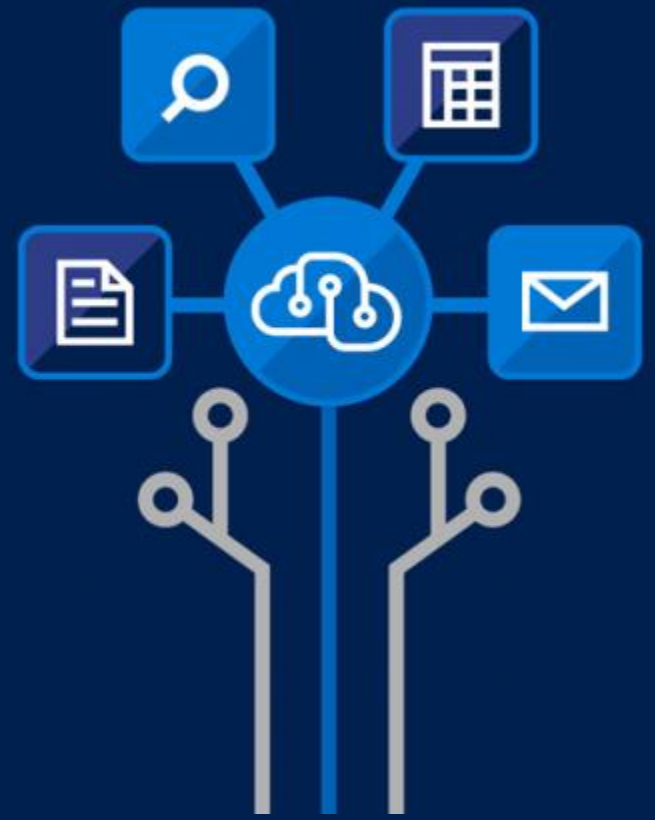
Leveraging Behavioral Data to Improve Workplace Productivity



Ahmed Hassan Awadallah
http://aka.ms/ahmed

Microsoft

Microsoft Research AI

What makes Information Personal?

Owned by “me”

About “me”

Relevant to “me”

Directed toward “me”

Experienced by “me”

Sent by “me”

Jones, 2008, Keeping Found Things Found

Sales Data

Code Reviews

Work Items

Business Data

Documents

Tasks

Communications

Calendar

Slides

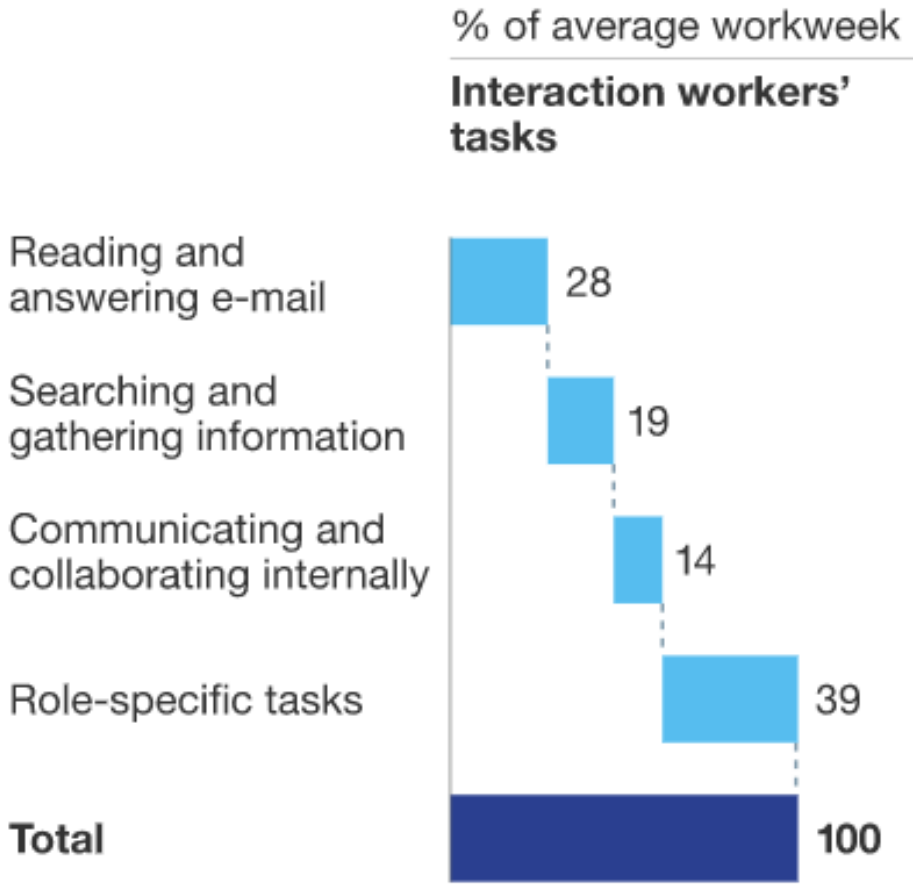
Enormous Volume of Information Exchanged in the Workplace

In 2018, the total number of business emails sent and received per day is **120.4 billion** by over **1 billion** users

Average office worker **receives 76 non-spam** emails a day, **sends out 32** business emails a day.

Source: Email Statistics Report. The Radicati Group INC.

Significant time spent interacting with this information



Source: International Data Corporation (IDC), McKinsey Global Institute analysis

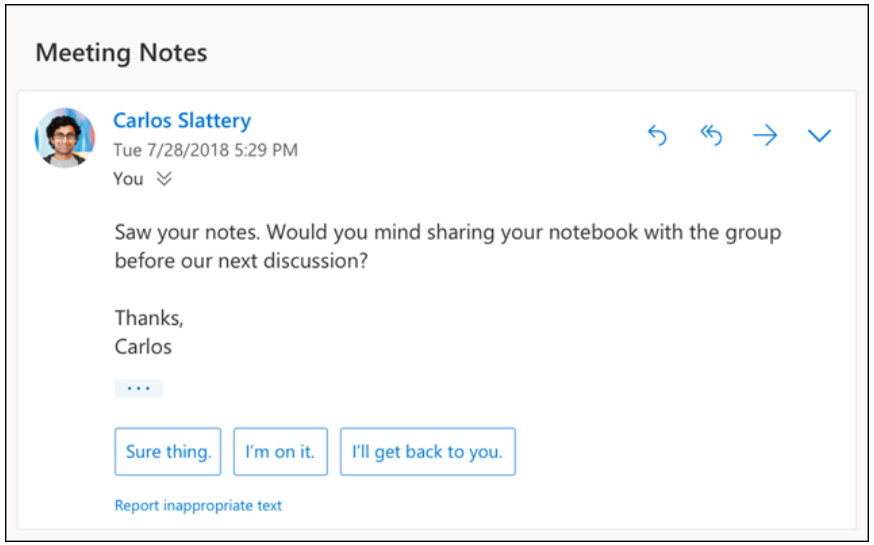
People should spend less time searching for and organizing information and more time acting on and drawing insights from it

Email requesting a document

Updated Proposal

"can you share the latest version of the proposal?"

Share Document **Open Email**

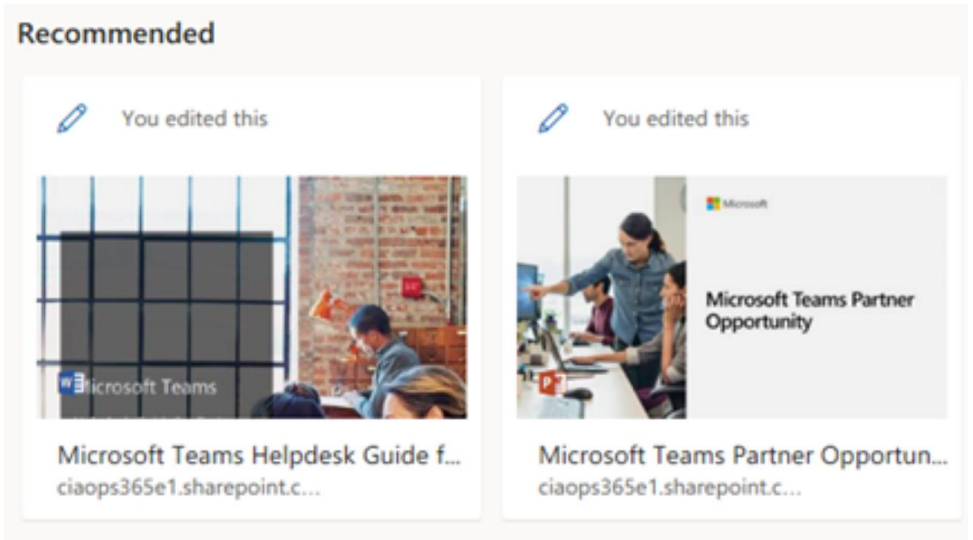


Email requesting a meeting

Re: Updated Proposal

"The proposal looks great. Let's meet to discuss the next steps next week?"

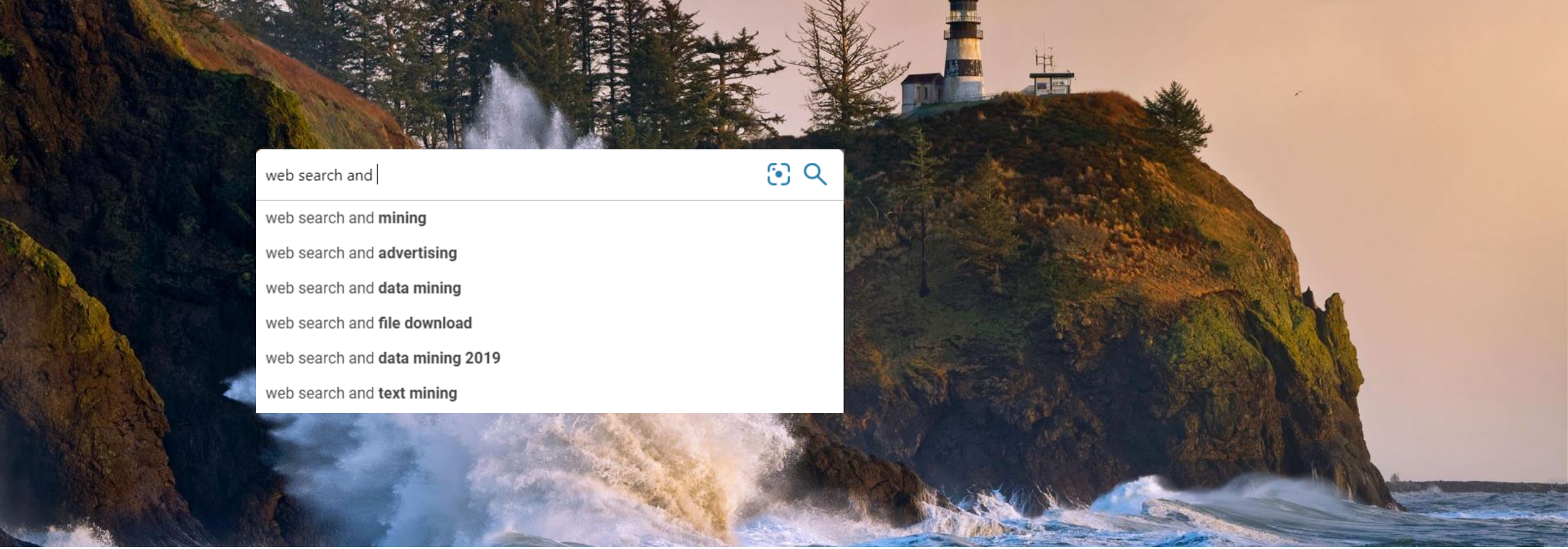
Create Meeting **Open Email**



The World Wide Web

Trillions of pages indexed.

Billions of queries per day.



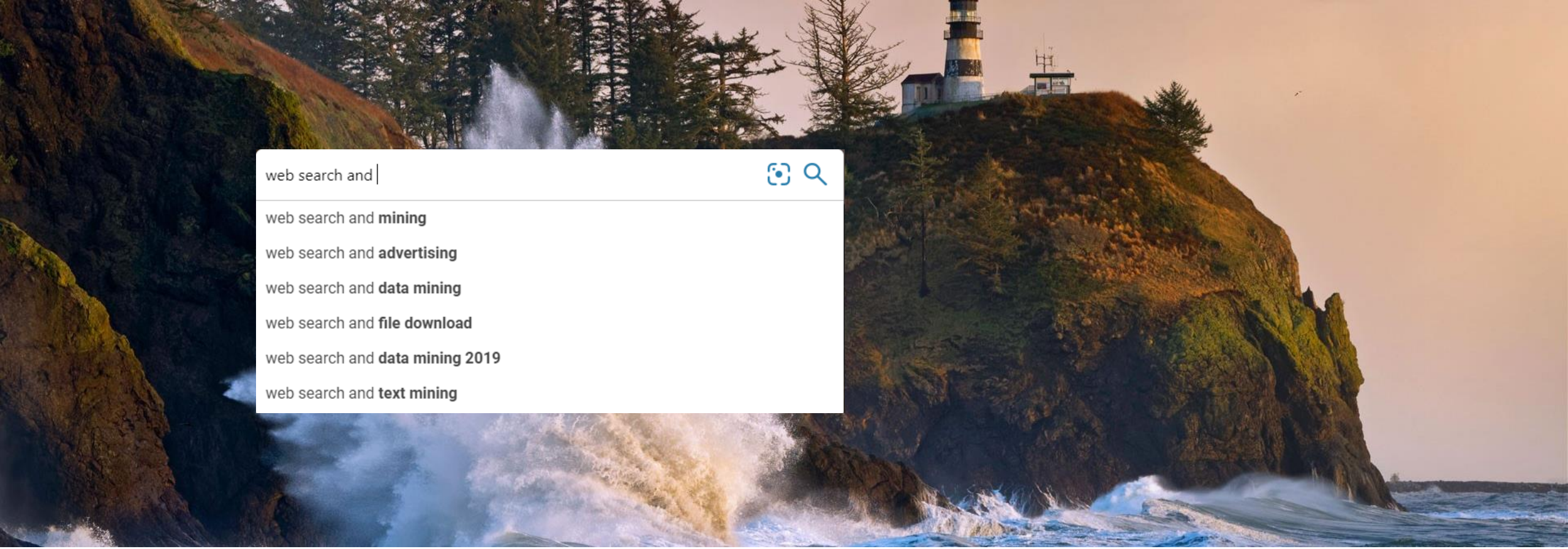
Modem Web Search Engines

User Behavioral Data

What is Behavioral data?

Behavioral logs are *traces of human behavior*

- seen through the lenses of whatever sensors we have
- Examples: utterance, queries, browsing, invoking apps, clicks on UI elements, keystrokes, gaze patterns, physiological responses



How can behavioral data help?

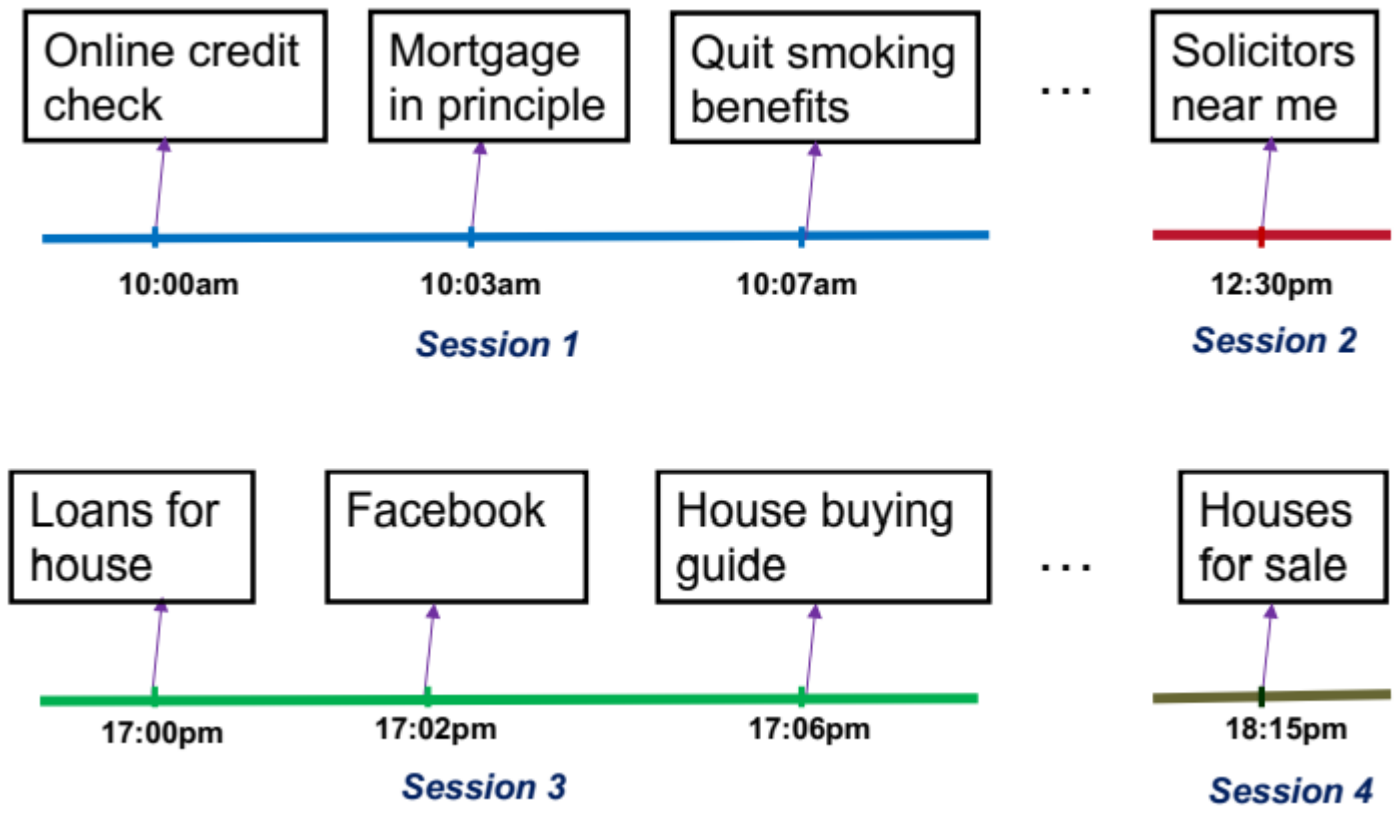
Understand users' intents and needs

Learn to build better ML models

Measure user satisfaction and engagement

How is Behavioral data used in Web Search?

