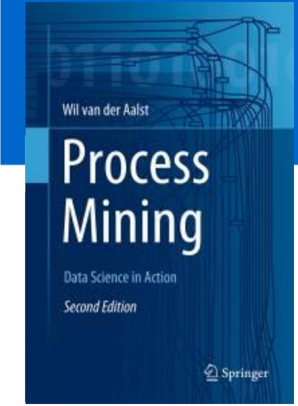
*Process Mining: Data Science in Action*

# How Process Mining Relates to Data Mining



Wil van der Aalst

Process Mining

Data Science in Action

*Second Edition*

Springer

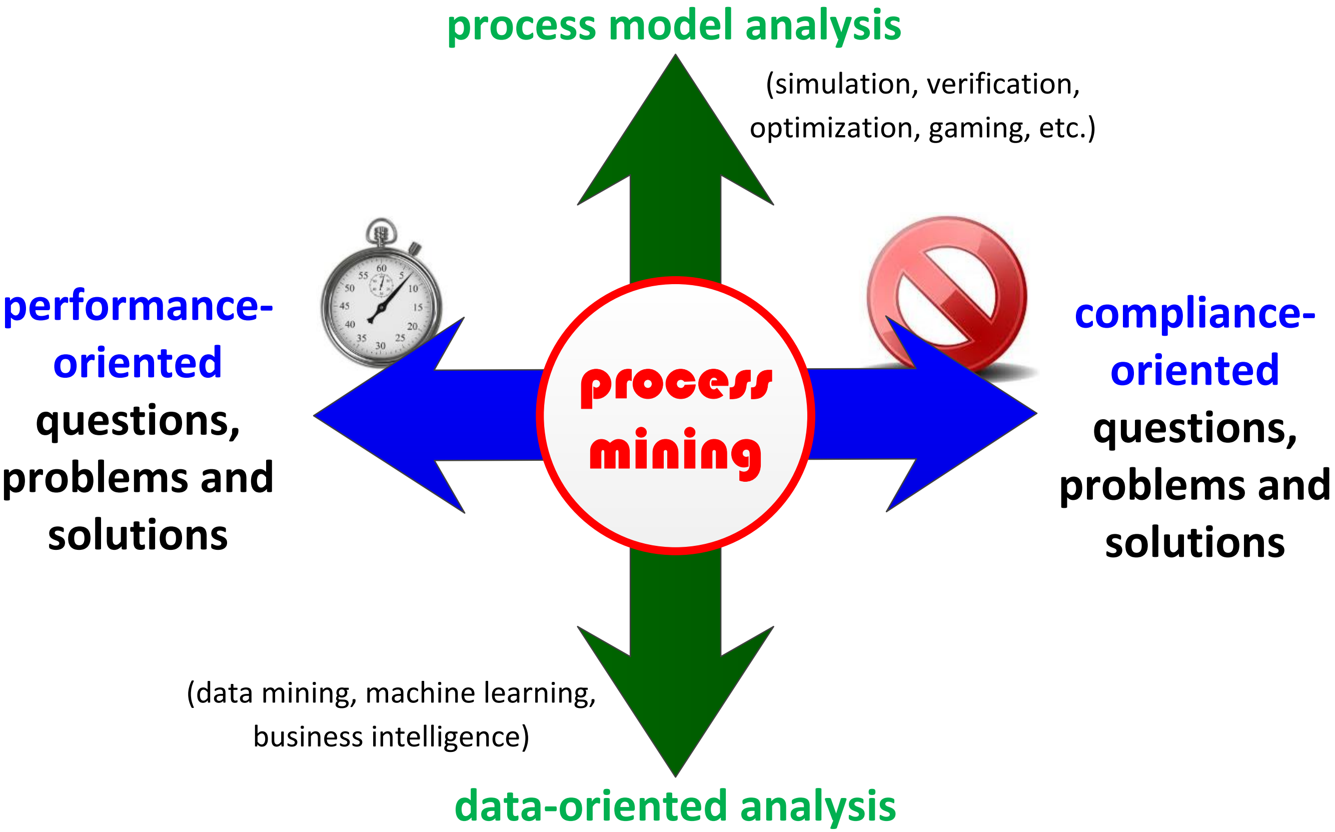
prof.dr.ir. Wil van der Aalst

www.processmining.org

TU/e Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

# Process mining: The missing link



process model analysis

(simulation, verification, optimization, gaming, etc.)

performance-oriented questions, problems and solutions

process mining

compliance-oriented questions, problems and solutions

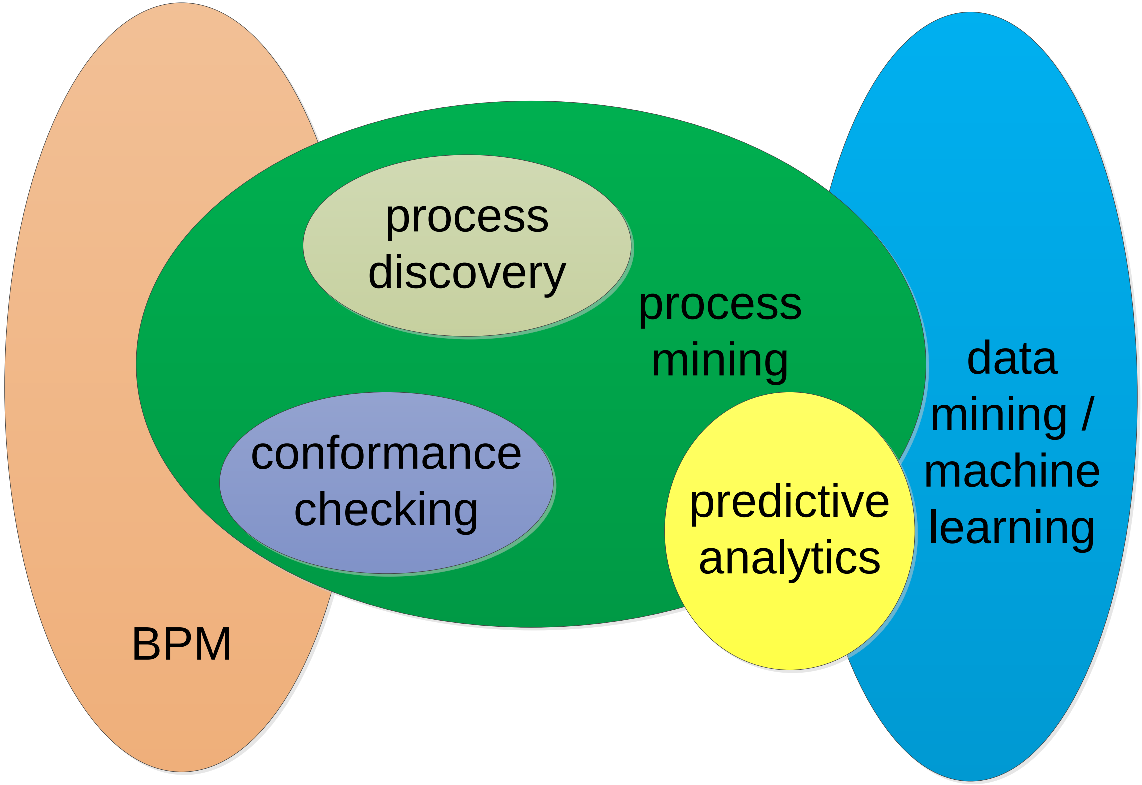
(data mining, machine learning, business intelligence)

