

Analyzing Sequential User Behavior on the Web

Tutorial @WWW2016

About Us



Philipp



Florian



Tutorial Website and Material

- Website:

sequenceanalysis.github.io

- Slides (to be uploaded)
- Jupyter notebooks:
 - Download and run/edit on your own computer
 - View the result on nbviewer
 - Virtual environment on mybinder

Structure of this Tutorial

- Introduction & Overview
- Sequential Pattern Mining



- Break -

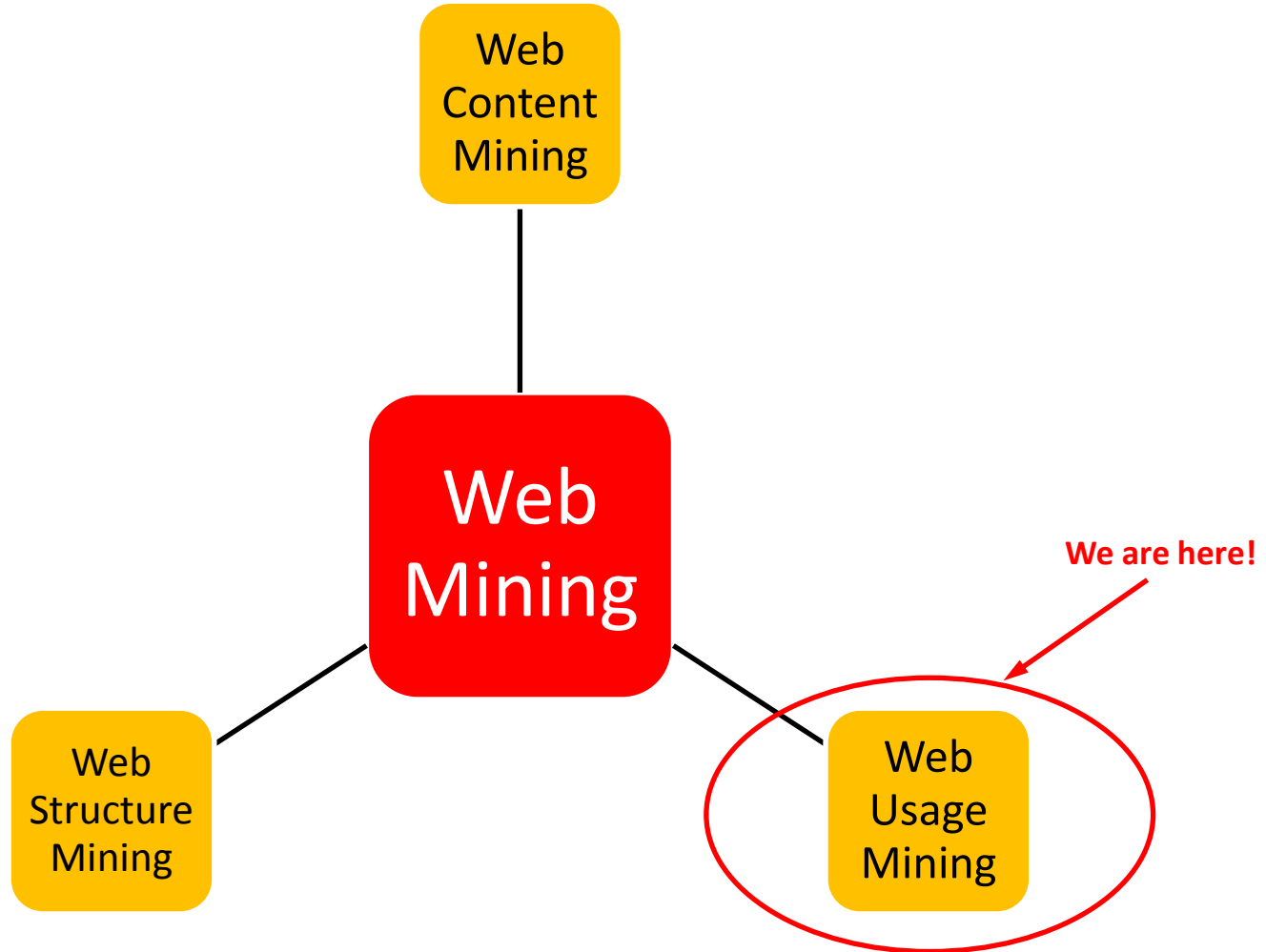
- Markov Chain Modeling
- Comparison of Hypotheses on Sequences