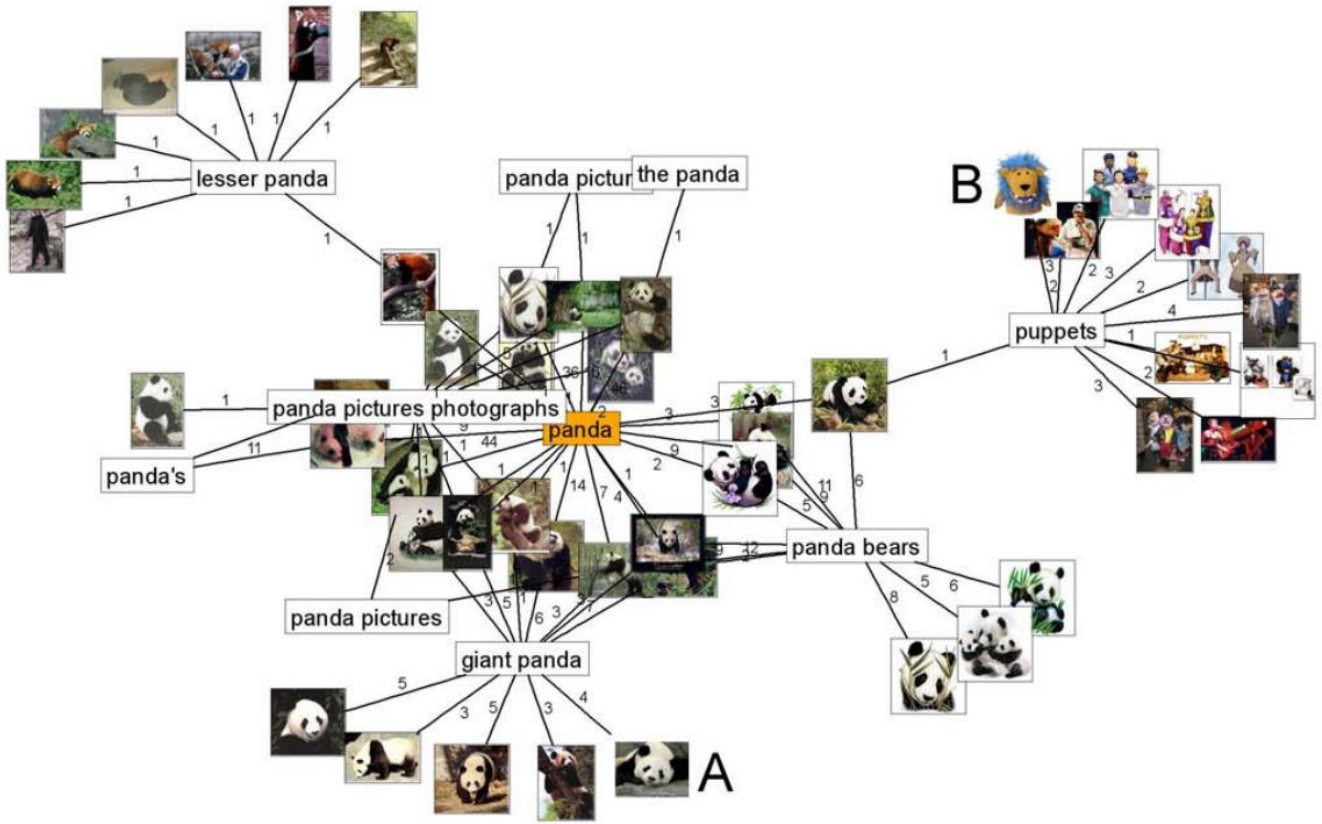
Understand Tasks and Intent from Query Logs



[Jones and Klinkner, 2008]

How is Behavioral data used in Web Search?

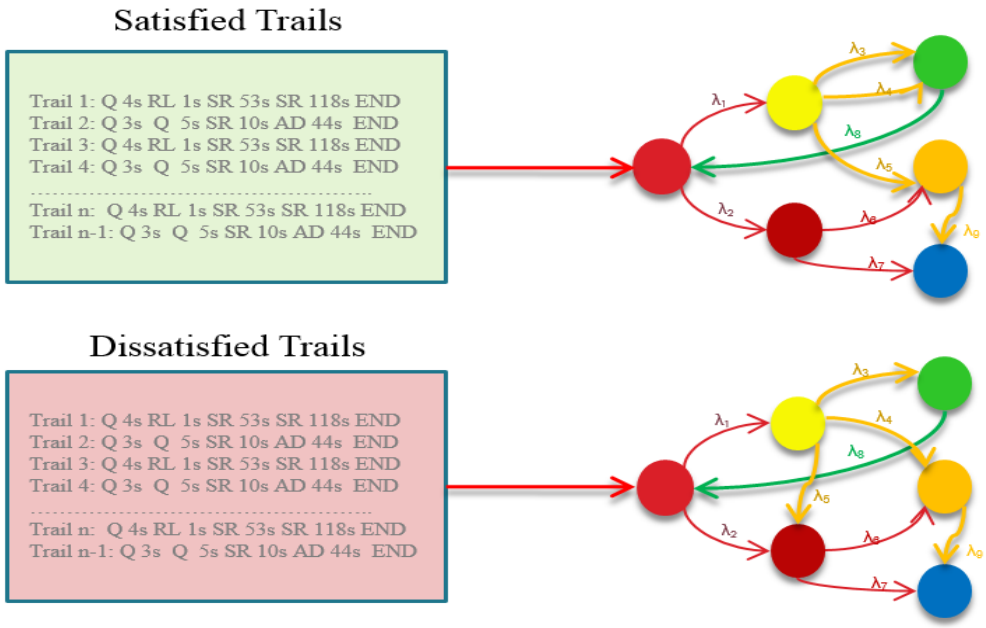
Learn to Rank using the Click Graph



[Joachims, 2002,Craswell and Szummer, 2007]

How is Behavioral data used in Web Search?

Measure User Satisfaction



[Hassan et al., 2010]

Improving Productivity by leveraging Behavioral Data

Three Productivity Applications

Communications

Help users respond to and take actions on email messages

Documents

Surface document recommendation relevant to the user's context

Tasks

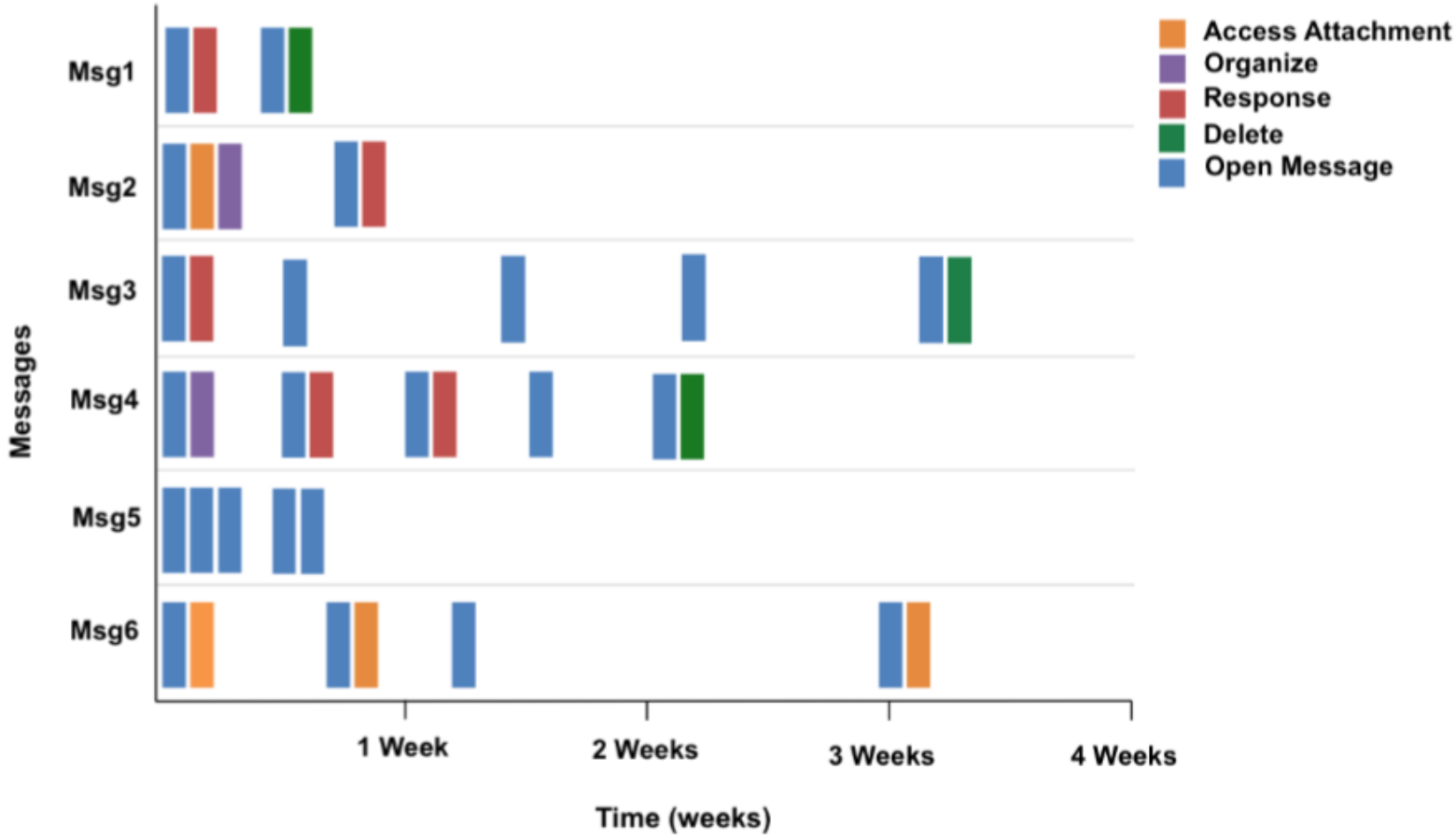
Assist users with task management and planning

Email Intelligence

The Lifetime of an Email Message

Different lifetime spans and different interaction patterns

Revisiting to **take an action** or **retrieve information**

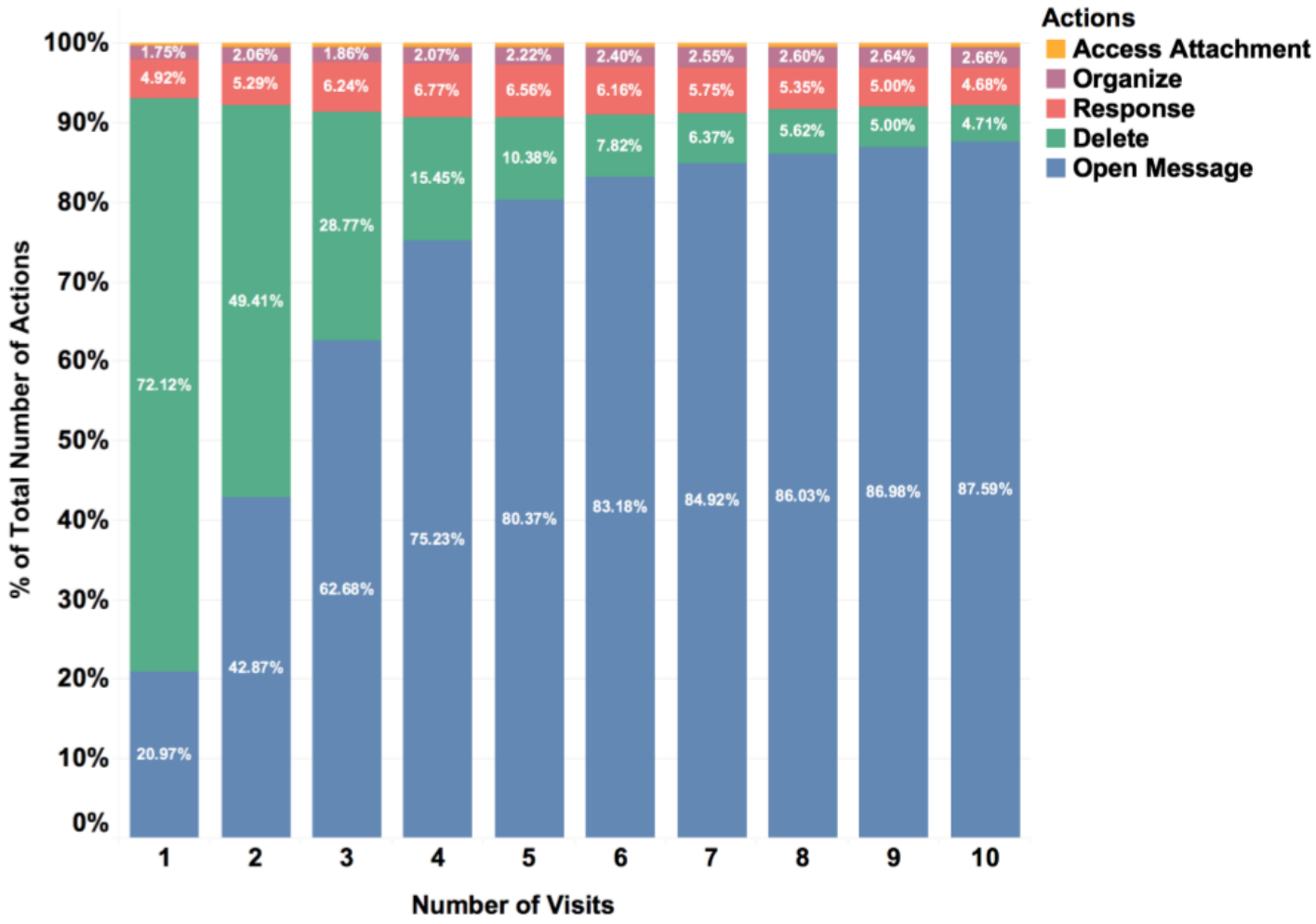


Alrashed et al., Sub, CHIIR 2018

The Lifetime of an Email Message

Many messages are deleted upon receipt, others are only read

Responding to a message is a frequent event and is most likely to happen during the first few visits

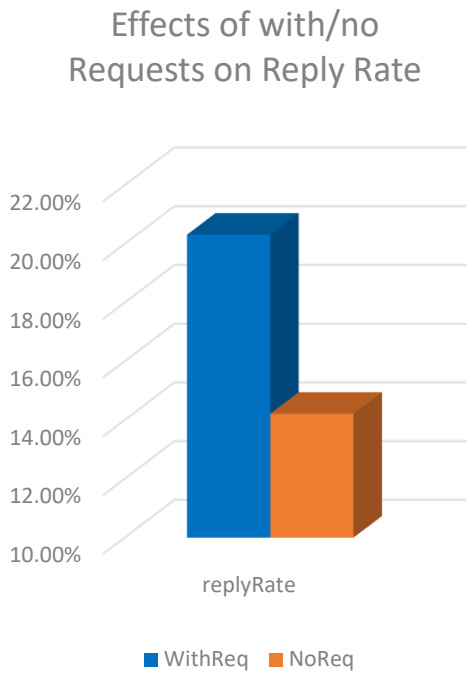


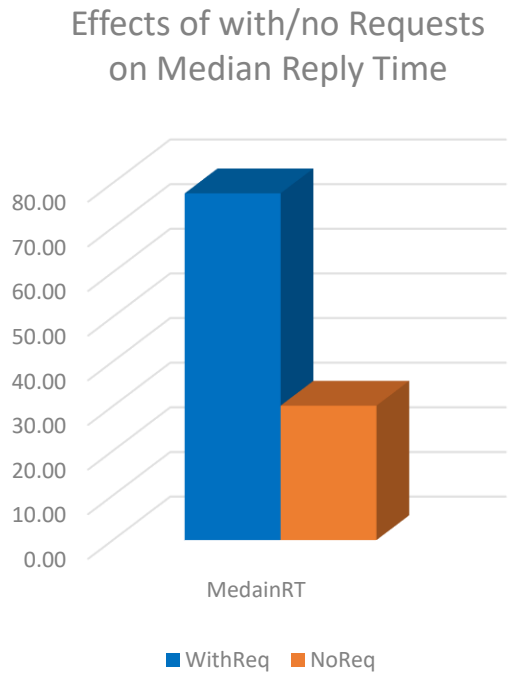
Alrashed et al., Sub, CHIIR 2018

Let’s focus on responding to a message

If the message is requesting an action from the recipient, it is more likely to receive a reply

