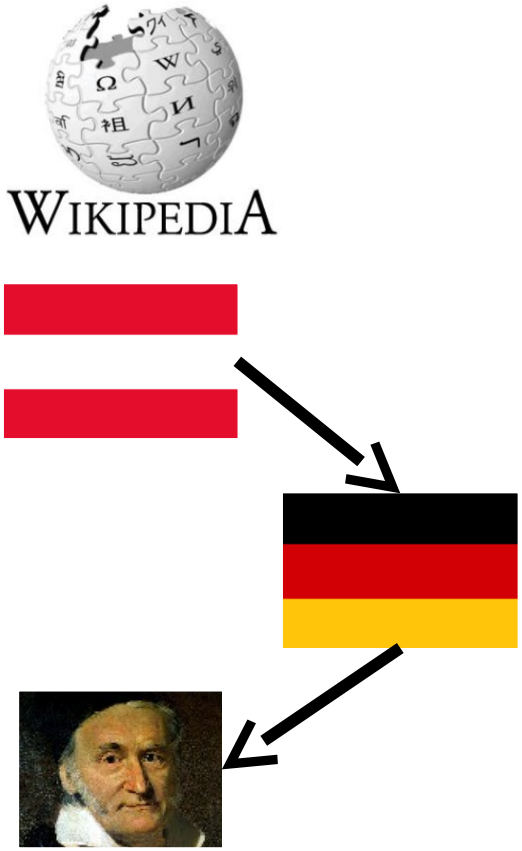
Part 4

Comparing Hypotheses about Sequential Data

Example: Human Navigation

- Humans prefer to navigate…
  - H1: over semantically similar websites
  - H2: via self-loops (e.g., refreshing)
  - H3: by using the structural link network
  - H4: by preferring similar categories
  - H5: by utilizing structural properties
  - H6: by information scent



[West et al. IJCAI 2009], [Singer et al. IJSWIS 2013],
[West & Leskovec WWW 2012], [Chi et al. CHI 2001]

Example: Human Navigation

- Humans prefer to navigate…
  - H1: over semantically similar websites
  - H2: via self-loops (e.g., refreshing)
  - H3: by using the structural link network
  - H4: by preferring similar categories
  - H5: by utilizing structural properties
  - H6: by information scent

What is the relative plausibility of these hypotheses given data?

[West et al. IJCAI 2009], [Singer et al. IJSWIS 2013],
[West & Leskovec WWW 2012], [Chi et al. CHI 2001]

Example: Human Navigation

- Humans prefer to navigate…
  - H1: over semantically similar websites
  - H2: via self-loops (e.g., refreshing)
  - H3: by using the structural link network
  - H4: by preferring similar categories
  - H5: by utilizing structural properties
  - H6: by information scent

HypTrails
[Singer et al. WWW 2015]

[West et al. IJCAI 2009], [Singer et al. IJSWIS 2013],
[West & Leskovec WWW 2012], [Chi et al. CHI 2001]

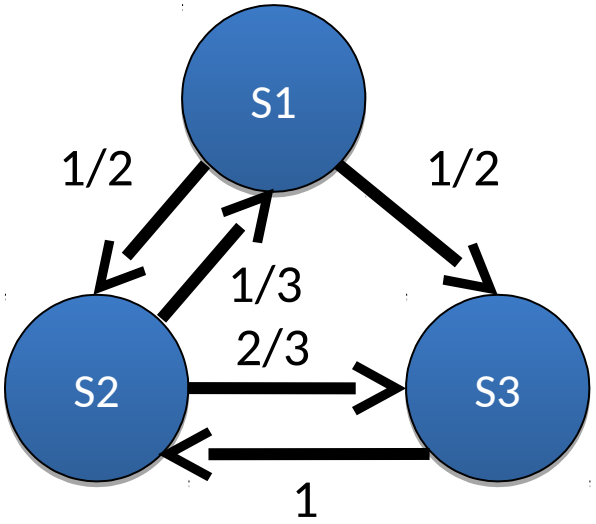
HypTrails in a nutshell

- Goal: Express and compare hypotheses about sequences in a coherent research approach
- Method:
  - First-order Markov chain model
  - Bayesian inference
- Idea:
  - Incorporate hypotheses as priors
  - Utilize sensitivity of marginal likelihood on the prior
- Outcome: Partial ordering of hypotheses

