Dialog Quality Assessment:

A Tool for Measurement and Classification of Dialog Management Systems

Presented by Emmett Coin

Who are we

- The AVIOS **Advanced Dialog Group**
- Explore and evangelize the leading edge of **human-computer** conversation
- This presentation represents the joint efforts of: **Bill Scholz, Lorin Wilde, John Tadlock, Marie Meteer, and Emmett coin**
- We have met regularly as a group for 15 Years. All of us have worked on multiple significant projects implementing **sophisticated** speech/multimodal dialogs.
- A large part of our method involves **serious but informal discussions** about recent and future technologies and how they can be used to improve human computer interaction.

- **We would like to extend an invitation to like-minded individuals to join our discussions!**

The Challenge of Measuring

- Measure **user perception** of the technical capabilities of the dialog management systems found in various products and services.
  - **Distinct** from a business assessment or overall product assessment.
  - This **specific** technical focus is our (AVIOS's) unique contribution
- Goal is to formulate a list of **technical parameters** that describe dialog management system capabilities.
- These measures will help to
  - Establish **technical targets** for high achieving conversations
  - Avoid **over-building** when minimal dialog management capabilities are required
  - Prioritize **improvements** in conversation skills in products and services

Note: We will present a list we have been working with over the last year. This is not intended as a final or exhaustive list.

Rational

- Some human computer interactions will ultimately become very conversational
- Conversational frameworks will soon become components incorporated into applications
- The mechanics of a conversational flow will evolve as an independent technology from other application components
- Developers will not have the time or skill to craft and debug sophisticated conversations
- Measurements will verify and reinforce the best frameworks

