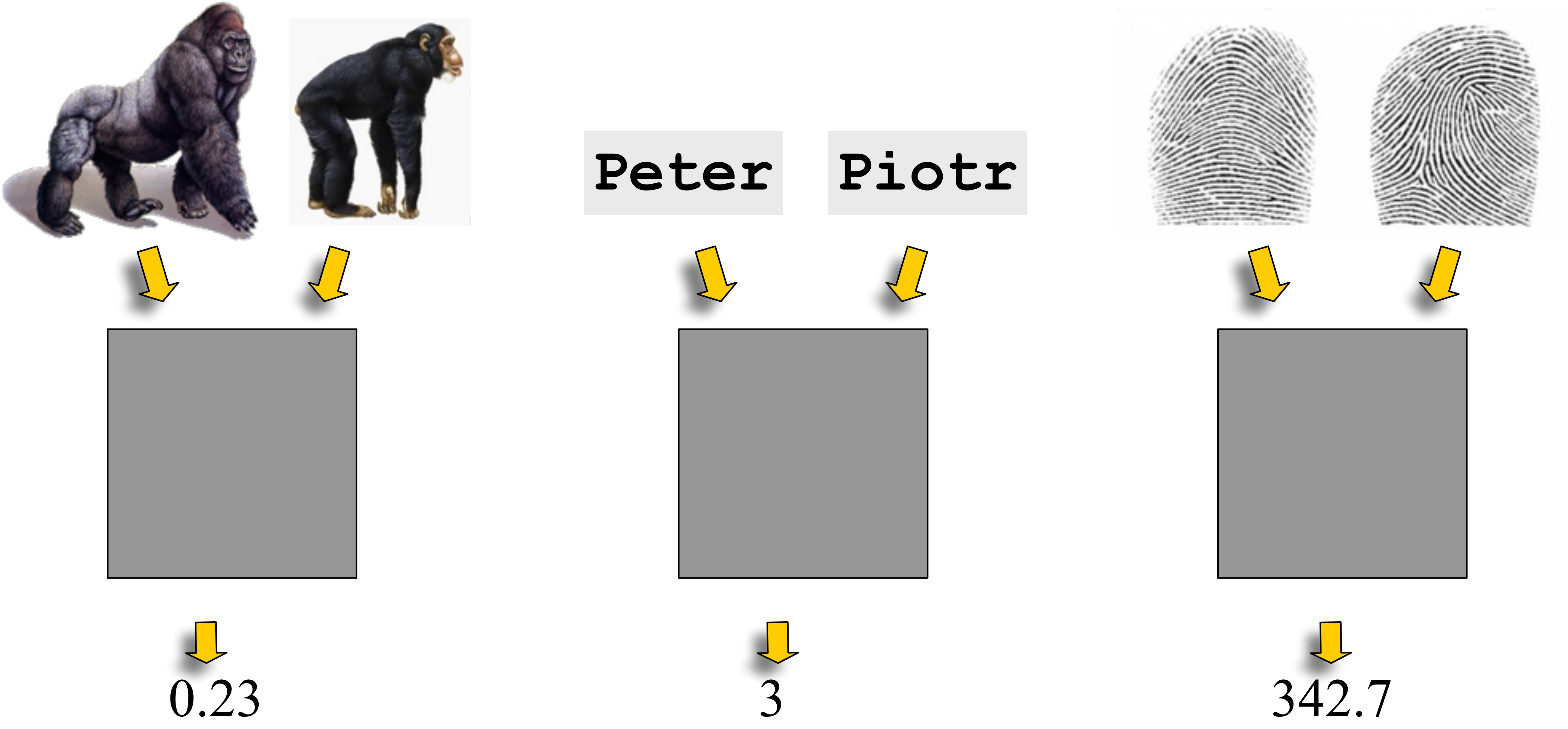


Females

Males

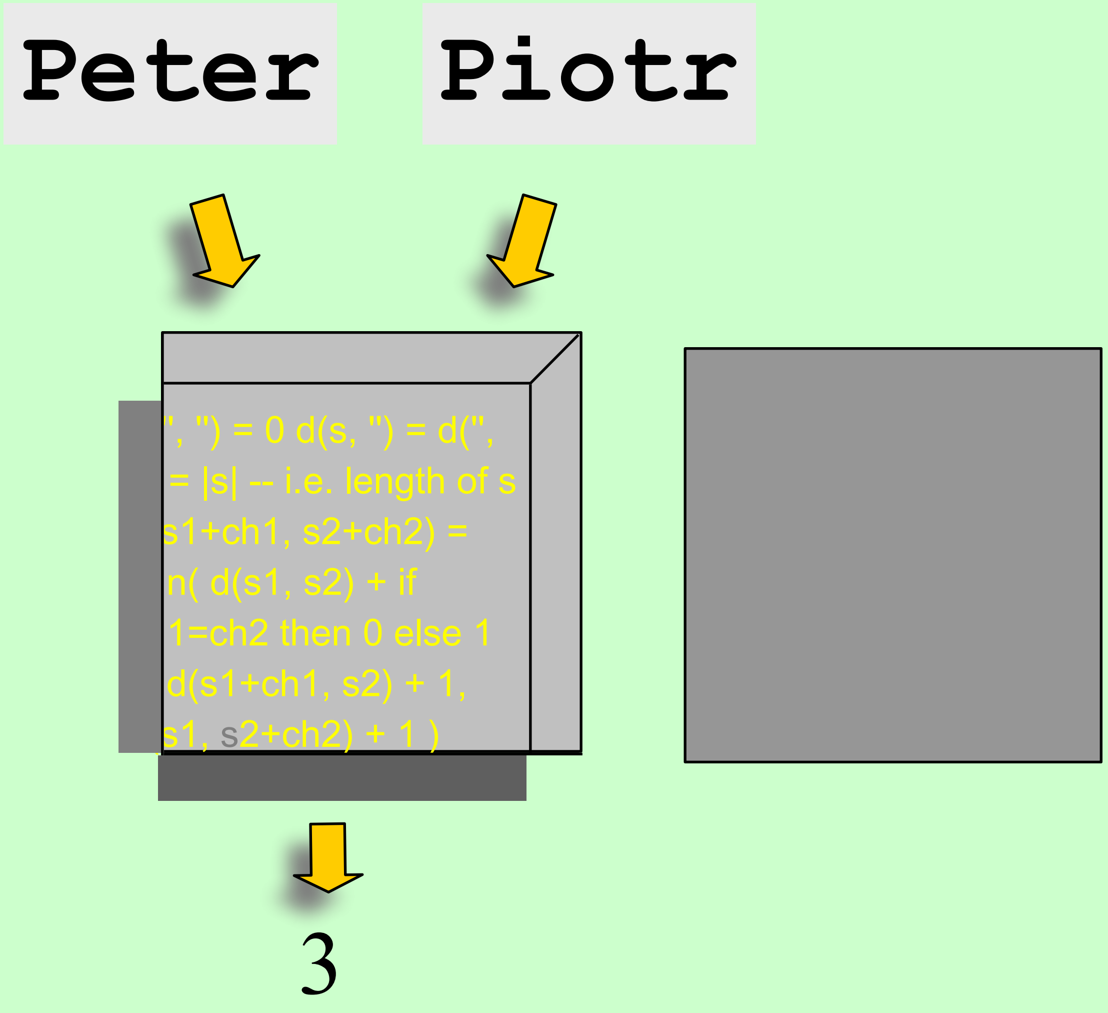
Defining Distance Measures

Definition: Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) between O_1 and O_2 is a real number denoted by $D(O_1,O_2)$



Peter Piotr

0.23 3 342.7



```
', '') = 0 d(s, '') = d('',
= |s| -- i.e. length of s
s1+ch1, s2+ch2) =
n( d(s1, s2) + if
1=ch2 then 0 else 1
d(s1+ch1, s2) + 1,
s1, s2+ch2) + 1 )
```

When we peek inside one of these black boxes, we see some function on two variables. These functions might very simple or very complex.
In either case it is natural to ask, what properties should these functions have?

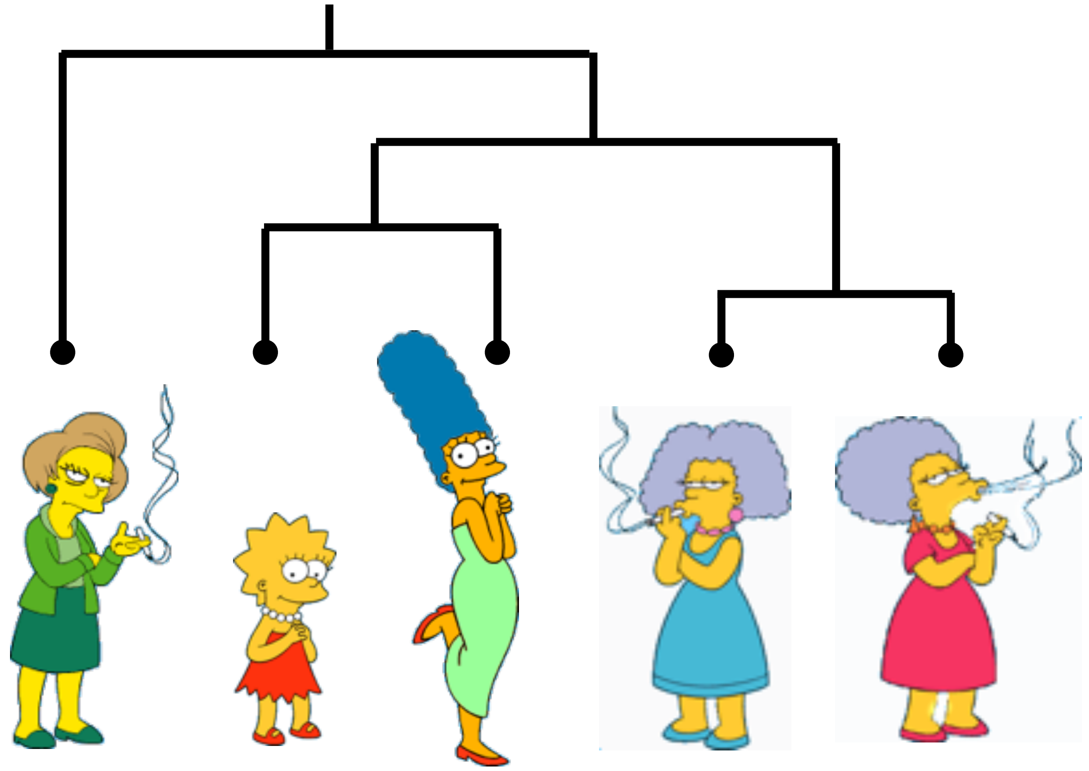
What properties should a distance measure have?

