Multimodal Voice Search for Interactive Media

Mazin Gilbert

Executive Director, AT&T Labs Research

Acknowledgement

Michael Johnston

Bernie Renger

Harry Chang

Linda Robert

Harry Blanchard

PinoDi Fabbrizio

JuergenSchroeter

BehzadShahraray

Benson Tang

New Trends in Interactive Media

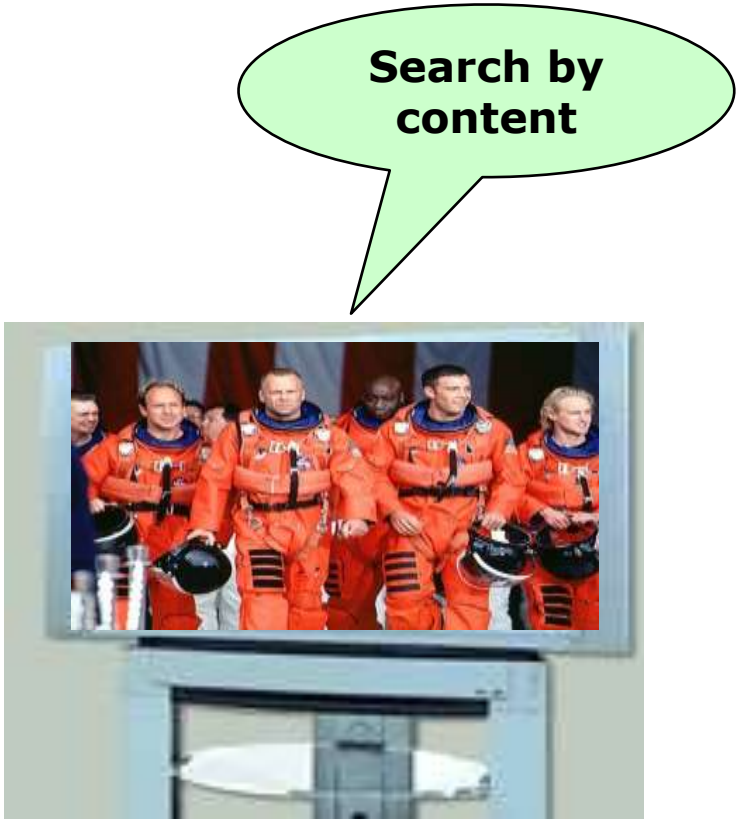
- **Media Consumption:**
  - On demand information (Anytime, Anywhere)
  - Multiple devices (TV, PC, Mobile)
- **Challenges:**
  - *Multimodal Inputs:* Typing text is cumbersome as devices get smaller; need for hands-free, eyes-free
  - *Search:* The abundance of multimedia content (speech, video, text, web) makes it difficult to search
  - *Multimedia Outputs:* The content in its original form may not be suitable for the consumer's device
- **Opportunities:**
  - Multimodal voice search technologies offer a more natural user experience
  - Differentiated services leading to new revenue generation

