Modeling Data

the different views on Data Mining

Views on Data Mining

- Fitting the data
- Density Estimation
- Learning
  - being able to perform a task more accurately than before
- Prediction
  - use the data to predict future data
- Compressing the data
  - capture the essence of the data
  - discard the noise and details

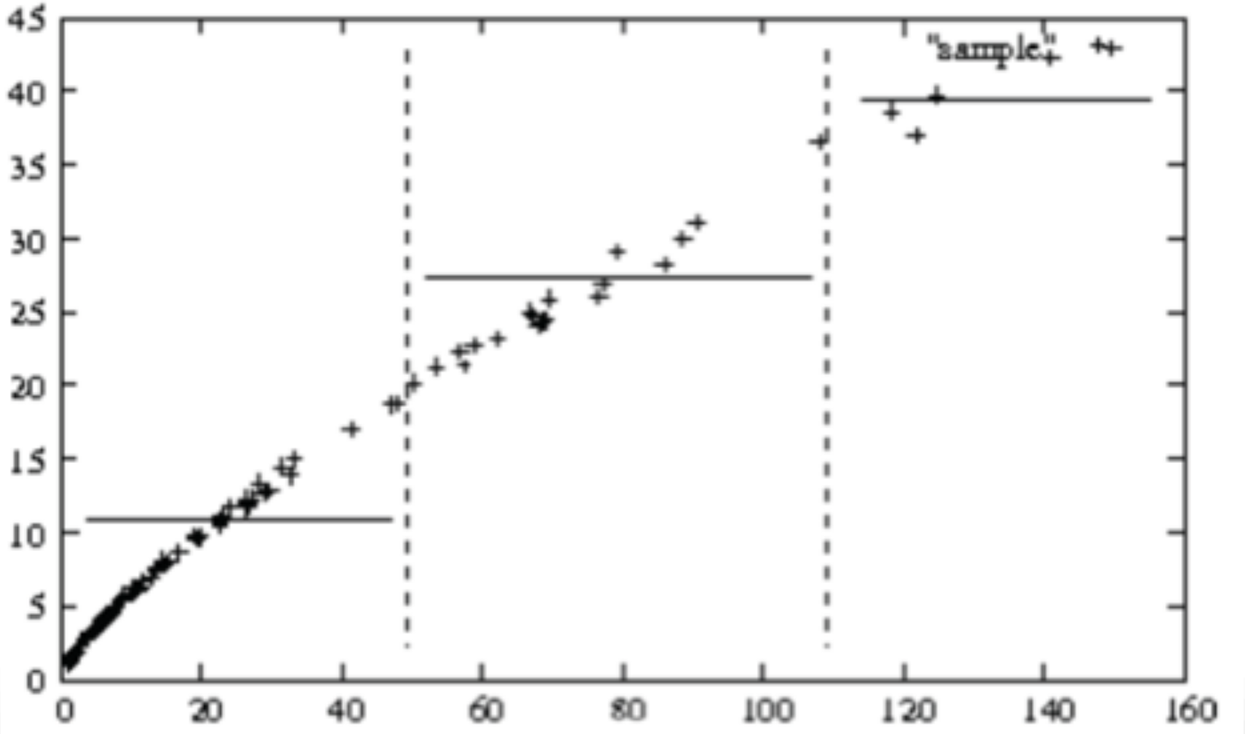
Views on Data Mining

- Fitting the data
- Density Estimation
- Learning
  - being able to perform a task more accurately than before
- Prediction
  - use the data to predict future data
- Compressing the data
  - capture the essence of the data
  - discard the noise and details

Data fitting

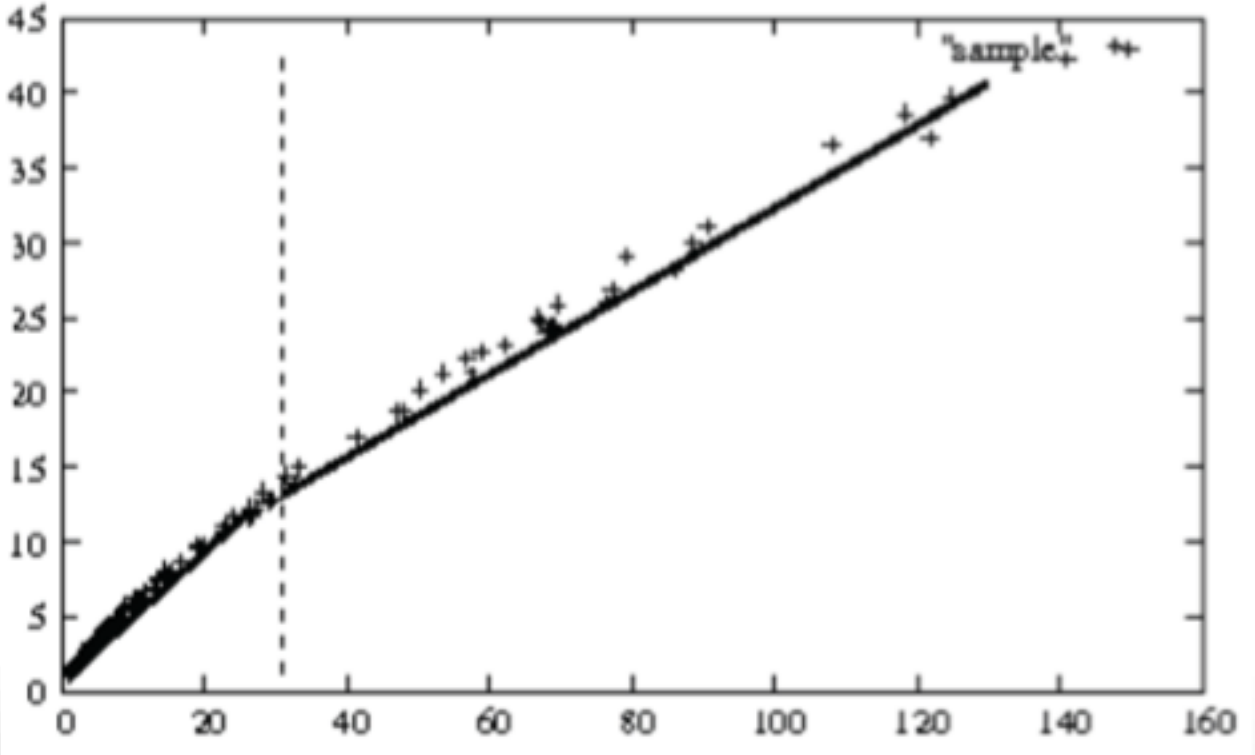
- Very old concept
- Capture function between variables
- Often
  - few variables
  - simple models
- Functions
  - step-functions
  - linear
  - quadratic
- Trade-off between complexity of model and fit (generalization)