data-oriented analysis

# Connecting things: Process mining as super glue

- Data – Process
- Business – IT
- Business Intelligence – Business Process Management
- Performance – Compliance
- Runtime – Design time
- …

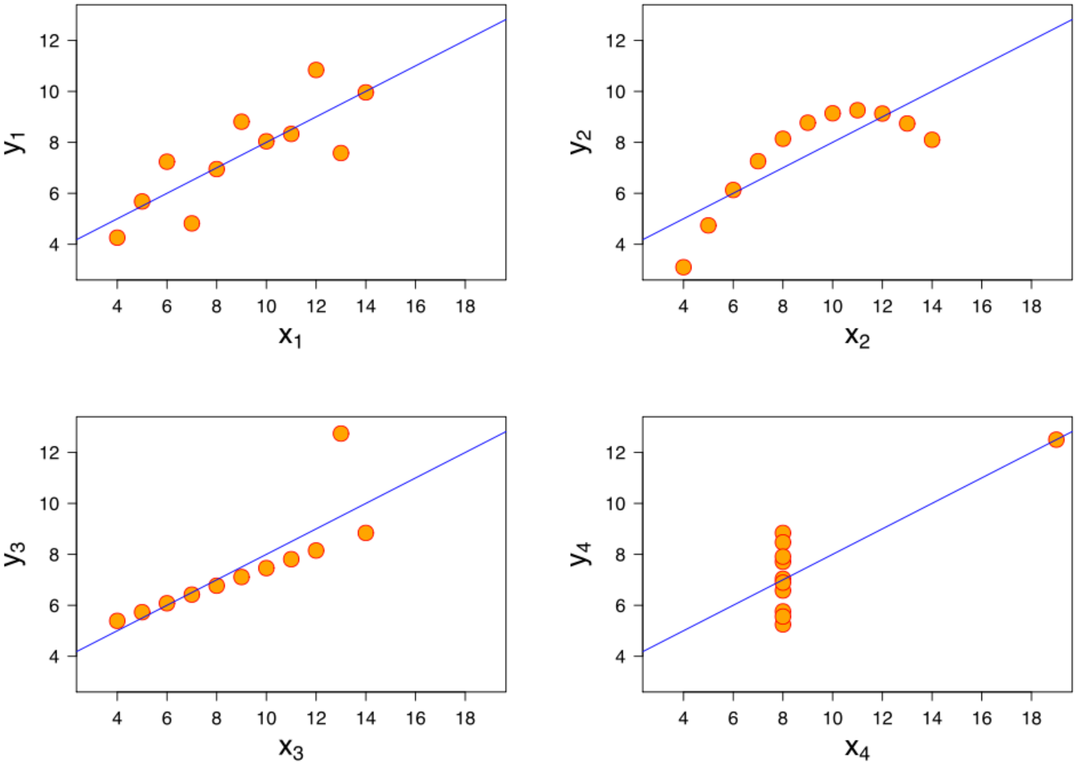
# Positioning Process Mining



**How about BI (Business Intelligence)?**

# Don't try to capture reality in a simple KPI!

**(Like BI tools do)**



**4 data sets of 11 elements**

## Anscombe's