MONASH University

FIT3175 - Usability

Voice User Interfaces and Multimodal Interfaces

Week 6 Lecture P1

Copyright Warning

Commonwealth of Australia Copyright Act 1968

Warning

This material has been reproduced and communicated to you by or on behalf of Monash University in accordance with section 113P of the Copyright Act 1968 (the Act).

The material in this communication may be subject to copyright under the Act. Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

Learning objectives

Designing Voice Interfaces

- Voice User Interfaces
- Voice interaction structure
- Voice interaction design guidelines

Multimodal Interface Concepts

- Unimodality and multimodality
- Human-machine interaction loop

Modality Design Beyond WIMP

- Variety of input and output modalities
- Emerging technologies

Preparing for Stage E Submission

Designing Voice Interfaces

Recap: Shneiderman's Interaction Styles

The interaction **styles** are a mechanism that allows interactions to occur:

- **Command language** Instructions manually specified with a required syntax.
- **Natural language** Imprecise instructions that are interpreted by the system.
- **Menu selection** User-selection from a range of system-defined options.
- **Form fill-in** User provides parameters to fit a supplied context.
- **Direct manipulation** Modification of objects with tightly-mapped feedback.

We typically associate **voice user interfaces** with the **natural language** style, however specific design of dialogues may implement commands, menus and fill-in-the-blanks.

IBM's Watson Destroys Humans in Jeopardy

Watson is a **NLP question answering system** developed by IBM.

The information retrieval abilities are combined with advance processing of natural language queries.

Watson made its debut in 2011 on the game show **Jeopardy!** against human champions.

https://www.youtube.com/watch?v=WFR3lOm_xhE

Voice User Interfaces

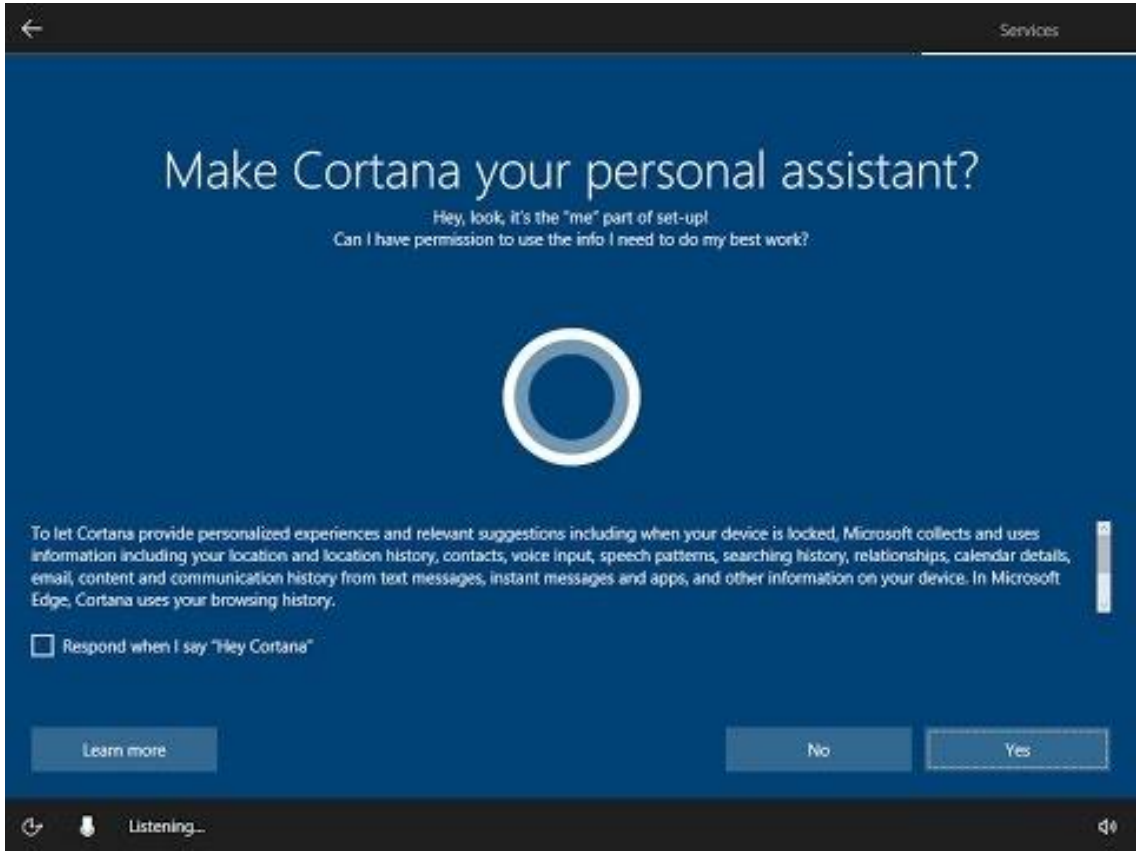
VUIs (also called Voice-Activated Personal Assistants) allow interactions through voice or speech commands Voice control is one of the most popular modes of interactions other than text.

Voice provides similar specificity of input compared to text

- Suitable for many people with physical impairments
- Suitable for people with limited visual attention
- Suitable for remote operation and multi-tasking

However, the **lack of visual signifiers and perceived affordances** requires application of different design guidelines.

Example: Windows installation



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

In the **Windows 10 Creators Update (2017),** the Windows first-time setup process can be completed using **Cortana**.

Many steps of the setup process can be completed using voice interactions.

Manual keyboard/mouse input is still required at a few points.

Accessibility impact of voice interfaces

Can you identify possible advantages or disadvantages of a voice user interface for each of the following accessibility issues?

