G53FIV: Fundamentals of Information Visualization

Lecture 10: Evaluation

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

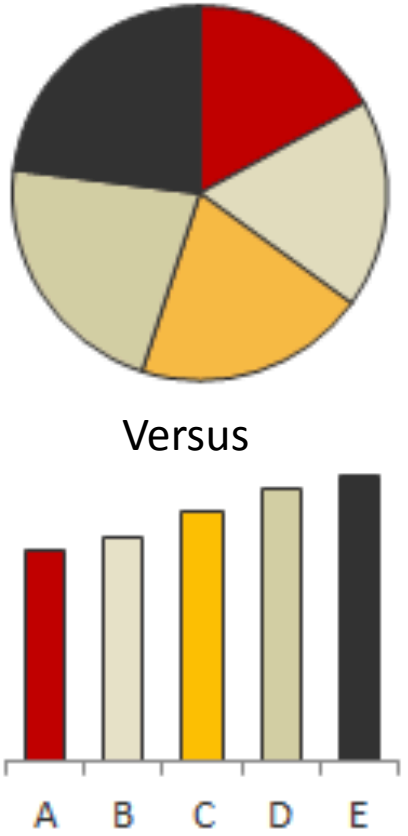
https://moodle.nottingham.ac.uk/course/view.php?id=68644

Overview

- Evaluation Methodologies
  - Controlled experiments
  - Subjective assessments
- Examples of evaluation

How do We Evaluate Visualizations?

- How do we evaluate visualizations?
  - Usability vs. Utility
- What evaluation techniques should we use?
- What do we measure?
  - What data do we gather?
  - What metrics do we use?



Evaluating Information Visualization in General

- Very difficult to compare “apples to apples”
  - Hard to compare System A to System B
  - Different tools were built to address different user tasks
- UI can heavily influence utility and value of visualization technique
- Utility vs. Aesthetics

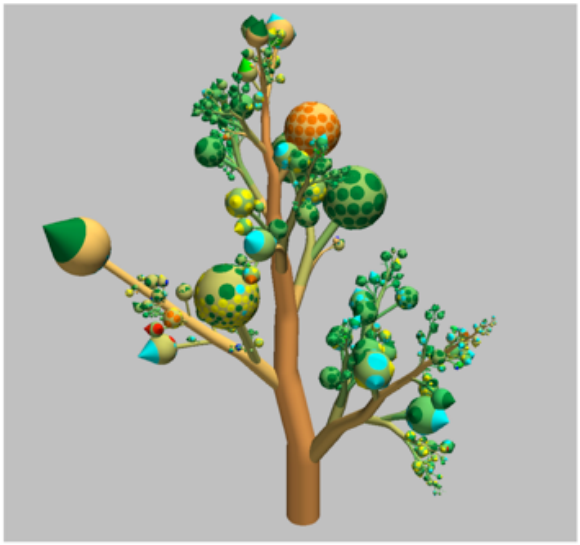


Figure 5: Botanic visualization contents of a hard disk [10, 27]. Useful or just a nice picture?

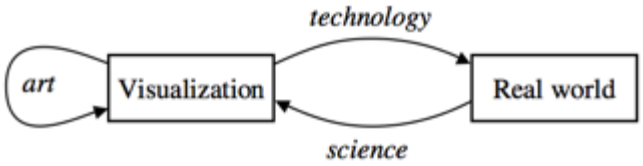


Figure 6: Views on visualization

Carpendale '08

- Challenges in information visualization evaluation
- Choosing an evaluation approach

Evaluating Information Visualizations

Sheelagh Carpendale

Department of Computer Science, University of Calgary,
2500 University Dr. NW, Calgary, AB, Canada T2N 1N4
sheelagh@ucalgary.ca

1 Introduction

Information visualization research is becoming more established, and as a result, it is becoming increasingly important that research in this field is validated. With the general increase in information visualization research there has also been an increase, albeit disproportionately small, in the amount of empirical work directly focused on information visualization. The purpose of this paper is to increase awareness of empirical research in general, of its relationship to information visualization in particular; to emphasize its importance; and to encourage thoughtful application of a greater variety of evaluative research methodologies in information visualization.

One reason that it may be important to discuss the evaluation of information visualization, in general, is that it has been suggested that current evaluations are not convincing enough to encourage widespread adoption of information visualization tools [57]. Reasons given include that information visualizations are often evaluated using small datasets, with university student participants, and using simple tasks. To encourage interest by potential adopters, information visualizations need to be tested with real users, real tasks, and also with large and complex datasets. For instance, it is not sufficient to know that an information visualization is usable with 100 data items if 20,000 is more likely to be the real-world case. Running evaluations with full data sets, domain specific tasks, and domain experts as participants will help develop much more concrete and realistic evidence of the effectiveness of a given information visualization. However, choosing such a realistic setting will make it difficult to get a large enough participant sample, to control for extraneous variables, or to get precise measurements. This makes it difficult to make definite statements or generalize from the results. Rather than looking to a single methodology to provide an answer, it will probably will take a variety of evaluative methodologies that together may start to approach the kind of answers sought.

The paper is organized as follows. Section 2 discusses the challenges in evaluating information visualizations. Section 3 outlines different types of evaluations and discusses the advantages and disadvantages of different empirical methodologies and the trade-offs among them. Section 4 focuses on empirical laboratory experiments and the generation of quantitative results. Section 5 discusses qualitative approaches and the different kinds of advantages offered by pursuing this type of empirical research. Section 6 concludes the paper.

