LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Understanding and Predicting *Client-side* User Clickstream

Ou Changkun

hi@changkun.us

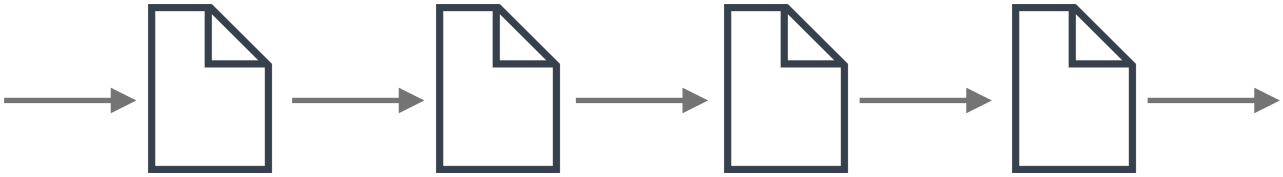
Masters Defense Presentation, LMU | Munich, Germany | 2018-01-08

Introduction

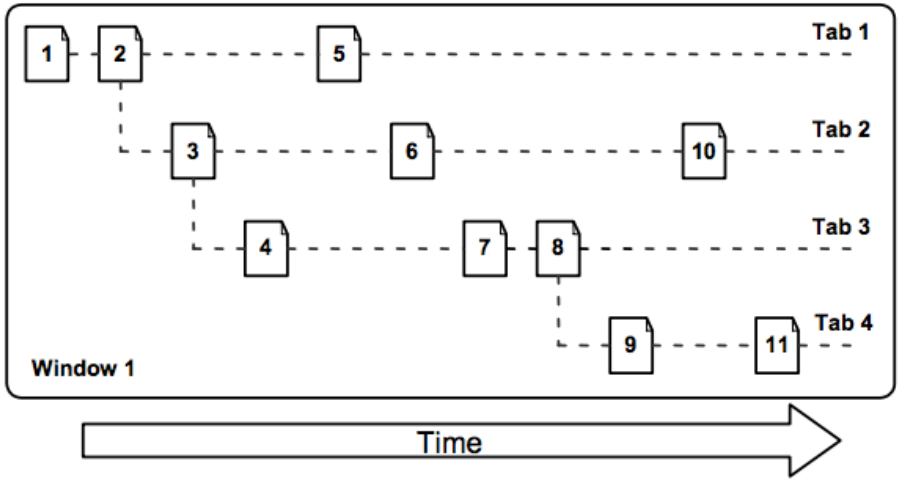
“Hello world!”

Background & Motivation

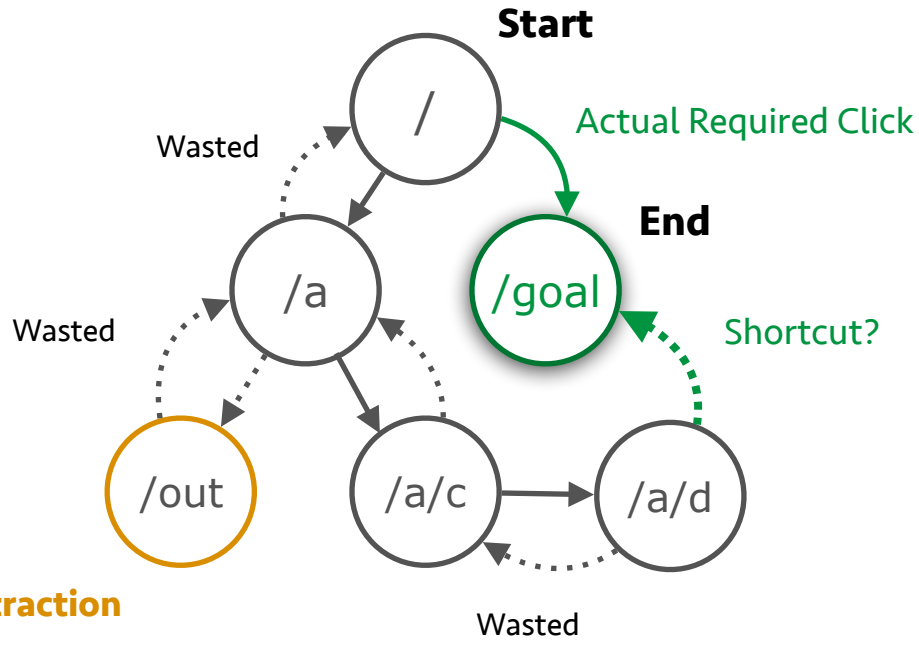
- Clickstream research: e.g. Privacy issue[SKOK, 1999]; Compromise Detection [WELLER, 2018], etc.



"Clickstream" [FRIEDMAN, 1995]



branching[HUANG, 2010] and backtrace[HUANG, 2012]



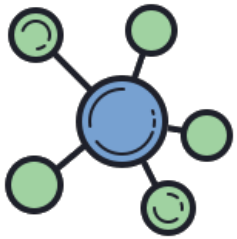
Research Questions

1. **Understanding**:
   - Why collecting clickstream on client-side differs from server-side collecting?
   - What are the most significant, identifiable user behaviors or activity patterns can be observed or detected in the context of web browsing that indicates information needs,
   - in which form of quantitative data can characterize a definitive boundary to distinguish browsing behaviors of a user?

2. **Classification**:
   - How accurate or how affirmative we can model or identify the proposed browsing behaviors progressively that makes an intelligent system serves proactively?
3. **Prediction**:
   - How much future movements of a user can be accurately inferred from the context of web browsing, and how much context is required for the prediction?

Agenda

Approaches



Evaluations

Applications

Approaches

"Stop talking. Just coding."

