Project understanding

General

- Problem formulation
- Mapping the problem formulation to a data analysis task
- Understanding the situation (available data, suitability of the data,...)

The 80-20 Rule!

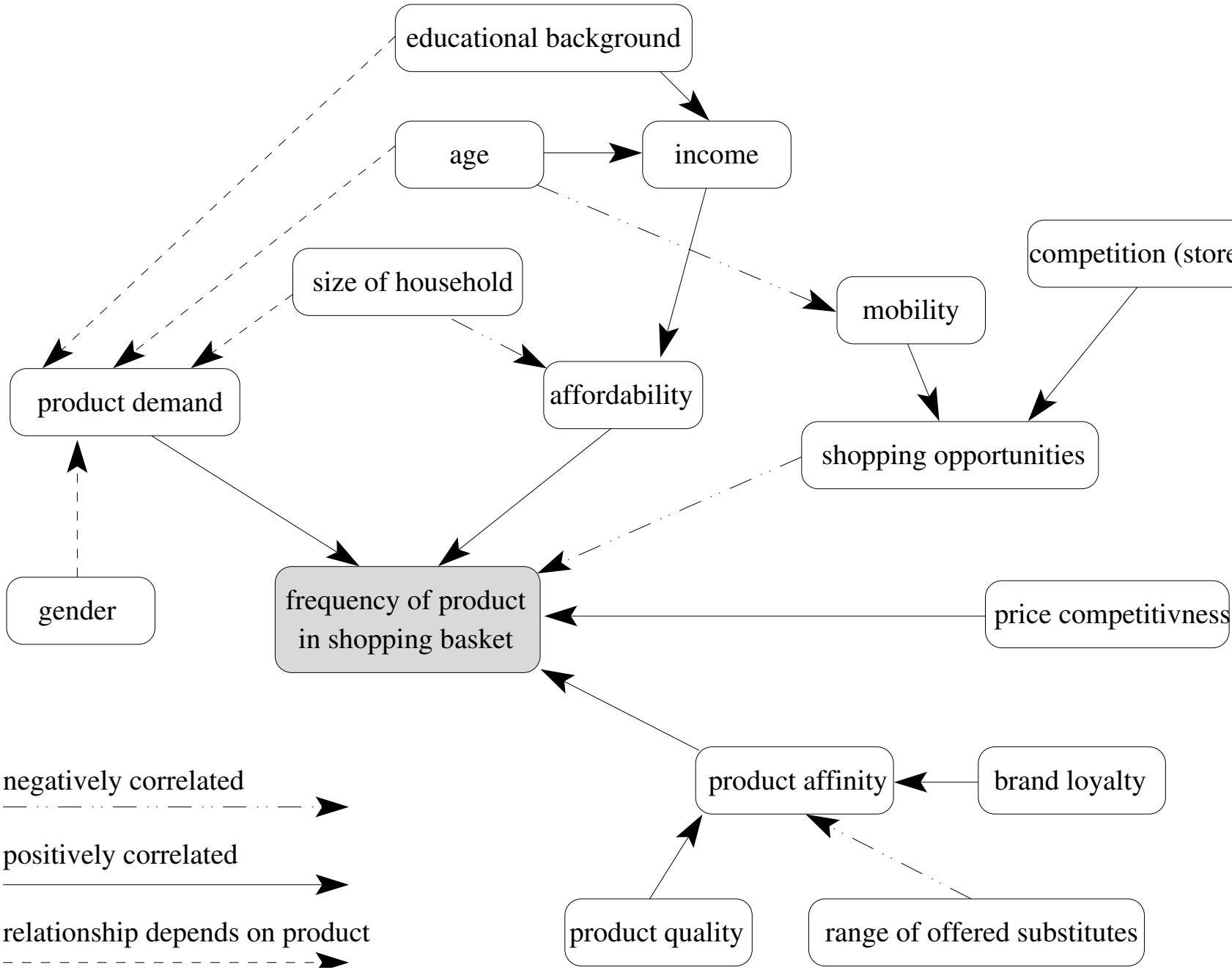
- Average time spent for project and data understanding within the CRISP-DM model: 20%
- Importance for success: 80%

1 Determine the Project Objective

problem source	project owner perspective	analyst perspective
communication	project owner does not understand the technical terms of the analyst	analyst does not understand the terms of the domain of the project owner
lack of understanding	project owner was not sure what the analyst could do or achieve models of analyst were different from what the project owner envisioned	analyst found it hard to understand how to help the project owner
organization	requirements had to be adopted in later stages as problems with the data became evident	project owner was an unpredictable group (not so concerned with the project)

Table: Problems faced in data analysis projects.

Cognitive Map



Determine project objective

- The aim of the project should be clearly defined.
- Criteria to measure the success of the project should be defined.

Example

objective:	increase revenues (per campaign and/or per customer) in direct mailing campaigns by personalized offer and individual customer selection
deliverable:	software that automatically selects a specified number of customers from the database to whom the mailing shall be sent, runtime max. half-day for database of current size
success criteria:	improve order rate by 5% or total revenues by 5%, measured within 4 weeks after mailing was sent, compared to rate of last three mailings

Asses the situation

- requirements and constraints
  - model requirements, e.g. model has to be explanatory
  - ethical, political, legal issues, e.g. variables such as gender, age, race must not be used
  - technical constraints, e.g. time limits
- assumptions
  - representativeness:
    The sample represents the whole.
  - informativeness:
    The model includes all important information.
  - good data quality
  - presence of external factors:
    The external world is not changing.

Determine analysis goals

Determine data mining tasks

(classification, regression, cluster analysis, finding associations, deviation analysis,...).

Determine analysis goals

Determine data mining tasks

(classification, regression, cluster analysis, finding associations, deviation analysis,...).

Specify the requirements for the models

Determine analysis goals

Determine data mining tasks

(classification, regression, cluster analysis, finding associations, deviation analysis,...).

Specify the requirements for the models

Determine analysis goals

- Interpretability
- Reproduceability/stability
- Model flexibility/adequacy
- Runtime
- Interestingness and use of expert knowledge