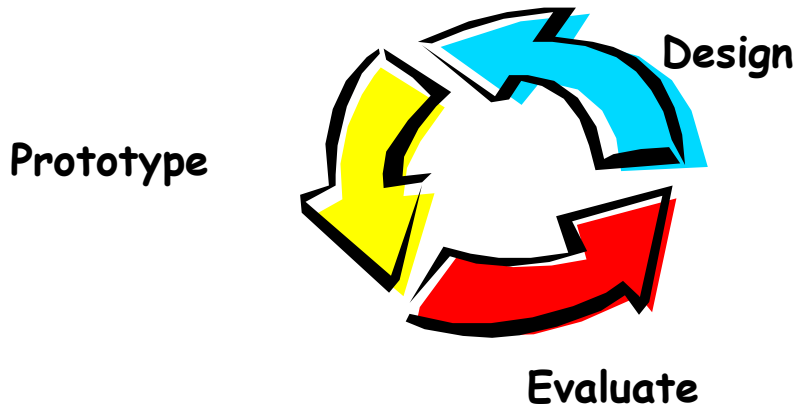
Qualitative Evaluation

Food for Thought

- Nest thermostat
  - https://youtu.be/oxOukh_Ma6o
- Programmable thermostats are no longer LEEDS certified
  - Why?
- And what is LEED?