- D(A,B) = D(B,A) *Symmetry*
- D(A,A) = 0 *Constancy of Self-Similarity*
- D(A,B) = 0 IIf A= B *Positivity (Separation)*
- D(A,B) ≤ D(A,C) + D(B,C) *Triangular Inequality*

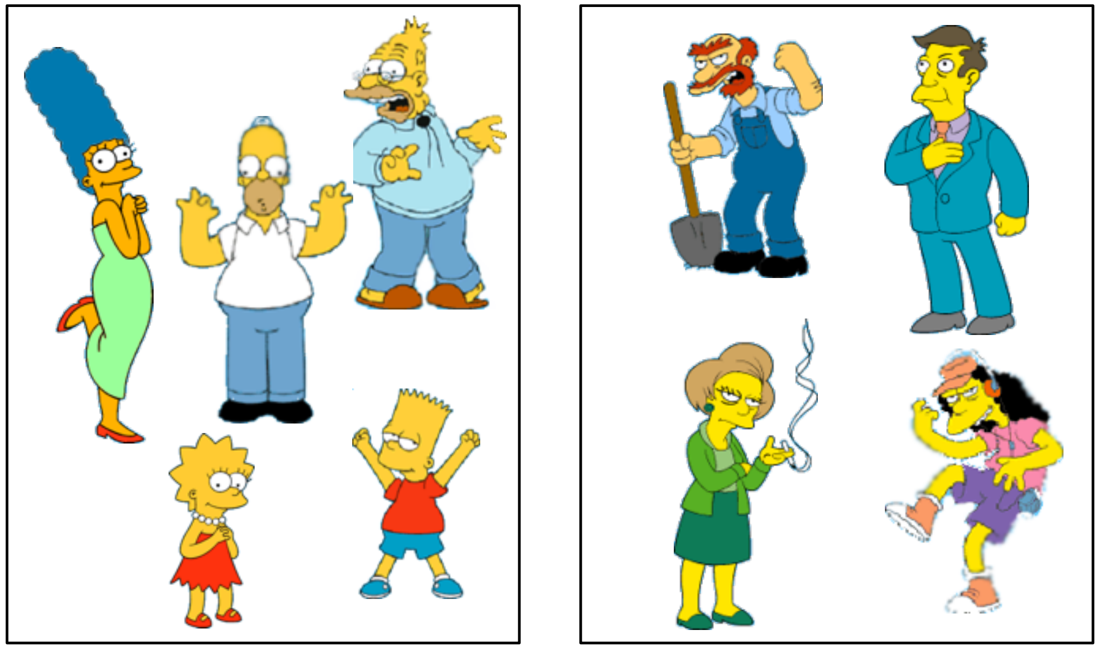
Two Types of Clustering

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion (we will see an example called BIRCH)
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion

Hierarchical



Partitional

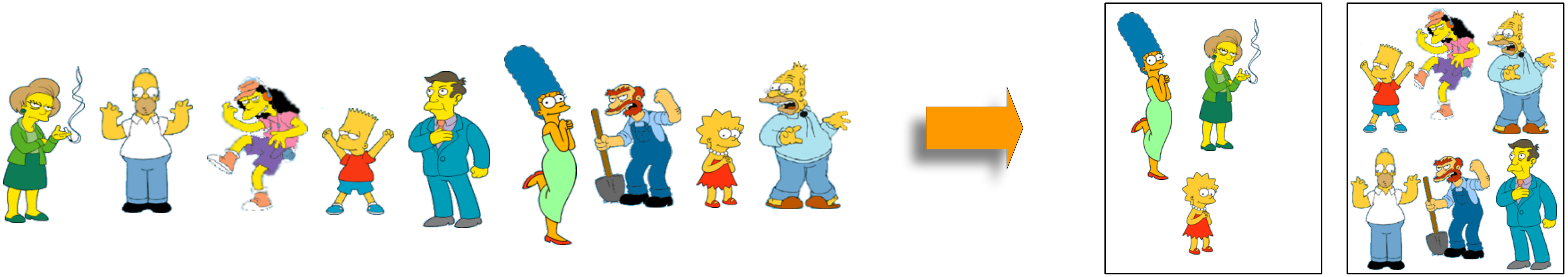


Desirable Properties of a Clustering Algorithm

- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- Incorporation of user-specified constraints
- Interpretability and usability

Partitional Clustering

- Nonhierarchical, each instance is placed in exactly one of K nonoverlapping clusters.
- Since only one set of clusters is output, the user normally has to input the desired number of clusters K.



Algorithm *k-means*

1. Decide on a value for *k*.

2. Initialize the *k* cluster centers (randomly, if necessary).

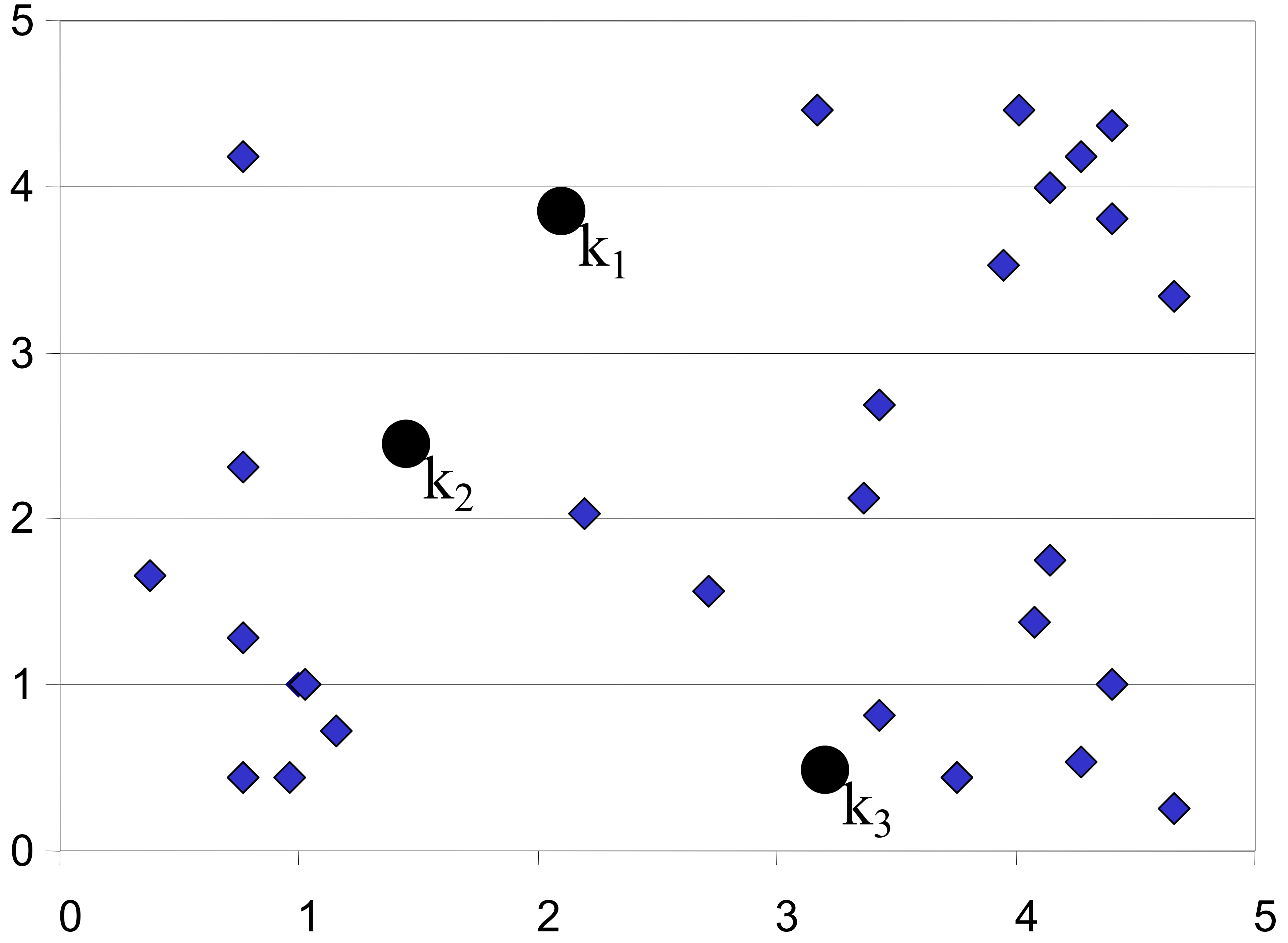
3. Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.

