Data Mining and Knowledge Discovery: An Introduction

Big Data Examples

Winter Corp. (2005) Commercial Database Survey:

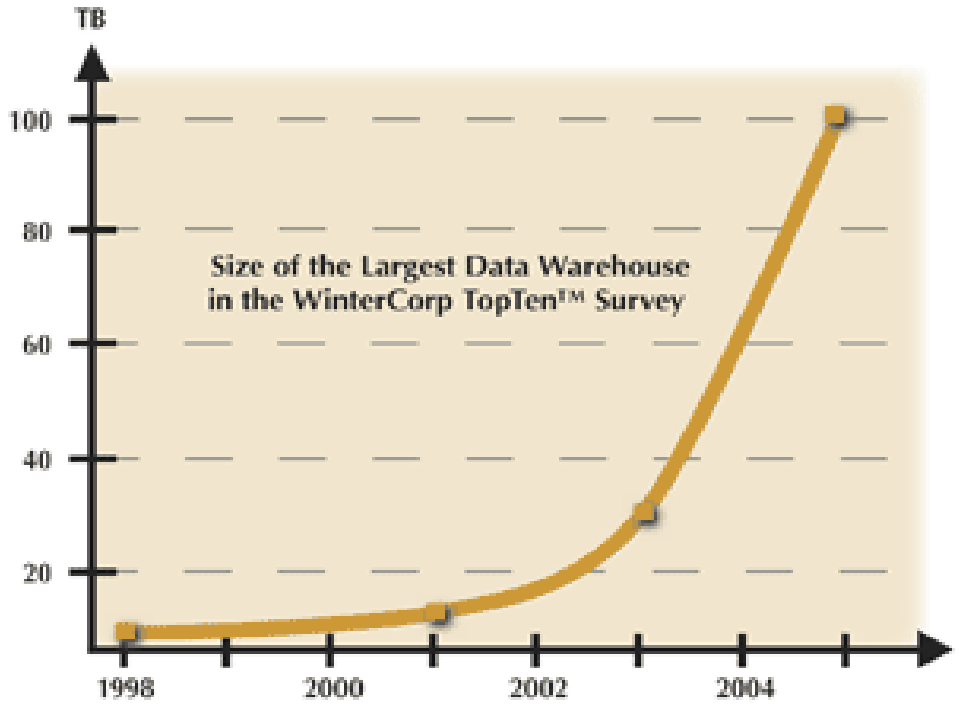
1. Max Planck Inst. for Meteorology , 222 TB
2. Yahoo ~ 100 TB (Largest Data Warehouse)
3. AT&T ~ 94 TB

www.wintercorp.com/VLDB/2005_TopTen_Survey/TopTenWinners_2005.asp

Big Data Examples

- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session
  - storage and analysis a big problem
- AT&T handles billions of calls per day
  - so much data, it cannot be all stored -- analysis has to be done “on the fly”, on streaming data

Data Growth



In 2 years, the size of the largest database TRIPLED! Knowledge Discovery is **NEEDED** to make sense and use of data.

Application Areas

What do you think are some of the most important and widespread business applications of Data Mining?

Application Areas

- Science
  - astronomy, bioinformatics, drug discovery, ...
- Business
  - fraud detection, profiling tax cheaters, robot learning, network security, targeted marketing...
- Web
  - text mining, Google “did you mean?”, targeted advertising, Netflix and Amazon

Assessing Credit Risk: Case Study

Person applies for a loan

- Task: Should a bank approve the loan?
- Note: People who have the best credit don't need the loans, and people with worst credit are not likely to repay. Bank's best customers are in the middle

Recommending Books: Case Study

- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"
- Recommendation program is quite successful

Network Security: Case Study

- IDS vs ADS
- Binary Classification Problem
- Recognize malicious packets/connections
- Storage impossible
- Online classification, speed is an issue!

