Introduction to Machine Learning for Civil & Architectural Engineers

November 7, 2017

Prof. Dr. Zoltan Nagy
nagy@utexas.edu
http://nagy.caee.utexas.edu

Plan for Today

- What is Machine Learning
- From Data to Wisdom
- A quick brush over typical ML algorithms
- Example of analyzing large amounts of time-series data

- Lecture notes are based on the course *Smart Buildings & Cities* from Spring 2017

What is Machine Learning?

Machine Learning

"Field of study that gives computers the ability to ***learn*** *without being explicitly programmed. "*

- Arthur Samuel (1959)

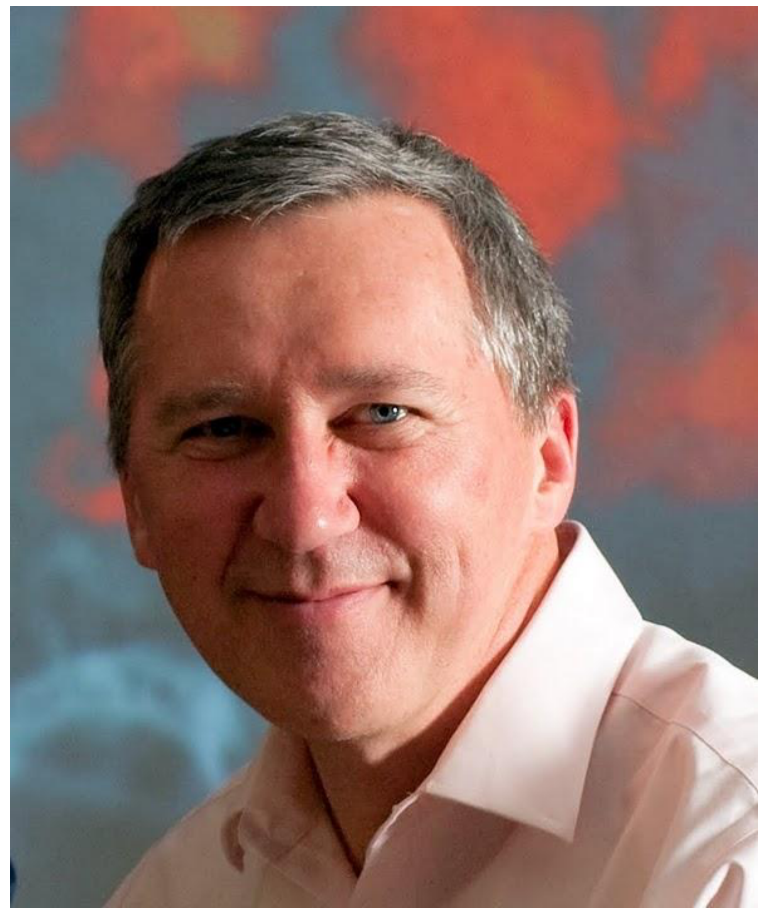
Machine Learning

"Field of study that gives computers the ability to ***learn*** *without being explicitly programmed. "*

- Arthur Samuel (1959)

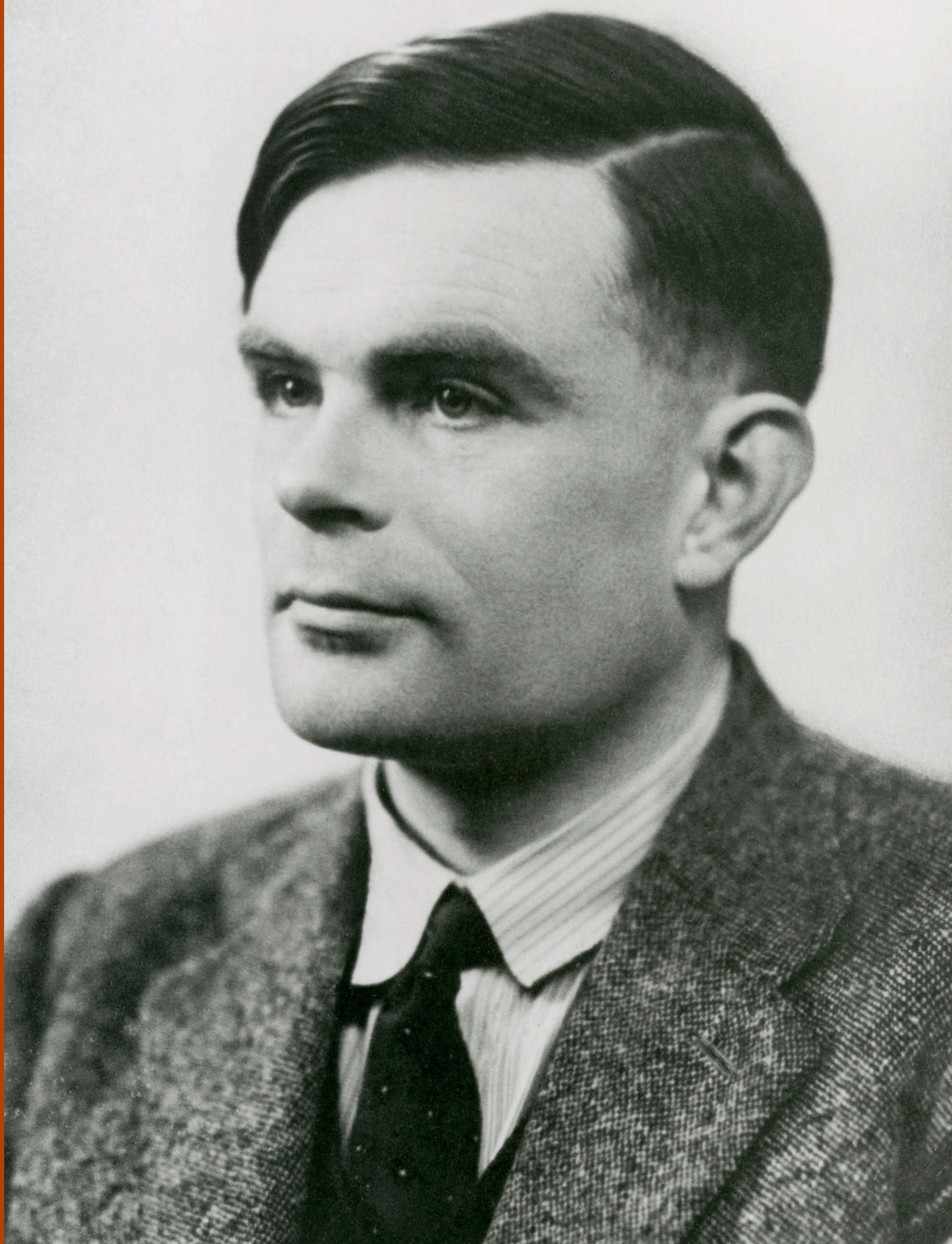
"A computer program is said to learn [...], if its ***performance*** *on a task* ***improves*** *with* ***experience****."*

- Tom Mitchell (1997)



“ Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? ”

Alan T~~o~~uring
Father of the Modern Computer
(1912-1954)

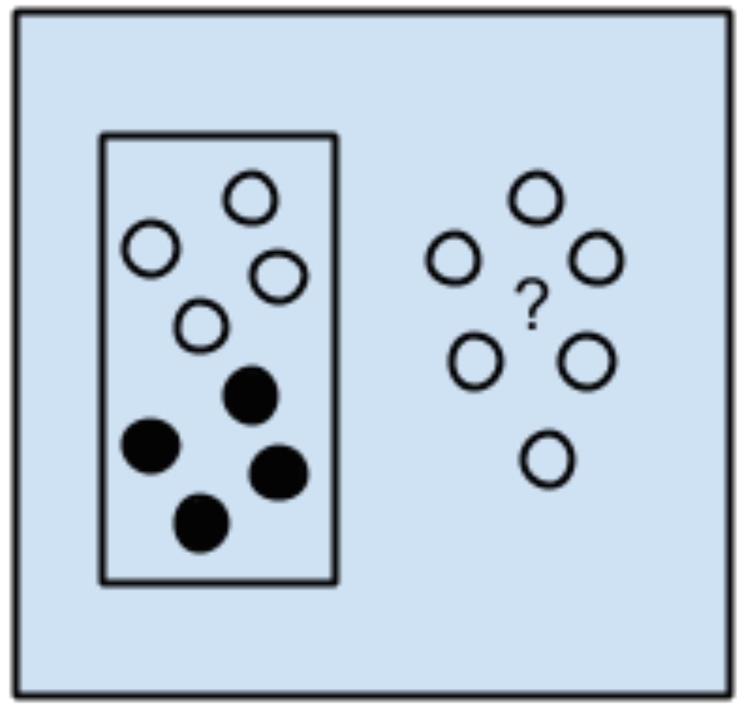


What machine learning algorithms do YOU know?

Machine Learning

$(x_1, x_2, x_3, \ldots) \longrightarrow (y_1, y_2)$

- Supervised Learning (95%)
  - " Learn **mapping** from input to output "
    - Classification ← ~~to~~ classes / groups
    - Regression / Ranking ↪ numeric data



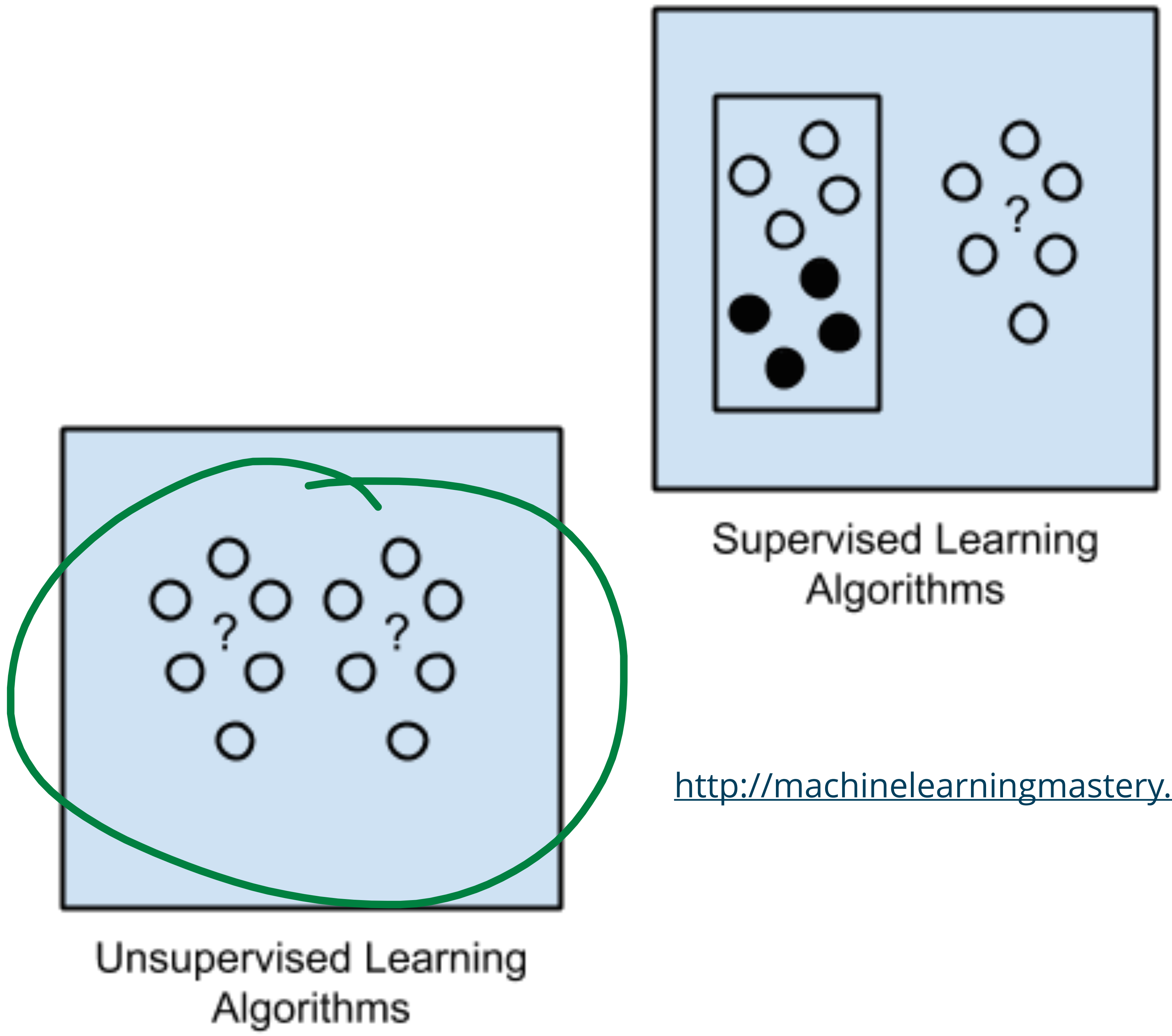
Supervised Learning Algorithms

http://machinelearningmastery.com

Machine Learning

GARBAGE IN = GARBAGE OUT

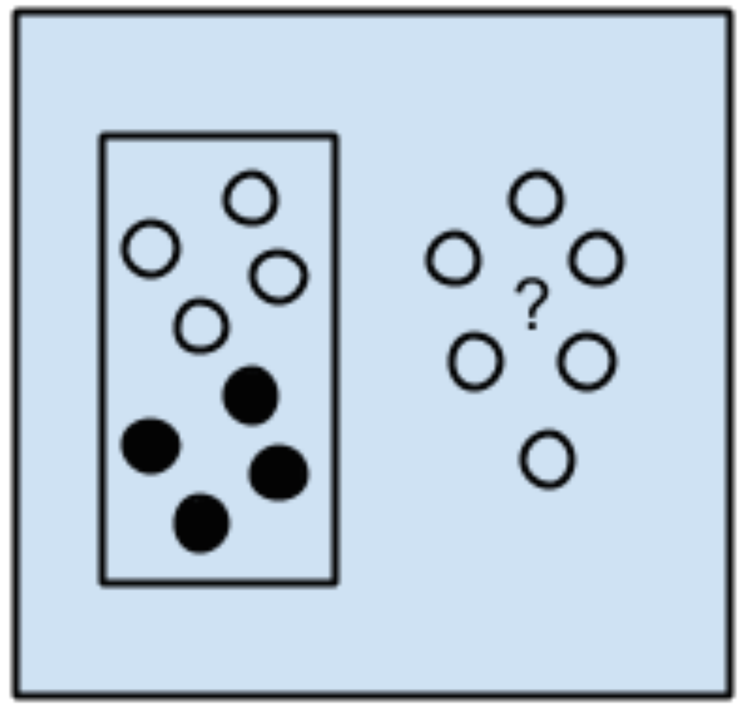
- Supervised Learning
  - " Learn **mapping** from input to output "
    - Classification
    - Regression / Ranking
- Unsupervised Learning
  - " Learn **patterns** in the input "
    - Clustering
    - Segmentation



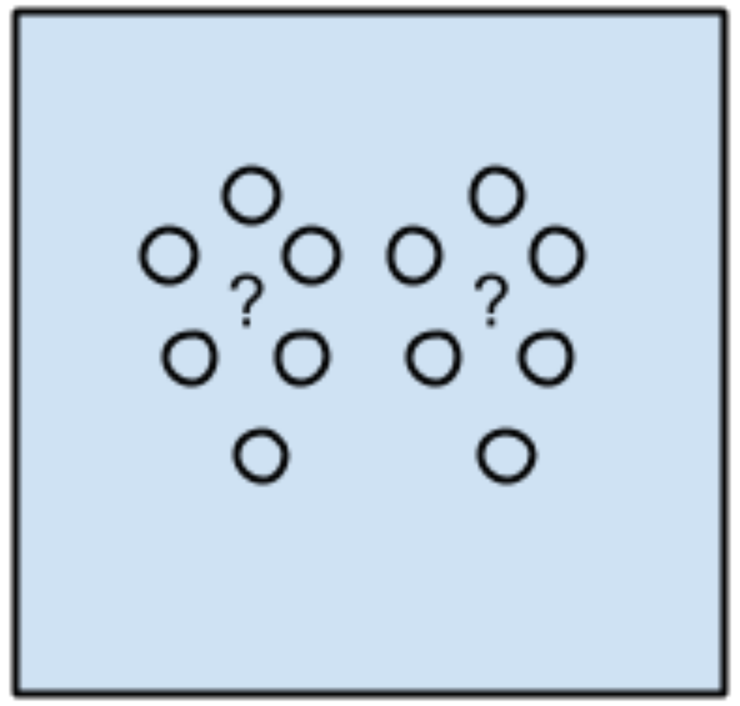
http://machinelearningmastery.com

Machine Learning

- Supervised Learning
  - " Learn **mapping** from input to output "
    - Classification
    - Regression / Ranking
- Unsupervised Learning
  - " Learn **patterns** in the input "
    - Clustering
    - Segmentation
- Reinforcement Learning / Deep Mind
  - " Learn to **act** from rewards in the absence of examples"
    - Decision process
    - Recommendation systems



