THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Towards Formalizing Data-Driven Decision-Making from Big Data: A Systematic Multi-Criteria Decision-Making Approach in Online Controlled Experiments

Jie “JW” Wu
Engineering Management and Systems Engineering
George Washington University

Agenda

Backgrounds

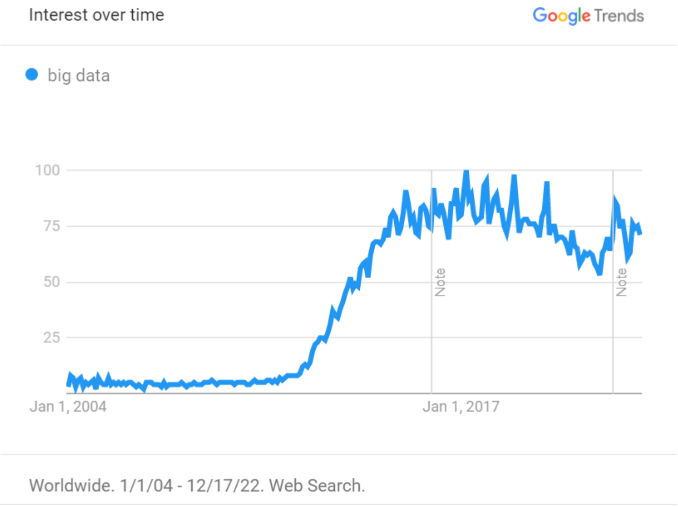
Problem Statement

Proposed Approach

Experiments

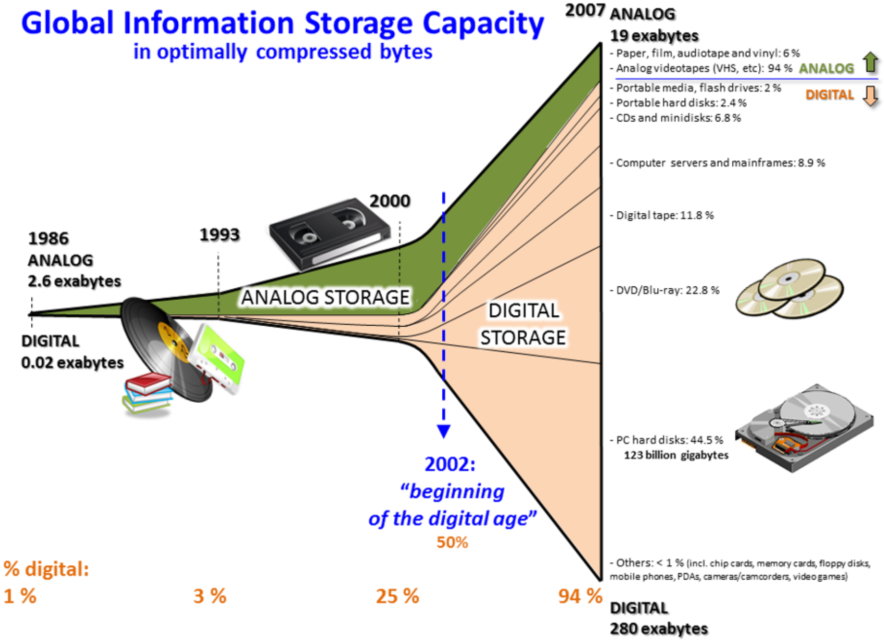
Conclusions

Background: Big data is changing our lives



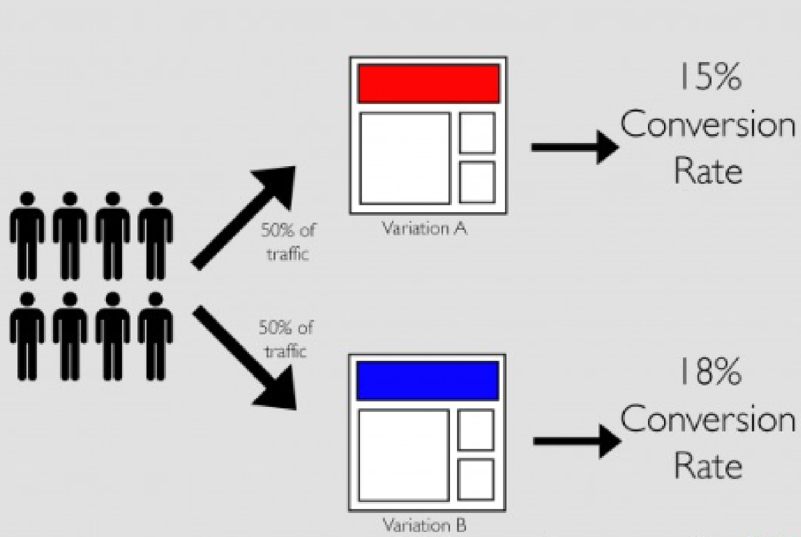
Worldwide. 1/1/04 - 12/17/22. Web Search.

