

873 final class

Note Title

12/2/2011

Exploring data mining

What is it good for?

SVD - clustering

comprehensible

good upstream

easy to use

S dimensionality reduction, & how much complexity

W

par with sparsity, normalization (non-G dists)

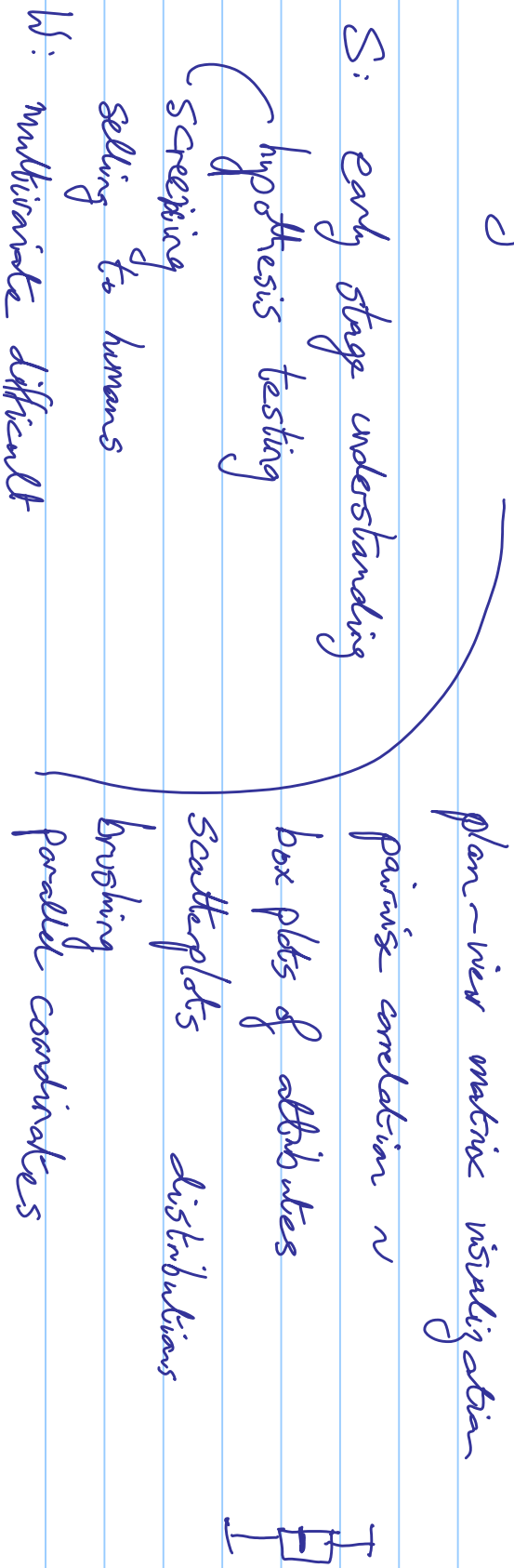
small data

this much variation



- D Symmetry, good for big structure
 - T categorical data problematic, low dimensional dataset
-

Visualization



no model
analyt dependent

D:

more accessible

more effort by analysts

T:

Scalability in space & time

Random Forests

S: efficient
good for wide datasets with outliers
no overfitting

W: not directly comprehensible
not for regression

O: attribute selection X 2

T:

Latent Dirichlet Allocation

S: dimension reduction

bicustering

really good for text

W: bad for numeric data

choosing K

not clear how/why it works

D:



T: assumes substitution-like property for attribute values.

Independent Component Analysis

- S: finds small irregularities / reduce "noise" / data validation pseudattributes with reduced variance
- W: hard to interpret is an easy kind of way cumbersome choosing k

o.

T: not for big structure

Support Vector Machines

S: fast even on hybrid data

W: 2 classes (more or less)
parameter choices / kernel choices

T:

O: ~~attempts~~ selection

Neural networks

S: non-linear boundaries

W: not for categorical data long run time

D: not for small n
no visible model

overfitting / cross layer size / threshold choice
not large # of classes

T:

GOLD STANDARD PROCESS

Given a dataset:

1. What's the problem?

2. Explore / clean the data

3. Visualization

4. Normalization
by attribute
by record (?)

5. Prediction
Clustering
Matching clusters with classes

SVD, PCA, LDA, Prediction

