Closing the Loop:
From Analysis to Design

Scott Klemmer

TAs: Marcello Bastea-Forte, Joel Brandt, Neil Patel, Leslie Wu, Mike Cammarano

15 November 2007

http://cs147.stanford.edu

Analyzing the results

- Quantitative data, which might include:
  - success rates
  - time to complete tasks
  - pages visited
  - error rates
  - ratings on a satisfaction questionnaire
- Qualitative data, which might include:
  - notes of your observations about the pathways participants took
  - notes about problems participants had (critical incidents)
  - notes of what participants said as they worked
  - participants' answers to open-ended questions

Source: Usability.gov

Using the Test Results

- Summarize the data
  - make a list of all critical incidents
    - positive & negative
  - include references back to original data
  - try to judge why each difficulty occurred
- What does data tell you?
  - UI work the way you thought it would?
    - users take approaches you expected?
  - something missing?

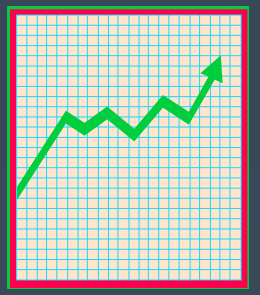
Using the Results (cont.)

- Update task analysis & rethink design
  - rate severity & ease of fixing CIs
  - fix both severe problems & make the easy fixes
- Will thinking aloud give the right answers?
  - not always
  - if you ask a question, people will always give an answer, even it is has nothing to do with facts
  - try to avoid specific questions

Measuring Bottom-Line Usability

- Situations in which numbers are useful
  - time requirements for task completion
  - successful task completion
  - compare two designs on speed or # of errors
- Ease of measurement
  - time is easy to record
  - error or successful completion is harder
    - define in advance what these mean
- Do not combine with thinking-aloud. Why?
  -

Analyzing the Numbers



- Example: trying to get task time <=30 min.
  - test gives: 20, 15, 40, 90, 10, 5
  - mean (average) = 30
  - median (middle) = 17.5
  - looks good!

Analyzing the Numbers (cont.)

- This is what statistics is for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55

Analyzing the Numbers (cont.)

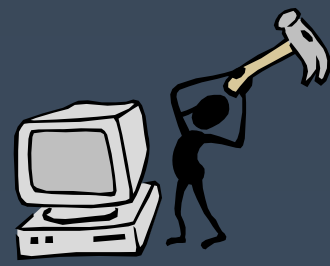
Web Usability Test Results						
Participant #		Time (minutes)				
1		20				
2		15				
3		40				
4		90				
5		10				
6		5				
	number of participants	6				
	mean	30.0				
	median	17.5				
	std dev	31.8				
	standard error of the mean	= stddev / sqrt (#samples)			13.0	
	typical values will be mean +/- 2*standard error			--> 4 to 56!		
	what is plausible? = confidence (alpha=5%, stddev, sample size)	25.4	--> 95% confident between 5 & 56			

Analyzing the Numbers (cont.)

- This is what statistics is for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55
- Usability test data is quite variable
  - need lots to get good estimates of typical values
  - 4 times as many tests will only narrow range by 2x
    - breadth of range depends on sqrt of # of test users
  - this is when online methods become useful
    - easy to test w/ large numbers of users

Measuring User Preference

- How much users like or dislike the system
  - can ask them to rate on a scale of 1 to 10
  - or have them choose among statements
    - "best UI I've ever…", "better than average"…
  - hard to be sure what data will mean
    - novelty of UI, feelings, not realistic setting …
- If many give you low ratings -> trouble
- Can get some useful data by asking
  - what they liked, disliked, where they had trouble, best part, worst part, etc. (redundant questions are OK)