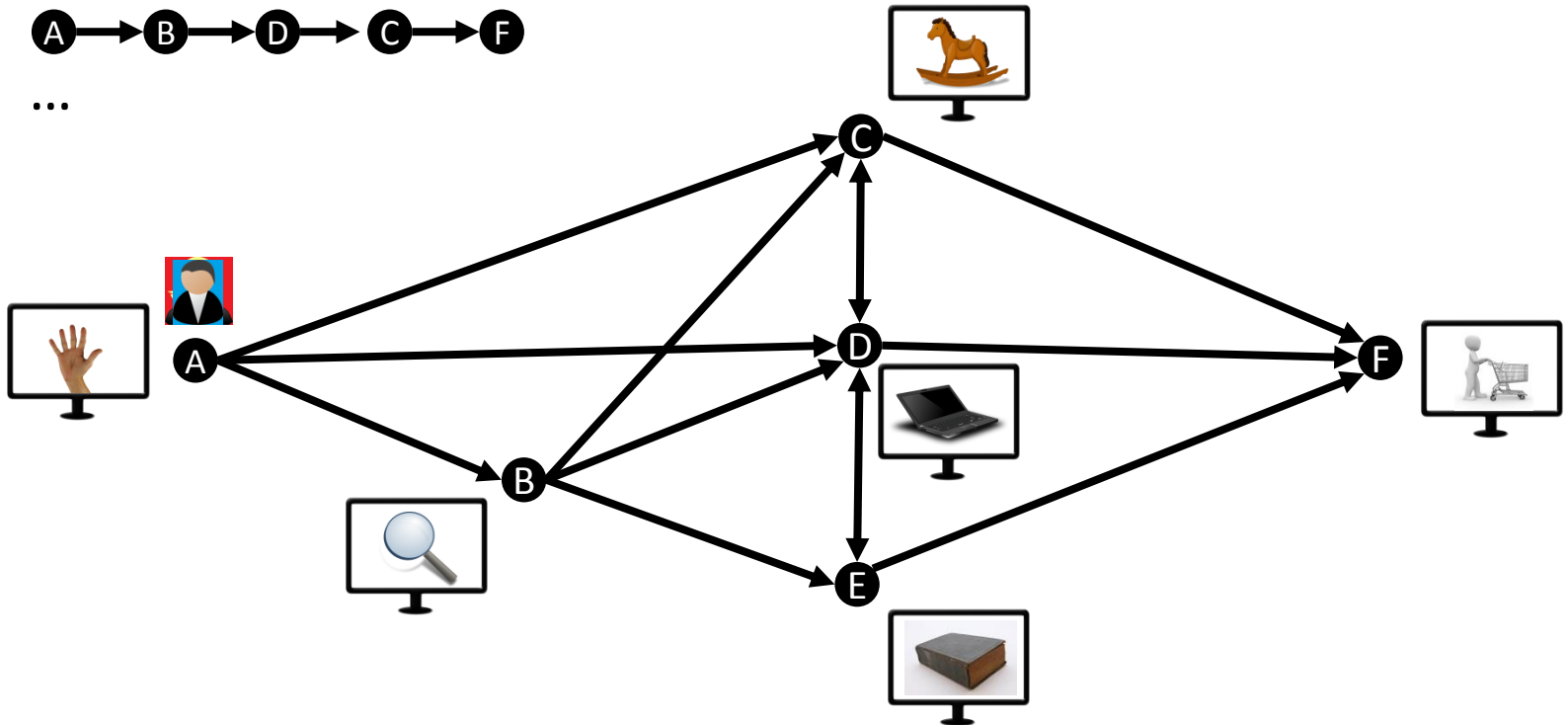
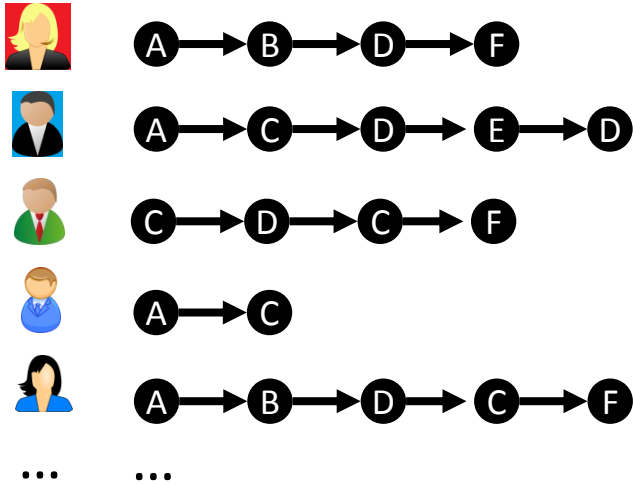
Part 1