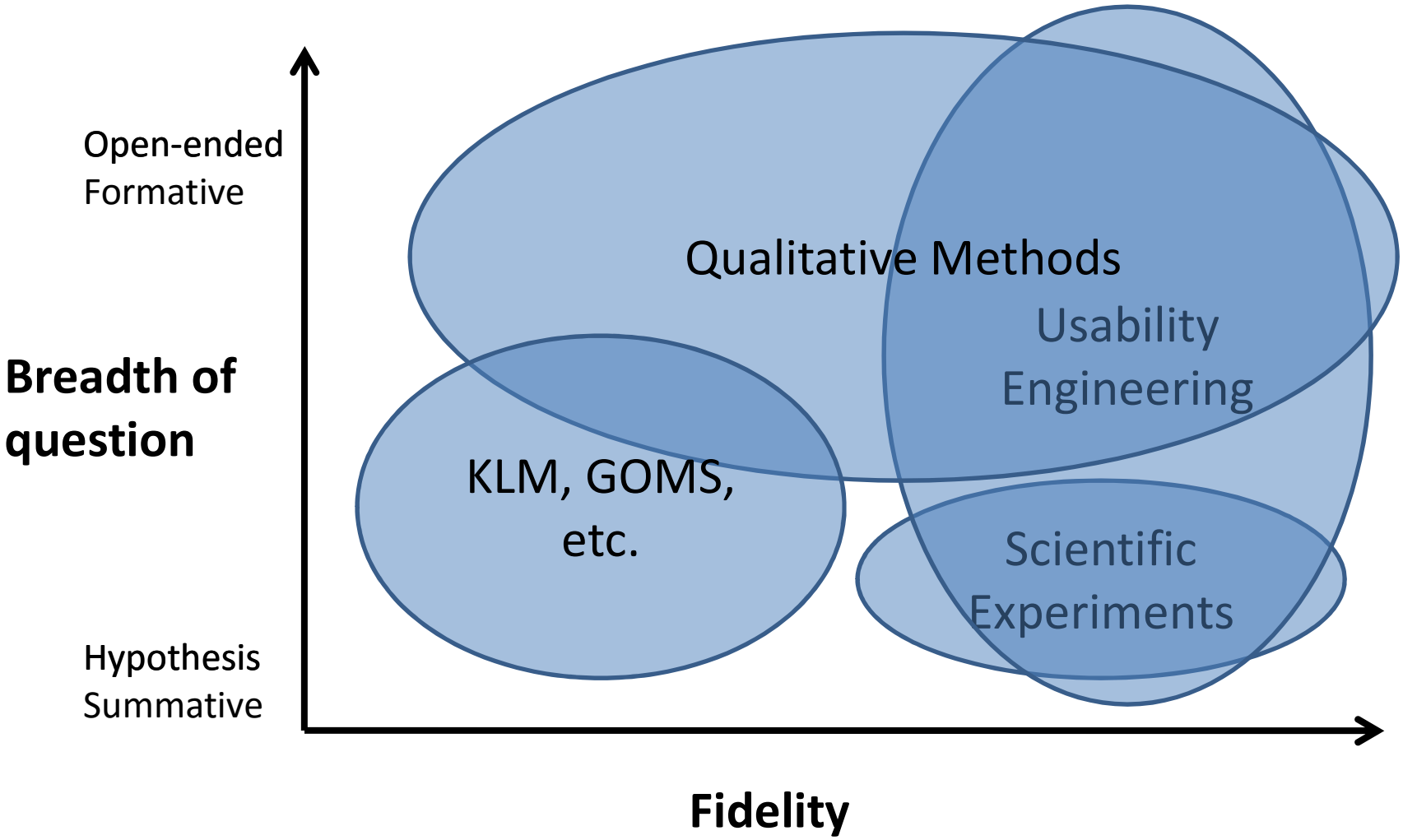
Evaluation overview

- Evaluation is concerned with gathering data about the usability of a design or product by a specified group of users for a particular activity within a specified environment or work context



