The Technology Powering Personal Digital Assistants

Ruhi Sarikaya
Microsoft Corporation

September 8th, 2015
Interspeech-2015
Dresden, Germany
Some slides (6-8) extended by Géza Németh

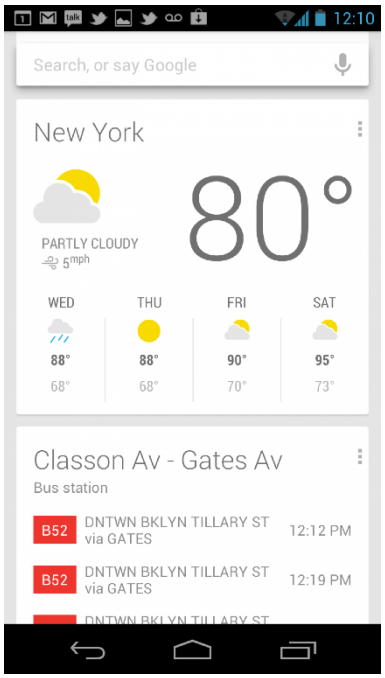
Talk Outline

- Personal Digital Assistants (PDAs)
  - What are they?
  - Why do we need them?
  - Why are companies investing into them?
- System architecture: proactive & reactive
- Proactive assistance and system components
- Reactive assistance and system components
- User Dissatisfactions (DSATs), Lessons Learned
- Metrics and Measurements
- Challenges and future

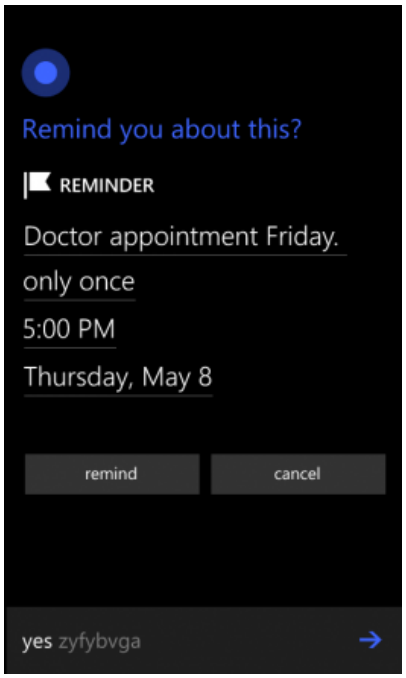
Siri vs. Google Now vs. Cortana vs. Echo

Siri (2011)

Google Now (2012)



Cortana (2014)



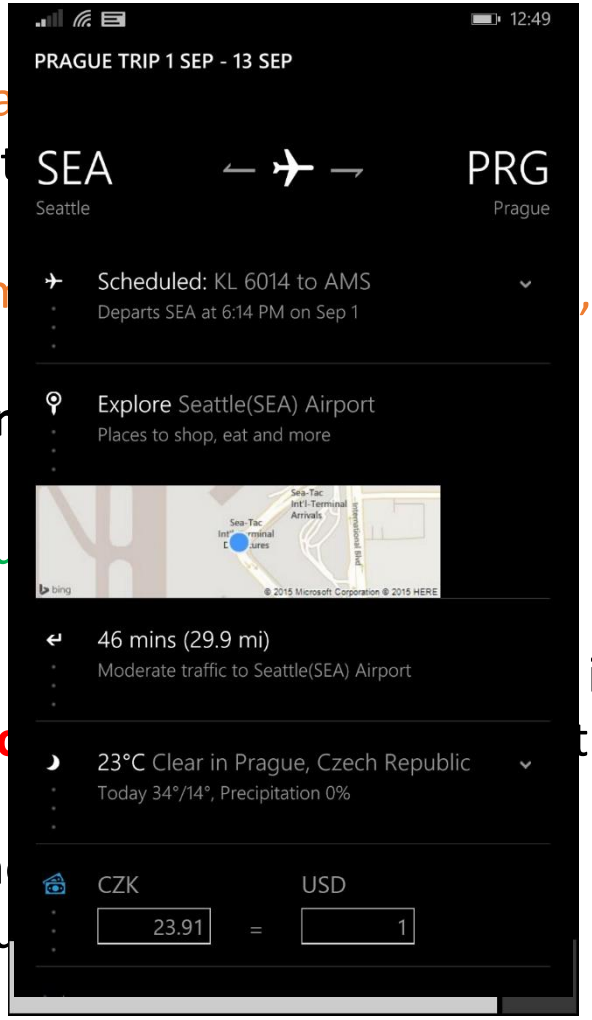
Alexa/Echo (2014)

What is a Personal Digital Assistant/Agent?

- **Meta layer of intelligence**
  - Sits on top of other services and applications
  - Performs actions using services/apps to fulfill the user's intent
  - Serves structured content and data to the user
  - Natural language interface
  - Relies on
    - machine learning, speech recognition, natural language understanding, dialog management, ranking, inference, personalization, etc..

Why do We Need a Personal Digital Assistant?

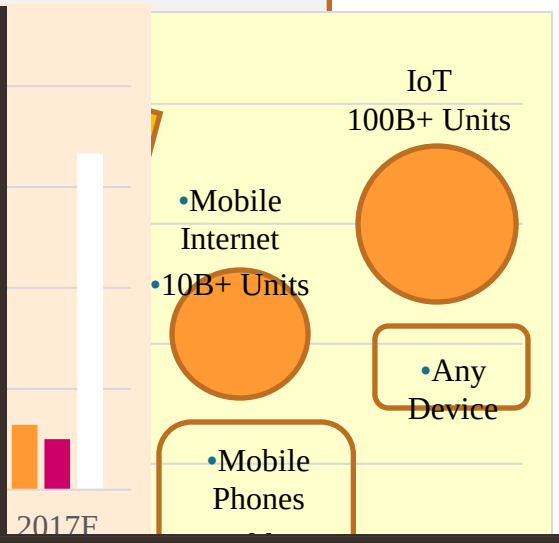
- Why?
  - Get things done (e.g. set up alarm/reminder, ta
  - Easy access to personal/external structured dat finding your docs/photos, finding restaurants)
  - Assist your daily schedule and routine (e.g. con meetings)
  - **Be more productive in managing your work ar**



- A real example scenario: “you booked a flight a cou
- Your Personal Assistant
  - **Scans (1)** your email and **extracts (2)** the flight it
  - **Computes (4)** your current location (GPS) and tions to the airport
  - **Tells (6)** you when to leave for the airport at th
  - **Checks (7)** the flight status and **updates (8)** you

- **Stitching together the steps can potentially mark a breakthrough in usefulness**
- Why are big tech companies investing into Personal Assistant Technology?

