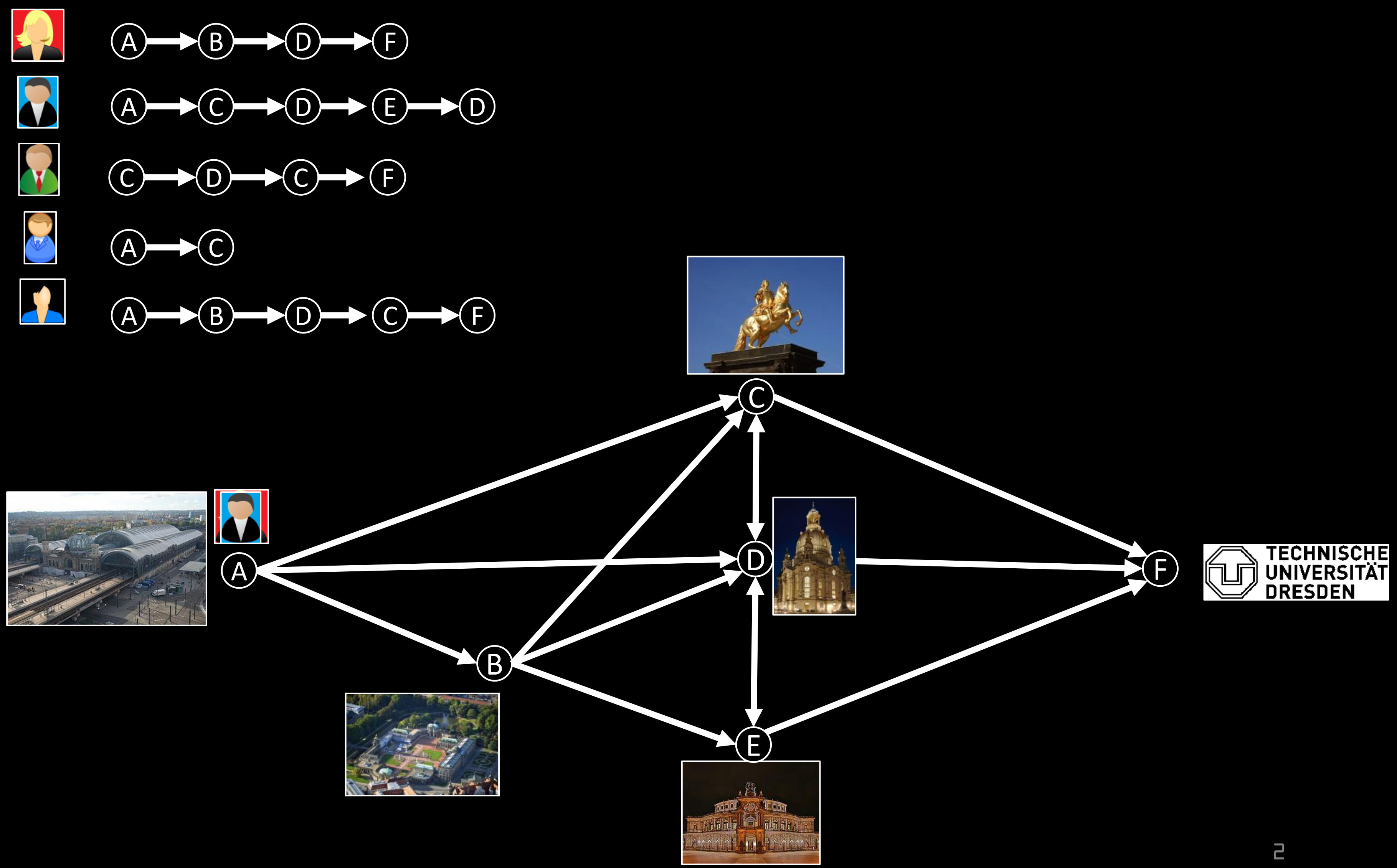
gesis Leibniz Institute for the Social Sciences

Bayesian hypothesis comparison in sequential data

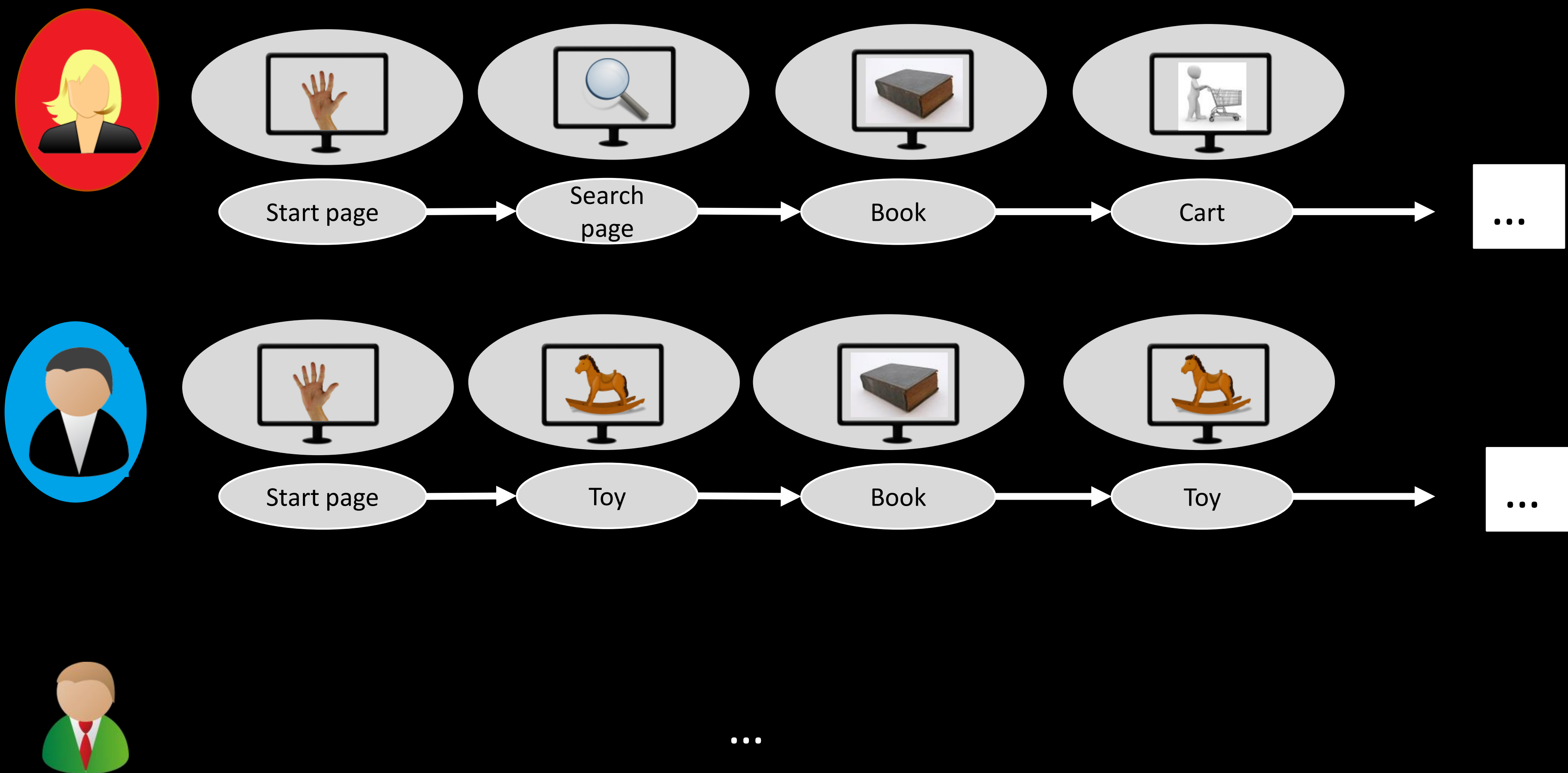
Florian Lemmerich

Dresden, 12.09.2017

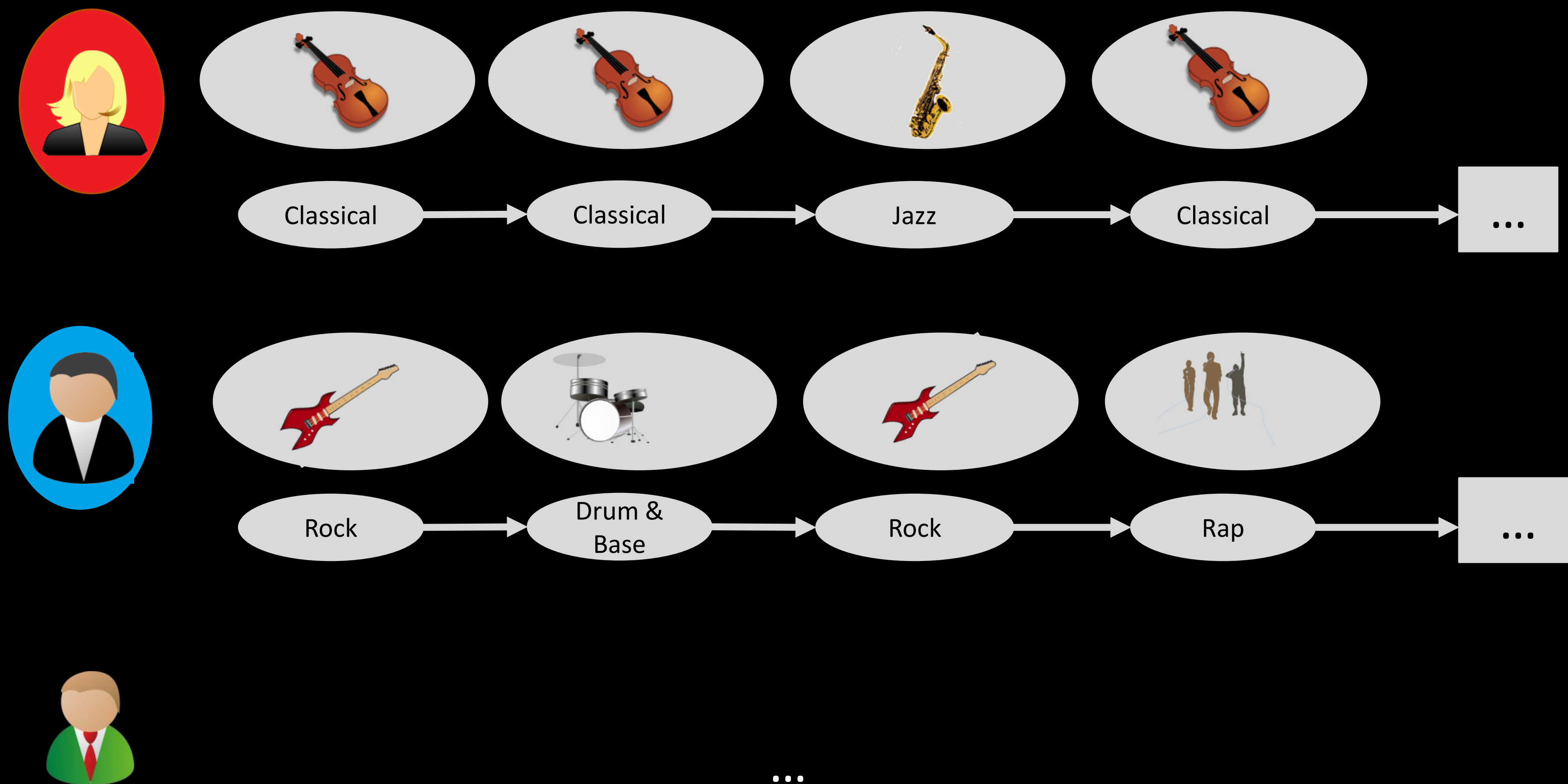
Urban navigation



Example: website navigation (online shop)



Example: listening history



What are the underlying mechanisms that generate this data?

