Datascope 2013

Data Mining

Fundamental Concepts

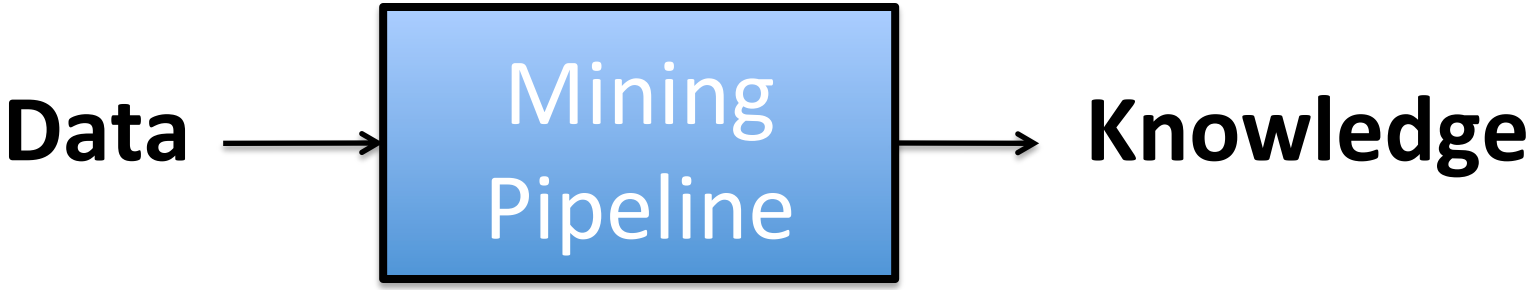
Ceyhun Burak Akgül, PhD

www.cba-research.com

In This Talk…

- What is data mining?
- What is it useful for?
- Where is it applied?
- How does it work?
- Does it always work?
- The Context: What is in it for us?

From Data to Knowledge



Data

Mining Pipeline

Knowledge

The objective of data mining is
to transform data into knowledge

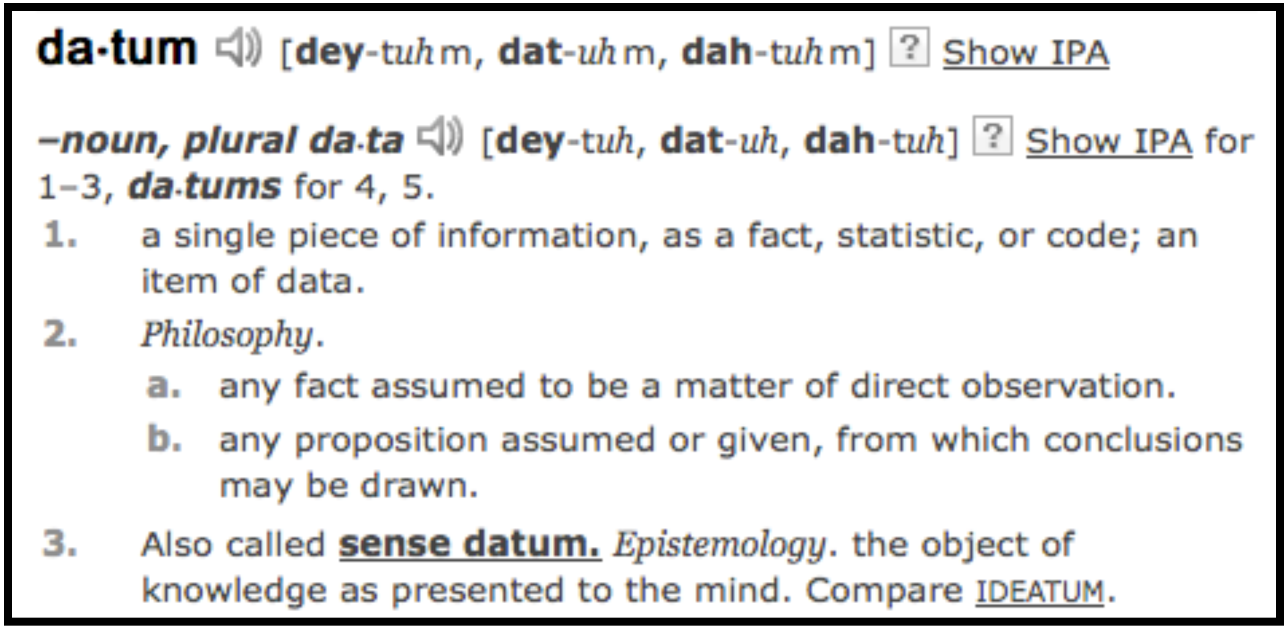
Preliminaries

What is data?

Preliminaries

What is data?

Dictionary.com definition



da·tum 🔊 [**dey**-t*uh*m, **dat**-*uh*m, **dah**-t*uh*m] ? Show IPA

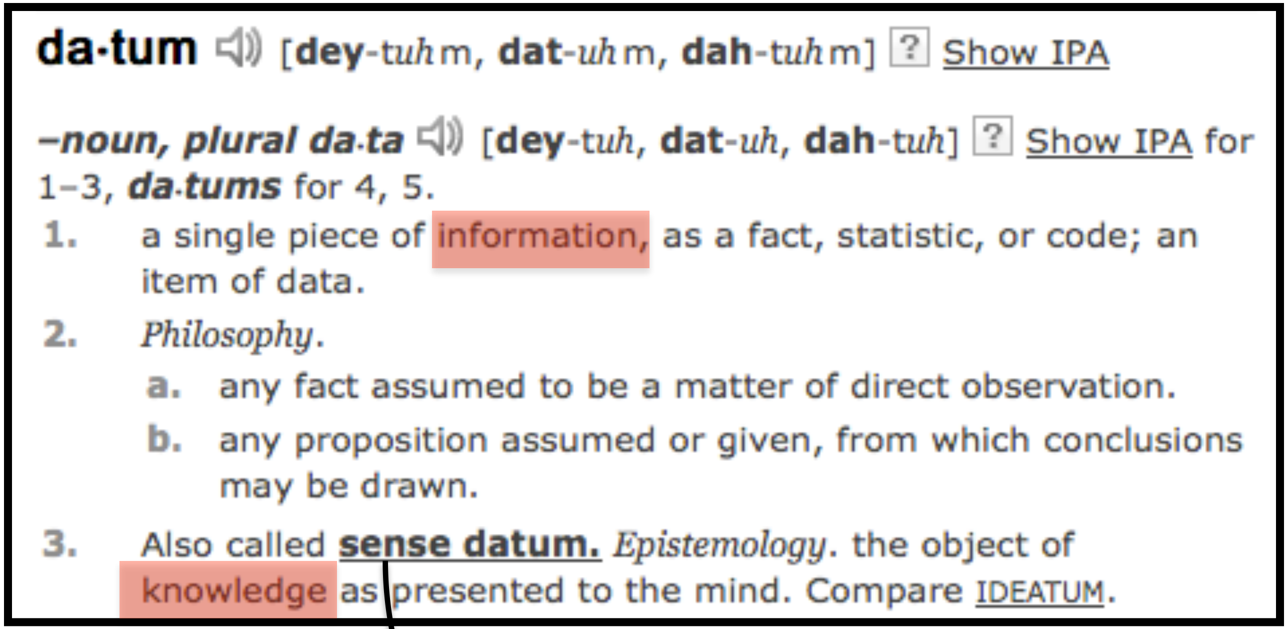
–noun, plural ***da·ta*** 🔊 [**dey**-t*uh*, **dat**-*uh*, **dah**-t*uh*] ? Show IPA for 1–3, ***da·tums*** for 4, 5.

1. a single piece of information, as a fact, statistic, or code; an item of data.
2. *Philosophy*.
   a. any fact assumed to be a matter of direct observation.
   b. any proposition assumed or given, from which conclusions may be drawn.
3. Also called **sense datum.** *Epistemology*. the object of knowledge as presented to the mind. Compare IDEATUM.

Preliminaries

What is data?