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



10M+ Units

- By 2017 mobile phone users are expected to reach 5.1B
- **No room for growth for connecting people to internet via smartphone (after 2017)**

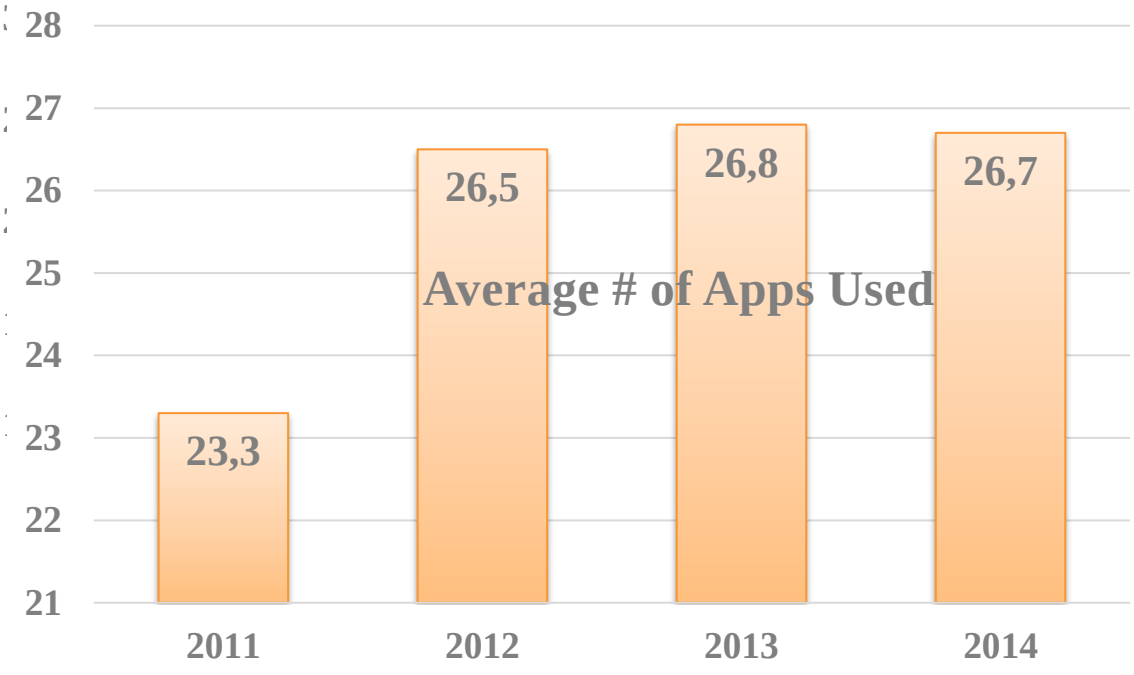
➔ **But companies have to grow – what is the next big opportunity?**

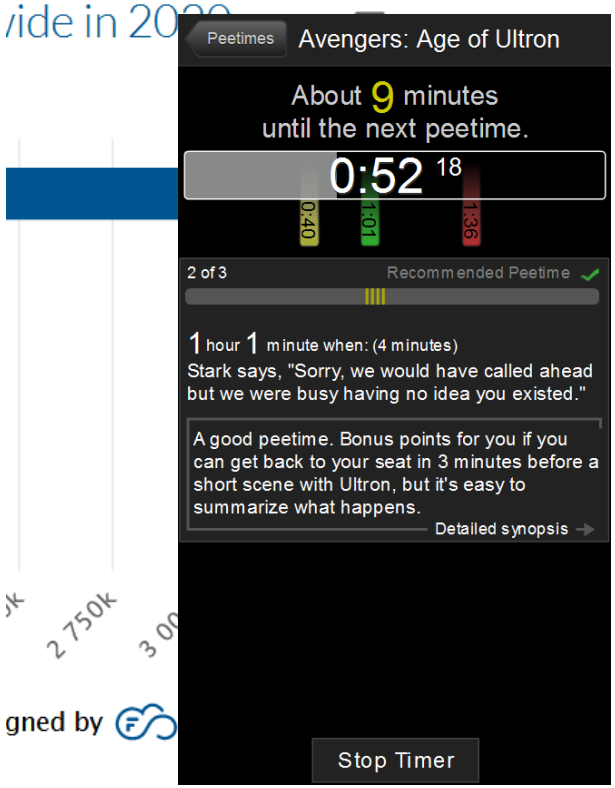
- Nev
- Mobile phones are dominating the device m
- About 2/3 of the world population (> 12 ye

Ultramobiles (Basic and Utility)** Mobile Phones

Forrás: Ruhi Sarikaya, Keynote paper, Interspeech 2015, Dresden, Sept. 2015

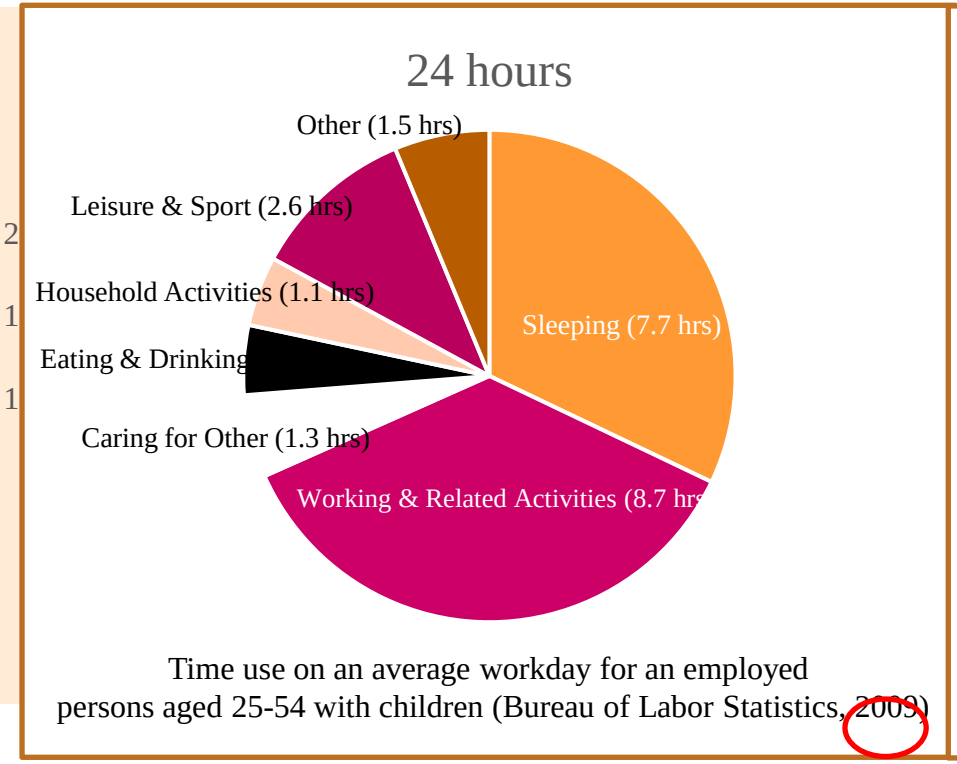
New Computing Cycle: Mobile Device & App Revolution (2)