“Big data” in Google Trends (2004-2022)



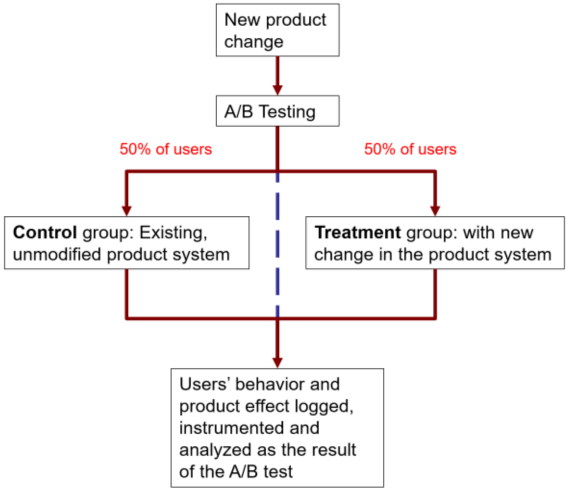
Background: Online Controlled Experiment (a.k.a. A/B Testing)

- A popular technique to analyze Big Data for data-driven decision making.
- Adopted by many web-facing companies (Facebook, Google, Amazon…) as a gold standard.
- Goal: understand how product works, identify bugs, make launch decisions.
- E.g.: Which has a higher conversion rate?



Background: A/B Testing in Industry

- Statistics:
  - # A/B tests: a few hundred annually at a mature company
  - # A/B metrics: hundreds of metrics (e.g., >6k metrics in Microsoft Bing)
- Decision-making process:
  1. defining goal metrics, along with secondary metrics and guardrail metrics,
  2. alerting, scorecards, and periodical diagnosis on these metrics,
  3. multiple approvals and discussions with stakeholders or experts before shipping features via A/B testing



New product change

A/B Testing

50% of users

50% of users

Control group: Existing, unmodified product system

Treatment group: with new change in the product system

Users' behavior and product effect logged, instrumented and analyzed as the result of the A/B test

Metric Category	Metric Name	Absolute Change in Treatment (over Control)	% Change in Treatment (over Control)
Engagement Metrics	App Open	88.409→ 88.621	+ 0.24% +/- 0.31% (p=0.015)
Engagement Metrics	Time Spent	5,101.722→ 5,108.36	+0.13% +/- 0.30% (p=0.1)
Performance Metrics	Network Success	452.978→ 340.14	-0.50% +/- 2.16% (p=0.03)
...	...	...	...

Sample A/B result dashboard comparing treatment variant over control variant

Problem Statement: Identified Issue

- The launch decision-making process of A/B tests is empirical and involves discussions and evaluations among experts.

Evidence from Literature

- Microsoft: analyzing the A/B testing results insights by hand to make informed decisions can be *cumbersome and challenging*.
- Netflix: A/B test results are used as an important source for making product decision, and yet interpreting A/B tests results remains *"partly art"*.
- Google: a process in place to *discuss with experts and agree on* 1) whether the experiment is a positive or negative user experience and 2) whether to launch this change.

Current Literature Gap

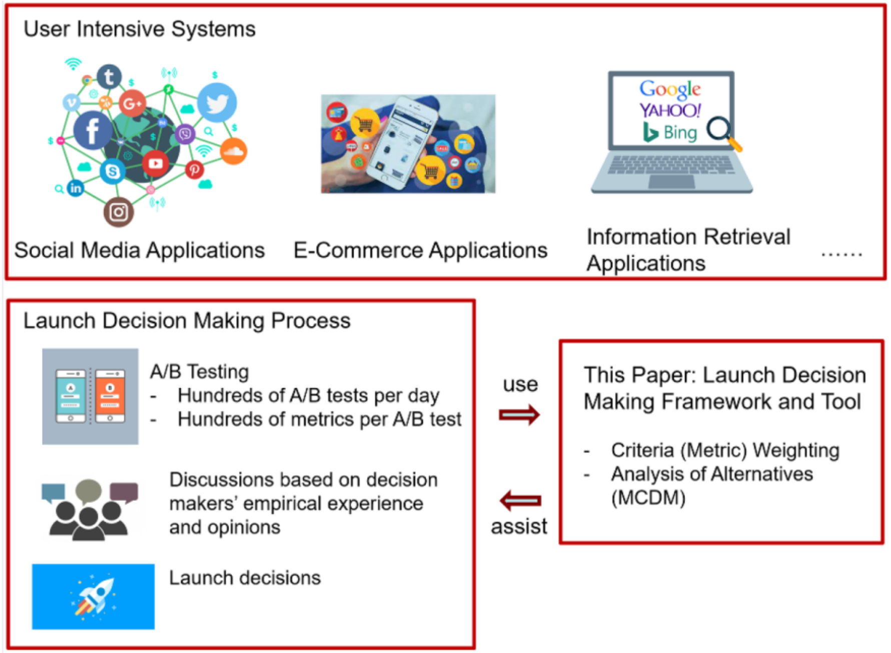
- There is no generalized or principled decision framework that suggests launch decisions with analysis based on the A/B testing results.