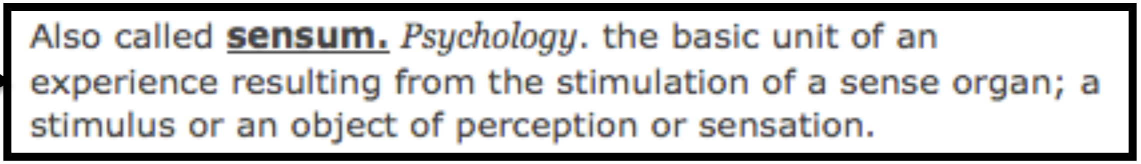
Dictionary.com definition



da·tum [**dey**-t*uh*m, **dat**-*uh*m, **dah**-t*uh*m] ? Show IPA

–noun, plural da·ta [**dey**-t*uh*, **dat**-*uh*, **dah**-t*uh*] ? Show IPA for 1–3, ***da·tums*** for 4, 5.

1. a single piece of information, as a fact, statistic, or code; an item of data.
2. *Philosophy*.
   - a. any fact assumed to be a matter of direct observation.
   - b. any proposition assumed or given, from which conclusions may be drawn.
3. Also called **sense datum.** *Epistemology*. the object of knowledge as presented to the mind. Compare IDEATUM.



Also called **sensum.** *Psychology*. the basic unit of an experience resulting from the stimulation of a sense organ; a stimulus or an object of perception or sensation.

Preliminaries

What is data?

- Objective facts about the perceived or sensed
- Measurements
- Physical instantiation/embodiment of *information*

Preliminaries

Where do data/information arise from?

- Humans
  - Our senses
  - Our likes/dislikes, favorite books, movies, music, …
  - Our identity, political views, web browsing sessions, …
  - Our genome, medical records, …
- Earth and Universe
  - Stars and planets, celestial motions, …
  - Climate, geography, oceans, …
- Organizations/Companies
  - Expenses and revenues, stock rates, …
- …

A Note on the Things to Come…

In this course, we will consider
The City in general, ***Tarlabaşı*** in particular
as our source of information (data).

Preliminaries

What is knowledge?

Preliminaries

What is knowledge?

Knowledge in general is the subject matter of epistemology:

- Plato: Knowledge is **justified true belief (JTB)**
  - Is it sufficient?
  - Gettier (1963): There are cases where JTB is not sufficient
- Aristotle considers knowledge from the **causality** perspective: we have knowledge when we know the cause on which a certain fact (data) depends.
  - How do we diagnose causality?

Preliminaries

"Information is not knowledge."
– Albert Einstein

"We are drowning in information and starving for knowledge."
– Rutherford D. Roger

What might these people have meant?

Preliminaries

"Information is not knowledge."

– Albert Einstein

"We are drowning in information and starving for knowledge."

– Rutherford D. Roger

What might these people have meant?

Information is the *preimage* (or raw form) of **knowledge**

What is Data Mining?

Data mining is

The semi-automatic process of extracting important patterns and trends from data arising in a specific domain in order to:

- **Predict** the future outcomes of a “system” from past observations (i.e., data)
- **Explain** the underlying rules and dynamics that generate the data
- **Understand** “what the data says” about a particular question (hypothesis) of interest

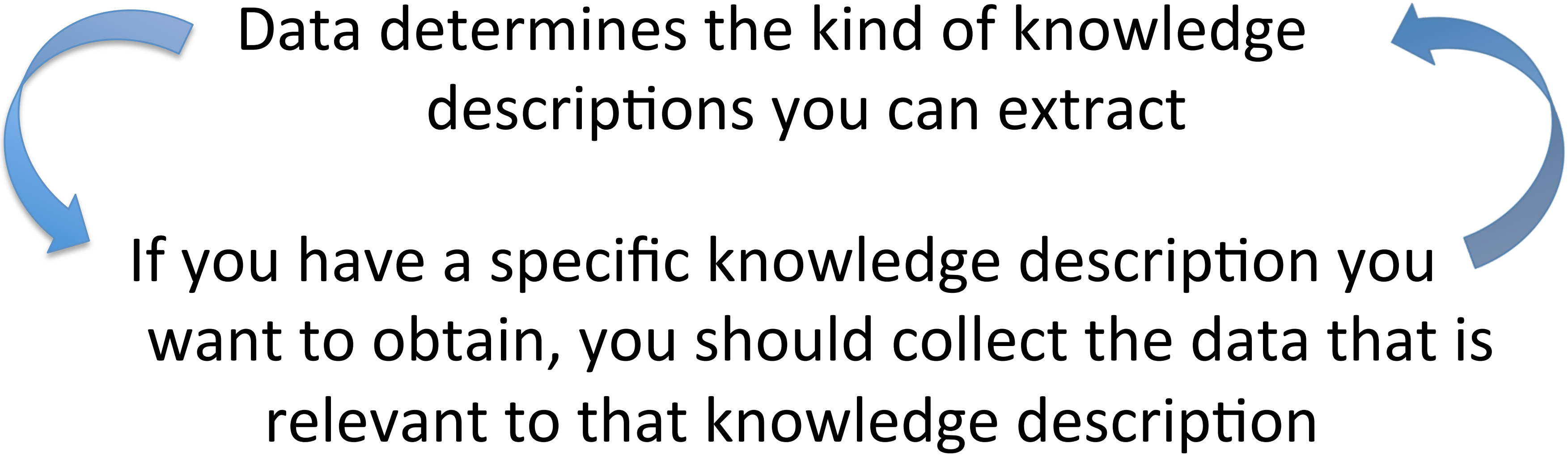
What is Data Mining?

- **Predict** the future outcomes of a "system" from past observations (data)
- **Explain** the underlying rules and dynamics that generate the data
- **Understand** "what the data says" about a particular question (hypothesis) of interest

*If we are able to (partially) perform one or more of these tasks, we're said to possess **(partial) operational knowledge** on that particular domain*

A Remark on Data Collection

Where/How do we get data?



Data determines the kind of knowledge descriptions you can extract

If you have a specific knowledge description you want to obtain, you should collect the data that is relevant to that knowledge description

What is Data Mining Useful for?

Prediction

- Predict the amount of rain that will fall in Istanbul during June 2013

  Data: monthly records over the past ten years

- Predict whether an elderly patient will develop Alzheimer's within the years to come

  Data: medical records (tests, MR scans, …) and similar about other Alzheimer's patients

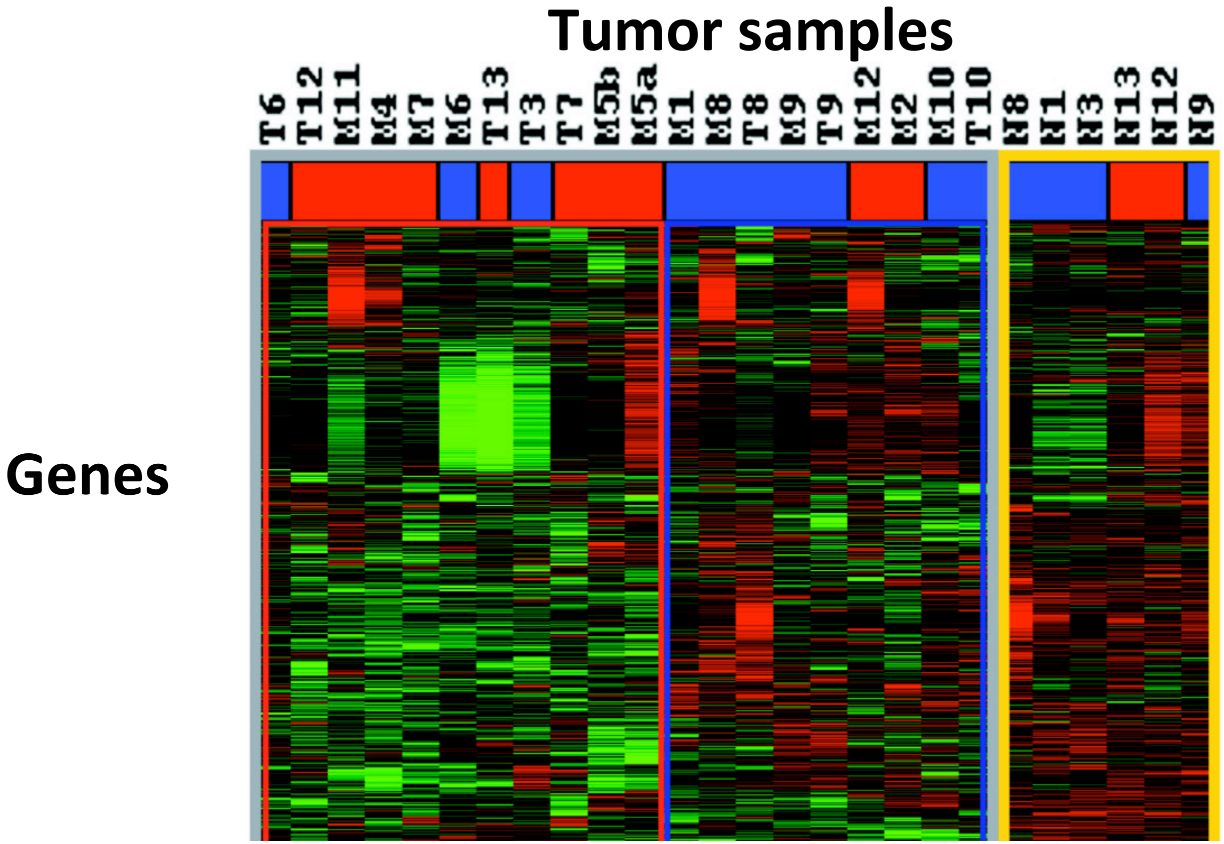
- Predict the price of a stock in six months