Artificial Vision: Case Study

- http://groups.csail.mit.edu/vision/TinyImages/

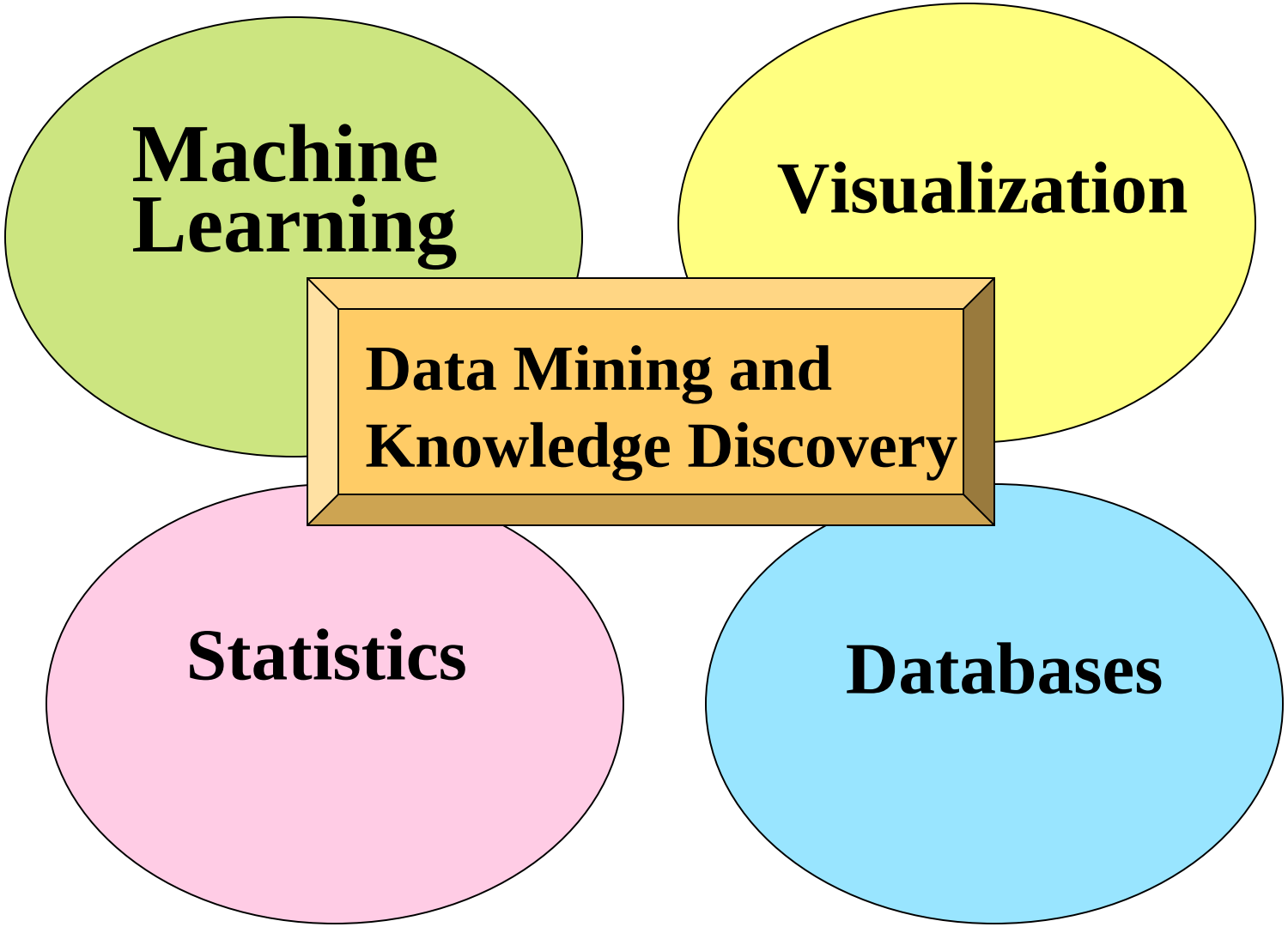
Knowledge Discovery Definition

Knowledge Discovery in Data is the *non-trivial* process of identifying:

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

Related Fields



Major Data Mining Tasks

Clustering: finding clusters in data

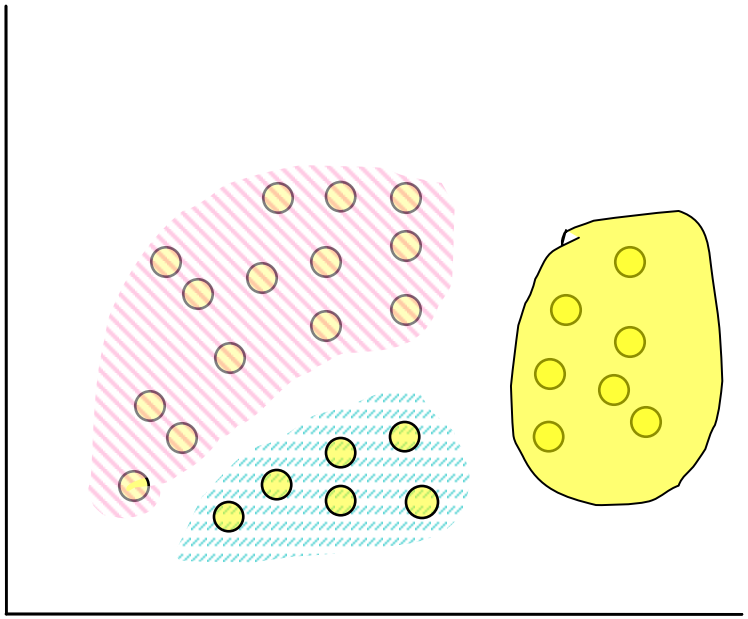
Classification: predicting an item class

Associations: e.g. A & B & C occur frequently

Visualization: to facilitate human discovery

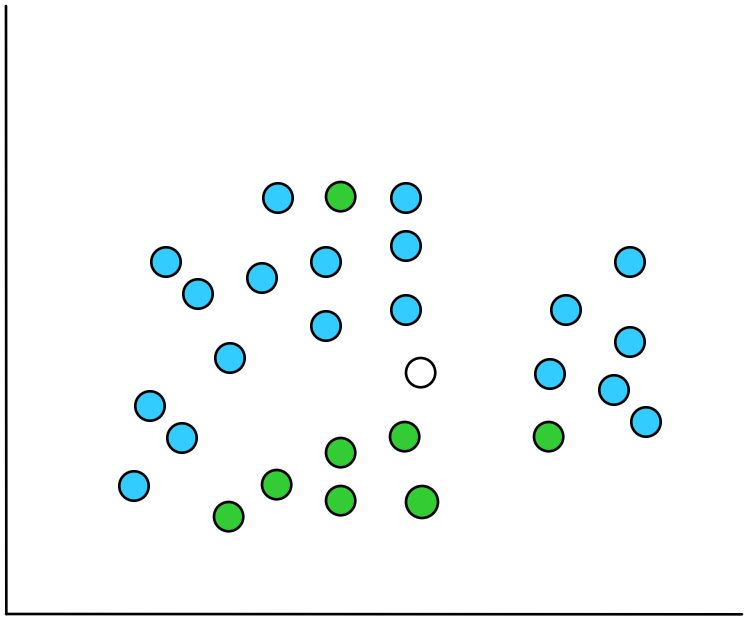
Clustering

Find “natural” grouping of instances given un-labeled data



Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances



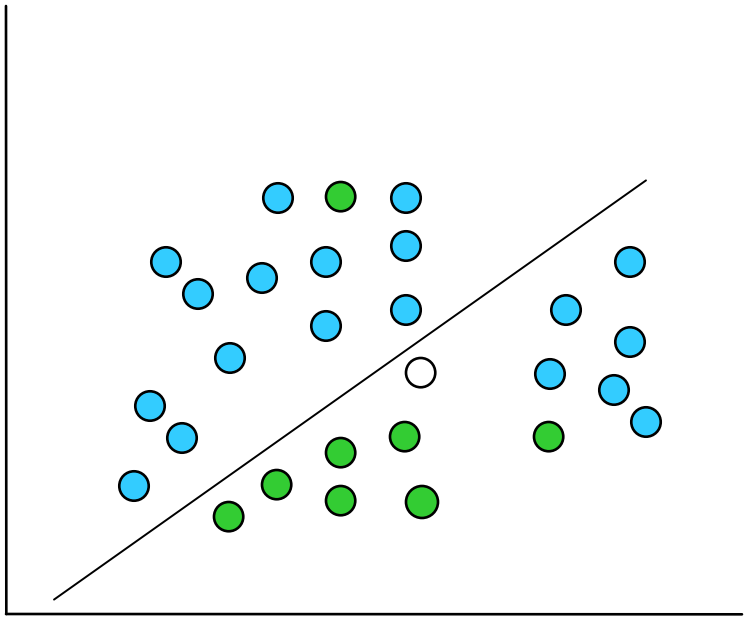
Many approaches:
Regression,
Decision Trees,
Bayesian,
Neural Networks,
...