- Similarity to many design tasks
  - Iterative nature

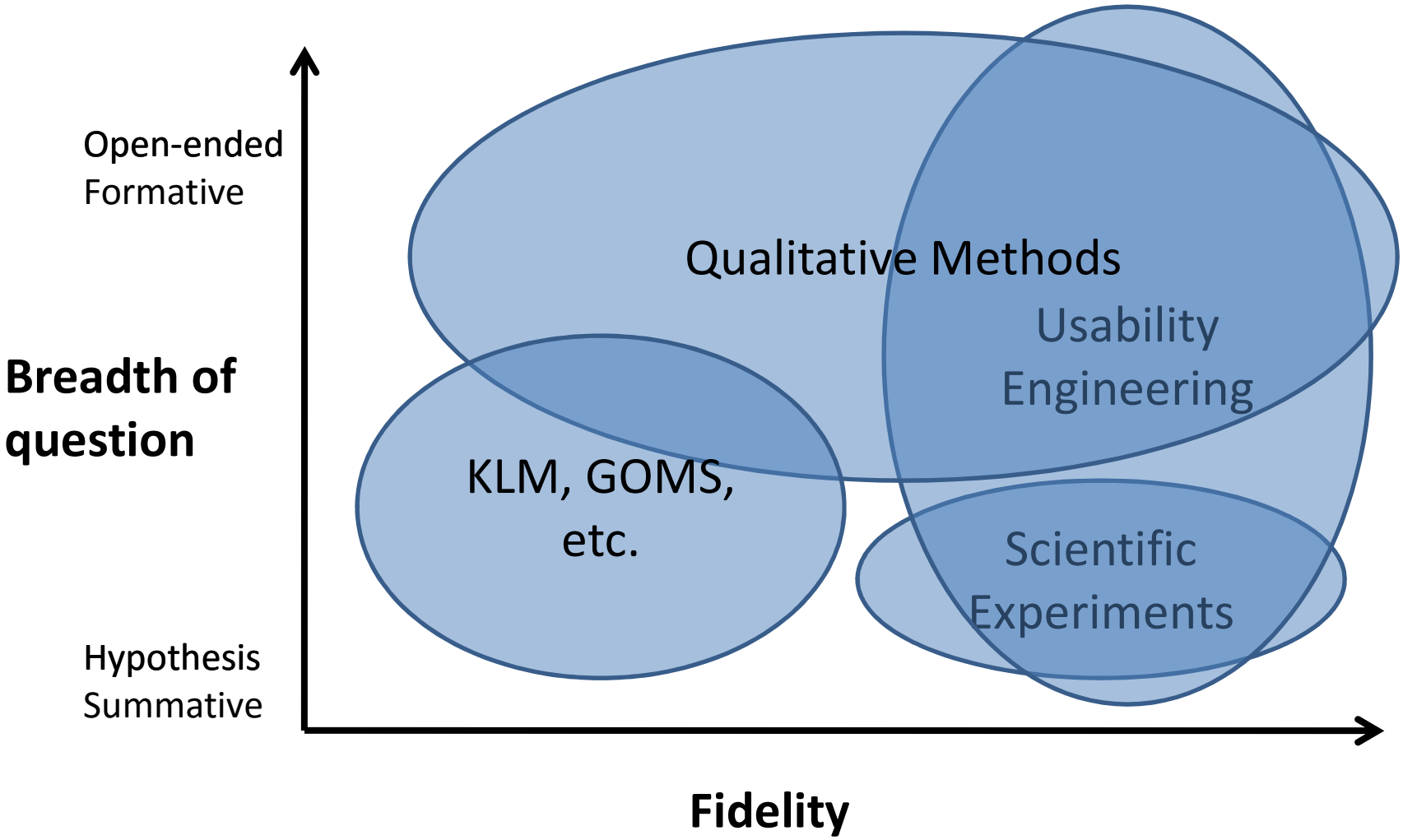
Recall: A Design Space for Evaluation



Recall

- Scientific Experiments
  - Useful for evaluating narrow features of software, e.g. a new interaction technique, a specific task
  - Measurements can include time, error rate, subjective satisfaction, clicks … anything quantitative
- Didn’t spend much time on qualitative evaluation
  - Beyond walkthroughs/thinkalouds for prototypes

A Design Space for Evaluation



Qualitative Evaluation

- Constructivist claims
- Very common in design
  - Can be used either during design or after design complete
  - Can also be used before design to understand world
- Broad categories
  - Walkthroughs/thinkalouds
  - Interpretive
  - Predictive

Recall Walkthroughs/Thinkalouds

- Variants include person-down-the-hall and with end-users
- Distinction?
  - Walkthroughs = you showing
  - Thinkalouds = user walkthrough while verbalizing what they are doing
  - Thinkalouds in two forms: concurrent and retrospective
- Advantages and disadvantages to walkthroughs versus thinkalouds

Qualitative Evaluation

- Constructivist claims
- Very common in design
  - Can be used either during design or after design complete
  - Can also be used before design to understand world
- Broad categories
  - Walkthroughs/thinkalouds
  - Interpretive
  - Predictive

Interpretive Evaluation

- Need real-world data of application use
- Need knowledge of users in evaluation
- Techniques (will revisit after talking about data collection)
  - Contextual Inquiry
    - Similar to for user understanding, but applied to final product
  - Cooperative and Participative evaluation
    - Cooperative evaluation allows users to walkthrough selected tasks, verbalize problems
    - Participative evaluation also encourages users to select tasks
  - Ethnographic methods
    - Intensive observation, in-depth interviews, participation in activities, etc. to evaluate
    - Master-apprentice is one restricted example of evaluation that can yield ethnographic data

Collecting usage data

- Observations
- Monitoring
- Collecting opinions

Observations

- Diaper 89: Not as straightforward as it seems
  - Are we seeing what we think we see?
  - Physiological and psychological reasons the eye produces a poor visual image:
    - You see what you want to see
- You want users to react to your ideas
  - Observation is one technique
  - Be aware of limitations
- Different types include:
  - Direct observation
  - Indirect observation
  - Collecting opinions

Direct observation

- Observe users as they perform tasks:
  - Problem: Your presence affects task
    - Called Hawthorne effect from study of plant workers in Hawthorne Illinois
      - Observation resulted in improved performance
  - Problem: Observations (even with notes) are incomplete
    - Consider evaluating the interface on an ATM
    - Consider evaluating a product with a kindergarten class

Direct observation notes

- Useful early in project
  - Insight into what users do
  - What users like
- To improve efficiency
  - Develop some shorthand notation
  - Create a checklist for common things
  - May want to record as well so you can refer back

Indirect observation

- Video recording is most common form
  - Can give very complete picture
  - Often coupled with some form of event logging
    - Keystroke logging
    - screen capture
    - multiple cameras
  - Need a lot of information
    - Facial features
    - Posture and body language
  - Can be awkward
    - In their workplace requires setup
    - Awareness of being filmed alters behavior (e.g. Hawthorne)

Analyzing video data

- Task-based analysis:
  - How users tackled given tasks
  - Where difficulties occurred
  - What can be done
- Performance-based analysis
  - Measure performance from data
  - Timing, frequency of errors, use of commands, etc.