Bridging the Gap

- We propose a Multi-criteria decision making (MCDM) based framework, called LDM framework, for the *launch decision making* of A/B testing results.
- *Motivation:* MCDM provide a formal approach to help decision makers improve analytic rigor, auditability, and conflict resolution.

Overview of LDM Framework

- Objective: develop a framework that provides the stakeholders with automatic decision analysis to assist, simplify and crosscheck the launch decision making process based on A/B testing results.



User Intensive Systems

Social Media Applications

E-Commerce Applications

Information Retrieval Applications

……

Launch Decision Making Process

A/B Testing
- Hundreds of A/B tests per day
- Hundreds of metrics per A/B test

Discussions based on decision makers' empirical experience and opinions

Launch decisions

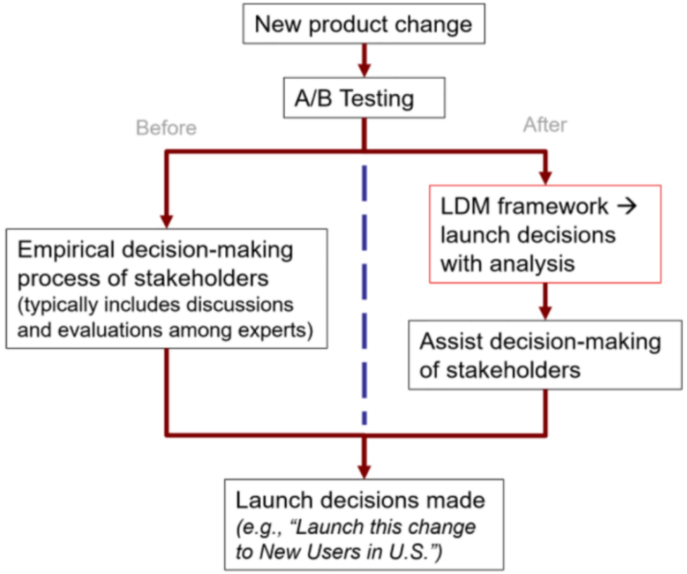
use

assist

This Paper: Launch Decision Making Framework and Tool

- Criteria (Metric) Weighting
- Analysis of Alternatives (MCDM)

The overall picture of the LDM framework



New product change

A/B Testing

Before

After

Empirical decision-making process of stakeholders (typically includes discussions and evaluations among experts)

LDM framework → launch decisions with analysis

Assist decision-making of stakeholders

Launch decisions made *(e.g., "Launch this change to New Users in U.S.")*

The flow of product launch decision-making process using A/B testing, with "before" and "after"

Processing A/B Test Result

Example Walkthrough:

A/B test is conducted with 2 variants: 1 control group and 1 treatment group.

Metrics (n=3): S_1 =App Open, S_2 =Time Spent, S_3 =Network Success

A/B test result for treatment over control $\{m_1, m_2, m_3\}$ = {0.24%, 0%, -0.5%}

Formulation:

We define an A/B test key metrics $S_1, S_2, \ldots, S_n$ and get the A/B test results for the treatment:

$$\{(m_1', CI_1, p_1), (m_2', CI_2, p_2) \ldots, (m_n', CI_n, p_n)\}$$

where m_i' is the raw % change of treatment group over control group on key metric S_i, p_i and CI_i are the p-value and confidence interval size of m_i'.

Next, we convert the results to $\{m_1, m_2, \ldots, m_n\}$, where m_i is normalized.

Question:

Given the A/B test result of treatment group: $\{m_1, m_2, m_3\}$ = {0.24%, 0%, -0.5%}, whether we want to launch this new change or not?

Metric Category	Metric Name	Absolute Change in Treatment (over Control)	% Change in Treatment (over Control)
Engagement Metrics	App Open	88.409→ 88.621	+ 0.24% +/- 0.31% (p=0.015)
Engagement Metrics	Time Spent	5,101.722→ 5,108.36	+0.13% +/- 0.30% (p=0.1)
Performance Metrics	Network Success	452.978→ 340.14	-0.50% +/- 2.16% (p=0.03)
…	…	…	…

Example of A/B test result comparing treatment group with control group

Mathematical Formulation

The goal is to maximize the function f, the positive benefit of launching variant t ∈ T,the set of possible variants (could be control or treatment group):

$$Max\ f(t) = Max\{m_1(t), m_2(t), \dots, m_n\ (t)\}$$

where m_i represents % of statistically significant change on variant t ∈ T over Control on A/B metric S_i

Solutions

One possible solution: Max $f(t) = \sum_i^n w_i\ m_i(t)$ (Weighted Goal Programming solution)

Example walkthrough:

$$T = \{a\ (control), b\ (treatment)\}$$
$$\{m_1(b), m_2(b), m_3(b)\} = \{0.24\%,\ 0\%,\ -0.5\%\}$$
$$m_1(a) = m_2(a) = m_3(a) = 0\%$$