response to new drug



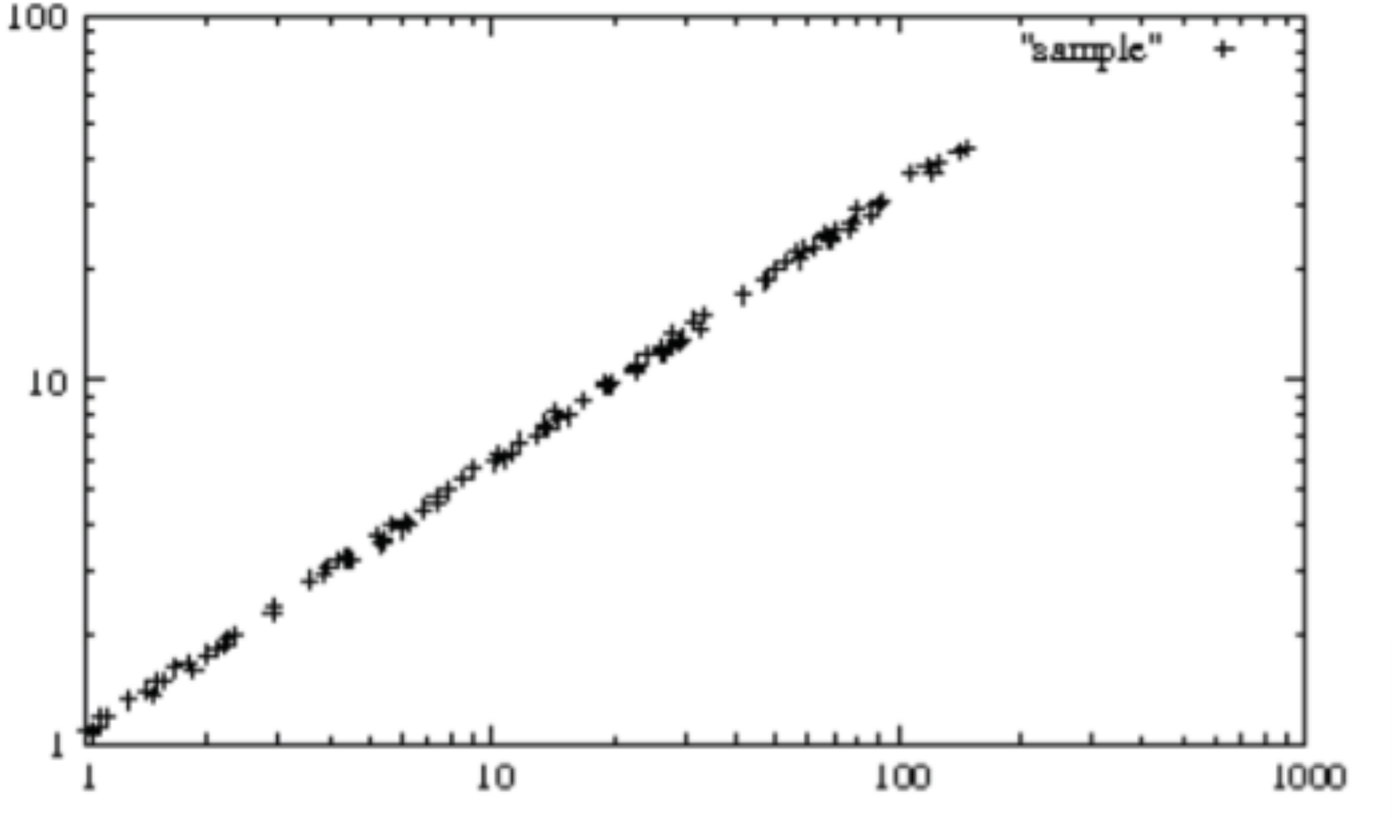
body weight

response to new drug



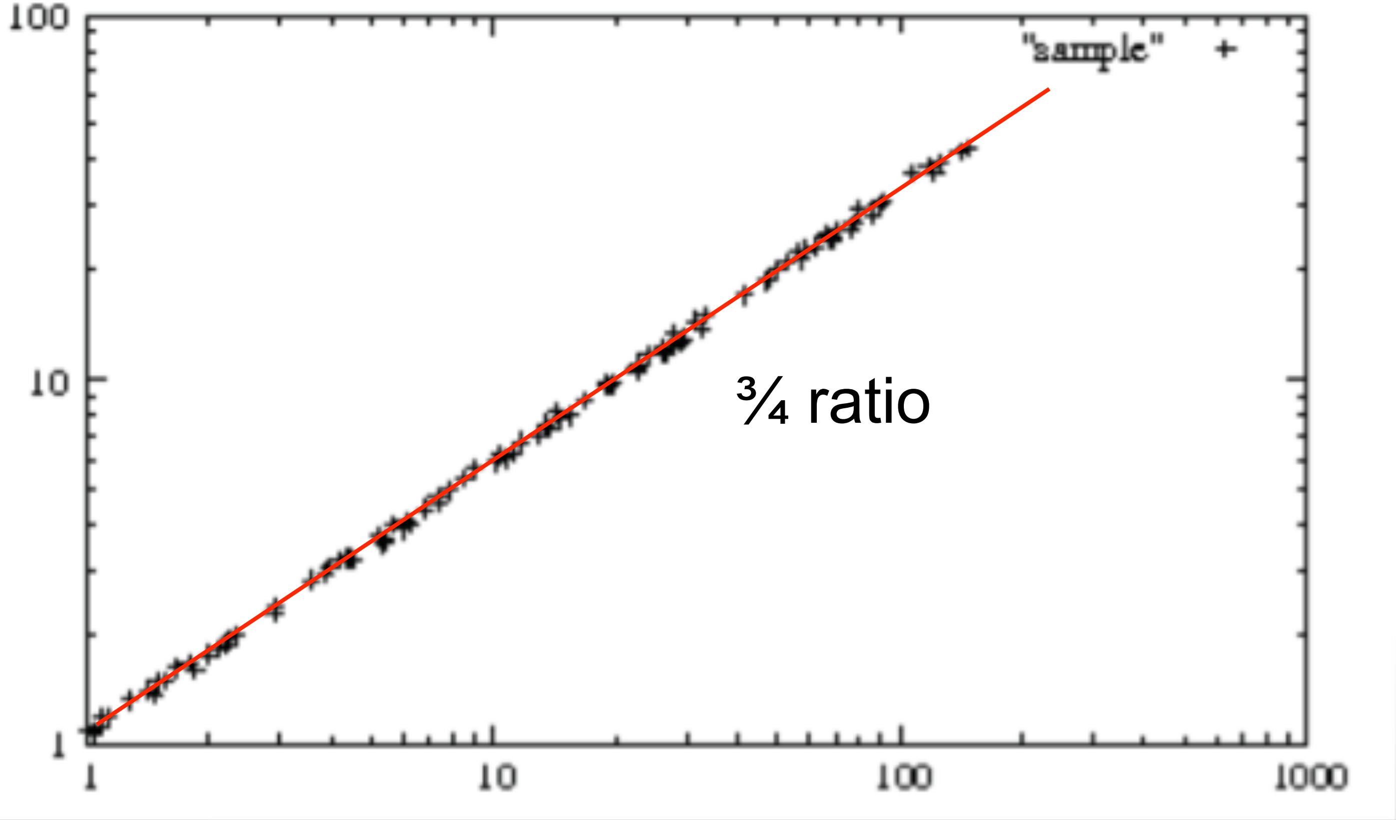
body weight

money spent



income

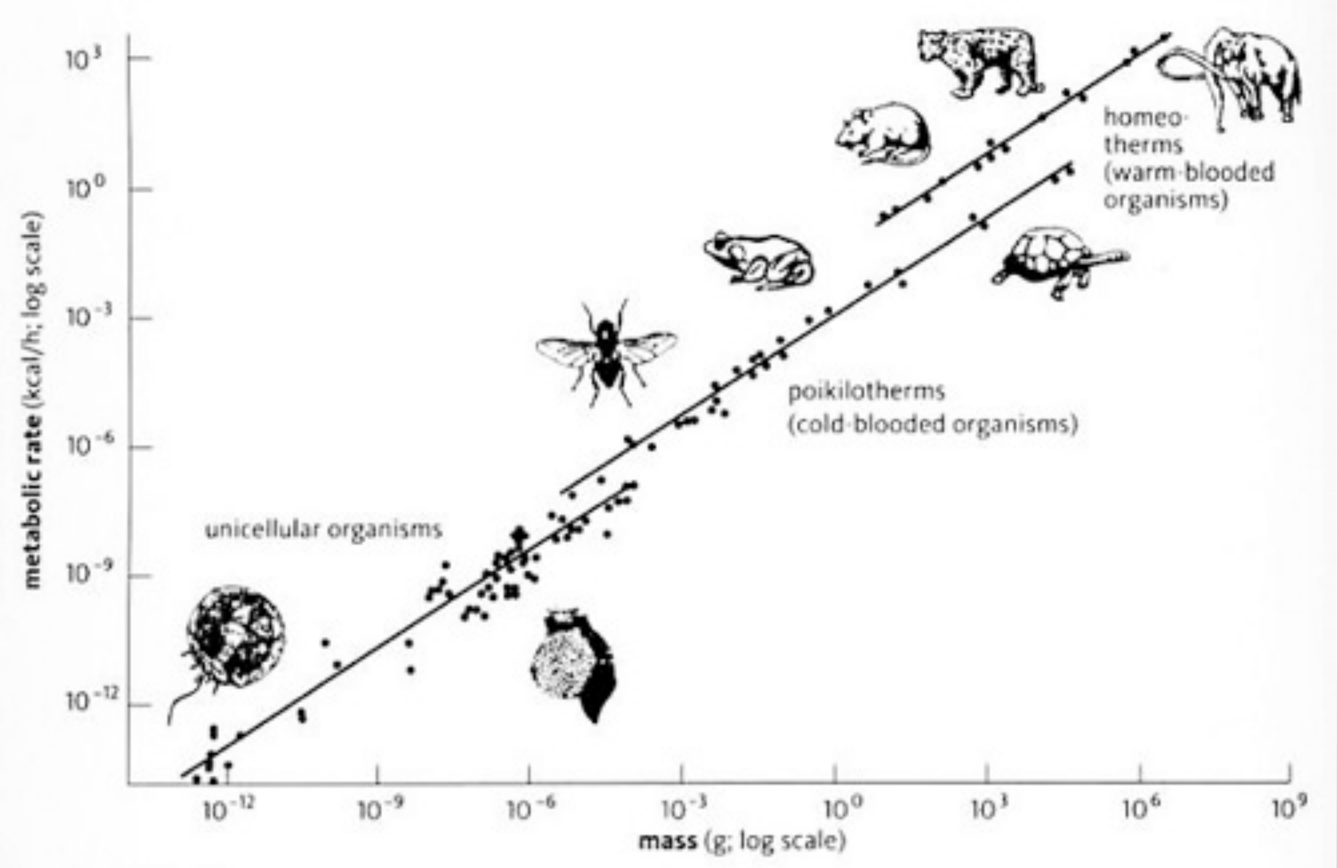
money
spent



¾ ratio

income

Kleiber’s Law of Metabolic Rate



1 kcal/h = 1.162 watts

