

# Leveraging Behavioral Data to Improve Workplace Productivity

Ahmed Hassan Awadallah

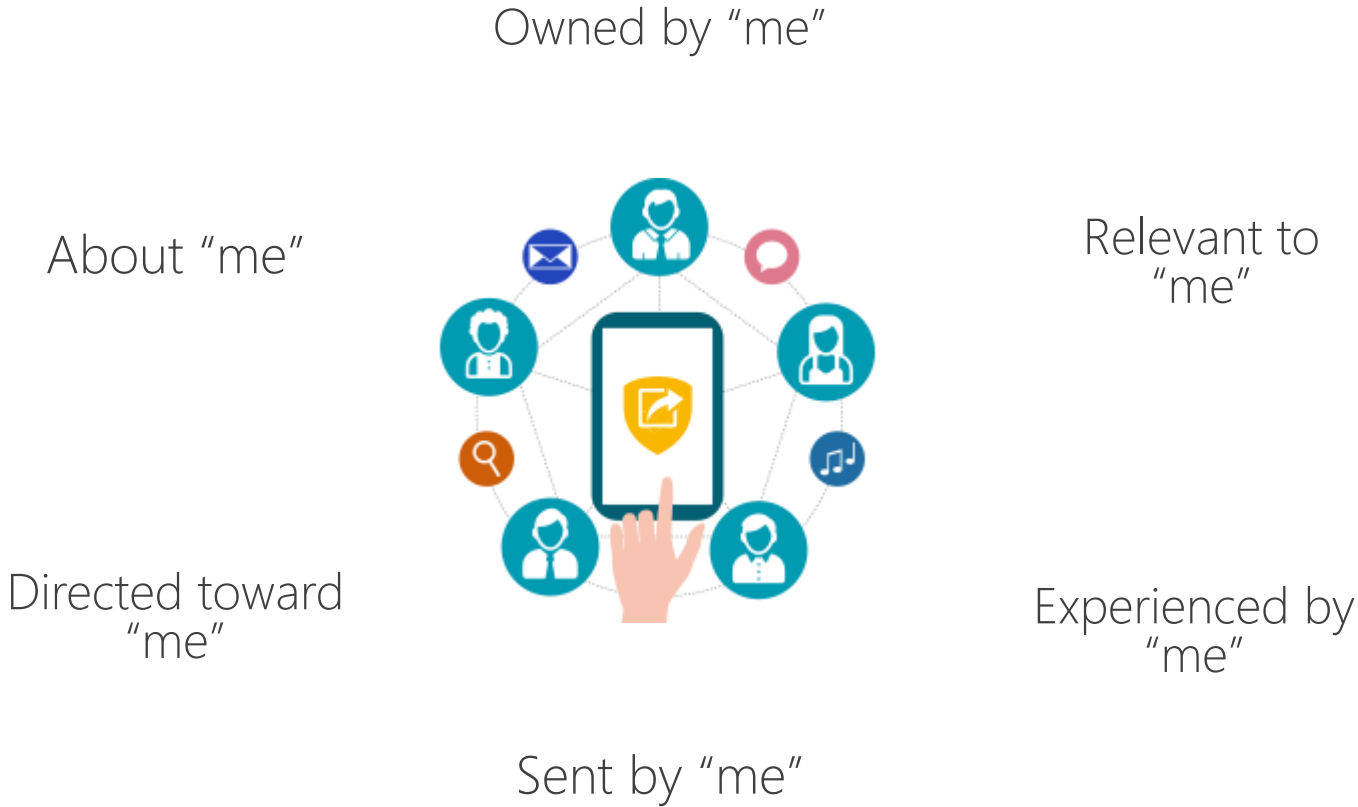
<http://aka.ms/ahmed>

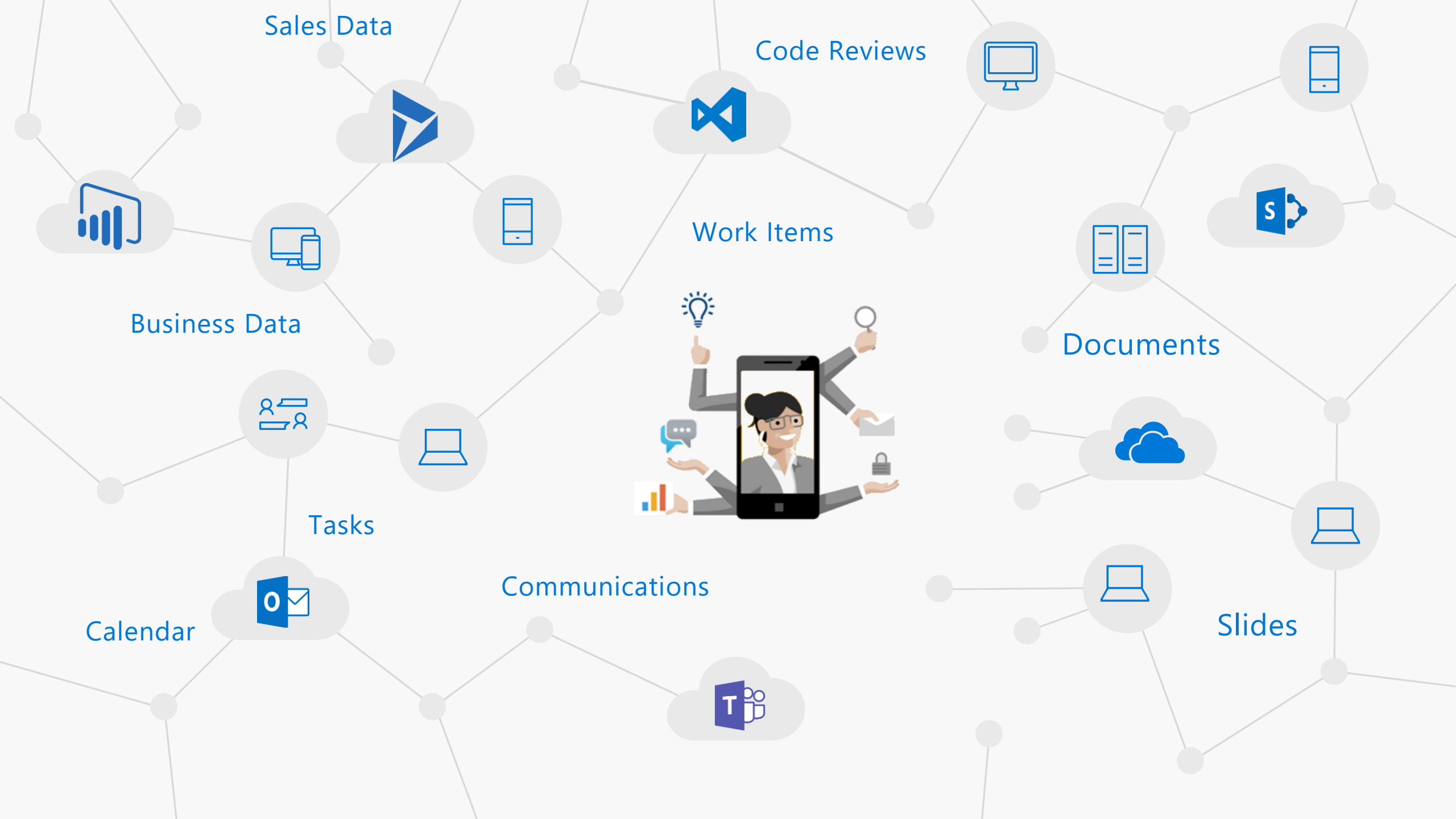


Microsoft Research AI



# What makes Information Personal?





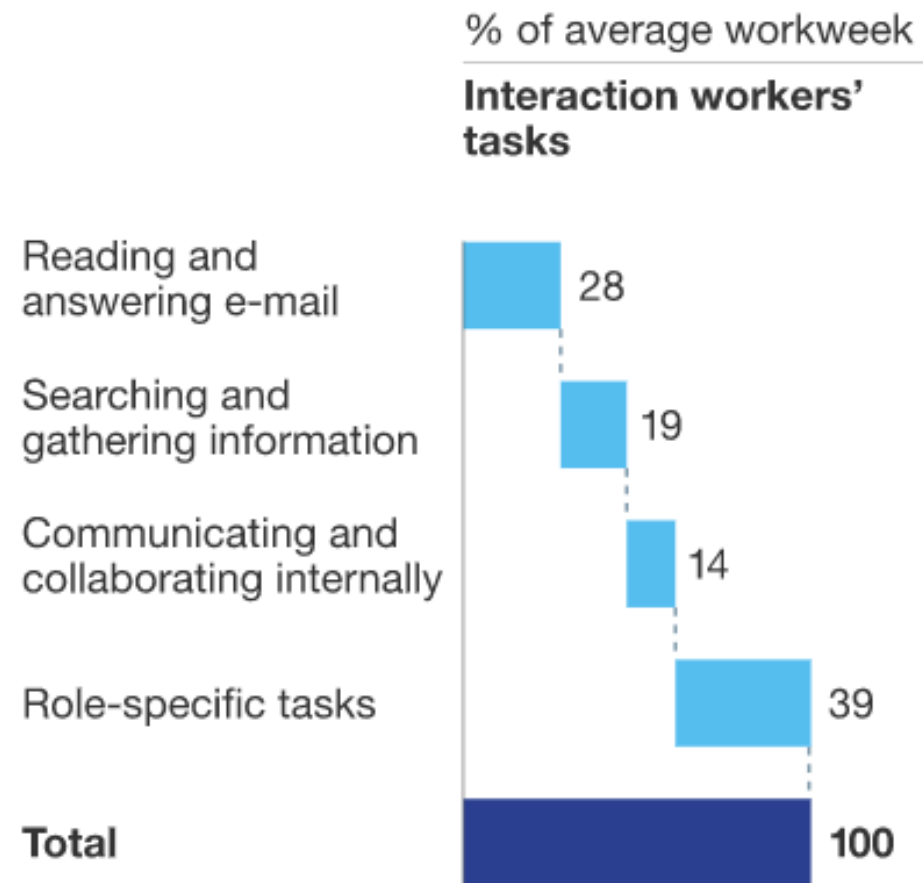
# Enormous Volume of Information Exchanged in the Workplace



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.

# Significant time spent interacting with this information



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

### Meeting Notes



Carlos Slattery

Tue 7/28/2018 5:29 PM

You



Saw your notes. Would you mind sharing your notebook with the group before our next discussion?

Thanks,  
Carlos



Sure thing.

I'm on it.

I'll get back to you.

Report inappropriate text

### Recommended



You edited this



Microsoft Teams Helpdesk Guide f...  
ciaops365e1.sharepoint.c...



You edited this



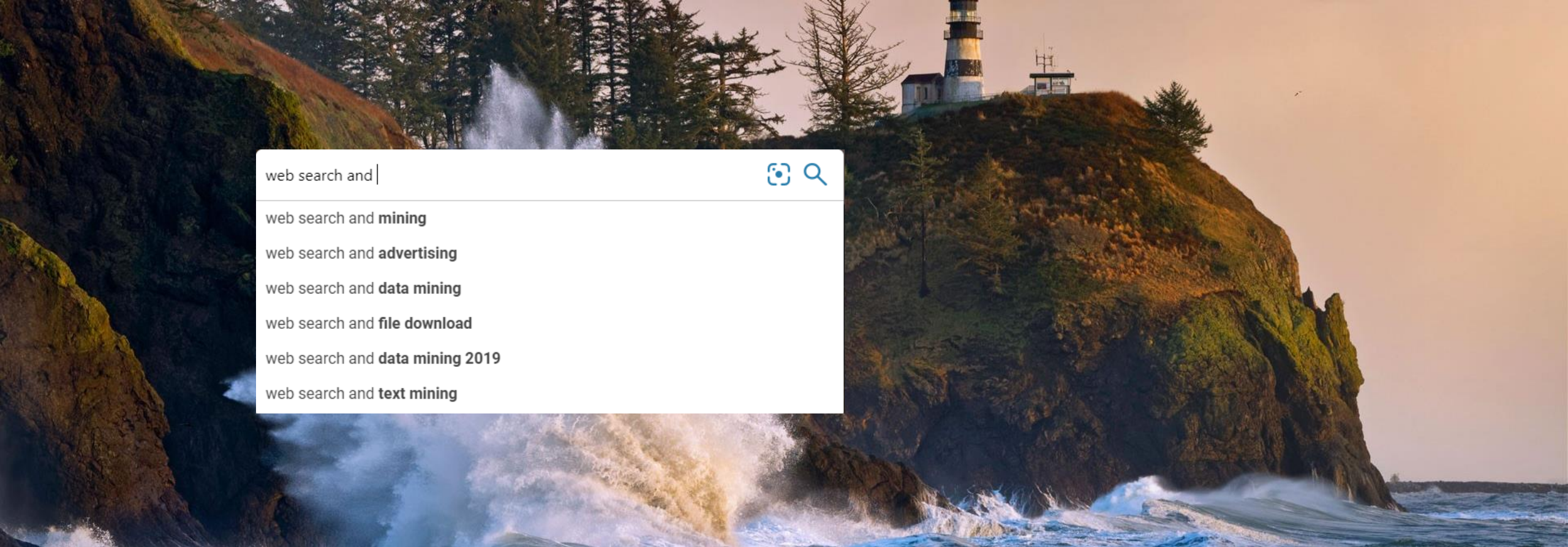
Microsoft Teams Partner Opportun...  
ciaops365e1.sharepoint.c...