4. Re-estimate the *k* cluster centers, by assuming the memberships found above are correct.

5. If none of the *N* objects changed membership in the last iteration, exit. Otherwise goto 3.

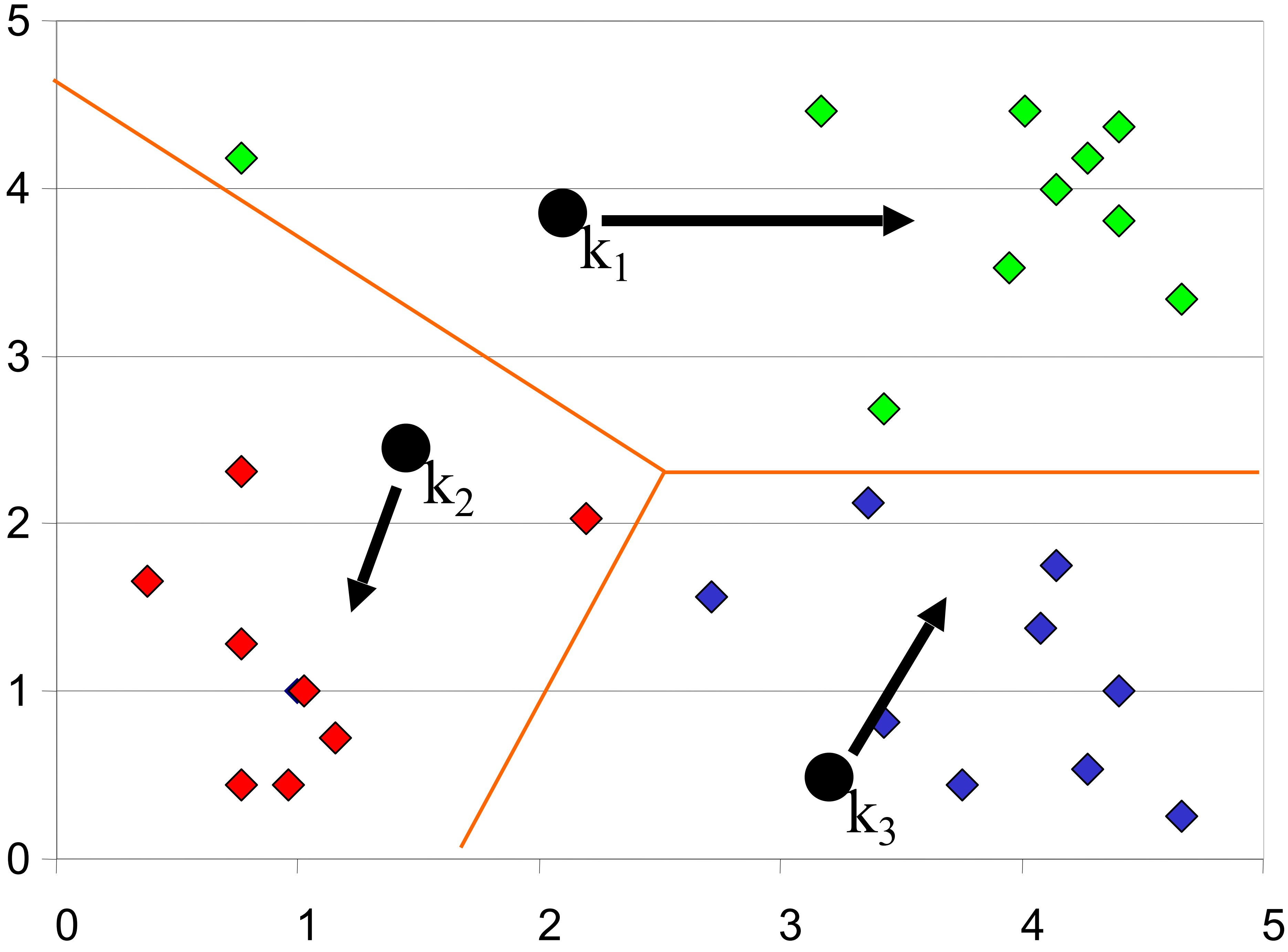
K-means Clustering: Step 1

Algorithm: k-means, Distance Metric: Euclidean Distance



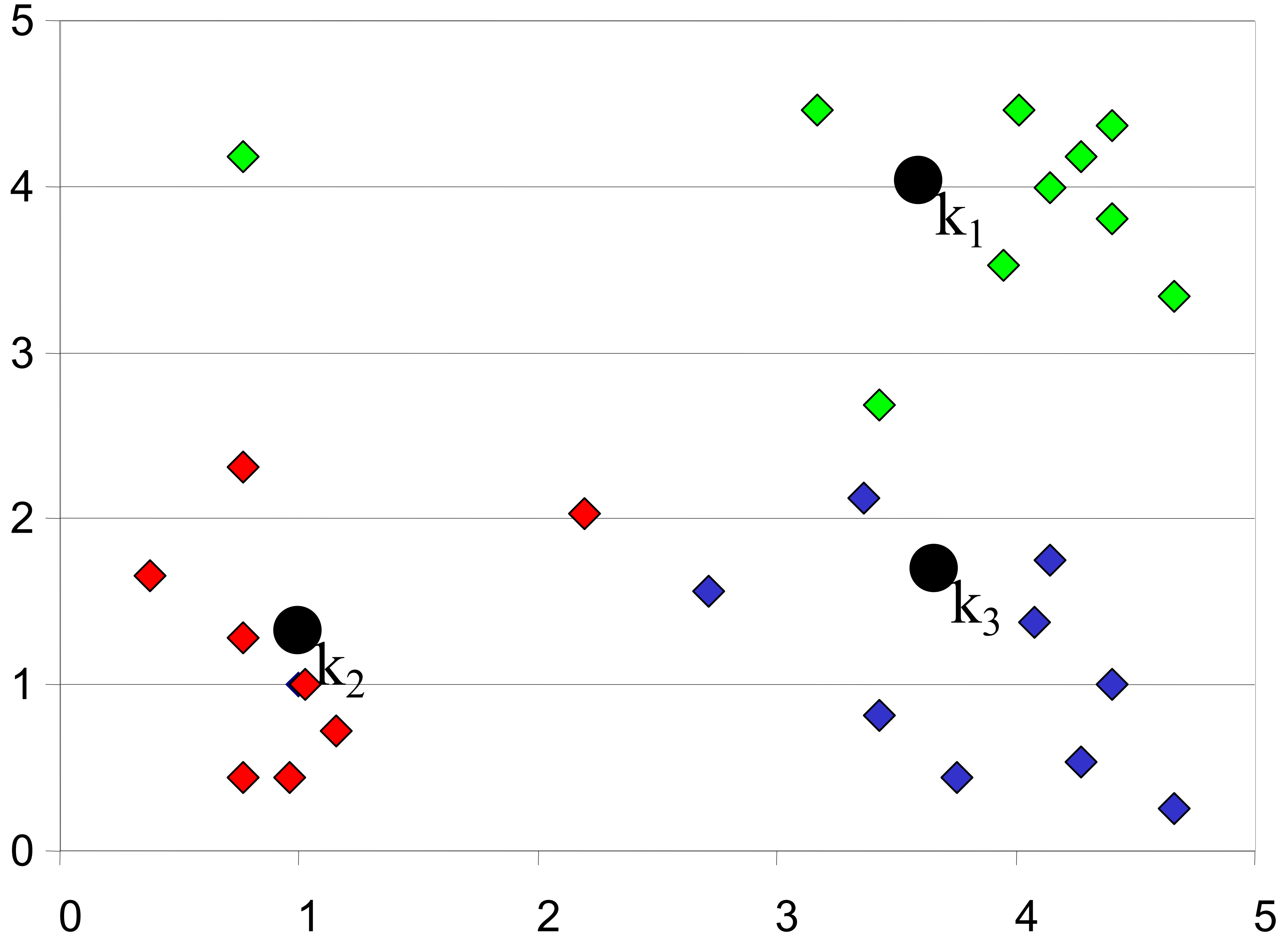
K-means Clustering: Step 2

Algorithm: k-means, Distance Metric: Euclidean Distance



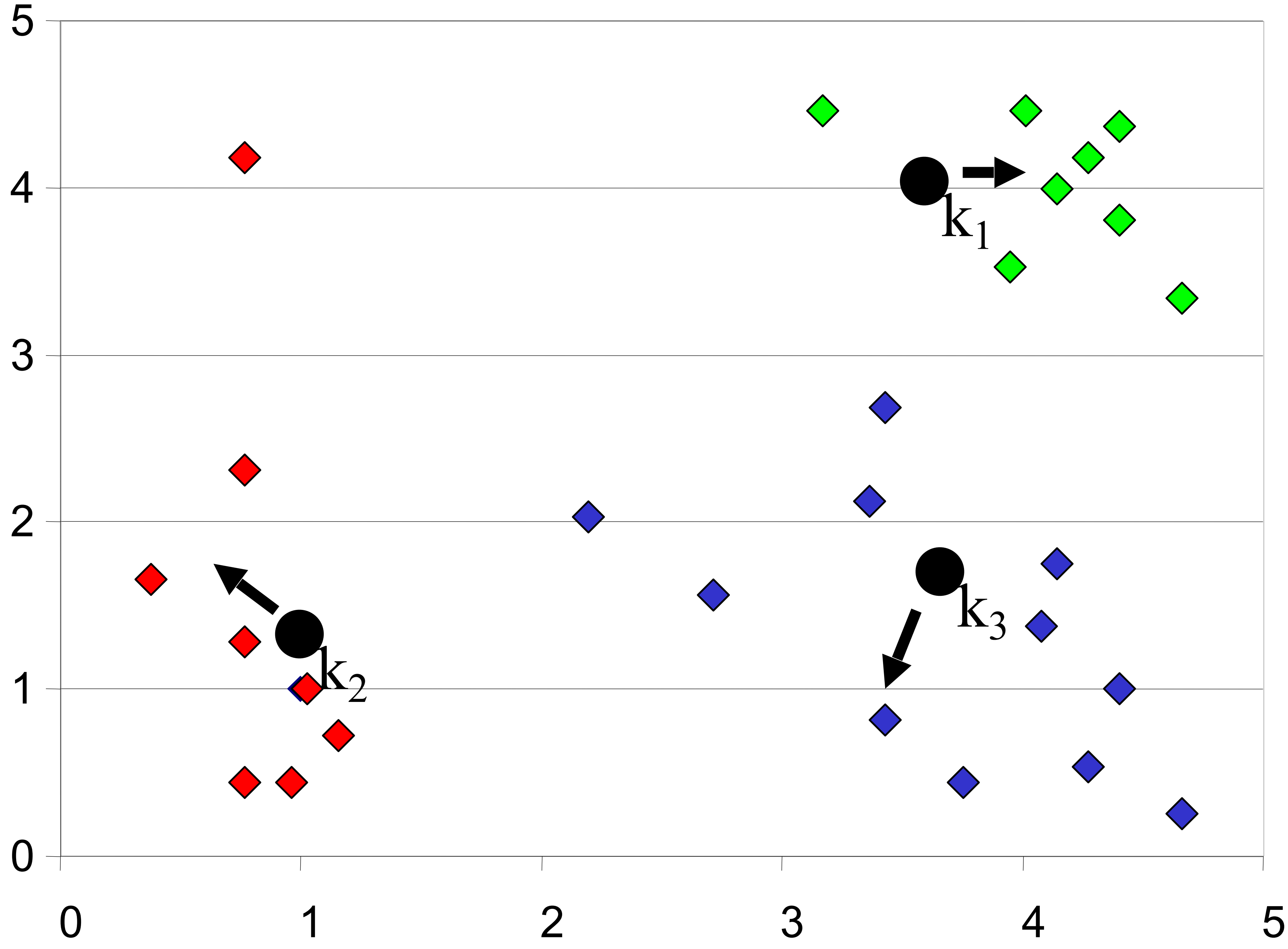
K-means Clustering: Step 3

Algorithm: k-means, Distance Metric: Euclidean Distance