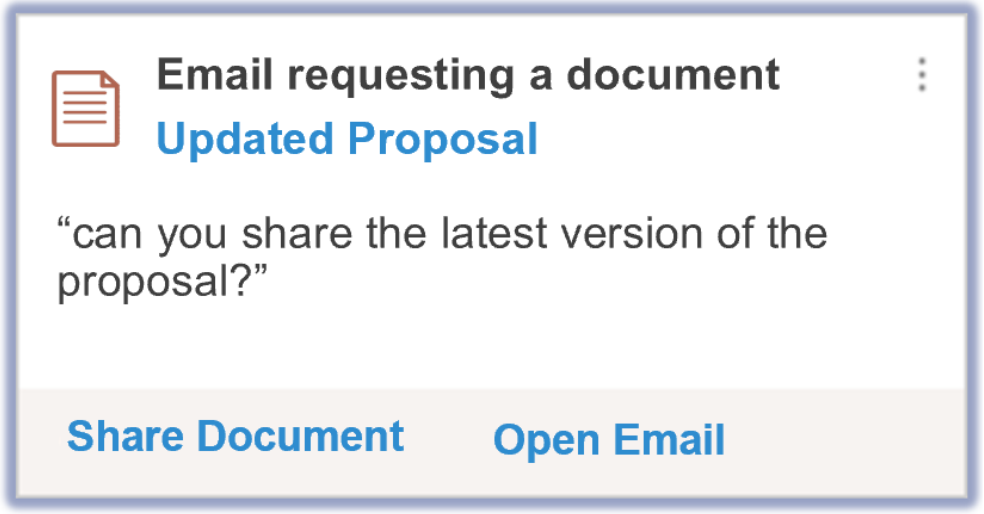
But it is also more likely to receive a slow reply

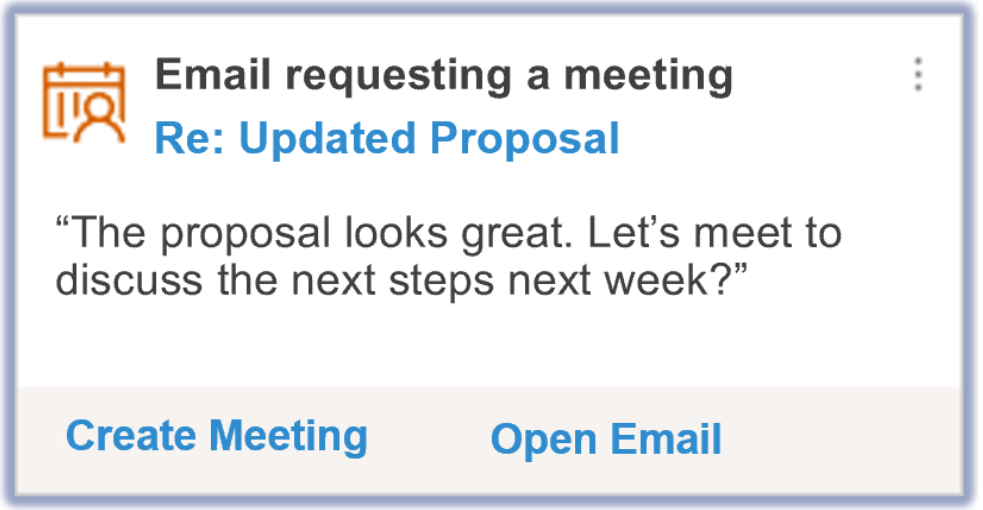




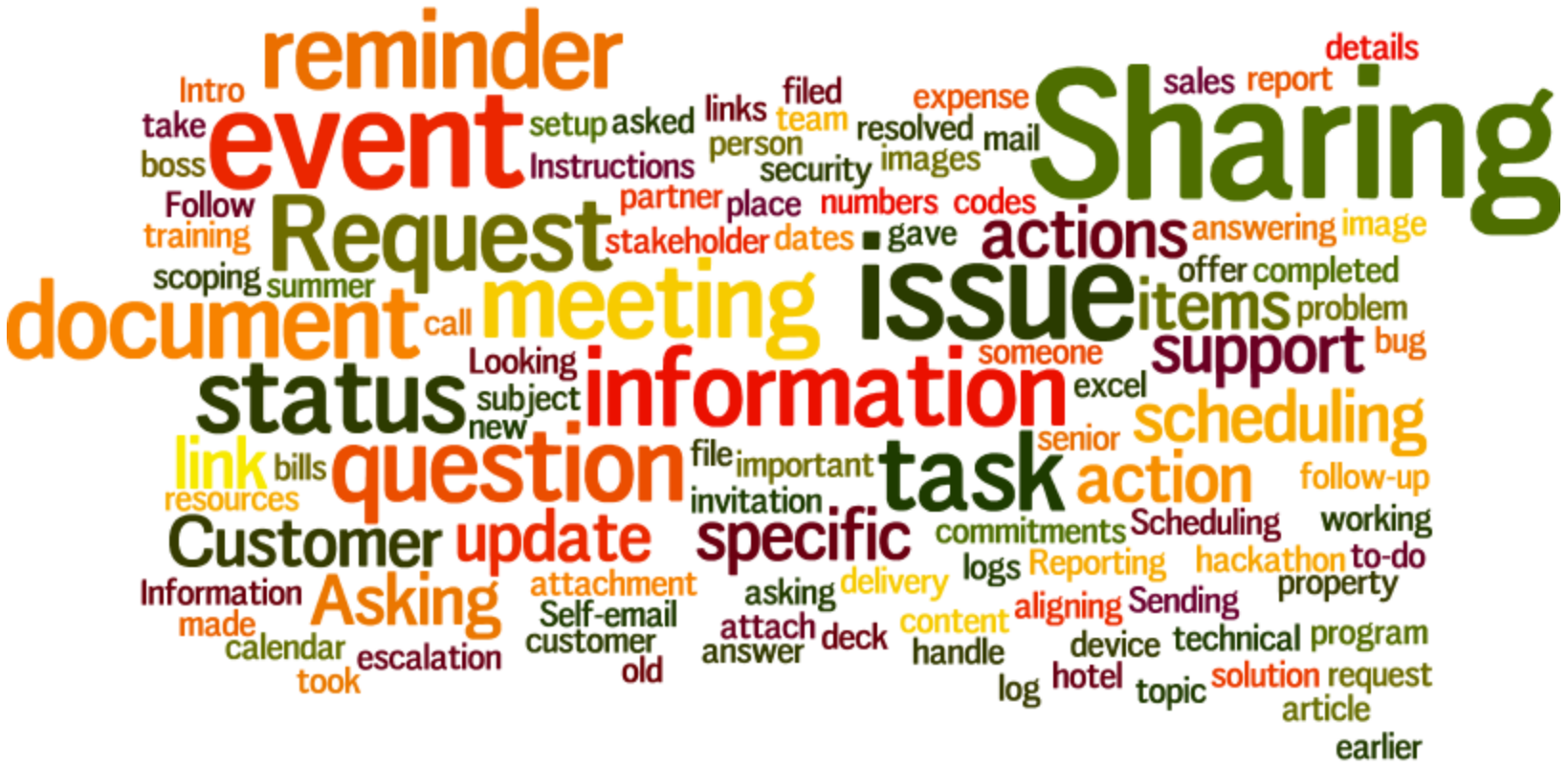
Yang et al., SIGIR 2017

Message to Action Recommendation



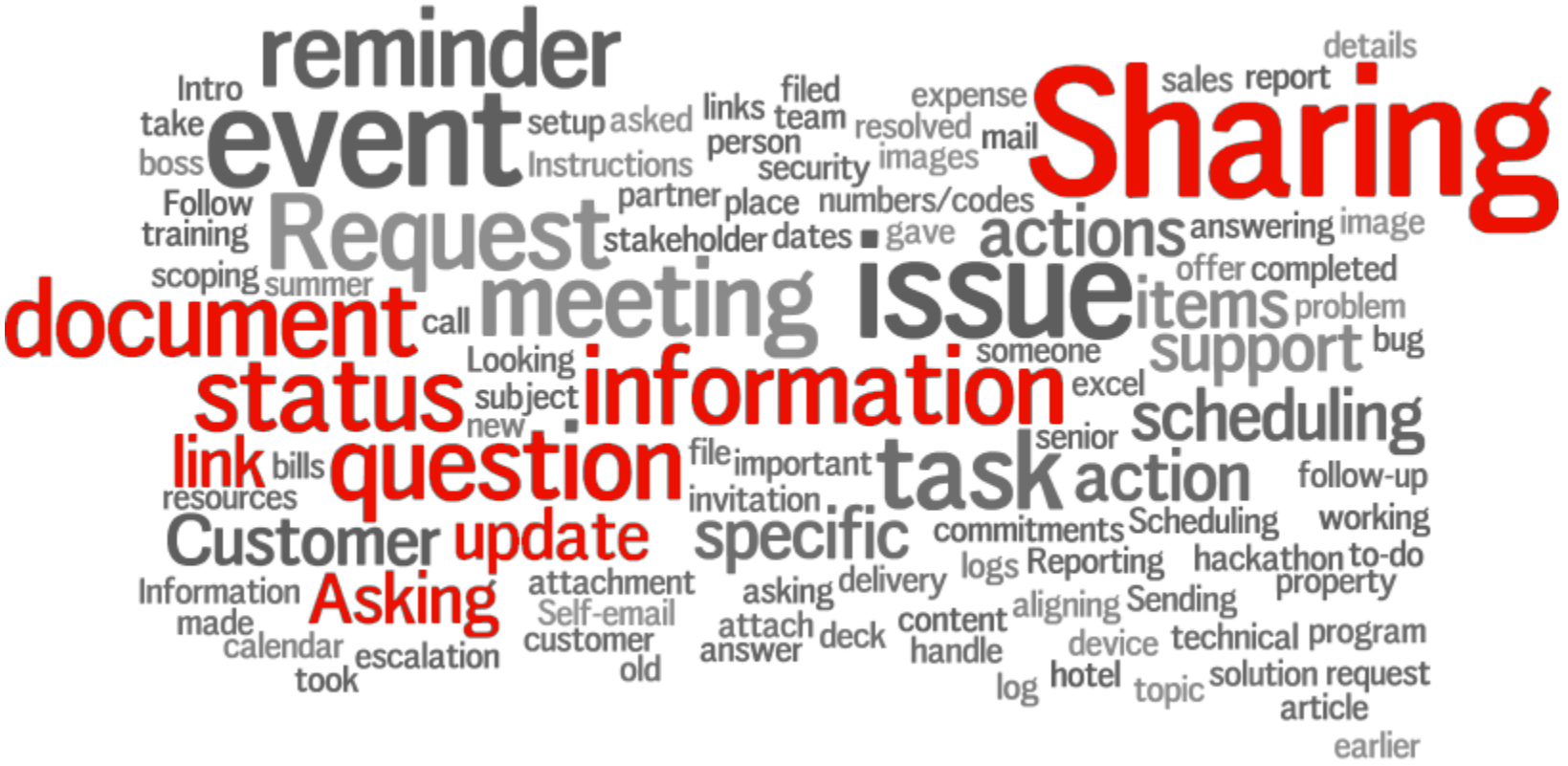


Why do we send email messages?



Why do we send email messages?

Information Seeking/Sharing



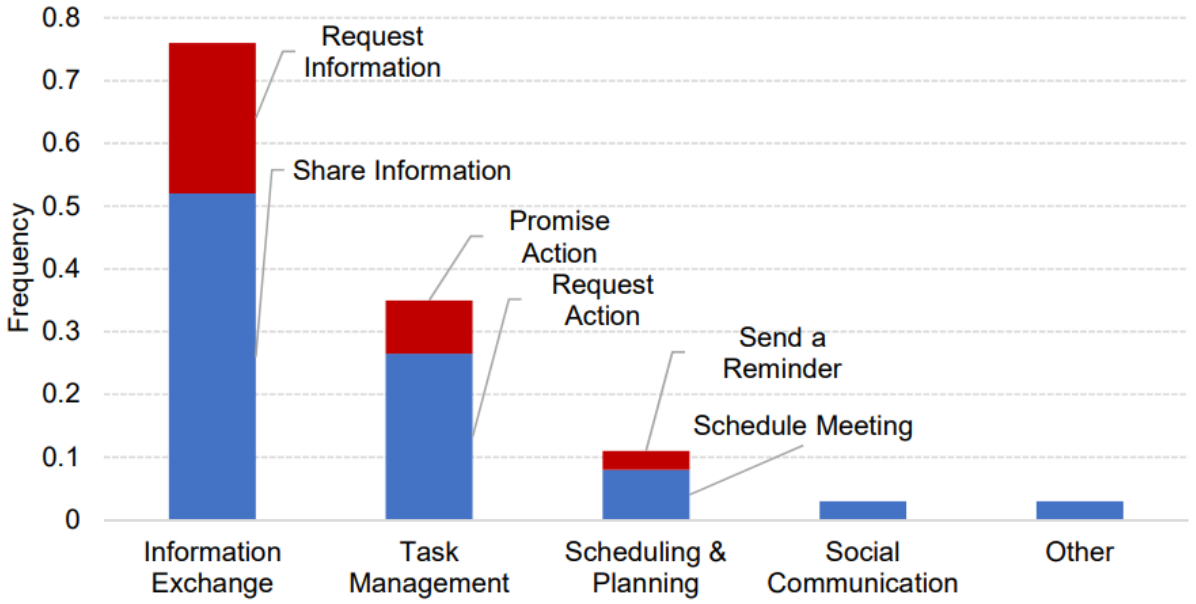
Why do we send email messages?

Scheduling/Coordinating

Why do we send email messages?