Structure of HypTrails

Structure of HypTrails

MC Model

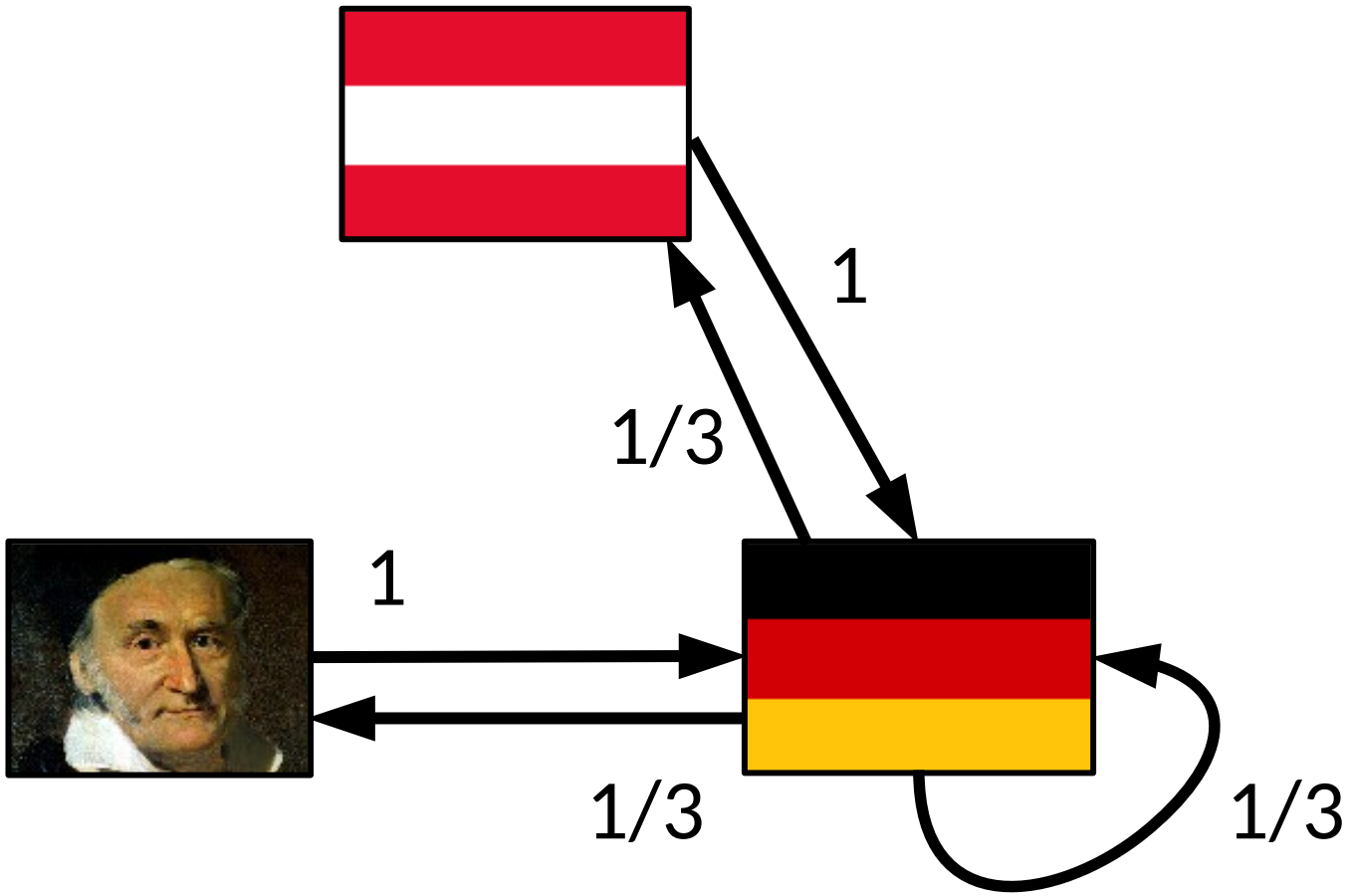


How to express hypotheses?

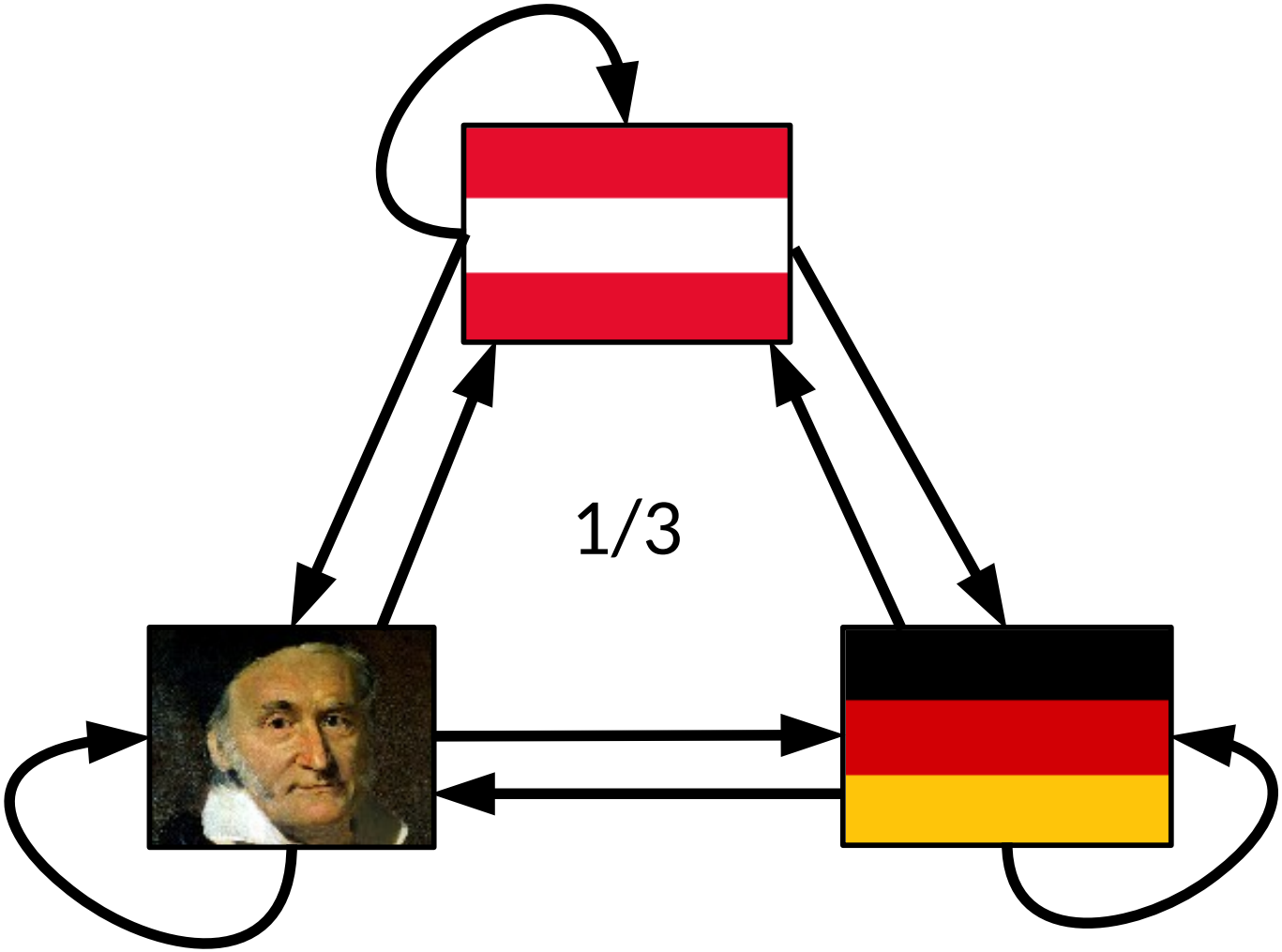
How to express hypotheses?

As assumptions in parameters of Markov Chain model.