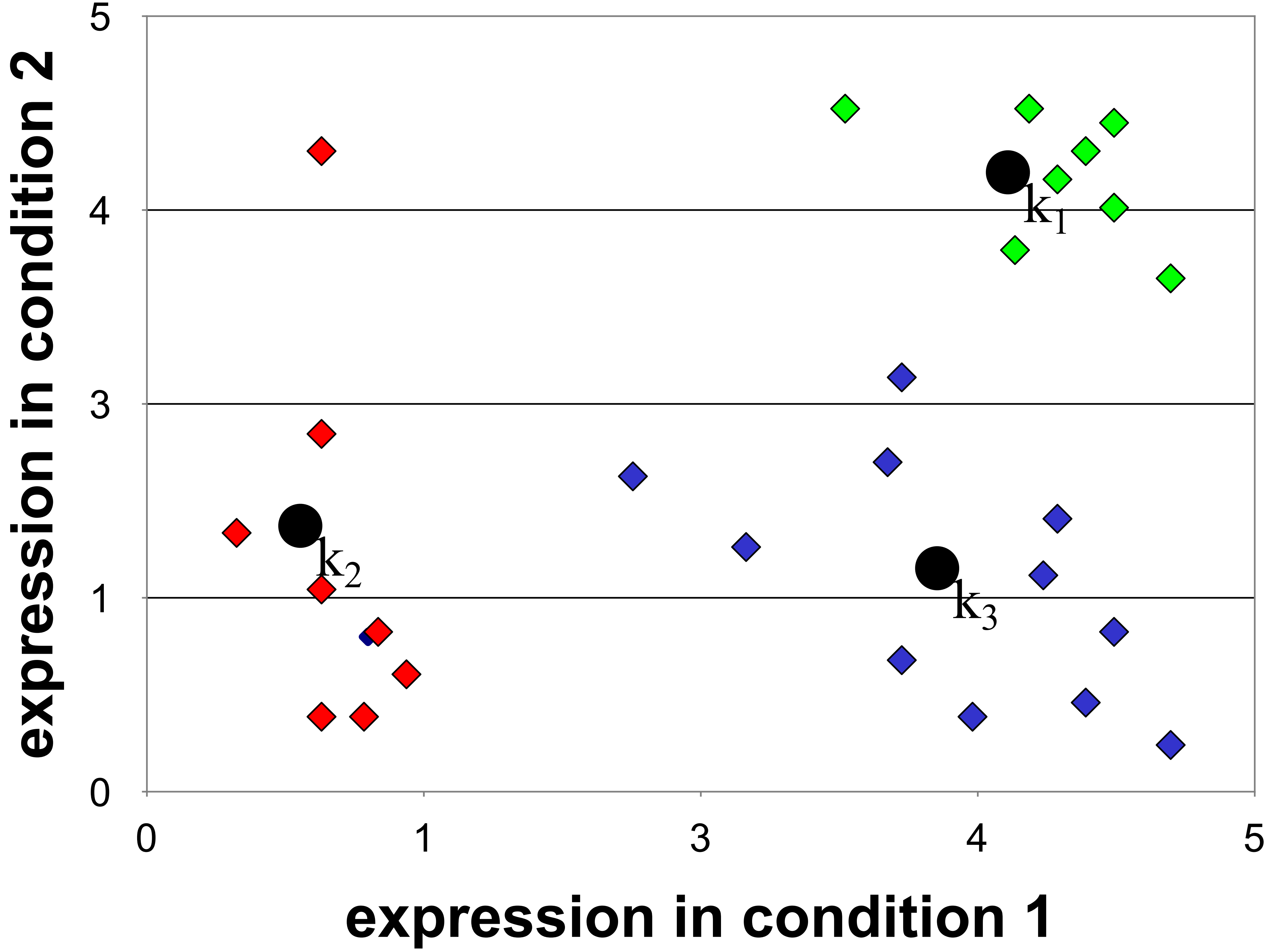
K-means Clustering: Step 4

Algorithm: k-means, Distance Metric: Euclidean Distance



K-means Clustering: Step 5

Algorithm: k-means, Distance Metric: Euclidean Distance



Comments on the *K-Means* Method

- Strength
  - *Relatively efficient*: $O(tkn)$, where *n* is # objects, *k* is # clusters, and *t* is # iterations. Normally, $k, t << n$.
  - Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- Weakness
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify *k,* the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

