Data Mining (Knowledge Discovery)

1) Process
   - → Interactive ←
   - → Iterative

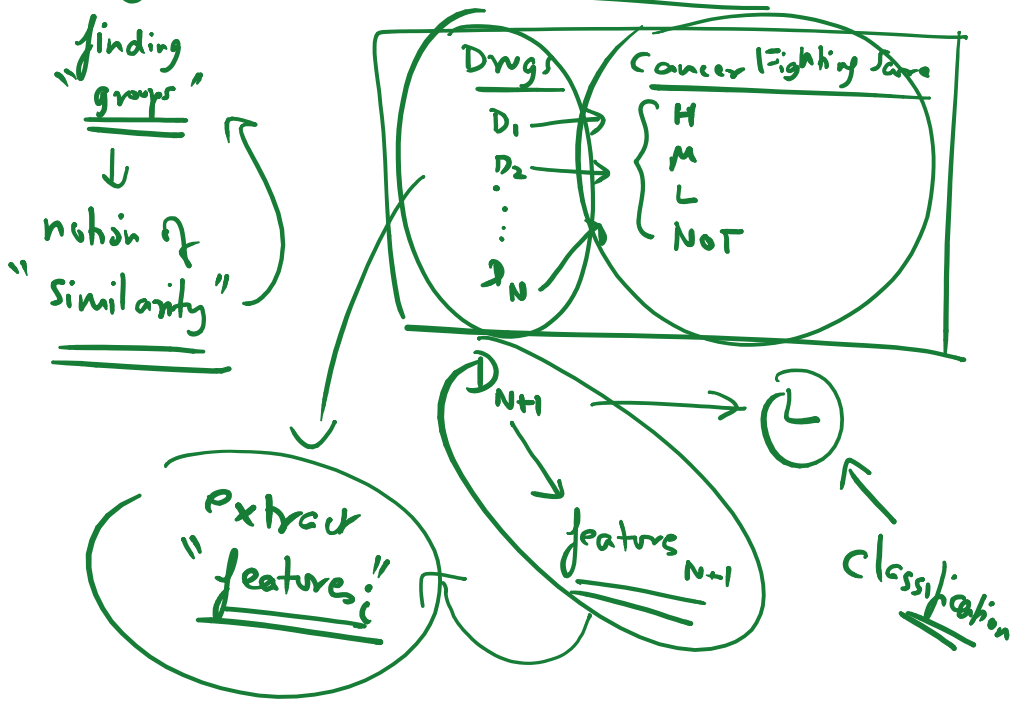
2) Discovery
   - → valid → generalizability.
   - → novel →
   - → useful