Assume equal weights:

$$w_1 = w_2 = w_3 = \frac{1}{3}$$

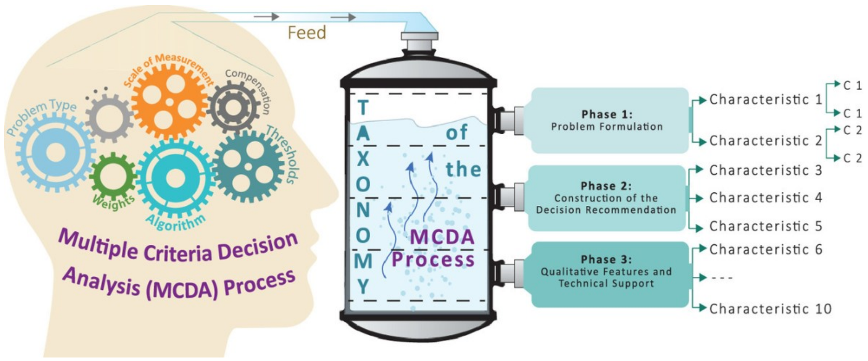
Then: $f(a) = \sum_{i=1}^3 w_i\ m_i(a) = 0,$

$f(b) = \sum_{i=1}^3 w_i\ m_i(b) = \frac{1}{3} * (0.24\% + 0\% \ - 0.5\%) = -0.09\%$ ← No launch for the treatment

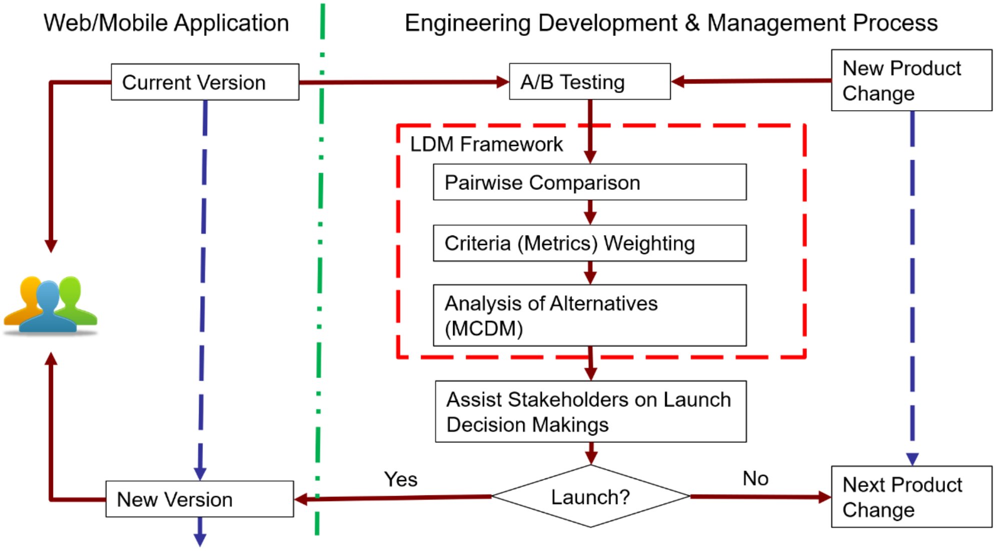
MCDM Approach: LDM Framework

We propose Multi-Criteria Decision-Making (MCDM) approach to address this multi-objective optimization problem for the launch decision making of A/B testing.

1. Framework Configuration Setup
2. Criteria Weighting
3. Pairwise Comparison between Criteria and Alternatives
4. Analysis of Alternatives (MCDM) Given Criteria Weights


Feed
Problem Type
Scale of Measurement
Compensation
Thresholds
Weights
Algorithm
Multiple Criteria Decision Analysis (MCDA) Process
TAXONOMY of the MCDA Process
Phase 1: Problem Formulation
Phase 2: Construction of the Decision Recommendation
Phase 3: Qualitative Features and Technical Support
Characteristic 1
Characteristic 2
Characteristic 3
Characteristic 4
Characteristic 5
Characteristic 6

Characteristic 10
C 1
C 1
C 2
C 2


Web/Mobile Application
Engineering Development & Management Process
Current Version
A/B Testing
New Product Change
LDM Framework
Pairwise Comparison
Criteria (Metrics) Weighting
Analysis of Alternatives (MCDM)
Assist Stakeholders on Launch Decision Makings
Yes
Launch?
No
New Version
Next Product Change

The proposed LDM framework in the engineering development process.

