

00111011000001010000
10101011111101110110
000000000000110011001
01011110110110010001
00000001011000001101
00110111011001100001
01001100011111111010
010101111100111111001
10110011111000111001
00101111100100101101
01110010111011111111
00110111111100010100
01110111011110010010
10000100100111110111
11101000111011111100
10000111100001001000
11110101110111000100
01011100010001001010
10100011000011000011
11000001011101011010
01010111011101010100
01111000111011011110
10101000110011001101
10010000111101011101
01111101111011101000
100001111110100010001
10010010100001010100
11111001100000100011
00010010100011001110
00110100101100101011
01110101001100101101
00101110010011111010
10011100100001001101
01010001110001110011
11000101011001000110
10010011001100101010
11101100001011011000
10110010101111010110
11010111110101100011
10011000110110110111
01010001001110010011
10110101100101111001
10100010111110100101
11110101100111101101
10011110011001111010
10111001101110111010
11110111110001100100
00011101000100011110
10001101111101011100
11101000000111000001
00111011000001010000
10101011111101110110
00000000000110011001
01011110110110010001
00000001011000001101
00110111011001100001
01001100011111111010
010101111100111111001
10110011111000111001
00101111100100101101
01110010111011111111
00110111111100010100
01110111011110010010
10000100100111110111
11101000111011111100
10000111100001001000
11110101110111000100
01011100010001001010
10100011000011000011

Everything You Always Wanted to Know About Data Mining But Were Afraid to Ask

or

Statisticians vs. Data Miners *"Did we finally lose the battle?"*

Martin Theus

martin.theus@math.uni-augsburg.de

Outline

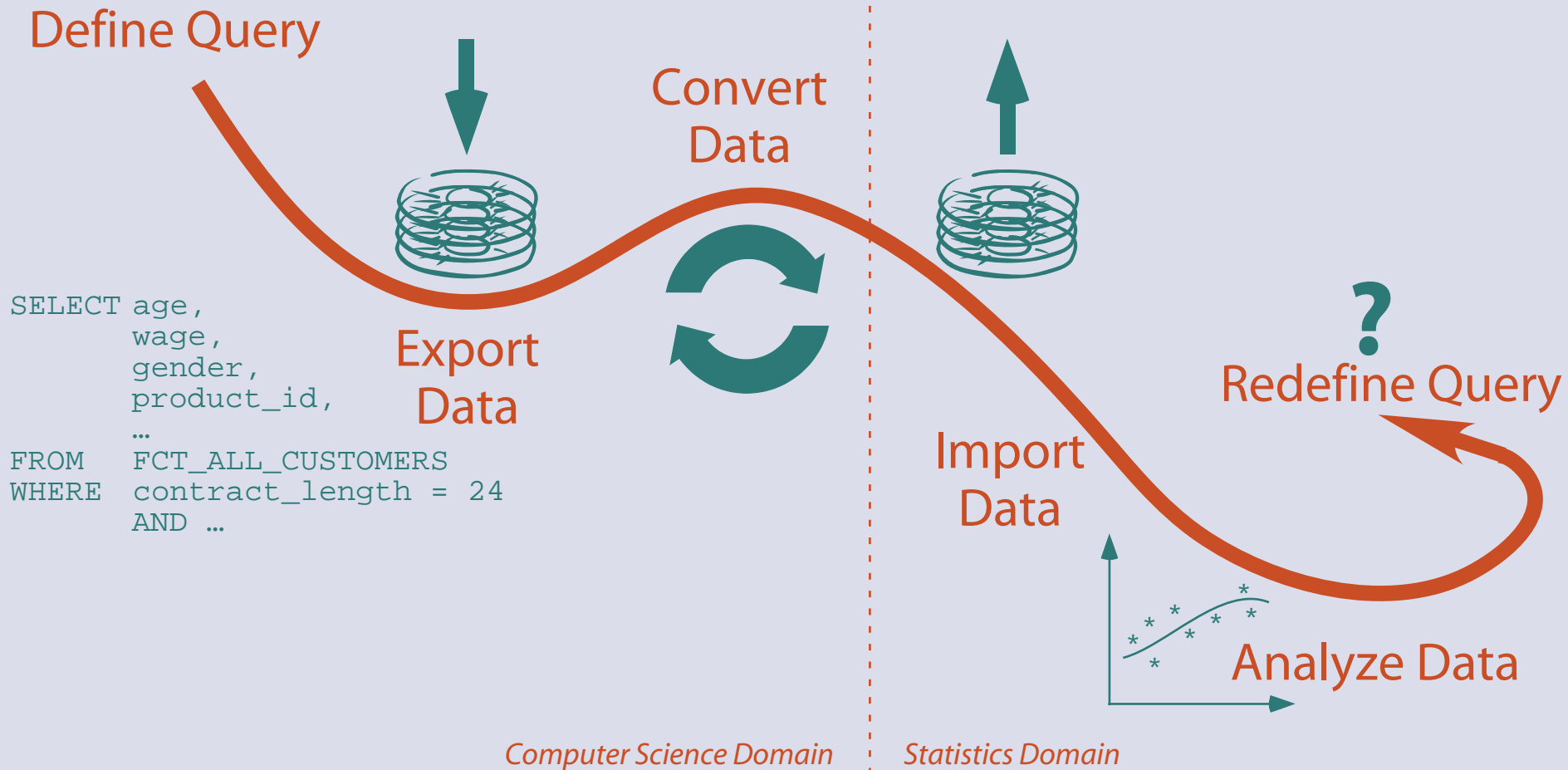
- **Tell me where your data sits**
What you should know about flat files
- **Selections**
“What you SELECT is what you get!”
- **A Day in the Life of a Statistician**
What has changed since 10, 20, 50 ... years
- **Do we need a new Discipline?**
Who takes the lead?