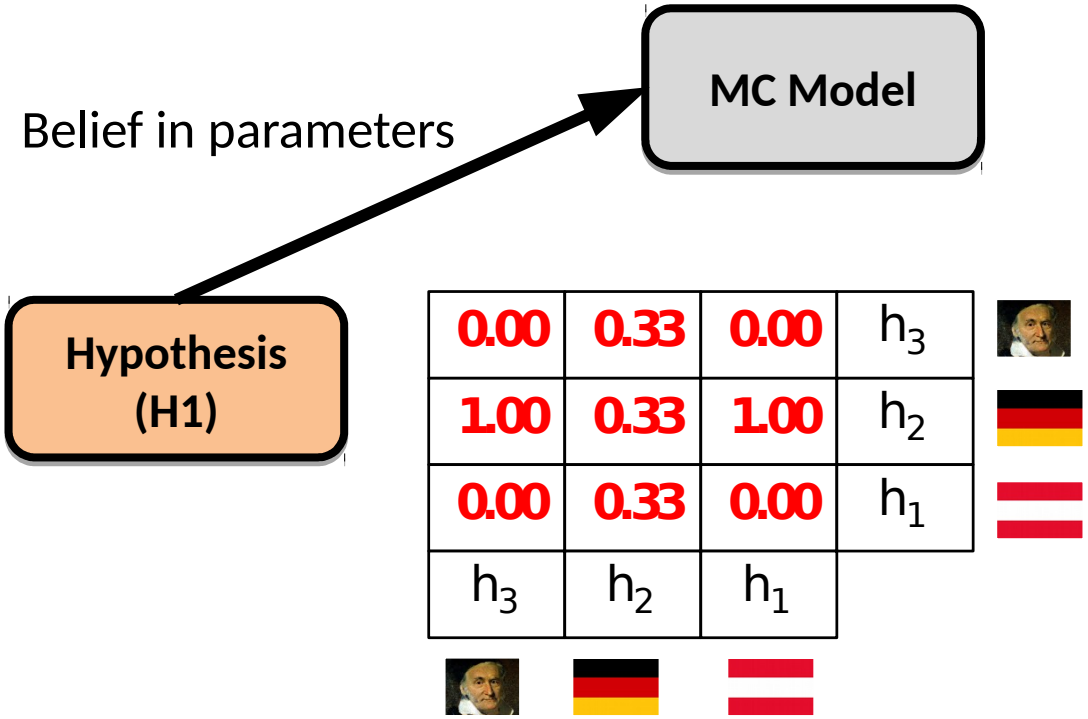
Structural hypothesis



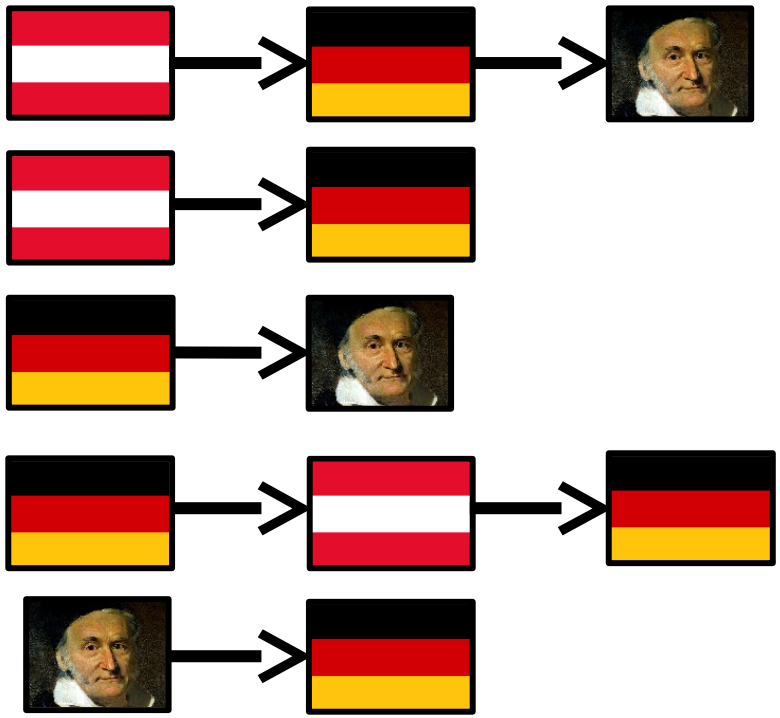
Uniform hypothesis

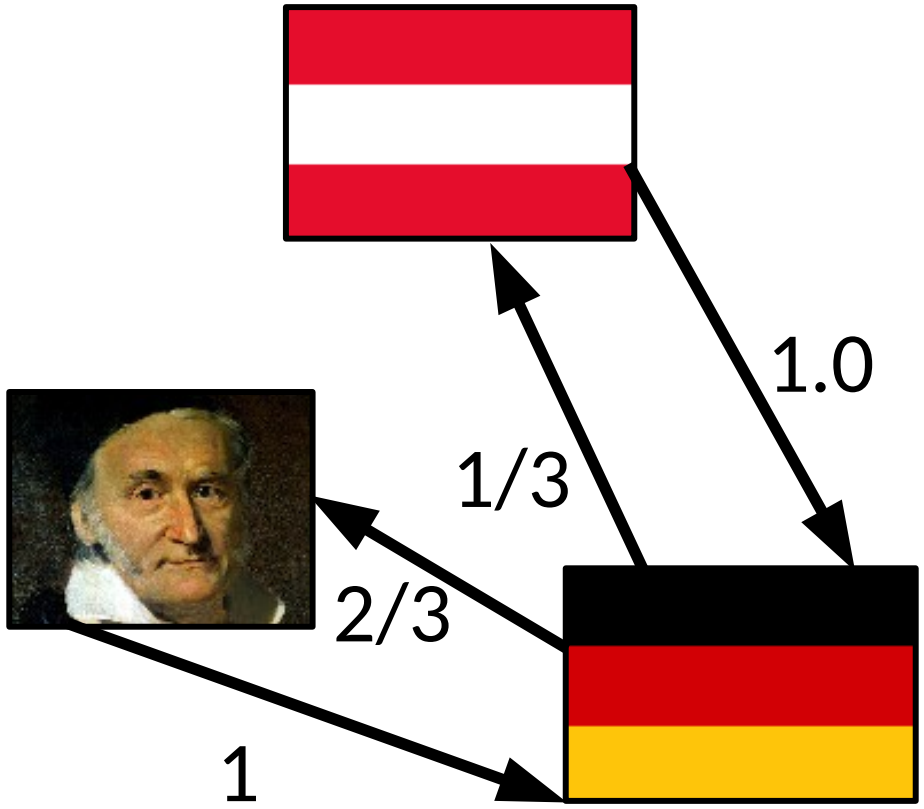


Structure of HypTrails



Empirical observations





Which hypothesis is the most plausible one?

Bayesian model comparison: marginal likelihood

$$\overbrace{P(D|M)}^{\text{marginal likelihood}} = \int \overbrace{P(D|\theta, M)}^{\text{likelihood}} \overbrace{P(\theta|M)}^{\text{prior}} d\theta$$

Bayesian model comparison: marginal likelihood

Probability of parameters before observing data

$$\overbrace{P(D|H)}^{\text{marginal likelihood}} = \int \overbrace{P(D|\theta)}^{\text{likelihood}} \overbrace{P(\theta|H)}^{\text{prior}} d\theta$$