Blindness or Low Vision

Deaf or Hearing Loss

Little or No Physical Mobility

Language Difficulties

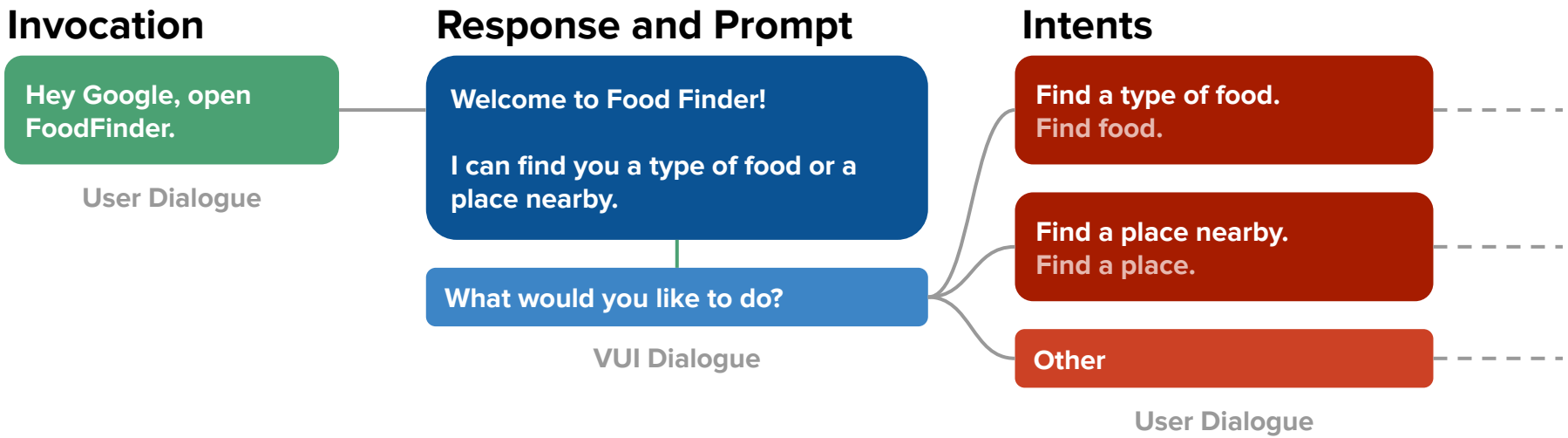
Cognitive or Attention Impairment

Some problems may require alternative solutions. However, many can be addressed in the way that interaction dialogues are designed.

https://uxdesign.cc/tips-for-accessibility-in-conversational-interfaces-8e11c58b31f6

Interaction structure

In most popular voice user interfaces, interaction dialogues are consistently structured using standard elements of conversational speech.



Information exchanged is best managed as simple turn-based conversations.

Invocation

An invocation requires a user to make a direct request for an action or application.

The invocation will typically include the following components:

- **Trigger phrase** recognised by the platform to start listening
- **Invocation name** that matches the action associated with your application
- Optional **invocation phrase** that links to a specific task within the application

"Ok Google, talk to FoodFinder to order lunch."

Trigger Phrase | Invocation Name | Optional Invocation Phrase

Responses and Prompts

The VUI's turn in a conversation allows it you respond to the user's request and request any required information for a task.

A simple response may be a single answer, but often continues the conversation to provide discoverability of the next action:

- Respond with a **greeting**, **acknowledgement**, **confirmation** or **apology**
- Continue the conversation with **question** or **suggestion** prompt.

"Welcome to FoodFinder.

I can search for food or a place. What would you like to do?"

Intents

An intents is a single task that the user requests from the VUI. This will involve processing user input to determine a command to execute.

An intent requires **utterances** for the different ways that a user may phrase a request.

"Find food."

"Search for food."

"I'm looking for food."

"I want some food."

Utterances should be designed to provide simulate intelligent language processing.

Utterance slots

An utterance can include data slots that behave similar to variable, but with a predefined list of acceptable values.

"I want to order a {food_type}."

Slots should be pre-populated with all values that a user is allowed to specify. This ensures that unknown/invalid values can be forwarded to an error-handling process.

{food_type} = "burger", "hamburger", "cheeseburger", "big mac", "whopper"...

Unknown values typically redirect to an **apology** response followed by a **re-prompt**.

Earcons - icons for your ears

Earcons are the various sound effects that are used in interactive systems to add aaural information and personality to a user experience.

The term is a fun ***play on words***:

- Icons (***"eye-cons"***) are small visual elements
- Therforem **"ear-cons"** are short aural elements

Earcons are not unique to VUIs - but play an important role in enhancing interactions.

Right: *Google provides a library of sounds that developers can programmatically insert into speech markup scripts.*

Select	Sound	Category
		all
☐	alarm clock	alarms
☐	assorted computer sounds	alarms
☐	beep short	alarms
☐	bugle tune	alarms
☐	digital watch alarm long	alarms
☐	dinner bell triangle	alarms
☐	dosimeter alarm	alarms
☐	mechanical clock ring	alarms
☐	medium bell ringing near	alarms
☐	phone alerts and rings	alarms
☐	radiation meter	alarms
☐	setting alarm clock	alarms
☐	spaceship alarm	alarms
☐	winding alarm clock	alarms
☐	ambient hum air conditioner	ambiences
☐	arcade room	ambiences
☐	barnyard with animals	ambiences
☐	carnival atmosphere	ambiences
☐	children group ambience	ambiences
☐	coffee shop	ambiences
☐	convention hall ambience noise	ambiences
☐	crickets with distant traffic	ambiences
☐	dmv background noise	ambiences
☐	daytime forrest bonfire	ambiences
☐	distant highway	ambiences
☐	factory background	ambiences
☐	factory manufacture background	ambiences
☐	factory morning start up	ambiences
☐	factory sounds	ambiences
☐	farm morning with sheep	ambiences
☐	fire	ambiences

https://developers.google.com/assistant/tools/sound-library

Recommendations from platform guidelines

Reading Between the Guidelines - a research survey carried out by Stacy Branham (Assistant Professor at the University of California, Irvine) identified common themes of 5 popular VUI guidelines in the market:

1. **Make conversations human**
2. **Make conversations personal**
3. **Make conversations efficient**
4. **Make conversations relational**
5. **Give the user a sense of control**

The research also explores the accessibility impact of these guidelines for blind users.

https://monash.hosted.exlibrisgroup.com/permalink/f/113cqvi/TN_cdi_acm_primary_3353797

1. Make conversations human

- **Model conversation after human speech**
  - Computers should adapt to the communication users learned first and know best. This helps create an intuitive and frictionless experience. **- Google**

- **Aim for natural conversation:**
  - Don’t emphasize grammatical accuracy over sounding natural. For example, ear-friendly shortcuts like ‘wanna’ or ‘gotta’ are fine. **- Microsoft**

- **Enact a persona**
  - If the voice skill to be designed is to tell users a joke, the voice image you set may be young and humorous; while for the technique of reading daily news, you need to use a more mature and stable voice image. **- Alibaba**

2. Make conversations personal

- **Set smart defaults when users are not explicit**
  - Use default values when the user is not specific. For example, if the user says, *"Make my room warmer",* Cortana should say, *"I've raised your room temperature to 72 degrees"* instead of *"Sure, what temperature?"* **- Microsoft**