Agenda

- Introduction
- Background: Markov Chain Models
- HypTrails: Comparing hypothesis about sequential data
  - Bayesian Hypothesis Testing
  - The Hyptrails approach
  - Applications
  - Extensions
- Conclusions

Background: Markov chain models

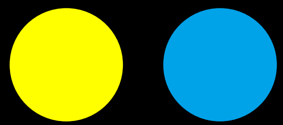
Markov chain model

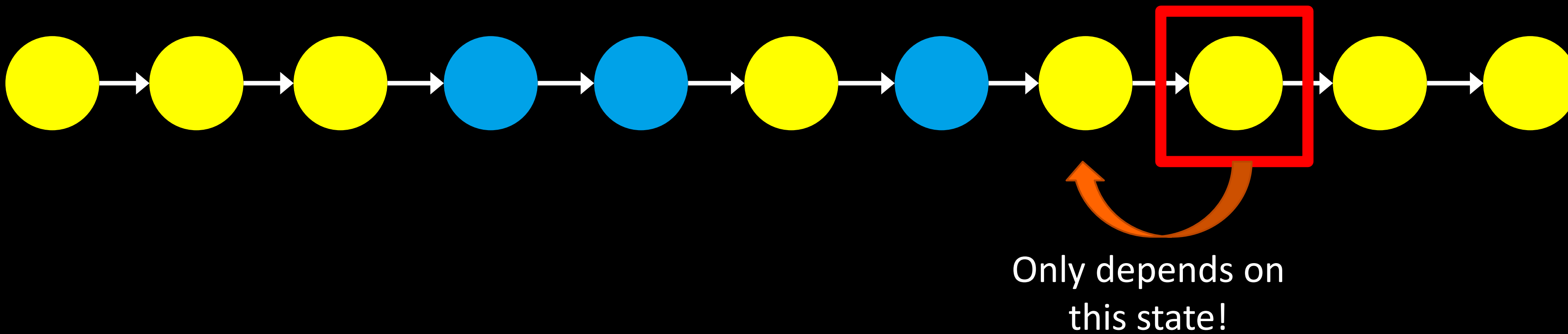
- Stochastic model for transitions between states
- State space $S = \{s_1, s_2, \ldots, s_m\}$
- Amounts to sequence of random variables $X_1, X_2, \ldots X_t$

- Markovian property:
  - Next state in a sequence only depends on the current one
  - Process is stable (constant) over time

$$P(X_{t+1} = s_j | X_1 = s_{i_1}, \ldots, X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}) = P(X_{t+1} = s_j | X_t = s_{i_t}) = p_{i,j}$$

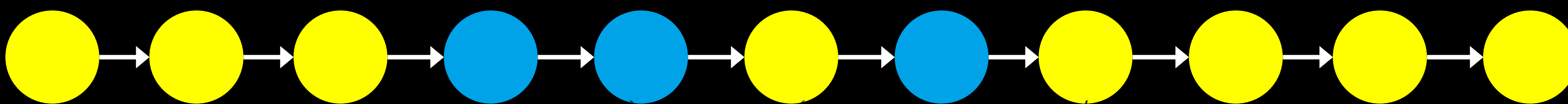
Example

Two States: 

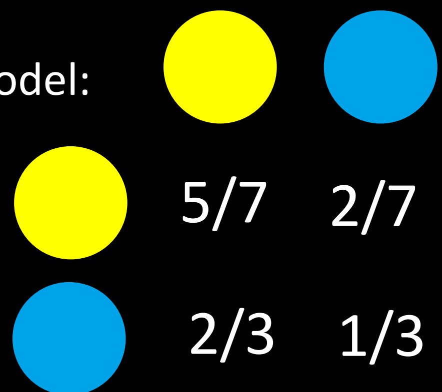


Computing the likelihood

- How good is a given model for some data?



Given a model:

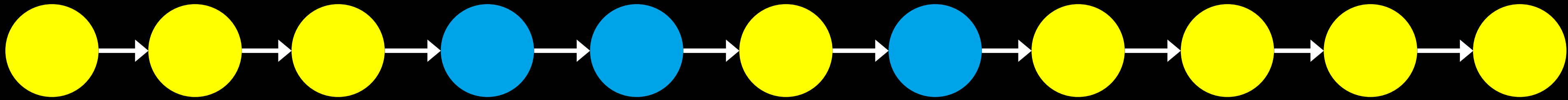


Likelihood: $$(5/7)^5 * (2/7)^2 * (2/3)^2 * (1/3)^1 = 0.002248$$

Log-Likelihood: $$5 * ln(5/7) + 2 * ln(2/7) + 2 * ln(2/3) \\ +1 * ln(1/3) = -6.0974$$

Fitting the model

- How to determine model parameters?



Transition count matrix

	5	2
	2	1

Transition probability matrix

	5/7	2/7
	2/3	1/3

= Model parameters

Extensions

- Higher order Markov chains
  - State depends on the last *n* states
- Variable order Markov chains
  - Order dependent on the context
  - Reduces parameter space of higher order Markov chains
- Hidden Markov models
  - There is an unobserved Markov chain sequence of variables that generates the observed sequence
- Semi-Markov chains
- Mixtures of Markov chains
- …

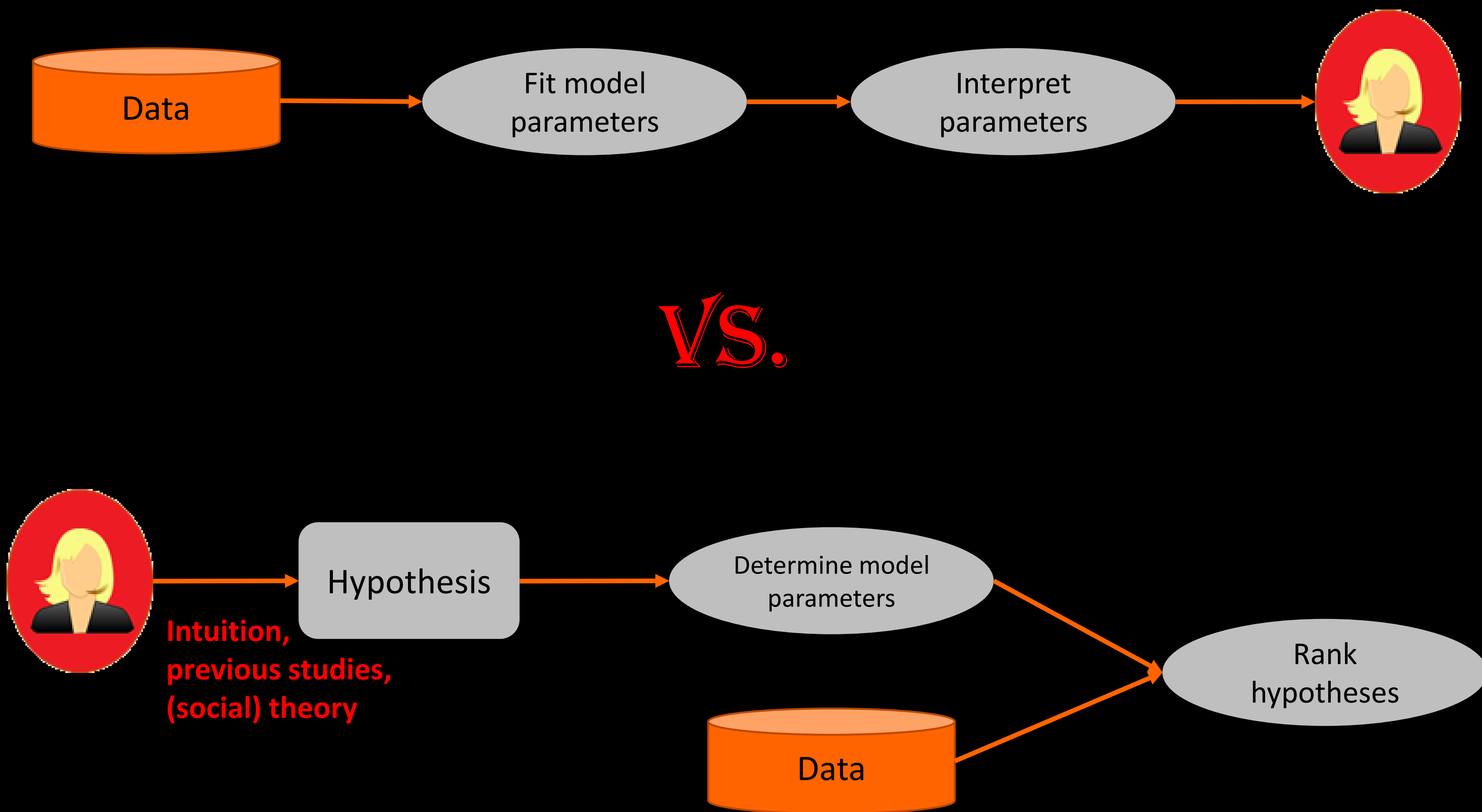
Applications

- Sequence of letters [Markov 1912, Hayes 2013]
- Web navigation, PageRank [Page et al. 1999]
- Speech recognition [Rabiner 1989]
- Weather data [Gabriel & Neumann 1962]
- Gene, DNA sequences [Salzberg et al. 1998]
- Computer performance evaluation [Scherr 1967]

- Markov Chain Monte Carlo (MCMC)

HypTrails

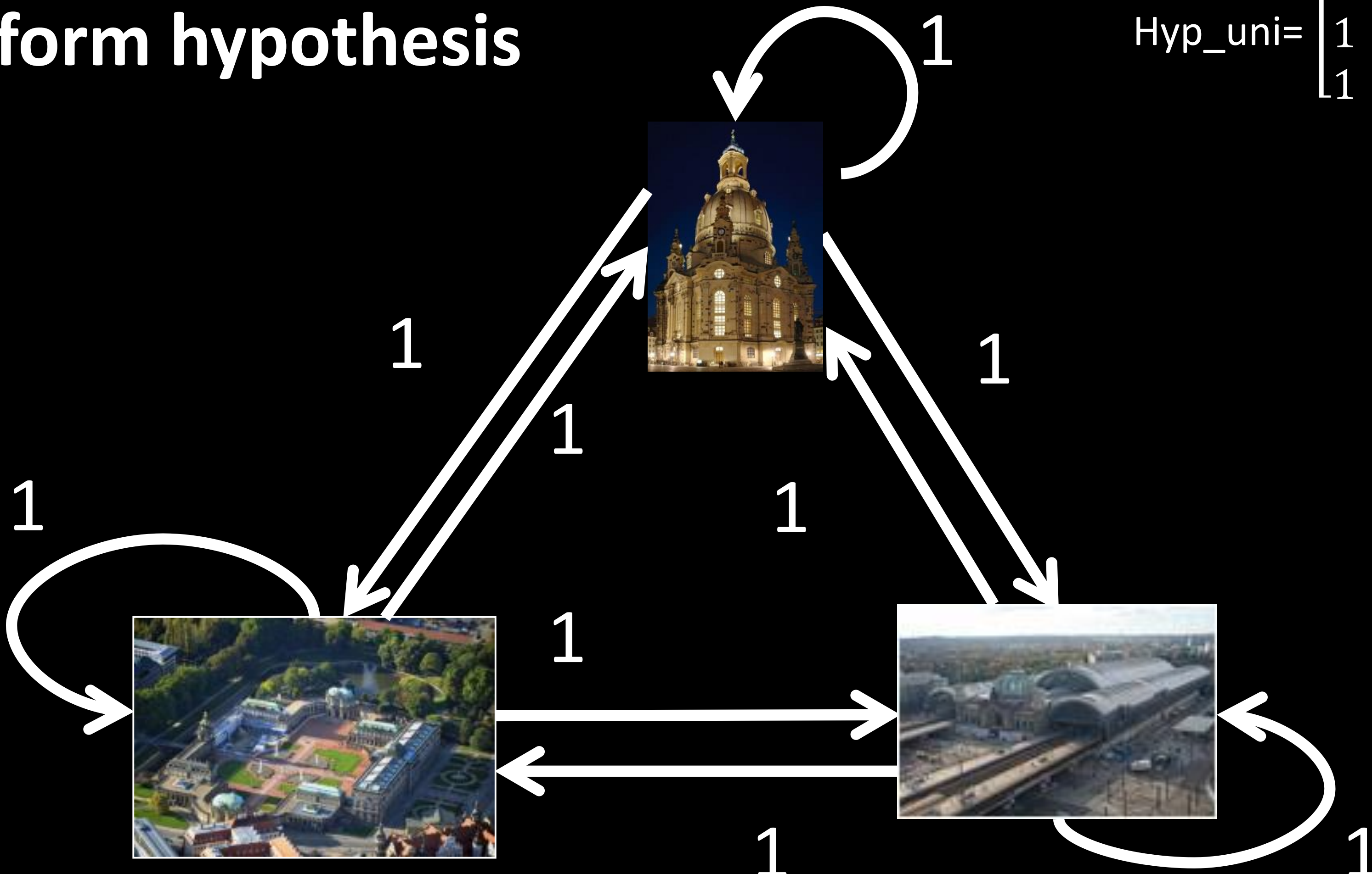
Parameter learning vs hypothesis testing



