Bilgisayarlarla Etkin Sözel İletişim için Yapay Zeka ve Elektronik Akıllı Yardımcı

Effective Interaction with Machines: Intelligent Personal Digital Assistants with Conversational AI

amazon alexa

Ruhi Sarikaya

Director of Applied Science

Outline

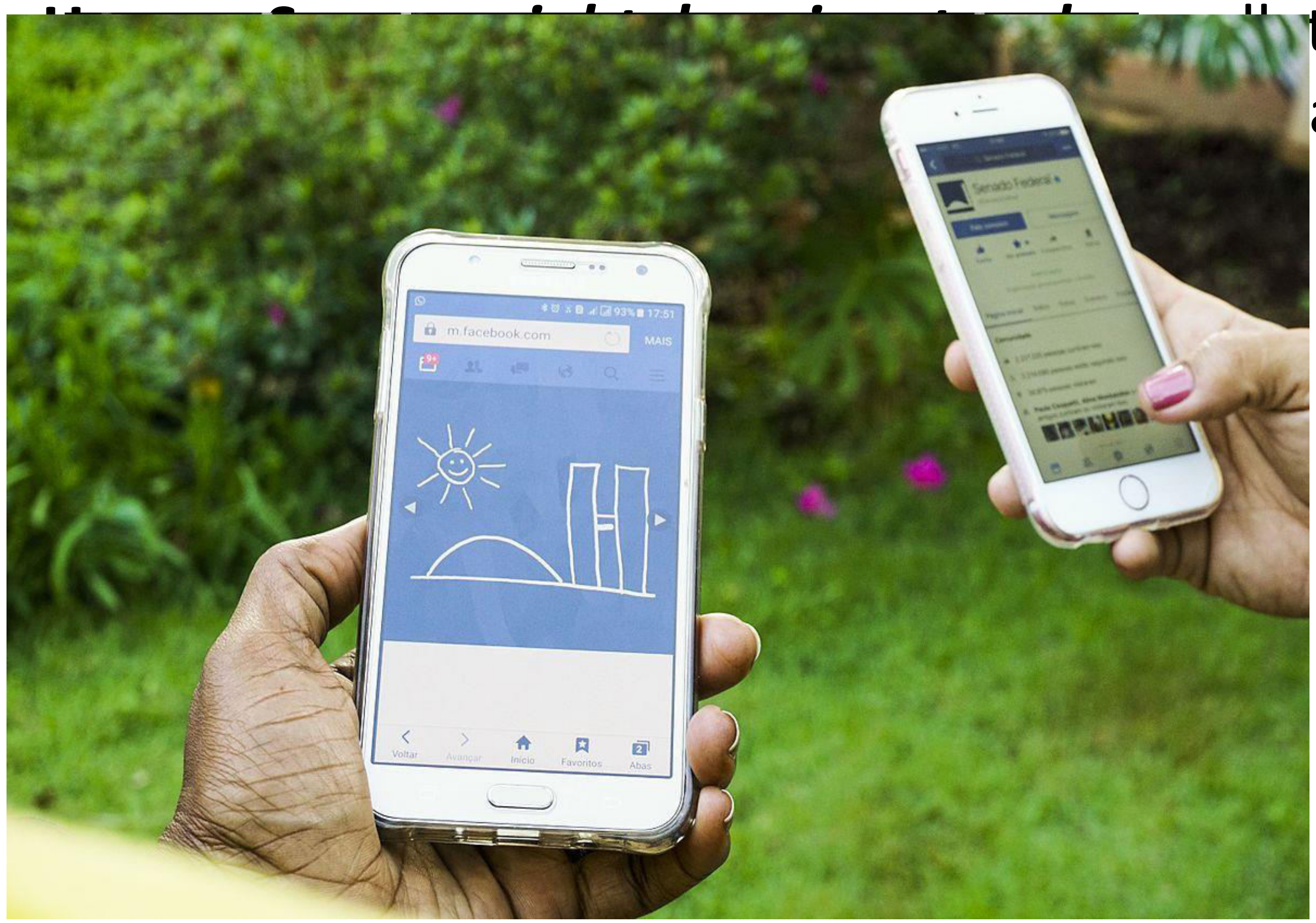
- Humans interaction with the physical and digital world
- Fundamental interaction frictions with digital devices, apps and services
- How does conversational AI work?
  - Natural Language Understanding (NLU)
  - Dialog Management
  - Natural Language Generation
  - Deep Learning Basics
- Solutions to some of the frictions with conversational AI
- Q & A

Frictions (and Amazon)

- What is friction?
  - Friction is any variable that is slowing down *(or entirely halting)* the progression towards achieving a goal.
  - Simple aggravations:
    - ✓ purchasing a product that is not ready to use out of the box (e.g. have to buy a separate cable or battery)
    - ✓ Overly complicated sign-up processes to buy something
    - ✓ Difficult to navigate menus
  - Errors and inefficient processes add to daily customer aggravations
- Reducing frictions in a transaction ➔ positive snowball effect
  - Business trip → buying luggage → tells you it would qualify as a carry-on, free shipping → customers tell these experiences to others
  - 1-click order, Amazon Prime, Amazon Go,….
- Removing friction is key to customer satisfaction
- What is the common theme/thread across these different frictions?

Human Interaction with the Physical and Digital World

taste
and



Computer 'Senses'

- No sight no hearing (until recently)
- Tactile & motor skill input (from humans)

Gap

- Computers (and backend services) are not yet designed for receiving **voice** input to operate

Problem

- It tethers you to a screen, 'immobilizes' you

➔ Friction!

Input → Processing → Outcome

→ Sensors → Perception → Cognition (Reasoning, Memory, Knowledge, Planning) → Action (Motor System) →

Alexa in the Media

The new and improved Echo Dot takes Amazon's best-in-class smart home speaker and wraps it in an ultra-affordable package.
- CNET

……the fact that the category Amazon created has become one of the hottest ones in tech is one reason why the next two years for these products promise to be even more eventful than their first two.
- Fast Company

Amazon's Alexa is the real star of the CES 2017: Here are your Alexa enabled devices
- TechTimes

Amazon's Alexa has more than 30,000 skills, and customers can control more than 4000 smart home devices from 1200 unique brands
- TheWrap (Feb' 2018)

Alexa is the first digital assistant that is actually helpful
- ZDNet (UK and DE review)

Amazon is doubling down on its voice-oriented "Alexa everywhere" strategy
- Business Insider (Sept' 2017)

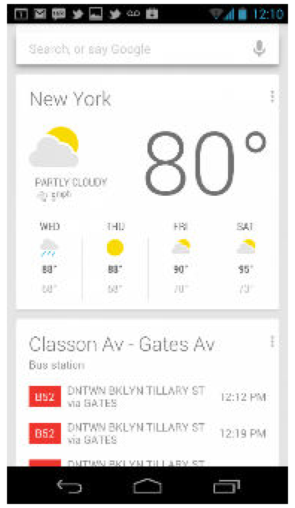
Amazon is Winning the Race to the Future
- Business Insider

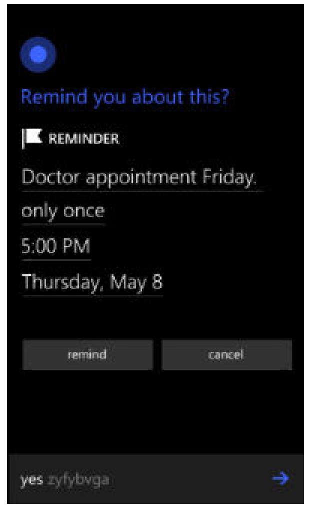
SONOS THE WIRELESS HiFi SYSTEM
vivint. SmartHome
ALCATEL
SAMSUNG
Ford
LG
GE
lenovo

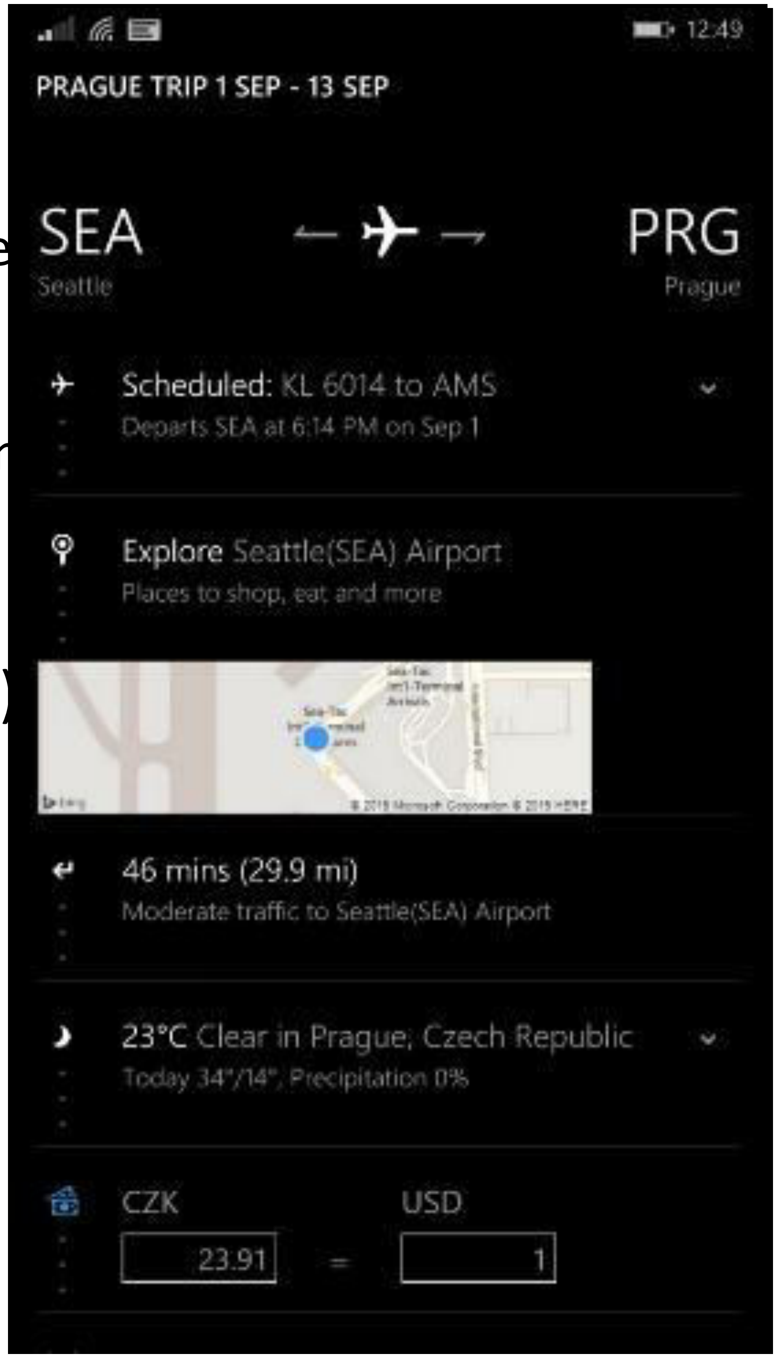
Intelligent Personal Assistants

- **Meta layer of intelligence**
  - Sits on top of other services and applications
  - Performs actions or serves content using services/apps to fulfill the
  - Natural language interface
  - Relies on
    - machine learning, AI, speech recognition, natural language ur management, ranking, inference, personalization, etc..
- **Major IPAs in the Market**

Siri (2011) Google Assistant (2012) Cortana (2014)







- A real example scenario: "flight card on your phone"
  - IPA **scans (1)** your email and **extracts (2)** the flight information and **stores (3)** it
  - **Computes (4)** your current location (GPS) and **checks (5)** the traffic conditions to the airport
  - **Tells (6)** you when to leave for the airport at the day of travel
  - **Checks (7)** the flight status and **updates (8)** you with that
- **Stitching together these steps can potentially mark a breakthrough in removing the cognitive friction!**

New Computing Cycle: Mobile Device & App Revolution (1)

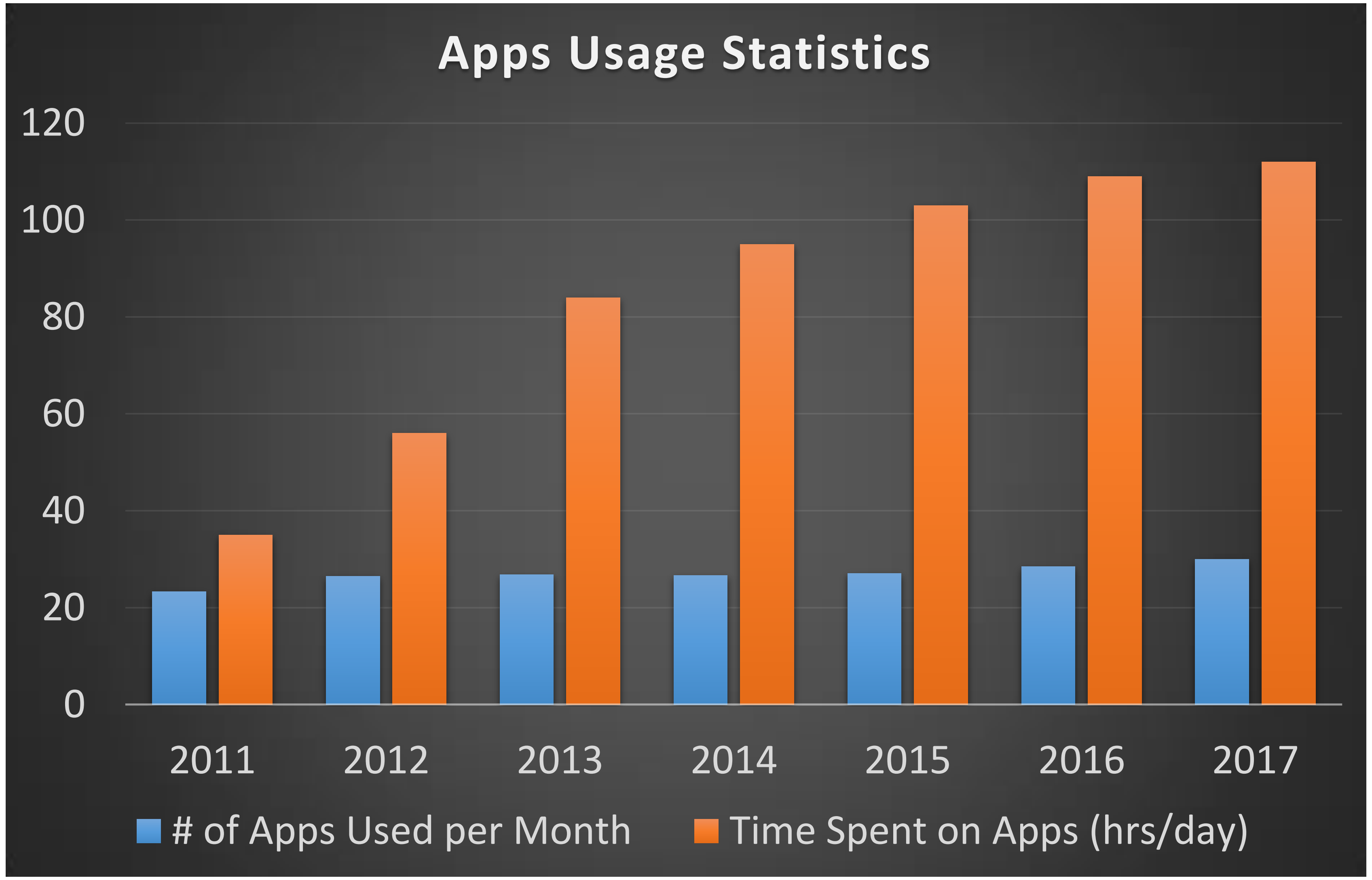
GLOBAL DIGITAL STATISTICS (January 2018)