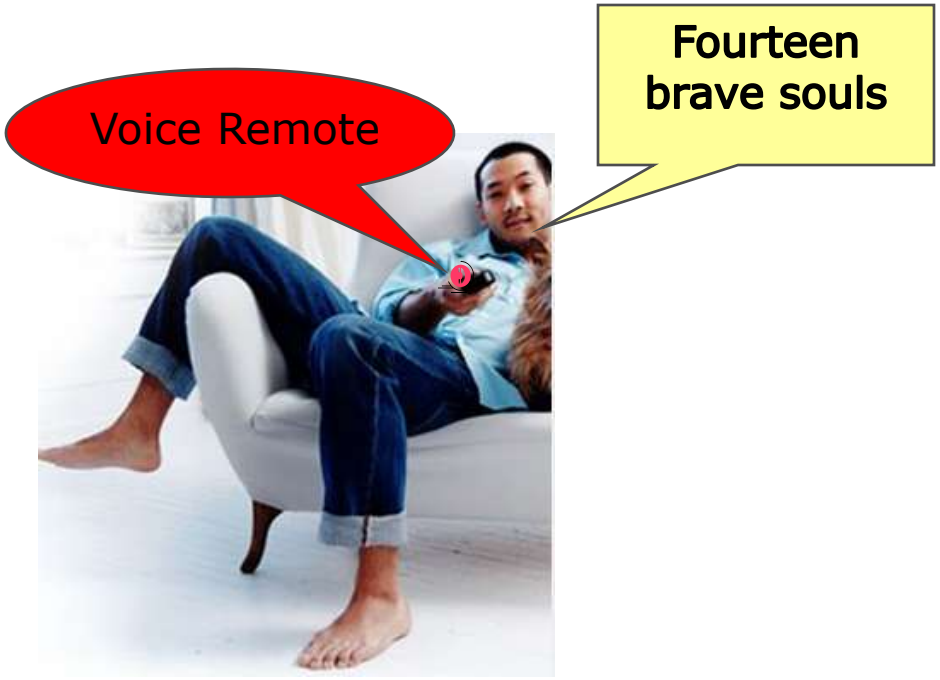
Media on Demand

“What if”

Movies On Demand

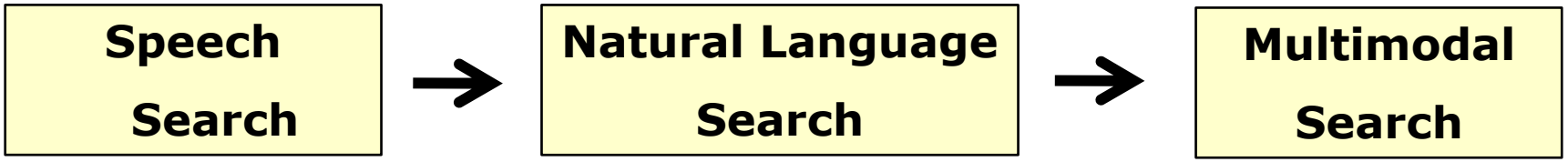
Movies On Demand





The Technology

Three Generations of voice search Technologies



Recognizing words	Understanding Meanings	Multimodal inputs
Dictation; web search; SMS; voice mail	Advanced search; question/answering; translation	Form filling; navigation; billing; ordering
"Italian restaurants"	"Cheap Italian Restaurants on market street"	"Italian restaurants*Here*"

AT&T WATSON Technologies

Types: "Lawyers open on Saturdays"

Natural Language Understanding

GPS

Preferences

Multimodal Search

Information Repository

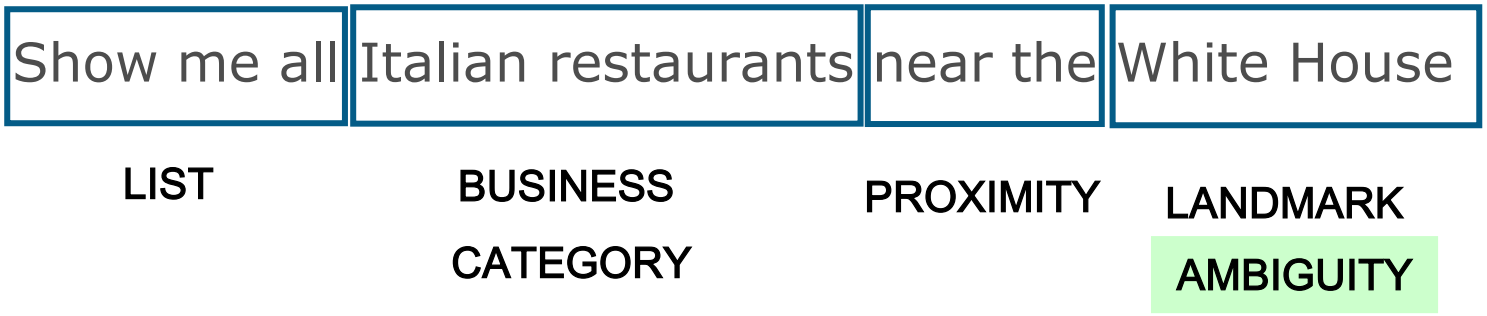
Lattices

speaks: "nearest pizza restaurant"

Speech Recognizer

at&t

Example: Semantics in Voice Search



GPS: Longitude/Latitude = Washington DC

Preferences: Price = $$$

Action: List

Category: Italian Restaurants

Constraints: Medium Price

Proximity: Near

Landmark: White House

Locality: Washington DC

Multimodal Search

Multiple input and output modes through multi-dimension finite state transducers

- Input: "HOW DO I GET HERE?"