A. Kerren et al. (Eds.): Information Visualization, LNCS 4950, pp. 19–45, 2008.

Evaluation Approaches

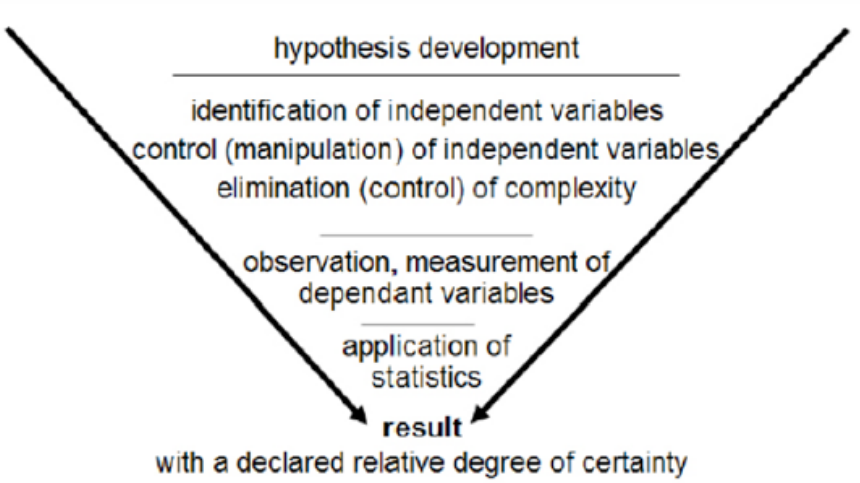
- Many different Forms
  - Qualitative, quantitative, objective, subjective, controlled experiments, interpretive observations, ...
- Two popular methodologies
  - Controlled experiments (Quantitative)
  - Subjective assessments (Qualitative)

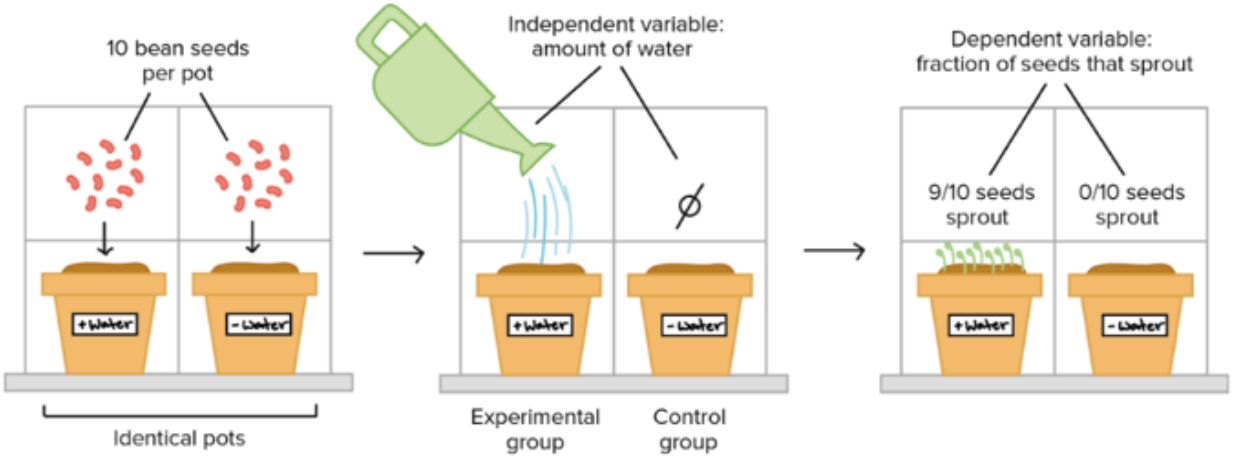
Quantitative Methods

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Quantitative Methods: Controlled Experiments

- Good for measuring performance or comparing multiple techniques
- What do we measure?
  - Performance, time, errors, …



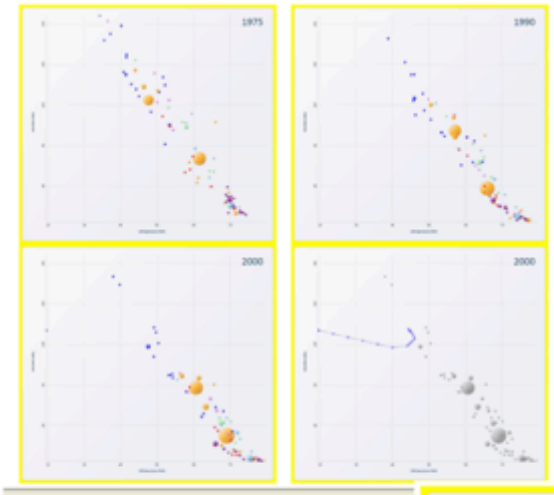


An Example: Controlled Experiment

- Run an experiment to evaluate three visualization strategies
  - Animation
  - Small multiples
  - Traces
- Especially interested in examining whether animated bubble charts are beneficial for analysis and presentation

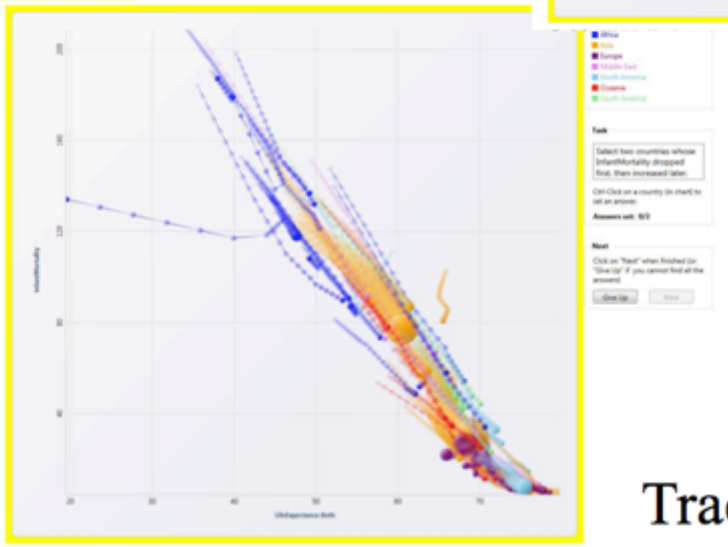
Do you remember Hans Rosling's TED talk? (Lecture 2)