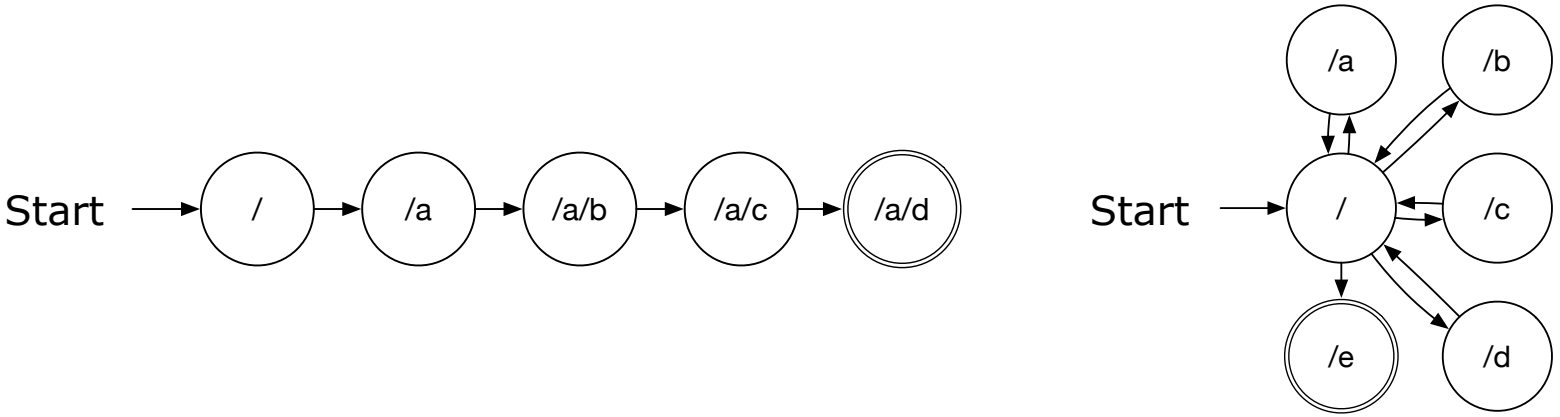
Data

- **An action path (URL+Duration):**

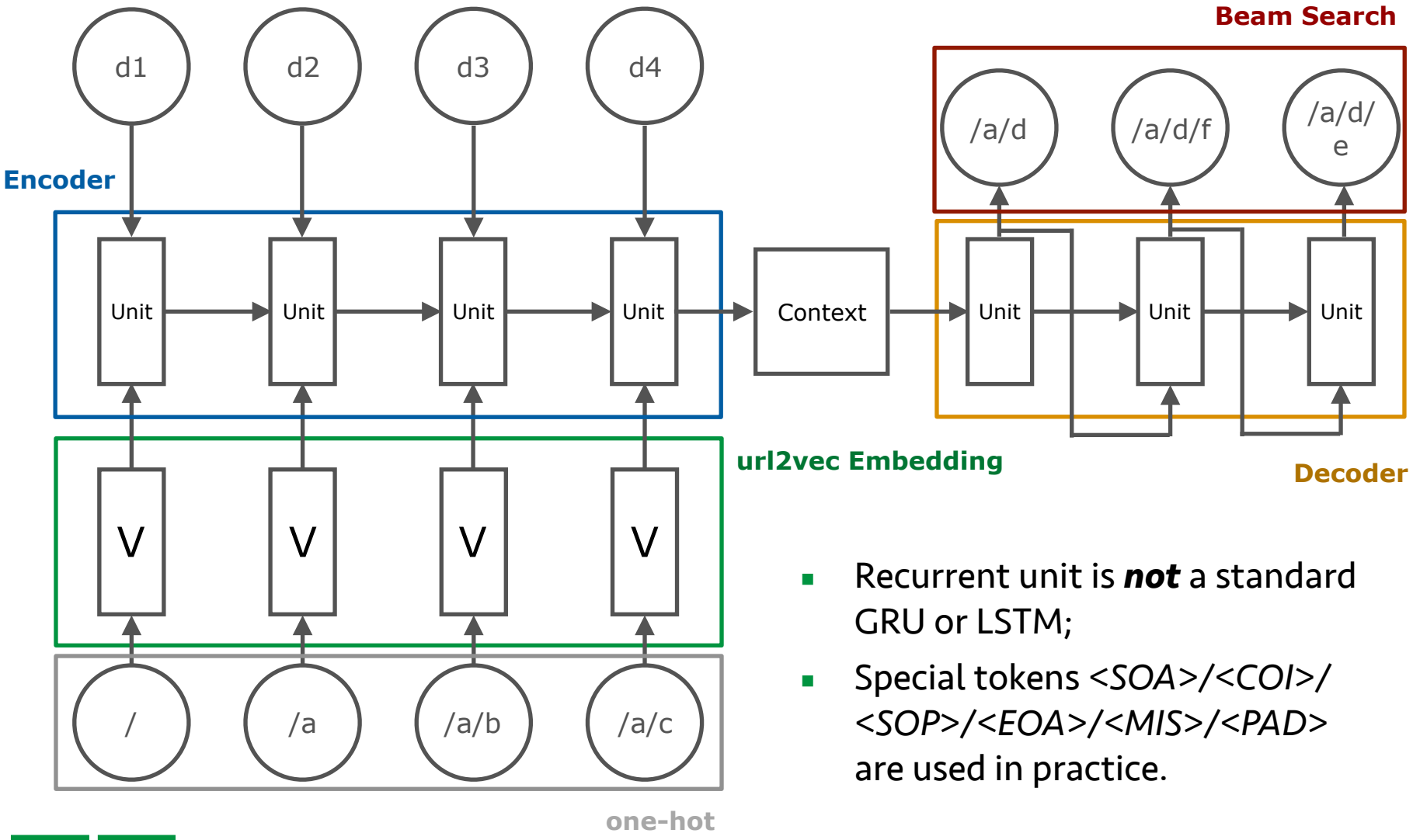
$$(\mathrm{URL}_1, d_1, \mathrm{URL}_2, d_2, ..., \mathrm{URL}_n, d_n)$$

where *n* is the total number of action in a browsing session.

- **Special cases:**



Model: *Action-Path* Model Overview

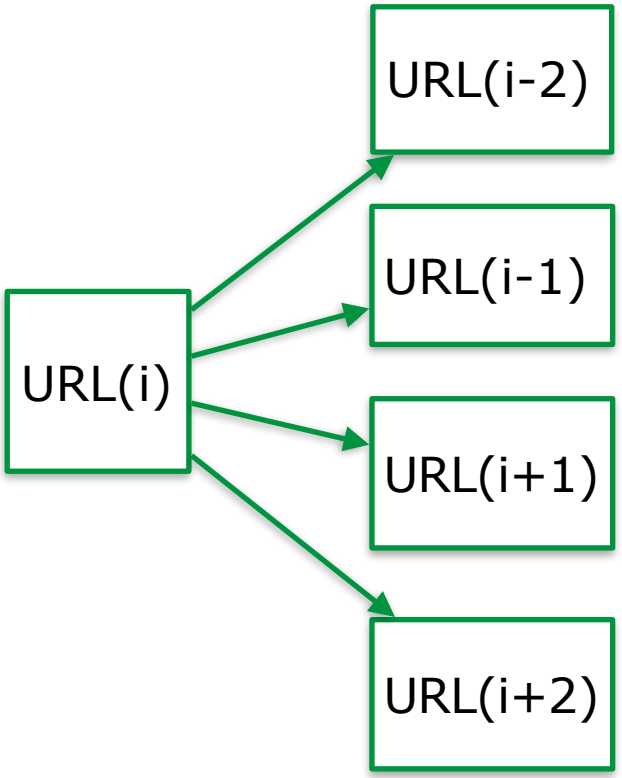


- Recurrent unit is ***not*** a standard GRU or LSTM;
- Special tokens *<SOA>/<COI>/<SOP>/<EOA>/<MIS>/<PAD>* are used in practice.

Model: *url2vec* Embedding

- *url2vec* construct URL presentation to better predict surrounding URLs

$$p(\mathrm{URL}_{t+i}|\mathrm{URL}_t) = \frac{\exp(v_{\mathrm{URL}_{t+i}}^{\top} v_{\mathrm{URL}_t})}{\sum_{i \in \text{all URLs}} \exp(v_i^{\top} v_{\mathrm{URL}_t})}$$



Model: *Recurrent Unit*

- Improved from **GRU or LSTM**;
- Feed not only URL embeddings but also stay duration

LSTM-like:

$$
\begin{aligned}
I_t &= \sigma(P^{(I)}U_t^{ij} + Q^{(I)}h_{t-1} + \boxed{\frac{d_t^{ij}}{d_t^{ij}+1}}) \\
F_t &= \sigma(P^{(F)}U_t^{ij} + Q^{(F)}h_{t-1} + b^{(F)}) \\
O_t &= \sigma(P^{(O)}U_t^{ij} + Q^{(O)}h_{t-1}) \\
C_t &= F^{(t)} \circ C_{t-1} + I_t \circ \tanh(P^{(C)}U_t^{ij} + Q^{(C)}h_{t-1}) \\
h_t &= O_t \circ \tanh(C_t)
\end{aligned}
$$

squashing

GRU-like:

$$
\begin{aligned}
Z_t &= \sigma(P^{(Z)}U_t^{ij} + Q^{(Z)}h_{t-1} + \boxed{\frac{d_t^{ij}}{d_t^{ij}+1}}) \\
R_t &= \sigma(P^{(R)}U_t^{ij} + Q^{(R)}h_{t-1}) \\
h_t &= (1 - Z_t) \circ \tanh(P^{(H)}U_t^{ij} + Q^{(H)}h_{t-1}) + Z_t \circ h_{t-1}
\end{aligned}
$$

squashing

User Study Design

"Data is priceless."