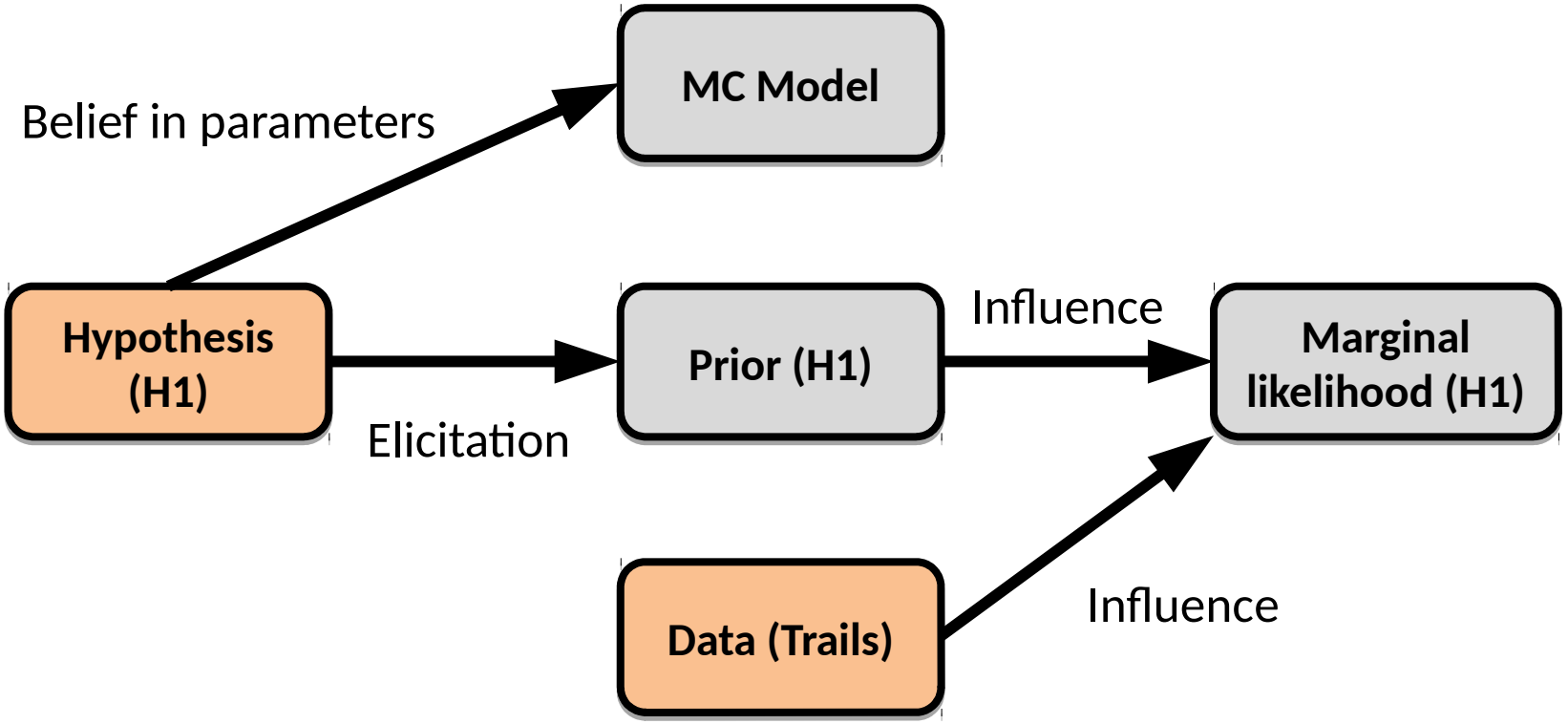
Bayesian model comparison: marginal likelihood

Probability of parameters before observing data

$$\overbrace{P(D|H)}^{\text{marginal likelihood}} = \int \overbrace{P(D|\theta)}^{\text{likelihood}} \overbrace{P(\theta|H)}^{\text{prior}} d\theta$$

Hypothesis

Structure of HypTrails



How to elicit priors from expressed hypotheses?

Conjugate Dirichlet prior

- Hyperparameters: pseudo counts

Conjugate Dirichlet prior

- Hyperparameters: pseudo counts

$$\begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,j} \\ p_{2,1} & p_{2,2} & \dots & p_{2,j} \\ \dots & \dots & \dots & \dots \\ p_{i,1} & p_{i,2} & \dots & p_{i,j} \end{bmatrix} \Rightarrow \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \dots & \alpha_{1,j} \\ \alpha_{2,1} & \alpha_{2,2} & \dots & \alpha_{2,j} \\ \dots & \dots & \dots & \dots \\ \alpha_{i,1} & \alpha_{i,2} & \dots & \alpha_{i,j} \end{bmatrix}$$

Hypothesis parameters

Dirichlet hyperparameters

Elicitation

- Multiply row-normalized hypothesis matrix with concentration parameter k
- Higher k → stronger belief
- Additional proto-prior

$$\begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,j} \\ p_{2,1} & p_{2,2} & \dots & p_{2,j} \\ \dots & \dots & \dots & \dots \\ p_{i,1} & p_{i,2} & \dots & p_{i,j} \end{bmatrix} \cdot k \Rightarrow \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \dots & \alpha_{1,j} \\ \alpha_{2,1} & \alpha_{2,2} & \dots & \alpha_{2,j} \\ \dots & \dots & \dots & \dots \\ \alpha_{i,1} & \alpha_{i,2} & \dots & \alpha_{i,j} \end{bmatrix}$$