Analyzing video data

- Huge tradeoff between time spent and depth of analysis
  - Informal can be undertaken in a few days
    - Often coupled with direct observation
  - Formal takes much longer
    - First analyze to determine performance measures
      - May take several play-throughs
    - Extraction of measures also requires multiple iterations
    - 5:1 or worse is often cited!

Monitoring

- Software logging
  - Complete systems, not low fidelity
  - Time-stamped keypresses gives record of each key user pushes
  - Interaction logging allows interaction to be replayed in real time
    - Often coordinated with video observation
  - Can skip through problem-free areas
  - Drawbacks include
    - Cost
    - Data volume

Soliciting opinions

- Interviews
- Questionnaires

Questionnaires and surveys

- Flexible means of gathering data
- Two possibilities:
  - Closed questions
    - Select from a list
    - Use scale to measure
    - E.g. yes/no/don't know
    - Easy to get statistical analysis
  - Open questions
    - Respondent provides own answer
- Can use pre and post
  - Measure changes in attitudes
  - Often limited correlation – Root and Draper, 83
    - Implies not good for eliciting design decisions

Interpretive Evaluation

- Take real world data and an understanding of users
- Then interpret that data to assess software
- Techniques (will revisit after talking about data collection)
  - Contextual Inquiry
    - Similar to for user understanding, but applied to final product
  - Cooperative and Participative evaluation
    - Cooperative evaluation allows users to walkthrough selected tasks, verbalize problems
    - Participative evaluation also encourages users to select tasks
  - Ethnographic methods
    - Intensive observation, in-depth interviews, participation in activities, etc. to evaluate
    - Master-apprentice is one restricted example of evaluation that can yield ethnographic data

Predictive Evaluation

- Avoid extensive user testing by predicting usability
- Includes
  - Inspection methods
  - Usage modeling
  - Person down the hall testing

Inspection methods

- Inspect aspects of technology
- Specialists who know both technology and user are used
- Emphasis on dialog between user and system
- Include usage simulations, heuristic evaluation, walkthroughs, and discount evaluation
  - Also includes standards inspection
    - Test compliance with standards
  - Consistency inspection
    - Test a suite for similarity

Inspection Methods: Heuristic evaluation

- Set of high level heuristics guide expert evaluation
  - High-level heuristics are a set of key usability issues of concern
- Guidelines are often quite generic
  - Simple natural dialog
  - Speaks users' language
  - Minimizes memory load
  - Consistent
  - Gives feedback
  - Has clearly marked exits
  - Has shortcuts
  - Provides good error messages
  - Prevents errors

Process

- Each review does two passes
  - Inspects flow from screen to screen
  - Inspects each screen against heuristics
- Sessions typically one to two hours
- Evaluators aggregate and list problems

How good is HE?

- Mean of six studies found that five reviewers found 75% of usability problems
  - Very cost effective
  - Compares favorably with other techniques

Usage simulations

- Review system to find problems
- Done by experts who simulate less experienced users
  - Also called expert reviews/evaluation
- Why not use regular users?
  - Efficiency
    - Many errors, one session (if they're good)
  - Prescriptive feedback
    - More forthcoming with feedback
    - Need less prompting
    - Detailed reports