Three Visualizations



Animation

Small multiples

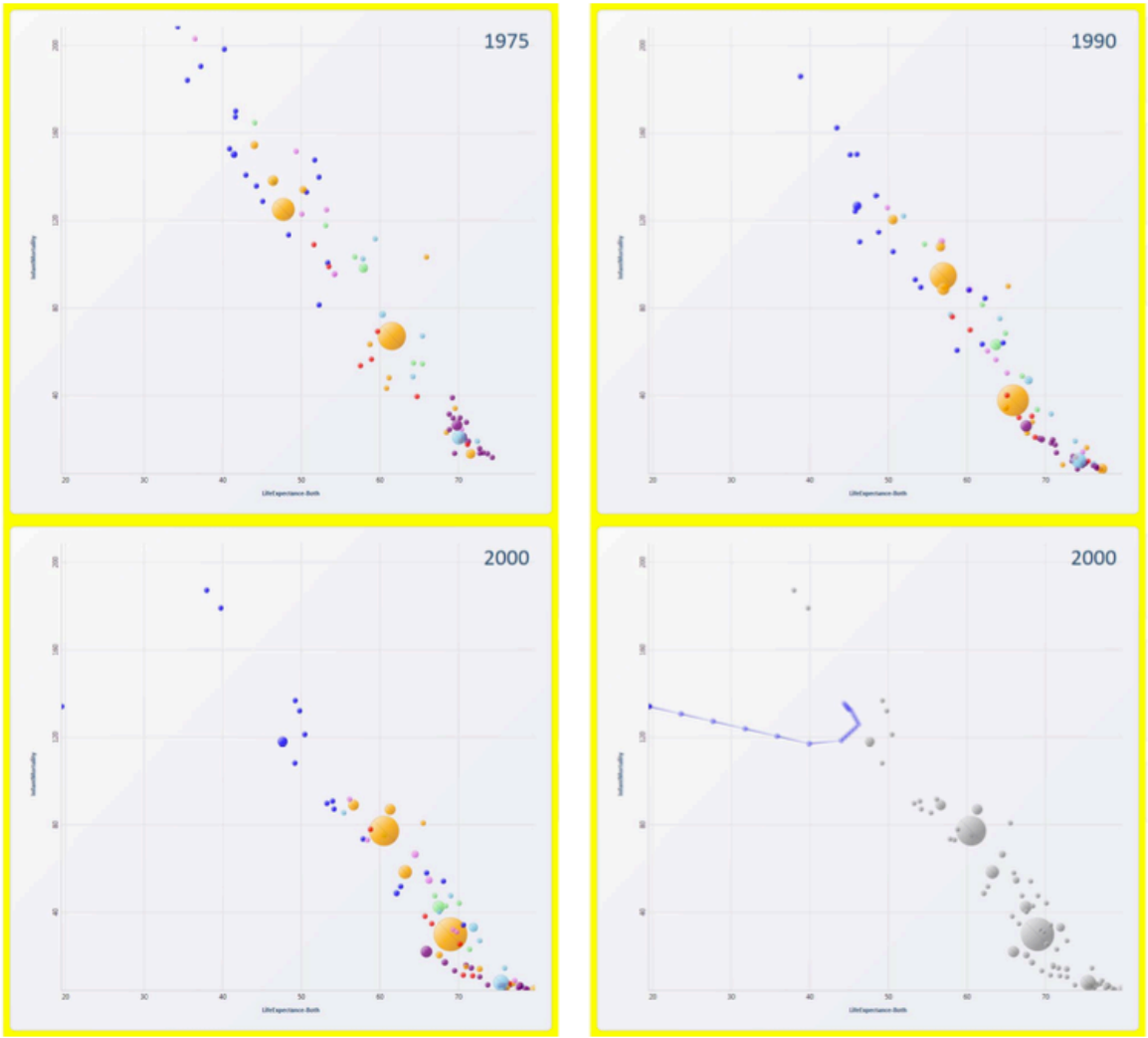


Traces

Robertson et al TVCG (InfoVis) ‘08

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

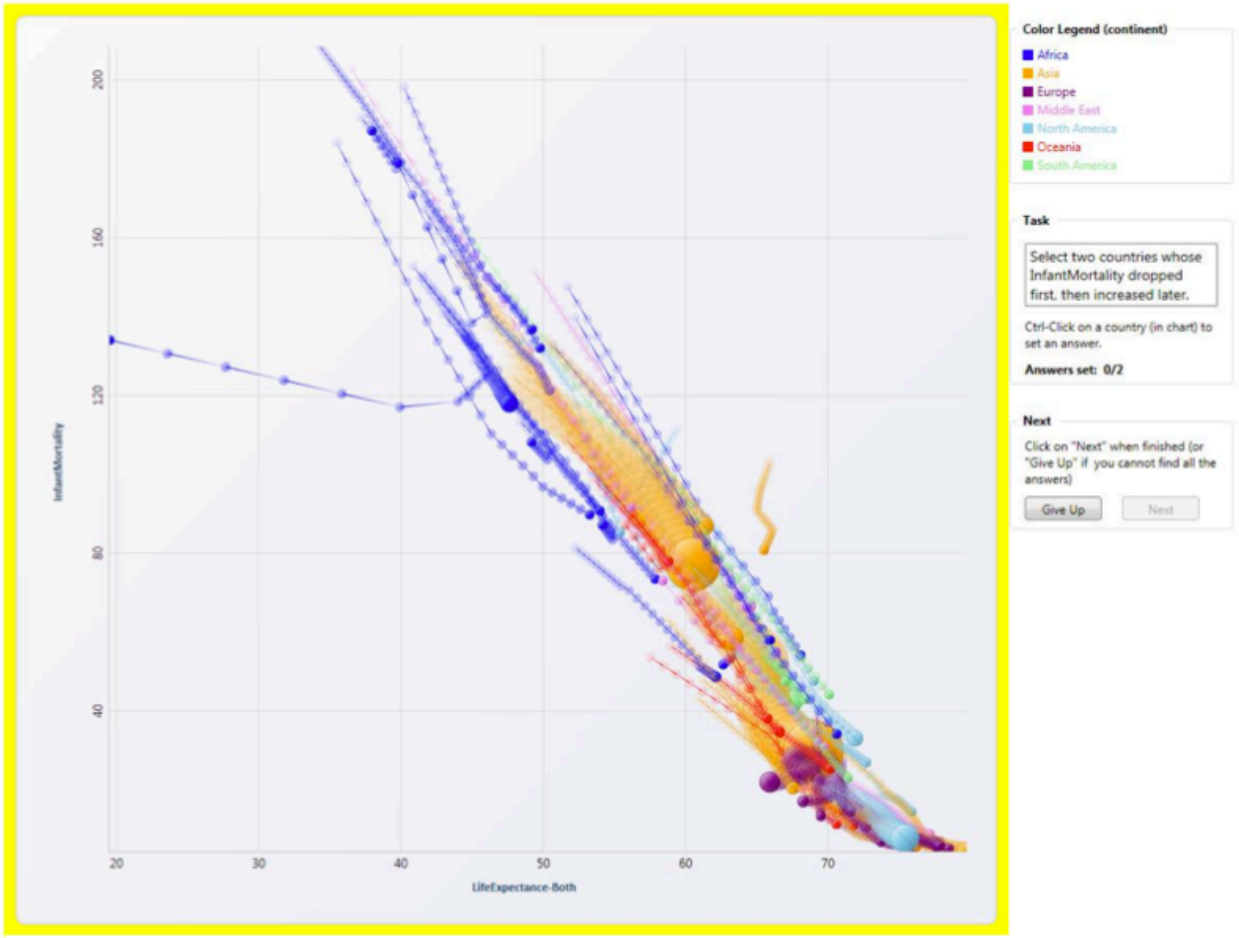
Three Visualizations: Animation



Robertson et al TVCG (InfoVis) ‘08

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Three Visualizations: Traces



Three Visualizations: Small Multiples

Experimental Design

- 3 (visualization types) x 2 (data size: small & large) x 2 (presentation vs. analysis)
  - Presentation vs analysis – between subjects
  - Others – within subjects

- Animation has 10-second default time, but user could control time slider

Experimental Design

- Data
  - Union Nations Common data about countries

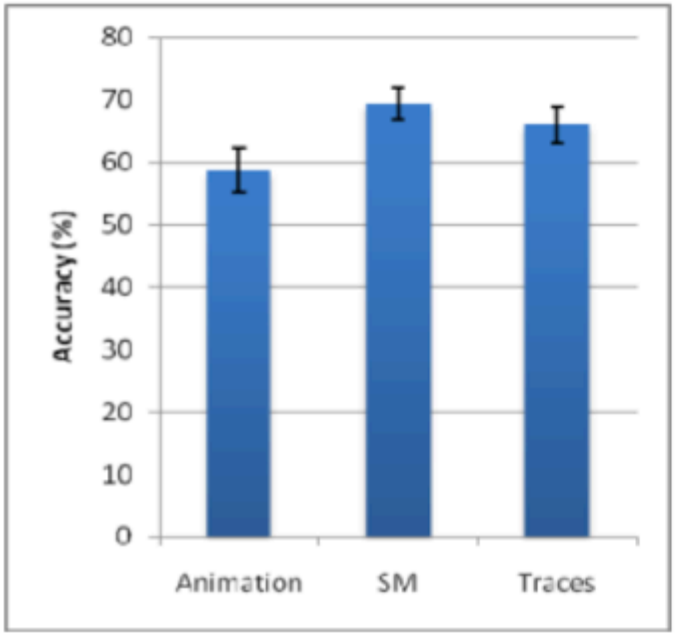
- Tasks
  - 24 tasks, 1-3 requires answers per
