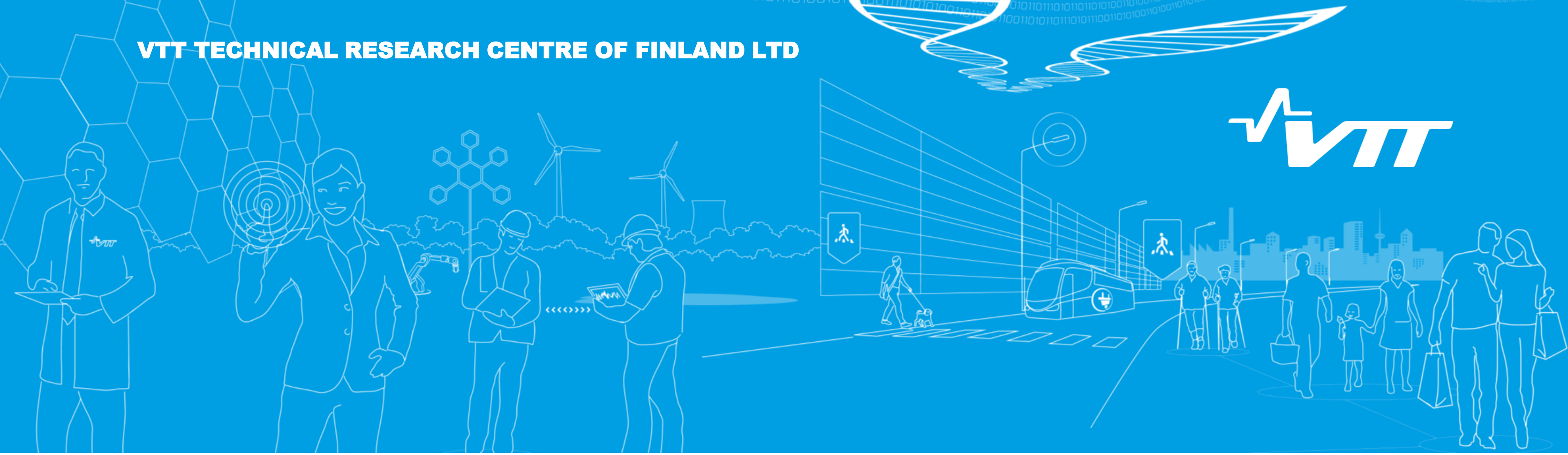
Title	Big Data in Media - results from a Finnish study
Author(s)	Bäck,Asta; Ollikainen,Ville; Södergård, Caj; Vainikainen, Sari
Citation	Nxt Media Conference, 17.11.2015, Trondheim, Norway
Date	2015
Rights	This presentation may be downloaded for personal use only.



VTT TECHNICAL RESEARCH CENTRE OF FINLAND LTD

Big Data in Media – results from a Finnish study

Nxt Media seminar 17.11.2015

Trondheim

Asta Bäck, Ville Ollikainen, Caj Södergård, Sari Vainikainen

What is needed (according to media houses) ?

- Current drawbacks
  - Log data is not collected systematically
  - Legal barriers for linking usage and users
  - Separate, unconnected data from each channel (Mobile, e-paper, PC)
- To be developed
  - Better understanding of subscribers
  - Better article recommendations
  - Adding metadata
  - Ad targeting
  - Predicting ad clicks

Aim

- Help media houses to use their own data for
  - Content recommendation
  - Customer segmentation
  - Predicting user behavior

- Define and set up a test environment

- Run trials on case data sets

- Funded by Finnish Media Fund (Viestintäalan tutkimussäätiö)

Two user data sets

- Data set 1: Cookies (one month, several services)