Information Behavior Theory on the Web

- Comparison (All based on Wilson's theory [WILSON, 1997] and Ellis's Model [ELLIS, 1989]):

Author	Terminologies	Terminologies	Terminologies	Main Factor
[CHOO et al, 1999]	Formal search	Conditioned viewing; Informal search	Undirected viewing	Moves of information seeking
[JOHNSON, 2017]	Directed browsing Known-item search	Semi-directed browsing Explorative seeking "You don't know what you need" Re-finding	Undirected browsing	Actions; Objectives
This thesis	**Goal-oriented**	**Fuzzy**	**Exploring**	**Information needs (purpose); Chaining; Differentiating, etc.**

- We conservatively grouped three distinguishable (evidence shows later) behaviors:
  - **Goal-oriented**: browsing on the web deterministically with systematically specified purpose
  - **Exploring**: browsing aimlessly without information needs and use
  - **Fuzzy**: "chaining" without "differentiating" from "starting"

Task Design & Collected Data

- 5~10 min/task, no visit restriction
- **9 tasks** from 35 tasks, **simulate** three browsing behavior.

Starting Point	Goal-oriented task	Fuzzy task	Exploring task
www.amazon.com	Assume your smartphone was broken and you have **1200 euros as your budget**. You want to buy **an iPhone**, a **protection case**, and a **wireless charging dock**. Look for these items and add them to your cart.	You want to buy **birthday present(s)** for **your best friend** Add **three** items to your cart.	Look for a product category that you are interested in and start browsing. Add **three** items to your cart that you would like to buy.

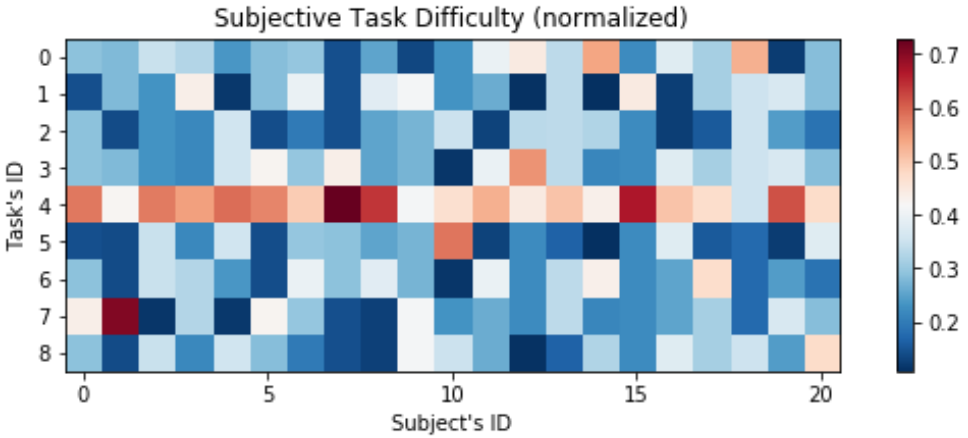
- Equipment of experiment
  - Software: Chrome; Hardware: Desktop & Laptop
  - Latin square
- 21 subjects, 189 action paths (clickstreams)
  - Age & Gender: min=18, max=29, median=23, mean=23.04, SD=3.22, male=10, female=11

Evaluations

"History tells our future."

Subjective Tasks Difficulty Score

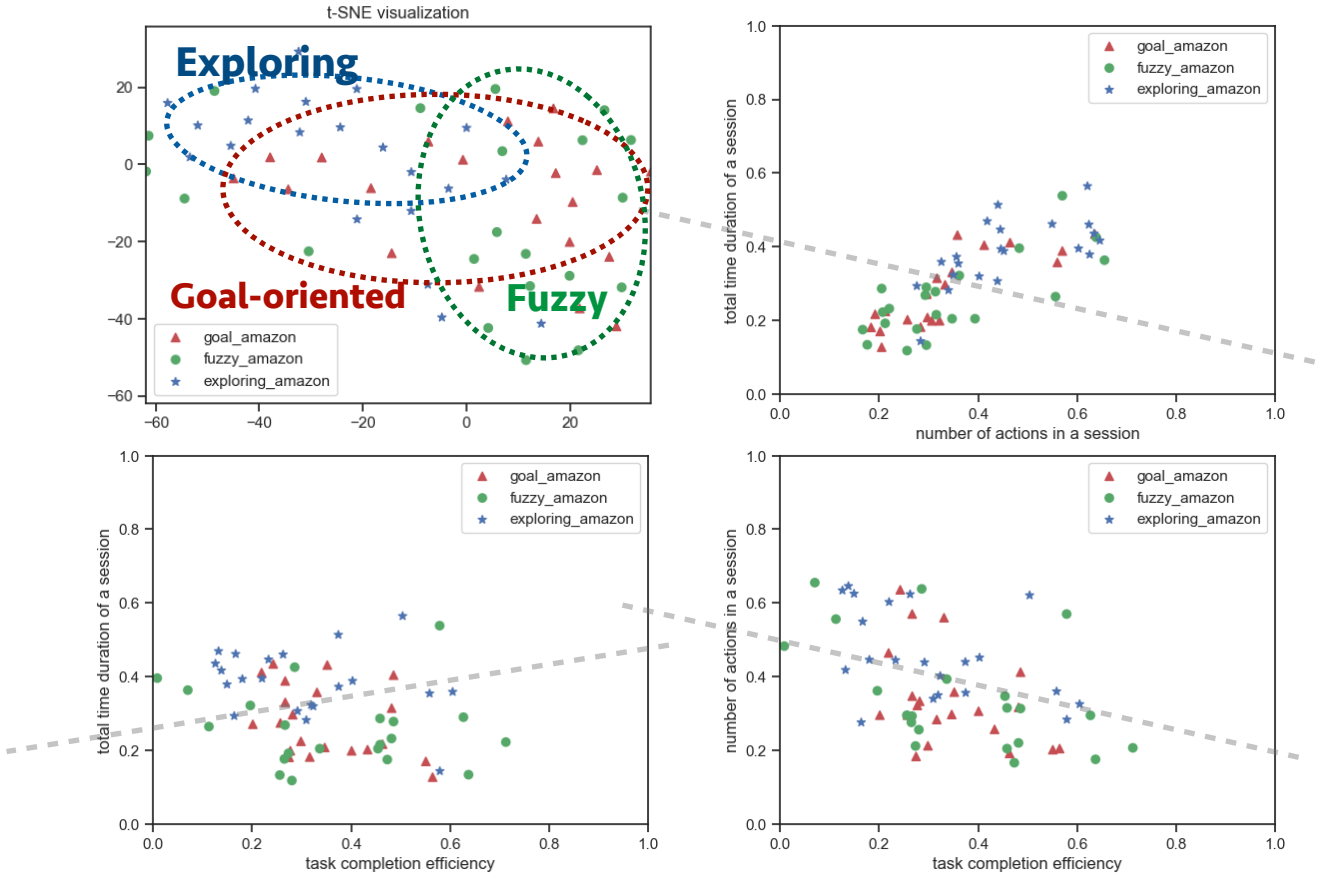
- Subjective difficulty score from participants



- One-tailed Mann-Whitney U test

- H0: the difficulty of fuzzy task is not greater than exploring task, $p < 0.05$, reject H0.
- For **difficulty:**
  - **Fuzzy task > Goal-oriented task > Exploring task**

“Efficiency”, Length, Duration: Classification?