- Output: "TAKE THE A TRAIN HEADING NORTH …"

Voice search for Television

Multimodal vs. Media Center Edition Remote

Microphone & 2 LED lights

Talk Button

Start Button Moved

MCE **Multimodal**

Multimodal Search for IPTV

Apply **spoken natural language** to easily access and search multimedia information, wherever they are

- **Technology Enablers:** WATSON speech recognition and search technologies provide accurate search of a variety of multimedia applications.
- **Preferred User Experience:** Using speech is simple and natural. Human Factors studies show 65% of users strongly prefer voice search and 50% are willing to pay for it.

Mobile Multimodal Search

Mobile Voice Search

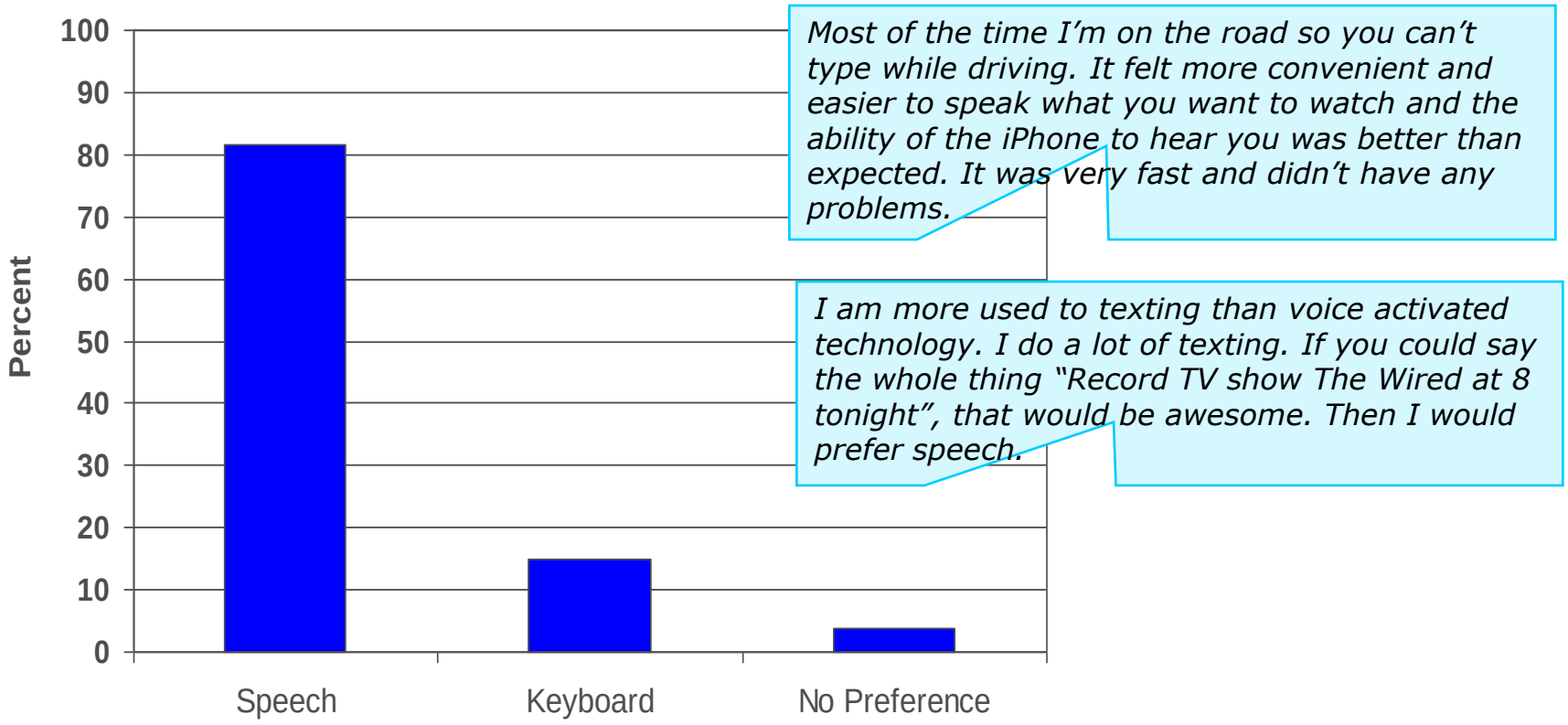
- **Differentiated User Experience:** Speech is faster, more accurate, easier, can multitask (e.g., speak and drive), fun to use, and meet regulatory mandates on hands-free, eyes-free.
- **Market transition:** Speech is a transitional technology for mobility. IC's faster, memory larger, phones smaller. But fingers are not getting smaller. Need screen, need to eliminate typing.