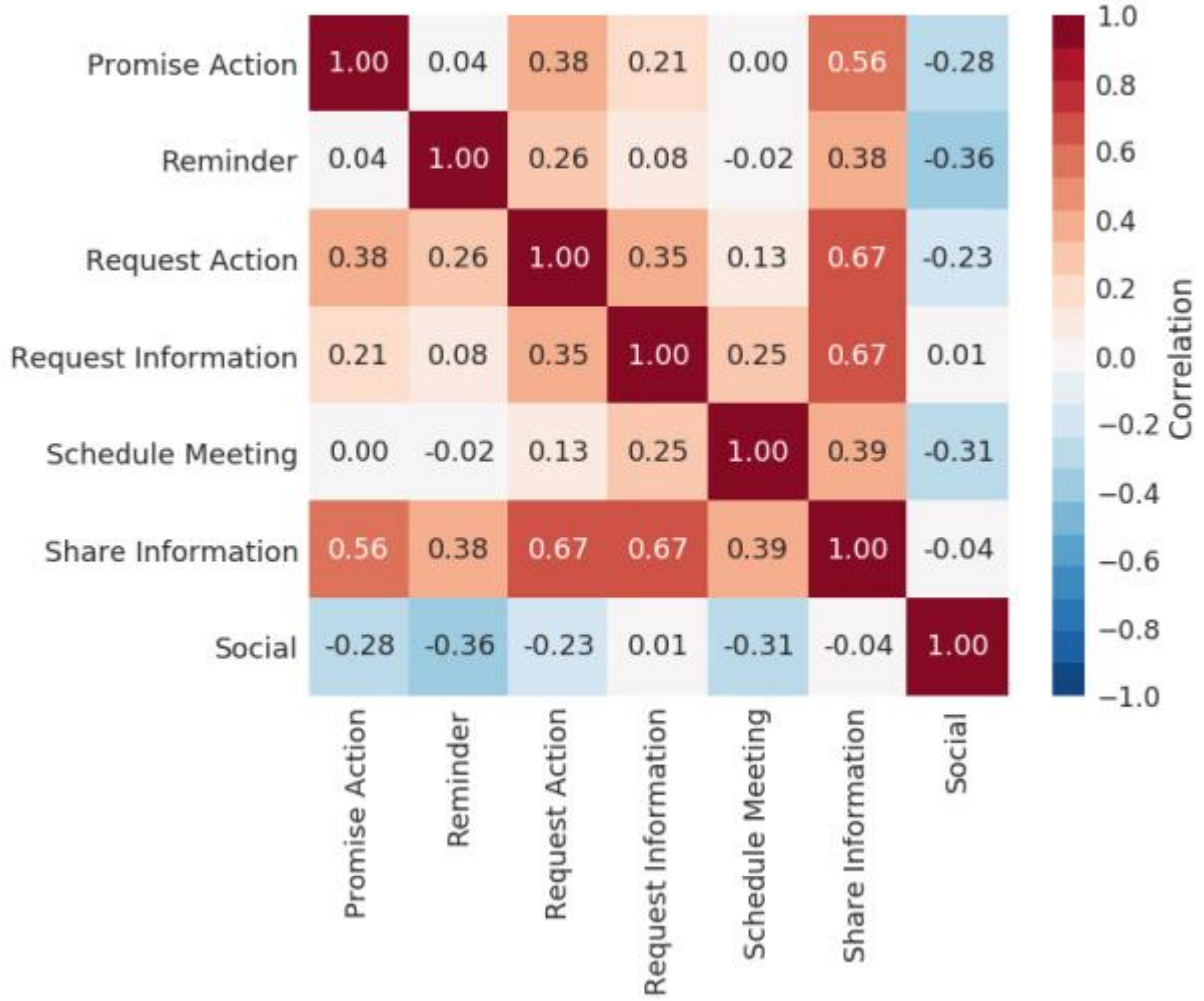
Tasks/Actions

Message to Action: Understanding Intent

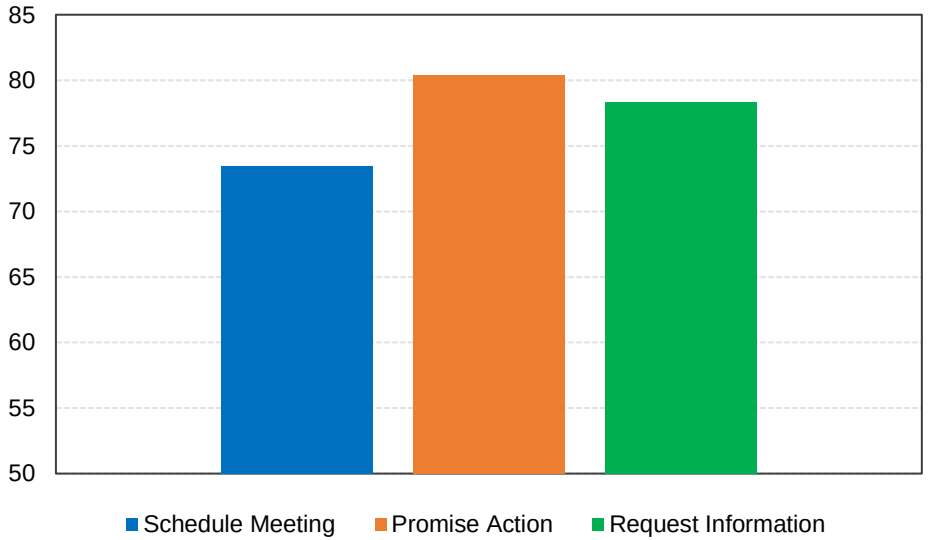


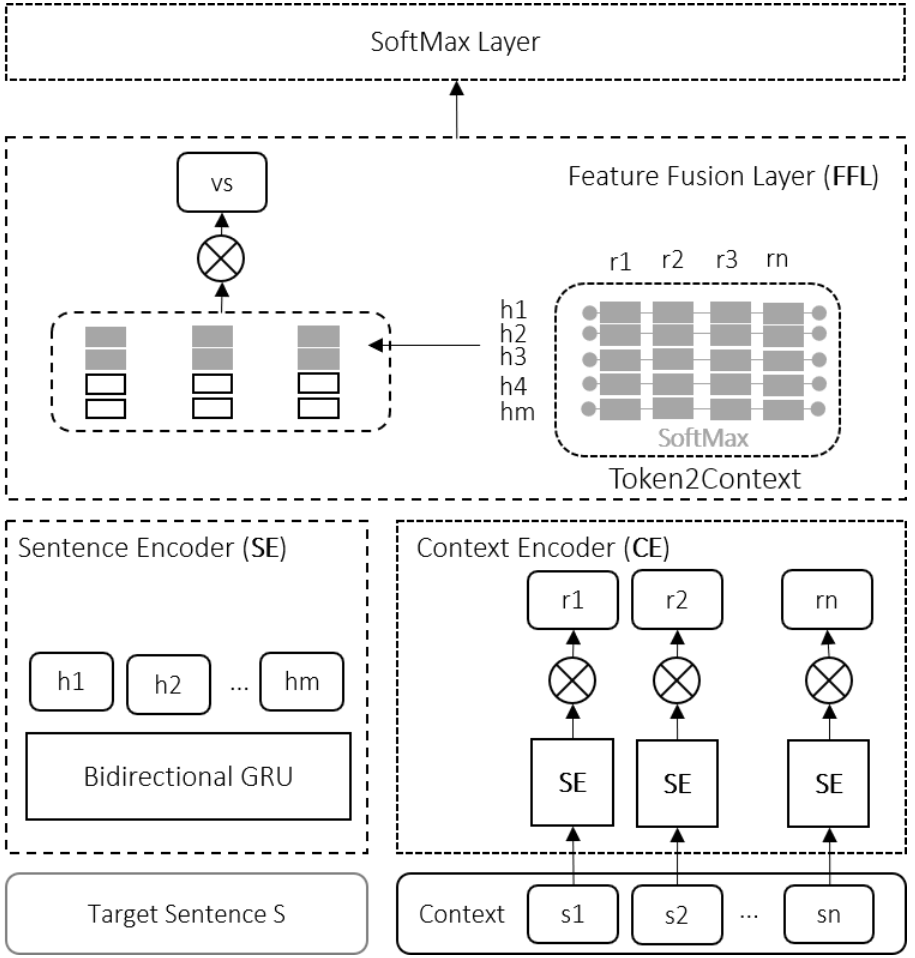
Wang et al., SIGIR 2019

Message to Action: Understanding Intent



Message to Action: Recommendation



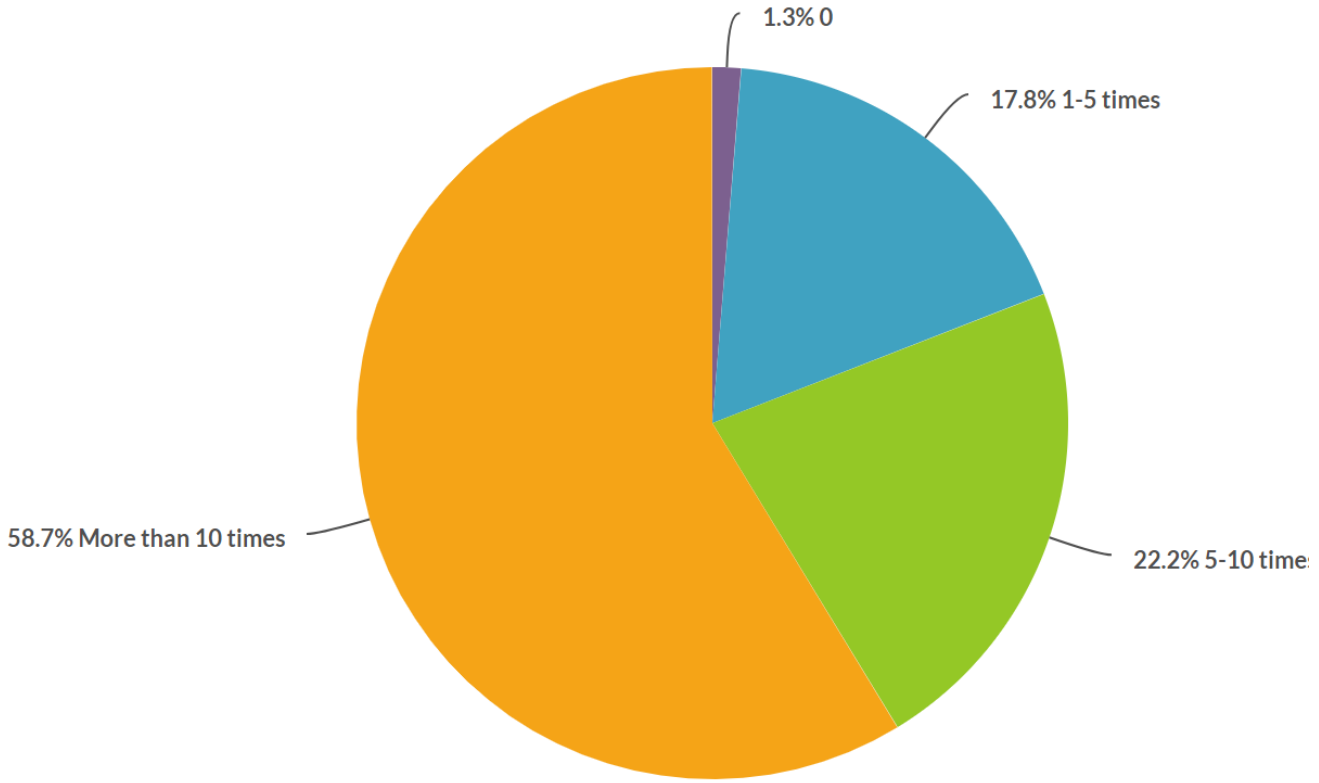


Wang et al., SIGIR 2019

Question/Answer Exchange over Email

People frequently exchange questions/answers over email and they often need to go back to such information

— more than 5 times a month for 80% of them



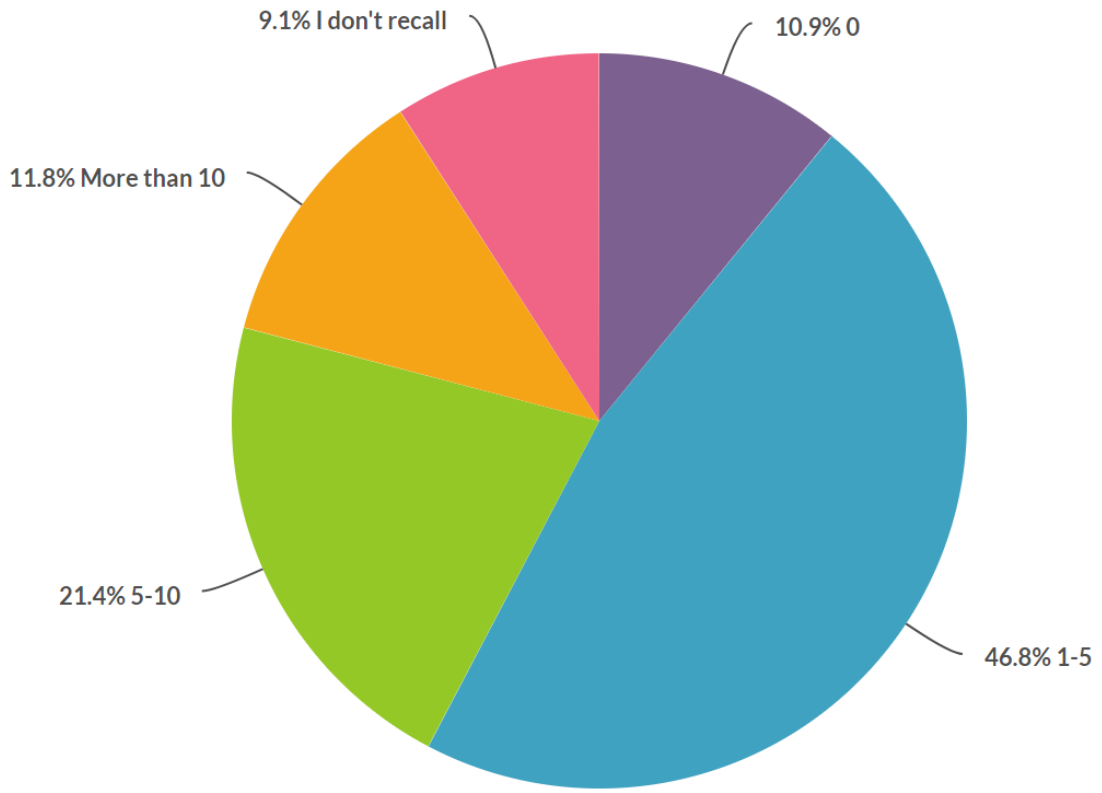
How often do you need to go back to threads because they contain Questions/Answers?

Yang et al., SIGIR 2018

Question/Answer Exchange over Email

Many of them receive the same question and have to answer it over and over again

— more than 5 times a month for 33% of them

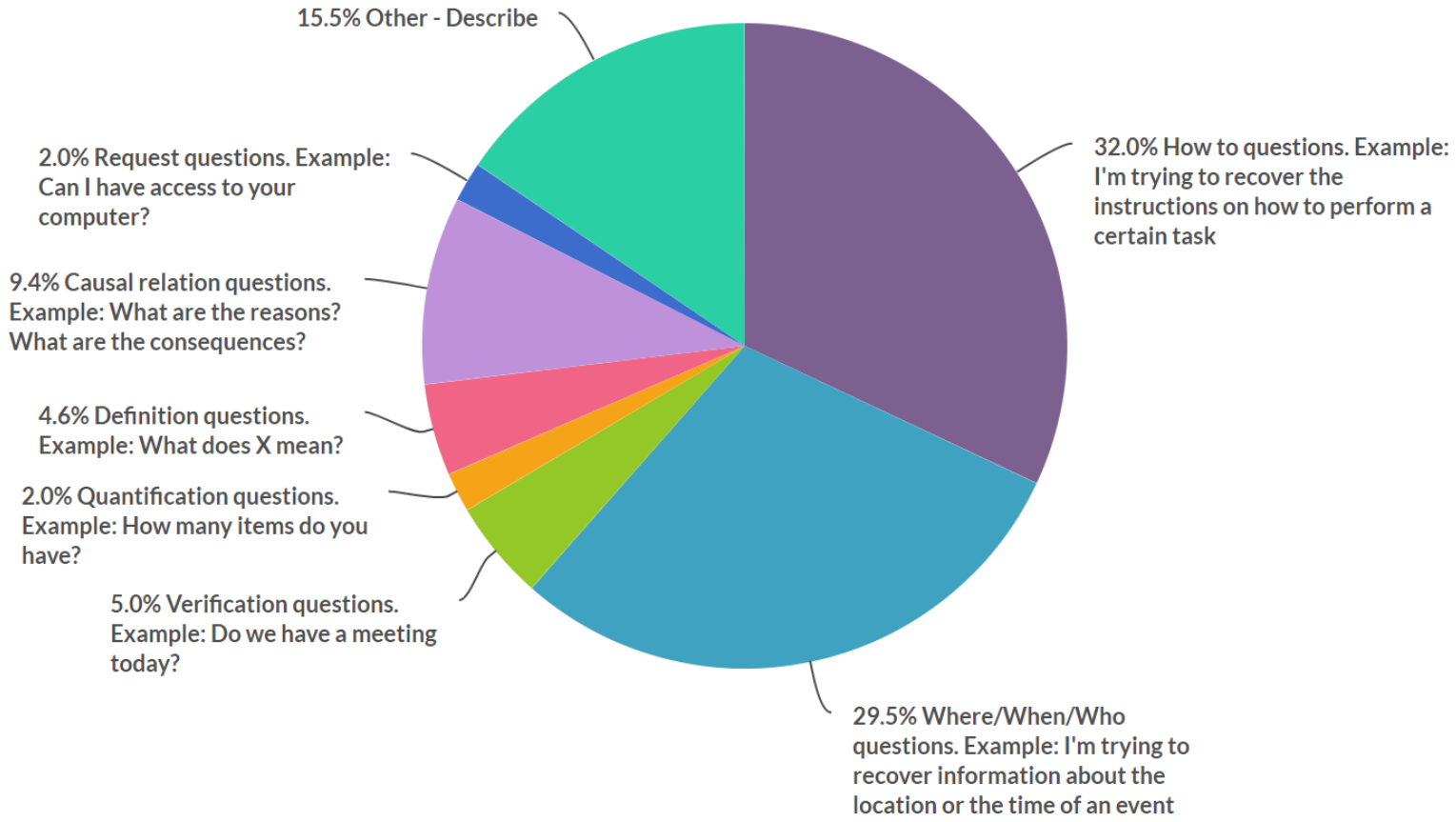


How often do you need to answer the same question multiple times?

Yang et al., SIGIR 2018

Question/Answer Exchange over Email

How-to and When/Where/Who questions are the most popular types people need to get back to

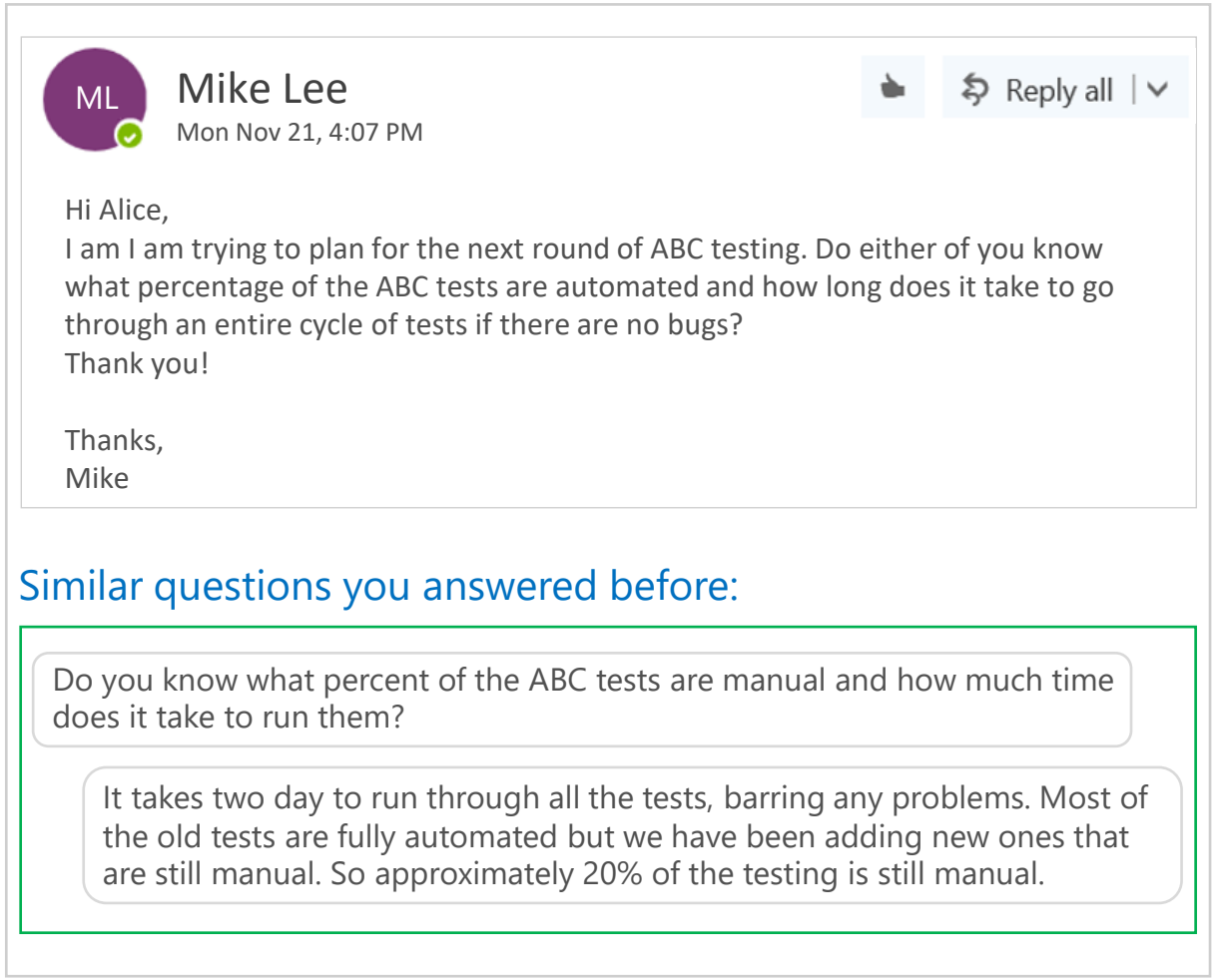


Yang et al., SIGIR 2018

Question/Answer Exchange over Email

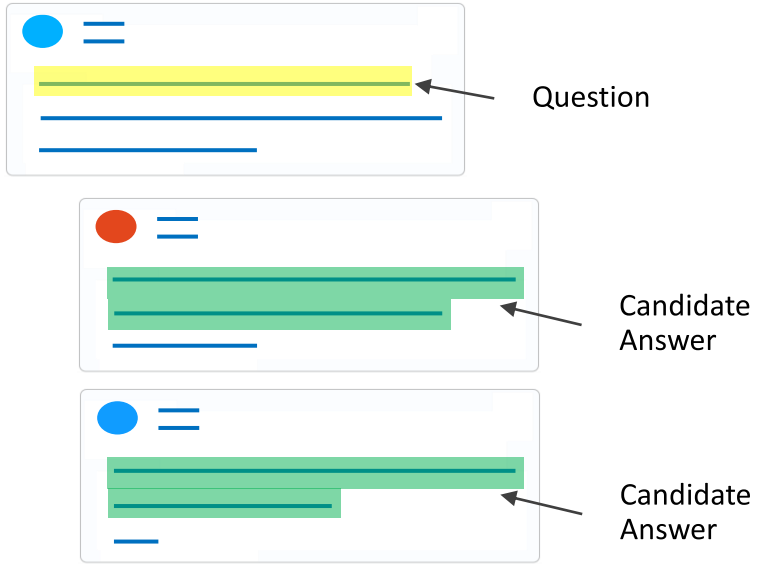
Reply Support: Help users answer questions they already answered before