Hypothesis parameters

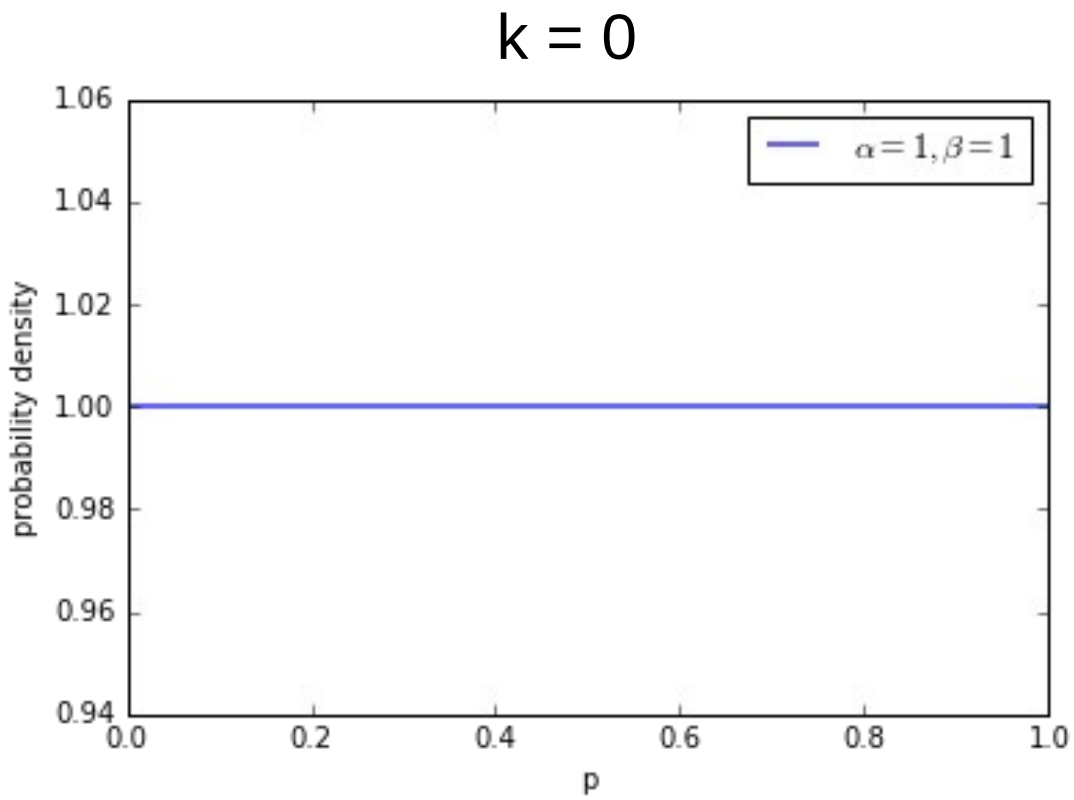
Dirichlet hyperparameters

2 state example: Beta prior

Hypothesis: $[0.7, 0.3]$

2 state example: Beta prior

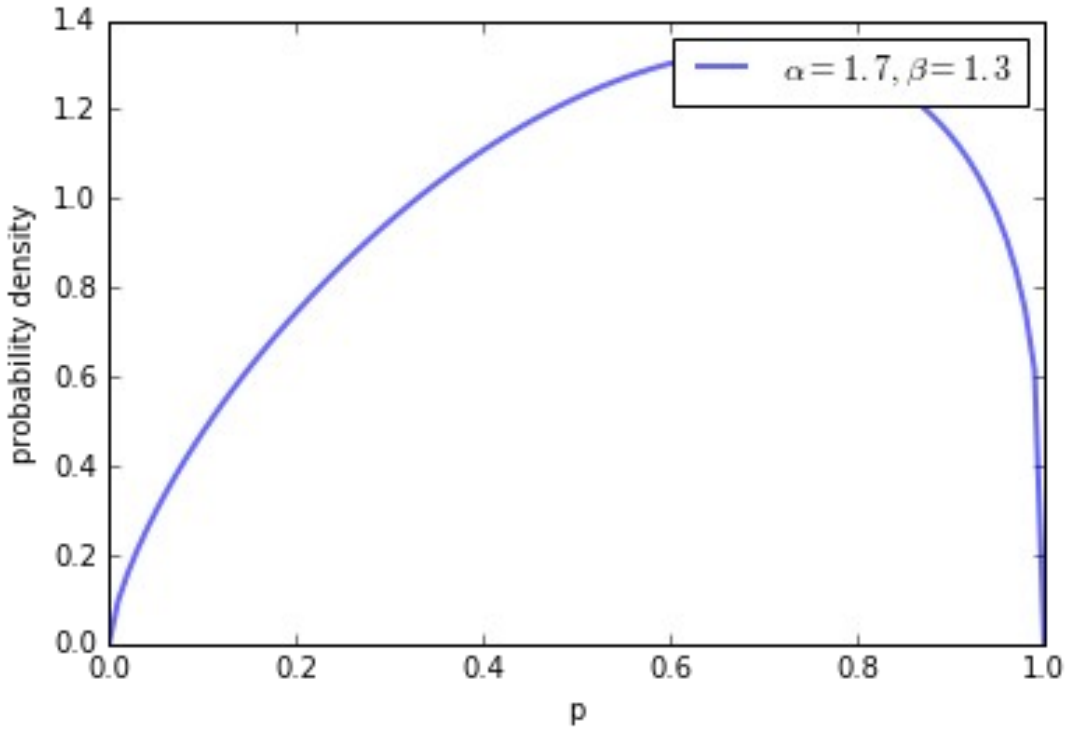
Hypothesis: $[0.7, 0.3]$



2 state example: Beta prior

Hypothesis: $[0.7, 0.3]$

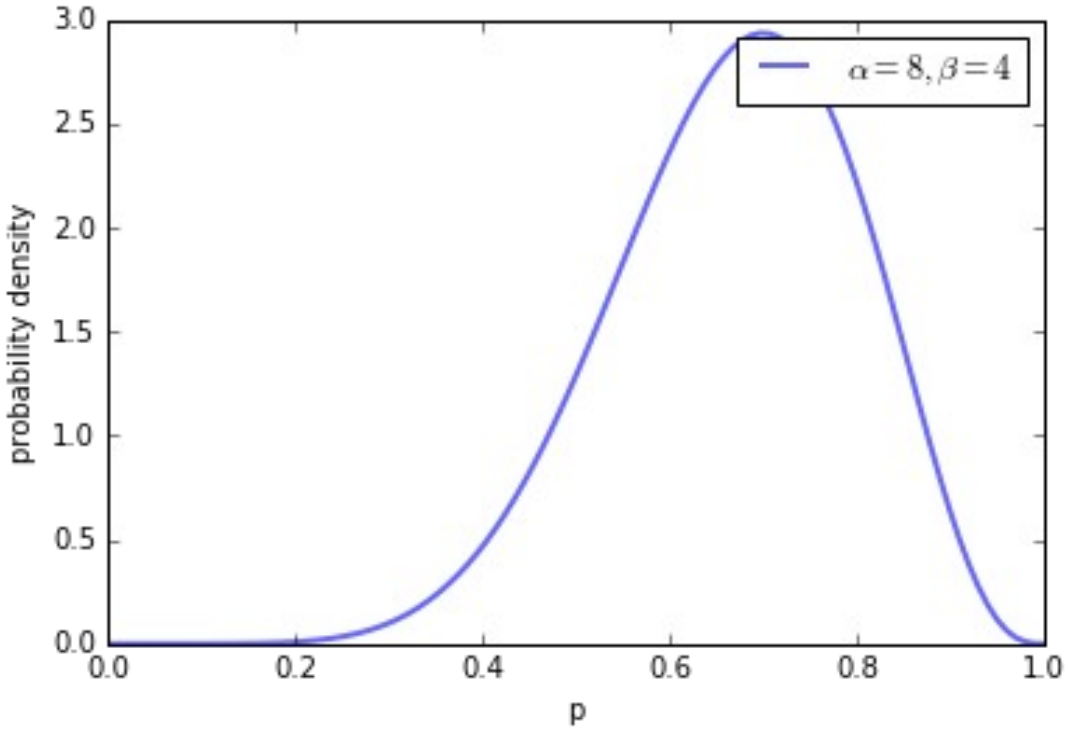
k = 1



2 state example: Beta prior

Hypothesis: $[0.7, 0.3]$

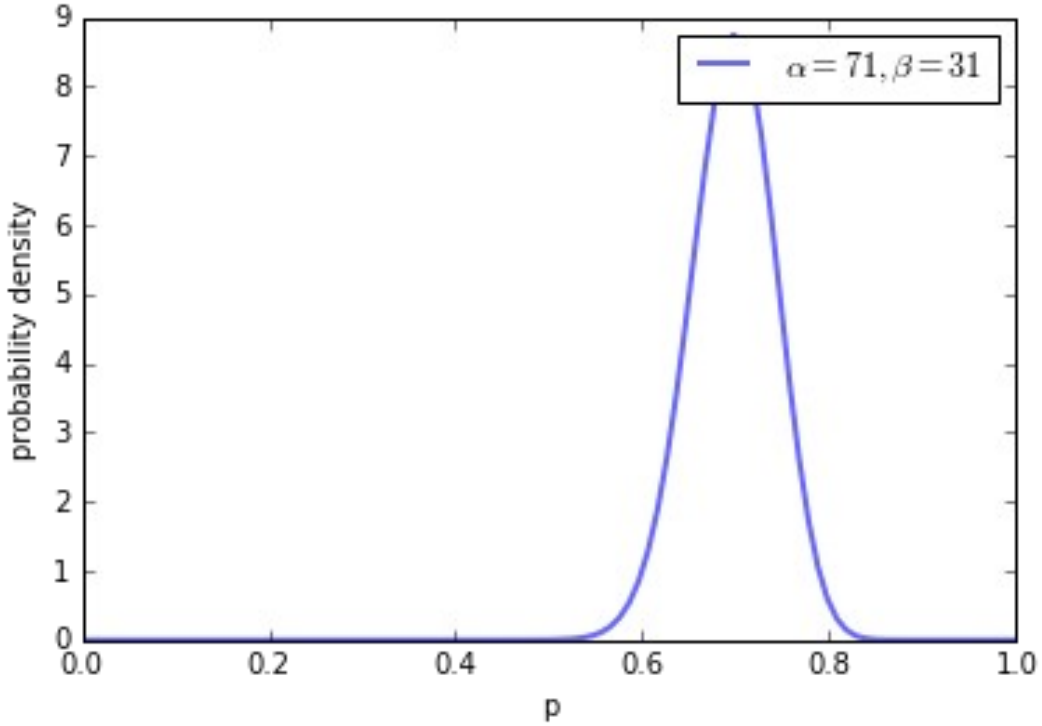
k = 10



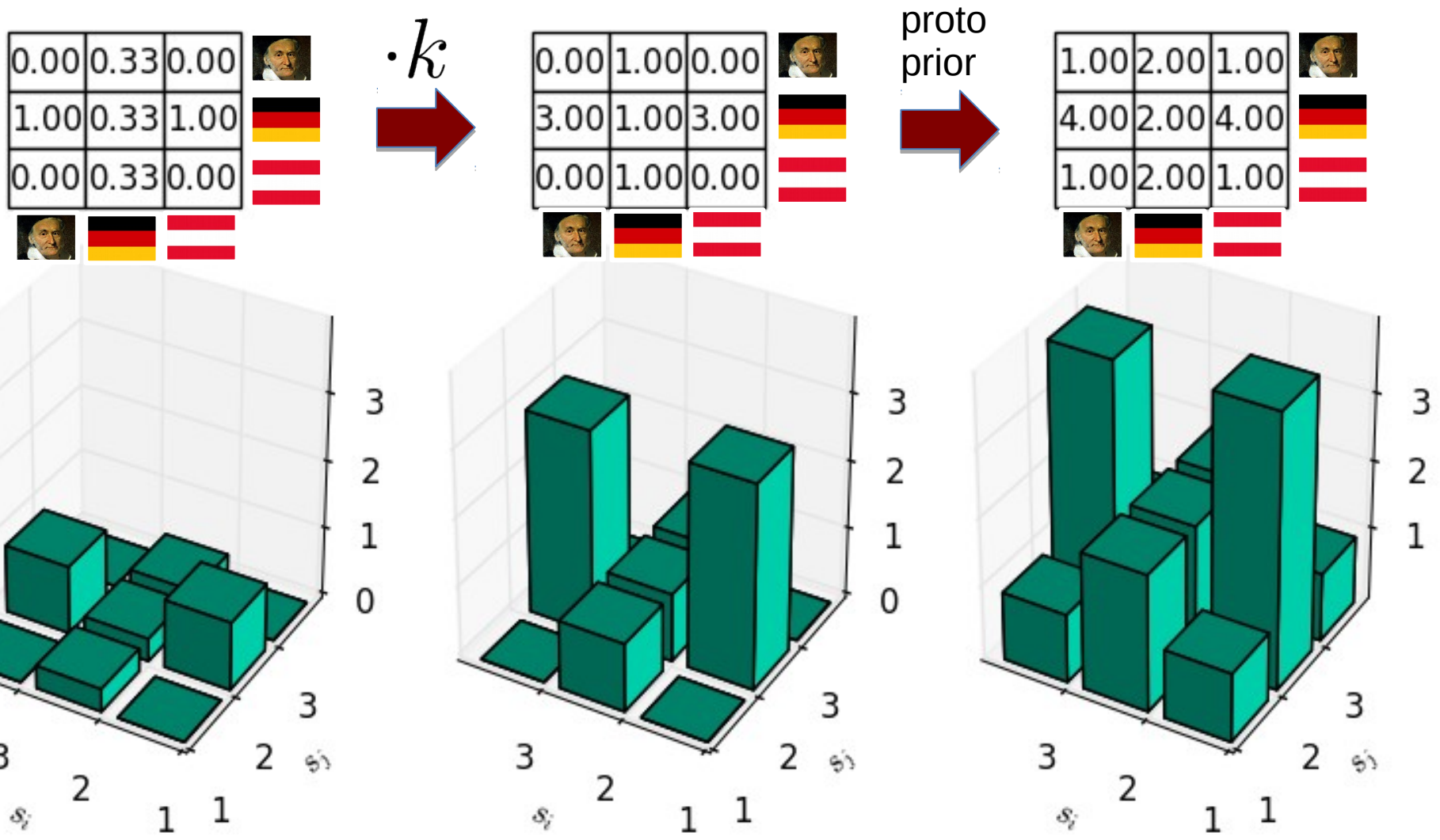
2 state example: Beta prior

Hypothesis: $[0.7, 0.3]$

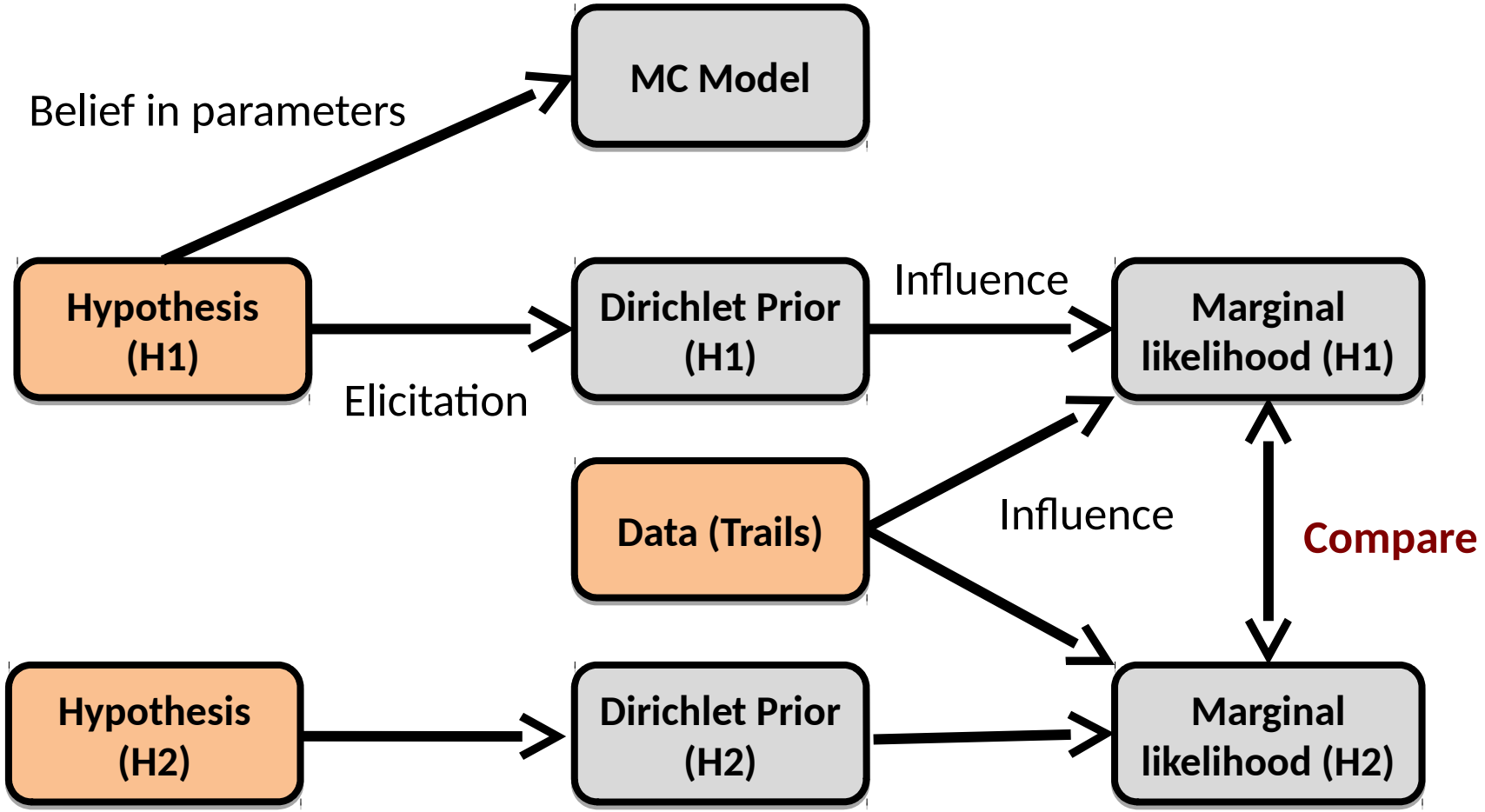
k = 100



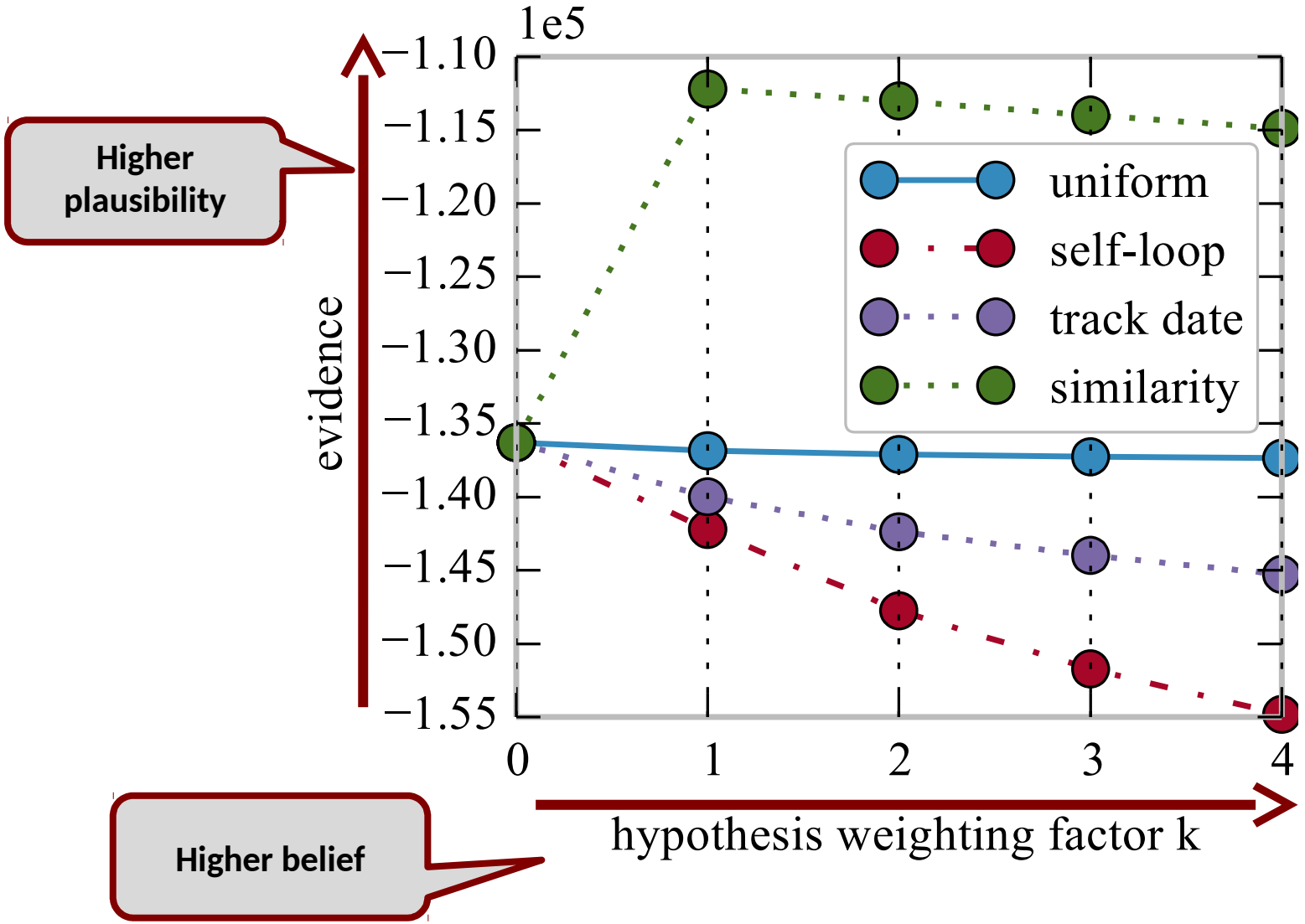
Example: Structural hypothesis



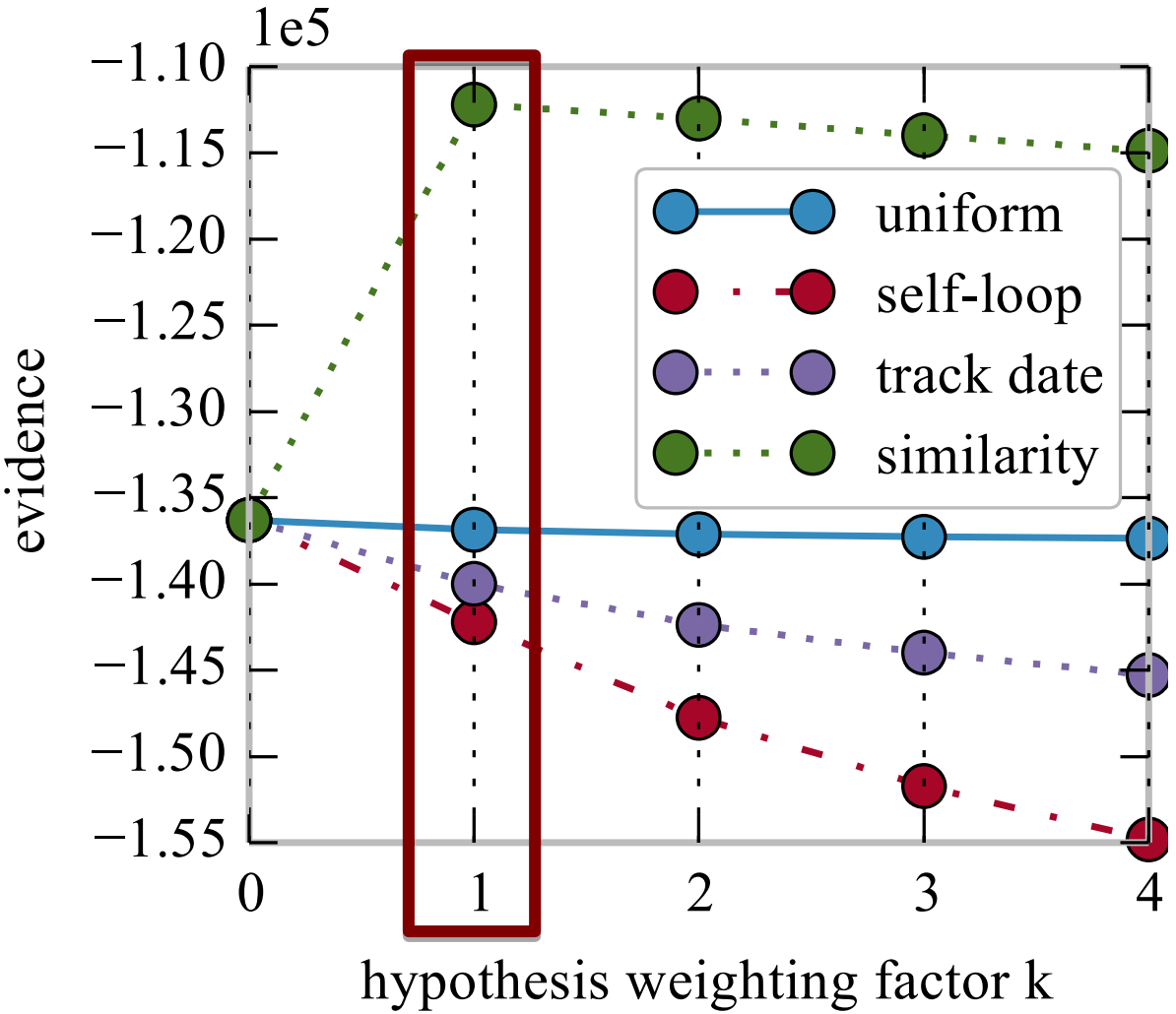
Structure of HypTrails



Example result: Last.fm



Example result: Last.fm



Hands-on jupyter notebook

Further applications

- Ontology engineering – edit sequences
  [Walk et al. ISWC 2015]
- Real-world navigational trails
  - Flickr [Becker et al. SocialCom 2015]