## The World Wide Web

Trillions of pages indexed.  
Billions of queries per day.





web search and | 

---

- web search and **mining**
- web search and **advertising**
- web search and **data mining**
- web search and **file download**
- web search and **data mining 2019**
- web search and **text mining**

# Modem Web Search Engines

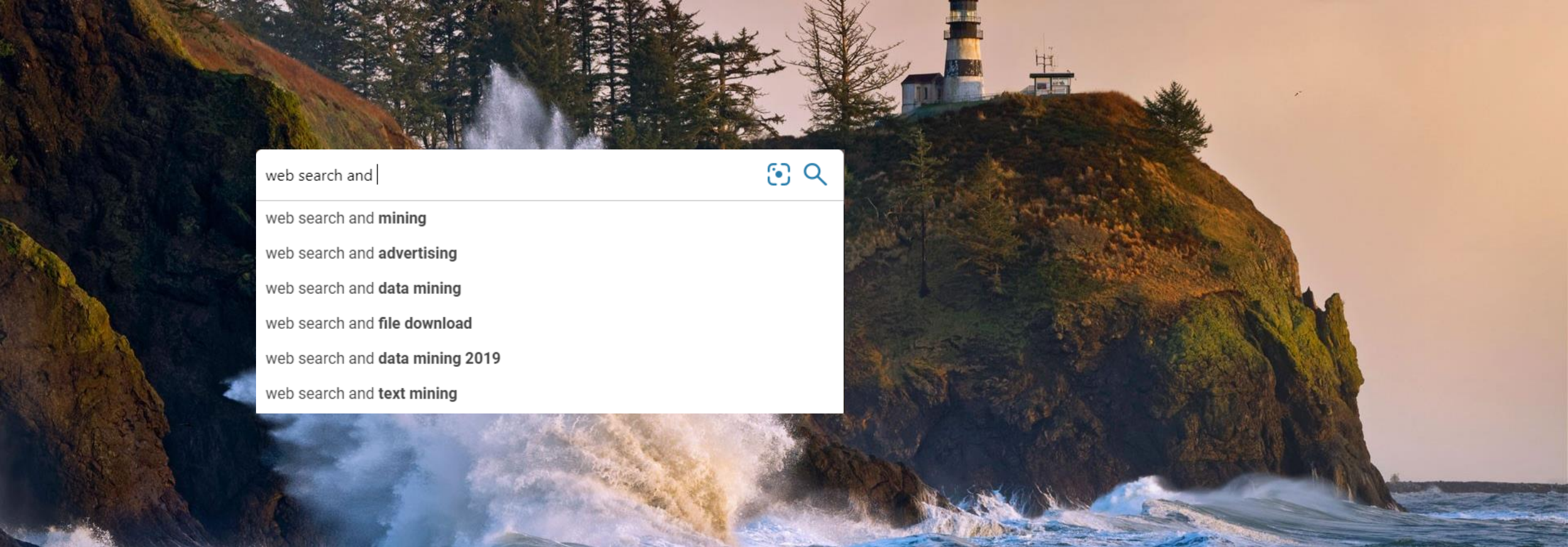
User Behavioral Data



# 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





web search and |  

---

web search and **mining**

web search and **advertising**

web search and **data mining**

web search and **file download**

web search and **data mining 2019**

web search and **text mining**

## How can behavioral data help?

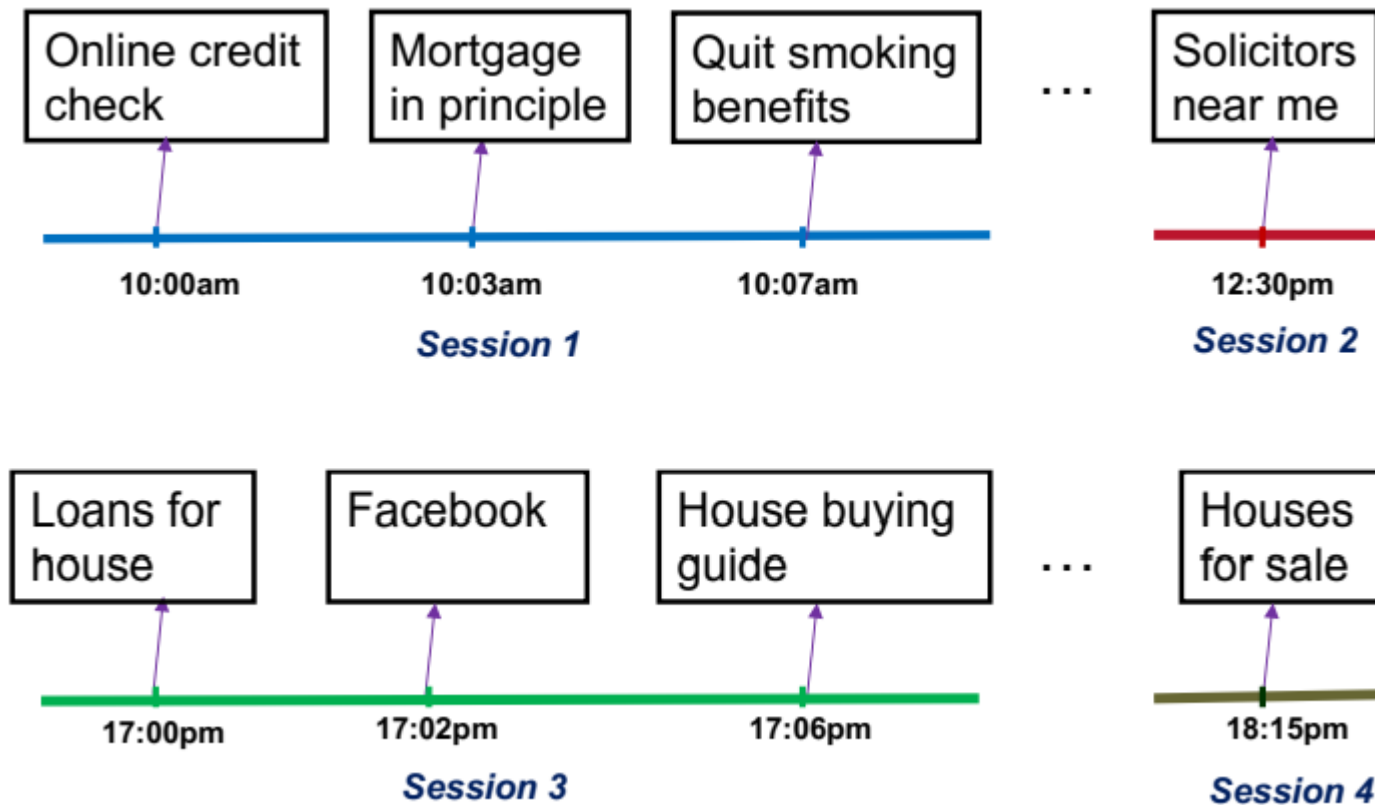
Understand users' intents and needs

Learn to build better ML models

Measure user satisfaction and engagement

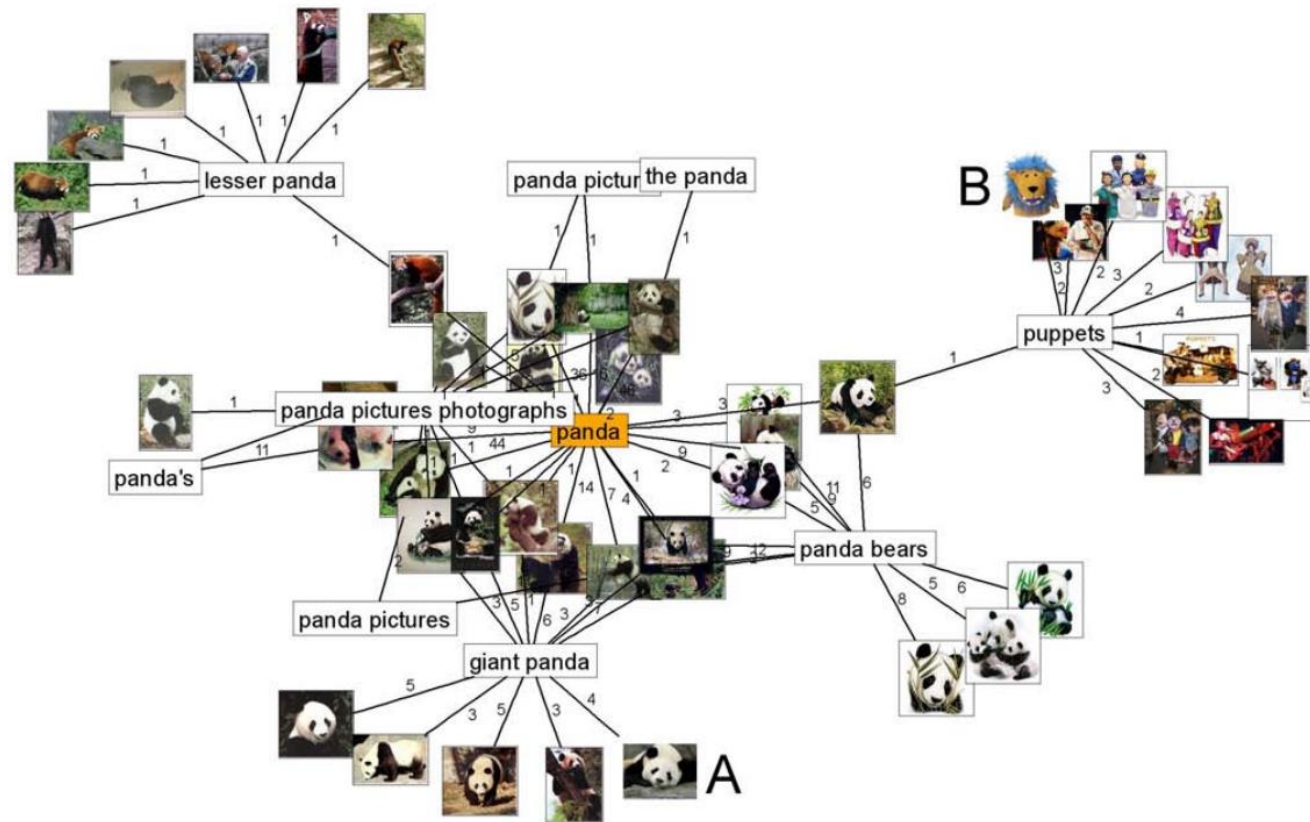
# How is Behavioral data used in Web Search?

## Understand Tasks and Intent from Query Logs



# How is Behavioral data used in Web Search?

## Learn to Rank using the Click Graph



# How is Behavioral data used in Web Search?

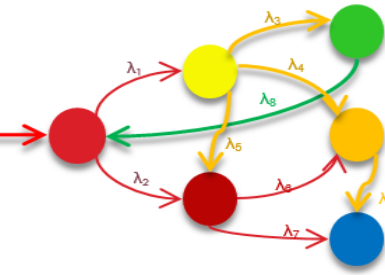
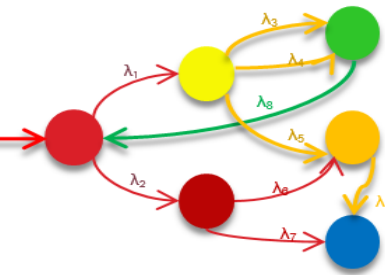
## Measure User Satisfaction

### Satisfied Trails

Trail 1: Q 4s RL 1s SR 53s SR 118s END  
Trail 2: Q 3s Q 5s SR 10s AD 44s END  
Trail 3: Q 4s RL 1s SR 53s SR 118s END  
Trail 4: Q 3s Q 5s SR 10s AD 44s END  
.....  
Trail n: Q 4s RL 1s SR 53s SR 118s END  
Trail n-1: Q 3s Q 5s SR 10s AD 44s END

### Dissatisfied Trails

Trail 1: Q 4s RL 1s SR 53s SR 118s END  
Trail 2: Q 3s Q 5s SR 10s AD 44s END  
Trail 3: Q 4s RL 1s SR 53s SR 118s END  
Trail 4: Q 3s Q 5s SR 10s AD 44s END  
.....  
Trail n: Q 4s RL 1s SR 53s SR 118s END  
Trail n-1: Q 3s Q 5s SR 10s AD 44s END