Patterns & models

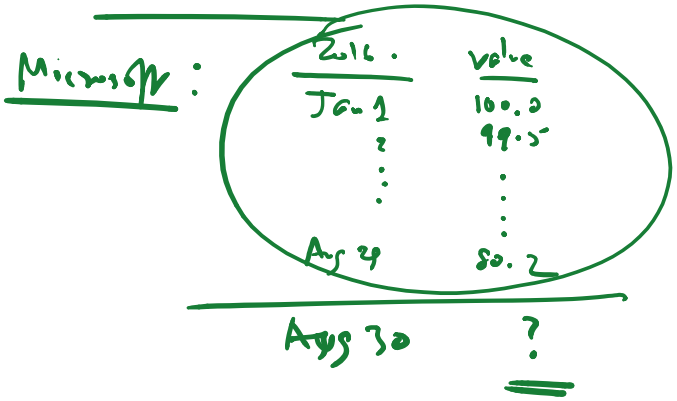
subset of the instance

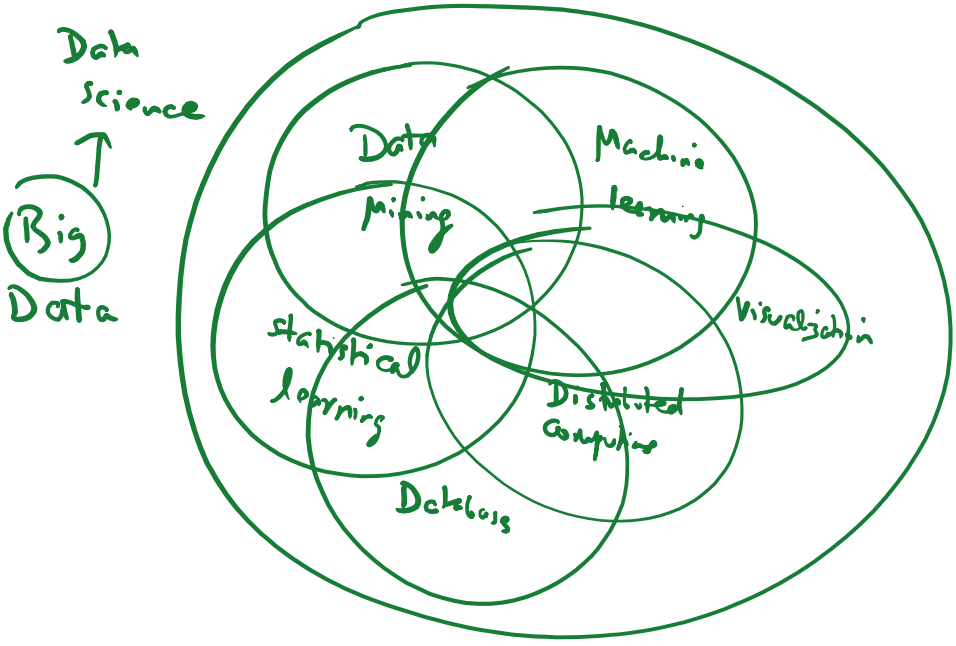
unsupervised
- clustering
- ↓
- finding "groups"
- ↓
- notion of "similarity"