Example

Uniform hypothesis

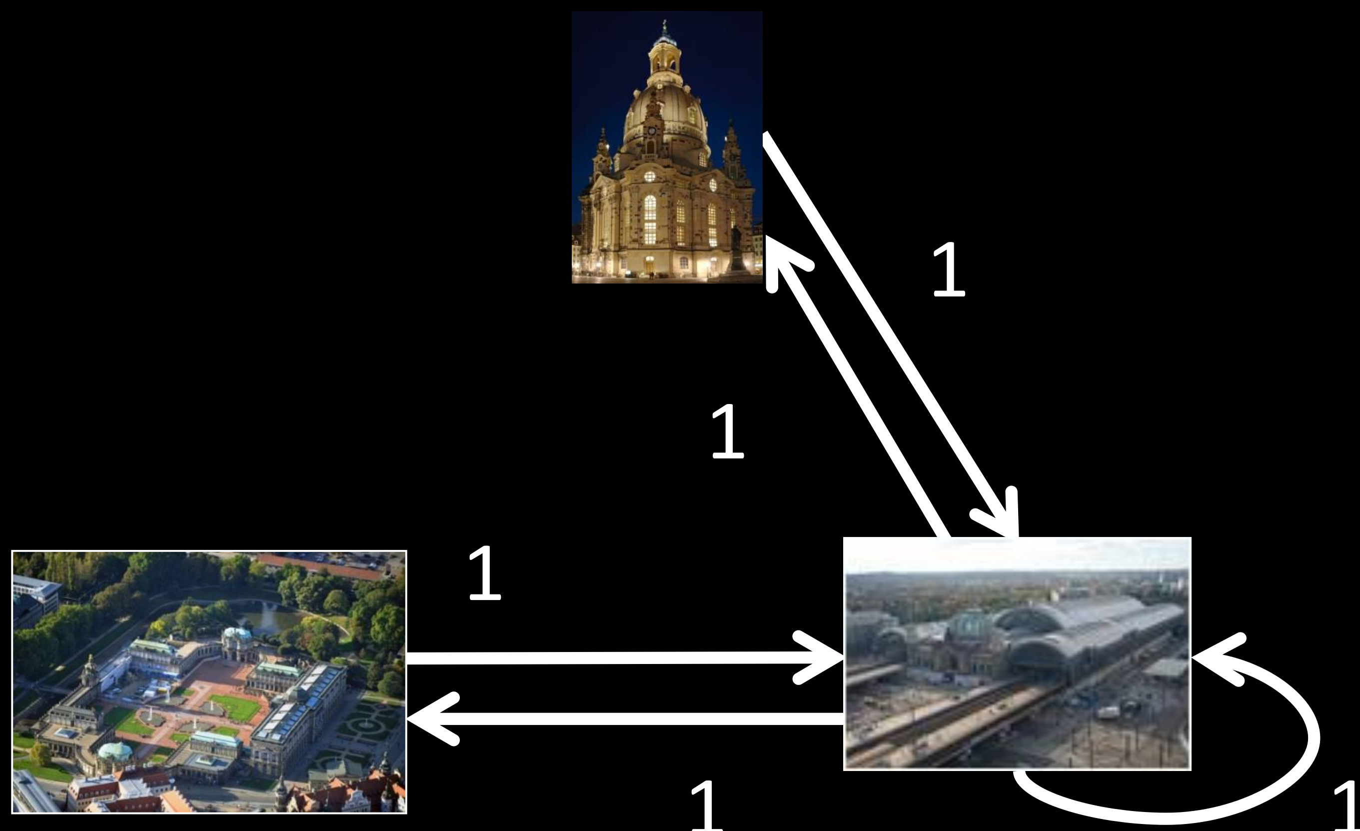
$$\text{Hyp\_uni} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$



Example

Bus route hypothesis

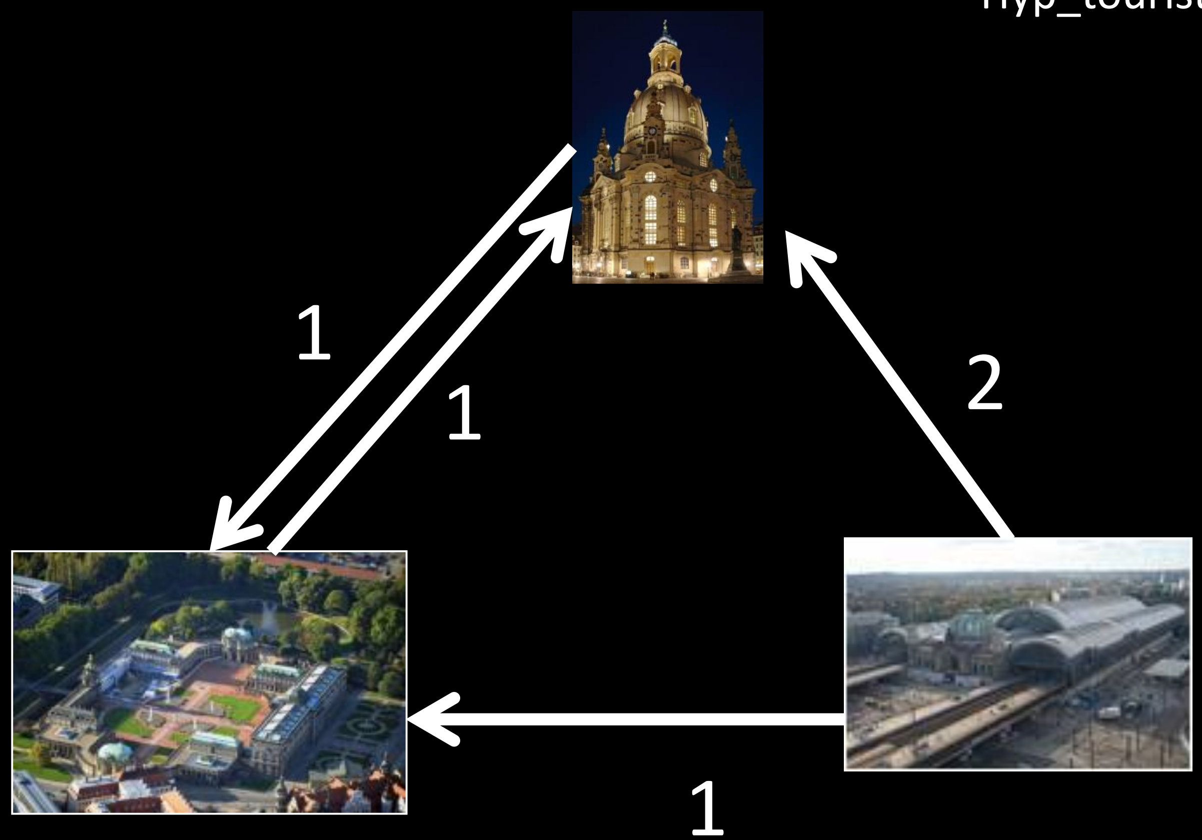
$$\text{Hyp\_bus} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$



Example

Tourist hypothesis

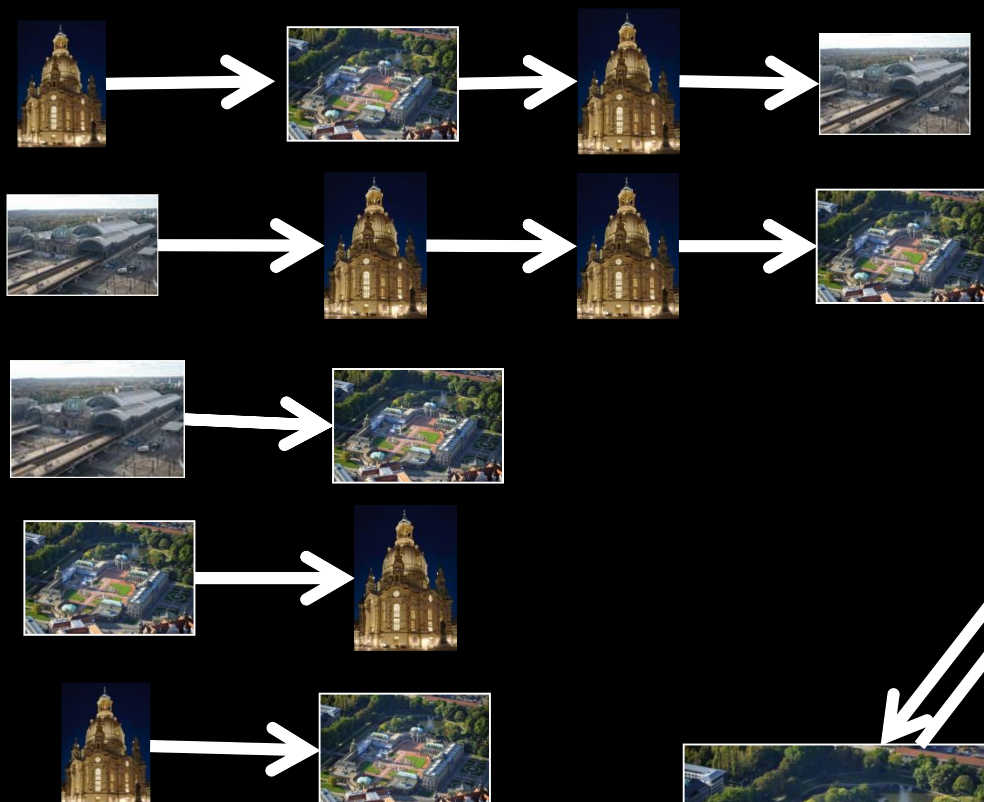
$$\text{Hyp\_tourist} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 2 & 1 & 0 \end{bmatrix}$$

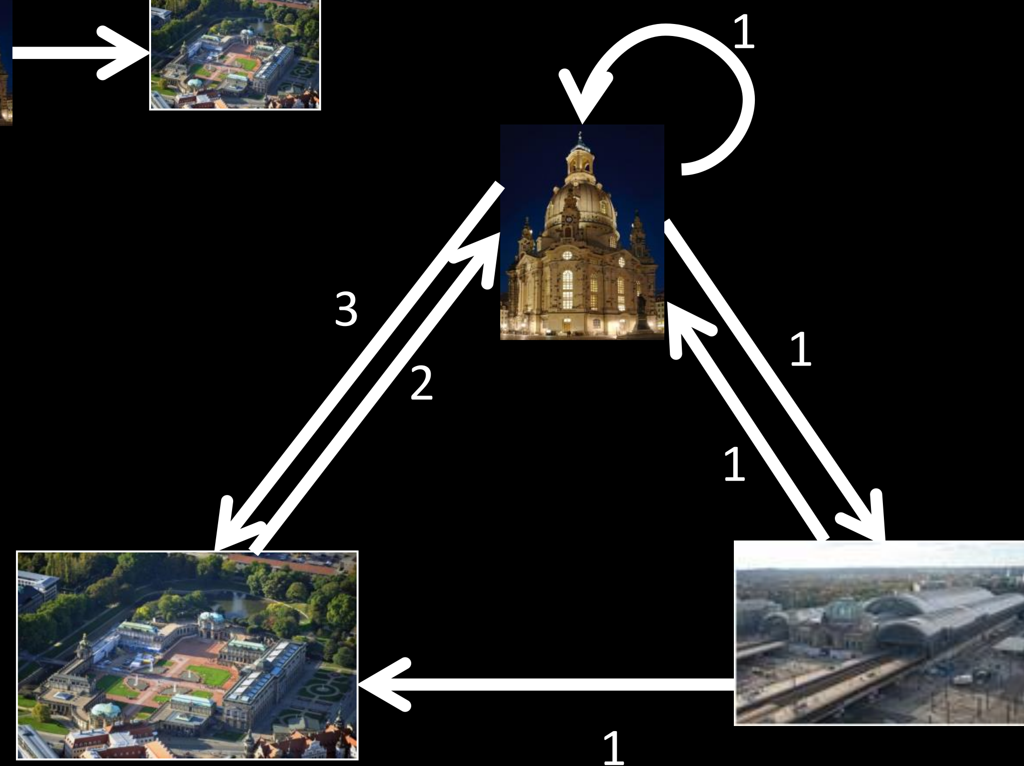


Example

Observed data

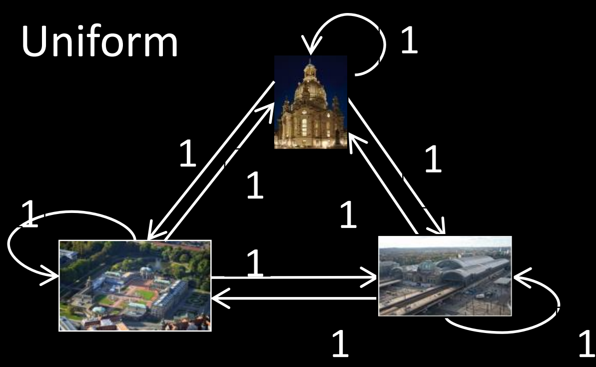
$$\text{Data} = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$



