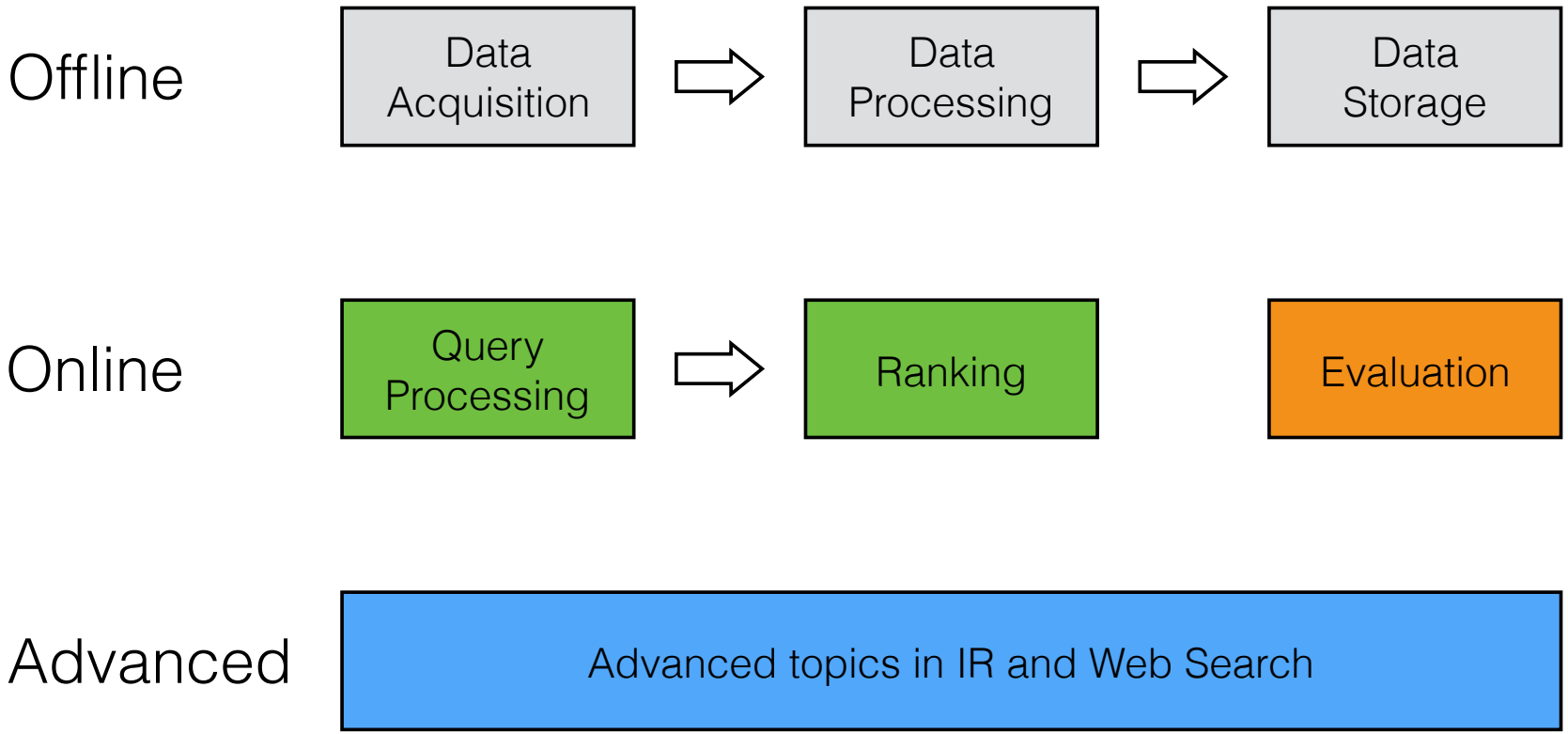
Information Retrieval

Click Models

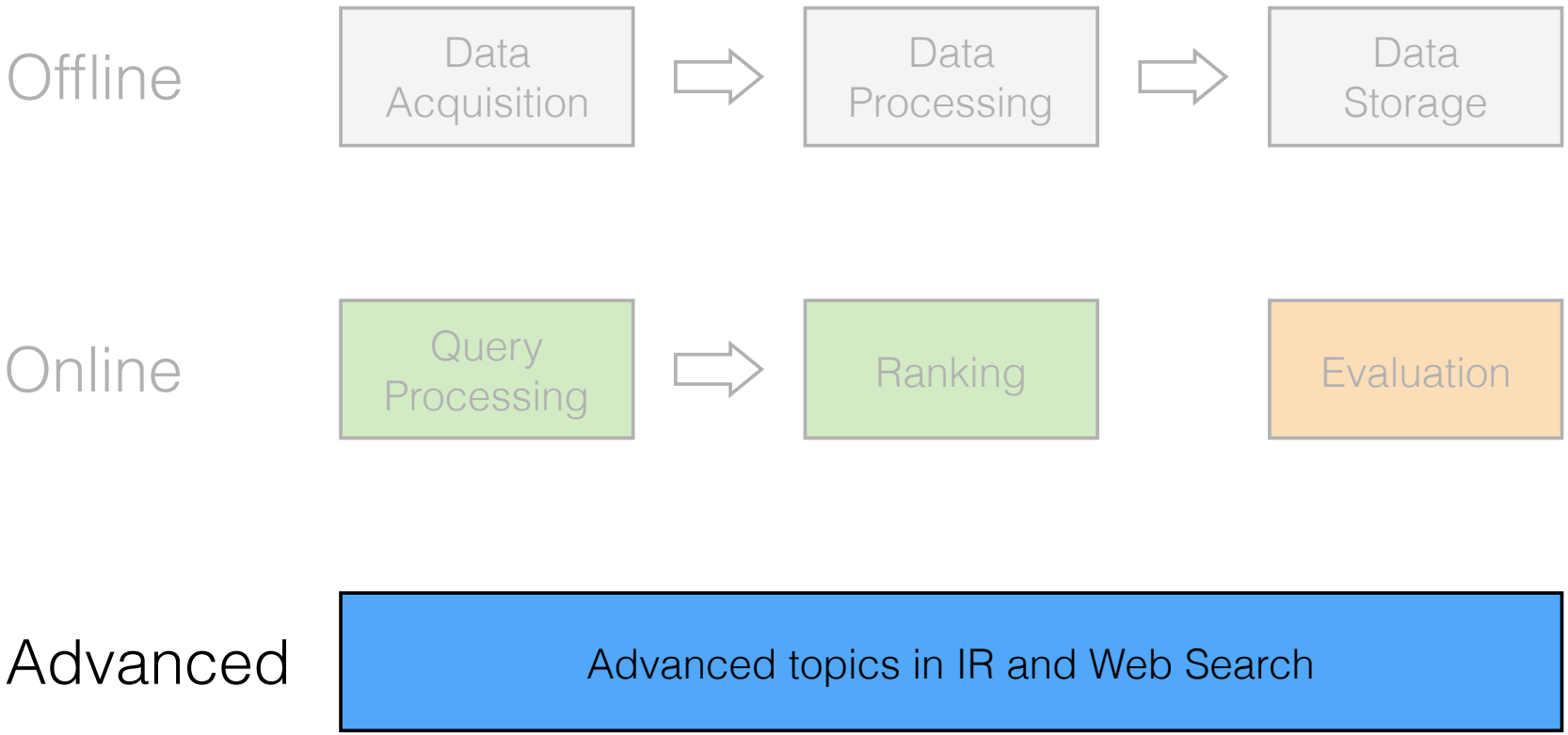
Ilya Markov
i.markov@uva.nl

University of Amsterdam

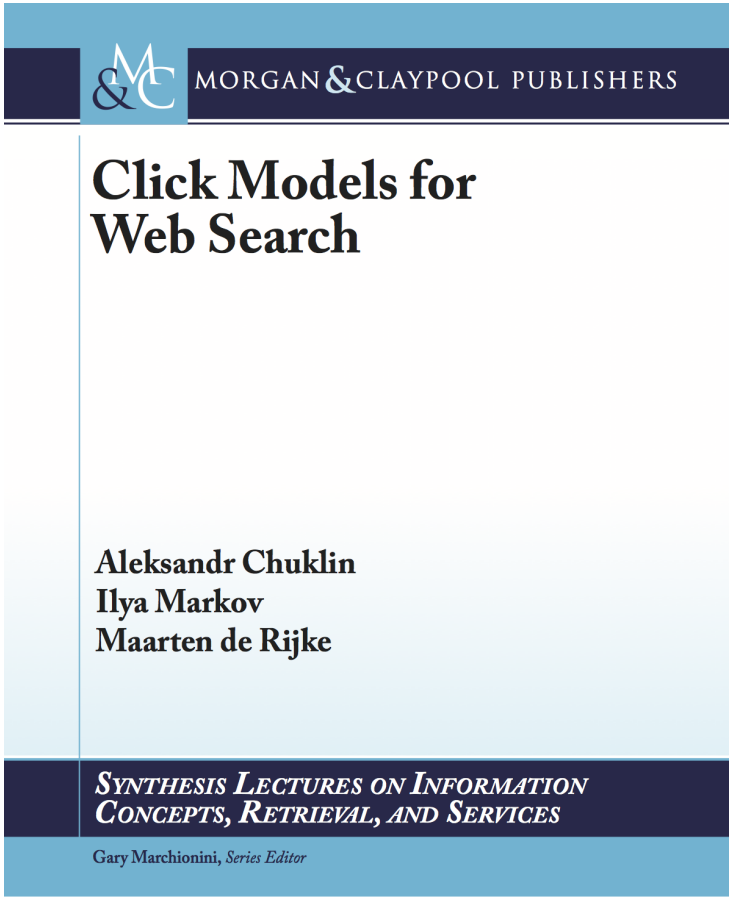
Course overview



Advanced topics in IR



The book



http://clickmodels.weebly.com/the-book.html

Tutorials

- SIGIR 2015, Santiago, Chile
- AINL-ISMW FRUCT 2015, St. Petersburg, Russia
- WSDM 2016, San Francisco, USA
- RuSSIR 2016, Saratov, Russia

`http://clickmodels.weebly.com/tutorials.html`

Outline

Outline

Incorporating clicks, attention and satisfaction

Incorporating clicks, attention and satisfaction into a search engine result page evaluation model