    - Select 3 countries whose rate of energy consumption was faster than their rate of GDP per capita growth
    - Select 2 countries with significant decreases in energy consumption
    - Which continent had the least changes in GDP per capita

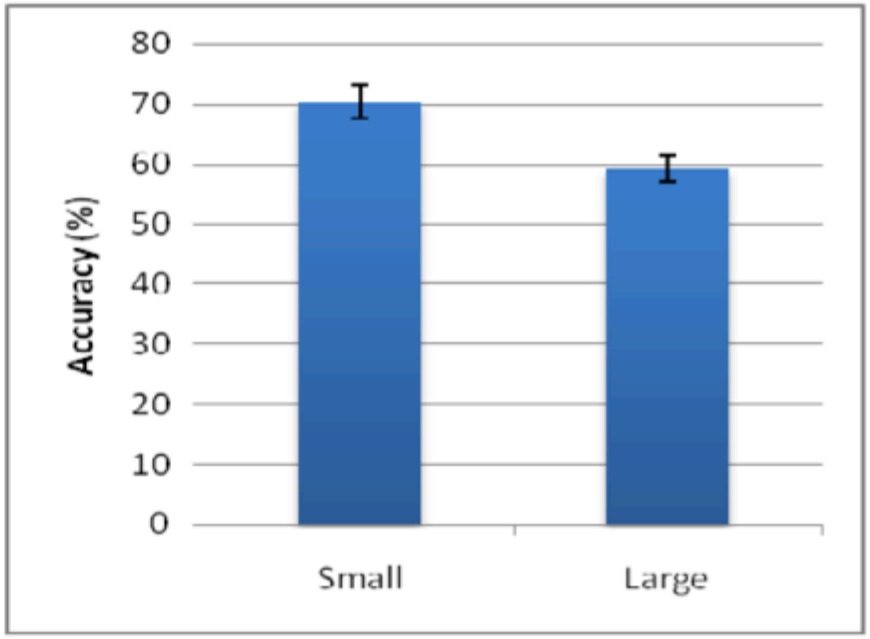
Conditions

- Analysis – straightforward, interactive

- Presentation
  - 6 participants at a time
  - Presenter described a trend relevant to task, but different
  - No interaction with system
    - In animation condition, participants saw last frame of animation (no interaction)

Results: Accuracy





- Summary
  - Small multiple is better than animation
  - Small data size is more accurate than large

Results: Speed

- Presentation
  - Animation faster than small multiples & traces 15.8 secs vs. 25.3 secs vs. 27.8 secs.