  Data: company performance and economic data

- …

What is Data Mining Useful for?

Explanation/Understanding

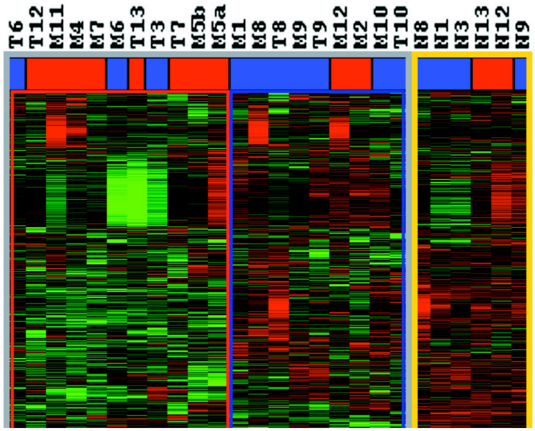


>>

Gene expression profile dataset

What is Data Mining Useful for?

Explanation/Understanding



Given a gene expression dataset:

- Which samples are most similar to each other, in terms of their expression profiles across genes?
- Which genes are most similar to each other, in terms of their expression profiles across samples?
- Do certain genes show very high (or low) expression for certain cancer samples?

Applications: Classical

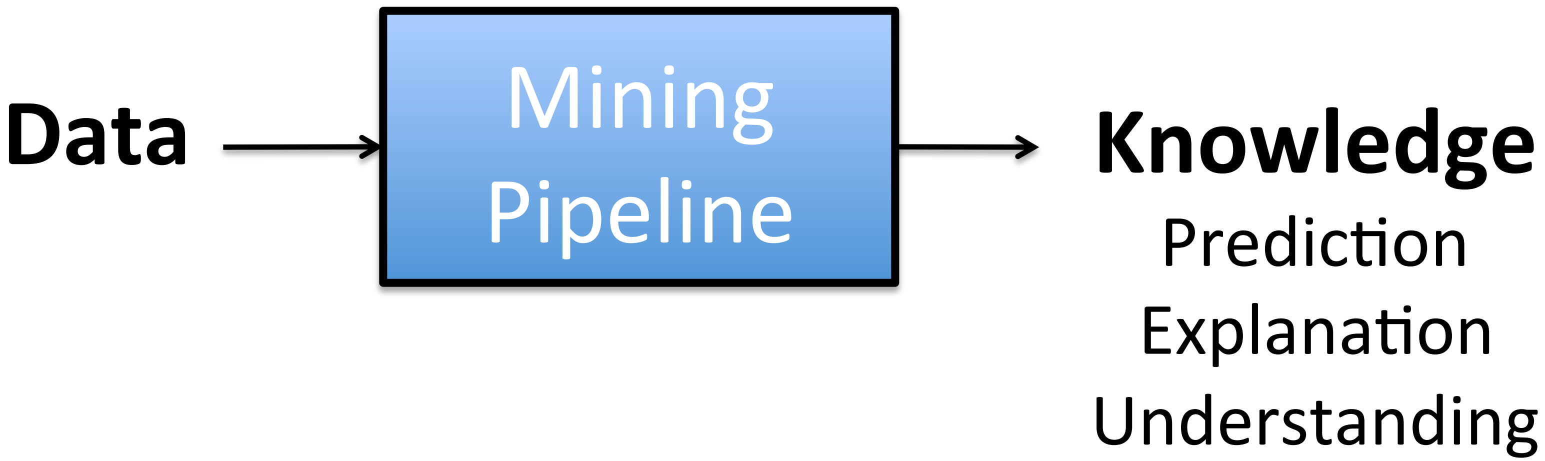
- Marketing (e.g., sales analysis)
- Banking (e.g., credit and loan approval)
- Medicine / Biology / Pharmacology
- Manufacturing (e.g., yield analysis)
- Finance (e.g., stock prediction)
- E-Commerce / Web (e.g., hits analysis)

>>

Applications: Less Classical

- Multimedia search engines
  - Content search by audiovisual similarity
- Multimedia content management
  - Automatic content categorization and annotation
  - Audiovisual concept detection
- Image-based medical diagnosis
  - Visual biomarker discovery from medical images

How Does Data Mining Work?



