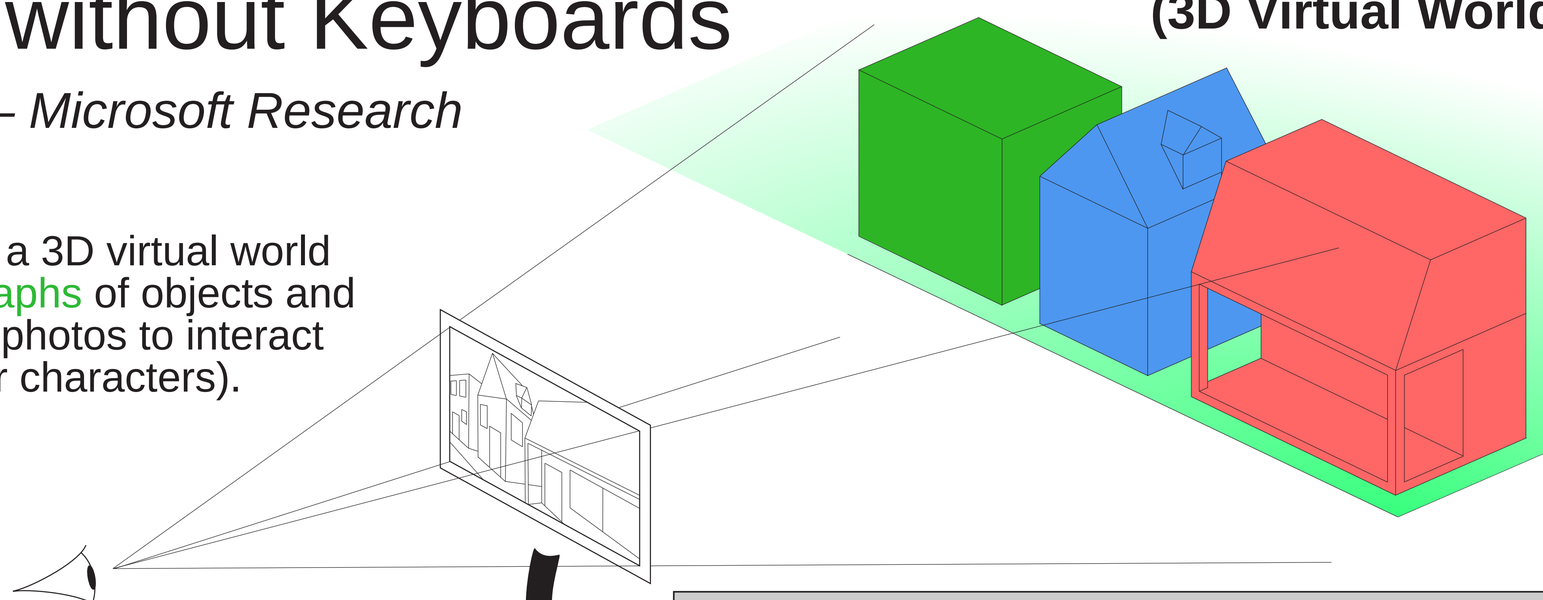
Multi-Modal Question-Answering: Questions without Keyboards

Gary Kacmarcik — Microsoft Research

Goal: Allow players in a 3D virtual world to take virtual photographs of objects and then use the resulting photos to interact with NPCs (non-player characters).



Interaction: Player shows photo to NPC and clicks on an item to bring up a menu of dynamically generated queries.

Selecting a query from this menu, e.g.:

```
be 1 ({ Verb } (?))
Tsub this 1 ({ Pron })
AppInfo 102
Lnom what 1 ({ Pron })
```

"What is this?"

selects the set of nodes in the KB that correspond to that object.

Treelets containing the selected nodes are combined to create:

```
be 1 ({ Verb } (.))
Tobj house 1 ({ Noun })
Tsub that 1 ({ Pron } <1>)
AppInfo 102
Possr Mr._Jones 1 ({ Noun })
AppInfo 23
```

"That is Mr. Jones' house."

From which we generate the final response.