Quartet

**mean x = 9**

**variance x = 11**

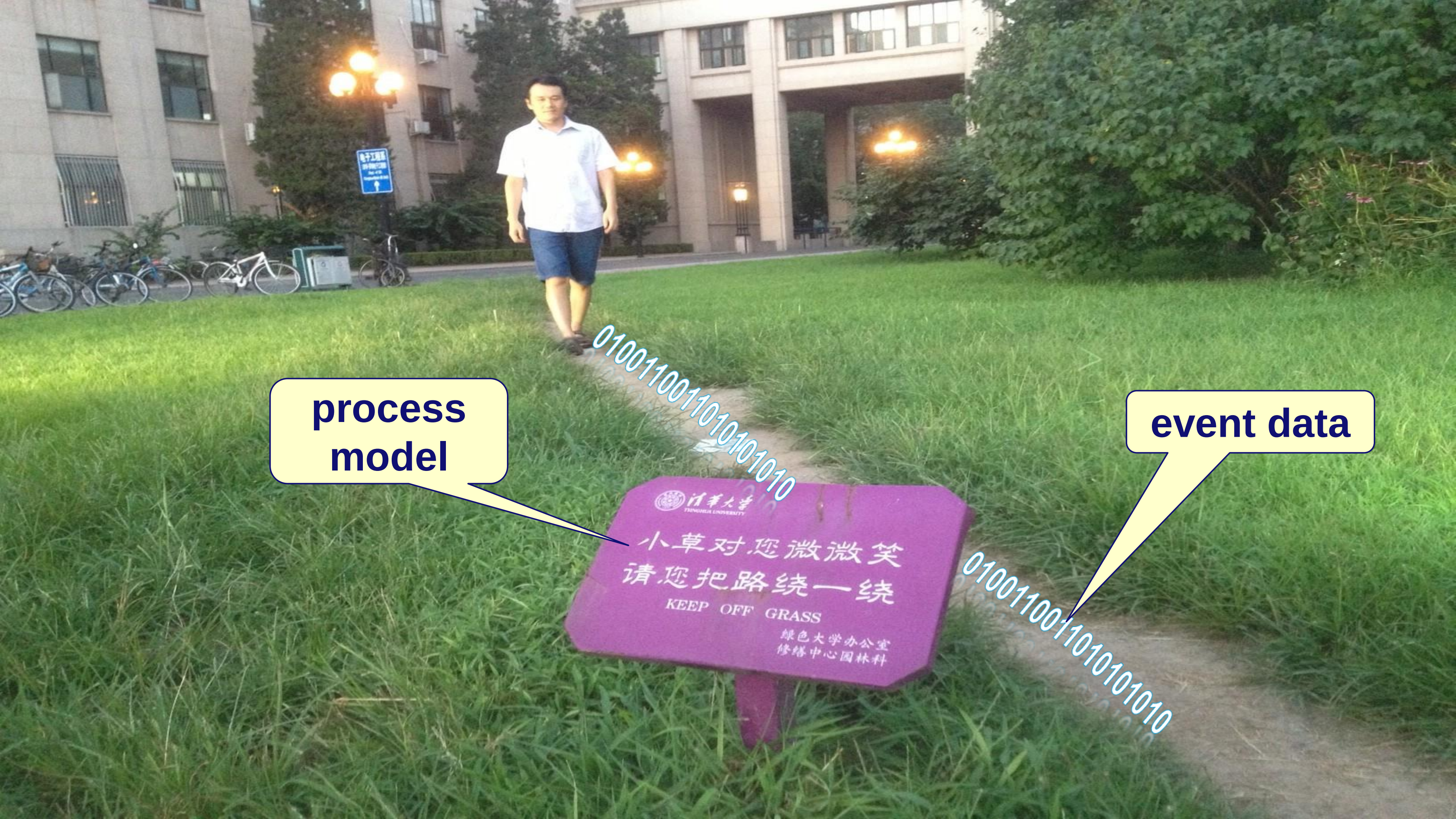
**mean y = 7.5**

**variance y = 4.12**

**correlation = 0.816**

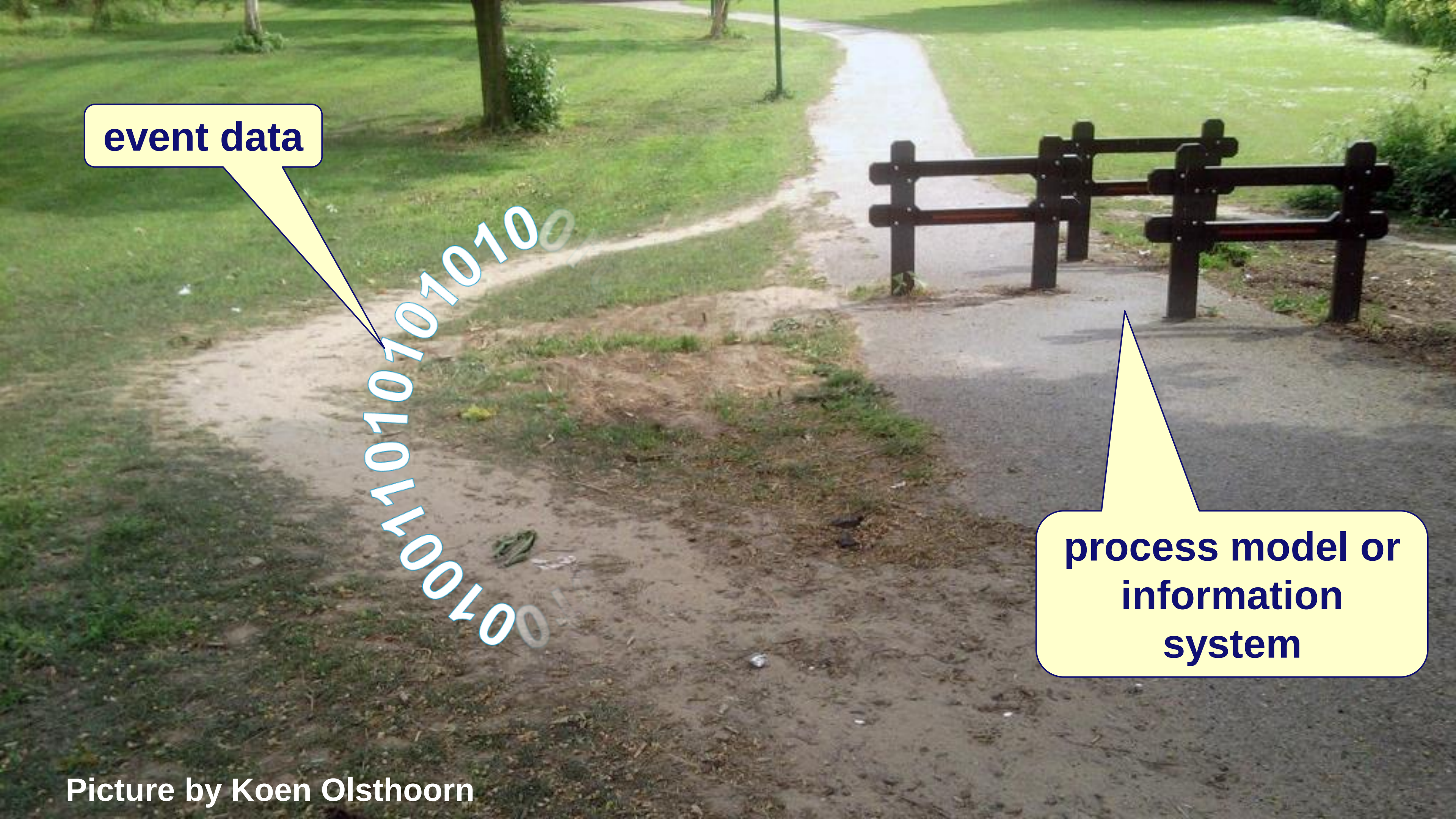
**same linear regression**

Francis Anscombe 1973, Figure by Schutz / CC BY



process model

event data



event data

process model or information system

Picture by Koen Olsthoorn

Demo

# Process discovery is like learning a language: By example

abc ?

ab(c|d) ?