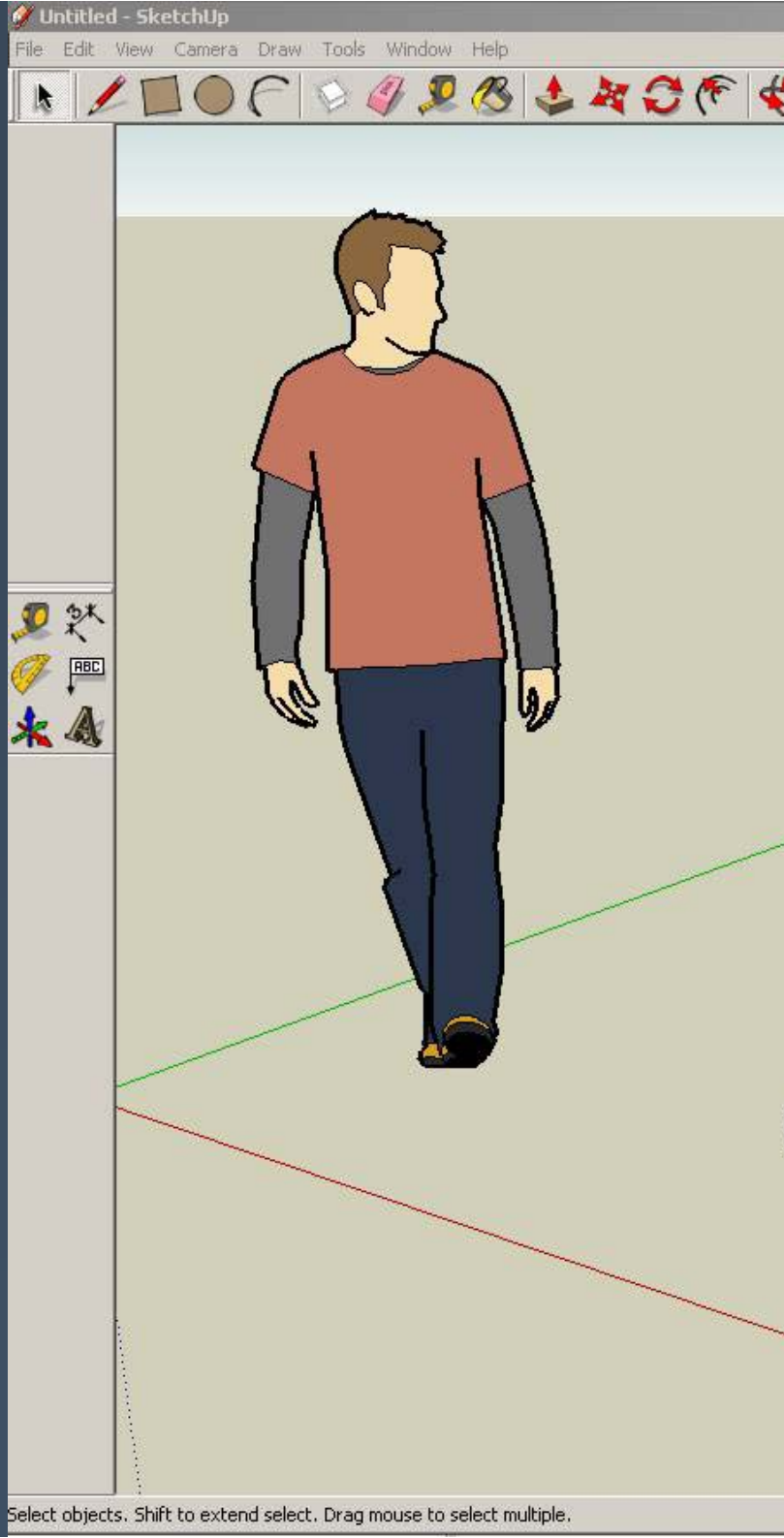


Reporting the Results

- Report what you did & what happened
- Images & graphs help people get it!
- Video clips can be quite convincing

CASE STUDY

David Akers evaluation of Google SketchUp



Study Goals

1. What individual differences (previous software used, computer use, spatial reasoning ability, etc.) best predict performance on simple modeling tasks?
2. What usage log metrics (e.g. frequency of undo operations, frequency of camera operations, etc.) best predict performance on simple modeling tasks?
3. What specific problem do novice SketchUp users encounter most frequently on simple modeling tasks?

n = 54

90% students

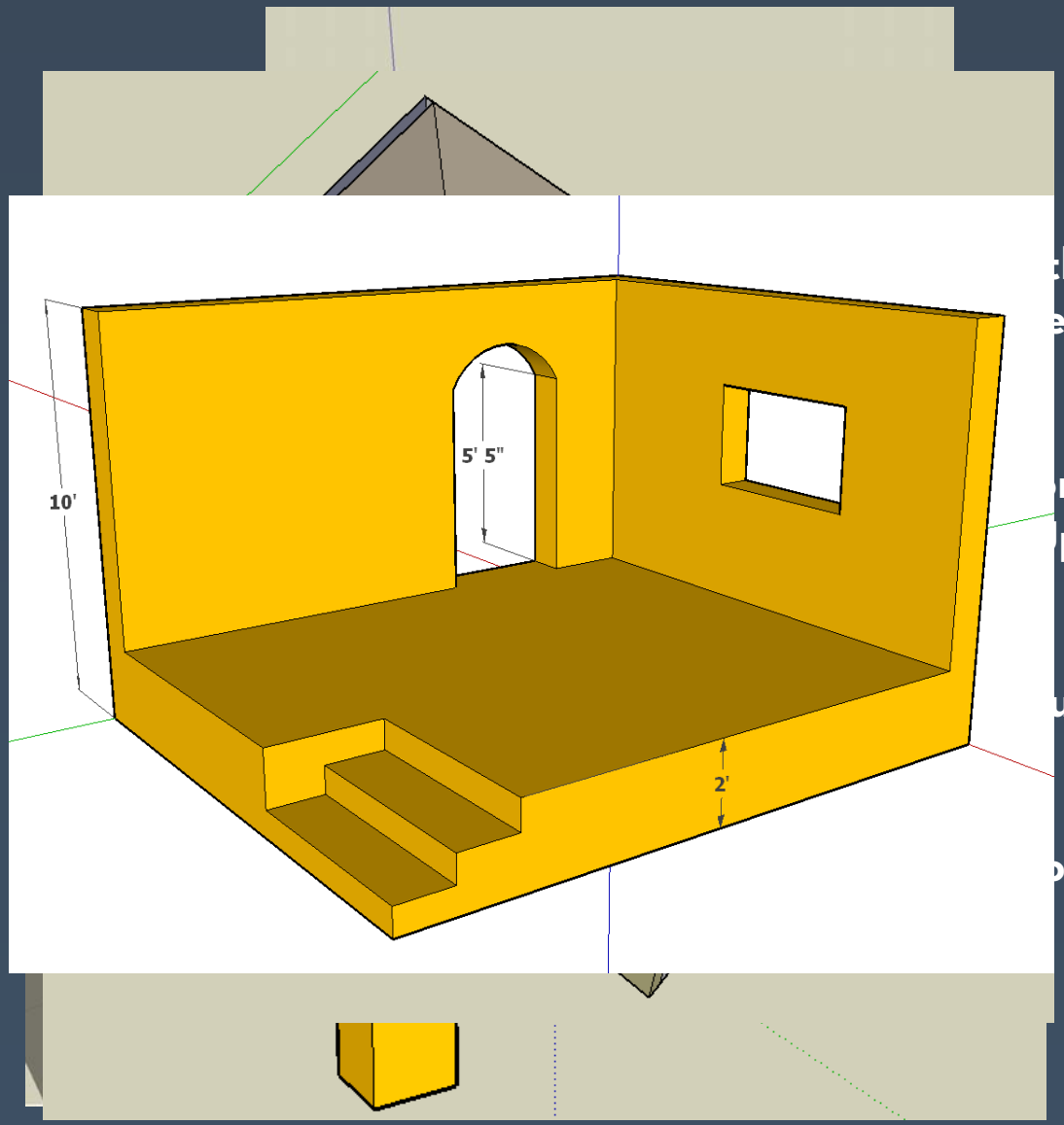
35% architecture
20% computer science
10% mechanical engineering
10% civil engineering
25% other (art, physics, etc.)