A Short Introduction to Categorical Sequences on the Web

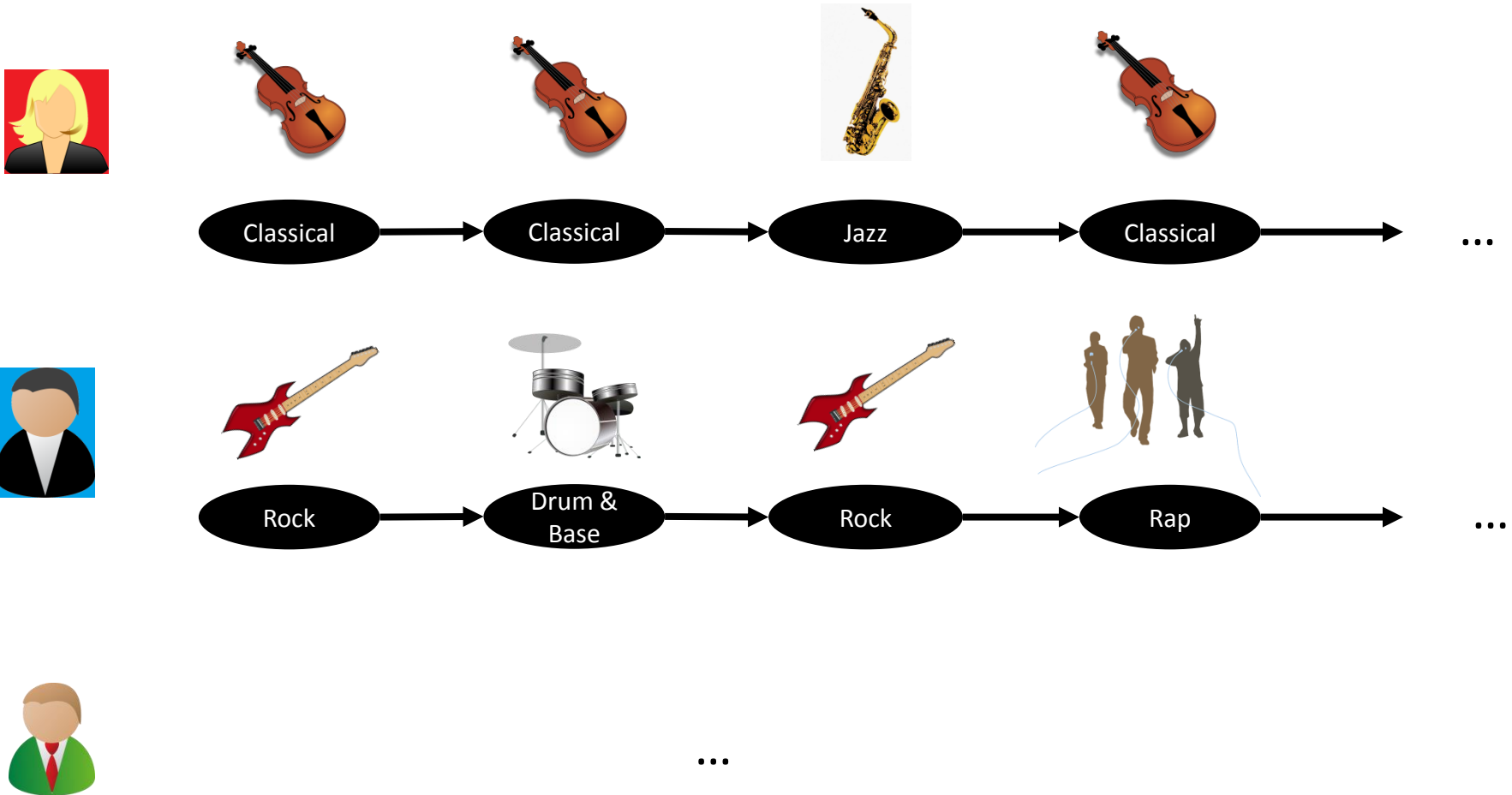
Web Mining [Srivastava 2000]