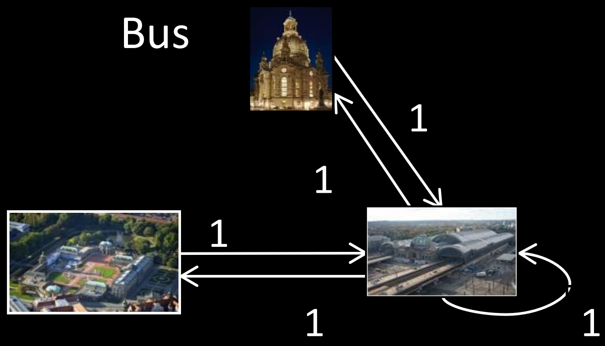


Summary

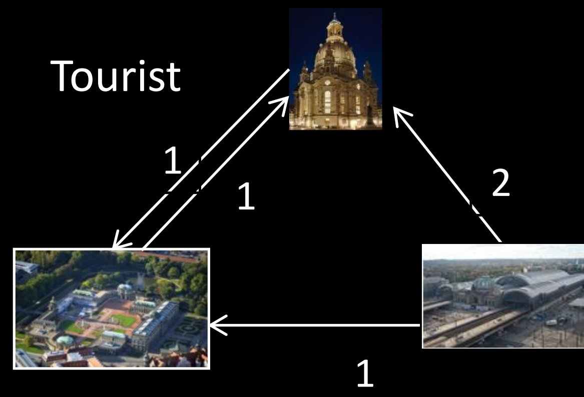
Uniform



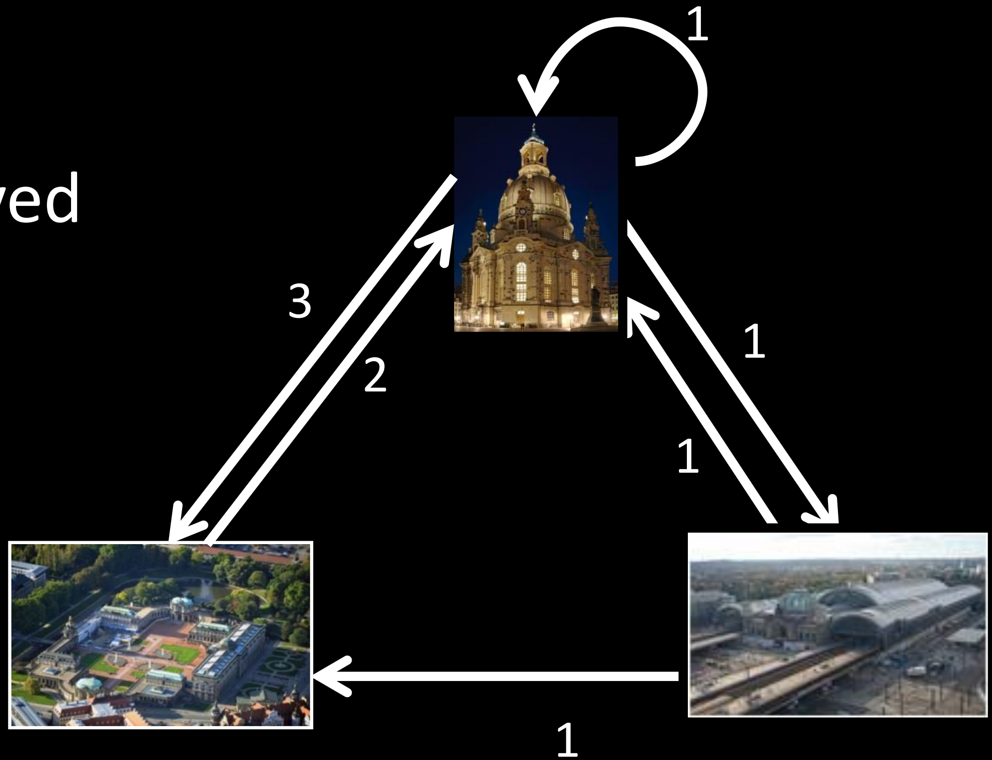
Bus



Tourist



Observed
Data



Which hypothesis is most plausible given the observed data?

Goal

- Come up with an ordering of such hypotheses with respect to plausibility to observed data
- Consider that hypothesis specifications are not precise/uncertain
- Compare the “significance” of a difference in plausibility between two hypotheses
- **NOT a goal:** come up with a good (but not interpretable) model

Model comparison

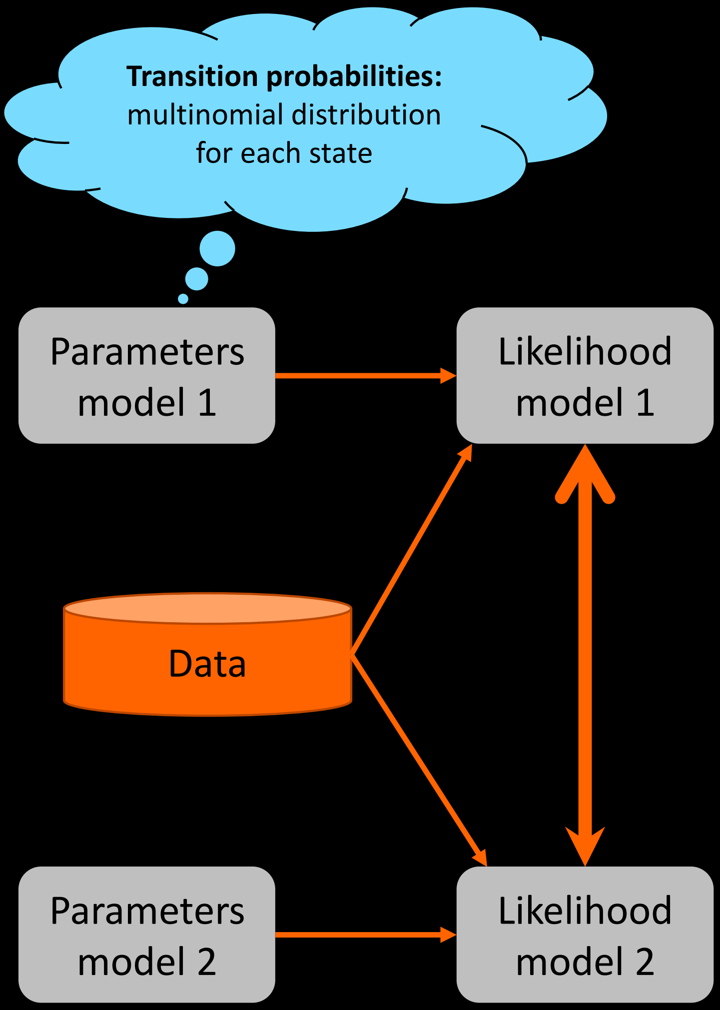
- Given two (parameterized) models, which model is better?

- Simple methods: compare the likelihoods

- Alternatives (for different types of models):
  - Akaike Information Criterion (AIC),
  - Bayesian Information Criterion (BIC),
  - Likelihood ratio test

  - **Bayes Factors**

Frequentist model comparison



Bayesian Statistics

- Random variables model *uncertainty* in the data
- Probability distributions model beliefs
- *Prior* beliefs get updated to a *posterior* belief once new data becomes available (with Bayes Formula)

$$P(A|B) = \frac{P\,(B|\,A)\,P(A)}{P(B)}$$

- Often a problem: dependency on the prior

Bayesian model selection

- Probability theory for choosing between models
- Posterior probability of model M given data D

Update parameters:

$$\overbrace{P(\theta|D,M)}^{\text{posterior}} = \frac{\overbrace{P(D|\theta,M)}^{\text{likelihood}}\,\overbrace{P(\theta|M)}^{\text{prior}}}{\underbrace{P(D|M)}_{\text{marginal likelihood}}}$$

Evidence

Evidence

Update belief in models:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

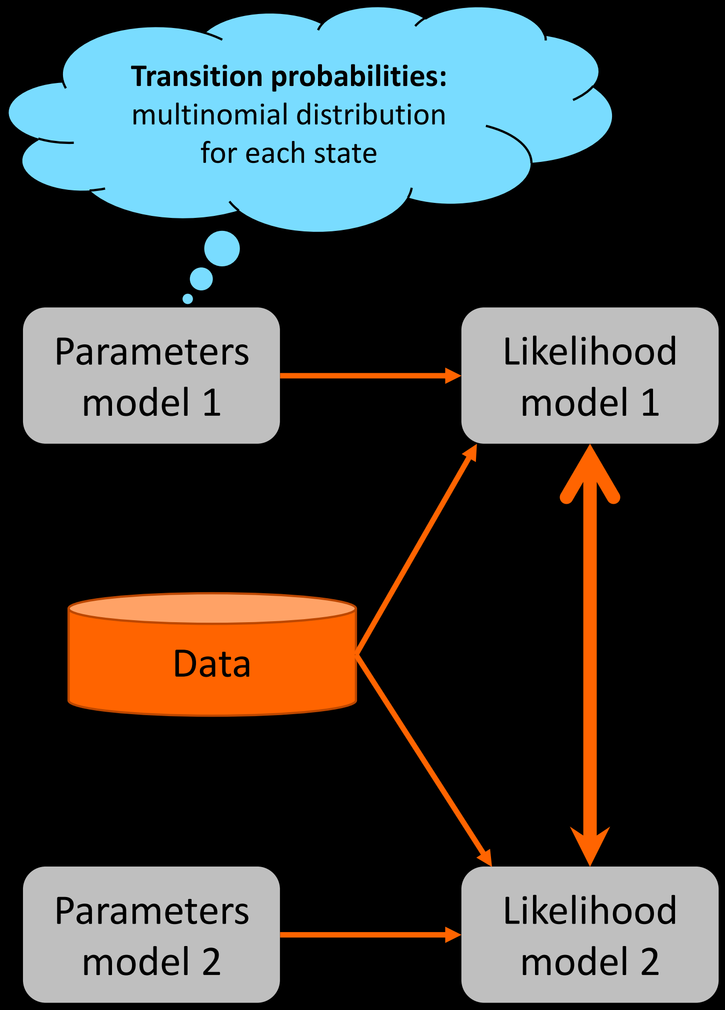
Bayes Factor

- Comparing two models $$\underbrace{\frac{\Pr(M_i|D)}{\Pr(M_j|D)}}_{\text{posterior odds}} = B_{i,j} \cdot \underbrace{\frac{\Pr(M_i)}{\Pr(M_j)}}_{\text{prior odds}}, \text{ with } \underbrace{B_{i,j} = \frac{\Pr(D|M_i)}{\Pr(D|M_j)}}_{\text{Bayes factor}}$$
- Bayes Factor

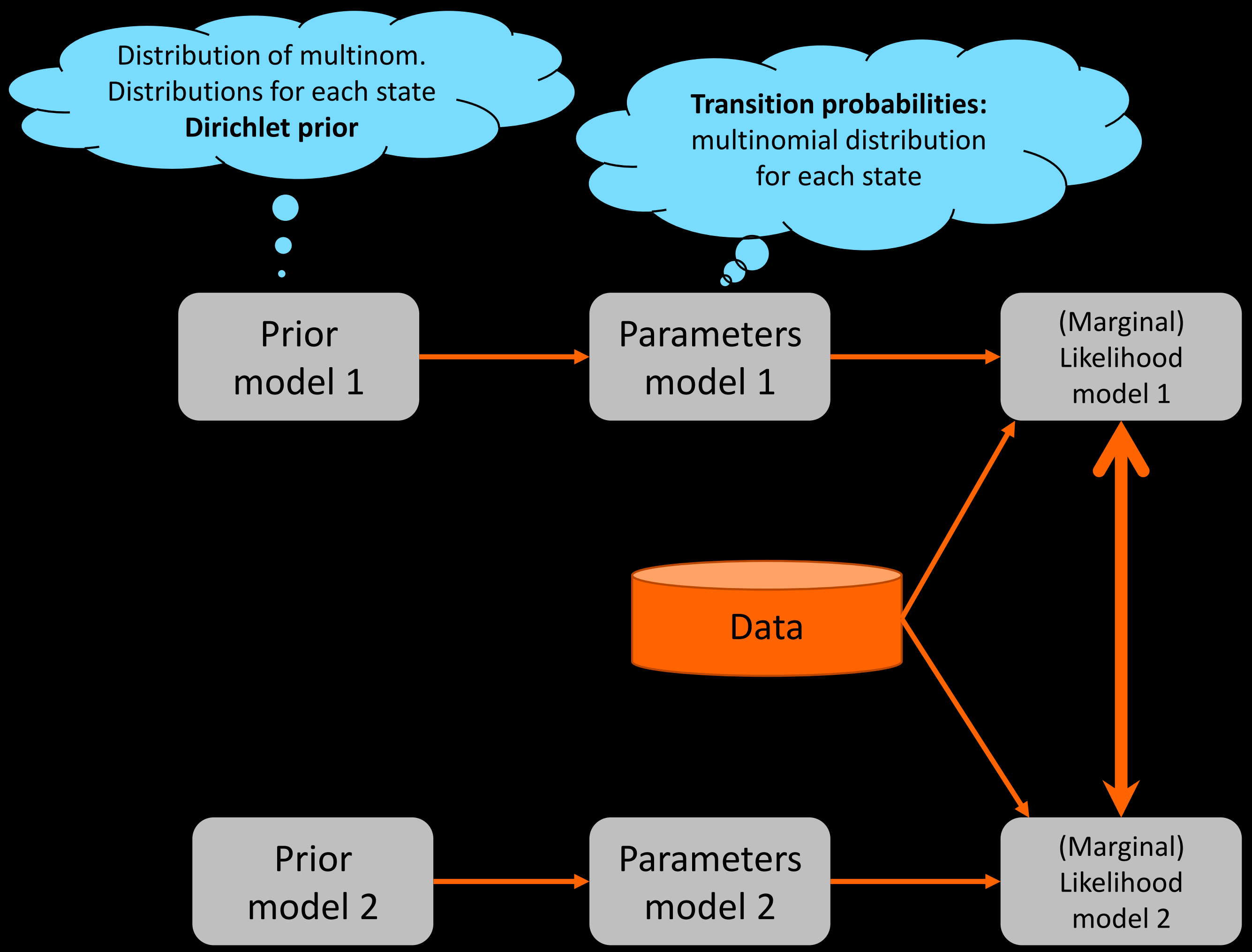
- Evidence: Parameters marginalized out
- Automatic penalty for model complexity (Occam's razor)
- Strength of Bayes factor: interpretation table

- It is a relative comparison!