41% never used
44% novice
15% intermediate


SketchUp Usability Study – Boulder, CO
September, 2007
Participant Instruction Packet
SketchUp

Study Design

1. Entry questionnaire (5 min.)
2. Mental rotation test (15 min.)
3. Video tutorials (15 min.)
4. Free exploration (10 min.)
5. Tasks (3 x 15 min.)
6. Exit questionnaire (5 min.)



Data Size

Event log data (450 MB)

Screen capture video (75 GB!)

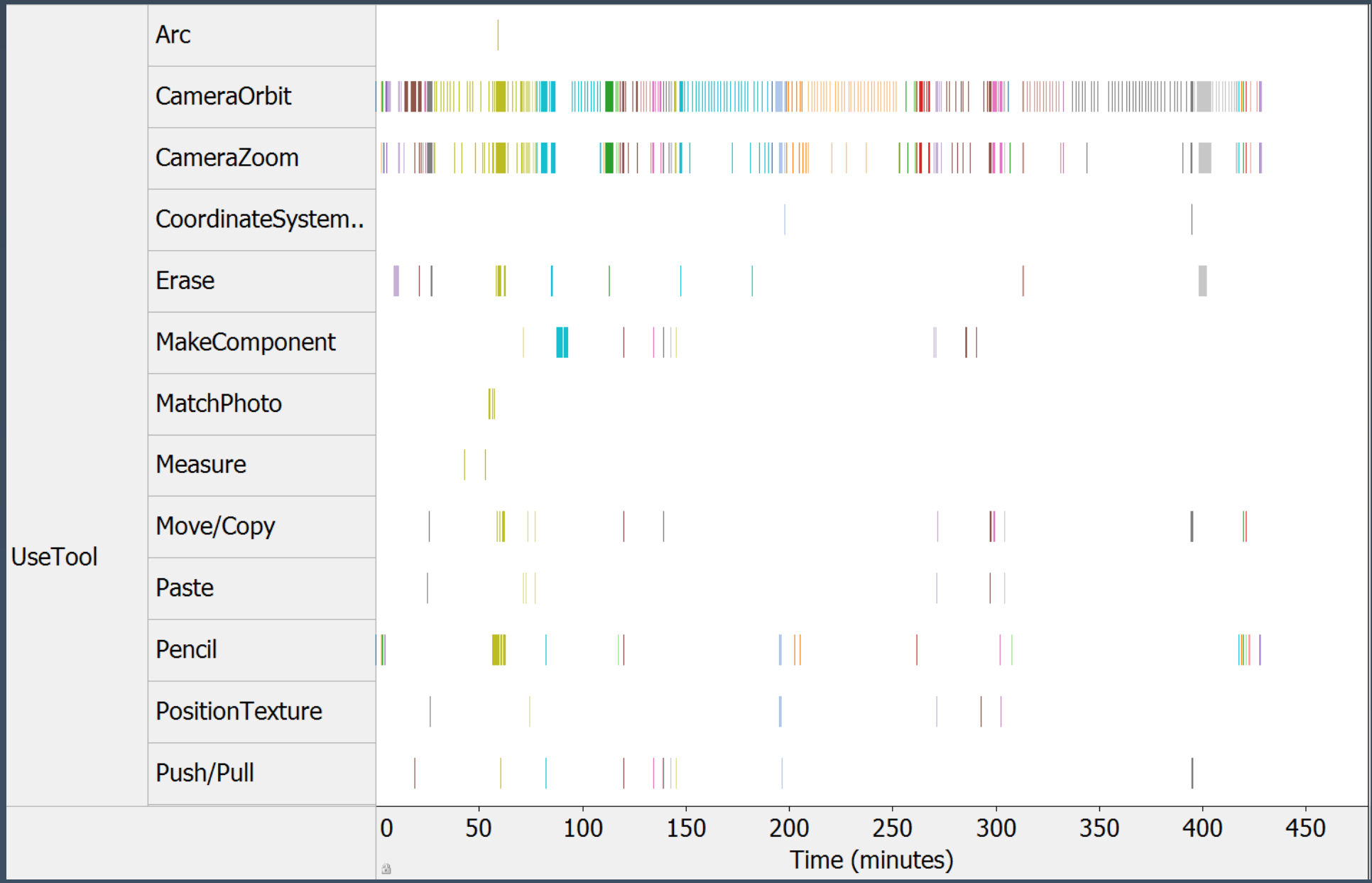
3D models (100 MB)