The Curse of Flat Files

- Statistics grew up 100 years ago with small data sets
- The era of mathematical statistics ended just before general computing arose ... still just small data sets
- Statistics is still a topic “owned” by math and not by computer science ...
- ... this is good and bad as well.
- Statistics software is good at reading flat files, but ...
... can you do data mining on flat files?
- Statisticians usually do not know what databases are
- ... and usually have a hard time converting formats

The Data Mining Process

(An Example ...)



Interactivity vs. Databases

- Interactions with data (graphical or numerical) demand quick response times from a system
- The usual interaction with a database is to select data
- Huge data sets cannot be handled outside a database
- Graphical representations of data often only need a subset or a summary of the data, e.g.
 - Barcharts and Mosaic plots
 - Boxplots
 - (Scatterplots)
- The “backpropagation” of attributes (selection flag, group information etc.) into the database is usually the bottleneck

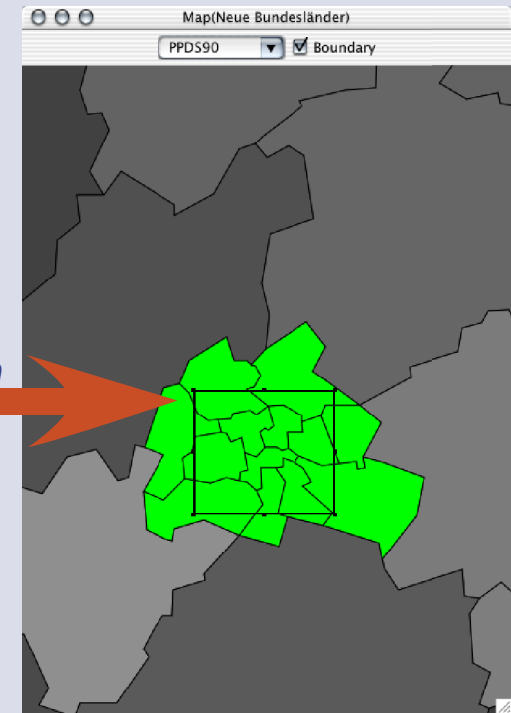
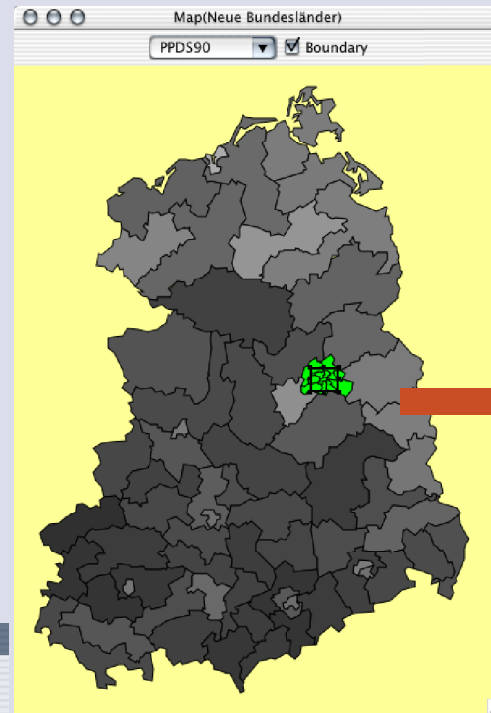
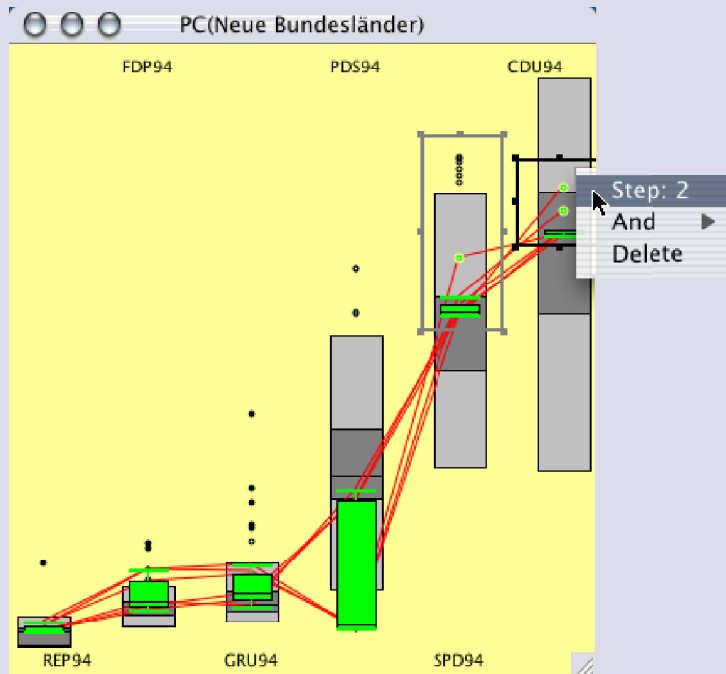
Is there a way out?

- There are “simple” ways to access databases:
 - ODBC, Windows standard, widely supported
 - JDBC, JAVAs build in Interface to databases
- “Openness” not even necessary. A native interface to Oracle and mySQL covers more than 90% of all DBs
- SAS system offers native interfaces for Oracle etc.
- DataDesk version 7 offers database connectivity (...getting hold of first betas as testers!)
- Early version of Mondrian did connect to DBs, new versions of Mondrian will support DB connections again
- Idea: Access the data on 2 levels ... avoid SQL-code

00111011000001010000
10101011111101110110
00000000000110011001
01011110110110010001
00000001011000001101
00110111011001100001
01001100011111111010
010101111100111111001
10110011111000111001
00101111100100101101
01110010111011111111
00110111111100010100
01110111011110010010
10000100100111110111
11101000111011111100
10000111100001001000
11110101110111000100
01011100010001001010
10100011000011000011
11000001011101011010
01010111011101010100
01111000111011011110
10101000110011001101
10010000111101011101
01111101111011101000
100001111110100010001
10010010100001010100
11111001100000100011
00010010100011001110
00110100101100101011
01110101001100101101
00101110010011111010
10011100100001001101
01010001110001110011
11000101011001000110
10010011001100101010
11101100001011011000
1011001010111010110
1101011110101100011
10011000110110110111
01010001001110010011
10110101100101111001
1010001011110100101
11110101100111101101
1001110011001111010
10111001101110111010
11110111110001100100
00011101000100011110
10001101111101011100
111010000011100001
00111011000001010000
10101011111101110110
00000000000110011001
01011110110110010001
00000001011000001101
00110111011001100001
01001100011111111010
01010111100111111001
10110011111000111001
00101111100100101101
01110010111011111111
00110111111100010100
01110111011110010010
10000100100111110111
11101000110111111100
10000111100001001000
11110101110111000100
010111000100101010
10100011000011001010

