Data Science (CDA)

Knowledge exploitation

- José Hernández Orallo, DSIC, UPV, jorallo@upv.es
- Fernando Martínez Plumed

  - Using material from Cèsar Ferri and Toby Segaran's book "Programming Collective Intelligence" (https://www.safaribooksonline.com/library/view/programming-collective-intelligence/9780596529321/ch02.html)
  - http://grouplens.org/datasets/movielens/

Content

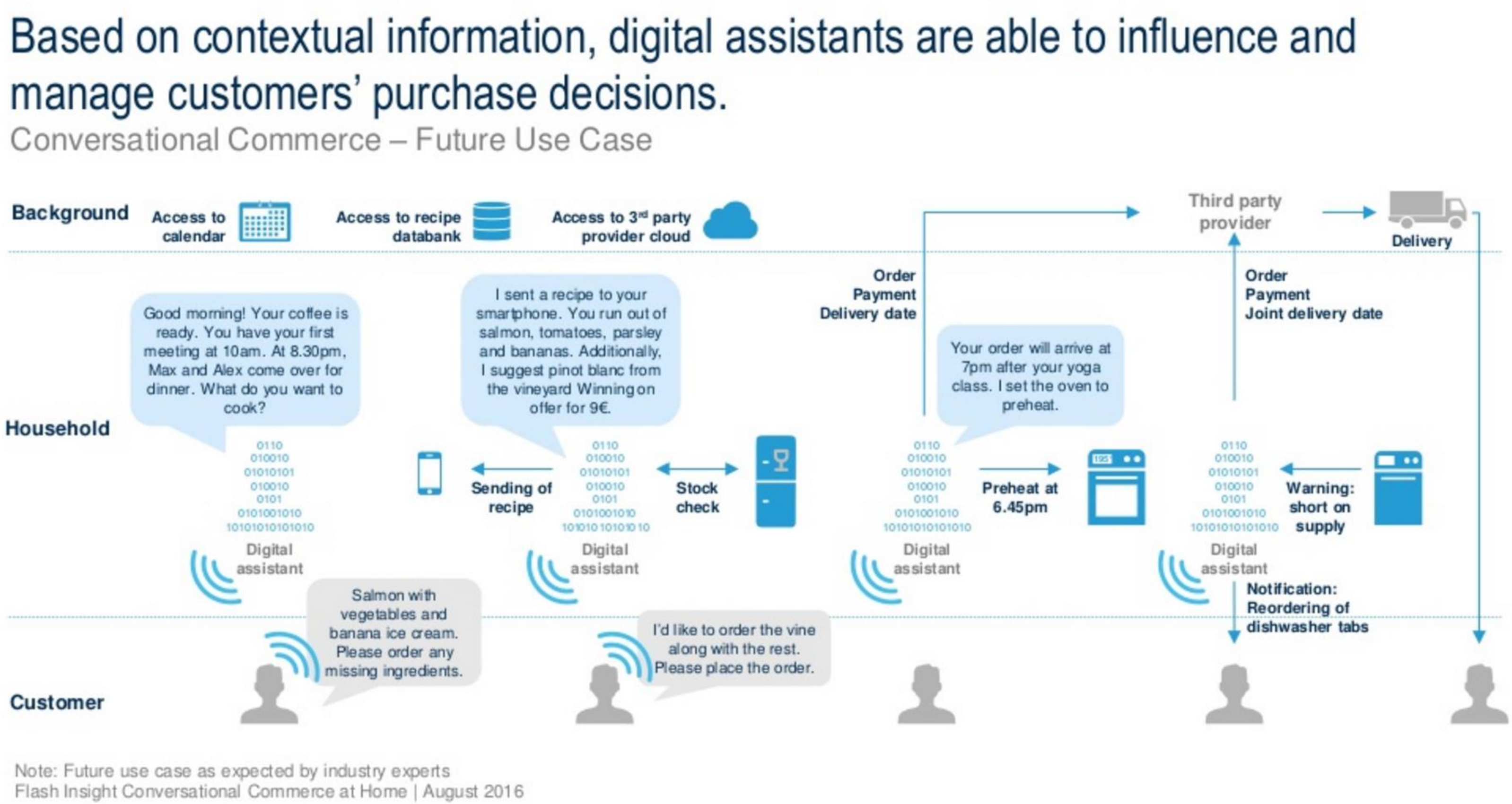
- **Unit 4**: Knowledge exploitation
    - Assistants, prescriptors and recommenders
    - Integration into decision making, dashboards and monitoring.