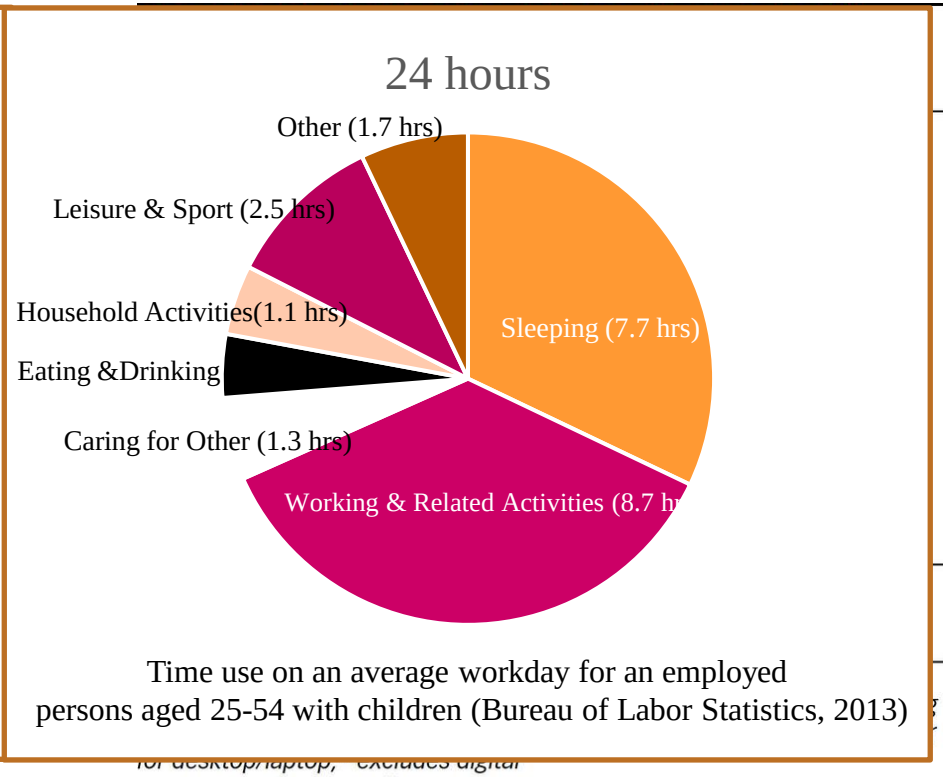


iOS App Store & Google Play Store have app. **1.5-1.6 million apps** (2015)
Avg # of InstalledApp=33 & UsedApp=26 per device per year is flat
Users typically use only 12 of them each month (80% are zombies)

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, 2013)

Source: eMarketer, April 2019

- TV=168min, Web (PC)=70min, MobilePhone=180min (2014)
- Smartphone(time) > TV(time) (Apps usage is 85% on Mobile Phones, 06/2015)
- Apps are after **you** and **your time (**# of users, how much time a user spends in each app)
- **Limited cognitive bandwidth to learn the apps (bottleneck #2)**
- Math does NOT add up ➔ where does the extra **~120mins** spent on smartphone come from?

Where do we use the smart phones?

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

Natural Way to Interact Personal Assistants: Voice Input

- Limited information flow into smartphones with typing/touch (bottleneck #4)
- Speech is expected to replace touch/typing as the primary input form
  - Pushed deeper into mobile platforms (e.g. Siri on iOS, Cortana on Windows 10, Google Now is integrated into Google Search App).
- 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
  - **Baidu**: The proportion of searches that are conducted by voice is around 10%
  - **Google**: # of spoken search query into their smartphones doubled last year
  - **Microsoft**: Voice queries to Cortana is a significant proportion of the overall volume

Why do big companies invest in Personal Assistant?

- Summary of Facts
  - Every person (>12 year old) (~5.1B people) will have a smartphone by 2017
  - You spend over 3hrs+/day (and increasing) on your smartphone
  - IoT is happening with no clear effective way to interact with them
- Challenges & Opportunities
  - [**Bottleneck #1**]: **App discovery**
  - [**Bottleneck #2**]: **Limited cognitive bandwidth to learn how each app works**
  - [**Bottleneck #3**]: **Information flow into small form factors**
  - [**Bottleneck #4**]: Your daily time budget is fixed: 1000 mins/day
- You need something that will:
  - **Provide a layer over the apps to complete the tasks**
  - **Increase your bandwidth: give you back your time**
    - Task delegation (e.g. track my flight, pay my bills, remember anything for me)
  - **On-demand assistance: "ask anything anytime"**
    - book a taxi, find transit schedule, find files, send email,...

  ➔ **Enhance productivity and lead to better management of your life!**
- Advances in ML, web services/apps and structured data makes it is the right time
- Next potential growth area ➔ possibly a $100B+ opportunity

Next Generation Personal Digital Assistant

1) Integrate potentially all the apps into Personal Assistants in a **scalable** way
2) "Service/app composition": draw information from different services/apps
3) No need to learn the exact query patterns to start a task
   - "get me a taxi" vs. "get me a taxi from Redmond town center to SeaTac airport at 2pm"

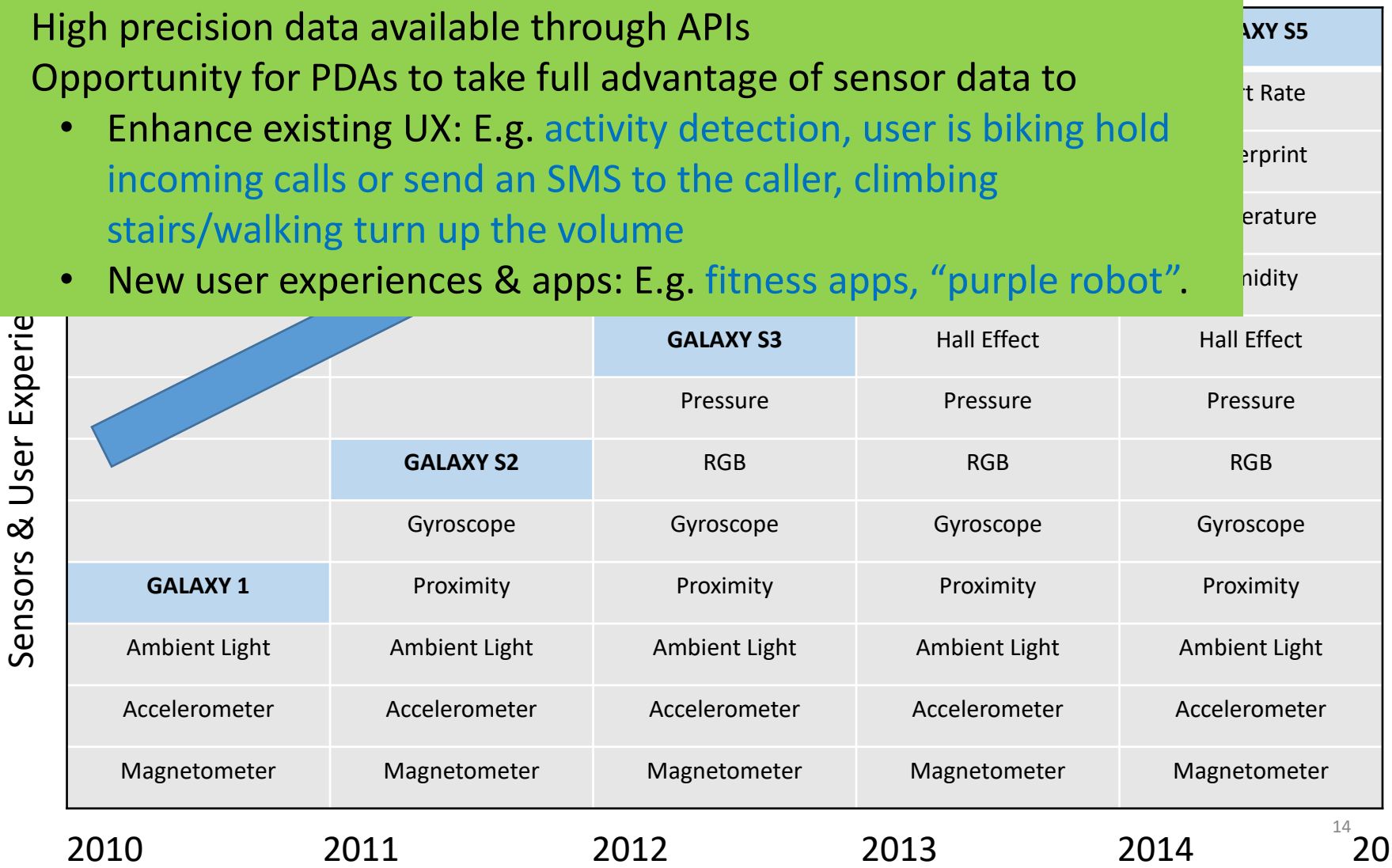
ZILLOW UBER TRANSIT

Personal Digital Assistant

- '**Personal**'
  - What do users do?
  - Interests?
  - What do they need?
  - When/where they need it?
- **Personalization Gaps**
  - Data, Computing, Interest, Action, Content
- **What has changed to make the digital assistants 'personal'?**
  - Device sensors
  - User data on device, services and apps