LDM Framework

1. Framework Configuration Setup
   A. *Select A/B metrics (criteria)*
      i. A/B test result includes hundreds of metrics (e.g., >6k metrics in Microsoft Bing)
      ii. Decision makers should review and select a set of key A/B metrics from the specific domain and experiment hypothesis
   B. *Determine launch decision candidates (alternatives)*
      i. Finite alternatives: variants in A/B test (control variant + treatment variants)
      ii. Infinite alternatives
2. Pairwise Comparison between Criteria and Alternatives
3. Criteria Weighting
4. Analysis of Alternatives (MCDM) Given Criteria Weights

LDM Framework

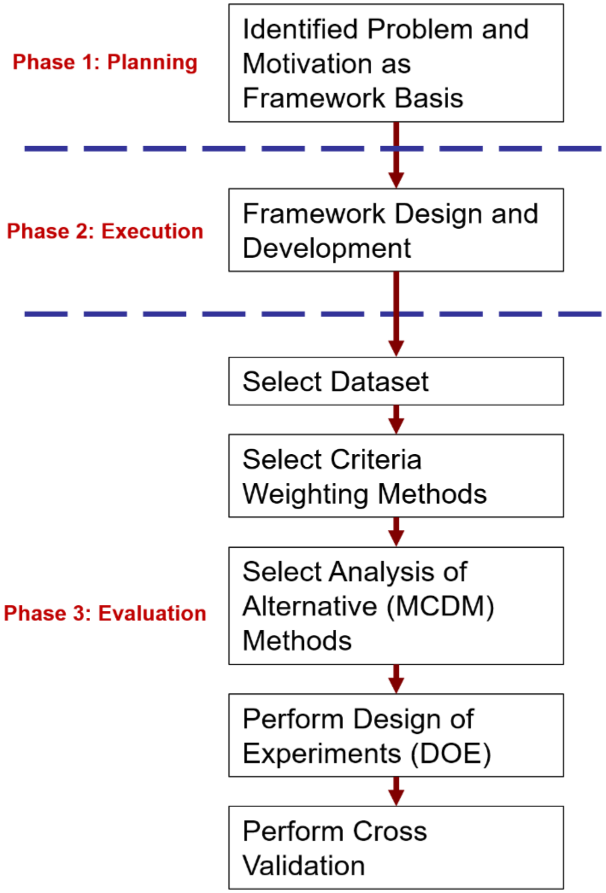
1. Framework Configuration Setup
2. Pairwise Comparison between Criteria and Alternatives
   i. we can get the pairwise comparison result $m_j(t)$ from the A/B testing results on variant *t*
      - Without human expert input, typically needed from traditional MCDM!
3. Criteria Weighting
   i. Objective methods that calculate the weights from objective information (e.g., the pairwise comparison matrix) without human judgments
      - No human input
   ii. Subjective methods that use human judgments and combine weights of stakeholders
      - Performs better
4. Analysis of Alternatives (MCDM) Given Criteria Weights
   i. produces a ranked list of alternatives (launch candidates) with scores. The launch candidate with the highest score is chosen by default.

- Step 1. Pairwise Comparison : For each alternative, obtain the Pairwise Comparison matrix between criteria and the alternative $\{m_1(x,t), m_2(x,t), \ldots, m_n\,(x,t)\}$, using the A/B test result.
- Step 2. Criteria weighting: obtain the weights (importance) of A/B metrics from pairwise comparison matrix in Step 1 (objective criteria weighting method). Alternatively, the weights can also be obtained from human expert judgements (subjective criteria weighting method).
- Step 3. Analysis of Alternatives (MCDM): calculate the score of each alternative, using the selected MCDM methodology

← Stepwise representation

Experiment: Outlined Research Process

- Phase 1: Planning
  - MCDM provide a formal approach to help decision makers improve analytic rigor, auditability, and conflict resolution.
- Phase 2: Execution
  - LDM framework is designed and implemented for launch decision making of A/B tests
- Phase 3: Evaluation
  - Select data set
  - Perform Design of Experiments (DoE)
  - Perform Cross Validation as verification

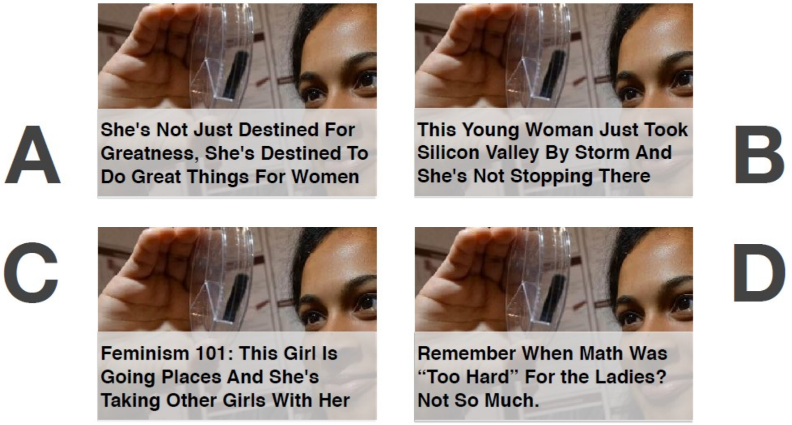

Phase 1: Planning
Identified Problem and Motivation as Framework Basis
Phase 2: Execution
Framework Design and Development
Phase 3: Evaluation
Select Dataset
Select Criteria Weighting Methods
Select Analysis of Alternative (MCDM) Methods
Perform Design of Experiments (DOE)
Perform Cross Validation

Outlined research process of LDM framework in experiments.