- Data set 2: Click data (one month, two services)
  - 800.000 data rows

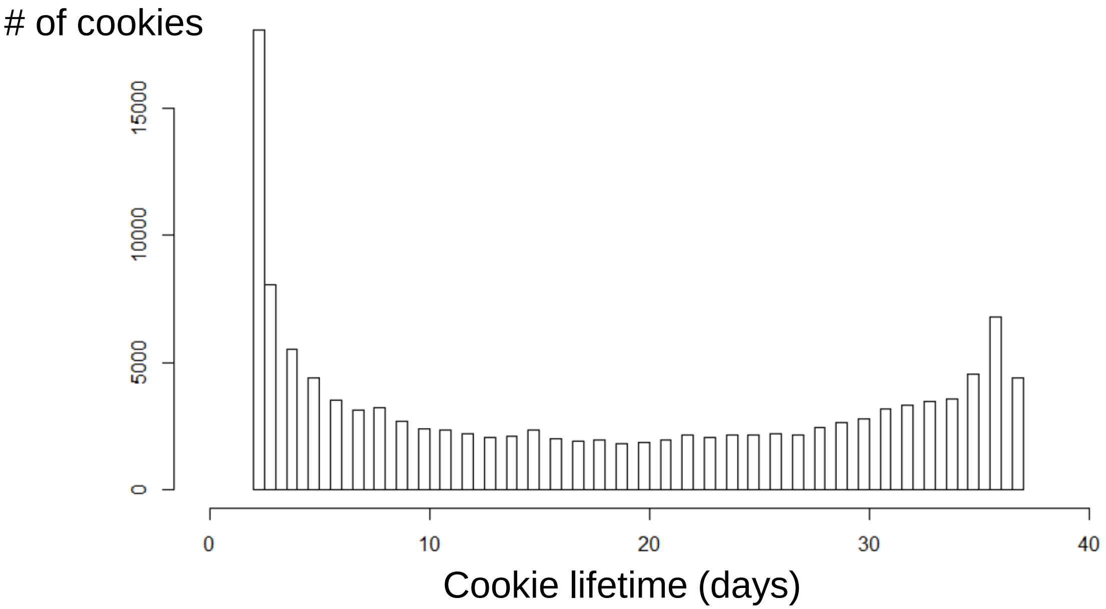
- Both sets contained:
  - Cookie / Anonym(ized) user ID
  - URL of the clicked page
  - Time of the click (1 hour/1 second accuracy)

We used several analyzing programs

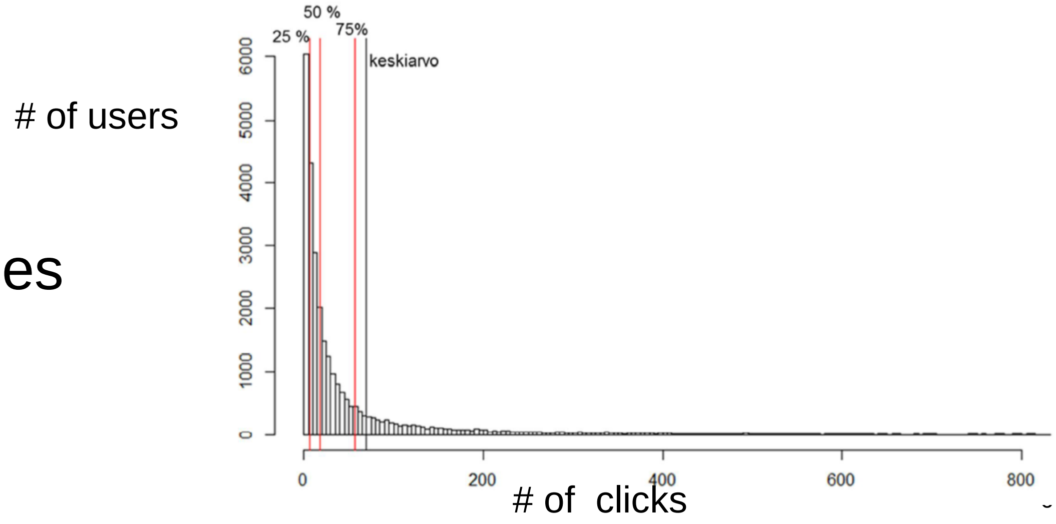
- R – Data exploration and visualization
- Weka – Association rules and k-clustering
- Microsoft Azure Machine Learning – usage prediction
- UPCV (developed by VTT)