**Improving Productivity by leveraging Behavioral Data**

# Three Productivity Applications



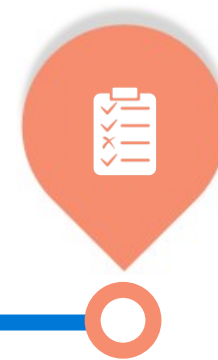
## Communications

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

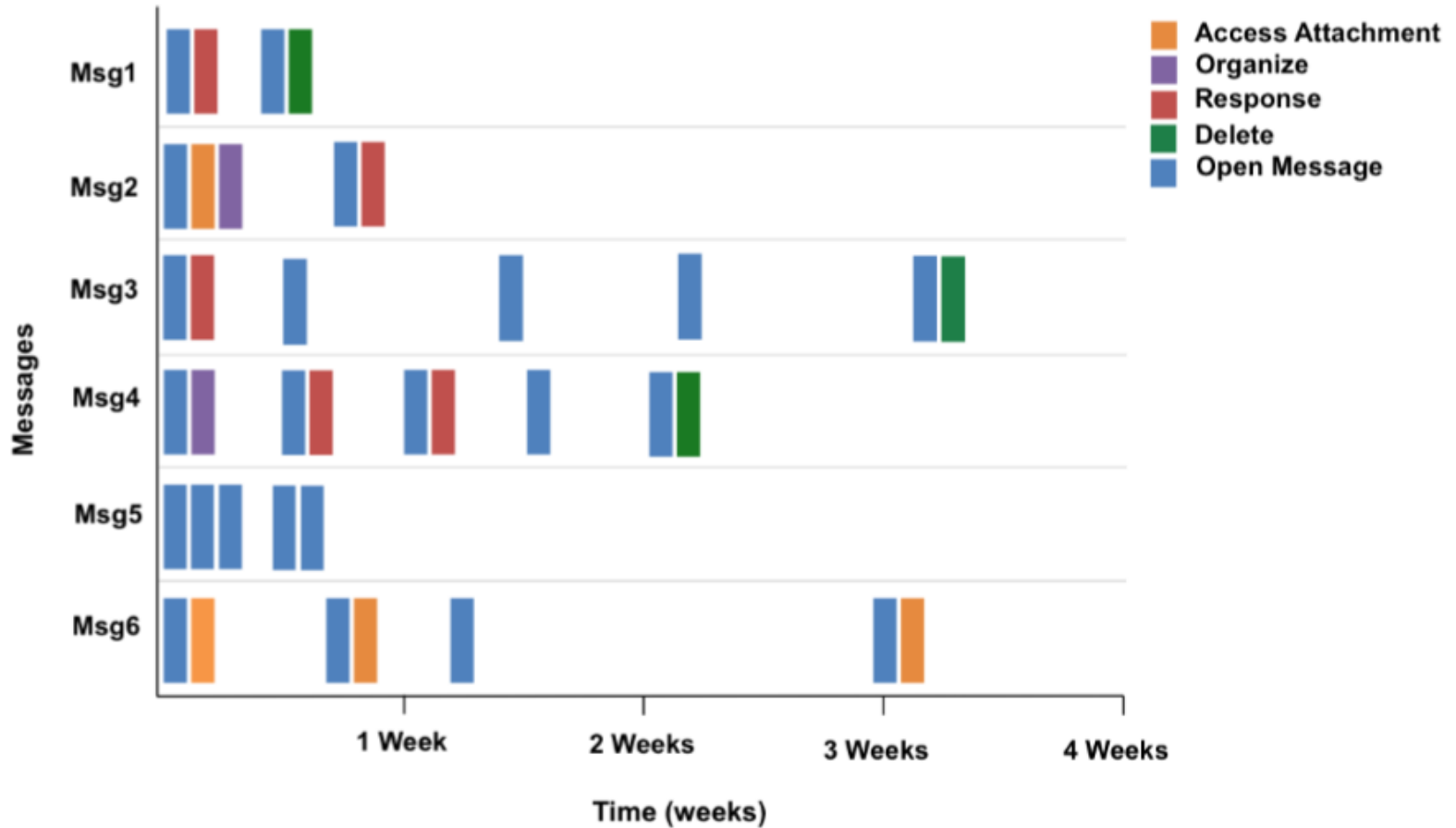


# Email Intelligence

# The Lifetime of an Email Message

Different lifetime spans and different interaction patterns

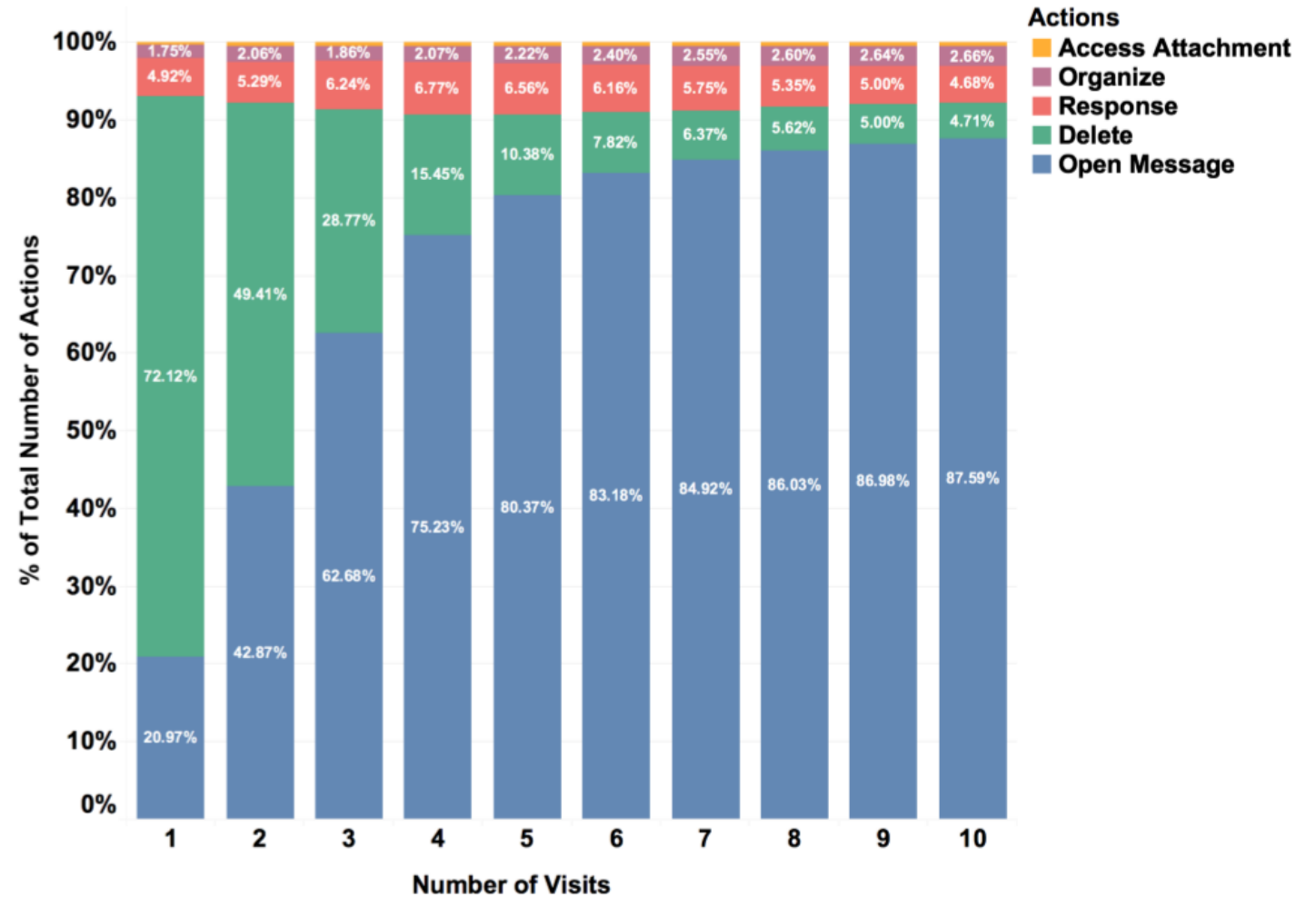
Revisiting to **take an action** or **retrieve information**



# 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

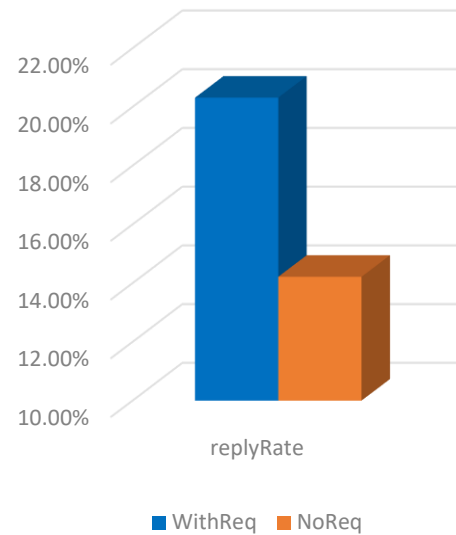


# Let's focus on responding to a message

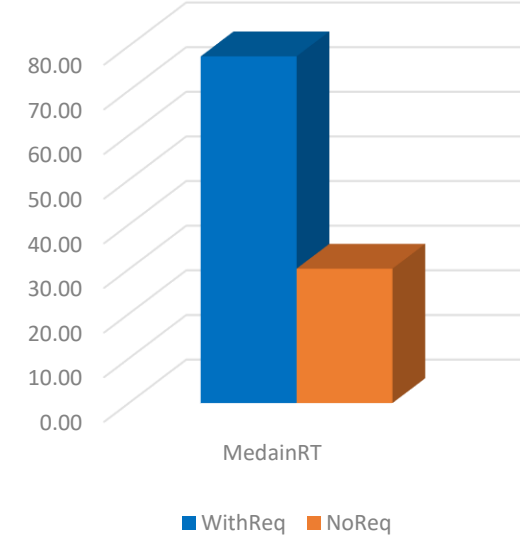
If the message is requesting an action from the recipient, it is more likely to receive a reply