- **Intelligently interpret and respond to users' with varied phrasing**
  - While the user might say *"plan a trip"*, he or she could just as easily say *"plan a vacation to Hawaii"* **- Amazon**

- **Use contextual data to cooperate with the individual**
  - Good error handling is context-specific. The conversational context of a request may be different on the second or third attempt **- Google**

3. Make conversations efficient

- **Reduce time and effort as compared to screens**
  - Saying *"Play the latest House of Cards"* is much easier than opening up an app, searching for "House of Cards", finding the latest episode, and pressing play. **- Microsoft**

- **Support multitasking with screen-free interaction**
  - Strive for a voice-driven experience that doesn't require touching or looking at the screen. **- Apple**

- **Keep conversations concise**
  - Longer prompts for novice users, shorter for experienced users. **- Google**
  - Can the utterance be stated in one breath? **- Amazon**

3. Make conversations efficient

- **Manage short-term memory load**
  - Most people can only remember a small amount of information when listening to instructions. Limit interactions to only what is absolutely required. For example, present only three items of a list at a time. **- Microsoft**

- **Scaffold Turn-taking**
  - Generally, end with a question before having the user respond. End the prompt right after the question. **- Amazon**

- **Don't launch into monologues**
  - Don’t go into heavy-handed details unless the user will clearly benefit from it.
  - Let users take their turn. **- Google**

4. Make conversations relational

"efficient communication relies on the assumption that there's an undercurrent of cooperation between conversational participants." **- Google**

- **Maintain friendly conversation but avoid niceties**
  - **NO:** *"Please accept the Ibento terms of service in order to proceed."*
  - **YES:** *"Sure, I can help with that. But first there's one thing you need to do: accept the Ibento terms of service."* **- Google**

- **Maintain transparency and take responsibility for errors**
  - Don't blame the user!
  - *"I can't reach your preferred florist right now to place your order. Should we wait a few minutes and try again, or order from another florist?"* **- Google**

5. Give the user a sense of control

- **Avoid teaching commands**
  - Teaching commands discourages experimentation and undermines trust. The implied message is that users have to say these exact phrases or they won't be understood. **- Google**

- **Agent dialogue should also be action-based, rather than prompt-based**
  - **Action-based:** *"Do you want to keep shopping or check out now?"*
  - **Prompt-based:** *"Now you can say keep shopping or check out."* **- Google**

- **Allow the user to make course corrections or discontinue a task immediately**
  - *"stop"*, *"cancel"*, *"shut up"* - always available without invocation.
  - Request confirmation depending on impact of cancellation outcome.

Limitations

Voice interfaces are difficult to perfect. Be aware of issues and possible solutions.

- **Failure to recognize voice commands results in undesirable actions.**
  - Recovery: ability to undo actions (may be costly depending on the task).
  - Prevention: read back feature (for users to the system's interpretation)
- **Learnability of new commands and actions**
  - Mitigation: provide tutorials and guides (may annoy user).
- **User uncertainty at which point their voice input is perceived**
  - Mitigation: provide ways for the user to know the status of the VUI

How can a voice user interface provide discoverable status to a user?

Limitations

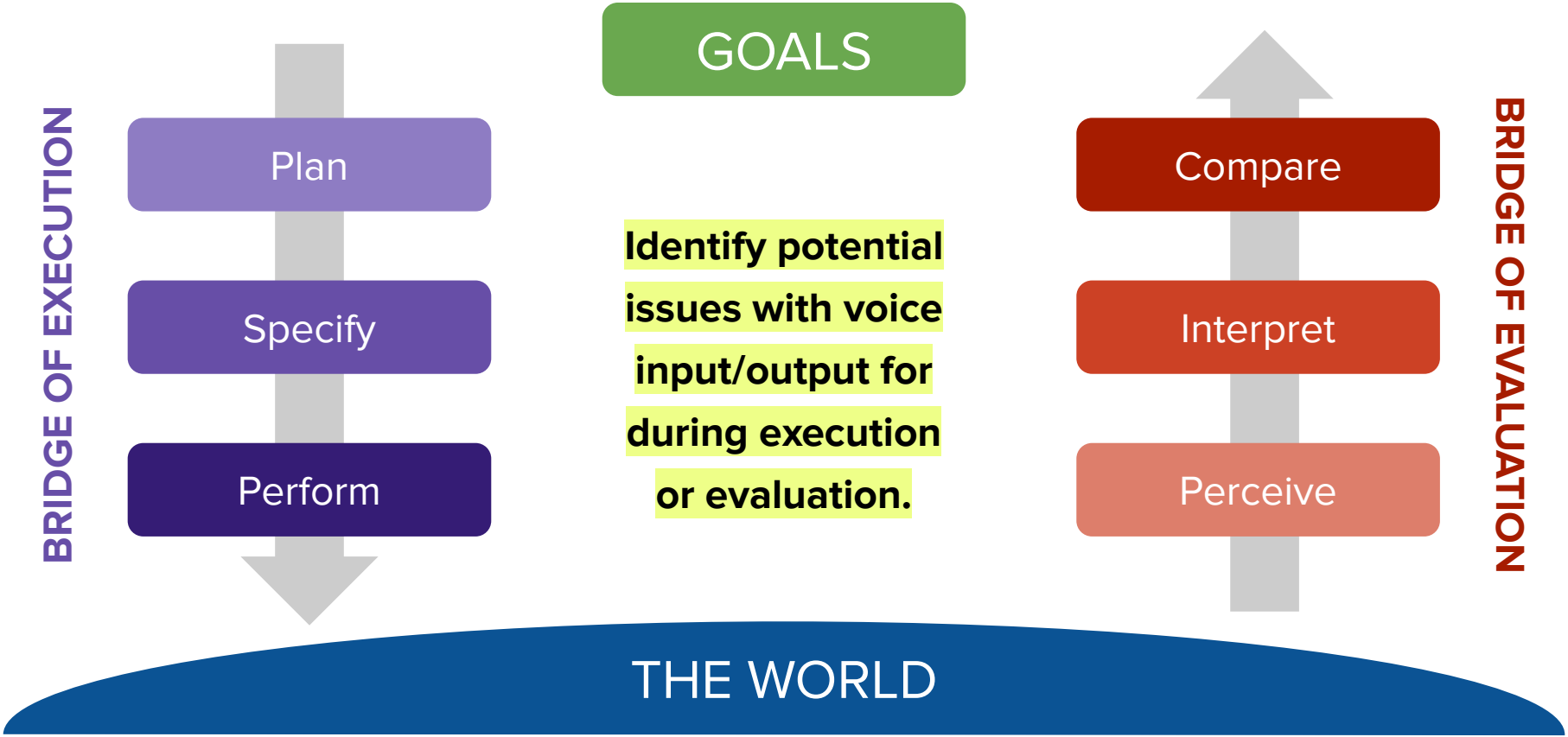
- **Users may be concerned with conducting voice dialogues in public settings.**
  - Mitigation: offer alternative interaction modes.