Given a set of points from classes ● ●
what is the class of new point ○?

Classification Process

- Collect Data
- Pre-Process Data
- Define Training and Testing sets (1/3 rule)
- Run classifiers
- Understand the Results

Classification: Linear Regression

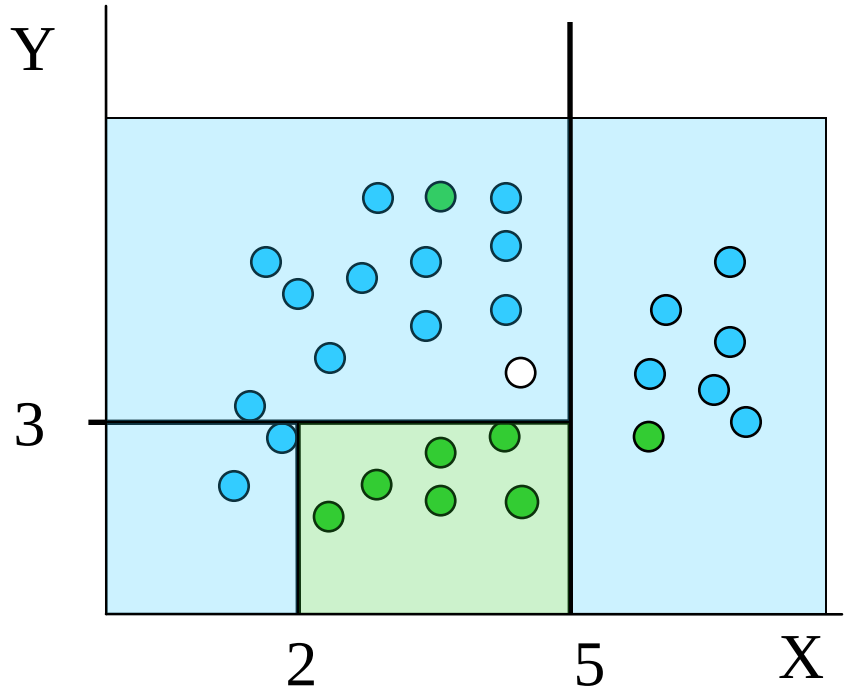


- Linear Regression

$w_0 + w_1 x + w_2 y >= 0$

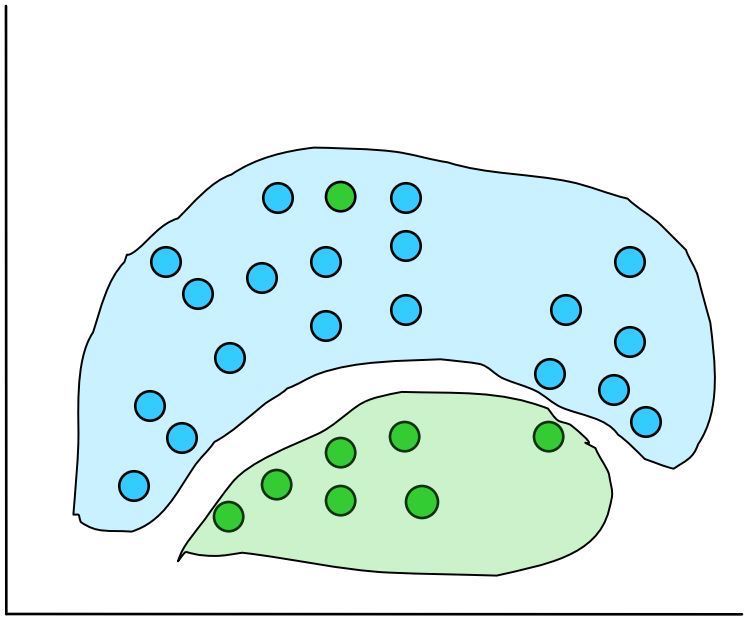
- Regression computes w_i from data to minimize squared error to ‘fit’ the data
- Not flexible enough