But it is also more likely to receive a slow reply


Effects of with/no Requests on Reply Rate



Effects of with/no Requests on Median Reply Time




# Message to Action Recommendation

 **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](#)



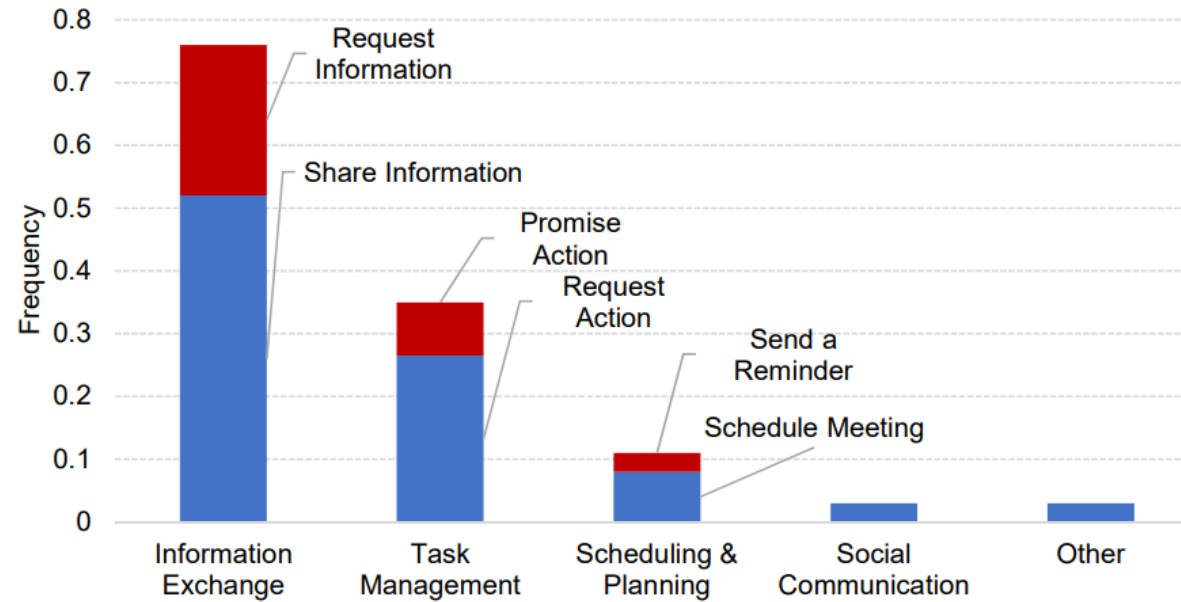




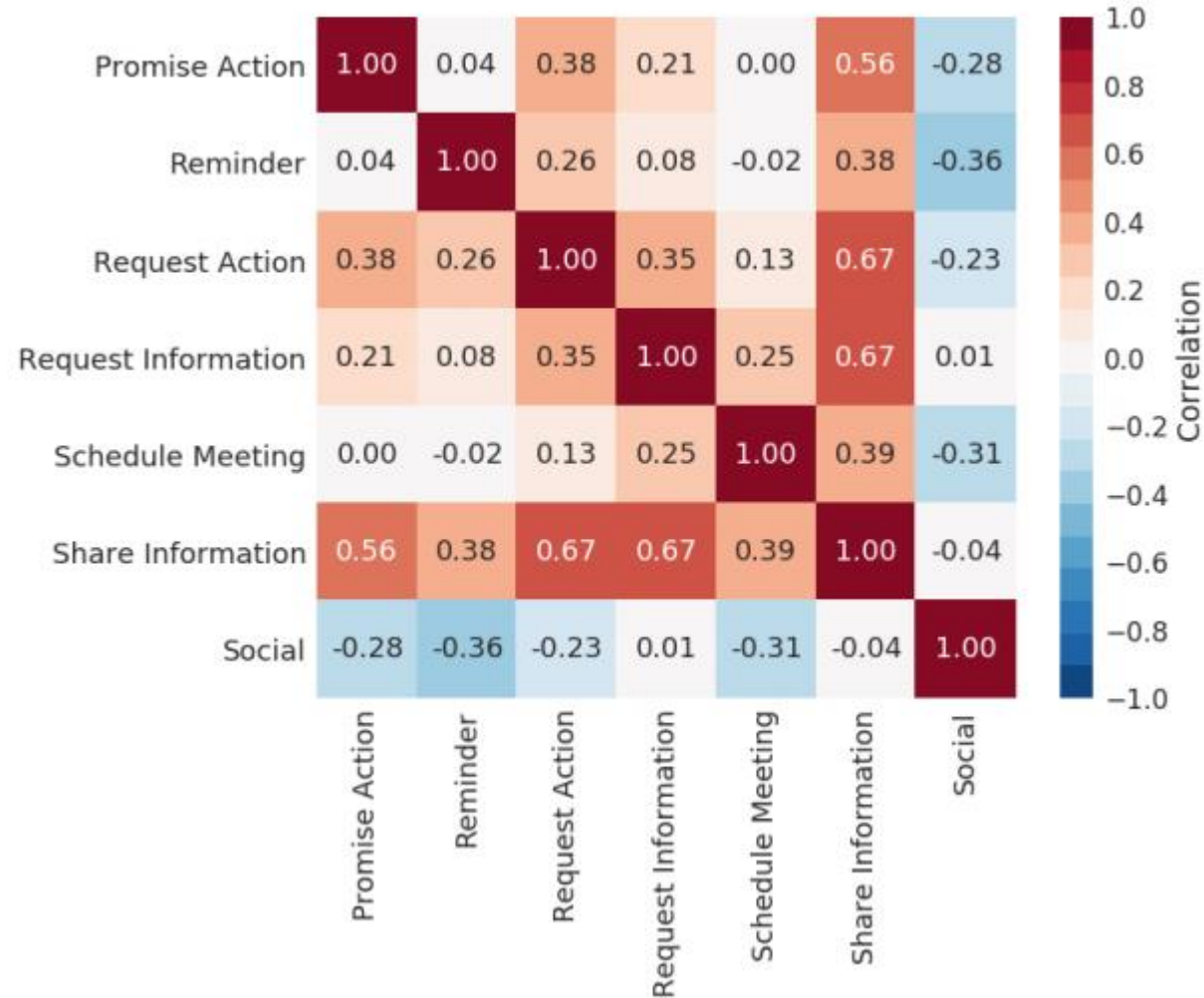




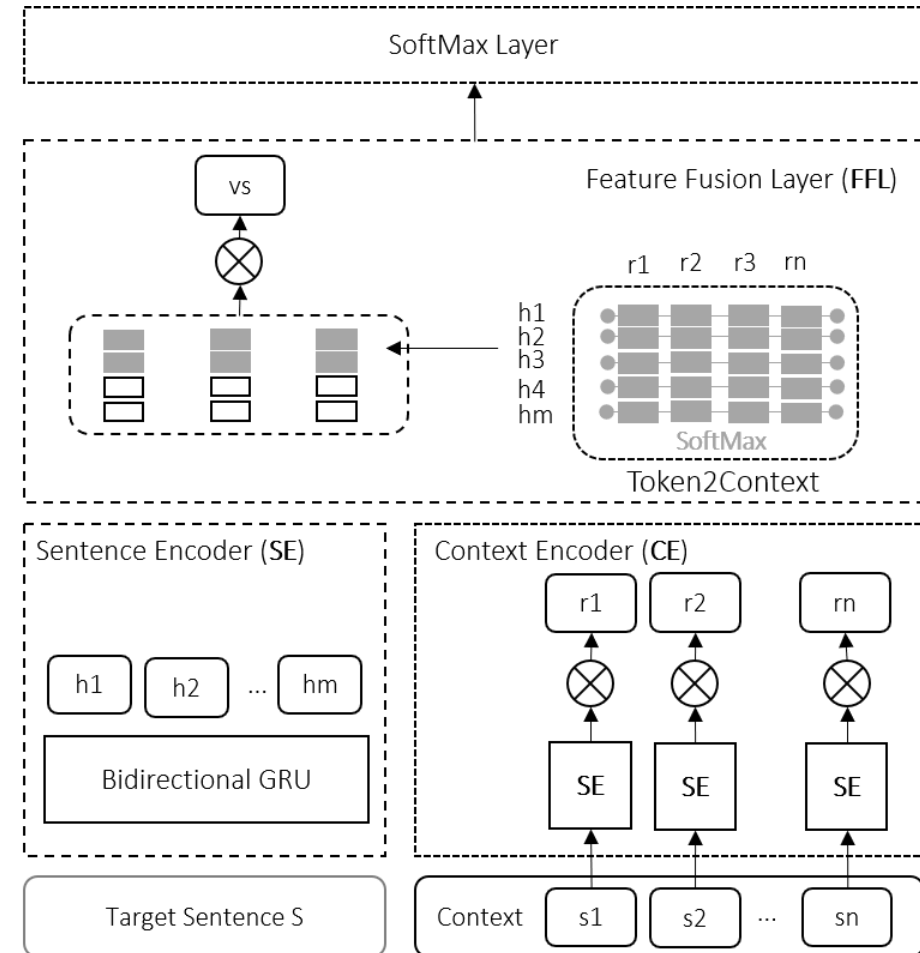
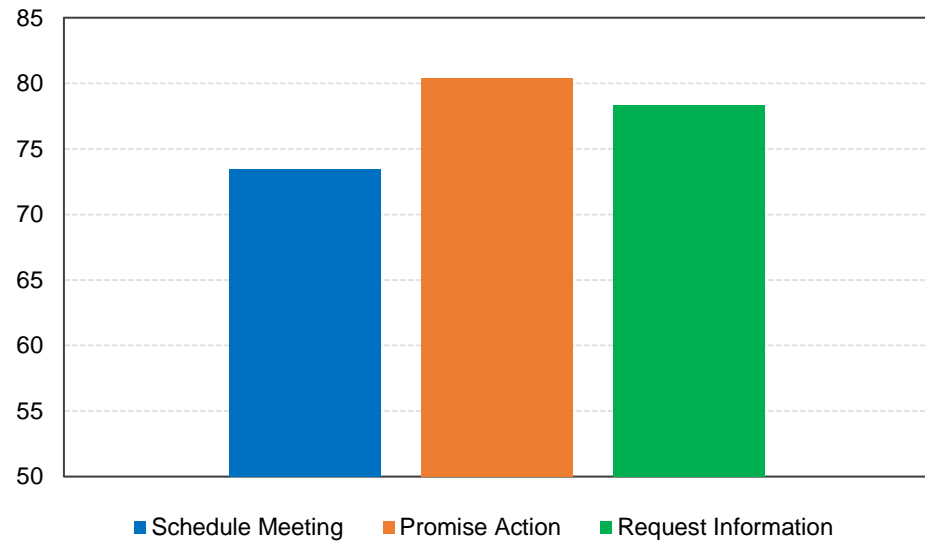
# Message to Action: Understanding Intent



# Message to Action: Understanding Intent



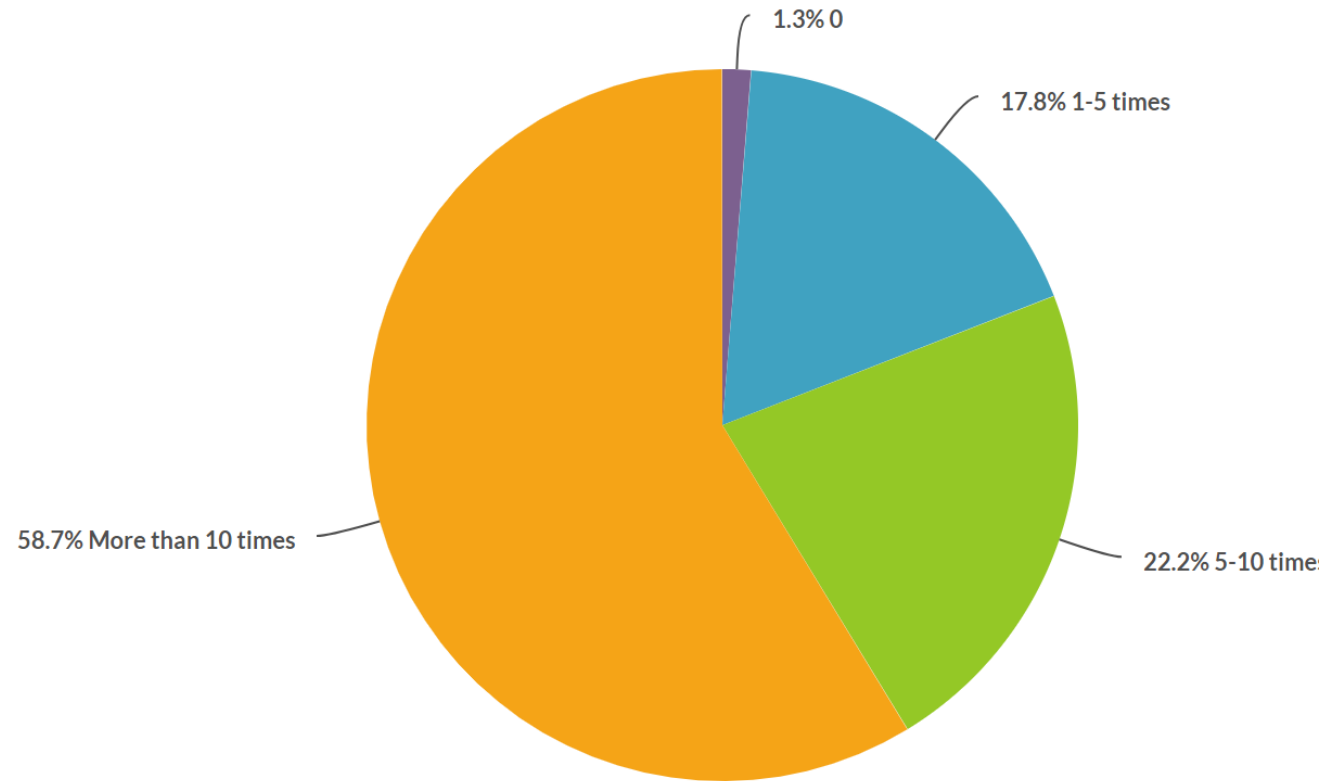
# Message to Action: Recommendation



# 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

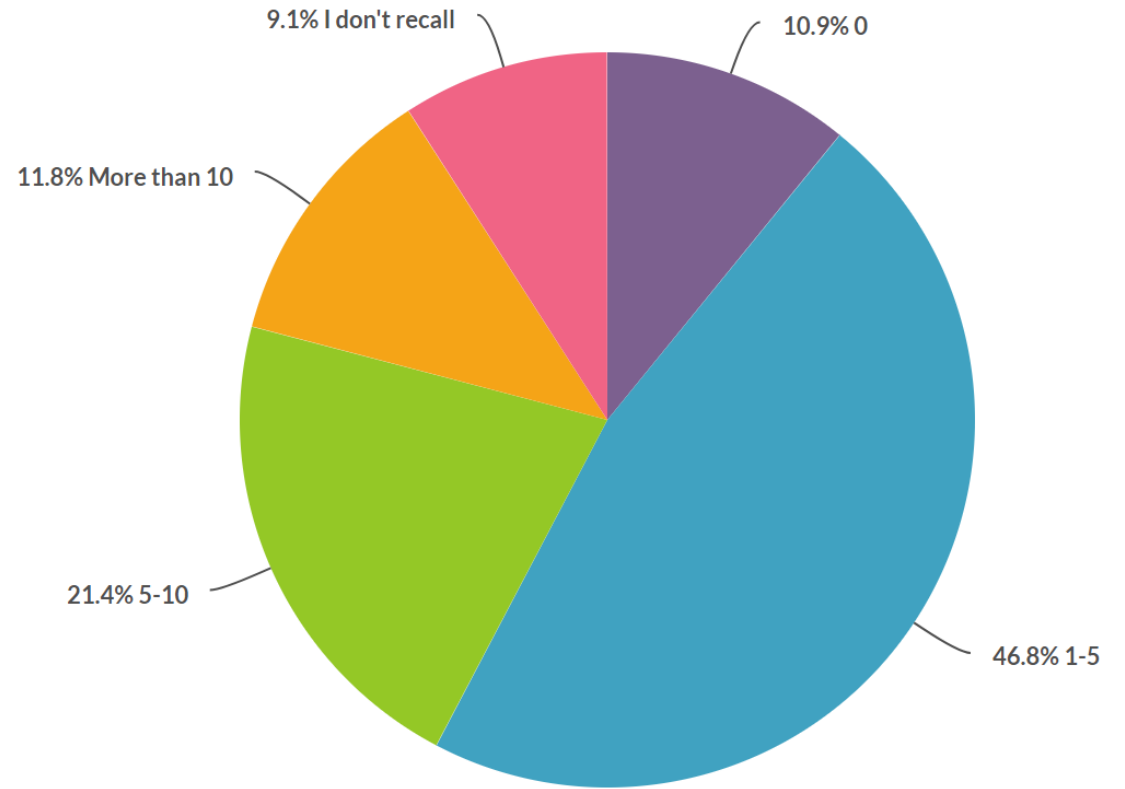


How often do you need to go back to threads because they contain Questions/Answers?

# Question/Answer Exchange over Email

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

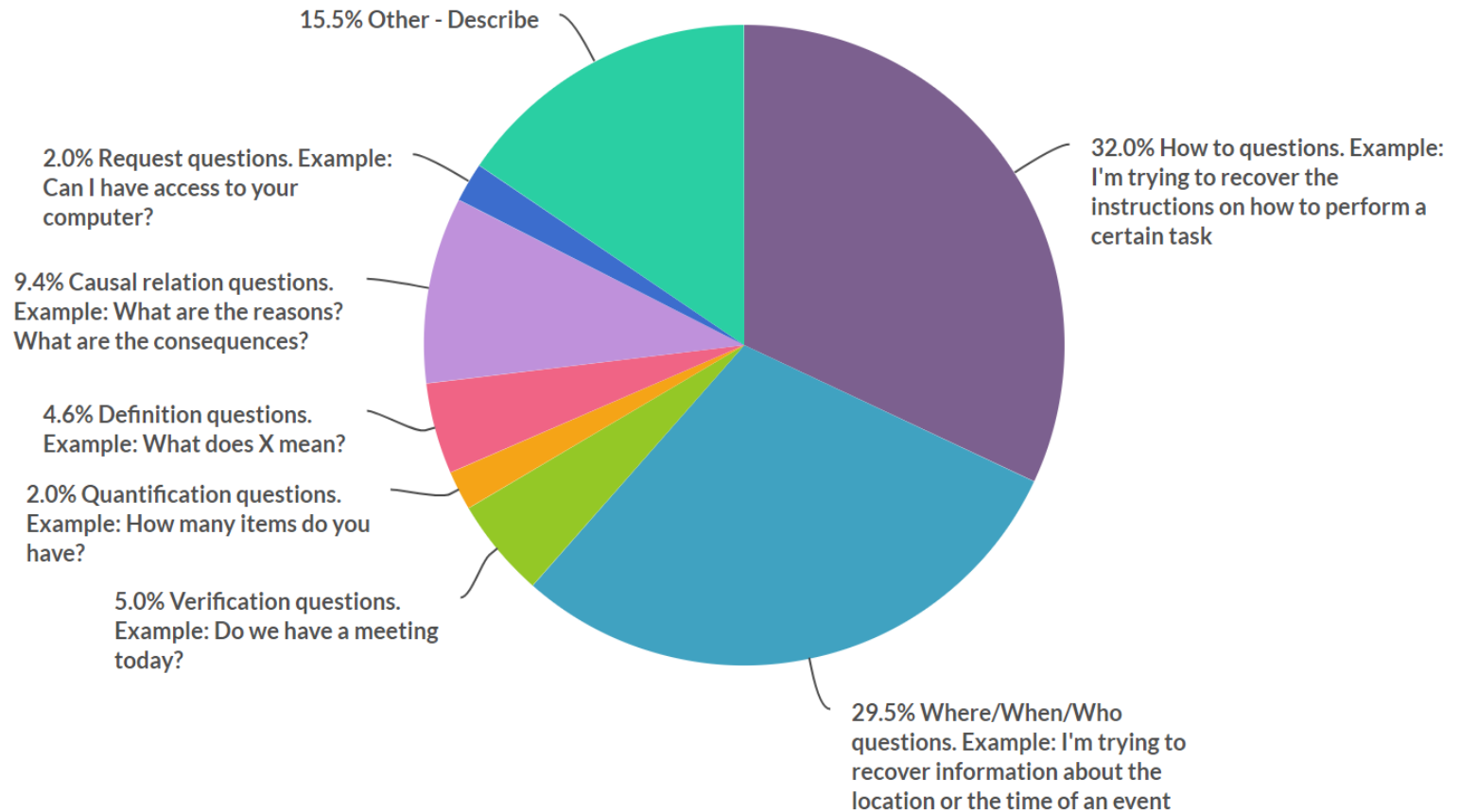
— more than 5 times a month for 33% of them