- **Users may be concerned with privacy of data collected by a voice interface.**
  - MItigation: do not make voice input mandatory.

- **Not offering a continuous listening mode.**
  - Mitigation: longer time-outs, feedback mechanisms

- **Personification increases engagement but may create unrealistic expectations.**
  - Mitigation: avoid using language that may give the impression of the VUI agent being a person; e.g. ‘I think’, ‘I feel’...

Consider: Norman's 7 Stages of Action



Example: Not-so-smart assistants

https://www.tiktok.com/@alisa.music_official

Consider the following conversation interaction with Siri:

User: Hey Siri, um... set timer for fifty minutes.

Siri: **Fifteen minutes. Counting down.**

User: Hey Siri, set timer for fif-***ty*** minutes.

Siri: **There's already a fifteen minute timer. But I can set another. What should I name this one?**

User: Hey Siri, cancel... both timers.

Siri: **It's cancelled.**

User: Hey Siri, set timer for ***FIF-TY*** minutes.

Siri: **Fifteen minutes. Starting now.**

User: How many minutes on my timer? ... Hey Siri, How many minutes on my timer?

Siri: **The timer has about fourteen minutes to go.**

User: Noooo! Hey Siri, cancel both timers.

Siri: **It's cancelled.**

User: Hey Siri, set timer for fifty-one minutes.

Siri: **Fifty-one minutes. Counting down.**

Multimodal Interface Concepts

What is "multimodality"?

Recall from Lecture 9:

> ***Enable multimodal input and output.***
> *Allow users to access content without having to rely on just one of their senses.*

Interaction **modalities** are the different methods of providing input or receiving output. Often there can be more than 1 way to do each.

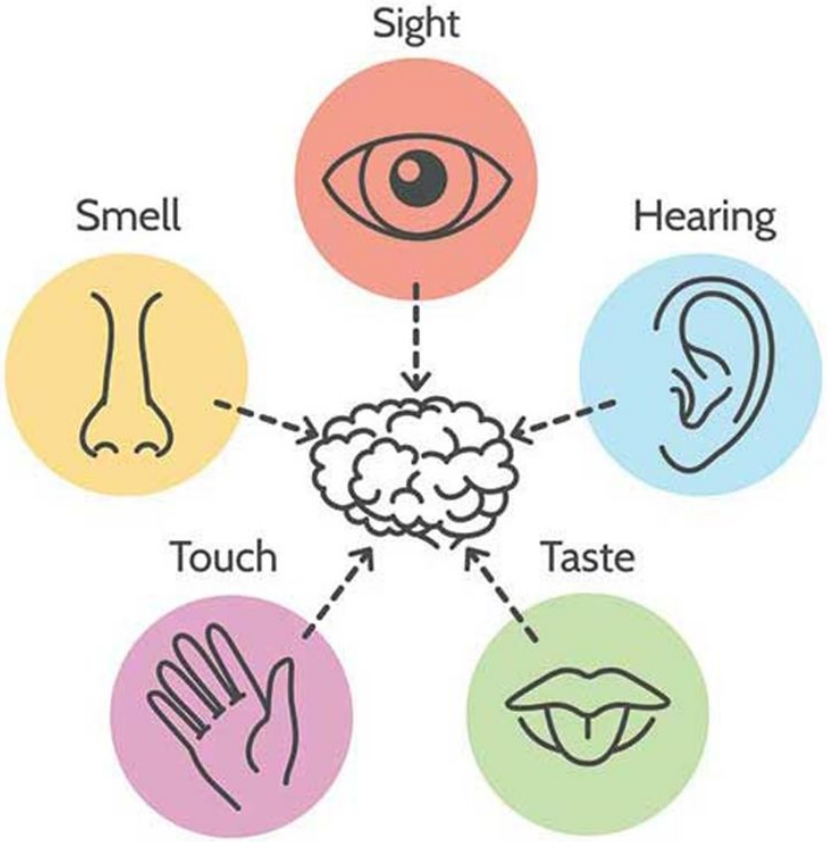
- **Unimodal** interfaces mainly rely on a single method of interaction
- **Multimodal** interfaces provide multiple methods that can be used in an integrated way - in **serial**, **parallel** or **both**.

Human interaction is multimodal

Human interaction with the real world regularly uses multiple senses, as well as physical and cognitive abilities.

Multisensory integration allows multiple sensory modalities to be combined into coherent and seamless representation.

Devices and interfaces that can take advantage of multiple senses and modalities of input have an opportunity to offer transformative experiences.



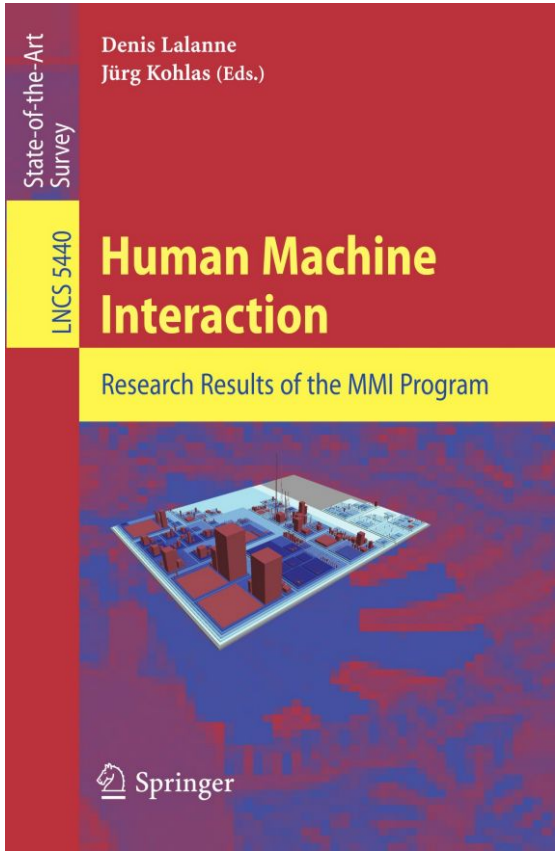
Objectives of MMI

According to [Dumas, Oviatt et al. 2009]:

> ***… Multimodal systems represent a new class of user- machine interfaces, different from standard WIMP interfaces.***

The goal of MMI is two-fold:

- to support and accommodate users' perceptual and communicative capabilities
- to integrate computational skills of computers in the real world, by offering more natural ways of interaction to humans



https://monash.hosted.exlibrisgroup.com/permalink/f/113cgvi/TN_cdi_springer_books_10_1007_978_3_642_00437_7_1

Richard Bolt's "Put That There" system

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

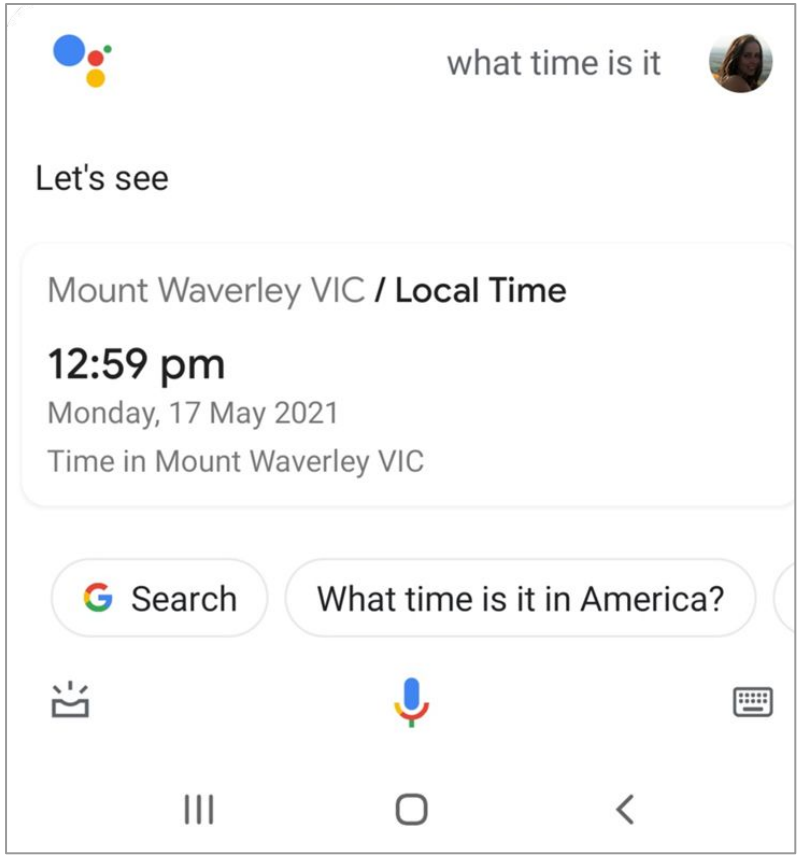
"Put That There" is a voice and gesture video interaction system developed at MIT in 1979.

Note the use of:

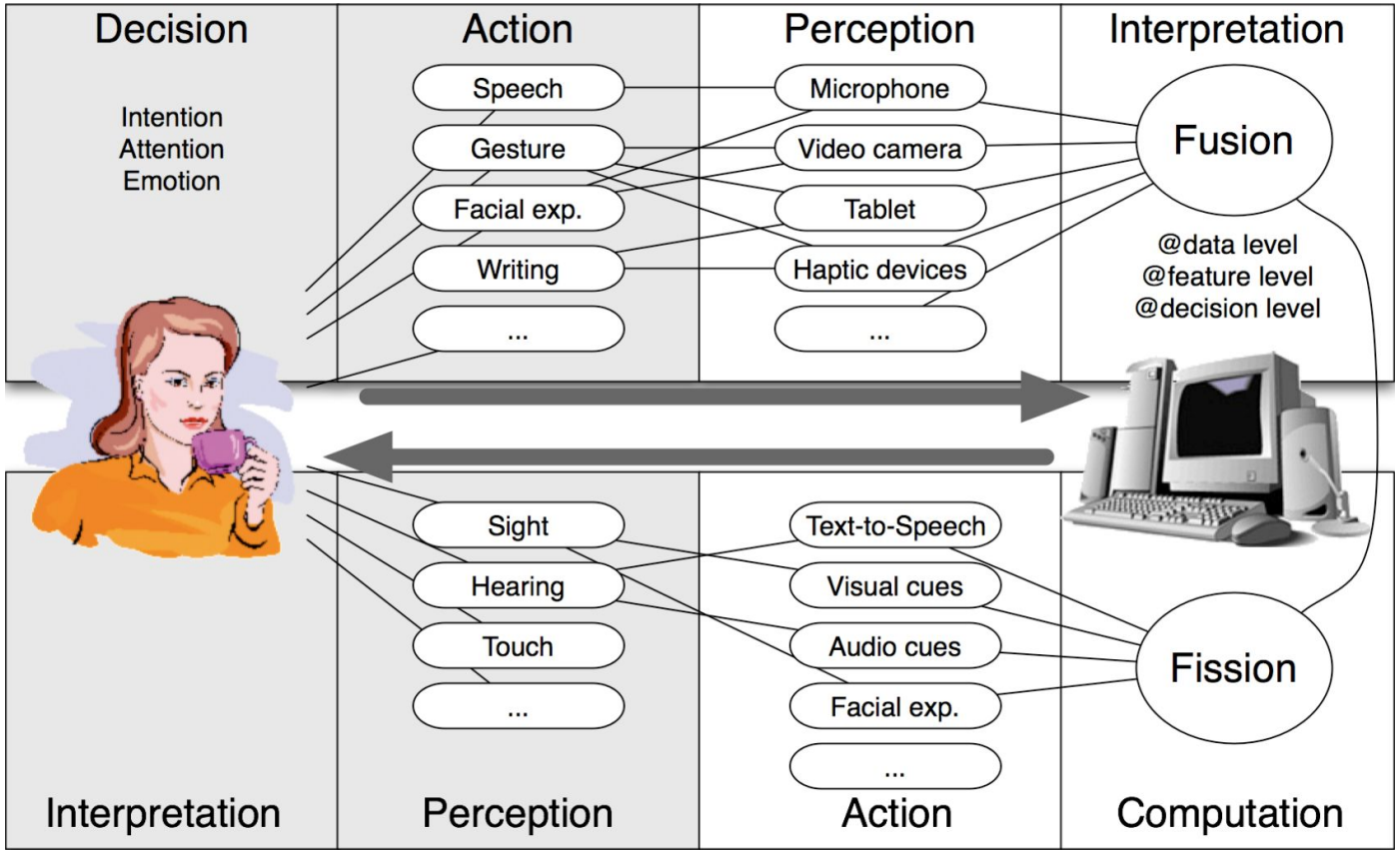
- Natural language
- Gestures
- Contextual queries
- Simultaneous input

Modern examples: assistant interfaces

Pointing, gesture and speech input. Visual and aural output.



