Continuous Experimentation and A/B Testing: A Mapping Study

Rasmus Ros and Per Runeson
A/B Testing

- Software changes
  - User interface tweaks
  - Algorithm parameters
  - "Old" vs "new" feature
A/B Testing

User data

A

B

A/B Testing


A/B Testing

A

4.3% X

B

4.4% X

A/B Testing

A

4.3% X

11% Y

B

4.4% X

10% Y

A/B Testing

Influence

A

4.3% X

11% Y

B

4.4% X

10% Y

A/B Testing

A/B Testing

A
B

4.3% X
11% Y
...
Z

4.4% X
10% Y
...
Z

A/B Testing

Wait, there’s more!
• Learning effects
• Impression effects
• Power calculations
• Multiple testing
• Early stopping
• …

4.3% X
11% Y
… Z

4.4% X
10% Y
… Z

A/B Testing

4.3%
11%
... Z

4.4%  X
10%  Y
... Z

Please stop!

A/B Test

- Controlled experiment
  - 2 groups
  - Metric
  - Hypothesis test

- Also A/B/n test and MVT

http://bit.ly/roose18ab
A/B Test

- Controlled experiment
  - 2 groups
  - Metric
  - Hypothesis test
- Also A/B/n test and MVT

Continuous Experimentation

- Continuous iterative process [2]
  1) Vision
  2) Business goals
  3) Experiment
  4) Learnings
- Roles, architecture, infrastructure, …
- Synergies with continuous *


Continuous Experimentation

- Continuous iterative process [2]
  1) Vision
  2) Business goals
  3) Experiment
  4) Learnings

- Roles, architecture, infrastructure, …

- Synergies with continuous *


Continuous Experimentation

- Continuous iterative process [2]
  1) Vision
  2) Business goals
  3) Experiment
  4) Learnings
- Roles, architecture, infrastructure, …
- Synergies with continuous *


Overview

- **Aim**
  - State of research
  - Applicability of CE

- **Method**
  - Phrase search and references
  - Thematic analysis

- **Result**
  - 62 papers

Overview

- **Aim**
  - State of research
  - Applicability of CE

- **Method**
  - Phrase search and references
  - Thematic analysis

- **Result**
  - 62 papers

Overview

- **Aim**
  - State of research
  - Applicability of CE

- **Search**
  - Phrase search and references
  - Thematic analysis

- **Result**
  - 62 papers

http://lup.lub.lu.se/search/ws/files/40009496/extracted_data.csv

Research Trend

Seminal paper:

Research Questions

- **RQ1** What are the main topics researched within CE and how are they studied?
- **RQ2** Which kind of organizations use CE and which sectors do they operate in?
- **RQ3** With what type of experiments have CE been applied?

Research Questions

- **RQ1** What are the main topics researched within CE and how are they studied?
- **RQ2** Which kind of organizations use CE and which sectors do they operate in?
- **RQ3** With what type of experiments have CE been applied?

Research Questions

• **RQ1** What are the main topics researched within CE and how are they studied?

• **RQ2** Which kind of organizations use CE and which sectors do they operate in?

• **RQ3** With what type of experiments have CE been applied?

## Research Topics (RQ1)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment process</td>
<td>7</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>10</td>
</tr>
<tr>
<td>Challenges</td>
<td>19</td>
</tr>
<tr>
<td>Benefits</td>
<td>3</td>
</tr>
<tr>
<td>Variability management</td>
<td>5</td>
</tr>
<tr>
<td>Metrics</td>
<td>6</td>
</tr>
<tr>
<td>Statistical methods</td>
<td>16</td>
</tr>
<tr>
<td>Design of experiments</td>
<td>8</td>
</tr>
<tr>
<td>Domain considerations</td>
<td>6</td>
</tr>
<tr>
<td>Ethics</td>
<td>1</td>
</tr>
</tbody>
</table>

Research Topics (RQ1)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment process</td>
<td>7</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>10</td>
</tr>
<tr>
<td>Challenges</td>
<td>19</td>
</tr>
<tr>
<td>Benefits</td>
<td>3</td>
</tr>
<tr>
<td>Variability management</td>
<td>5</td>
</tr>
<tr>
<td>Metrics</td>
<td>6</td>
</tr>
<tr>
<td>Statistical methods</td>
<td>16</td>
</tr>
<tr>
<td>Design of experiments</td>
<td>8</td>
</tr>
<tr>
<td>Domain considerations</td>
<td>6</td>
</tr>
<tr>
<td>Ethics</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Evaluation**
  - Evaluation research
  - Experience report