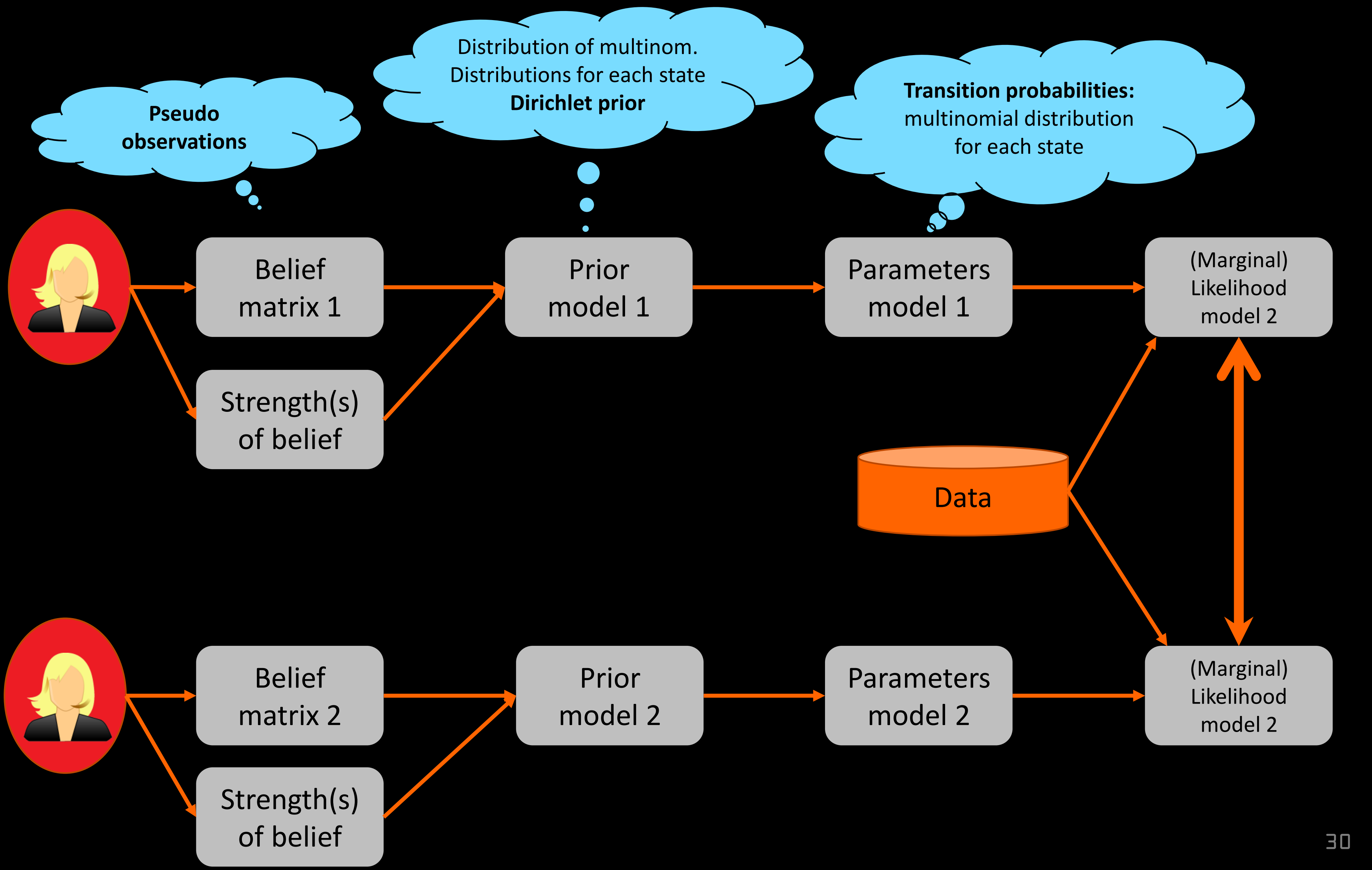
Frequentist model comparison



Bayesian model comparison



HypTrails



Pseudo observations
Distribution of multinom. Distributions for each state Dirichlet prior
Transition probabilities: multinomial distribution for each state
Belief matrix 1
Prior model 1
Parameters model 1
(Marginal) Likelihood model 2
Strength(s) of belief
Data
Belief matrix 2
Prior model 2
Parameters model 2
(Marginal) Likelihood model 2
Strength(s) of belief

HypTrails

- Conjugate Prior: Dirichlet distribution (belief in parameters)

$$Dir(\boldsymbol{x}, \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_j x_j^{\alpha_j - 1} = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_j x_j^{\alpha_j - 1}$$

- Marginal Likelihood (Evidence)

$$P(D|M) = \prod_i \frac{\Gamma(\sum_j \alpha_{i,j})}{\prod_j \Gamma(\alpha_{i,j})} \frac{\prod_j \Gamma(n_{i,j} + \alpha_{i,j})}{\Gamma(\sum_j (n_{i,j} + \alpha_{i,j}))}$$

- Usually we compute (and plot) log (marginal likelihoods)!

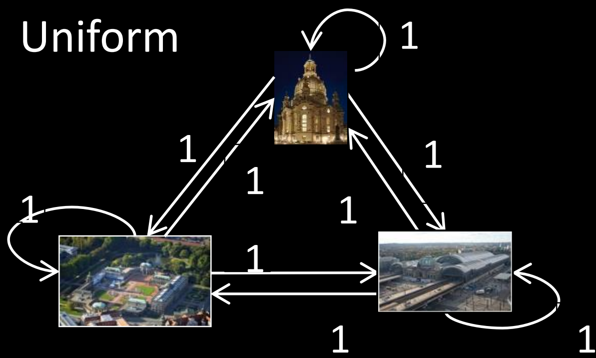
HypTrails

- Input:
  - A set of belief matrices
  - A set of parameters k for the strength of belief
  - Observed data

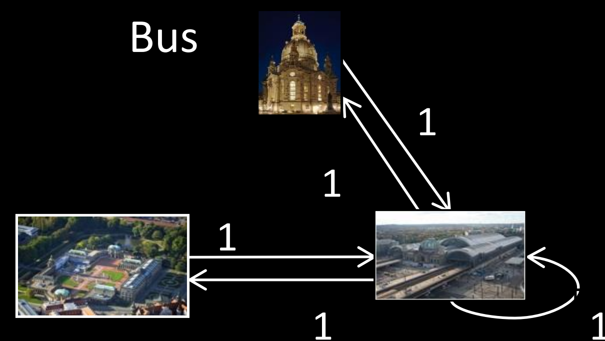
- Output:
  - A marginal likelihood for each hypothesis and each k
  - → Ordering of the hypotheses with respect to their plausibility for the data
  - → A Bayes Factor to compare two hypotheses (substitute for a p-value)