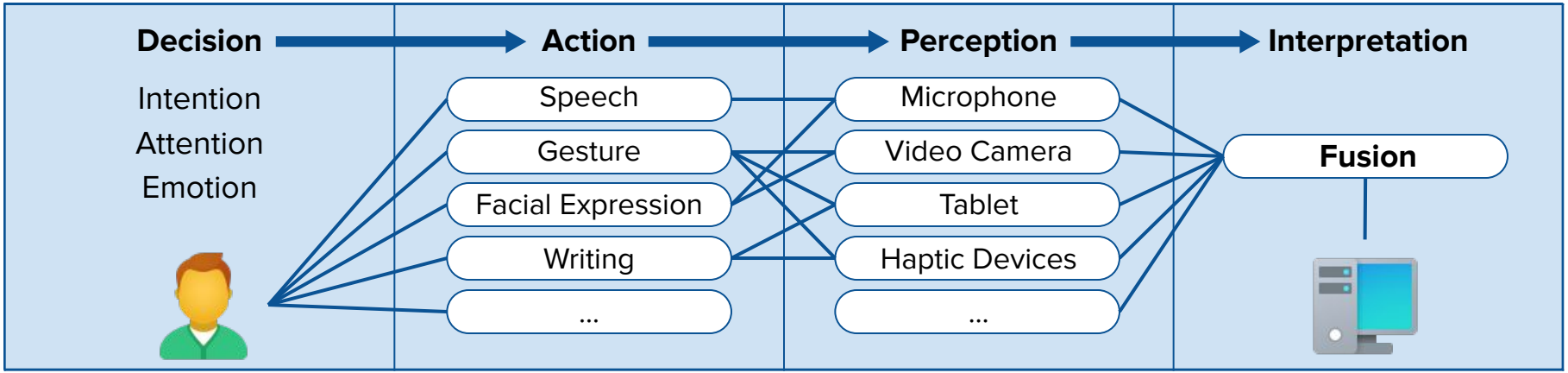
Multimodal human-machine interaction loop



Note: *Consider the similarities between this interaction model and Norman's 7 Stages of Action.*

Dumas, Oviatt et al. 2009

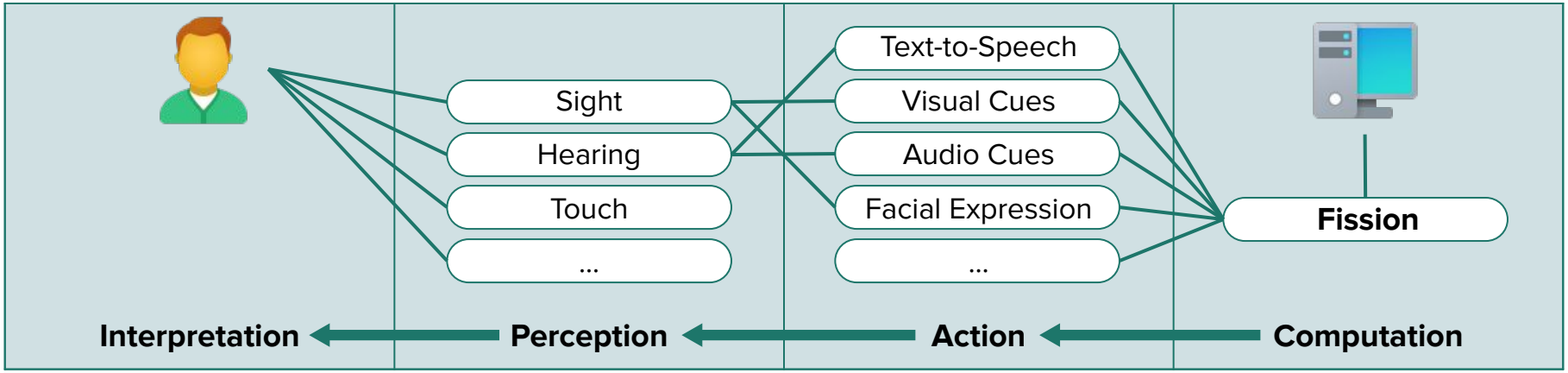
Human-to-machine interaction: Fusion



When an interaction is performed, the modalities are parsed to synthesize commands.

1. **Decision state:** The user plans a task using various thought processes.
2. **Action state:** Execution methods are selected and performed by the user.
3. **Perception:** The machine grasps signals from the available input sources.
4. **Interpretation:** Various signals are fused together and meaning is extracted.

Machine-to-human interaction: Fission



Once the action and meaning are known, compute and return a result to the user.

1. **Computation state:** Prepare feedback dialogue, determine output modalities.
2. **Action state:** Send relevant message signals to available output modalities.
3. **Perception:** The user detects the various signals generated by the computer.
4. **Interpretation:** The user understands the combined meaning of the signals.

Advantages

- **Flexibility in input/output modes**
  - Deliver natural and efficient interaction
  - May increase task efficiency
  - Help prevent overuse of single mode
- **Alternatives to meet the needs of diverse users**
  - Wider range of users, tasks and scenarios
  - More robust
- **Greater expressive power**
  - User may process information faster
  - Improve error handling & reliability
  - More engaging

Challenges

- **Real-time continuous processing and feedback**
  - Processing multiple inputs is computationally expensive
  - Security and privacy issues
- **Size and type of recognition vocabularies**
  - Mapping from input modes and specific user actions
  - Compatibility of input and output modalities
- **Learnability**
  - Taking into account cognitive and physical abilities
- **Multimodal Input integration**
  - Fusion mechanisms still in their infancy

Ambiguous interpretation

The **fusion** of input signals presents some one of the largest challenges in effective multimodal interaction:

Users' multimodal inputs are subject to **interpretation** - where input signals contain more **implicit** information, the result is potential **ambiguity**.