  - Taxi data [Espín-Noboa et al. WWW 2016]
  - Car data [Atzmüller et al. WWW 2016]
- Wikipedia co-editing patterns
  [Samoilenko et al. 2016]

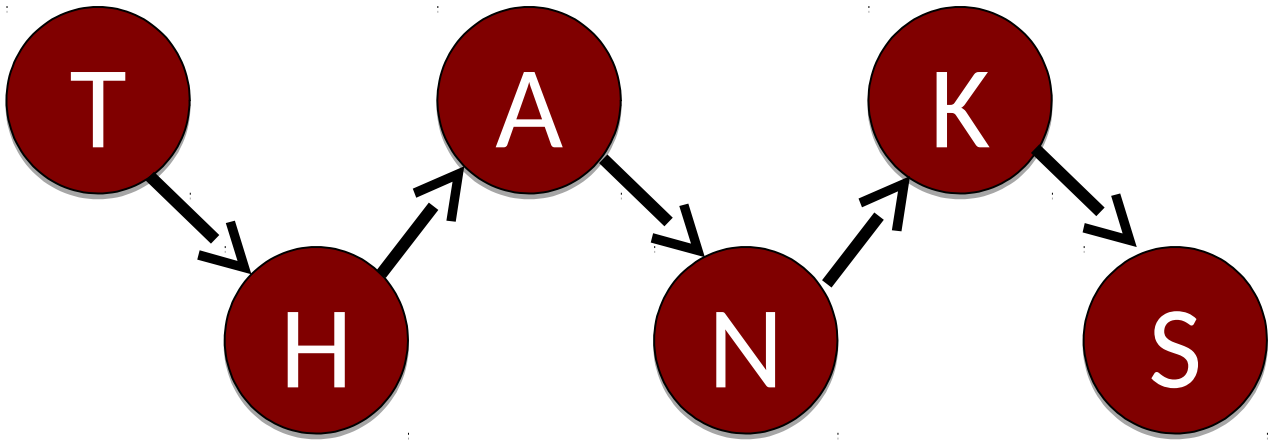
Methodological extensions

- Detect and model heterogeneity in data
- Higher-order Markov chain models
- Adaption for other models

What have we learned?

- Comparing hypotheses about sequential data
- Bayesian approach: HypTrails
- Applications

Questions?



for your attention!

@ph_singer
www.philippsinger.info

florian.lemmerich.net

References 1/2

[West et al. WWW 2015] Robert West, Ashwin Paranjape, and Jure Leskovec: Mining Missing Hyperlinks from Human Navigation Traces: A Case Study of Wikipedia. 24th International World Wide Web Conference (WWW'15), Florence, Italy, 2015.

[Singer et al. IJSWIS 2013] Philipp Singer, Thomas Niebler, Markus Strohmaier and Andreas Hotho, Computing Semantic Relatedness from Human Navigational Paths: A Case Study on Wikipedia, International Journal on Semantic Web and Information Systems (IJSWIS), vol 9(4), 41-70, 2013