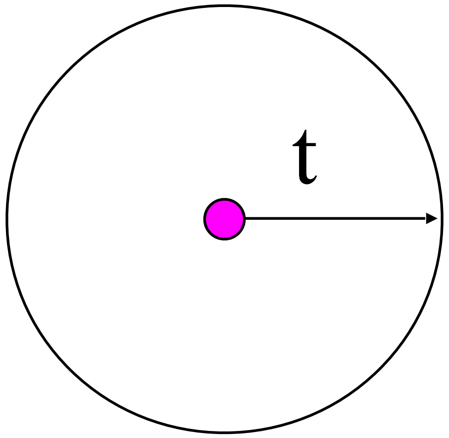
What happens if the data is streaming…

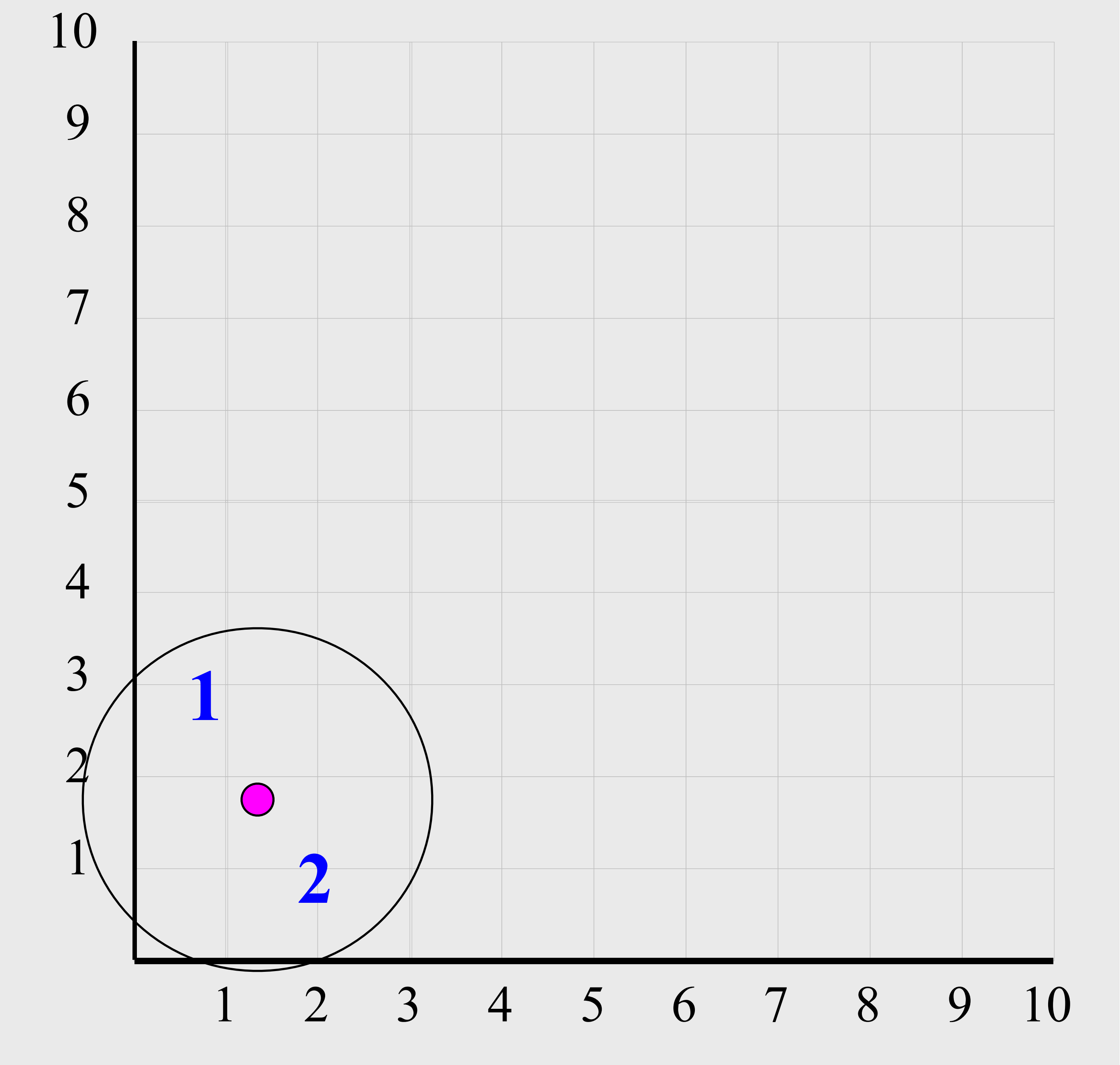
Nearest Neighbor Clustering

Not to be confused with Nearest Neighbor **Classification**

- Items are iteratively merged into the existing clusters that are closest.
- Incremental
- Threshold, t, used to determine if items are added to existing clusters or a new cluster is created.

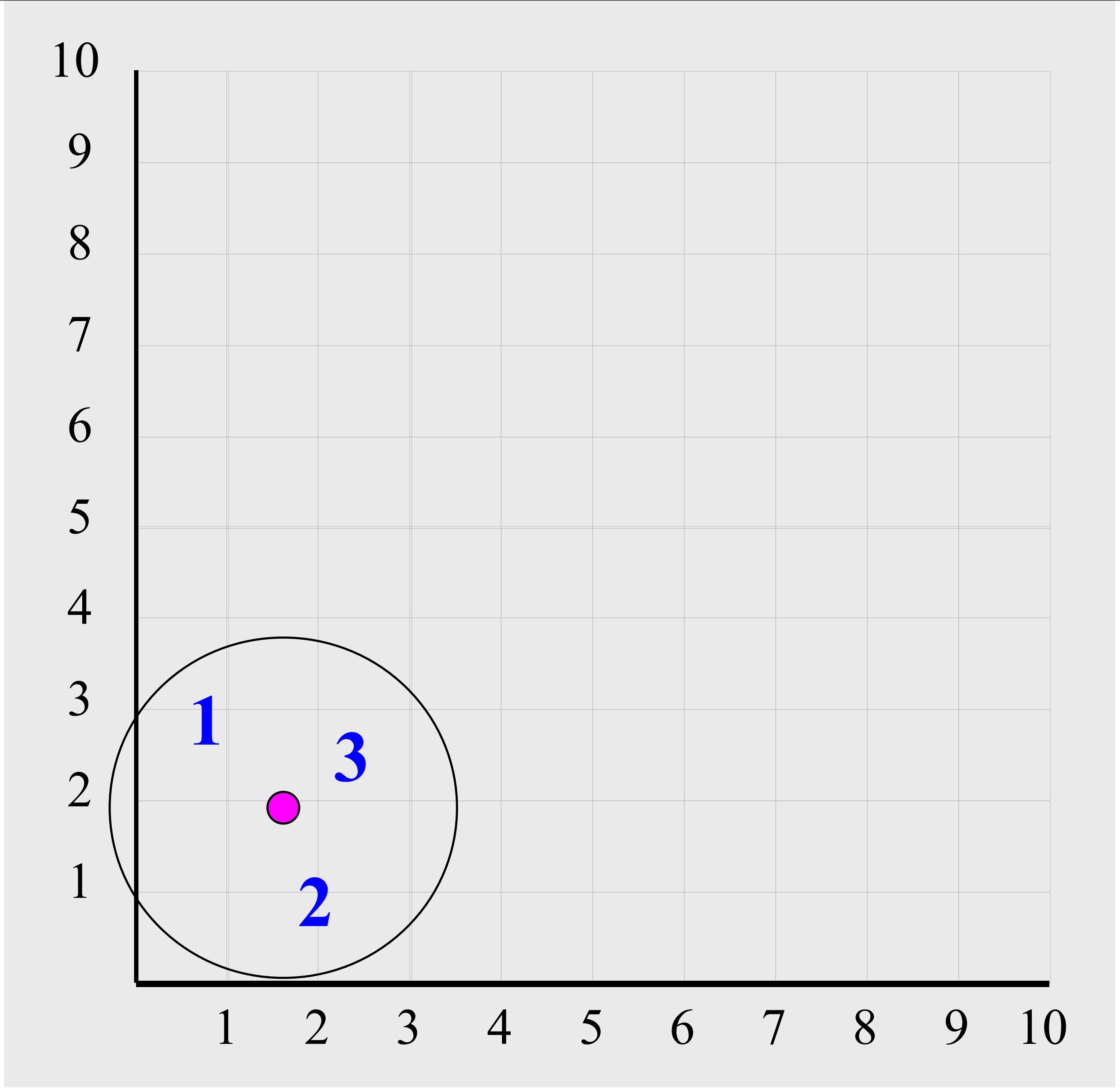
Threshold t





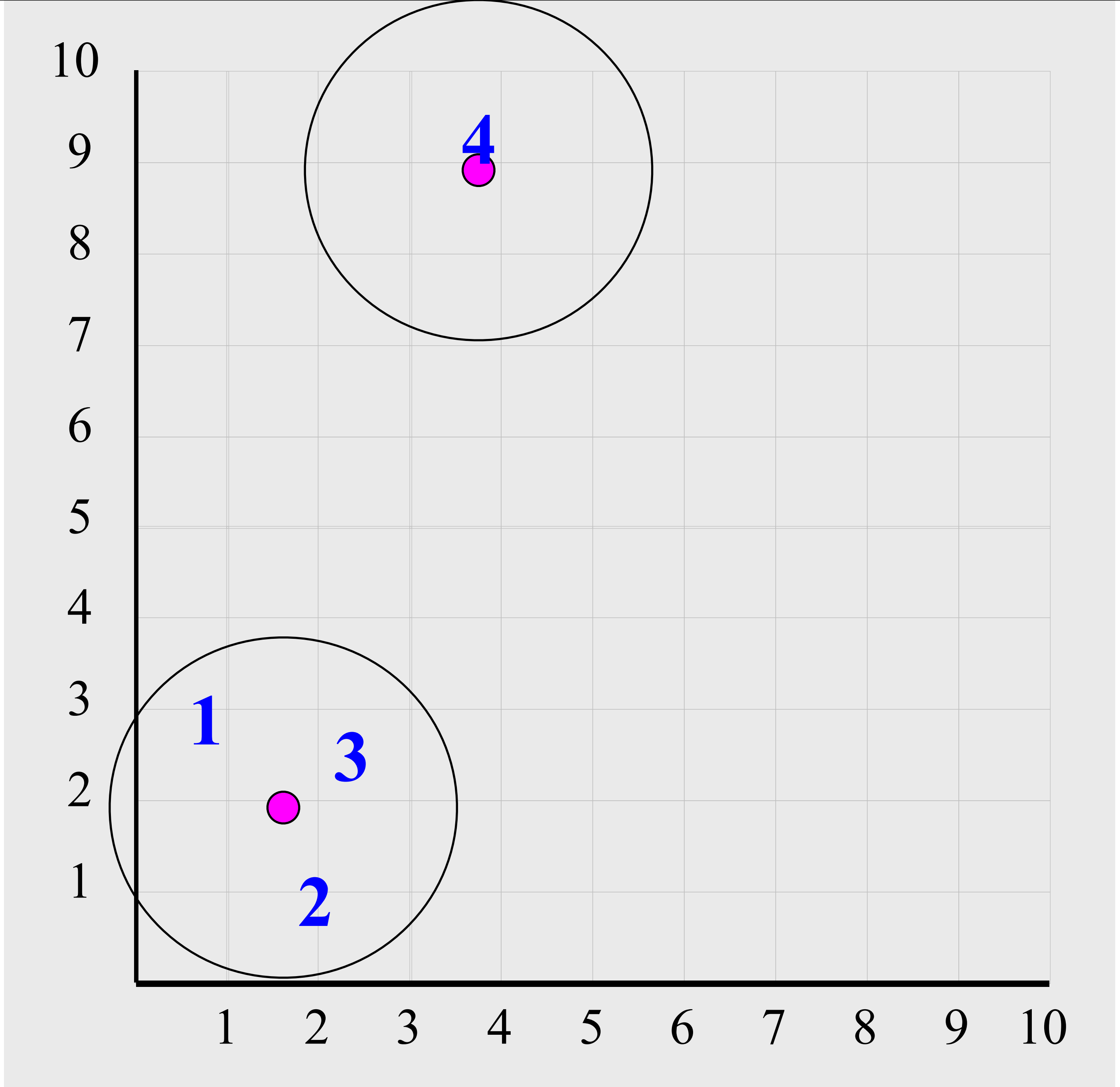
New data point arrives…

It is within the threshold for cluster 1, so add it to the cluster, and update cluster center.