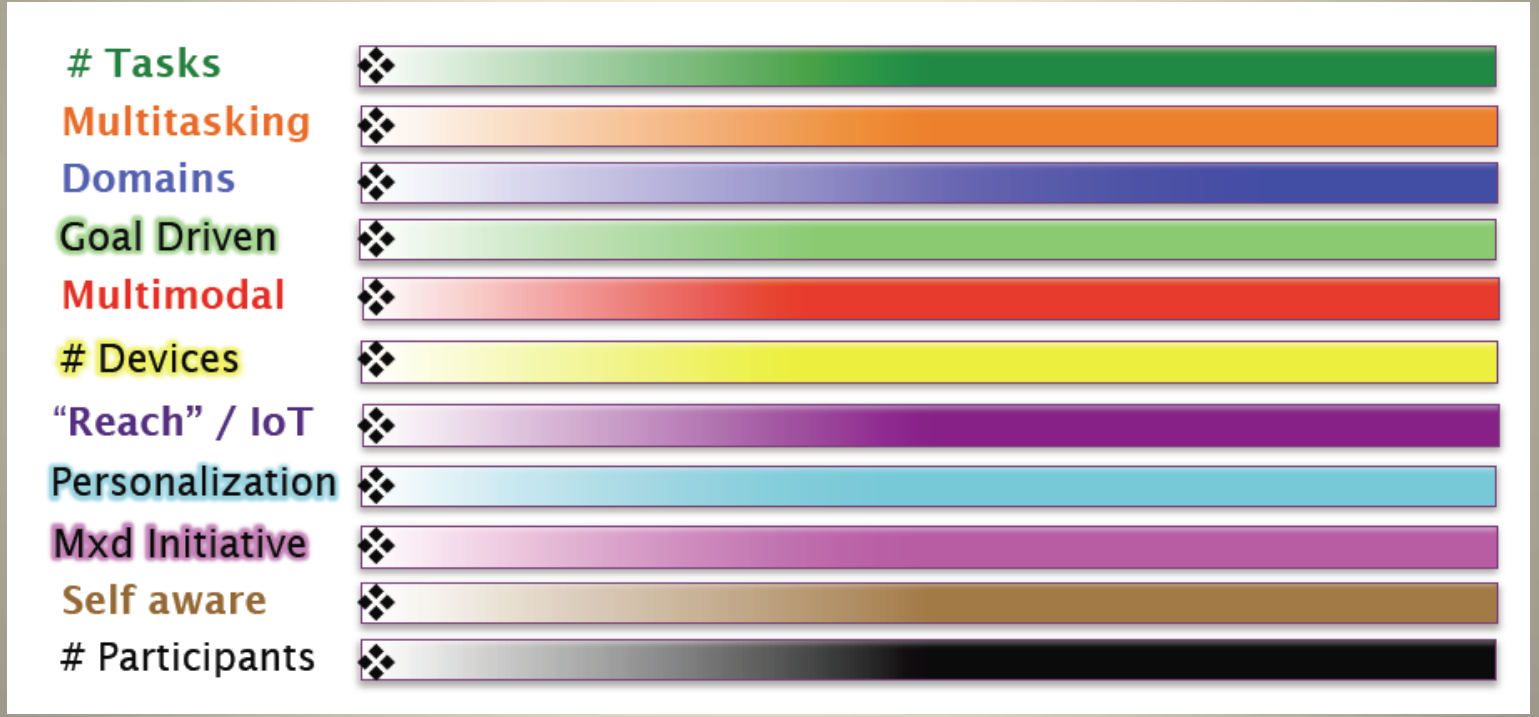
Building on last year's presentation

- Number and complexity of tasks
- Multitasking
- Number of different domains
- Goal Driven
- Multimodality
- Reach into the IoT