- **Ambiguity** exists when more than one interpretation of input is possible.
- A **multimodal ambiguity** may also be created when a combination of inputs produces incoherent meanings.

Effective systems must include methods to manage ambiguity.

Ten Myths of Multimodal Interaction

How true are the following, based on what you have learned and experienced?

Myth 1: If you build a multimodal system, users will interact multimodally.

Myth 2: Speech and pointing is the dominant multimodal integration pattern.

Myth 3: Multimodal input involves simultaneous signals.

Myth 4: Speech is the primary input mode in any multimodal system that includes it.

Myth 5: Multimodal language does not differ linguistically from unimodal language.

Sharon Oviatt, https://dl.acm.org/doi/10.1145/319382.319398

Ten Myths of Multimodal Interaction (continued)

Myth 6: Multimodal integration involves redundancy of content between modes.

Myth 7: Individual error-prone recognition technologies combine multimodally to produce even greater unreliability

Myth 8: All users' multimodal commands are integrated in a uniform way.

Myth 9: Different input modes are capable of transmitting comparable content.

Myth 10: Enhanced efficiency is the main advantage of multimodal systems

Sharon Oviatt, https://dl.acm.org/doi/10.1145/319382.319398

Modality Design
Beyond WIMP

Moving beyond the WIMP metaphor

What is the most unique interface that you use?

Increasingly multimodal input

Multimodal systems will make use of a variety of input modes, which may be conventional or unconventional.

Pointing **Typing** **Multi-touch** **Gestures** **Haptics**

Accelerometers **Gyroscopes** **Proximity** **Tangible objects**

Speech **Non-speech sounds** **Eye gaze** **Facial expression**

Body movement **Head position** **Brain activity** **Biometrics**

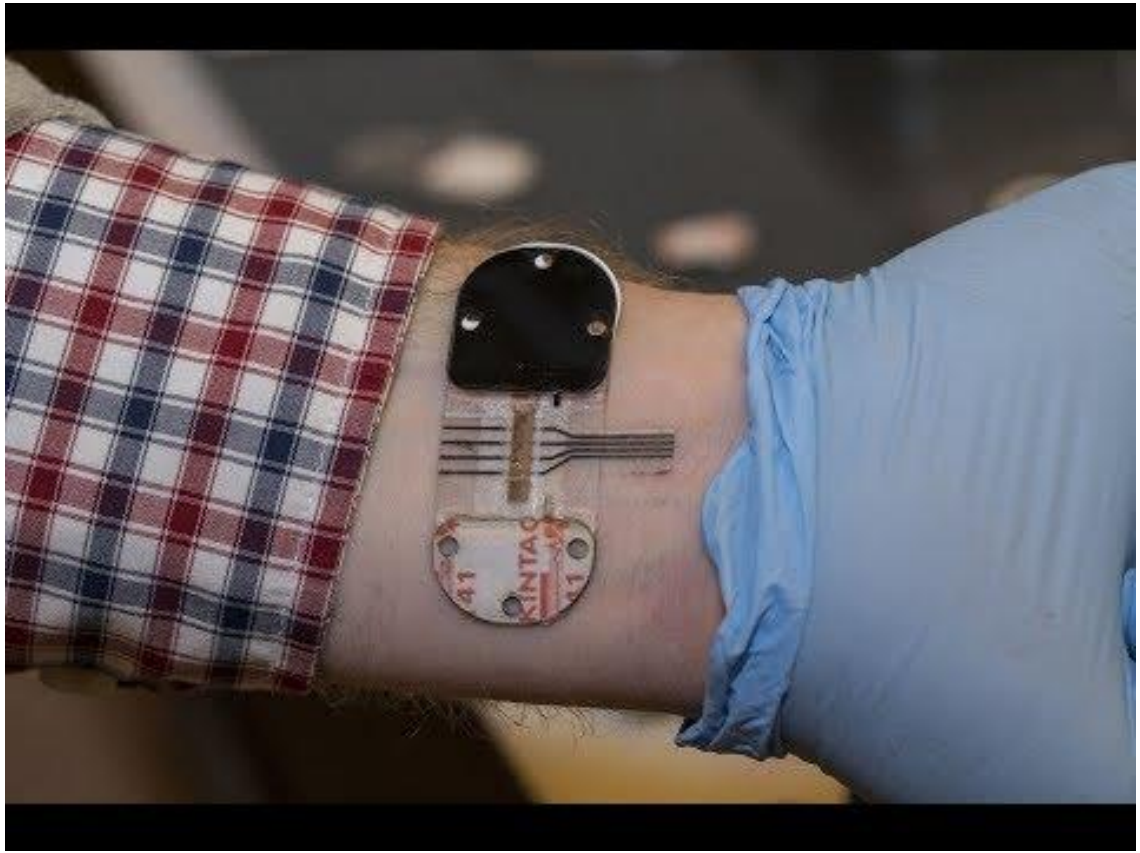
The selection of input modes depends on characteristics of the tasks, contexts, user ability. The usability of each mode must be considered.

Variety of output modalities

A multi-modal system can also make use of a wide variety of output modes.

- **Visual**
  - Text, graphics, animations, virtual and augmented reality
- **Auditory**
  - Speech, non-speech sound, earcons
- **Haptics**
  - Vibration, force-feedback
- **Emerging technologies**
  - Scent, taste, temperature

Biometric interfaces



In current use, we typically think of biometrics in terms of security.

However, biometric sensors can provide a continuous data stream that can be integrated with other data, creating greater contextual interaction.

https://www.youtube.com/watch?v=Wrf4LiXPt6E

Brain control interfaces

Neo-Noumena from Monash's **Exertion Games Lab** combines BCI, AR and AI to present dynamically rendered representations of emotions.

The project explores both the interaction between people and computers, as well as the interpersonal interactions of users.

Haptic interfaces

Haptic interfaces provide a sense of touch for interactive visualisations.

Beyond simple vibrations, haptic input and output devices allow provide active tactile feedback when interacting with virtual objects.

https://www.youtube.com/watch?v=OK2y4Z5IkZ0

Olfactory interfaces

Olfactory interfaces that can detect smell or emit scents.