Data

Mining Pipeline

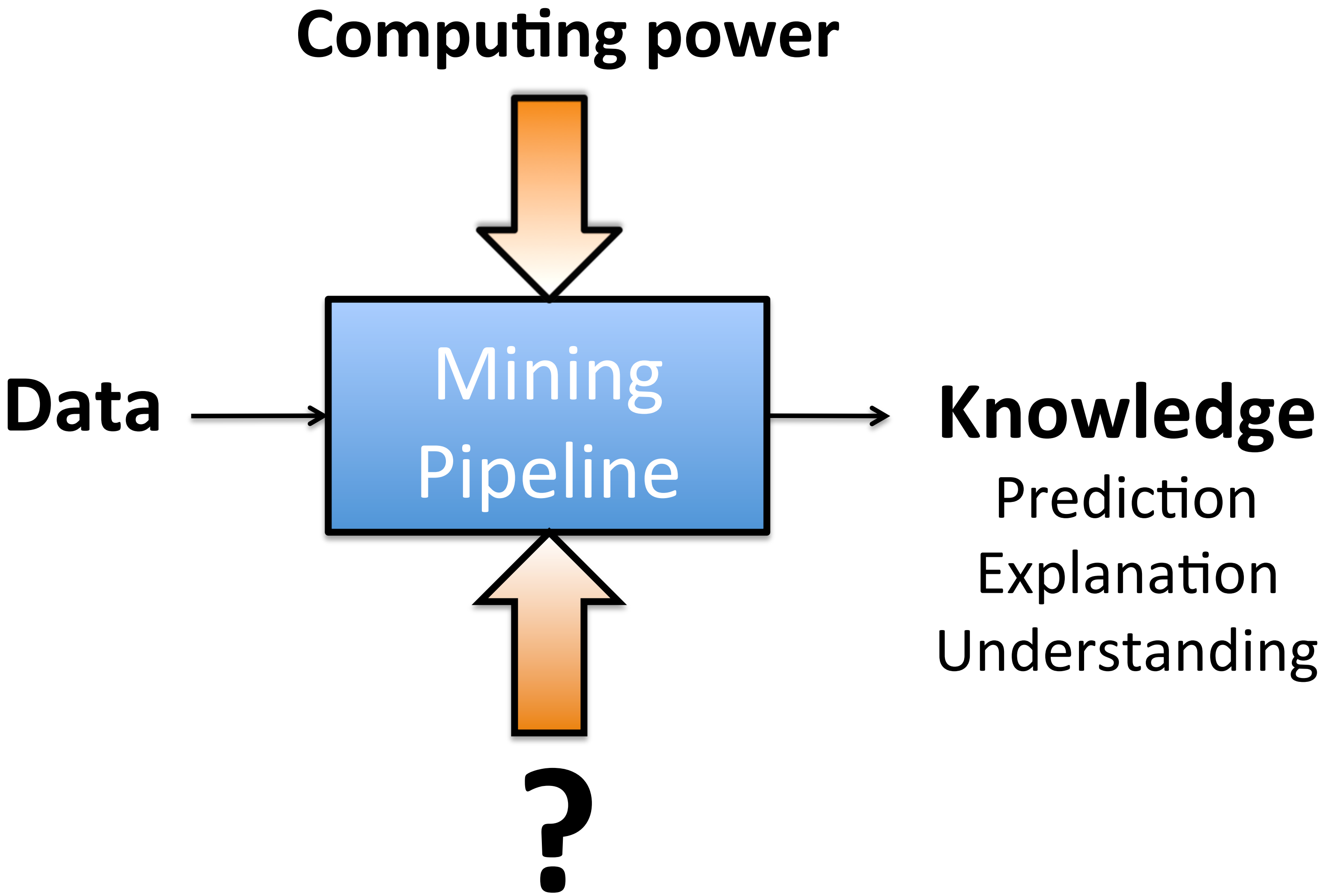
Knowledge

Prediction

Explanation

Understanding

How Does Data Mining Work?



Computing power

Data

Mining Pipeline

Knowledge

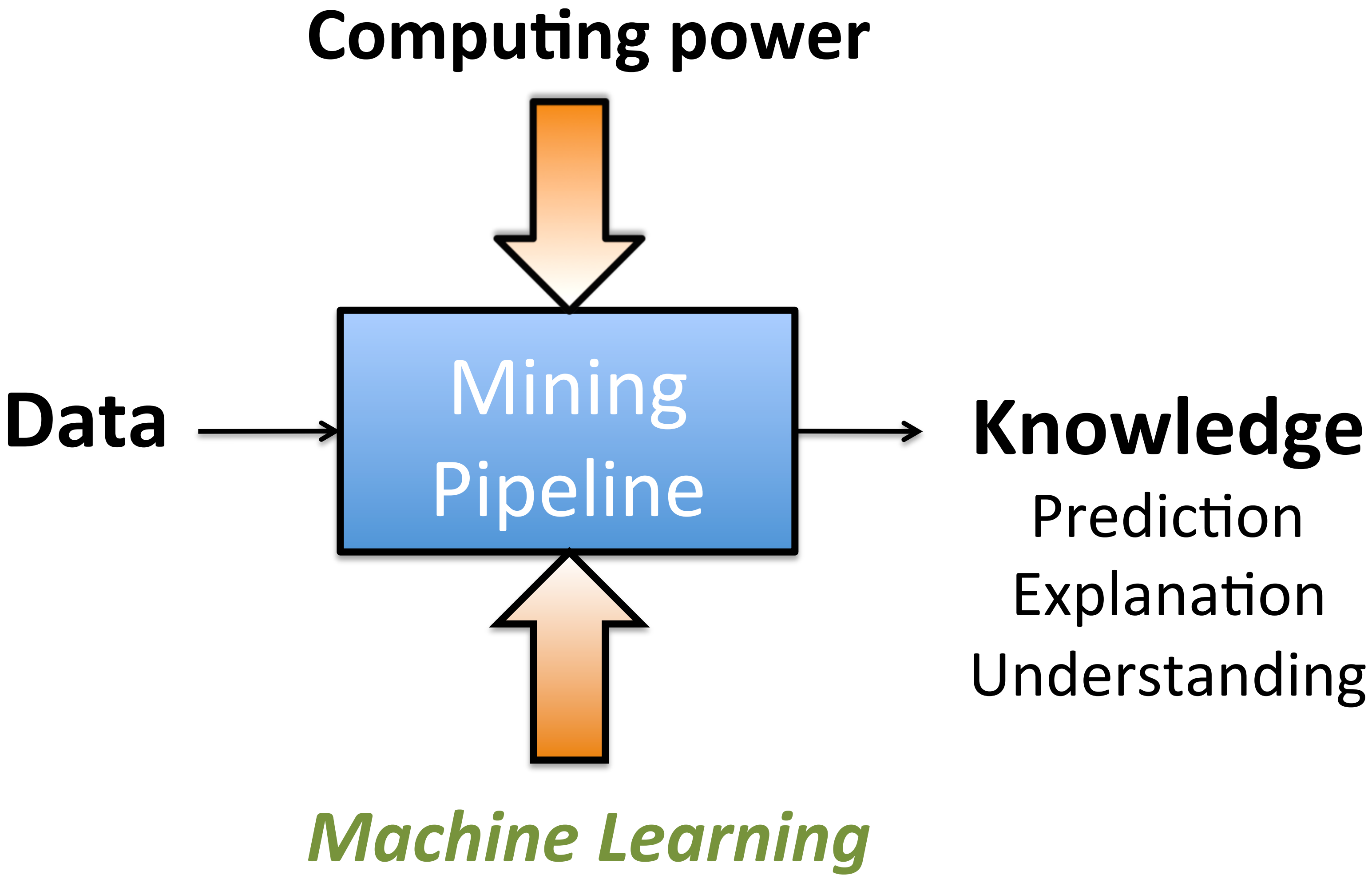
Prediction

Explanation

Understanding

?

How Does Data Mining Work?



Computing power

Data

Mining Pipeline

Knowledge

Prediction

Explanation

Understanding

Machine Learning

How Does Data Mining Work?

If data are the fuel of data mining,
***Machine Learning** is its engine.*

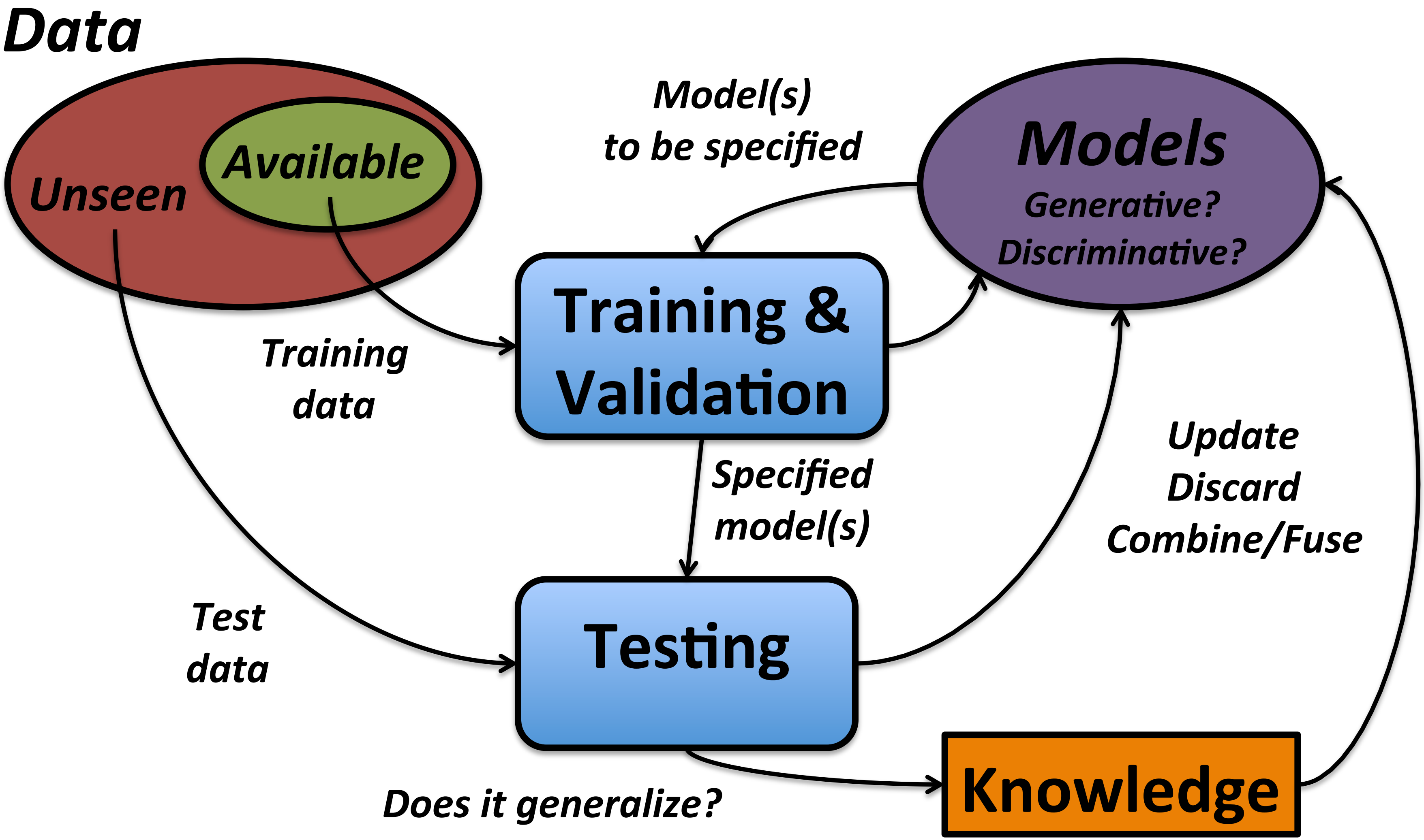
- Not quite like human learning:

  Computers have no awareness!
- Though not quite like a calculator neither

Machine learning involves

- Specification of a *statistical* model (generative, discriminative, or both)
- *Training* the model with available data
- *Testing* the model with new (unseen) data

Machine Learning in One Picture

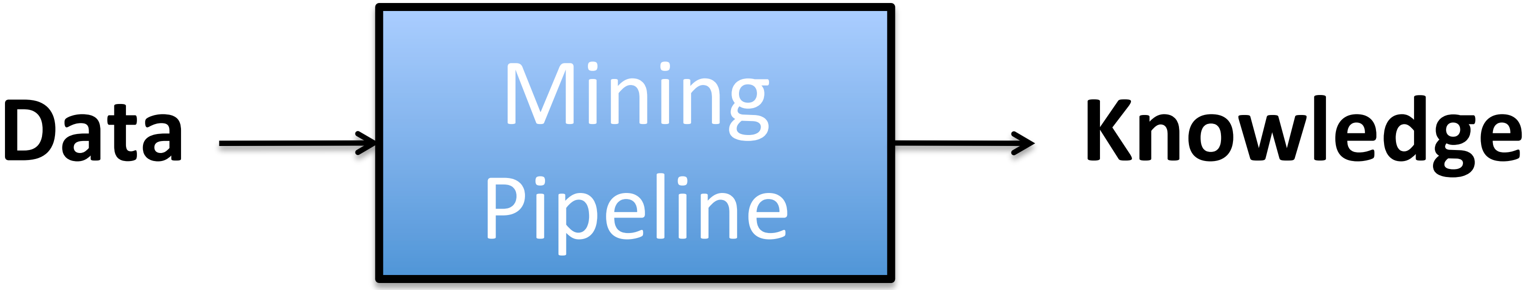


Does it always work?

Data mining doesn't work when/if

- You don't have (sufficient) data
- You ask the wrong question
- You just focus on training
- You rely on just one technique
- You mix apples with bananas
- You don't use your intuition
- You use your intuition
- You try to answer every question
- …

Data Mining Pipeline – *again*

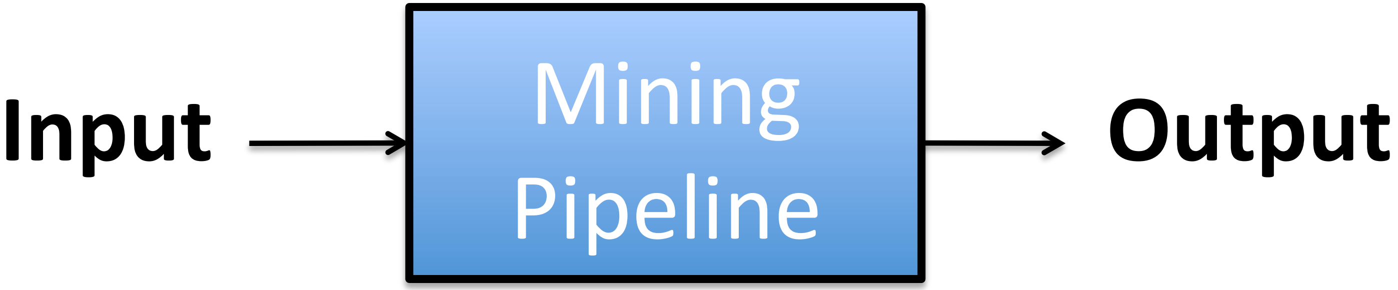


Data

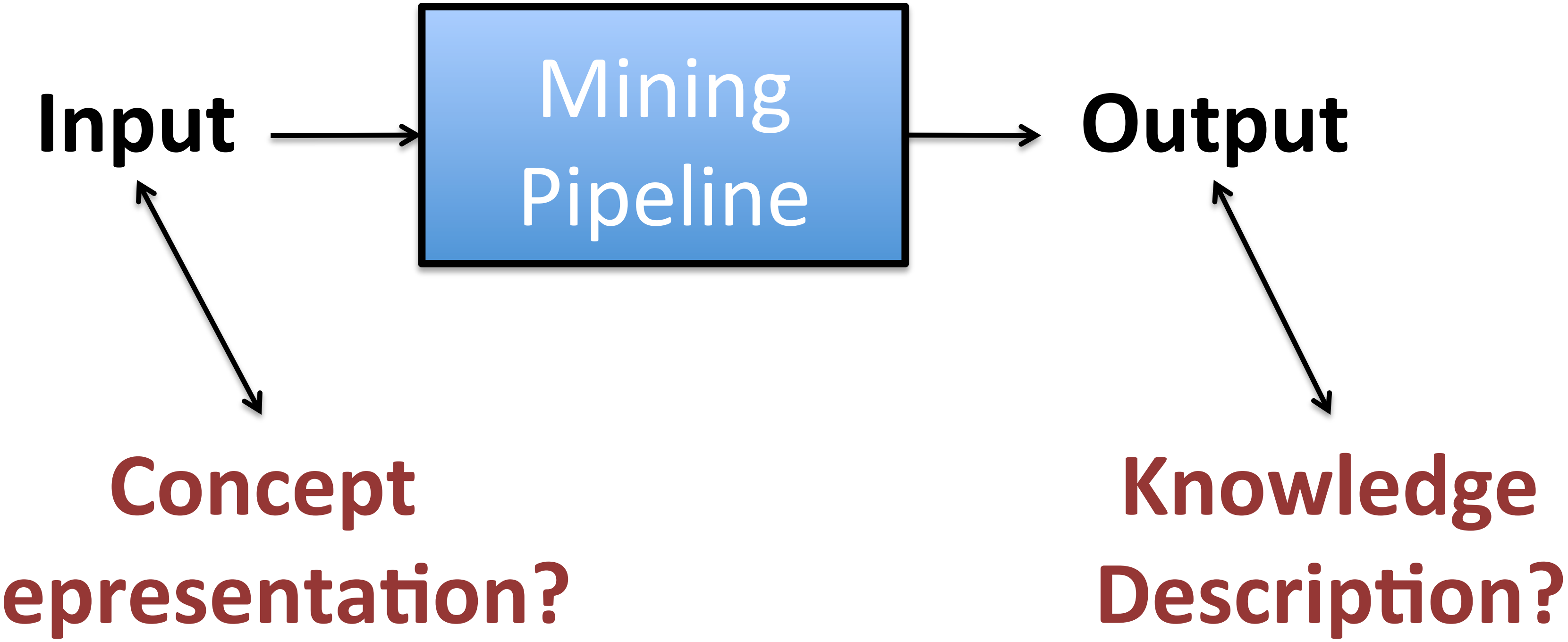
Mining Pipeline

Knowledge

Data Mining Pipeline – *again*



Data Mining Pipeline – *again*



Input

Mining Pipeline

Output

Concept Representation?