(ad)|(ab(c|d)) ?

ab*(c|d) ?

abc

abd

ad

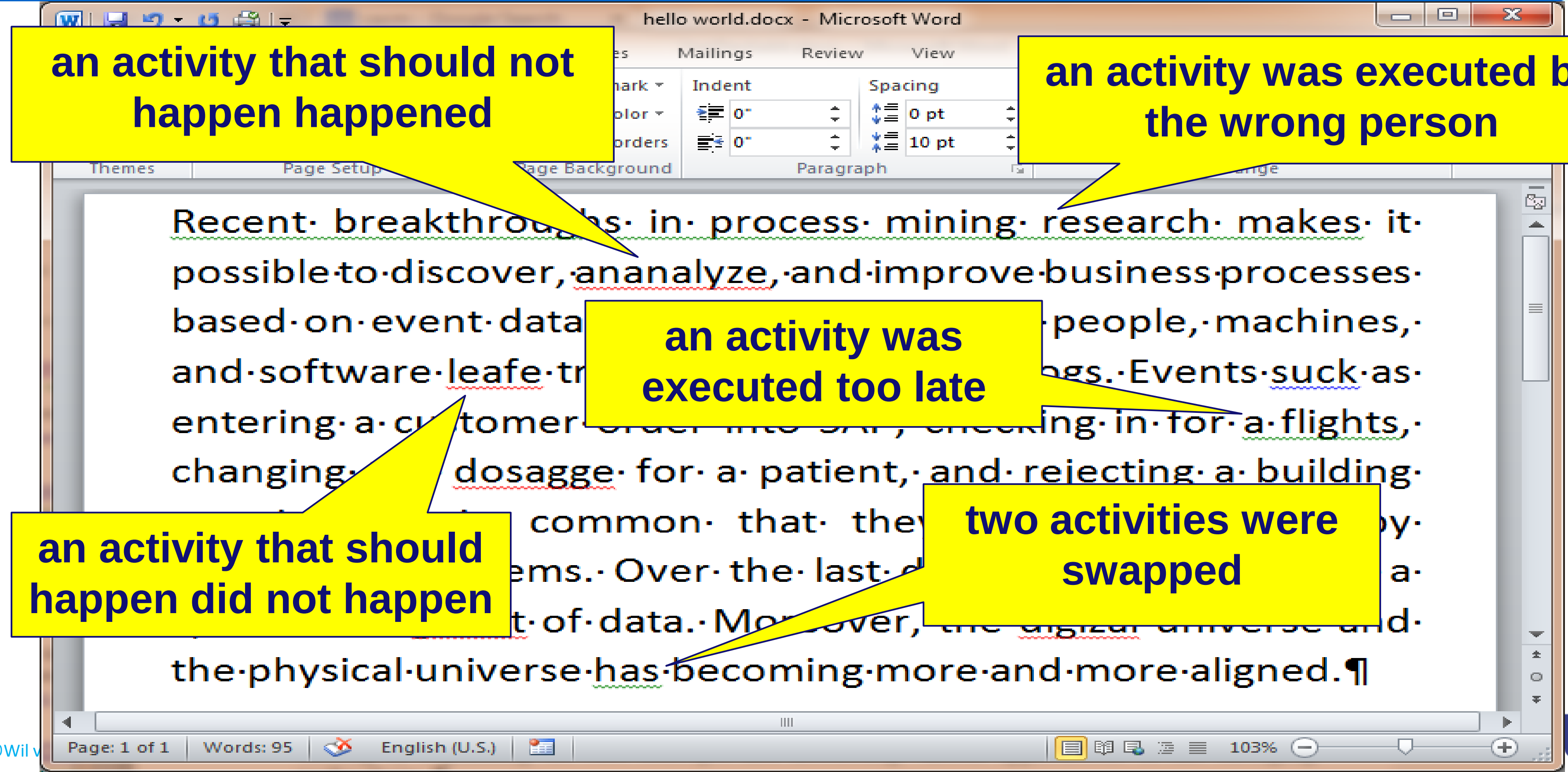
abbc

ac

sentence ≅ trace in event log …

language ≅ process model

# Conformance checking is like spell checking



Indent 0" 0"

Spacing 0 pt 10 pt

Recent breakthrou...s in process mining research makes it possible to discover, ananalyze, and improve business processes based on event data ... people, machines, and software leafe tr... ogs. Events suck as entering a cu...tomer order into SAP, checking in for a flights, changing ... dosagge for a patient, and rejecting a building ... common that the... by ... ems. Over the last d... a ... t of data. Moreover, the digital universe and the physical universe has becoming more and more aligned.¶