- Analysis
  - Animation slower than small multiples & traces 83.1 secs. vs. 45.69 secs. vs. 55.0 secs

Results: Subjective Ratings

Table 3. Average ratings for seven questions for each visualization. * indicates significant differences (p<.05).

	Animation	SM	Traces
Q1. The visualization was helpful to me in answering the questions.	4.6 *Traces	4.2	4.1
Q2. For the smaller dataset, I found the tasks easy using this visualization.	4.6 *SM	4.2	4.5
Q3. For the larger dataset, I found the tasks easy using this visualization.	2.6	3.4 *Traces	2.3
Q4. I enjoyed using this visualization.	4.3 *SM *Traces	3.7	3.5
Q5. I found this visualization exciting.	4.3 *SM *Traces	3.1	3.0
Q6. For the smaller dataset, I found the screen too cluttered.	1.8	1.5	2.0
Q7. For the larger dataset, I found the screen too cluttered.	4.4	2.8 *Animation *Traces	4.7

Table 4. Average ratings for a few general questions.

	Presentation	Analysis	Overall
G1. I found the Traces view enjoyable.	3.8	2.9	3.4
G3. I found the Small Multiples view enjoyable.	4.1	3.4	3.7
G5. I found the Animation view enjoyable.	4.6	5.0	4.8
G7. The animation went too fast for me.	3.2	2.8	3.0
G8. The animation went too slow for me.	1.6	1.3	1.4
G9. I lost track of some data points as they moved.	4.9	4.6	4.8

Presentation, small: Animation (9) > SM (6) > Traces (3)
Presentation, large: Traces (8) > SM (6) > Animation (4)
Analysis, small: Animation (7) > SM (6) > Traces (5)
Analysis, large: Animation (8) > SM (6) > Traces (4)

Likert: 0-strongly diagree, 6-strongly agree

Summary of Results

- People rated animation more fun, but small multiples was more effective.
- As data grows, accuracy becomes an issue
  - Traces & animation get cluttered
  - Small multiple gets tiny
- Animation:
  - “fun”, “exciting”, “emotionally touching”
  - Confusing, “the dots flew everywhere”

Controlled Experiments at Large

- Online A/B testing in the commercial world
  - https://www.coursera.org/learn/ui-testing/lecture/pMhKt/industry-practice-massive-a-b-testing-interview-with-ronny-kohavi

- A very widely used methods in evaluating developed new systems/algorithms

Quantitative Challenges

- Conclusion Validity
  - Is there a relationship?
- Internal Validity
  - Is the relationship causal?
- Construct Validity
  - Can we generalize to the constructs (ideas) the study is based on?
- External Validity
  - Can we generalize the study results to other people/places/times?
- Ecological Validity
  - Does the experimental situation reflect the type of environment in which the results will be applied?

Qualitative Methods

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Qualitative Methods

- Types
  - Nested methods
    - Experimenter observation, think-aloud protocol, collecting participant opinions
  - Inspection evaluation methods
    - Heuristics to judge
- Observational context
  - In situ, laboratory, participatory
  - Contextual interviews is important

Subjective Assessments

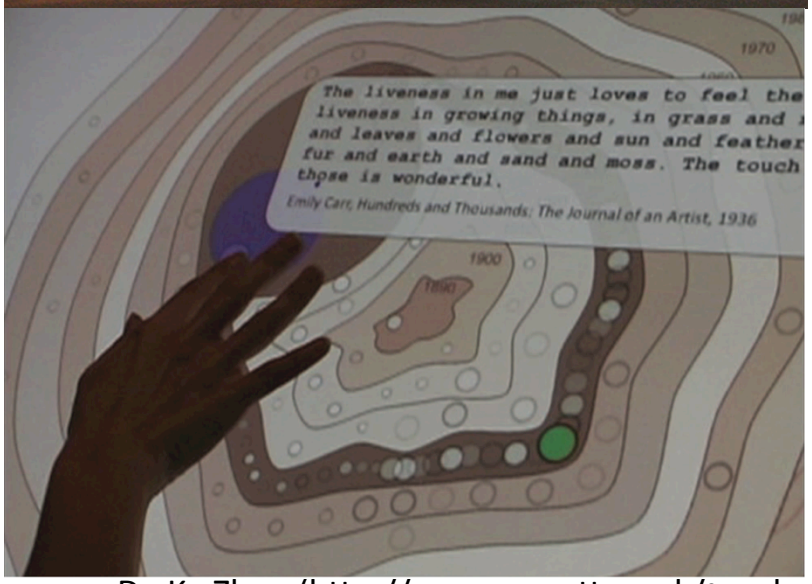
- Learn people's subjective views on tool
  - Was it enjoyable, confusing, fun, difficult, ...?
- This kind of personal judgment strongly influence use and adoption, sometimes even overcoming performance deficits