Questionnaires (17 KB)

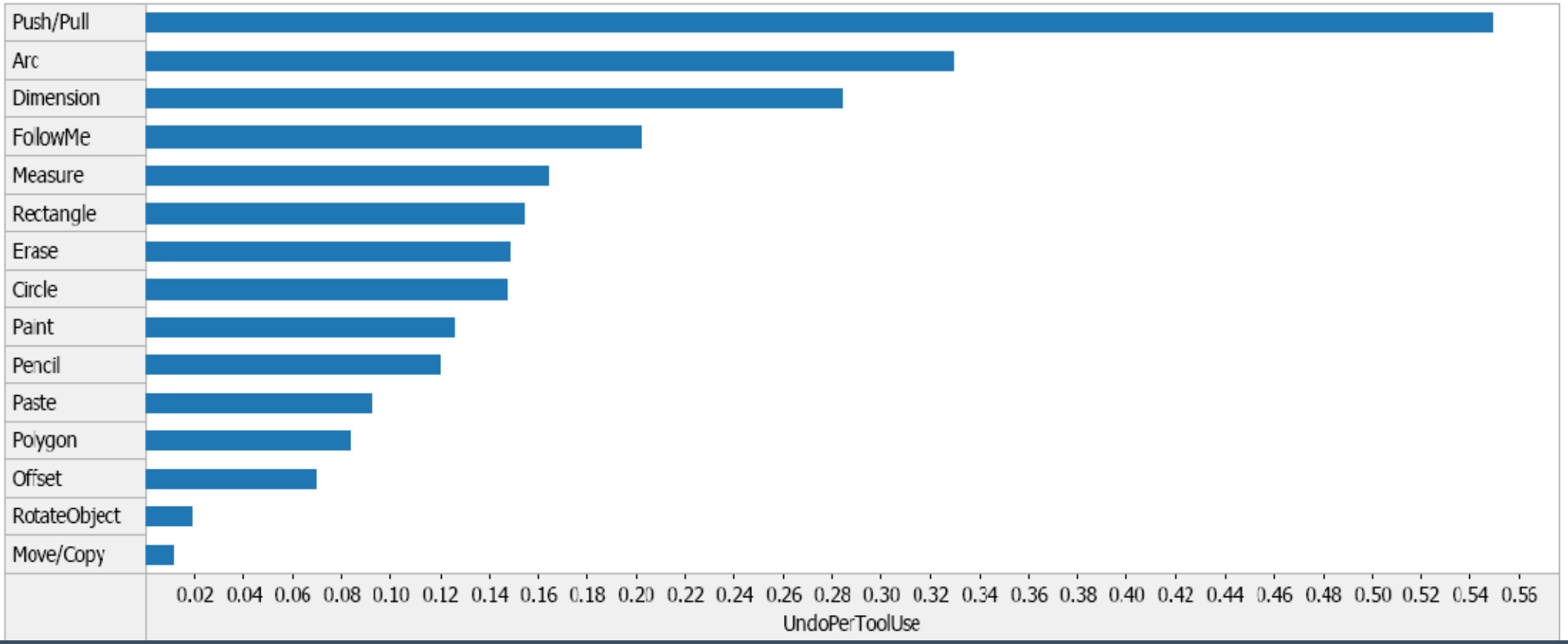
Study Goals

1. What individual differences (previous software used, computer use, spatial reasoning ability, etc.) best predict performance on simple modeling tasks?
2. What usage log metrics (e.g. frequency of undo operations, frequency of camera operations, etc.) best predict performance on simple modeling tasks?
3. **What specific problem do novice SketchUp users encounter most frequently on simple modeling tasks?**

