“Automated” Speech Recognition and Dialog Systems

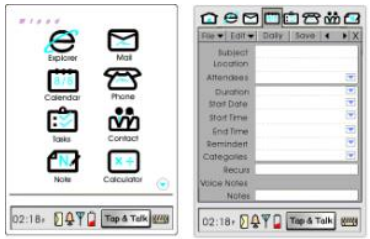
CS136 Speech Recognition
January 14, 2020
Professor Meteer

Speech → Dialog

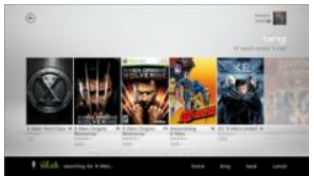
- Speech: Audio → words
- Dialog systems
  - What did the speaker mean?
  - What should the dialog system do?
  - How should it reply?
  - How do you build a task oriented dialog system, where the speaker is trying to get something done?
  - How do you build a “conversational” dialog system, where the goal is to keep the conversation going?
  - We know context is key, but what does that mean?

+ Brief History of Dialog Systems

Multi-modal systems
e.g., Microsoft MiPad, Pocket PC



TV Voice Search
e.g., Bing on Xbox



Virtual Personal Assistants

Task-specific argument extraction
(e.g., Nuance, SpeechWorks)
User: "I want to fly from Boston to New York next week."

Early 20

Siri

2017

IBM WATSON

Early 1990s

Intent Determination
(Nuance's Emily™, AT&T HMIHY)
User: "Uh…we want to move…we want to change our phone line from this house to another house"

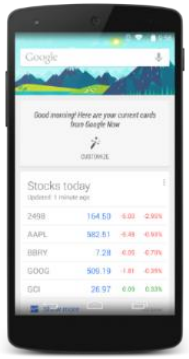
DARPA
CALO Project

Keyword Spotting
(e.g., AT&T)
System: *"Please say collect, calling card, person, third number, or operator"*

Thanks to VIvian Chen, Asli Celikyilmaz, & Dilek Hakkani-Tur: http://deepdialogue.miulab.tw

So many things to talk to!

Apple Siri (2011)

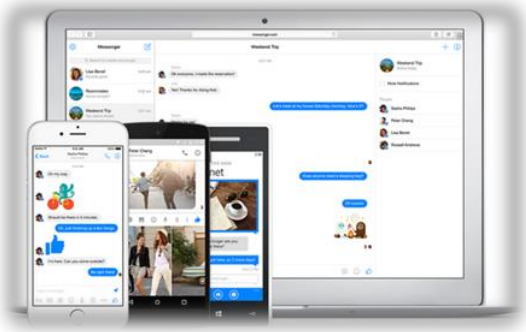


Google Now (2012)

Google Assistant (2016)

Microsoft Cortana (2014)

Amazon Alexa/Echo (2014)



Facebook M & Bot (2015)

Google Home (2016)

Apple HomePod (2017)