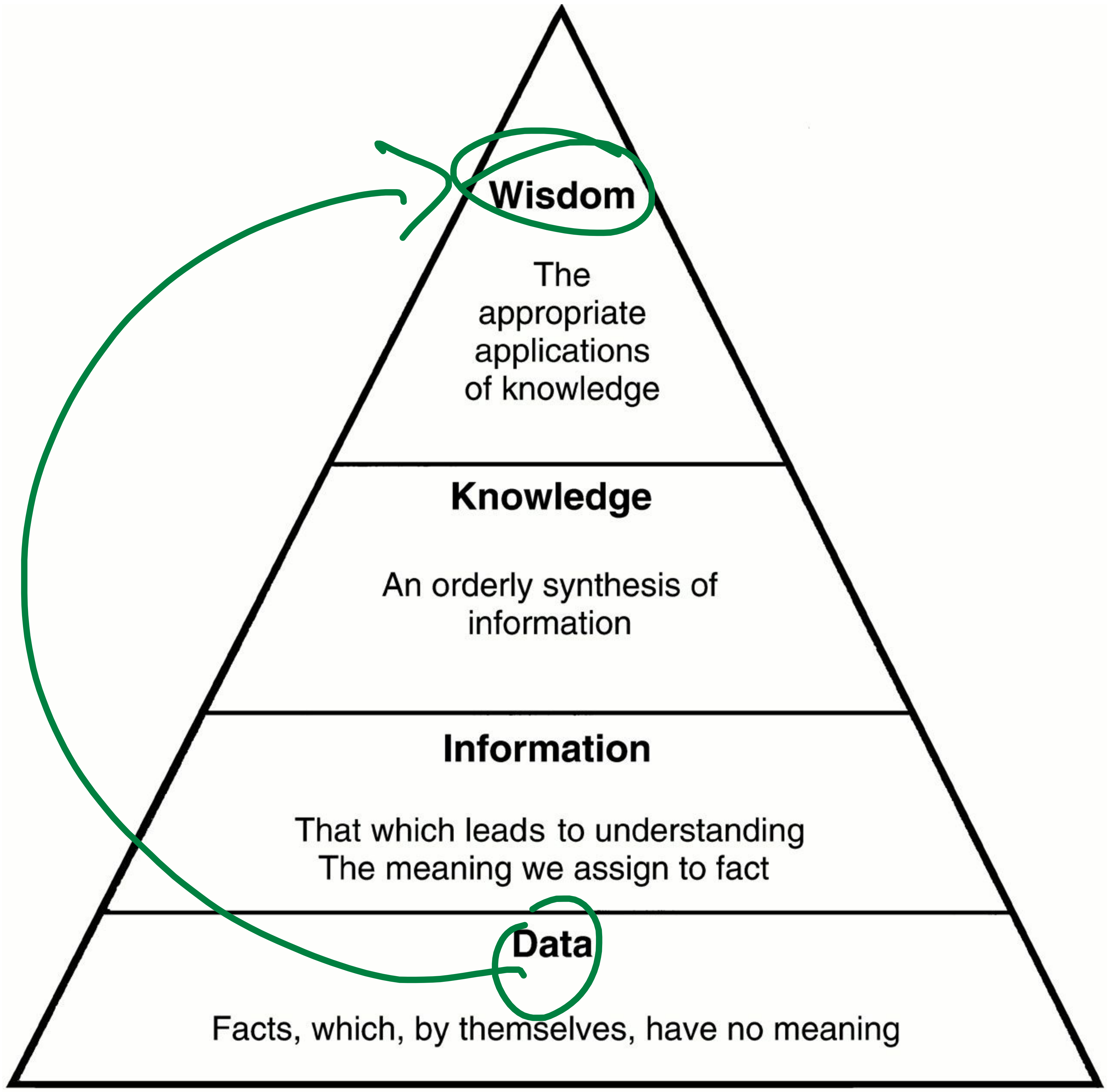
Supervised Learning Algorithms



Unsupervised Learning Algorithms

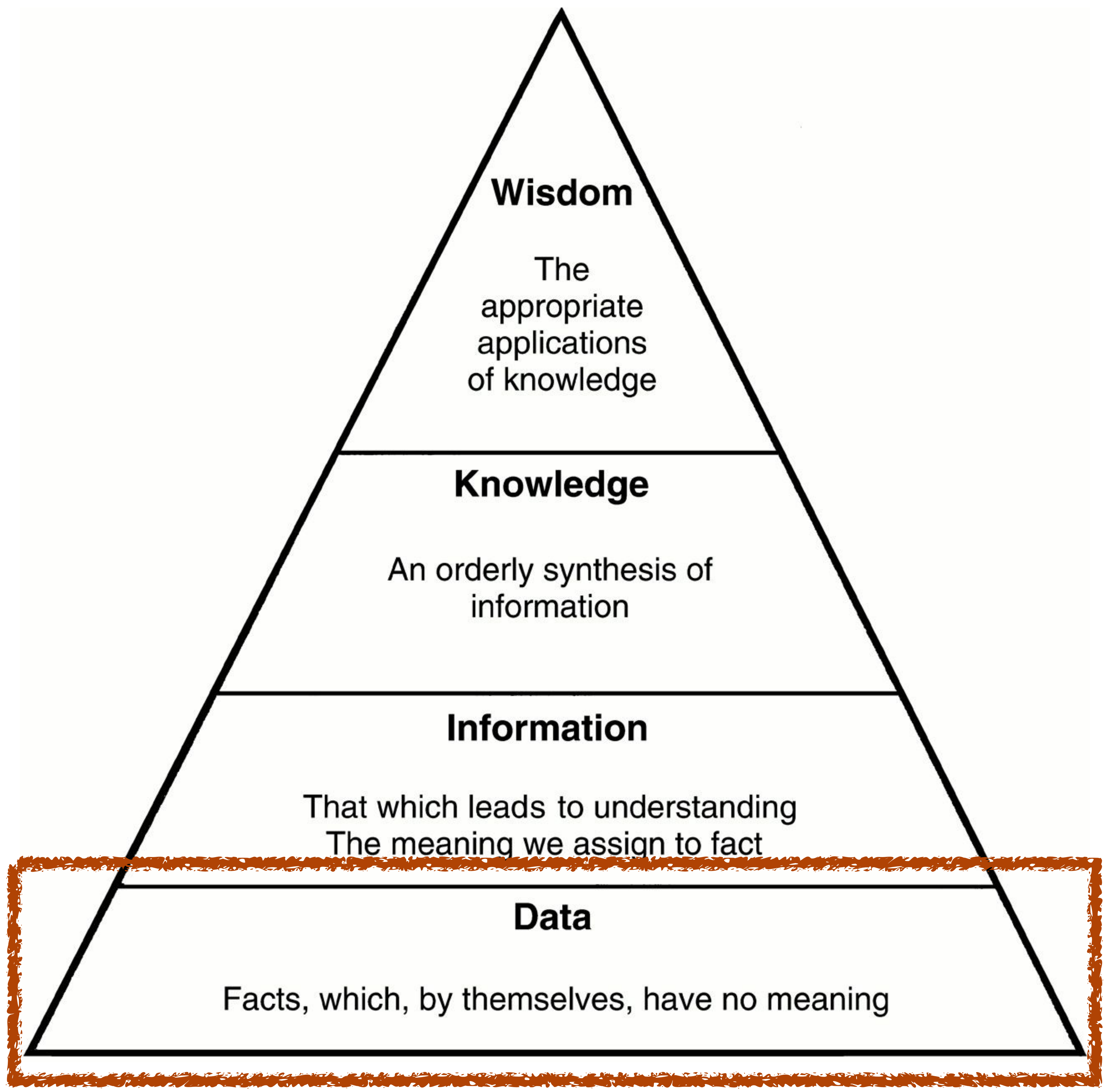
http://machinelearningmastery.com

The DIKW Pyramid



http://jbjs.org/content/82/6/888

The DIKW Pyramid

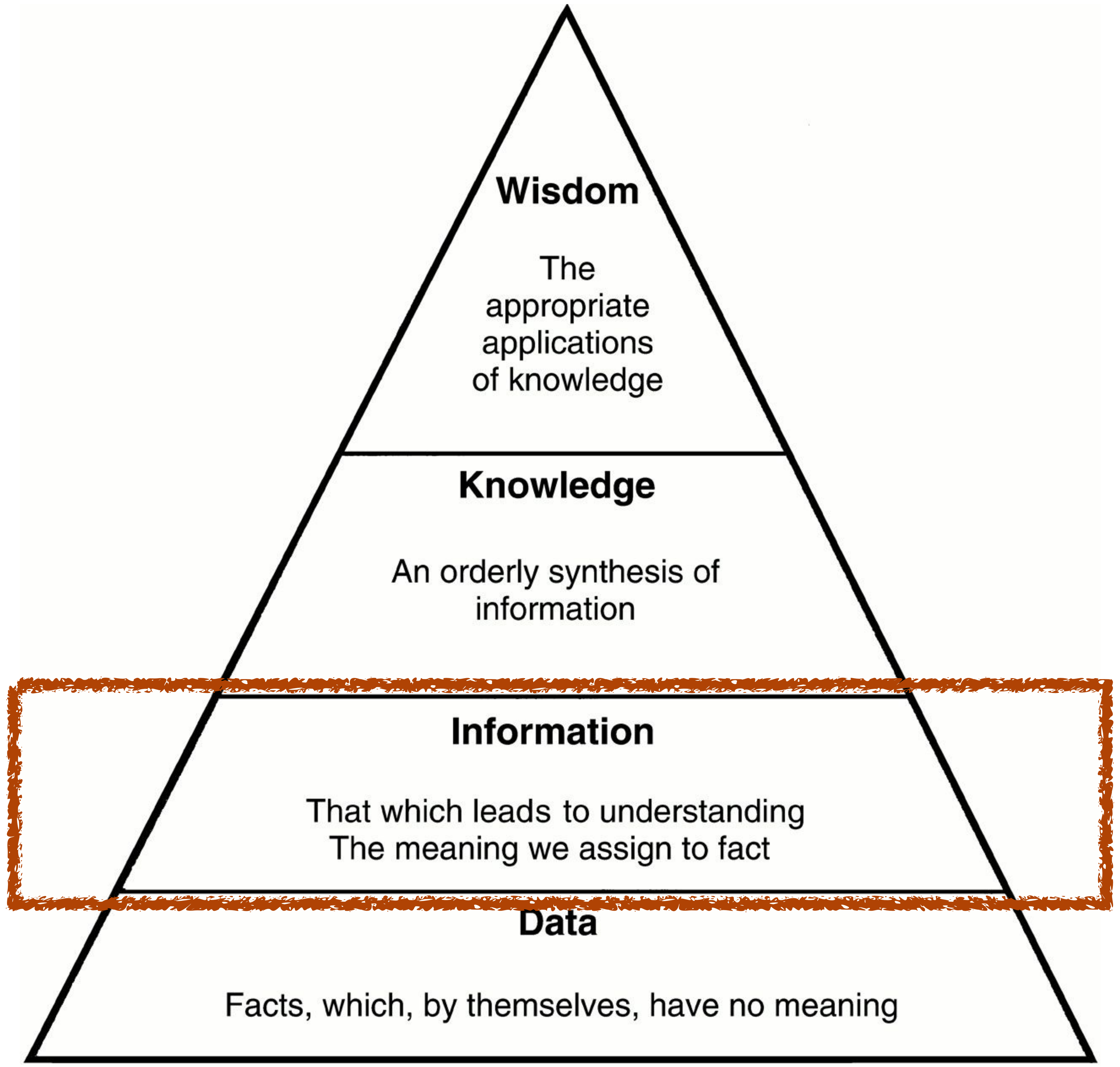


http://jbjs.org/content/82/6/888

Data

Facts, which, by themselves have no meaning

The DIKW Pyramid



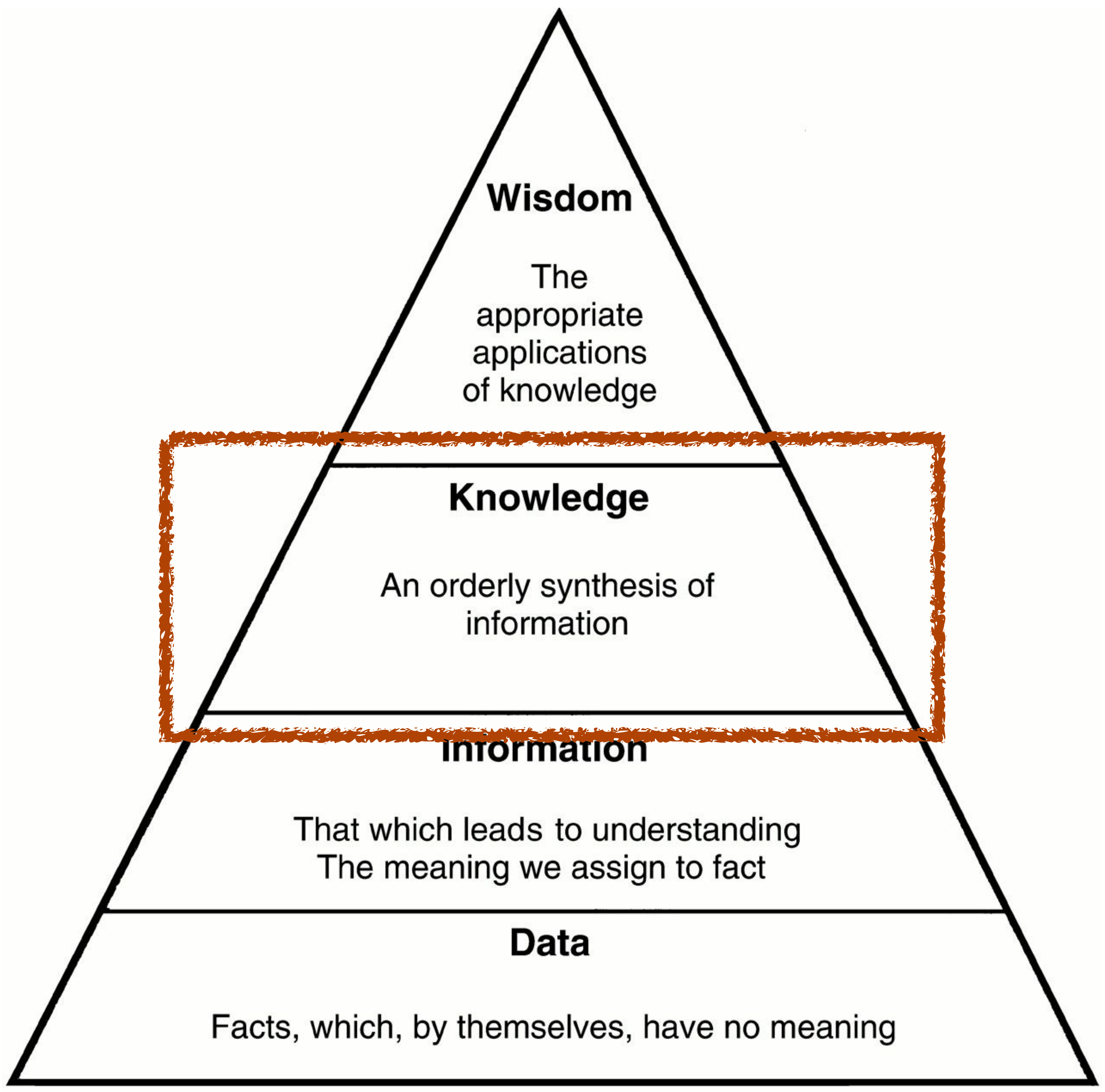
http://jbjs.org/content/82/6/888

Information

That which leads to understanding

The meaning we assign to facts

The DIKW Pyramid

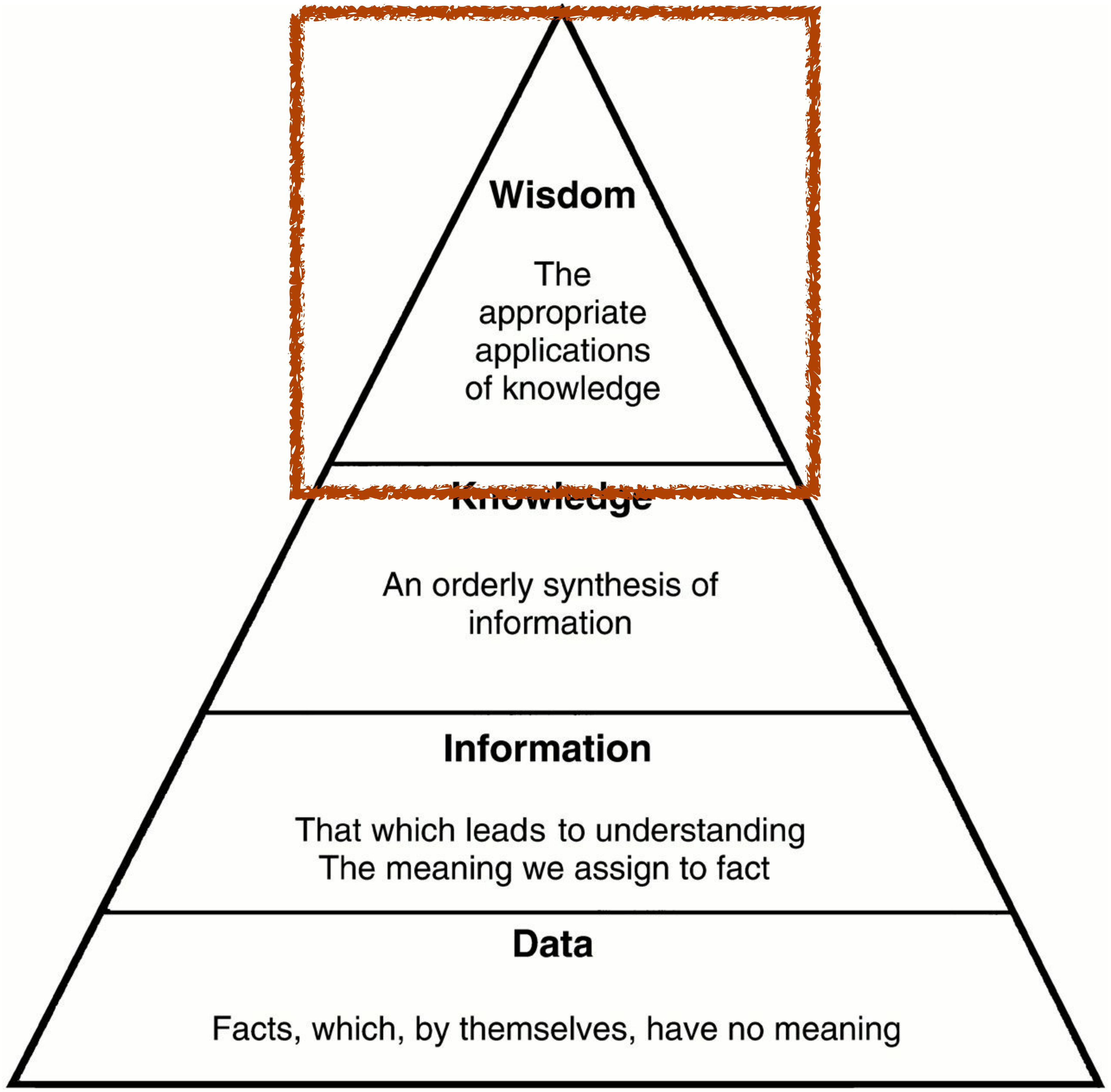


http://jbjs.org/content/82/6/888

Knowledge

An orderly synthesis of information

The DIKW Pyramid

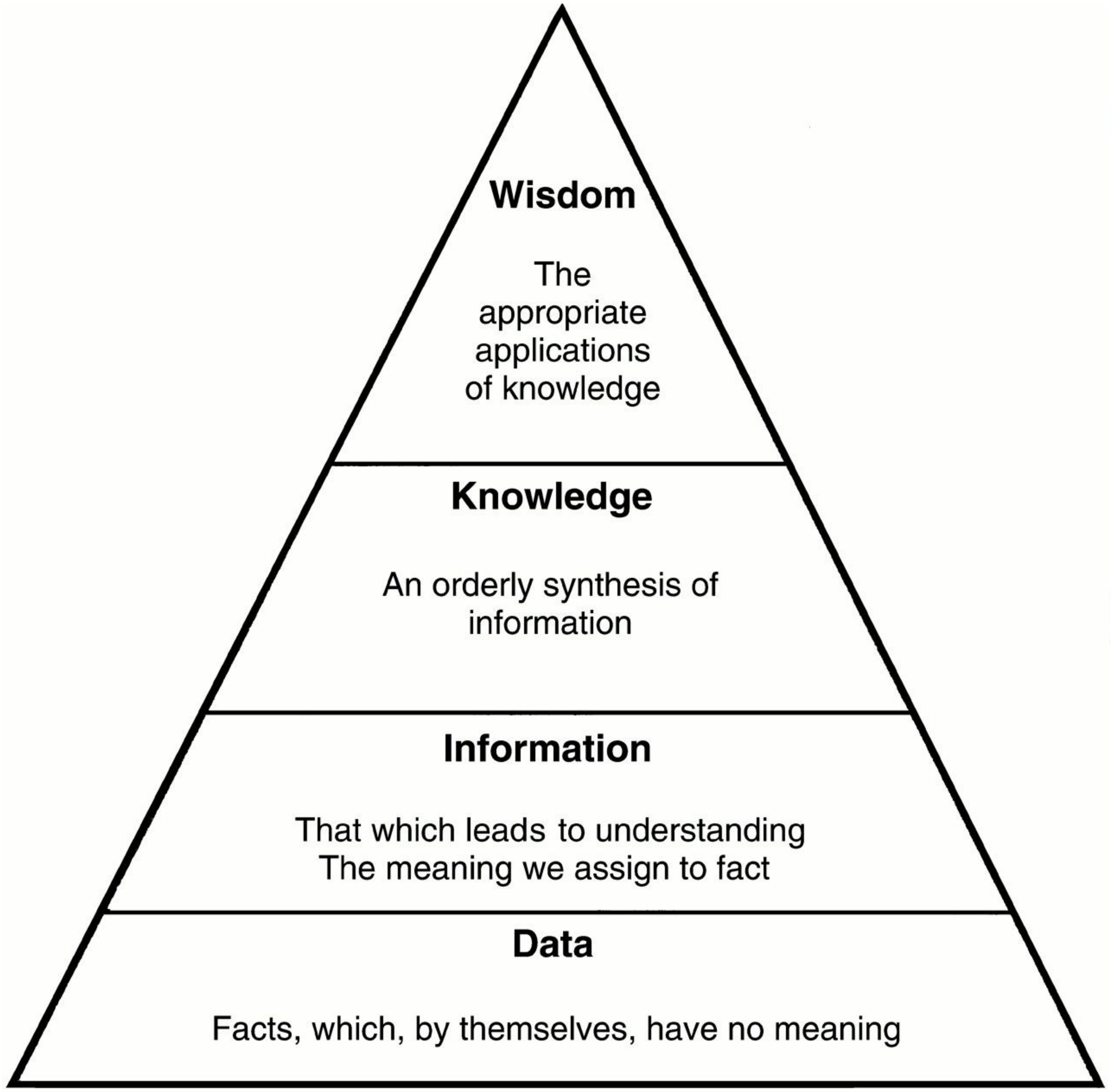


http://jbjs.org/content/82/6/888

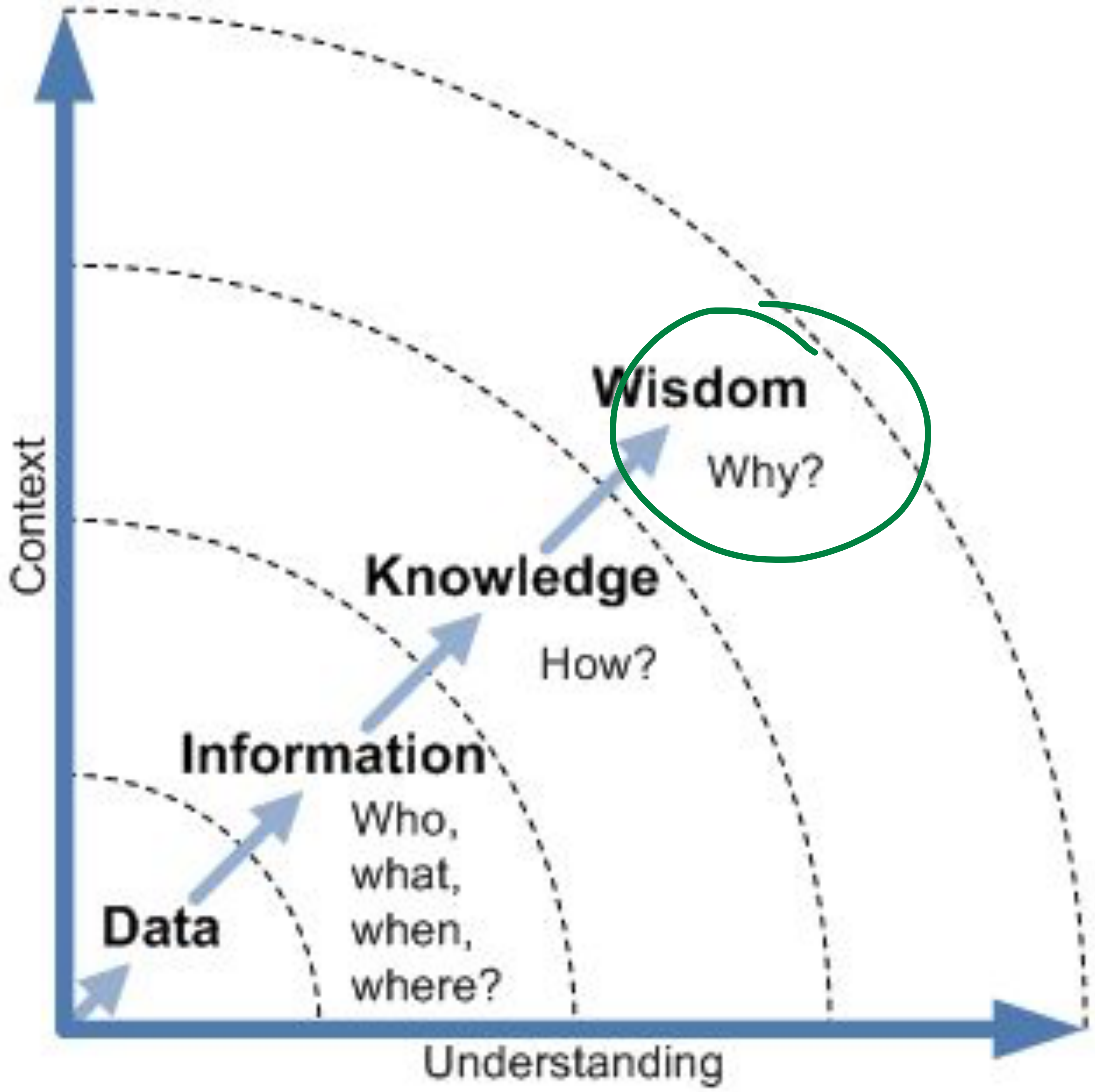
Wisdom (Insight)

The appropriate applications of knowledge

The DIKW Pyramid



http://jbjs.org/content/82/6/888



http://www.theitsmreview.com

Goal of Machine Learning

- A good ML algorithm / model should provide a good **generalization**

Goal of Machine Learning

- A good ML algorithm / model should provide a good **generalization**
- Consistency with data is a necessary but not a sufficient condition