- Pros and Cons?
  - Compared to controlled experiments

An Example: Subjective Assessments

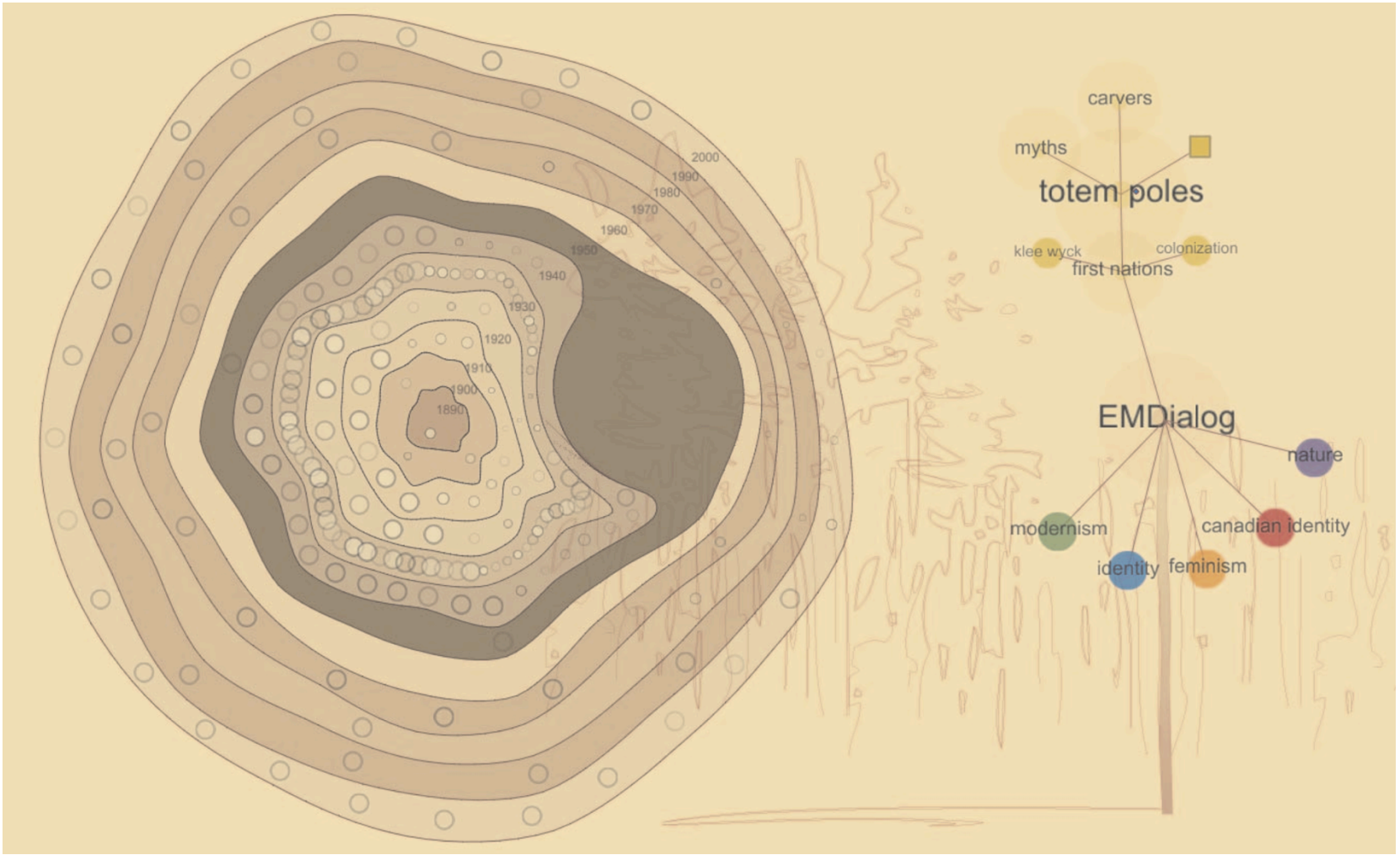
- Evaluating a newly developed visualization system (EMDialog) at the museum (Emily Carr exhibit)