Smart Selection Sequences

- Transformation invariant
 - Location is determined from the data

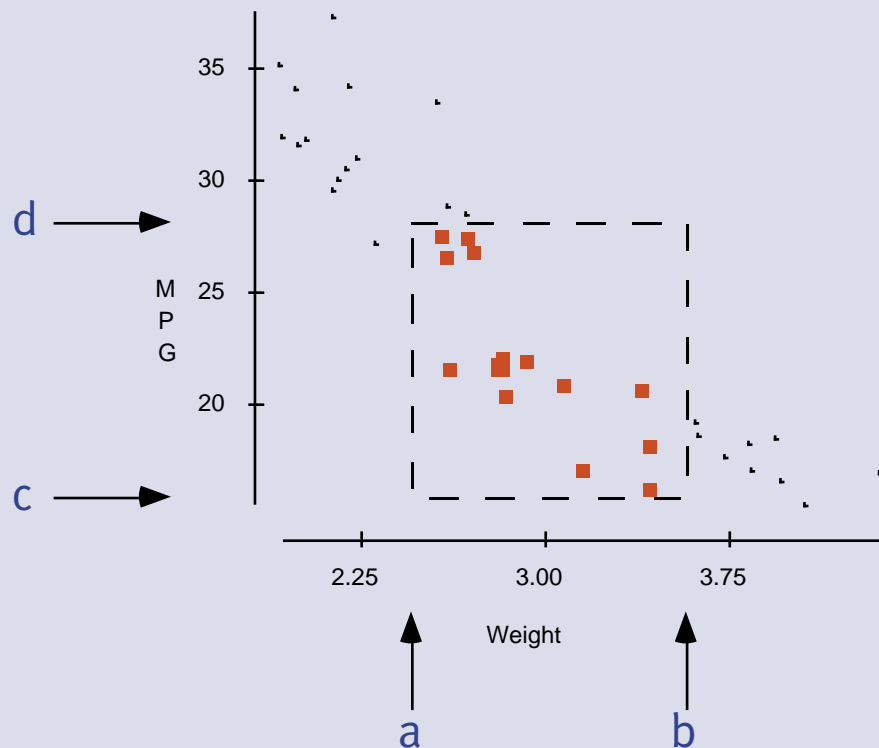


- Multiple Selections per Window
 - A window can hold as many selection rectangles as necessary

Selection Sequences translate well into SQL

Example: Drag-Box

• Plot



• Formal description

a) **mathematical definition:**

$$U = \{ e_i \mid a < x_i < b \wedge c < y_i < d, e_i \in E \}$$

b) **relational algebra:**

$$E \left(a < x_i < b \wedge c < y_i < d(E) \right)$$

c) **SQL**

```
SELECT weight,  
        MPG  
FROM CARS_DATA  
WHERE weight between a and b  
      and MPG between c and d
```