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

# Question/Answer Exchange over Email

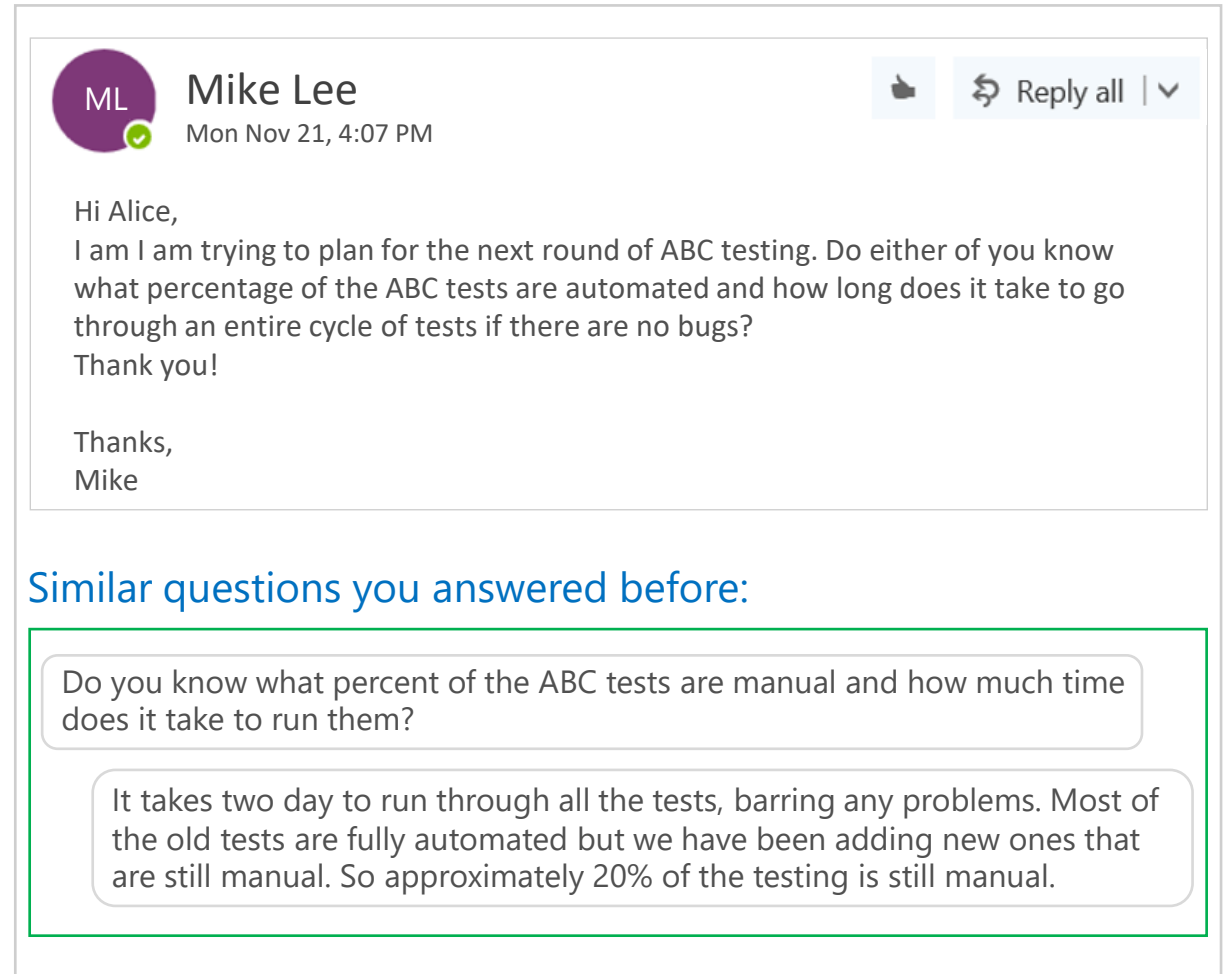
How-to and When/Where/Who questions are the most popular types people need to get back to



# Question/Answer Exchange over Email

Reply Support: Help users answer questions they already answered before

Search Support: Help users locate answers they received before



The screenshot displays an email interface. At the top, a purple circular profile picture with the initials 'ML' and a green checkmark is next to the name 'Mike Lee' and the timestamp 'Mon Nov 21, 4:07 PM'. To the right are icons for thumbs up and a 'Reply all' button with a dropdown arrow. The main body of the email contains the following text:

Hi Alice,  
I am I am trying to plan for the next round of ABC testing. Do either of you know what percentage of the ABC tests are automated and how long does it take to go through an entire cycle of tests if there are no bugs?  
Thank you!

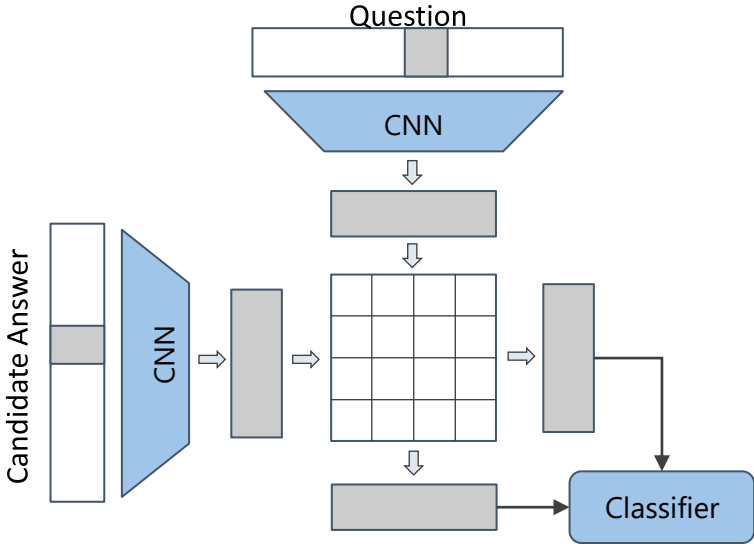
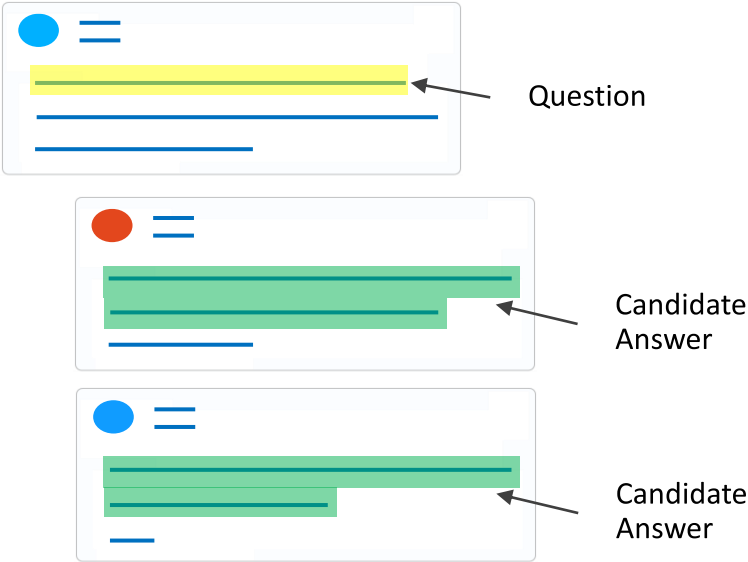
Thanks,  
Mike

Below the email content is a section titled 'Similar questions you answered before:' in blue text. This section is enclosed in a green border and contains two rounded rectangular boxes. The first box contains the question: 'Do you know what percent of the ABC tests are manual and how much time does it take to run them?'. The second box contains the answer: 'It takes two day to run through all the tests, barring any problems. Most of the old tests are fully automated but we have been adding new ones that are still manual. So approximately 20% of the testing is still manual.'



# Question/Answer Exchange over Email

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



# Document Recommendation

# Document Recommendations

The screenshot displays the Office 365 home page. At the top, there's a search bar and navigation icons. Below, the 'Apps' section shows icons for Outlook, OneDrive, Word, Excel, PowerPoint, OneNote, SharePoint, Teams, and Yammer. The 'Documents' section features 'Upload and open...' and 'New' buttons. A red dashed box highlights the 'Recommended' section, which contains four document cards with red boxes around their headers:

- Card 1:** "You edited this Thu at 1:13 PM" with a pencil icon. Below is a document thumbnail for "Data for the WWW 2020 Paper".
- Card 2:** "Jack share this file with you Wed at 11:41 AM" with a share icon. Below is a document thumbnail for "Slides for the WWW 2020 Paper".
- Card 3:** "Tom commented on this file Mon at 3:25 PM" with a comment icon. Below is a document thumbnail for "Draft of the WWW 2020 Paper".
- Card 4:** "You recently opened this Jun 20" with a folder icon. Below is a document thumbnail for "Sample of a WWW 2019 Paper".

A red dashed box labeled "Recommended Document Pane (RDP)" encompasses these cards. To the right, a "Recommendation Explanation" pane is shown, containing a detailed text explanation of the recommendation logic, including sections for "ABSTRACT" and "CONCEPTS". A "Navigation Arrow" (a double-headed arrow) is positioned between the RDP and the explanation pane. At the bottom, the "Recent" section lists documents like "Slides for the research group meeting" and "Slides for brain storming".

# Types of Behavioral Data

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

# Types of Behavioral Data

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

# Types of Behavioral Data

	Observational	Experimental
<b>Lab Studies</b> <i>Controlled tasks, in laboratory, with detailed instrumentation</i>	In-lab behavior observations	In-lab controlled tasks, comparisons of systems
<b>Panel Studies</b> <i>In the wild, real-world tasks, ability to probe for detail</i>	Ethnography, case studies, panels (e.g., Nielsen)	Clinical trials and field tests
<b>Log Studies</b> <i>In the wild, no explicit feedback but lots of implicit feedback</i>	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



## A Brief History of the Middle Earth

Author is Samwise Gamgee  
SharePoint

Shared with you by Gandalf the Grey







## A Brief History of the Middle Earth

Author is Samwise Gamgee  
SharePoint


Shared with you by Gandalf the Grey

### Overview of past actions

-  Last time, you edited this  
2 days ago
-  **Charlie Davidson** and **you** edited this  
4 hours ago
-  Eve Frank, Charlie Davidson and 4 others commented on this  
3 days ago
-  **Eve Frank, Ivan Judy** and **2 others** mentioned **you**  
a day ago


Do you recognize this document?

NO  YES




Had you planned on accessing this file today or in the near future?

Yes  No



Now that you have seen the document, would you be interested in accessing it?

Yes  No





# Field Study



100 +  
PARTICIPANTS



~2000  
DOCUMENTS



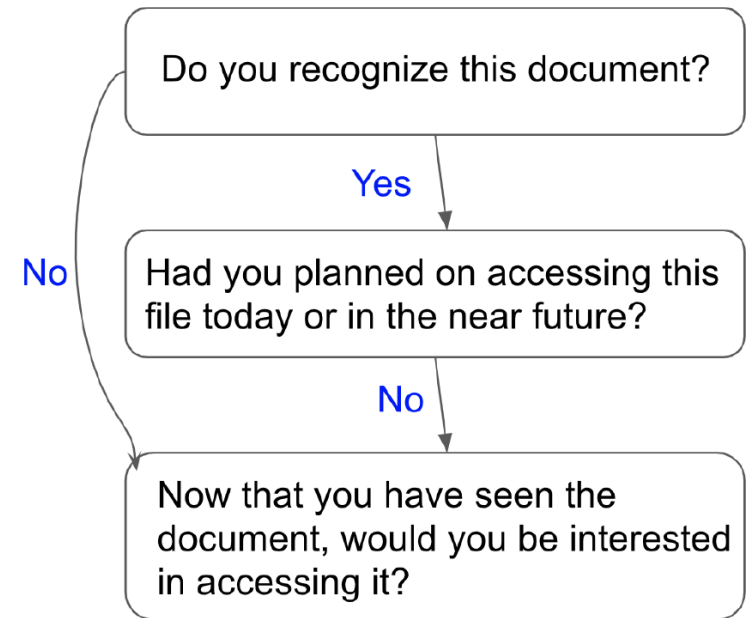
Recency of Access



Richness of Interactions

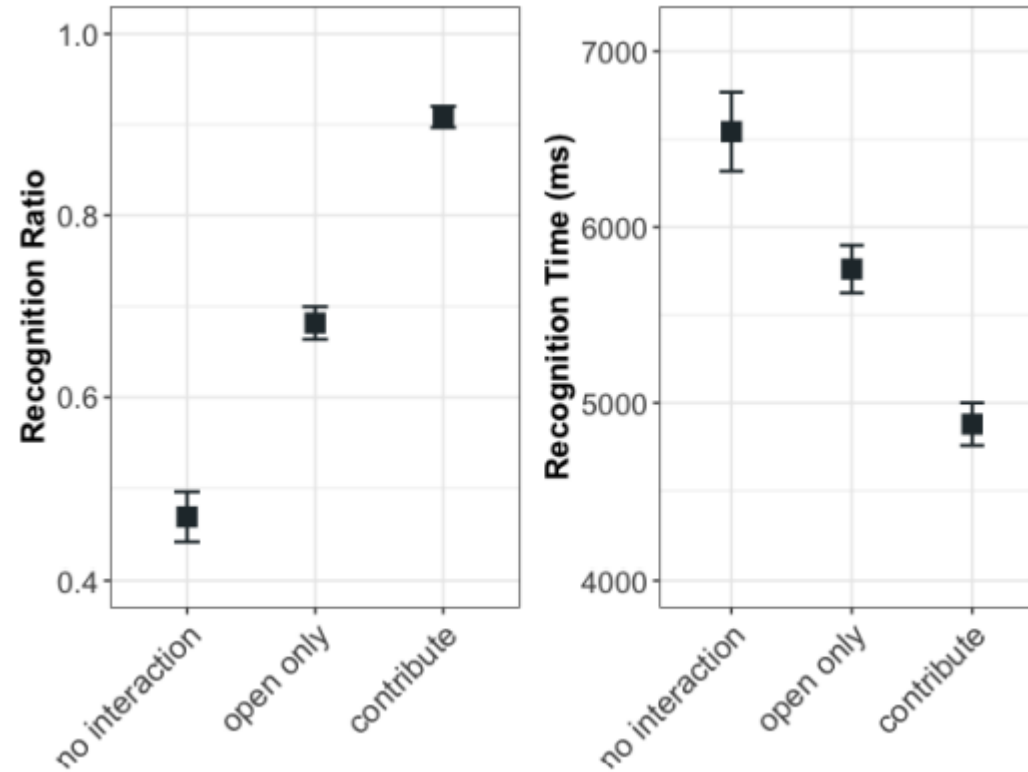


Presentation



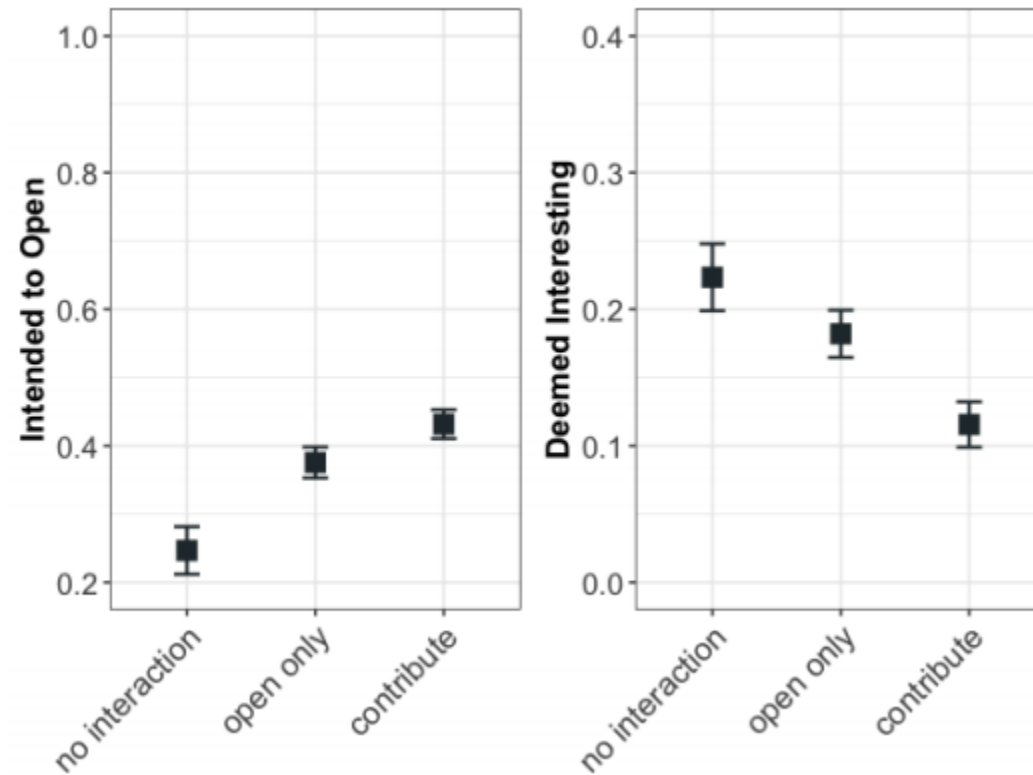


Familiar documents are easier and faster to recognize compared to less familiar ones