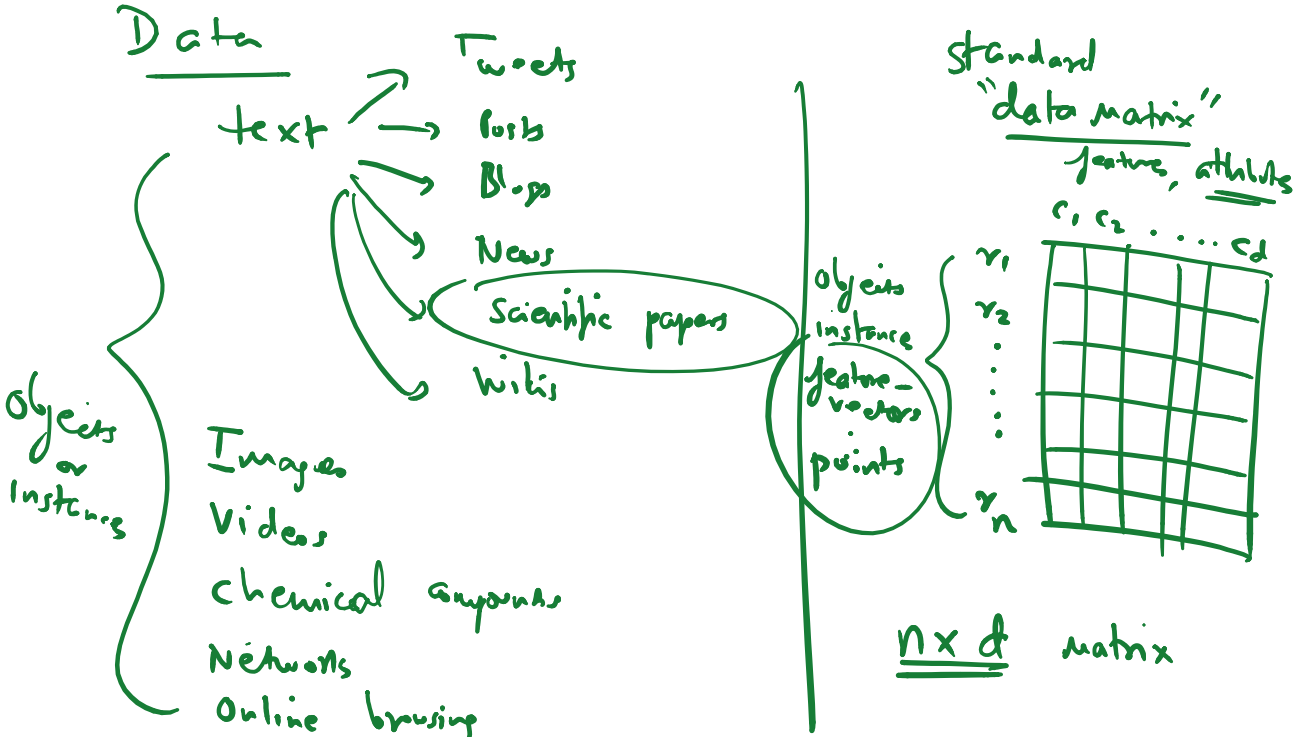
supervised
- classification/regression



Regression

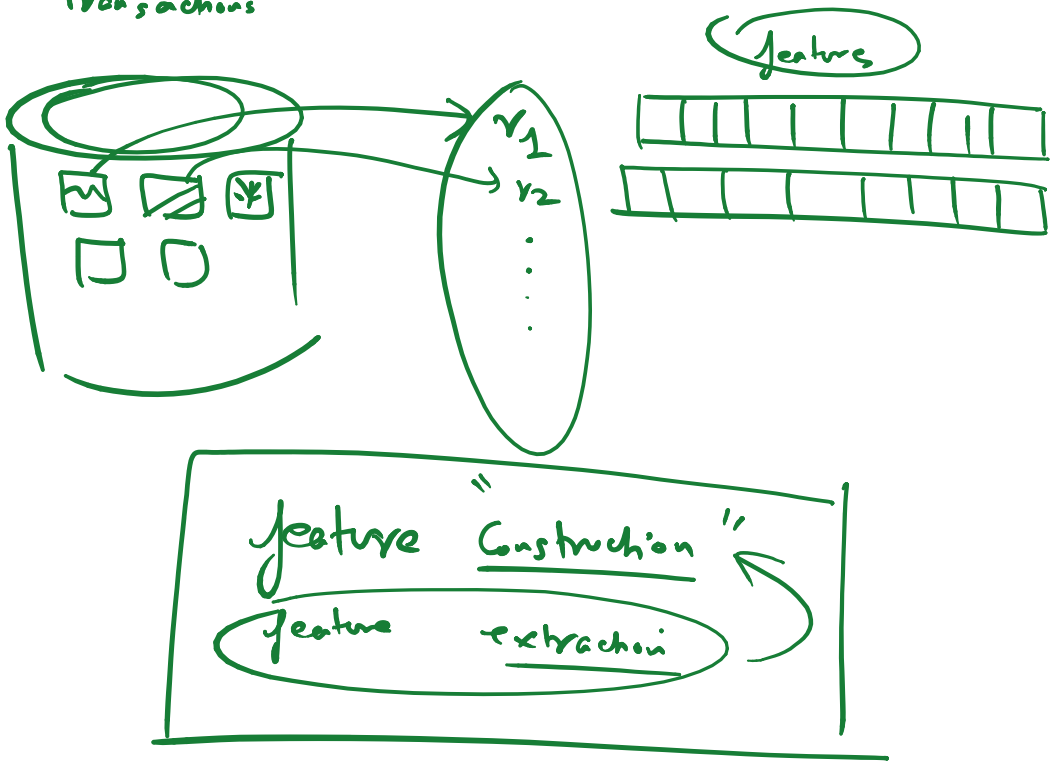

Microsoft :
Date
value
Jan 1
100.2
99.5
Aug 29
80.2
Aug 30
?


Data science
Big Data
Data mining
Machine learning
Statistical learning
Visualization
Distributed computing
Databases

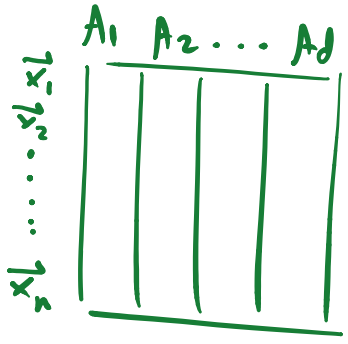

Data
text
Tweets
Posts
Blogs
News
Scientific papers
Wikis
Objects or Instances
Images
Videos
chemical compounds
Networks
Online browsing
standard "data matrix"
feature, attributes
c_1 c_2 ... c_d
Objects
instance
feature-vectors
points
r_1
r_2
r_n
$n \times d$ matrix