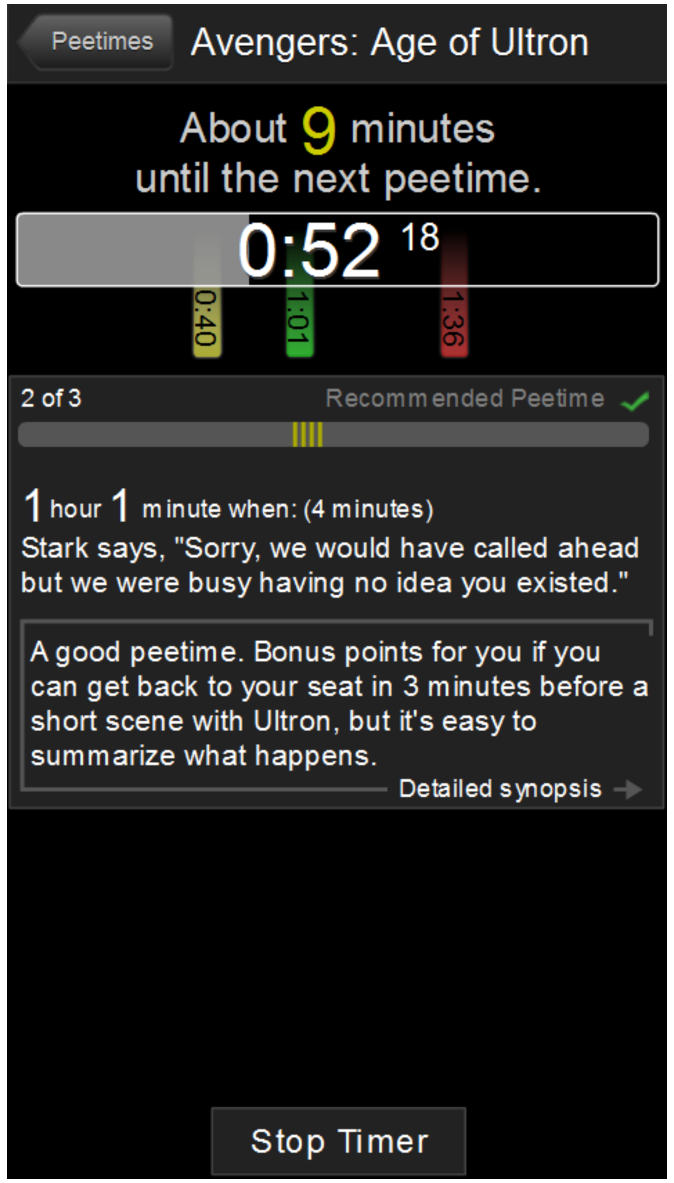
Total Population	Internet Users	Active Social Media Users	Unique Mobile Users
			
7.59B	**4.02B**	**3.19B**	**5.13B**

- New computing cycle of mobile and connected devices [Meeker, Morgan Stanley, 2014]
- Mobile phones are dominating the device market [Gartner-2015]
- About 70% of the world population will have a mobile phone (~5.5B people) by 2020 [CISCO-2016]
  - **No room for growth for connecting people to internet via smartphone (after 2020)**

➔ **What is the next growth opportunity?**

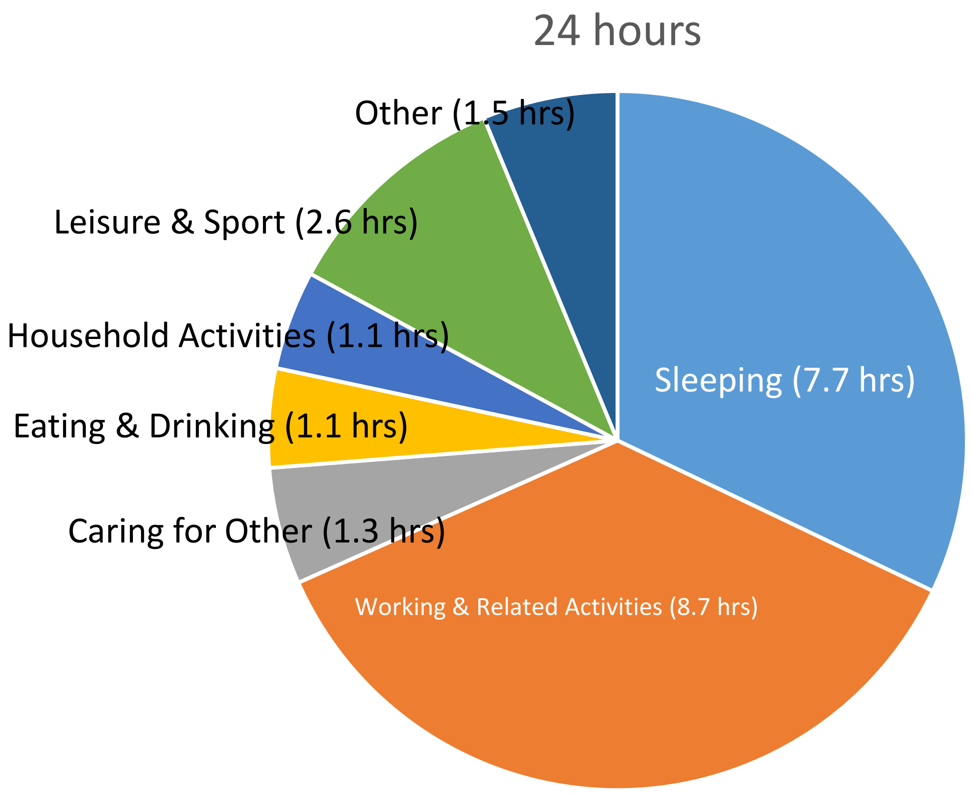
New Computing Cycle: Mobile Device & App Revolution (2)



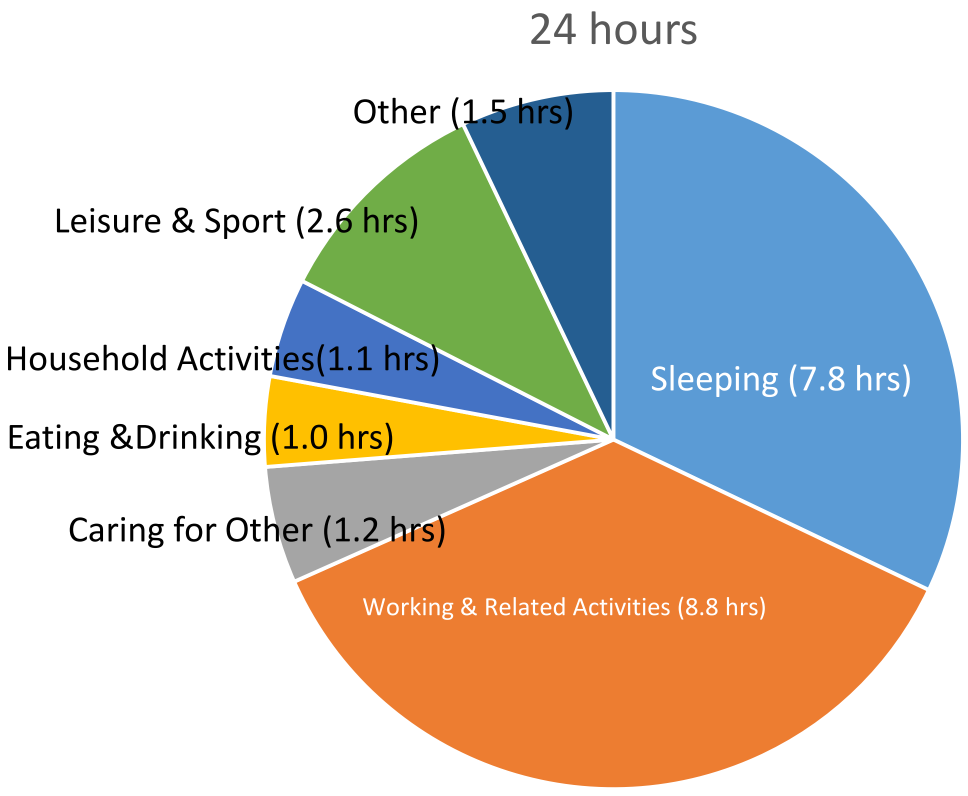


- Apple App Store/Google Play Store have around **2M+/3M+ apps (Jan 2018)**
- Avg # of **UsedApp about 9/30** per day/month is flat [App Annie, 2017]
- About 80% of the apps are zombie apps
- **App discovery challenge (friction #1)**
- **Limited cognitive bandwidth to learn the apps (friction #2)**

Time: 1000 minutes



Time use on an average workday for an employed persons aged 25-54 with children (Bureau of Labor Statistics, 2009)



Time use on an average workday for an employed persons aged 25-54 with children (Bureau of Labor Statistics, 2015)

- TV=168min, Web (PC)=70min, Mobile Phone=180min (2014)
- Smartphone(time) > TV(time)
- Apps are after **you** and **your time**
- Math does NOT add up ➔ where does the extra **110mins** spent on smartphone come from?
- **The budget is fixed: ~1000 min/day (constraint)**

So, what does that mean?

- Apps penetrated into anywhere/anytime/anything we do
- Separation between work and personal life got fuzzy and ‘problematic’
- <u>**You are well over your 1000mins budget! You need help for managing your life.**</u>
- “One needs a machine to beat a machine” → Intelligent Personal Assistant could be that machine to give you your time back

Natural Way to Interact with Personal Assistants: Voice Input

- Limited information flow into smartphones/devices with typing/touch (friction #3)
  - People can speak up to 4 times faster than they can type
- Speech is expected to replace touch/typing as the primary input form
  - Pushed deeper into platforms (e.g. Siri on iOS, Cortana on Windows 10, Google Now is integrated into Google Search App)
  - By 2018 30% of all interactions with devices will be voice based (Gartner)
  - By 2020 50% of all searches will be voice searches (comScore)
  - By 2020 about 30% of searches will be done without a screen (Mediapos)

- Deep Learning had a tremendous impact on speech recognition accuracy
  - **Google**: WER for recognizing words in a mobile apps < 8%
    - Practical alternative to entering text in a box
  - **IBM/Microsoft:** Speech recognition is on the verge of super-human accuracy (2017)
  - **Amazon Alexa:** entirely voice driven and adoption and usage keeps increasing

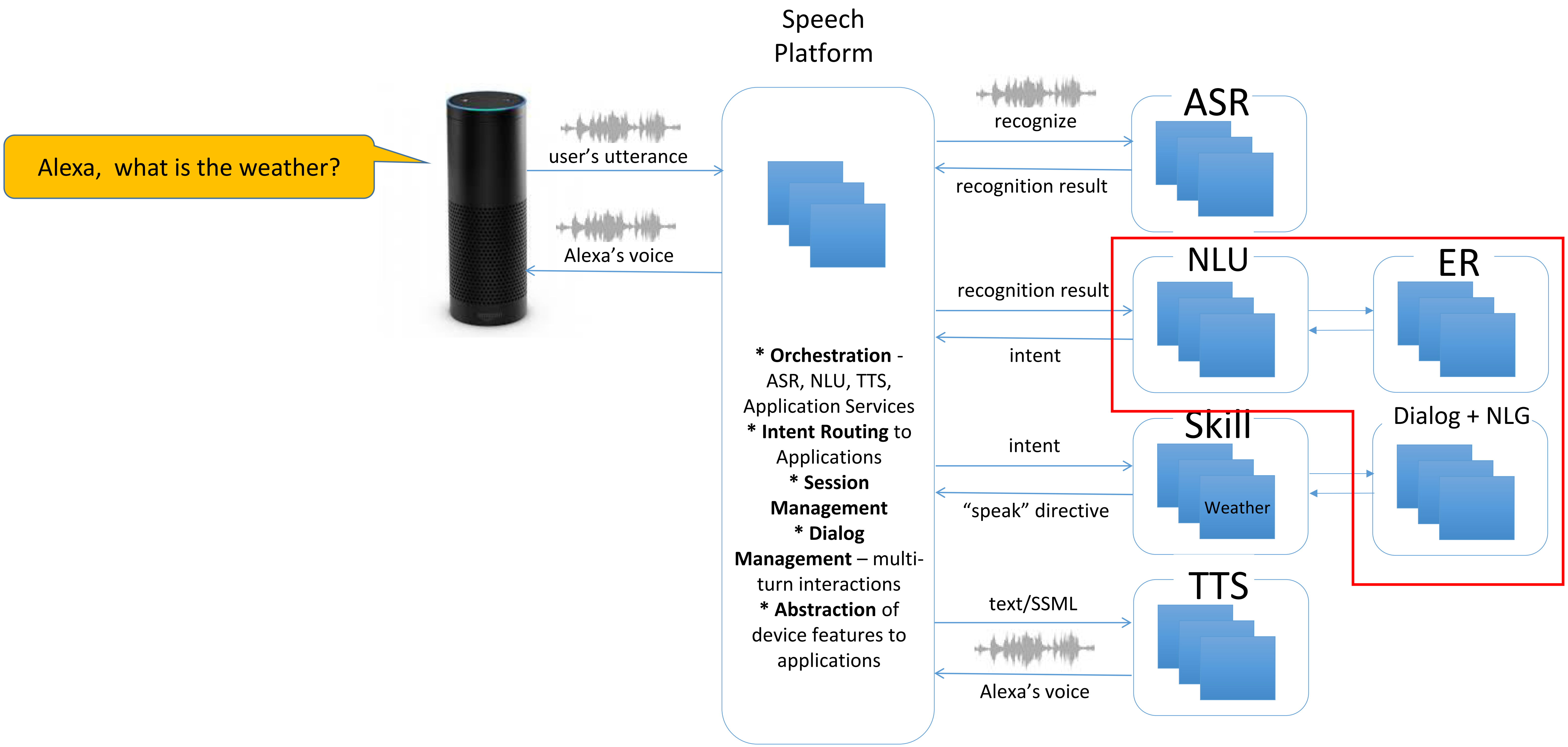
What is the Promise of Conversational AI? Intelligent Personal Assistants

- <u>Summary of Facts</u>
  - Almost every person (>10 year old) (~5.5B people) will have a smartphone by 2020
  - You spend over 3hrs+/day (and increasing) on your smartphone
  - IoT is happening with no clear effective way to interact with them
- <u>Frictions with Apps and Services on Digital Devices: Challenges and Opportunities</u>
  - [Friction #1]: App discovery
  - [Friction #2]: Limited cognitive bandwidth to learn how each app works
  - [Friction #3]: Information flow into small form factors and IoT devices
  - [Constraint]: Your daily time budget is fixed: 1000 mins/day

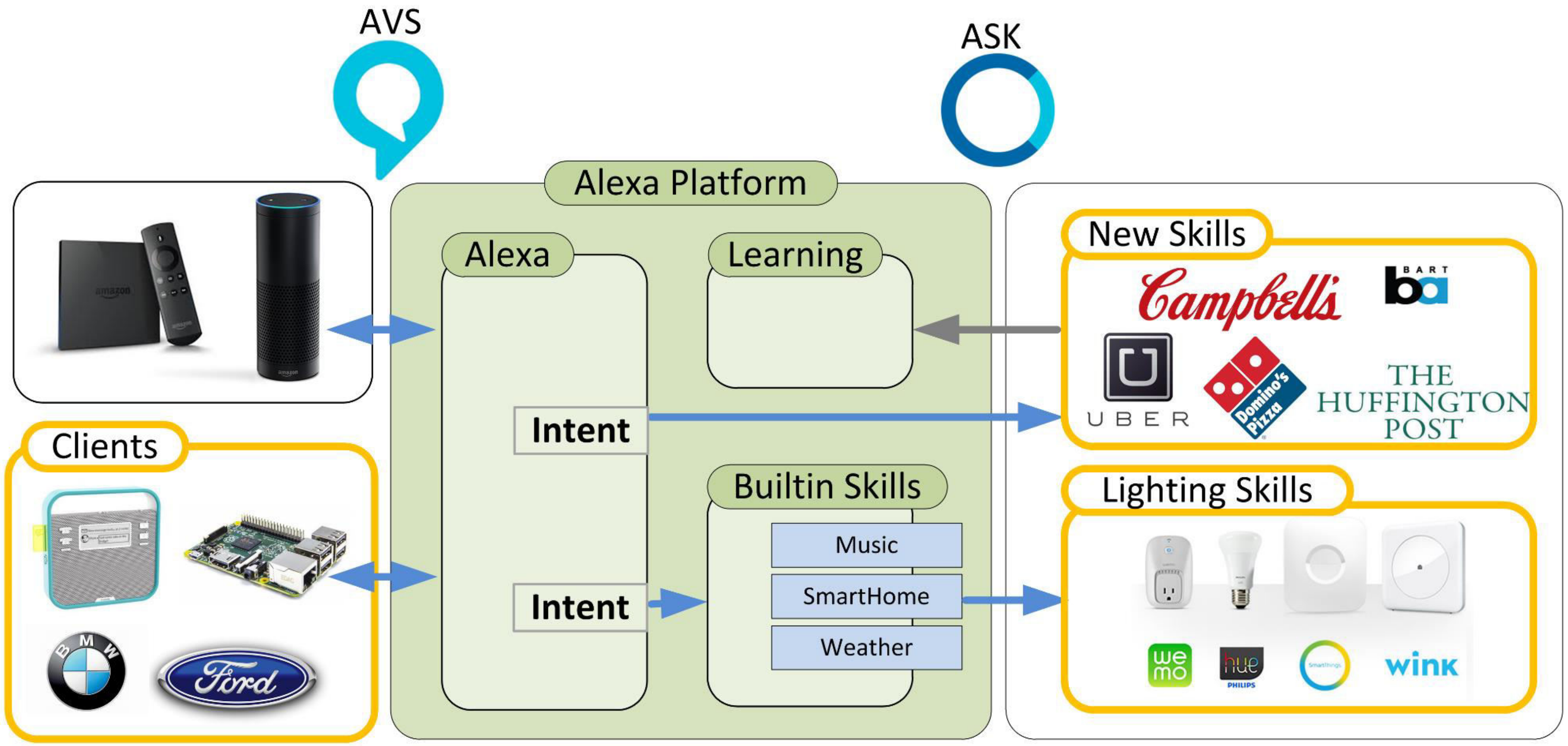
- **Alexa aims to address these frictions through Machine Learning and AI!**
  - Provide a layer over the apps/services, find the right app for the task
  - **Proactively**
    - Completing the tasks and notifying you
  - **Reactively**
    - On-demand assistance: “ask for anything anytime anywhere” through voice

➔ Increase your bandwidth by easing information access and task completion.