Observations

- Large variation in life span of cookies

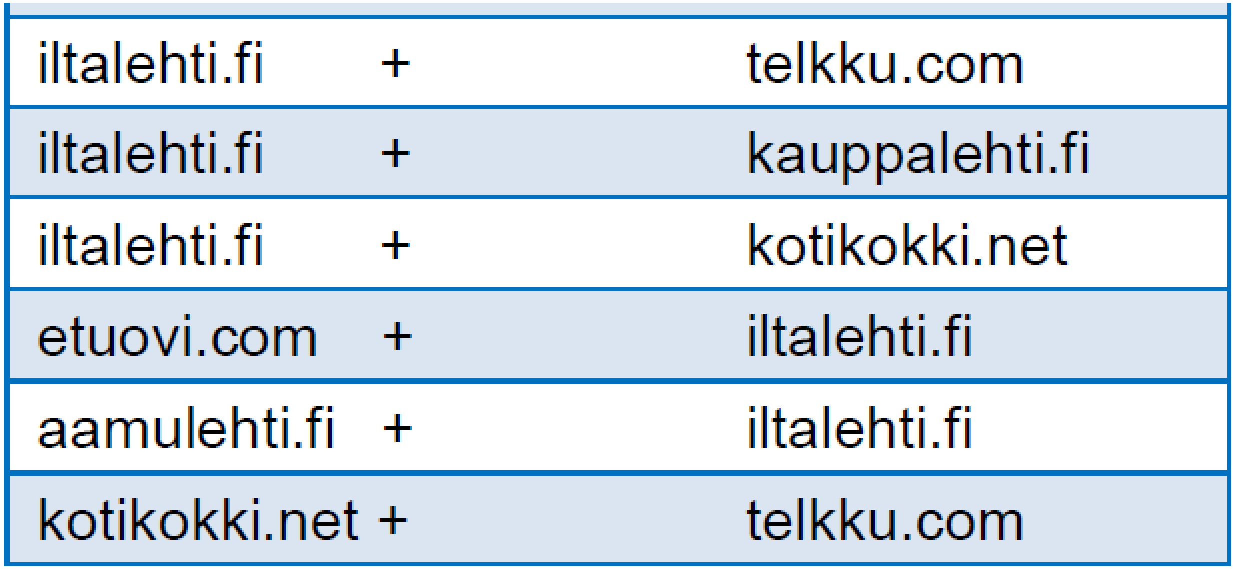


- Most users had only a small number of clicks



- Hardcore users:
    - Used many content services
    - More clicks per service

Surfing paths – evening news is the entry point

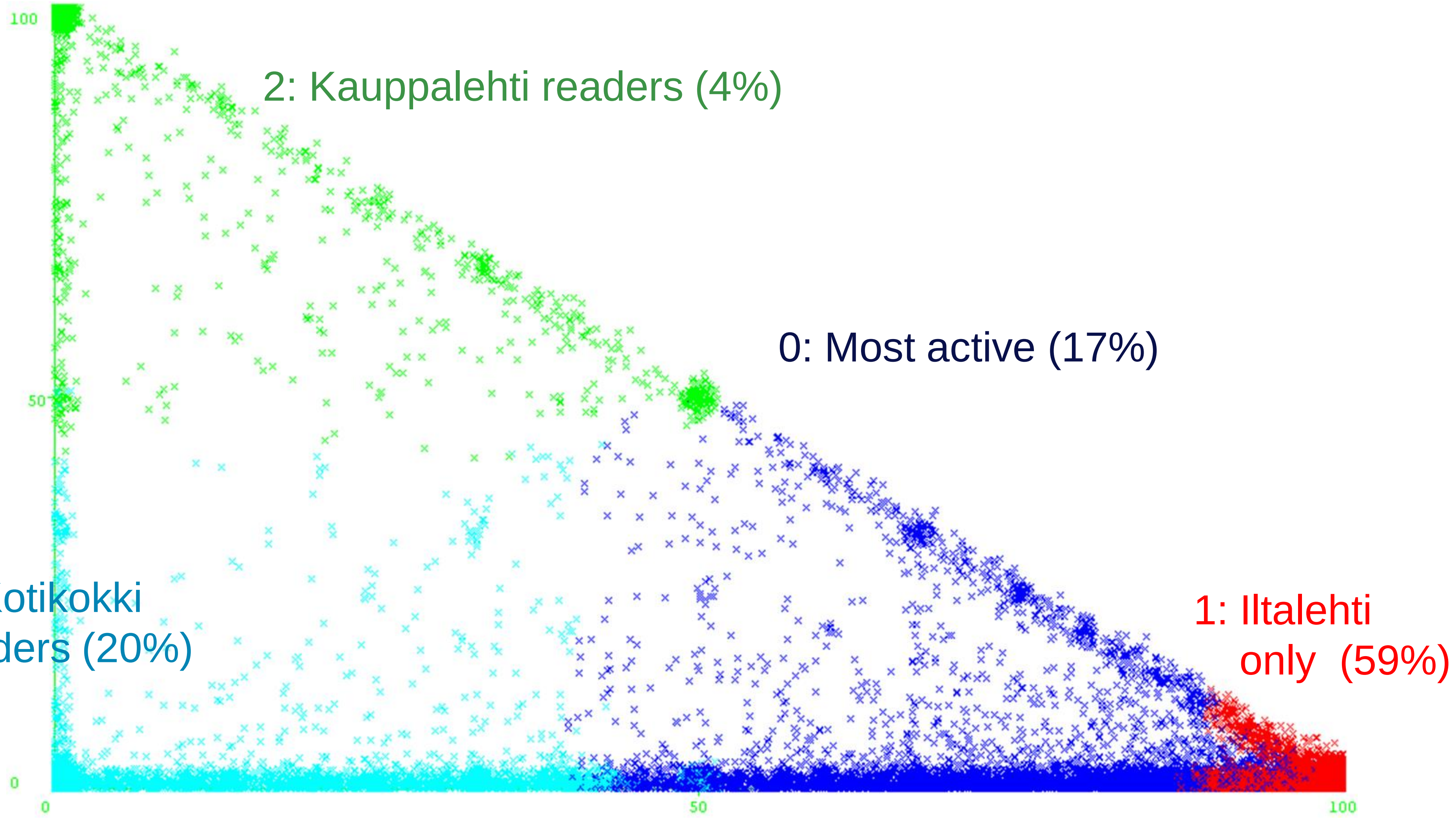


iltalehti.fi	+	telkku.com
iltalehti.fi	+	kauppalehti.fi
iltalehti.fi	+	kotikokki.net
etuovi.com	+	iltalehti.fi
aamulehti.fi	+	iltalehti.fi
kotikokki.net	+	telkku.com

Iltalehti =Evening news
Telkku = TV guide
Etuovi = Apartment advertisements
Aamulehti = Morning news
Kauppalehti=Financial news
Kotikokki = Food and recepes