Goal of Machine Learning

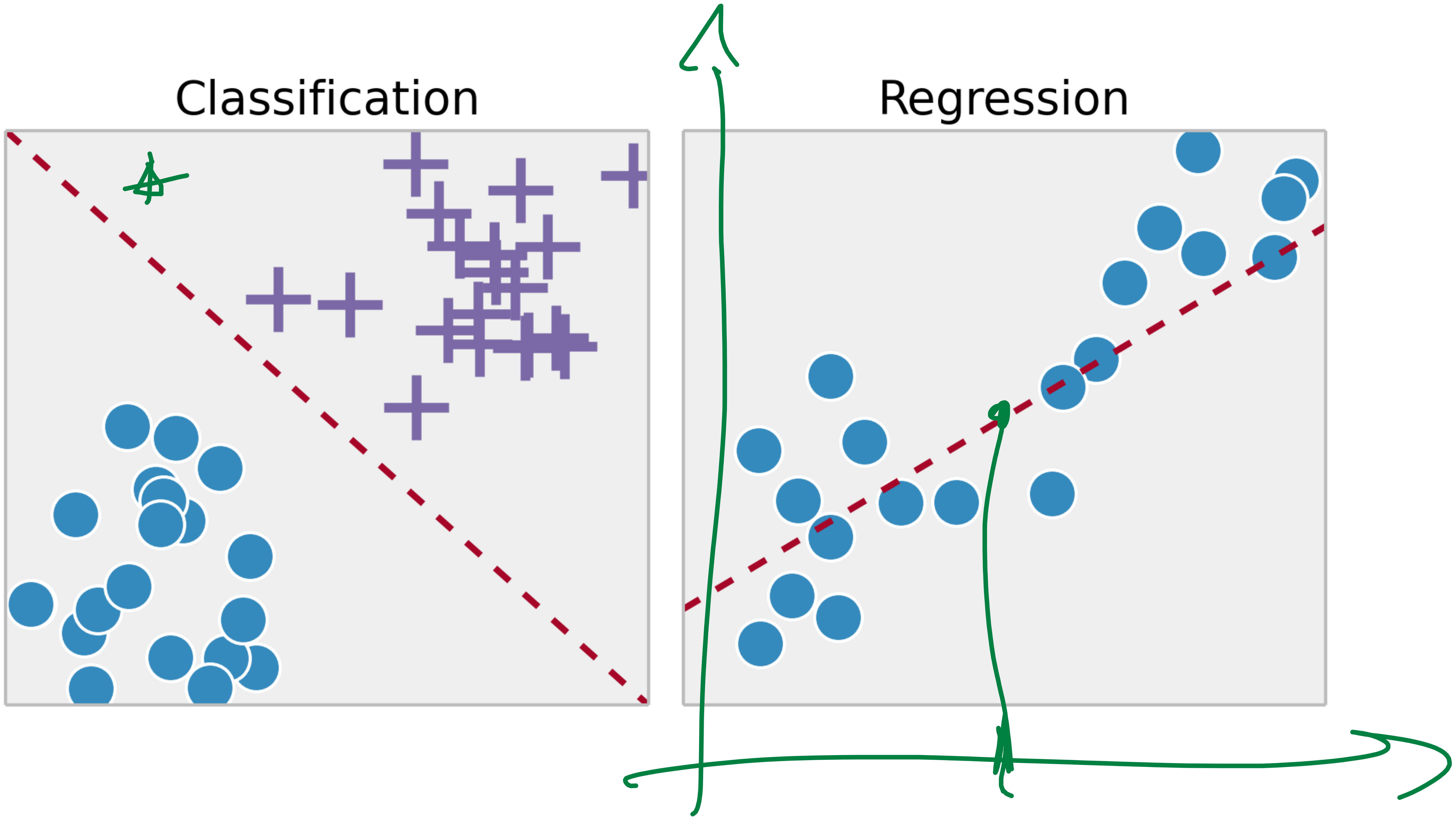
- A good ML algorithm / model should provide a good **generalization**
- Consistency with data is a necessary but not a sufficient condition
- Important! All models/algorithms have their own set of assumptions: **Inductive bias**

Goal of Machine Learning

- A good ML algorithm / model should provide a good generalization
- Consistency with data is a necessary but not a sufficient condition
- Important! All models/algorithms have their own set of assumptions: Inductive bias

1. Restriction bias: constrain learning to certain models
2. Preference bias: guide learning to prefer certain models over others

Supervised Learning



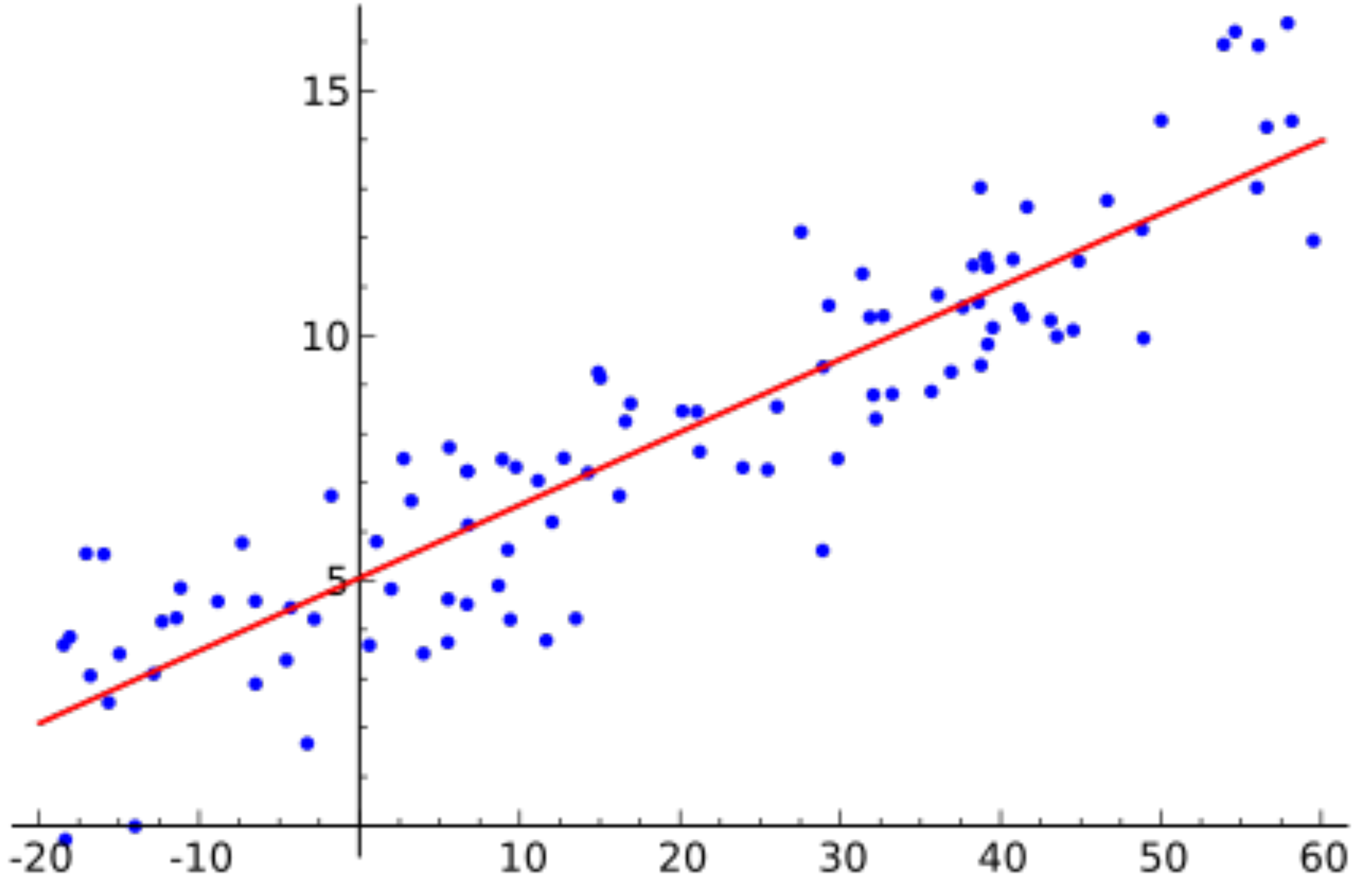
http://ipython-books.github.io/images/ml.png

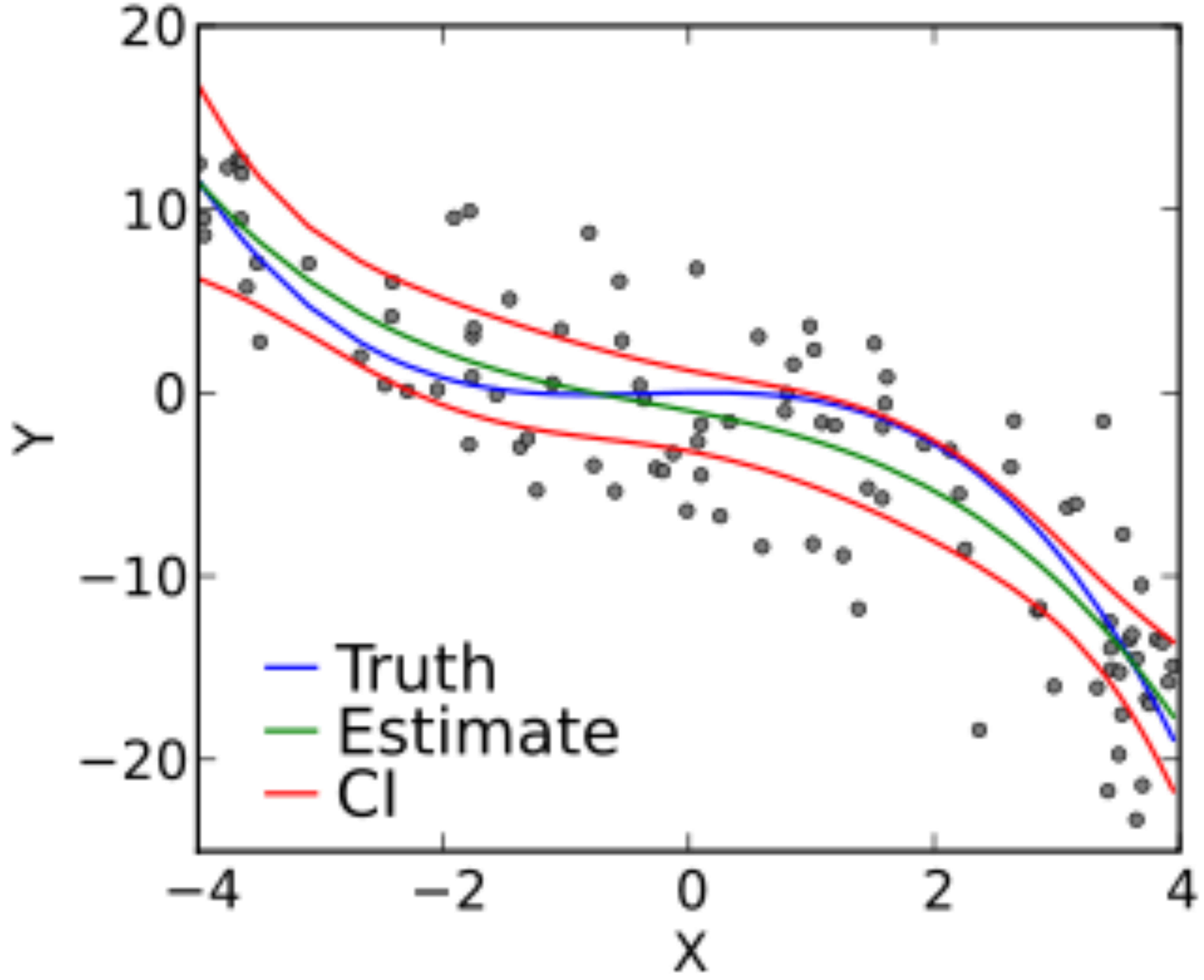
Supervised Learning - Algorithms

- Linear Regression
- Logistic Regression
- KNN (*k* nearest neighbor)
- Naïve Bayes Classifier
- Decision Trees
- Support Vector Machines
- Ensemble Methods

Linear Regression

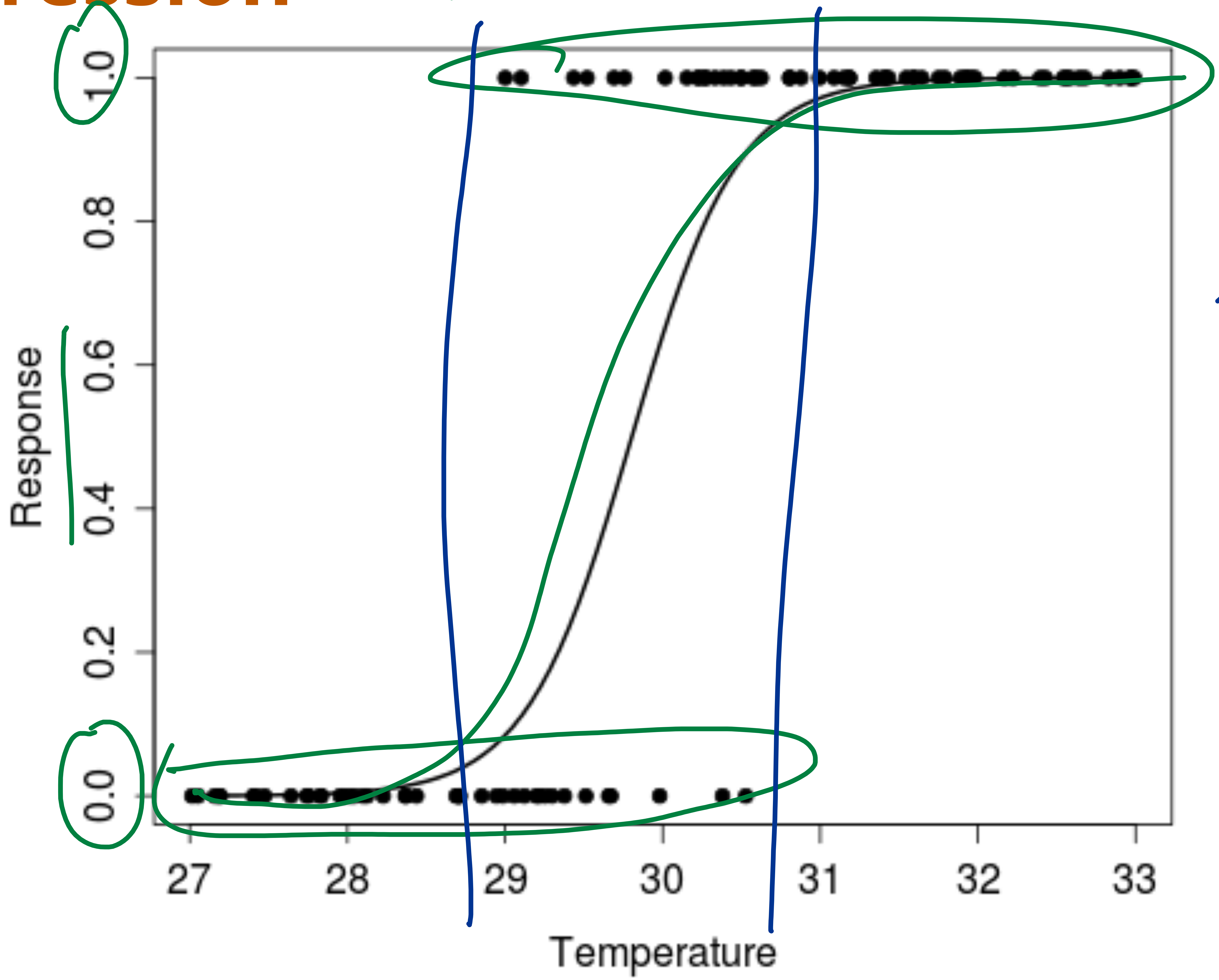
$$y = \underline{\underline{\beta}} \cdot \underline{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

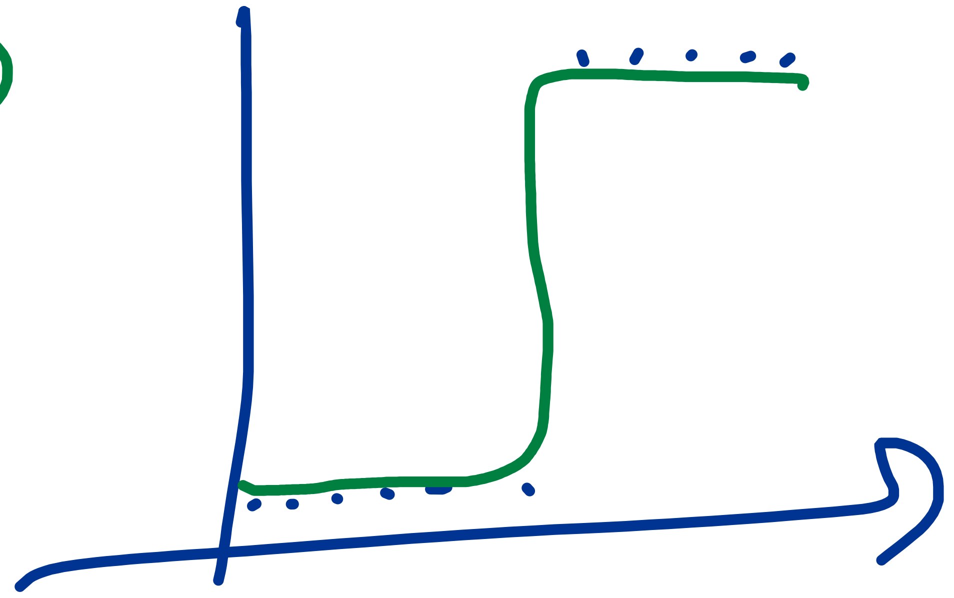




Logistic Regression → probabilities

$p \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots +$





k Nearest Neighbors

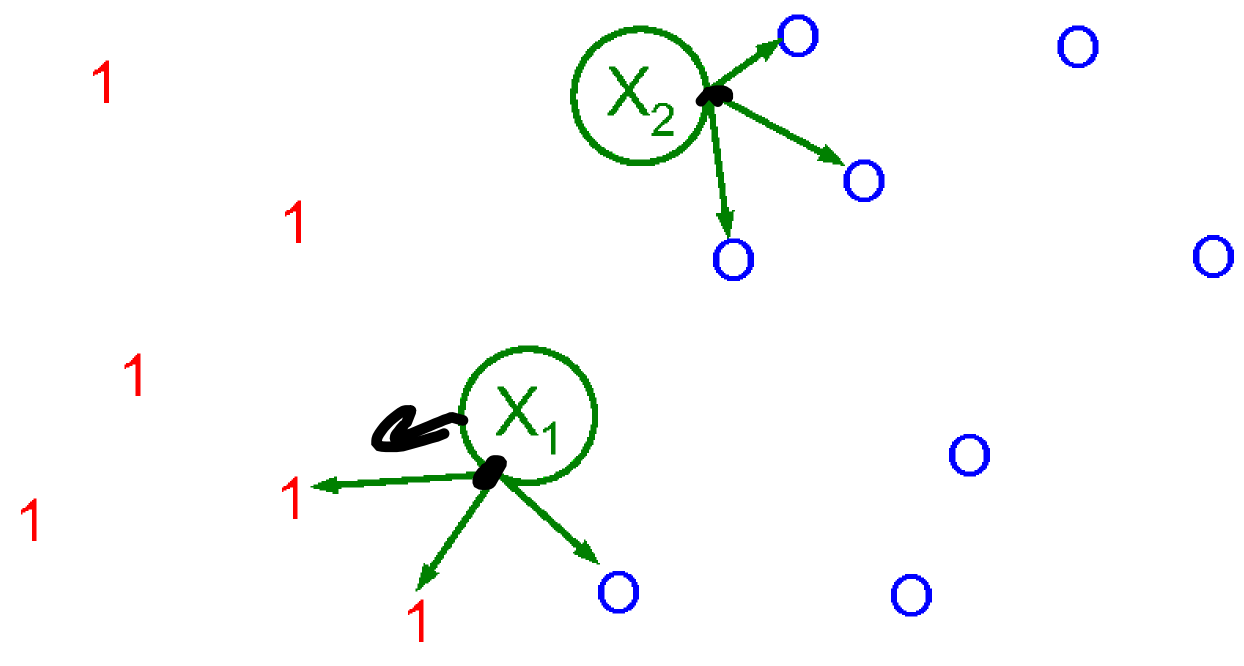
K-NN

K=?

k=1 ... N

data points

Example 1 (k=3)



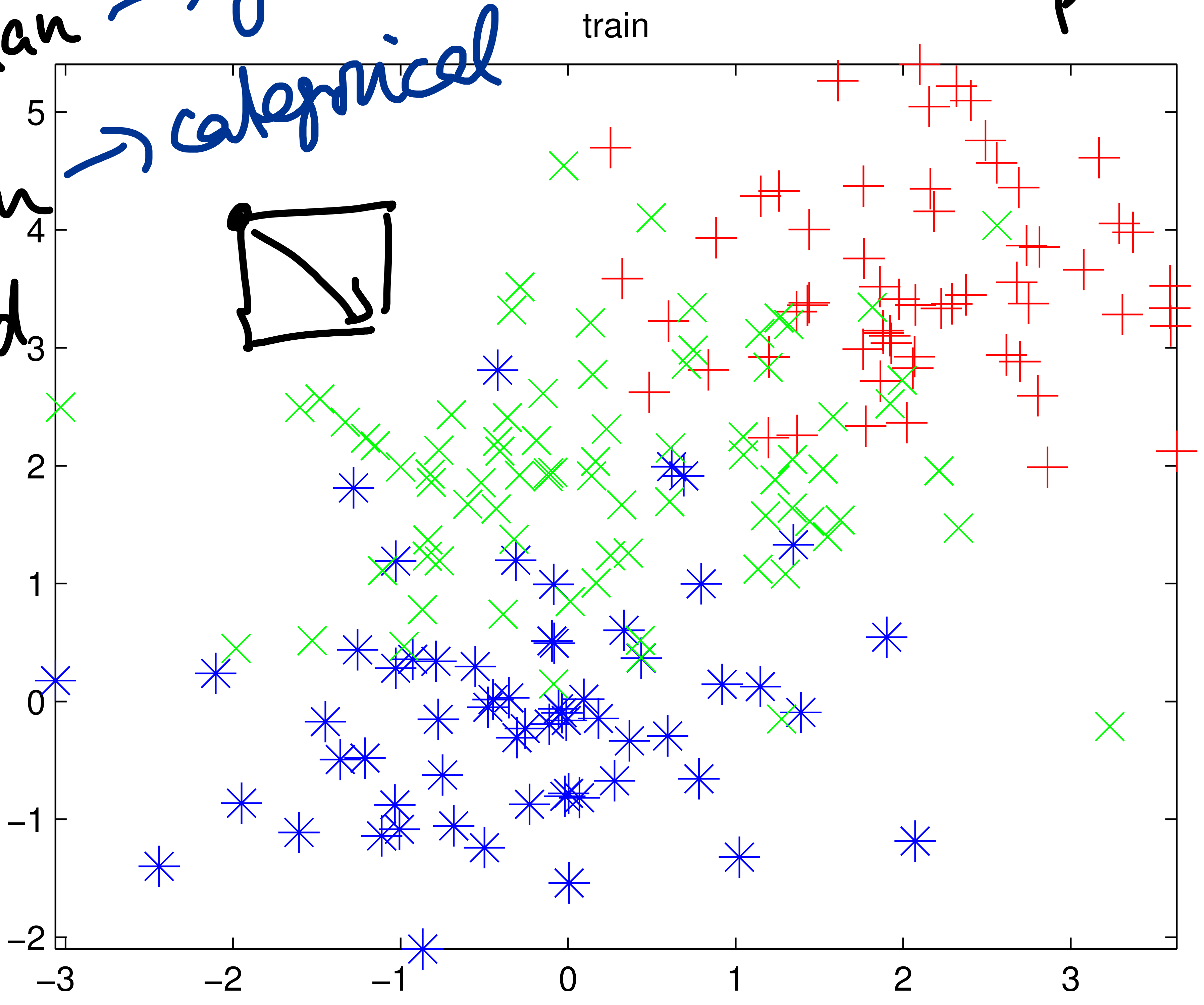
Example 2 (k=10)