Summary

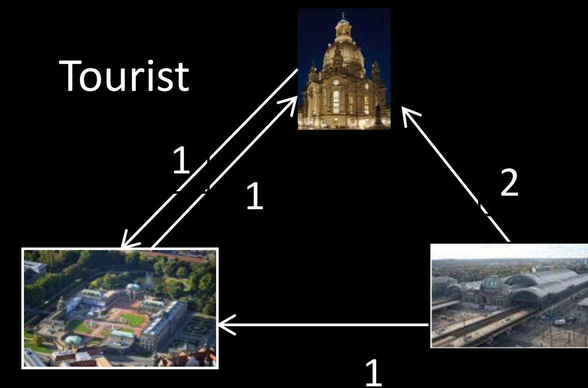
Uniform



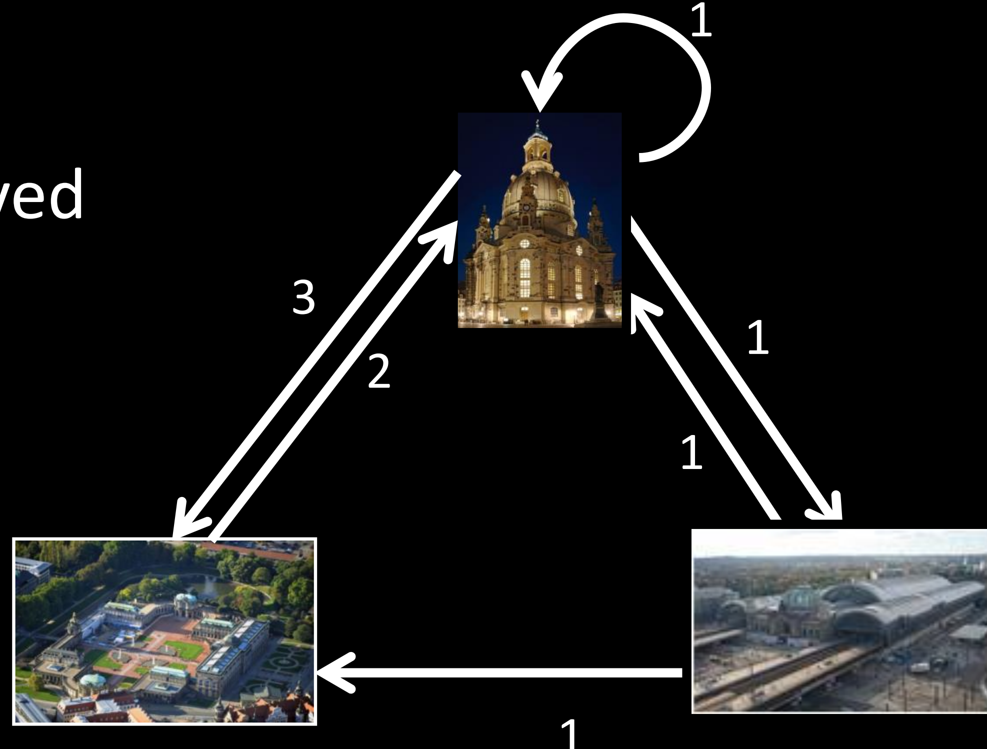
Bus



Tourist

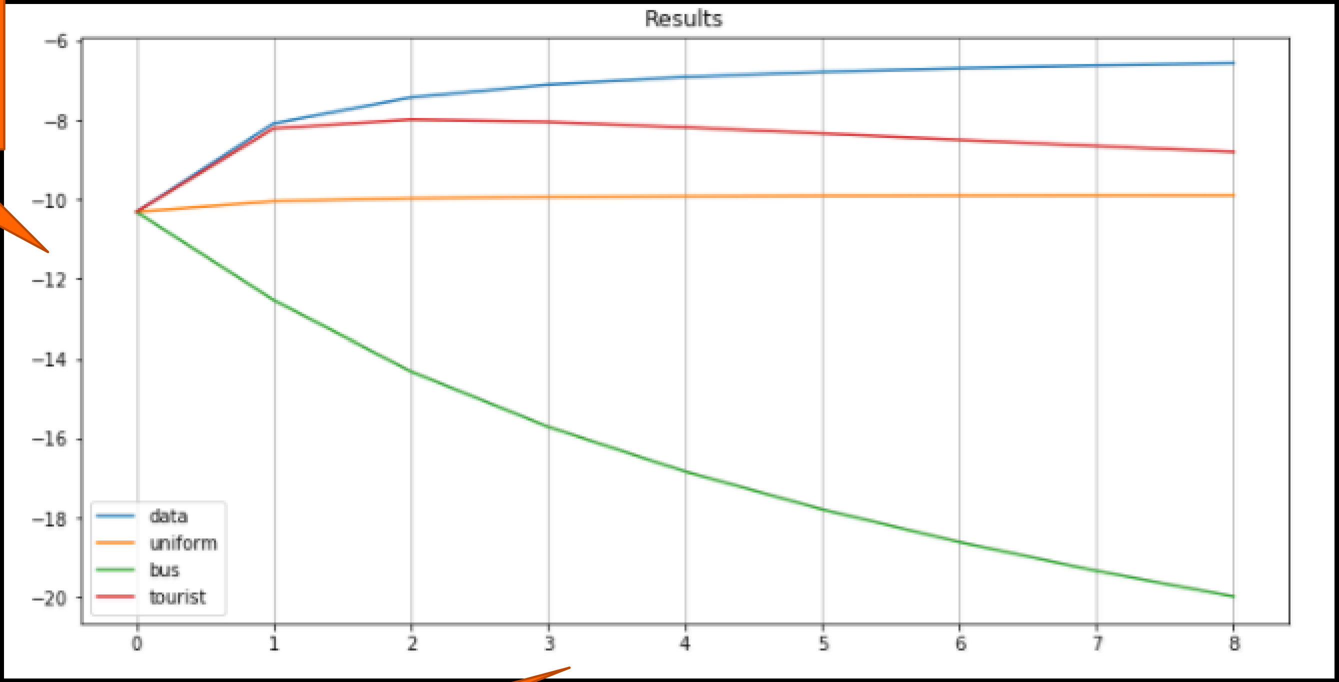


Observed Data



Example results

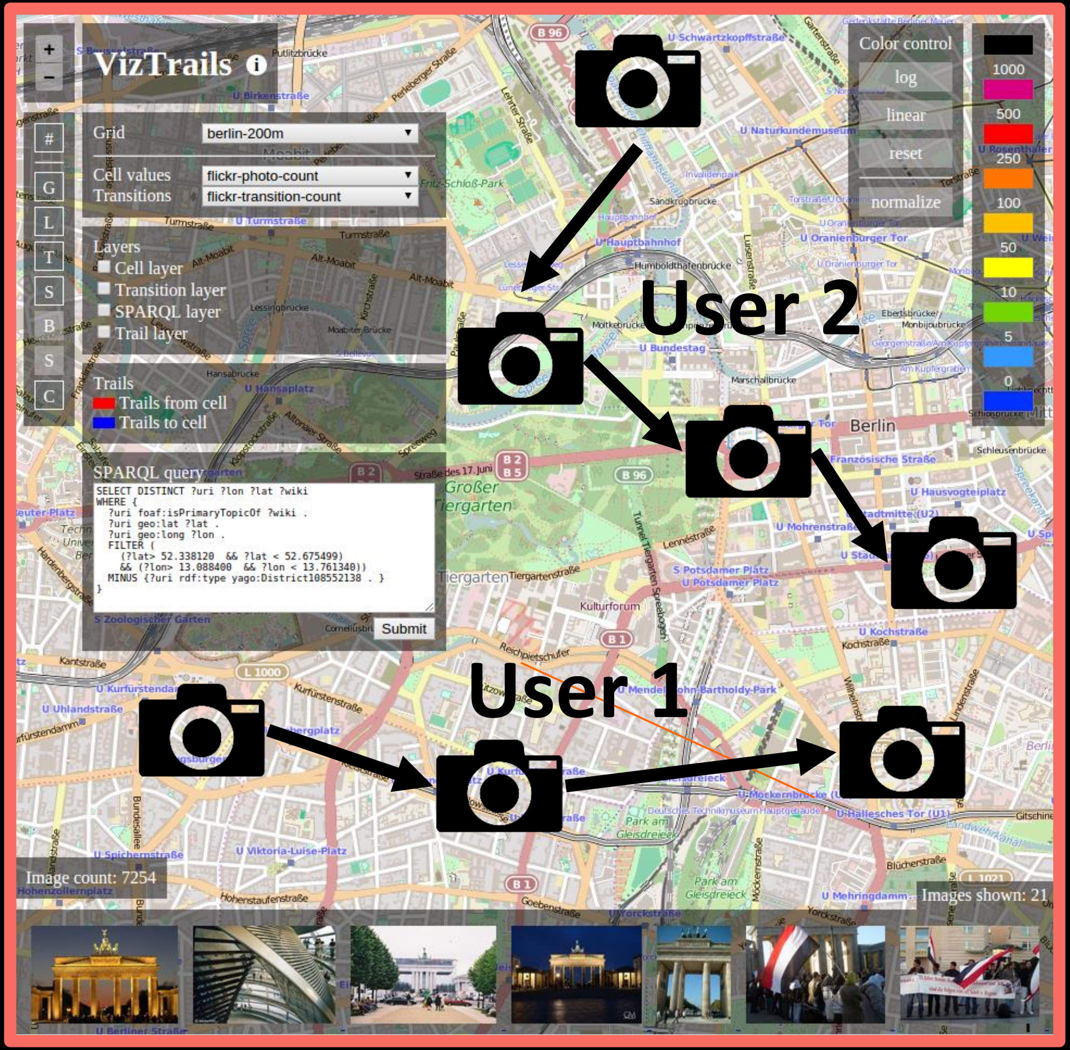
Higher plausibility (marginal likelihood)



Stronger belief in parameters

Applications

FlickrTrails

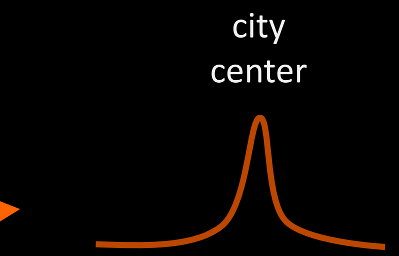


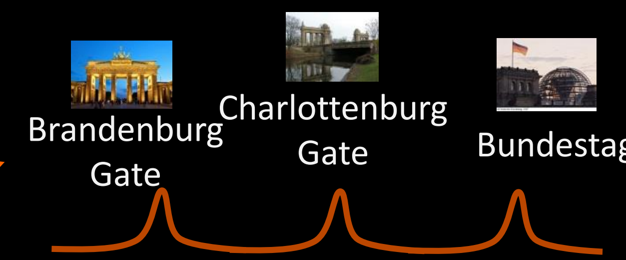
FlickrTrails

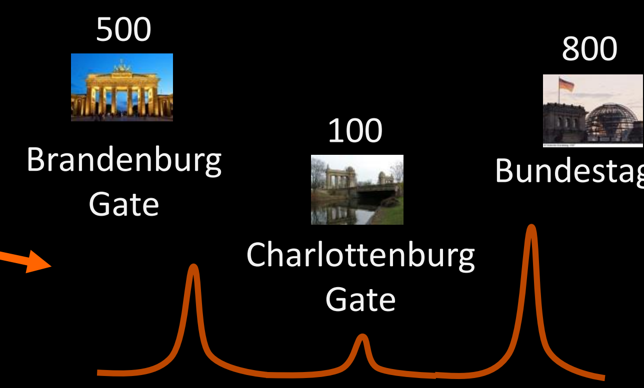
- Crawled all pictures with geo-tags in 4 major cities
- Generated user paths for each user within the city
- Used grid to obtain a discrete state space
- Where will a user take his next picture?
- Details:
  - Only photos with accuracy 16 (street level)
  - 200 x 200m grids
  - One trail per user
  - No self transitions
  - Minimum trail length 2

Flickr Hypotheses

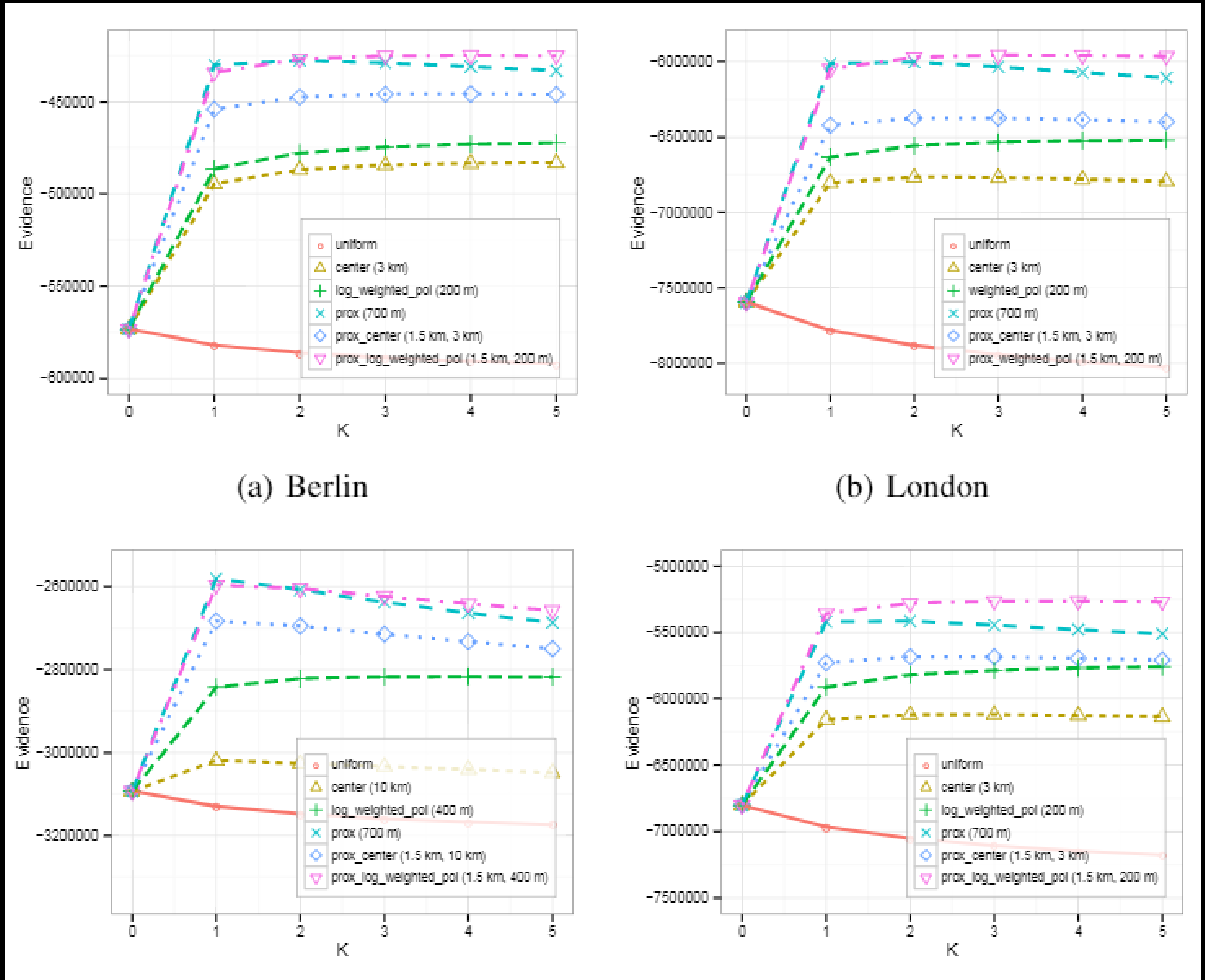
- Uniform
- Center
- Proximity (several)
- Points-of-interest
- Weighted points-of-interest
- Mixtures of hypotheses