Log Analysis of Tool Usage



Undo Rates

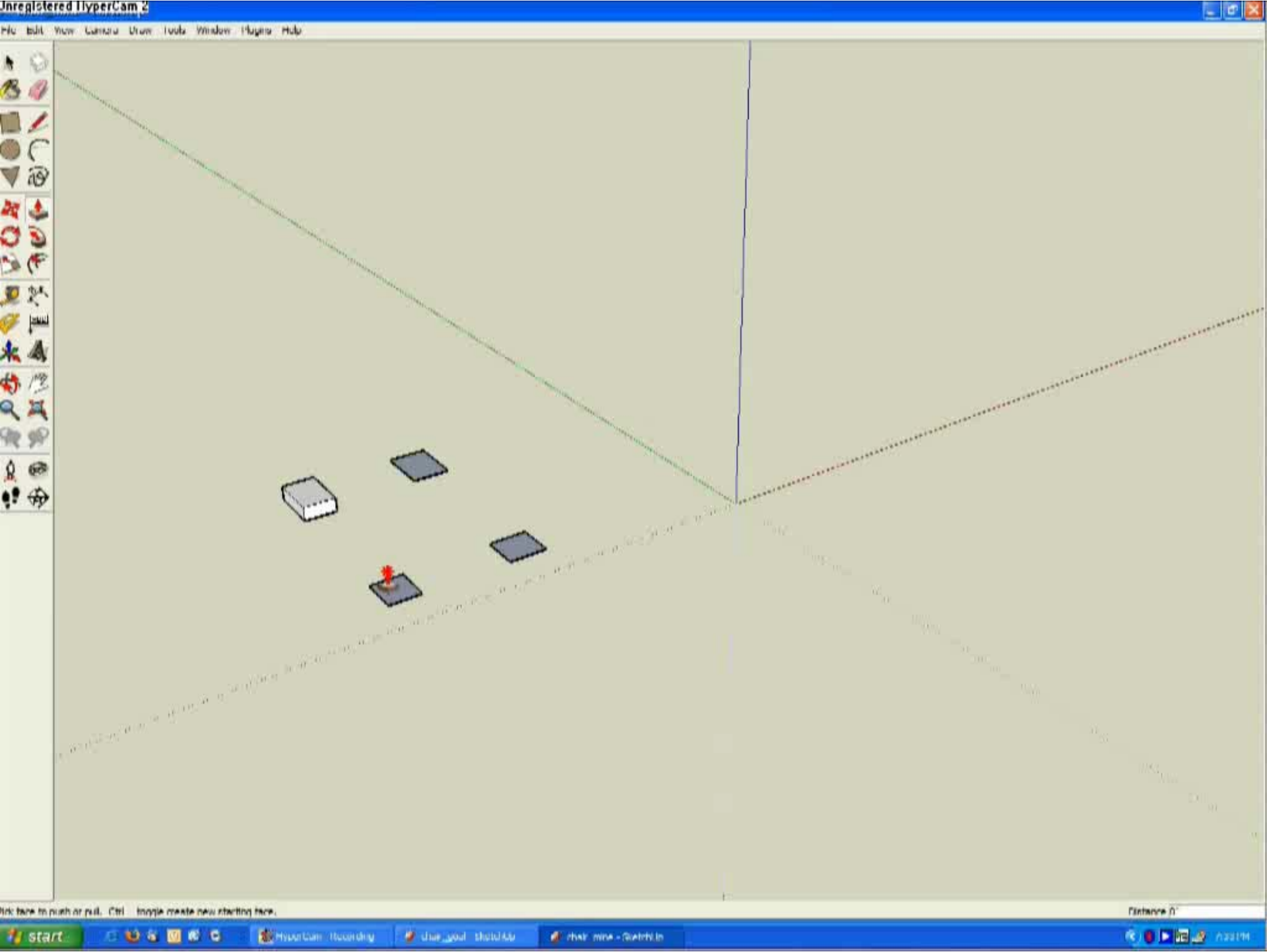

Push/Pull
Arc
Dimension
FolowMe
Measure
Rectangle
Erase
Circle
Paint
Pencil
Paste
Polygon
Offset
RotateObject
Move/Copy
0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 0.32 0.34 0.36 0.38 0.40 0.42 0.44 0.46 0.48 0.50 0.52 0.54 0.56
UndoPerToolUse

Developer Hypotheses (Wrong)

For the Push/Pull tool:

90% of undo operations are caused by bugs in SketchUp.

10% of undo operations are caused by difficulties with inferencing.