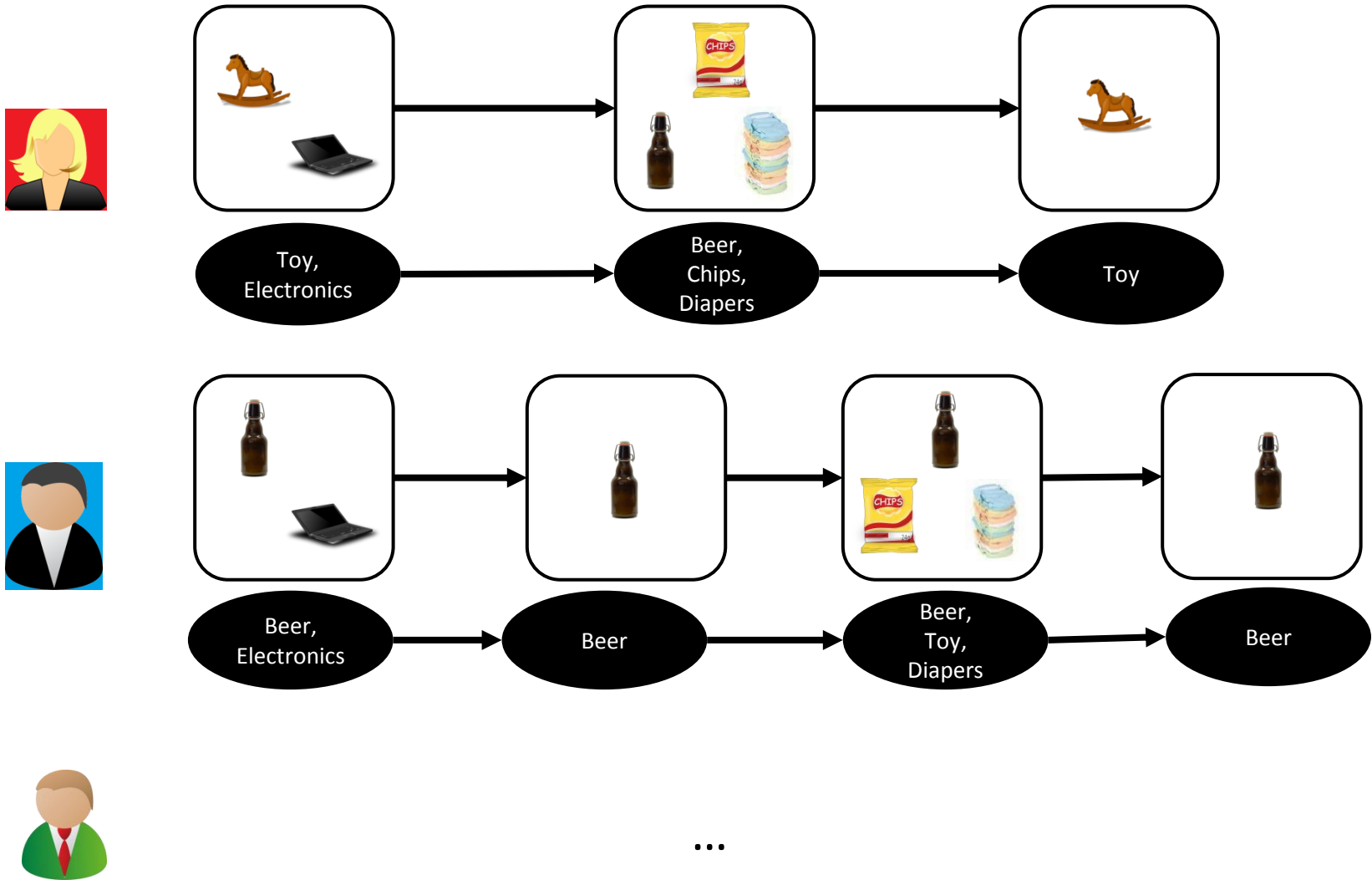
Example: Navigation through the Web



Example II: Listening History

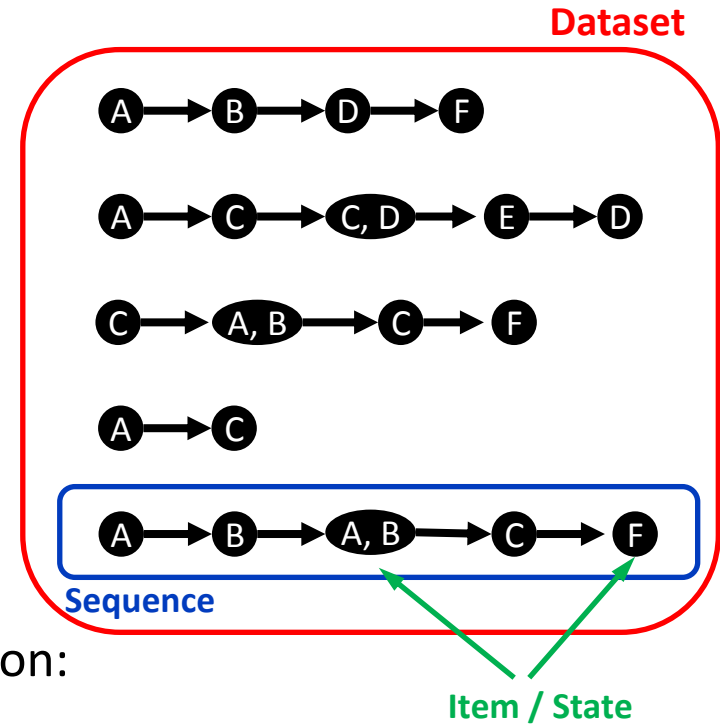


Example III: Shopping History



Data Covered in this Tutorial

- Dataset is given by a set of sequences
- Each sequence contains several events X
- Each event in a sequence has...
 - Exactly one categorical variable (state) (Modeling, Hypotheses Comparison)
 - Multiple Binary variables (items) (Sequential Pattern Mining)
- We do not cover methods using more information:
 - Numeric/ordinal variables each event
 - No time stamps (only ordering)
 - == *NO time series analysis*
 - Text



Data Sources

- Web Server Logs (e.g., Apache logs)

```
186.62.86.163 -- [31/Mar/2016:21:49:45 -0700] "GET /component2/assets/skins/skin1/right.png HTTP/1.1" 200 2967 "http://www.feraval.com/" "Mozilla/5.0 (Windows NT 6.1; rv:15.0) Gecko/20100101 Firefox/15.0"
186.62.86.163 -- [31/Mar/2016:21:49:45 -0700] "GET /component2/assets/transitions/3D.xml?uid=1459486267468 HTTP/1.1" 200 900 "http://www.feraval.com/" "Mozilla/5.0 (Windows NT 6.1; rv:15.0) Gecko/20100101 Firefox/15.0"
186.62.86.163 -- [31/Mar/2016:21:49:48 -0700] "GET /component2/assets/transitions/favorites-text.xml?uid=1459486267809 HTTP/1.1" 200 2941 "http://www.feraval.com/" "Mozilla/5.0 (Windows NT 6.1; rv:15.0) Gecko/20100101 Firefox/15.0"
31.184.238.174 -- [31/Mar/2016:21:54:41 -0700] "GET /logs/access.log HTTP/1.1" 200 32753 "http://gravatar.com/orderpyridostigmineonlinenoprescription" "Mozilla/5.0 (Windows NT 6.1; rv:15.0) Gecko/20100101 Firefox/15.0"
31.184.238.174 -- [31/Mar/2016:22:04:26 -0700] "GET /logs/access.log HTTP/1.1" 200 32991 "http://gravatar.com/buypiroxicamwithoutrx" "Mozilla/5.0 (Windows NT 6.0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.118 Safari/537.36"
77.247.22.51 -- [31/Mar/2016:22:11:25 -0700] "GET /logs/access_141104.log HTTP/1.1" 200 107309 "http://all4webs.com/edjhife/hkjlh.htm" "Mozilla/5.0 (Windows NT 6.1; rv:15.0) Gecko/20100101 Firefox/15.0"
```