How Does Conversational AI work?



Alexa Skills Kit (ASK)



NLU: Domain Identification

Requires Predefined Domain Ontology

User

find a good eating place for turkish food

Intelligent Agent

Restaurant DB

Music DB

Movie DB

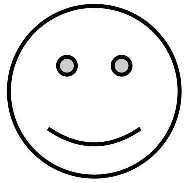
Organized Domain Knowledge (Database)

Classification!

NLU: Intent Detection

Requires Predefined Schema

User



find a good eating place for turkish food

Intelligent
Agent

Restaurant DB

FIND_RESTAURANT
FIND_PRICE
FIND_TYPE
:

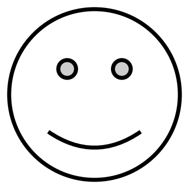
Classification!

NLU: Slot Filling

Requires Predefined Schema

O O B-rating O O O B-type O

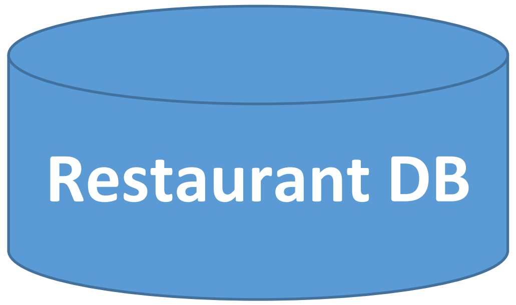
User



find a good eating place for turkish food

Restaurant	Rating	Type
Rest 1	good	Turkish
Rest 2	bad	Turkish
:	:	:

Intelligent Agent



FIND_RESTAURANT
rating="good"
type="turkish"

Semantic Frame

SELECT restaurant {
rest.rating="good"
rest.type="turkish"
}

Sequence Labeling

NLU: Entity Resolution

Problem: There are multiple ways of referring to real world entities

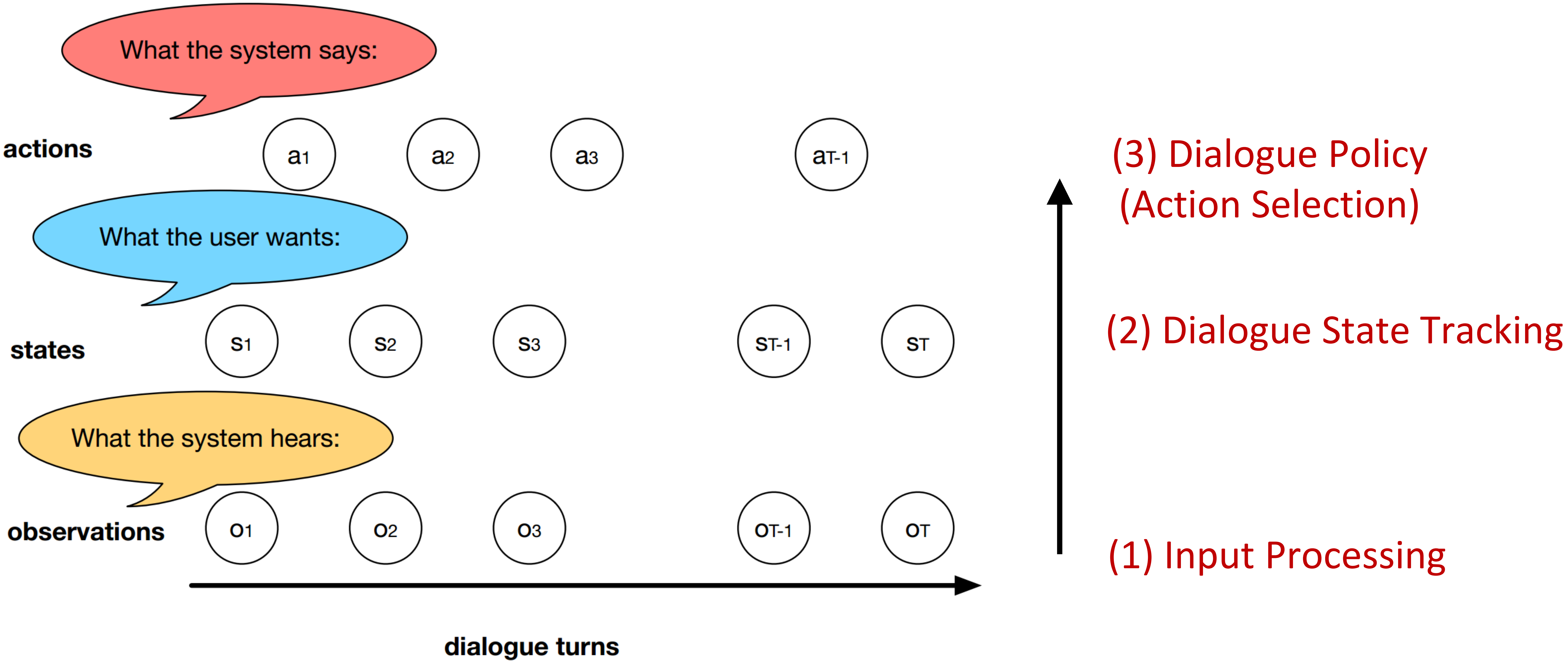
- Kara Kartal / Besiktas Spor Kulubu / BJK
- iPAD Two 16 GB WiFi White / iPAD 2nd Generation 16GB WiFi White

Solution: Entity Resolution - It is the task of resolving different surface forms to the same entity (e.g. Database ID)

Slot/Entity Canonicalization

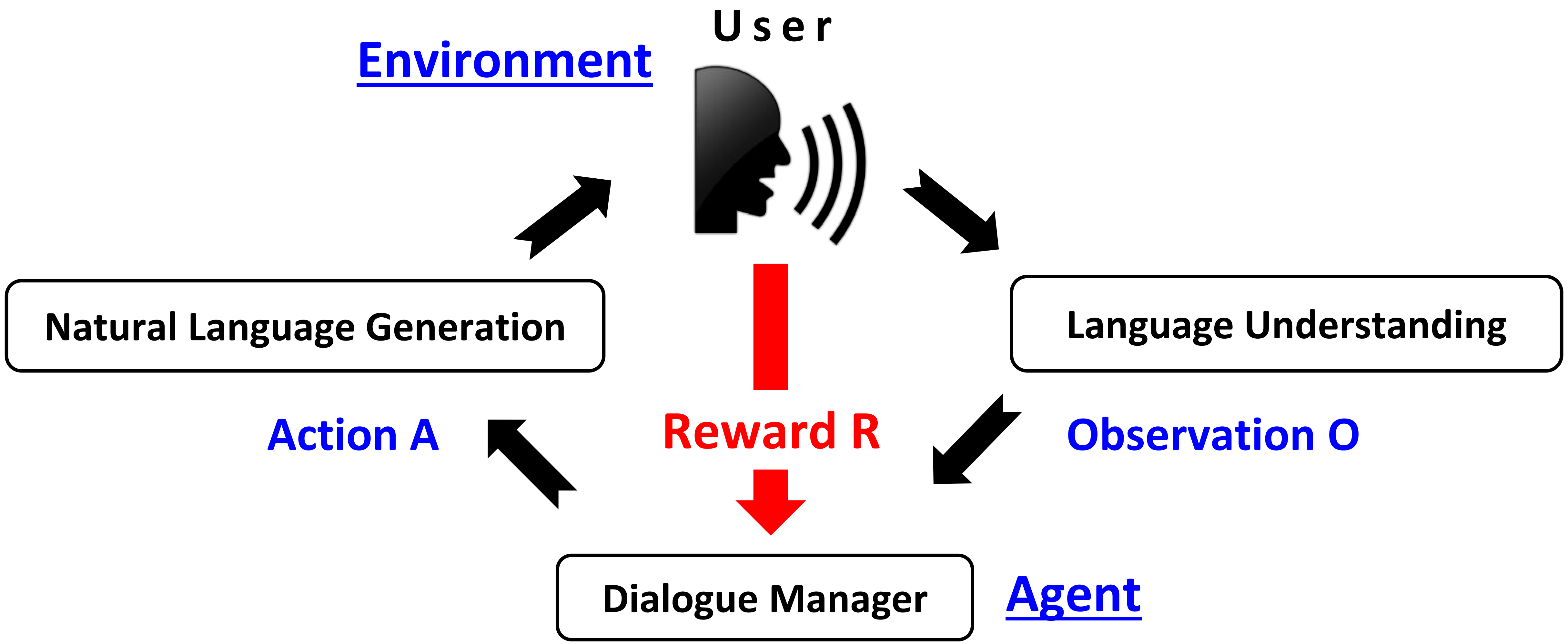
- Mapping “highly” to “good” (canonical form)
- Time: Date/Time resolution, Location: latitude/longitude mapping

Components of Dialogue Management



Dialogue Policy Optimization

- Dialogue management in a RL framework



Optimized dialogue policy selects the best action that can maximize the future reward.
Correct rewards are a crucial factor in dialogue policy training

Natural Language Generation (NLG)

- Mapping system action into natural language to communicate to the user

inform(name=Seven_Days, foodtype=Chinese)

↓

Seven Days is a nice Chinese restaurant

Template-Based NLG

- Define a set of rules to map frames to NL

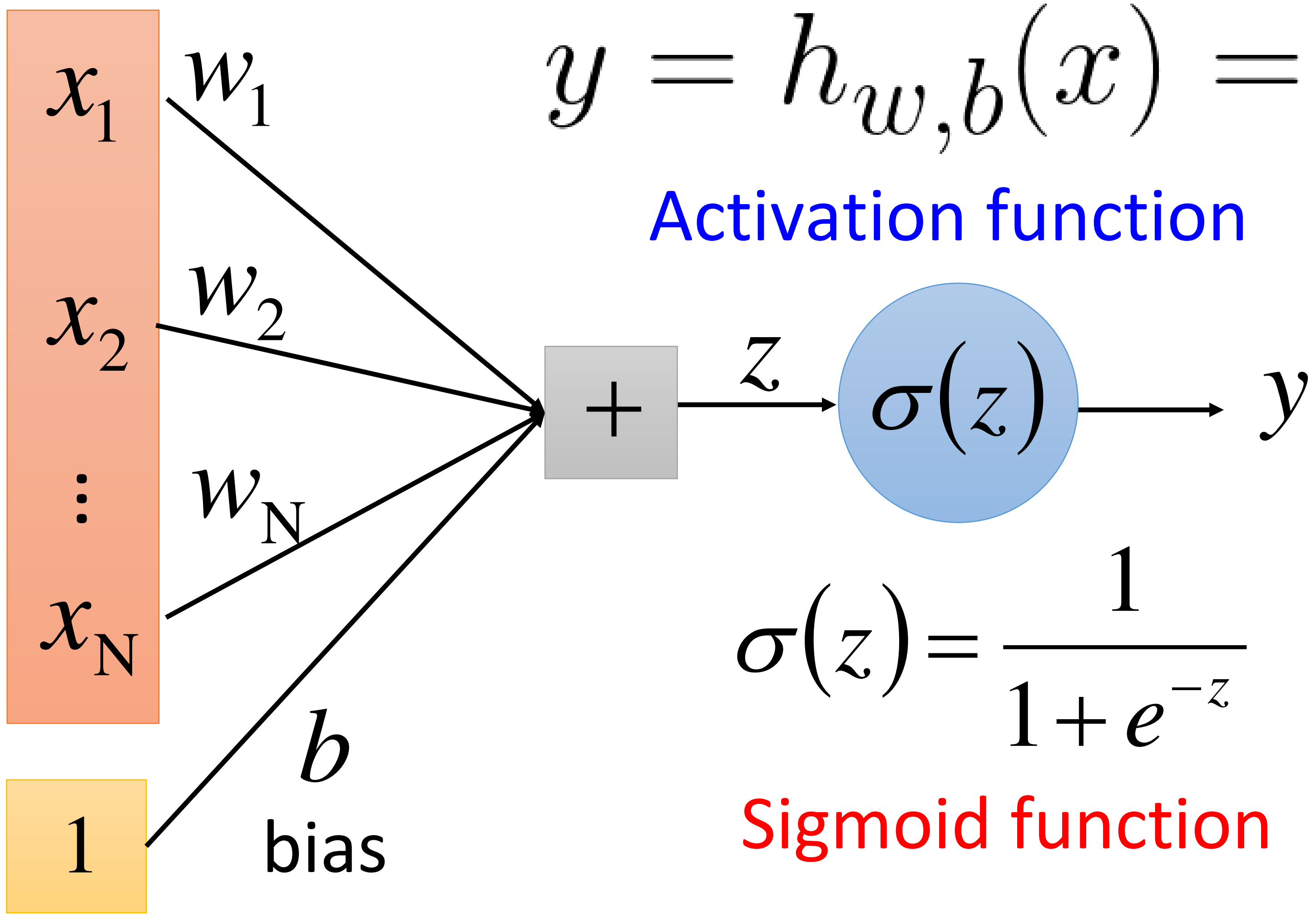
Semantic Frame	Natural Language
confirm()	"Please tell me more about the product your are looking for."
confirm(area=$V)	"Do you want somewhere in the $V?"
confirm(food=$V)	"Do you want a $V restaurant?"
confirm(food=$V,area=$W)	"Do you want a $V restaurant in the $W."