Networks
Online browsing
Transactions

n x d matrix



US Census
Edu Sal . . .
Cat. Numeric

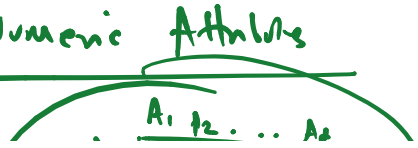


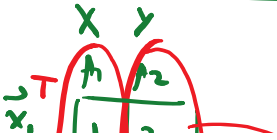
"d" ← # of dimensions
attribute
feature

n : # of points in the data matrix

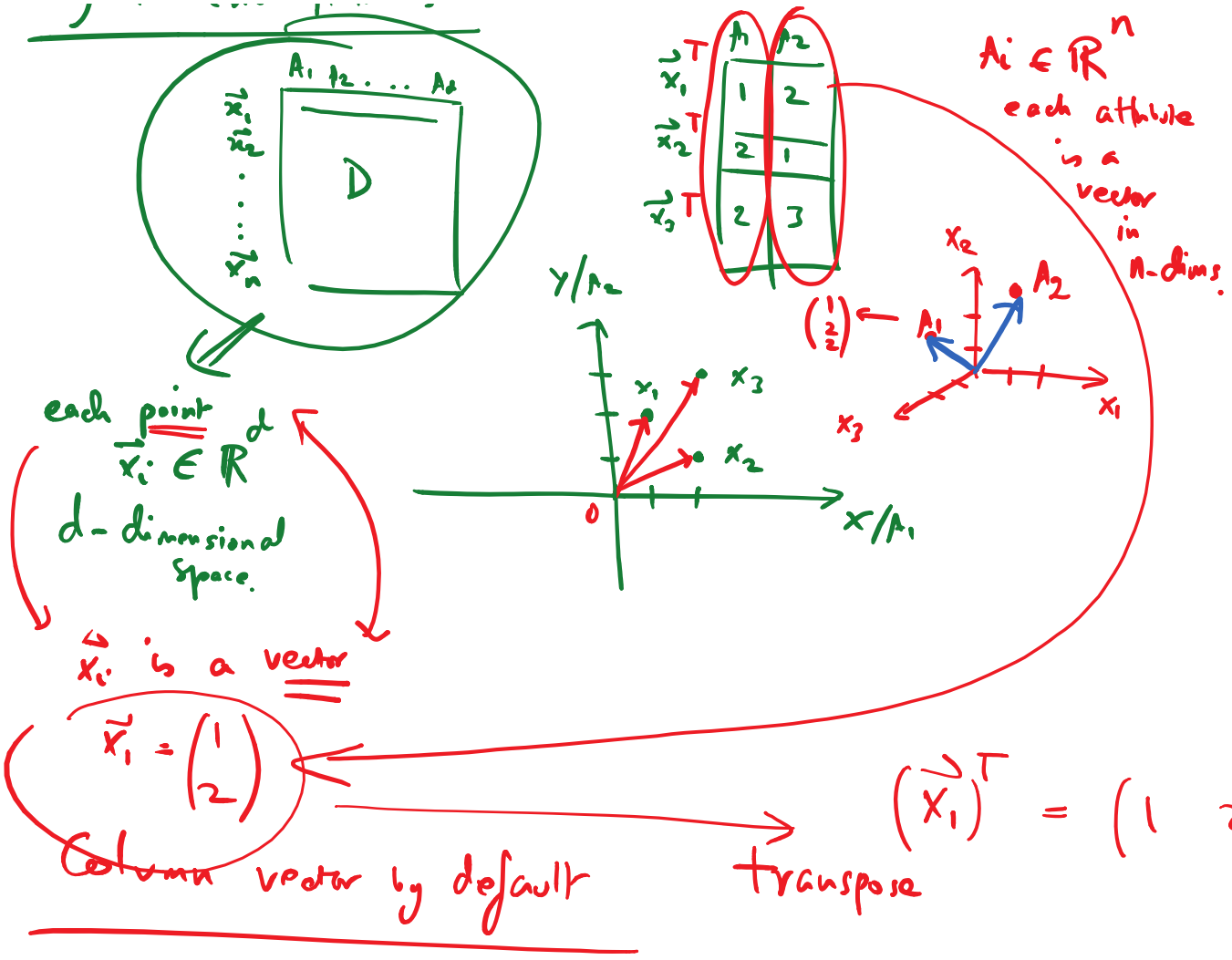
Attribute → Numeric
→ Categorical → nominal (Red, Yellow, ...)
→ ordinal (H, M, L)

Only Numeric Attributes





$A_i \in \mathbb{R}^n$


A1 A2 ... Ad
D
A1 A2
1 2
2 1
2 3
Ai ∈ R^n
each attribute is a vector in n-dims.
each point
d
d - dimensional space.
Y/A2
X/A1
x1
x2
x3
0
A1
A2
x1
x2
x3
is a vector
Column vector by default
transpose
(1 2)