Assistants and prescriptors

- **Assistants** and **prescriptors** don't give predictions, but must give advice or prescriptions to let the user take the decision:
  - This goes beyond the idea of having soft classifiers, and minimise cost or risk functions.
    - An **assistant** can suggest to take an action and not only be asked about which action to take.
    - A **prescriptor** suggests decisions that are usually accompanied by the implications of those decisions (risks, costs, etc.), and alternative choices.
    - One may (or may not) have a decision made between several courses of action, but never take it!

Assistants and prescriptors

- **Digital assistants** (Siri, Google Now, Cortana, …):



Mücke Sturm Company, http://www.slideshare.net/muecke-sturm/conversational-commerce-65314019

Assistants and prescriptors

- **Prescriptions** (risk or utility management)

Probability		Very Low (0.05)	Low (0.1)	Medium (0.2)	High (0.4)	Very High (0.8)
0.9	Very High 71-90%	0.045	0.09	0.18	0.36	0.72
0.7	High 51-70%	0.035	0.07	0.14	0.28	0.56
0.5	Medium 31-50%	0.025	0.05	0.10	0.20	0.40
0.3	Low 11-30%	0.015	0.03	0.06	0.12	0.24
0.1	Very Low up to 10%	0.005	0.01	0.02	0.04	0.08

Impact

0.1: Minor
0.3: Moderate
0.5: Major
0.7: Critical
0.9: Catastrophic

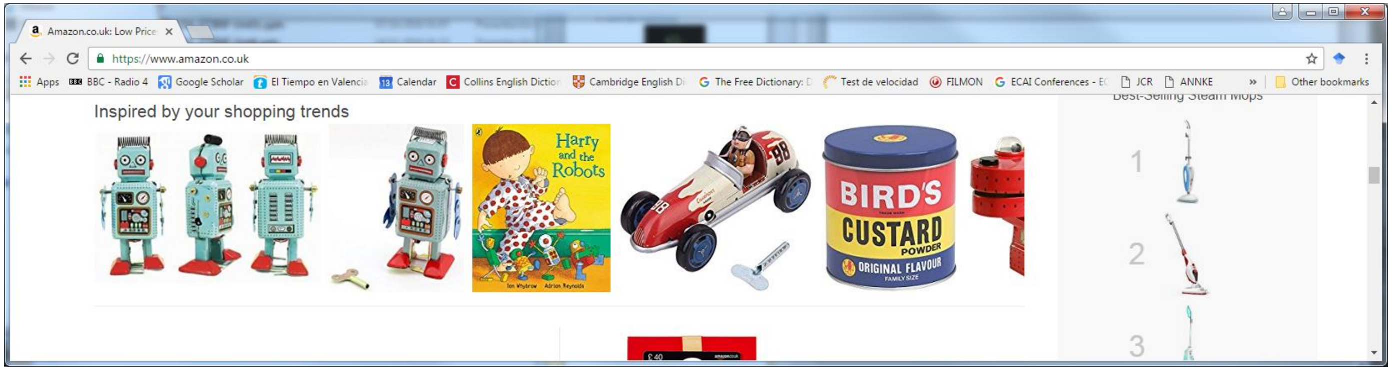
0.05: Very Unlikely
0.1: Possible
0.2: Likely
0.4: Very Likely
0.8: Almost Certain

http://www.ruleworks.co.uk/riskguide/images/risk-tolerance-en.gif

- Risk evaluation is concerned with assessing probability and impact of individual risks.
  - Some risks, such as financial risk, can be evaluated in numerical terms.
  - Others, such as adverse publicity, can only be evaluated in subjective ways.