Page: 1 of 1 | Words: 95 | English (U.S.) | 103%

an activity that should not happen happened

an activity was executed by the wrong person

an activity was executed too late

an activity that should happen did not happen

two activities were swapped

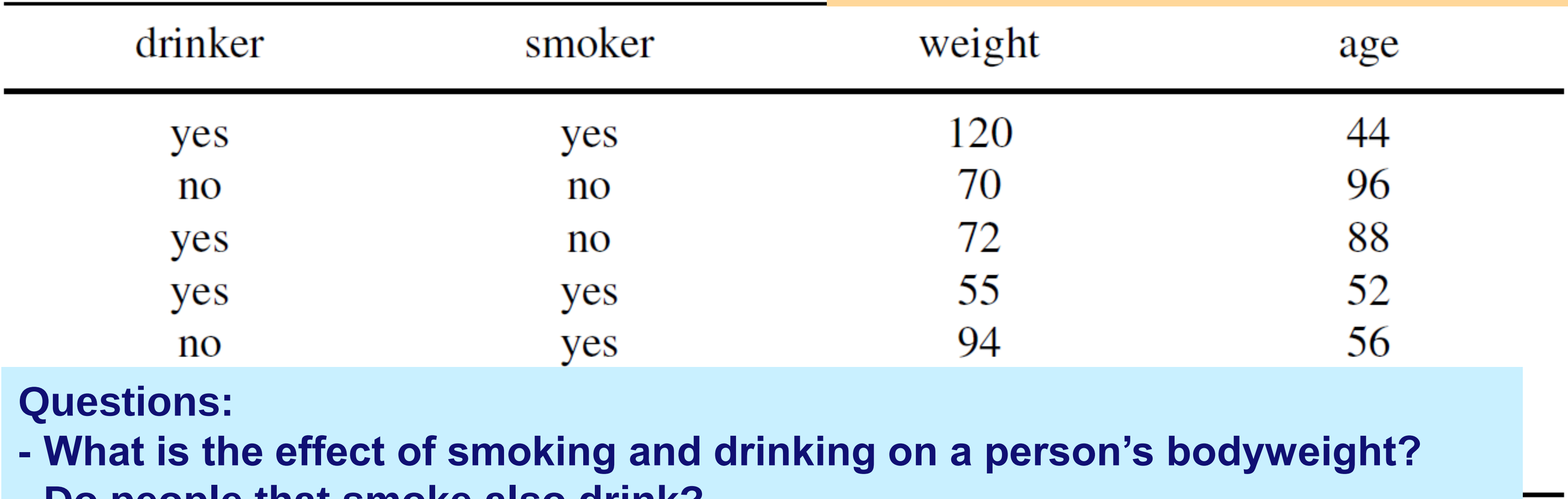
Data Mining
Straight Ahead

# Data mining

- The growth of the "digital universe" is the main driver for the popularity of data mining.
- Initially, the term "data mining" had a negative connotation ("data snooping", "fishing", and "data dredging").
- Now a mature discipline.
- Data-centric, not process-centric.

# Data set 1

**Data about 860 recently deceased persons to study the effects of drinking, smoking, and body weight on the life expectancy.**



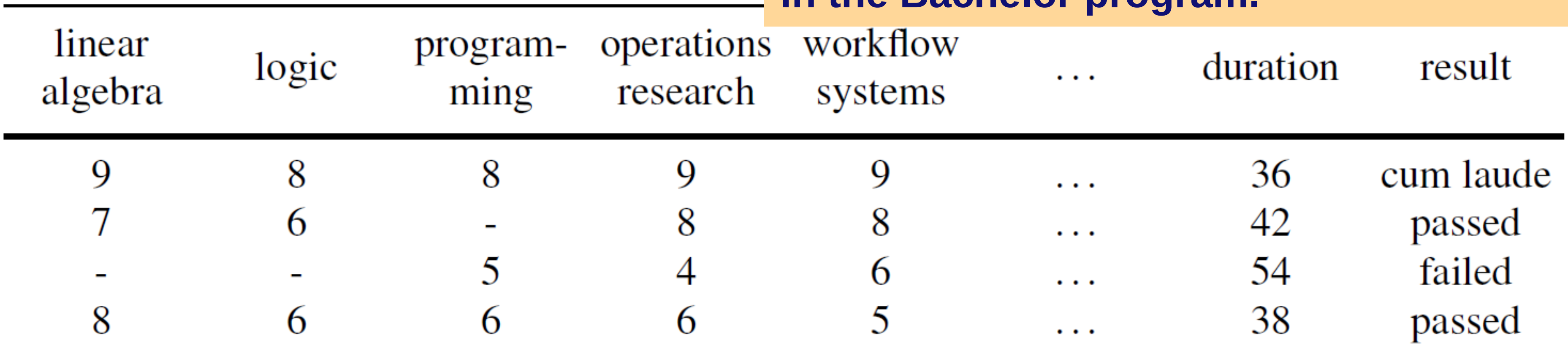
| drinker | smoker | weight | age |
|---|---|---|---|
| yes | yes | 120 | 44 |
| no | no | 70 | 96 |
| yes | no | 72 | 88 |
| yes | yes | 55 | 52 |
| no | yes | 94 | 56 |

**Questions:**

- **What is the effect of smoking and drinking on a person’s bodyweight?**
- **Do people that smoke also drink?**
- **What factors influence a person’s life expectancy the most?**
- **Can one identify groups of people having a similar lifestyle?**

# Data set 2

**Data about 420 students to investigate relationships among course grades and the student's overall performance in the Bachelor program.**