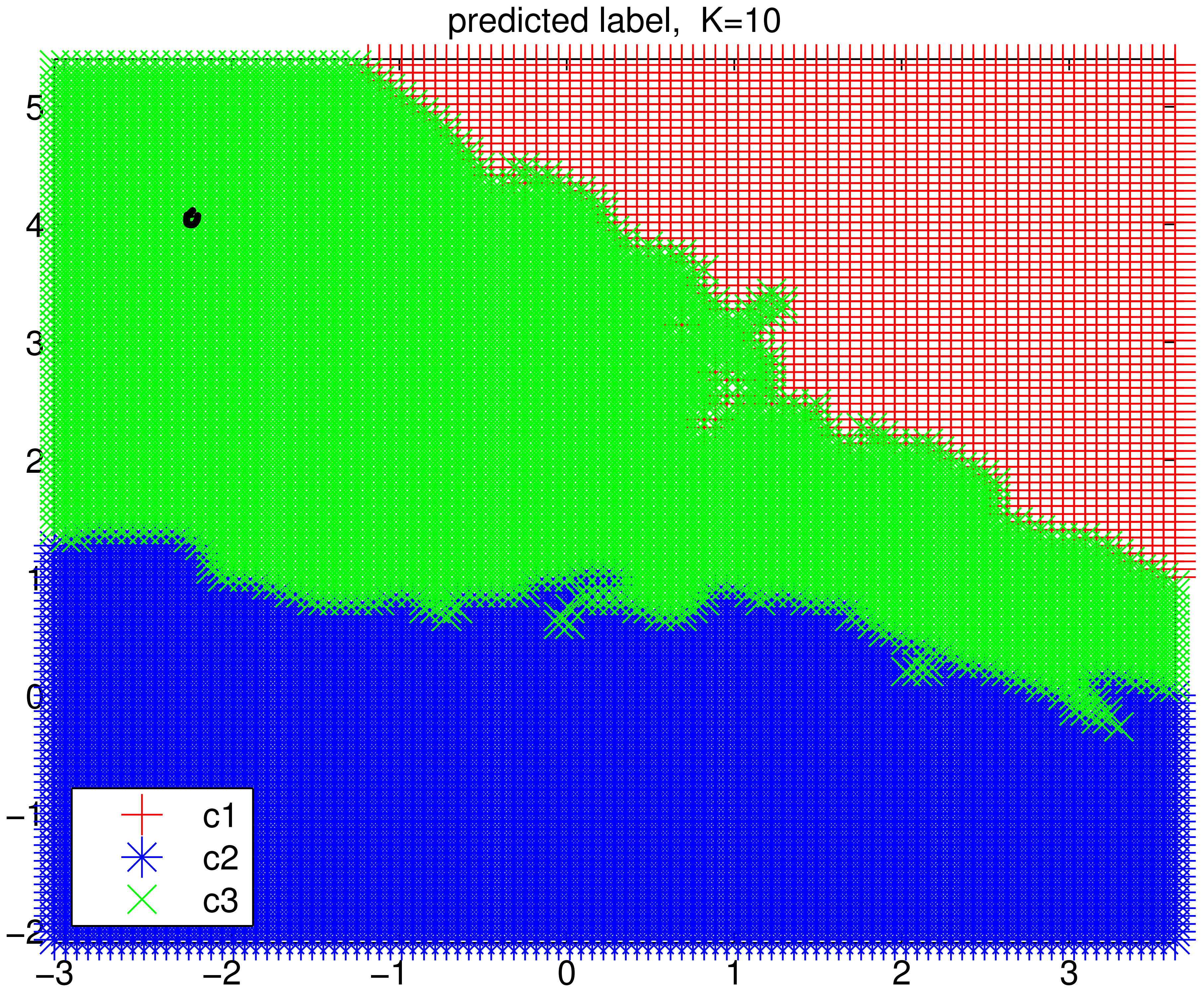
· Euclidian → geometric

· Manhattan → categorical

· Jaccard

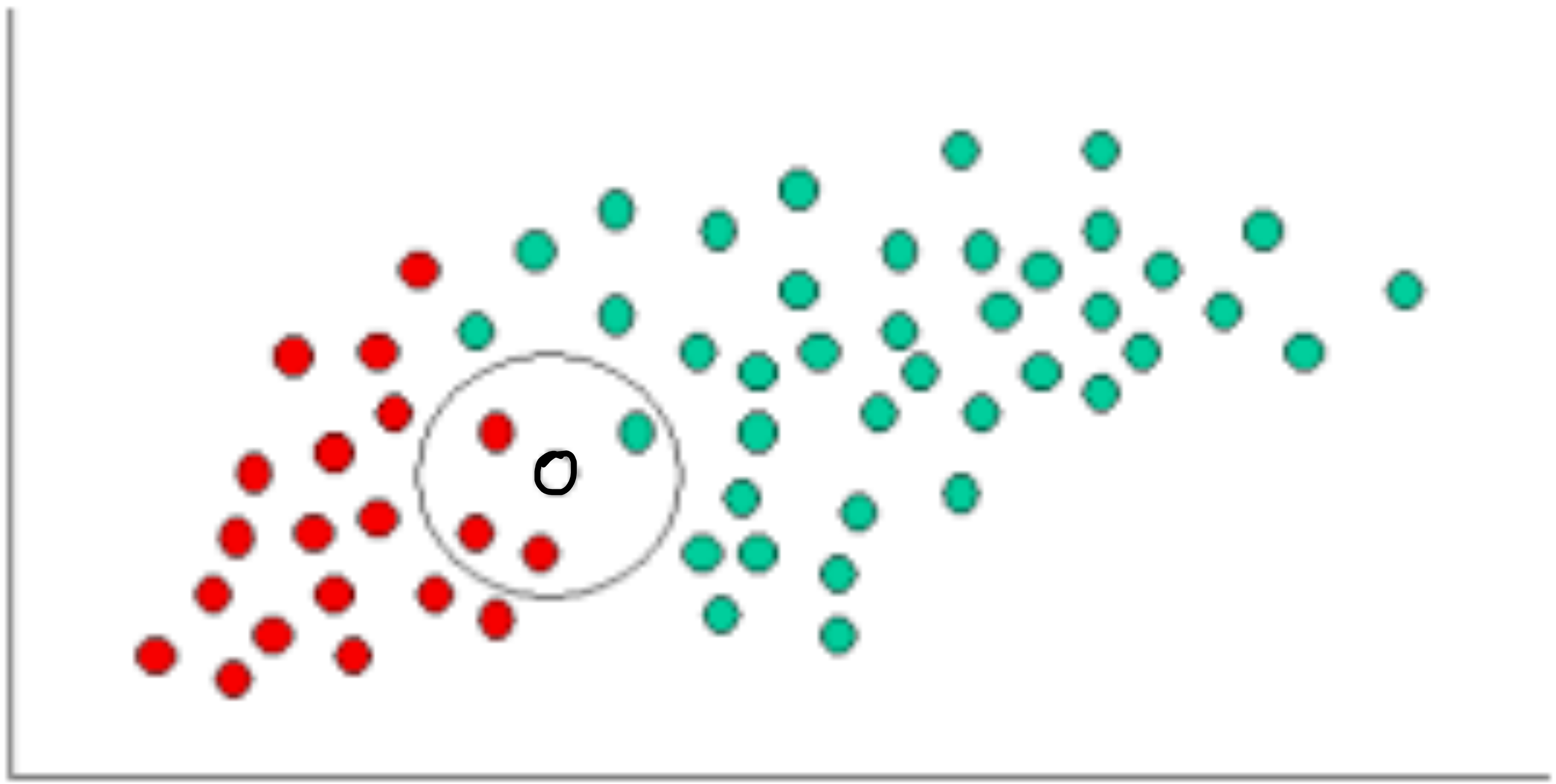
→ Text analysis





Naïve Bayes Classifier

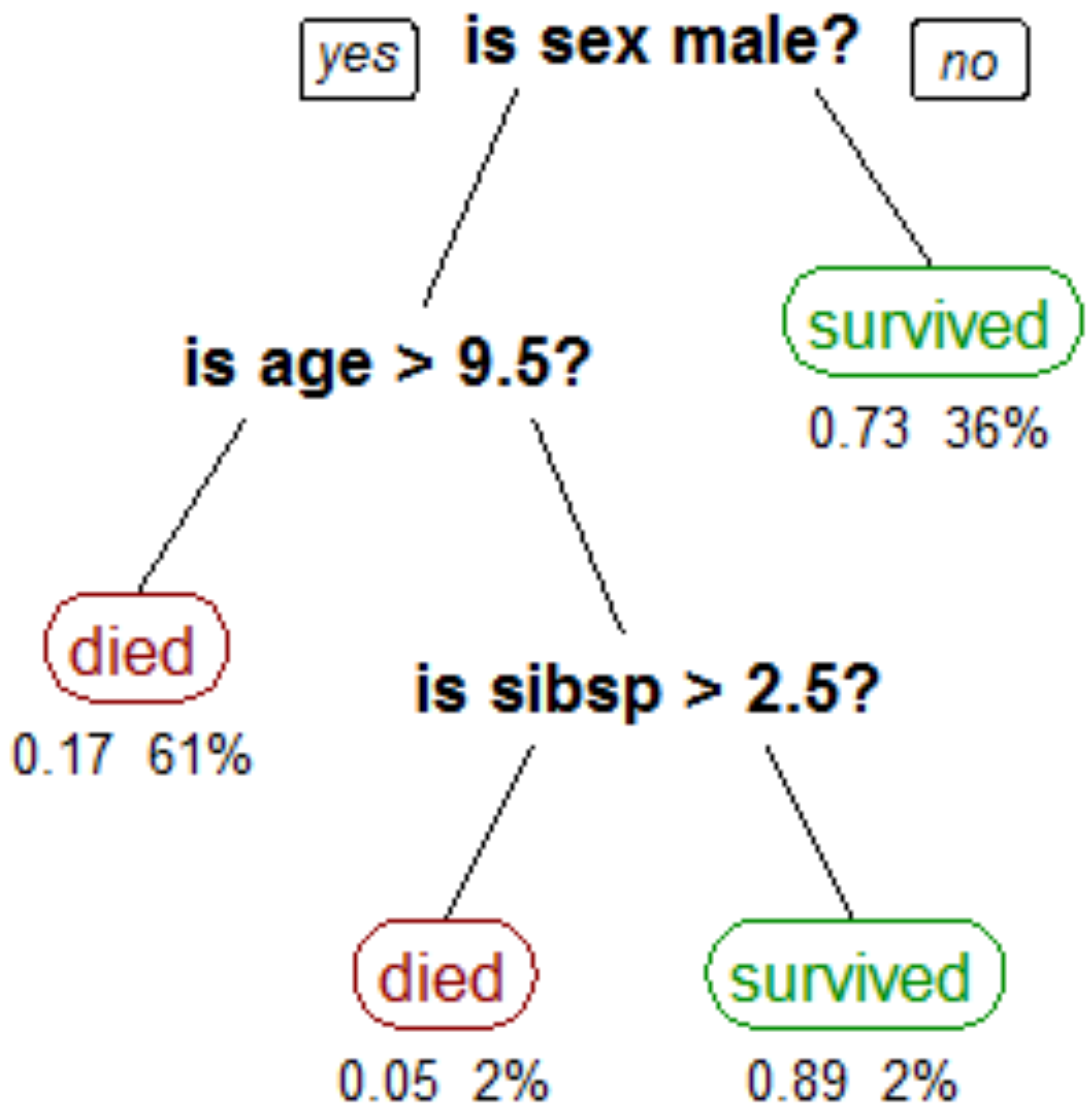
Probabilities

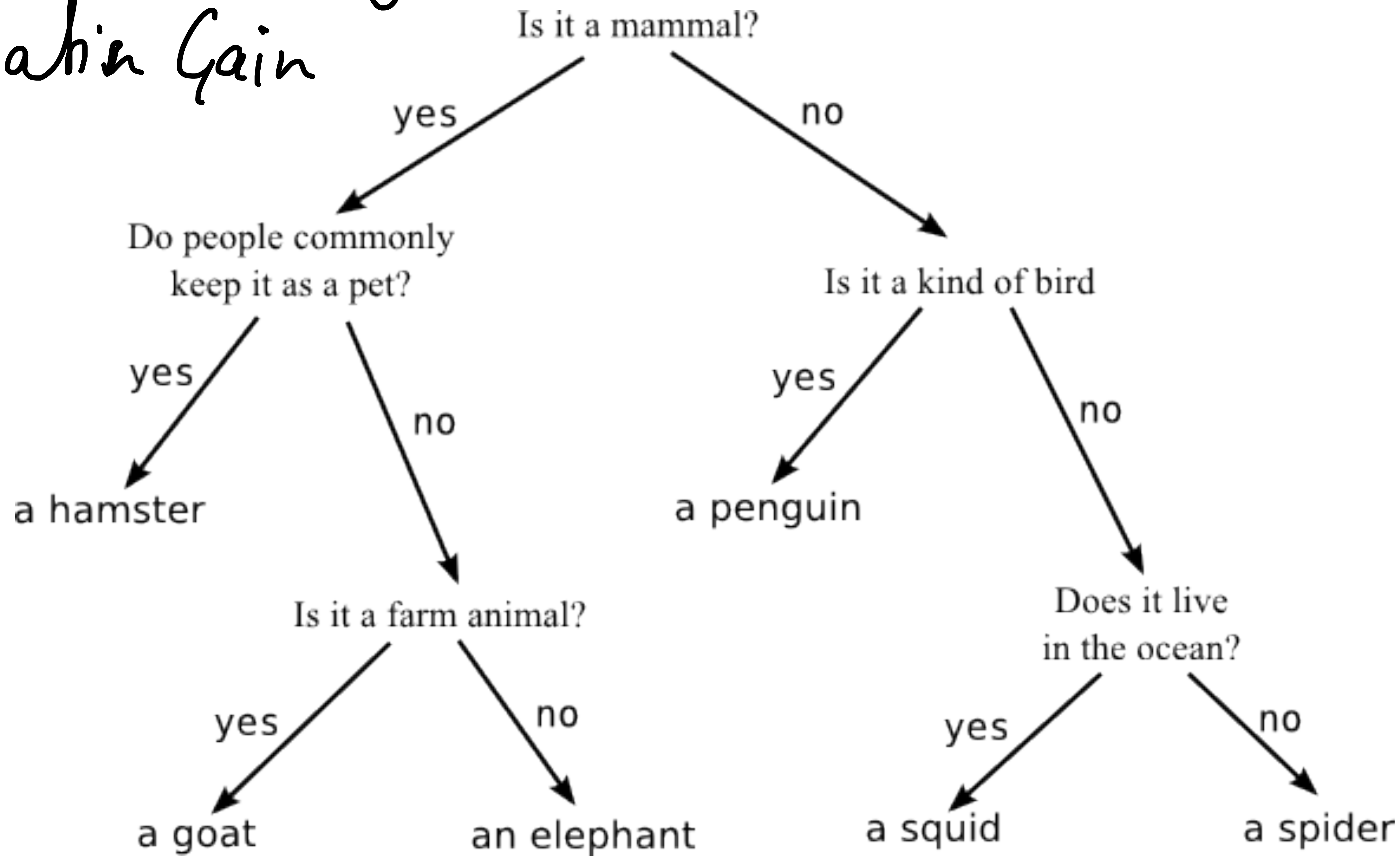


$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Decision Trees

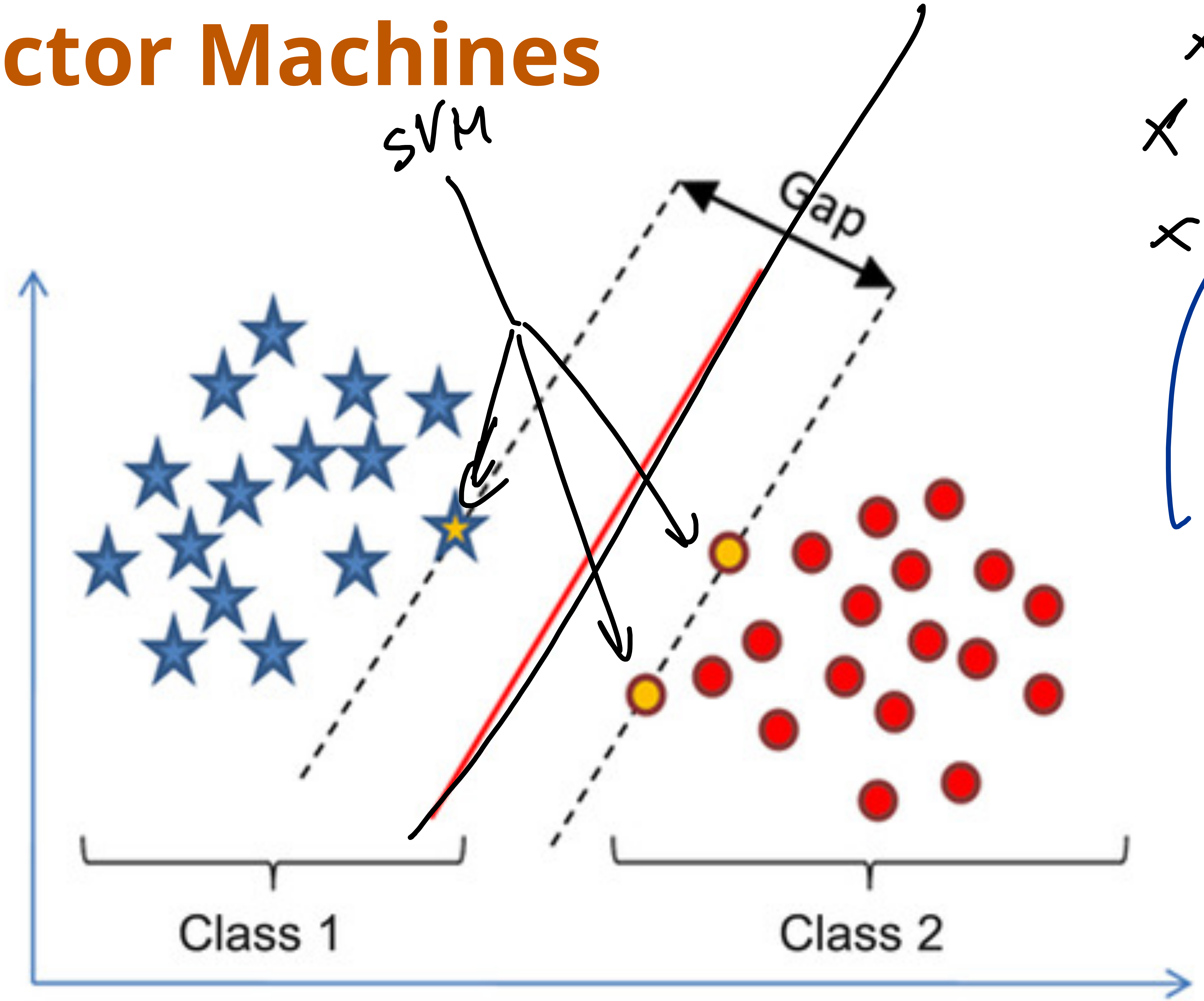
Information Theory
Information Gain





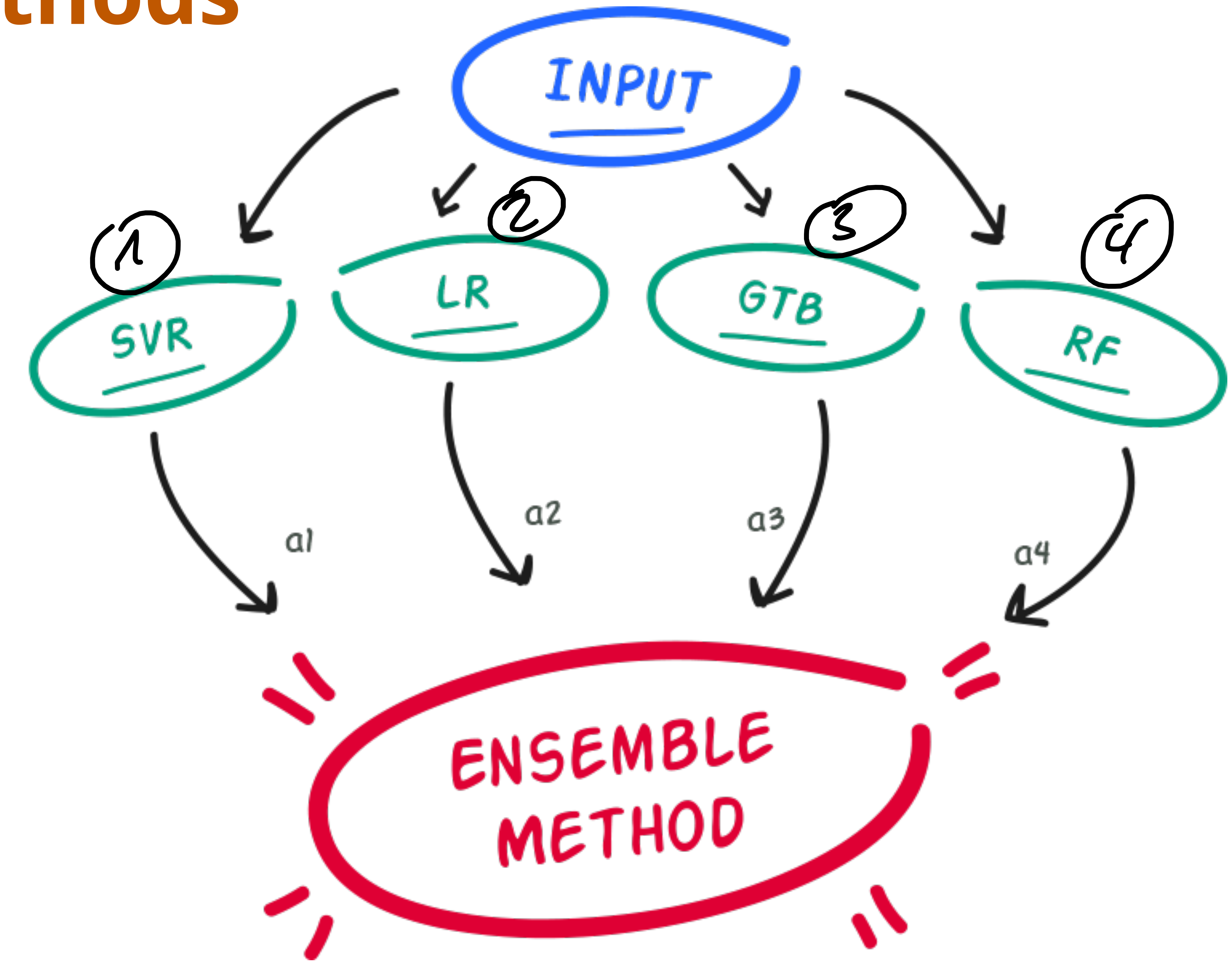
Support Vector Machines

SVM



Ensemble Methods

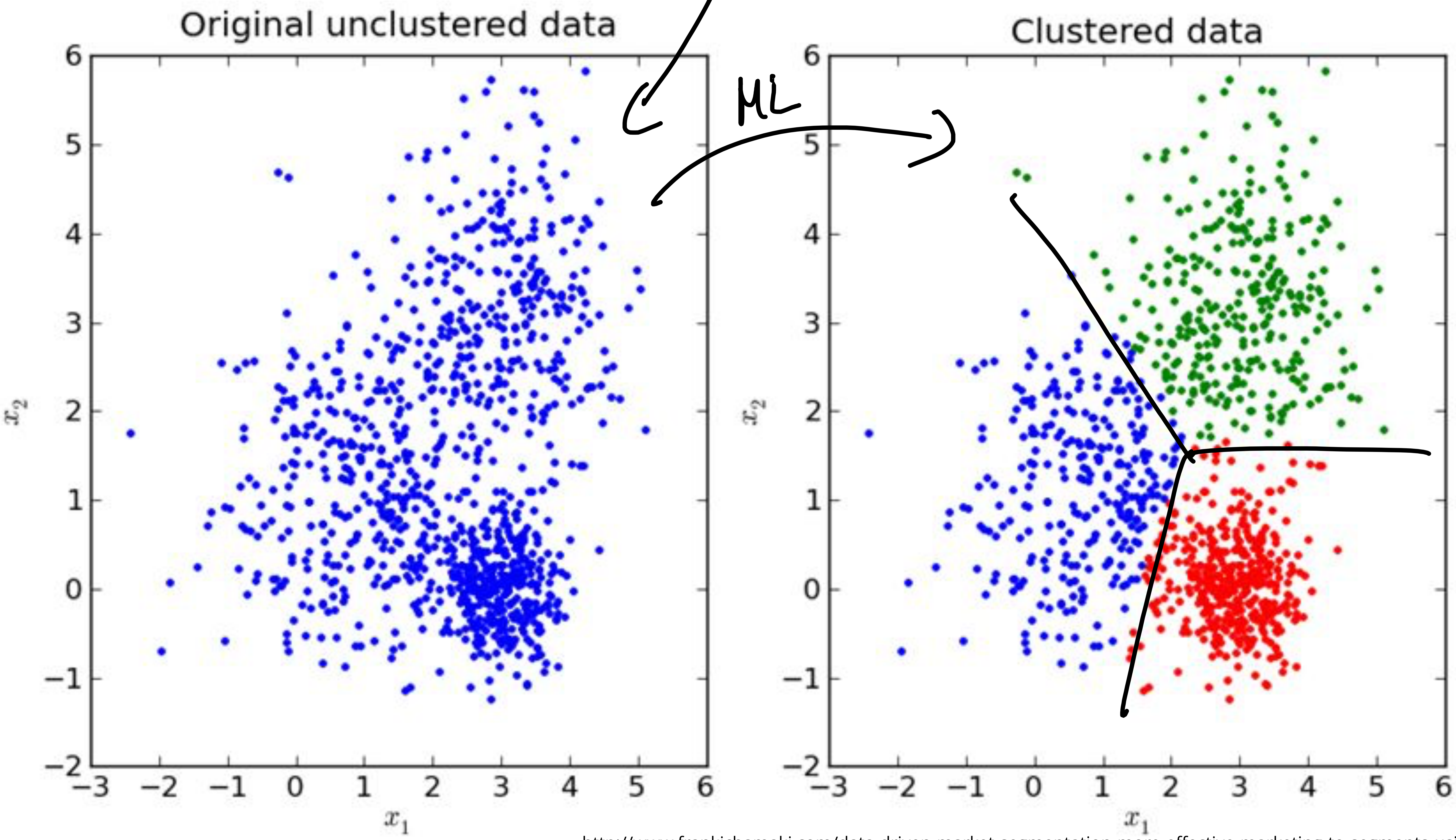
No Free Lunch Theorem



Unsupervised Learning

input

ML



http://www.frankichamaki.com/data-driven-market-segmentation-more-effective-marketing-to-segments-using-ai/

Unsupervised Learning - Algorithm

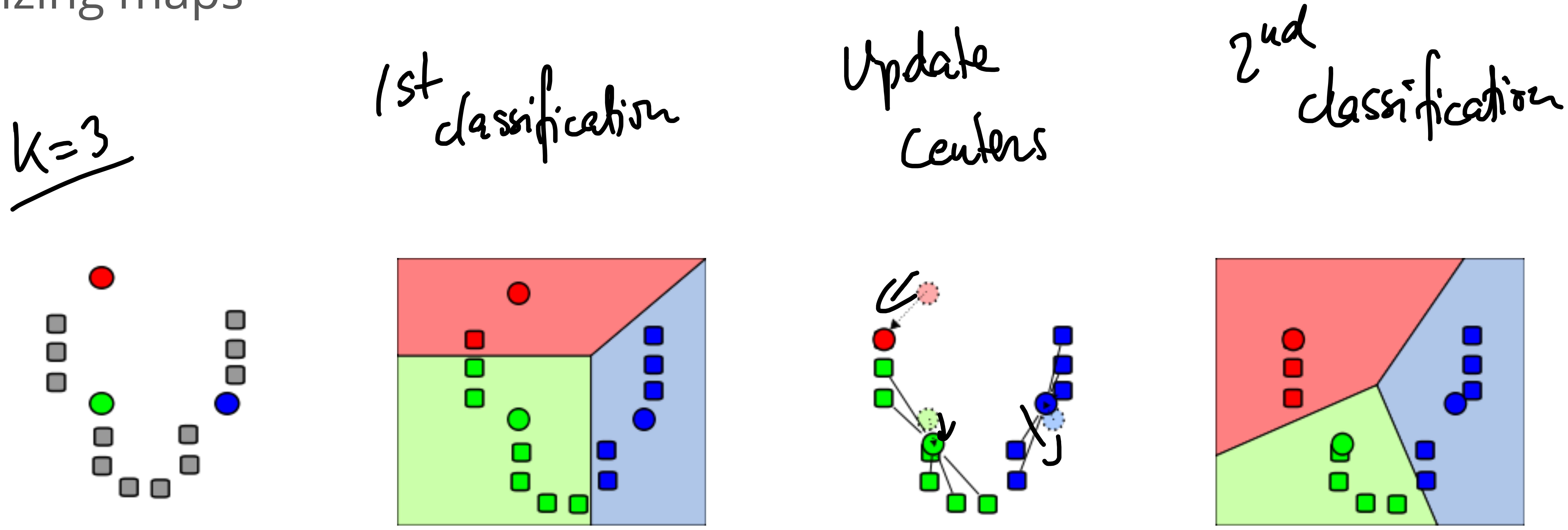
- k-means clustering
- (self organizing map)

- Similar to Human & Animal learning
- No need for data labeling
- Little information