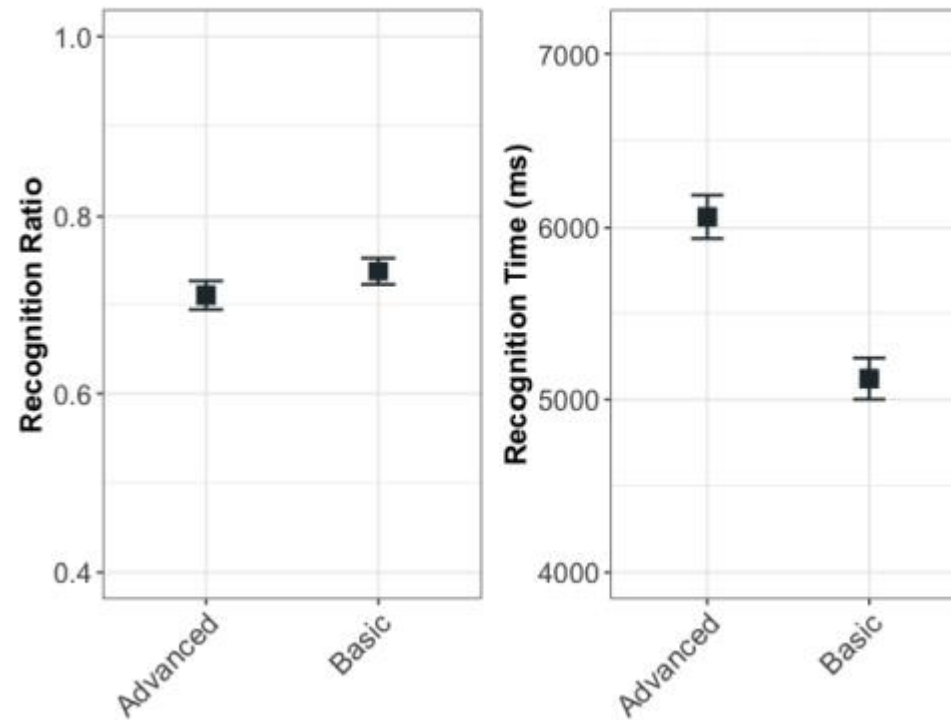


Two distinct intents: *refinding* known and recently accessed document and *discovering* new documents





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



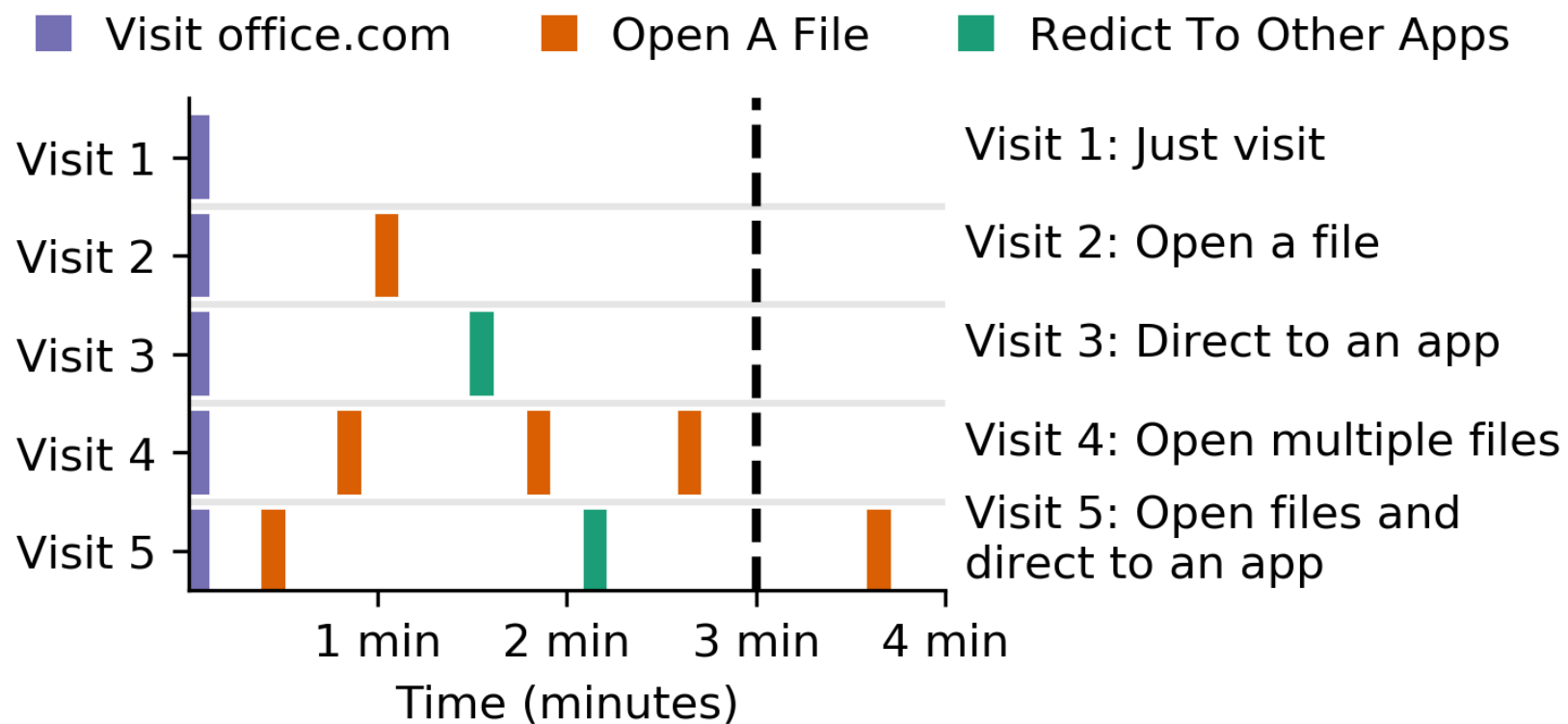
# Large-Scale Log Study



800K  
USERS



8.3M  
VISITS



# Document Recommendations: How to measure utility?

	Shown in RDP	Not shown in RDP
Open in RDP	47.5 % <span>Ⓒ</span>	—
Open Elsewhere	17.2% <span>Ⓑ</span>	35.3%

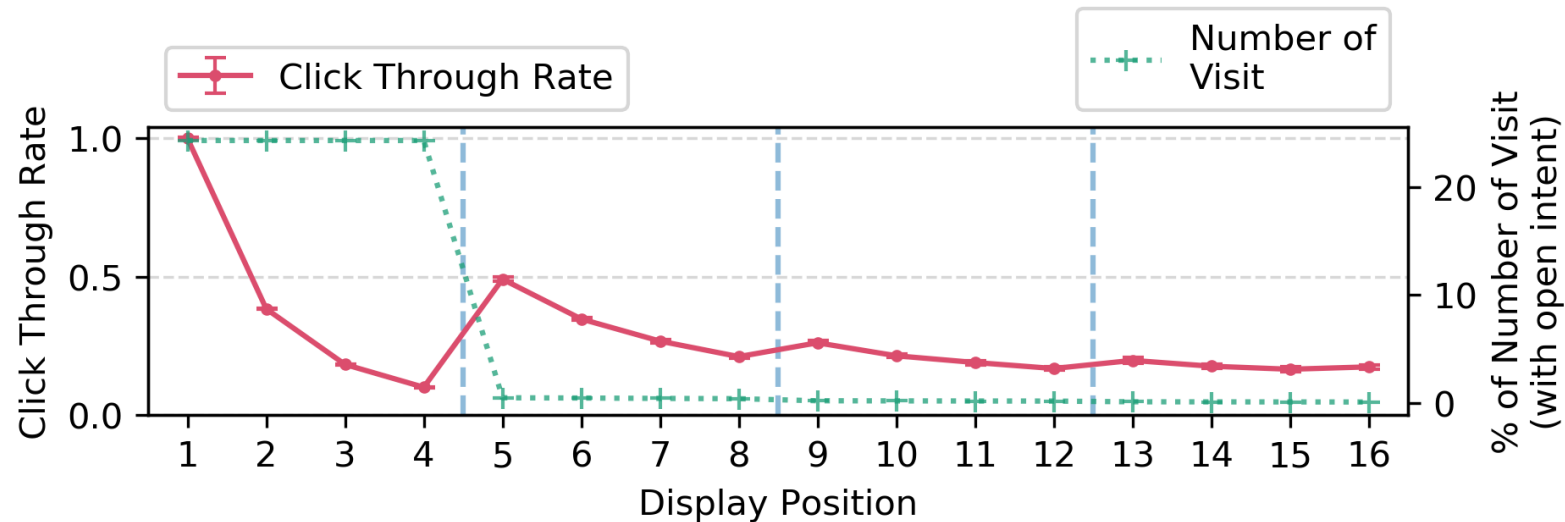
$$\text{Click Through Rate (CTR)} = \frac{\text{Number of Opening a File from RDP}}{\text{Number of Visits}}$$

Time to Open (TTO) = Time Elapse between Visit and File Open from RDP

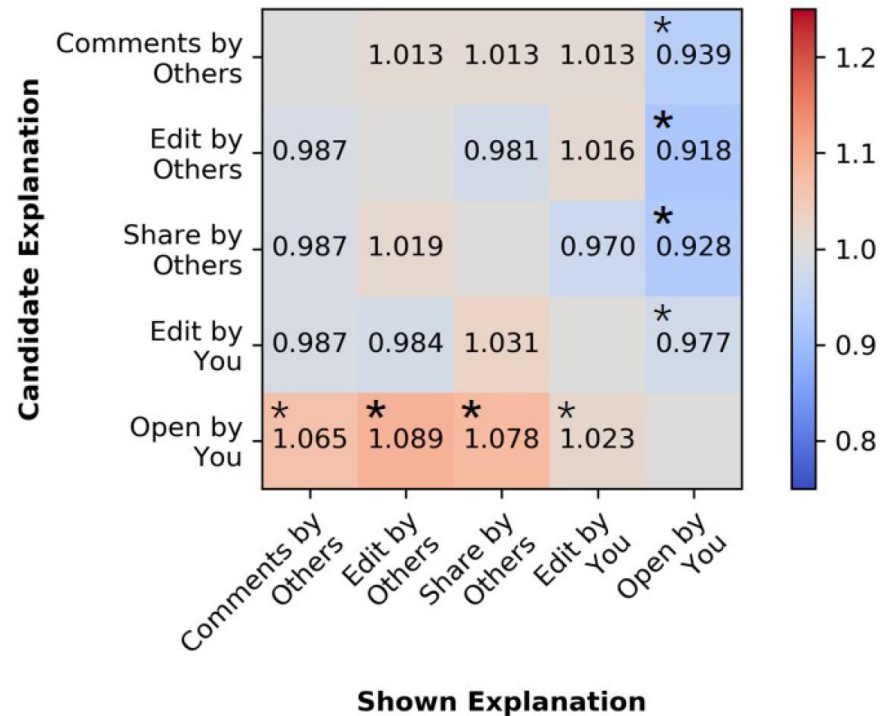
$$\text{Recognize Rate (RR)} = \frac{\text{Number of Opening a File from RDP} \text{ Ⓒ}}{\text{Number of Opening a File shown in RDP} \text{ Ⓑ}}$$

# Document Recommendations: Interactions

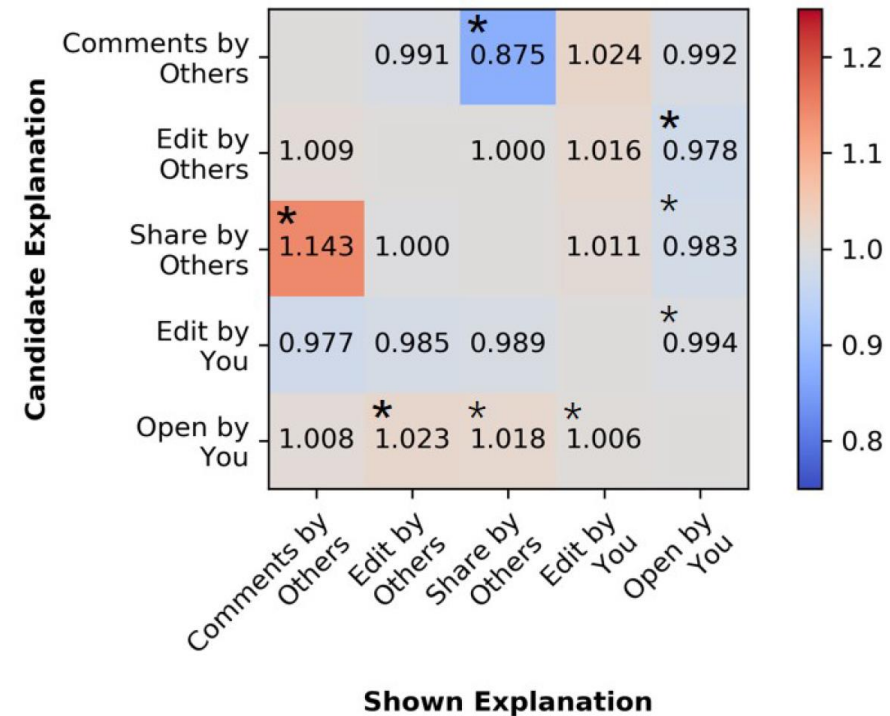
- CTR decrease from the Position 1 (leftmost) to 4 (right most)
- Most of the people do not visit the “next page” of the RDP, but once they go there, there is a jump up of the CTR