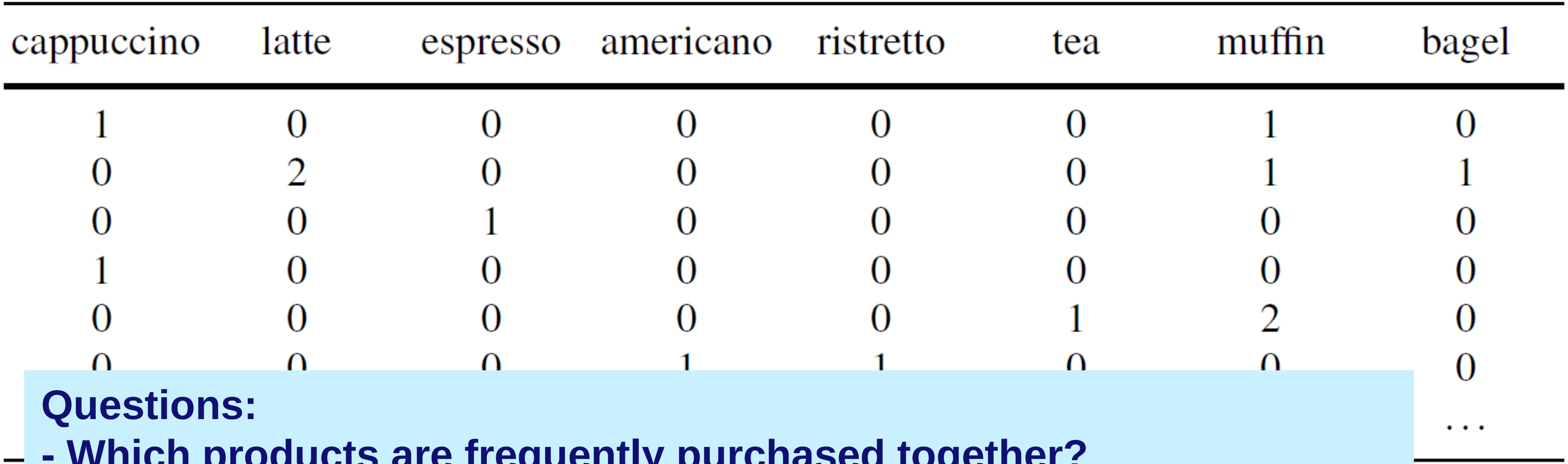
| linear algebra | logic | programming | operations research | workflow systems | ... | duration | result |
|---|---|---|---|---|---|---|---|
| 9 | 8 | 8 | 9 | 9 | ... | 36 | cum laude |
| 7 | 6 | - | 8 | 8 | ... | 42 | passed |
| - | - | 5 | 4 | 6 | ... | 54 | failed |
| 8 | 6 | 6 | 6 | 5 | ... | 38 | passed |

**Questions:**

- **Are the marks of certain courses highly correlated?**
- **Which electives do excellent students (cum laude) take?**
- **Which courses significantly delay the moment of graduation?**
- **Why do students drop out?**
- **Can one identify groups of students having a similar study behavior?**

# Data set 3

**Data on 240 customer orders in a coffee bar recorded by the cash register.**



| cappuccino | latte | espresso | americano | ristretto | tea | muffin | bagel |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 2 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
|  |  |  |  |  |  |  | ... |

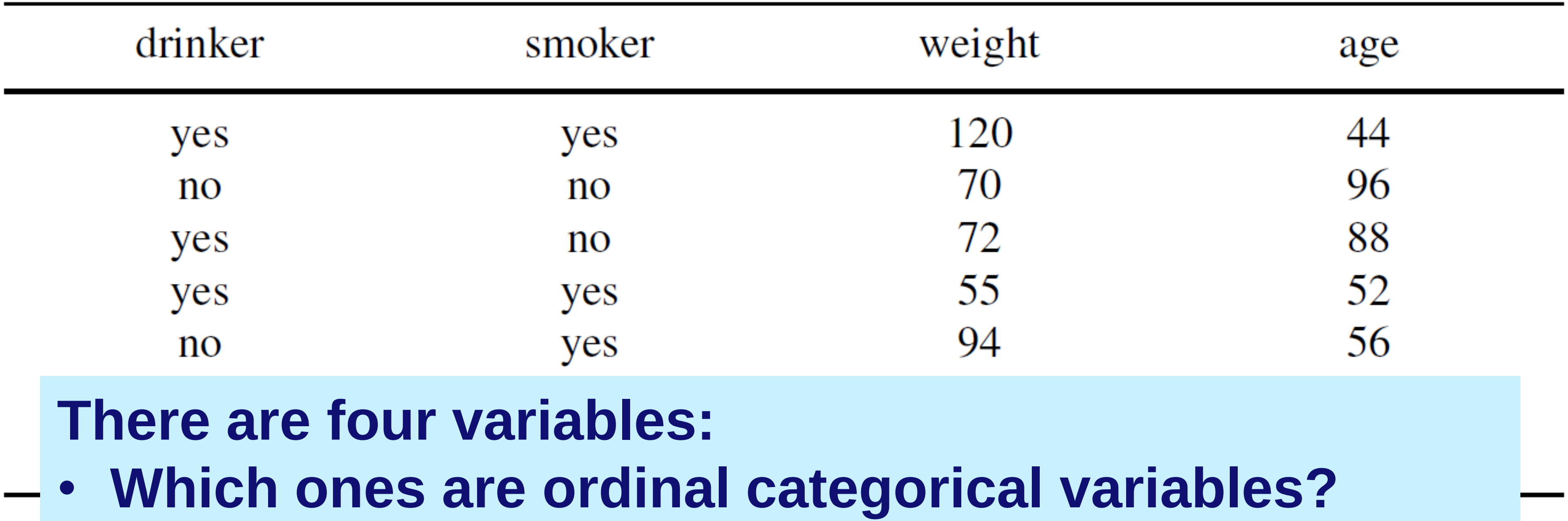
**Questions:**

- **Which products are frequently purchased together?**
- **When do people buy a particular product?**
- **Is it possible to characterize typical customer groups?**
- **How to promote the sales of products with a higher margin?**

# Variables

- Data set (sample or table) consists of instances (individuals, entities, cases, objects, or records).
- Variables are often referred to as attributes, features, or data elements.
- Two types:
  - categorical variables:
    - ordinal (high-med-low, cum laude-passed-failed) or
    - nominal (true-false, red-pink-green)
  - numerical variables (ordered, cannot be enumerated easily)

