Designing Interactive Systems 2

Lecture 11: Multimedia & Multimodal Interfaces

Prof. Dr. Jan Borchers
Media Computing Group
RWTH Aachen University

hci.rwth-aachen.de/dis2

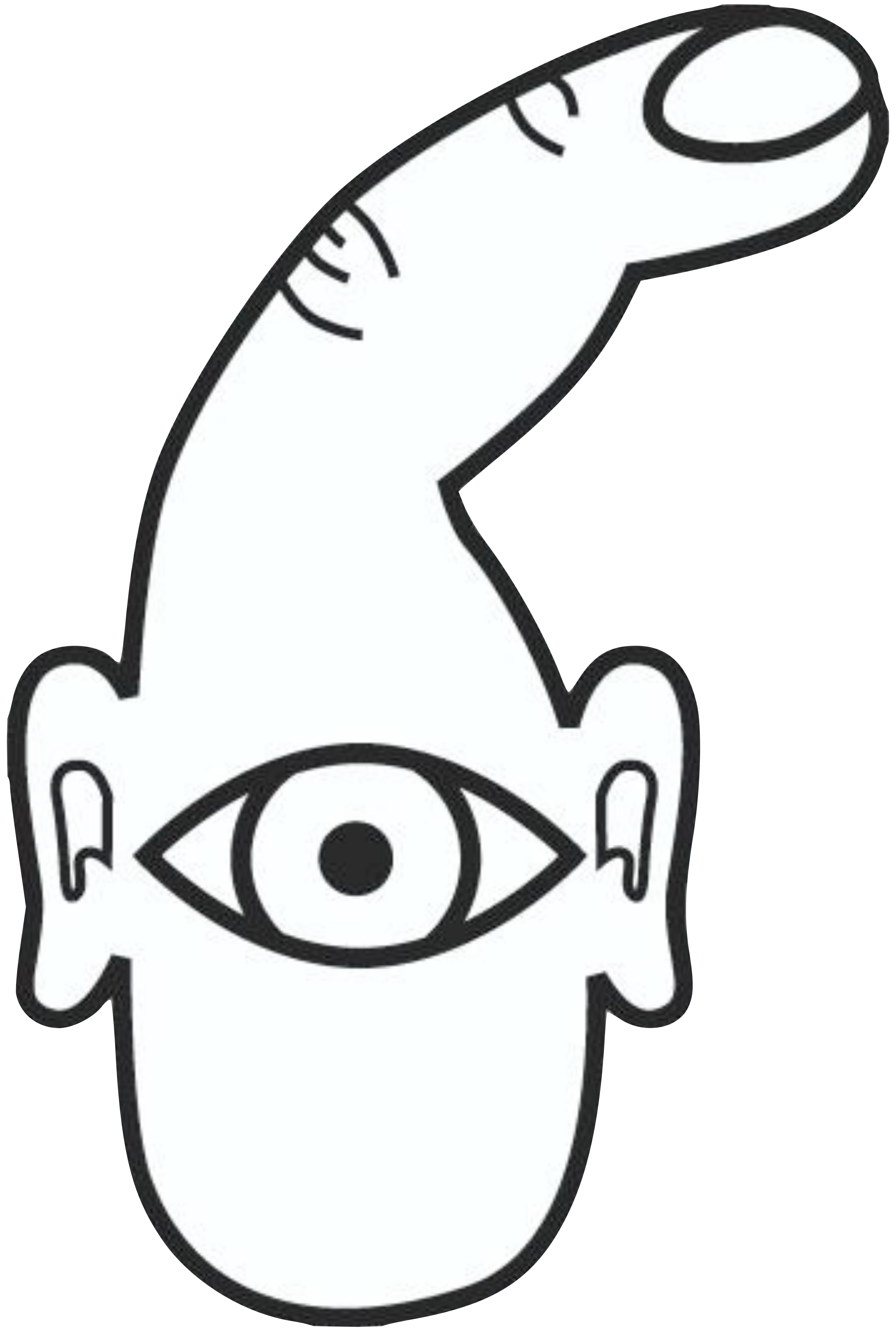
CHAPTER 32

Multimedia & Multimodality

In-Class Experiment

- Which senses have you used when you have interacted with your computer? Give examples!

How the desktop PC used to see us



From: O'Sullivan, Igoe: Physical Computing

using more
than one

$\overbrace{\text{Multi}}\underbrace{\text{modality}}$

Multimodality

type of communication channel used to convey or acquire information

using more
than one

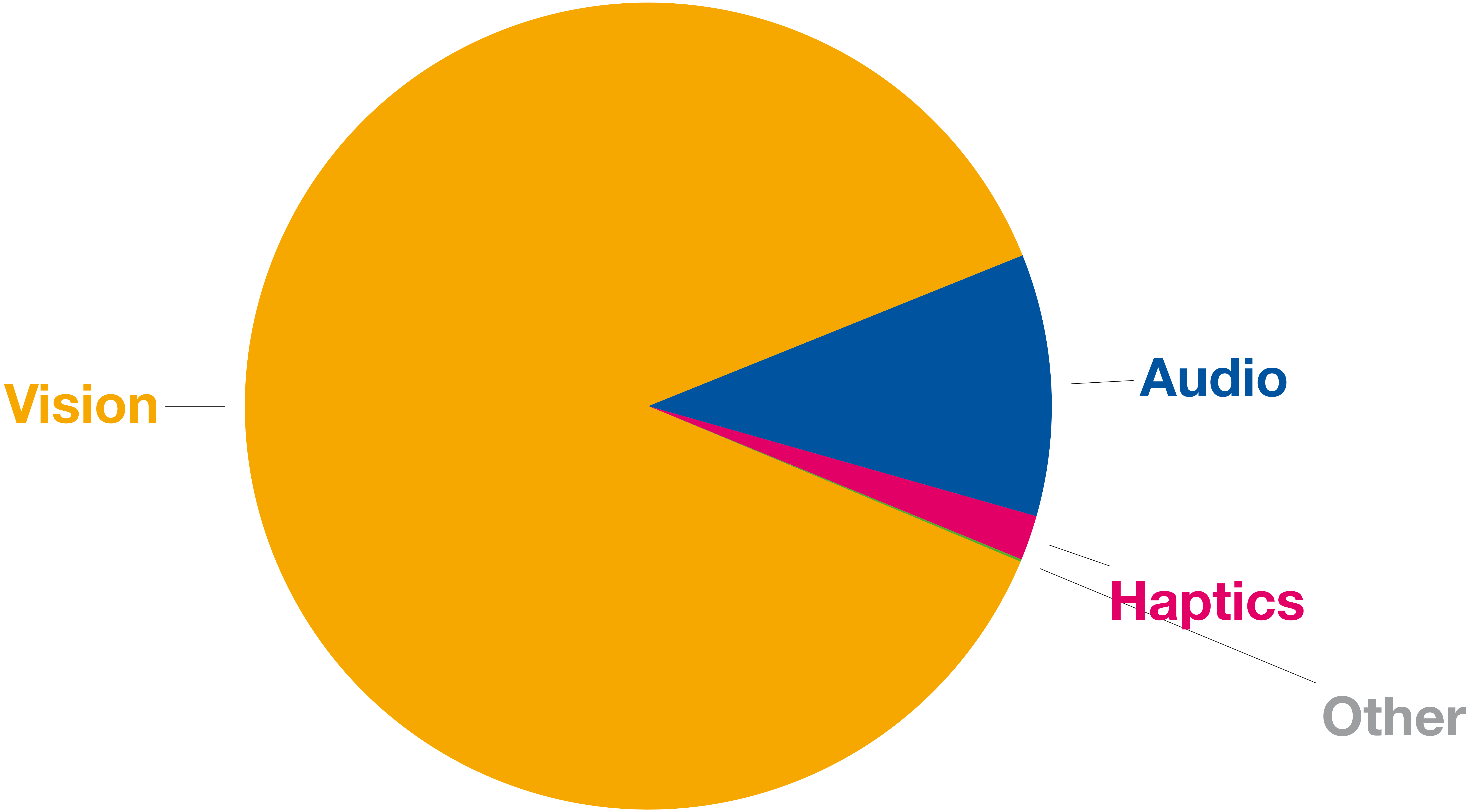
Multimedia

text, images, music,
video, animation, …


- LUNCH :

Multimodal interactions are natural.

Usage of Modalities in Computer Systems



Put That There

Advantages of Multimodal Systems

Input

- Fallback input techniques
  - Increased usability
  - Increased accessibility
- Prevent errors, increase robustness
- Bring more bandwidth to the communication

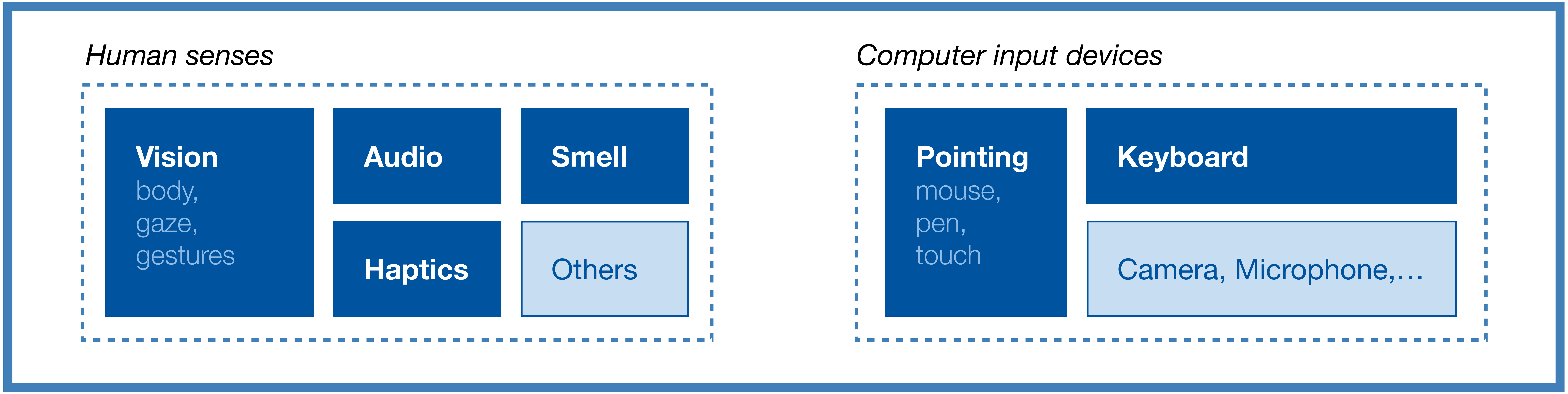
Output

- Redundancy
- Synergy effects
- Increased bandwidth
- Realism

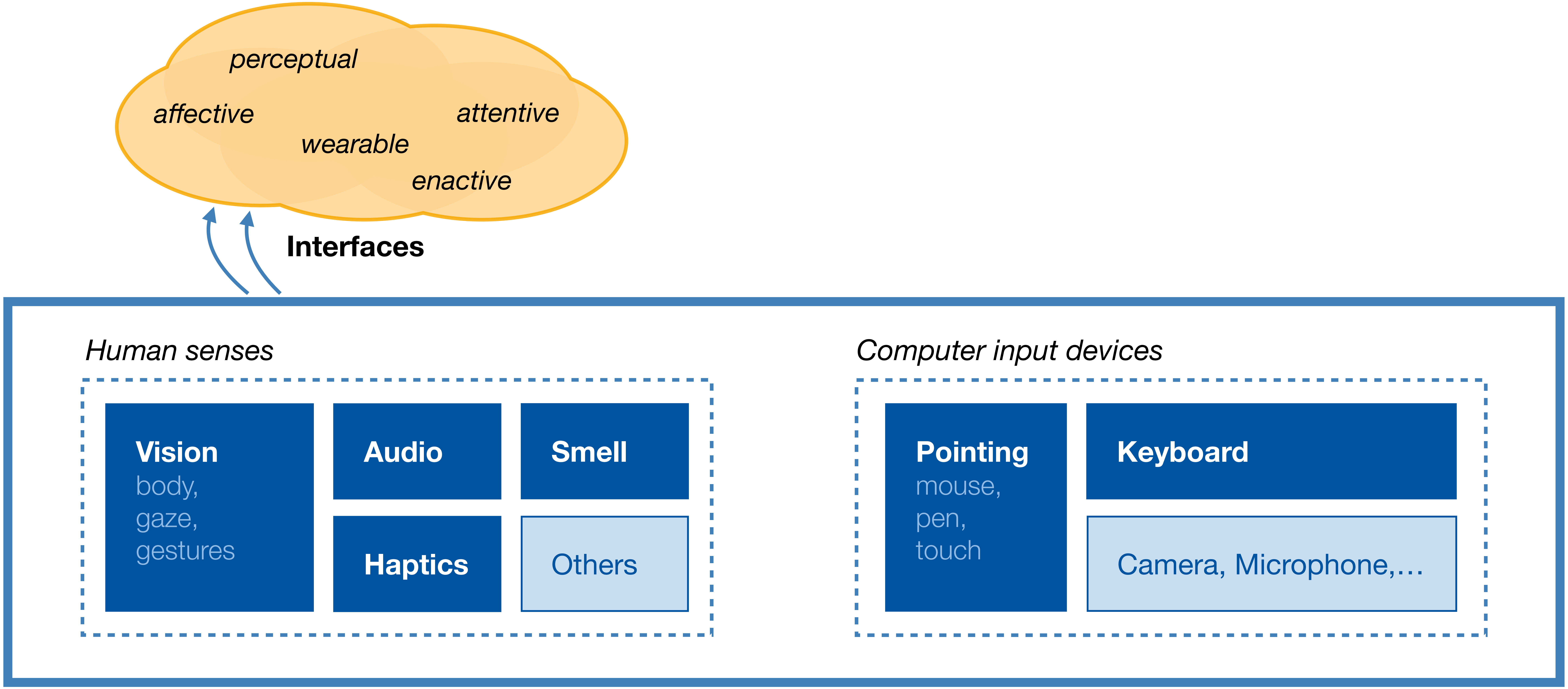
In-Class Exercise: Design Space of Multimodality

	Visual		Auditory		Haptic	
	Input	Output	Input	Output	Input	Output
Control						
Data						

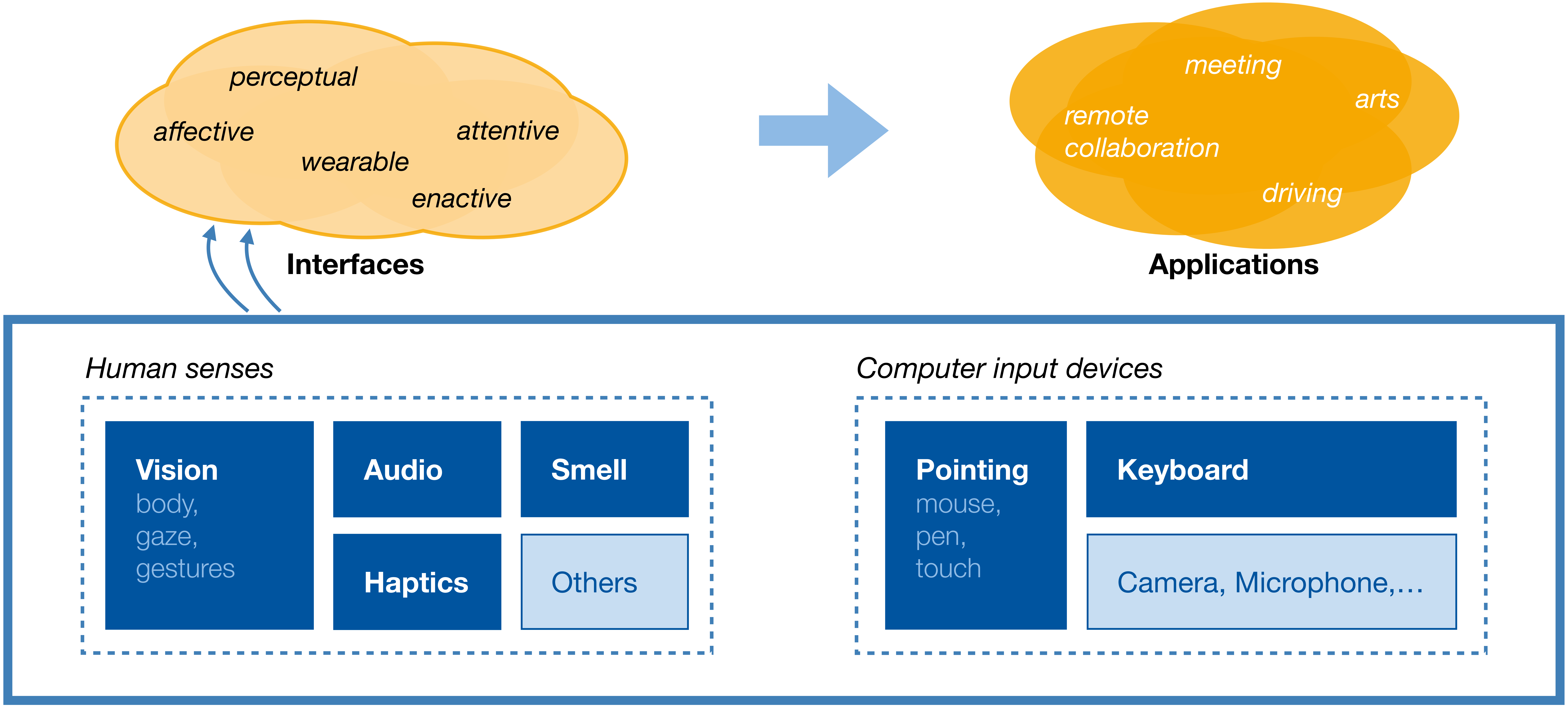
Interfaces for Multimodal Interaction



Interfaces for Multimodal Interaction



Interfaces for Multimodal Interaction



CHAPTER 33

Audio

In-Class Exercise: Audio Output

- **Advantages**
  - Undirected
  - Needs no screen space
  - No visual contact necessary
  - Gets attention even when distracted
  - Reaches entire group
- **Disadvantages**
  - Undirected
  - Attention-grabbing
  - Annoying user and others
  - Transient
  - Dictates speed
  - Cannot overlap easily

Audio Output: Types

Noise	Volume, duration
Beep	Volume, pitch, duration
Melody	Volumes (emphasis), pitches, durations, sequence
Chords	Melody, plus harmony (minor/major, dissonance)
Speech	Textual contents

Audio Output: Noise

- Only audio output that may be "natural"
- Also used artificially in applications: **Auditory Icons**
  - Play everyday sounds along with actions of the system
  - Particularly useful in complex situations (too much to watch)
  - But: not everything maps to a sound

In-Class Exercise: SonicPhotoshop

- What natural sounds could represent the standard drawing operations in Photoshop (draw, move, copy, delete, rotate)?