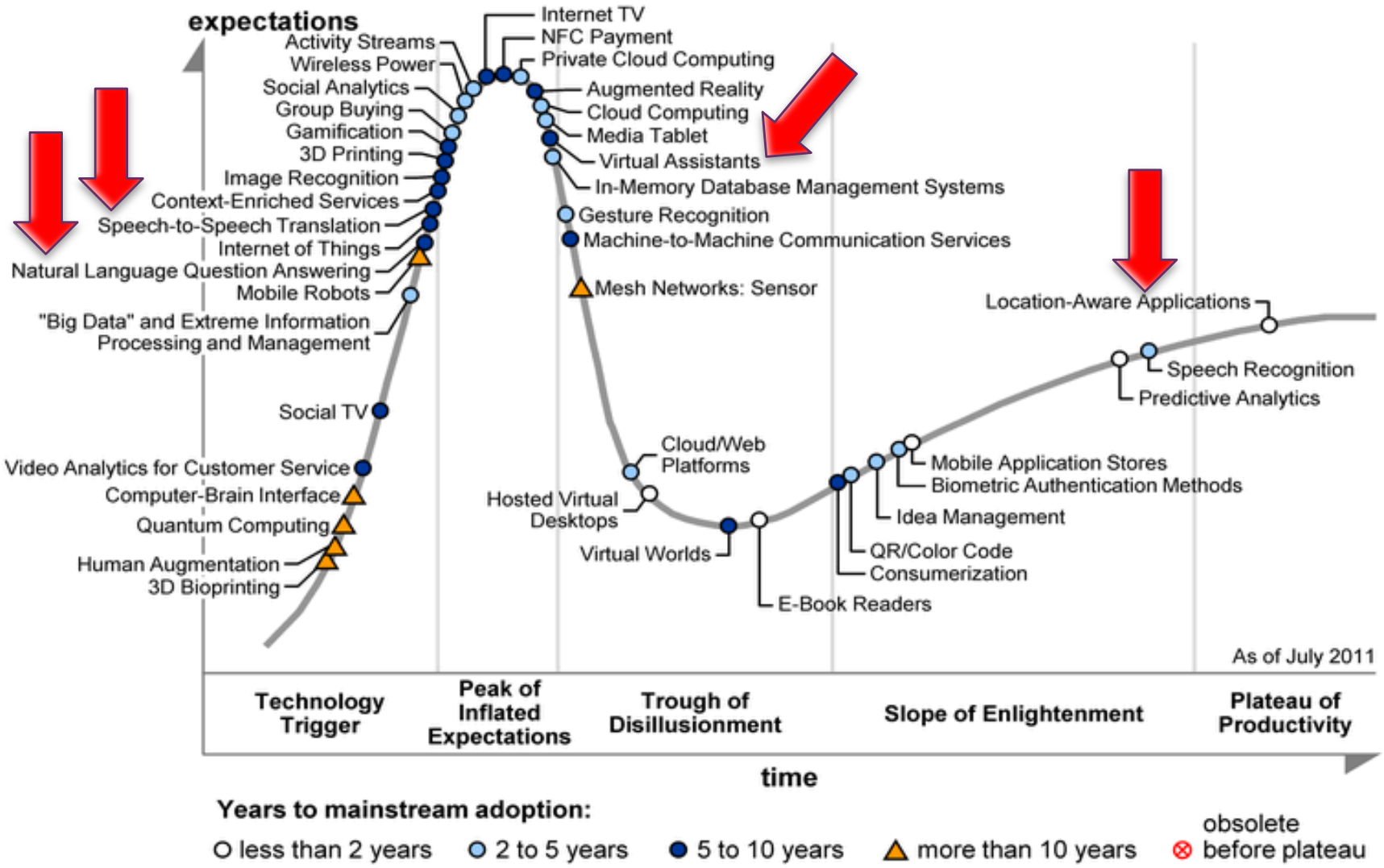
What devices do we have and what do we want?

- What do you talk to?
  - Devices
  - Tasks
  - Information access
- What do you want to talk to?
  - Devices
  - Tasks
  - Information access
- How would that change if …
  - You were texting, not talking
  - If talked to you first
  - …