Media Search By Content

- WATSON speech recognition for video transcription and multimodal search
- Natural language understanding for parsing the speech output
- AT&T MIRACLE for video search of media programs

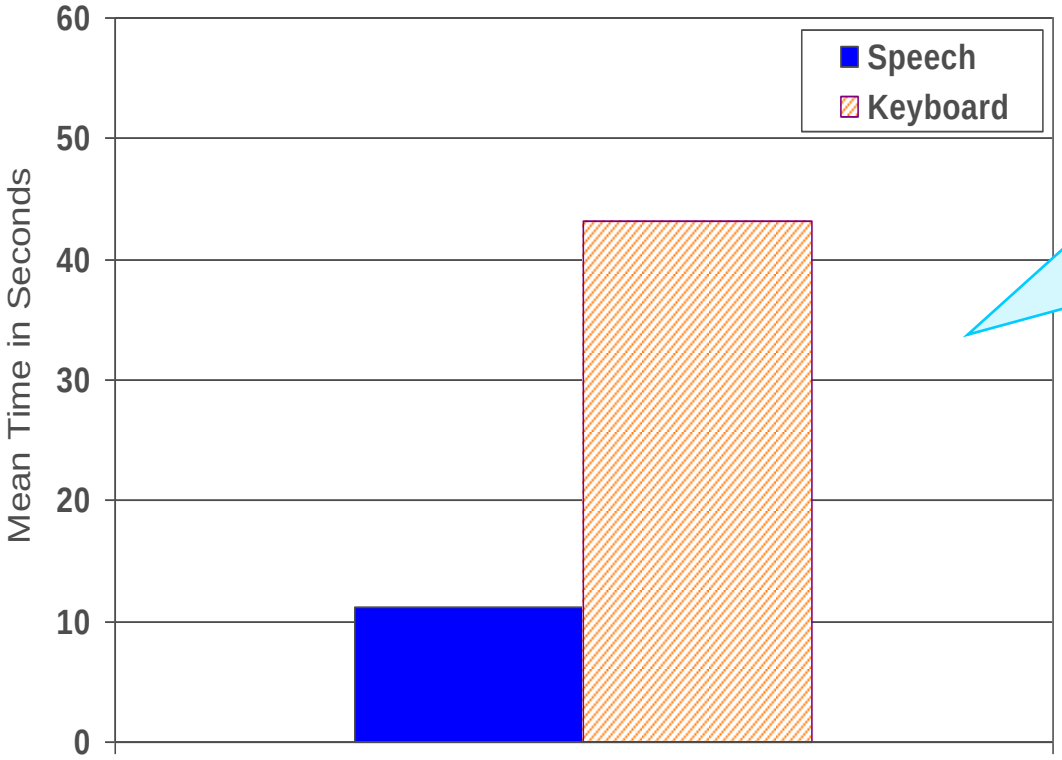
Searching "Academy awards" for "comedy shows on NBC past week"

Human Factors Study



Speech was markedly preferred over keyboard

HF Study: Speech is Faster



It's [speech] quicker and less mistakes due to hitting the wrong button and spelling. Probably more fun too. If I thought I forgot to record a show, I would definitely use it in my car.