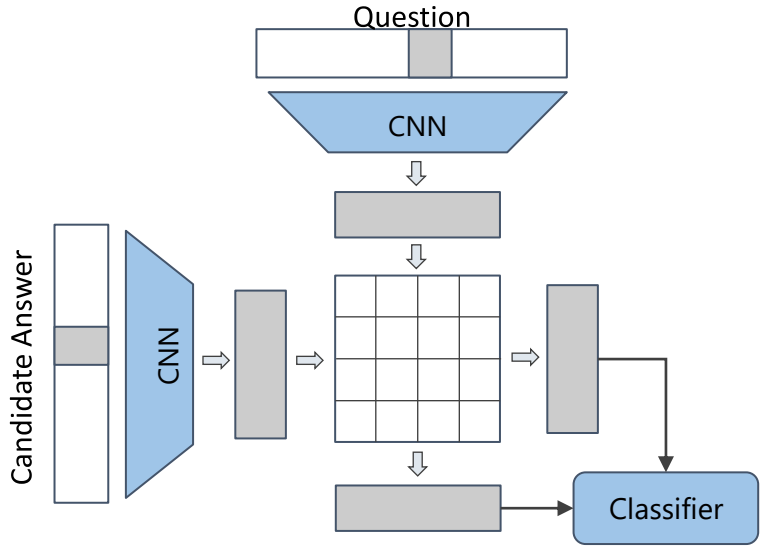
Search Support: Help users locate answers they received before



Question/Answer Exchange over Email

Learn to identify question/answer pairs using content and user interaction information

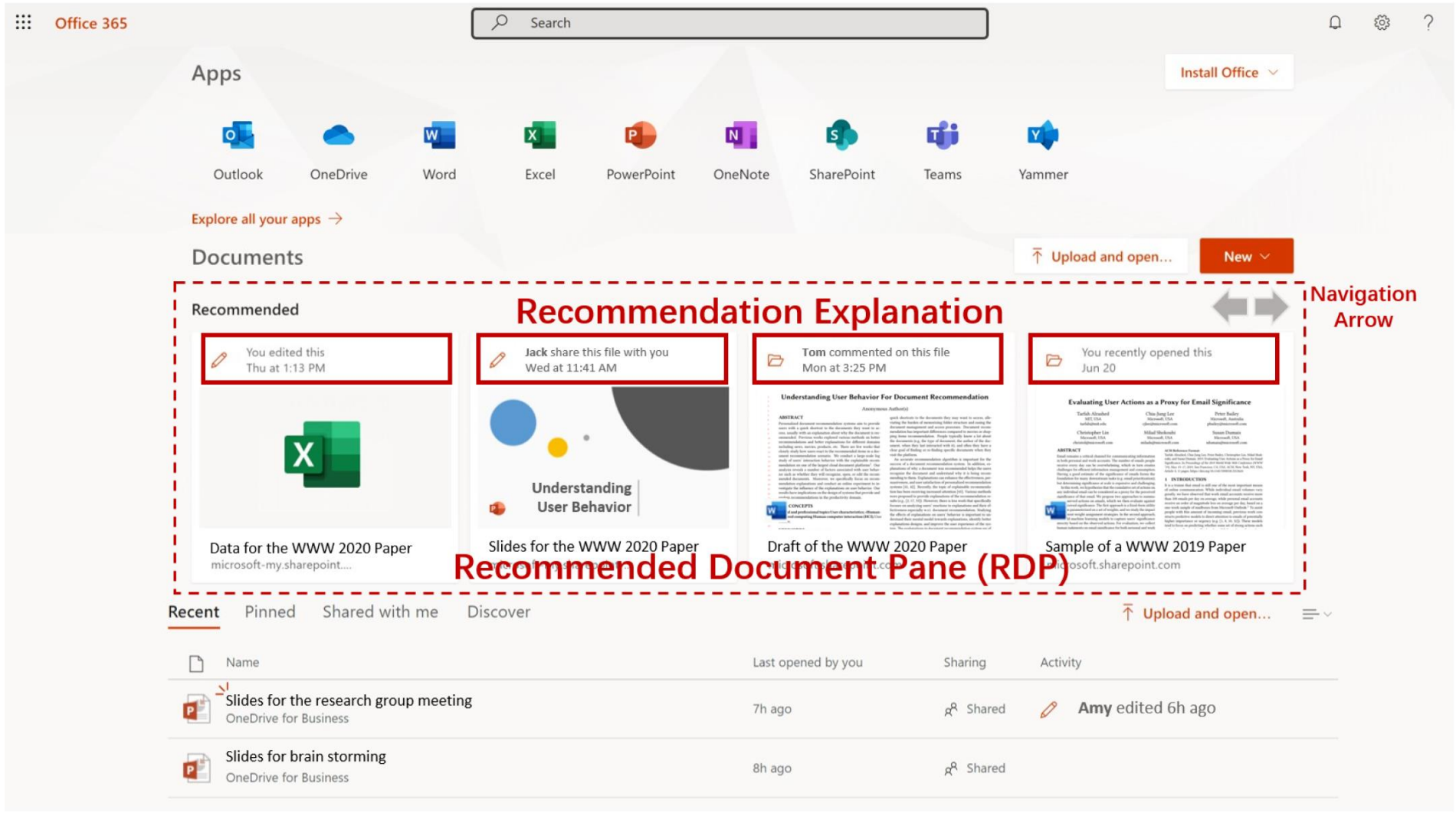




Yang et al., AAAI 2019

Document Recommendation

Document Recommendations



Types of Behavioral Data

	Observational	Experimental
Lab Studies *Controlled tasks, in laboratory, with detailed instrumentation*	In-lab behavior observations	In-lab controlled tasks, comparisons of systems
Panel Studies *In the wild, real-world tasks, ability to probe for detail*	Ethnography, case studies, panels (e.g., Nielsen)	Clinical trials and field tests
Log Studies *In the wild, no explicit feedback but lots of implicit feedback*	Logs from a single system	A/B testing of alternative systems or algorithms

Types of Behavioral Data

	Observational	Experimental
Lab Studies *Controlled tasks, in laboratory, with detailed instrumentation*	In-lab behavior observations	In-lab controlled tasks, comparisons of systems
Panel Studies *In the wild, real-world tasks, ability to probe for detail*	Ethnography, case studies, panels (e.g., Nielsen)	Clinical trials and field tests
Log Studies *In the wild, no explicit feedback but lots of implicit feedback*	Logs from a single system	A/B testing of alternative systems or algorithms

Types of Behavioral Data

	Observational	Experimental
Lab Studies *Controlled tasks, in laboratory, with detailed instrumentation*	In-lab behavior observations	In-lab controlled tasks, comparisons of systems
Panel Studies *In the wild, real-world tasks, ability to probe for detail*	Ethnography, case studies, panels (e.g., Nielsen)	Clinical trials and field tests
Log Studies *In the wild, no explicit feedback but lots of implicit feedback*	Logs from a single system	A/B testing of alternative systems or algorithms

Lab Study

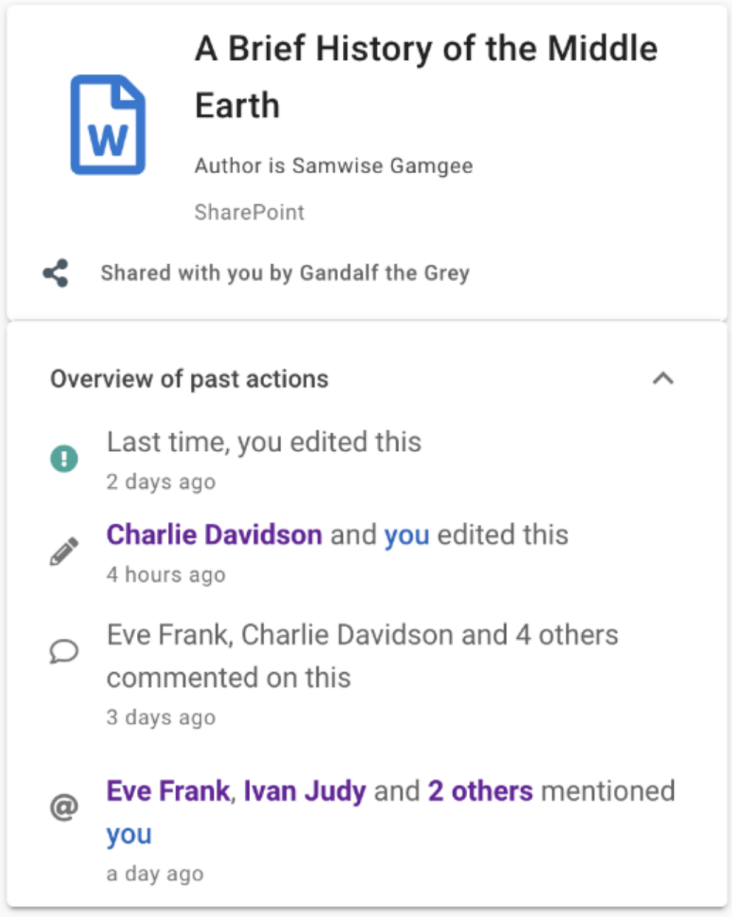
10
PARTICIPANTS

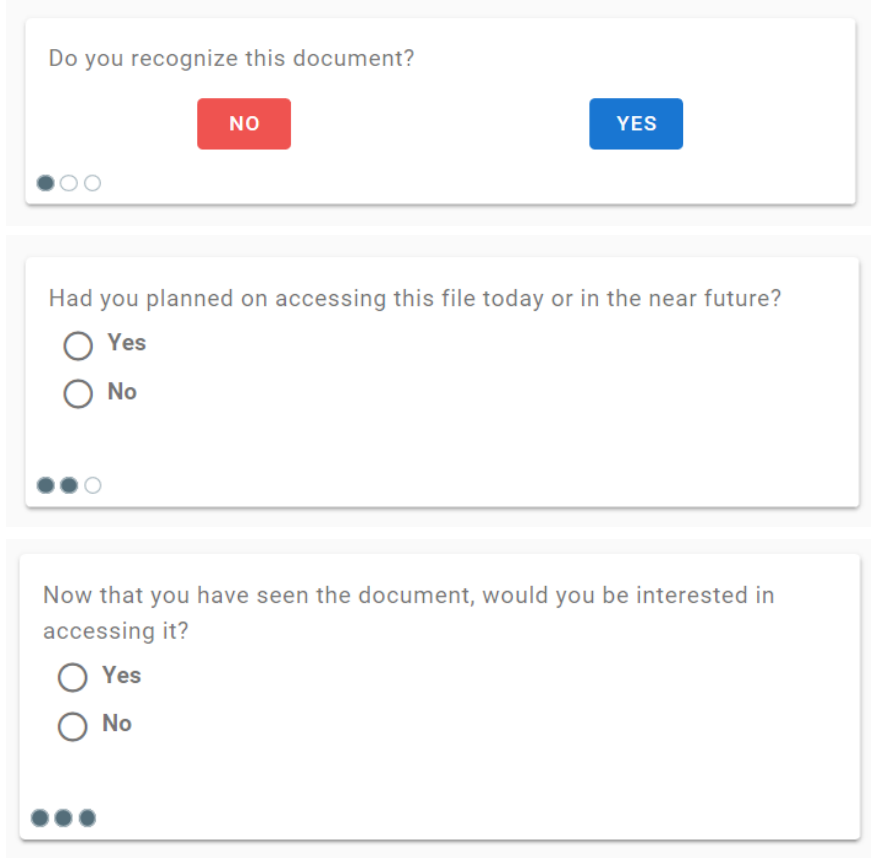
~100
DOCUMENTS

Contextual inquiry, where the participants were asked to look through their document recommendations and share some information about it

Variety of questions with free form text answers to understand *what* and *why* they interact with recommendations