[West & Leskovec WWW 2012] Robert West and Jure Leskovec: Human Wayfinding in Information Networks 21st International World Wide Web Conference (WWW'12), pp. 619–628, Lyon, France, 2012.

[Chi et al. CHI 2001] Chi, Ed H., et al. "Using information scent to model user information needs and actions and the Web." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2001.

[Singer et al. WWW 2015] Singer, P., Helic, D., Hotho, A., and Strohmaier, M. (2015, May). Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In Proceedings of the 24th International Conference on World Wide Web (pp. 1003-1013). International World Wide Web Conferences Steering Committee.

[Walk et al. ISWC 2015] Simon Walk, Philipp Singer, Lisette Espín Noboa, Tania Tudorache, Mark A. Musen and Markus Strohmaier, Understanding How Users Edit Ontologies: Comparing Hypotheses About Four Real-World Projects, 14th International Semantic Web Conference, Betlehem, Pennsylvania, USA, 2015

References 2/2

[Becker et al. SocialCom 2015] Martin Becker, Philipp Singer, Florian Lemmerich, Andreas Hotho, Denis Helic and Markus Strohmaier, Photowalking the City: Comparing Hypotheses About Urban Photo Trails on Flickr, 7th International Conference on Social Informatics, Beijing, China, 2015

[Espín-Noboa et al. WWW 2016] Lisette Espín-Noboa, Florian Lemmerich, Philipp Singer and Markus Strohmaier, Discovering and Characterizing Mobility Patterns in Urban Spaces: A Study of Manhattan Taxi Data, 6th International Workshop on Location and the Web at WWW2016, Montreal, Canada, 2016

[Samoilenko et al. 2016] Samoilenko, A., Karimi, F., Edler, D., Kunegis, J., & Strohmaier, M. (2016). Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. EPJ Data Science, 5(1), 1., 2016

[Atzmüller et al. WWW 2016] Atzmueller, M., Schmidt, A., & Kibanov, M. (2016). DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails. In Proceedings of the World Wide Web Conference Companion, 2016