User IP

Date / Time

Requested Page

Referrer

Browser / OS

- Cookies
- Explicit user input
- Client-side tracking (modified browsers, eye-tracking)
- Web APIs (e.g.,  reddit or  Wikipedia) or scraping:
 - Maybe not capture user actions directly
 - Results/edits form sequences

Data Pre-processing of Web Logs [Chitraa et al. 2010]

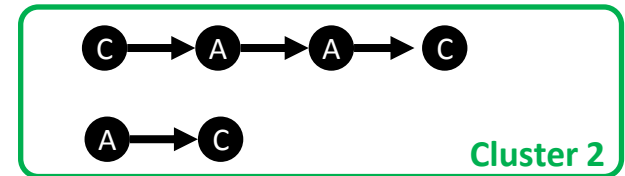
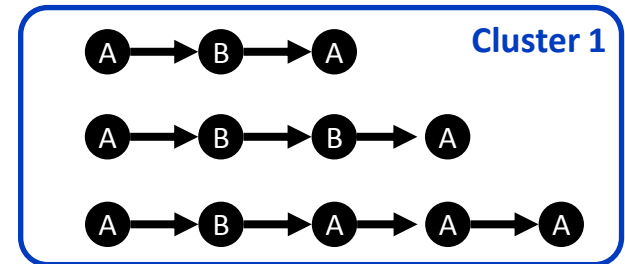
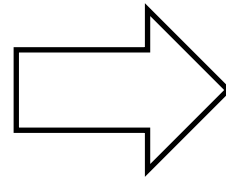
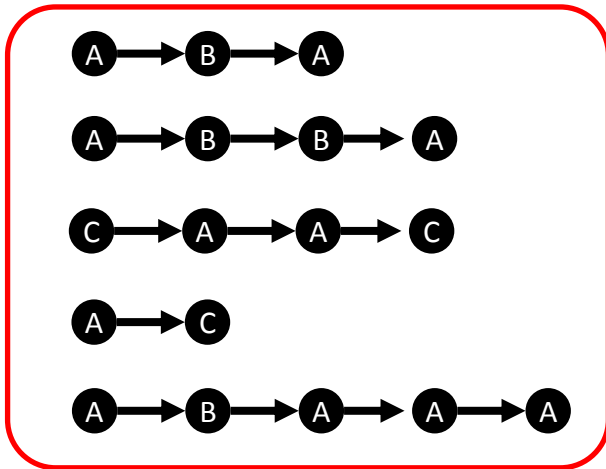
- Data Cleaning, e.g.
 - Remove access to single images
 - Erroneous requests (http errors)
- User identification (usually based on IP address)
- Session identification
 - Time-oriented heuristics
 - Navigation-oriented heuristics
- Path completion: accounts for proxy / caching effects

Tasks for Sequential Data

- Sequence Clustering
- Sequence Classification
- Sequence Prediction
- Sequence Labeling
- Sequence Segmentation
- Sequential Pattern Mining
- Sequence Modeling
- Hypotheses Comparison on Sequences

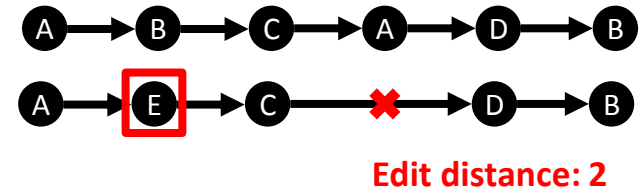
Sequence Clustering: Task

“Find groups in the sequence dataset such that sequences within one group are similar and sequences in different groups are dissimilar”



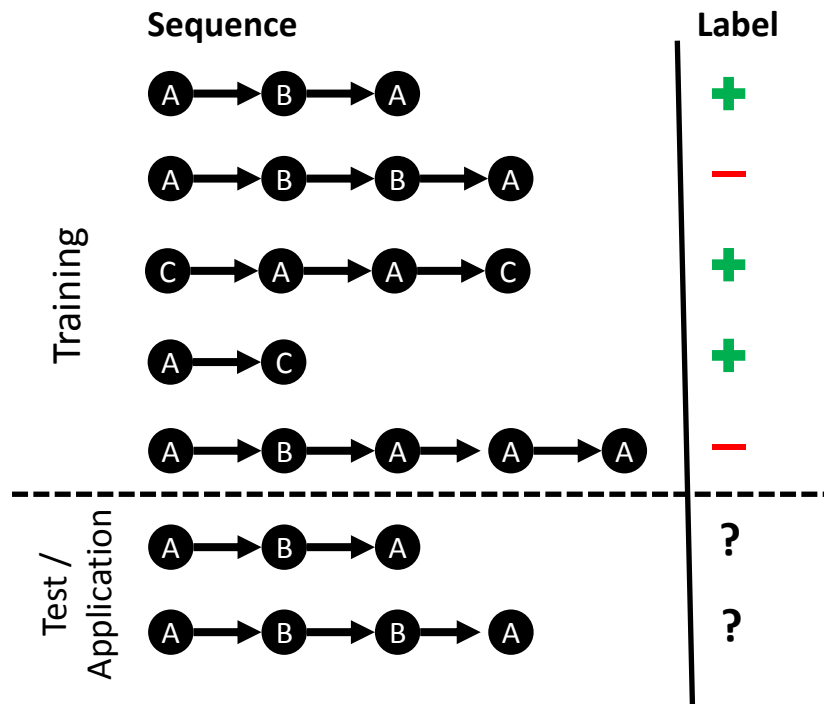
Sequence Clustering: Method Overview [Xu & Wunsch, 2005]