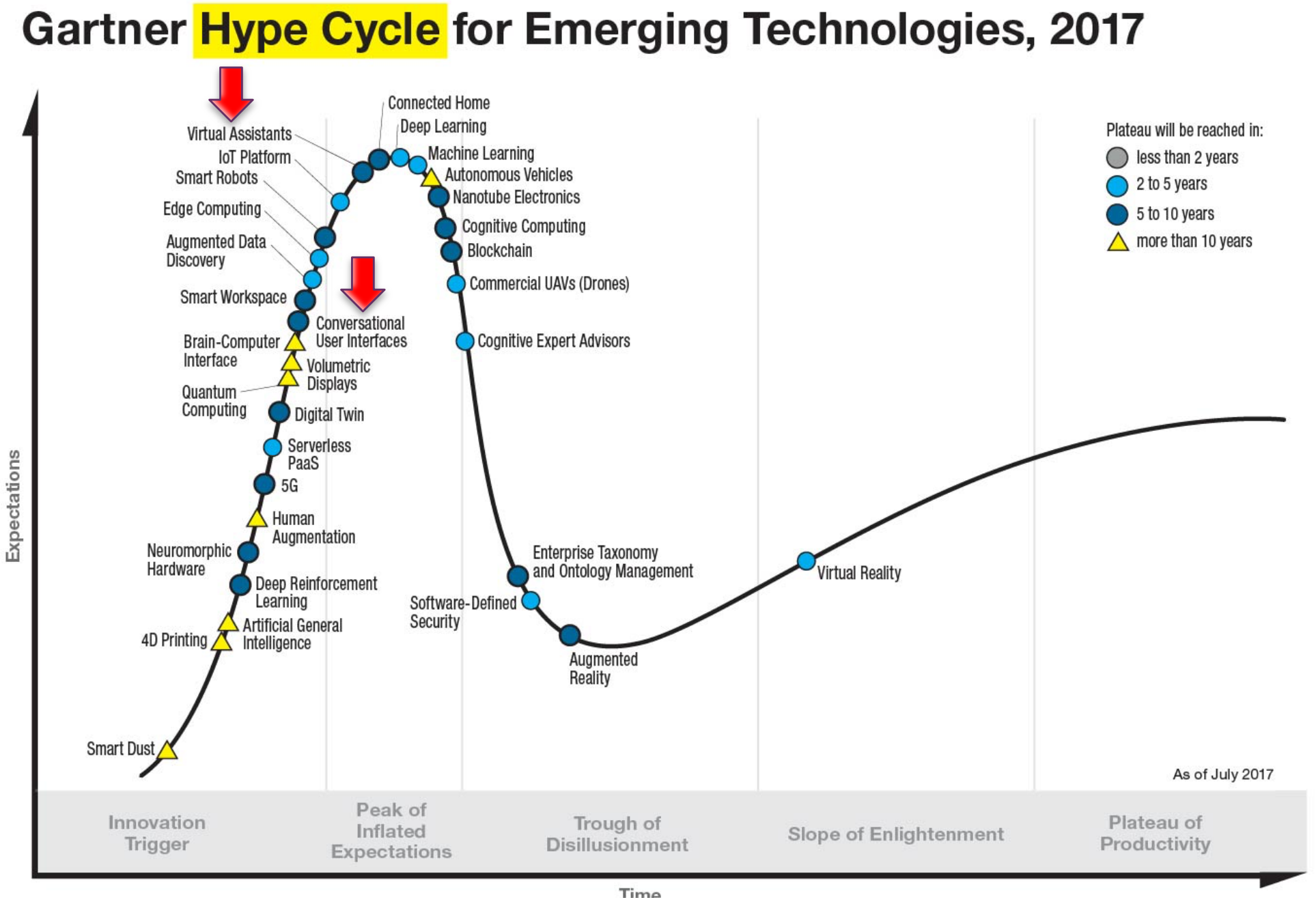
Beware the Hype Cycle

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

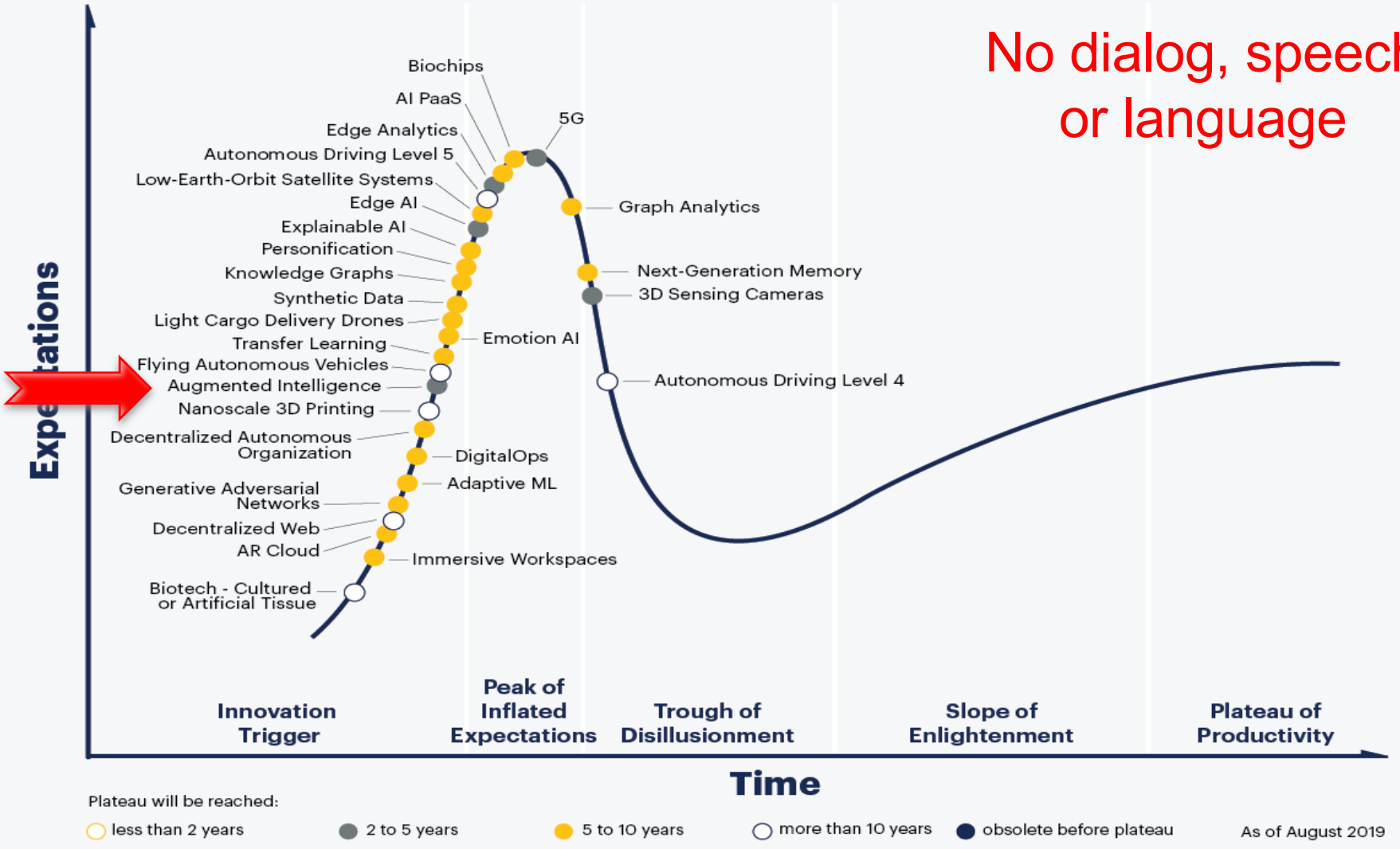
Gartner's 2011 Hype Cycle



+ And now?


Gartner Hype Cycle for Emerging Technologies, 2017
Virtual Assistants
IoT Platform
Smart Robots
Edge Computing
Augmented Data Discovery
Smart Workspace
Brain-Computer Interface
Quantum Computing
Neuromorphic Hardware
4D Printing
Smart Dust
Conversational User Interfaces
Volumetric Displays
Digital Twin
Serverless PaaS
5G
Human Augmentation
Deep Reinforcement Learning
Artificial General Intelligence
Connected Home
Deep Learning
Machine Learning
Autonomous Vehicles
Nanotube Electronics
Cognitive Computing
Blockchain
Commercial UAVs (Drones)
Cognitive Expert Advisors
Enterprise Taxonomy and Ontology Management
Software-Defined Security
Augmented Reality
Virtual Reality
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
Expectations
As of July 2017
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time

Gartner Hype Cycle for Emerging Technologies, 2019



No dialog, speech or language

gartner.com/SmarterWithGartner

Source: Gartner

Gartner

Learning goals

- Premise
  - Everything you're learning today is going to change
  - Progress will be made and papers written about it at a much faster rate than you can keep up with
  - Areas outside your expertise may become critically important to what you're doing
- Active learning
  - Programming
    - Using systems and tools without being "taught" them
    - Tackling problems that don't have answers
  - Content material
    - Reading papers you might not have adequate background for
      - How to figure out which papers to read
      - How to pull out what you need from a paper and leave the rest behind
- Team learning
  - Your colleagues will be your classroom
    - How to teach them what you know
    - How to learn from them
  - The strength of a group is dependent on communication