Experiment: Select Dataset

Upworthy headline A/B tests dataset. It has 4,873 A/B tests of headlines conducted by Upworthy from January 2013 to April 2015. Each package (treatment) in an A/B test includes:

- created_at: time the package was created
- test_week: week the package was created
- clickability_test_id: test the package was in
- headline
- impressions: # who viewed the package
- clicks: # who clicked the package



A: She's Not Just Destined For Greatness, She's Destined To Do Great Things For Women

B: This Young Woman Just Took Silicon Valley By Storm And She's Not Stopping There

C: Feminism 101: This Girl Is Going Places And She's Taking Other Girls With Her

D: Remember When Math Was "Too Hard" For the Ladies? Not So Much.

A/B metrics:

1) #impression,
2) #clicks,
3) click-through-rate

Candidate variants (alternatives):

- treatment variants or control variant (no ship)

Evaluation:

- LDM framework outputs a ranked list of variants for each A/B test
- Ground truth: winning variant (labeled in the dataset)
- Precision@K is reported: do the top K variants has the winning variant (true label)?

Experiment: Design of experiments (DOE)

- DOE is a systematic, efficient method to study the relationship between multiple factors and their responses.
- DOE was performed to understand the accuracy of the decision-making framework with respect to
  1. criteria weighting method,
  2. analysis of alternative method.

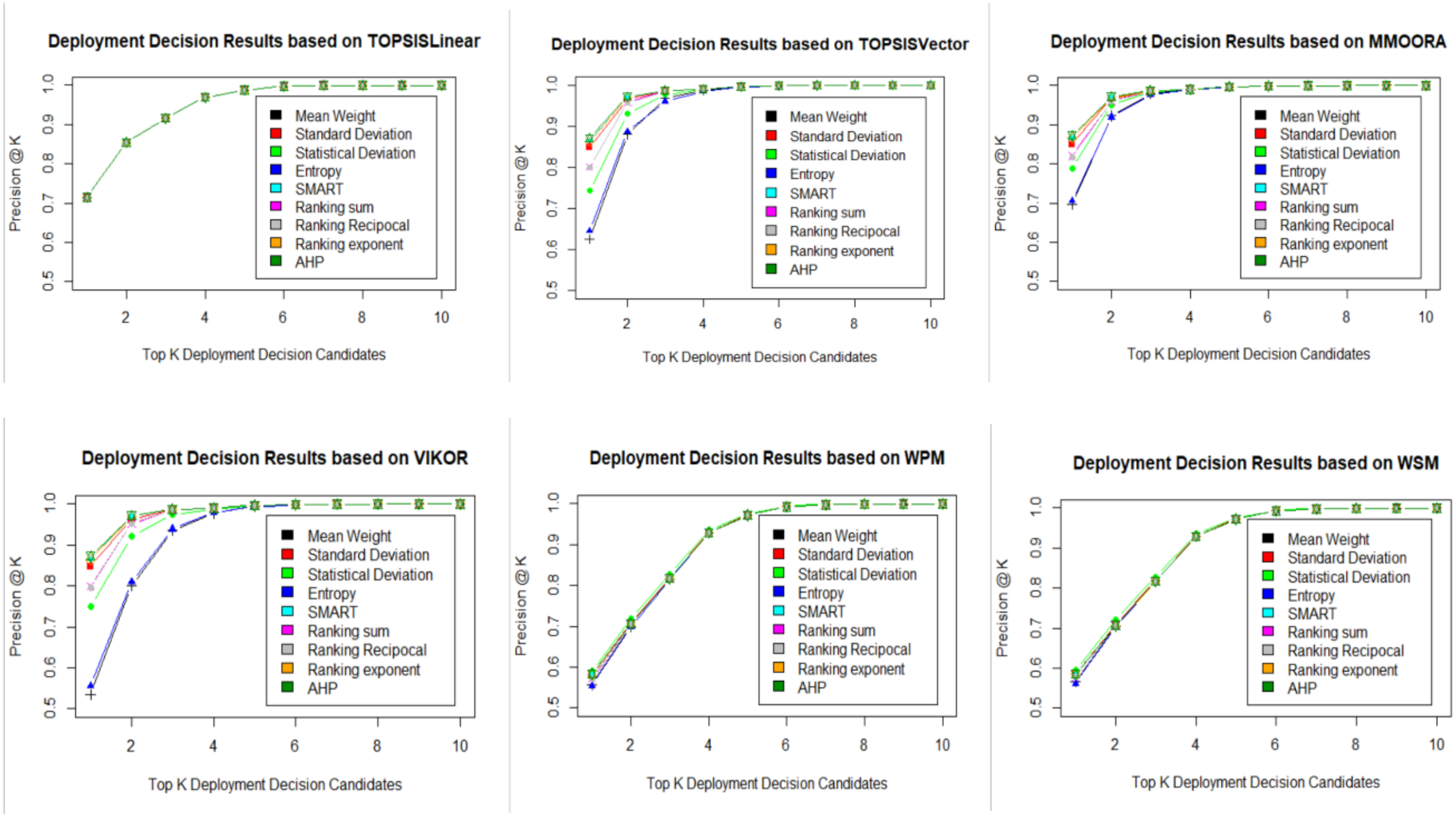
List of Criteria Weighting Methods	Category
Mean Weight	Objective
Standard Deviation	Objective
Statistical Deviation	Objective
Entropy	Objective
SMART	Subjective
Ranking Sum	Subjective
Ranking Reciprocal	Subjective
Ranking Exponent	Subjective
Pairwise comparison (AHP)	Subjective