Analysis with Weka´s Apriori algorithm

Clustering users based on their reading habits



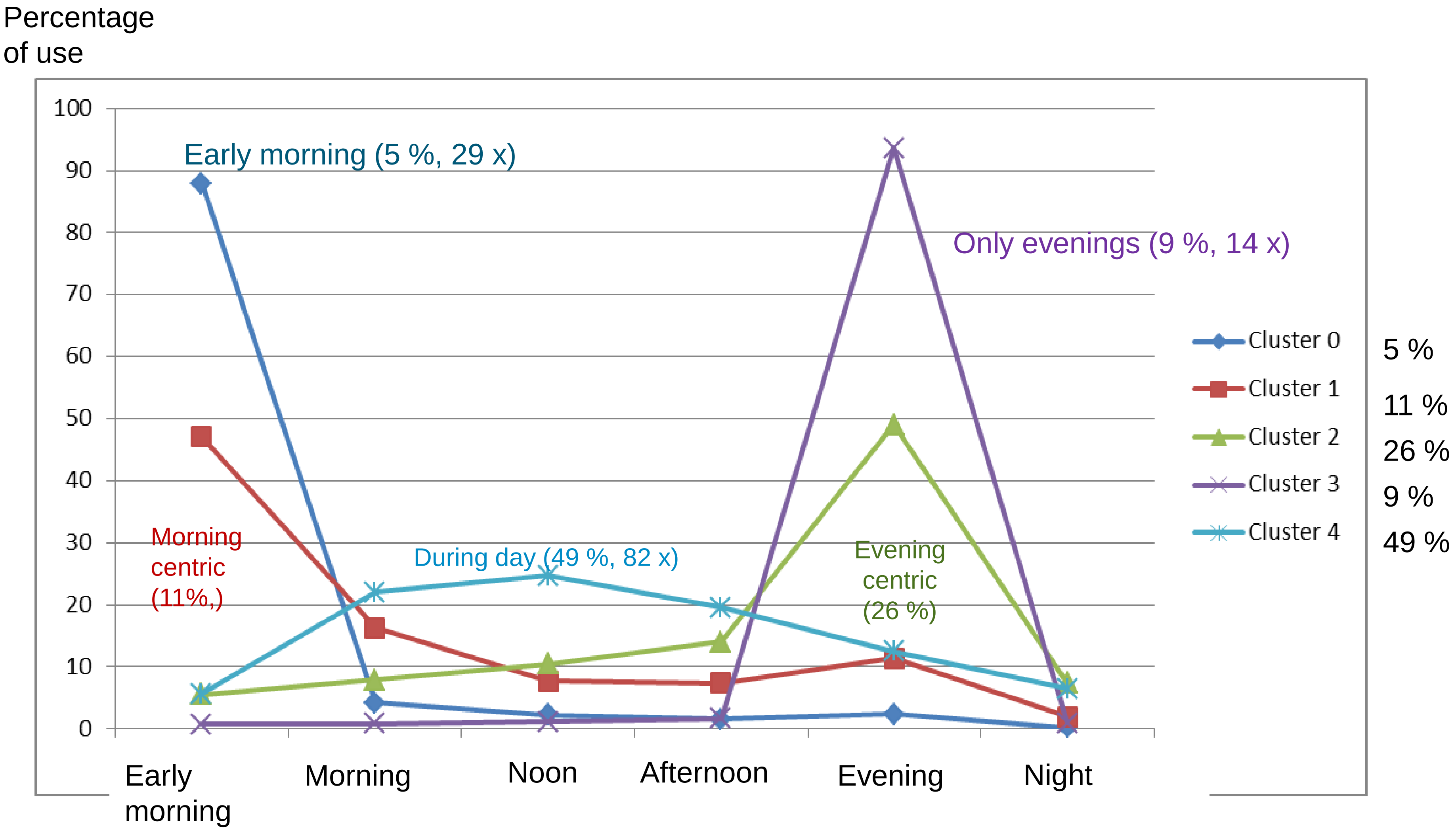
2: Kauppalehti readers (4%)

0: Most active (17%)

3: Kotikokki readers (20%)

1: Iltalehti only (59%)

Most users read daytime



Early morning (5 %, 29 x)

Only evenings (9 %, 14 x)

Morning centric (11%,)

During day (49 %, 82 x)

Evening centric (26 %)