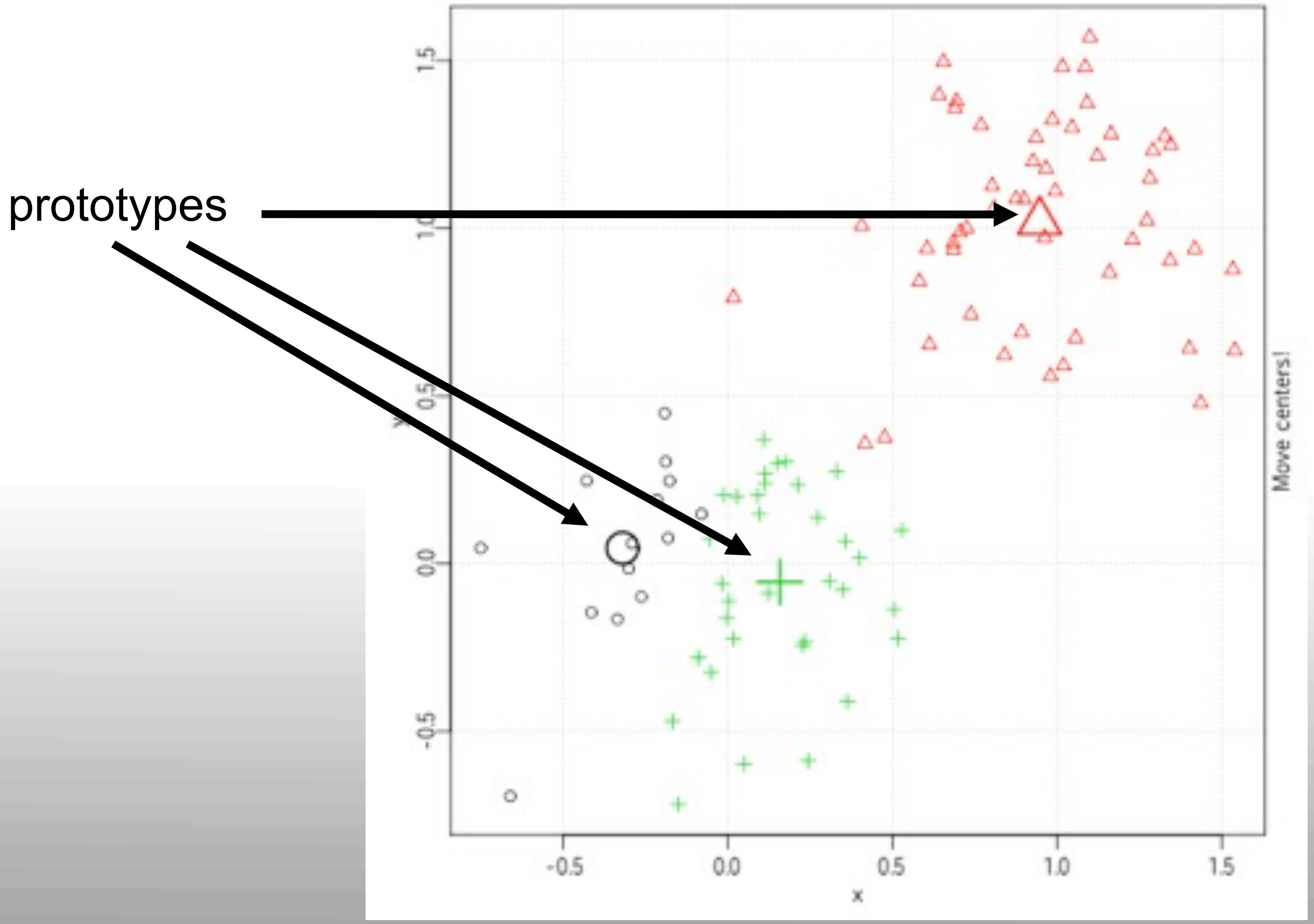
Universiteit Leiden

Views on Data Mining

- Fitting the data
- Density Estimation
- Learning
  - being able to perform a task more accurately than before
- Prediction
  - use the data to predict future data
- Compressing the data
  - capture the essence of the data
  - discard the noise and details