New data point arrives…

It is **not** within the threshold for cluster 1, so create a new cluster, and so on..

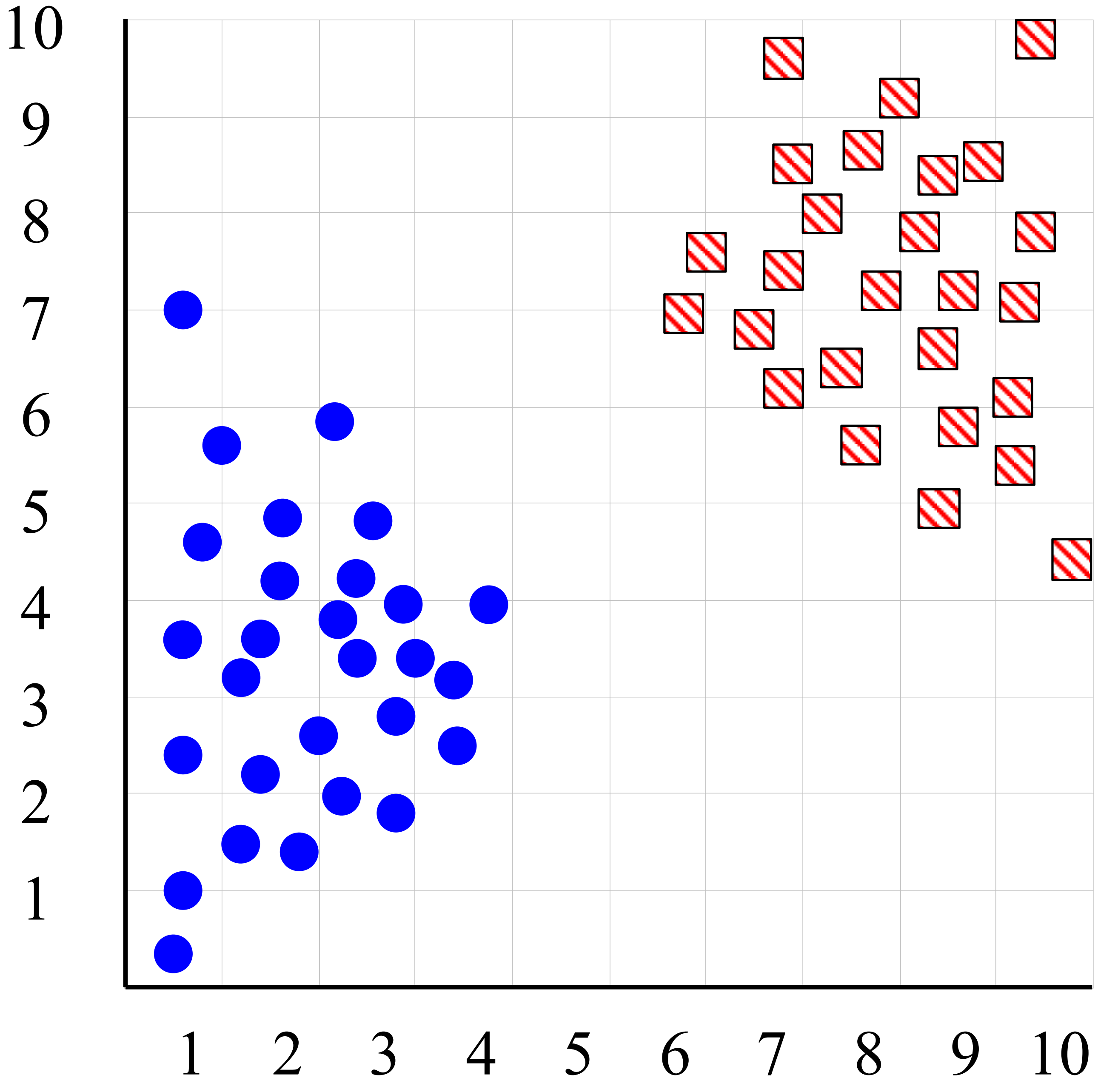


Algorithm is highly order dependent…

It is difficult to determine t in advance…

How can we tell the *right* number of clusters?

In general, this is a unsolved problem. However there are many approximate methods.



Exercício 2

- Defina uma tarefa de clusterização para clientes da Amazon.com. Identifique features relevantes e indique que tipo de agrupamentos poderiam ser encontrados.

Association Rules (market basket analysis)

- Retail shops are often interested in associations between different items that people buy.
  - Someone who buys bread is quite likely also to buy milk
  - A person who bought the book *Database System Concepts* is quite likely also to buy the book *Operating System Concepts*.
- Associations information can be used in several ways.
  - E.g. when a customer buys a particular book, an online shop may suggest associated books.
- **Association rules:**

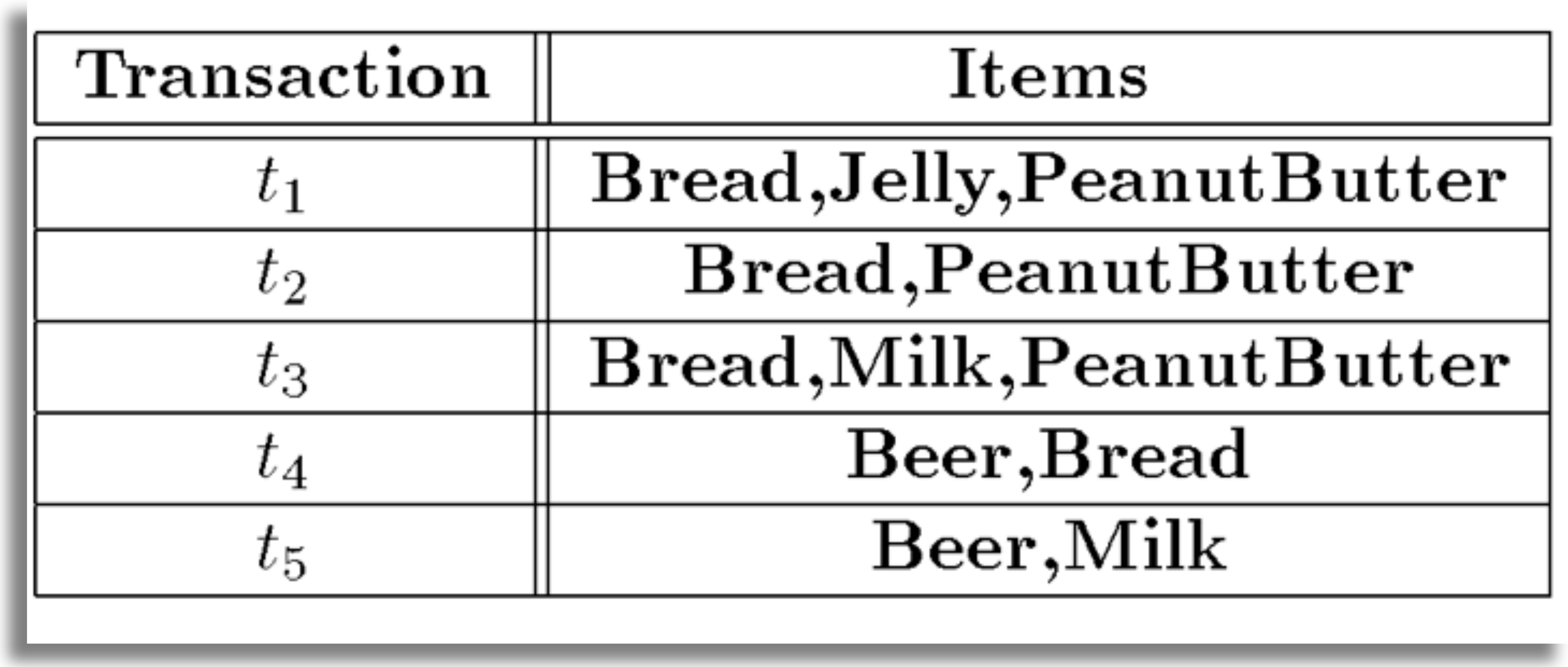
  bread ⇒ *milk*   *DB-Concepts, OS-Concepts* ⇒ Networks

  - Left hand side: antecedent,   right hand side: consequent
  - An association rule must have an associated population; the population consists of a set of instances
    - E.g. each transaction (sale) at a shop is an instance, and the set of all transactions is the population

Association Rule Definitions

- ***Set of items:*** $I=\{I_1,I_2,\ldots,I_m\}$
- ***Transactions:*** $D=\{t_1,t_2, \ldots, t_n\}$, $t_j\subseteq I$
- ***Itemset:*** $\{I_{i1},I_{i2}, \ldots, I_{ik}\} \subseteq I$
- ***Support of an itemset:*** Percentage of transactions which contain that itemset.
- ***Large (Frequent) itemset:*** Itemset whose number of occurrences is above a threshold.