When we're learning to see, nobody's telling us what there might answers are - we just look. (Geoffrey Hinton, 1996)

k-means clustering

- k-means clustering
- Self organizing maps



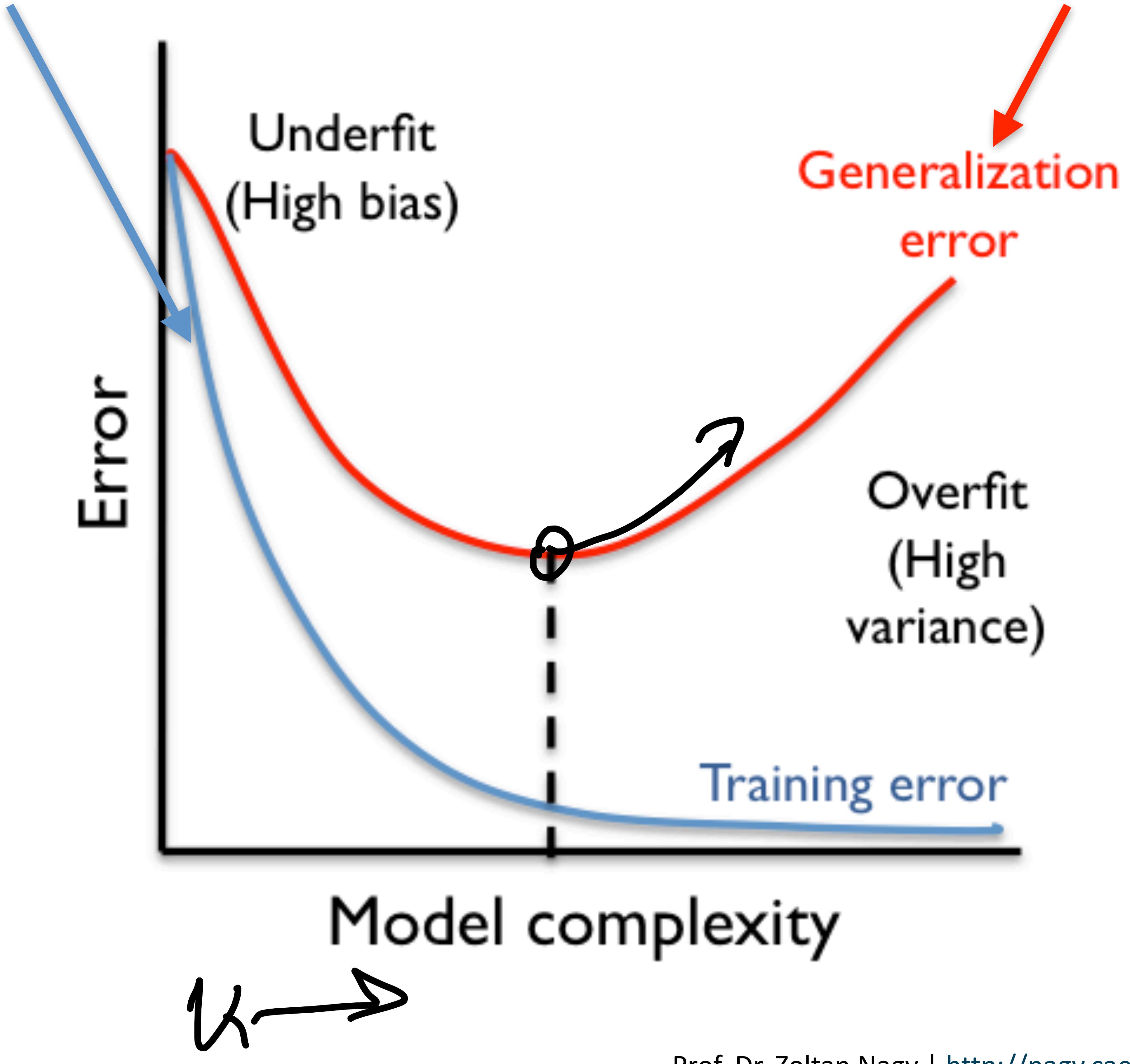
http://ipython-books.github.io/images/ml.png

Bias-variance tradeoff in ML

- *Bias*: error due to assumptions
  - High bias = underfitting
- Variance: *error* due to sensitivity of model
  - High variance = overfitting

Approach: Split input data

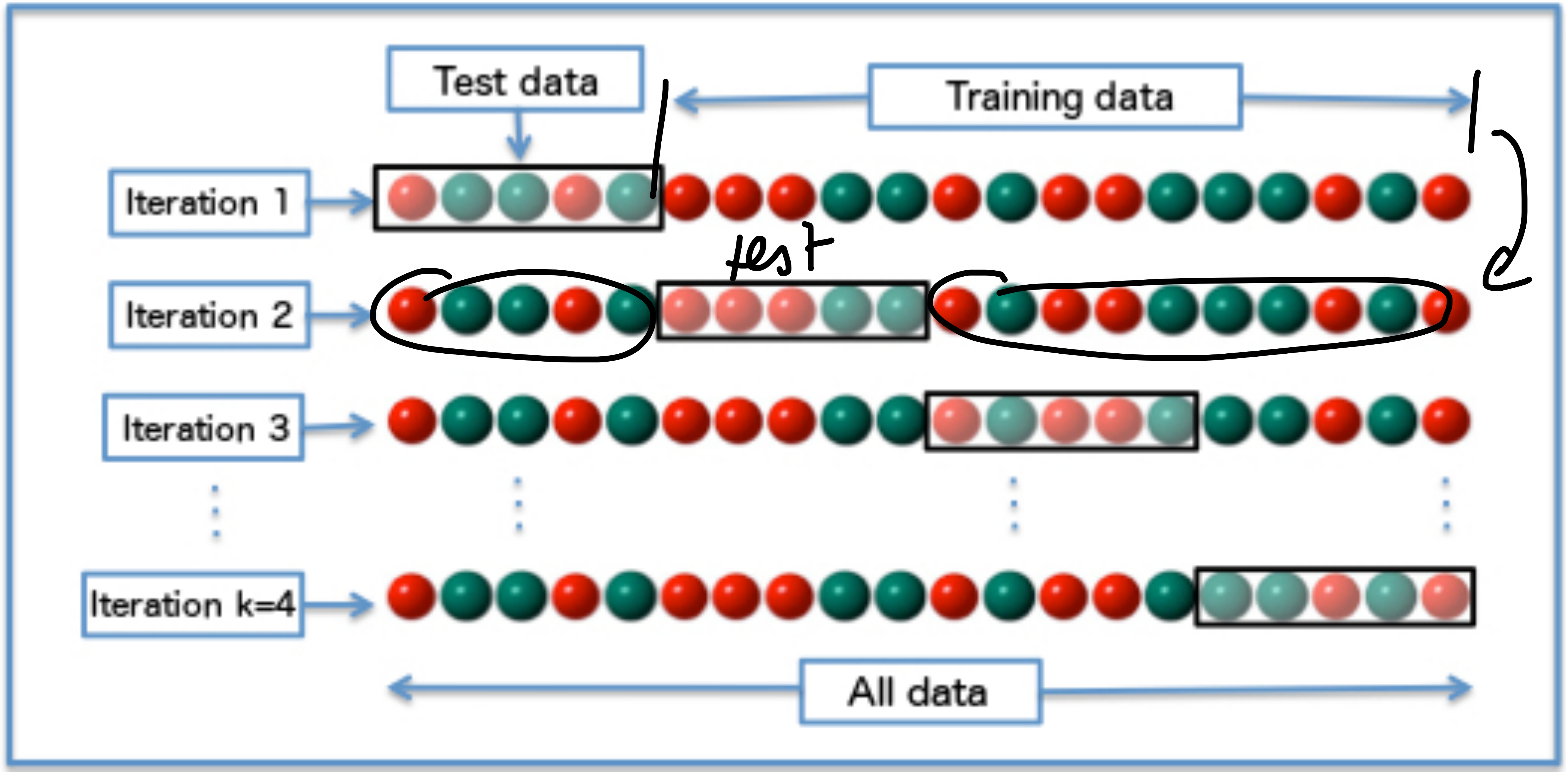
Training set (70 %)	Validation set (30 %)



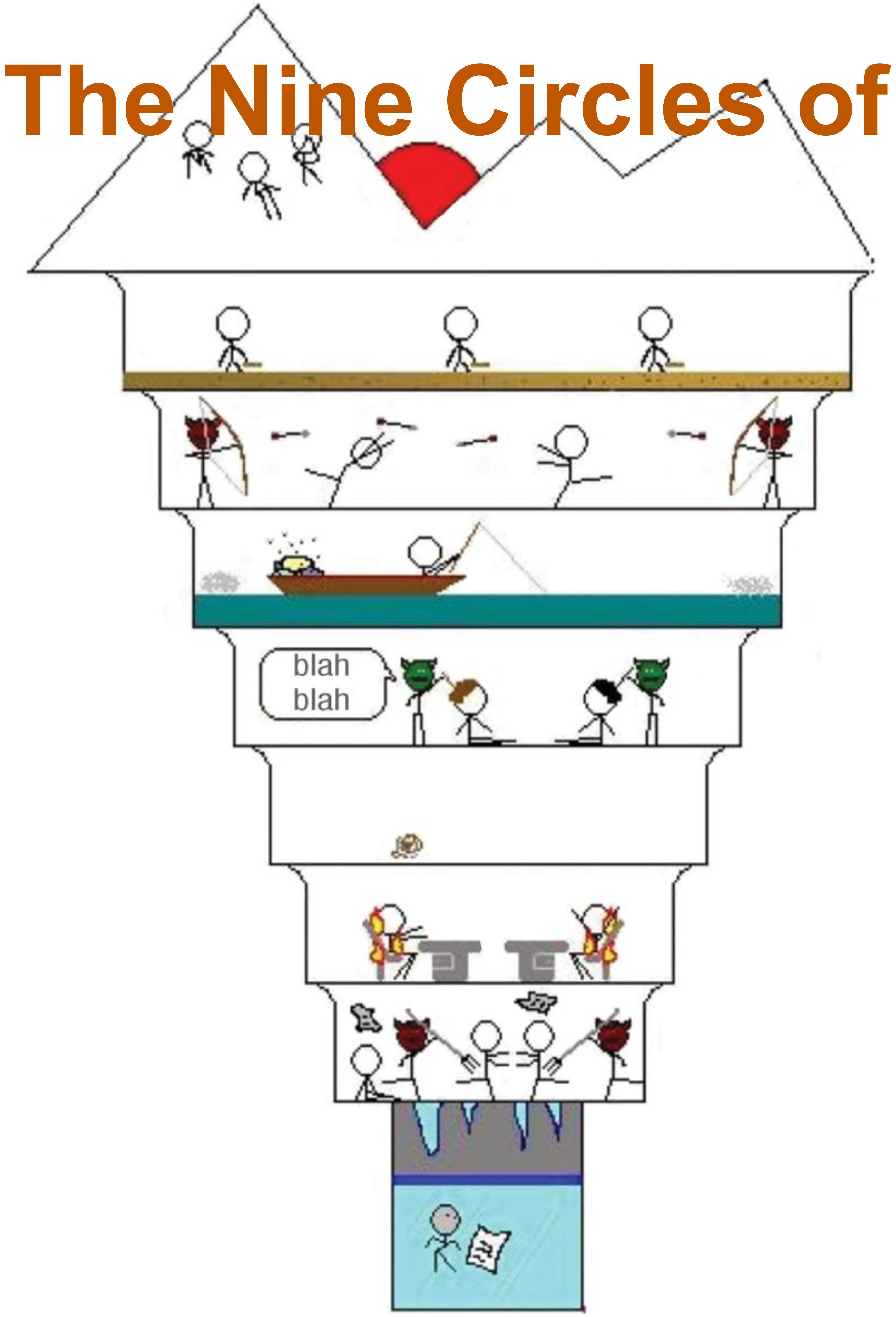
Note:
If choosing between rival models, one often uses 3-fold split
Training: train
Validation: compare models
Test: estimate accuracy (no model tuning anymore here)

Cross-Validation: Split & Repeat

10 fold



The Nine Circles of Scientific Hell

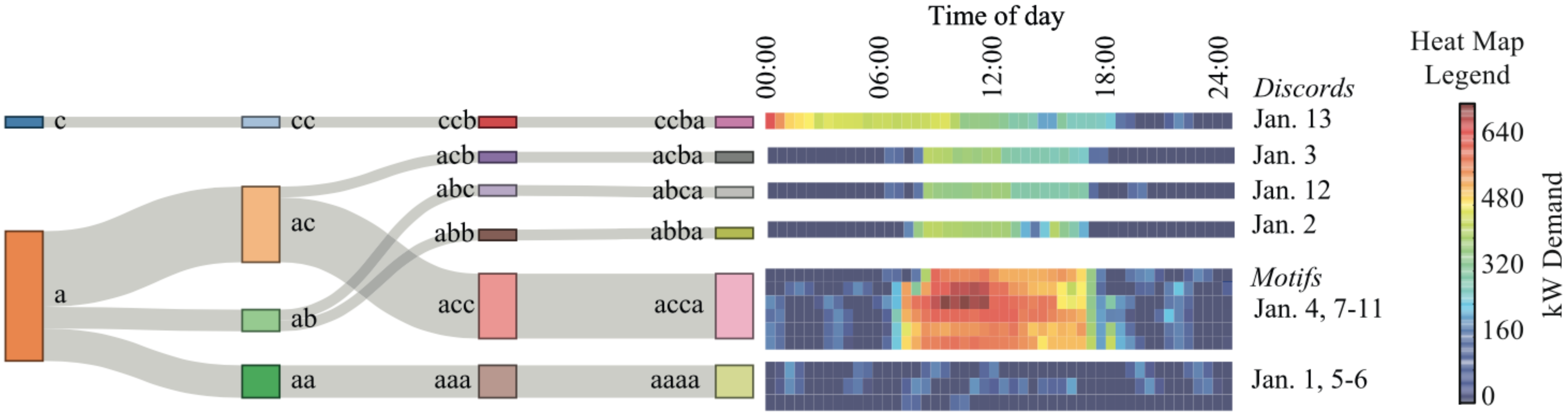


I	Limbo
II	Overselling
III	Post-Hoc Storytelling
IV	P-Value Fishing
V	Creative Outliers
VI	Plagiarism
VII	Non-Publication
VIII	Partial Publication
IX	Inventing Data