Pros: simple, error-free, easy to control
Cons: time-consuming, poor scalability

Neural Network Basics

A single neuron

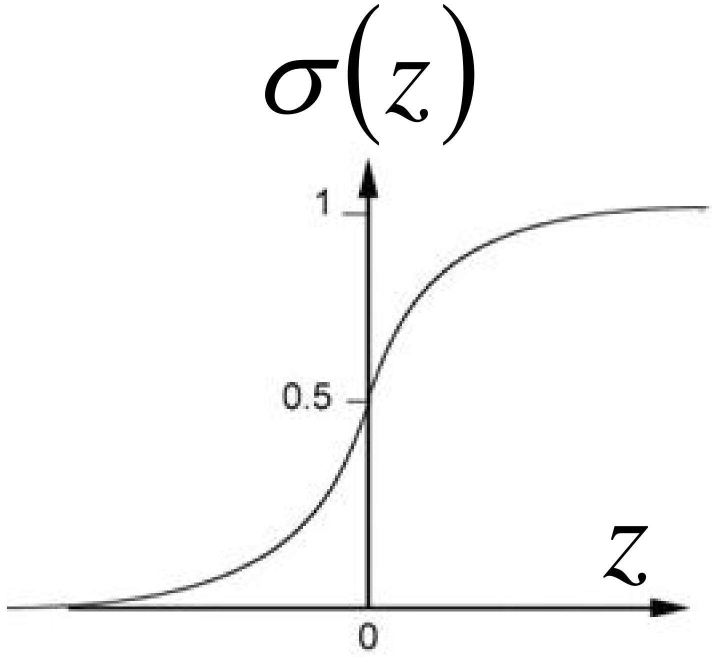
$$y = h_{w,b}(x) = \sigma(w^T x + b)$$



Activation function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

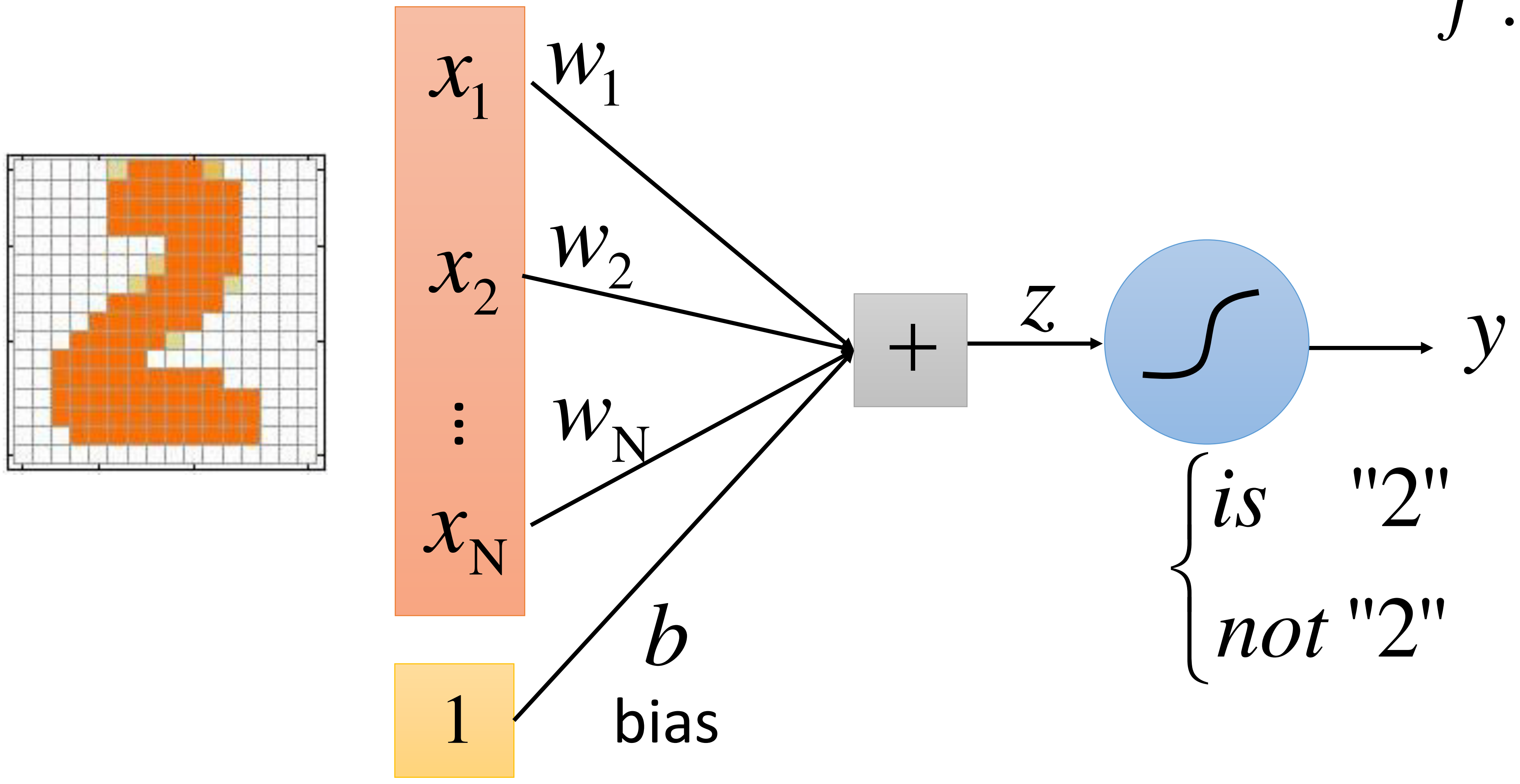
Sigmoid function



w, *b* are the parameters of this neuron

A Single Neuron

$$f : R^N \rightarrow R^M$$



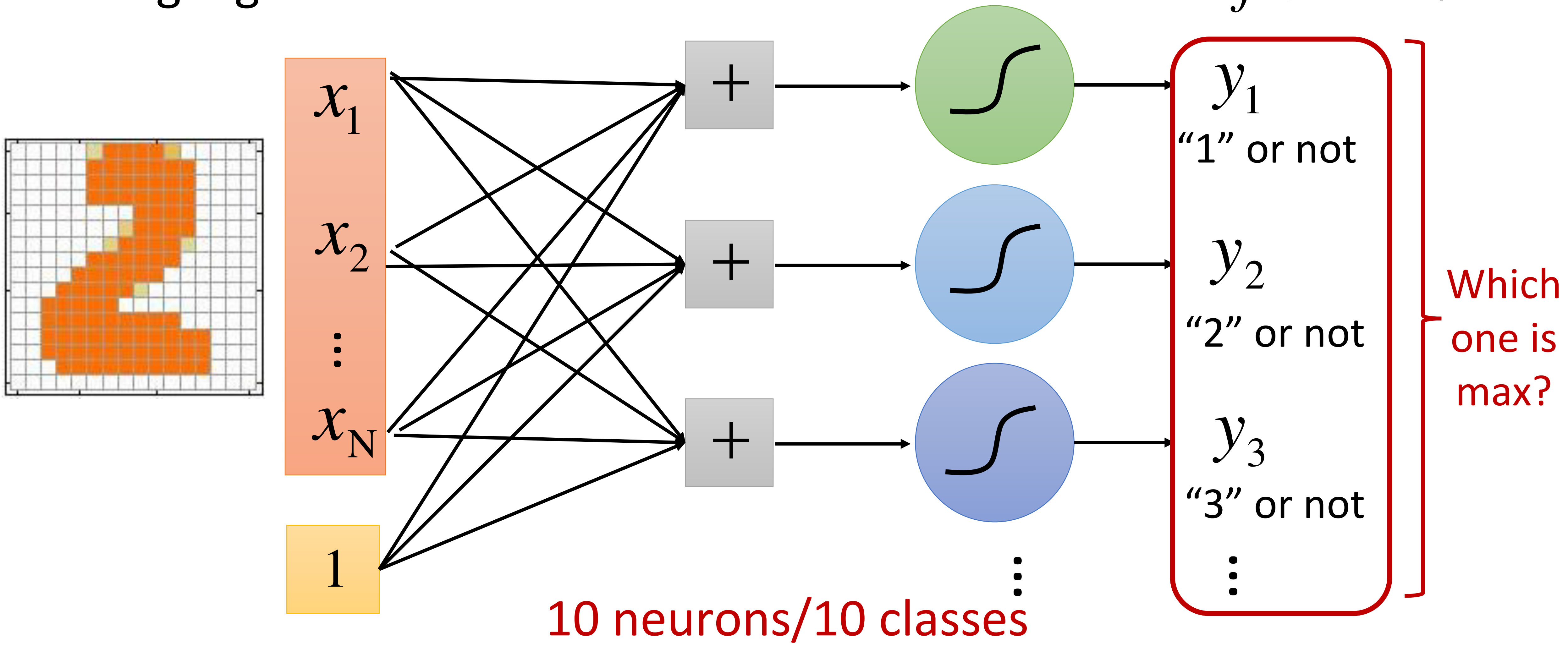
$$\begin{cases} is \quad "2" & y \geq 0.5 \\ not \; "2" & y < 0.5 \end{cases}$$

A single neuron can only handle binary classification

A Layer of Neurons

- Handwriting digit classification

$$f: R^N \rightarrow R^M$$

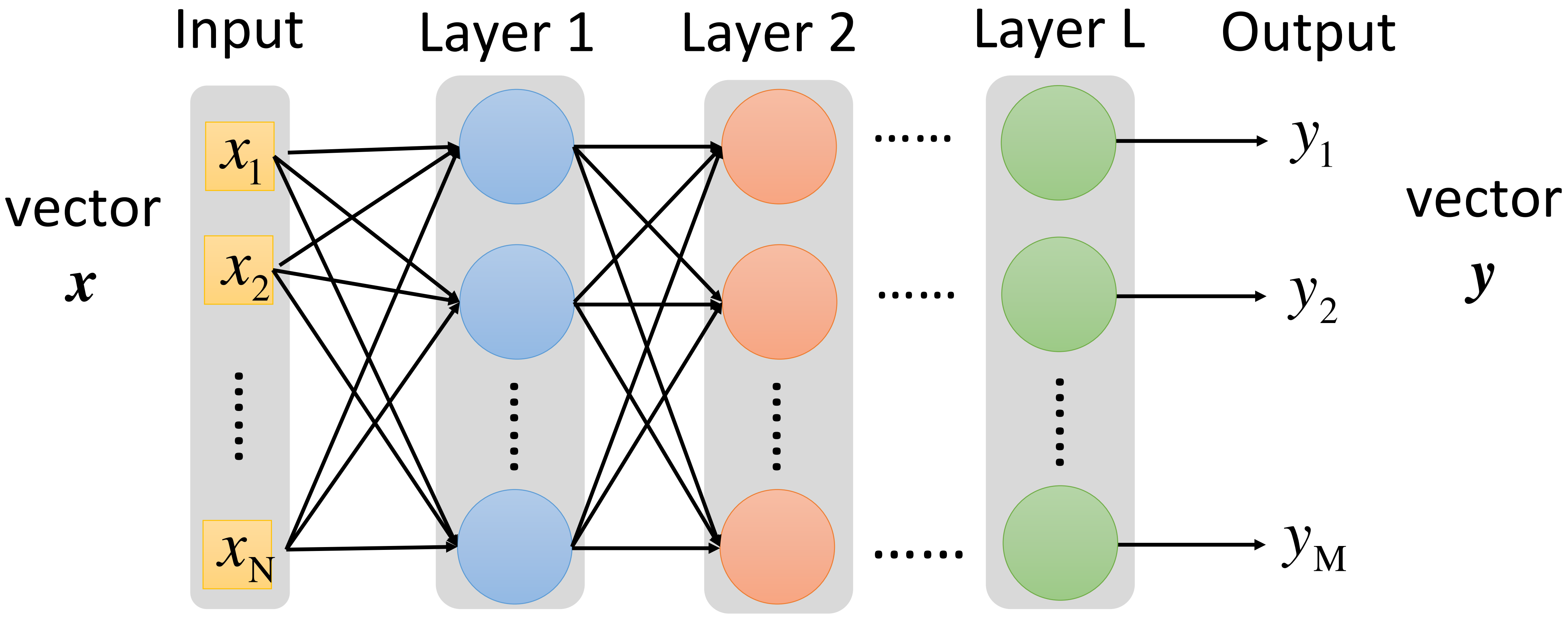


A layer of neurons can handle multiple possible output, and the result depends on the max one

Deep Neural Networks (DNN)

- Fully connected feedforward network

$$f : R^N \rightarrow R^M$$

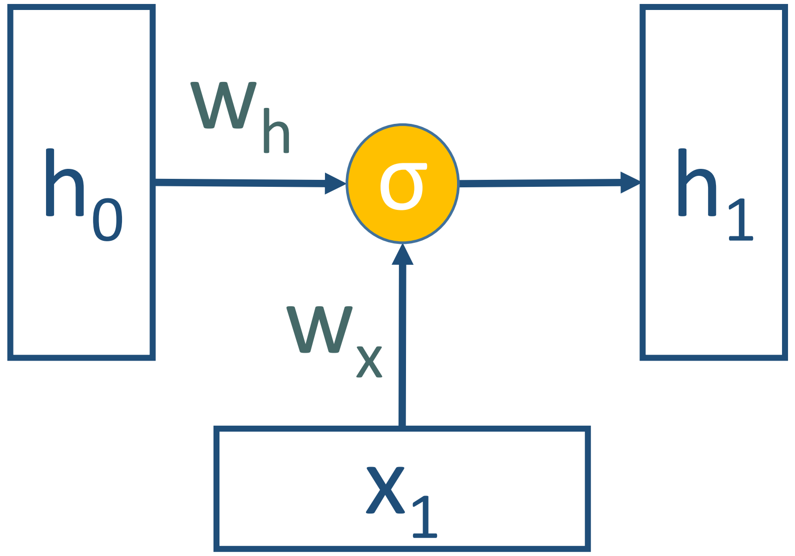


Deep NN: multiple hidden layers

Recurrent Neural Networks (RNNs)

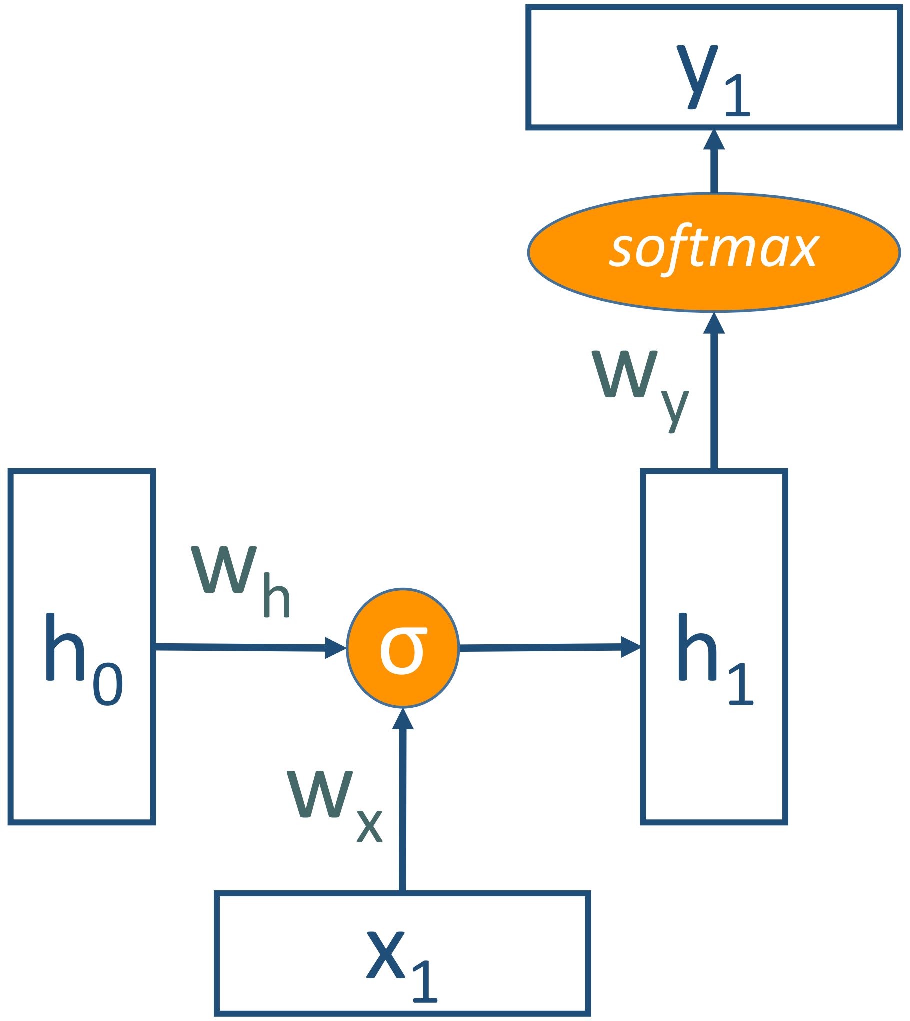
$$h_t = \sigma(W_h h_{t-1} + W_x x_t)$$

$\sigma(\cdot)$: tanh, ReLU



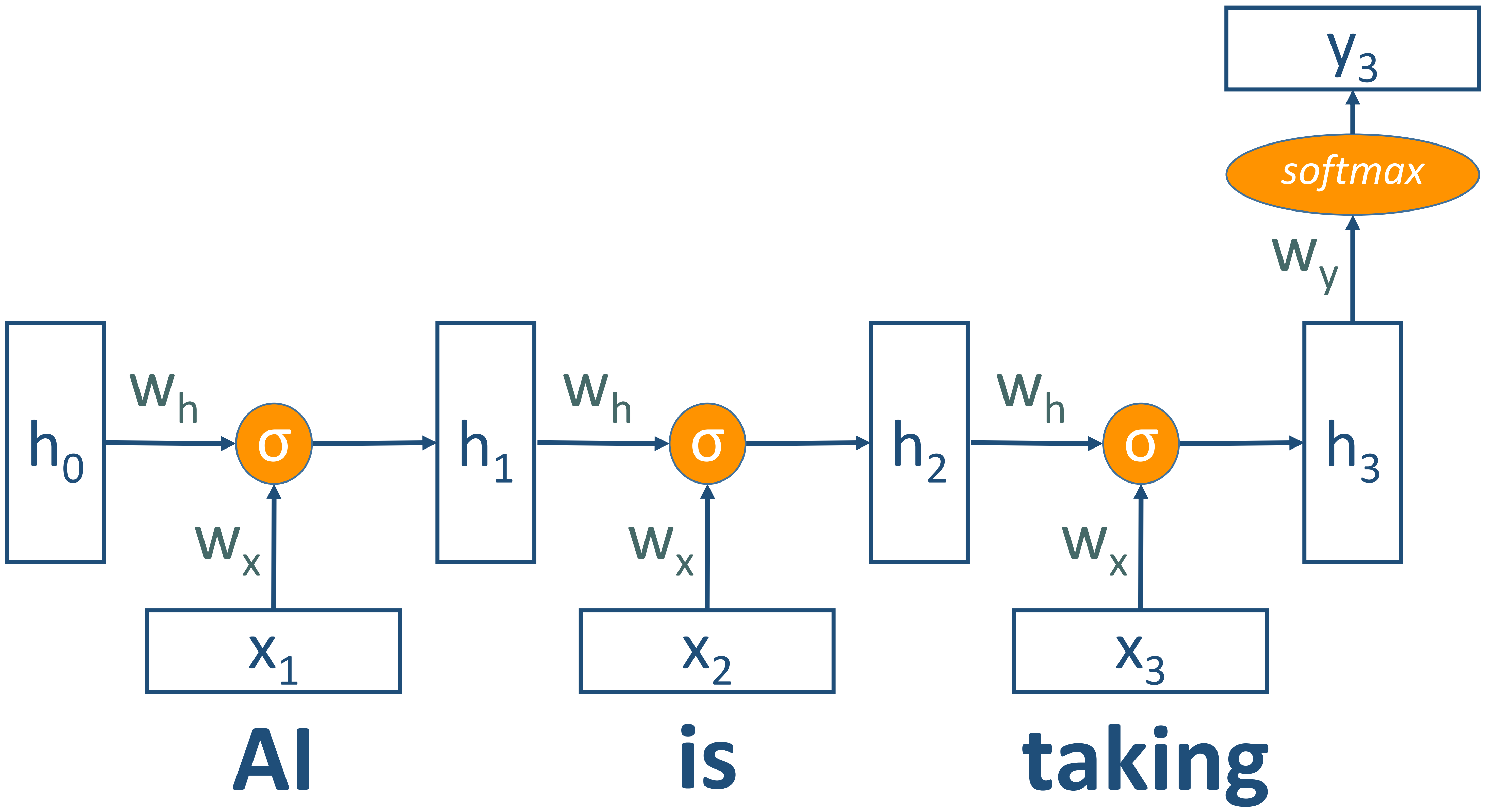
RNN can learn accumulated sequential information (time-series)