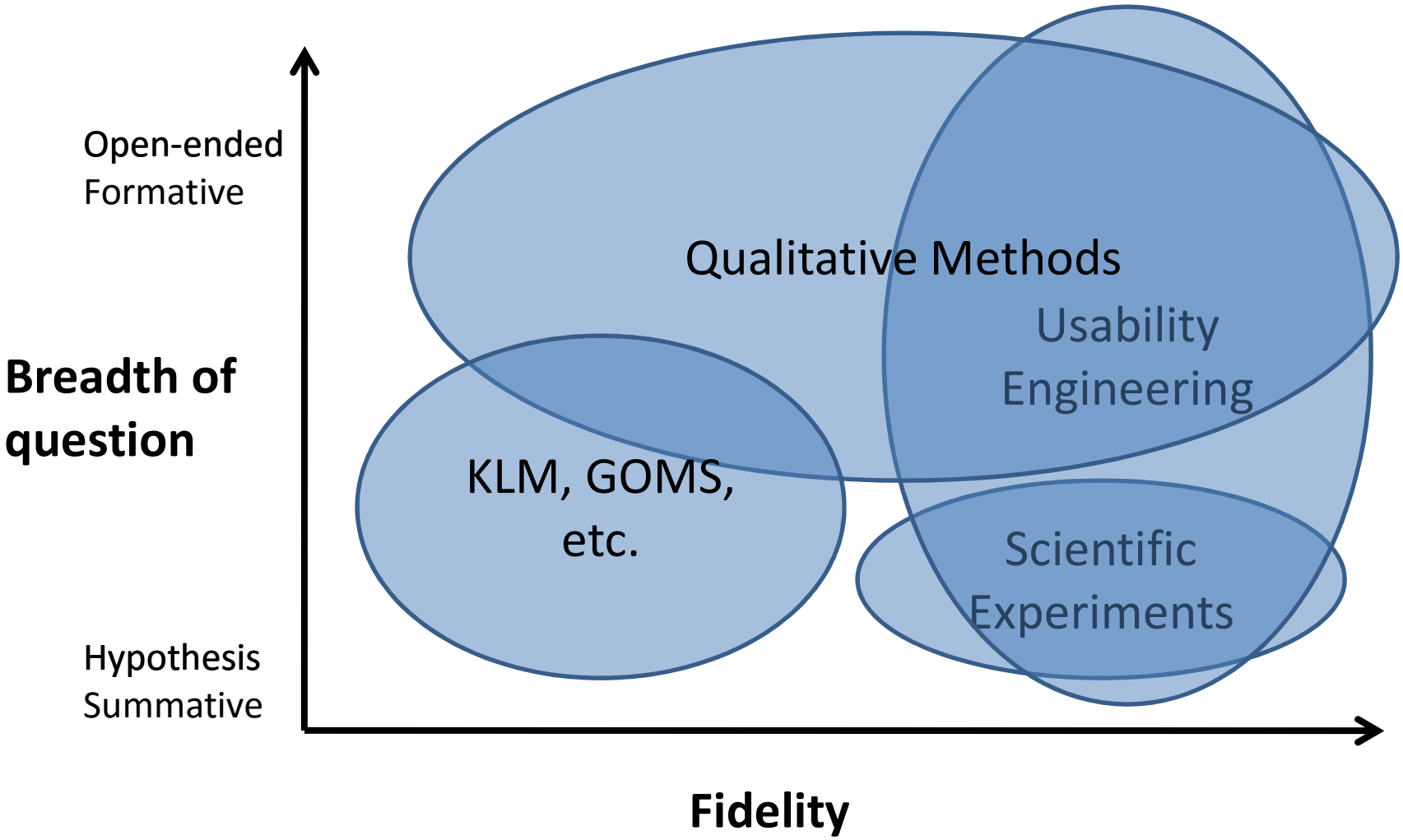
Usage simulation caveats

- Reviewers should not have been involved previously
- Reviewers should have suitable experience
  - In HCI and in Media/creative design for some systems
  - May be difficult to find!
- Role of reviewers needs to be clearly defined
  - Want them to adopt correct level of knowledge
  - Intermediate user is difficult
- Need common tasks and system prototype
- Need several experts to avoid bias
  - Different people have different opinions
- Won’t capture the full variety of real user behavior
  - It’s always surprising how bad real users are