Device Side Sensors/Signals

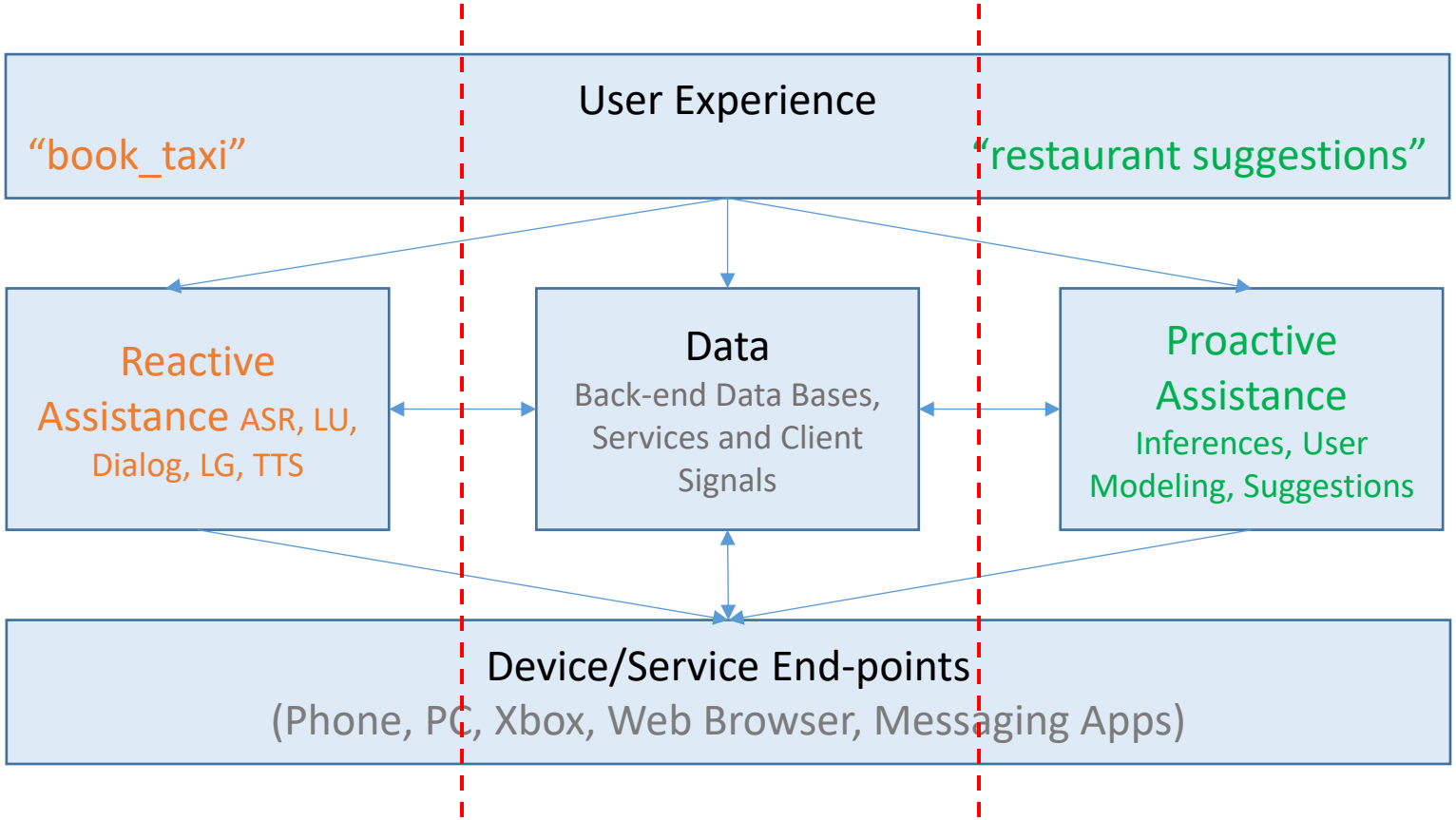
- Smart-phone sensors measure motion/orientation, and various user and environmental conditions
- High precision data available through APIs
- Opportunity for PDAs to take full advantage of sensor data to
  - Enhance existing UX: E.g. activity detection, user is biking hold incoming calls or send an SMS to the caller, climbing stairs/walking turn up the volume
  - New user experiences & apps: E.g. fitness apps, “purple robot”.



How to Make PDAs 'Personal'

- **You (the User): dimensions**
  - User **Profile** vs. **Digital Activity** vs. **Space** vs. **Time**
- **User Profile**
  - You, your work/home and (explicit) interests & preferences
  - Your people graph
  - Files, documents and photos stored (device and cloud)
- **Digital Activity**
  - Everything you do to with your device in the digital world. E.g.:
    - Calendar/emails
    - Apps used on the device and service
    - Social media activity: posts, likes etc.
    - Movies/games/music/… watched/played/listened/…
    - Web searches
    - Inferred user preferences/interests
- **Space**
  - Places you have been
- **Time**
  - When did the digital and physical (i.e. real world) activity take place
- Privacy & security

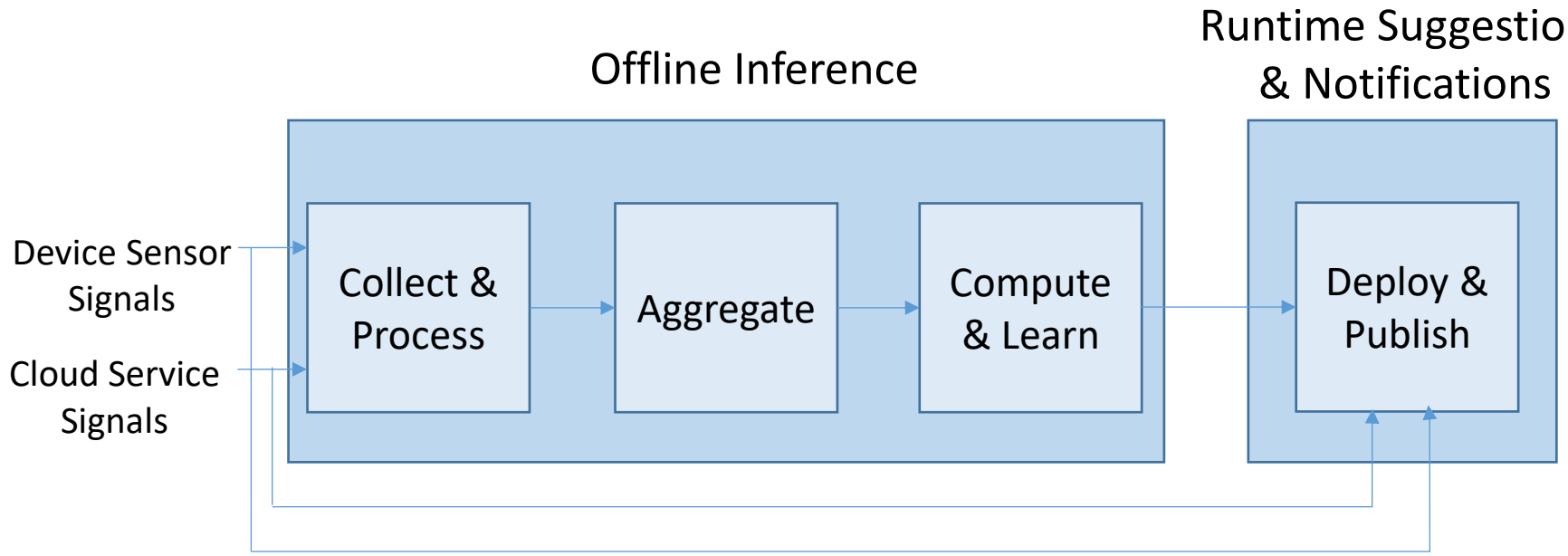
Top Level Personal Assistant Architecture