5 %

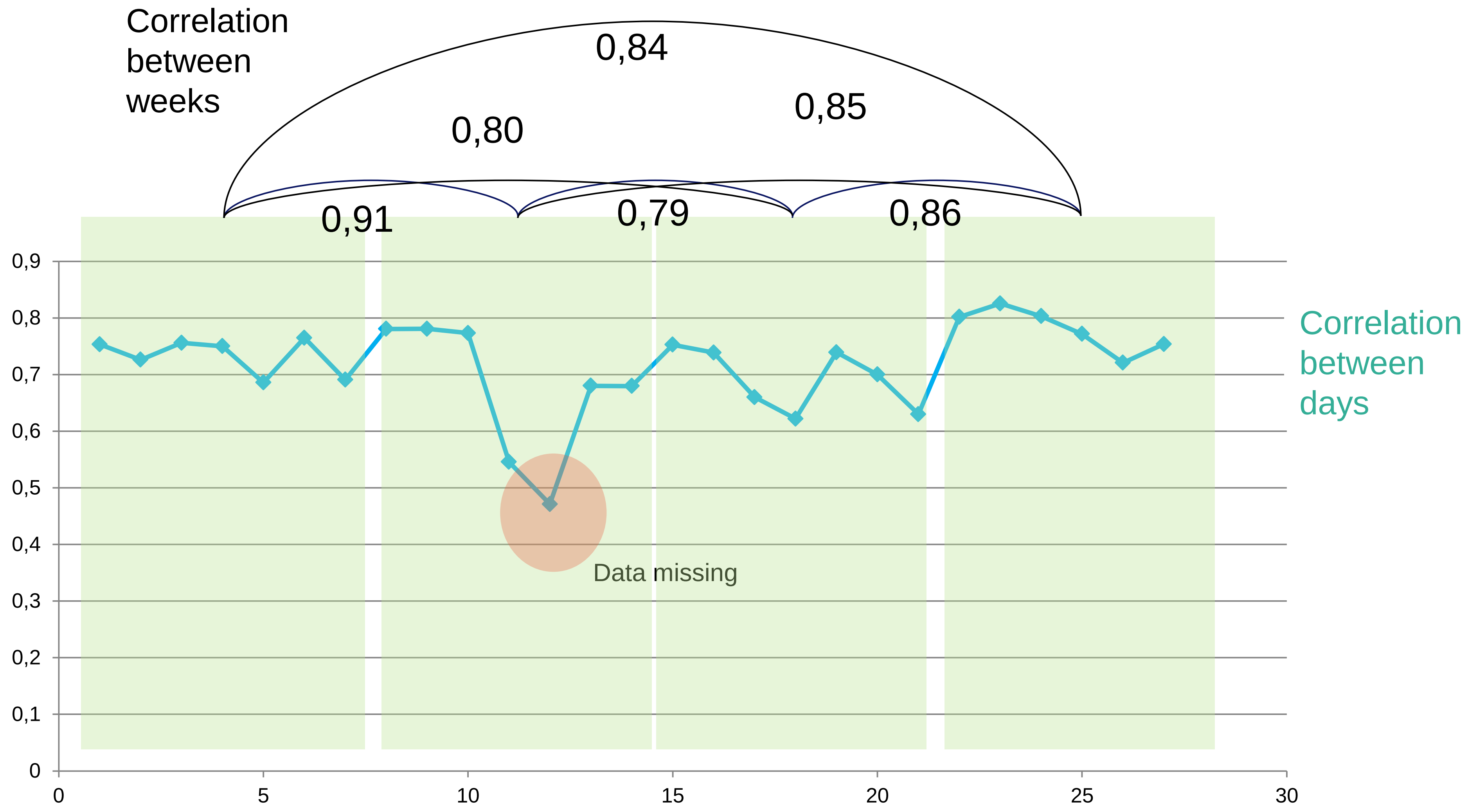
11 %

26 %

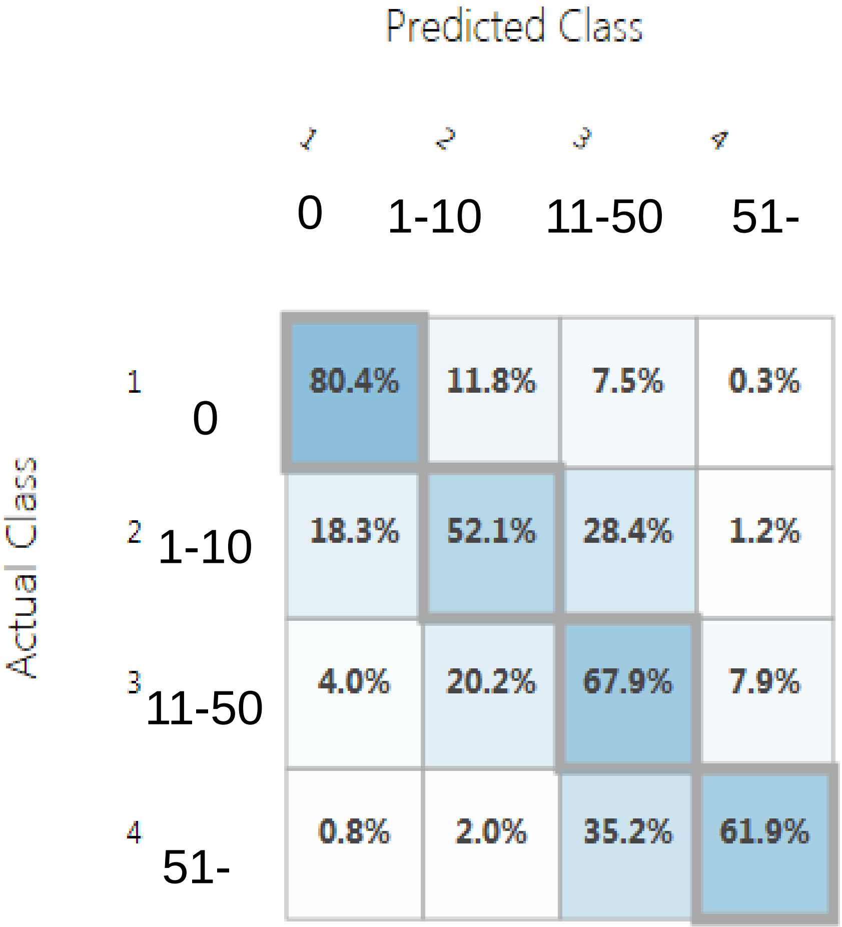
9 %

49 %

Clicks between sequential days and weeks correlate



Predicting the clicks for the week 4



Predicted Class

Actual Class		1	2	3	4
		0	1-10	11-50	51-
1	0	80.4%	11.8%	7.5%	0.3%
2	1-10	18.3%	52.1%	28.4%	1.2%
3	11-50	4.0%	20.2%	67.9%	7.9%
4	51-	0.8%	2.0%	35.2%	61.9%

- Clicks during the week 4 were predicted based on the clicks and usage times in weeks 1…3
- 4 classes
  - 0: 1-10 clicks (2181 users)
  -1-10 clicks (1968 users)
  -11-50 clicks (2183 users)
  >50 clicks (847 users)
- Machine learning method
- Fairly high overall accuracy