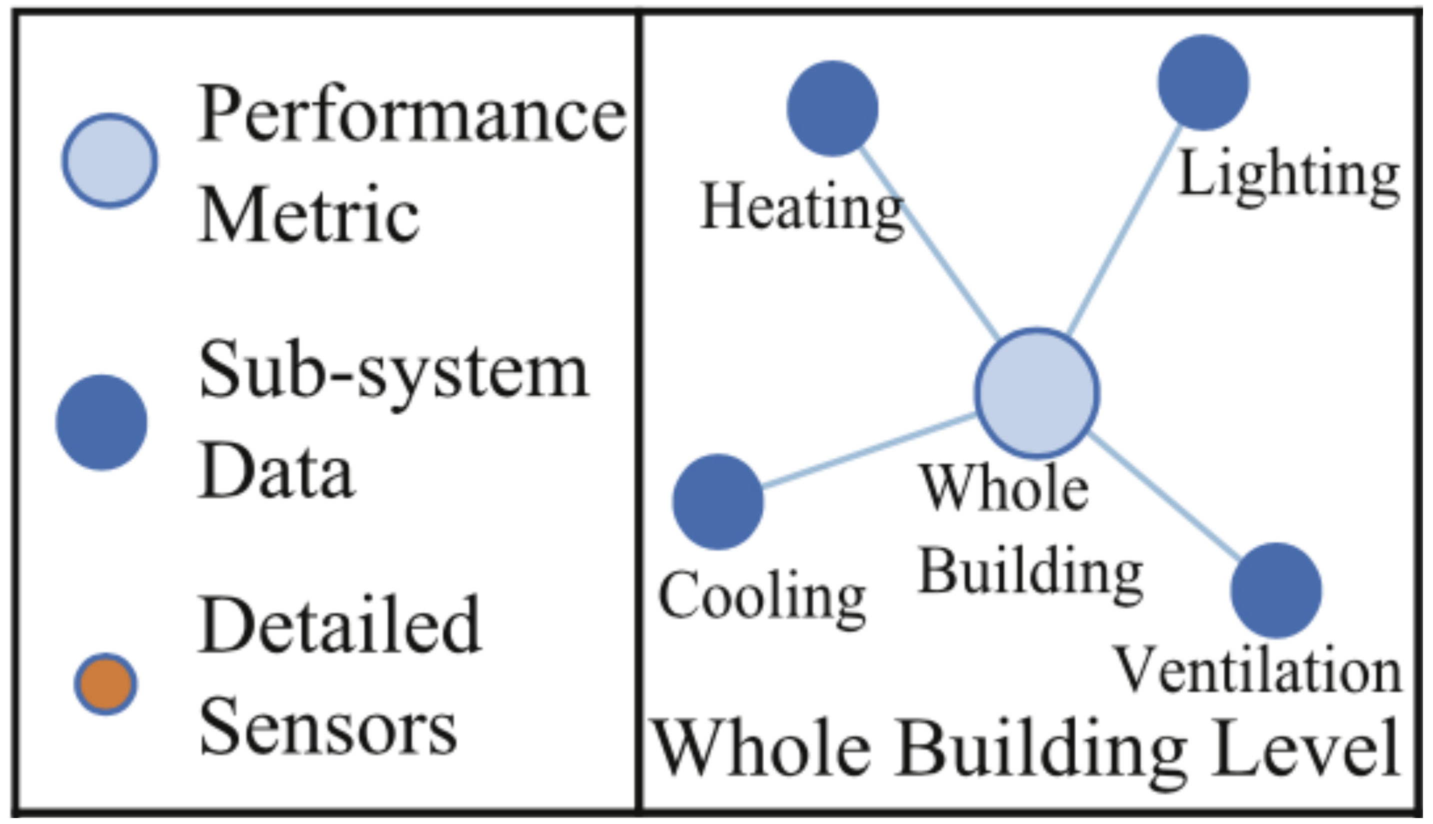
4 out of 9 deal with data analysis !!!

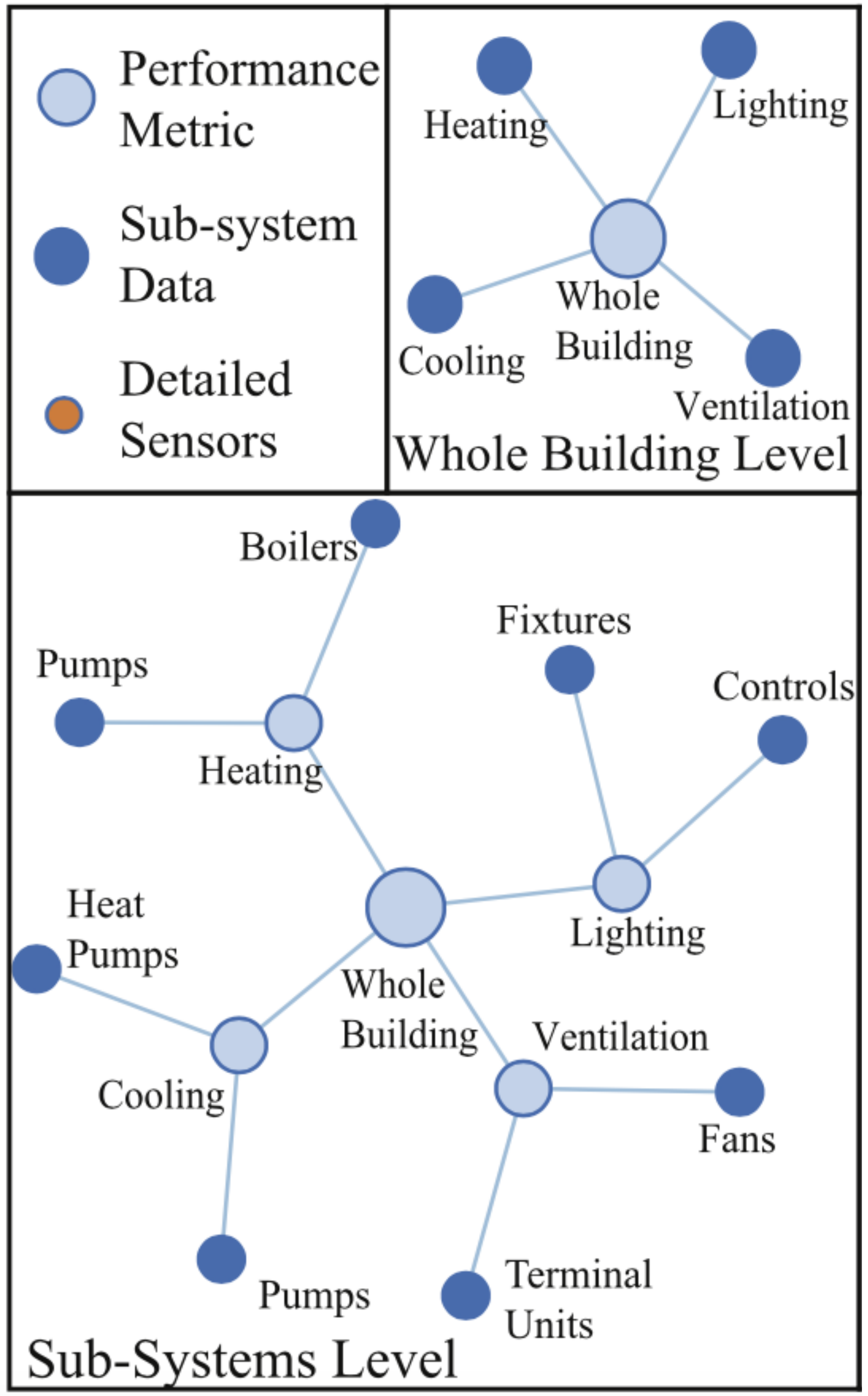
Perspectives on Psychological Science
Vol 7, Issue 6, pp. 643 - 644

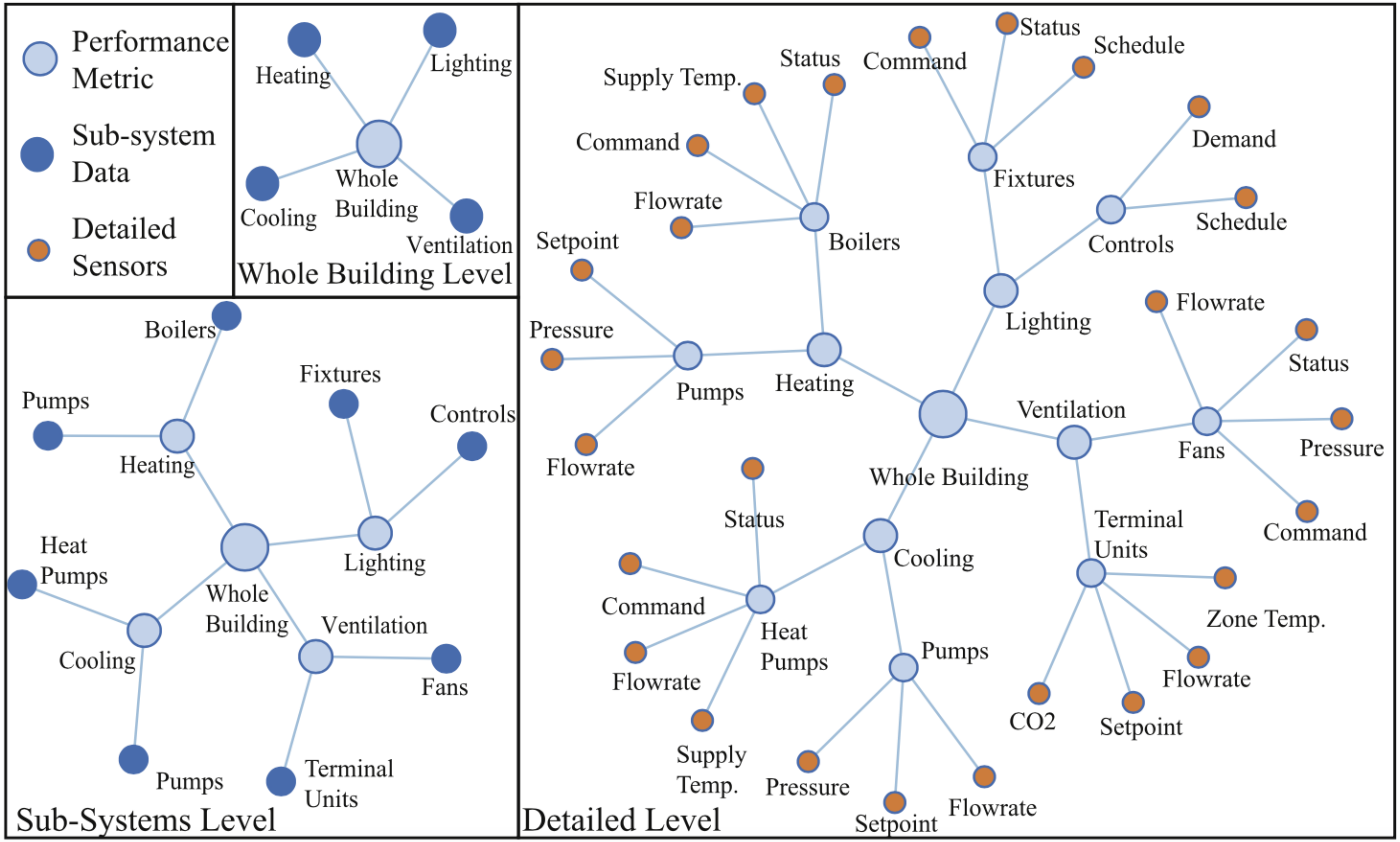
Automated daily pattern filtering of measured building performance data

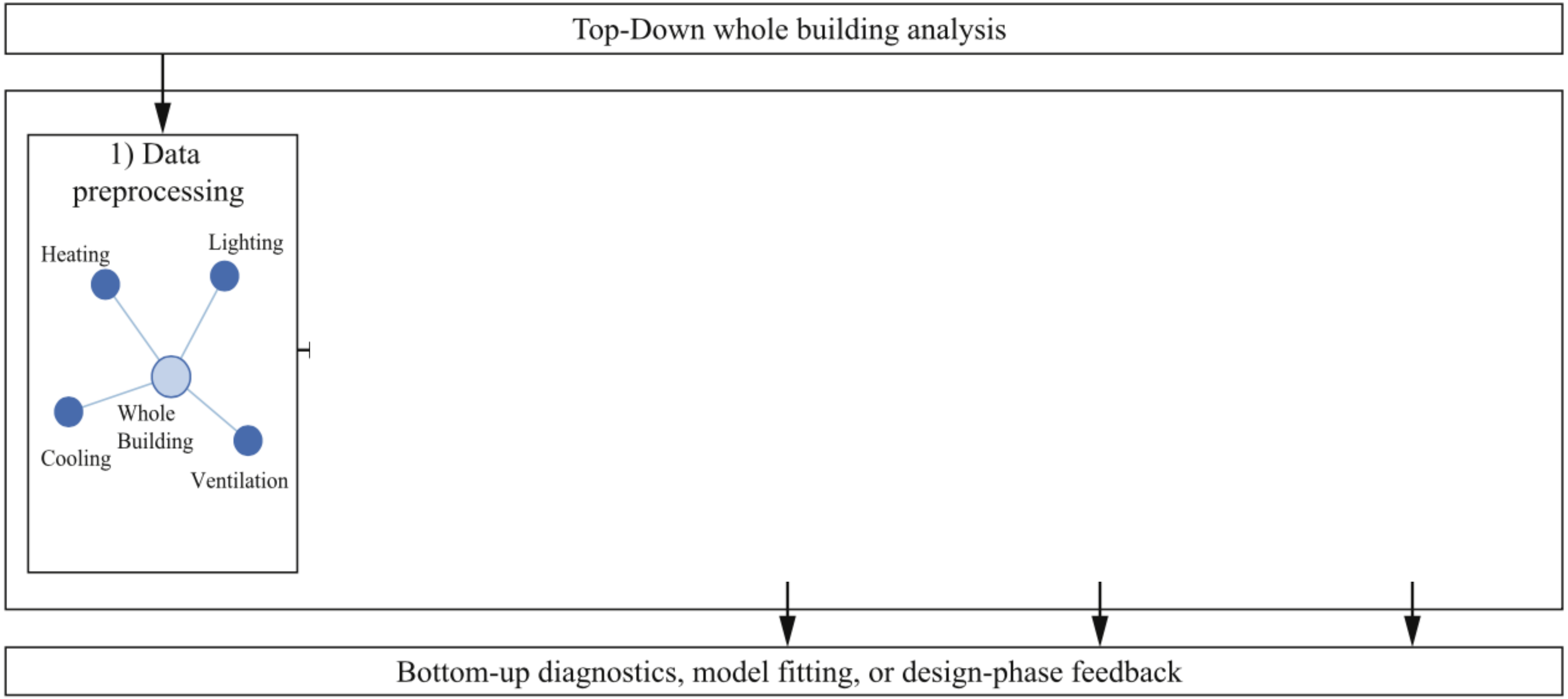


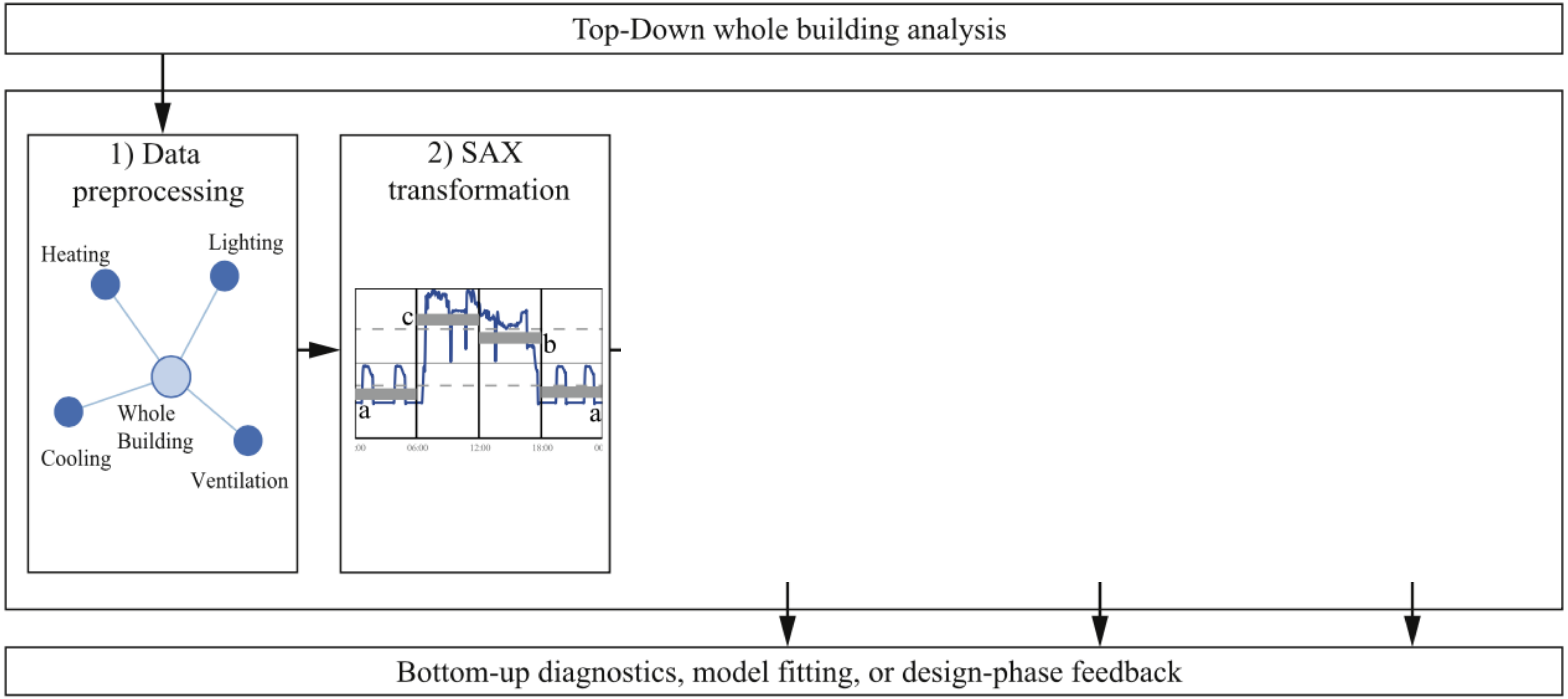
C. Miller, Z. Nagy, and A. Schlueter, *Automated Daily Pattern Filtering of Measured Building Performance Data*, Automation in Construction, vol. 49, no. A, pp. 1–17, Jan. 2014.

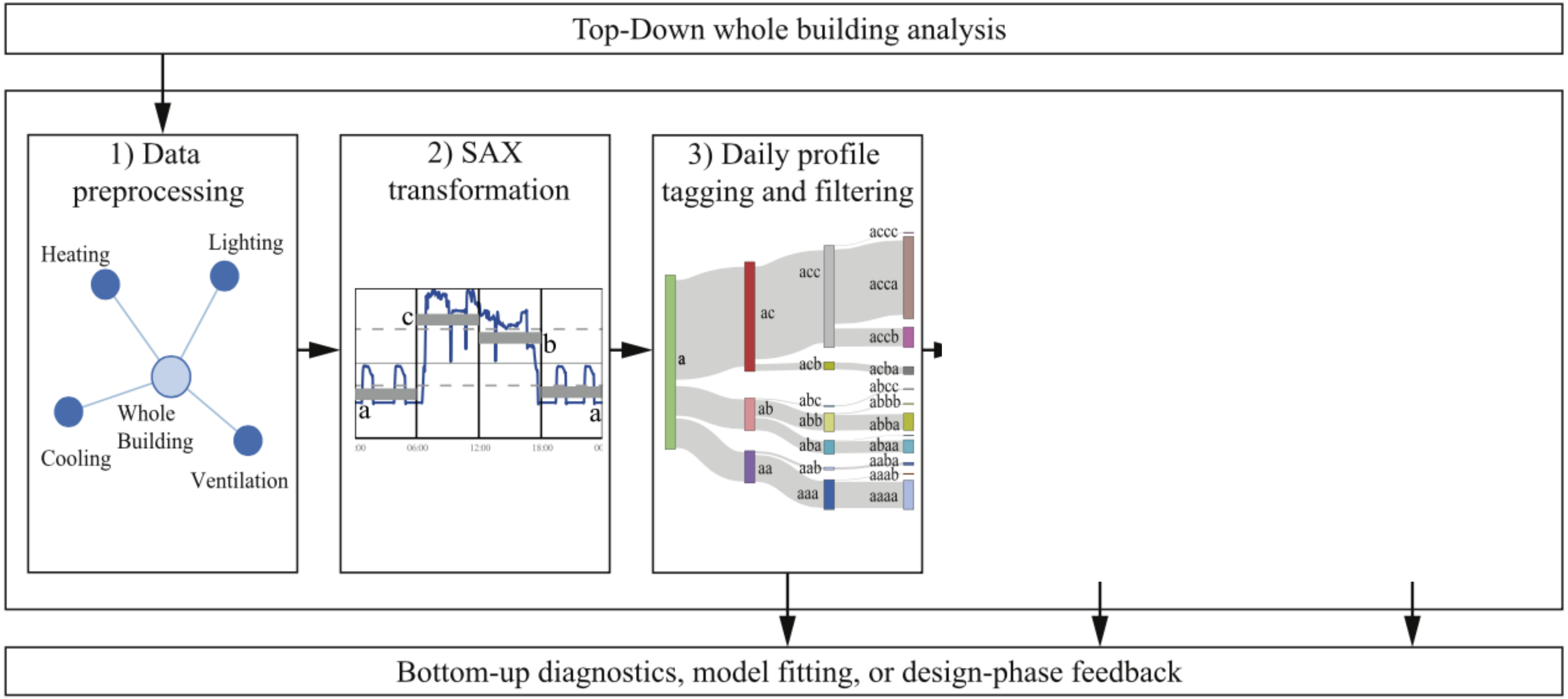

Performance Metric
Sub-system Data
Detailed Sensors
Heating
Lighting
Whole Building
Cooling
Ventilation
Whole Building Level