Proactive Assistance

- Proactive Assistance is based on the 'theory of proactivity'
  - How/when to provide assistance: costs vs. benefits of the potential actions
- Proactivity continuum ranges from zero to full automation
  - "Do it yourself" (no help)
  - *"Tells you what to pay attention to"*
  - *"Makes suggestions"*
  - "Makes decisions and takes actions"
- Proactive Agents' Attributes
  - **Valuable**: advances user's interests and tasks
  - **Does not interfere** with user's activities w/o user's approval
  - **Unimposing**
  - **Transparent** in what it knows about the user
  - **Anticipatory**: know about the future needs of the user and surface the opportunities when they show up
  - **Reason/learn** on continuous basis with signals it is receiving

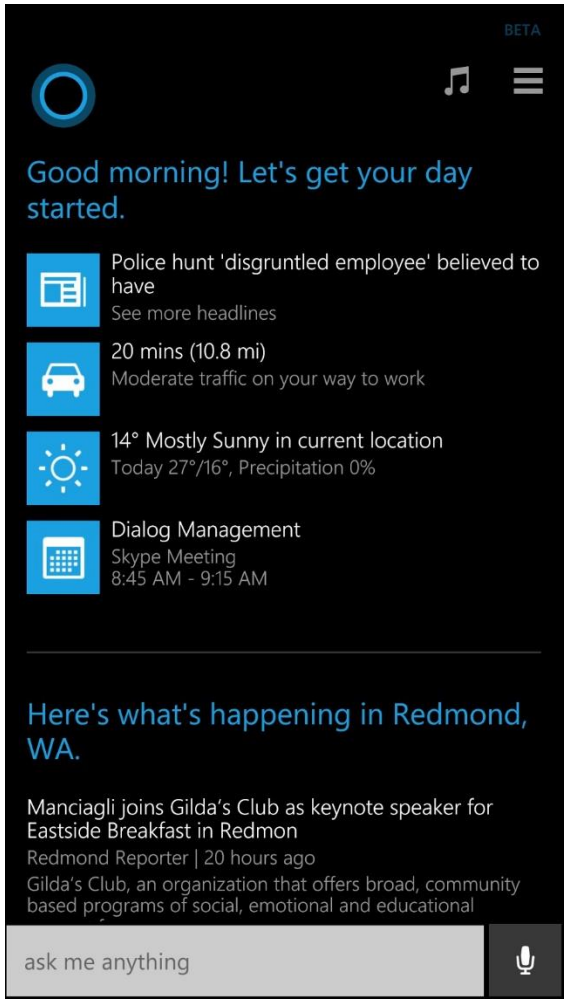
Proactive Inference



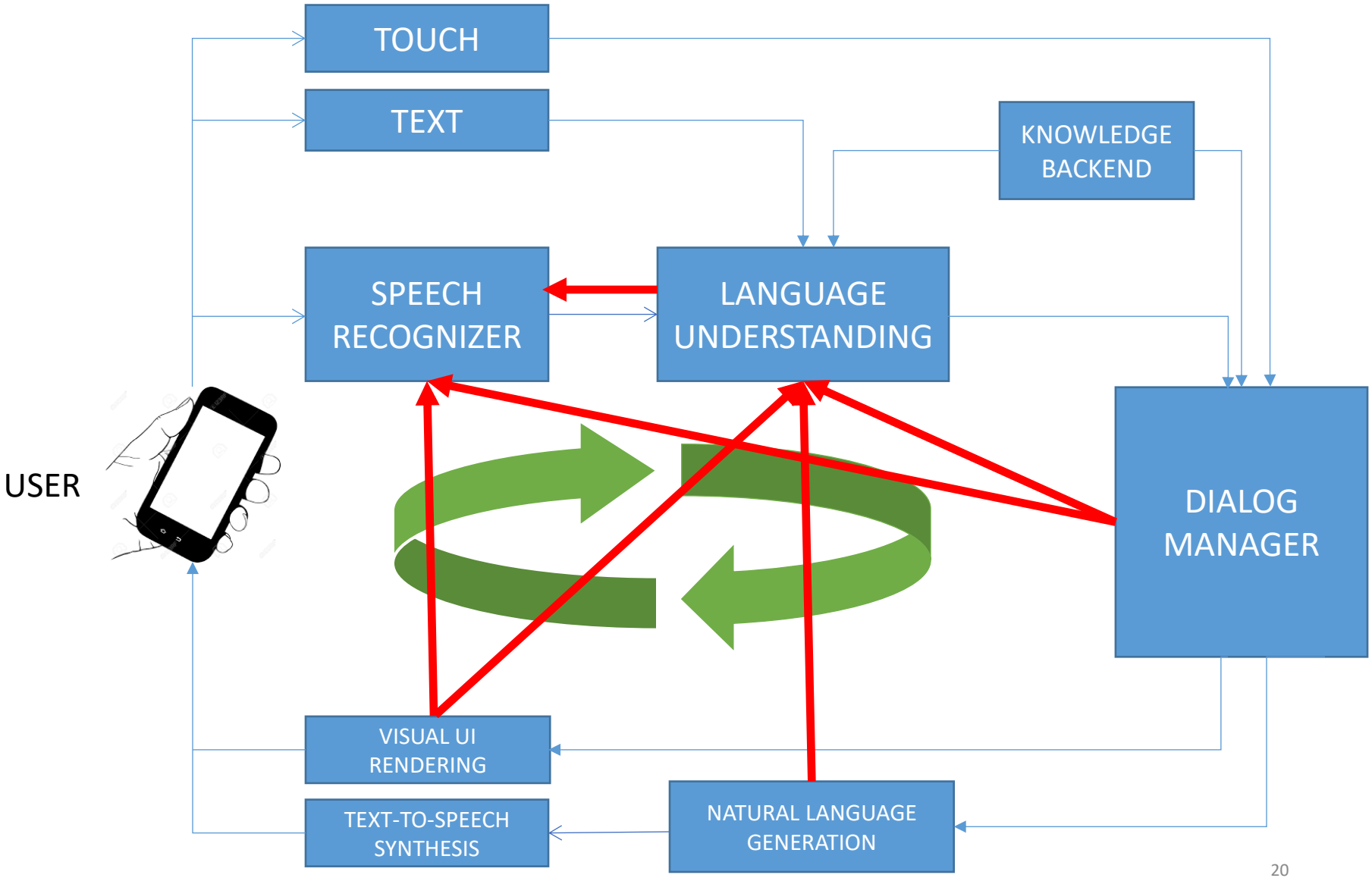
- Inferences are offline part
- Modeling user (e.g. habits, activities) across real and digital world activities. E.g.
  - commute hours
  - team you support
  - movies you like
- Signals are extracted from both from real and digital world
  - Device Sensors, Services/Apps and/or explicitly provided by the user
- The learned models
  - Simple as histograms
  - Advanced ML models (incorporate context, meta signals and dependencies etc.)