# Large-Scale Log-Based Experimental Study: Document Explanations



(a) Pairwise Comparison on CTR



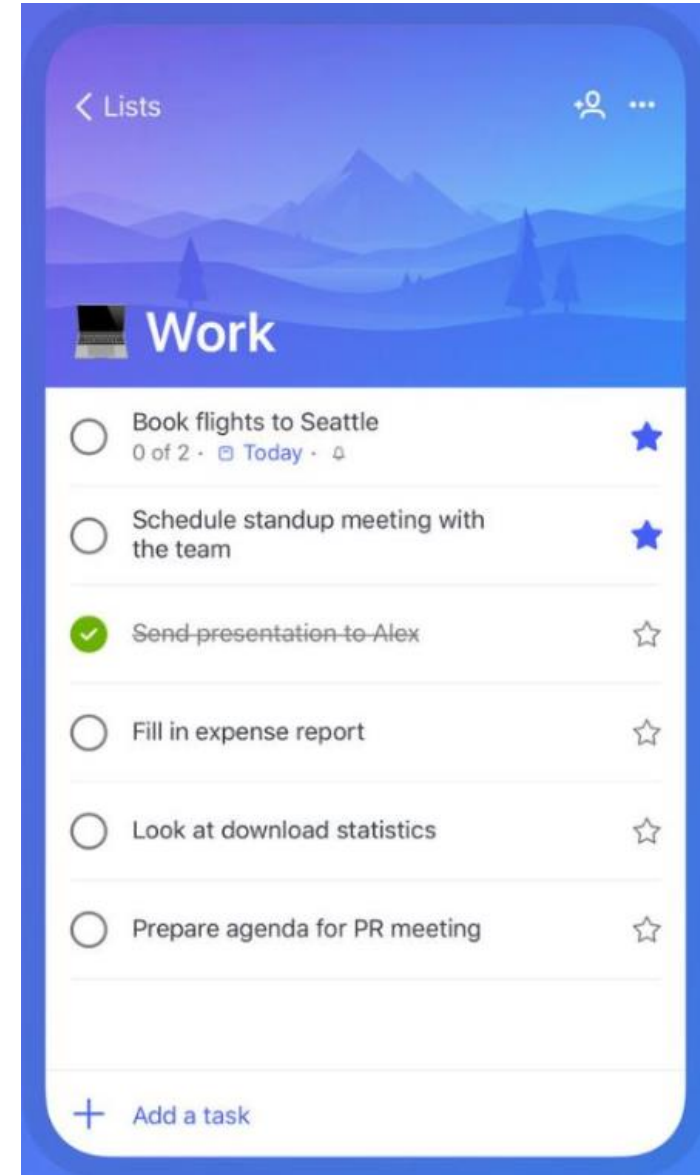
(b) Pairwise Comparison on RR



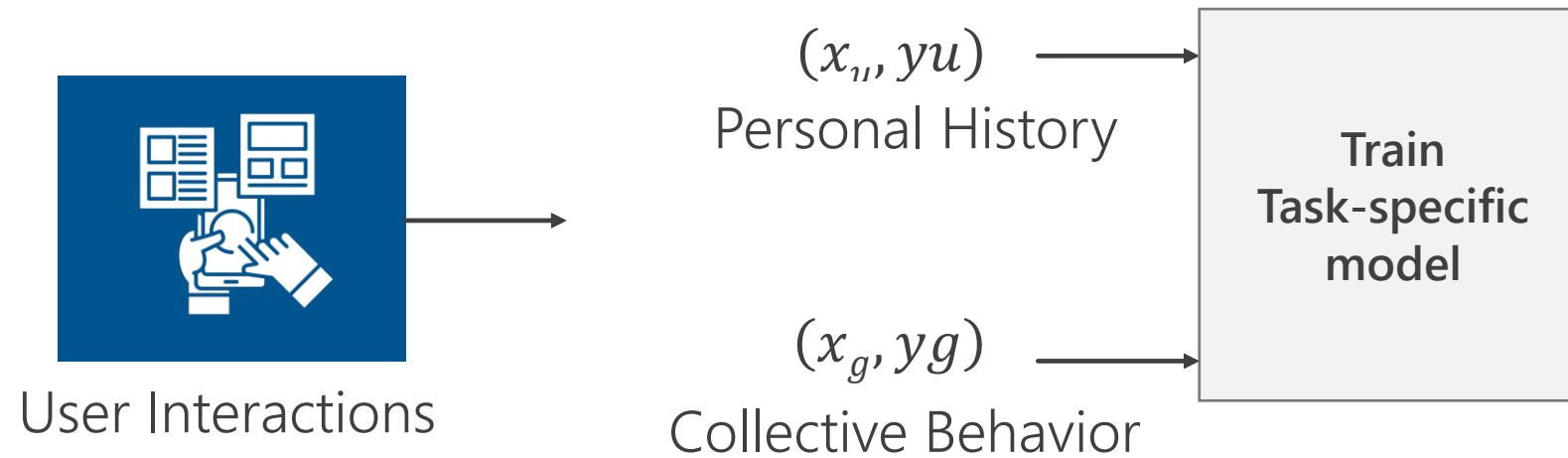
# Task Intelligence

# 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

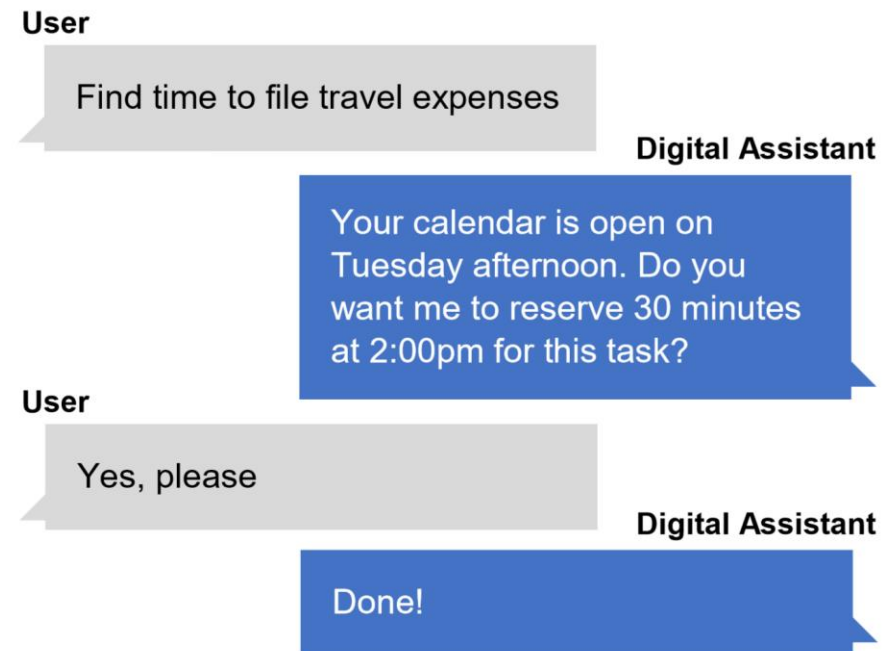


# User Interactions as labels for ML Models



# Task Duration Estimation

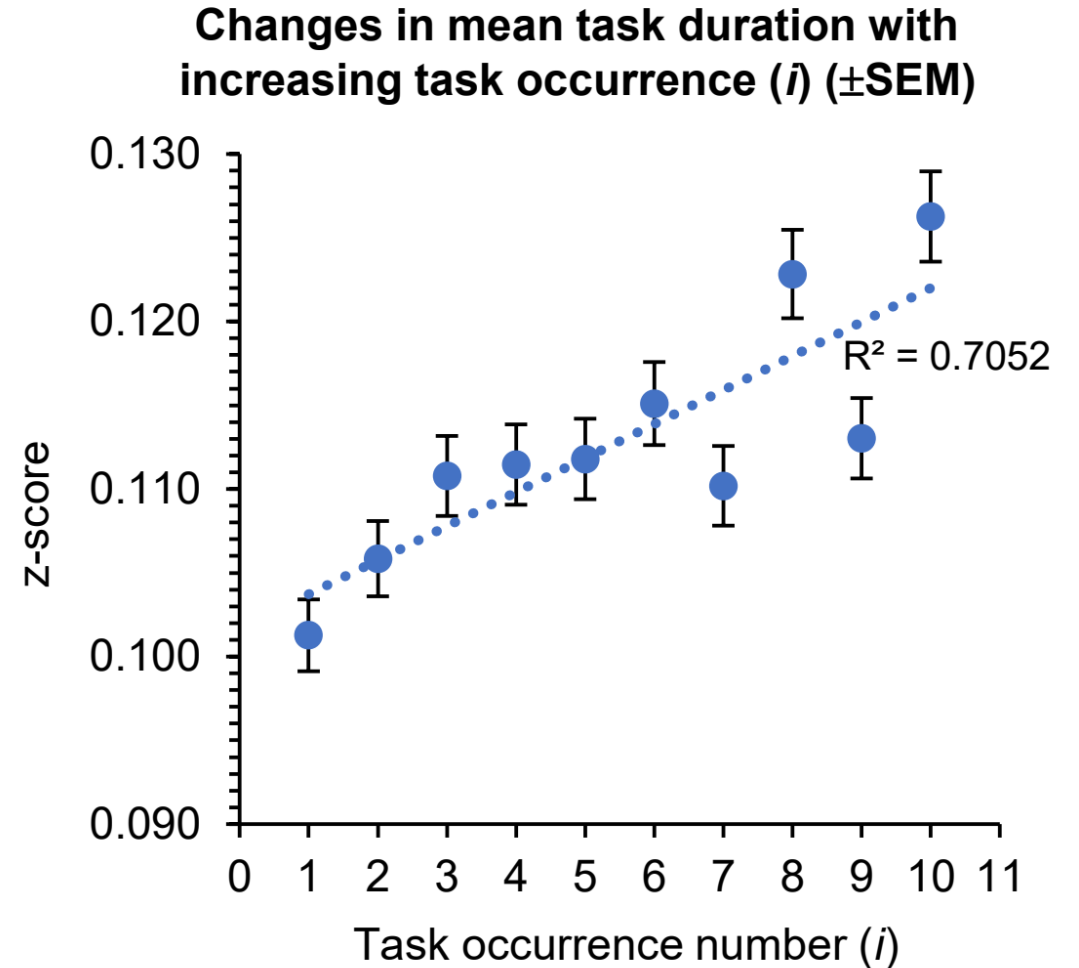
- Time estimation occurs often (every task needs this)
- People struggle to estimate how long tasks take, esp. for new tasks
- Forecast durations using large-scale data



Aspirational goal is building such experience

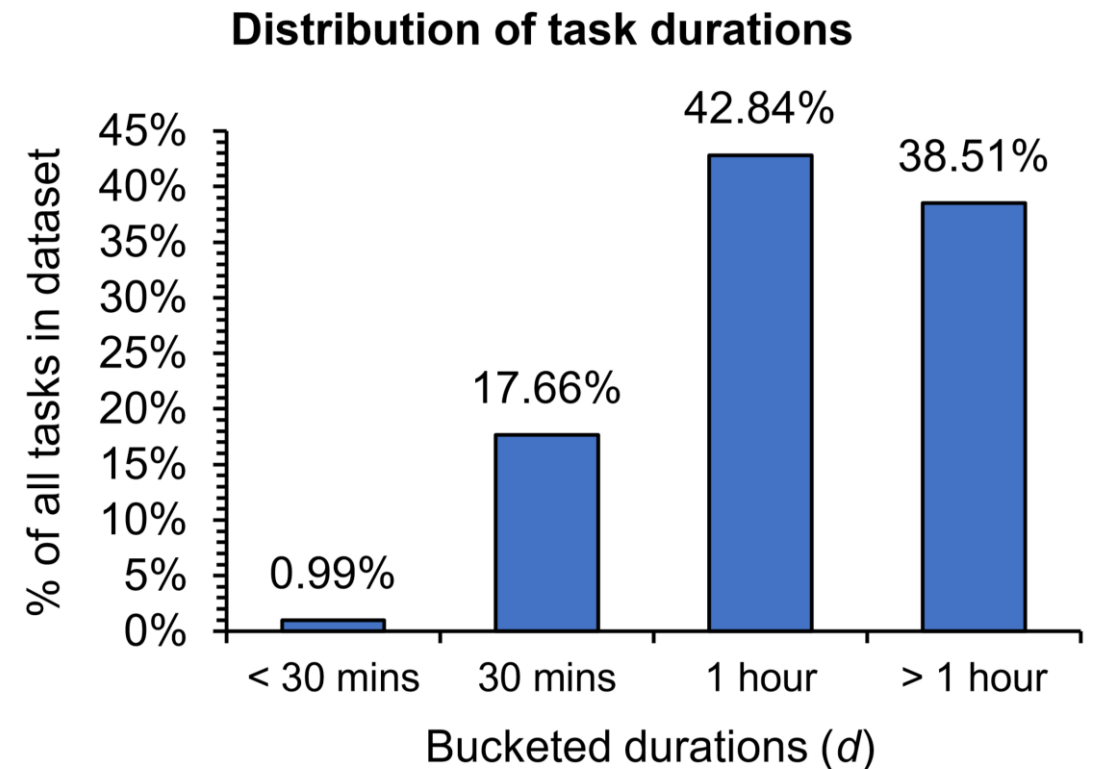
# Duration Estimation is hard even for people

- 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	<i>r</i>
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