00111011000001010000
10101011111101110110
00000000000110011001
01011110110110010001
00000001011000001101
00110111011001100001
01001100011111111010
010101111100111111001
10110011111000111001
00101111100100101101
01110010111011111111
0011011111100010100
01110111011110010010
10000100100111110111
11101000111011111100
10000111100001001000
11110101110111000100
01011100010001001010
10100011000011000011
11000001011101011010
01010111011101010100
01111000111011011110
10101000110011001101
10010000111101011101
0111101111011101000
10000111110100010001
10010010100001010100
11111001100000100011
00010010100011001110
00110100101100101011
01110101001100101101
00101110010011111010
10011100100001001101
01010001110001110011
11000101011001000110
10010011001100101010
11101100001011011000
10110010101111010110
11010111110101100011
10011000110110110111
01010001001110010011
10110101100101111001
10100010111110100101
11110101100111101101
10011110011001111010
10111001101110111010
11110111110001100100
00011101000100011110
10001101111101011100
11101000000111000001
00111011000001010000
10101011111101110110
00000000000110011001
01011110110110010001
00000001011000001101
00110111011001100001
01001100011111111010
01010111100111111001
10110011111000111001
00101111100100101101
01110010111011111111
00110111111100010100
01110111011110010010
10000100100111110111
11101000111011111100
10000111100001001000
11110101110111000100
01011100010001001010
10100011000011000011
10100011000011000011

A Day in the Life of a Statistician

Task

- 7:30: check jobs of last night
- 8:00: fix problem of job
- 9:00: check SMS for board
- 9:30: meeting with marketing
- 11:00: work on churn model
- 12:30: lunch
- 13:00: weekly team meeting
- 14:30: monthly mngmt. report
- 16:00: managmt. info system
- 17:00: ad hoc question