- Clustering based on sequence similarity
 - E.g., edit distance (Levenshtein distance):
Number of transformation operations
 - Can apply hierarchical clustering, density-based clustering, ...
- Indirect clustering: Extract features first
 - Features: all n-grams, sequential patterns
 - Use (classical) vector-spaces clustering on these features
- Statistical sequence clustering / model based clustering
 - Use set of Hidden Markov Models (HMM)
 - Each model “generates” the sequences of one cluster
 - EM algorithm optimizes clusters and sequence-cluster mapping



Sequence Classification: Task

“Given a training dataset of labeled sequences, predict the labels of future sequences”

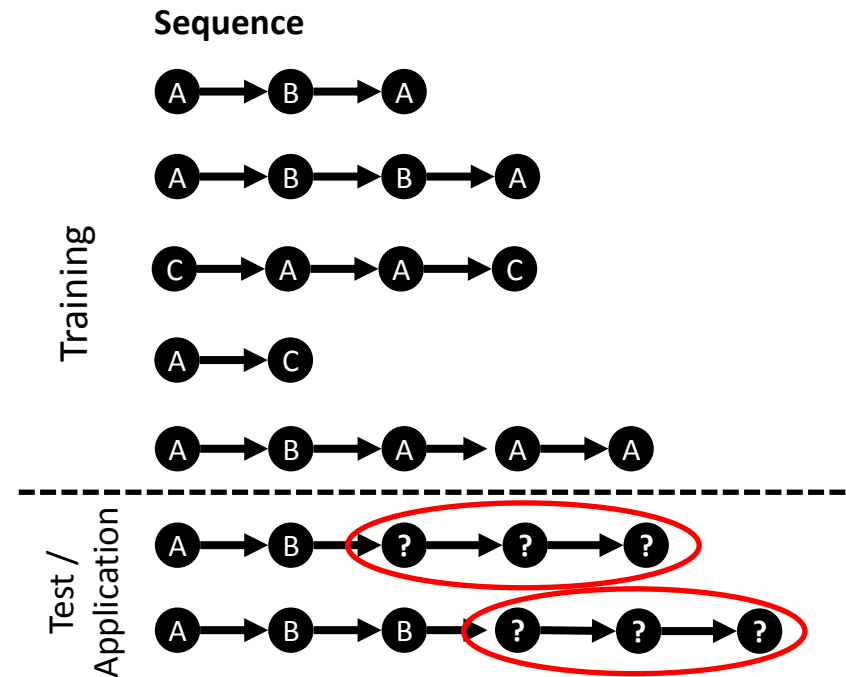
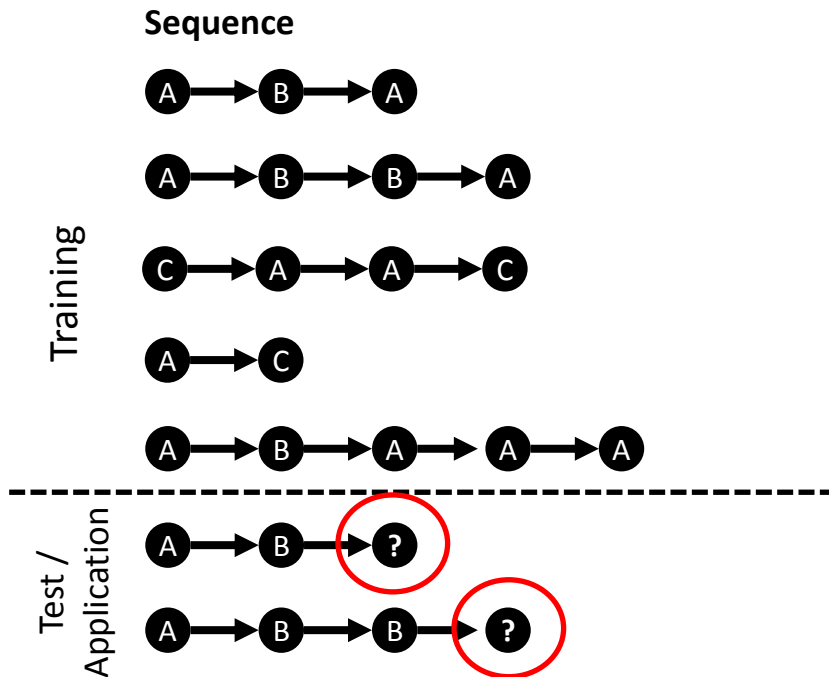


Sequence Classification: Methods [Xing et al. 2010]

- Use sequence similarity measure
 - See sequence clustering
 - Apply k-nearest-neighbor for classification
- Indirect classification: extract features first
 - See sequence clustering
 - Apply any classification method
 - SVM with string kernels:
do not compute the features explicitly, but only use a kernel instead
- Model-based classification
 - Discriminatively trained Markov Models
 - Different variations of Hidden Markov Models

Sequence Prediction / Sequence Generation: Task

“Given a set of sequences and some incomplete sequences, how will the new sequences continue?”