- The group 1-10 times per week was the most difficult one

Recommending content

- Collaborative method, no data analysis

UPCV is a novel collaborative recommendation method preserving users' privacy and supportting distributed computing and distributed data reposotories.

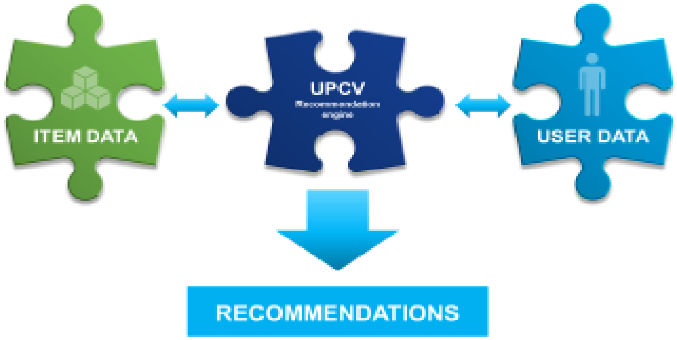
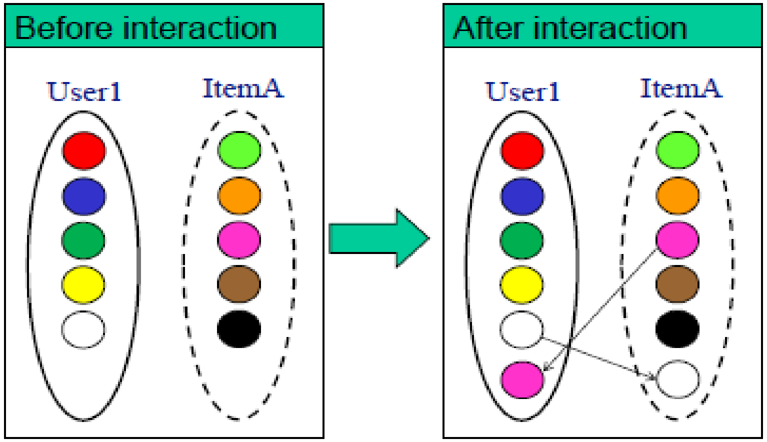


Figure 1. Item data is separated from User data. This allows syncronising data from several services safely.

Figure 2. In UPCV similarity will spread from users to items and from items to users making it possible to inherently provide user-item, item-item, item-user and user-user recommendations..