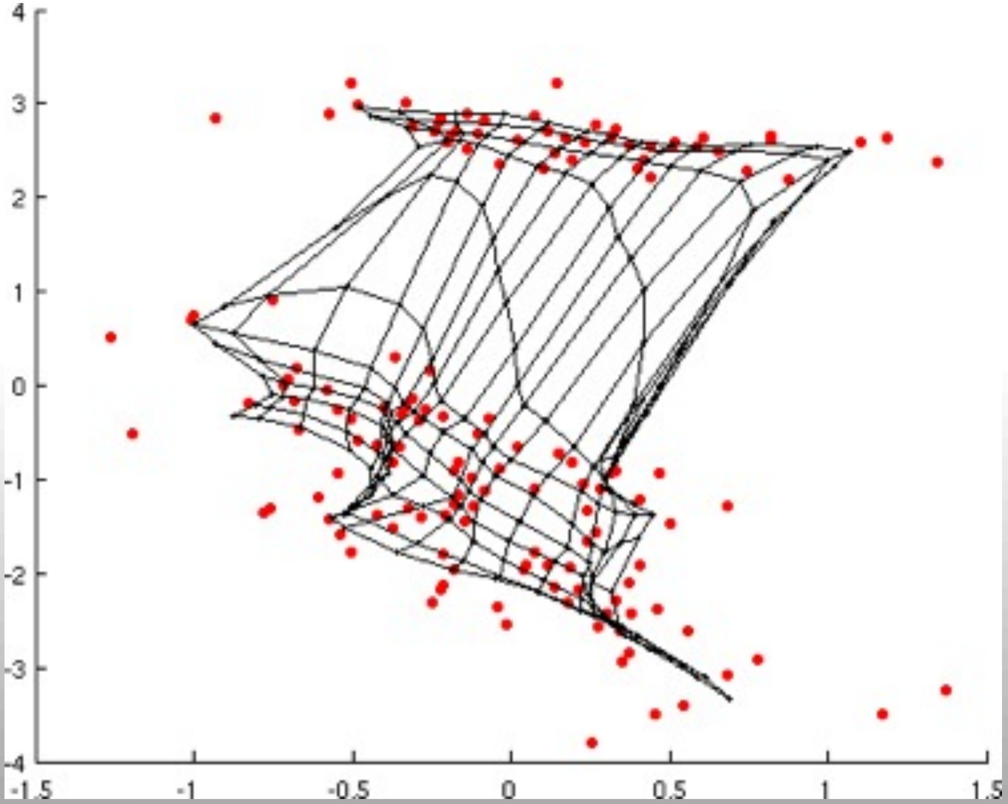
Density Estimation

- Dataset describes a sample from a distribution
- Describe distribution is simple terms



Density Estimation

- Other methods also take into account the spatial relationships between prototypes
- Self-Organizing Map (SOM)



Universiteit Leiden

Views on Data Mining

- Fitting the data
- Density Estimation
- Learning
  - being able to perform a task more accurately than before
- Prediction
  - use the data to predict future data
- Compressing the data
  - capture the essence of the data
  - discard the noise and details

Learning

- Perform a task more accurately than before
- Learn to perform a task (at all)

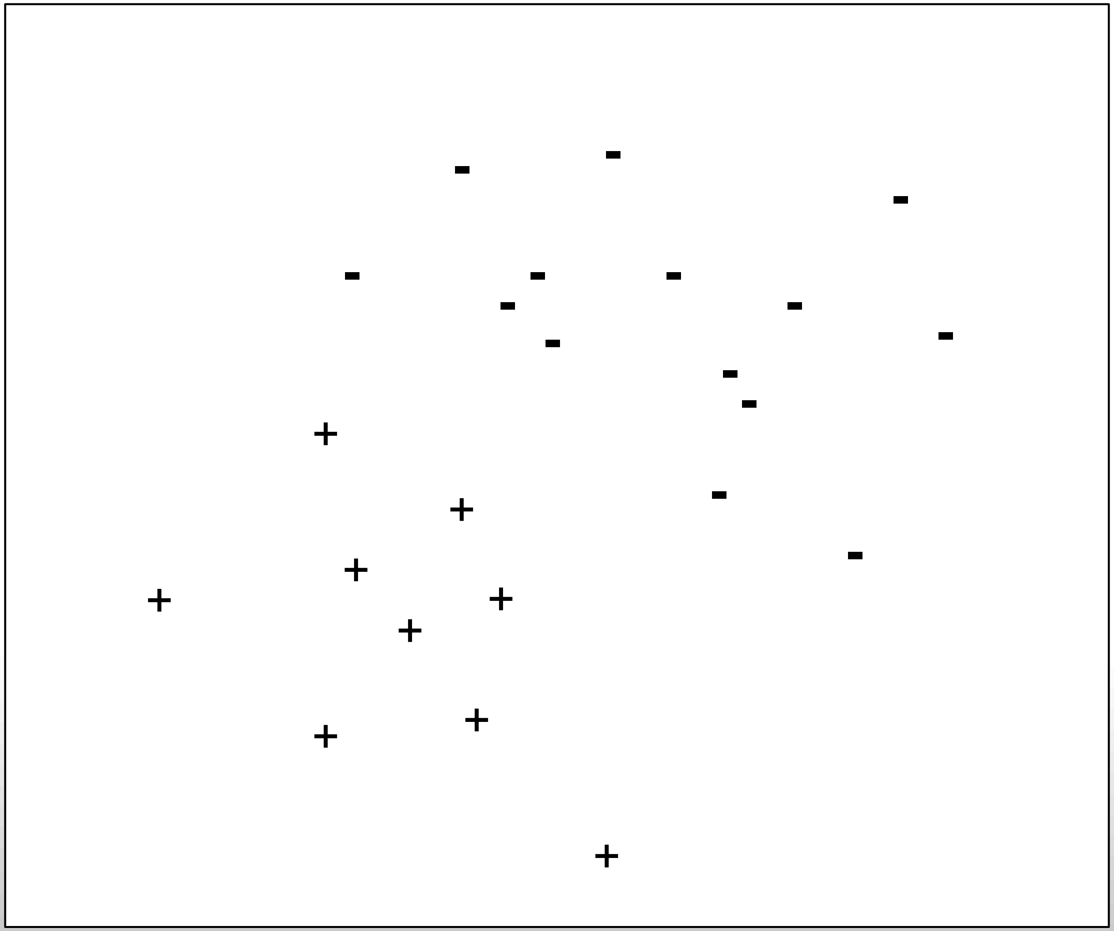
- Suggests an interaction between model and domain
  - perform some action in domain
  - observe performance
  - update model to reflect desirability of action
- Often includes some form of experimentation

- Not so common in Data Mining
  - often static data (warehouse), observational data