RNN



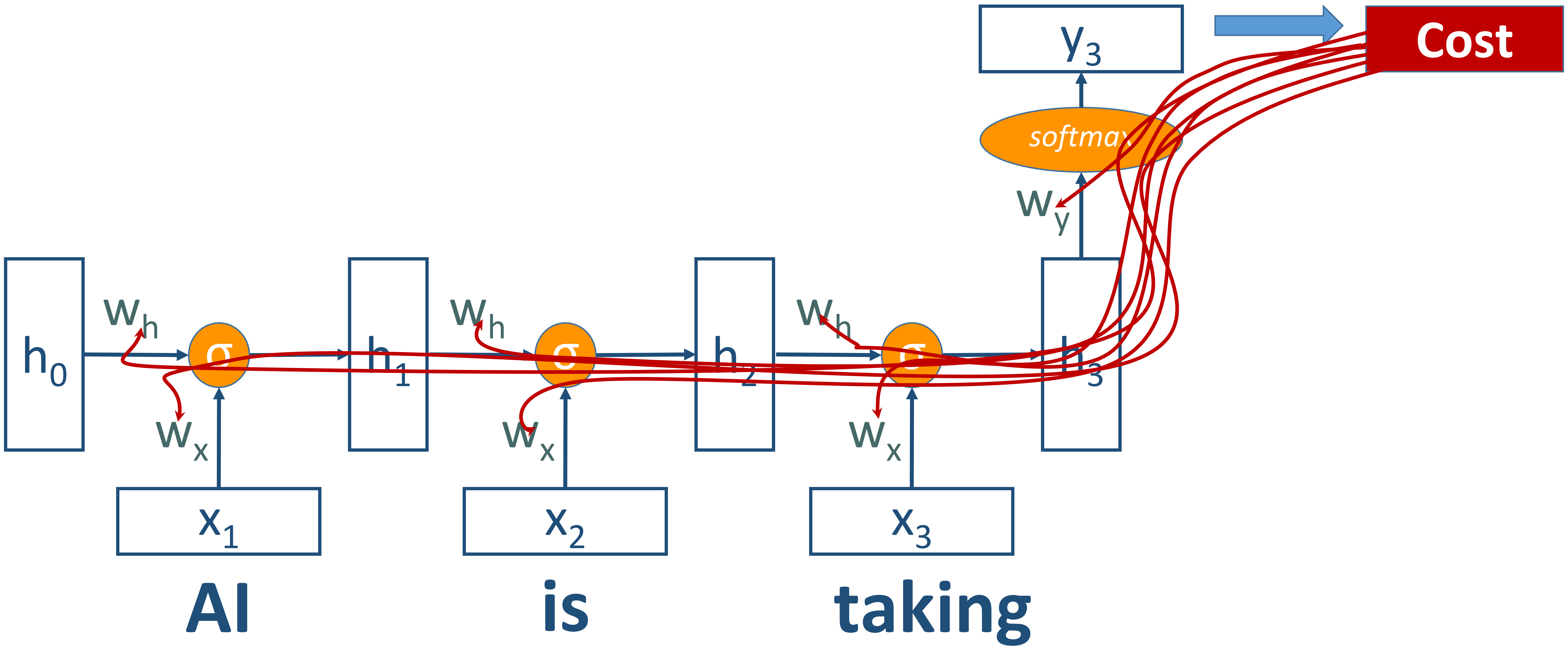
$$h_t = \sigma(W_h h_{t-1} + W_x x_t)$$
$$y_t = softmax(W_y h_t)$$

RNN

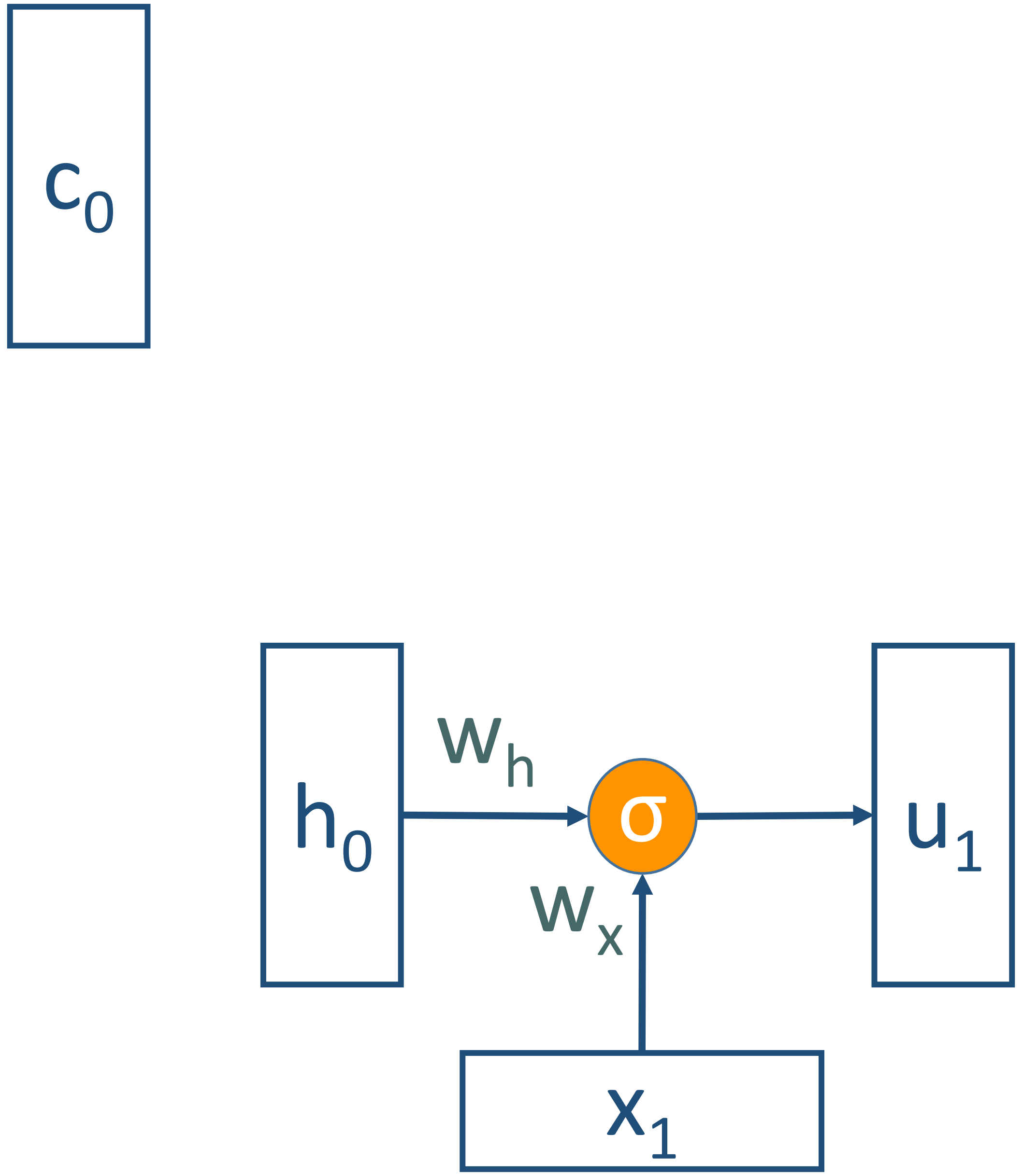


y_3
softmax
w_y
h_0
w_h
σ
h_1
w_h
σ
h_2
w_h
σ
h_3
w_x
w_x
w_x
x_1
x_2
x_3
AI
is
taking