+ Topic Areas → Tools

- Speech recognition:
  - Kaldi, OpenFST, SCLite Evaluation software
- Natural language understanding & frame-based dialog systems:
  - Dialogflow (Google), LUIS & Azurebot (Microsoft), Watson Conversation (IBM), and of course Alexa.
- Statistical Dialog Modeling:
  - Various DNN toolkits and applications to the Dialog State Tracking Challenges and AlexaPrize.

Course Requirements

- Course work
  - Homework
    - Using tools to solve specific tasks
  - Readings, quizzes, group work
  - Presentations
  - Project
    - Groups of 3 (+/- 1)
    - Topic ideas will be on the “Projects” page
- Systems
  - Website for schedule and assignments
  - Git for programming
  - GoogleDocs for group work on papers and presentations
  - Slack for communication
  - Latte for quizzes and some other things

Assignment 1: "State of the Art in Continuous Speech Recognition" by John Makhoul and Rich Schwartz

- **Read and reflect on what you know and don't know.**
- Highlight areas you think are important and that fall into the following 3 categories. I suggest using highlighters, but use whatever works for you:
  - (green) This is something I'm familiar with, enough that I could explain it to someone else.
  - (yellow) I recognize the phrases and know I've studied it in the past, but couldn't actually explain it.
  - (pink) This is like a different language. I can't readily connect this with anything I know.
- You don't need to highlight every paragraph, just those that strike you as important.
- Save this out and submit it before class Tuesday, Aug. 30th. We will be discussing it in class.

NLP Commercialisation in the last 25 Years

Robert Dale, in Journal of Natural Language Engineering.

- Where were we in1995?
  - Windows 95 and Netscape became available
  - Java 1.0 appeared, and JavaScript was developed
  - DVDs were introduced
  - Sony released the PlayStation in North America
  - NSFNet was decommissioned, removing the last restrictions on the commercialisation of the Internet
  - Yahoo.com, eBay.com and Amazon.com all launched
  - BM unveiled Deep Blue, the computing system that went on to beat world chess champion Garry Kasparov
- On the downside
  - All we had were flip phones

NLP Commercialisation in 1995

Ken Church & Lisa Rau

	Well-understood	State-of-the-art	Forward-looking
NLP	NL interfaces	Entity extraction and generation from databases	Event extraction
Word processing	Simple string matching	Spell checkers	Grammar checkers
Information retrieval	Keyword search	NL search	Conceptual search
Machine translation	Glossary look-up	Translation memories and direct transfer	Conceptual search

Only speech recognition used statistical methods in commercial products

+ Six key areas in NLP

Machine translation	The meaning-preserving translation of linguistic content in a source natural language into a target natural language.
Speech technologies	The conversion of a wave form corresponding to a linguistic utterance into a textual rendering of that utterance (speech recognition); or, conversely, the conversion of a textual representation of a linguistic utterance into a corresponding audio form (speech synthesis or TTS).
Dialog interfaces	Interactive interfaces that permit a user to ask a natural language question (either via speech or text) and receive an answer, possibly in the context of a longer multi-turn dialog.
Text analytics	The detection of useful content or information in textual sources, typically either via classification or via extraction.
Natural language generation	The generation of linguistic output from some underlying non- linguistic representation of information.
Writing assistance	Technologies that embody some knowledge of language for the purposes of improving the quality of human-authored linguistic content.

Speech Recognition

- Getting to 1995
  - 1970's: Speech adopts HMMs and statistical modeling
  - By the end of the 70's: Vocabularies of 1000 words, but isolated word recognition
  - Speech goes commercial
    - Early 1990's
      - Dragon Dictate $9000, but only isolated word recognition
      - Telephone IVR
        - Nuance out of ISI
        - Speechworks out of MIT
        - Hark out of BBN
      - All constrained grammars, limited vocabulary
- Today
  - Unlimited vocabulary (rare words will still be wrong)
  - Accuracy for native speakers intending to be is amazing!