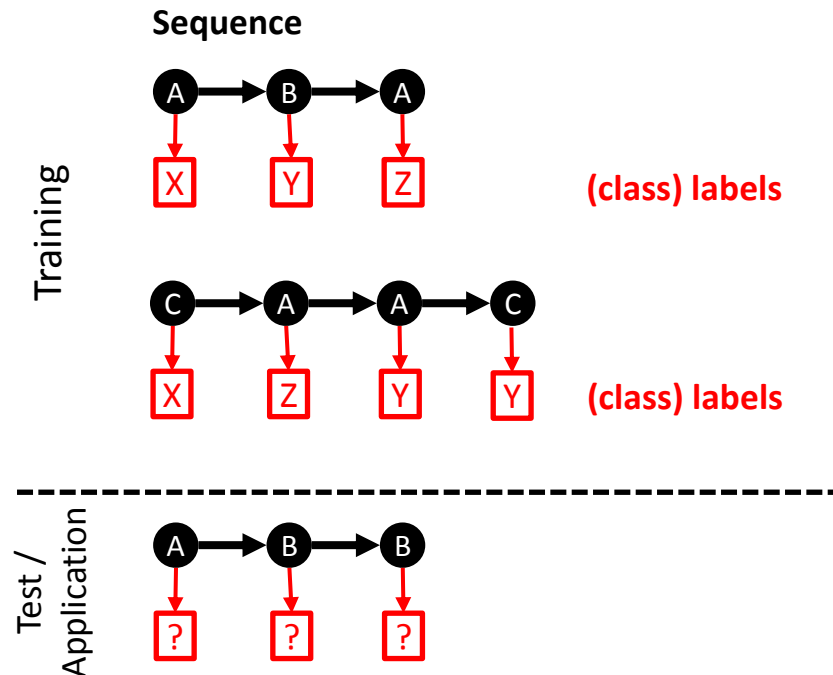


Sequence Prediction: Methods

- Apply (Hidden) Markov Models
- (Partially ordered) Sequential rules (based on sequential patterns)
- Recurrent Neural Networks (RNNs)

Sequence Labeling: Task

“Given a set of sequences with labels for each event, predict the labels of new (unlabeled) events”



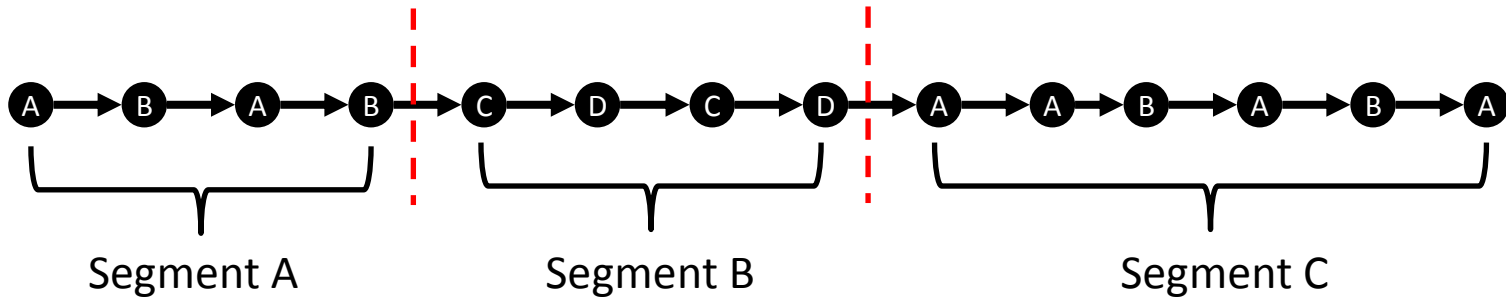
Sequence Labeling [Nguyen & Guo 2007]

- More typical for Natural Language Processing
E.g., part of speech tagger, reference extraction, ...

- Methods:
 - Hidden Markov Models [Rabiner 1989]
 - Conditional Random Fields [Laferty et al. 2001]
 - SVM-Struct [Tsochantaridis et al. 2005]
 - ...

Sequence Segmentation

“Partition a sequence into segments such that the segments are as homogeneous as possible”



Sequence Segmentation: [Terzi & Tsaparas 2006]

- Applications:
 - Detect behavioral stages of web users
 - DNA segmentation
 - Text segmentation

- Methods:
 - Given time information: similar to discretization
 - Models + MDL [Kiernan & Terzi 2009]
 - Set of models, optimizes (log-) likelihood [Yang et al. 2014]

Tasks for Sequential Data

- Sequence Clustering
- Sequence Classification
- Sequence Prediction
- Sequence Labeling
- Sequence Segmentation
- **Sequential Pattern Mining**
- **Sequence Modeling**
- **Hypotheses Comparison on Sequences**

Human Navigation

- User Navigation from Web logs [Catledge & Pitkow 1995]
- Strong regularities in WWW surfing [Huberman et al. 1998]
- Mining longest repeating subsequences for prediction [Pitkow & Pirolli 1999]
- Information scent theory [Chi et al. 2001]

- Navigation in Wikipedia
 - Human wayfinding in information networks [West & Leskovec 2012]
 - Automatic vs. Human Navigation [West & Leskovec 2012-2, Trattner et al. 2012]
 - Memory and structure [Singer et al 2014]

Detecting a-typical Surfing Behavior

- Characterizing (a-)typical user behavior [Sadagopan & Li 2008]
 - Model sequences with Markov chains
 - Detect improbable sequences
 - Characterize outliers manually
- Sybil (Fake identity) [Wang et al 2013]
 - Visualize transition probabilities in Markov chains
 - Use SVM/similarity based approaches for classification

Further Application Areas [Facca & Lanzi 2005]

- Improved website design
- Personalization of web content [Pehtaa et al 2012, Andersson 2002, Eiriniki et al 2003]
 - Recommending links
 - Personalized site maps
- Pre-fetching and caching [Patil & Patil 2015, Wu & Chen 2002]
- E-commerce / customer relation ship management [Bounsaythip & Rinta-Russala 2001, Ansari et al. 2001,]
- Identifying relevant websites [Bilenko & White 2008]
- ...