- Feature Importance (RDT approach): **Length > Duration > Efficiency**
- One-tailed Mann-Whitney U test:
  - **completion efficiency: {Goal,Fuzzy} > Exploring**
  - **browsing action length: {Fuzzy,Exploring} > Goal**
  - **total stay duration: Exploring > {Goal, Fuzzy}**

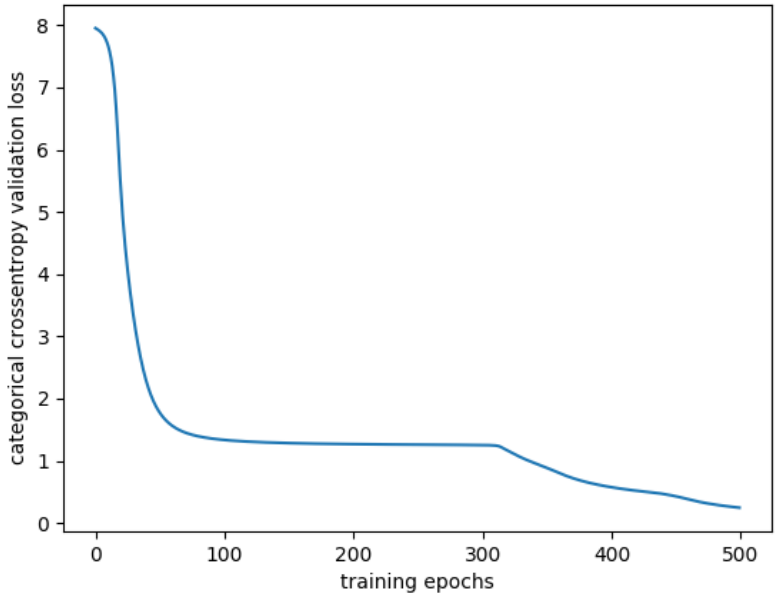
Model Performance: on Prediction and Classification

- One-layer model
  - 90323 params, 1500 epochs, 10 latent dim, 32 batch size, Adam optimizer
  - categorical cross-entropy loss, L1&L2 regularizer in decoder with early stopping
  - Total samples: 189, train/val/test: 132/38/19 (0.7/0.2/0.1 ratio)
- **Split ratio >0.9 achieves** excellent performance, **>60% accurate (still can be optimized)**
- Best prediction steps = average length of sequence * 0.95, **in our tiny dataset ≈ 3~5 steps**

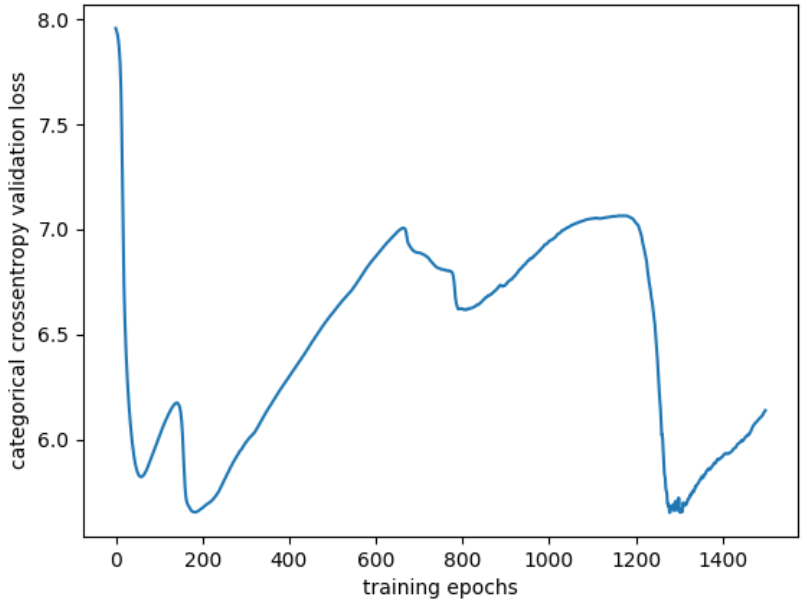
Split ratio: len(input)/len(input+output)

Model Performance: More Observations

- **Classification** is a **special case** of prediction in the model (**predict last step**)
  - Ending marks as classes: <EOA_GOAL>, <EOA_FUZZY>, <EOA_EXPLORE>
  - **Classification accuracy: ~100.0%!**
  - **Prediction accuracy: >60%**
- Validation loss:

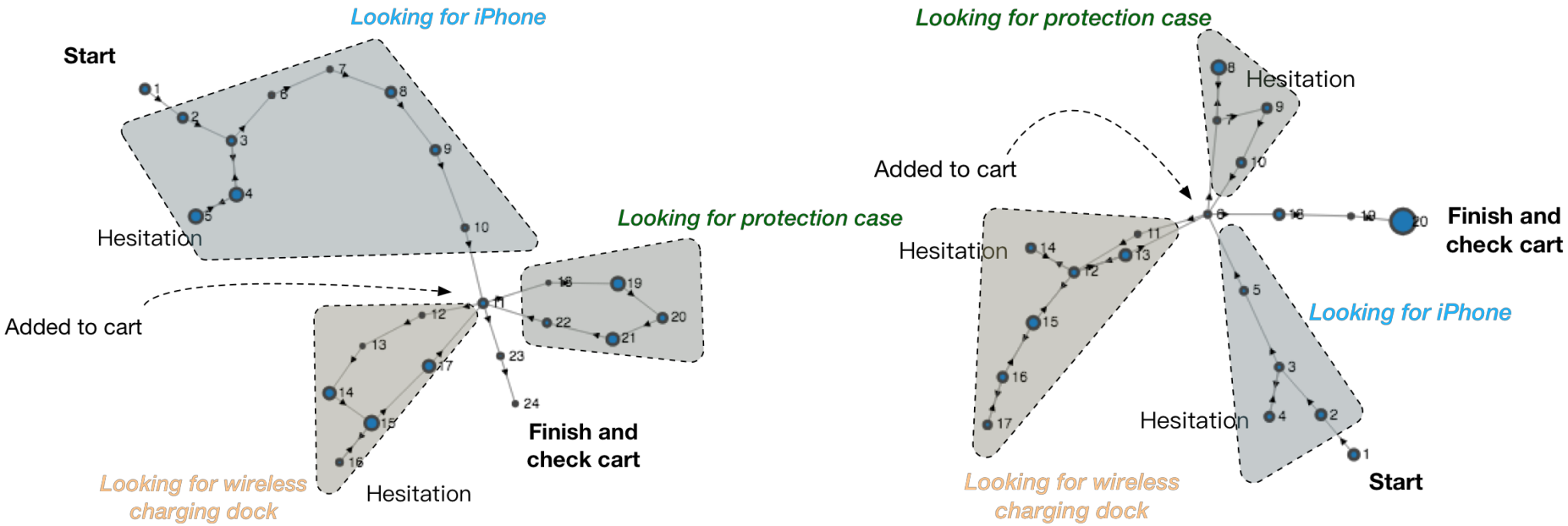


Classification



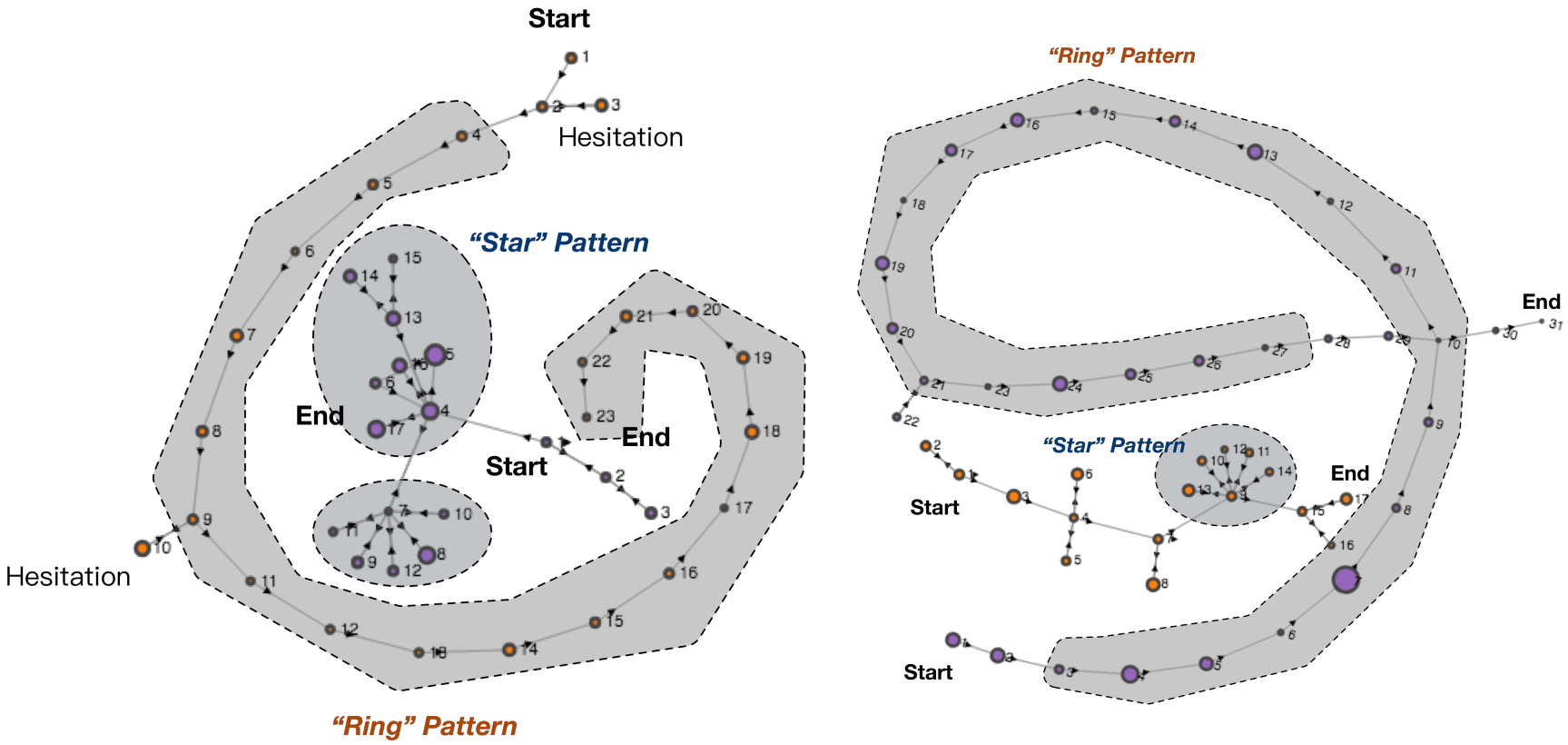
3 future steps prediction

Patterns: Goal-oriented Tasks



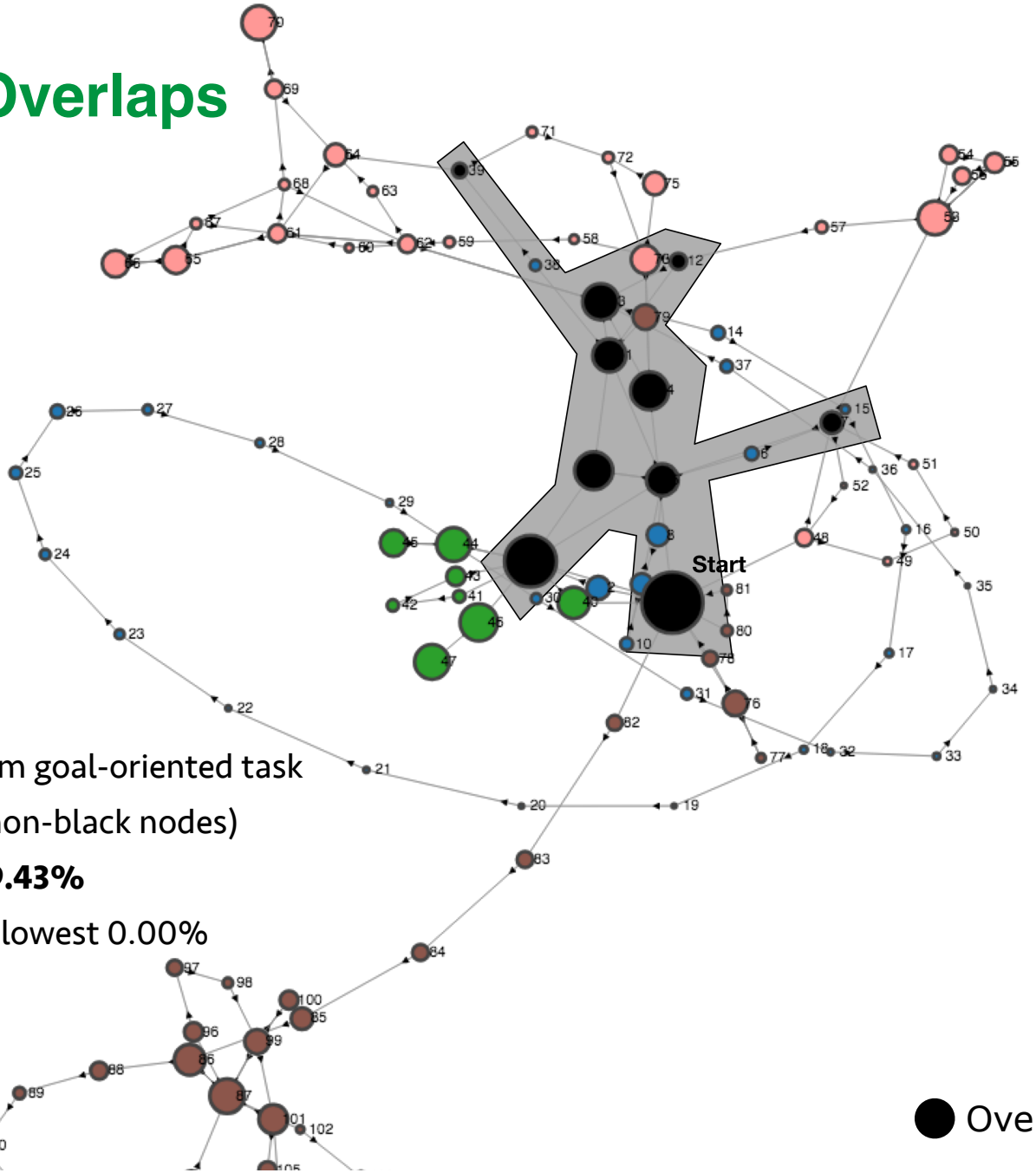
Example: Amazon's goal oriented task

Patterns: Fuzzy & Exploring Tasks



Example: Two participants, Amazon and Dribbble's fuzzy and exploring task

Patterns: Overlaps

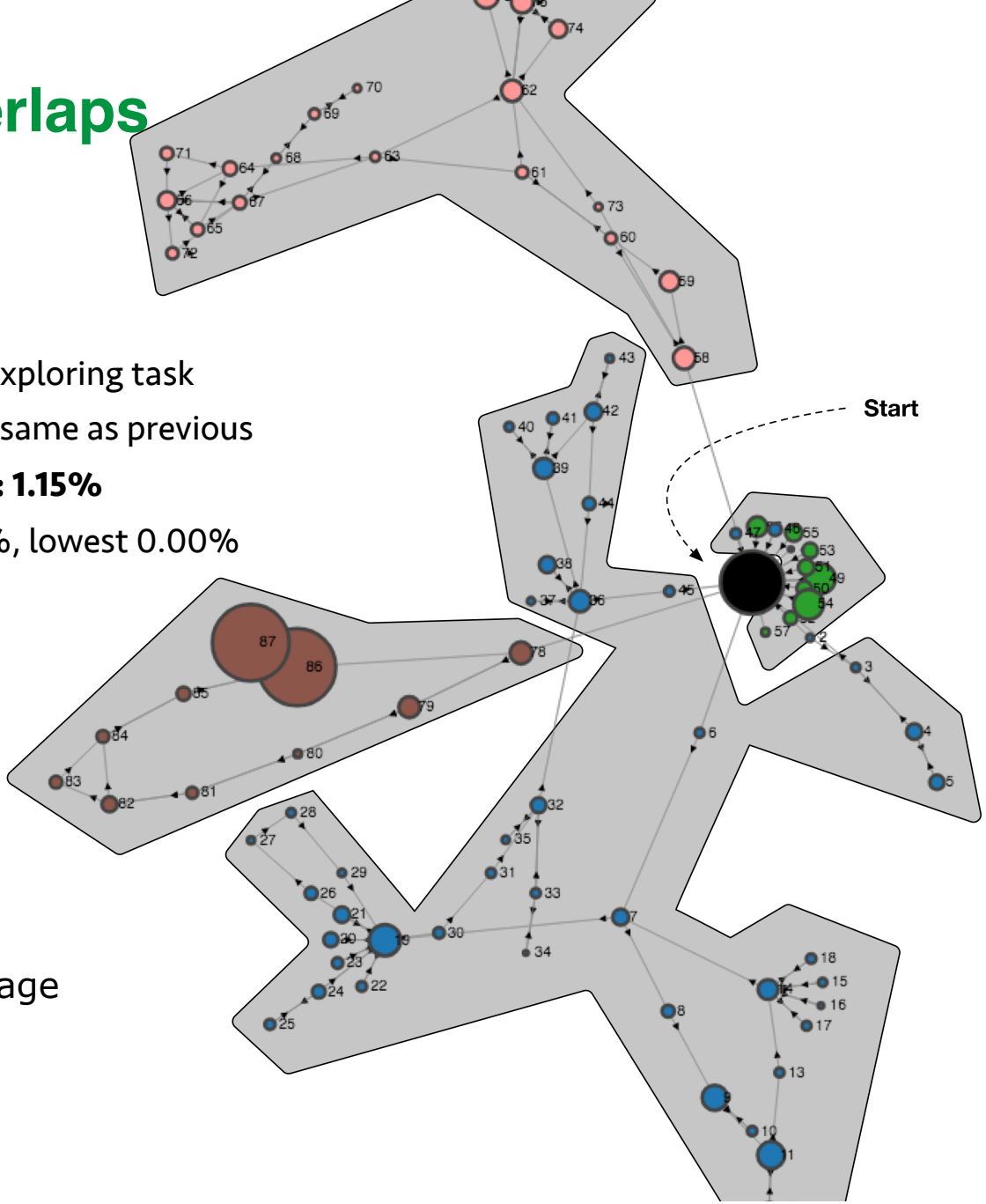


- Example: Medium goal-oriented task
- 4 participants (non-black nodes)
- **Overlap ratio: 9.43%**
- Highest 11.84%, lowest 0.00%