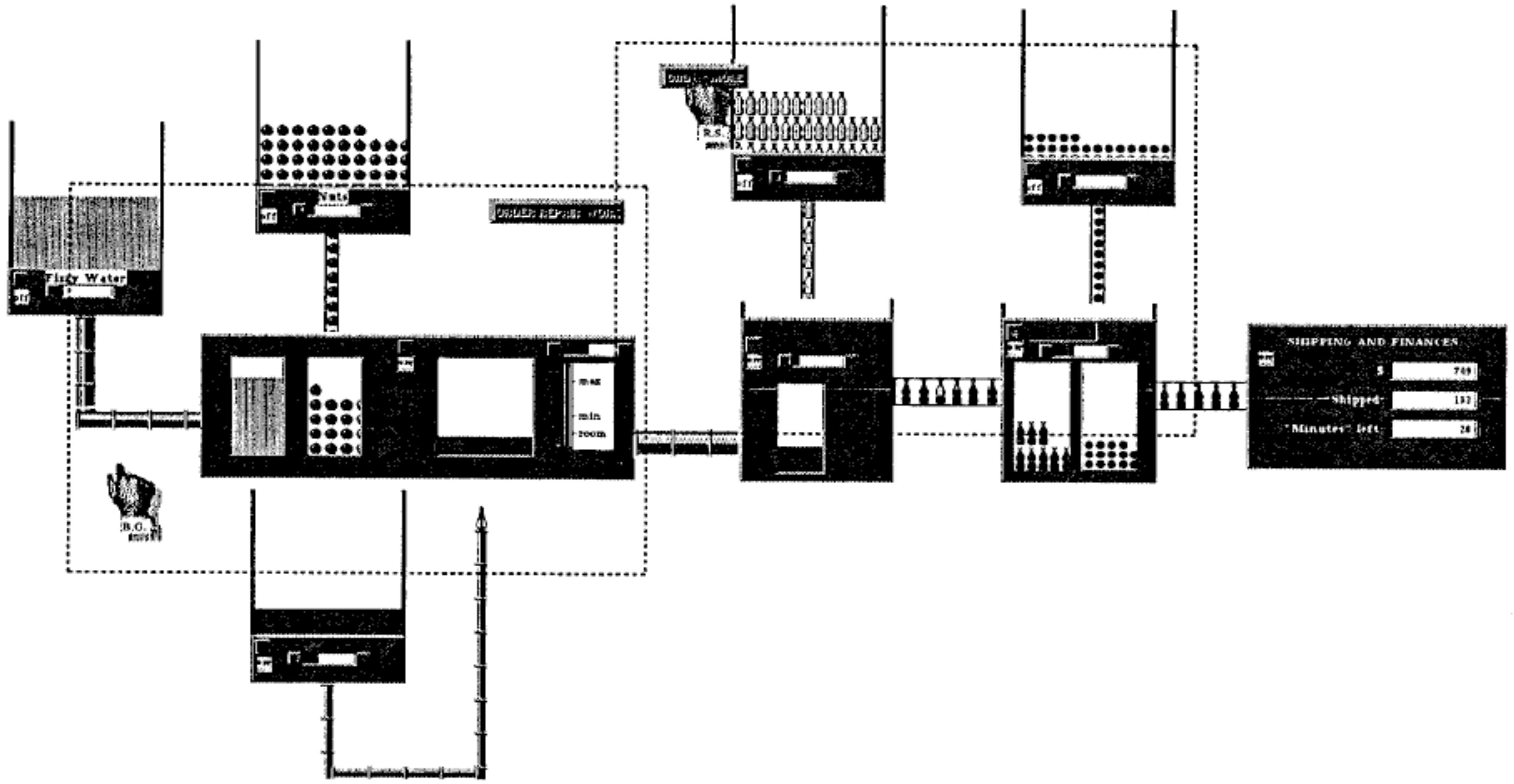
The SonicFinder (Gaver, 1989)

- Enhanced Apple Mac Finder
- Auditory icons using noises for desktop objects and operations
  - Files wooden, folders papery, applications metallic, bigger = deeper sound
  - Throw into wastebasket: smashing
  - Copying: a problem!
  - Filling jug with water, rising pitch = progress
- Mostly satisfied users, but natural analogies are limited
- Also used in experimental Mosaic web browser and commercial products
  - Since Mac System 7 Finder, MS Office,…

Gaver, W. W.: *The SonicFinder: An interface that uses auditory icons.* Human-Computer Interaction 4(1), 67–94, 1989.

ARKola (Gaver, 1991)

- Simulation of Coke bottling plant
- 2 remote users collaborating
- Noises reflect status (bottles clink when released, liquid splashes if wasted, bottles break if wasted)
- Result: improved collaboration



Gaver, W. W., Smith, R. B., and O'Shea, T.:
Effective sounds in complex systems: the ARKOLA simulation. In Proc. CHI '91, 85–90.

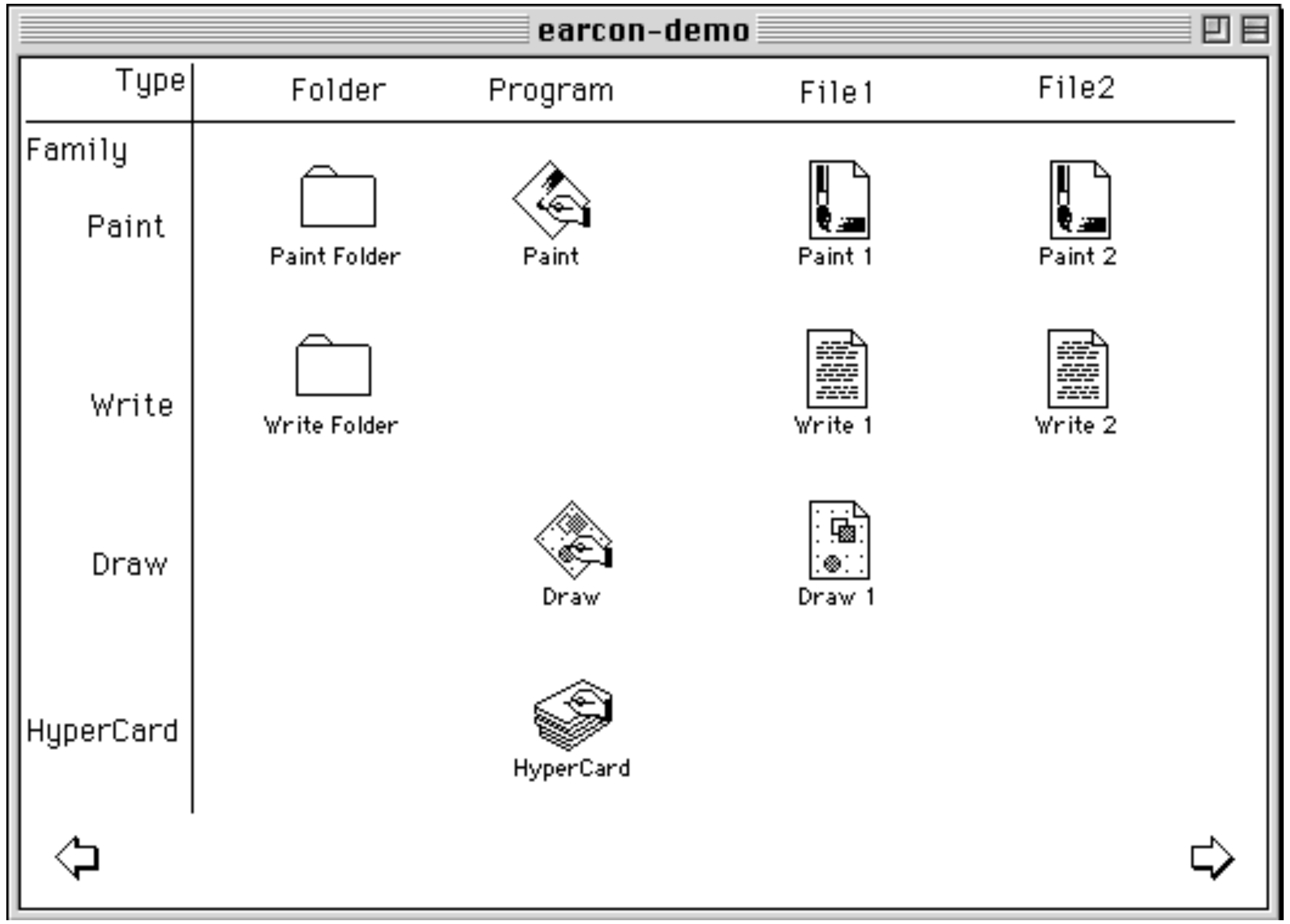
Audio Output: Melodies

- **Earcons** (Blattner 1989)
  - Abstract, synthetic tones, non-verbal
  - Building block: motif
    - Short, rhythmic pitch sequence with variable intensity, timbre and register
  - Combine into complex messages

Demo: Earcons

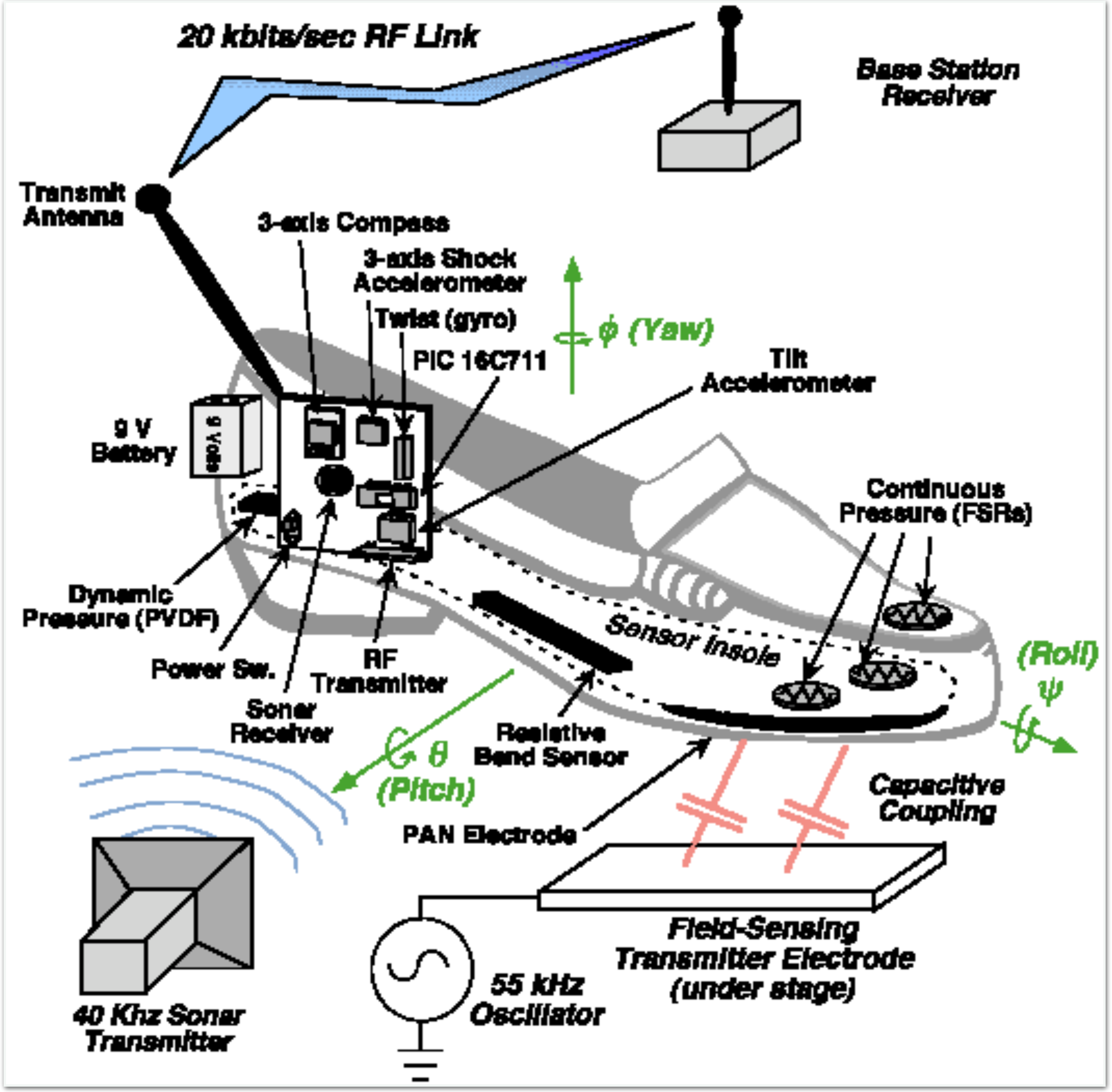
Earcons: Guidelines

- Use instrument timbres for distinction
- Pitch & register don't work well on their own, unless with large differences, use pitches 150Hz–5kHz
- Vary rhythms greatly, do not use notes shorter than 8ths, vary number of notes. Can parallelize.
- Intensity: Bad for distinction. Keep in small range (10–20db over background noise), user may adjust volume. Use localization.
- Join motifs with .1s gaps



Gait Shoe

- Example of melodic audio output
- Paradiso, MIT Media Lab, 2004
- Shoes detect bad walking habits in patients and alert them through dissonance in musical feedback stream



Earcons vs. Auditory Icons

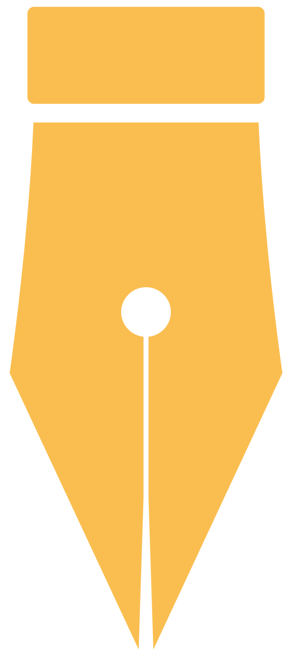
Musical Listening

Perceiving Notes

Earcons vs. Auditory Icons

Musical Listening

Perceiving Notes



Everyday Listening

Perceiving Events

Audiolization

- Visualization using audio
- Example: Algorithm audiolization
  - Hear how QuickSort works
  - May help with debugging (Alty, CHI 1997)
- Example: Scientific audiolization
  - Listening to turbulence (Blattner 1992)

CHAPTER 33

Audio Output: Speech

Audio Output: Speech

- Also known as Text-to-speech (TTS) systems
- Using recorded chunks of different size
- 3 Levels of quality:
  Phonemes (flexible, small database) < words < sentences (more realistic)

Speech Output

- **Advantages**
  - Natural, familiar, emotional
  - For the visually impaired
  - Eyes-free

- **Disadvantages**
  - Slower than visual (bandwidth)
  - Transient/ephemeral
  - Hard to browse/search
  - Synthetic missing "prosody" (melody,…)
  - Unlike noise, cannot fade into background (hard to ignore)

Speech in Humans

The primary way of human communication

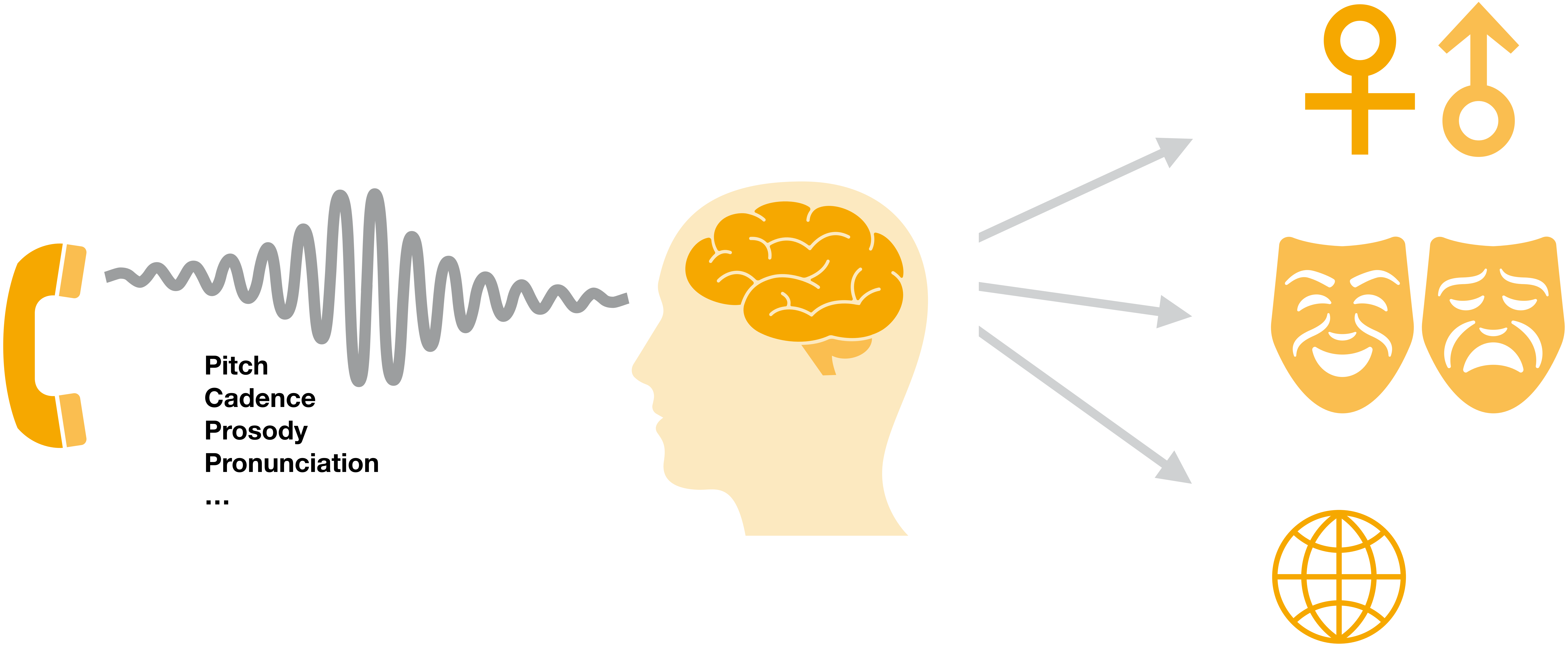
Speech implicates more parts of the brain than other functions