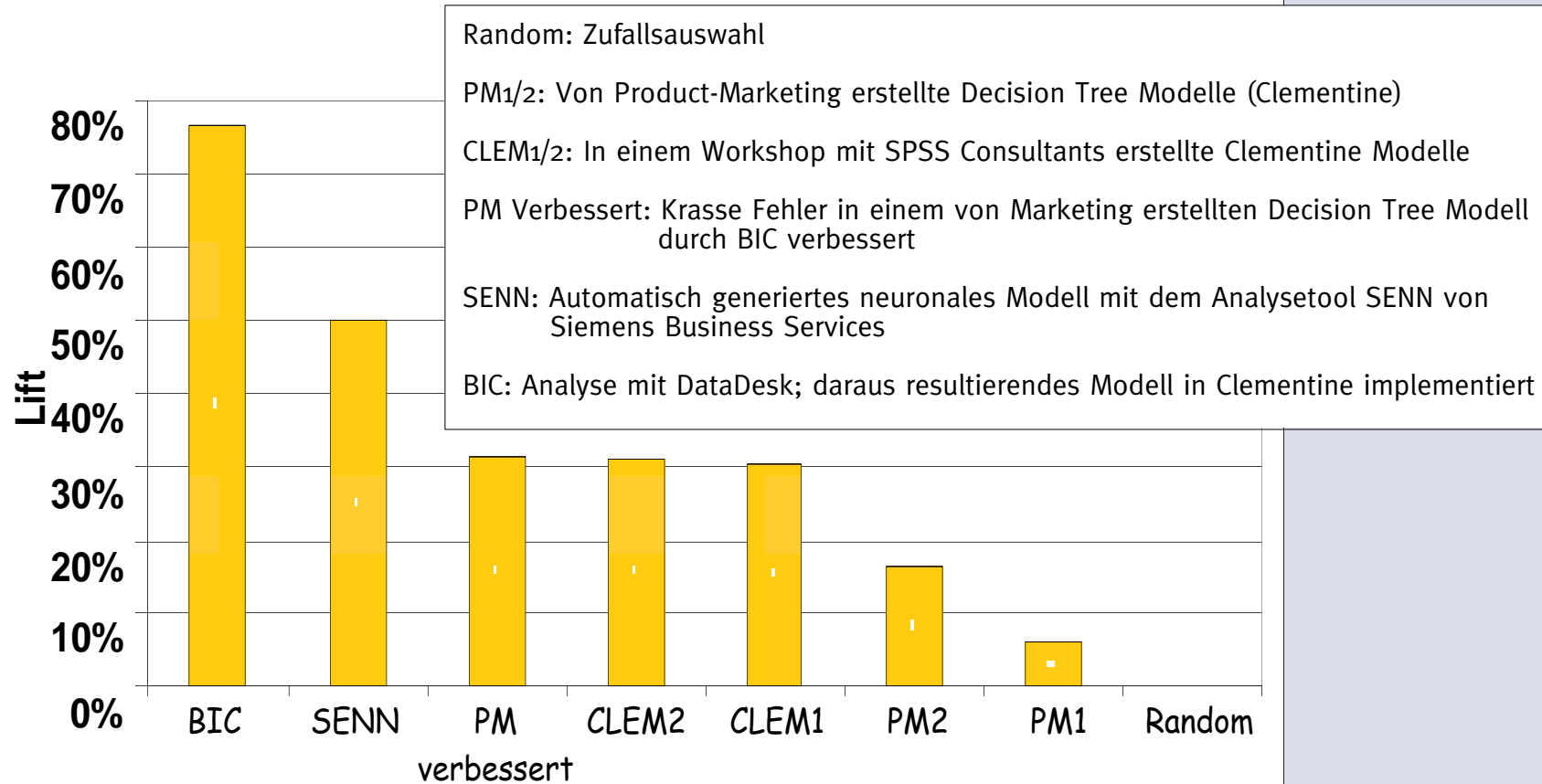
Skills

- UNIX, WWW
- Perl
- UNIX, WWW
- Common Sense, Patience
- Statistics, DataDesk, Clementine
- ...
- ...
- Excel, Powerpoint
- SQL, Perl, HTML
- SQL, MapInfo

A Case Study

Churn Prediction at a German Mobile Carrier

Vergleich des Lifts der Churn-Modelle auf neuen Daten