- Discourse visualization
  - Time
  - Context



Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Visualization

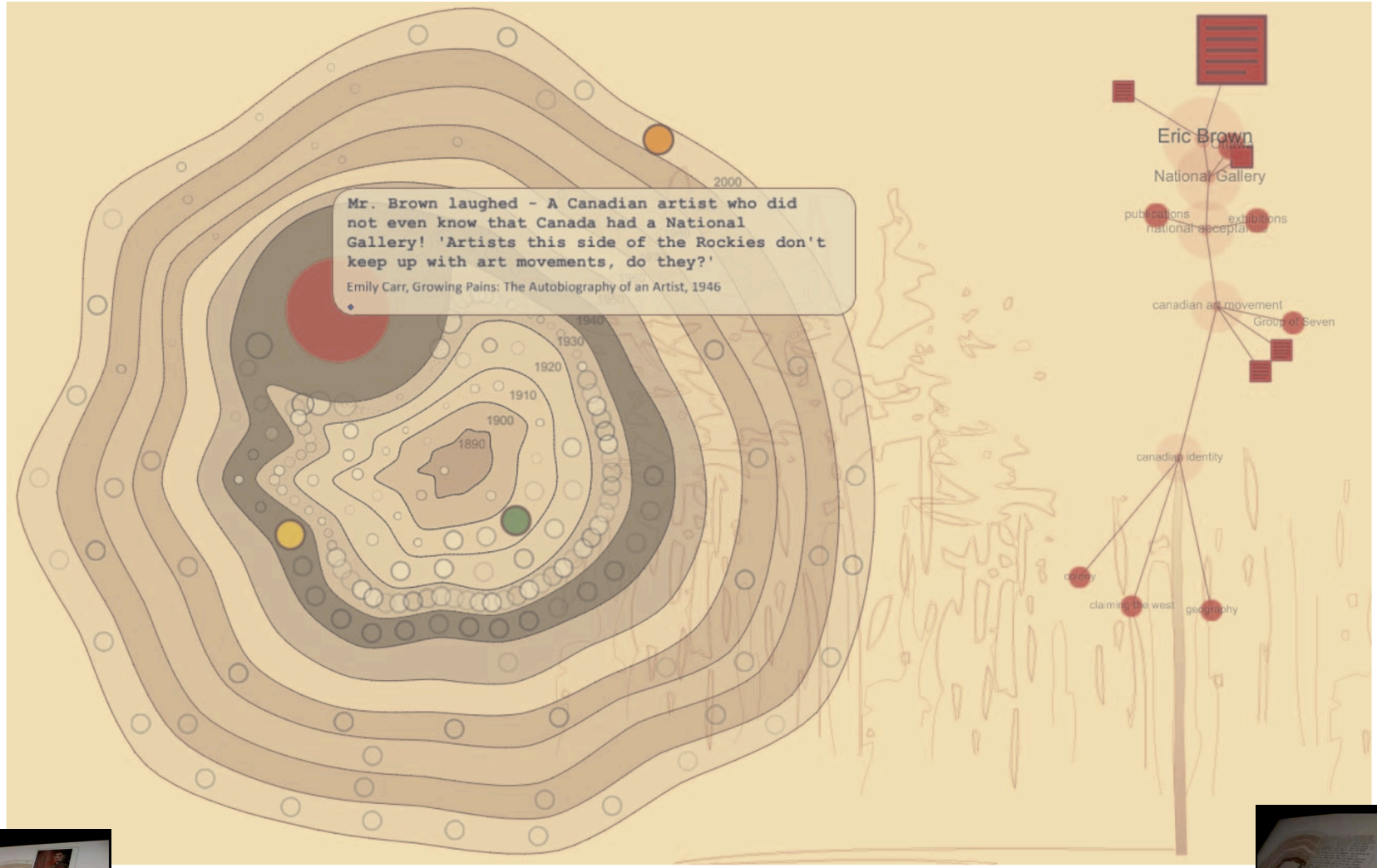


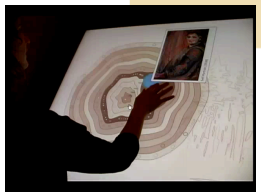
Cut section

Tree section

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

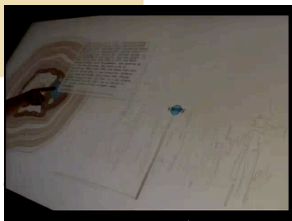
Visualization





Cut section

Tree section



Evaluation Goals

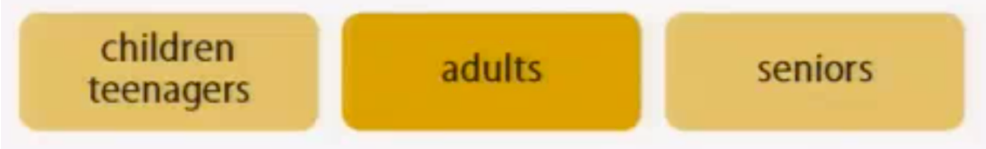
- Approach
  - How would people approach EMDialog? What would draw them toward the installation?
- Exploration techniques
  - How would they explore the information visualizations?
- Acceptance
  - What would visitors generally think of this type of information presentation in the museum context?

Experimental Design

- Emily Carr exhibition floor at the Glenbow Museum for around a month
- Open observation
  - Non-intrusive observation
  - Field notes
- Open-ended Questionnaires (voluntarily)
  - What participants liked or disliked about the installation

Approach

- Age



- Number

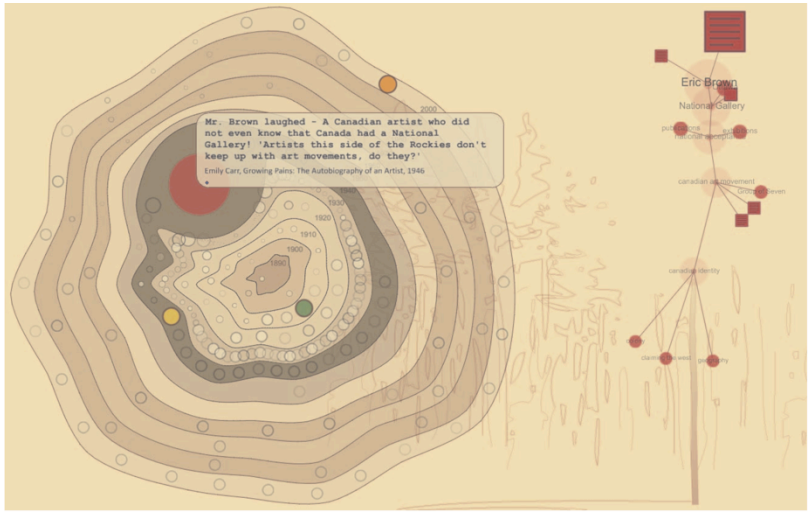
- Interaction time

- Motivation
  - Physical setup
  - Interaction of other people

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Exploration Styles

- All information at once
- Broad exploration

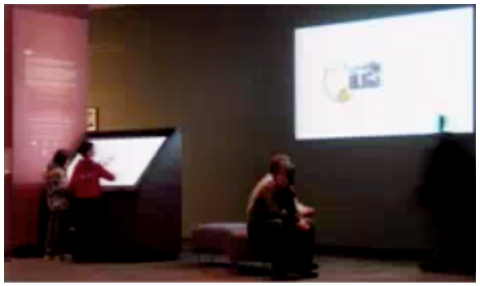


- Structured
- Interest-based exploration

- Switch between two sections
  - Visual cues
  - Missing feedback

Performance

- Visible interaction
  - Evokes curiosity
  - Teaches interaction techniques
- Awareness of being observed
  - Performing in front of strangers
  - Taking over control



I like watching other people interacting!

I am uncertain about the performance aspect, I'm kind of an introvert!

I felt guilty interacting with the display not knowing whether or not someone was in the middle of reading the projected screen!