Left brain hemisphere is specialized to process speech

Speech in Humans

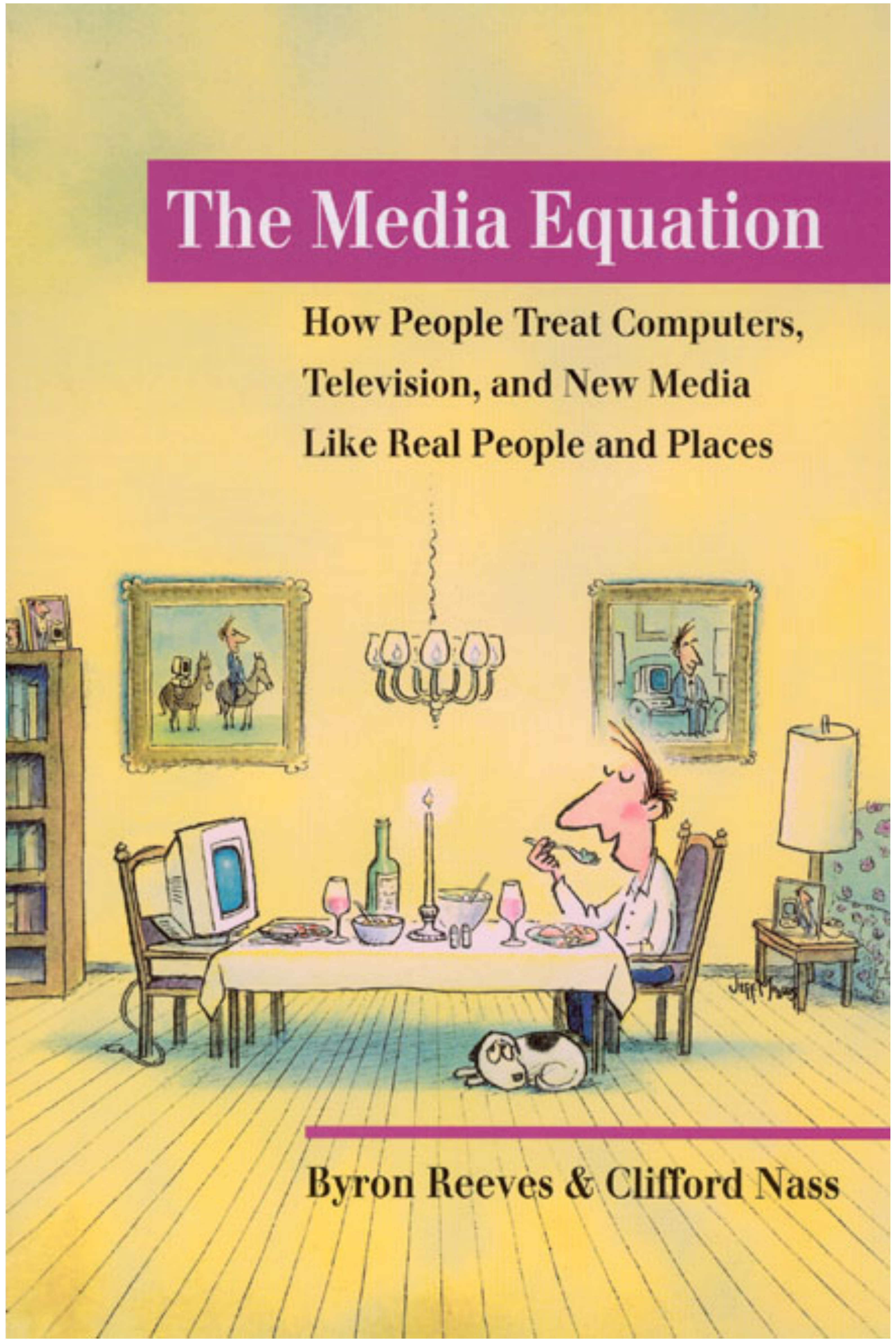
- Cliff Nass (Stanford): Speaking fundamental, starts early in human development
    - IQ > 50, brain > 400g
    - At 18months through adolescence 1 new word every 2h
    - 1-day old differentiates speech vs. other sounds
    - 4-day olds differentiate native from foreign language
    - Adults differentiate 40–50 phonemes/sec (other sounds < 20)
    - Cope with cocktail parties

Speech Output: The Social Aspect



The Media Equation

- Clifford Nass, Byron Reeves (Stanford University)
- Overall message:
  Users treat computers and other interactive media like humans
  - Computers are social actors, HCI = HHI
- What does this mean for speech output?
- Following sample data from Clifford Nass

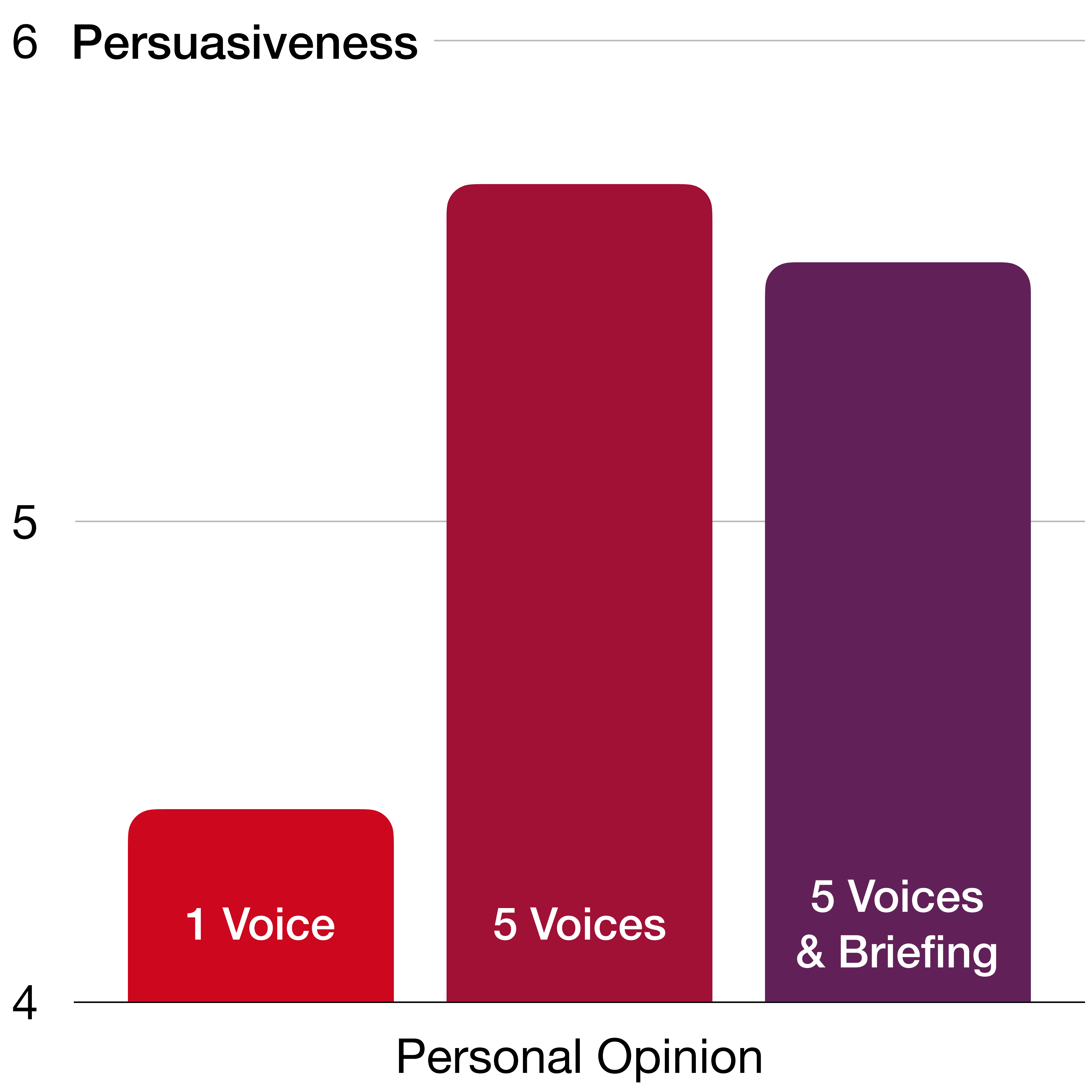


The Media Equation: Persuasiveness

- Sample Experiment:
  Five positive Amazon customer recommendations read
  by one vs. several computer voices

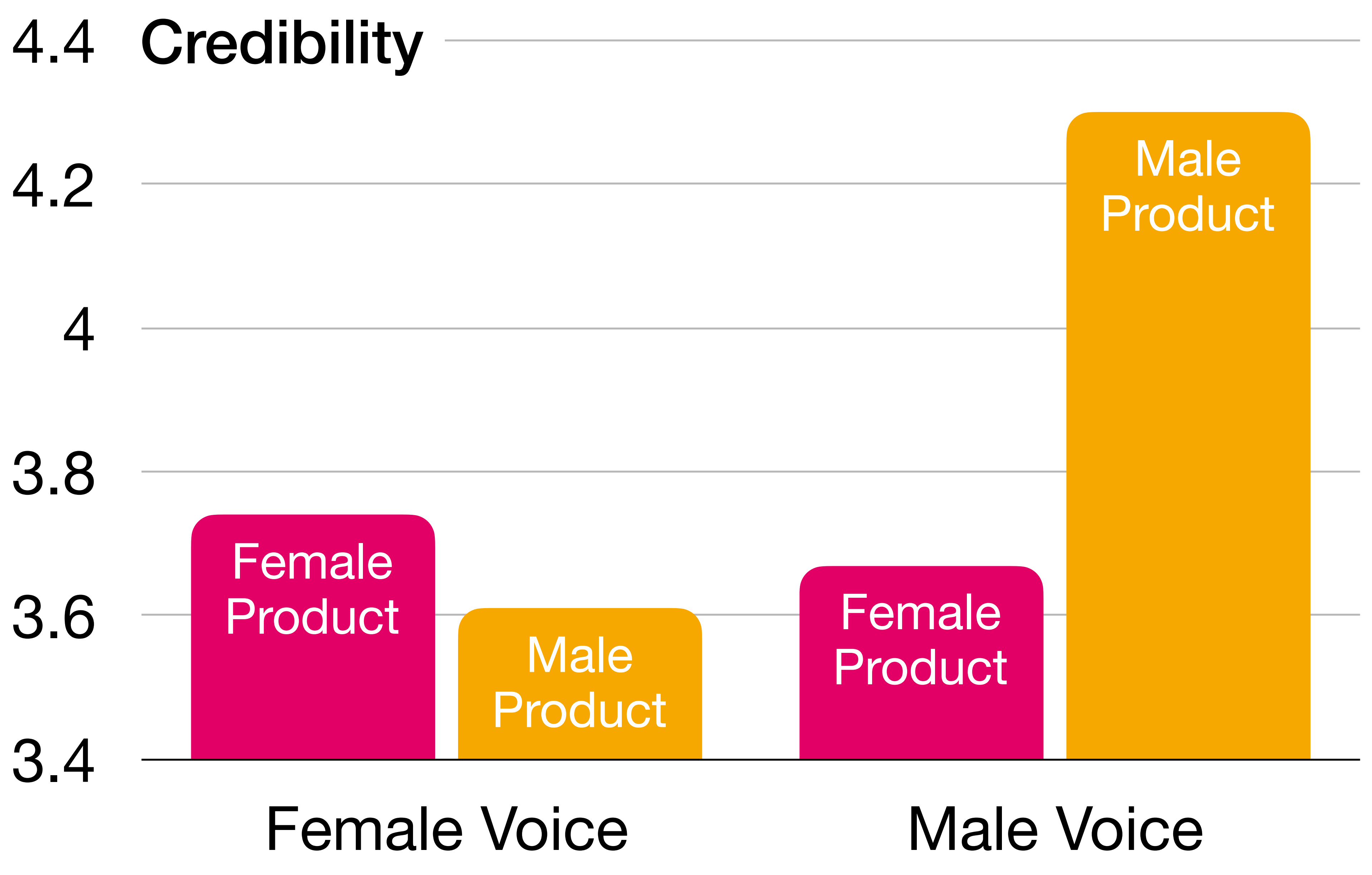
- Multiple voices more persuasive, even when participants were briefed about the use of computer voices beforehand

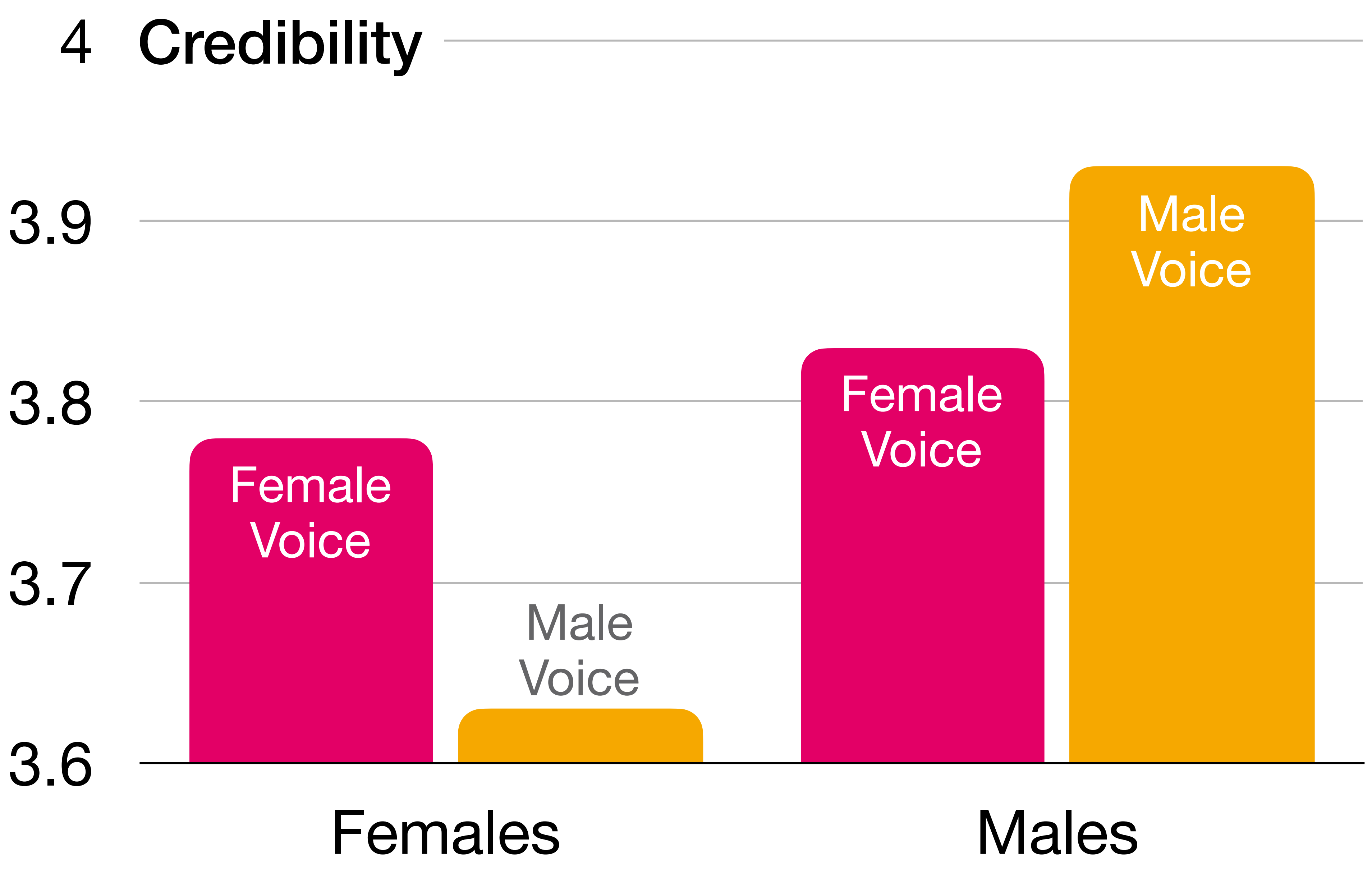
- Does not work with multiple fonts, for example



The Media Equation: Credibility

- Product reviews are most credible when gender of product, voice, and user match





Audio Input: Noise

- Use raw audio data to trigger events or control system
- Examples
  - Clap to turn on an interactive room
  - Monitor noise level to switch between remote sites displayed in multi-party video conferences
  - Localize current speaker to pan video conferencing camera
  - Monitor noise level to adjust phone ringer volume
- Fairly simple, used in commercial applications (and devices) today

Audio Input: Melodies

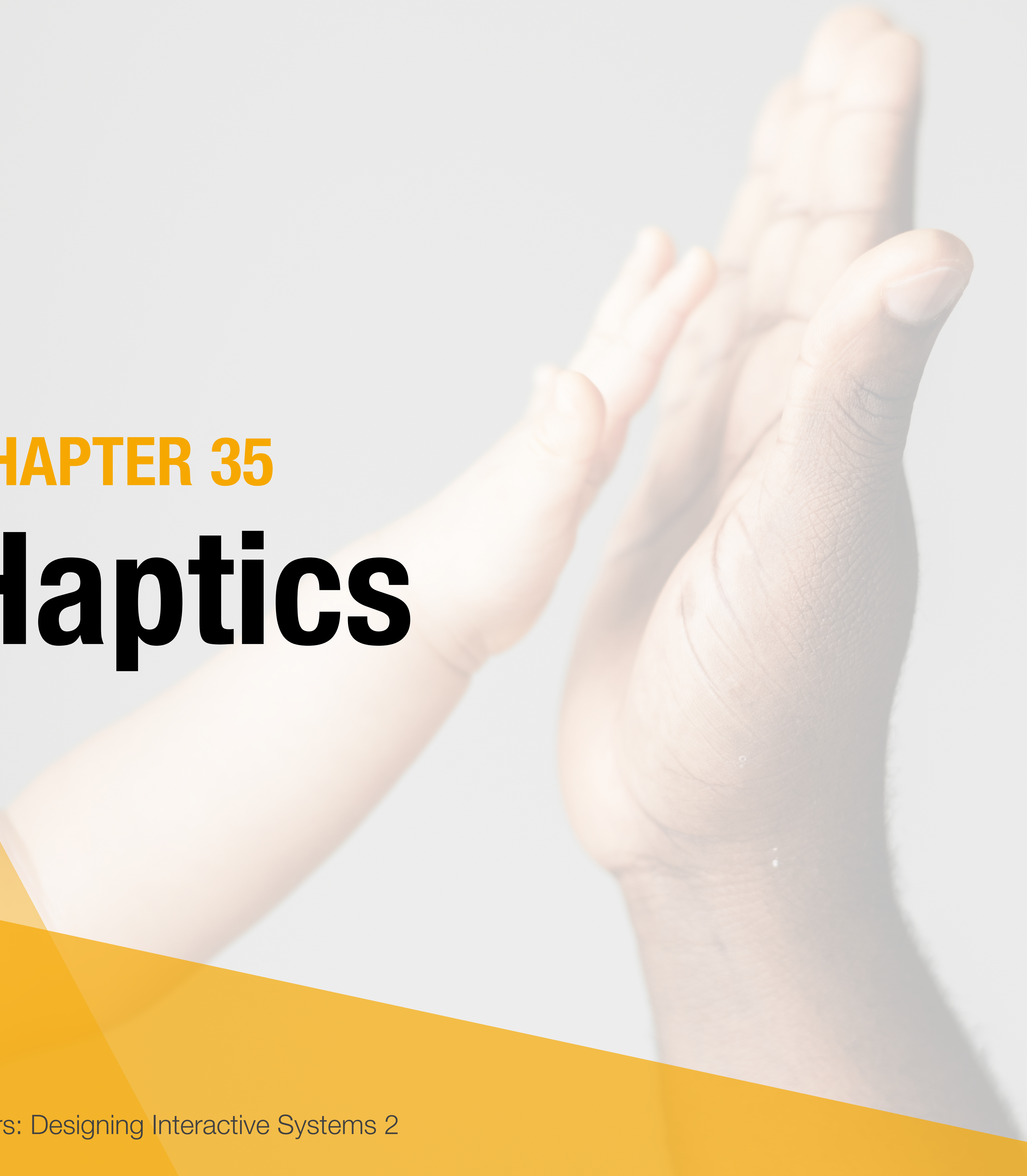
- **Data**: Musical Input
  - Record music
  - Synthetic (MIDI) or real (audio) data
- **Search**: Query By Humming (QbH)
  - Given a database of music, let users hum a melody to find the corresponding piece in the database
  - Difficult algorithmic problem

Speech Input: Problems and Challenges

- Speaker dependency (accent, intonation, stress, volume,…)
- Vocabulary dependency
- Background noise very critical
- Detection precision is high, but getting the semantics out of a sentence still an issue
- Many syntax combinations for same semantics
- Continuous speech
  - Humans disambiguate blurred word boundaries and multiple semantics
  - "How to wreck a nice beach"
- Higher cognitive load
  - Hand-eye coordination can happen in parallel to planning and problem solving
  - Speaking while thinking is more difficult

CHAPTER 35

Haptics



Do You Know How It Feels?