# Question

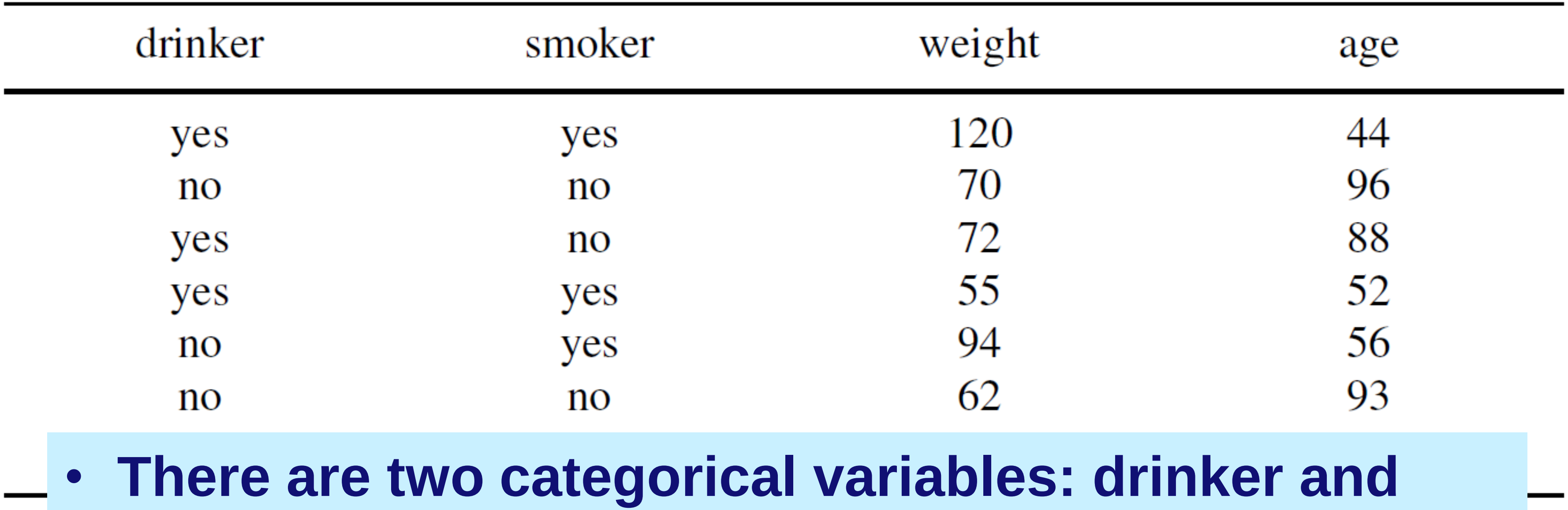


| drinker | smoker | weight | age |
|---|---|---|---|
| yes | yes | 120 | 44 |
| no | no | 70 | 96 |
| yes | no | 72 | 88 |
| yes | yes | 55 | 52 |
| no | yes | 94 | 56 |

**There are four variables:**

- **Which ones are ordinal categorical variables?**
- **Which ones are nominal categorical variables?**
- **Which ones are numerical variables?**

# Answer



| drinker | smoker | weight | age |
|---|---|---|---|
| yes | yes | 120 | 44 |
| no | no | 70 | 96 |
| yes | no | 72 | 88 |
| yes | yes | 55 | 52 |
| no | yes | 94 | 56 |
| no | no | 62 | 93 |

- **There are two categorical variables: drinker and smoker. Both are nominal.**
- **There are two numerical variables: weight and age.**

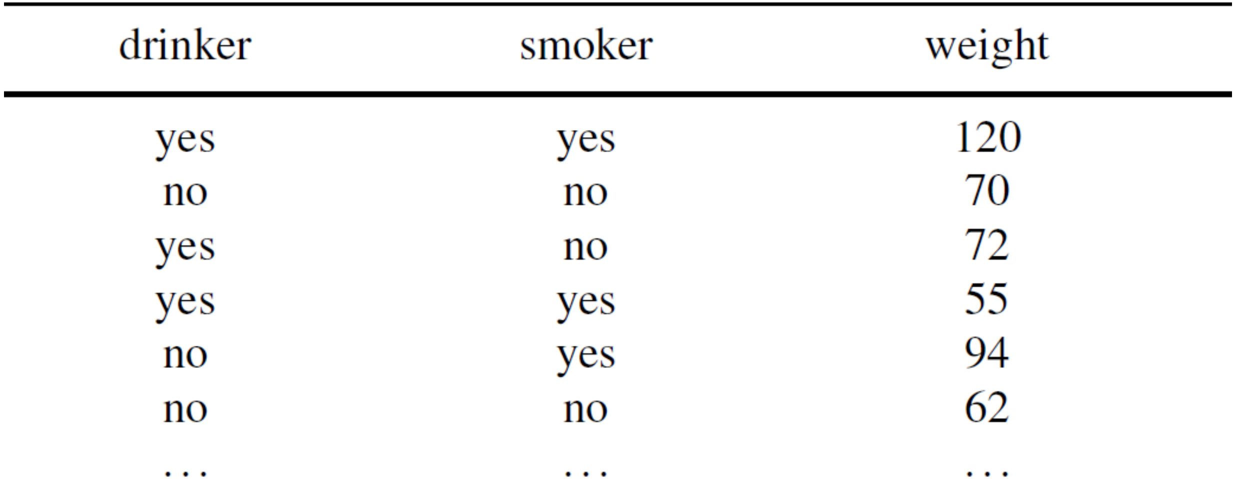
# Supervised Learning

- Labeled data, i.e., there is a response variable that labels each instance.
- Goal: explain response variable (dependent variable) in terms of predictor variables (independent variables).

# Supervised Learning

- Classification techniques (e.g., decision tree learning) assume a categorical response variable and the goal is to classify instances based on the predictor variables.
- Regression techniques assume a numerical response variable. The goal is to find a function that fits the data with the least error.