Knowledge Description?

Inputs: Concept Representation

“What is it in a name? That which we call a rose
By any other name would smell as sweet”

Shakespeare

Inputs: Concept Representation

- **A concept is an abstraction of a physical thing**

  [Recall Plato's *allegory of cave*]

- Inputs are **instantiations** of a concept

Inputs: Concept Representation

- **A concept is an abstraction of a physical thing**

  [Recall Plato's *allegory of cave*]

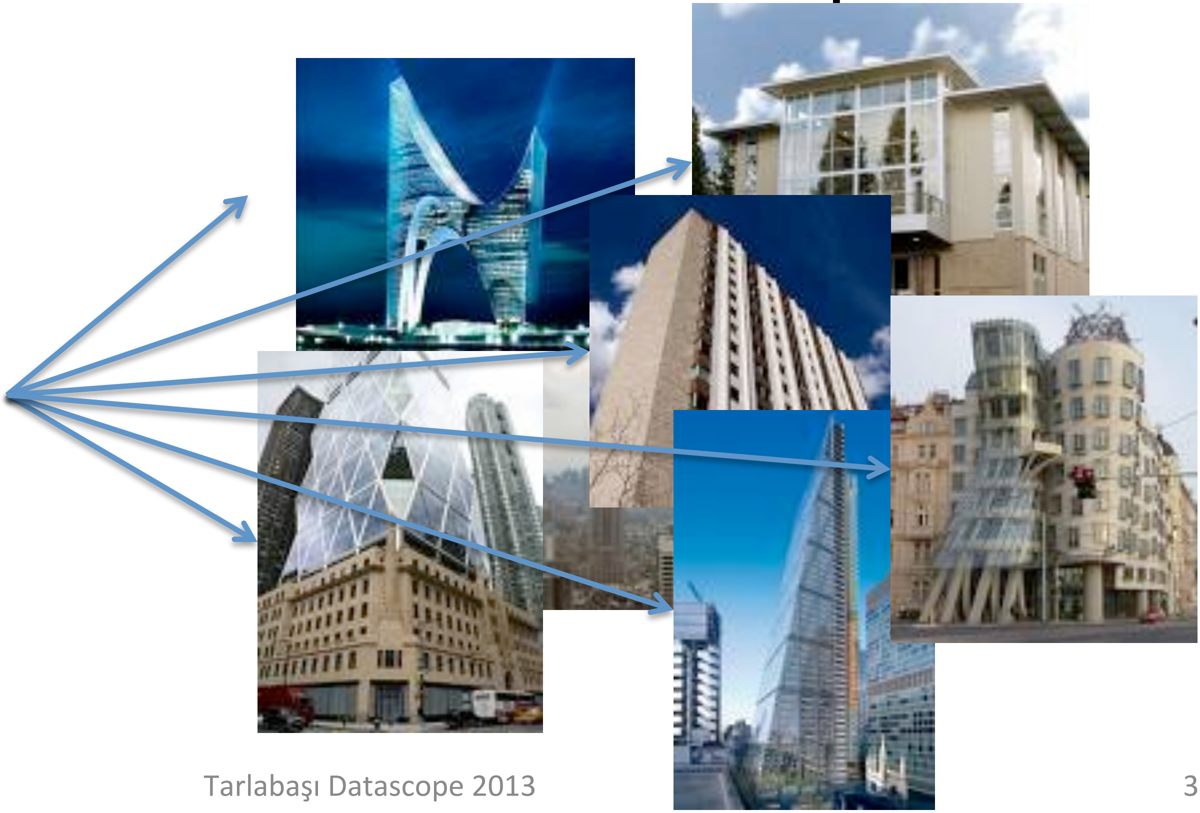
- Inputs are **instantiations** of a concept

Building

Inputs: Concept Representation

- **A concept is an abstraction of a physical thing**

  [Recall Plato’ s *allegory of cave*]
- Inputs are **instantiations** of a concept



Building

Inputs: Concept Representation

How do we represent a concept?

Inputs: Concept Representation

How do we represent a concept?

Consider the concept “Karnıyarık”

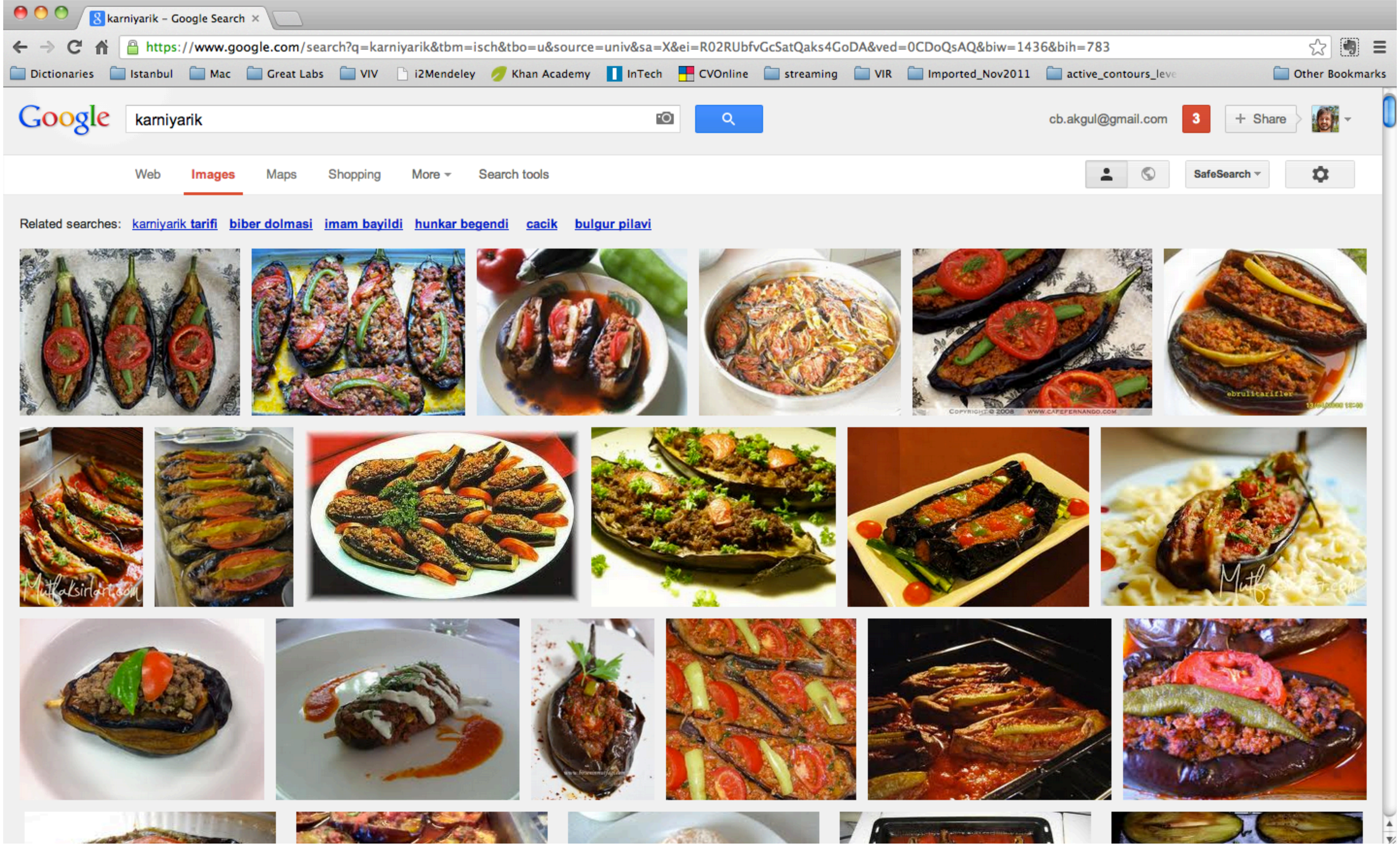
A traditional Turkish hot dish prepared by stuffing an eggplant with minced meat, finely sliced tomatoes, onion, garlic; flavored with spices…

It tastes good when made with competence…

It’s a summer dish…

…

Googling for “karnıyarık”…



https://www.google.com/search?q=karniyarik&tbm=isch&tbo=u&source=univ&sa=X&ei=R02RUbfvGcSatQaks4GoDA&ved=0CDoQsAQ&biw=1436&bih=783

karniyarik

Inputs: Concept Representation

How do we represent a concept?