FlickrTrails: Results



(a) Berlin

(b) London

(c) Los Angeles

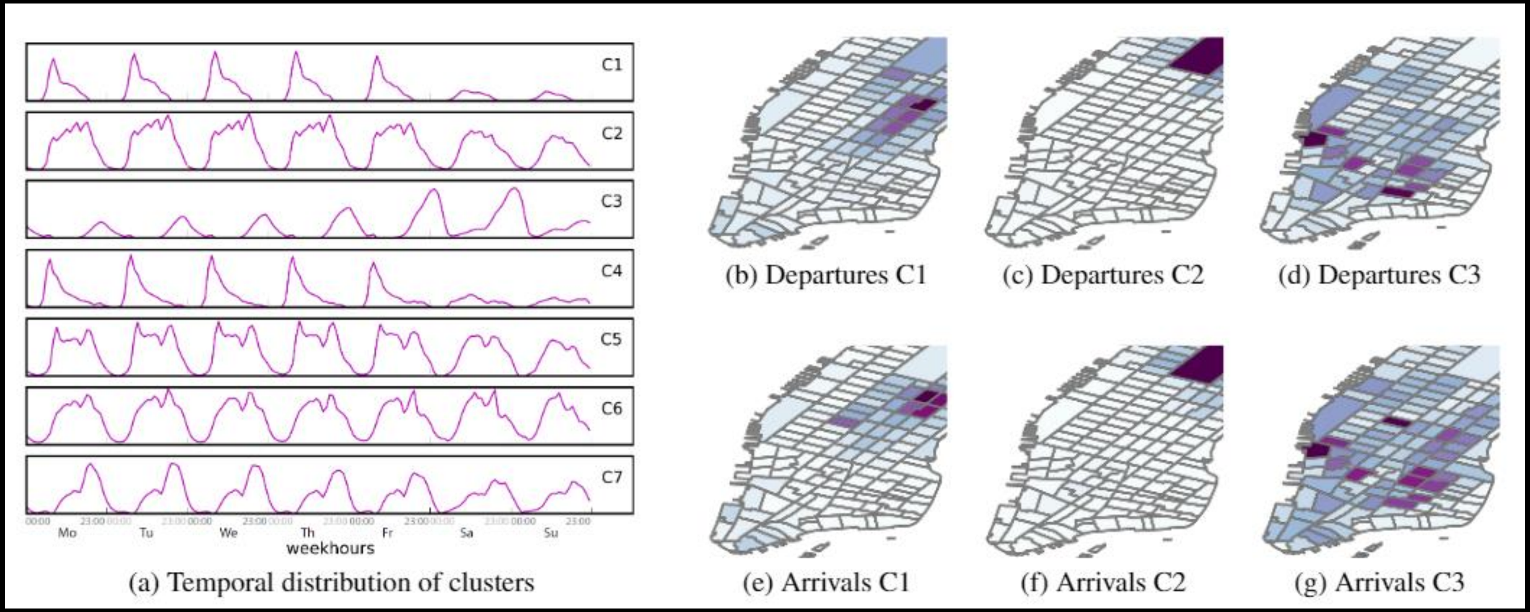
(d) New York City

TaxiTrails

- Data on ~170 million taxi rides in New York City in 2013
- Mapped each start and stop location to its NYC tract
- Focused on rides within Manhattan
- Features to build hypotheses:
  - Distance-based: Geographical Center, Flatiron Building, Times Square
  - Census-data: Population size, percentage of white people, black people, People in labor force, people below poverty level, number of theaters, number of libraries, % occupied by parcs, …
  - Foursquare-data: # venues/checkins overall, and filtered on types of venues (nightlife, sport, food, shops)…
- Overall 70 hypotheses

Clustering of taxi rides

- Additional:
  Spatio-temporal clustering of data (by tensor-factorization)



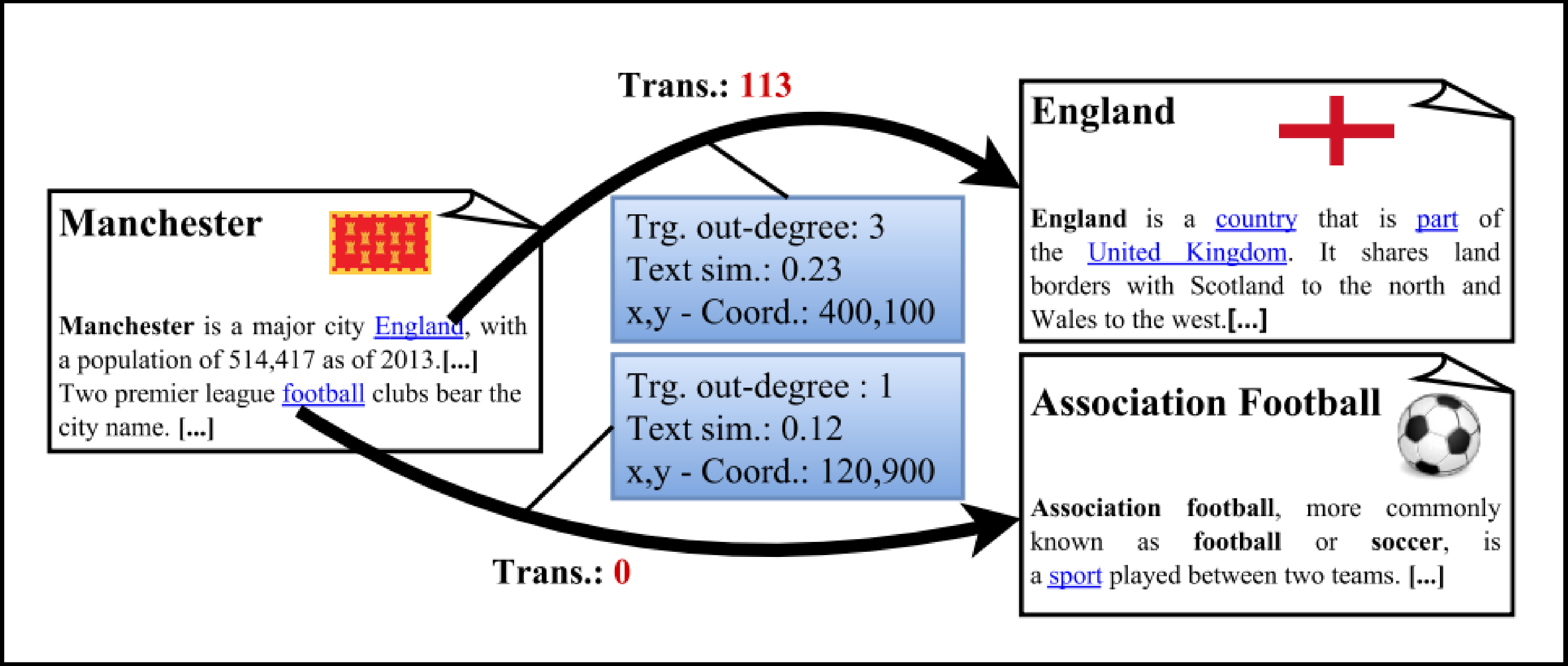
Taxi data results

- Apply HypTrails separately on each cluster, rank hypotheses

Table 4: Ranking of Hypotheses. This table shows the ranking of 23 out of 70 hypotheses evaluated with HypTrails over 3 different groups. *Overall* represents all 143*M* taxi rides in Manhattan 2013, clusters C_i are clusters identified by NTF. Numeric cells represent the ranks of the hypotheses in respective clusters. For the distance-based hypotheses, we only show results for the best parameter of the standard deviation σ (parameter in parentheses). Green cells highlight all hypotheses that outperform the uniform hypothesis.

HYPOTHESES	Overall *2013*	C1 *Workdays 9am*	C2 *Workdays 6pm*	C3 *Weekends 1am*	C4 *Workdays 7am*	C5 *Workdays 9am, 6pm*	C6 *Mo-Sa 6pm Sa-Su 2pm*	C7 *Workdays 6pm*
Baseline								
Uniform	42	56	56	56	56	55	62	59
Distance-based (σ)								
Proximity	14 (3.0)	7 (1.0)	2 (0.5)	14 (1.0)	10 (1.0)	19 (1.0)	10 (0.01)	13 (1.0)
Centroid (Geographical Center)	38 (5.0)	50 (5.0)	25 (1.0)	58 (5.0)	52 (5.0)	58 (5.0)	51 (3.0)	51 (5.0)
Centroid (Flatiron Building)	29 (5.0)	32 (2.0)	51 (5.0)	17 (1.0)	2 (0.01)	4 (0.5)	44 (3.0)	20 (0.5)
Centroid (Times Square)	22 (3.0)	1 (0.5)	43 (3.0)	46 (3.0)	1 (0.5)	43 (2.0)	2 (0.01)	1 (0.01)
Foursquare								
Gravitational (All venues)	1	12	14	10	14	10	14	9
Gravitational (Check-ins)	9	3	30	2	11	5	4	3
Gravitational (Work)	2	5	12	24	8	6	13	11
Gravitational (Food)	5	4	31	4	12	15	11	4
Gravitational (Party)	7	17	37	1	19	9	20	5
Gravitational (Recreation)	15	21	10	9	17	13	7	33
Venue Similarity	39	53	53	53	53	52	58	53
Census								
Gravitational (Population)	21	61	28	20	59	46	42	25
Gravitational (Tract Area)	23	34	8	26	24	20	24	38
Gravitational (%White people)	6	24	13	28	28	27	35	27
Gravitational (Residential zoning)	50	65	19	35	65	61	67	49
Gravitational (Commercial zoning)	13	8	32	22	9	24	19	15
Gravitational (Art Galleries)	46	23	1	38	5	2	52	54
Gravitational (Museums)	54	13	3	40	6	7	26	58
Gravitational (Parks)	63	62	4	44	63	59	6	64
Race Similarity	32	48	50	52	50	50	59	50
Poverty Similarity	37	55	52	54	55	53	61	56
Employment Similarity	40	57	54	55	58	54	60	58

What makes a link succesful in Wikipedia

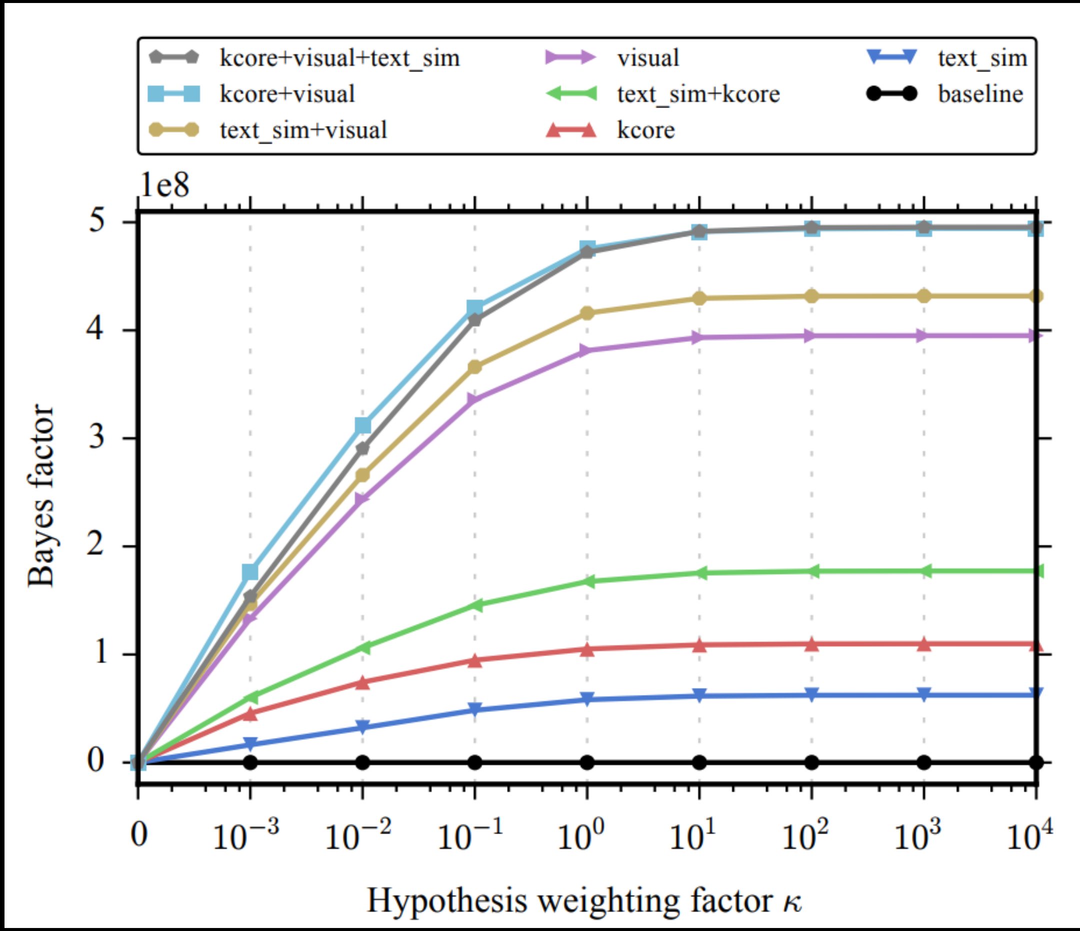


What makes a link succesful in Wikipedia

- Data:
  - Once month of viewer data
  - Source page -> target page

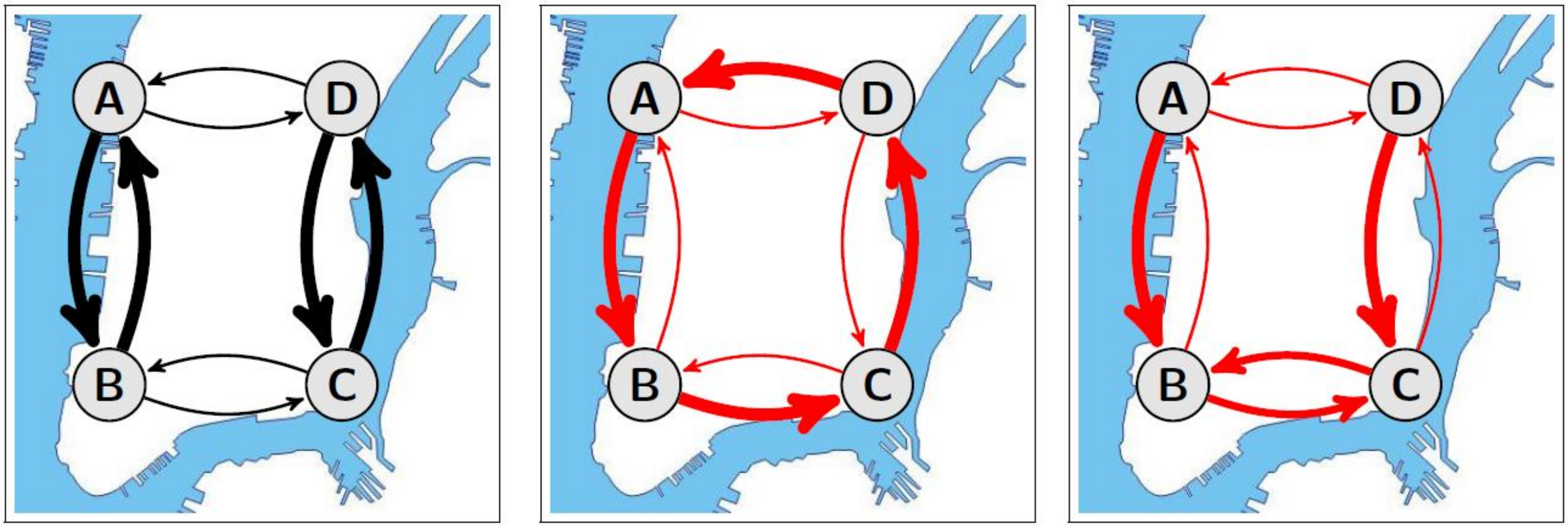
- Features to form hypotheses:
  - Network-based: degree, centrality (k-core), page rank
  - Similarity-based features: text similarity, category similarity
  - Link-position features: head, body, info-box, nav-bar, …