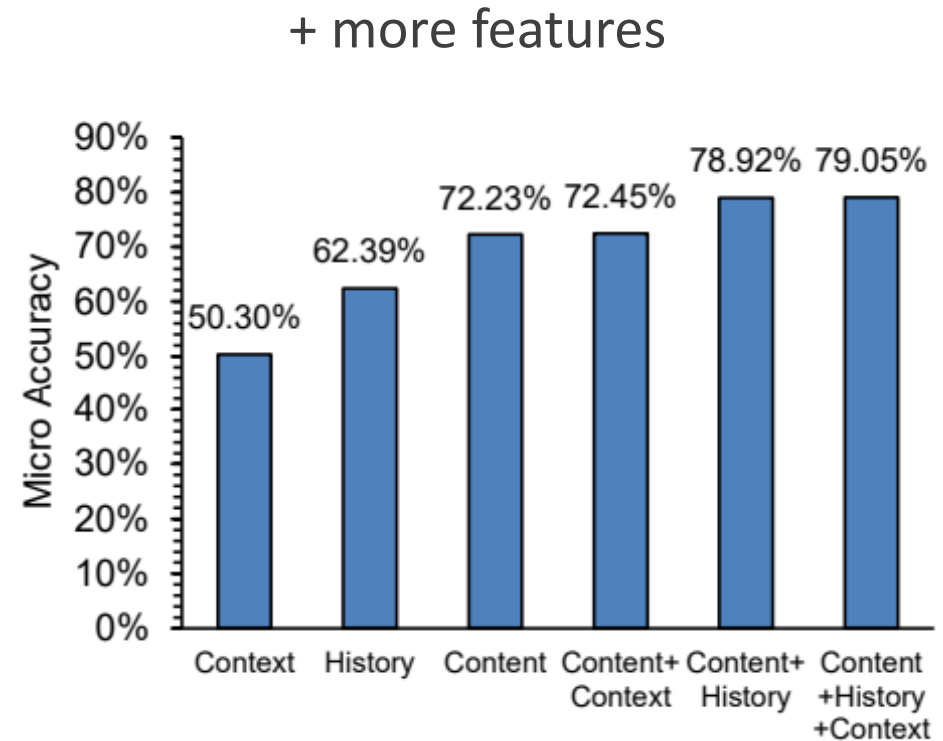
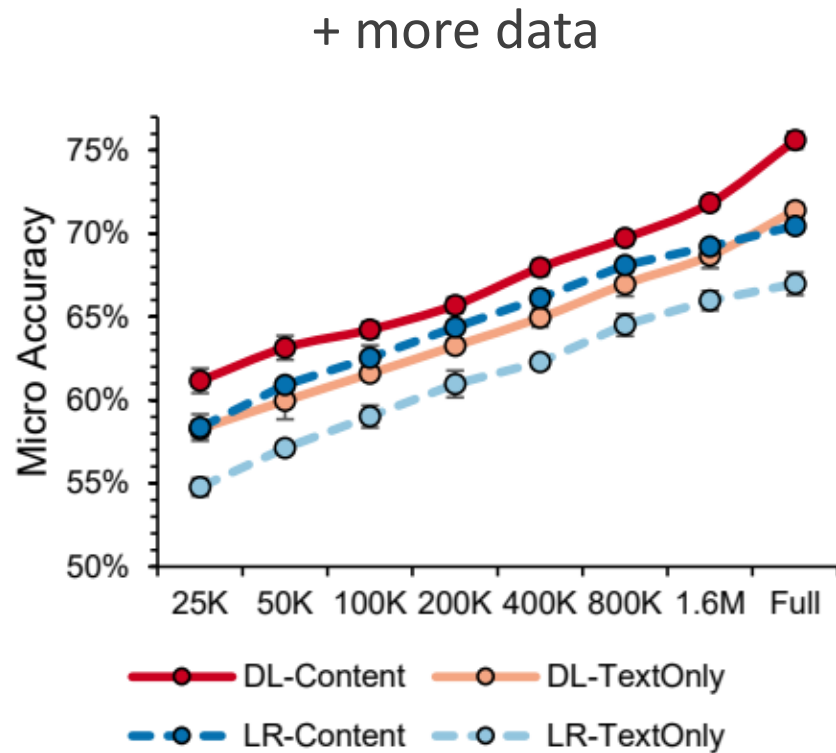
user history  
is important

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

$r$  = correl with appt duration

# Learning to efficiently predict task duration

- Performance changes with -





# Task Auto-Deprecation

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

## Pending Tasks

**“I’ll work on that later.”**

Sent to: Gregg Newton — 8/21/2018, 12:43pm



Snooze



View email



Completed

**“I will find out what else they have.”**

Sent to: Clayton Jones — 8/25/2018, 09:01am



Snooze



View email



Completed

**! It looks like this task is already complete ...**

~~“I will send you the file by end of day.”~~

Sent to: Norma Saunders — 8/16/2018, 10:54am



Snooze



View email



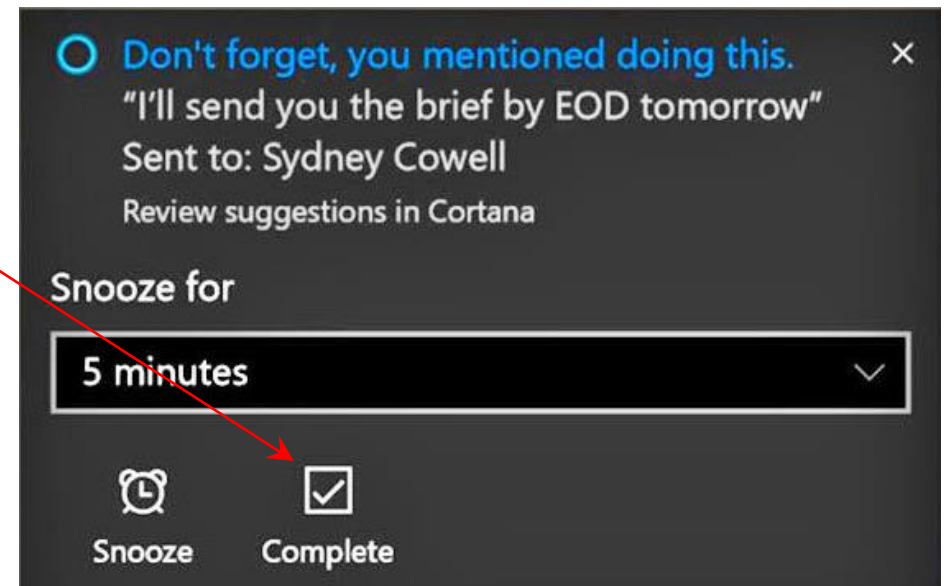
Completed



Not completed

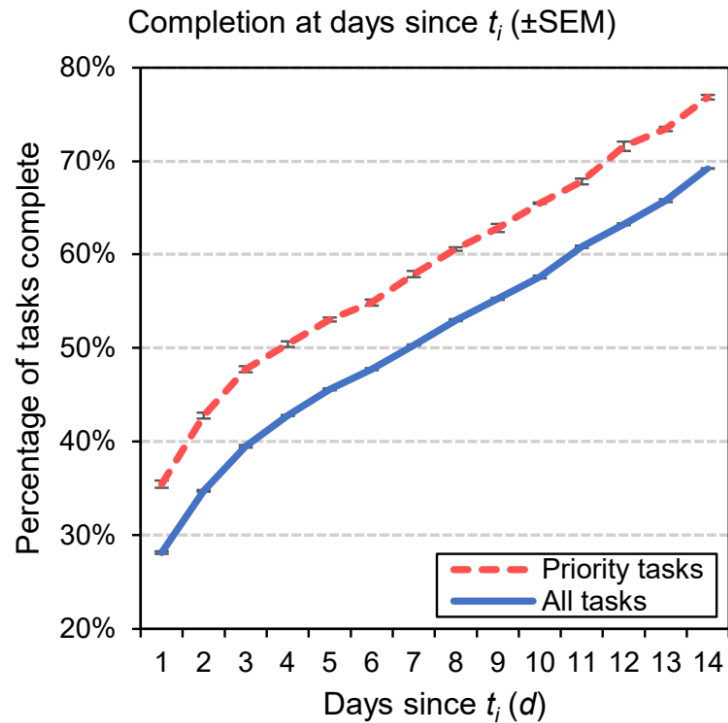
# Learning from User Behavior Data

- 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

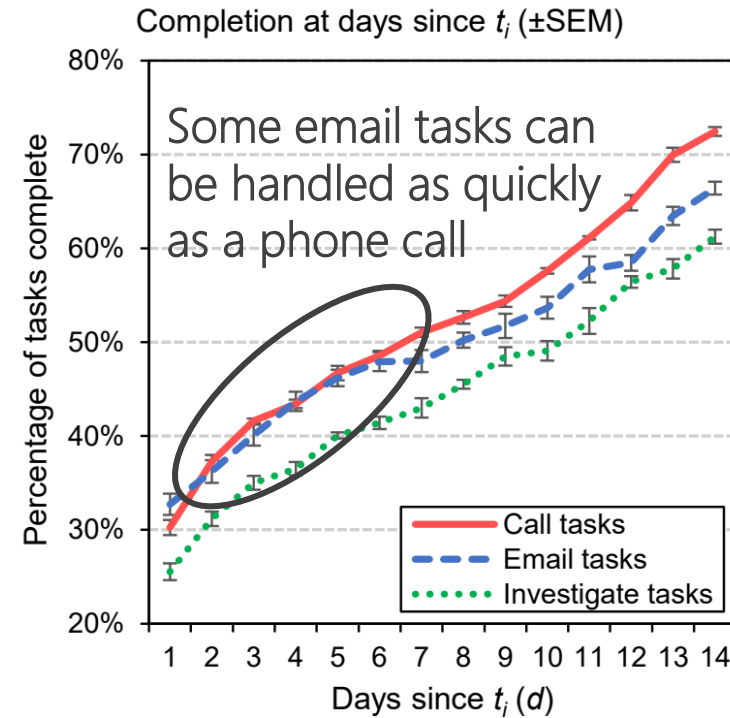


# Task Completion Over Time

- 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

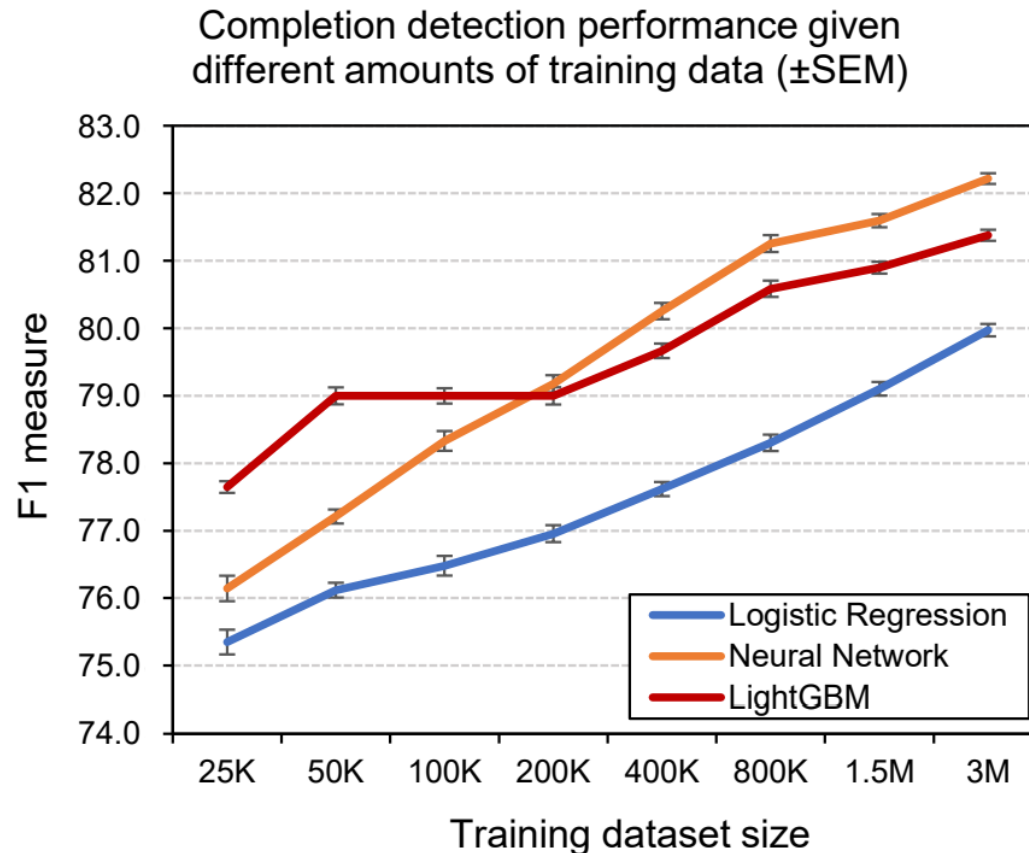
# Learning to efficiently predict task completion

- 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.

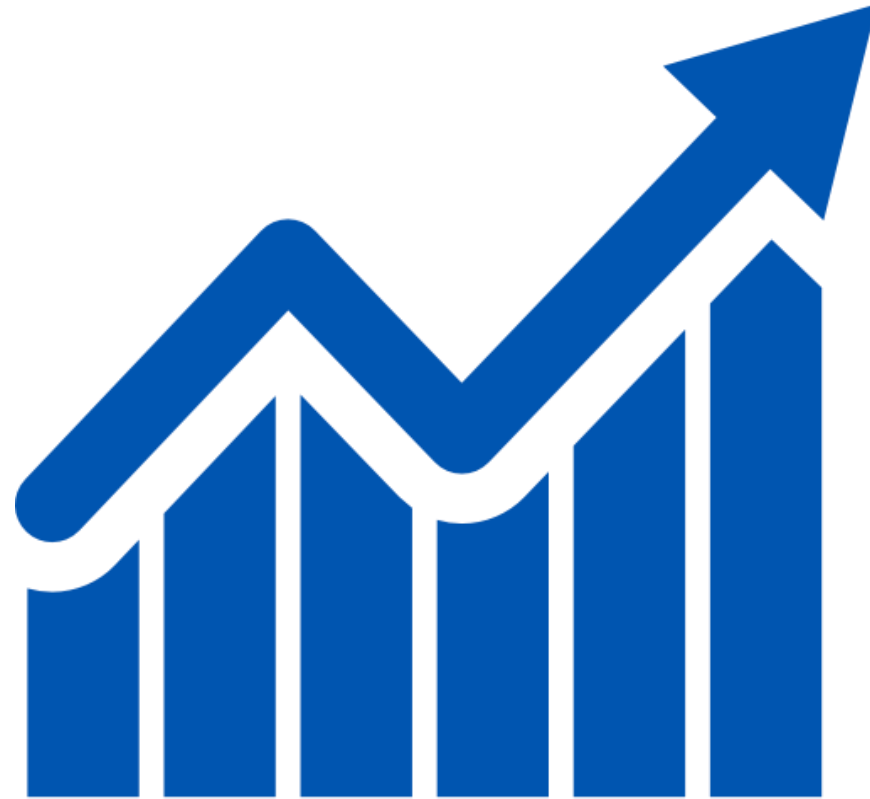
Leverage both content and behavioral features

# Learning to efficiently predict task completion

- Strong performance and improves with more data



# Challenges and Opportunities



You get what you **measure**





**Privacy** powers **Trust**

**Trust** enables **Data**

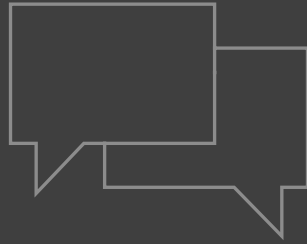
**Data** powers **Innovation**





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>