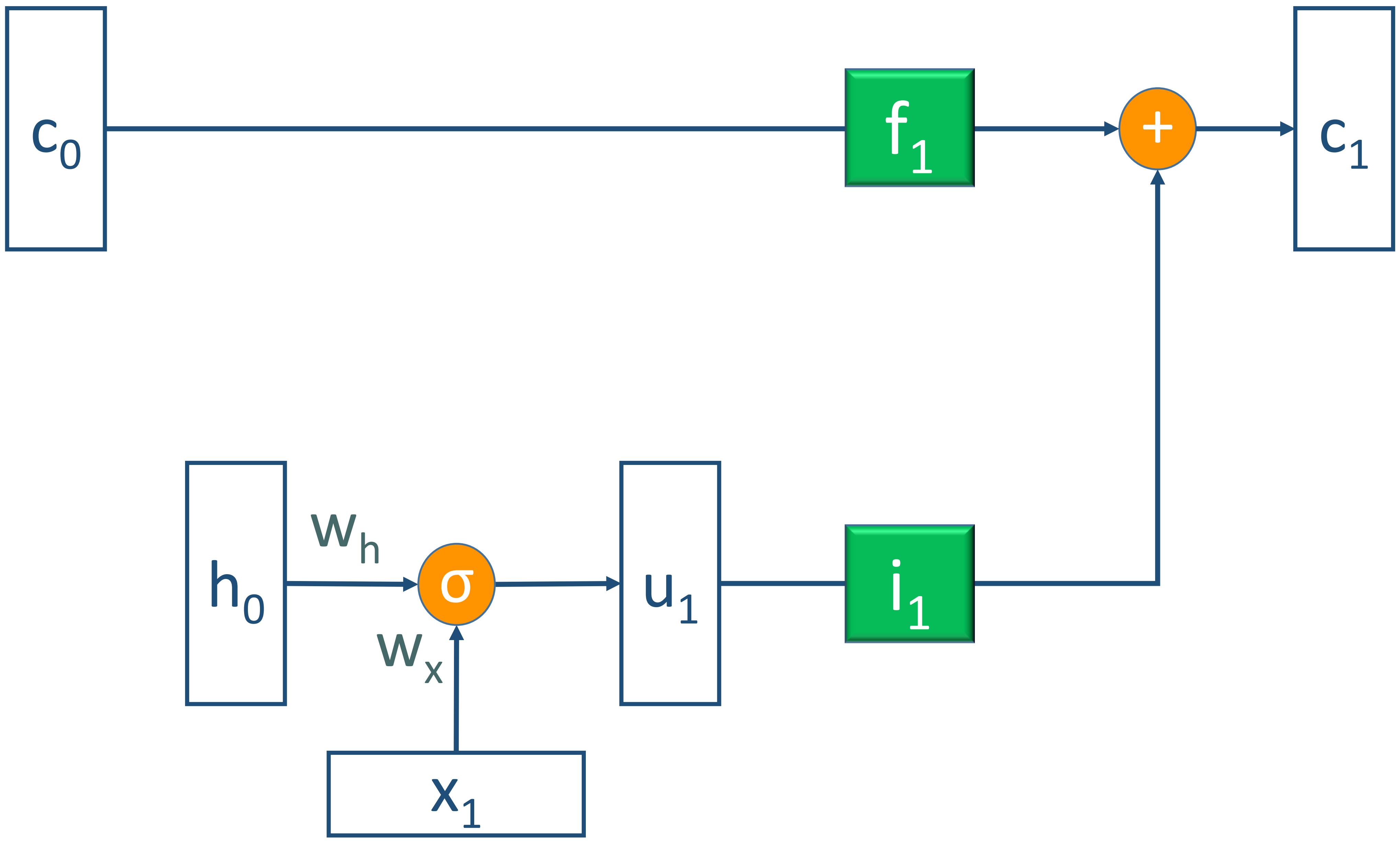
Updating Parameters of an RNN



Converting RNN to LSTM

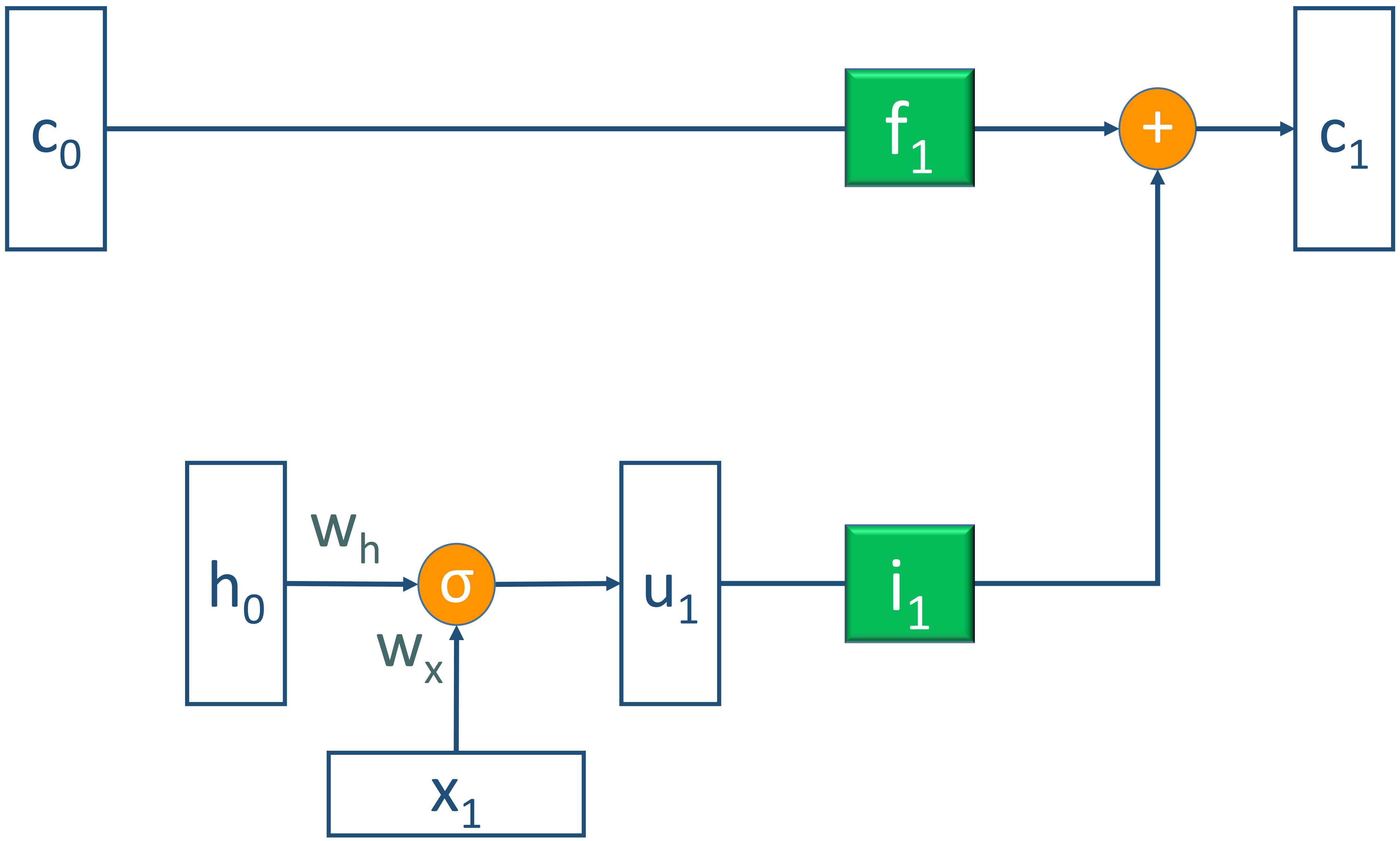


Transforming RNN to LSTM



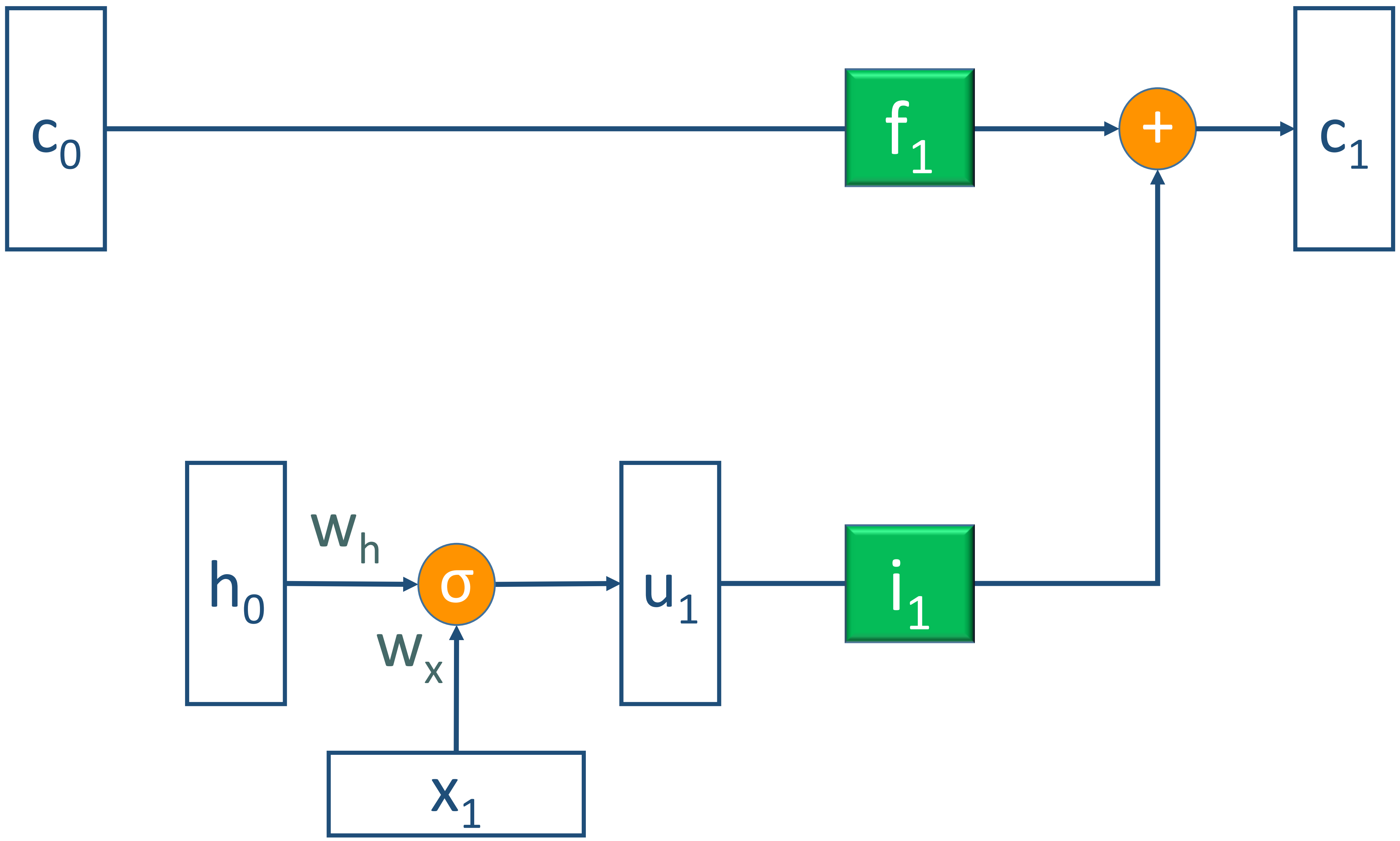
$$c_t = f_t \odot c_{t-1} + i_t \odot u_t$$

Transforming RNN to LSTM



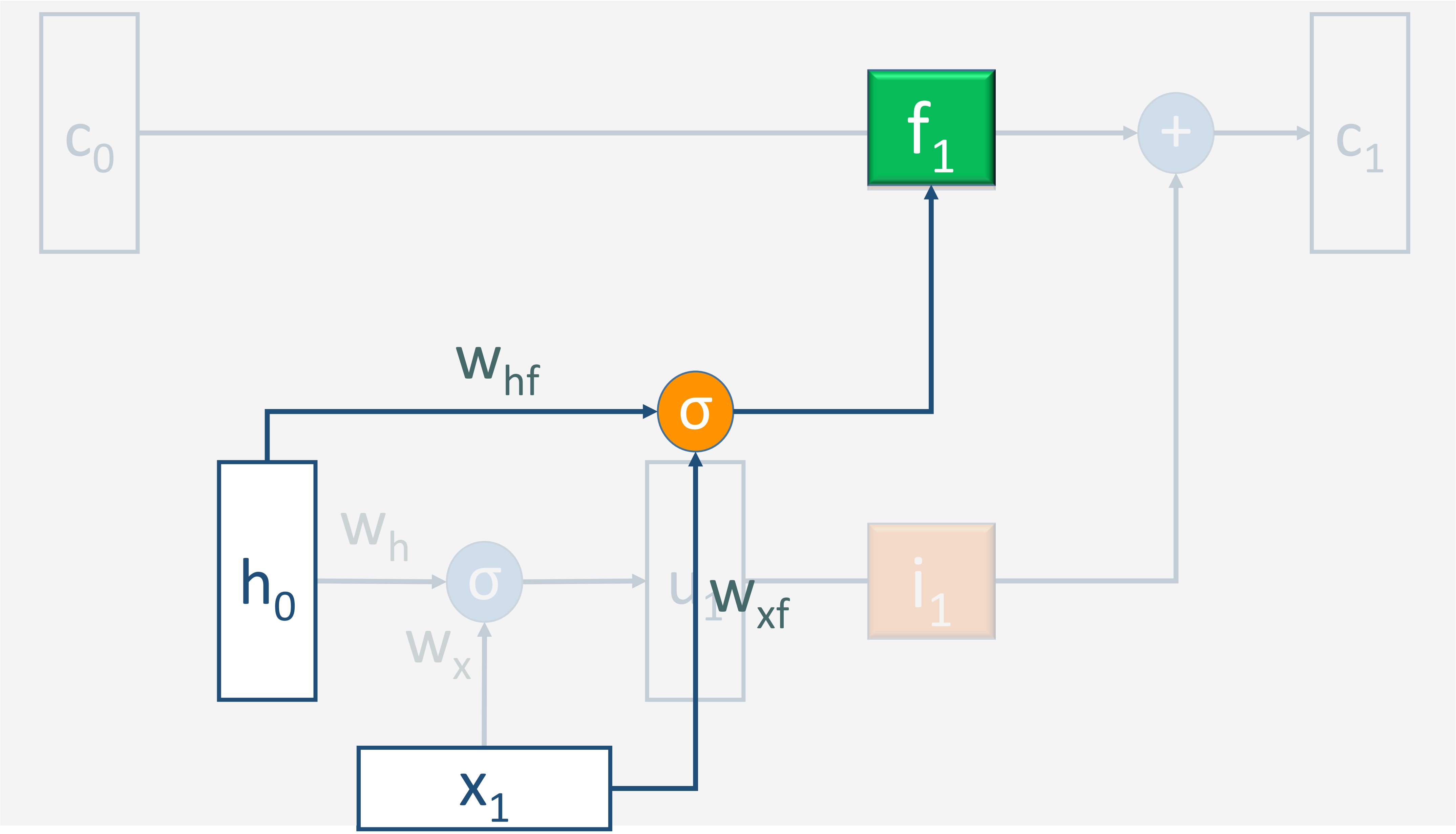
$$c_t = f_t \odot c_{t-1} + i_t \odot u_t$$

Transforming RNN to LSTM



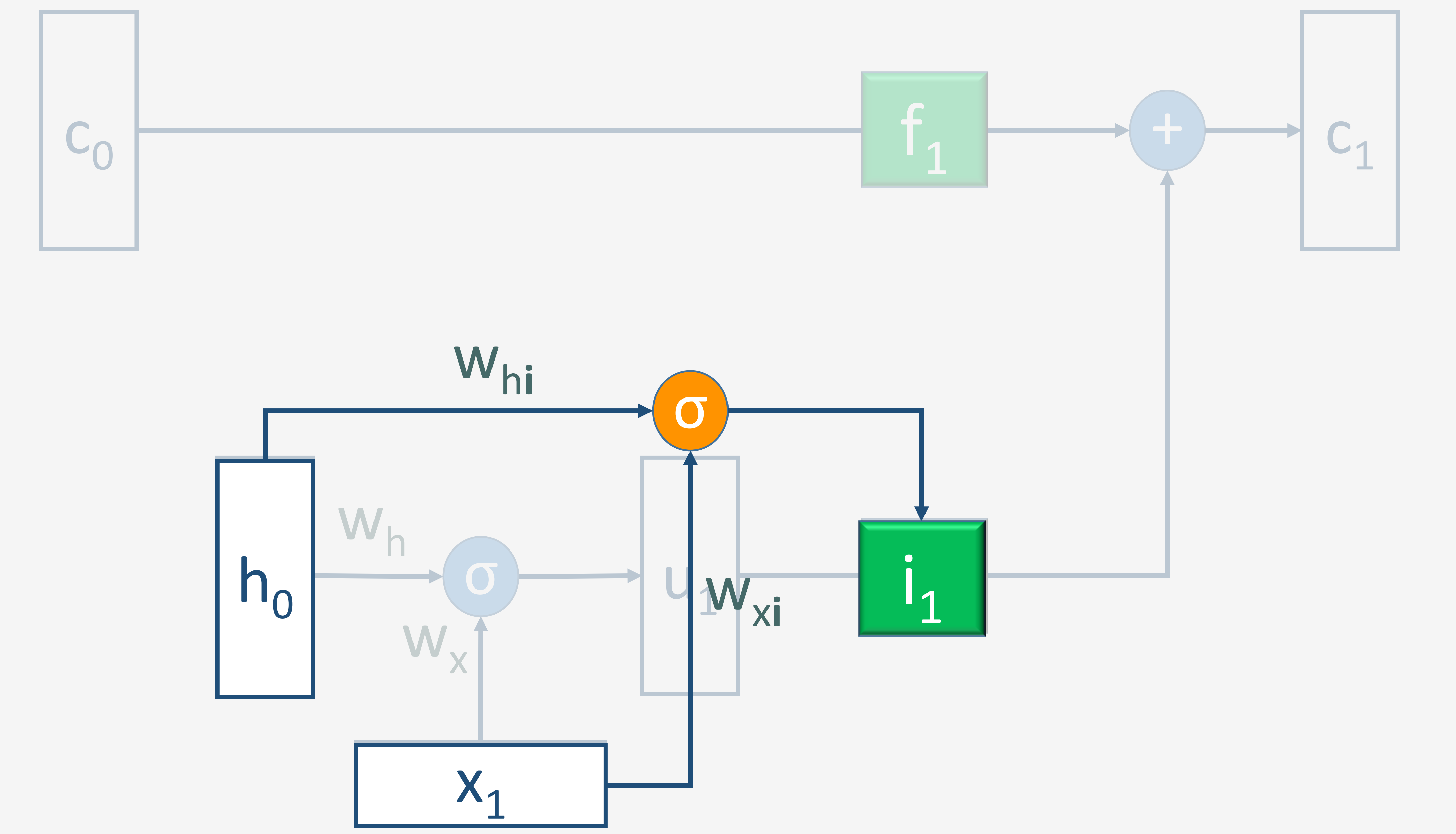
$$c_t = f_t \odot c_{t-1} + i_t \odot u_t$$

Transforming RNN to LSTM



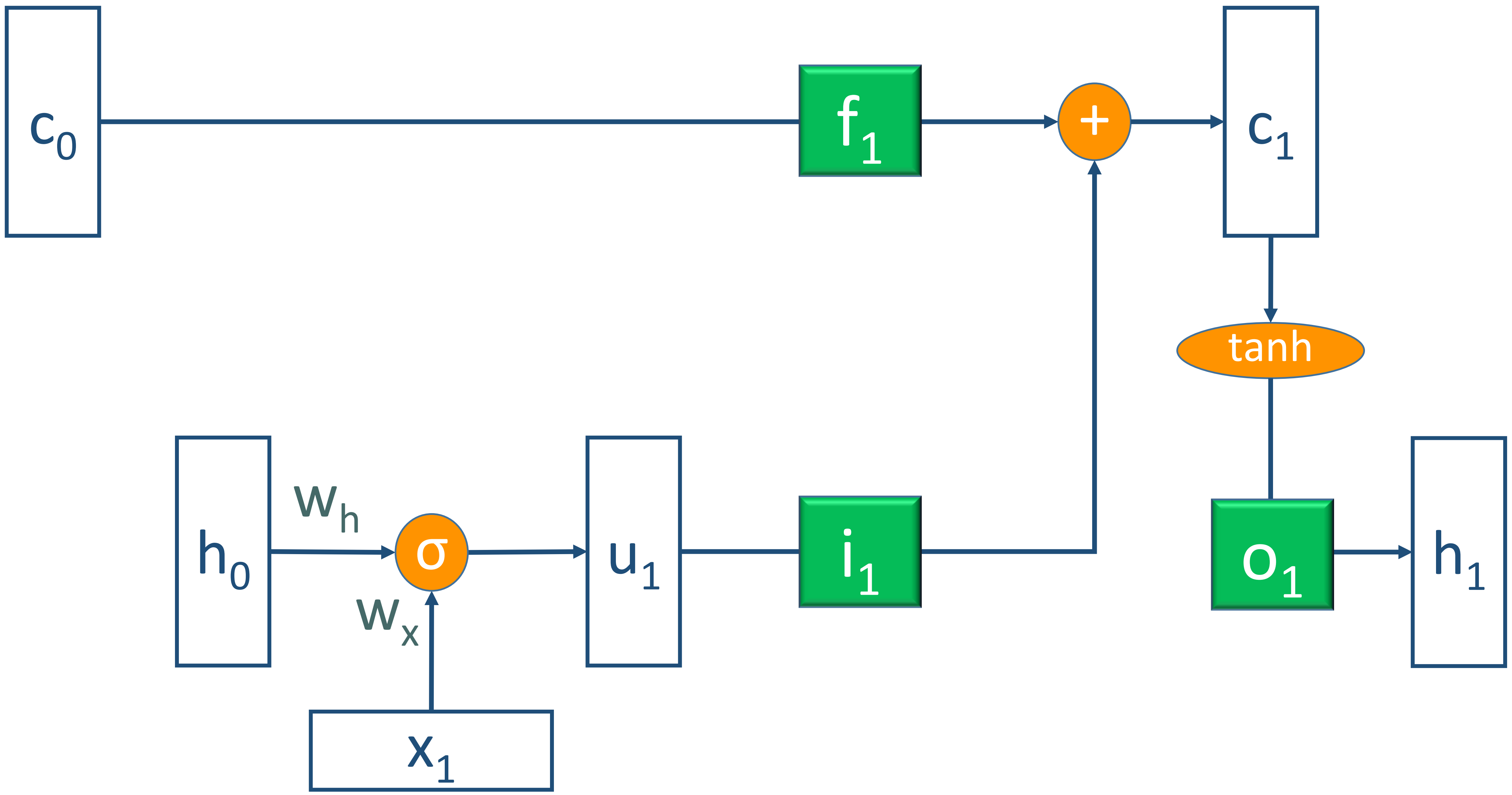
$$f_t = \sigma(W_{hf} h_{t-1} + W_{xf} x_t)$$

Transforming RNN to LSTM



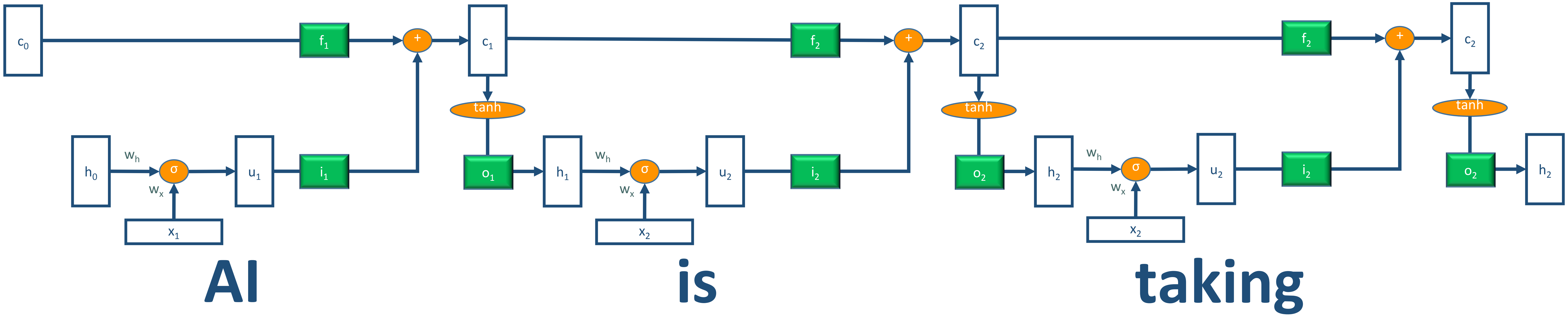
$$i_t = \sigma(W_{hi}h_{t-1} + W_{xi}x_t)$$