Classification: Decision Trees



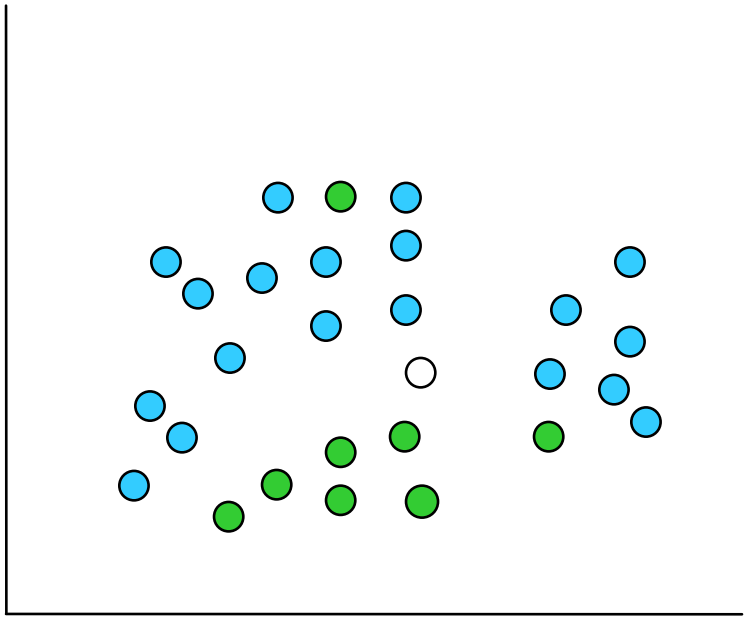
if X > 5 then blue
else if Y > 3 then blue
else if X > 2 then green
else blue

Classification: Neural Nets



- Can select more complex regions
- Can be more accurate
- Also can overfit the data – find patterns in random noise

Classification: K-NN



- Very natural approach
- Problem dependent
- Can be very accurate
- Does not require any training
- Slow

Attribute types

- Nominal, e.g. eye color=brown, blue, ...
  - only equality tests
  - important special case: boolean (True/False)
- Ordinal, e.g. grade=1,2,..,12

Why specify attribute types?

- ***Q: Why algorithms need to know about attribute type?***
- A: To be able to make right comparisons and learn correct concepts, e.g.
  - **Outlook > "sunny"** does not make sense, while
  - **Temperature > "cool"** or **Humidity > 70** does
- Additional uses of attribute type: check for valid values, deal with missing, etc.

The weather problem

- Given this data, what are the rules for play/not play?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...	...	...	...	...

The weather problem

- Conditions for playing

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...	...	...	...	...

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

Weather data with mixed attributes

- Rules with mixed attributes

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```

Classifying iris flowers

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```

Predicting CPU performance

- Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- Linear regression function

```
PRP =  -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
+ 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

Preparing the input

- Problem: different data sources (e.g. sales department, customer billing department, ...)
  - Differences: styles of record keeping, conventions, time periods, data aggregation, primary keys, errors
  - Data must be assembled, integrated, cleaned up
  - “Data warehouse”: consistent point of access

The ARFF format

```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

Missing values

- Frequently indicated by out-of-range entries
  - Types: unknown, unrecorded, irrelevant
  - Reasons:
    - malfunctioning equipment
    - changes in experimental design
    - collation of different datasets
    - measurement not possible
- Missing value may have significance in itself (e.g. missing test in a medical examination)
  - Most schemes assume that is not the case
    ⇨ “missing” may need to be coded as additional value

Inaccurate values

- Reason: data has not been collected for mining it
- Result: errors and omissions that don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes $\Rightarrow$ values need to be checked for consistency
- Typographical and measurement errors in numeric attributes $\Rightarrow$ outliers need to be identified
- Errors may be deliberate (e.g. wrong zip codes)
- Other problems: duplicates, stale data

Getting to know the data

- Simple visualization tools are very useful
  - Nominal attributes: histograms (Distribution consistent with background knowledge?)
  - Numeric attributes: graphs
    (Any obvious outliers?)
- 2-D and 3-D plots show dependencies
- Need to consult domain experts
- Too much data to inspect? Take a sample!