Aleksandr Chuklin, Maarten de Rijke

Proceedings of CIKM 2016, Indianapolis, USA

A neural click model for web search

A neural click model for web search

Alexey Borisov, Ilya Markov, Maarten de Rijke, Pavel Serdyukov

Proceedings of WWW 2016, Montreal, Canada

A context-aware time model for web search

A context-aware time model for web search

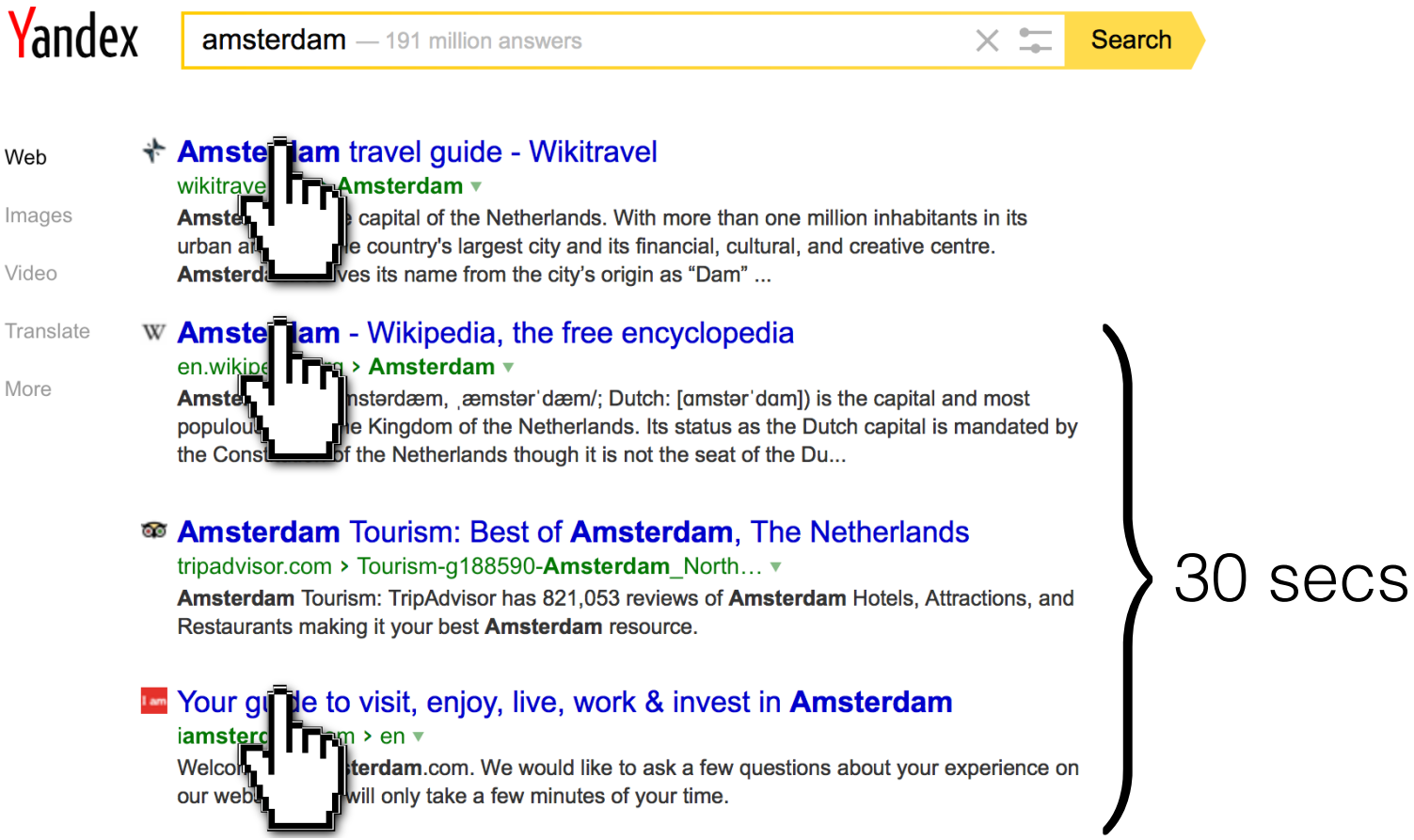
Alexey Borisov, Ilya Markov, Maarten de Rijke, Pavel Serdyukov

Proceedings of SIGIR 2016, Pisa, Italy
best student paper award

Time between user actions

- **Time between clicks**
- Time to first click
- Time to last click
- Time between queries

Time between clicks



there is *context bias*

Modeling time

- Average

$$Time(\text{“}Amsterdam\text{”}, \text{“}wikipedia.org\text{”}) = \frac{120 + 60 + 30}{3}$$

- Probability distribution

$$Time(\text{“}Amsterdam\text{”}, \text{“}wikipedia.org\text{”}) \sim Gamma(\boldsymbol{k}, \boldsymbol{\theta})$$

where $(\boldsymbol{k}, \boldsymbol{\theta})$ are estimated from 120, 60, 30

context bias is not modeled

Context-aware time modeling

$$Time(\text{“}Amsterdam\text{”}, \text{“}wikipedia.org\text{”}, context_1) \sim Gamma(\boldsymbol{k_1}, \boldsymbol{\theta_1})$$
$$Time(\text{“}Amsterdam\text{”}, \text{“}wikipedia.org\text{”}, context_2) \sim Gamma(\boldsymbol{k_2}, \boldsymbol{\theta_2})$$
$$Time(\underbrace{\text{“}Amsterdam\text{”}, \text{“}wikipedia.org\text{”}}_{\text{user action}}, \underbrace{context_3}_{\text{context}}) \sim Gamma(\boldsymbol{k_3}, \boldsymbol{\theta_3})$$