Field Study





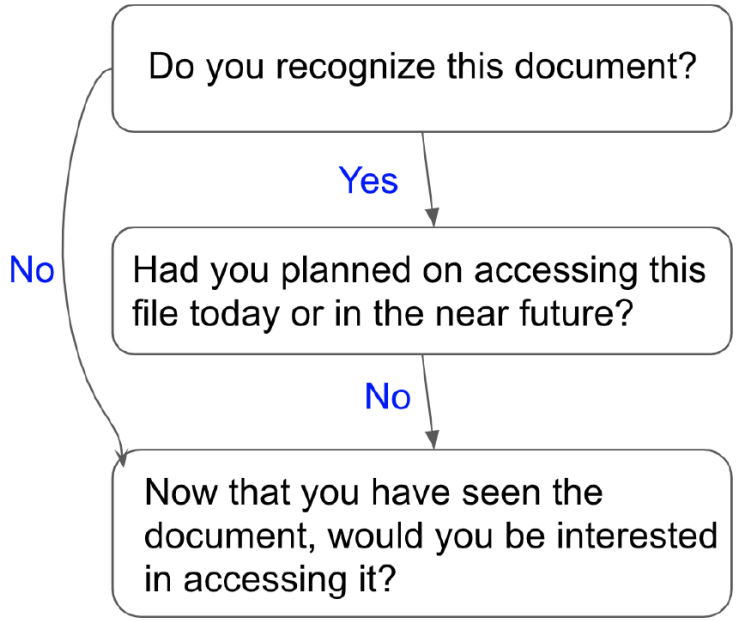
Jahanbakhsh et al., CHIIR 2020

Field Study

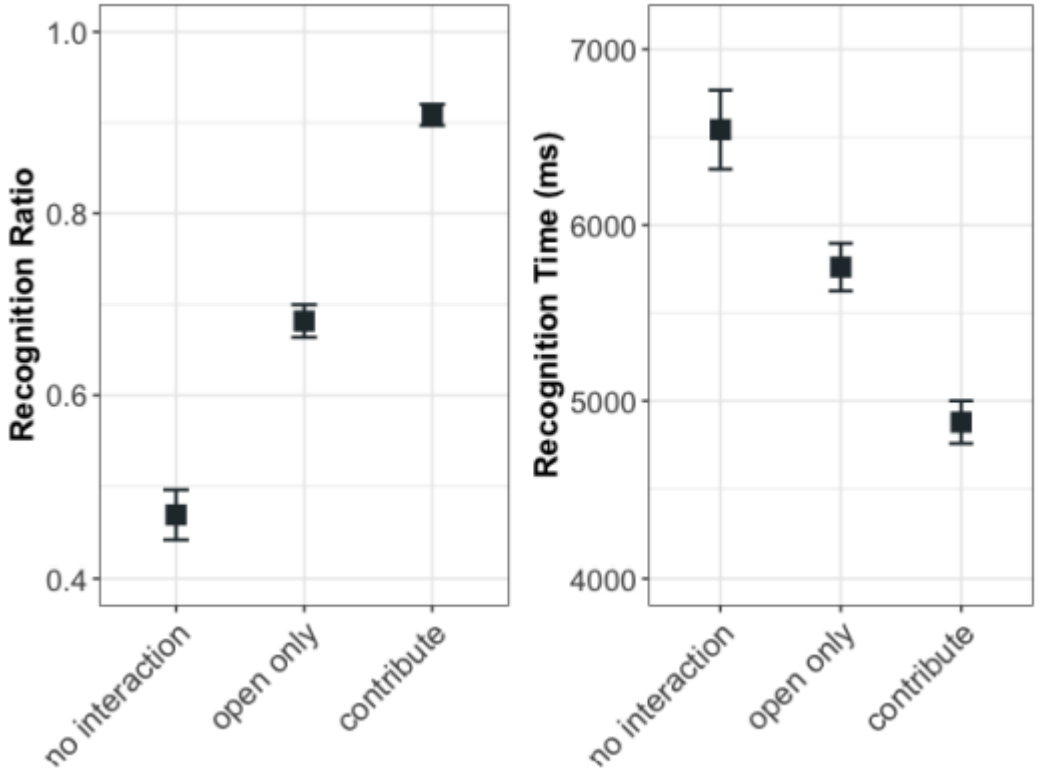
100 +
PARTICIPANTS

~2000
DOCUMENTS

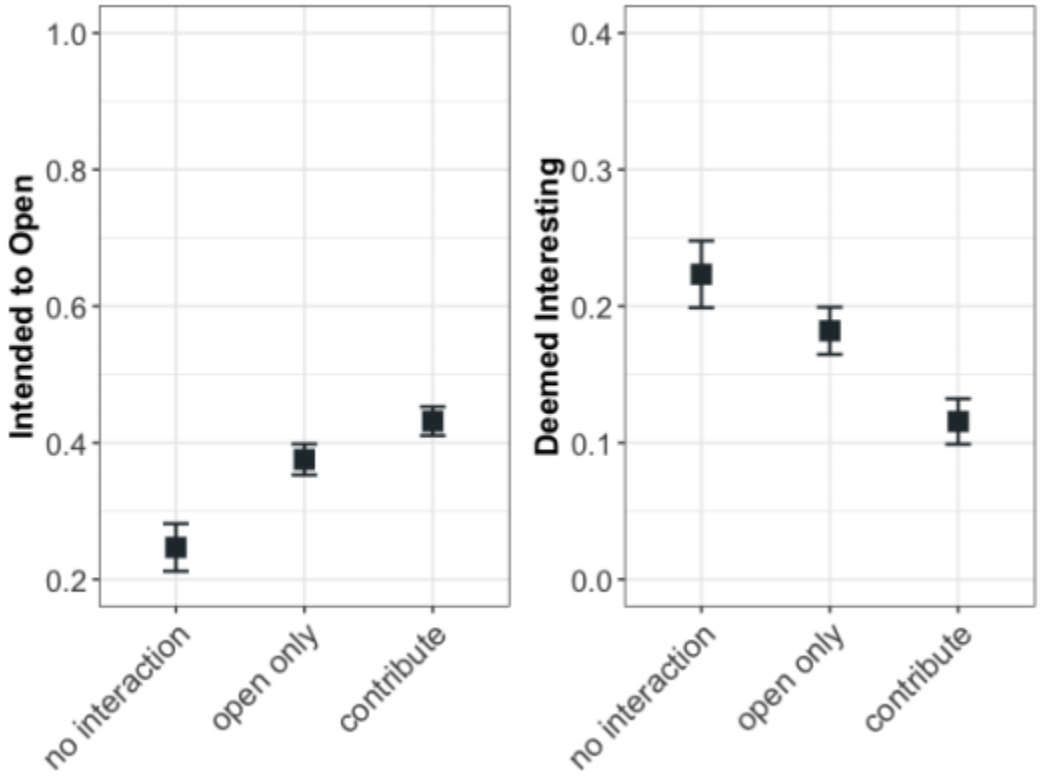
- Recency of Access
- Richness of Interactions
- Presentation



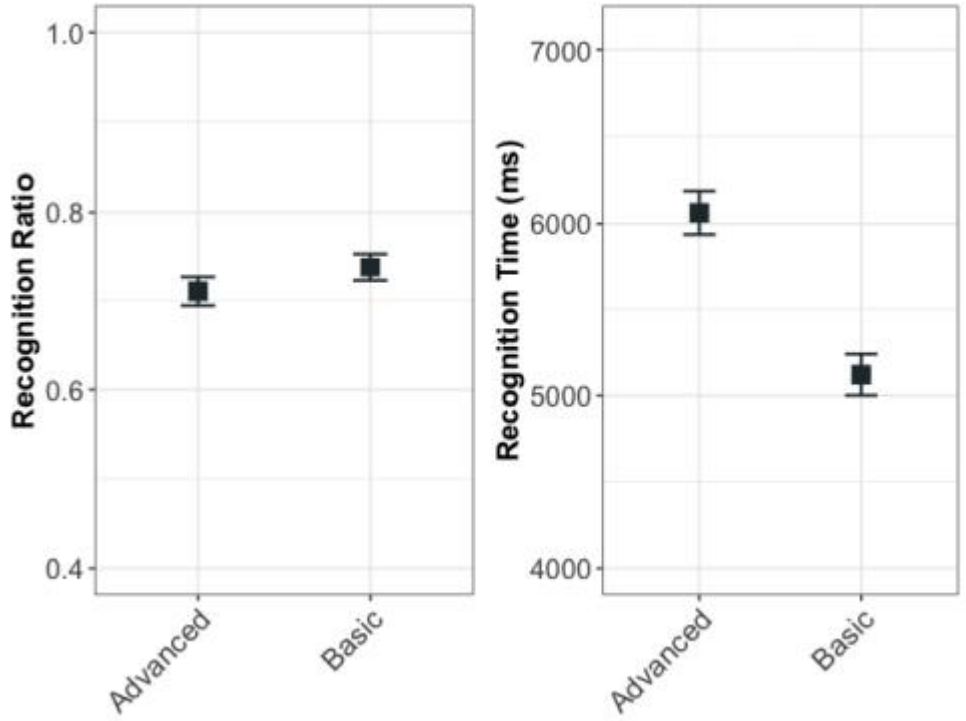
Familiar documents are easier and faster to recognize compared to less familiar ones



Two distinct intents: ***refinding*** known and recently accessed document and ***discovering*** new documents



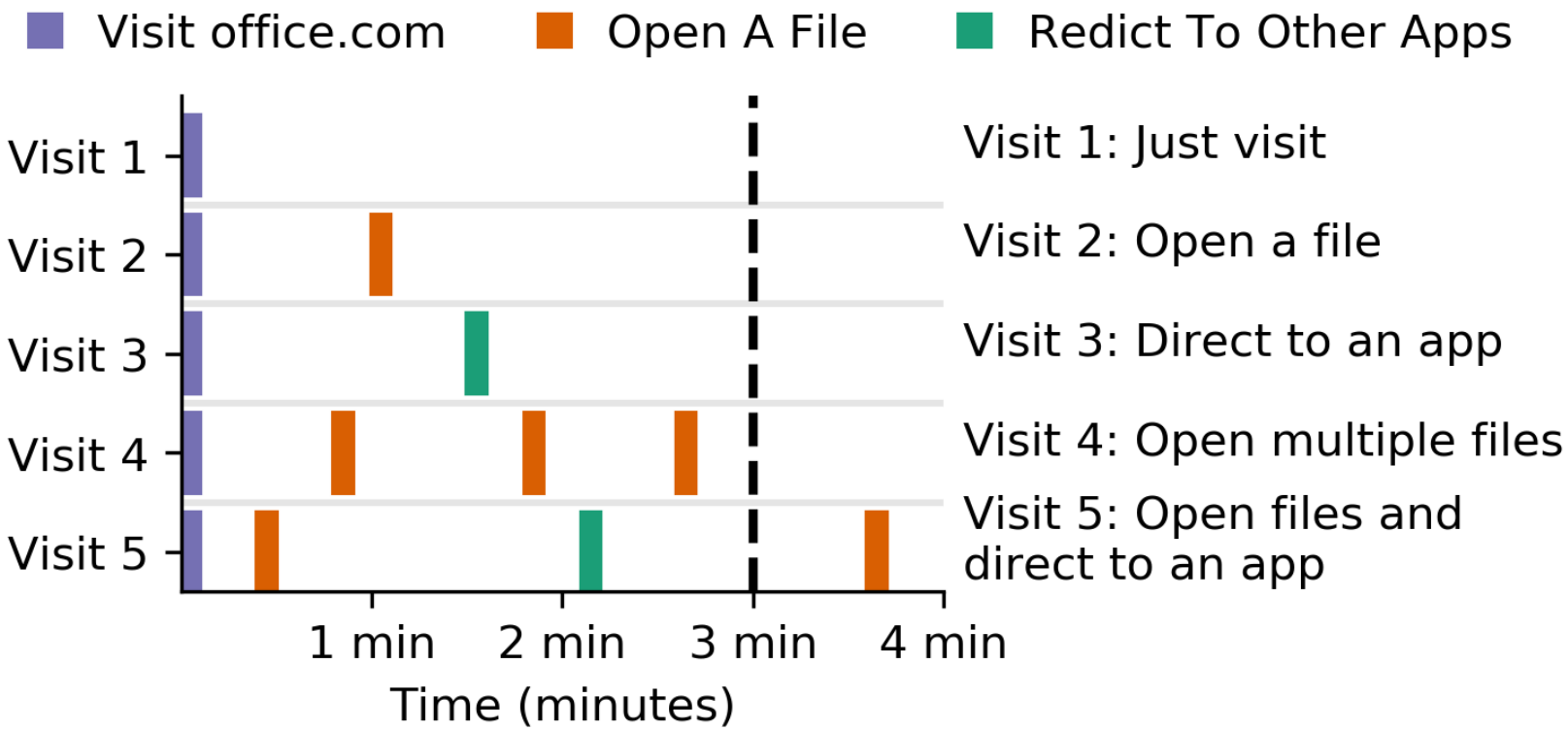
Richer presentations are helpful (verbatim feedback) but take more time to process



Large-Scale Log Study

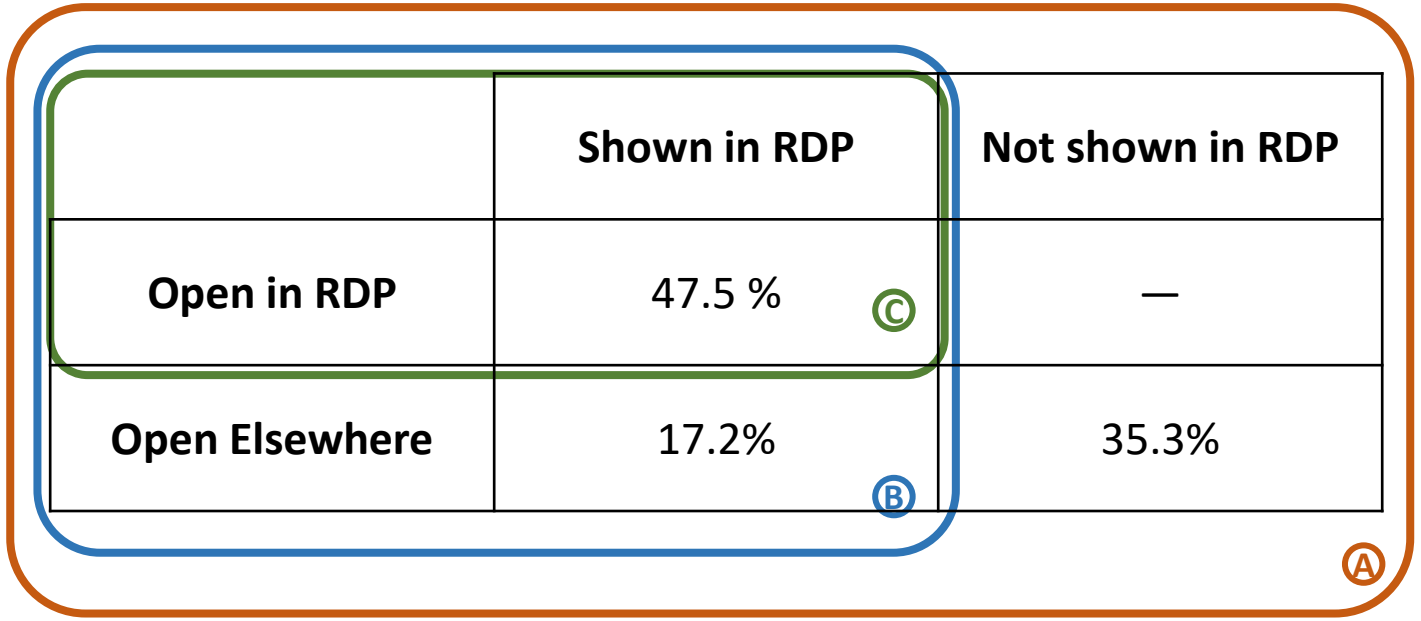
800K
USERS

8.3M
VISITS



Xu et al., WWW 2020

Document Recommendations: How to measure utility?



	Shown in RDP	Not shown in RDP
Open in RDP	47.5 % Ⓒ	—
Open Elsewhere	17.2% Ⓑ	35.3%

Ⓐ

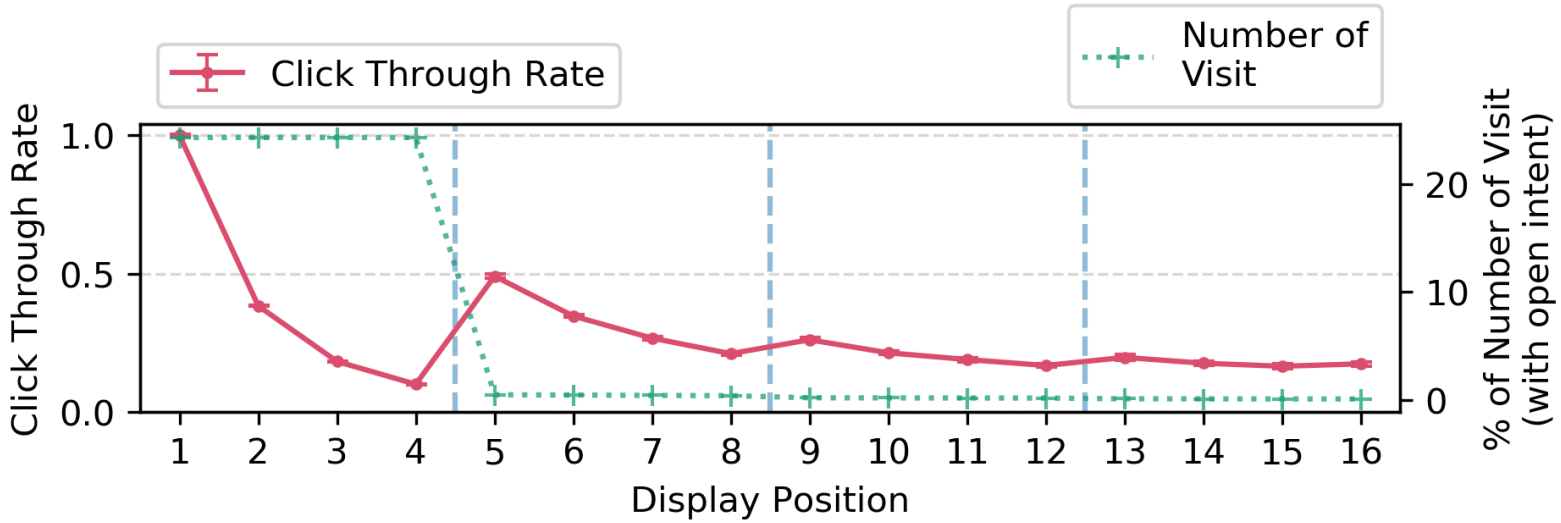
$$\text{Click Through Rate (CTR)} = \frac{\text{Number of Opening a File from RDP}}{\text{Number of Visits}}$$

$$\text{Time to Open (TTO)} = \text{Time Elapse between Visit and File Open from RDP}$$

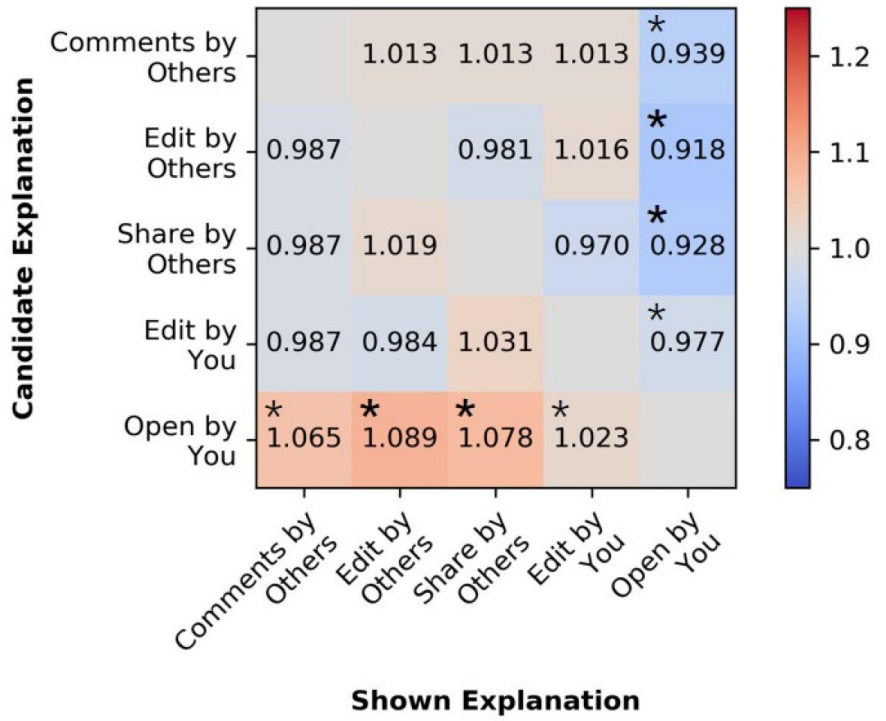
$$\text{Recognize Rate (RR)} = \frac{\text{Number of Opening a File from RDP}}{\text{Number of Opening a File shown in RDP}} \quad \frac{Ⓒ}{Ⓑ}$$