- Hardness
- Height maps
- Temperature
- Damping
- Friction
- Flexibility

Touch is Special

- Bidirectional
- Socially intentional-committing, invasive
- Gestural-expressive (functional and emotional signals)
- Many parameters: force, texture, temperature, moisture,…
- Poor absolute but high relative resolution
- Touch to do, probe, poke, fidget, communicate, verify, enjoy, connect,…
- Inhibitions: dirty, painful, forbidden, too intimate,…

Main Types of Haptic Interfaces

Cutaneous stimuli

- On the skin, i.e. tactile
- E.g., heat, pressure, vibration, slip, pain

Kinesthetic stimuli

- Bodily movements
- Detected in muscles, tendons, and joints
- E.g., limb position/motion/force

RWTHAACHEN UNIVERSITY

Haptic Output

- **Advantages**
  - Realistic
  - Intimate
  - Eyes-free
  - Needs no screen space
- **Disadvantages**
  - Limited resolution
  - Intimate
  - Unexpected

Designing Interactive Systems 2

Lecture 12: Multimodality II

Prof. Dr. Jan Borchers
Media Computing Group
RWTH Aachen University

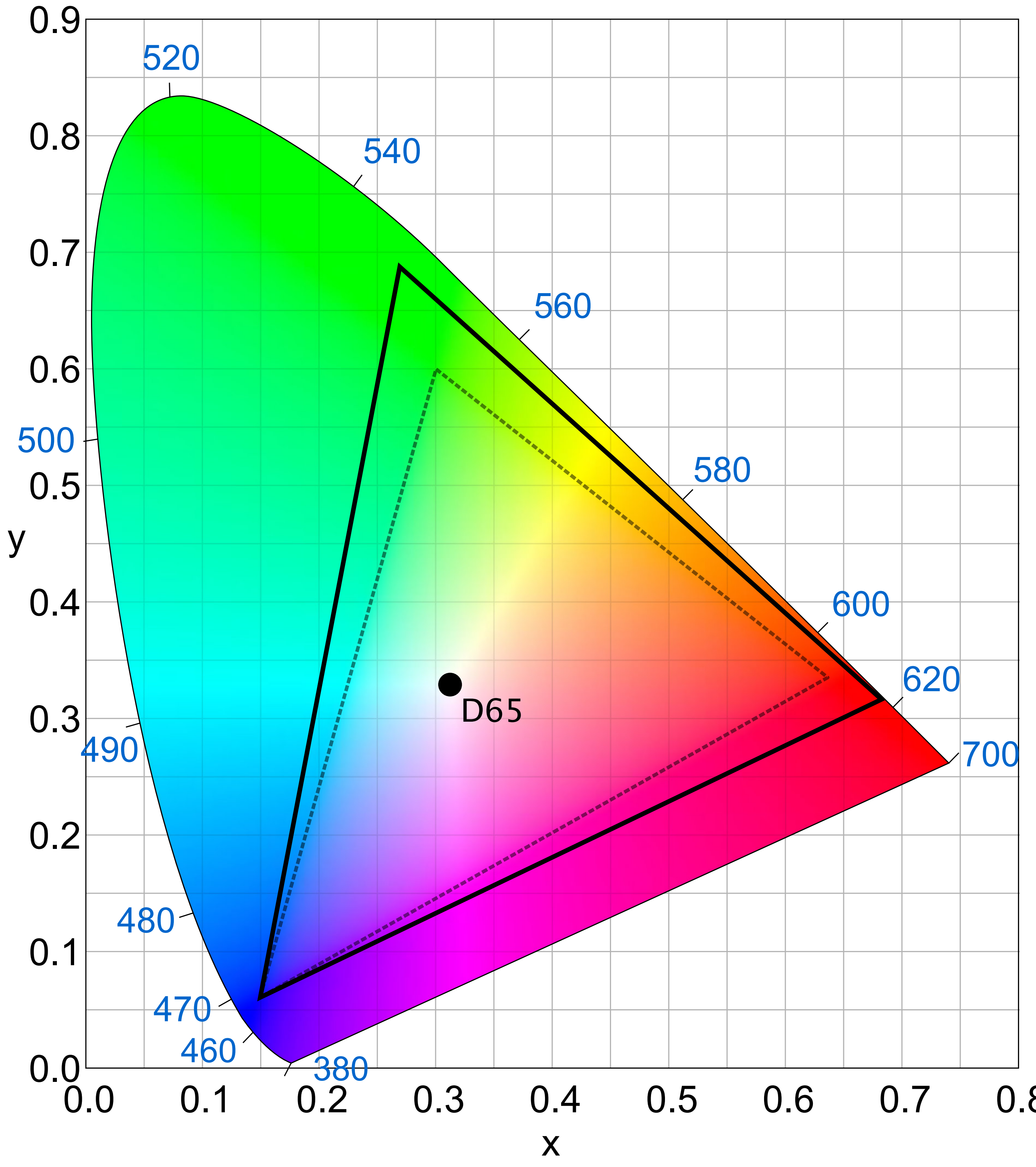
hci.rwth-aachen.de/dis2

CHAPTER 35

Vision

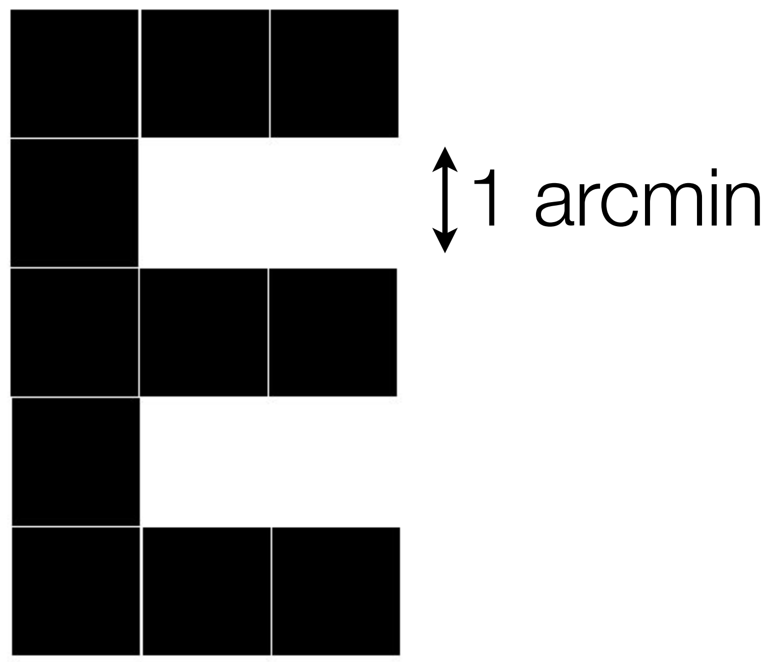
Visual Output: Display Characteristics

- Physical dimensions (diagonal, depth)
- Resolution
- Brightness (luminance in *cd/m²* = *nits* for displays, *(ANSI) lumen* for projectors)
  - Current laptop display: 500 nits;
    projections require 250 (indoor) to 500 (outdoor) *ANSI lumen/m²*
- Contrast (luminance ratio)
- Glare (workplace safety, glossy vs. matte displays)
- Color range and calibration (sRGB, P3)
- Refresh rates (Hz; difference to latency)
- Viewing angles, portability, reliability, power consumption, cost,…



Derived from: Source

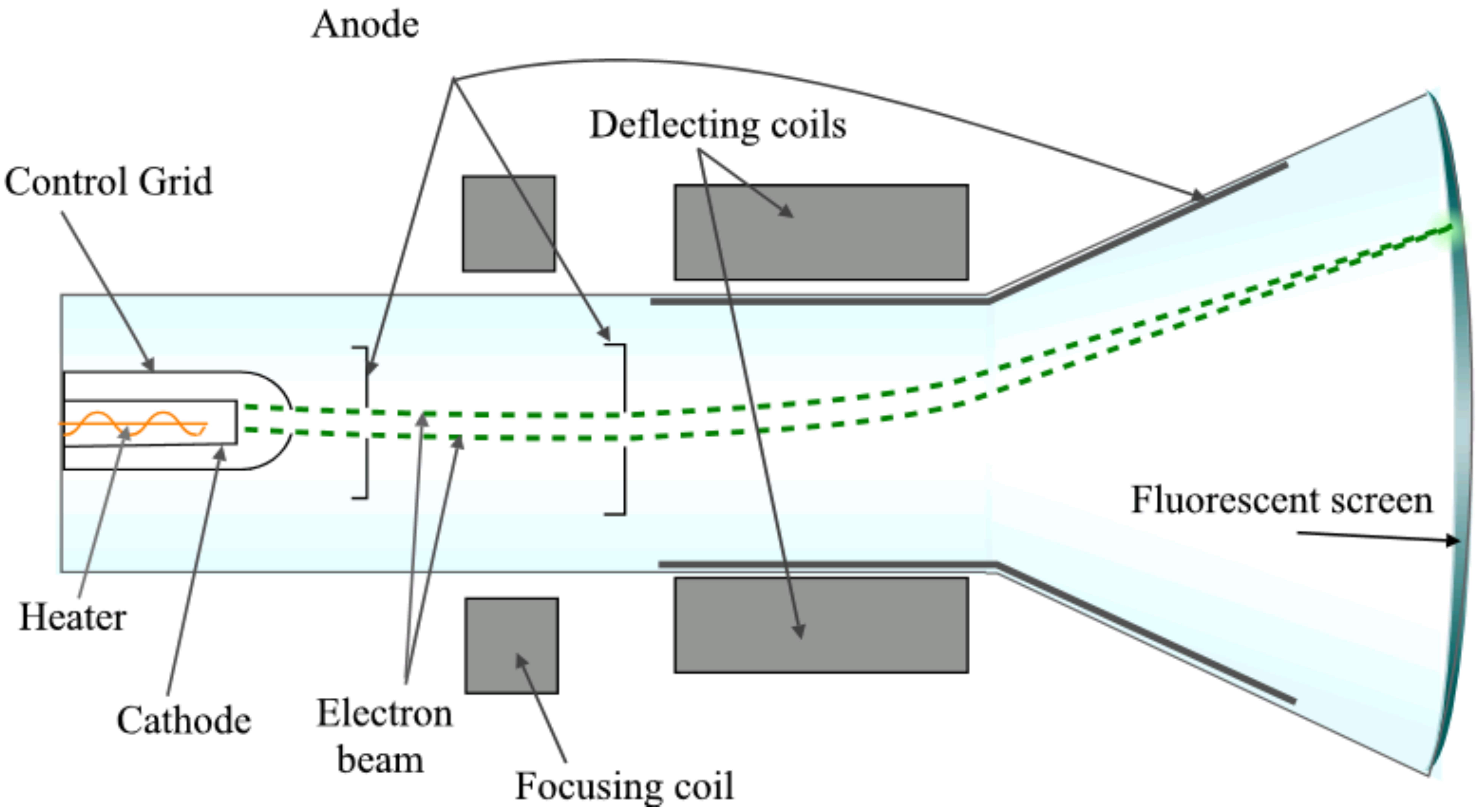
Visual Output: Resolution



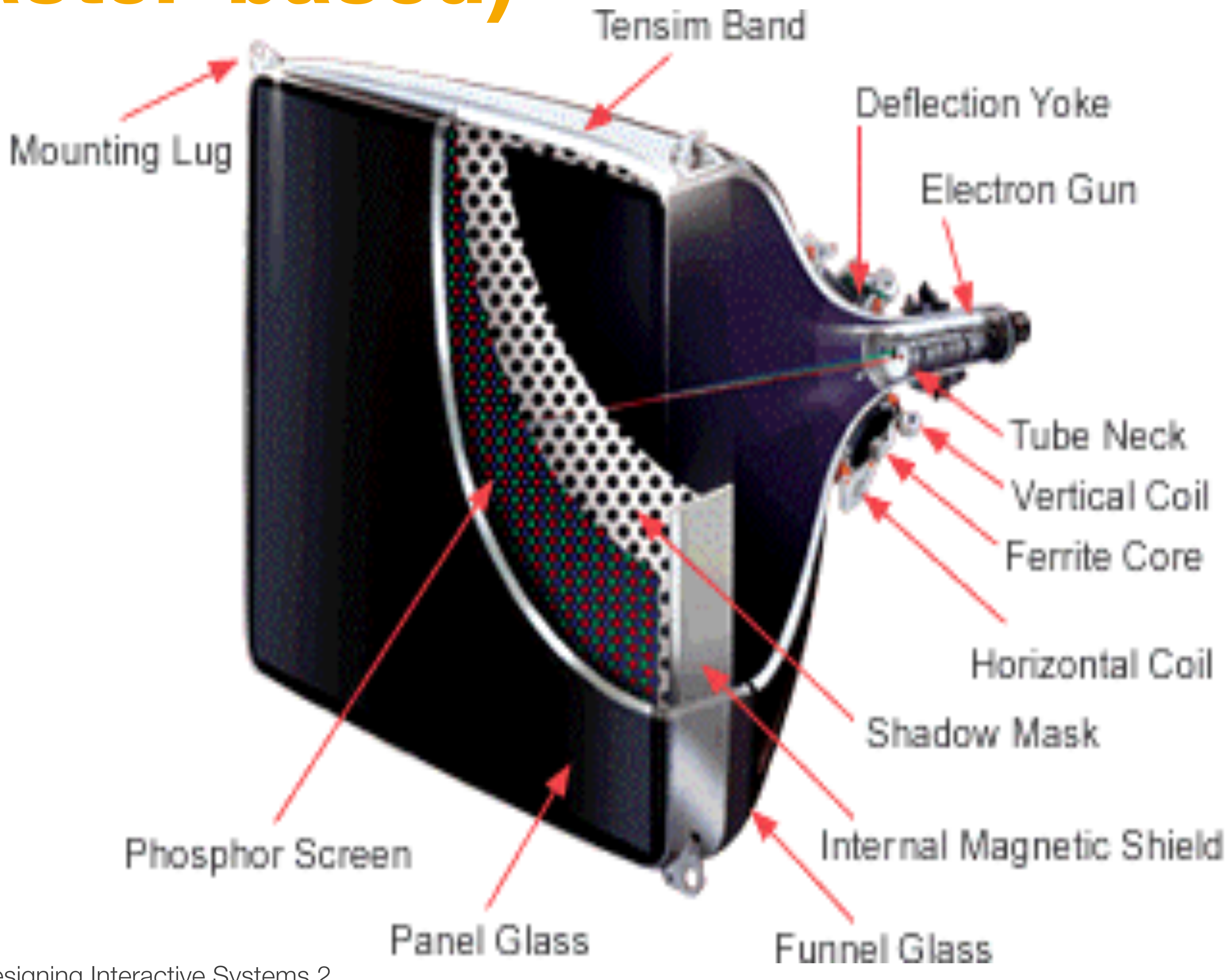
- Printer output resolution: **dots per inch (dpi)**
- Display output and digital media resolution: **pixels per inch (ppi)**
- Display resolution at display's typical viewing distance: **pixels per degree (ppd)** — *this matters*
- Resolution of the human eye (retina & lens):
  - "Normal": 6/6 (20/20, 1.0) vision = 1 arcminute = **60 ppd** = 1.75 mm @ 6 m
  - But: population avg. 6/4.5 (80 ppd), mid-twenties 6/4 (90 ppd), practical limit 6/3 (120 ppd), theoretical limit 6/2 (180 ppd) [Ohlsson 2005]
- Smartphones: viewing distance 30 cm ⇒ 60 ppd = 287 ppi (iPhone 11: 68 ppd, iPhone 11 Pro: 97 ppd)
- Laptops: viewing distance 60 cm ⇒ 60 ppd = 144 ppi (MacBook Pro 16" 2019: 97 ppd)
- TVs: viewing distance 3 m ⇒ 60 ppd = 29 ppi (FullHD 50" TV: 92 ppd)

Displays: CRTs (Vector-based)

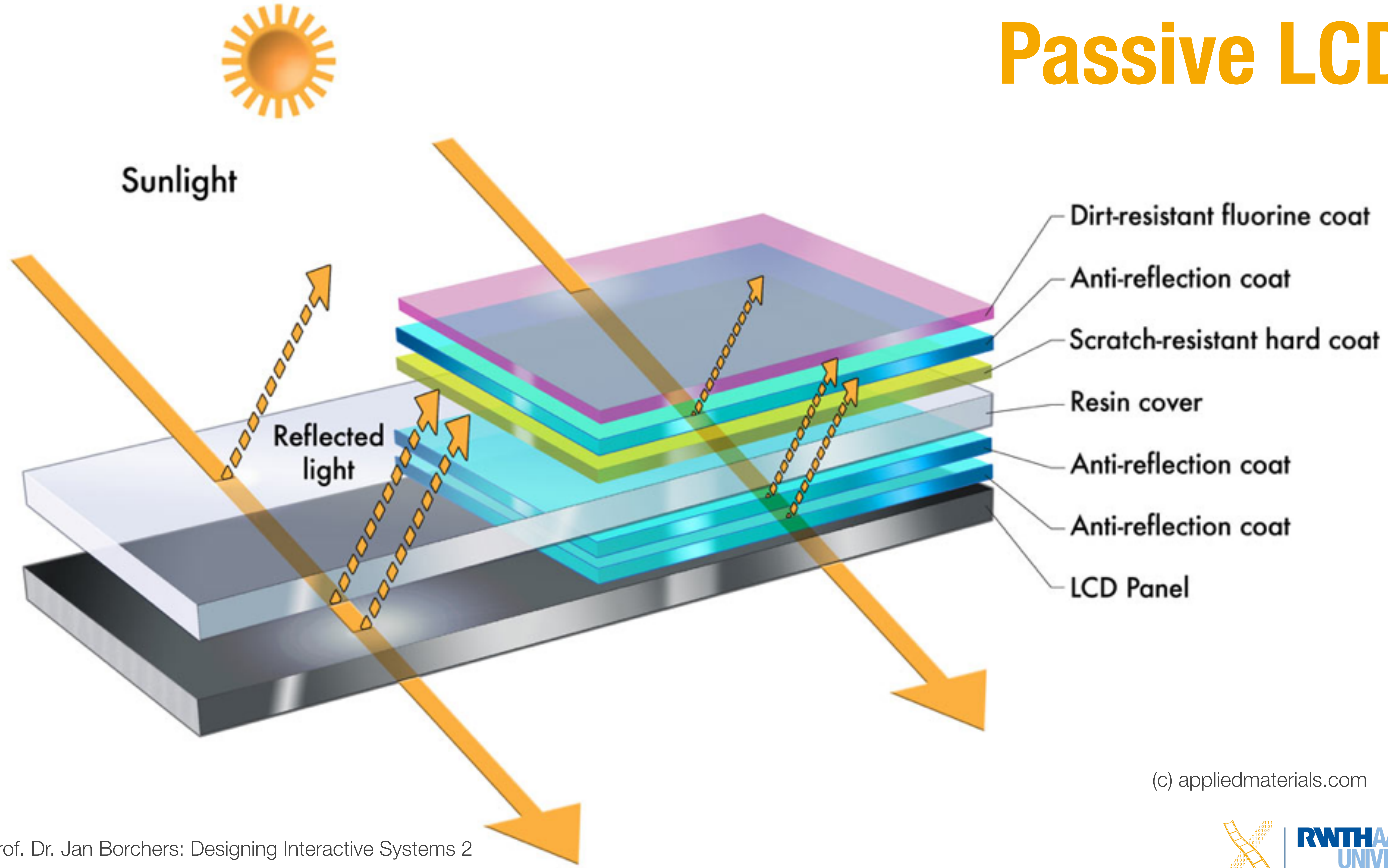
Displays: CRTs (Vector-based)