Risk Log - Example of Security Hazards			Tolerability Level = 0.20
Hazard	**Impact**	**Probability**	**Risk Rating**
Computer system failure	0.05	0.9	0.045
Building damage	0.8	0.1	0.08

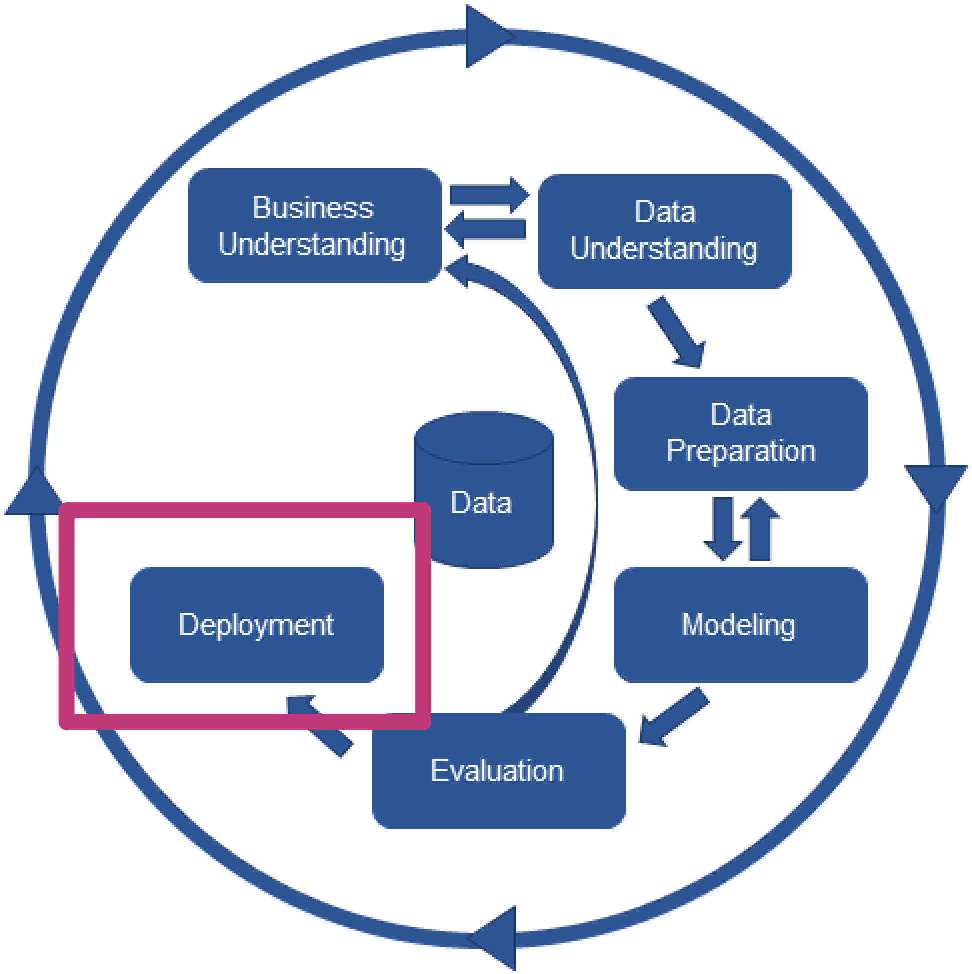
- **Recommenders** are the prototypical case of a prescriptor → **Predict the subjective evaluation a user will give to an item.**
- How things are chosen now?
  - New products are recommended exploiting the evaluations or ratings provided by user(s) for previously viewed or purchased items.
    - E.g., Amazon, netflix, Ciao, Booking, Trip advisor, Be2, Delicious, Youtube