Proactive Suggestions/Notifications

- Suggestions/Notifications Service is the runtime part
- Provides the most likely content that user would like given the user's context. E.g.
  - commute hours → show traffic card (notification)
  - team you support → show Arsenal's match score (notification)
  - favorite actor → show the new movies by him (suggestion)
  - cuisine habits → show a new restaurant in our neighborhood (suggestion)
- Provides correct sequencing of different yet correlated (through time or location) set of events
  - Night out: dinner → parking → movie (suggestions)
- Receive explicit feedback from the user to validate (or invalidate) the learned reasoning



(REACTIVE) DIALOG SYSTEM ARCHITECTURE



Basic Language Understanding Models

Goal: Extract precise meaning of the spoken/typed query

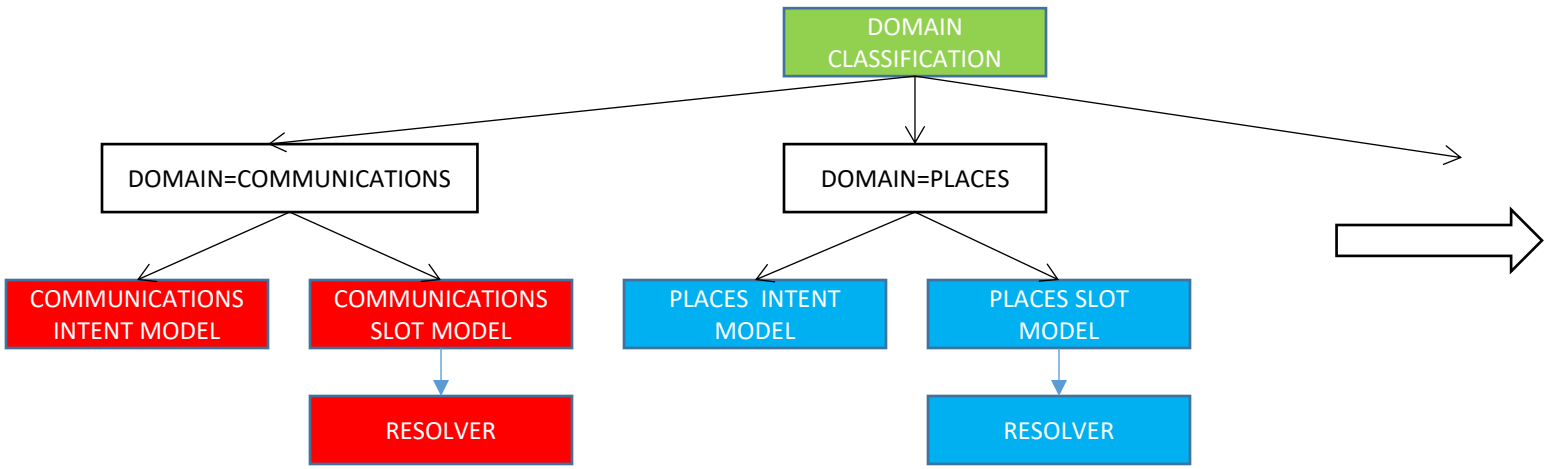
Basic LU Modeling: slot filling, intent detection, domain detection

- Domain Models: Binary classification problem: Does input belong to domain? True/False
- Intent Models: Multi-class classification problem: Determine which one of the N intents the input belongs to within a given domain
- Slot Models: Sequence classification problem: Identify entities key words in the query
- "find closest highly rated indian restaurants"
  - Domain: PLACES
  - Intent: find_place
  - Slots={NEARBY="closest", RATING="highly", COUSINE="indian", PLACE_TYPE="restaurants"} ➔ queries to the backend

Slot/Entity Canonicalization/resolution

- Mapping entities to a database ID
- Mapping "highly" to "good" (canonical form)
- Time: Date/Time resolution, Location: lat/long mapping,…

State-of-the-art Statistical LU Modeling Approach



Contextual Language Understanding Modeling

- Machine Learned Models

Improvements (Domain)

what's my schedule tomorrow **calendar**
what time is the first one **places → calendar**

what's the weather like today **weather**
what about tomorrow **web → weather**
and for the weekend **calendar → weather**

tacos places in seattle **places**
select rancho bravo tacos **web → places**

who went to the final four last year **web**
when is Michigan's next basketball game **calendar → web**

Improvements (Intent)

directions to home depot **get_route**
The one in Bellevue **find_place → get_route**

opening hours for home depot **get_hours**
The one in Bellevue **find_place → get_hours**

User Goal Tracking (a.k.a. Slot Carry Over)

- Machine Learned Model
- Slot Carry Over (SCO) is a first stage user goal (state) tracking capability
  - Crucial to enable multi-turn conversational behaviour
  - SCO decides which slots from the previous turns are still relevant in the current turn for query building:

    Turn 0: "**find french restaurants in seattle**"

    State 0: {cuisine="french", place type="restaurants", absolute_location="seattle"}

    Turn 1: "**how about chinese**"

    State 1: {cuisine="chinese", place_type="restaurants", absolute_location="seattle"}

  - Implicit start-over conditioned on domain switch
  - SCO can operate both within and across domains
- SCO reduced to a supervised learning binary classification task
  - For each slot from previous turns, decide to carry-over or drop
  - Slots detected by tagger in current turn unconditionally kept

Flexible Selection: See-it-Say-it (SISI)

- Machine Learned Model

[Turn-1]: *"Hamburger places near me"*

(1) Five Guys Burgers and Fries

(2) Kidd Valley Burgers and Shakes

(3) Wibbley's Gourmet Hamburgers

(I) Explicit Referential (Turn 2)

"show five guys menu"

"directions to the five guys and fries"

(II) Implicit Referential

"the one on 24th street"

"the one in Redmond"

(III) Explicit and Implicit (Mixed) Referential

"five guys in redmond"

"the burger shop in redmond"

(IV) Explicit Locational

"look up hours for the top one"

"call the first (burger) place"

(V) Implicit Locational

"what time does the first one open"

"call the 1st one"

Hypothesis Ranking/Selection (HRS)

HRS is the brain of the system – final decision maker

- Ranker: Machine Learned GBDT (Gradient-Boosted Decision Trees) model
- Augments LU analysis with additional features & signals from knowledge, session and external sources
- Operates on the dialog hypothesis representing the state
- Performs ranking of dialog hypothesis coming from different domain WFs
- Provides a mechanism to deal with ambiguous/erroneous speech reco results, LU results (domain/intent/slots)
- Allows dialog state tracking