CRTs (Raster-based)

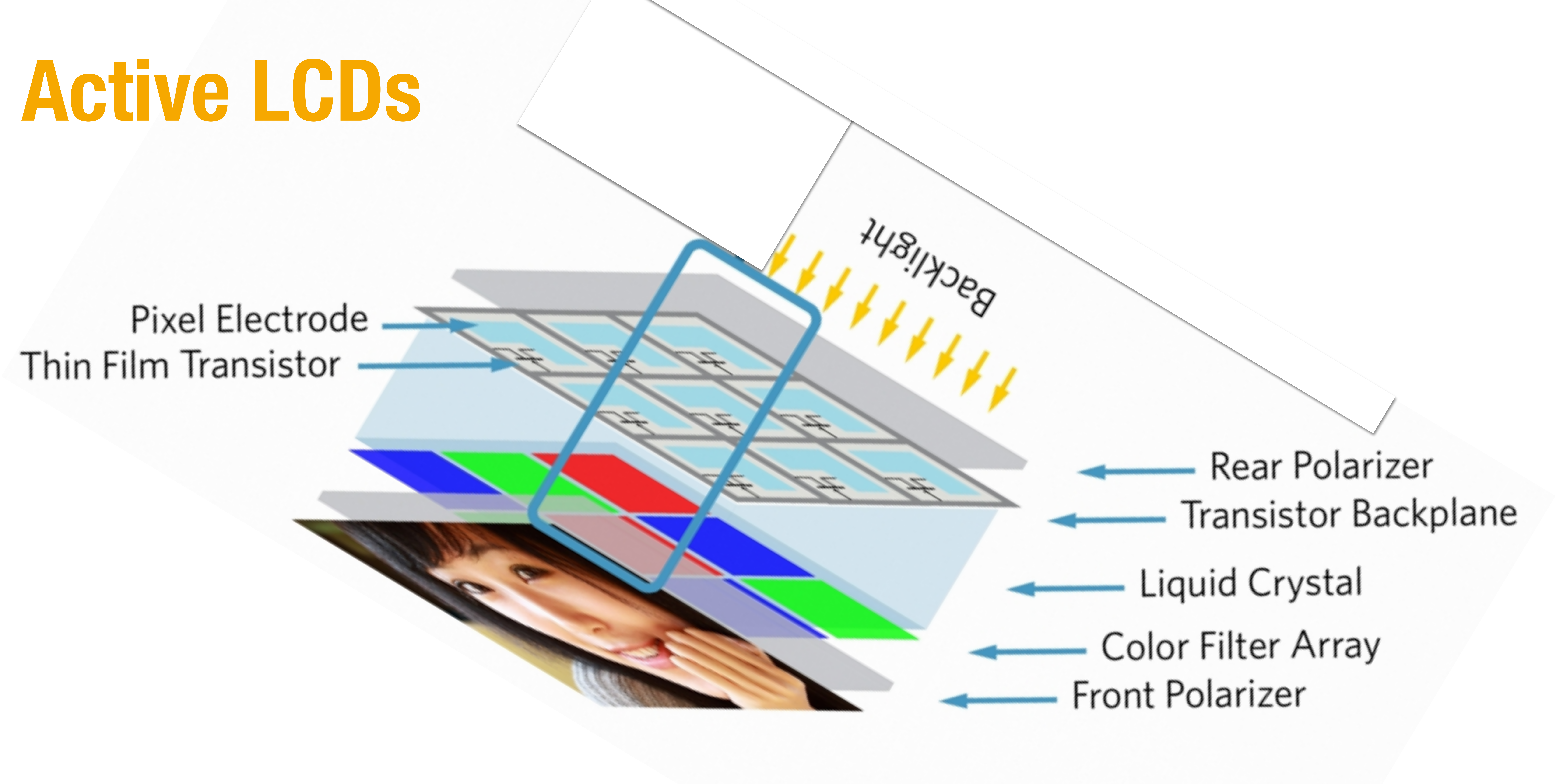


Passive LCDs



(c) appliedmaterials.com

Active LCDs



(c) appliedmaterials.com

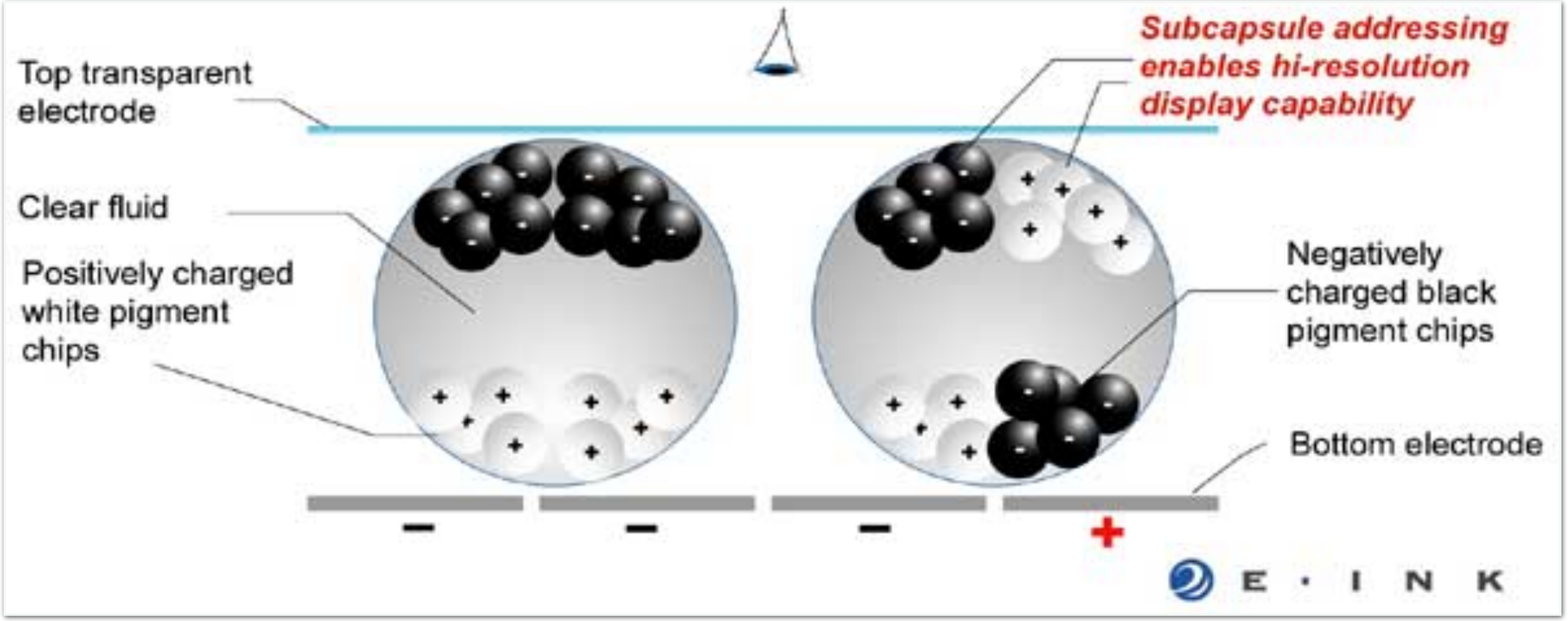
LED Displays

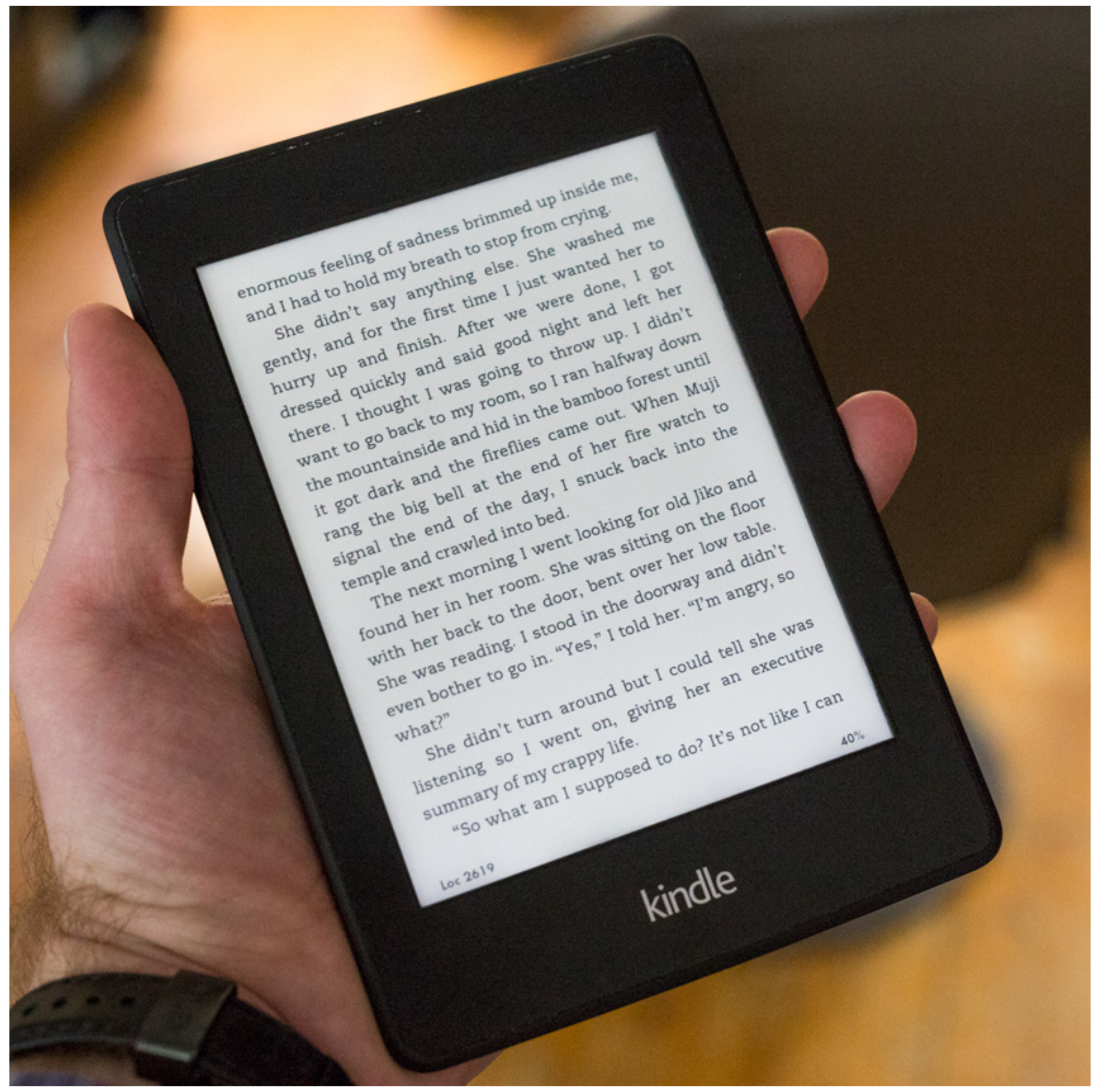
RWTH AACHEN UNIVERSITY

OLED Displays

Projection Displays

Electronic Ink






SEIKO
10
08

VR & AR Displays

3D Displays

An Interactive 360° 3D Display

Andrew Jones Ian McDowall* Hideshi Yamada†
Mark Bolas‡ Paul Debevec

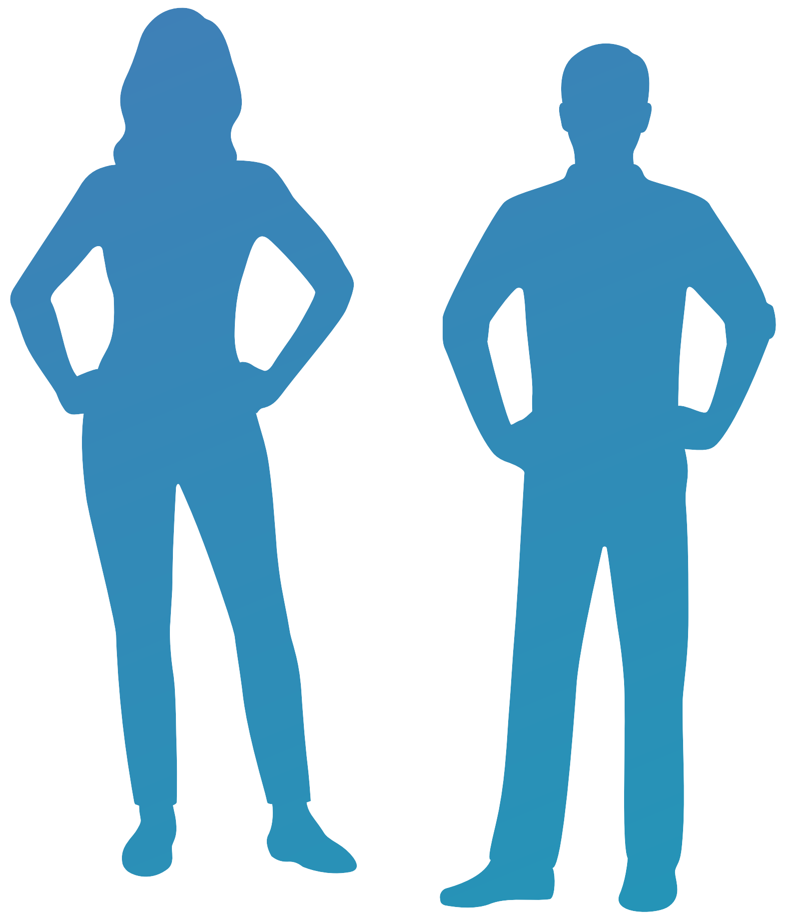
USC Institute for Creative Technologies Fakespace Labs*
Sony Corporation† USC School of Cinematic Arts‡

http://gl.ict.usc.edu/Research/3DDisplay

Visual Input

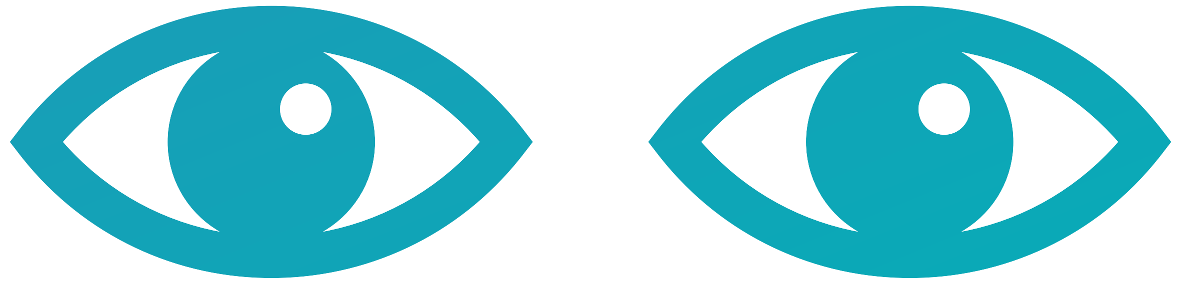
- What can the computer “see”?
  - Brightness
  - Color
  - Edges
  - Shapes
  - Objects

Visual Input: Tracking



Body Movements & Posture

Hand Gestures



Gaze Tracking

Visual Tracking

Visual Tracking

1

Capture frame and apply basic filtering to enhance image

2

Image / motion **segmentation** to identify regions, and features

3

Object classification to identify parts in the image, match with world model, and labeling them

4

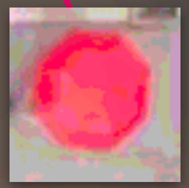
Tracking
Where do the objects of interest move along?