List of Analysis of Alternatives Methods (MCDM)
WSM
WPM
TOPSIS-Linear
TOPSIS-Vector
MMOORA
VIKOR

Experiment: Results

Precisions stay mostly the same (<75% when K=1) for TOPSIS-Linear, WPM and WSM with different Criteria Weighting methods.

Precisions change significantly for TOPSIS-Vector, MMOORA, and VIKOR. Top performing objective method is Standard Deviation. Top performing subjective method is AHP.



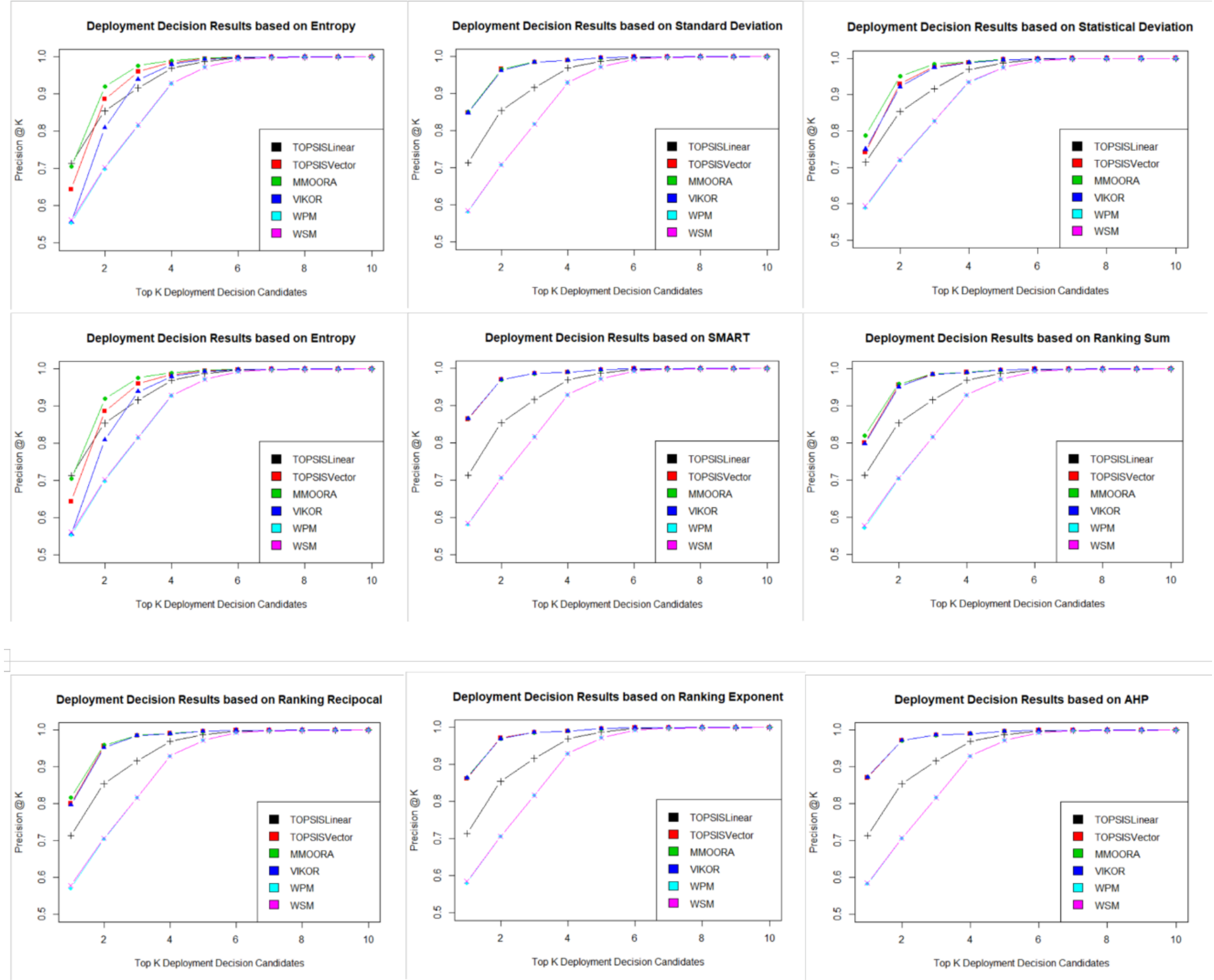
Launch decision results of varying Criteria Weighting method, while the Analysis of Alternative method (MCDM) was held constant.