- Number of devices it works with
- Degree of personalization
- Mixed Initiative
- Self Aware
- Number of participants.

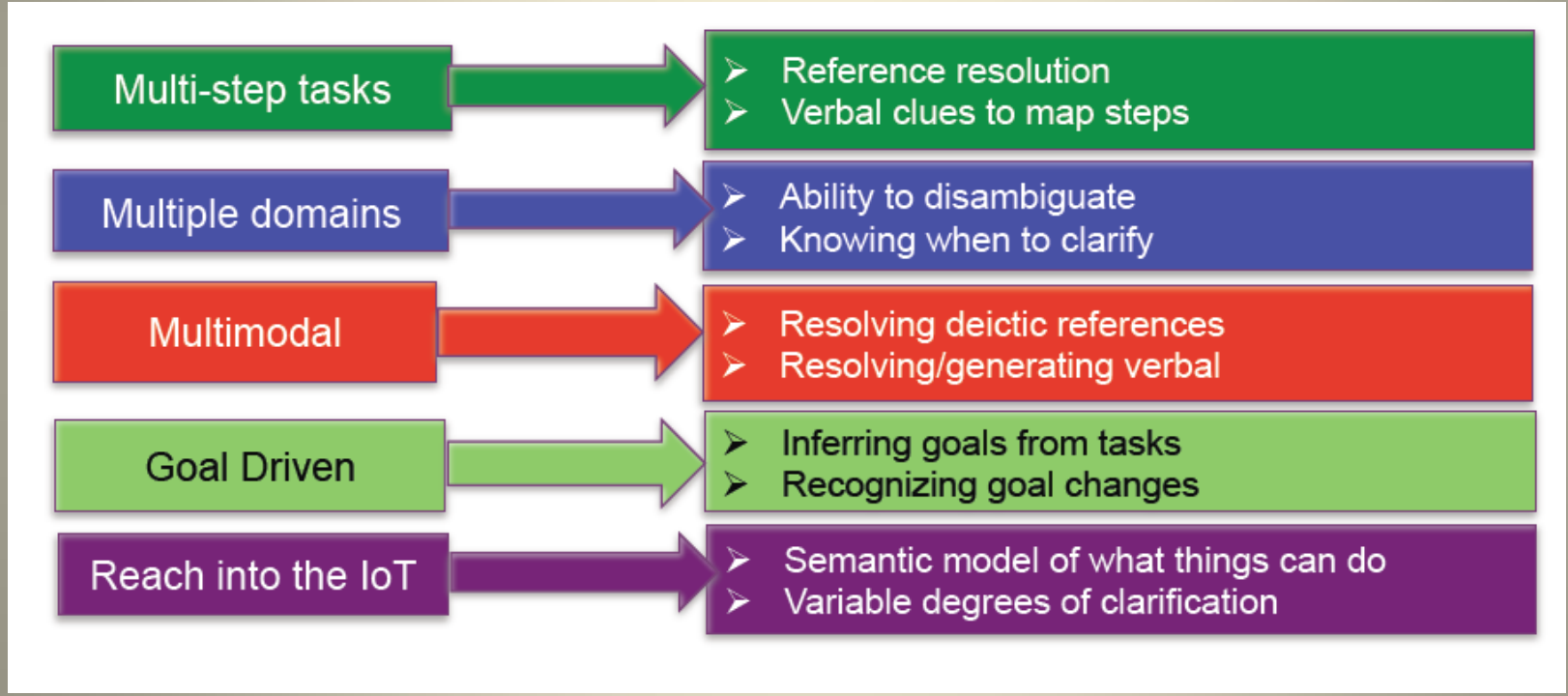
This is the list is from our 2016 Presentation:

Assessing Dialog Management Systems

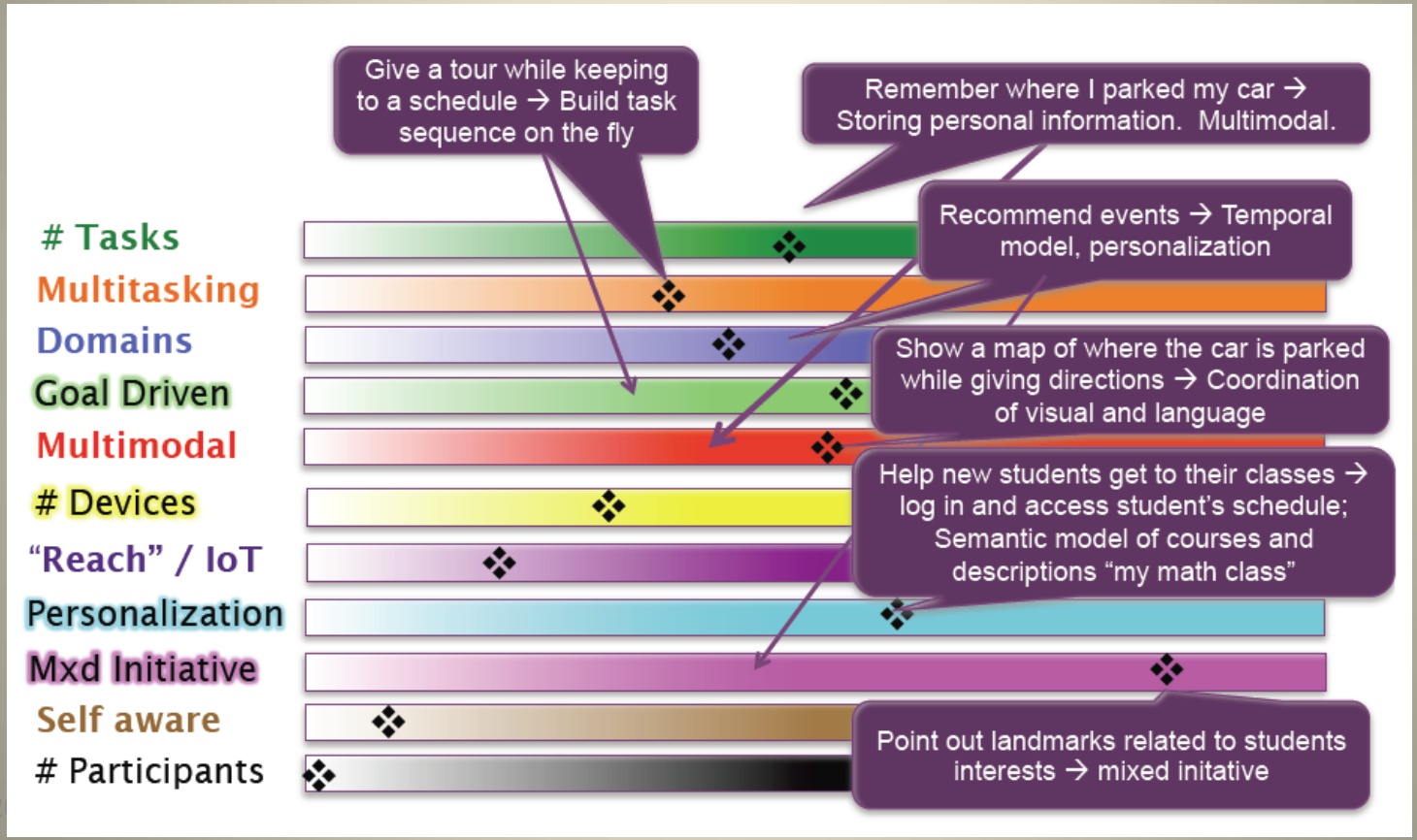
Presented as Sliders



So, you had to understand this



And think about things like this



Considerations

- Developers need to measure and **adjust** parameters
- Experts in the linguist parameters are **not typical** users
- Enlisting **trained** experts is expensive
- Real users (lay people) can consistently rate agreement with simple **general** statements
- But, general statements do not represent **technical** parameters
- Good analysis requires **lots** of data
- **Therefore**: We need an easy to fill out survey from a large user community

Complex Conversations

- Every interaction is a unique loop
- Loops are similar but never the same
- Parameters present in different proportions
- Each encounter results in a different score

Evaluating Complex Loops

Approach

Lay people may not understand the technical metrics themselves
So we need to …

- Create a list of statements to be rated on a discreet point scale
- Elicit high level impressions which embody several parameters
- Design statements to be easily and quickly rated
- Analyze those results to produce *pure* technical metrics

Note: We have drawn inspiration from “Standardized Questionaires for Voice Interaction Design” by James R. Lewis in The Journal of AVID, April 2016

Some General Statements

- The interaction with the system was a natural conversation.
- It was hard to know when the system was slow and when it didn't hear.
- It took concentration to use the system.
- The system was helpful in completing a complex task.
- The system easily accepted multiple pieces of information.
- The system remembered details that were discussed before (e.g. dates, amounts)
- …
- The system sometimes volunteered additional useful information.
- Sometimes the system stopped paying attention at odd times.
- After multiple sessions the system worked more smoothly.
- Achieving the goal seemed like a team effort.
- The language seemed confusing at times.

Parameters mapped to Statements

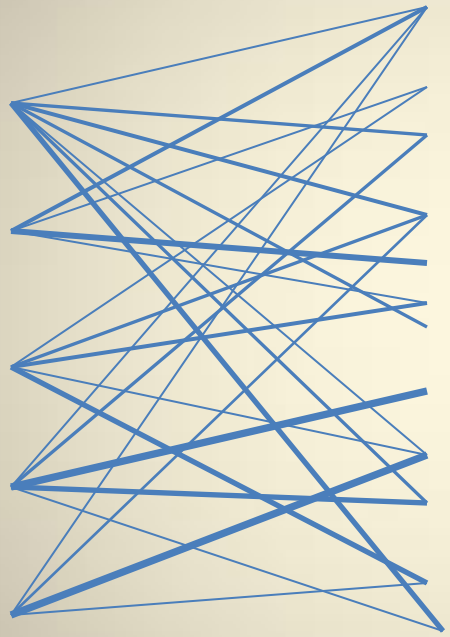
Goal Driven

Ellipsis

Personalization

Mixed Initiative

Context



The interaction was a natural conversation

It was hard to know if it was slow or if it didn't hear me

I needed to concentrate to use the system

It helped me complete a complex task

I could give it multiple pieces of information at once

The system figured out the goal as soon as it had enough information

The system was too repetitive

The conversation caused me to smile

I understood how the system was approaching the goal

The system controlled the flow of the interaction

It took specific words or phrases to make the system work.

It remembered details we discussed before (e.g. dates, amounts)

The system sometimes volunteered additional useful information

Sometimes the system stopped paying attention at odd times.