Update model with new position and features

5

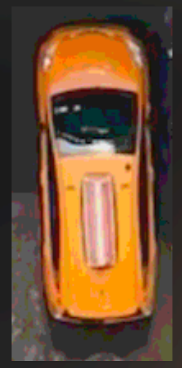
Interpretation
React to observed changes, continue with next frame

Call callback in your application



Model from last frame

Umbrella 

 Car

 Tree

Visual Tracking

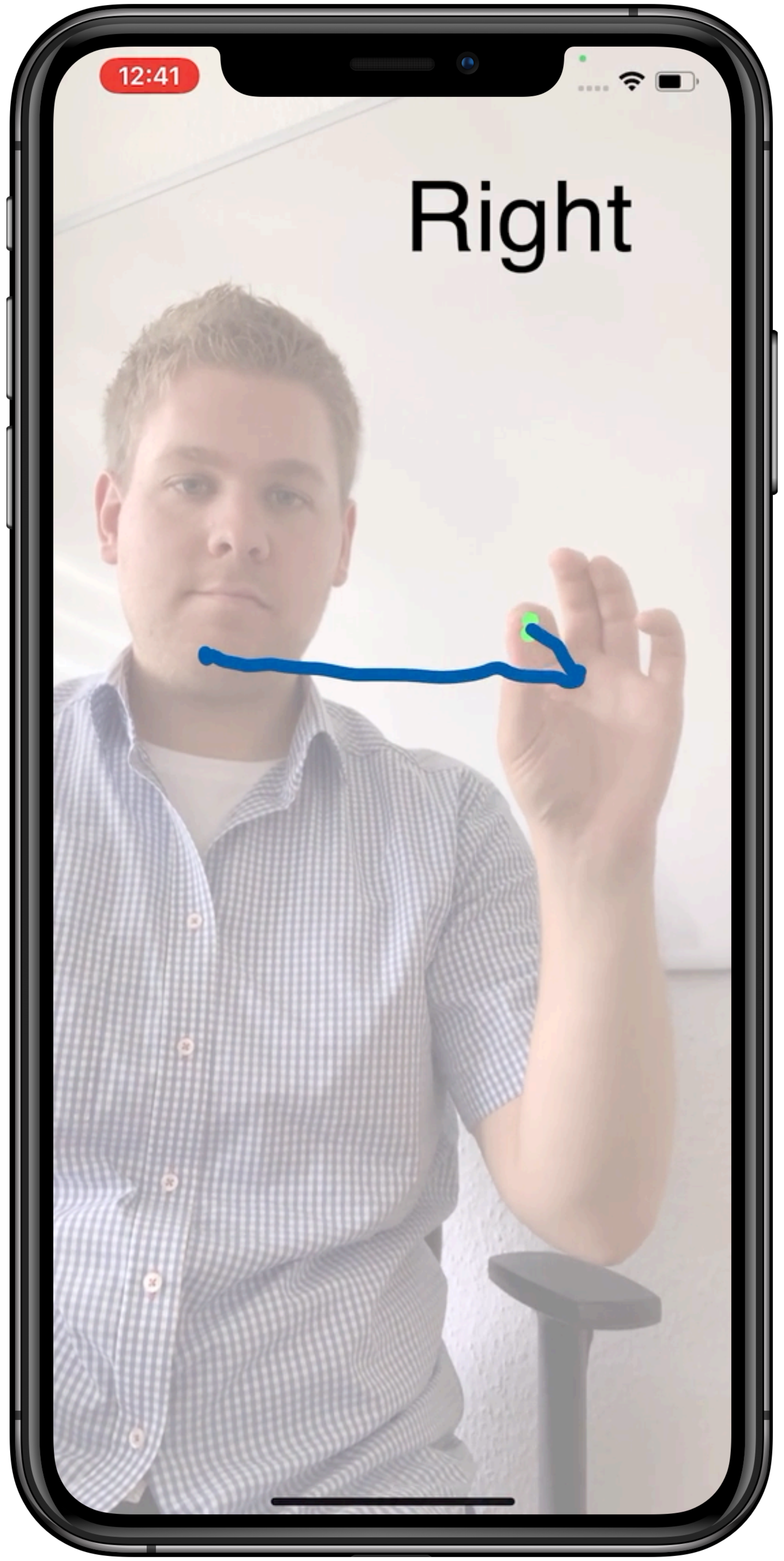
Example: Mid-Air Gestures

1: Camera
Camera captures frame

2+3: Use a CV Framework
Specify a tracking request
and receive prediction and position

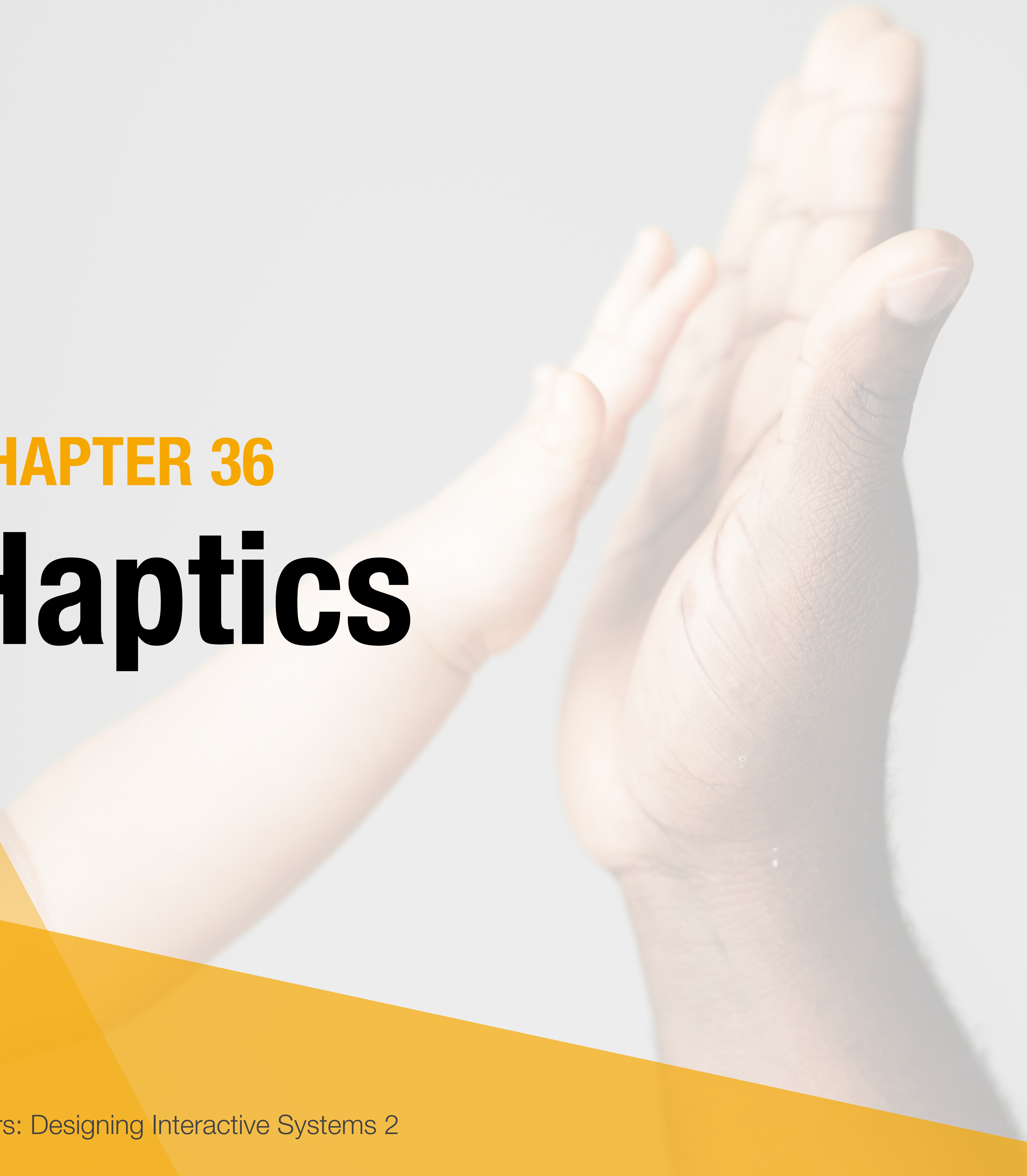
4: Create Semantics
Process and interpret the incoming data stream

5: Update UI
E.g. drawing are reacting to a gesture



CHAPTER 36

Haptics



Exercise: What makes these feel different?

- Hardness
- Height maps
- Temperature
- Damping
- Friction
- Flexibility

Touch is Special

- Bidirectional
- Socially intentional-committing, invasive
- Gestural-expressive (functional and emotional signals)
- Many parameters: force, texture, temperature, moisture,…
- Poor absolute but high relative resolution
- Touch to do, probe, poke, fidget, communicate, verify, enjoy, connect,…
- Inhibitions: dirty, painful, forbidden, too intimate,…

Touching Provides Information

- Assess properties, verify completion, monitor activity/progress, building mental models for invisible systems, judging others

- Can help to deal with complex interfaces
  - Distinguish buttons, dents offer cues, physical interfaces
  - Muscle memory

Main Types of Haptic Interfaces

Cutaneous stimuli

- On the skin, i.e. tactile
- E.g., heat, pressure, vibration, slip, pain

Kinesthetic stimuli

- Bodily movements
- Detected in muscles, tendons, and joints
- E.g., limb position/motion/force

Haptic Output

- **Advantages**
  - Realistic
  - Intimate
  - Eyes-free
  - Needs no screen space
- **Disadvantages**
  - Limited resolution
  - Intimate
  - Unexpected

Haptics Hardware Challenges

- Compare to input devices
- Device cost
- Size
- Weight (for performance)
- Robustness
- Bandwidth
- Technology-centered view
- Innovation needed

Research Domains

- UI design: Application interface design, concept prototyping
- Psychology: perception and cognition studies, user experimentation and analysis, biomechanics and kinesiology
- CS: building multisensory displays and controls, realtime software architectures, rendering algorithms, physical system modeling

Example: Tactons (Lorna Brown 2007)

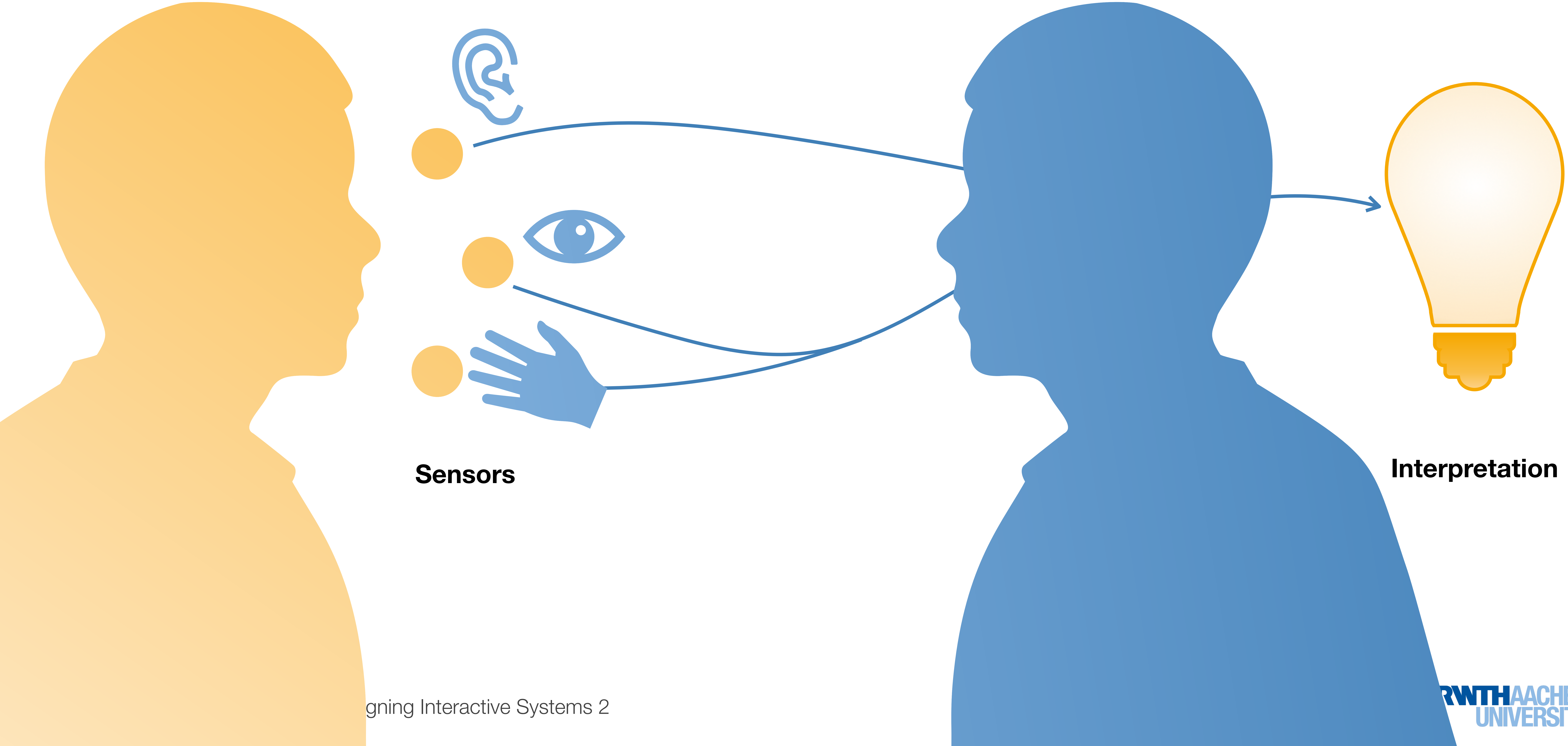
- Tactile Icons = subset of Haptic Icons
- Similar to earcons, just using touch sense
- Encode multi-dimensional information by using an abstract mapping
- Parameters: spatial location, waveform (roughness), rhythm and intensity change over time
  - other parameters (e.g., just intensity) not that well suited
  - Resolution ≥ 3

CHAPTER 37

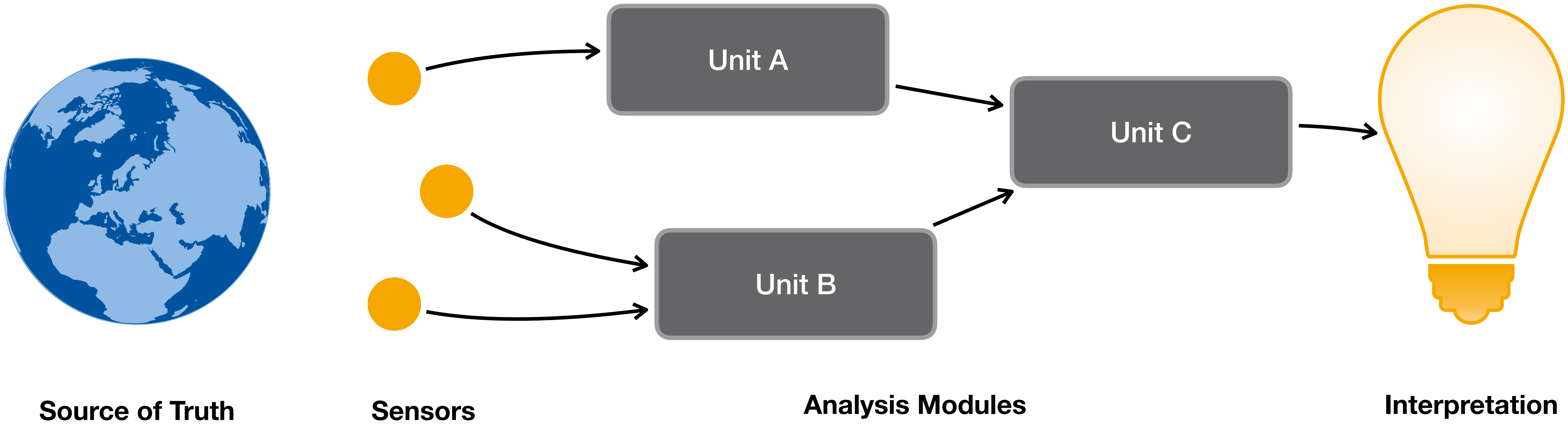
Creating Multimodal Interaction

How to **process** inputs from multiple modalities?