Document Recommendations: Interactions

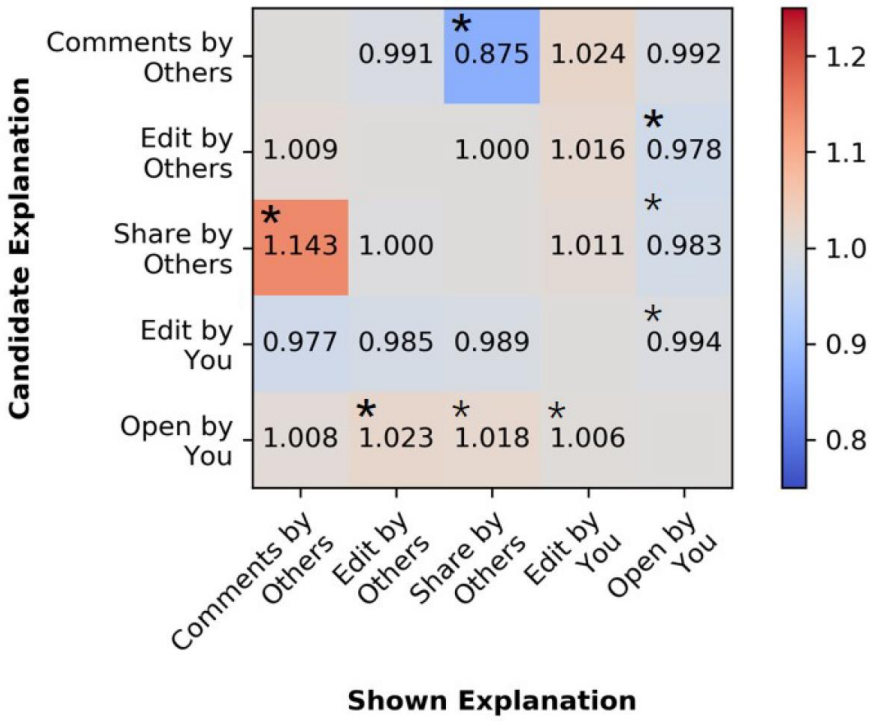
- CTR decrease from the Position 1 (leftmost) to 4 (right most)
- Most of the people do not visit the “next page” of the RDP, but once they go there, there is a jump up of the CTR



Large-Scale Log-Based Experimental Study: Document Explanations



(a) Pairwise Comparison on CTR



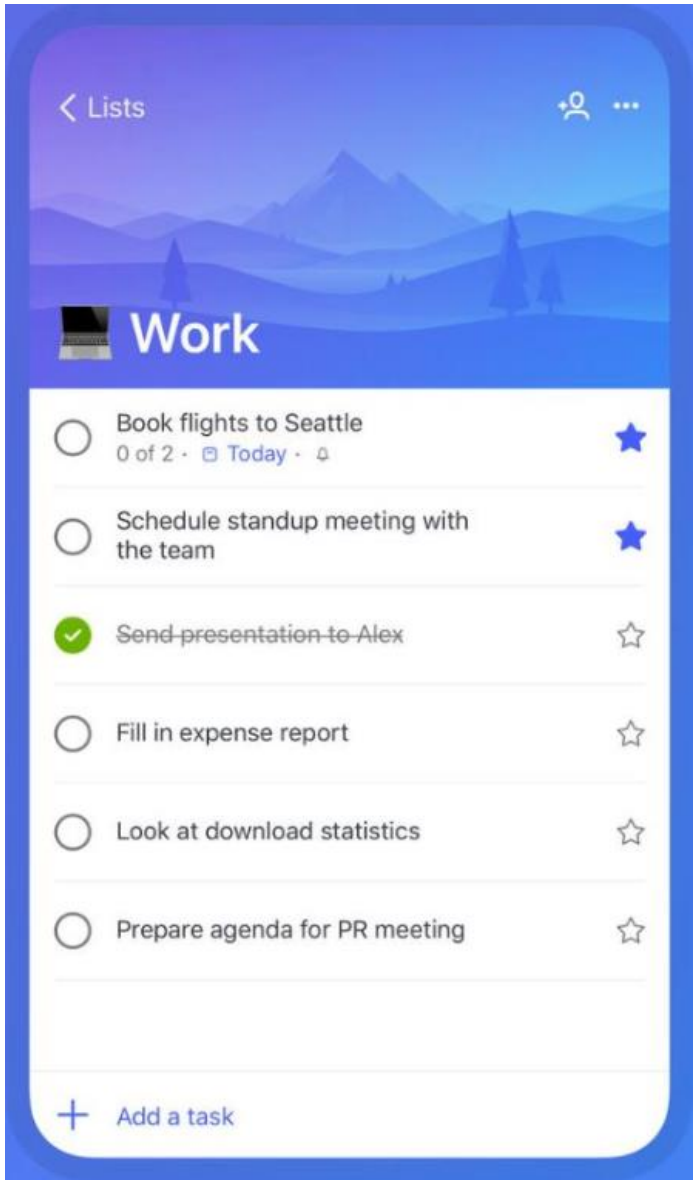
(b) Pairwise Comparison on RR

Task Intelligence

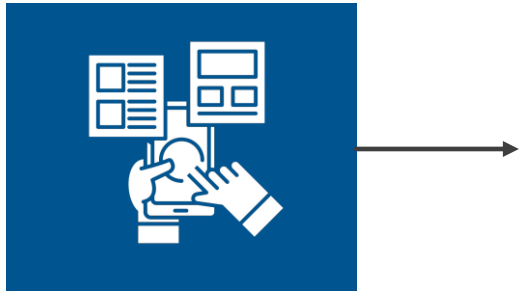
Challenges in Task Management

- Intelligent systems (digital assistants, etc.) store / remind users about tasks
- Tasks can be explicitly specified or inferred
- Users still face several challenges. For example:

    1. Task Scheduling – Duration Estimation
    2. Task status Tracking – Task Auto Deprecation



User Interactions as labels for ML Models



User Interactions

(x_u, yu)

Personal History

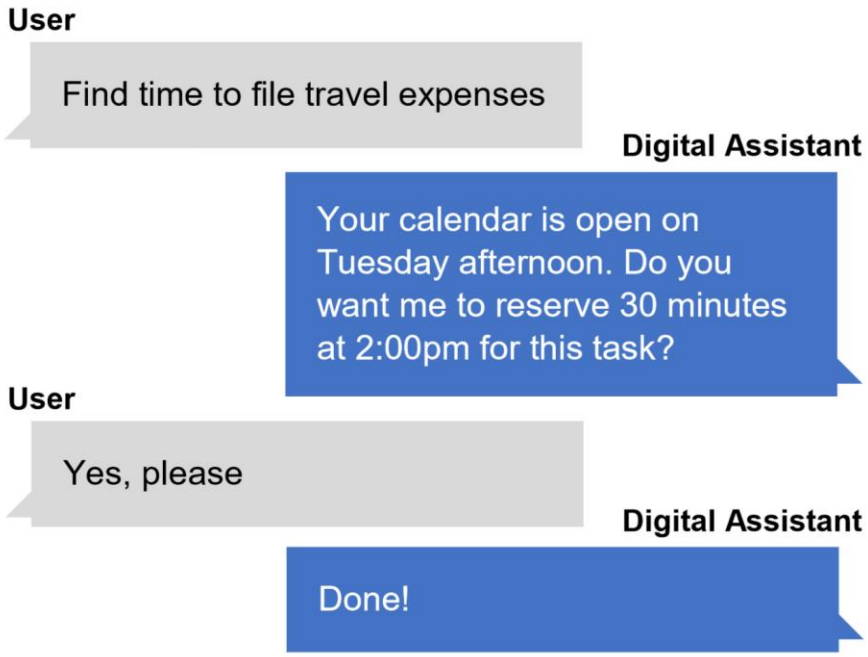
(x_g, yg)

Collective Behavior

Train Task-specific model

Task Duration Estimation

- Time estimation occurs often (every task needs this)
- People struggle to estimate how long tasks take, esp. for new tasks
- Forecast durations using large-scale data

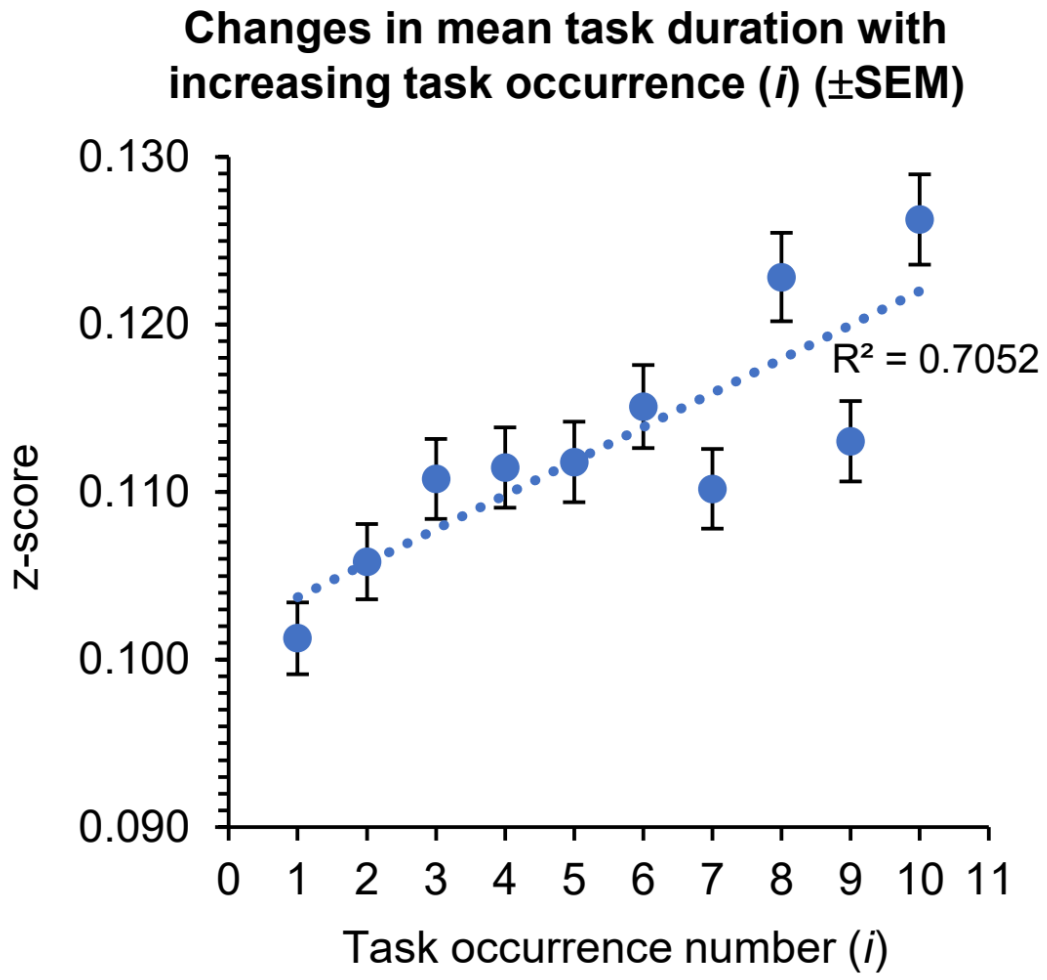


Aspirational goal is building such experience

White and Awadallah, WSDM 2019

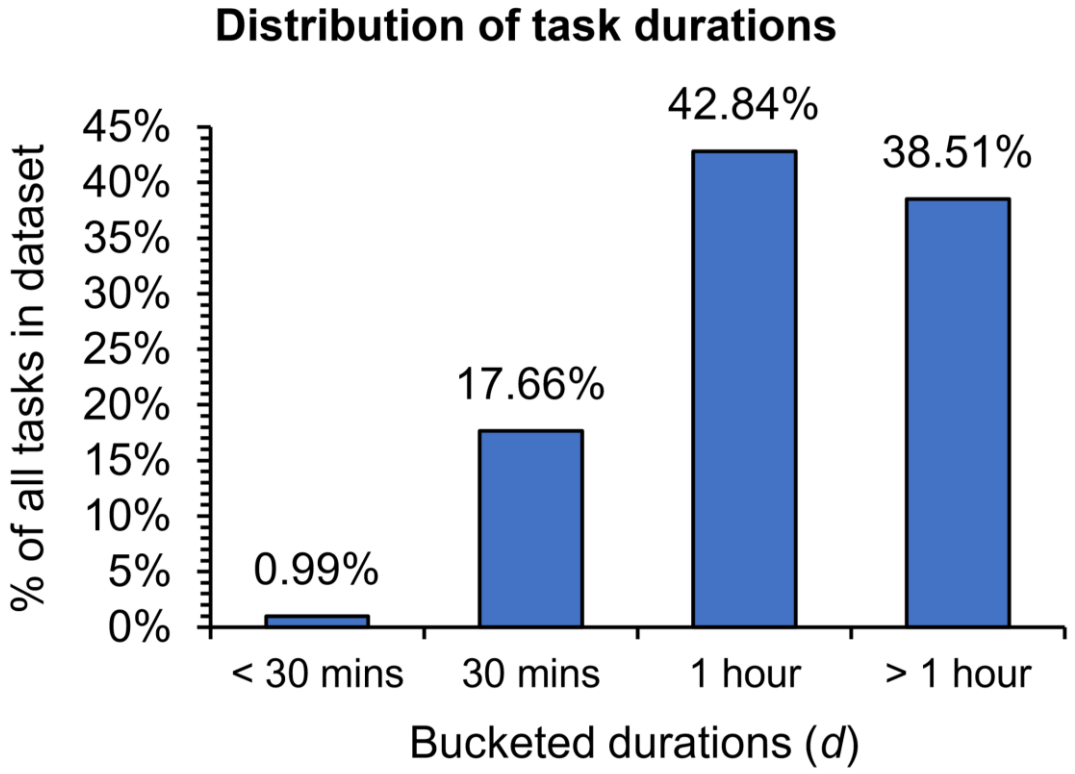
Duration Estimation is hard even for people

- Duration increases with recurrence
- Connections to planning fallacy, etc.



Learning from Personal & Collective User Behavior Data

- Anonymized appointment data
- >3M appointments from 700k users
- Appt subjects are hashed
- Time durations blocked by users
- Grouped into four classes



Personal behavior data is important when available

- History is important
- Locations correlate with longer durations while phone numbers correlate with shorter ones
- More basic language is used to describe clearer/simpler tasks

Generated task features:

Attribute	*r*
Mean duration (per user-task)	+0.41487
Median duration (per user-task)	+0.41056
Mean duration (per task)	+0.35767
Median duration (per task)	+0.35035
Mean duration (per user)	+0.32666
Median duration (per user)	+0.30167
Has location †	+0.23894
Std deviation duration (per user-task)	+0.14959
Std deviation duration (per user)	+0.14557
Maximum token length †	+0.05715
Has country †	+0.05117
Has address †	+0.05024
Minimum token length †	-0.05258
Start minute	-0.05719
Has phone number †	-0.06061
Fraction text stop words (a, the, etc.) †	-0.07673
Total number of stop words in text †	-0.08378
Number of unique stop words in text †	-0.08741
Task popularity (across all users)	-0.13104
Number of action verbs †	-0.14705

user history is important

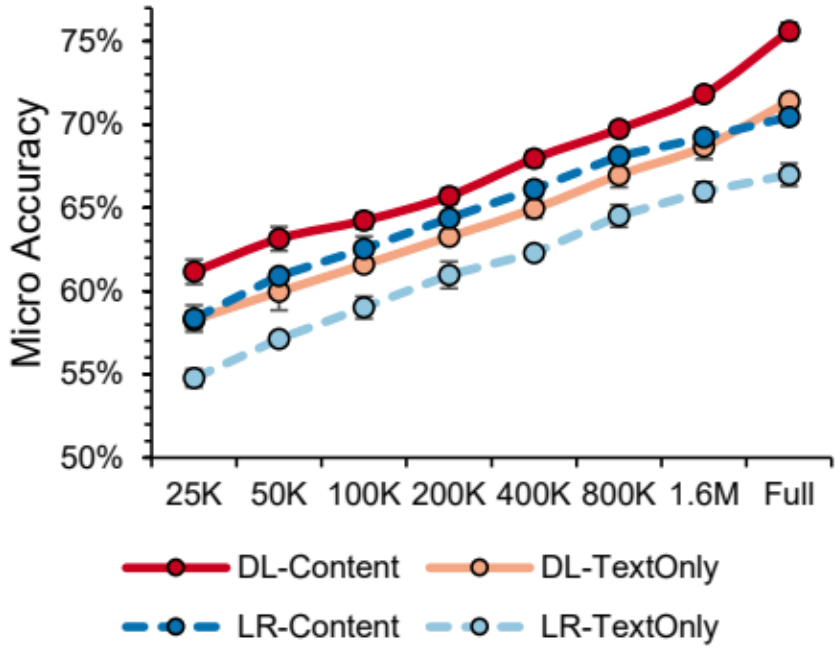
† Non-(time/duration/history) attributes are based on task description

r = correl with appt duration

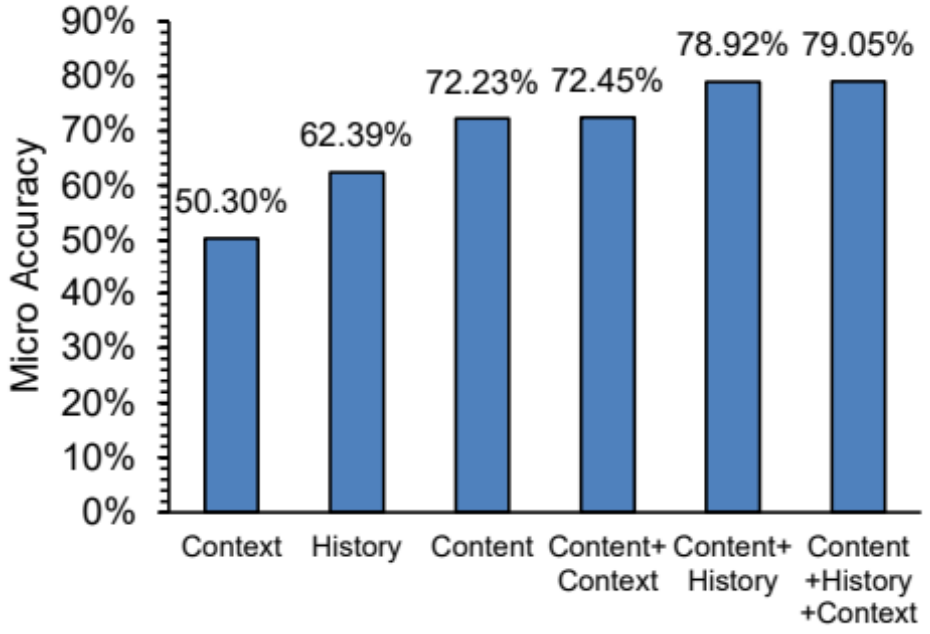
Learning to efficiently predict task duration