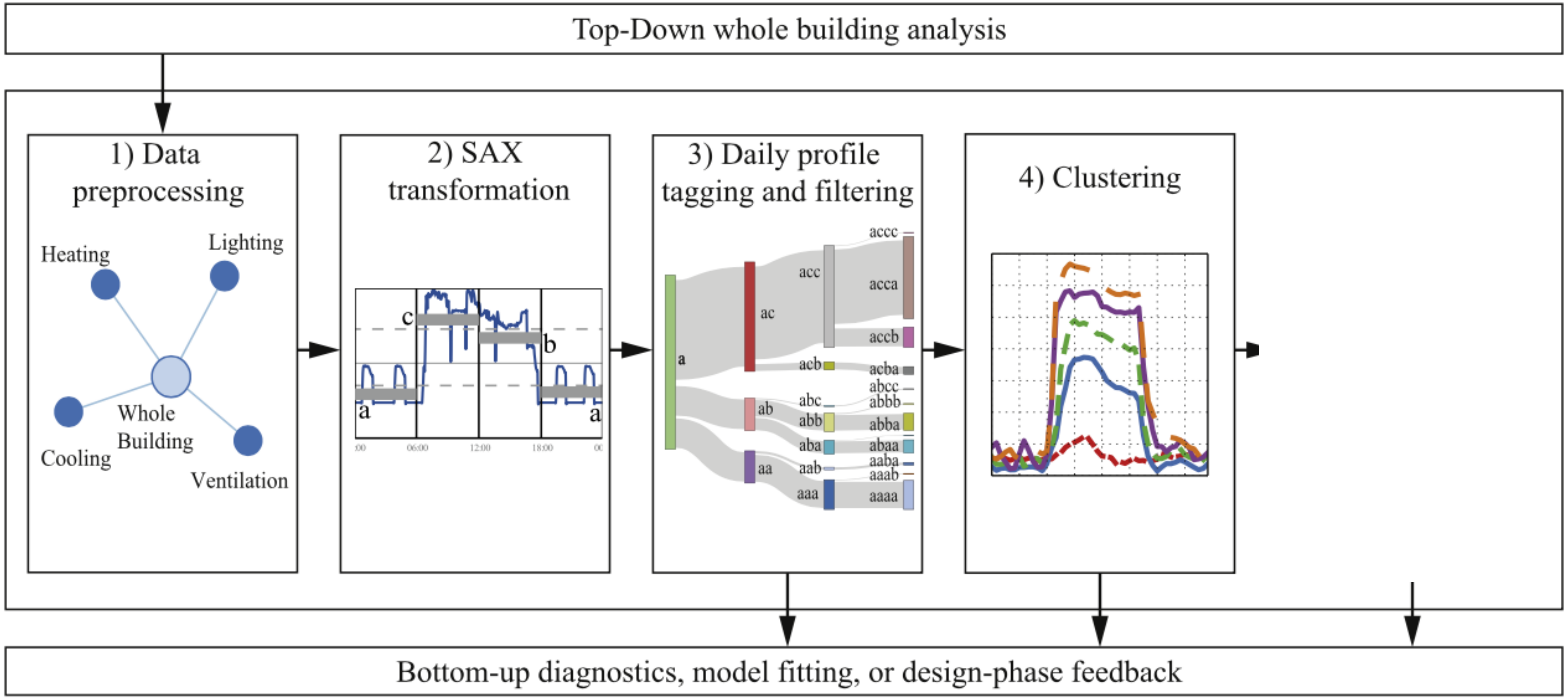

Performance Metric
Sub-system Data
Detailed Sensors
Heating
Lighting
Whole Building
Cooling
Ventilation
Whole Building Level
Boilers
Fixtures
Pumps
Controls
Heating
Heat Pumps
Lighting
Whole Building
Ventilation
Cooling
Fans
Pumps
Terminal Units
Sub-Systems Level

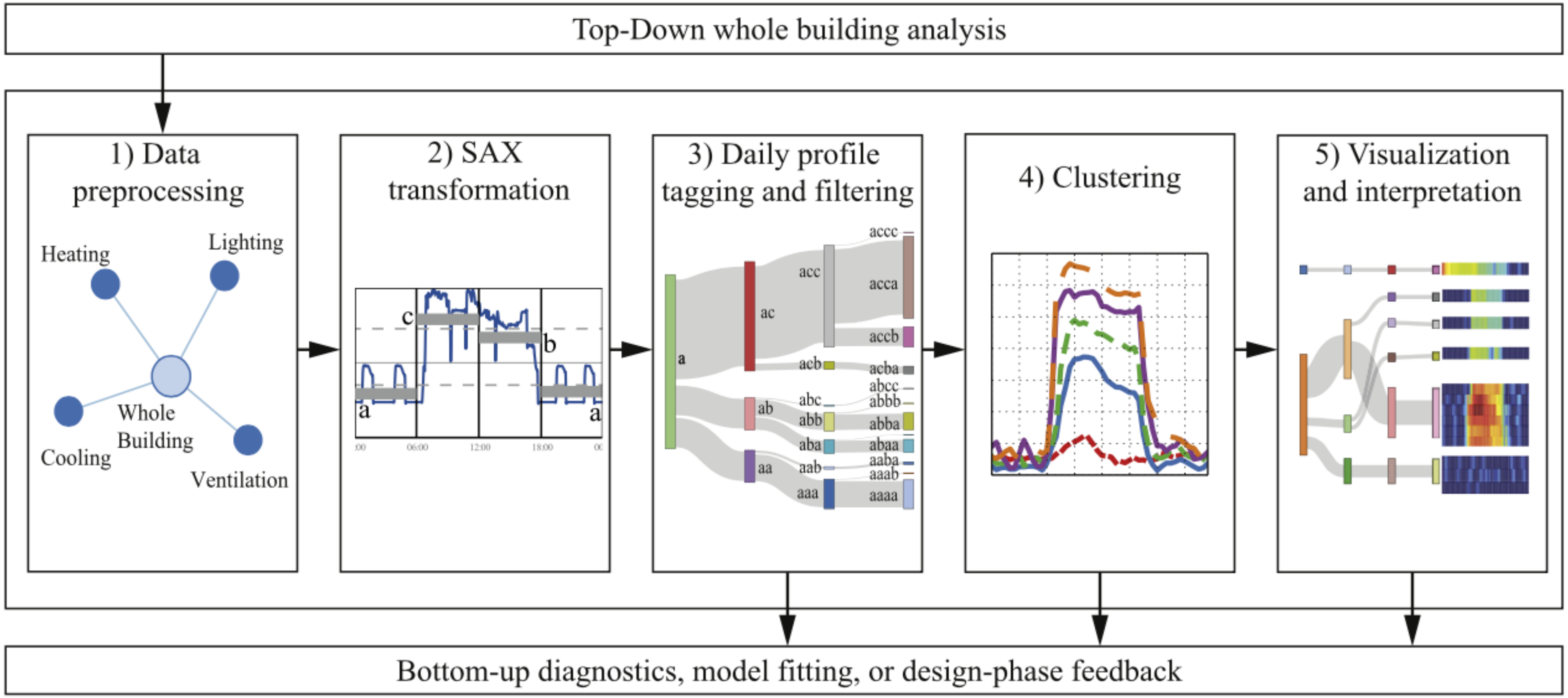

Performance Metric
Sub-system Data
Detailed Sensors
Heating
Lighting
Whole Building
Cooling
Ventilation
Whole Building Level
Boilers
Pumps
Heating
Fixtures
Controls
Heat Pumps
Lighting
Whole Building
Ventilation
Cooling
Fans
Pumps
Terminal Units
Sub-Systems Level
Supply Temp.
Status
Command
Status
Schedule
Command
Demand
Fixtures
Flowrate
Boilers
Schedule
Setpoint
Controls
Pressure
Lighting
Flowrate
Pumps
Heating
Status
Ventilation
Fans
Pressure
Flowrate
Whole Building
Status
Terminal Units
Command
Cooling
Command
Zone Temp.
Heat Pumps
Flowrate
Pumps
Flowrate
CO2
Setpoint
Supply Temp.
Pressure
Setpoint
Flowrate
Detailed Level


Top-Down whole building analysis
1) Data preprocessing
Heating
Lighting
Whole Building
Cooling
Ventilation
Bottom-up diagnostics, model fitting, or design-phase feedback


Top-Down whole building analysis
1) Data preprocessing
Heating
Lighting
Whole Building
Cooling
Ventilation
2) SAX transformation
c
b
a
a
Bottom-up diagnostics, model fitting, or design-phase feedback

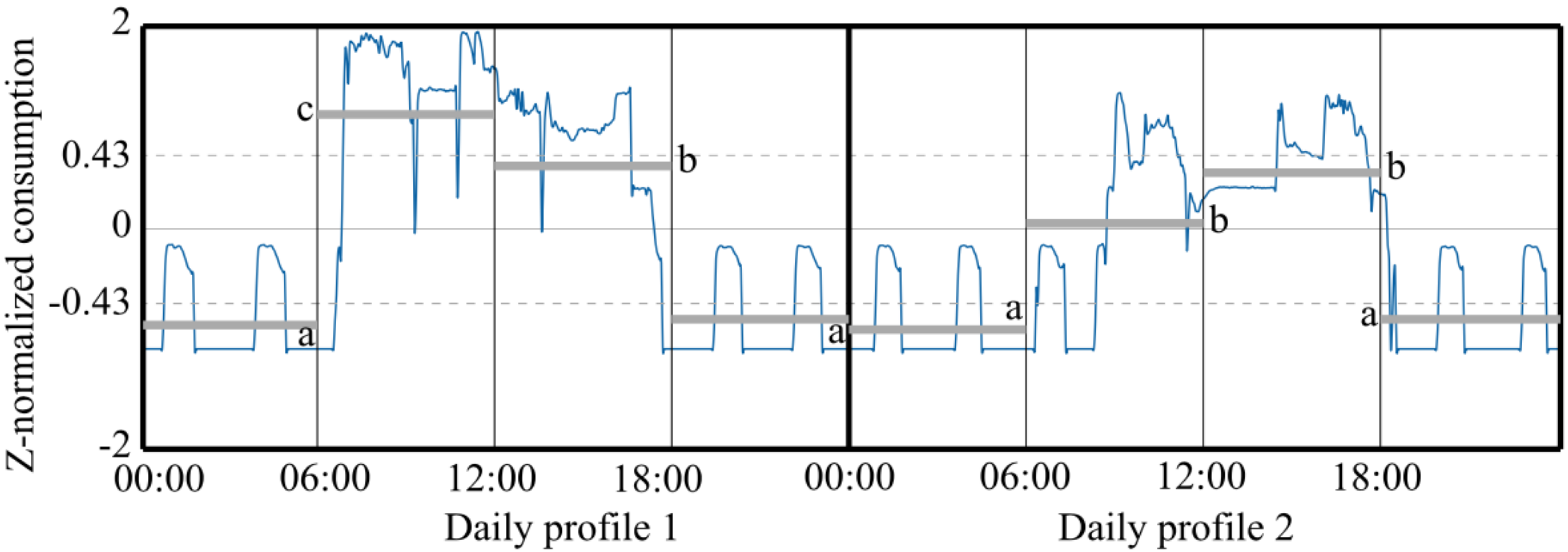

Top-Down whole building analysis
1) Data preprocessing
Heating
Lighting
Whole Building
Cooling
Ventilation
2) SAX transformation
c
b
a
a
3) Daily profile tagging and filtering
accc
acc
acca
ac
accb
a
acb
acba
abcc
abc
abbb
ab
abb
abba
aba
abaa
aaba
aab
aaab
aa
aaa
aaaa
Bottom-up diagnostics, model fitting, or design-phase feedback


Top-Down whole building analysis
1) Data preprocessing
Heating
Lighting
Whole Building
Cooling
Ventilation
2) SAX transformation
c
b
a
a
3) Daily profile tagging and filtering
a
ac
acc
accc
acca
accb
acb
acba
abcc
abc
abbb
ab
abb
abba
aba
abaa
aaba
aab
aaab
aa
aaa
aaaa
4) Clustering
Bottom-up diagnostics, model fitting, or design-phase feedback


Top-Down whole building analysis
1) Data preprocessing
Heating
Lighting
Whole Building
Cooling
Ventilation
2) SAX transformation
c
b
a
a
3) Daily profile tagging and filtering
a
ac
ab
aa
acc
acb
abc
abb
aba
aab
aaa
accc
acca
accb
acba
abcc
abbb
abba
abaa
aaba
aaab
aaaa
4) Clustering
5) Visualization and interpretation
Bottom-up diagnostics, model fitting, or design-phase feedback

SAX Transformation

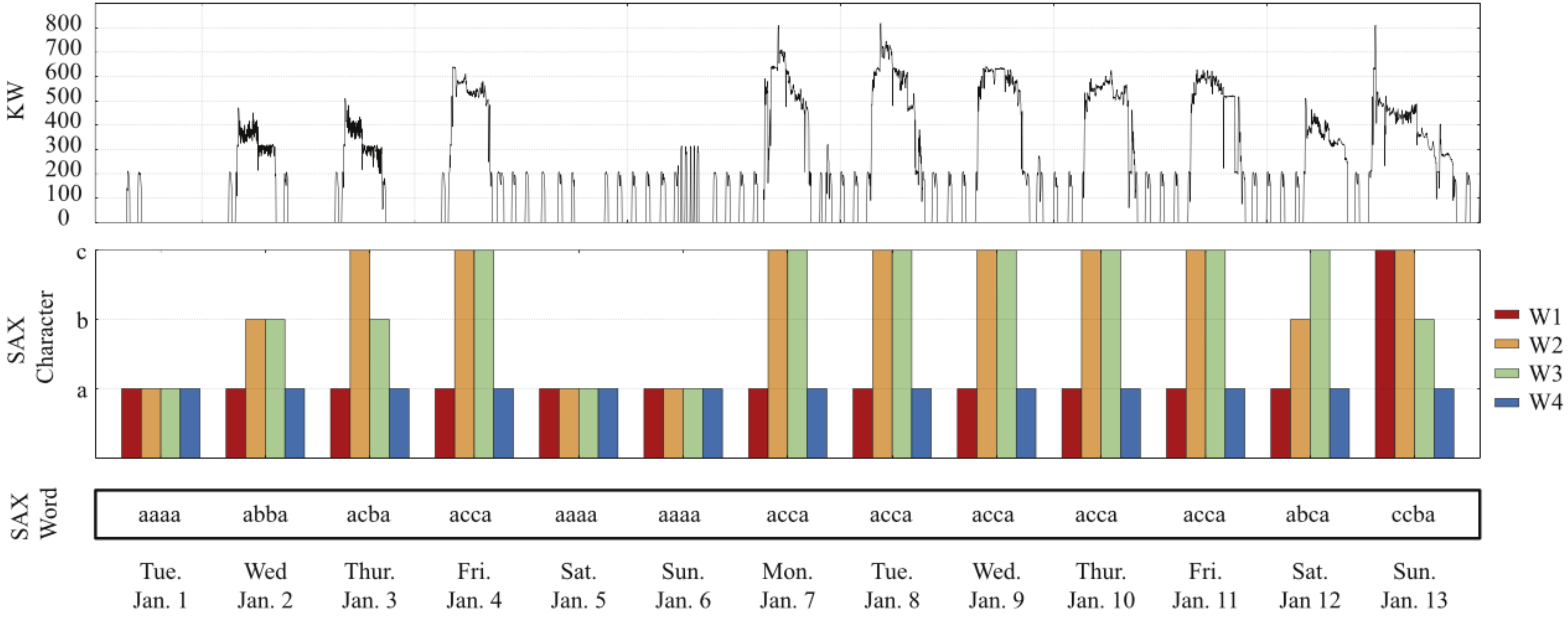
- Symbolic Aggregate approXimation



- Why is this a
- Why is this a bad idea?