Personal Assistant Design Gaps

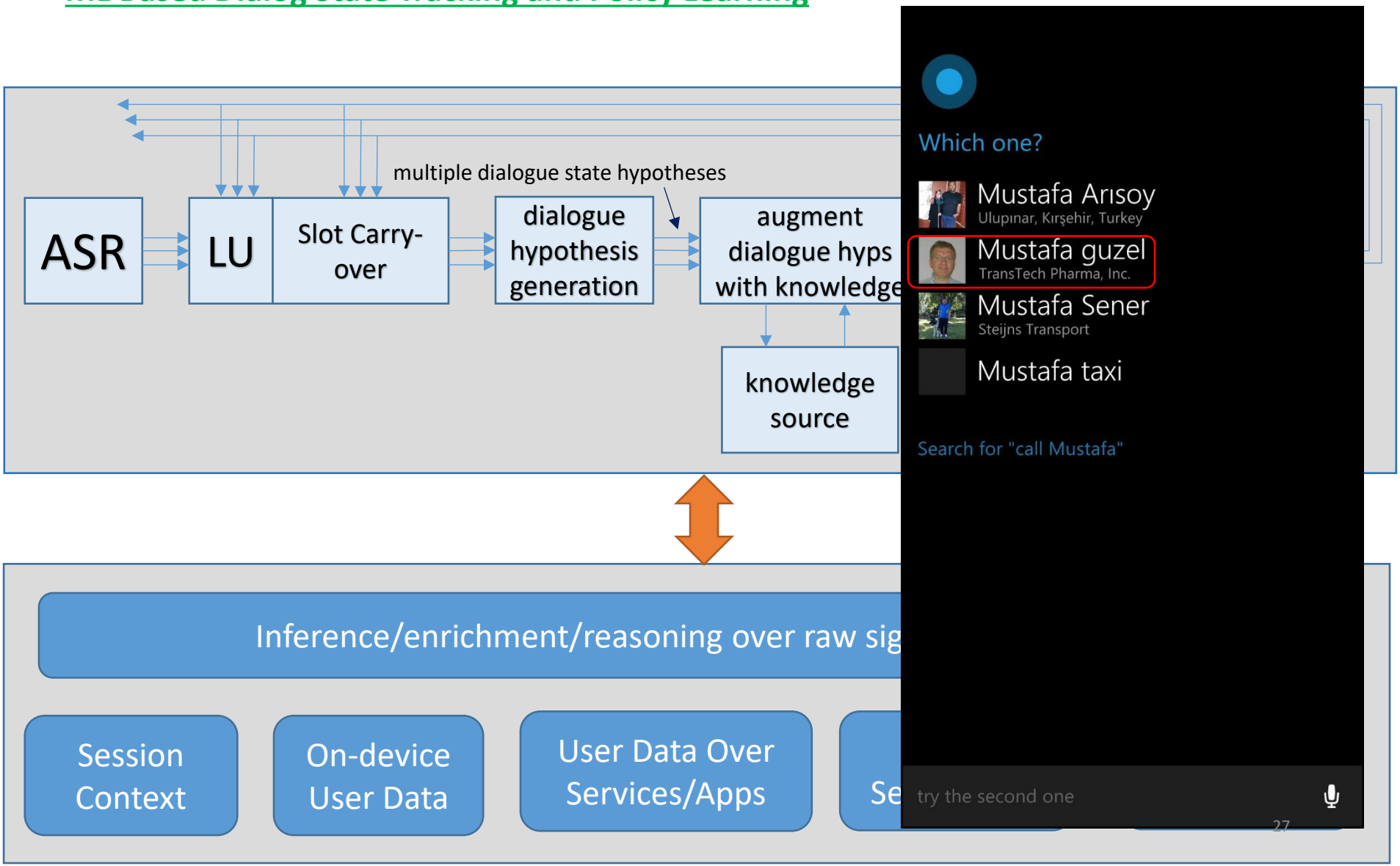
- (Current) Reactive Assistance is not ‘intelligent’:

https://www.youtube.com/watch?v=Ue0fZfwHfzo

- The Personal Assistants should be the ‘right person’ to ask questions to get the results you are seeking – currently they are not!
- Currently reactive and proactive stacks are built in silos and there is not bidirectional interaction between them.
  - Reactive should use all the proactive (and additional inference) signals
- 3rd party app integration into personal assistants has to be front and center
  - Limiting that and designing for a set of in-house apps will be a mistake
  - Currently this does not have a clear scalable path
  - Critical for domain/scenario coverage

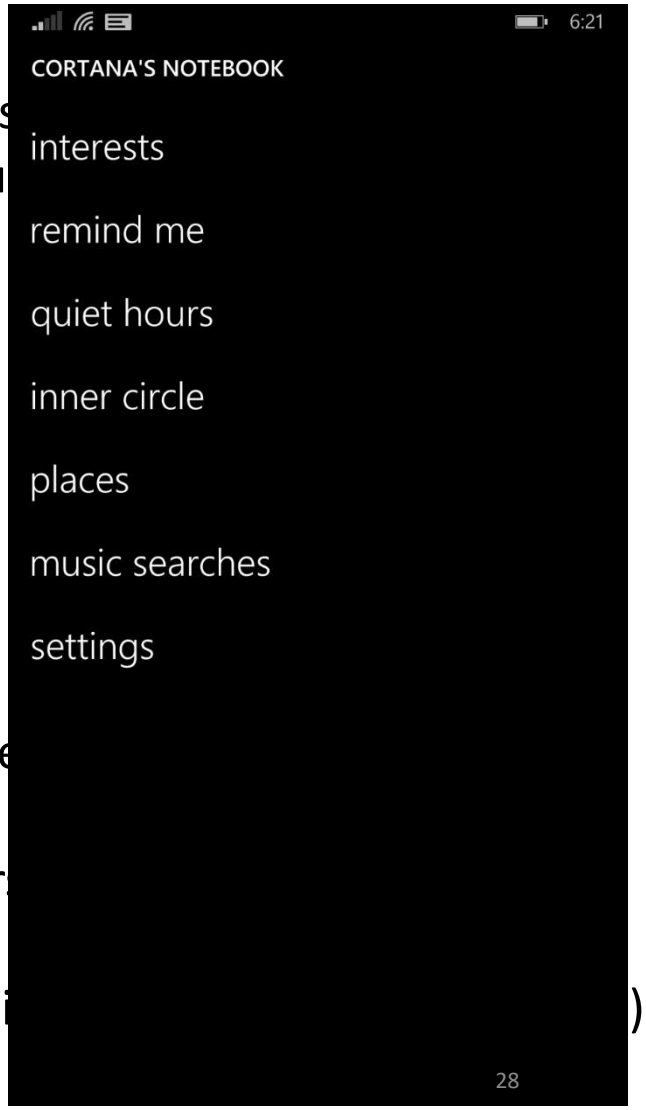
NEXT GENERATION DIALOG SYSTEM ARCHITECTURE

- **ML Based Dialog State Tracking and Policy Learning**



User Facing Challenges for Personal Digital Assistants

- **Operation errors:**
  - Discrepancy between user's mental model of the s
  - Lack of proper UI: limited information about the u
    - Intuitive sequence of operations to complete
  - Limited adaptation to different user's profiles and
- **Lack of competence:**
  - PDAs are not at the level to reliably decide:
    - When to help the user
    - What to help the user with
    - How to help
- **Trust**
  - Whether the user feels comfortable delegating the
  - Privacy and security of the user's data and profile
  - How much PDA can/should know about their user mechanism?
  - Autonomous vs. progressive intelligence (compani)