- **Solution**
  - Validation
  - Proposed solution

## Research Topics (RQ1)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Total</th>
<th>Research approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Evaluation</td>
</tr>
<tr>
<td>Experiment process</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Challenges</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>Benefits</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Variability management</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Metrics</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Statistical methods</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Design of experiments</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Domain considerations</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Ethics</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

## Research Topics (RQ1)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Total</th>
<th>Research approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Evaluation</td>
</tr>
<tr>
<td>Experiment process</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Challenges</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>Benefits</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Variability management</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Metrics</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Statistical methods</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Design of experiments</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Domain considerations</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Ethics</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Anna Karenina principle

# Research Topics (RQ1)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Total</th>
<th>Research approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Evaluation</td>
</tr>
<tr>
<td>Experiment process</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Challenges</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>Benefits</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Variability management</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Metrics</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Statistical methods</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Design of experiments</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Domain considerations</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Ethics</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Technical topics**
- **Ethics guidelines**

## Research Topics (RQ1)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Total</th>
<th>Research approach</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Evaluation</td>
<td>Solution</td>
</tr>
<tr>
<td>Experiment process</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Infrastructure</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Challenges</td>
<td>19</td>
<td>17</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Benefits</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Variability management</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Metrics</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Statistical methods</td>
<td>16</td>
<td>1</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Design of experiments</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Domain considerations</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Ethics</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

- No data sets
- One open source tool
Organizations – Where is A/B Testing Used? (RQ2)

- **Sectors**
  - E-commerce
  - Search engine
  - Other

- **Business model**
  - Subscribed or free: Σ 29
    - E-commerce: 9
    - Search engine: 5
    - Other: 15

- **Company size**
  - Perpetual: Σ 9
    - Finance: 4
    - Gaming: 2
    - Other: 4

- **Embedded**
  - Σ 4

Organizations – Where is A/B Testing Used? (RQ2)

- Sectors
- Business model
- Company size

<table>
<thead>
<tr>
<th>Business model</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business to consumer (B2C)</td>
<td>32</td>
</tr>
<tr>
<td>Business to business (B2B)</td>
<td>10</td>
</tr>
</tbody>
</table>

Quality increase

Vs.

Quality increase

Reputation

Organizations – Where is A/B Testing Used? (RQ2)

- Sectors
- Business model
- Company size (employees)

<table>
<thead>
<tr>
<th>Company size</th>
<th>≥ 250</th>
<th>&lt; 250</th>
<th>&lt; 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td></td>
<td></td>
<td>29</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Small</td>
<td></td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>

Experiments – What and How? (RQ3)

- **Treatment**
- **Goal**
- **Experiment design**

<table>
<thead>
<tr>
<th>Treatment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual change</td>
<td>64</td>
</tr>
<tr>
<td>Algorithmic change</td>
<td>23</td>
</tr>
<tr>
<td>New feature</td>
<td>4</td>
</tr>
</tbody>
</table>

Experiments – What and How? (RQ3)

- Treatment
- Goal
- Experiment design

<table>
<thead>
<tr>
<th>Goal</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td>58</td>
</tr>
<tr>
<td>Revenue</td>
<td>26</td>
</tr>
<tr>
<td>Knowledge</td>
<td>7</td>
</tr>
</tbody>
</table>
Experiments – What and How? (RQ3)

- Treatment
- Goal
- Experiment design

<table>
<thead>
<tr>
<th>Experiment design</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/B</td>
<td>77</td>
</tr>
<tr>
<td>A/B/n</td>
<td>3</td>
</tr>
<tr>
<td>MVT</td>
<td>6</td>
</tr>
<tr>
<td>Optimization</td>
<td>3</td>
</tr>
</tbody>
</table>

Slicing Experiments

Take Aways

1) Research gaps
   - Tool evaluations
   - Ethics guidelines
   - Embedded

2) Diverse organisations do A/B testing

3) Simple experimental designs – (Best) practice or bias?

Take Aways

1) Research gaps
   - Tool evaluations
   - Ethics guidelines
   - Embedded

2) Diverse organisations do A/B testing

3) Simple experimental designs – (Best) practice or bias?

Take Aways

1) Research gaps
   - Tool evaluations
   - Ethics guidelines
   - Embedded

2) Diverse organisations do A/B testing

3) Simple experimental designs – (Best) practice or bias?

"Out-of-Scope” A/B Testing Highlights

- Qubit meta analysis

- Errors in statistics
  - https://www.aarondefazio.com/tangentially/?p=83

- Continuous Experimentation origin
  - http://mcfunley.com/design-for-continuous-experimentation

- A/B testing superfluous changes
  - https://iterativepath.wordpress.com/2012/10/29/testing-40-shades-of-blue-ab-testing/

- Ethics of experiments
  - https://doi.org/10.1073/pnas.1320040111