In data mining:

- We need operational representations
- We need a way to **encode the available information**
- **Attributes** are a means to encode the information
- A concept is thus represented by a set of attributes
- Attributes may not necessarily cover the whole semantic field of the concept*

Inputs: Concept Representation

How do we represent a concept?

Let’s represent “Karnıyarık”:

Inputs: Concept Representation

How do we represent a concept?

Let’s represent “Karnıyarık”:

- Meat
- Eggplant
- Tomato
- Onion
- Garlic
- Spice
- Cooking time
- Region
- Taste
- …

Attributes

An attribute is an atomic property of a concept

Inputs: Concept Representation

Input → Concept → Instances

An example is a particular instance of a concept

The attributes of an instance are specified

- – Meat (grams) = 250 ***–numeric***
- – Eggplant (piece) = 1 ***–numeric***
- – Tomato (piece) = 3 ***–numeric***
- – Onion (piece) = 2 ***–numeric***
- – Garlic (piece) = 1 ***–numeric***
- – Spice (spoon) = 1 ***–numeric***
- – Cooking time (minute) = 30 ***–numeric***
- – Taste (good/bad/mediocre) = good ***–nominal (categorical)***

Inputs: Concept Representation

Summary

(1) A **concept** is an abstraction of a physical thing

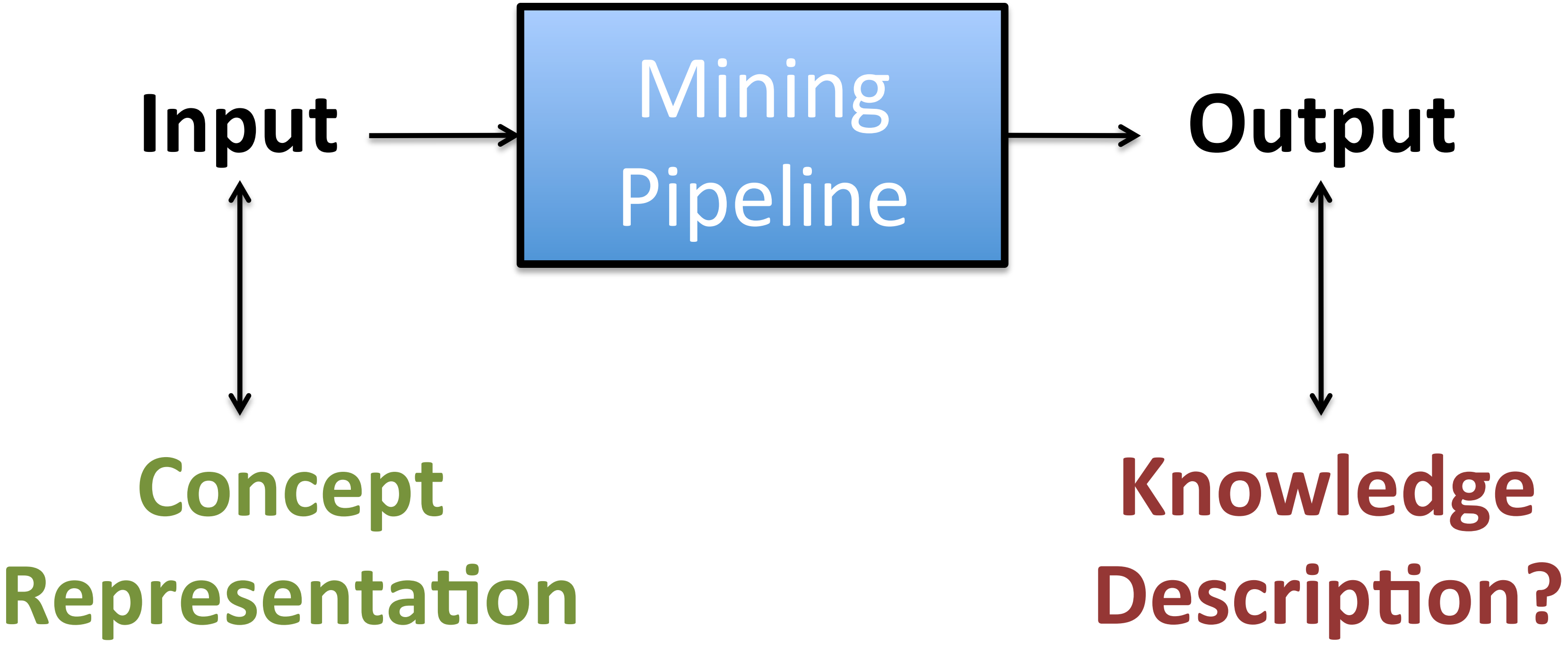
(2) Information about a concept is encoded by a set of attributes (or features)

(3) An **attribute** is an atomic property of a concept

(4) An **instance** is a particular realization of a concept

(5) Input to the data mining pipeline is a set of instances with specified attributes

Data mining pipeline – *again*



Input

Mining Pipeline

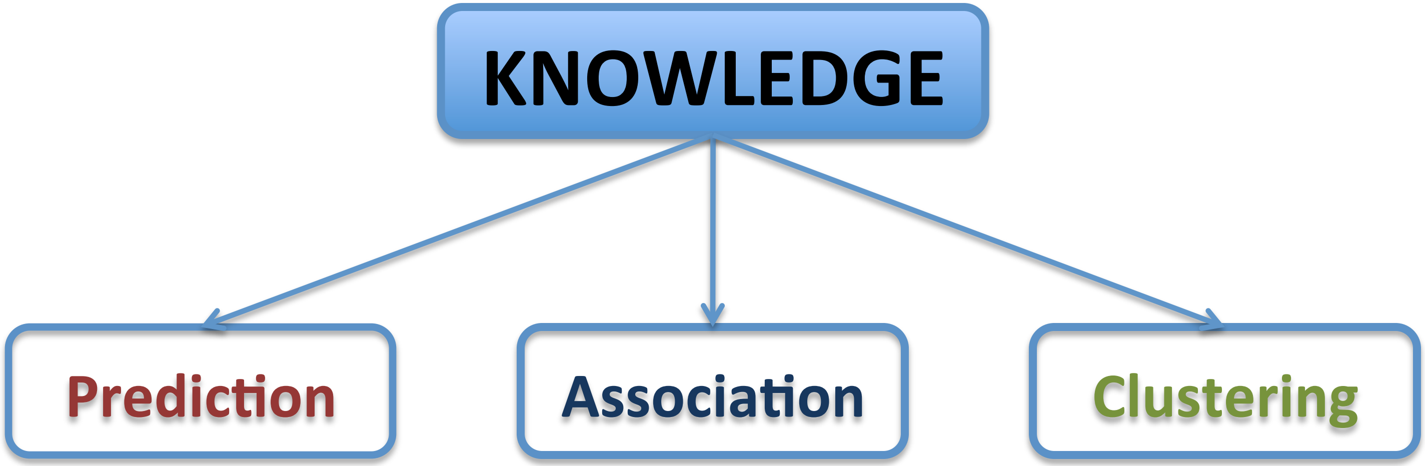
Output

Concept Representation

Knowledge Description?

Outputs: Knowledge Description

What kind of knowledge do we want to extract?



Outputs: Knowledge Description

How do we describe knowledge?

Prediction rules

- Classification

  Given the amount of each ingredient, cooking time, etc. find a rule predicting whether the "karnıyarık" will taste good, bad, or mediocre.

- Numeric prediction (Regression)

Outputs: Knowledge Description

How do we describe knowledge?