Transforming RNN to LSTM



$$h_t = o_t \odot \tanh c_t$$

LSTM for Sequences



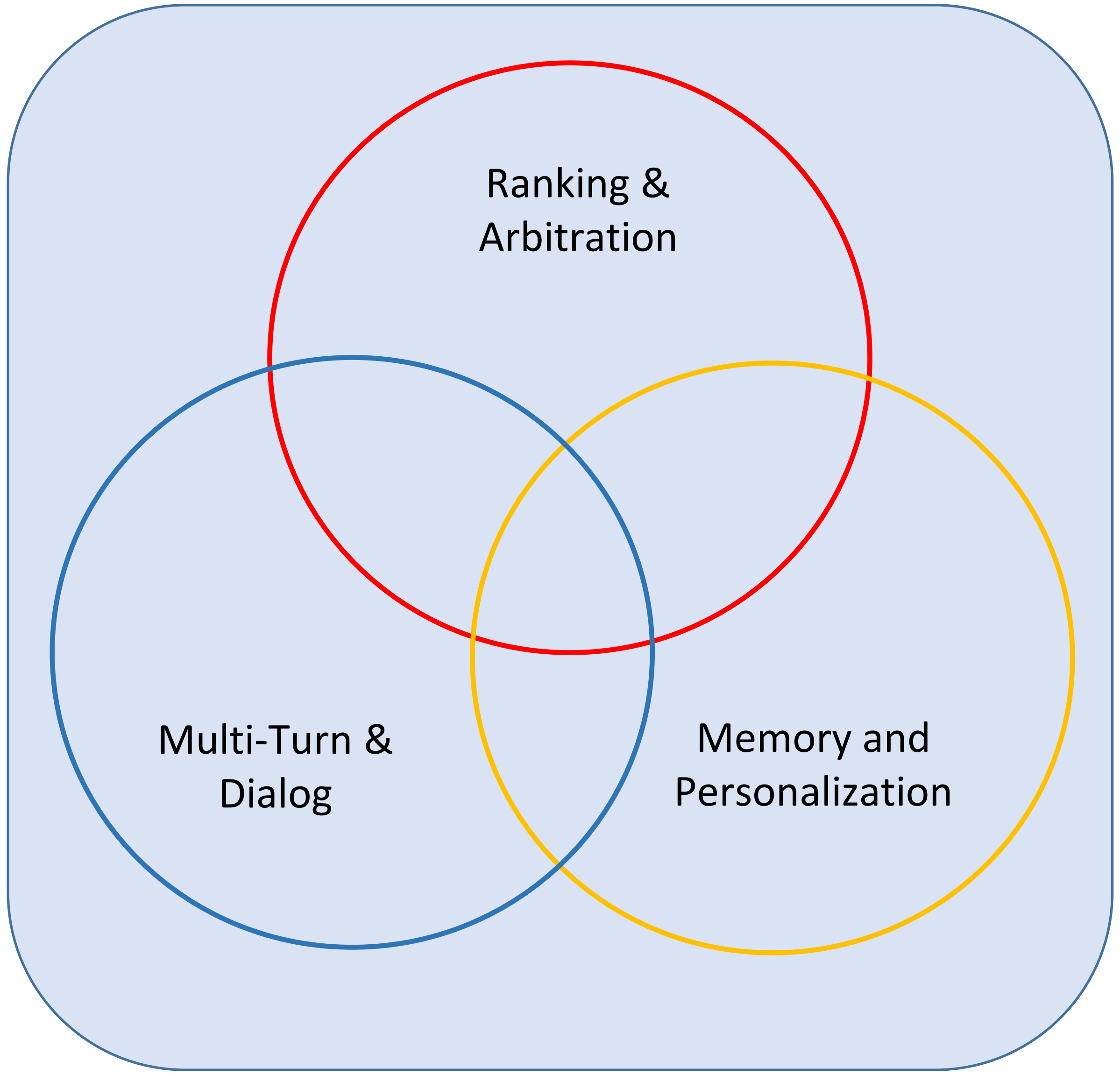
Alexa Brain

Removing Natural Language Interaction Frictions

- Skill arbitration and direct interaction
  - [Friction #1]: App discovery
  - [Friction #2]: Limited cognitive bandwidth to learn how each app works
- Context carryover across multiple turns
- Alexa Memory: “Remember This!”

- Official Amazon Blog Posts:

https://amzn.to/2JqrjyI
https://developer.amazon.com/blogs/alexa/post/352e9834-0a98-4868-8d94-c2746b794ce9/improve-alexa-skill-discovery-and-name-free-use-of-your-skill-with-canfulfillintentrequest-beta



Natural Language Skill Interaction Challenges

Ambiguity

- “Alexa, what should I do for dinner?”
  - What is the user’s intent?
    - book_restaurant? order_food? find_recipe?
- “Alexa, find hunger games?”
  - What is the user’s intent?
    - play_music? play_video? play_audiobook? buy_item?
- “Alexa, schedule a package pickup for 6pm”
  - Which domain/skill should handle this request?
    - Fedex? UPS?
- “Alexa, order me lunch without delivery fee?”
  - Which domain/skill should handle it?
    - GrubHub? Peach? Amazon Restaurants?
- “Alexa, what is X?” (e.g. X is Flu, AI, GPU)
  - Which domain/skill has an answer?
    - QnA (Evi)? WebMD? Wikipedia?

3P Skill Interaction Model

- “Alexa, ask X to do Y”
- Natural interaction with 40K+ 3rd party skills (similar to 1st party domain experience)

1st time use

”Alexa, how can I remove the oil stain”

”Tide stain remover can help you with that, do want to enable it?”

”Yes, please”

2nd time use

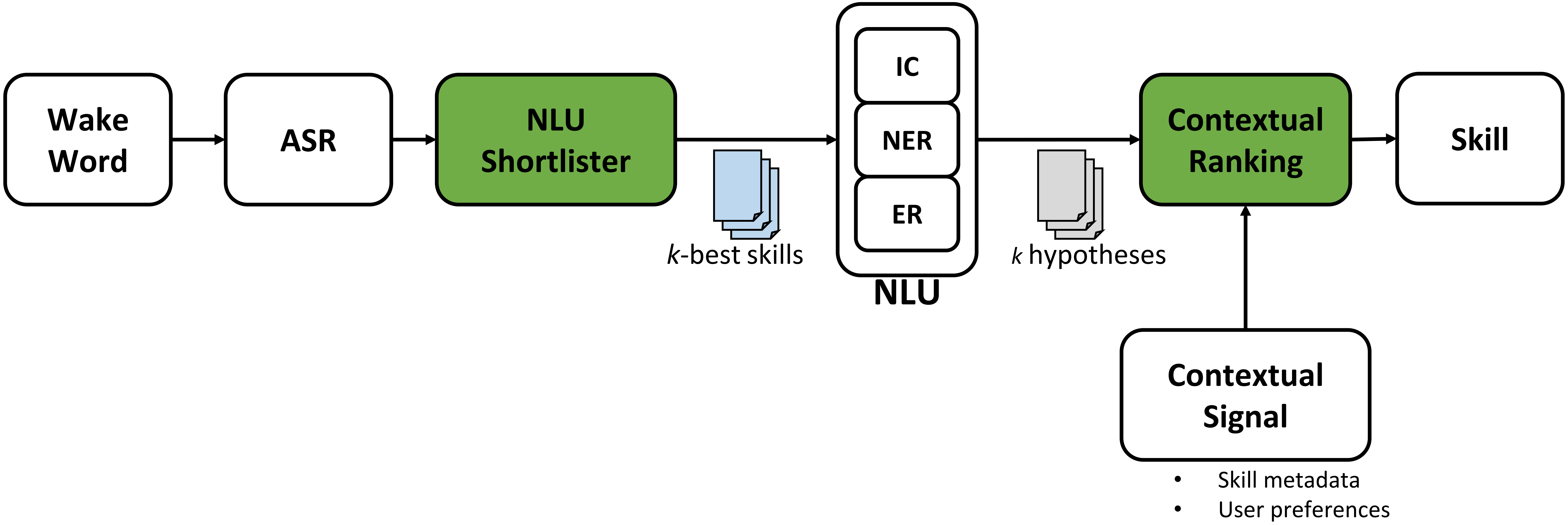
”Alexa, how can I remove the oil stain out of my shirt

Ok, Tide Stain Remover can help you with that. → Oil stain, I will guide you removing the oil stain ....”

- 81% of low skill usage (relative to 1P domains) reasons are attributed to frictions around skills discovery, awareness, invocation and recall.

Solutions to Natural Language Interaction Frictions

High-Level System Architecture



NLU Shortlister with Personalized Attention

(YB Kim et al., ACL2018)

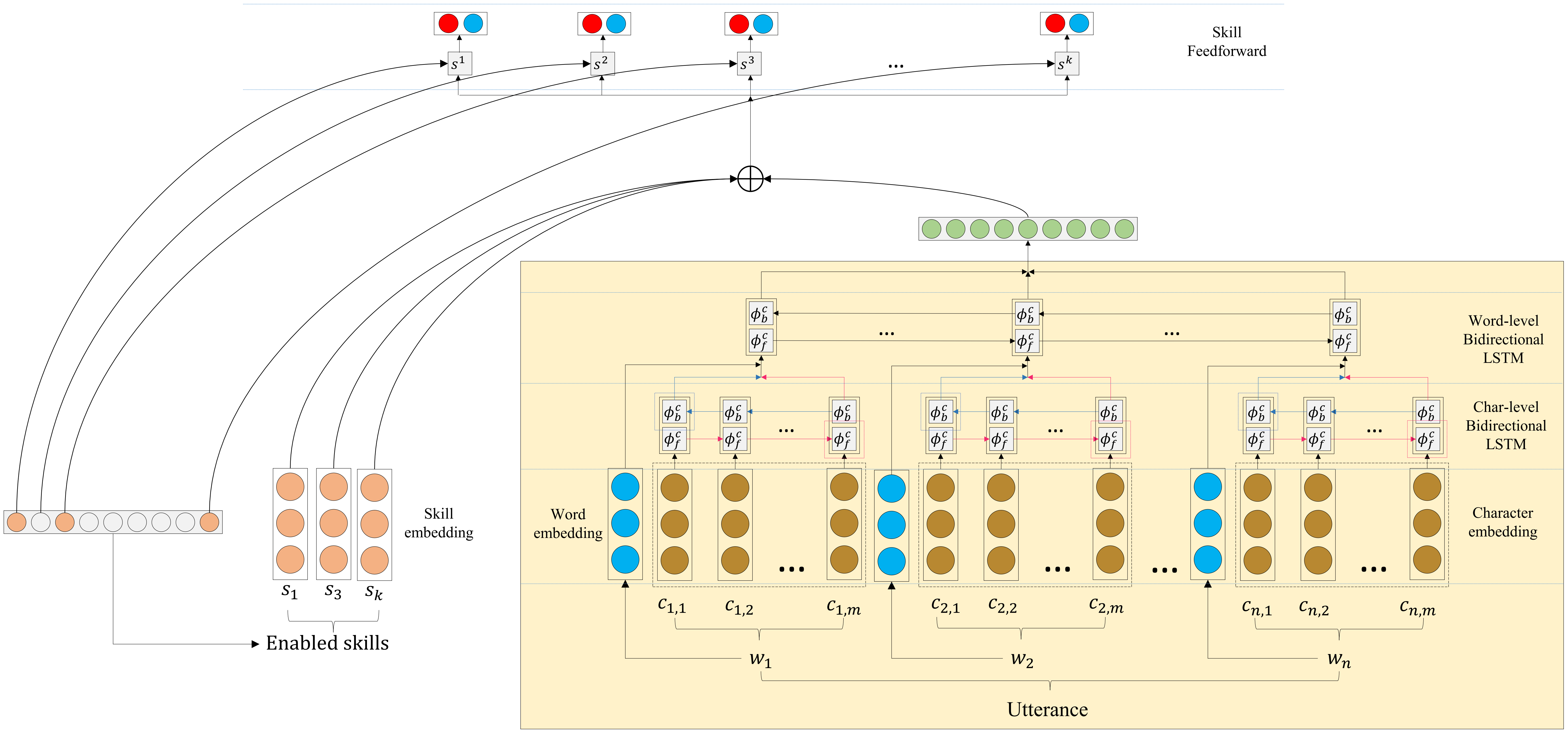
k-dimensional enabled skill n-hot vector

100-dimensional encoded vector

1 Shared encoder across all domains

Binary vector

s^k 201 × 2



NLU Skill Shortlisting Accuracy

(YB Kim et al., ACL2018)

Modeling (1500 skills)	TestSet1			TestSet2		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Binary	78.29	87.90	88.92	73.79	85.35	86.45
MultiClass	78.58	87.12	88.11	73.78	84.54	85.55
MultiTask	80.46	89.27	90.16	75.66	86.48	87.66
1-Bit Flag	91.97	95.89	96.68	86.50	92.47	93.09
Attention*	94.83	97.11	98.35	89.64	95.39	96.70
1-Bit + Att	**95.19**	**97.32**	**98.64**	**89.65**	**95.79**	**96.98**

- **Binary** trains a separate binary classifier for each skill.
- **MultiClass** has a shared encoder followed by a softmax.
- **MultiTask** replaces the softmax with per-skill classifiers.
    - **1-Bit Flag** adds a single bit for personalization to each skill classifier in MultiTask.
    - **Attention** extends MultiTask with personalized attention. The last 3 models are personalized.
- Personalization signals carry strong disambiguating information

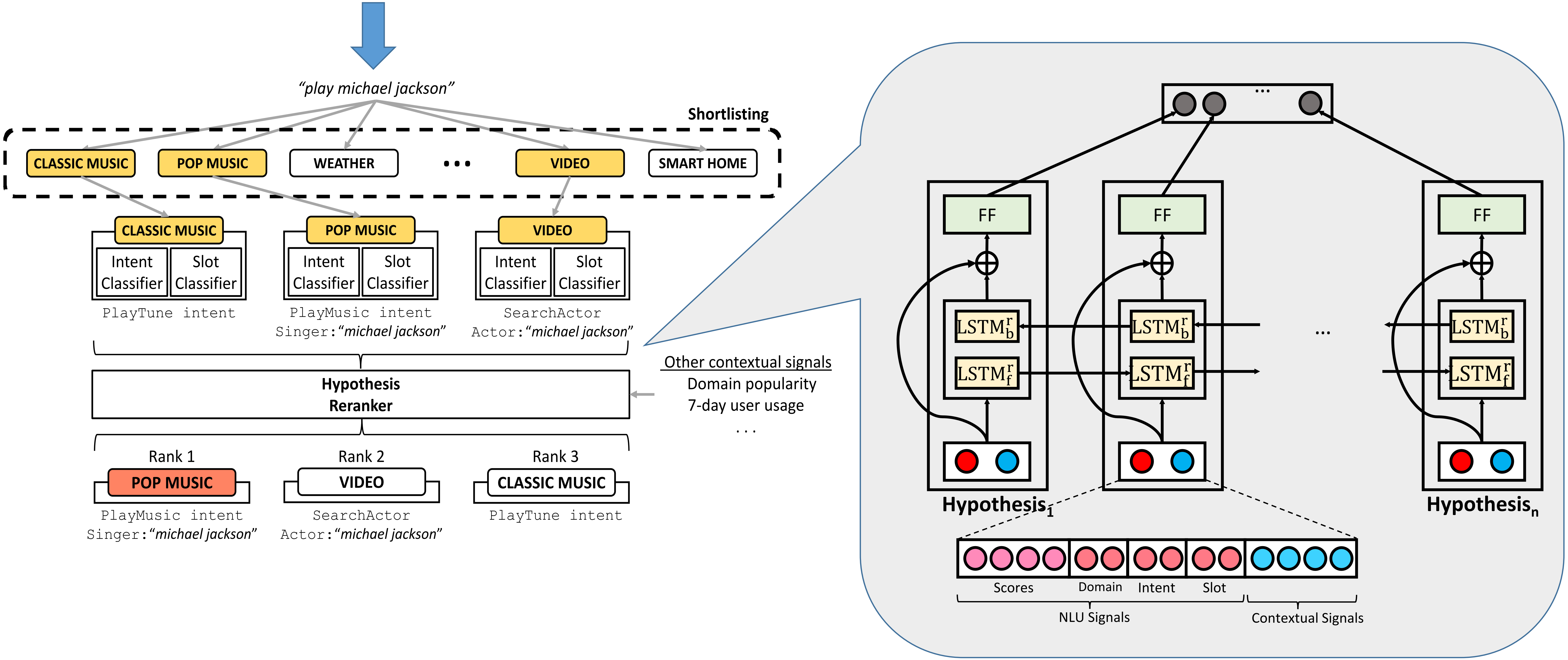
Modeling (Adding 20 new skills)	Time	Accuracy
Binary	34.81	78.13
Expand	30.34	94.03
Refresh	5300.18	94.58

Comparison of per-epoch training time (seconds) and top-1 accuracy (%) on an NVIDIA Tesla M40 GPU.