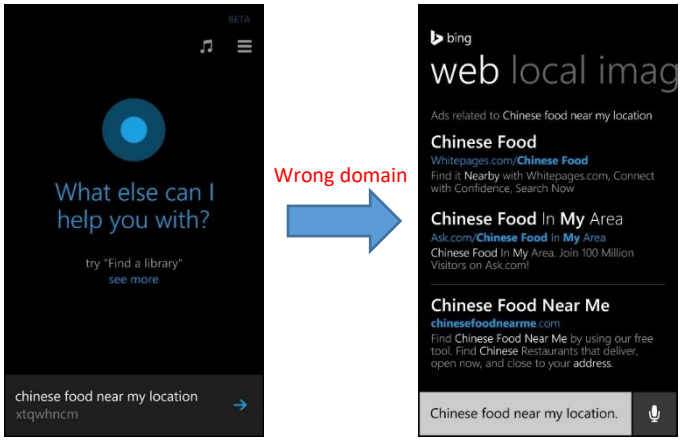


Domain Classification Errors

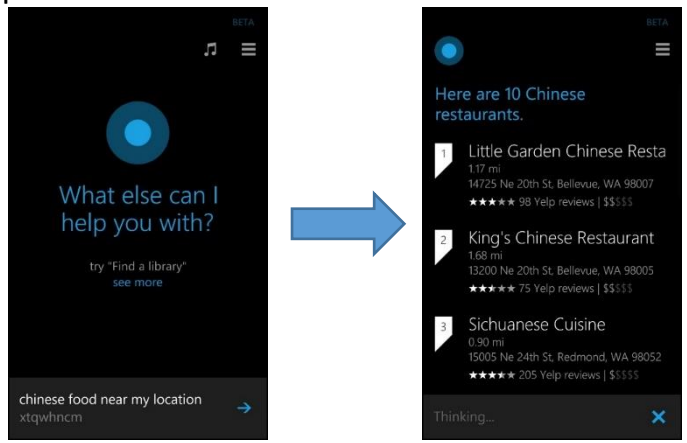
Example

- User: "Chinese food near my location"
- We expect answers (list of restaurants) but instead web results are shown
- Cause: LU domain detection error (places → web)

DSAT Result



Expected Result

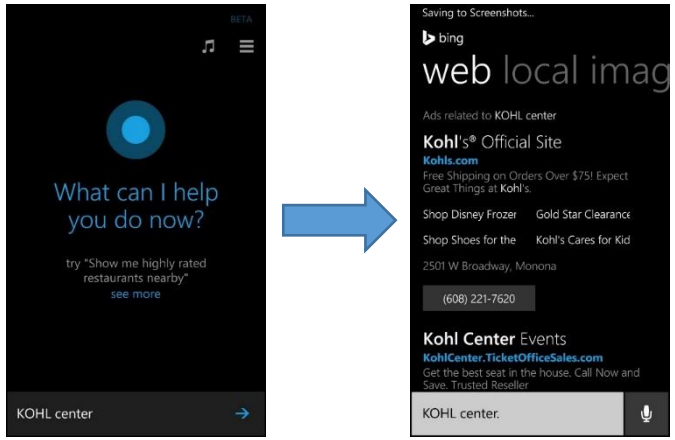


Example

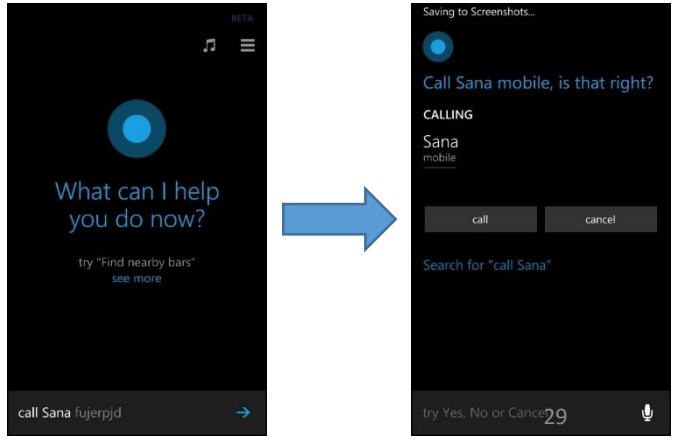
- User: "Call Sana" [ASR: Kohl center]
- Communications domain → Web
- Immediately the user tried the task again to successfully complete it.

Cause: ASR Error

DSAT Result



Expected Result

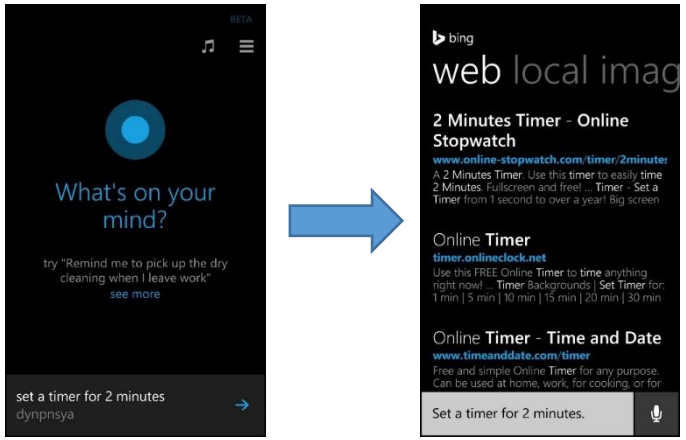


Unsupported Scenario

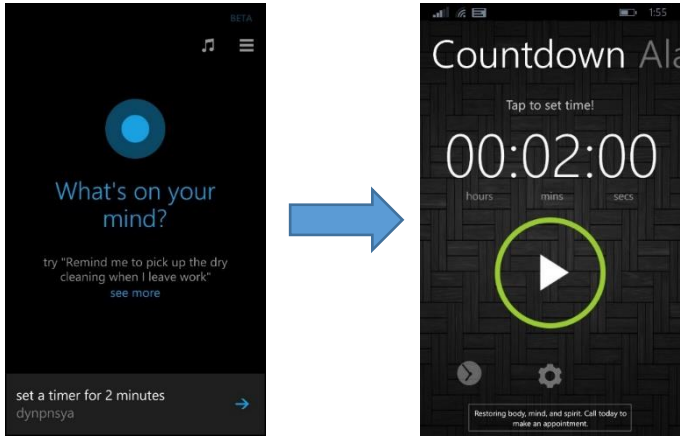
Example

- TIMER
  - ➢"set a timer for 2 minutes" results in web search
    - ➢Cause: Timer functionality is not supported
- EMAIL
  - ➢ "send an email to John Hansen"
  - ➢ Cause: Email app is not supported
- Unsupported action can be anything that the user expects the system to handle:
  - device action
  - fact answers
  - proactive card answer etc.

DSAT Result



Expected Result



Personal Assistant Product Metrics

- How often users use PDA?
  - Daily Active Users (DAU)
  - Monthly Active Users (MAU)
  - DAU/MAU → overall engagement
- # queries handled (reactive)
- # page views took place (proactive)
- E2E Accuracy
  - Query/SystemResult (i.e. rendered UI) 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
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

Technology Challenges

Experience scaling

3rd party integration/tools/infrastructure

Speech Recognition Challenges

- Background Noise, Speaker Accent, Bluetooth
- Side Speech, Pocket Dial, Unintentional Wake up Voice
- Open domain unlimited vocabulary (e.g. user's contact list)

Language Understanding Challenges

- Domain Scaling: rapid model development
- Open Domain LU, Domain scaling
- Scaling quality
- Difficulty of building reusable models (e.g. no shared schema)

Dialog Management

- Heterogeneous back-ends, interfaces
- Lack of reusability

NLG

- Localization scaling

Proactive

- Experience Scaling

Future

- PDA war may set the balance of power in the next phase of the internet.
- Too early to call it “an inflection point” for PDA technology.
- Talking to gadgets becomes second nature soon.
- Will 3rd party apps be able to integrate with the PDAs as deeply as they would like?
- How will the PDAs decide what information/apps to put in front of a user?
  - If I ask PDA to find me a taxi, it may not use Uber ➔ Consequences could be profound for the companies that rely on customers accessing their services over smartphones
- The walls between apps will start to break down
- ML and deep learning to truly understand the user and world is to be seen.
- Once computers truly understand text, speech, images and sounds, they will become our indispensable assistants.
  - This will revolutionize the way we interact with computers, helping us live more conveniently in our day-to-day lives and perform more effectively at work.