vom Stichtag 01.01.2001



Are Computer Science Skills most important?

- No!
- But they matter!
- They are indispensable but not at all sufficient!

Are Data Analysis Skills most important?

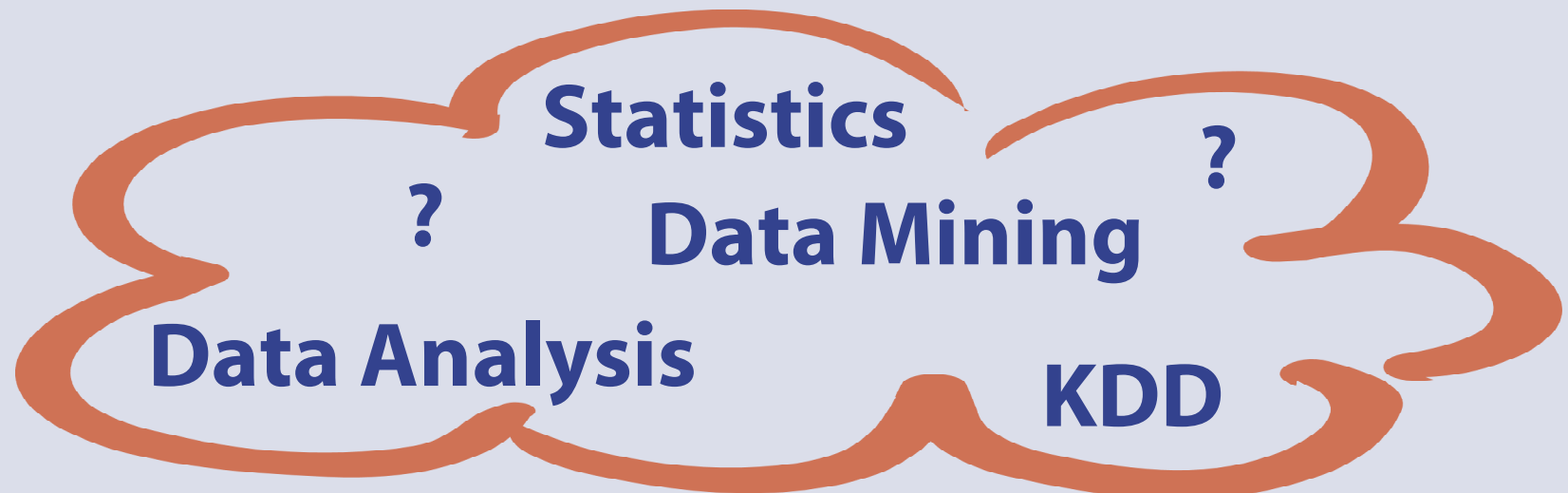
- Yes!
- Common sense is important but not enough!
- They are indispensable **and** sufficient!