● Overlapped page

Patterns: Overlaps

- Example: Dribbble exploring task
  - 4 participants same as previous
  - **Overlap ratio: 1.15%**
  - Highest 11.84%, lowest 0.00%

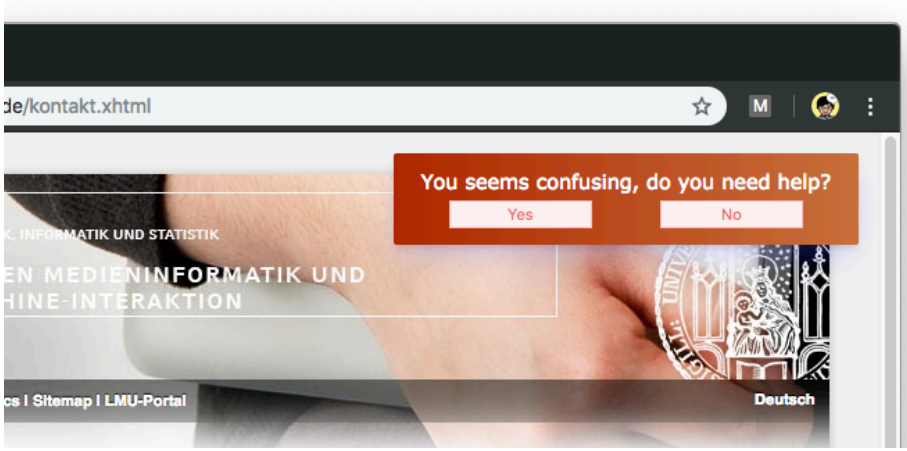


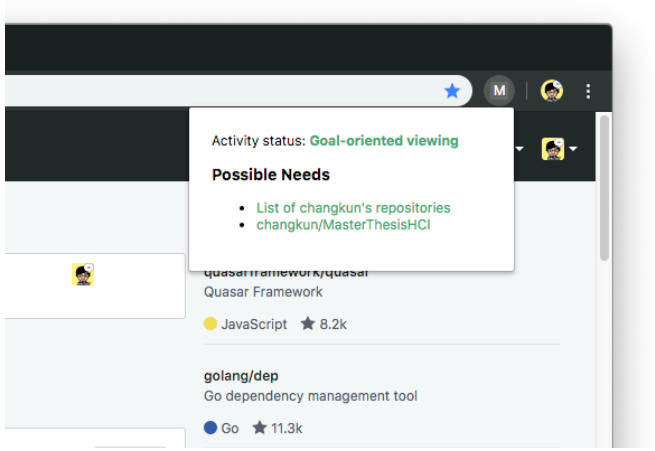
● Overlapped page

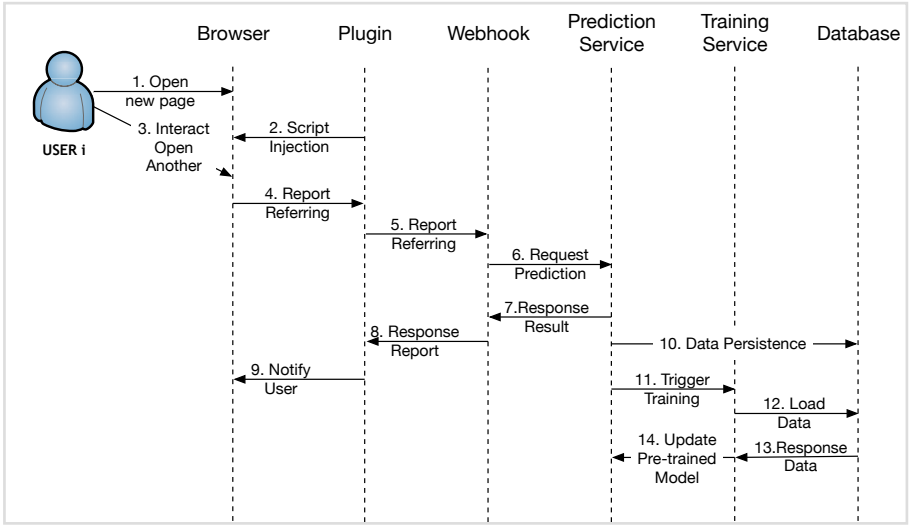
Applications

"Let's make a story"

Possible Applications: Browser Plugin & Standardization

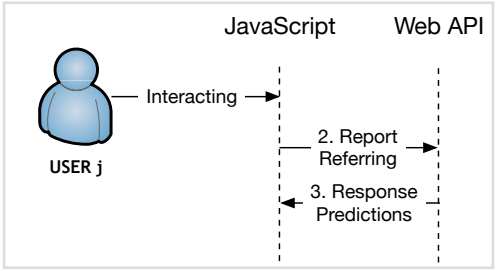






Standardize

As Browser Web APIs



Conclusion

"To the higher truth."