Association rules

Any regularity between two or more attributes can be expressed as an association rule

Ex1. If meat is X grams, then there should be Y pieces of eggplants (such that Y = aX)

Ex2. If tastes bad, then there were too much onion (>>2pieces) and cooking time was not enough (<10 mins)

…

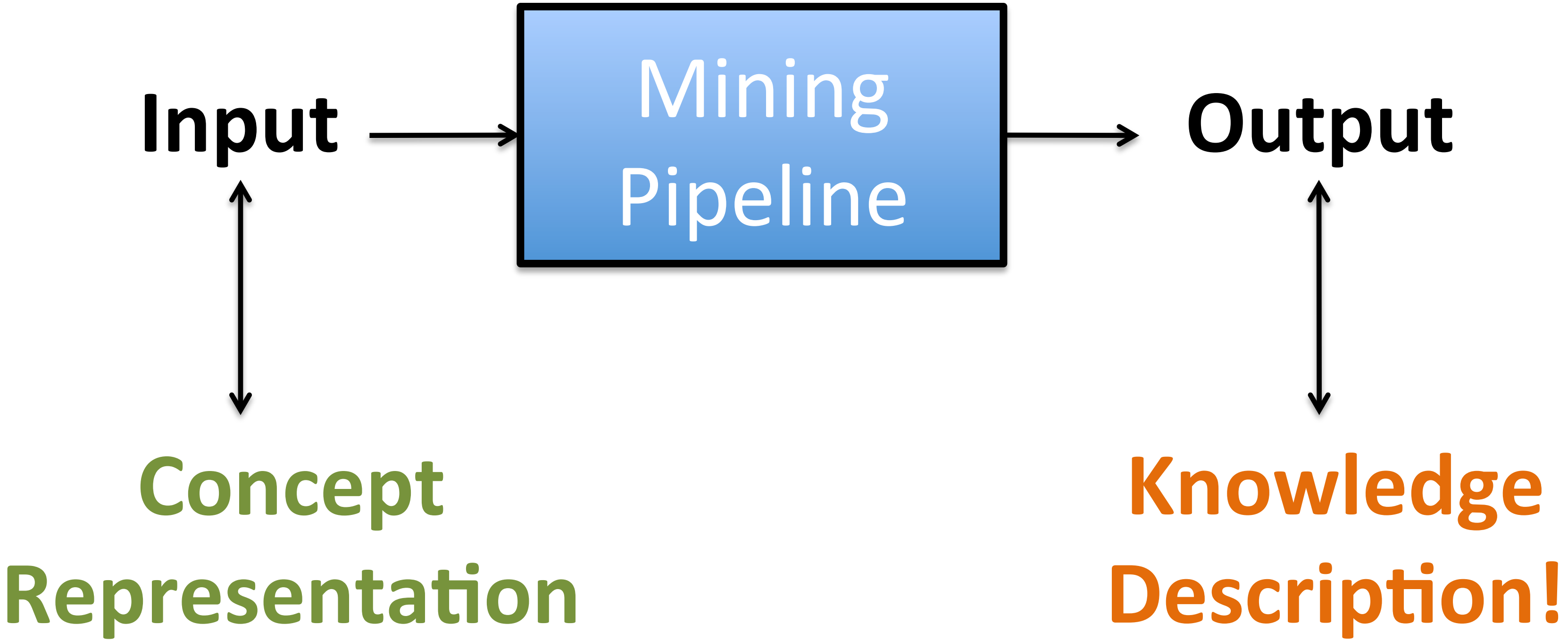
Outputs: Knowledge Description

How do we describe knowledge?

Clustering rules

Suppose you have all the attributes in place except the *taste* or *region* information, clustering automatically *regroups similar karnıyarık instances in the attribute space.*

Data Mining Pipeline – *finally!*



Input

Mining Pipeline

Output

Concept Representation

Knowledge Description!

The Context

ID	X	Y	LNDUSE_1	LNDUSE_2	LNDUSE_3	EMPTY_FLR	NBR_FLOORS	LND_PRICE	Shape_Leng	Shape_Area
24	414173.76	4545521.54	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	65	19.67644139	20.6855014
47	414546.937	4545568.77	Business-Shopping	Business-Shopping	Business-Shopping	%1-19 empty	7	323	47.26726314	111.9283511
49	414522.306	4545570.56	Business-Shopping	Empty	Empty	%80-99 empty	6	215	44.75515759	79.23590761
51	414516.73	4545579.21	Business-Shopping	No 2nd Floor	No 3rd Floor	%60-79 empty	2	215	44.78549084	108.1408138
54	414510.006	4545590.6	Business-Shopping	Business-Shopping	Empty	%40-59 empty	4	215	36.19779414	51.4612925
112	414473.822	4545533.28	Residential	Residential	Residential	%20-39 empty	4	75	31.01216642	48.53906077
148	414475.002	4545686.75	Business-Shopping	No 2nd Floor	No 3rd Floor	%40-59 empty	2	81	28.91314444	48.32320615
218	414412.732	4545636.38	Residential	No 2nd Floor	No 3rd Floor	%20-39 empty	2	81	37.94099886	89.22532668
303	414359.249	4545615.84	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	54	25.31028882	39.90526661
360	414339.192	4545535.45	Business-Shopping	No 2nd Floor	No 3rd Floor	%20-39 empty	2	118	27.04306332	43.09465423
605	414398.229	4545419.92	Business-Shopping	Residential	Residential	%20-39 empty	5	75	27.0247108	38.84584943
719	414359.223	4545449.15	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	48	41.04082385	84.93521138
1051	413850.002	4545418.87	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	269	19.03025298	21.15196311
1307	414127.711	4545474.8	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	108	23.52865716	31.60648204
1631	414299.567	4545531.73	Business-Shopping	Business-Shopping	Business-Shopping	%40-59 empty	4	183	45.14404758	82.31604095
1682	414304.009	4545479.39	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	75	25.59570944	32.38807975
1704	414078.417	4545579.82	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	118	26.86361993	40.13686342
1758	414314.541	4545597.35	Residential	Residential	No 3rd Floor	%20-39 empty	3	75	26.96805139	40.34437397
1872	414182.114	4545567.74	Business-Shopping	Residential	Residential	%20-39 empty	4	118	33.20154348	56.34492073
1899	414138.347	4545605.27	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	118	23.51417851	32.9281061
2013	413939.65	4545299.65	No 1st Floor	No 2nd Floor	No 3rd Floor	%40-59 empty	1	75	31.49221144	47.61425107
2150	414333.677	4545392.69	Business-Shopping	Residential	Empty	%40-59 empty	4	48	31.89077123	43.24782143
2163	414338.045	4545611.89	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	54	23.10194745	32.71690436
2231	414467.629	4545685.17	Business-Shopping	Empty	Residential	%40-59 empty	5	81	29.67665084	49.86289799
2344	413867.016	4545456.22	Business-Shopping	Residential	No 3rd Floor	%40-59 empty	3	75	26.44983088	37.41656042
2381	414228.078	4545469.4	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	65	19.6780004	19.01515631
2389	414094.932	4545525.71	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	65	17.15585481	18.1162578
2427	414245.613	4545574.51	No 1st Floor	No 2nd Floor	No 3rd Floor	All Empty	1	81	16.55497042	17.09602541

A snapshot of Tarlabaşı data

The Context

Some example uses of data mining in architecture and urban research

- Archetypal office building layouts (Hannah, 2007)
- Urban block morphology in terms of shape and density (Laskari, 2007)
- Arabic house typologies (Reffat, 2008)
- Spatio-temporal urban growth patterns and trends for modeling and prediction of urban growth (Liu and Seto, 2008)
- Urban typologies focusing on the aspects of morphology and density for blocks, mobility for streets (Gil *et al.*, 2009).

Thought exercise [for lunch ☺]

Think of the "City" as a concept:

- Designate a set of attributes related to the city
- Instantiate the "City" concept with examples
- Specify the attributes of your "City" examples

What kind of knowledge descriptions can you extract with your chosen set of attributes?

Do the reverse*