Privacy: Ethical and Legal issues

- Ethical issues:
 - Web Usage Mining exploits user data, often no (conscious) agreement
 - User are judged based on group characteristics instead of individual merit

[Van Wel & Royakkers 2004]

- Legal issues: [Velazquez 2013]
 - Depends on the country
 - *“contracts are the main legal tool to protect users’ privacy, therefore affirming **informed consent** as being the key concept in deploying a suitable privacy policy.”*
 - Key question:
Is the IP address personal data (personally identifiable information):
Example Germany: “potentially personal data“
 - Also different regulations for academic and commercial use

References 1/2

- Anderson, C. R. (2002). *A machine learning approach to web personalization* (Doctoral dissertation,
- Ansari, S., Kohavi, R., Mason, L., & Zheng, Z. (2001). Integrating e-commerce and data mining: Architecture and challenges. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on* (pp. 27-34). IEEE.
- Bilenko, M., & White, R. W. (2008). Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *Proceedings of the 17th international conference on World Wide Web* (pp. 51-60). ACM.
- Bounsaythip, C., & Rinta-Runsala, E. (2001). Overview of data mining for customer behavior modeling. *VTT information Technology*, 18, 1-53.
- Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery*, 7(4), 399-424.
- Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN systems*, 27(6), 1065-1073.
- Chierichetti, F., Kumar, R., Raghavan, P., & Sarlos, T. (2012). Are web users really markovian?. In *Proceedings of the 21st international conference on World Wide Web* (pp. 609-618). ACM.
- Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001, March). Using information scent to model user information needs and actions and the Web. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 490-497). ACM.
- Chitraa, V., Davamani, D., & Selvdoss, A. (2010). A survey on preprocessing methods for web usage data. *arXiv preprint arXiv:1004.1257*.
- Deshpande, M., & Karypis, G. (2004). Selective Markov models for predicting Web page accesses. *ACM Transactions on Internet Technology (TOIT)*, 4(2), 163-184.
- Eirinaki, M., Vazirgiannis, M., & Varlamis, I. (2003). Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process. In *Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'03), Washington DC*.
- Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering*, 53(3), 225-241.
- Mehtaa, P., Parekh, B., Modi, K., & Solanki, P. (2012). Web personalization using web mining: concept and research issue. *International Journal of Information and Education Technology*, 2(5), 510.
- Patil, N. V., & Patil, H. D. (2015). Prediction of Web Users Browsing Behavior: A Review.
- Perkowski, M., & Etzioni, O. (2000). Towards adaptive web sites: Conceptual framework and case study. *Artificial intelligence*, 118(1), 245-275.
- Pirolli, P. L., & Pitkow, J. E. (1999). Distributions of surfers' paths through the World Wide Web: Empirical characterizations. *World Wide Web*, 2(1-2), 29-45.
- Pitkow, J., & Pirolli, P. (1999). Mining longest repeating subsequences to predict world wide web surfing. In *Proc. USENIX Symp. On Internet Technologies and Systems* (p. 1).

Icons in this slide set are CC0 Public Domain, taken from pixabay.com

References 2/2

- Sadagopan, N., & Li, J. (2008, April). Characterizing typical and atypical user sessions in clickstreams. In *Proceedings of the 17th international conference on World Wide Web* (pp. 885-894). ACM.
- Singer, P., Helic, D., Taraghi, B., & Strohmaier, M. (2014). Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PloS one*, *9*(7), e102070.
- Trattner, C., Singer, P., Helic, D., & Strohmaier, M. (2012, September). Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* (p. 14). ACM.
- Van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, *6*(2), 129-140.
- Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., & Zhao, B. Y. (2013). You are how you click: Clickstream analysis for sybil detection. In *Proc. USENIX Security* (pp. 1-15).
- West, R., & Leskovec, J. (2012, May). Automatic Versus Human Navigation in Information Networks. In *ICWSM*. An, J., Quercia, D., & Crowcroft, J. (2014, October). Partisan sharing: facebook evidence and societal consequences. In *Proceedings of the second ACM conference on Online social networks* (pp. 13-24). ACM.
- West, R., & Leskovec, J. (2012, April). Human wayfinding in information networks. In *Proceedings of the 21st international conference on World Wide Web* (pp. 619-628). ACM.
- Wu, Y. H., & Chen, A. L. (2002). Prediction of web page accesses by proxy server log. *World Wide Web*, *5*(1), 67-88.
- Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, *12*(1), 40-48.
- Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, *16*(3), 645-678.
- Huberman, B. A., Pirolli, P. L., Pitkow, J. E., & Lukose, R. M. (1998). Strong regularities in world wide web surfing. *Science*, *280*(5360), 95-97.
- Yang, J., McAuley, J., Leskovec, J., LePendu, P., & Shah, N. (2014, April). Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd international conference on World wide web* (pp. 783-794). ACM.