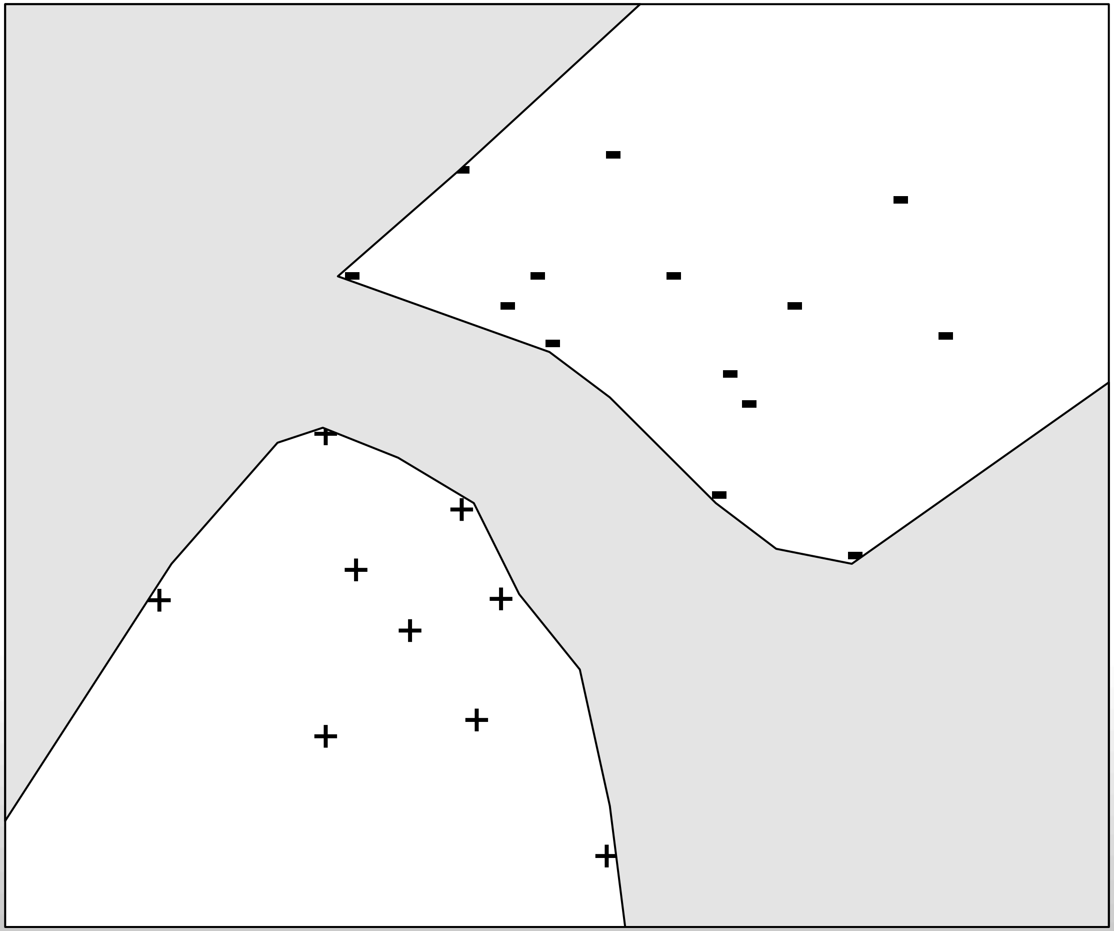
Views on Data Mining

- Fitting the data
- Density Estimation
- Learning
  - being able to perform a task more accurately than before
- Prediction
  - use the data to predict future data
- Compressing the data
  - capture the essence of the data
  - discard the noise and details