Fusion



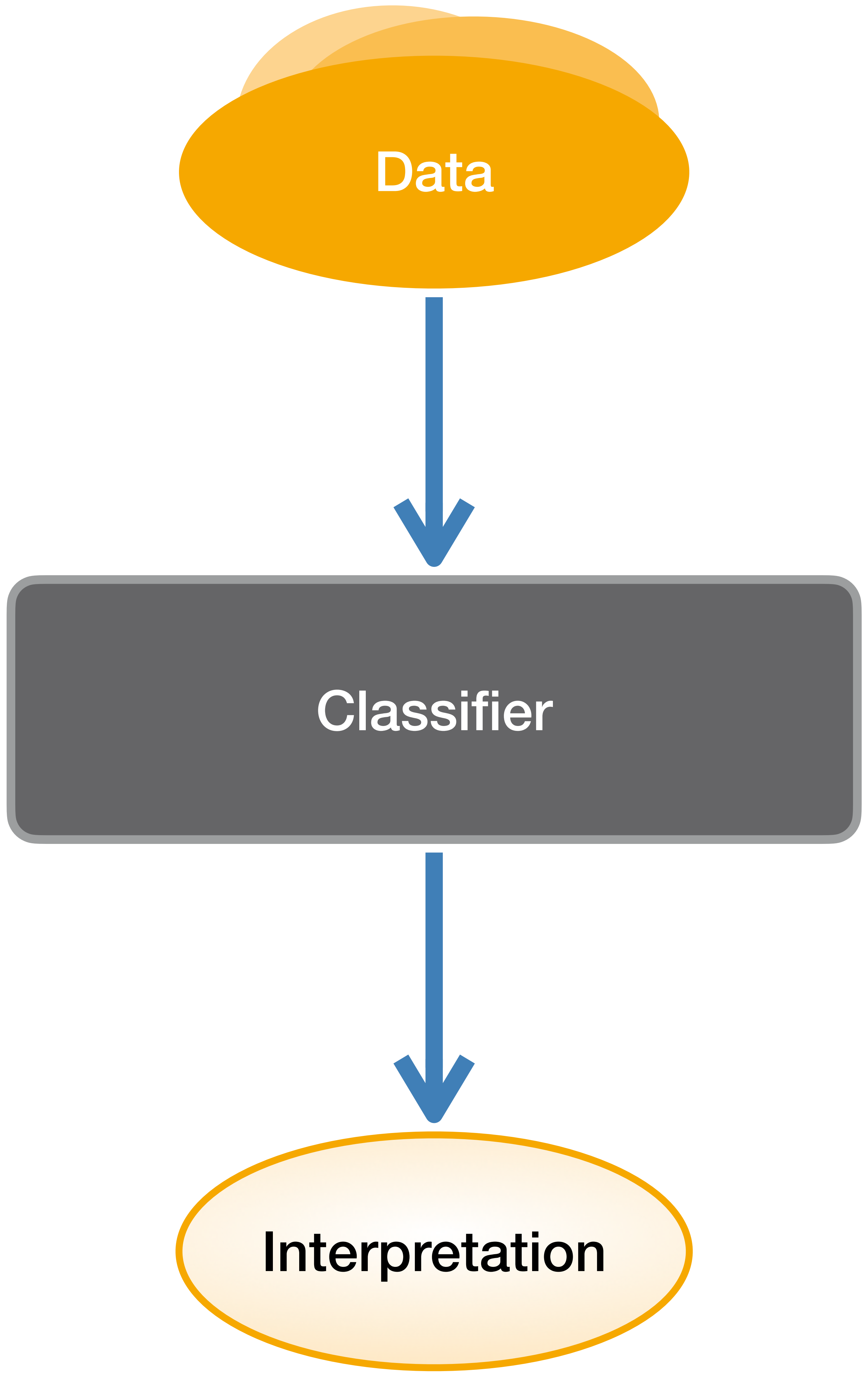
Fusion



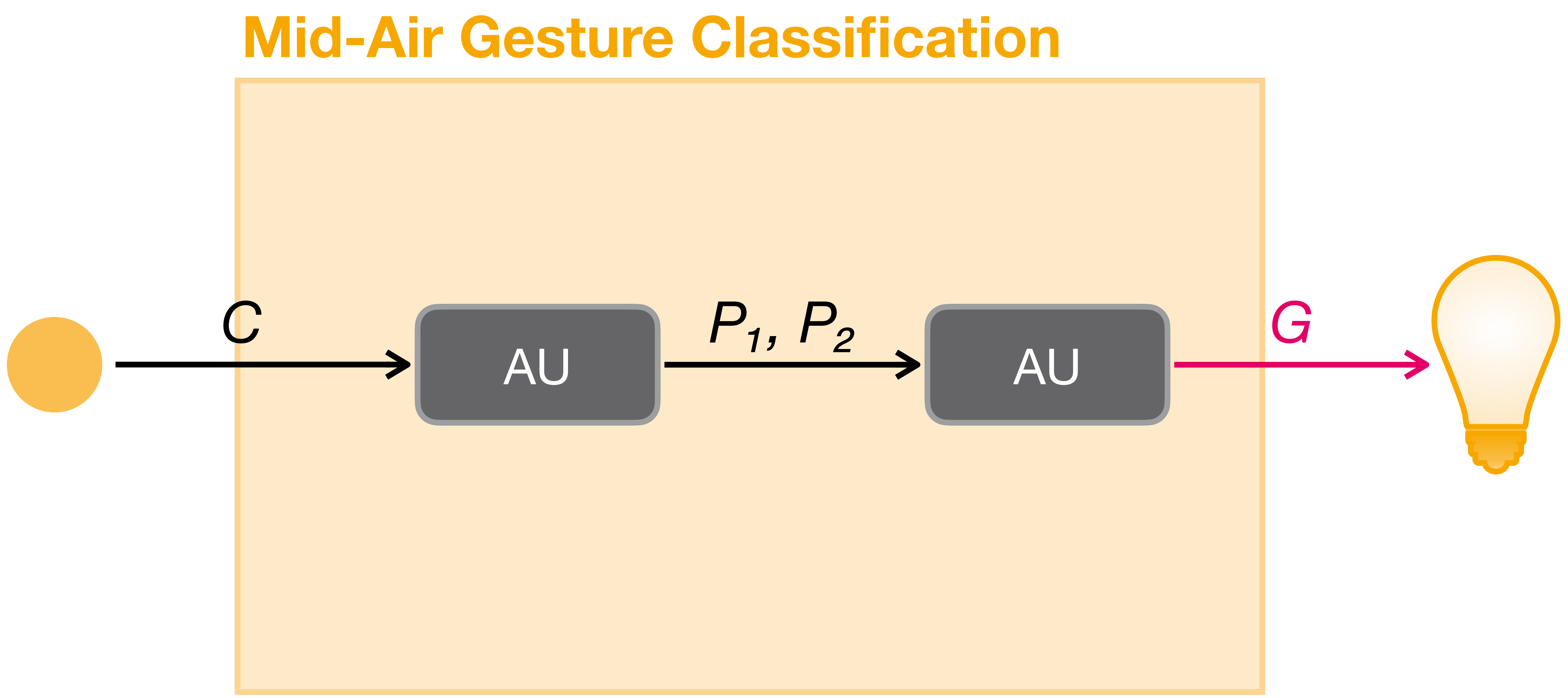
Unit A
Unit B
Unit C
Source of Truth
Sensors
Analysis Modules
Interpretation

Issues for Interpretation

- For a **single** sensor
  - Noise (sensor, channel, modality-specific)
  - Non-universality
- For **multiple** sensors
  - Ambiguity due to contradicting information
  - Different formats and sampling rates



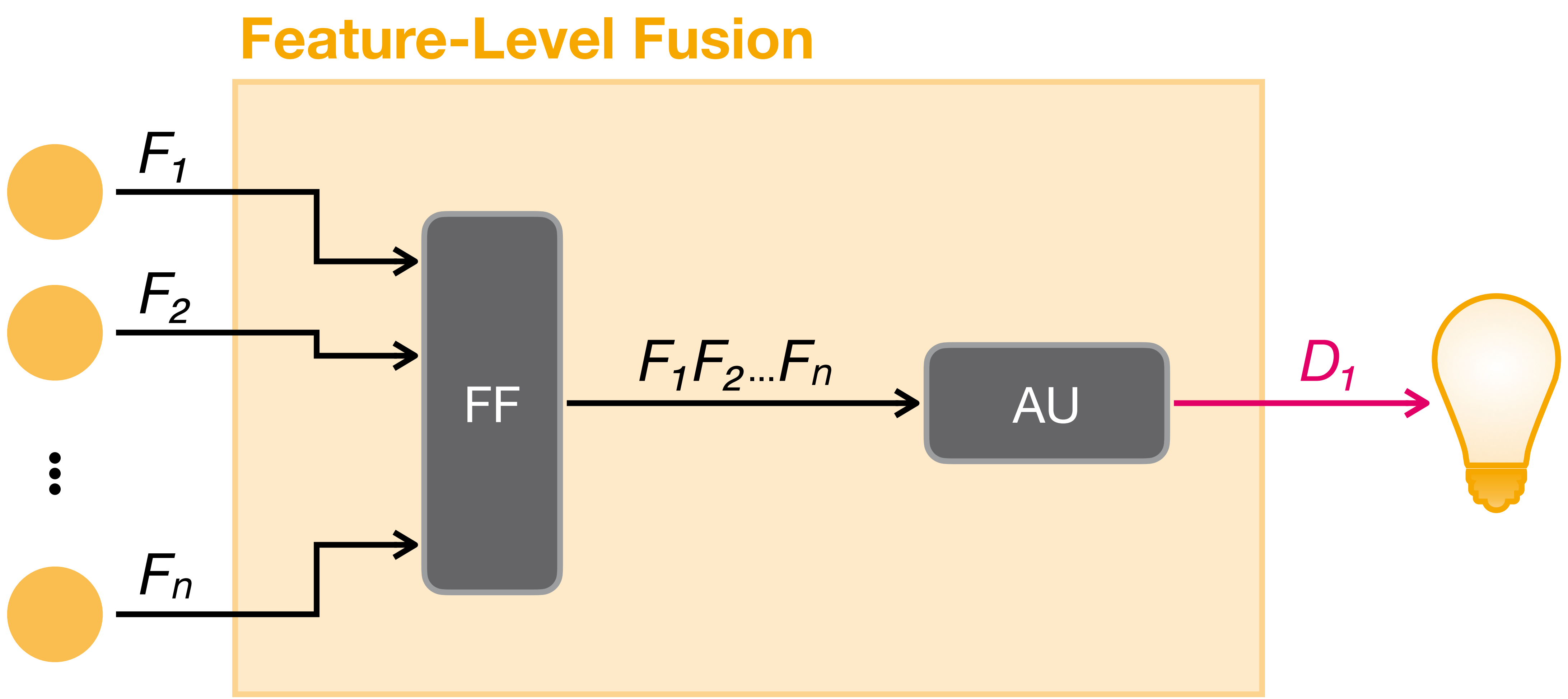
Example: Mid-Air Gestures



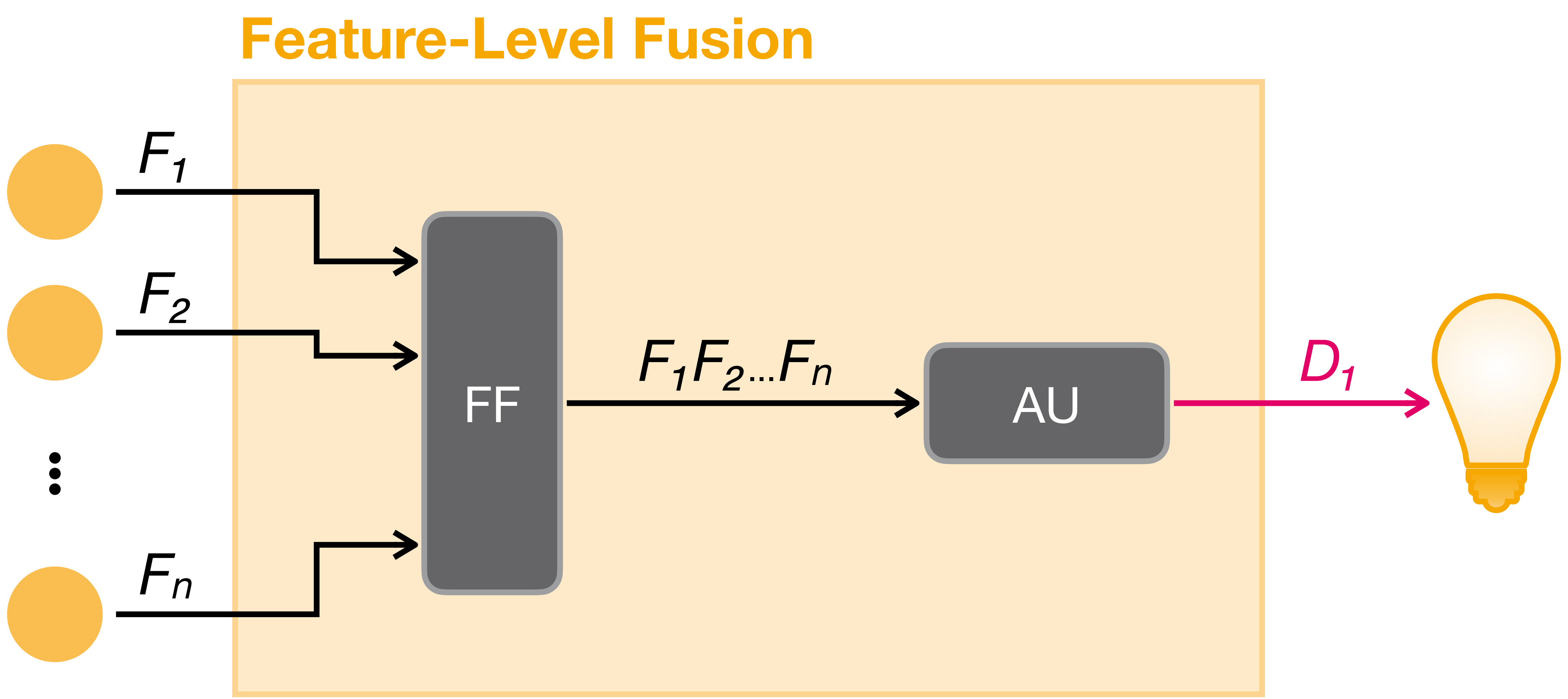
Fusion Levels

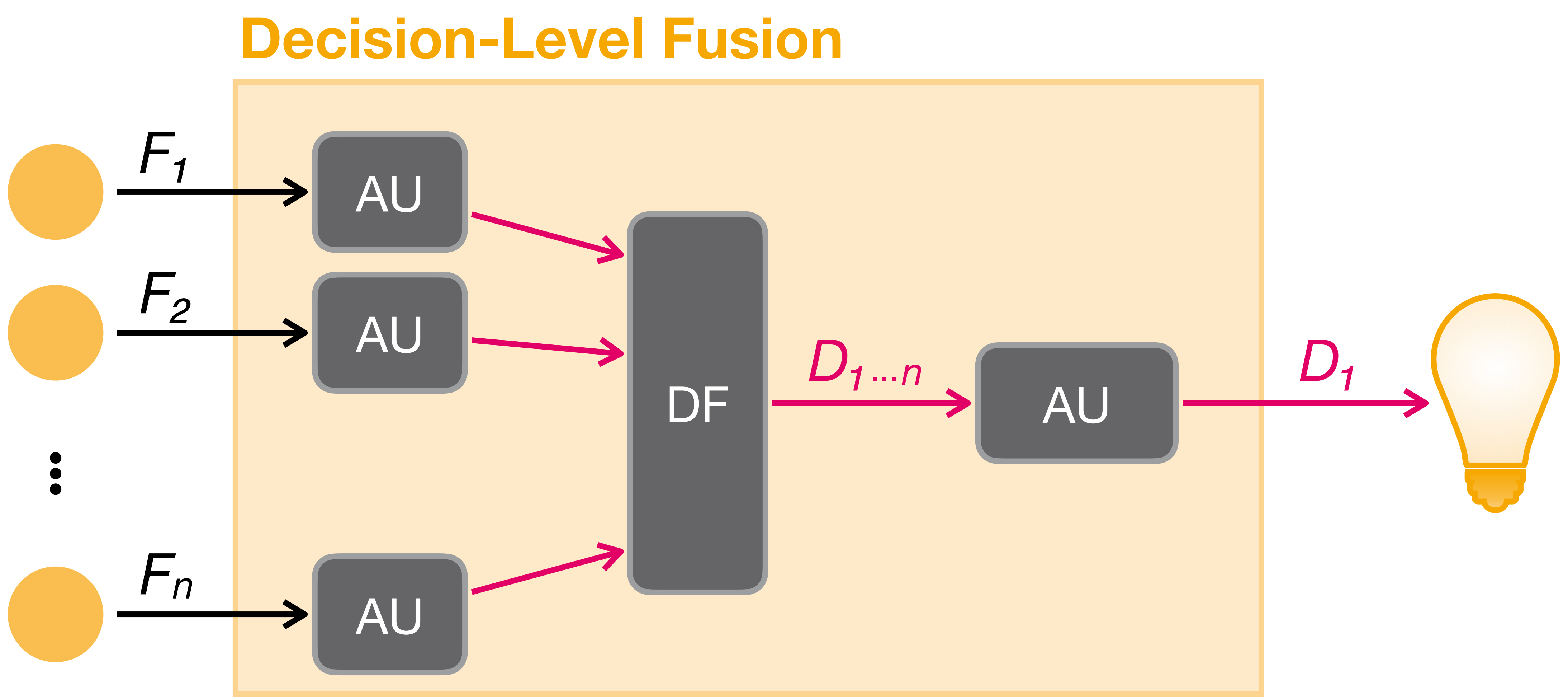
- How do we implement our multimodal system?
  - Which data from which modalities belongs together?
  - How can we increase the confidence of our decisions?
  - Performance
- Fusion can take place at different levels
  - At the feature level
  - At the decision level
  - A hybrid of both

Fusion Levels



Fusion Levels





Fusion Methods

Rule-based methods

Linear Weighted Fusion
Majority Voting Rule
...

Classification-based methods

Support Vector Machine
Bayesian Inference
...

Estimation-based methods

Kalman Filter
Particle Filter
...

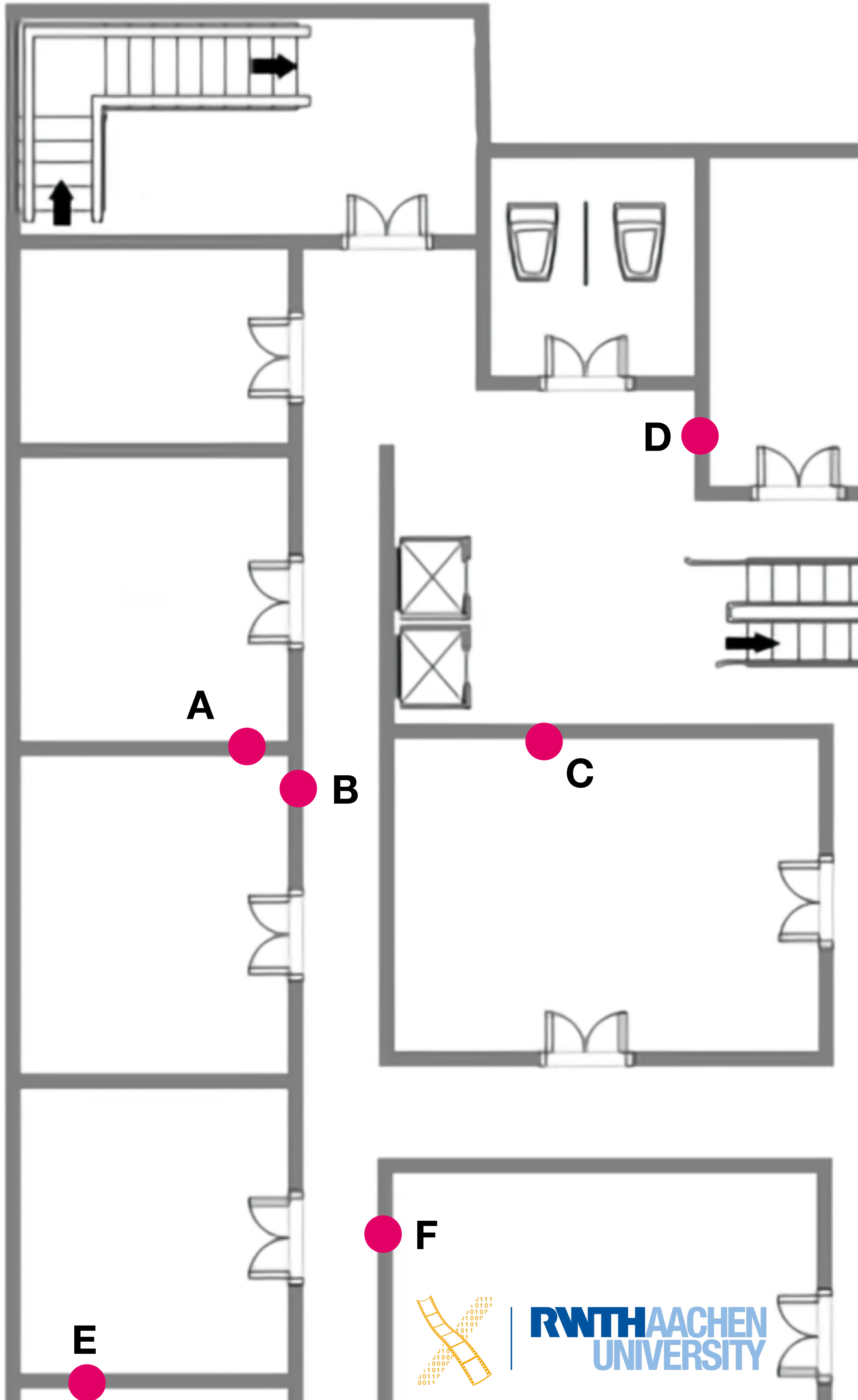
Example: Indoor Location with Beacons

- The signal strength of different Bluetooth beacons can be used for indoor triangulation
- Early fusion:
  The signal vector of multiple beacons can be used to derive a position

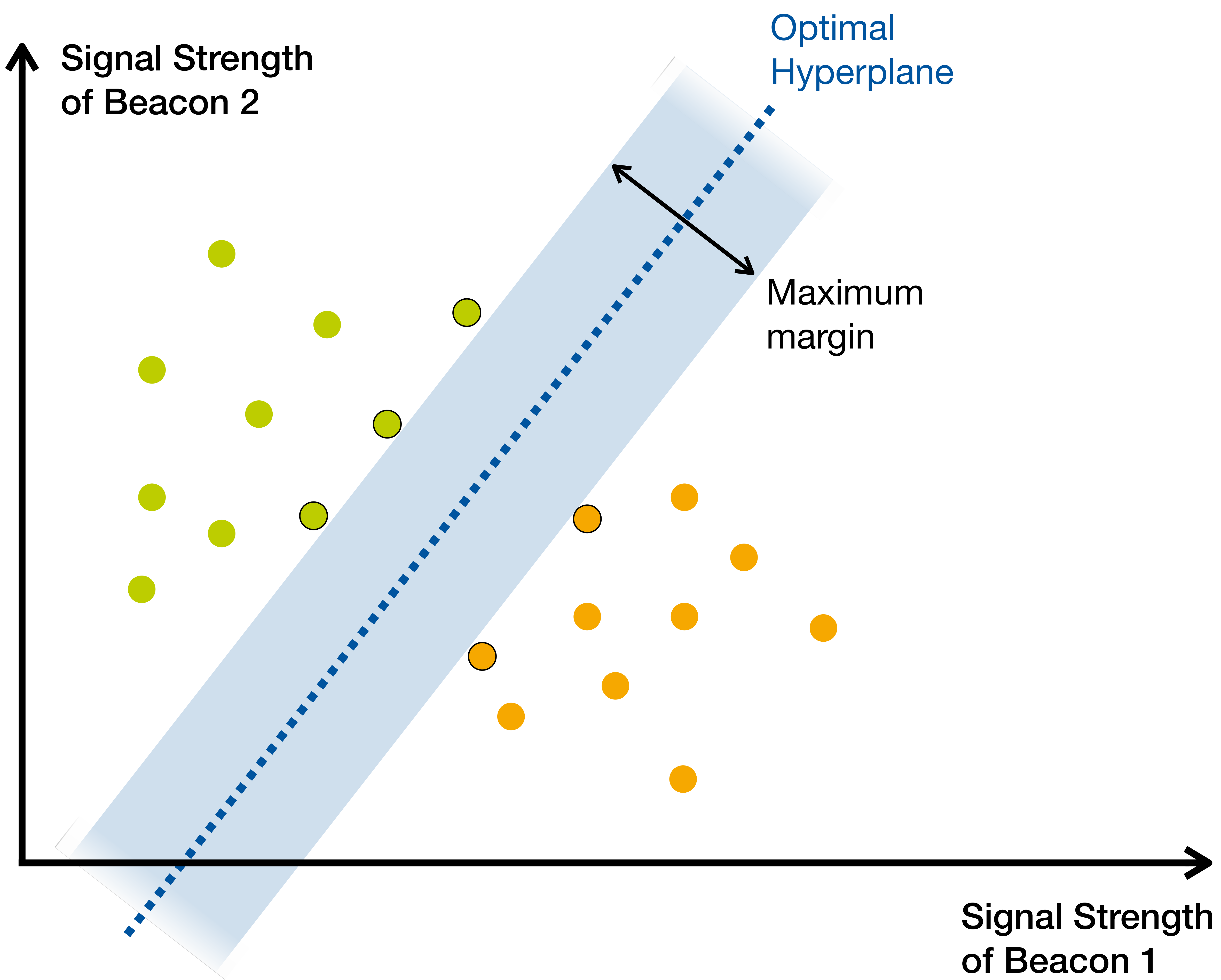
Indoor Location

- In which shop of the mall is the user when measuring these signal strengths?