Are Statistical Skills most important?

- No!
- But they are **essential** as the basis for data analysis

Is KDD a concurrent Discipline to Statistics?

- Are R-DBMS interfaces the successors of statistical tools?
- Obviously Knowledge Discovery in Databases works on databases
- KDD is “owned” by computer scientists, but is about :
gathering data, sampling, experimental design, analyzing data,
modeling, discovery, visualization, forecasting, classification



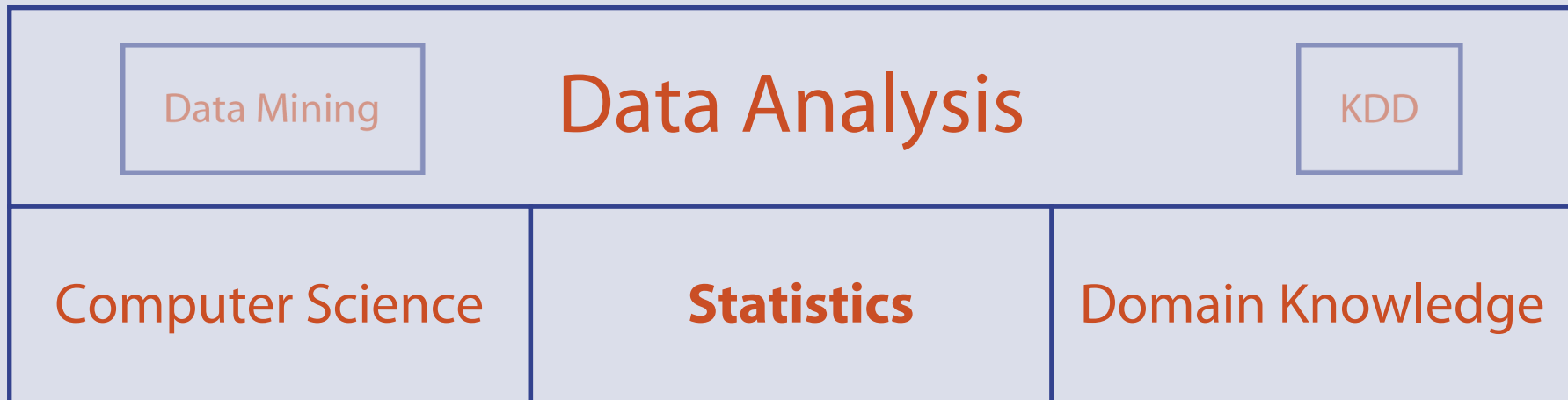
Do we need a new Discipline?

- **Example:**

- Biology and Chemistry -> Biochemistry

- Statistics and Computer Science -> ? Data Mining ?, ? KDD ?

- **A Model:**



What Others say ...

- **John Tukey (1962)**

- Statistics - concerned with data analysis - should be defined in terms of a set of *problems* (as are most fields) rather than a set of *tools*, namely those problems that pertain to data.

- **Brad Efron (19??)**

- "Statistics has been most successful in information science."
- "Those who ignore statistics are condemned to reinvent it."

- **Daryl Pregibon (1999)**

- KDD = statistics at scale & speed

- **Jerry Friedman (1999)**

- "Every time the amount of data increases by a factor of ten, we should completely rethink how we analyze data."
- "We will also have to expand our curriculum to include current computer oriented data analysis methodology ..."

- **Bill Cleveland (1999)**

- Data Science: Expanding the technical Areas of Statistics

Summary

- Data no longer only sits in flat files but in databases
- Statistics software must work on databases in the same way as it used to on flat files
- Convenient selection mechanisms are crucial for an analysis of large/complex data sets - graphics matters
- We need to think about how statistics defines itself in the future ... to assure a future
- It may be better to take the lead in a new discipline, rather than to diminish over the next 20 years