Unregistered HyperCam 2
File Edit View Camera Draw Tools Window Plugins Help

EYE TO THE FUTURE
Instrumenting Applications

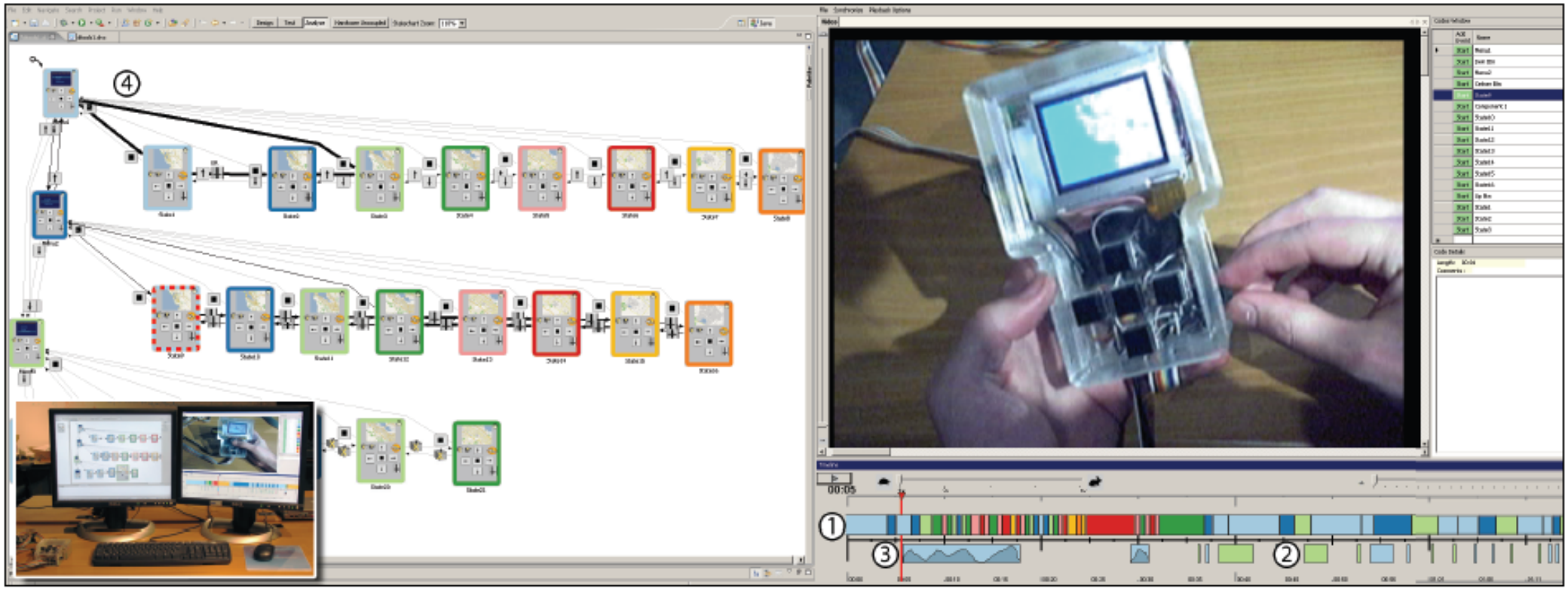


Figure 3 In Analysis mode, statechart and recorded video are synchronized and each can be used to access the other. *Inset:* simultaneous interaction with statechart and video editing is possible on a dual-screen workstation.

d.Tools physical prototyping captures user tests

Source: Hartmann, B., Klemmer, S.R., Bernstein, M., Abdulla, L., Burr, B., Robinson-Mosher, A., Gee, J. Reflective physical prototyping through integrated design, test, and ana
Proceedings of UIST 2006, October 2006

Challenges (1/8 – *from Grudin*)

- **Disparity of Work and Benefit**
  Groupware applications often require additional work from individuals who do not perceive a direct benefit from the use of the application

Challenges (2/8)

- **Critical Mass and Prisoner's Dilemma**
  Groupware may not enlist the "critical mass" of users required to be useful, or can fail because it is never to any one individual's advantage to use it


Group Calendaring
Calendar Microsoft Outlook
Type a question for help
Full week
October 31-November 04, 2005
Search...
All Calendar Items
My Calendars
Calendar in Rob Young
People's Calendars
Mark Hassall
Dennis Bye
Roan Kang
Other calendars
TeamSite-team calendar
Calendar Actions
Open a Shared Calendar...
Browse Calendars Online
Share My Calendar...
Send Calendar via E-mail...
Add New Group
Mail
Calendar
Contacts
Tasks
Calendar
TeamSite-team calendar
31 Monday
1 Tuesday
2 Wednesday
3 Thursday
4 Friday
Elena visit
OOF
Halloween; United States
Election Day; United Stat
8 am
9 00
10 00
11 00
12 pm
1 00
2 00
3 00
4 00
5 00
Sales update
36/2078
Mark Hassall
Budget review
Updated:
16/3043
Jeff Smith
Research Review -- Tin
Conf Room 36/2013 (14) A
Guy Bilbert
FW: Weekly WSS Updat
16/3043
Burke Fewel
Design Checkpoint: Ren
Conf Room 16/2369 (16) A
Florian Voss
FW: Progress update
16/3043
Luis Sousa
1-1; Jeff's office; Rob Yo
Outloc Beta 1 Team 36/401 ; Ric Gray
Outloc Beta 1 Team 36/401 ; Chris Gray
Go ove
Finalize
Liz Keys
Sales Team Meeti 1 LOCAT CHANG 36/337
Design Office Hours 16/Atric (under the stairs) Jeff Teper
Budget Review for Marketing team, 36/4013 Chris Gray
FW: Tea Lun TBD Aar
Offi 36/: Nicle
New Outlook Chrome: W
Rob's office
Josh Edwards
Mike's 1:1
Rob's Office
Rob Young
Marketing PM Meeting
36/1339
Leads Meeting
Design Lab
Joe Andreshak
FW: Office I
36/1339
Joe Andres
Sales meeting
McKinley
Employee Meeting
33
Updated: O
Conf Room 3
Giovanni M
Common n
Mike's soccer game
Pick up
68 Items
All folders are up to date.
Connected

Challenges (3/8)