Prediction: learning a decision boundary

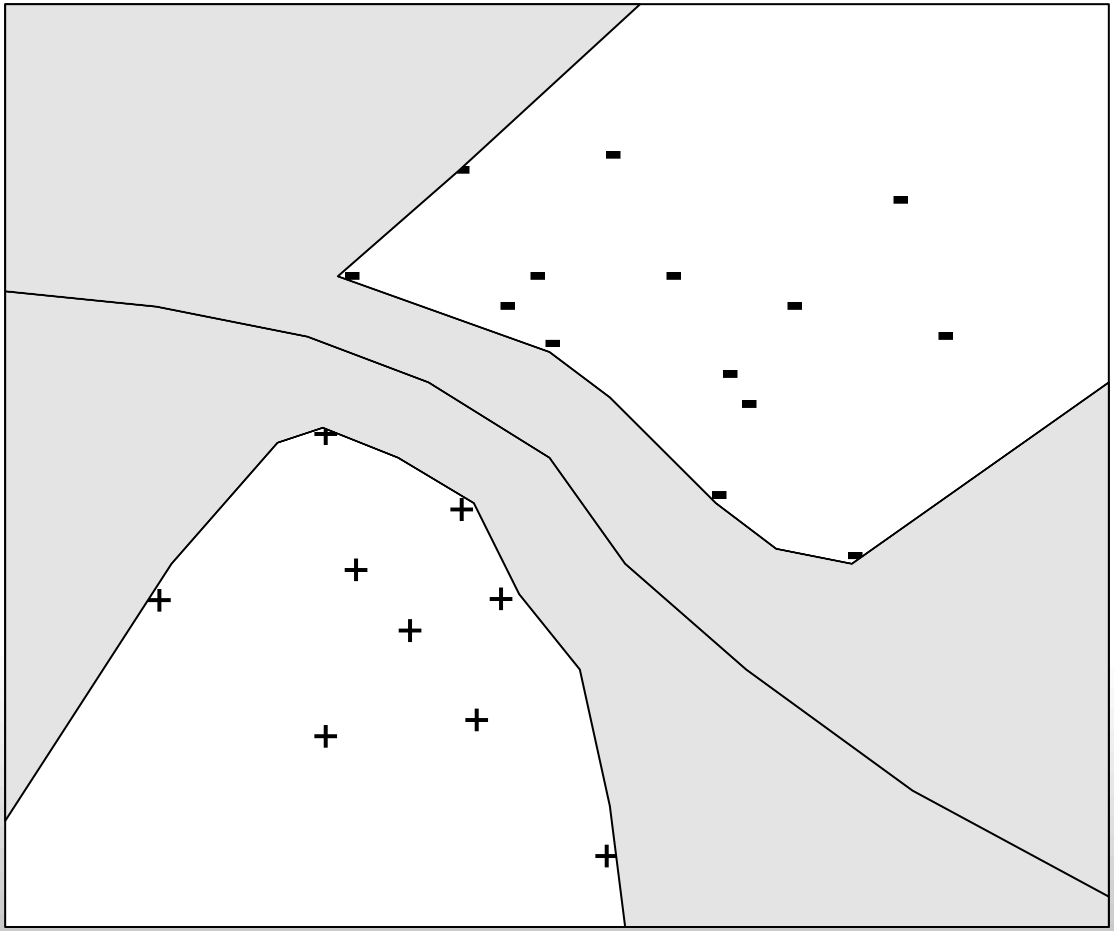


Prediction: learning a decision boundary

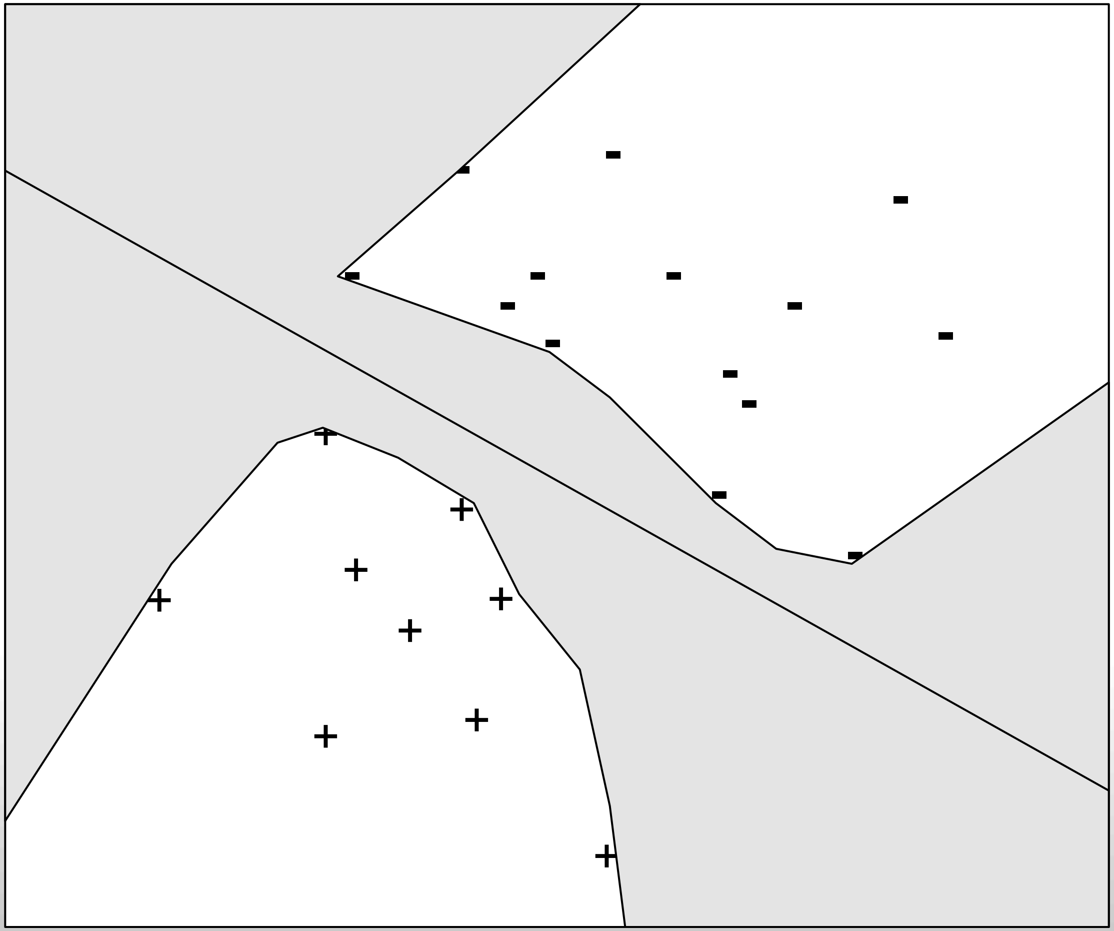


Universiteit Leiden

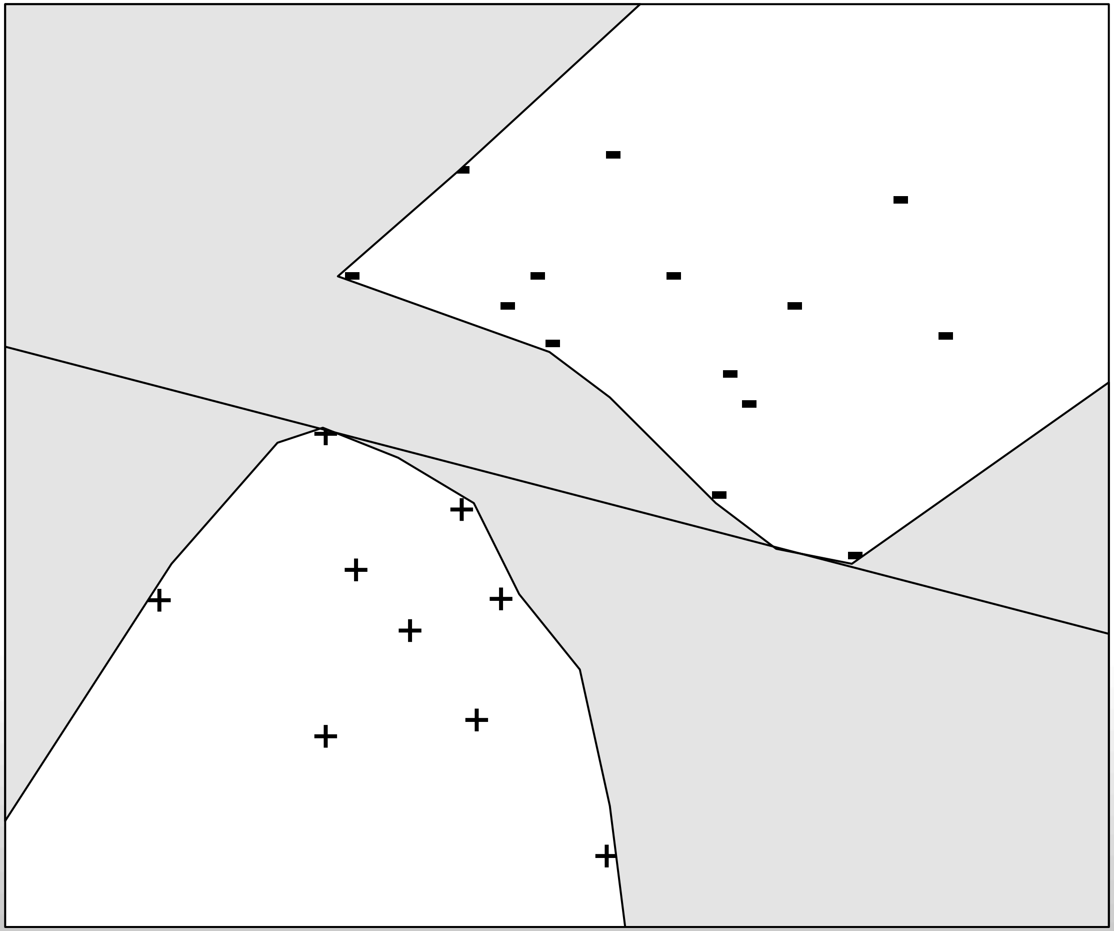
Prediction: learning a decision boundary



Prediction: learning a decision boundary



Prediction: learning a decision boundary



Views on Data Mining

- Fitting the data
- Density Estimation
- Learning
  - being able to perform a task more accurately than before
- Prediction
  - use the data to predict future data
- Compressing the data
  - capture the essence of the data
  - discard the noise and details

Compression

- Compression is possible when data contains structure (repeting patterns)
- Compression algorithms will discover structure and replace that by short code
- Code table forms interesting set of patterns

A	B	C	D	E	F
1	0	1	1	0	0
1	1	1	1	1	0
0	1	0	1	1	0
1	1	1	1	0	1
…	…	…	…	…	…

Compression

- Compression is possible when data contains structure (repeting patterns)
- Compression algorithms will discover structure and replace that by short code
- Code table forms interesting set of patterns

A	B	C	D	E	F
1	0	1	1	0	0
1	1	1	1	1	0
0	1	0	1	1	0
1	1	1	1	0	1
…	…	…	…	…	…

Compression

- Compression is possible when data contains structure (repeting patterns)
- Compression algorithms will discover structure and replace that by short code
- Code table forms interesting set of patterns

A	B	C	D	E	F
1	0	1	1	0	0
1	1	1	1	1	0
0	1	0	1	1	0
1	1	1	1	0	1
…	…	…	…	…	…

•Pattern ACD appears frequently

•ACD helps to compress the data

•ACD is a relevant pattern to report

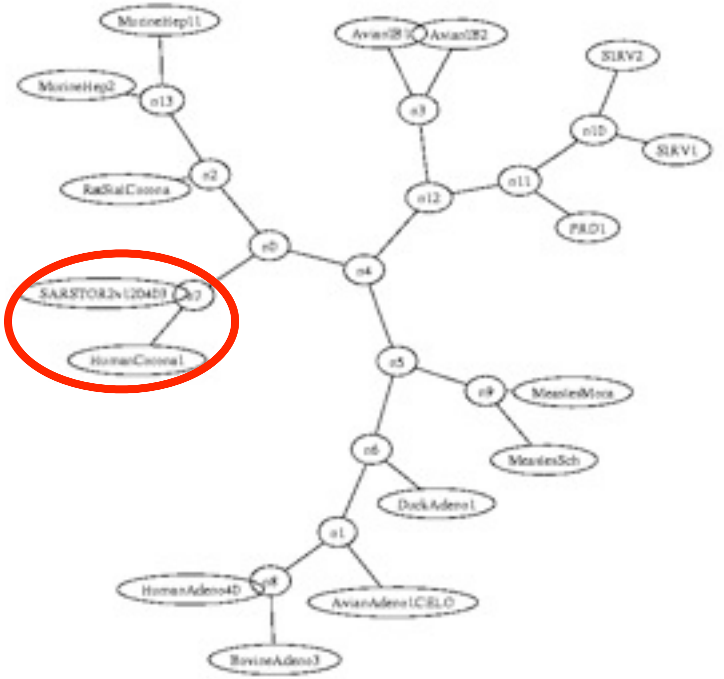
Compression

Paul Vitanyi (CWI, Amsterdam)