Summary

- A study regarding **action path,** a.k.a. *client-side* clickstreams; contributes to:
  - Understanding of browsing behavior on client-side:
    - Three classifiable behavior: **goal-oriented**, **fuzzy**, and **exploring**
    - **Historical URLs** & **stay duration** are the most important features
    - Browsing intents tend to distributed and clustered individually
    - Browsing behavior is **not user-specific** but intersection
    - "cluster"/"hesitation"/"ring"/"star"/"overlaps" patterns
    - Goal-oriented behavior tent to overlap because no overlaps
  - One model to learn them all (Classification & Accuracy):
    - Browsing behavior **classification**: **~100.0% accurate**
    - **Page-level** 5 future steps universal **prediction**: **>60% accurate**
  - Possible Applications:
    - As browser plugin proactively improves user actions
    - As standard API helps developers improve their product
- Limitations & Future works:
  - Privacy & Trustiness & Security Issues
  - More data for more precise behavior categories
  - Performance optimization & & more applications

changkun.de/master.pdf

References

- [FRIEDMAN, 1995]: Friedman, Wayne and Weaver, Jane. Calculating cyberspace: tracking “clickstreams.”. February 1995.
- [SKOK, 1999]: Skok, Gavin. "Establishing a legitimate expectation of privacy in clickstream data." Mich. Telecomm. & Tech. L. Rev. 6 (1999): 61.
- [WALSH et al., 2000] Walsh, John, and Sue Godfrey. "The Internet: a new era in customer service." European Management Journal 18.1 (2000): 85-92.
- [SCHONBERG. 2000] Schonberg, Edith, et al. "Measuring success." Communications of the ACM 43.8 (2000): 53-57.
- [MOBASHER et al., 2001] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. 2001. Effective personalization based on association rule discovery from web usage data. In Proceedings of the 3rd international workshop on Web information and data management (WIDM '01). ACM, New York, NY, USA, 9-15.
- [WATERSON et al, 2002] Sarah Waterson, James A. Landay, and Tara Matthews. 2002. In the lab and out in the wild: remote web usability testing for mobile devices. In CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02). ACM, New York, NY, USA, 796-797.
- [SARAH, 2002] Sarah J. Waterson, Jason I. Hong, Tim Sohn, James A. Landay, Jeffrey Heer, and Tara Matthews. 2002. What did they do? understanding clickstreams with the WebQuilt visualization system. In Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '02), Maria De Marsico, Stefano Levialdi, and Emanuele Panizzi (Eds.). ACM, New York, NY, USA, 94-102.
- [ANALIA et al., 2006] Anália G. Lourenço and Orlando O. Belo. 2006. Catching web crawlers in the act. In Proceedings of the 6th international conference on Web engineering (ICWE '06). ACM, New York, NY, USA, 265-272.
- [SCHEIDER et al., 2009] Fabian Schneider, Anja Feldmann, Balachander Krishnamurthy, and Walter Willinger. 2009. Understanding online social network usage from a network perspective. In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement (IMC '09). ACM, New York, NY, USA, 35-48.
- [MEIER et al., 2016] Florian Meier and David Elsweiler. 2016. Going back in Time: An Investigation of Social Media Re-finding. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '16).
- [WANG et al., 2016] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y. Zhao. 2016. Unsupervised Clickstream Clustering for User Behavior Analysis. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). ACM, New York, NY, USA, 225-236.
- [WELLER, 2018] Tobias Weller. 2018. Compromised Account Detection Based on Clickstream Data. In Companion Proceedings of the The Web Conference 2018(WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 819-823.
- .... And More!

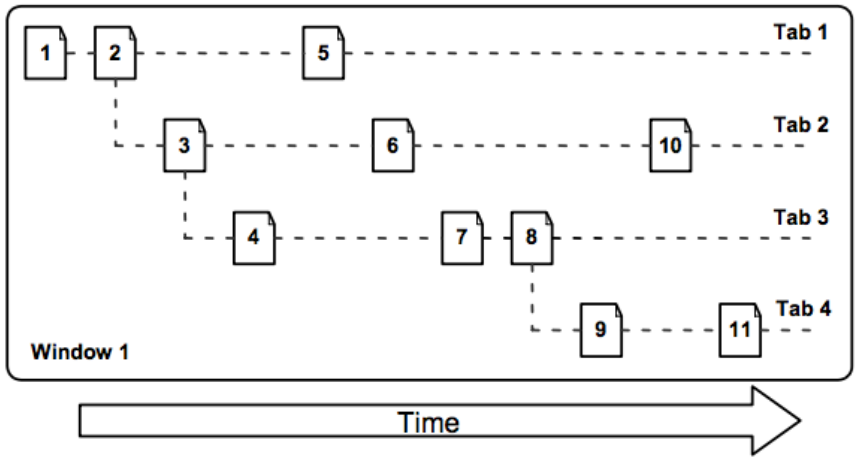
BACKSTAGE

"You found a secret place."

A Brief History of "Clickstream"

- **A clickstream [FRIEDMAN, 1995] (informally) contains a sequence of hyperlinks clicked by users on the web over time.** Early clickstream research emerges for privacy discussion [SKOK, 1999], customer services [WALSH et al., 2000], business decisions[SCHONBERG, 2000], personalization web services [MOBASHER et al., 2001], remote usability testing [WATERSON et al., 2002],etc.
- Recent works take clickstream on web user behavior analysis for social characterizing [SCHEIDER et al., 2009], media re-finding [MEIER et al., 2016], user clustering[WANG et al., 2016], account security [WELLER, 2018] and more…
- Closest research to ours is branching[HUANG, 2010] and backtrace[HUANG, 2012]**, but still a server side analysis and individually analyzed.**

- Questions:
  - Making difference if collect from user side?
  - Branching & backtrace effect?
  - Matters to user?
  - Benefits to user?



For business and on server side:
+ branching [HUANG, 2010]
+ backtrace [HUANG, 2012]

Ref. #ch 4

Information Behavior Theory

- Information behavior theory developed by T.Wilson [WILSON, 1981] and evolved many major versions [WILSON, 1997, 2000, 2010] for general purpose.
- Four information behavior [CHOO et al. 1999] on the Web are discussed for information need, information seeking and information use and uses Ellis' Model [ELLIS, 1989].