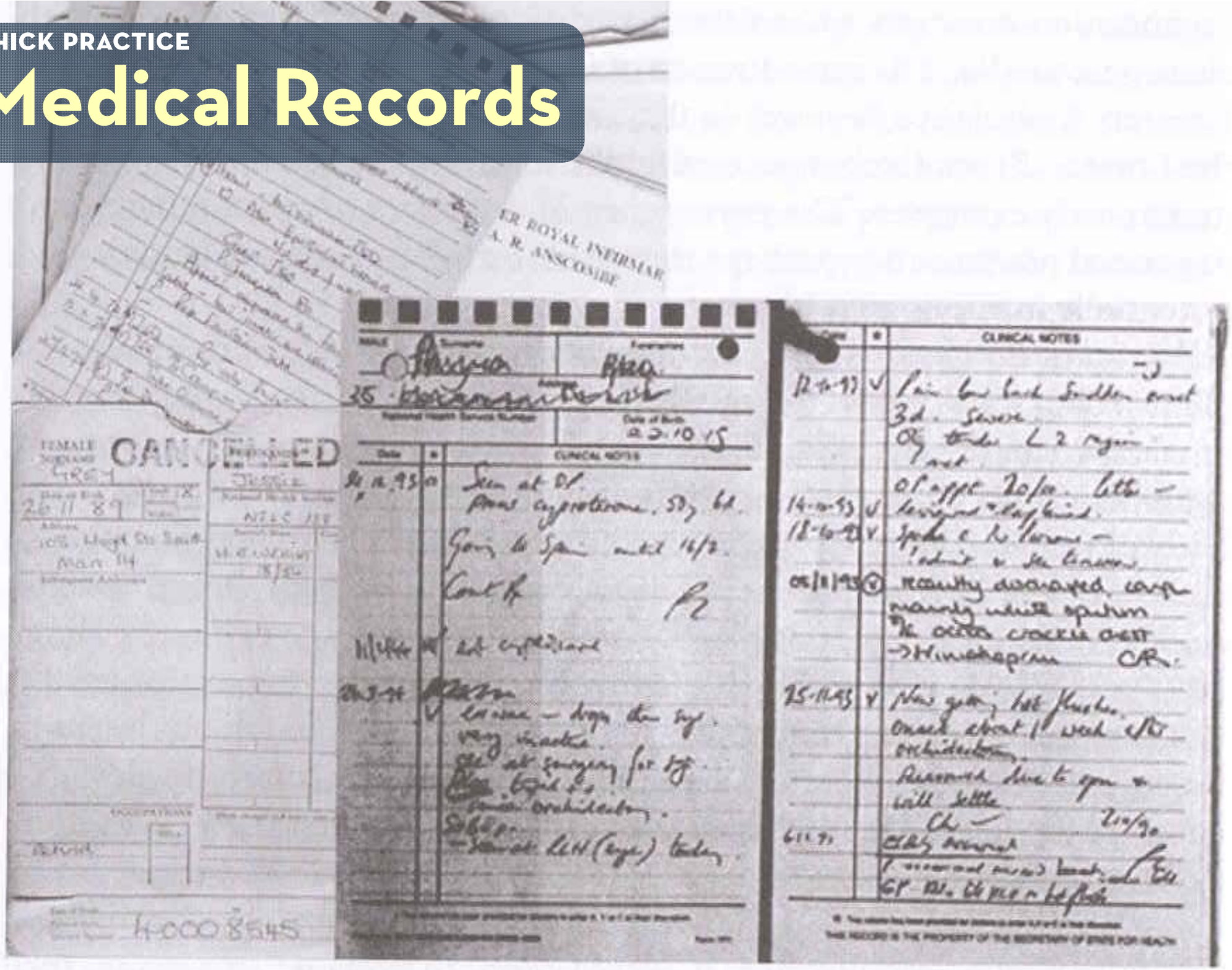
- **Disruption of Social Processes**
  Groupware can lead to activity that violates social taboos, threatens existing political structures, or otherwise demotivates users crucial to its success

Challenges (4/8)

- **Exception Handling**
  Groupware may not accommodate the wide range of exception handling and improvisation that characterizes much group activity

THICK PRACTICE

Medical Records



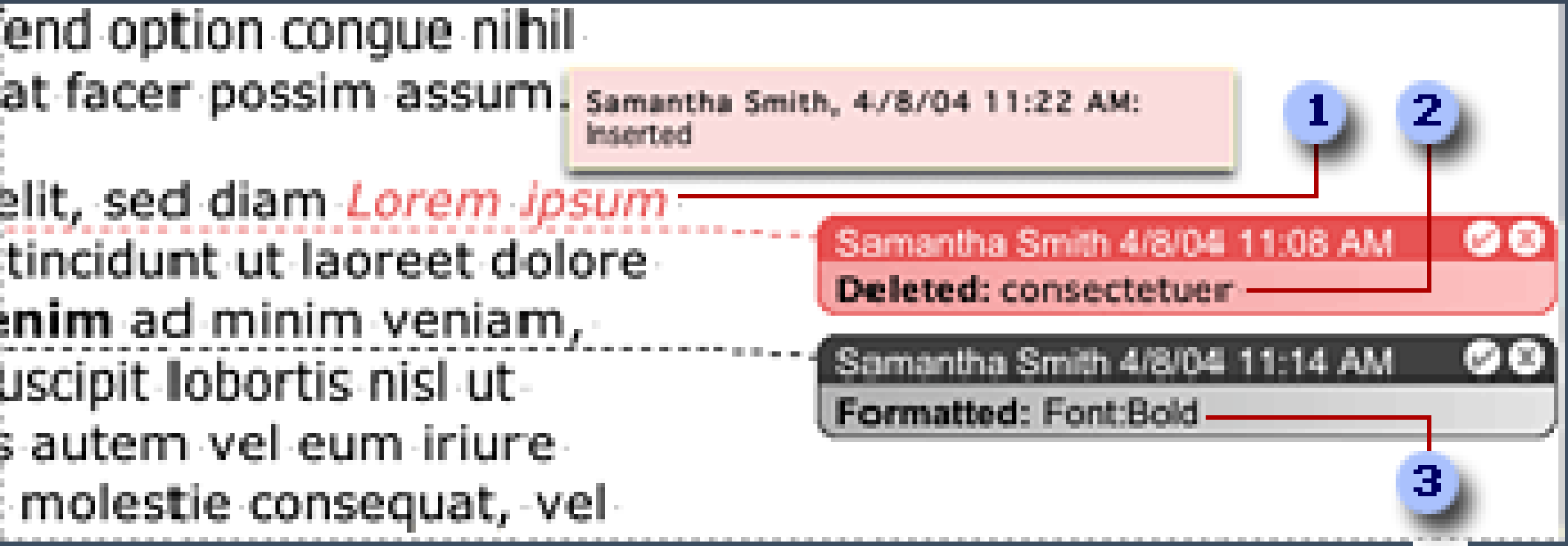
Challenges (5/8)

- **Unobtrusive Accessibility**
  Features that support group processes are used relatively infrequently, requiring unobtrusive accessibility and integration with more heavily used features.