- Adding a new skill is two orders of magnitude faster than retraining the entire model
- Adding a new skill adds only around 1K parameters (4KB to runtime memory footprint)

Final Ranker: Contextual Ranking

(YB Kim et al., NAACL 2018)



Contextual Ranking Results

(YB Kim et al., NAACL 2018)

Model	Small-Scale (21 Alexa Domains)	Large-Scale (1500 Alexa Skills)
Nonpersonalized Shortlister (LSTM)	95.56	81.49
Logistic Regression (LSTM)	95.59	87.50
$LSTM^{C}$	**97.55**	**93.83**
$LSTM^{CH}$	97.34	93.46
UPPER BOUND	98.77	95.93

Logistic Regression
Hypothesis vector (HYP) → ranking

$LSTM^{C}$: (HYP → LSTM) ⊕ HYP → FF → ranking
$LSTM^{CH}$: (HYP → LSTM) ⊕ HYP ⊕ cross-hypothesis features → FF → ranking

UPPER BOUND: Upper bound of shortlister 5-best accuracy

Cross-hypothesis features

- Ratio of Shortlister scores to the maximum score
- Relative number of slots across all hypotheses
- …..

- Our LSTM-based contextual ranker comes close to the 5-best accuracy (i.e. UPPER-BOUND) from the Shortlister
- Better than the approach that uses manual cross-hypothesis features.

Context Carryover and Resolution

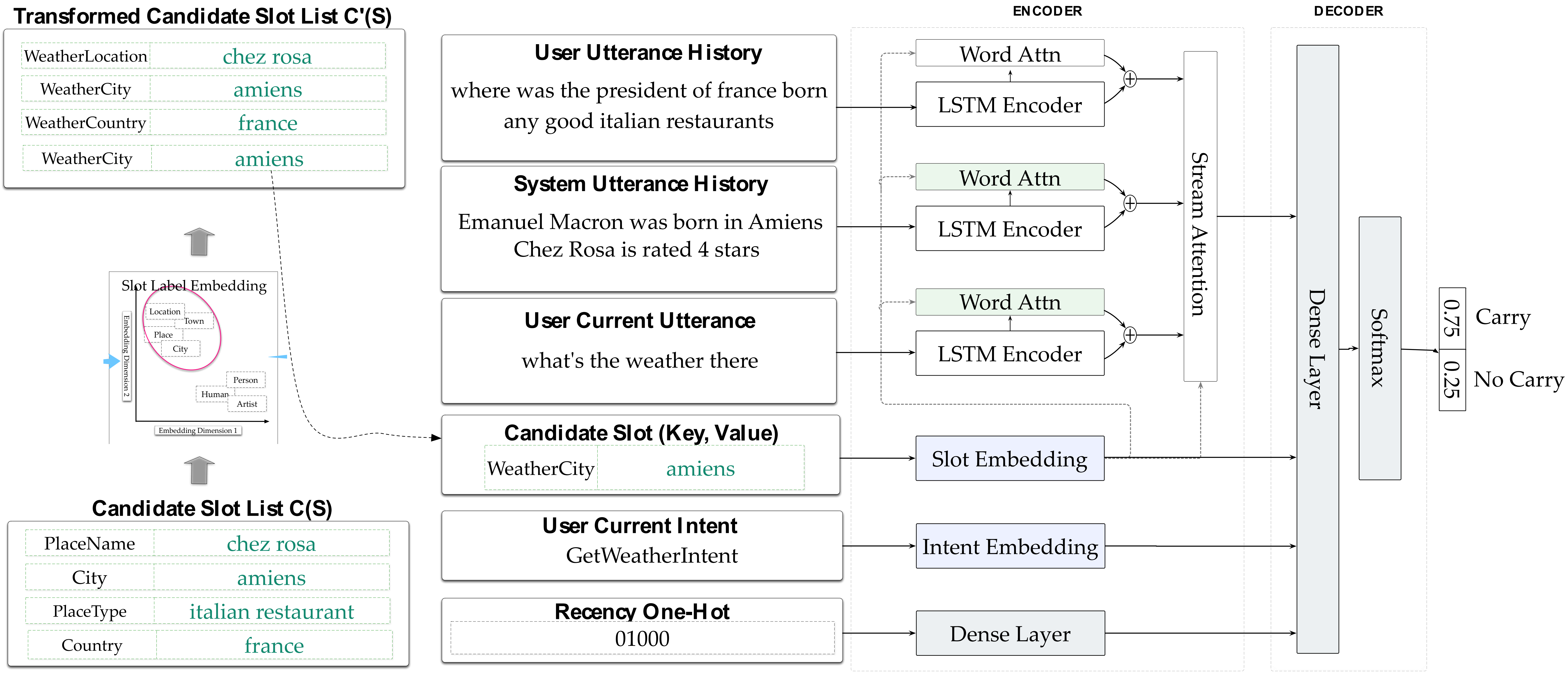
- Carrying/dropping both intents and slots across user and system turns
  - Within and cross domain
- Slot Carryover

Turn	Utterance	Current Turn Slots	Carried Slots
User T1	Where was the president of France born	GeographicalArea:France	
Alexa T1	Emmanuel Macron was born in Amiens	Person:Emmanuel Macron TownOrCity: Amiens	
User T2	Any good Italian restaurants	PlaceType:Italian restaurants	City:Amiens Country:France
Alexa T2	Chez Rosa is rated 4 star	PlaceName: Chez Rosa	
User T3	What's the weather there		WeatherCity: Amiens WeatherCountry:France

- Intent Carryover
  - "Alexa, how is the weather in Seattle?" → "Alexa, how about this weekend?"
  - "Alexa, how is my schedule?" → "Alexa, how about this weekend?"

- Challenge: Disparate schemas across domains
  - "Alexa, how is the weather in Redmond?" [Weather: WeatherLocation=Redmond]
  - "Alexa, what are the nice restaurants there?" [Local Search: City=Redmond]

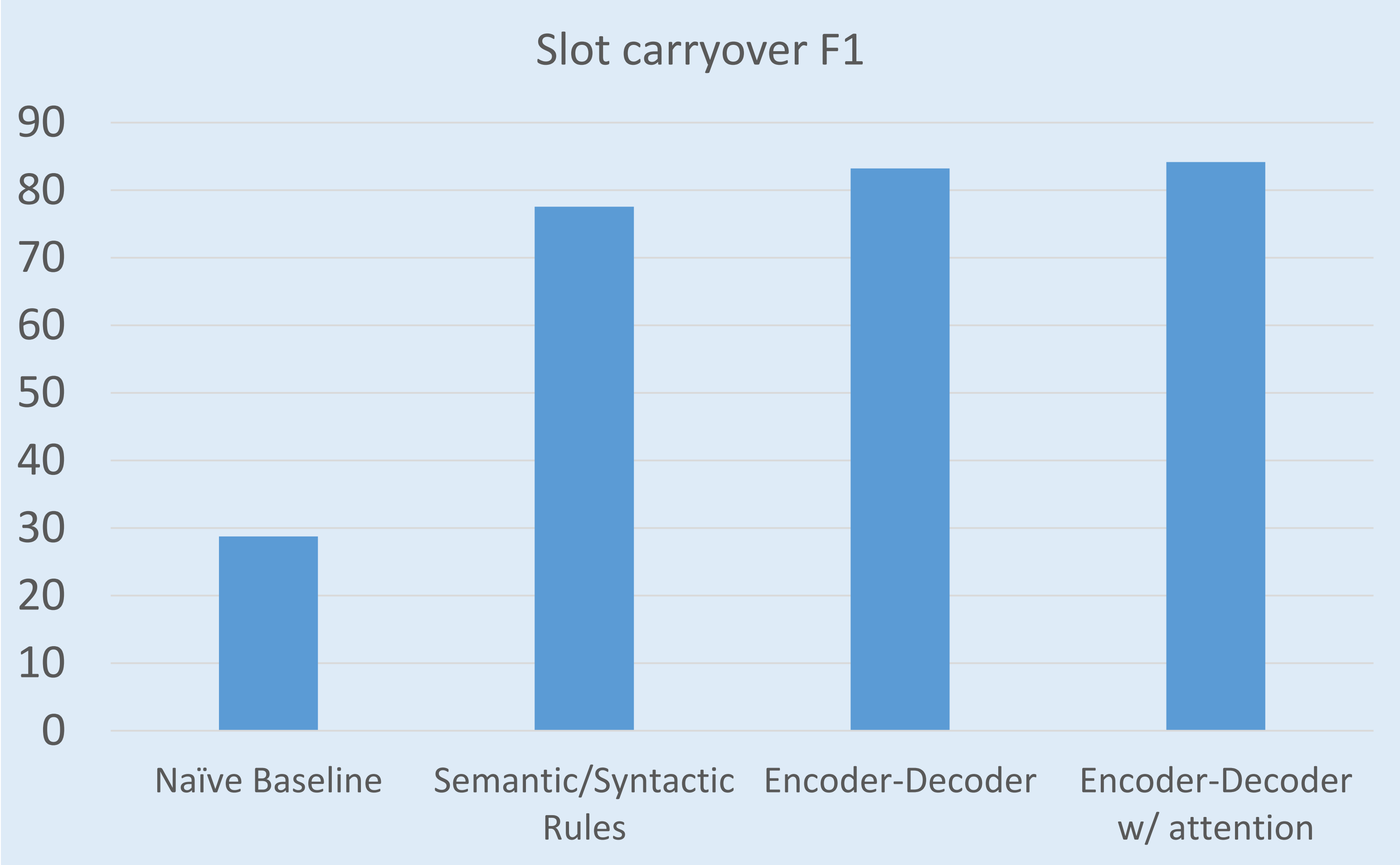
Context Carry Over Model Architecture

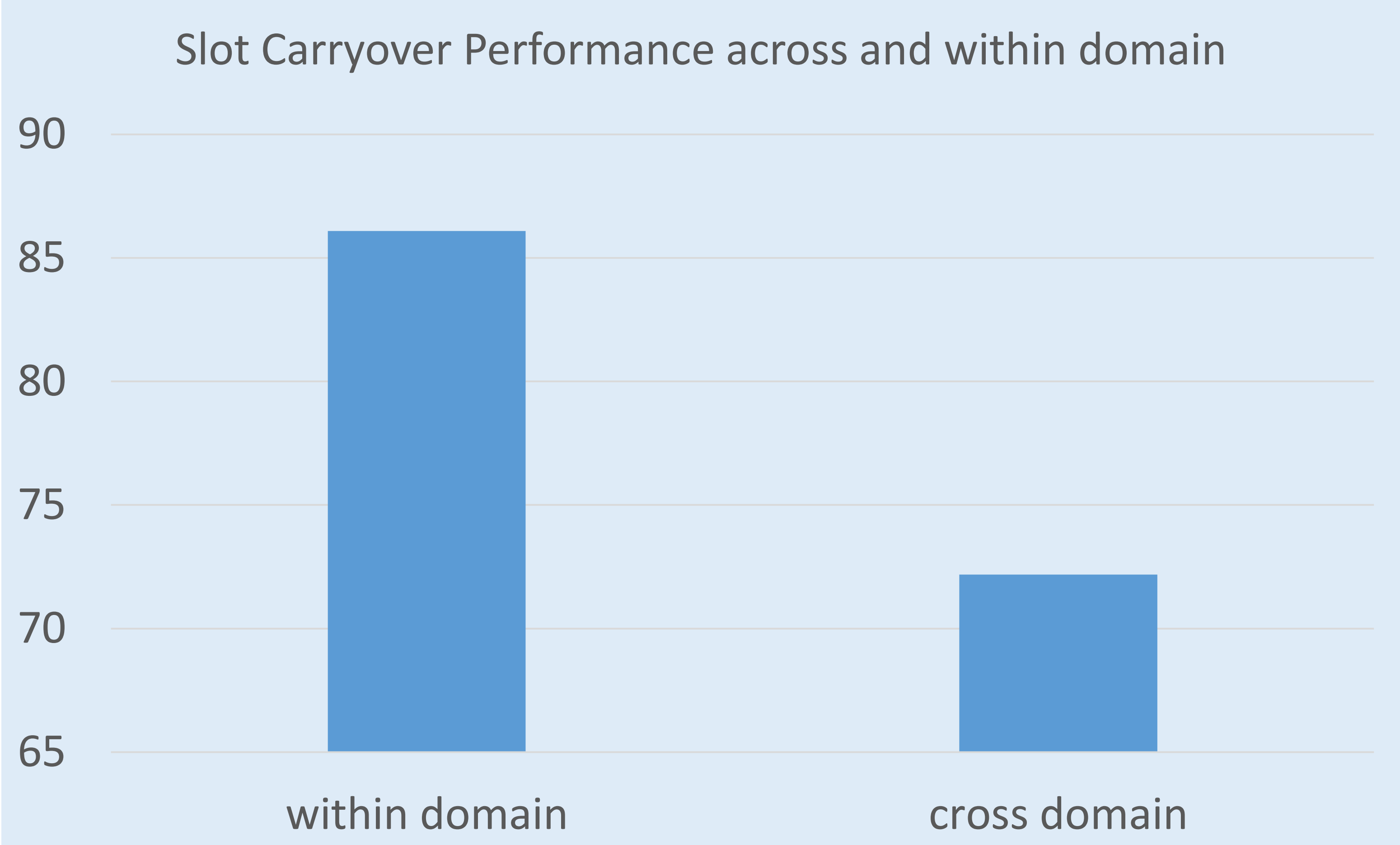


Experiment Results

Data: 16K training, 2K dev and 2K test

- 8 domains: Weather, Music, Video, Books, HomeAutomation, LocalSearch, CinemaShowtimes, Knowledge
- On average 2.2 turns per session
- 20% of cross-domain interactions across disparate schemas
- Highly imbalanced data – on average 1 out of 4 candidates per turn positive





Key Observations

- Encoder-decoder gives significant improvements over a strong rule baseline
- Cross-domain carryover is significantly harder than within domain

Alexa Memory: “Alexa Remember This!”

- Motivation:
  - A spoken scratch pad.
  - Remembers any arbitrary piece of information.
  - Reduces the customer’s cognitive friction.
- Storing information:
  1) Alexa – remember that I gave my nephew a Batman Lego set for his 5th birthday.
  2) Alexa - remember that John's basketball practice is at 6pm on Saturdays.
  3) Alexa - remember that my nephew likes pistachio ice cream.
- Retrieval:
  - Alexa – What birthday present my nephew had for his 5th birthday? (retrieves 1)
  - Alexa - Which Lego set I bought my nephew last year? (another way to retrieve 1)
  - Alexa - What did I tell you about my nephew? (in this case Alexa should retrieve 1 and 3)

	Test Set (3497 questions)	
	Precision	Recall
Lexical search	0.97	0.70
+ Semantic matching	0.95	0.75
+ Phonetic representation	0.93	0.81

Intelligent Personal Digital Assistant Product Metrics

- How often users use PDA?
  - Daily Active Users (DAU)
  - Monthly Active Users (MAU)
  - DAU/MAU → overall engagement
- # queries handled (reactive)
- # suggestions & notifications and conversion rate (proactive)
- E2E Accuracy
  - Query/SystemResult (i.e. rendered UI) accuracy
  - Skill Routing Accuracy
- Competitive Analysis
  - Side-by-Side
- Revenue/Profit

Component Metrics

- Measurements are based either 1) offline human judgment, 2) online

	Metric	Description
LU	Domain classification P/R	Precision/recall of domain classification
	Intent classification accuracy	Accuracy of intent classifier
	Slot tagging P/R	Precision/recall of slot extraction + labeling
	Semantic frame accuracy	Accuracy of the whole semantic frame
Dialog	System Action with Parameters	Dialog contract accuracy
ASR	WER	Word Error Rate
	Display WER	
	SER	Sentence Error Rate
Routing Accuracy	Skill/answer selection	Answer Rate, Hit Rate, False Positive Rate
LG	Human Judgment, BLEU	
TTS	MOS	
Reactive	Offline: Defect Rate, SBS (relative metric) Online: CTR, action execution, time spend on per pixel	Measures the system E2E
Proactive	Defect Rate, SBS (relative metric) Online: CTR, time spend per pixel	Measures the product E2E

High-Level Technology Challenges

Experience scaling

- 1st party domain expansion
- 3rd party integration, tools, infrastructure
- Locale Expansion
- Alexa on many end-points (e.g. hotel room, car)

Arbitration among domains

- NLU ambiguity
- Answer relevance

End-to-End Measurement

- Testing and feedback loop
- Measuring the experience quality

Component Level Technology Challenges

Speech Recognition Challenges

- Noise, Speaker Accent, Speaker Tracking
- Side Speech, Unintentional Wake up Voice
- Open domain unlimited vocabulary

Spoken Language Understanding Challenges

- Rapid model development
- Open domain SLU, contextual SLU
- Quality scaling
- Difficulty of building reusable models (e.g. no shared schema)

Dialog Management

- Heterogeneous back-ends, custom interfaces
- Complex business logics, Lack of reusability

Natural Language Generation

- Agent Persona, Localization scaling

Proactive

- Personalization

It is still Day 1!

amazon alexa

Q & A