It took 4 times longer to input the keyboard search (vs. speech entry)

Mobile Local Search

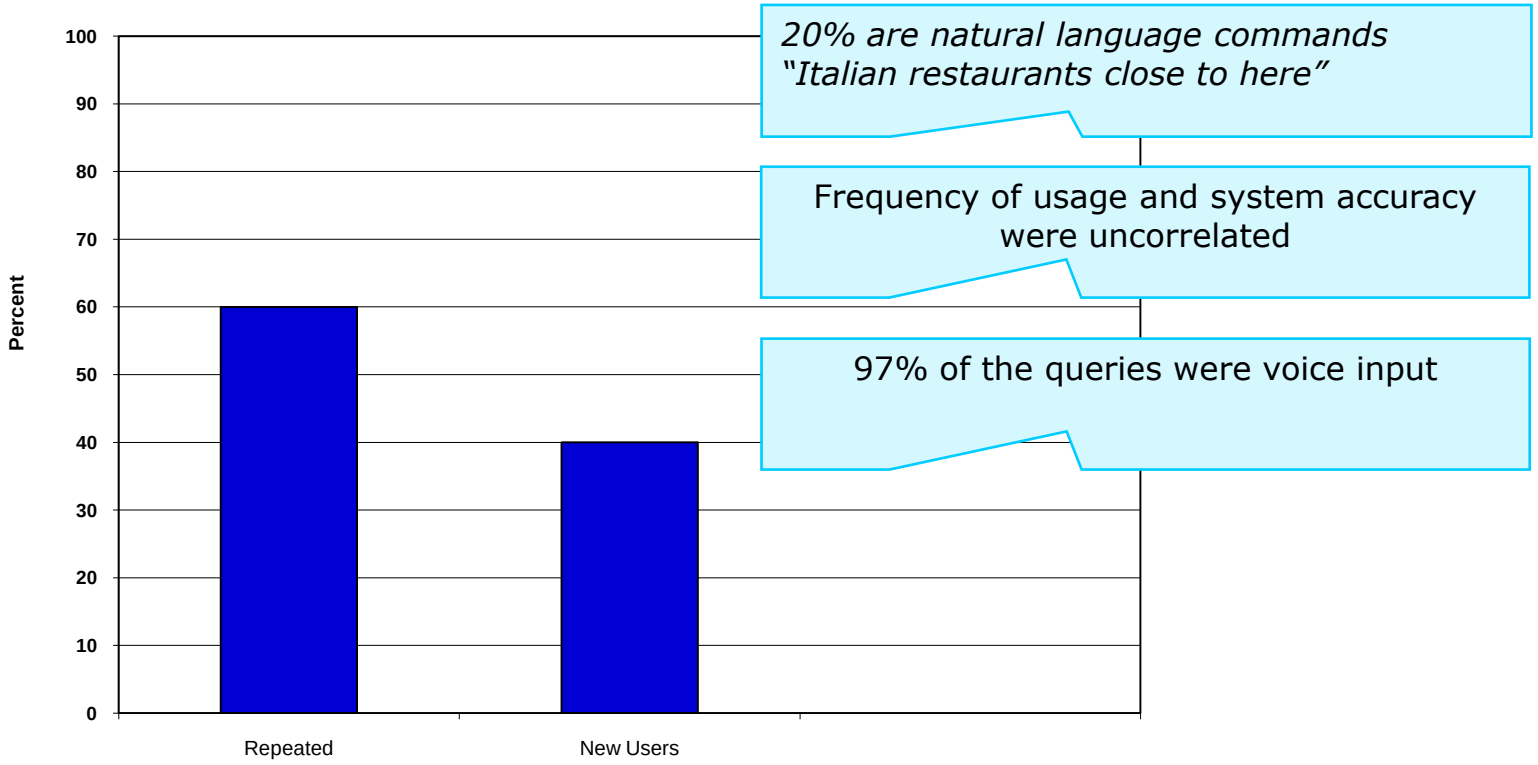
Speak4It: Multimodal Voice Search for Local Businesses

"Real estate lawyers in Madison New Jersey"