- Data noise makes it difficult
  - Infrequent amount of beacons
  - Bluetooth signal strength is volatile



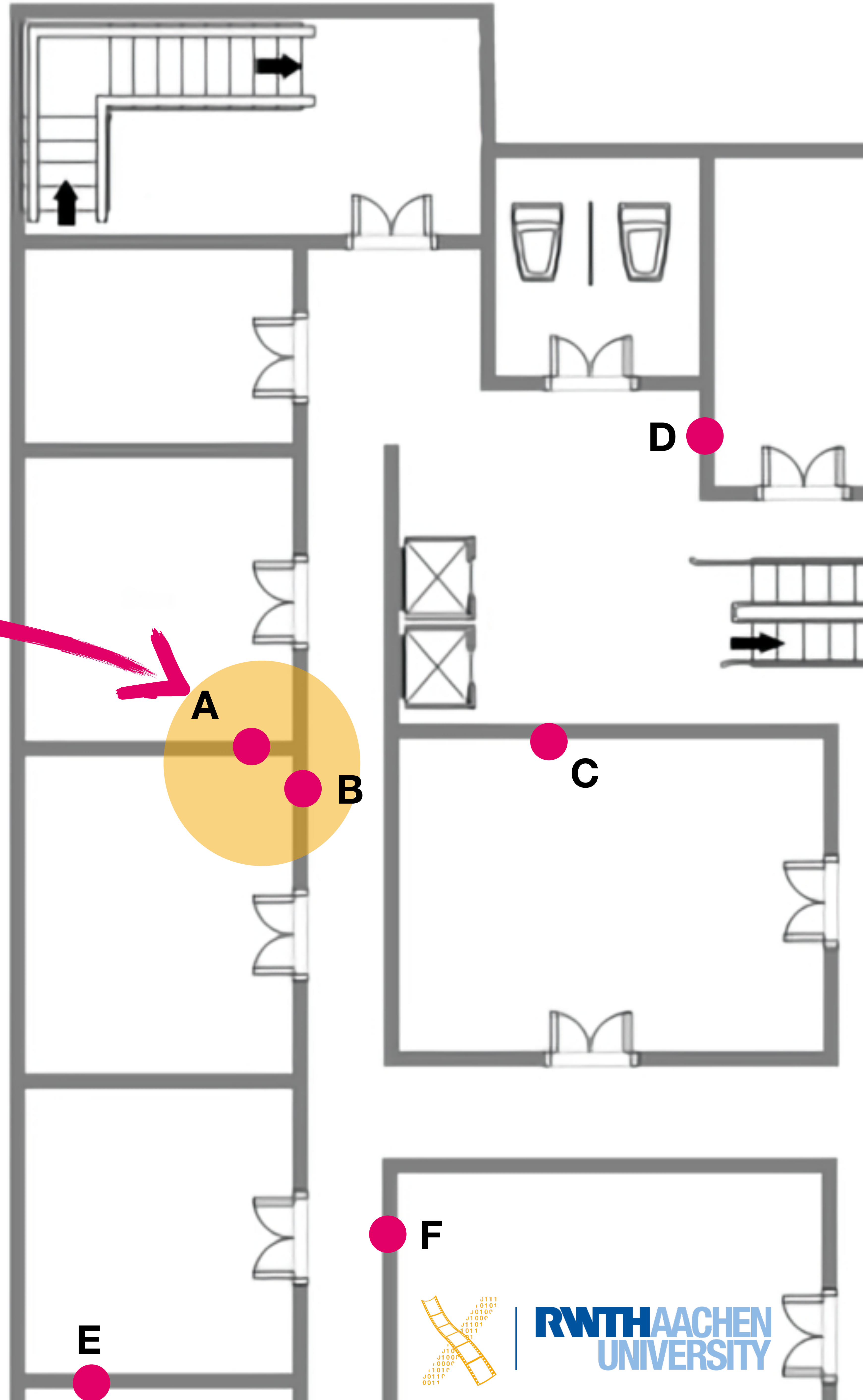
Example: Support Vector Machines



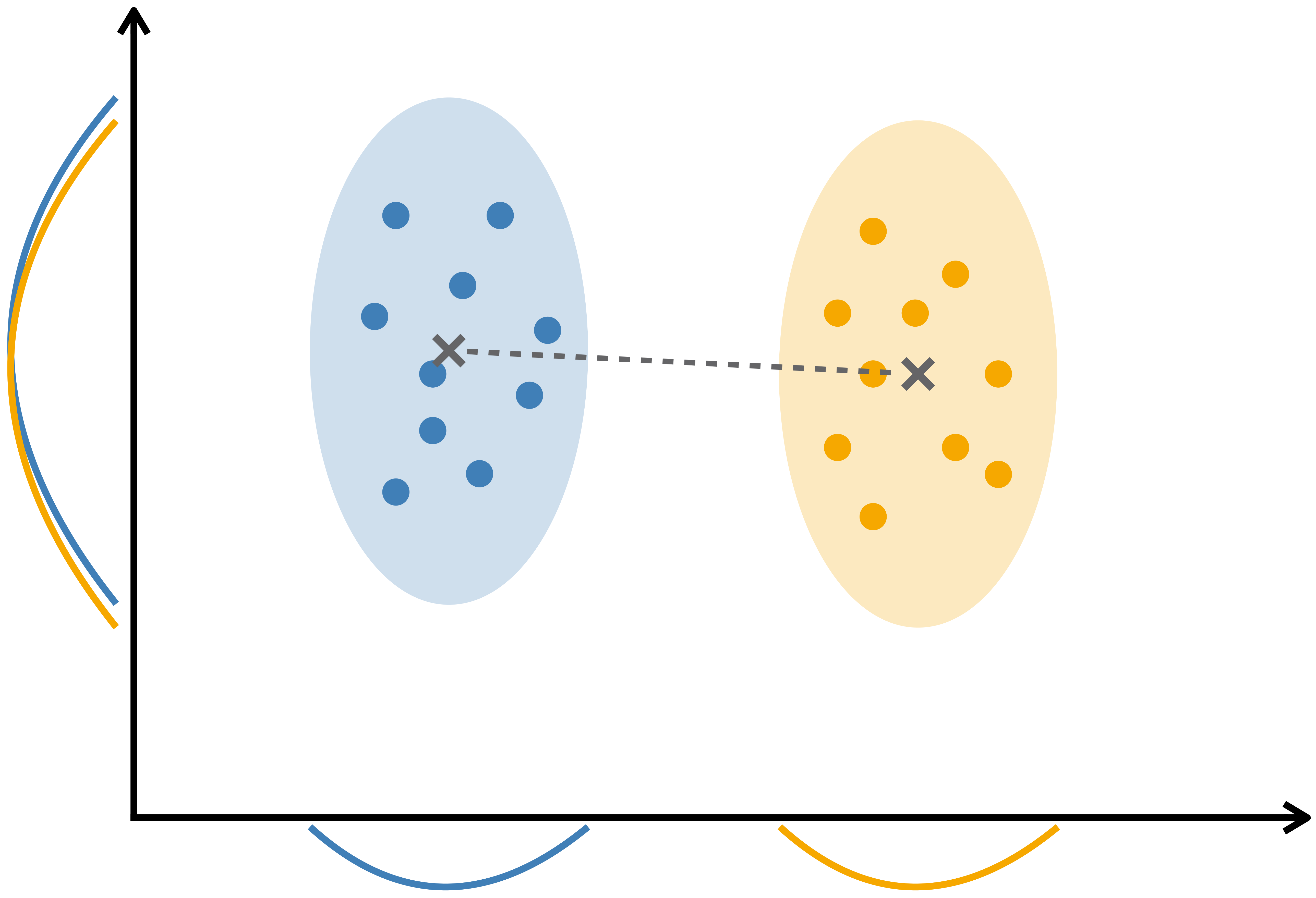
- SVMs are supervised ML models
- Based on the training data a hyperplane is determined that separates the data
- The support vectors are the data points that influence the position of the hyperplane

Shortening Data Vector

Probably **B** does not help that much to determine our location…

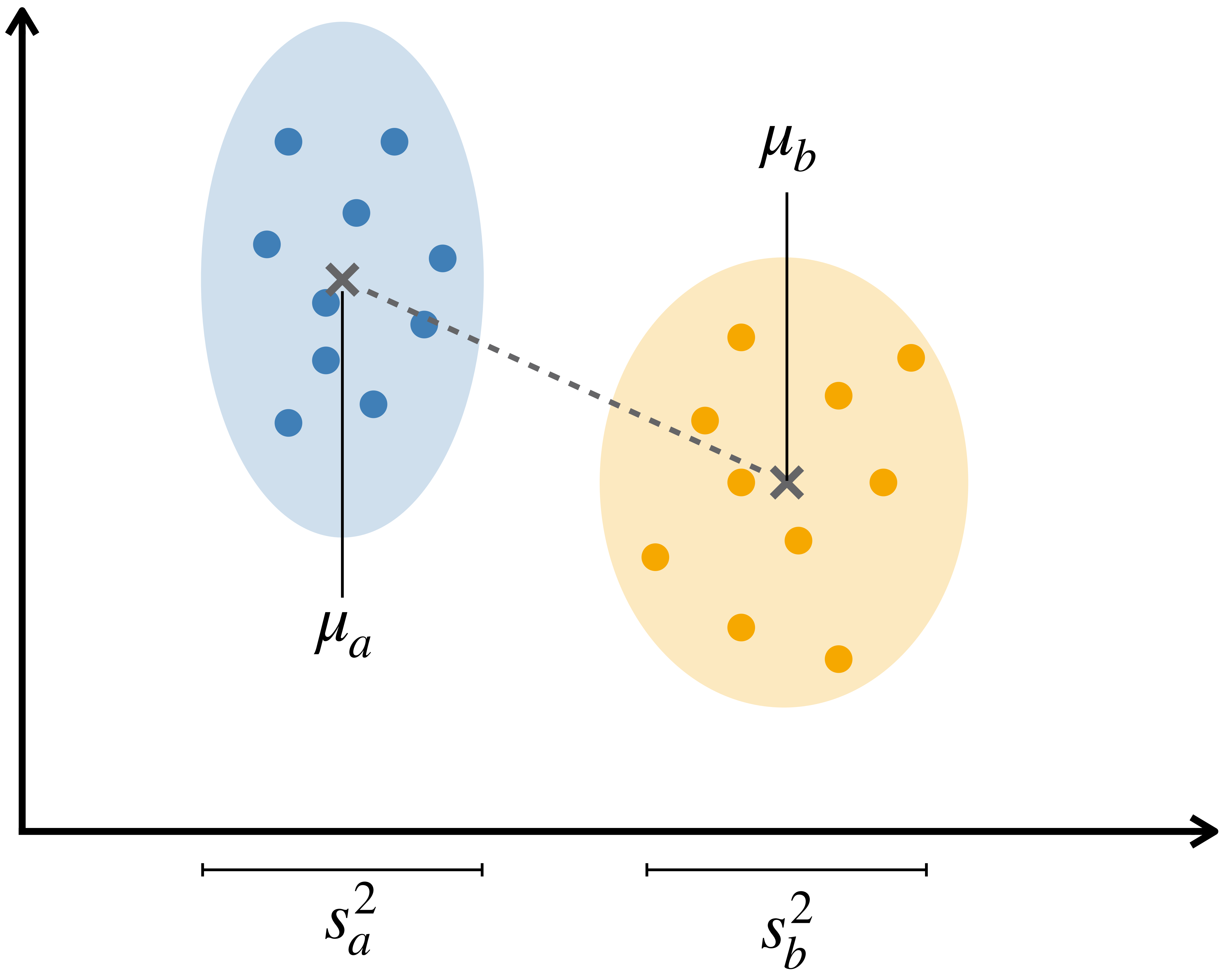


Example: Linear Discriminant Analysis



- LDA allows to reduce the feature vector length by finding a good projection axis
- Here, the x information is definitely more helpful to differentiate between the two categories
- But what happens if it is not that clear?

Example: Linear Discriminant Analysis

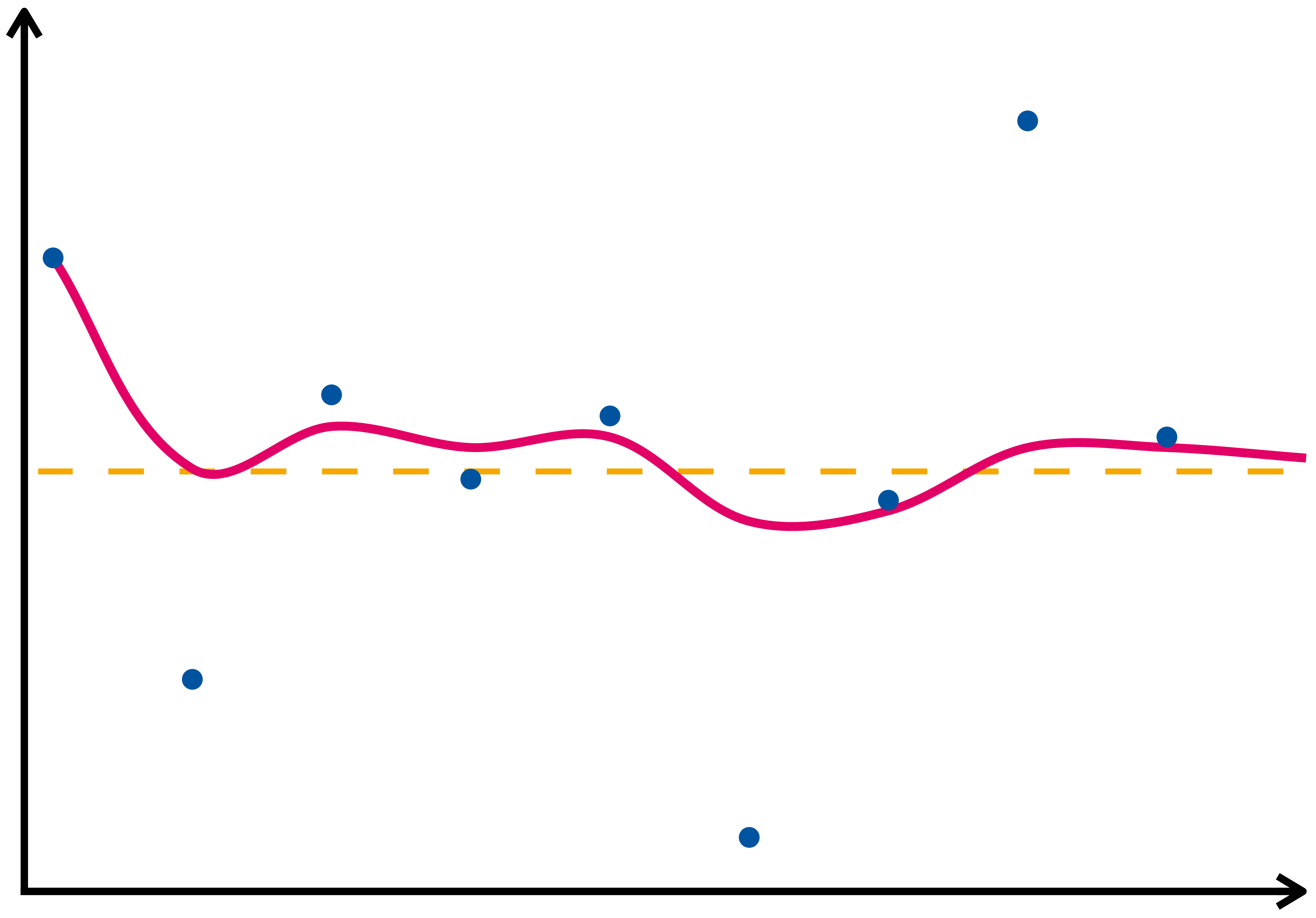


- LDA finds the ideal axis so that the mean difference is maximized and the scatter, i.e. variation, is minimized in each category:

$$\frac{(\mu_a - \mu_b)^2}{s_a^2 + s_b^2}$$

Smooth Tracking and Predict Movements

Example: Kalman Filter



- The Kalman filter predicts values based on the measurements from previous time periods
- Smoothing of values makes it ideal for scenarios that track the user

Multimodality: Summary

	Visual		Auditory		Haptic	
	Input	Output	Input	Output	Input	Output
Control						
Data						