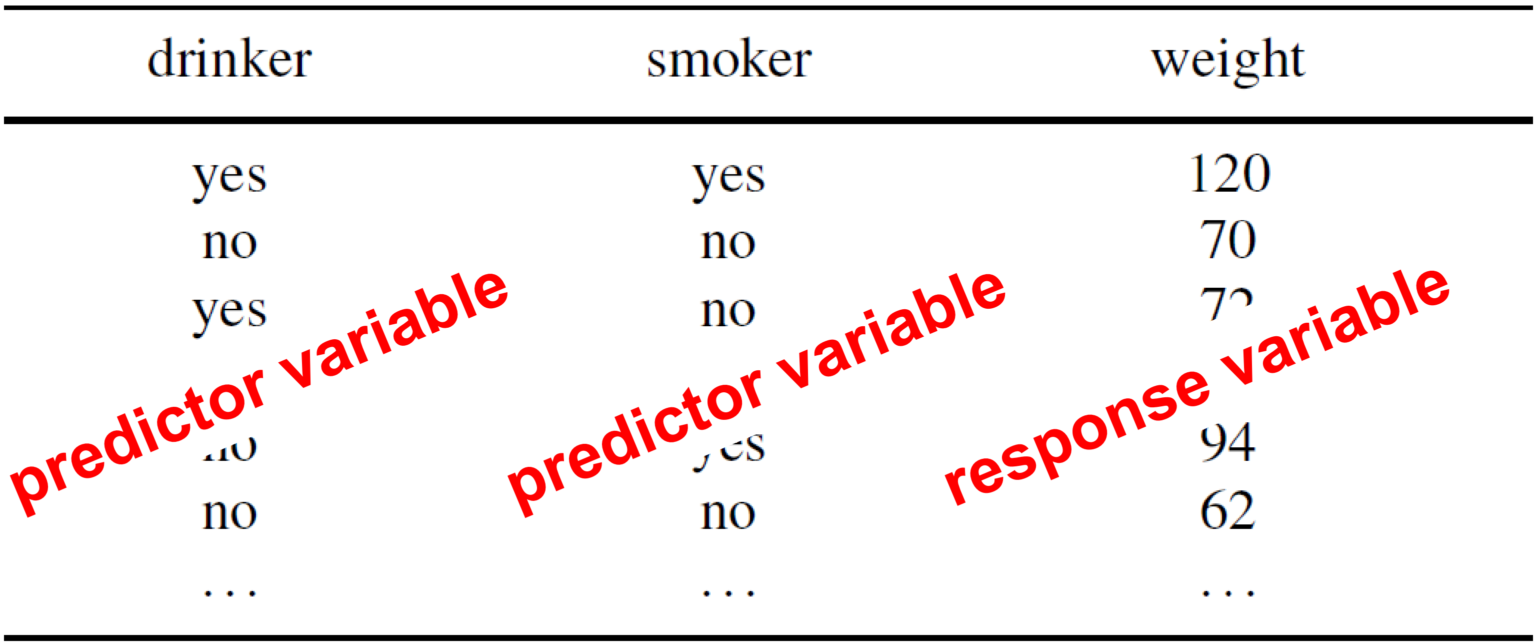
# Question



| drinker | smoker | weight |
|---|---|---|
| yes | yes | 120 |
| no | no | 70 |
| yes | no | 72 |
| yes | yes | 55 |
| no | yes | 94 |
| no | no | 62 |
| ... | ... | ... |

**We would like to learn the influence of drinking and smoking on someone's body weight. What are the response  and predictor variables?**

# Answer



| drinker | smoker | weight |
|---|---|---|
| yes | yes | 120 |
| no | no | 70 |
| yes | no | 72 |
| [illegible] | [illegible] | [illegible] |
| no | yes | 94 |
| no | no | 62 |
| ... | ... | ... |

predictor variable

predictor variable

response variable

# Unsupervised Learning

- Unsupervised learning assumes unlabeled data, i.e., the variables are not split into response and predictor variables.
- Examples: clustering (e.g., k-means clustering and agglomerative hierarchical clustering)  and pattern discovery (association rules)

# Data Mining Tools

- **RapidMiner** (rapidminer.com, partly commercial)
- **R** (r-project.org, free)
- **Weka** (www.cs.waikato.ac.nz/ml/weka/, GNU)
- **KNIME** (knime.org, partly commercial)
- **SAS** (sas.com, commercial)
- **IBM**
- **IBM**
- **QlikView** (qlikview.com, commercial)
- **SAP BusinessObjects/HANA** (www.sap.com/pc/analytics/, commercial)

**We will use RapidMiner to illustrate classical data mining techniques.**

# Process Mining Versus Data Mining

- Both start from data.
- Data mining techniques are typically not process-centric.
- Topics such as process discovery, conformance checking, and bottleneck analysis are not addressed by traditional data mining techniques.

# Process Mining Versus Data Mining

- End-to-end process models and concurrency are essential for process mining.
- Process mining assumes event logs where events have timestamps and refer to cases (process instances).
- Process mining and data mining need to be combined for more advanced questions.

*Part I: Introduction*

**Chapter 1**
Data Science in Action

**Chapter 2**
Process Mining: The Missing Link

*Part II: Preliminaries*

**Chapter 3**
Process Modeling and Analysis

**Chapter 4**
Data Mining

*Part III: From Event Logs to Process Models*

**Chapter 5**
Getting the Data

**Chapter 6**
Process Discovery: An Introduction

**Chapter 7**
Advanced Process Discovery Techniques

*Part IV: Beyond Process Discovery*

**Chapter 8**
Conformance Checking

**Chapter 9**
Mining Additional Perspectives

**Chapter 10**
Operational Support

*Part V: Putting Process Mining to Work*

**Chapter 11**
Process Mining Software

**Chapter 12**
Process Mining in the Large

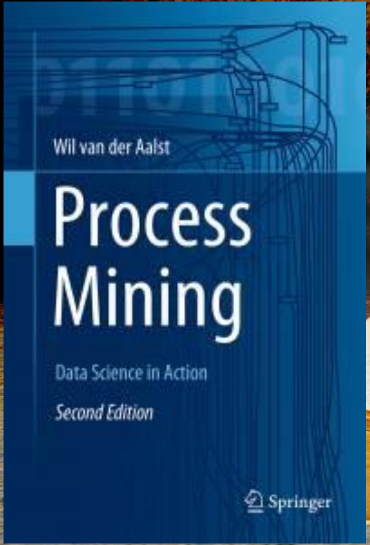
**Chapter 13**
Analyzing “Lasagna Processes”

**Chapter 14**
Analyzing “Spaghetti Processes”

*Part VI: Reflection*

**Chapter 15**
Cartography and Navigation

**Chapter 16**
Epilogue



Wil van der Aalst

Process Mining

Data Science in Action

Second Edition

Springer