- Approaches:
  - **Collaborative filtering**: They collect and analyse data about the preferences of the users. They recommend new items by observing preferences of <u>similar users</u>
    - Facebook, Linkedin, Last.Fm…
  - **Content-based filtering**: They collect information about items and then similarities are computed among them. When a user expresses preference about an item, the system recommends the most <u>similar items</u>
    - Imdb, Rotten tomatoes, Pandora,…
  - **Hybrid systems**: Combine both techniques
    - Netflix, seethisnext…

- **Unit 4**: Knowledge exploitation
  - Assistants, prescriptors and recommenders
  - Integration into decision making, dashboards and monitoring.

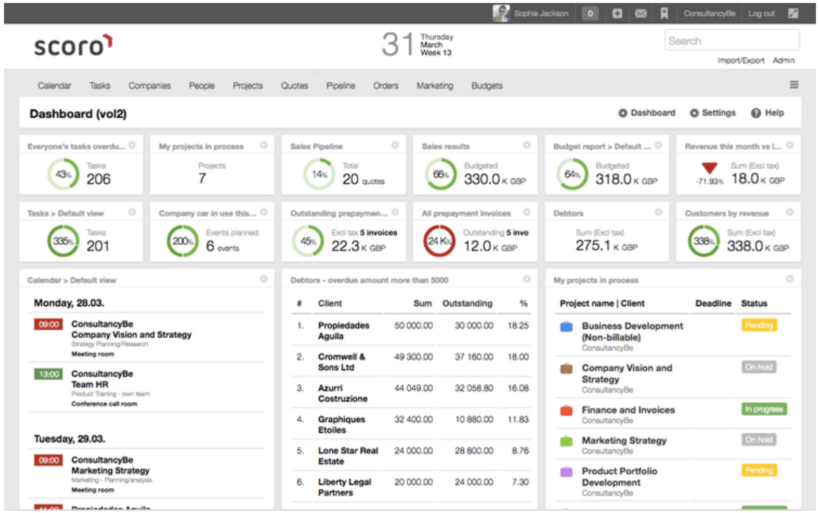
Decision Making



- The idea is to **exploit** the potential of the extracted models, **integrate** them into the decision-making processes of the organisation, **spread** reports about the extracted knowledge, etc.

- In order to **apply a model** we need to:
  - Interpret and understand it.
    - Contrast it with the previous knowledge.
  - Combine it with other models of the same problem or combine them globally (simulation).
  - Exchange and spread it.
    - Integrate it, if necessary, in software applications, DSS and the organisation structure.
  - Adapt it to new application context.

- **Understand the knowledge:**
  - Black-box models can be made more understandable:
    - Converted into rules by rule extraction or distillation.
    - Represented graphically.
  - Models in general can be presented as (interactive) reports on the web.
    - Dashboards may include model prescriptions and simulation.



- **Combining models and simulation**
  - (Not talking about boosting, bagging or random forest)
  - We're talking about using several models (of different characteristics, including manual models).
    - Through weighting majority, stacking or cascade
  - Outputs of a model can be input of another
    - (predictions of the customer affluence model are used as input for sales model).
- **Exchange and Integration.**
  - How to insert models into our applications?
  - How to export models to other platforms?
    - Need for standards:
      - PMML: predictive model markup language
      - http://dmg.org/pmml/products.html

PMML
Predictive Model Markup Language

- **Monitoring**
  - Detects if the model doesn't work as well as it usually did (or as it should).
  - Periodic evaluation using fresh data
    - Alarms
  - Open to comments from the users.
  - Detect context changes.
- **Revision**
  - If the model cannot be adapted to the new context or change of data, revision is needed:
    - Partial: part of the model is changed (e.g., obsolete rules) and part is preserved.
    - Total: the new model changes drastically.
  - If revision is not possible:
    - The model is retrained completely.