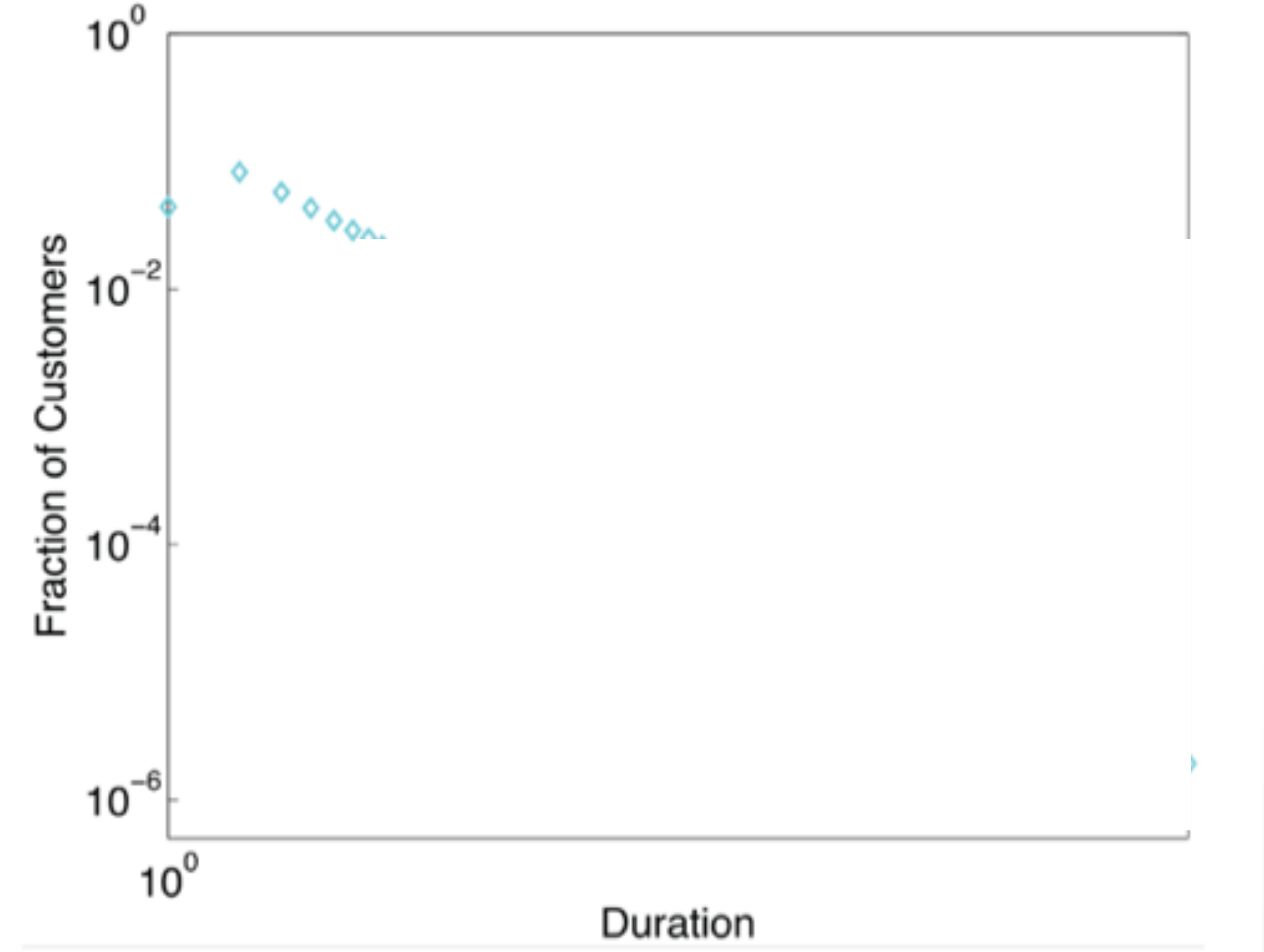
- Software to unzip identity of unknown composers
  - Beethoven, Miles Davis, Jimmy Hendrix

- SARS virus similarity

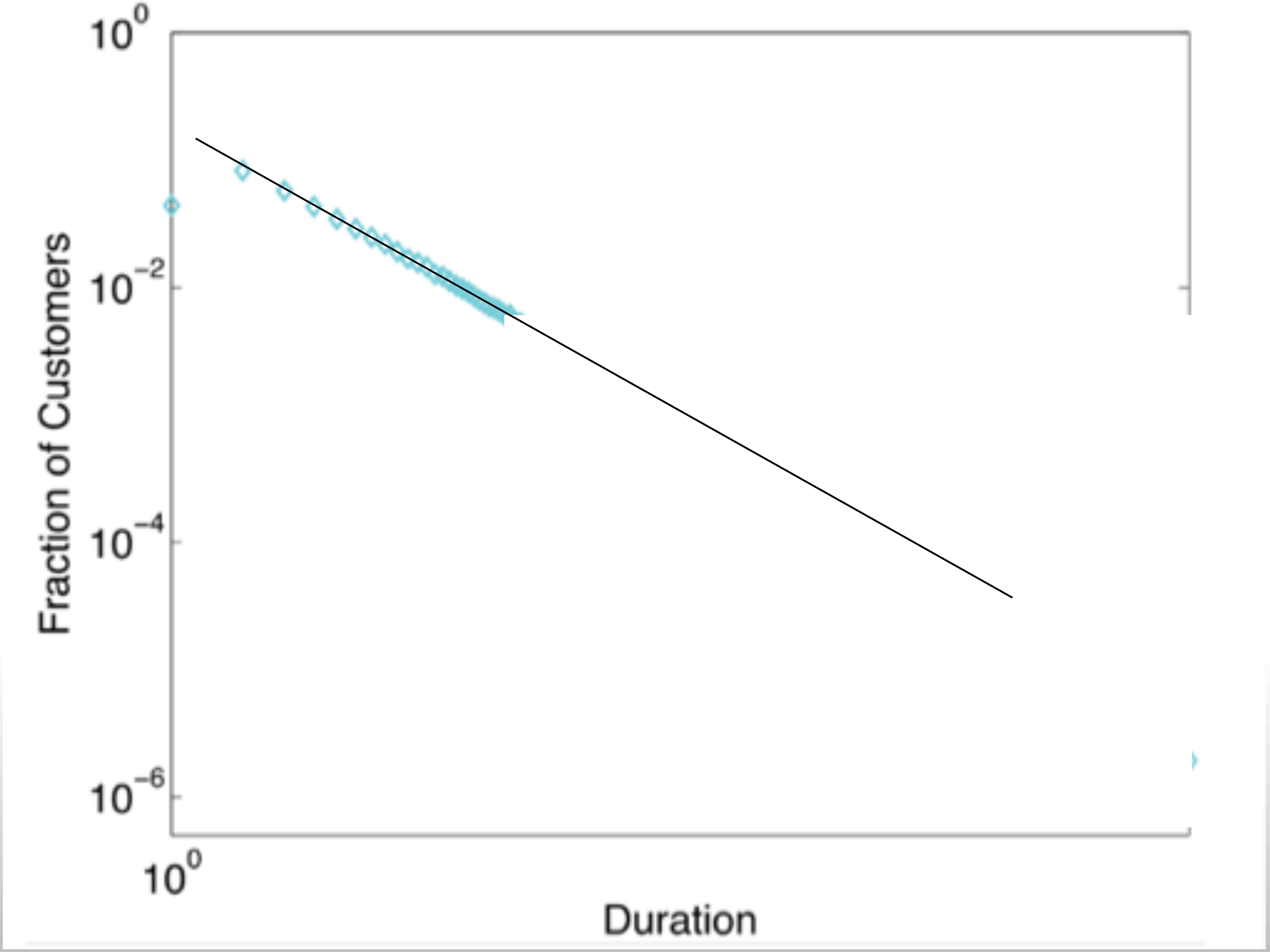
- internet worms, viruses
- intruder attack traffic
- images, video, …