Dialog Interfaces

- Range of technologies
  - Early 1960s' and 1970s' experiments in text-based conversational interaction, exemplified by Joseph Weizenbaum's Eliza
  - Text-based natural language database interfaces of the kind that were a focus of attention in the 1970s and 1980s, allowing users to type questions like 'How many sales people are earning more than me?' rather than having to formulate an SQL queries
  - Voice dialog systems of the 1990s, where finite-state dialog modelling was combined with grammar-based speech recognition in telephony-based applications.

+ Dialog

- Getting to 1995
  - 1992: AT&T launched “How may I help you?”, applying statistical topic modeling nto call routing
  - DARPA “ATIS” (Air Travel Information Services) program is an early “shared task” focused on dialog modeling.
  - Commercial speech IVR systems, “Directed dialog”
- Where we are
  - Today’s virtual assistants, such as Siri, Alexa and Google Assistant.
  - Today’s text-based chatbots, found on many websites and messaging apps, displaying a wide range of capabilities (or, often, lack of capability)
  - Analysis on one chatbot indicates that they move quickly from open dialog to very constrained answer “selection”

+ Text Generation

- 1990: I did my PhD thesis in Text Generation.
  - At BBN, text generation work was only in research
- 1995, there was virtually no commercial NLG activity
  - CoGenTex (http://cogentex.com), had been founded in 1990, and survived mostly on government contracts
- Recent commercial revival after long hiatus
  - Narrative Science (https://narrativescience.com/)
  - Automated Insights (https:// automatedinsights.com)
- Recent research relies on automated scoring
  - My opinion on this can't be printed here
- Robert Dale's commentary on the state of the art:
  - There's little evidence that commercial NLG offerings are leveraging the richer linguistic concepts, like aggregation and referring expression generation, that have been developed in the research community;
  - My sense is that the use cases being explored so far don't warrant the use of these ideas.

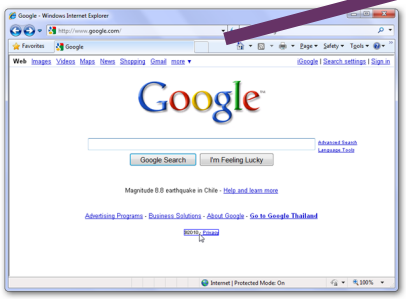
+ Why use language?

- More natural and convenient
- More efficient
  - Typing: 40 words/min
  - Speaking: 120 words/min
  - Reading: 80 words/min
  - Image (Is a picture really worth 1000 words?)
  - Brain: ?? Thoughts/min
- Devices are getting smaller
  - What does this do to typing speed? Reading ability?
- 3.65B Active unique mobile users

GUI vs. VUI

- Web-based application: What you see is your world
- Impact on programming: All the “context” is on the page

Type a URL



Type a search query.
Select a link from results

Select a link from “history”