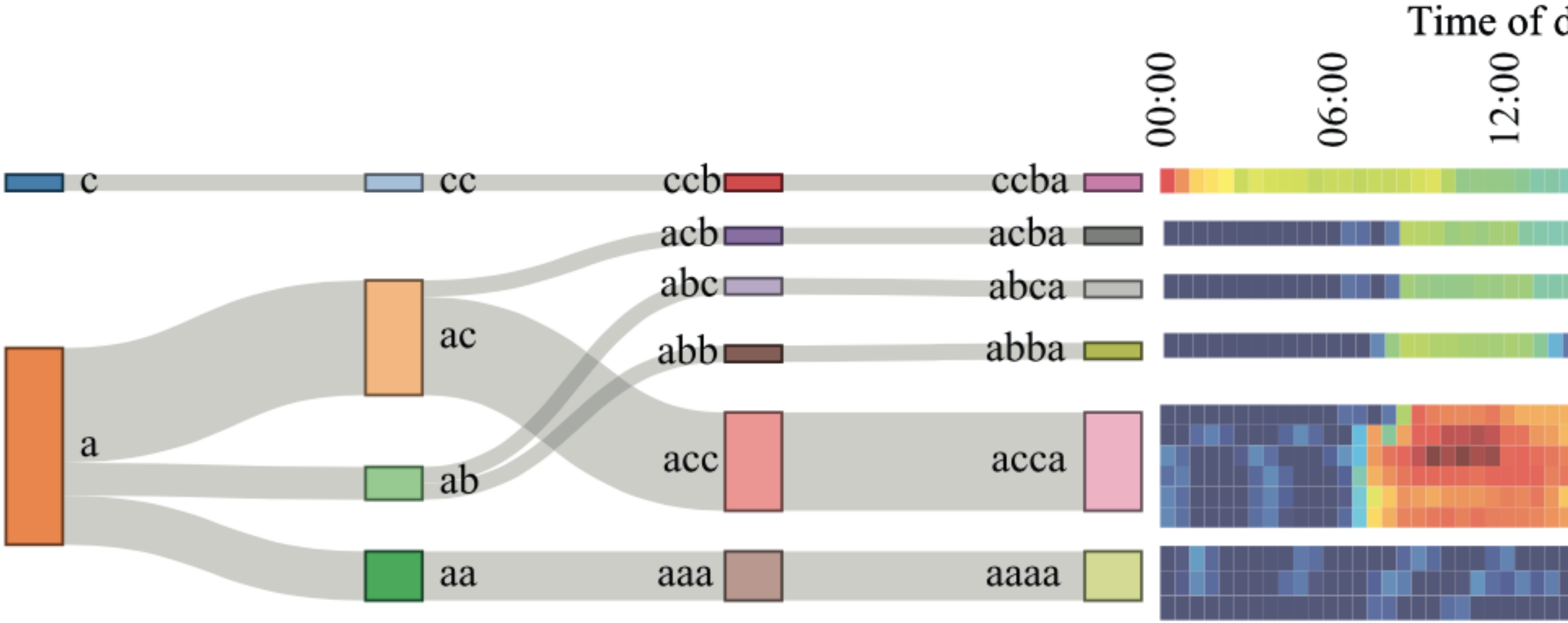
SAX Transformation

- Symbolic Aggregate approXimation



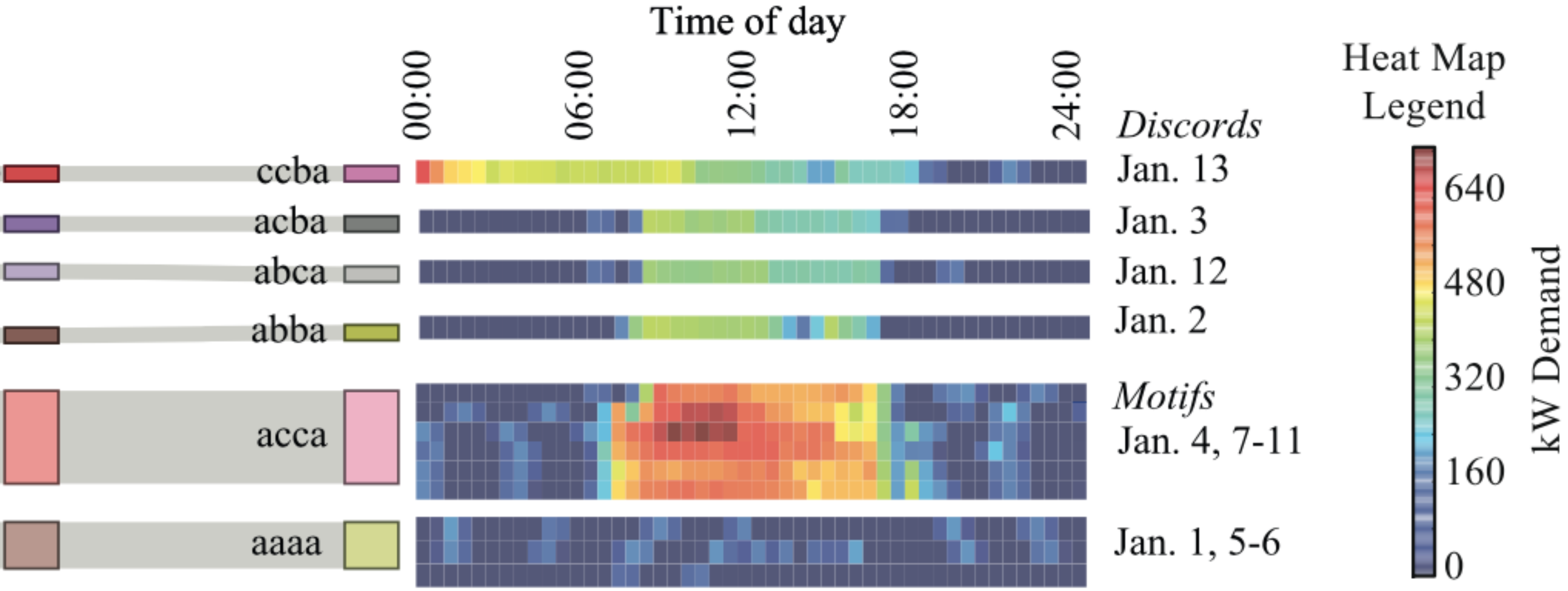
Motifs & Discords

- Symbolic Aggregate approXimation



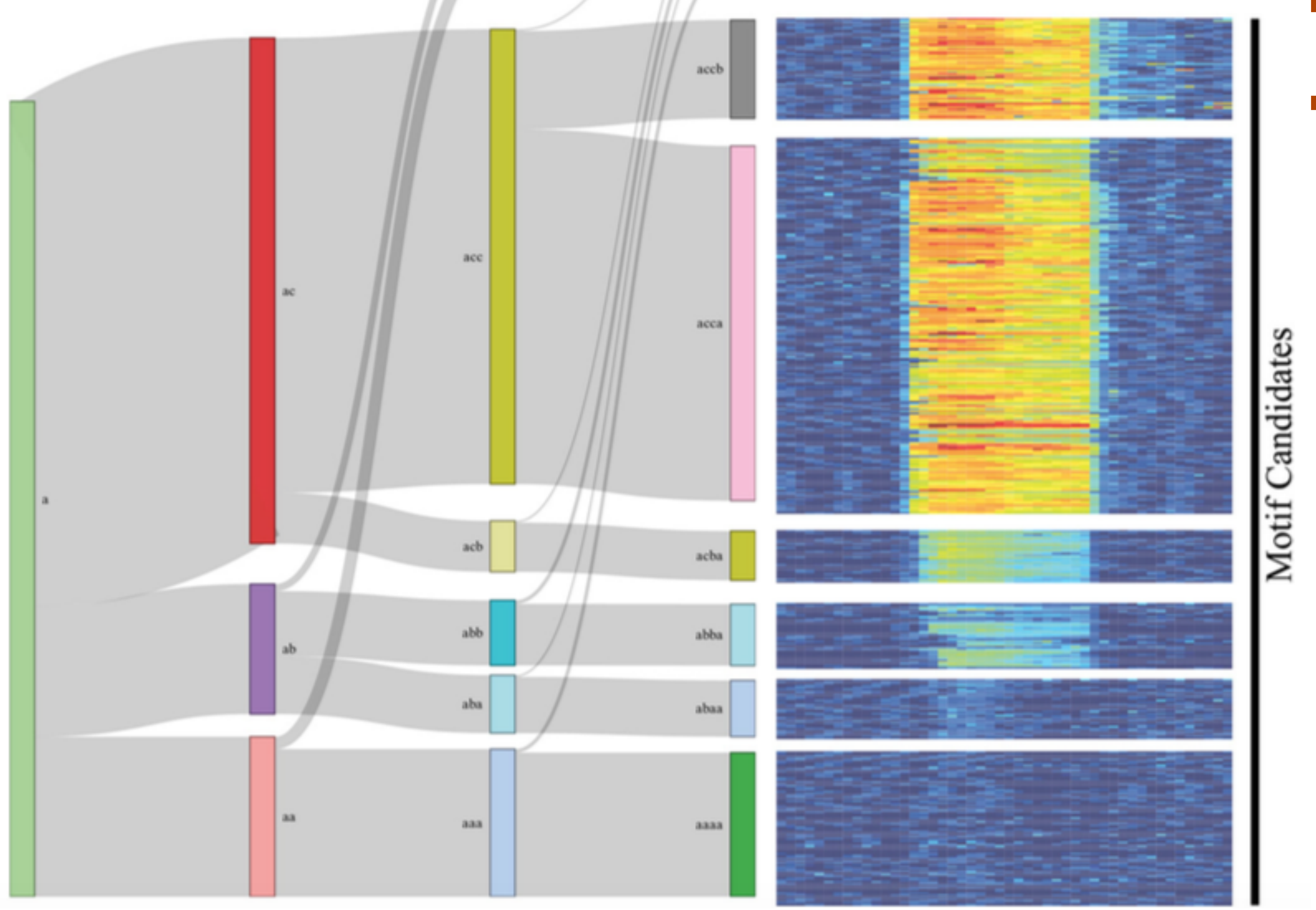
Motifs & Discords

- Symbolic Aggregate approXimation



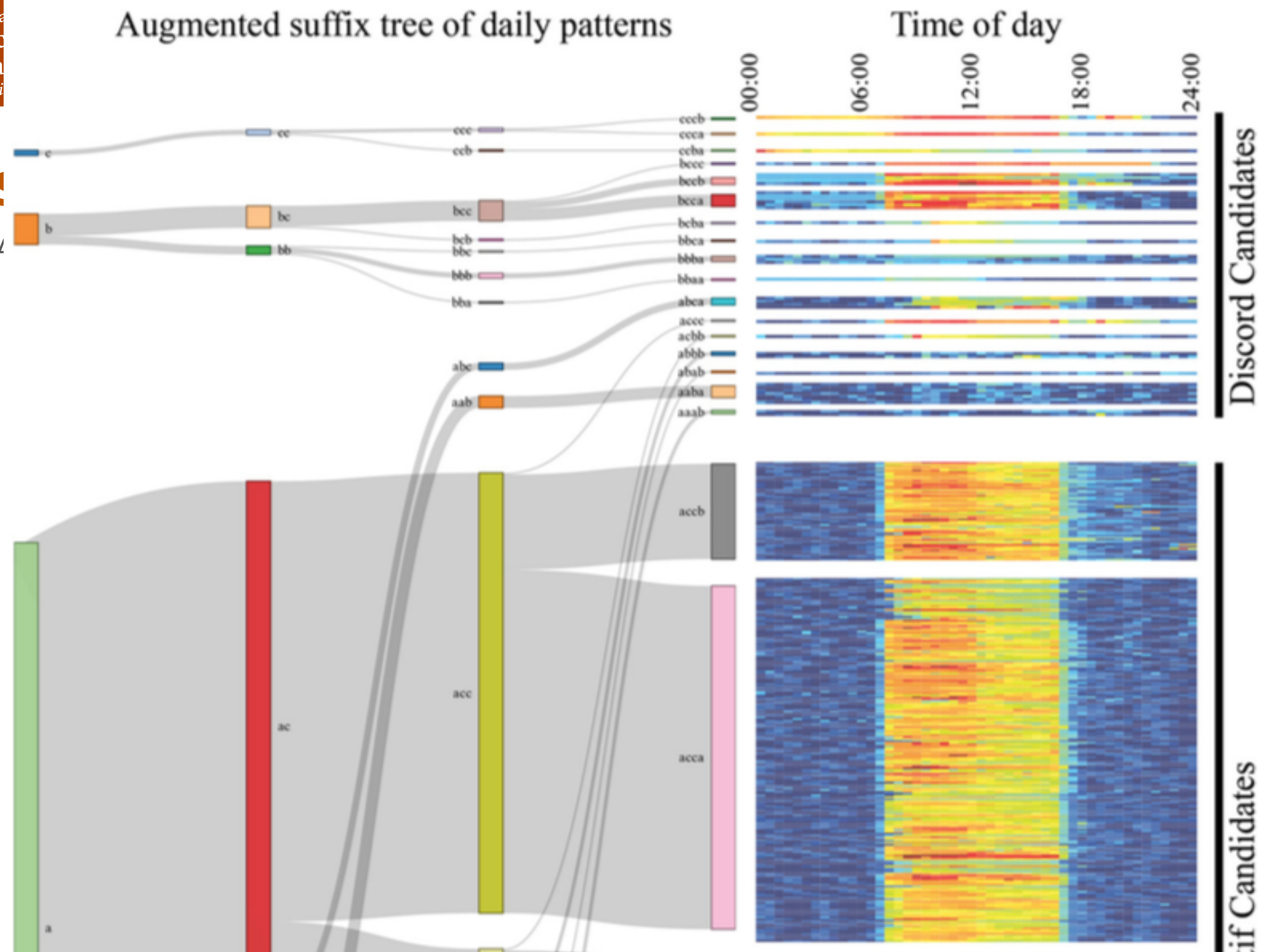
Motif

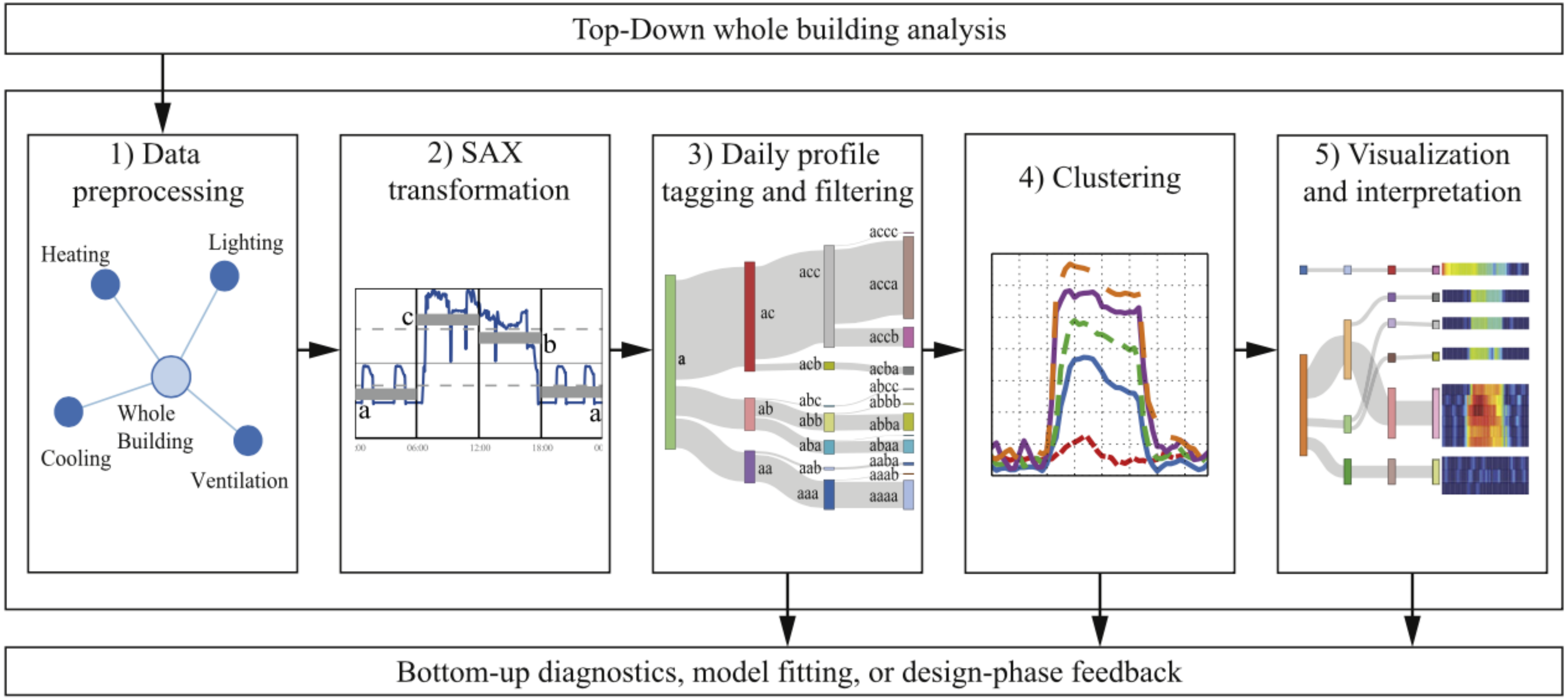
- Symbolic

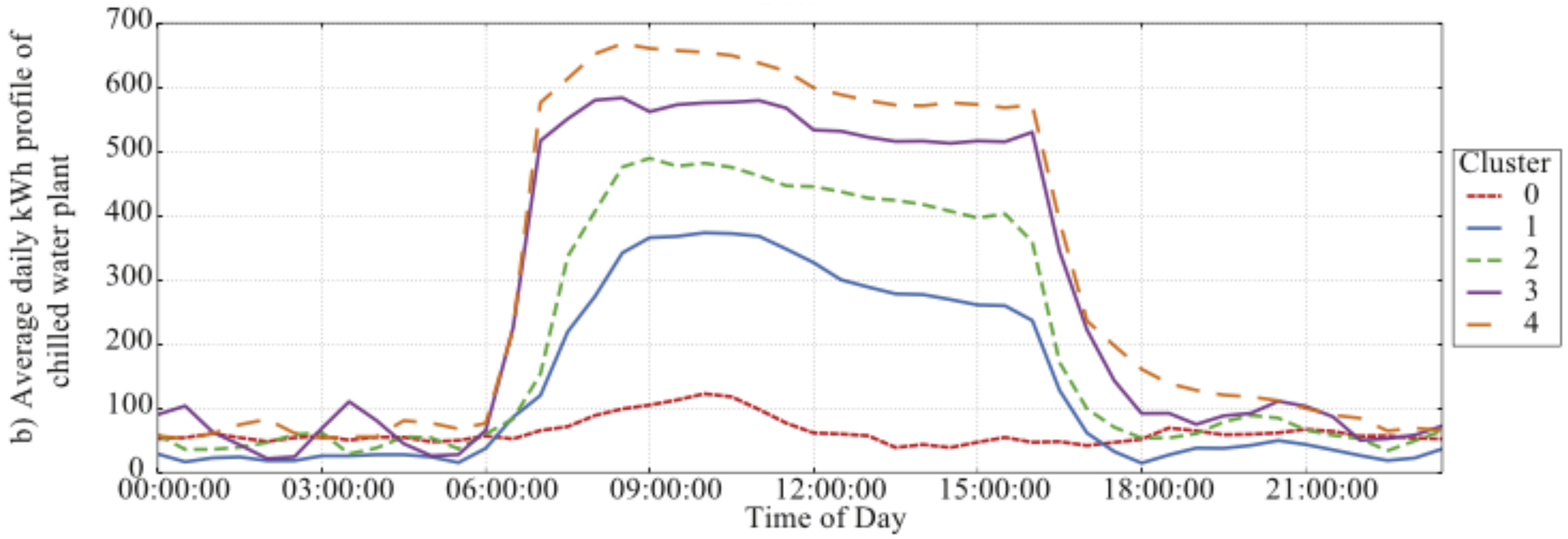


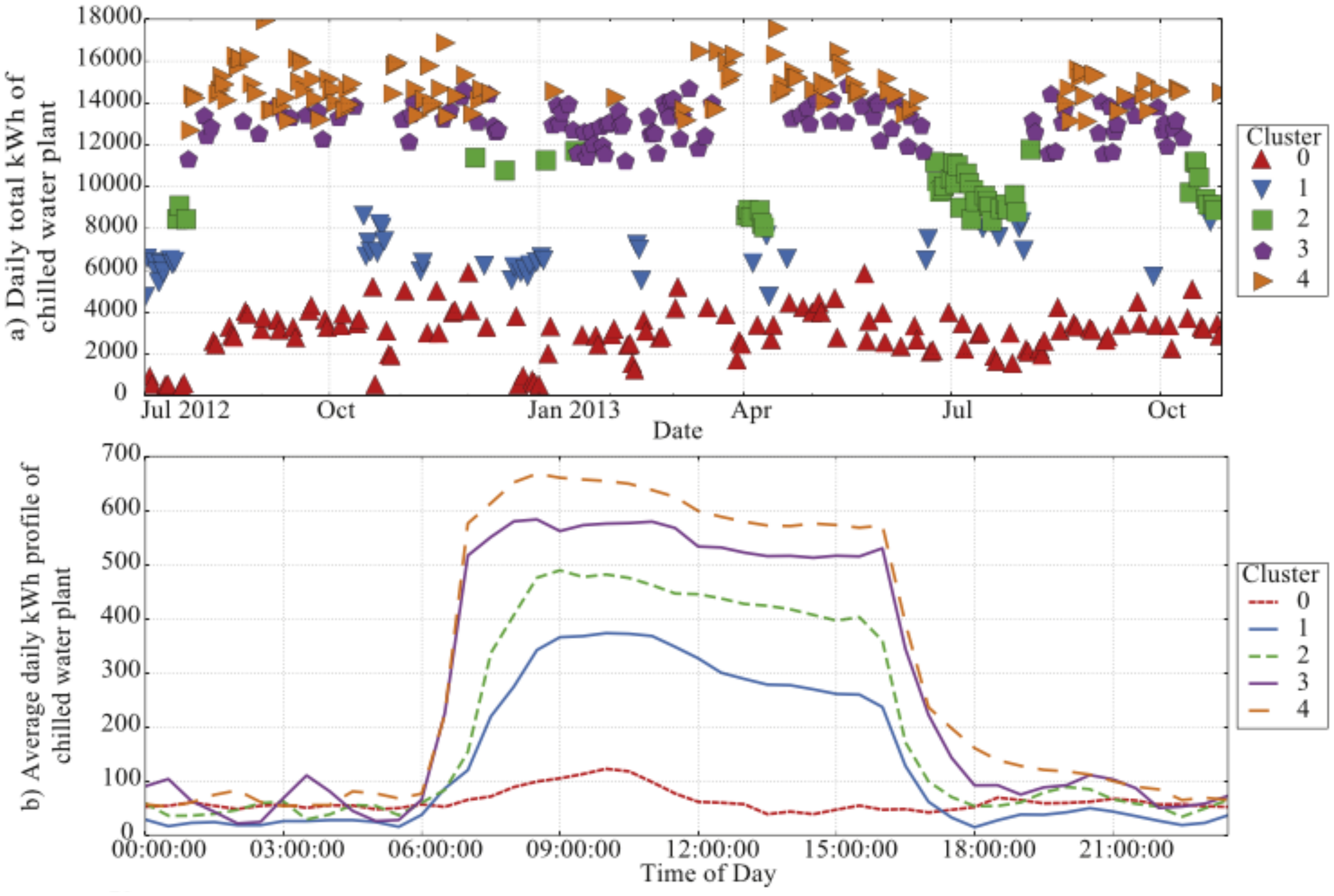
Motif

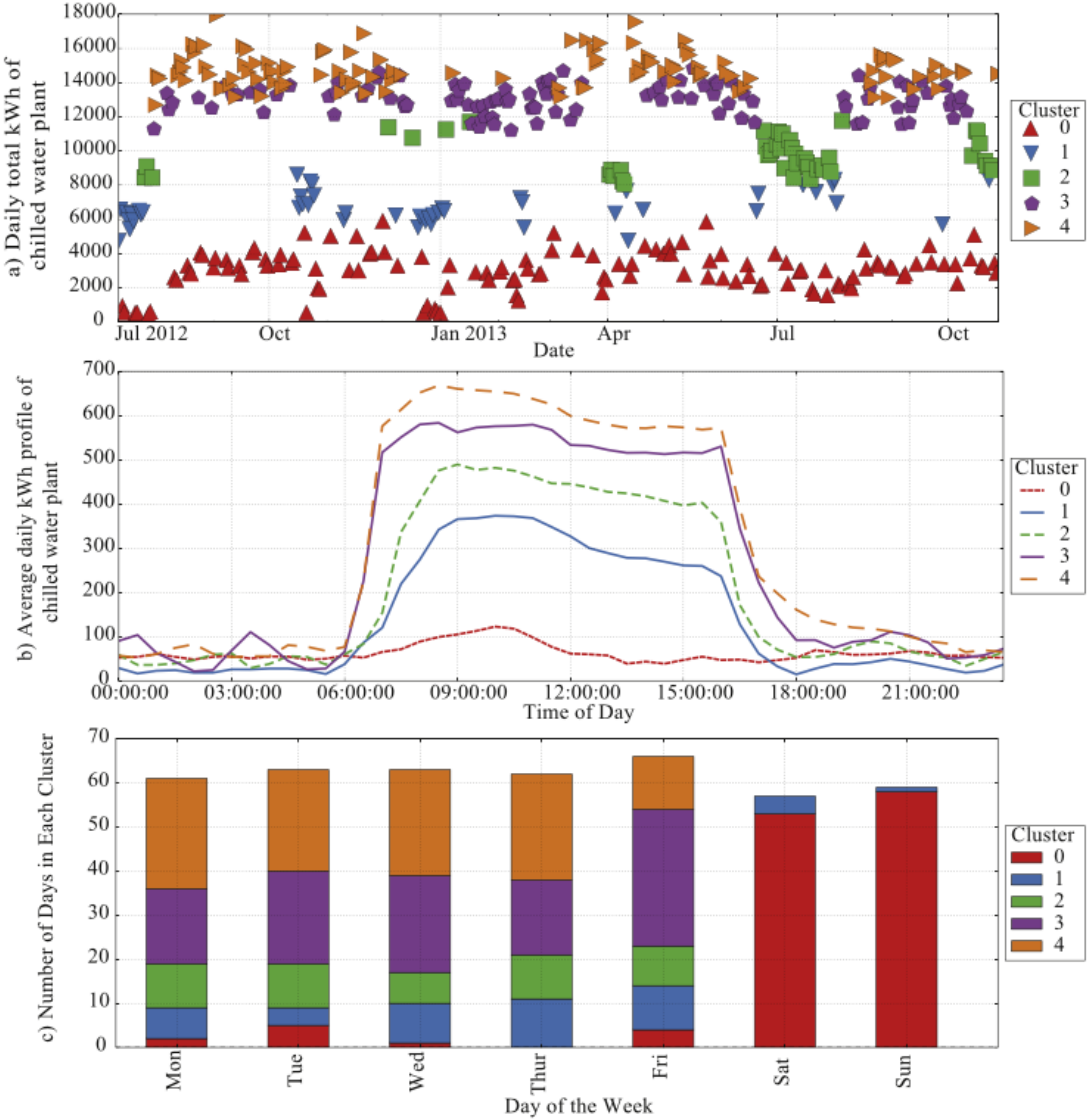
- Symbolic




Top-Down whole building analysis
1) Data preprocessing
Heating
Lighting
Whole Building
Cooling
Ventilation
2) SAX transformation
c
b
a
a
3) Daily profile tagging and filtering
4) Clustering
5) Visualization and interpretation
Bottom-up diagnostics, model fitting, or design-phase feedback


700
600
500
400
300
200
100
0
b) Average daily kWh profile of chilled water plant
00:00:00
03:00:00
06:00:00
09:00:00
12:00:00
15:00:00
18:00:00
21:00:00
Time of Day
Cluster
0
1
2
3
4


a) Daily total kWh of chilled water plant
Date
Jul 2012
Oct
Jan 2013
Apr
Jul
Oct
Cluster
0
1
2
3
4
b) Average daily kWh profile of chilled water plant
Time of Day
00:00:00
03:00:00
06:00:00
09:00:00
12:00:00
15:00:00
18:00:00
21:00:00


a) Daily total kWh of chilled water plant
Date
Jul 2012
Oct
Jan 2013
Apr
Jul
Oct
Cluster
0
1
2
3
4
b) Average daily kWh profile of chilled water plant
Time of Day
00:00:00
03:00:00
06:00:00
09:00:00
12:00:00
15:00:00
18:00:00
21:00:00
c) Number of Days in Each Cluster
Day of the Week
Mon
Tue
Wed
Thur
Fri
Sat
Sun