Performance: mixed visitor reaction

- *[EMDialog] allowed me to **put Carr's work into context**.*
- ***Enhanced the museum's experience** by linking chronology and concept*
- *It allowed me to **focus on one aspect** / period of her work.*

- *It **took me a while to get the idea** (and resist fatigue after spending two hours in the exhibit) but **it quickly engaged me and was really neat and fun to use**.*

- ***Too much reading** / not enough pictures.*
- ***Totally confusing***

Take-Away for EMDialog

- Appealing information representation
- Interactivity
  - Short- and long-term exploration
  - Collaborative information exploration
  - Various exploration styles
- Leave traces in the visualization

Qualitative Challenges

- Sample sizes
- Subjectivity
- Analyzing qualitative data

Meta-evaluate Evaluation Approaches

- Desirable features
  - Generalizability
  - Precision
  - Realism

Methodology vs. Desirable Features

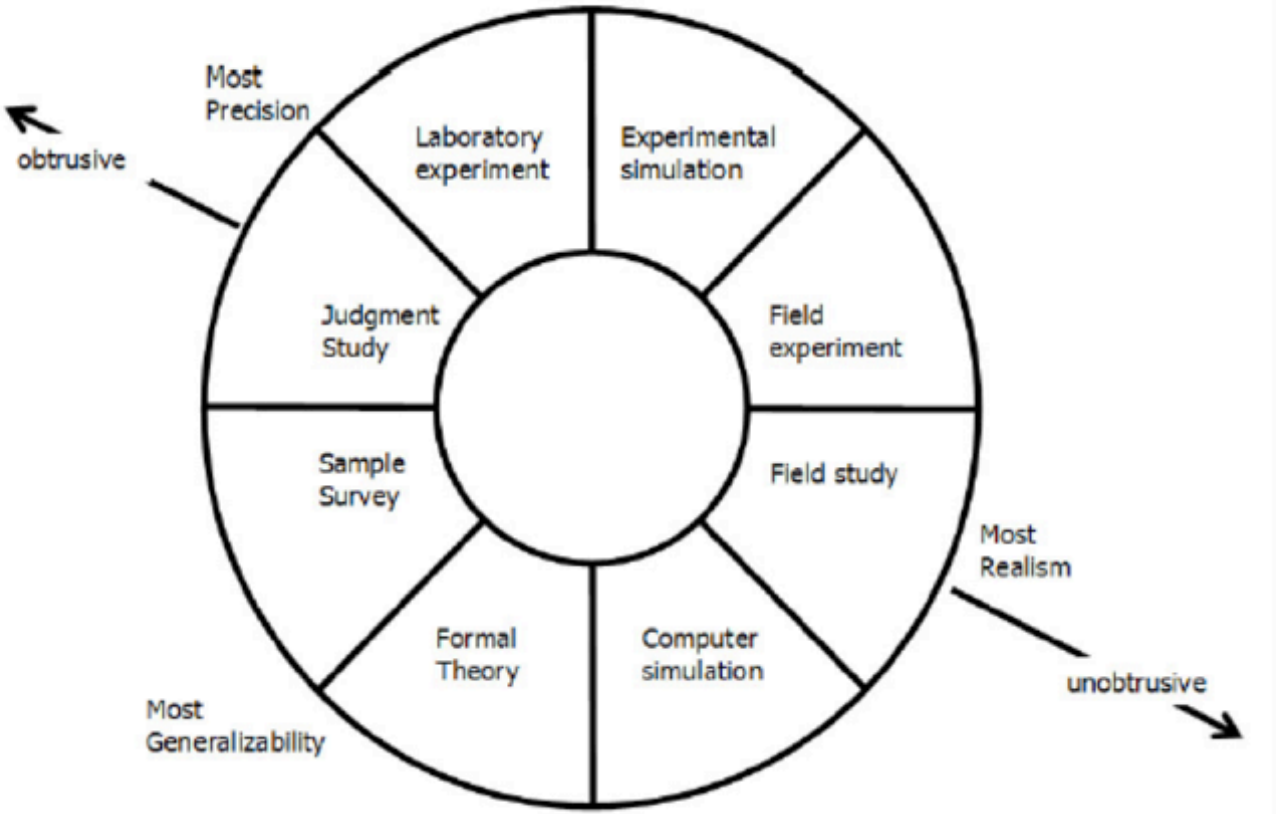


Fig. 1. Types of methodologies organized to show relationships to precision, generalizability and realism. (adapted, simplified from McGrath 1995)

Summary

Research Aspect	Quantitative	Qualitative
Common Purpose	Test Hypotheses or Specific Research Questions	Discover Ideas, used in Exploratory Research with General Research Objects
Approach	Measure and Test	Observe and Interpret
Data Collection Approach	Structured Response Categories Provided	Unstructure, Free-Form
Research Independence	Researcher Uninvolved Observer. Results Are Objective.	Researcher Is Intimately Involved. Results Are Subjective.
Samples	Large Samples to Produce Generalizable Results	Small Samples – Often in Natural Settings
Most Often Used	Descriptive and Causal Research Designs	Exploratory Research Designs

Crowd-Based Evaluation

- e.g. Amazon Mechanical Turk
- Emerging Method that enables scale
- Lots of issues

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

How to Conduct Evaluation Studies

- You can learn more about those in the following courses.
  - (G52HCI) Introduction to Human Computer Interaction
  - (G54HCI) Individual Project: Human-Computer Interaction

Summary

- Why do evaluation of InfoVis systems?
  - We need to be sure that new techniques are really better than old ones
  - We need to know the strengths and weaknesses of each tool; know when to use which tool
- Challenges
  - There are no standard benchmark tests or methodologies to help guide researchers
  - Defining the tasks is crucial
  - What about individual differences?
  - Controlled experiments vs. subjective assessments

Next Lecture

- Topic:
  - Text and Document

- The next Monday (4 Mar)
  - 12:00 - 14:00
  - A25, Business South, Jubilee Campus