- Performance changes with -

+ more data

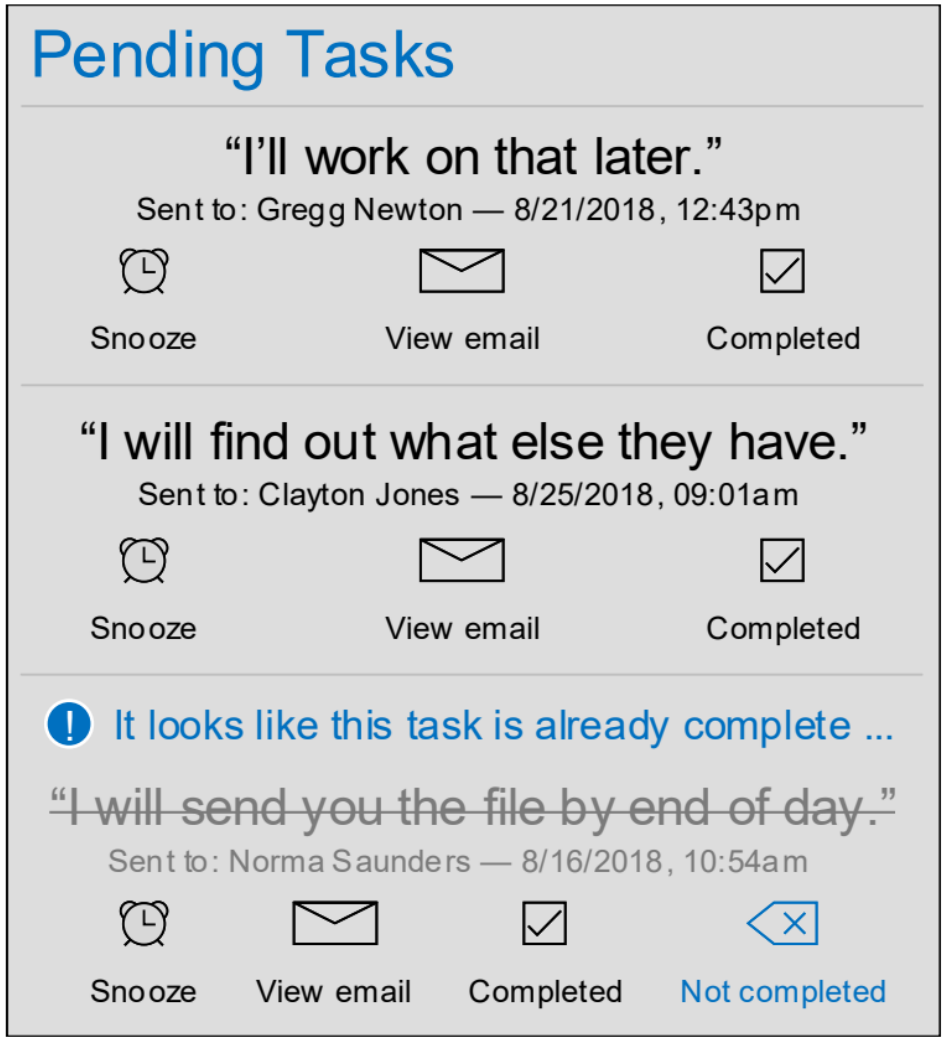


+ more features



Task Auto-Deprecation

- Show pending tasks
- Flag or deprecate completion candidates
- Suppress notifications
- Other applications possible, incl. task ranking, task prioritization, etc.

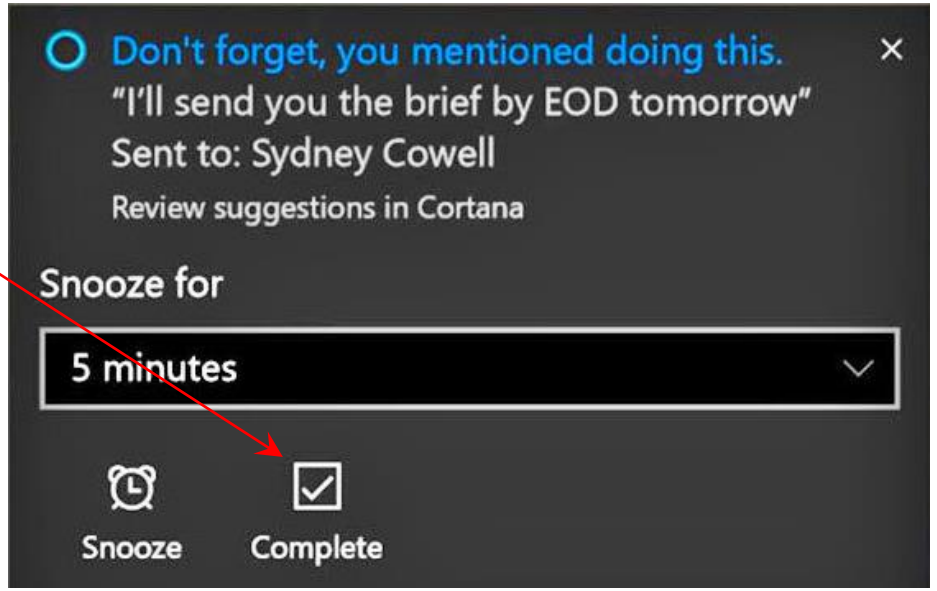


White et al., SIGIR 2019

Learning from User Behavior Data

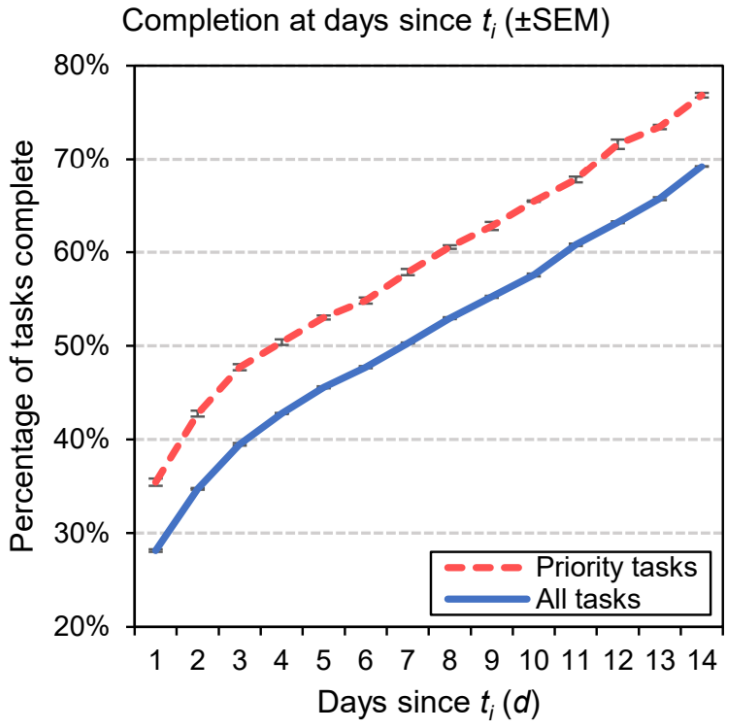
- Offer a feedback affordance for users to indicate task completion

- "Complete" clicks help form ground truth
  - Only says task was completed BY some time, not WHEN the task completion occurred

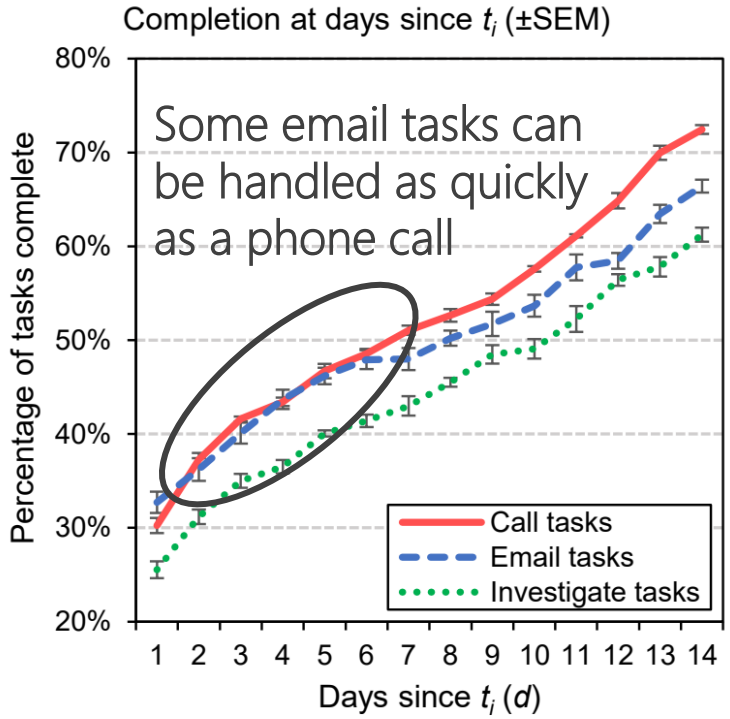


Task Completion Over Time

- Compute fraction of tasks completed at t_n, all tasks and per task type
  - Task type by priority (high-pri language) and by activity (call, email, investigate)



High priority tasks are completed faster



Relative completion timing: Call < Email < Investigate
Connected to avg relative complexity

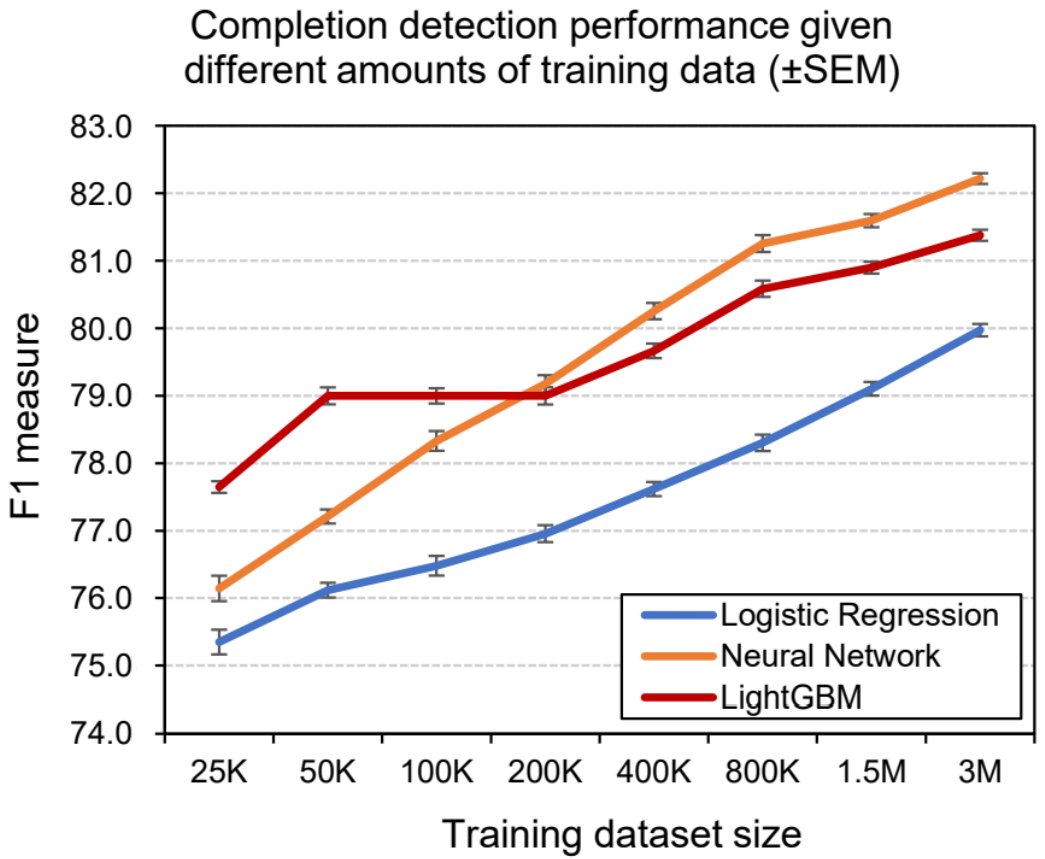
Learning to efficiently predict task completion

- Five feature classes:
  - **Time:** time elapsed since task created
  - **Content:** n-grams, verbs, priority, due date, is conditional, intent, etc.
  - **Email:** subject n-grams (no email body), is reply, number of recipients, etc.
  - **Notifications:** logged Cortana notifications (16% of tasks), num notifications, etc.
  - **User History:** (38% of users), historic tasks, completion time/rates, etc.

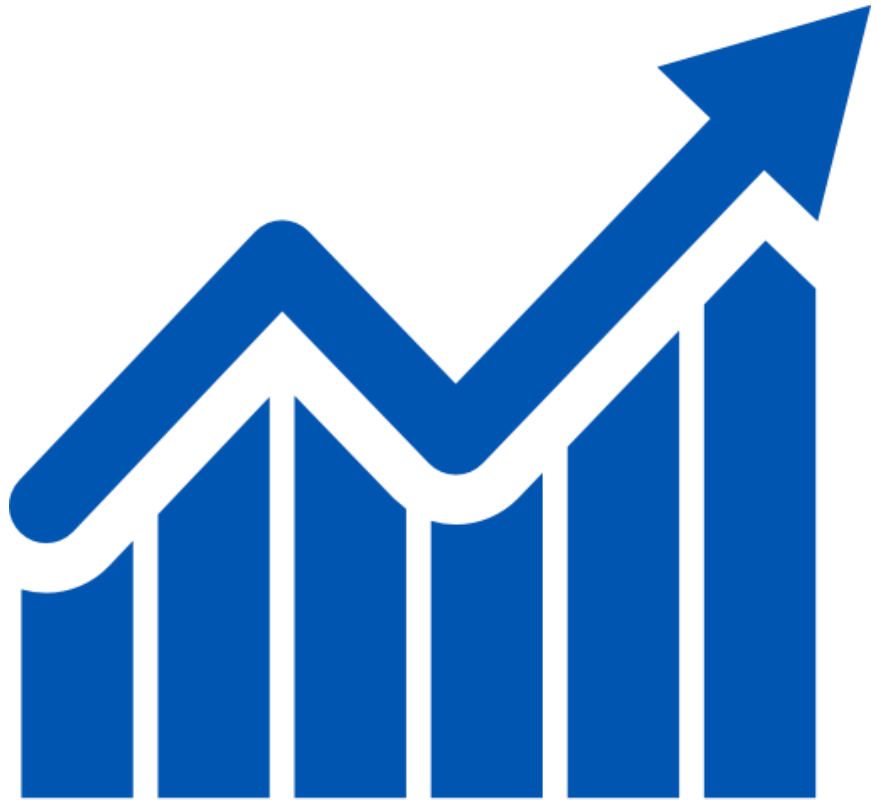
Leverage both content and behavioral features

Learning to efficiently predict task completion

- Strong performance and improves with more data



Challenges and Opportunities



You get what you measure


Experience
Feedback Loop
Data
Engagement

Privacy powers **Trust**

Trust enables **Data**

Data powers **Innovation**

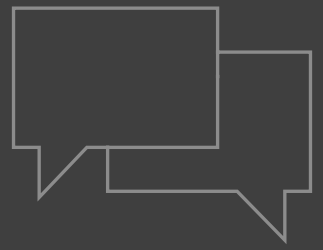
Microsoft

Microsoft Research AI

LANGUAGE AND INFORMATION TECHNOLOGIES

Contributors: Ahmed Awadallah, Paul Bennett , Susan Dumais , Saghar Hosseini, Subho Mukhrejee, Farheen Omar, Bogdan Popp, Robert Rounthwait, Robert Sim, Milad Shokouhi, Wei Wang, Ryen White, Guoqing Zheng

Interns: Farnaz Jahanbakhsh, Yang Liu, Yang Xiao, Orson Xu



Thank you

http://aka.ms/ahmed

Microsoft