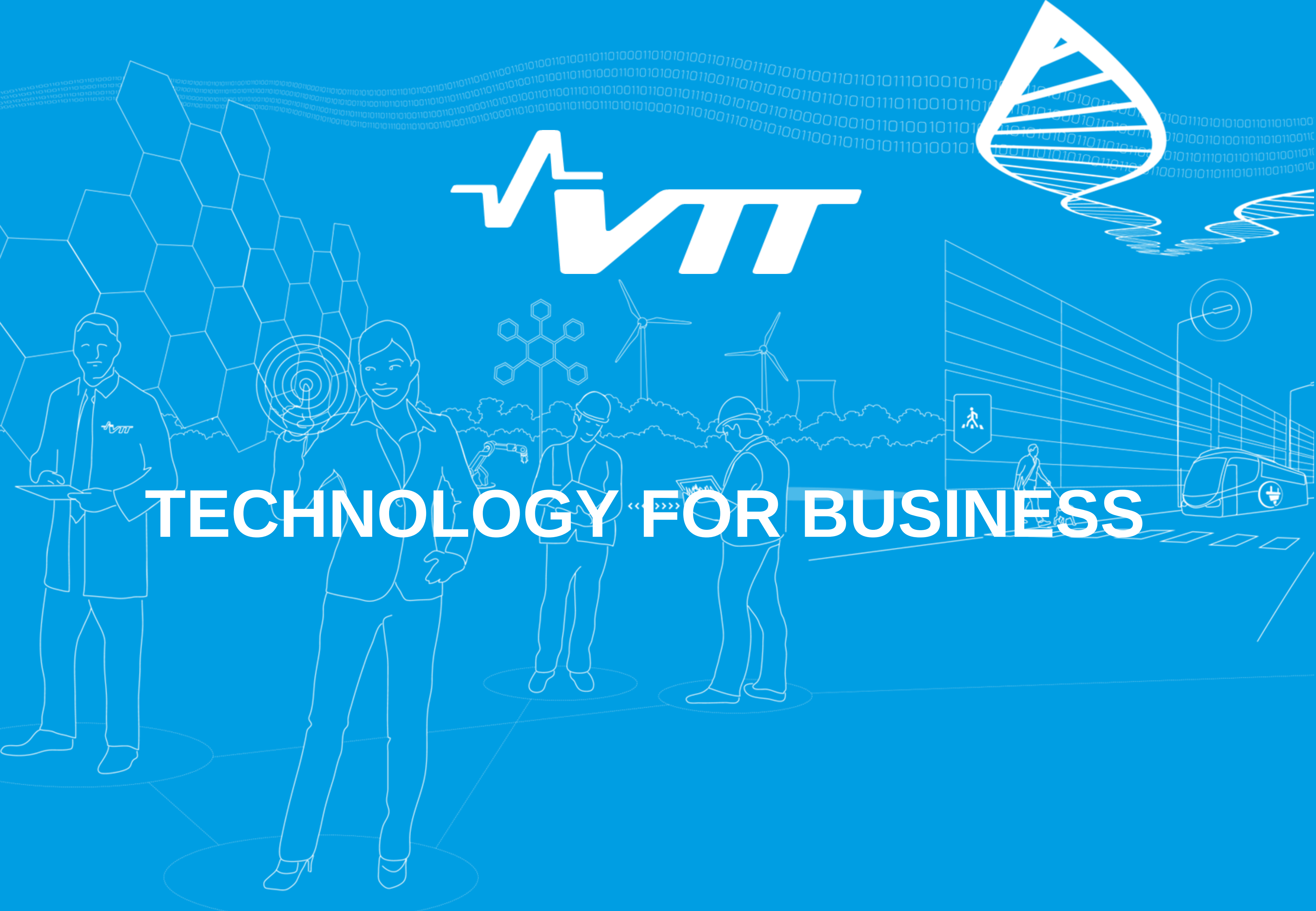


Results from Collaborative Recommendations

- Large data set:
  - 5,6 million users (i.e. cookies) & 308 million rows,
  - 118.000 articles
- Recommendation quality depends heavily on the user
  - Having accessed only popular articles, or only a few articles => few and rather noisy recommendations
  - Having any pattern of reading articles of a **rare topic => relevant recommendations**
    - e.g. among other readings two articles on cyber security and one article on security in general => one Data Security Officer job ad and several IT job ads, not recommended to anyone else in the evaluated subset.
    - (N.B. Pure collaborative approach: no data analysis)

Summary

- Media houses want to understand user behavior (subscription, reading patterns), ad targeting and content recommendations
- User information is mostly limited in scope
  - Cookies have very varying lifetime & users have many cookies
  - Users use various accounts or do not log in
- However, also partial information of user behavior gives insights
- Clusters of users emerged -> develop services for these segments
- Usage pattern is stabile from day to day and week to week
- Subscriber and demographic user data would improve the services
- Quality of collaborative recommendations depend on user behavior



VTT

TECHNOLOGY FOR BUSINESS