Association Rules Example



Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

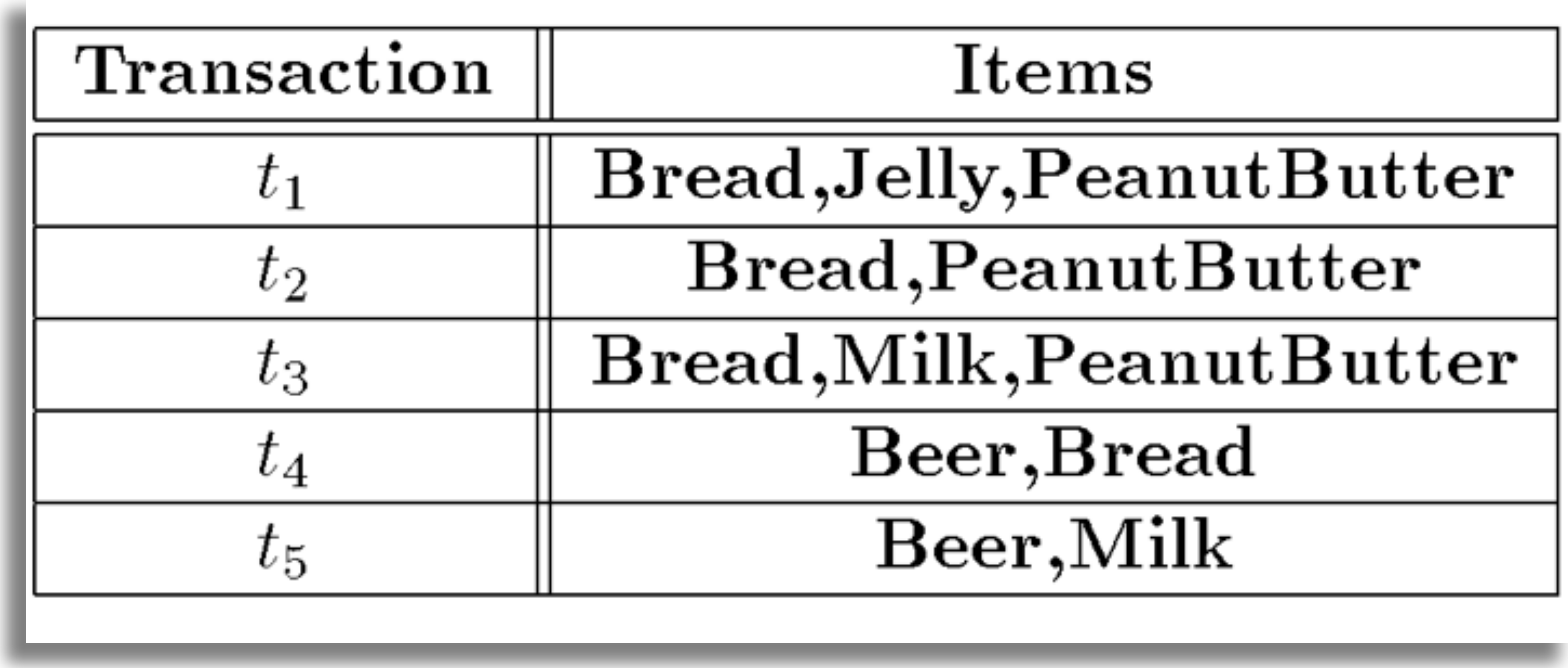
I = { Beer, Bread, Jelly, Milk, PeanutButter}

Support of {Bread,PeanutButter} is 60%

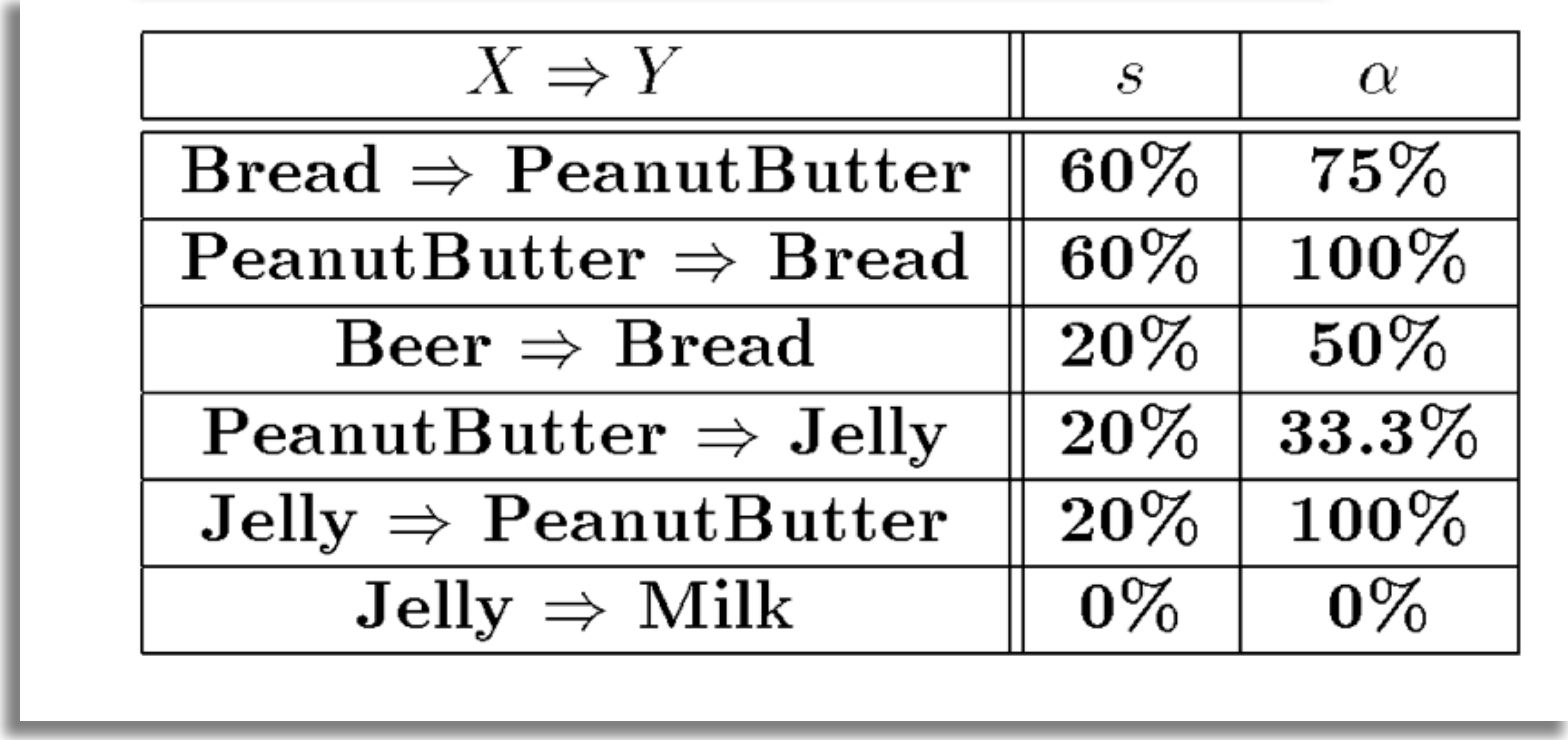
Association Rule Definitions

- ***Association Rule (AR):*** implication $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y =$ the null set;
- ***Support of AR (s) $X \Rightarrow Y$***: Percentage of transactions that contain $X \cup Y$
- ***Confidence of AR (*** α ***) $X \Rightarrow Y$:*** Ratio of number of transactions that contain $X \cup Y$ to the number that contain X

Association Rules Example

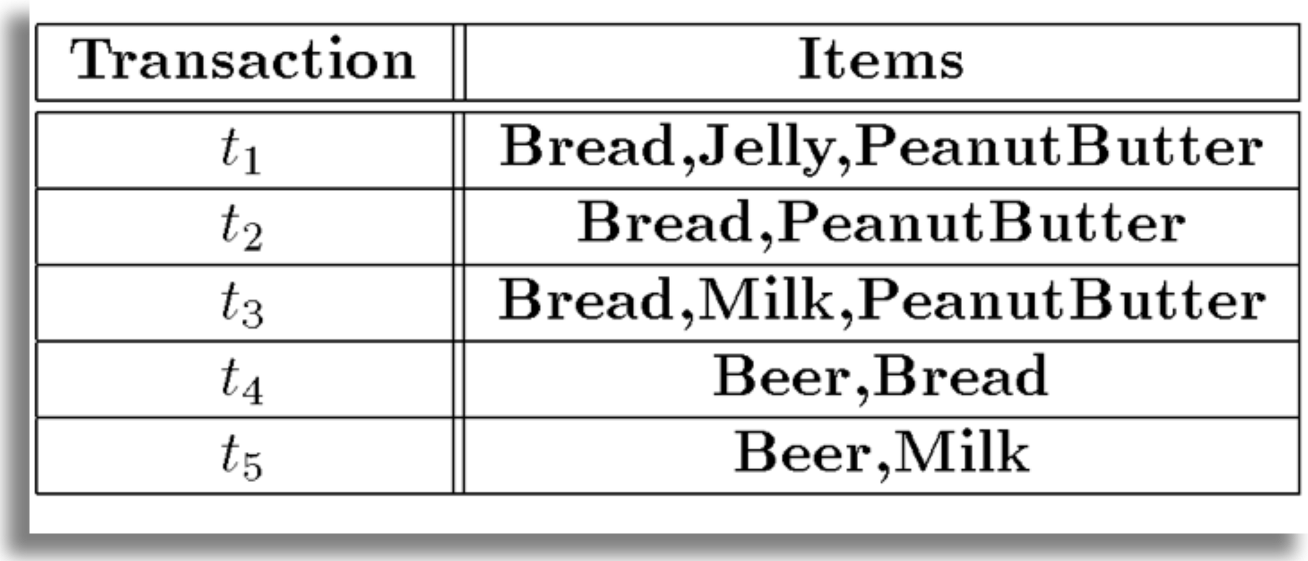


Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk



$X \Rightarrow Y$	s	α
Bread ⇒ PeanutButter	60%	75%
PeanutButter ⇒ Bread	60%	100%
Beer ⇒ Bread	20%	50%
PeanutButter ⇒ Jelly	20%	33.3%
Jelly ⇒ PeanutButter	20%	100%
Jelly ⇒ Milk	0%	0%