- Natural-language large-vocabulary voice search for local business.
- No application-specific audio data were initially available.
- Recognition and understanding models were bootstrapped using heterogeneous data; web-based local search data, customer care speech data, and YP database.
- Models were adapted through semi-supervised training once the service went live.

at&t

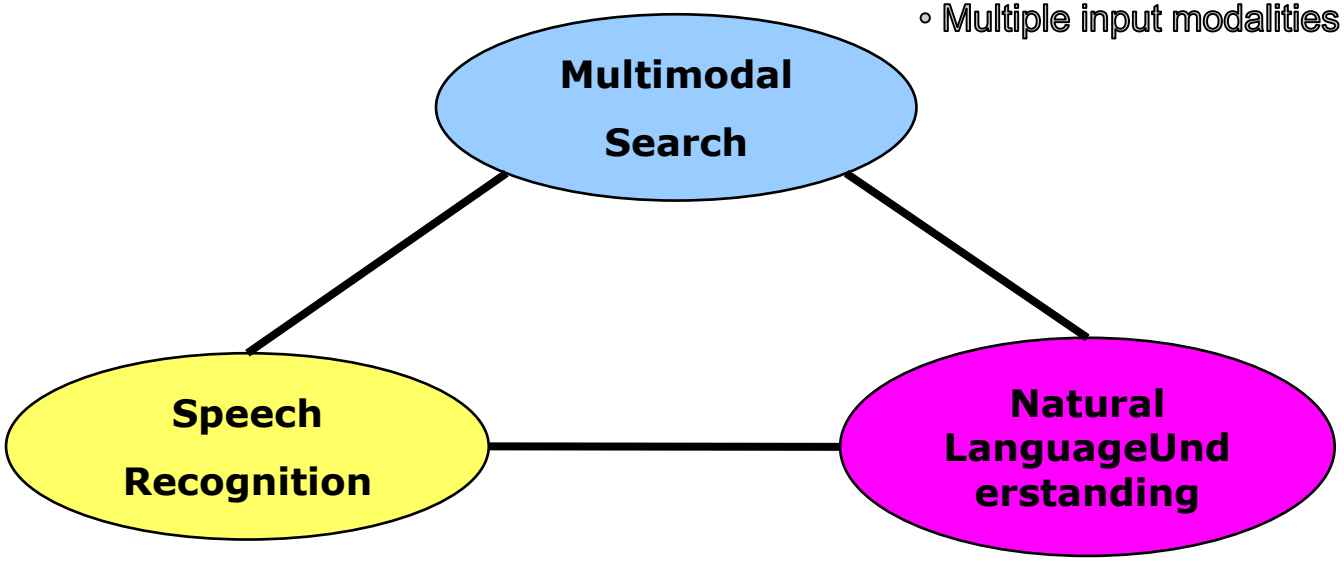
Speak4It Results



ABCNews TECH Bytes (September 16, 2008)

AT&T WATSON Technologies

- Lattice- based input
- Multiple input modalities



- Adaptive acoustic/language models
- Trained on heterogeneous large corpora
- Robust to noise

- Stochastic query parsing
- Training on heterogeneous large corpora
- Robust to speech recognition errors

New Challenges & Opportunities

- **Providing natural communication**
  - Combining speech recognition, text to speech, and natural language understanding to provide users the ability to speak naturally
- **Supporting advanced search**
  - Including video content, speech, text, metadata, description, etc.
- **Adaptive learning and personalization**
  - Customize the application to the users and their
- **Repurposing content for use on different devices**
  - Mobile phones, desktops, TVs, hand-held devices

Advanced Search with Natural Language Understanding

True single search box

- Show me ice cream stores in Pasadena California

Temporal qualification

- Pizza places open after midnight in Florham Park

Geographic qualification

- Seafood restaurants on Market street

Attribute qualification

- Restaurants in New York with outdoor dining
- Cheap Moroccan places around here

Voice Activated Remote Control (VARC)

Marketing Survey

- **Marketing Study**
  - 1000+ TV households
  - Participants were shown the "VARC marketing video”
- **Main Results**
  - >50% likely to purchase or seek more info
  - >20% are *very likely* to purchase or seek more info