On-Face Olfactory Interfaces research considers the utility of olfactory sensors as wearable accessories.

An on-face scent delivery approach has advantages over other body locations.

https://www.youtube.com/watch?v=Wv0VcjBS0ro

Gesture interfaces

Car navigation design balances needs of complex system functionality with the constraints of the operating context.

A current solution offered by BMW is a navigation system controlled using a variety of hand gestures.

https://www.youtube.com/watch?v=IIh9w_K7LHE

Metaverse

https://about.facebook.com/meta/

In context: Technology 20 years from now

The adoption of emerging technologies is difficult to predict. Technologies are always under active development. **Consider the development of mobile communication.**

1980

1990

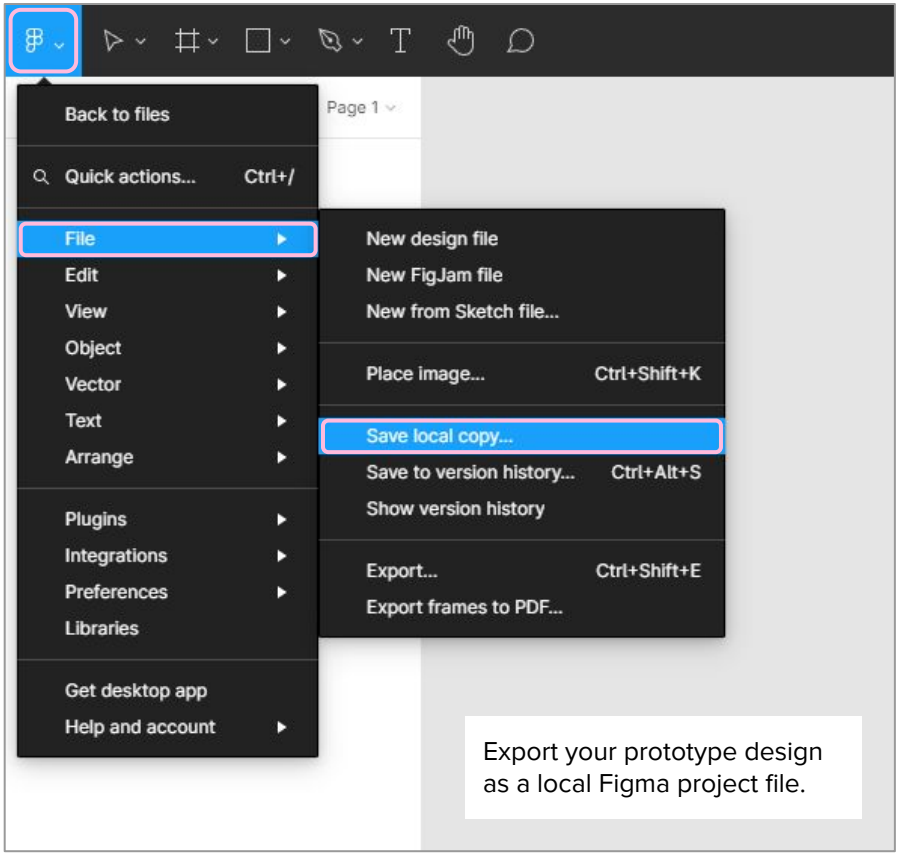
2000

2010

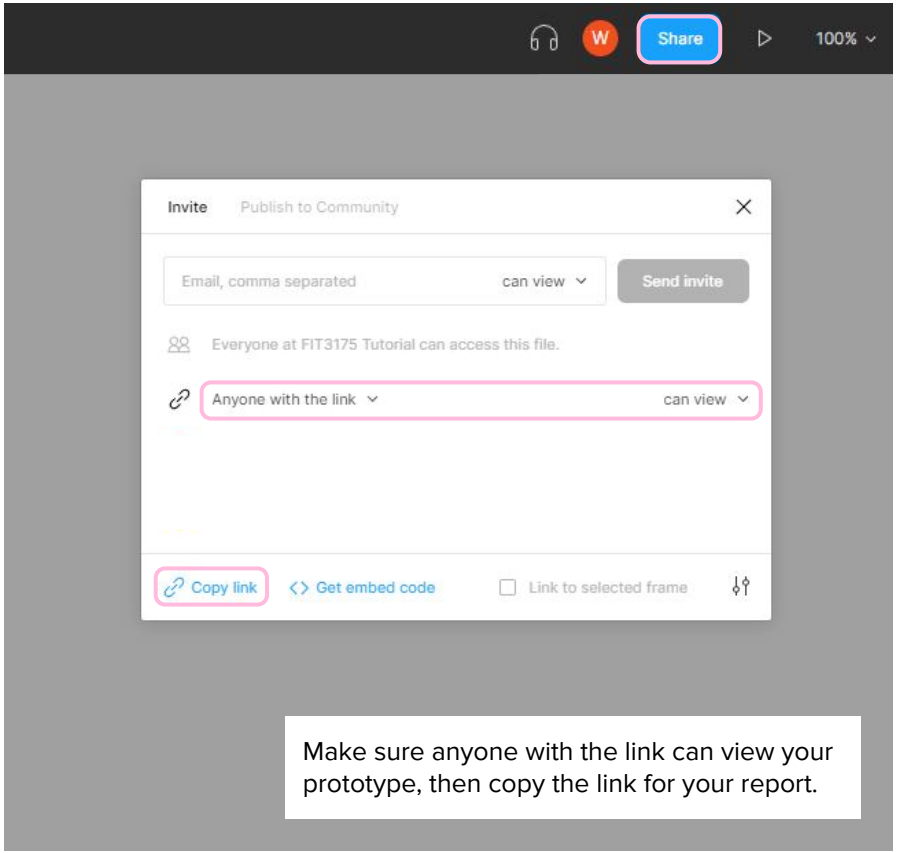
2020

Preparing for Stage E Submission

Stage E - Figma prototype submission



Export your prototype design as a local Figma project file.



Make sure anyone with the link can view your prototype, then copy the link for your report.

Next session

- UX beyond traditional interfaces
- Stage F in-class presentations

Reminders

- **Peer Evaluation for Stage D closes on Tuesday**
  - Rate your group members and provide feedback to staff.
- **Stage E + F are submissions next week**
  - Stage E: Group high fidelity prototype (due Friday week 6, 11:55PM)
  - Stage F: Individual presentation (in-class, Week 6 Tutorial P2)