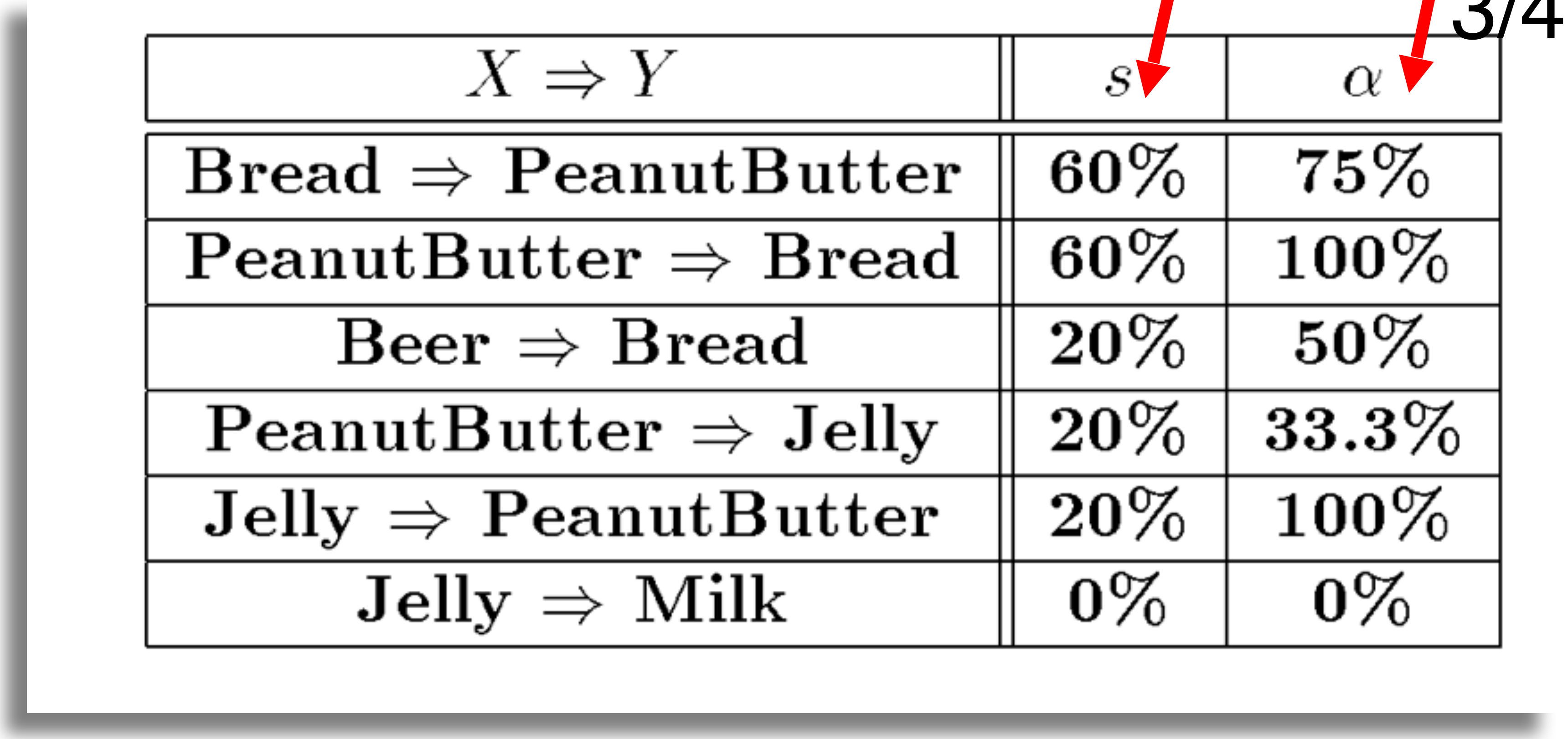
Association Rules Example



Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- Of 5 transactions, 3 involve both Bread and PeanutButter, 3/5 = 60%

- Of the 4 transactions that involve Bread, 3 of them also involve PeanutButter 3/4 = 75%



$X \Rightarrow Y$	s	α
Bread ⇒ PeanutButter	60%	75%
PeanutButter ⇒ Bread	60%	100%
Beer ⇒ Bread	20%	50%
PeanutButter ⇒ Jelly	20%	33.3%
Jelly ⇒ PeanutButter	20%	100%
Jelly ⇒ Milk	0%	0%

Association Rule Problem

- Given a set of items $I=\{I_1,I_2,\ldots,I_m\}$ and a database of transactions $D=\{t_1,t_2, \ldots, t_n\}$ where $t_i=\{I_{i1},I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, the ***Association Rule Problem*** is to identify all association rules $X \Rightarrow Y$ with a minimum support and confidence (supplied by user).

- ***NOTE:*** Support of $X \Rightarrow Y$ is same as support of $X \cup Y$.

Association Rule Algorithm (Basic Idea)

1. Find Large Itemsets.
2. Generate rules from frequent itemsets.

This is the simple naïve algorithm, better algorithms exist.

Exercício 3

- Encontre possíveis regras de associação para atividades de usuários de uma rede social como Facebook (use algum outro serviço online caso você não esteja familiarizado com redes sociais). O objetivo deve ser sugerir novas atividades a partir de uma atividade executada por um usuário. Exemplos de atividades são adicionar usuário, postar, ver profile, enviar mensagem, sugerir amigos, etc.

Cuidado! Paradoxo de Rhine

- Um parapsicologista nos anos 50 acreditava que certas pessoas tinham Percepção Extra-Sensorial (PES)

- Ele criou um experimento onde sujeitos tinham que adivinhar a cor (vermelho ou azul) de 10 cartas escondidas

- Ele descobriu que cerca de 1 em 1000 tinham PES - eles acertaram todas as 10!

- Ele então disse as pessoas que elas tinham PES e as chamou para outro teste idêntico

- Surpresa! Ele descobriu que as pessoas perderam seus poderes!

- Ele concluiu: não se deve dizer para as pessoas que ela têm PES pois elas perdem os poderes ao descobrir!

Data Mining hoje e no futuro

- Sistemas de recomendação
- Diagnósticos de saúde
- MapReduce
- Web Mining/Extração de Informação
- Streams de dados
- Machine Learning em larga escala
- Análise de links (webgraph, PageRank)
- Redes Sociais
- Análise de sentimento . . .

Data mining/Machine learning toolkits

WEKA
The University
of Waikato

Conclusions

- Data Mining is an important procedure in strategic knowledge discovery
- We have seen three main tasks: classification, clustering, association rules
- Many other tasks and variations, but most based on the three above
- Machine Learning techniques are becoming a major part or DM

References

- Dr Eamonn Keogh DM Course Slides
- Ramakrishnan and Gehrke 2003 - Database Management Systems - 3rd Edition
- Elmasri and Navathe 2003 - Fundamentals of Database Systems - 4th Edition
- Han, Kamber 2011 - Data Mining - Concepts and Techniques - 3rd Edition
- Stanford CS345A Data Mining course notes, 2010

Exercício 1

- Considere o banco de dados da empresa Toyota contendo informações sobre funcionários, vendas, fornecedores, concessionárias, modelos, concorrência, etc.

- Descreva três tarefas de classificação que poderiam ser usadas para guiar as estratégias de negócio da empresa. Use fontes de dados diversificadas.

- Resposta: classificar funcionários de confiança (features: tempo de trabalho, idade, faltas, notificações…), fornecedores preferenciais (features: tempo de parceria, volume de negócios interno, volume total, tempo médio de entrega…), concessionárias de luxo (renda per capta da região, valor médio de compras, quantidade de opcionais, etc…)

Exercício 2

- Defina uma tarefa de clusterização para clientes da Amazon.com. Identifique features relevantes e indique que tipo de agrupamentos poderiam ser encontrados.

- Resposta: Clientes por período preferencial de compra (compradores noturnos, clientes comerciais, etc…)

- Importante: função de distância deve ser bem definida da tarefa!

Exercício 3

- Encontre possíveis regras de associação para atividades de usuários de uma rede social como Facebook (use algum outro serviço online caso você não esteja familiarizado com redes sociais). O objetivo deve ser sugerir novas atividades a partir de uma atividade executada por um usuário. Exemplos de atividades são adicionar usuário, postar, ver profile, enviar mensagem, sugerir amigos, etc.

- Resposta: adicionar usuário -> postar no mural, ser marcado em um álbum -> adicionar outras pessoas marcadas como amigos…