After multiple sessions the system worked more smoothly

Achieving the goal seemed like a team effort

The language seemed confusing at times

Parameters mapped to Statements

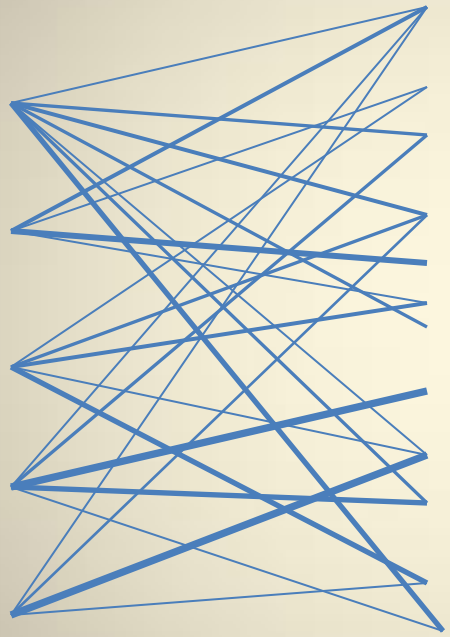
Goal Driven

Ellipsis

Personalization

Mixed Initiative

Context



The interaction was a natural conversation

It was hard to know if it was slow or if it didn't hear me

I needed to concentrate to use the system

It helped me complete a complex task

I could give it multiple pieces of information at once

The system figured out the goal as soon as it had enough information

The system was too repetitive

The conversation caused me to smile

I understood how the system was approaching the goal

The system controlled the flow of the interaction

It took specific words or phrases to make the system work.

It remembered details we discussed before (e.g. dates, amounts)

The system sometimes volunteered additional useful information

Sometimes the system stopped paying attention at odd times.

After multiple sessions the system worked more smoothly

Achieving the goal seemed like a team effort

The language seemed confusing at times

Parameters mapped to Statements

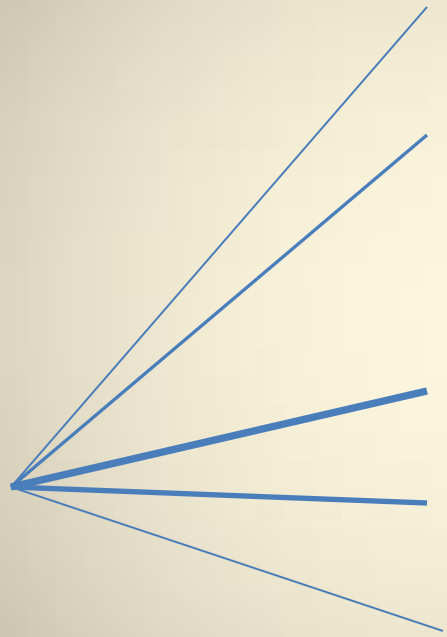
Goal Driven

Ellipsis

Personalization

Mixed Initiative

Context



The interaction was a natural conversation

It was hard to know if it was slow or if it didn't hear me

I needed to concentrate to use the system

It helped me complete a complex task

I could give it multiple pieces of information at once

The system figured out the goal as soon as it had enough information

The system was too repetitive

The conversation caused me to smile

I understood how the system was approaching the goal

The system controlled the flow of the interaction

It took specific words or phrases to make the system work.

It remembered details we discussed before (e.g. dates, amounts)

The system sometimes volunteered additional useful information

Sometimes the system stopped paying attention at odd times.

After multiple sessions the system worked more smoothly

Achieving the goal seemed like a team effort

The language seemed confusing at times

Key Points

- Scoring conversation by parameters is hard
- Users do not understand the mechanics behind the interaction
- Linguists and developers *who do understand* are not normal users
- We need many evaluations to score an agent

Conclusions

- Complex conversational agents are inevitable
- Measuring performance is essential to product success
- Measurement informs us on what features need work … and what doesn't
- We must continue exploring approaches to measure conversational interactions

Thank You!

The AVIOS Advanced Dialog group