Context-aware time modeling

$$Time(action, context) \sim Gamma(\\ \boldsymbol{a}_k(ctx) \cdot \boldsymbol{k}(act) + \boldsymbol{b}_k(ctx), \\ \boldsymbol{a}_\theta(ctx) \cdot \boldsymbol{\theta}(act) + \boldsymbol{b}_\theta(ctx))$$

Parameter estimation

$$Time(action, context) \sim Gamma(\boldsymbol{a}_k(ctx) \cdot \boldsymbol{k}(act) + \boldsymbol{b}_k(ctx), \boldsymbol{a}_\theta(ctx) \cdot \boldsymbol{\theta}(act) + \boldsymbol{b}_\theta(ctx))$$

1. Fix context-independent parameters
2. Optimize context-dependent parameters using *neural networks*
3. Fix context-dependent parameters
4. Optimize context-independent using *gradient descent*
5. Repeat until convergence

Parameter estimation

- We do not know the form of context-dependent parameters $\Longrightarrow$ neural networks
- We know the form of context-independent parameters (Gamma distribution) $\Longrightarrow$ direct optimization

Context

General	
Is query (Q-action)	(0: no, 1: yes)
Is click (C-action)	(0: no, 1: yes)
$\log(1 + \text{observed time since previous action})$	(0: undefined)
$\log(1 + \text{average time since previous action})$	(0: undefined)
Q-action	
Is new search session	(0: no, 1: yes)
Number of terms in issued query	(0: undefined)
BM25 (issued query, previous query)	(0: undefined)
BM25 (previous query, issued query)	(0: undefined)
C-action	
Is click on the 1^{st} position	(0: no, 1: yes)
...	...
Is click on the 10^{th} position	(0: no, 1: yes)

Dataset

3 months of log data from Yandex search engine

Time between actions	Max time	# Observations
Time-to-first-click	1 min	30,747,733
Time-between-clicks	5 min	6,317,834
Time-to-last-click	5 min	30,446,973
Time-from-abandoned-query	1 min	11,523,351

Evaluation tasks

Task1. Predict time between clicks
- Log-likelihood
- Root mean squared error (MSE)

Task2. Rank results based on time between clicks
- nDCG@$\{1, 3, 5, 10\}$

Task 1. Predicting time

Time model	Distribution	Log-likelihood	RMSE
Basic	exponential	−4.9219	60.73
	gamma	−4.9105	60.76
	Weibull	−4.9077	60.76
Context-aware	exponential	−4.8787	58.93
	gamma	−4.8556	58.98
	Weibull	−4.8504	58.94

Task 2. Ranking results

		NDCG			
Time model	Distribution	@1	@3	@5	@10
Average	—	0.651	0.693	0.728	0.812
Context-aware	exponential	0.668	0.710	0.743	0.820
	gamma	0.675	0.715	0.748	0.822
	Weibull	0.671	0.709	0.745	0.821

Other times

- Time to first click
- Time to last click
- Time between queries

Summary

- Removed context bias from time between actions
- Predicted user search interactions better (**Task 1**)
- Used the context-independent component for better document ranking (**Task 2**)

Outline

Future research

- Keep on adding new variables – not a good idea
- Parameter estimation
  - Efficiency
  - Online learning
- Other interactions and environments
  - Interactions beyond clicks
  - Devices beyond desktop computers

Future research

Model's output	Evaluation	Application
Conditional click probs	Log-likelihood	User simulation
Full click probs	Perplexity	Model-based metrics
Parameter values	Ranking evaluation	Ranking

- Why use intermediate evaluation?
  - Evaluate applications, not models
- Why maximize log-likelihood?
  - Optimize models for specific applications

Outline

Outline

Materials

- Aleksandr Chuklin, Ilya Markov, Maarten and de Rijke
  Click Models for Web Search
  Morgan & Claypool, 2015

Advanced topics in IR