What makes a link succesful in Wikipedia



Extensions

Subgroup Behavior

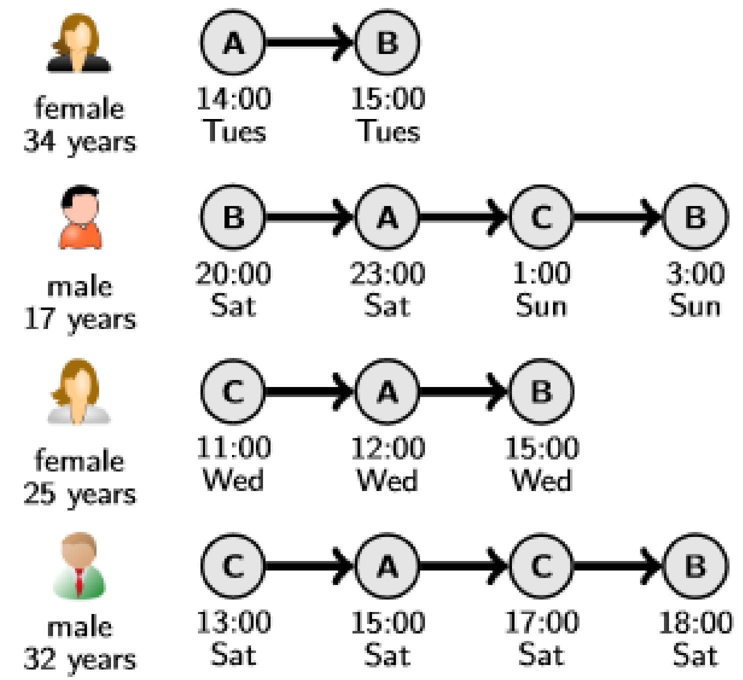


(a) All transitions

(b) Transitions of tourists

(c) Transitions of locals in the evening hours

Data Preparation



(a) Sequence data with background knowledge

A_M		A_D				
Source State	Target State	Gender	Age	Hour	Weekday	# Visits of user
A	B	f	34	14	Tue	2
B	A	m	17	20	Sat	4
A	C	m	17	23	Sat	4
C	B	m	17	1	Sun	4
C	A	f	25	11	Wed	3
A	B	f	25	12	Wed	3
C	A	m	32	13	Sat	4
A	C	m	32	15	Sat	4
C	B	m	32	17	Sat	4

(b) Transition dataset

SubTrails

- Based on Subgroup Discovery / Exceptional Model Mining

- Find interpretable descriptions of subsets in the data that
  - …have significantly different transition behavior than the entire dataset
  - … specifically match a hypothesis **or**
  - **… specifically contradict a hypothesis**

Results: Subtrails (Flickr)

(a) Comparison to the overall dataset

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
# Photos > 714	76,859	103.83 ± 2.41	42,277	106.68
# Photos $\leq$ 25	78,254	88.83 ± 2.07	37,555	141.78
Tourist = True	76,667	75.42 ± 1.79	33,418	148.64
Tourist = False	310,314	75.00 ± 1.60	33,418	16.92
Country = US	163,406	64.47 ± 1.39	44,822	70.97
# Photos = 228-715	77,448	46.10 ± 1.02	33,214	115.65
Country = Mexico	2,667	33.22 ± 0.82	3,575	122.83
# PhotoViews > 164	79,218	31.58 ± 0.74	31,461	107.84
# PhotoViews < 12	76,573	30.54 ± 0.71	30,881	110.83

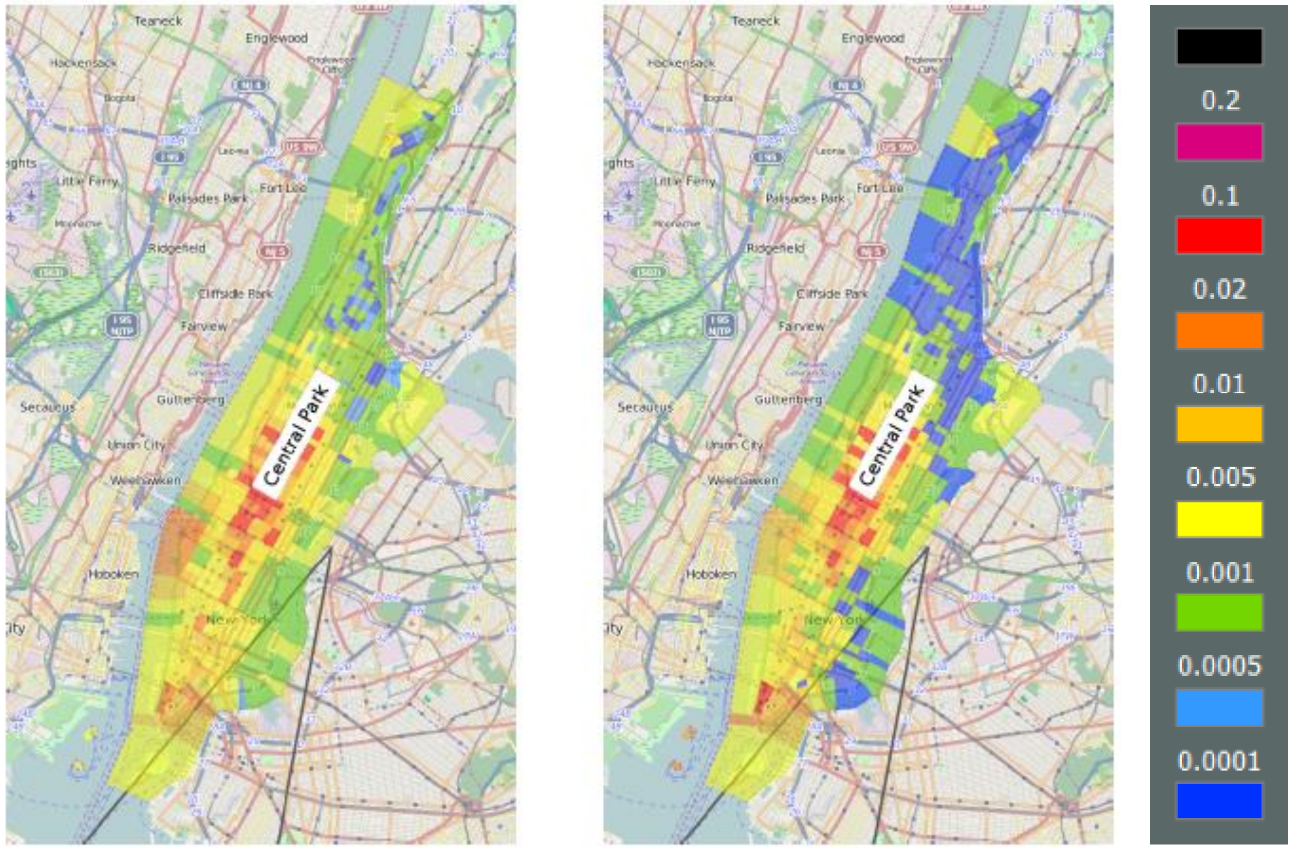
(b) Comparison to the *Proximate-PoI* hypothesis, contradicting

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
# Photos $\leq$ 25	78,254	64.85 ± 1.37	110,124	221.07
# Photos = 26–81	77,003	23.41 ± 0.53	99,646	207.21
Hour = 22h–23h	14,944	18.26 ± 0.43	20,526	215.69
Hour = 23h–0h	11,726	17.42 ± 0.37	16,404	208.91
Hour = 21h–22h	17,806	16.52 ± 0.33	23,951	211.34
Tourist = False	310,314	16.09 ± 0.35	379,676	185.13
Hour = 0h–1h	9,693	15.12 ± 0.33	13,590	215.42

(c) Comparison to the *Proximate-PoI* hypothesis, matching

Description	# Inst.	$-q_{tv}$ (score)	ω_{tv}	Δ_{tv}
# Photos > 714	76,859	58.59 ± 1.30	80,690	164.16
# PhotoViews < 12	76,573	21.56 ± 0.50	88,948	185.78
Hour = 12h–13h	25,022	14.04 ± 0.32	29,590	187.84
# Photos = 228–714	77,448	10.63 ± 0.23	91,877	193.57
Tourist = True	76,667	10.60 ± 0.24	91,214	197.79
Hour = 14h–15h	27,420	10.51 ± 0.25	33,028	194.40
Hour = 11h–12h	20,323	9.18 ± 0.21	24,613	196.99

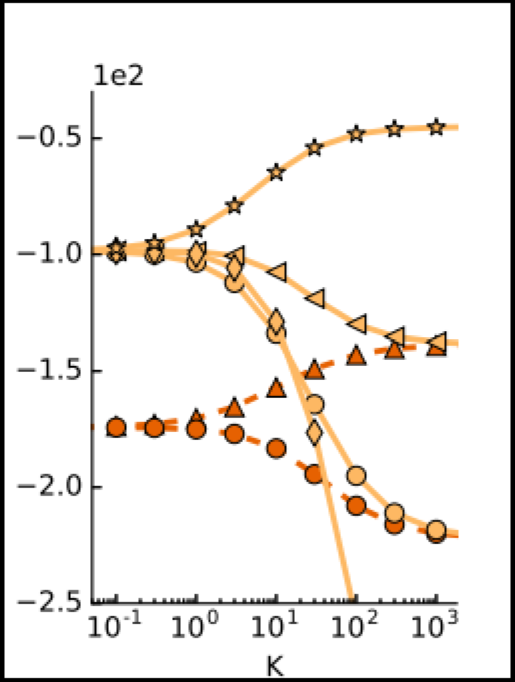
Results: Subtrails (Flickr)



(a) All transitions (b) Transitions of tourists

Mixed Trails

- Allow to specify different hypothesis for different parts of the data
- E.g., tourist go to PoI, non-tourists stay in their neighborhood
- Probabilistic assignment to groups



Summary & Outlook

Summary

- HypTrails:
  - A novel combination of methods
  - Try to explain underlying mechanisms that generate data
  - Bayesian hypothesis testing and ranking on sequential data
  - Easy and efficient to apply

- Example applications:
  - Flickr: explain sequences of locations a user took pictures
  - Taxi: explain destinations of taxi rides
  - Wikipedia: explain the popularity of links on a page

Image sources

- https://upload.wikimedia.org/wikipedia/commons/0/02/Dresden-Frauenkirche-night.jpg
- https://tu-dresden.de/++theme++tud.theme.webcms2/img/tud-logo.svg
- https://upload.wikimedia.org/wikipedia/commons/thumb/7/7a/Dresden-Germany-Main_station.jpg/290px-Dresden-Germany-Main_station.jpg
- https://upload.wikimedia.org/wikipedia/commons/a/a8/Aerial_view_of_the_Zwinger%2C_Dresden.jpg
- https://upload.wikimedia.org/wikipedia/commons/thumb/f/f0/Semperoper_at_night.jpg/220px-Semperoper_at_night.jpg
- Icons from pixabay