Usage simulation reporting

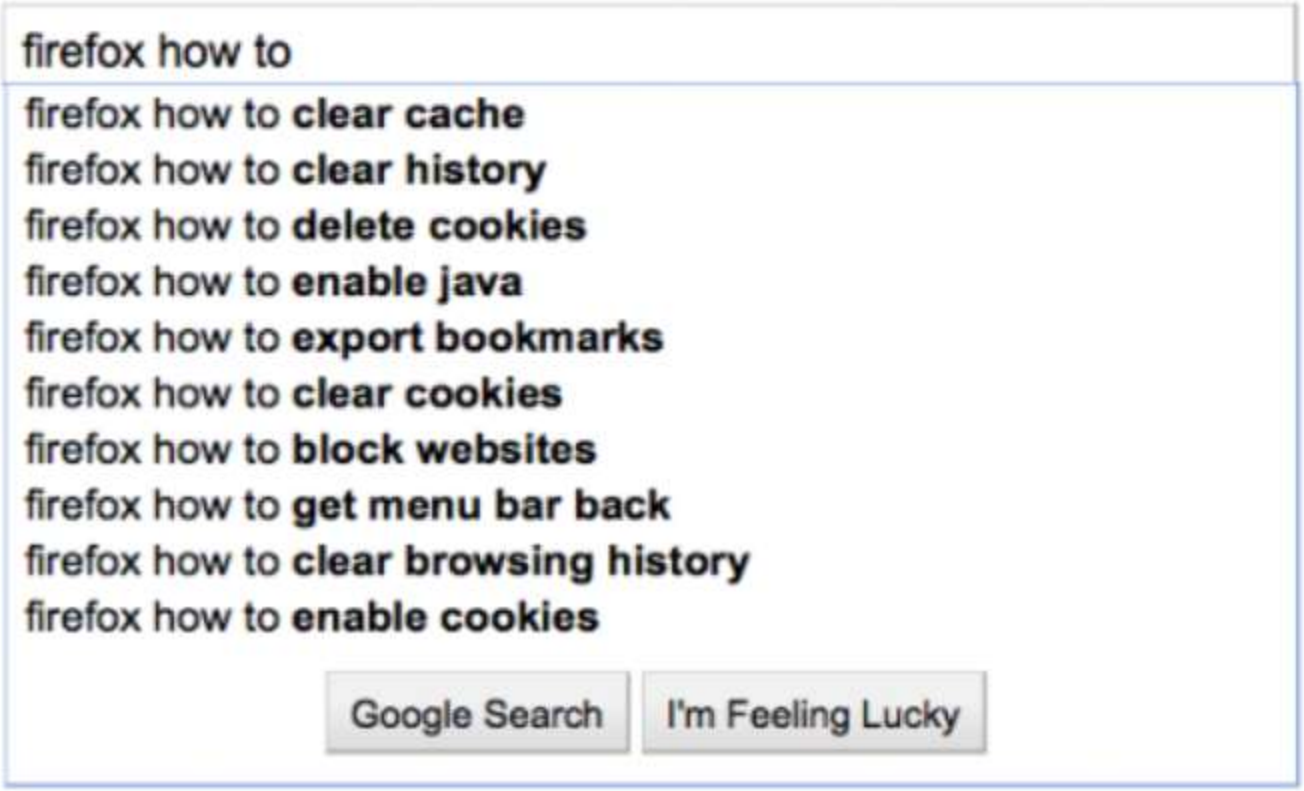
- Structured reporting
  - Specify nature of problems, source, and importance for user
  - Should also include remedies
- Unstructured reporting
  - Just report observations and categorization of problem areas reported afterwards
- Predefined categorization
  - Start out with list of problem categories and get experts to report problems in these categories

Recall: A Design Space for Evaluation



Some UWaterloo Research

- Adam Fourney and Mike Terry
  - Mine Google suggest



Recall: A Design Space for Evaluation