Challenges (6/8)

- **Difficulty of Evaluation**
  The almost insurmountable obstacles to meaningful, generalizable analysis and evaluation of groupware prevent us from learning from experience

Track Changes



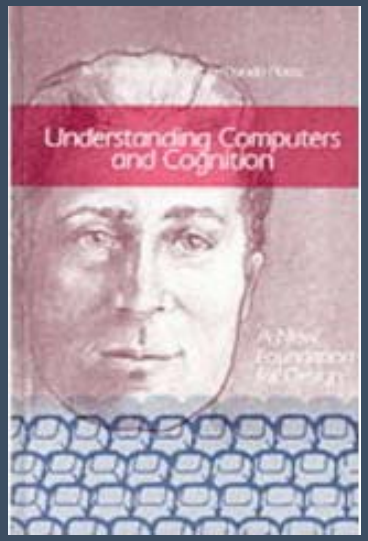
Challenges (7/8)

- **Failure of Intuition**
  Intuitions in product development environments are especially poor for multiuser applications, resulting in bad management decisions and error-prone design process.

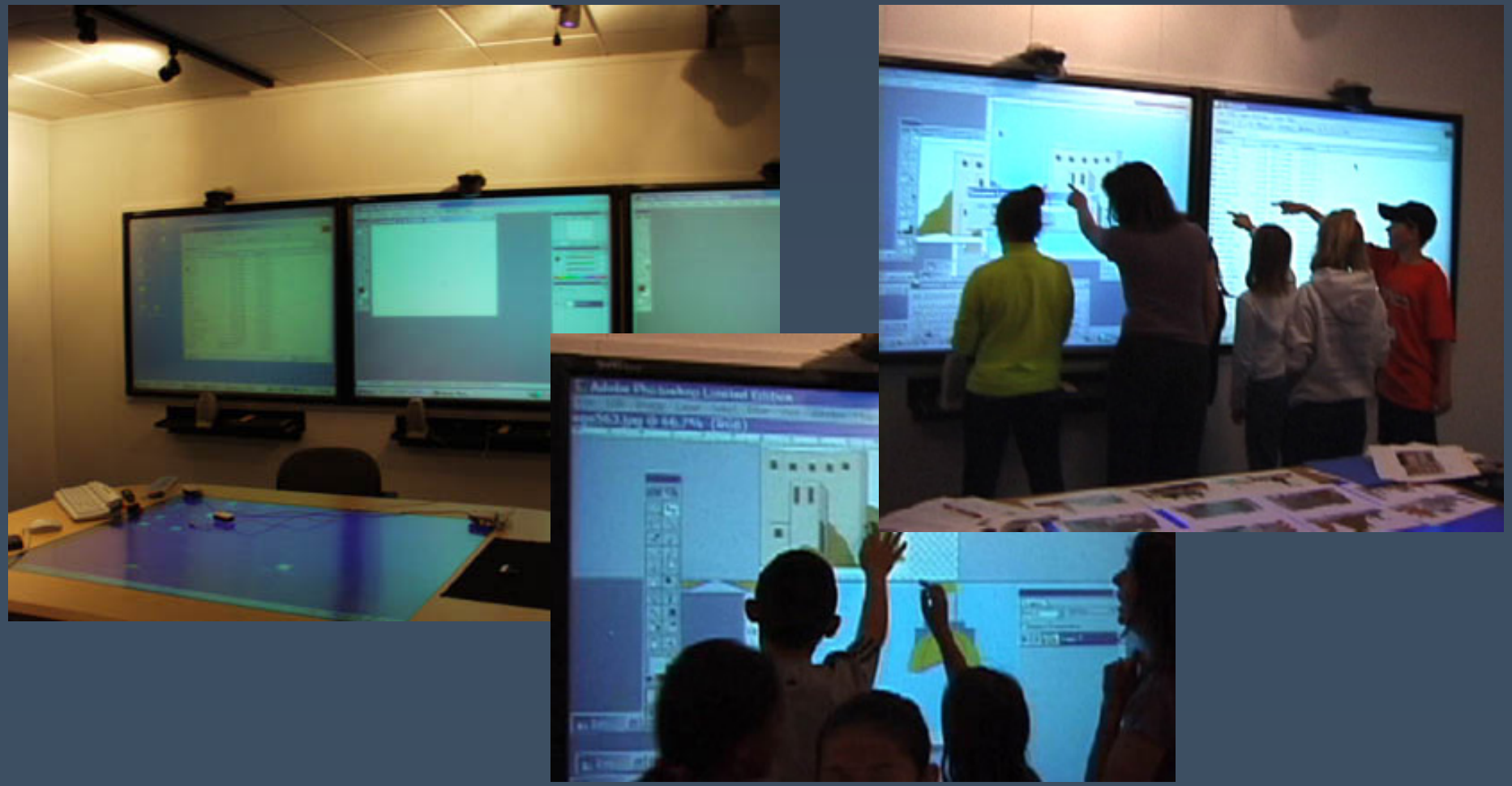
Challenges (8/8)

- **The adoption process**
  Groupware requires more careful implementation in the workplace than product developers have confronted

The Communicator



Eye to the future: iRoom



Source: Johanson, Brad and Armando Fox and Terry Winograd. “The Stanford iRoom and Interactive Workspaces Project”. Stanford.