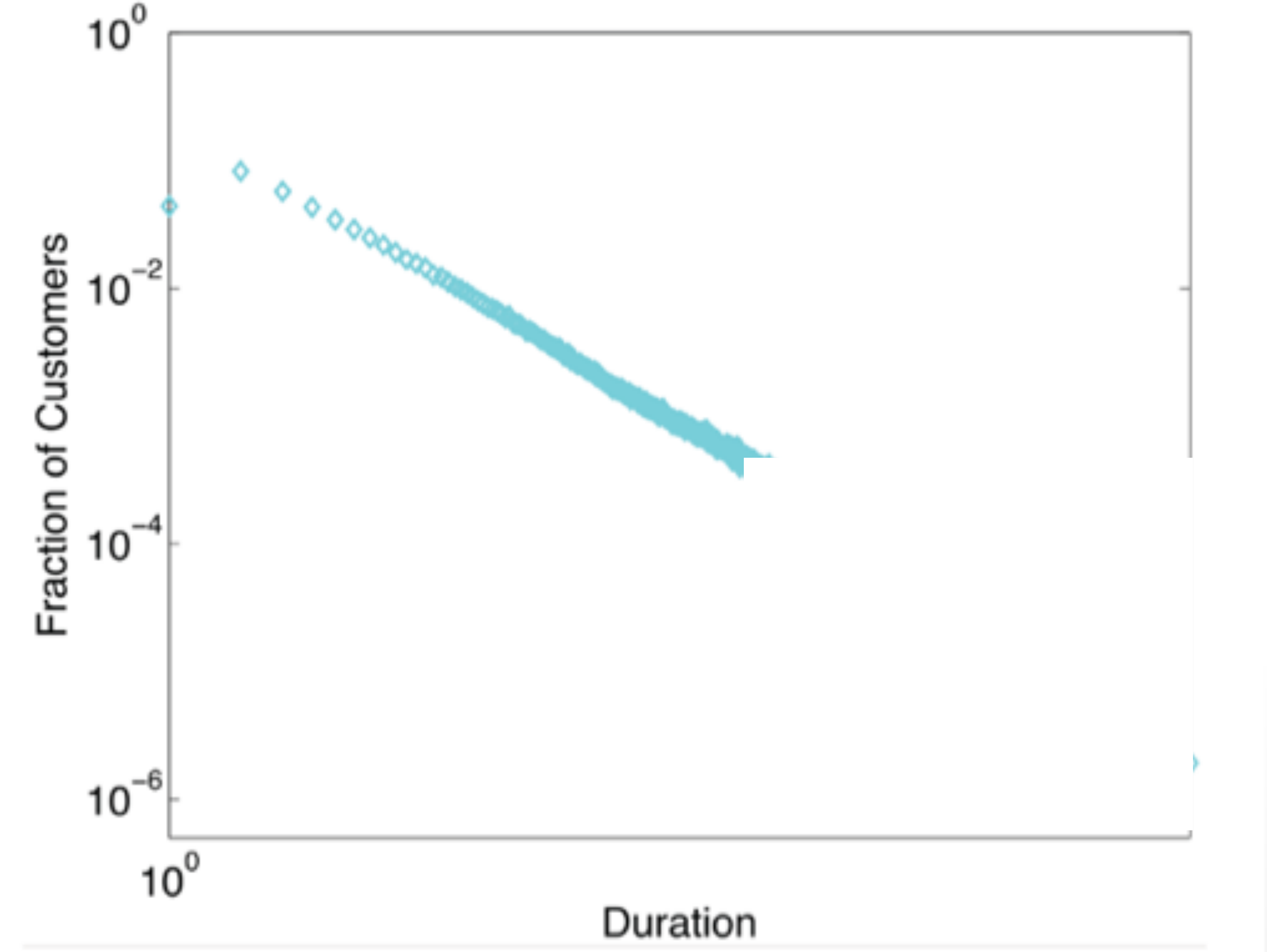
Mobile calls: modeling duration of calls



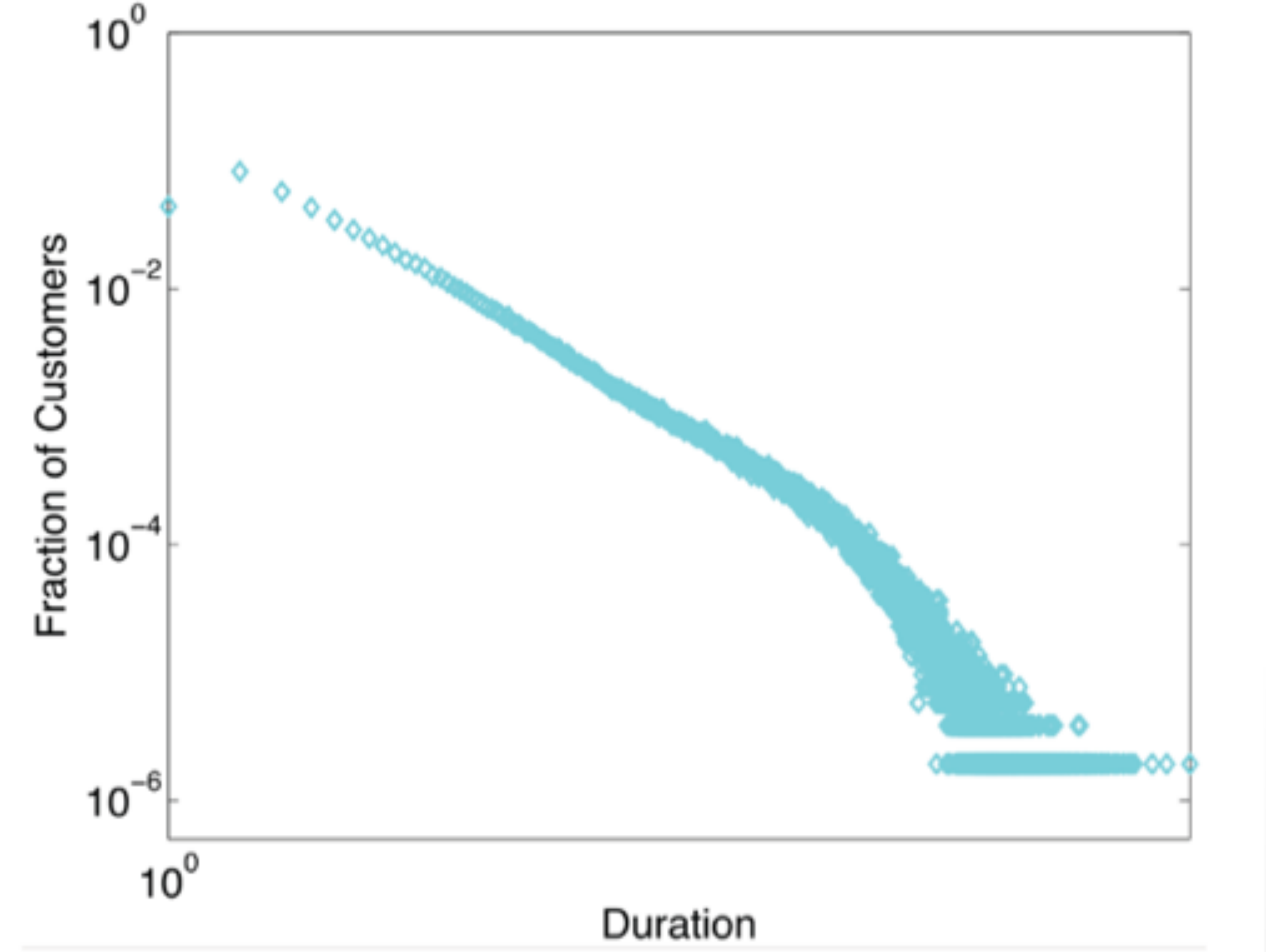
More data: linear model



Even more data: still linear?



Hmmm


Fraction of Customers
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{0}
Duration