Select a link on a page

Select a link on a page

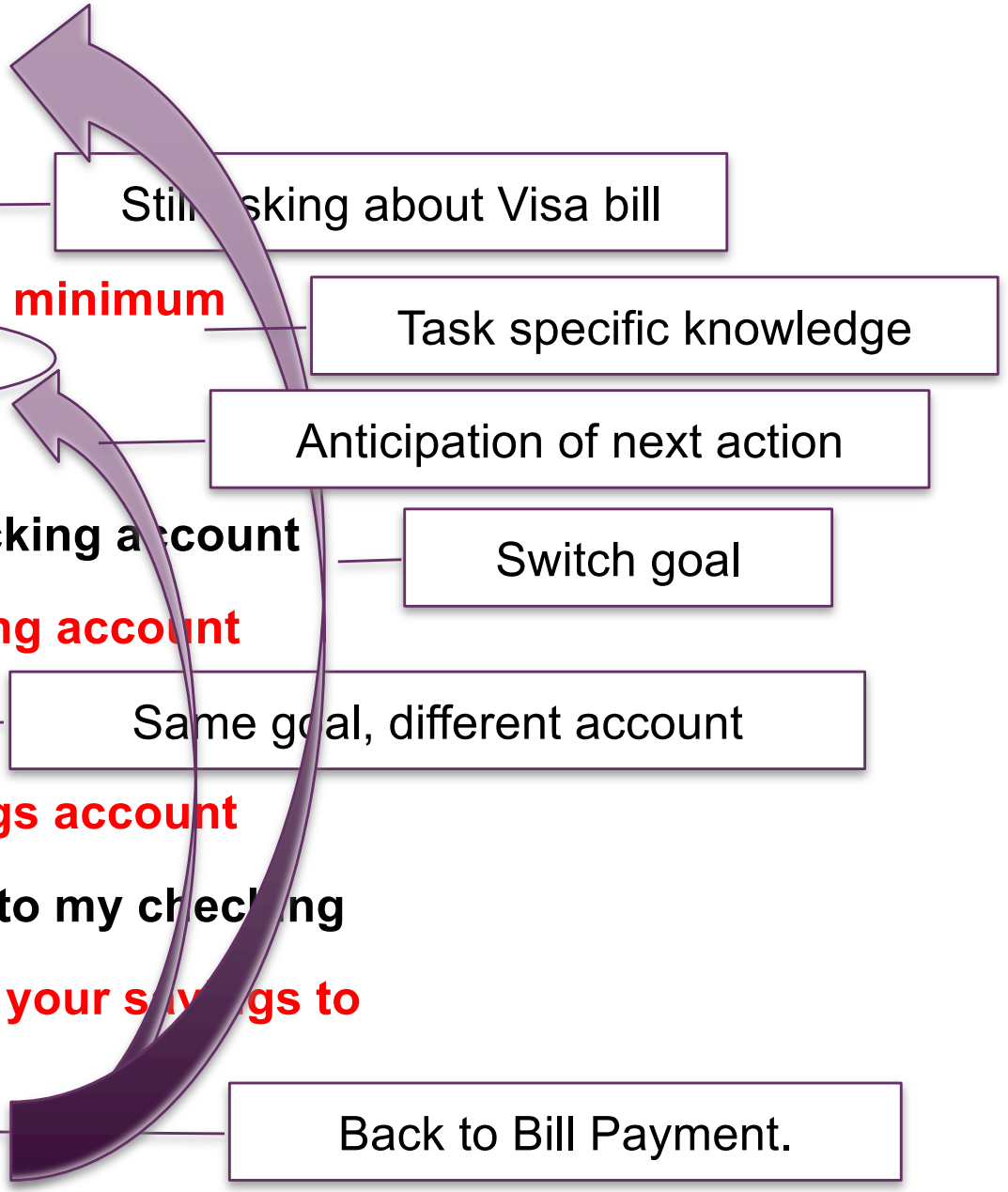
Welcome Marie!

You have 3 items in your shopping cart.

Log In provides profile and some context such as the shopping cart.

+ Example

- **When is my Visa bill due**
- **Monday, April 15th**
- **How much do I owe?**
- **Your balance is $583.24 and the minimum payment is twenty five dollars.**
- **Would you like to pay your bill?**
- **How much do I have in my checking account**
- **You have 164.25 in your checking account**
- **How about my savings**
- **You have 2019.97 in your savings account**
- **Transfer $500 from my savings to my checking**
- **$500 has been transferred from your savings to your checking account**
- **OK. Now pay the minimum**



GUI VS. CUI (conversational UI)

	Website/APP's GUI	Msg's CUI
Situation	Navigation, no specific goal	Searching, with specific goal
Information Quantity	More	Less
Information Precision	Low	High
Display	Structured	Non-structured
Interface	Graphics	Language
Manipulation	Click	mainly use texts or speech as input
Learning	Need time to learn and adapt	No need to learn
Entrance	App download	Incorporated in any msg-based interface
Flexibility	Low, like machine manipulation	High, like converse with a human