Author	Passive Attention	Passive Search	Active Search	Ongoing Search
[WILSON, 1997]	No info-seeking intended, but acquisition take place	Occasionally relevant to individuals	Actively seeking	Update basic framework of ideas
[CHOO et al. 1999]	Undirected viewing	Conditioned viewing	Formal search	Informal search

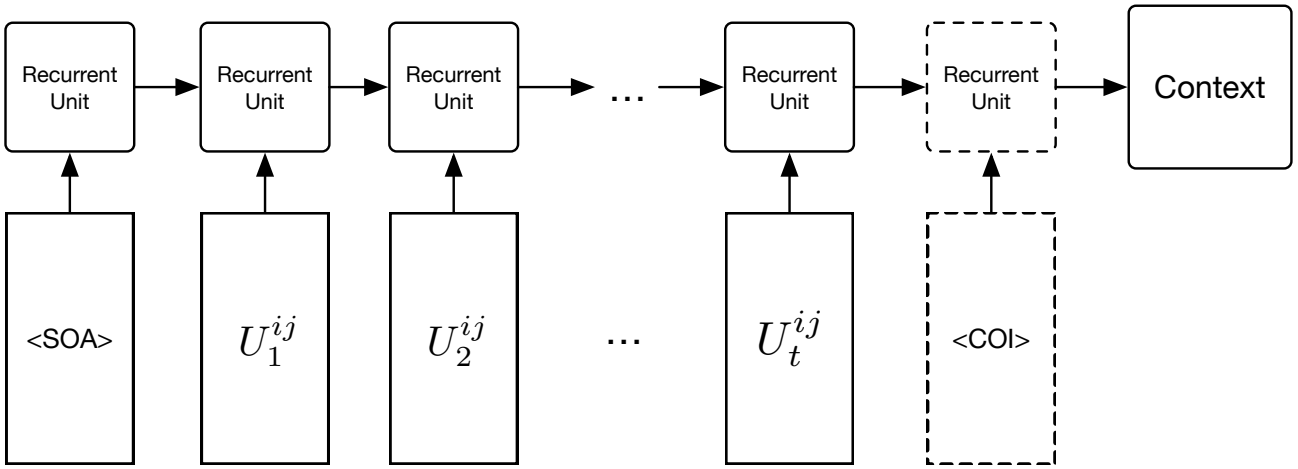
- More browsing patterns are discussed [JOHNSON, 2017], and we selected three categories:
  - Directed browsing
  - Semi-directed browsing
  - Undirected browsing
  - Known-item search
  - Exploratory seeking
  - "You-don't-know-what-you-need"
  - Re-finding

Analysis based on Wilson and Ellis' Model

Behaviors / Patterns	Information Need	Information Seeking						Information Use
		Starting	Chaining	Browsing	Differentiating	Monitoring	Extracting	
Goal-oriented		Yes						
Fuzzy		Yes						
Exploring		Yes						
"cluster"	Observed				Yes			
"star"			Yes					
"ring"		Yes						
"hesitation"	Observed				Yes			
"overlap"	Observed						Yes	

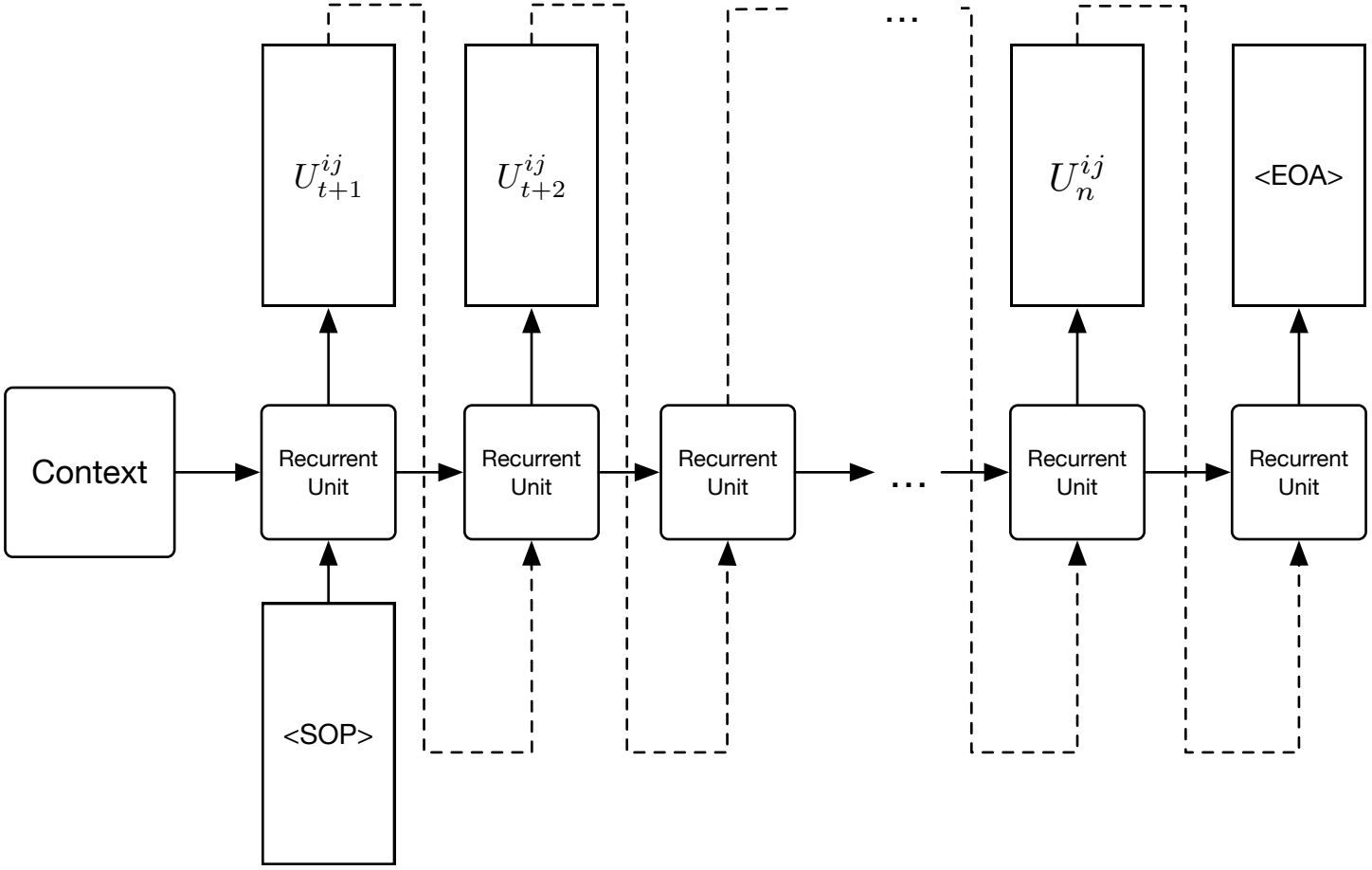
Ref. #ch 3.4.1

Model: *Context Encoder*



Ref. #ch 3.4.1

Model: *Context Decoder*



Task Design: Collect Data from User

- Participants are allowed to visit any pages during the task, even out side the starting point domain.
- 5~10 minutes for each of the task, 80min in total.

Starting Point	Goal-oriented task	Fuzzy task	Exploring task
www.amazon.com	Assume your smartphone was broken and you have **1200 euros as your budget**. You want to buy **an iPhone**, a **protection case**, and a **wireless charging dock**. Look for these items and add them to your cart.	You want to buy **a gift** for **your best friend** as a birthday present.. Add three items to your cart.	Look for a product category that you are interested in and start browsing. Add any items to your cart that you would like to buy.
www.medium.com	Assume you were making plans for your **summer vacation**. You want to **visit Tokyo, Kyoto, and Osaka**. You want to find out what kind of experience other people made when traveling to these three places in Japan. Your task is to **find three posts for traveling tips** regarding these cities. **Elevate (👏)** a post if it is one of your choices.	Assume you got an occasion to **visit China for business**. You are **free to travel** to China for **a week**. You want to make a travel plan for touring China within a week. Your task is to find out what kind of experience other how people made when going to secondary cities or towns in China, then decide on three cities you want to visit. Elevate (👏) if a post helped you make a decision.	Visit a category you are interested in and elevate (👏) the post you like.
www.dribbble.com	You are hired to a **Cloud Computing** startup company. You get an assignment to designing the logo of the company. Search for existing **logos** for inspiration and **download three candidate logos you like the most**.	You are preparing a presentation and need one picture for each of these animals: **cat**, **dog**, and **ant**. Download the three pictures you like the most.	Explore dribbble and download **images you like the most** while you browse.