Experiment: Results

Insight: Analysis of Alternative method is <u>*critical and sensitive*</u> for the performance of the LDM framework.

The MMOORA, TOPSIS-Vector and VIKOR achieved the top 3 in precision, with relatively small difference.

No matter which Criteria Weighing method we pick, the results for WSM, WPM and TOPSIS-Linear are of the lowest precision overall.



Deployment Decision Results based on Entropy

Deployment Decision Results based on Standard Deviation

Deployment Decision Results based on Statistical Deviation

Deployment Decision Results based on Entropy

Deployment Decision Results based on SMART

Deployment Decision Results based on Ranking Sum

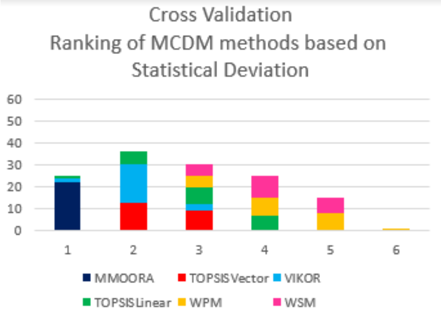
Deployment Decision Results based on Ranking Recipocal

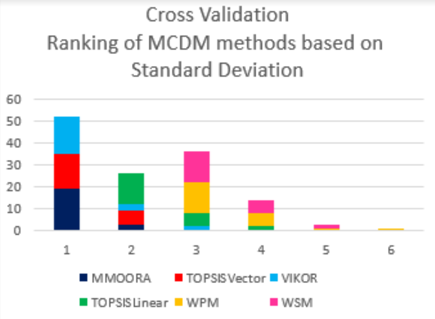
Deployment Decision Results based on Ranking Exponent

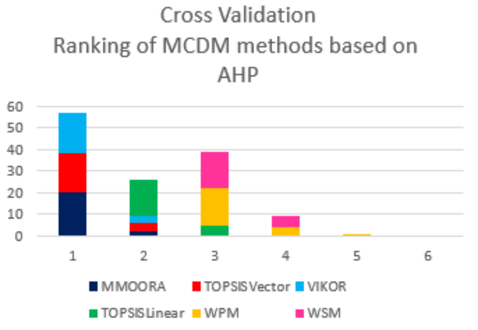
Deployment Decision Results based on AHP

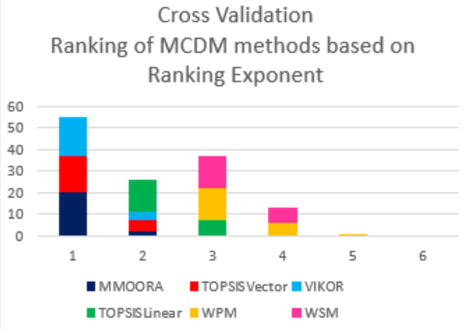
Launch decision results of varying Analysis of Alternative method (MCDM) while the Criteria Weighting method was held constant.

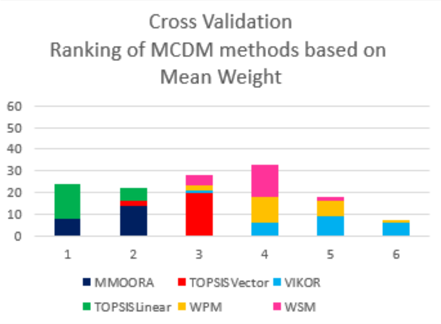
Experiment: Cross Validation

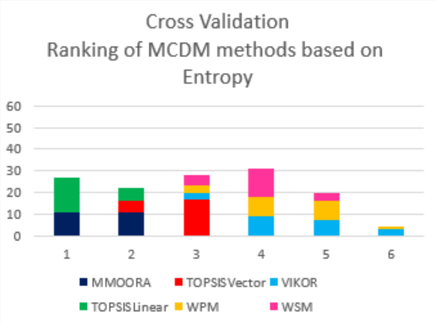

Cross Validation
Ranking of MCDM methods based on Statistical Deviation

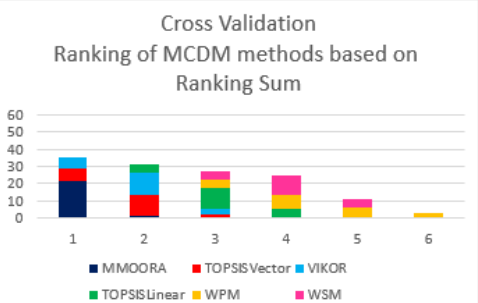

Cross Validation
Ranking of MCDM methods based on Standard Deviation

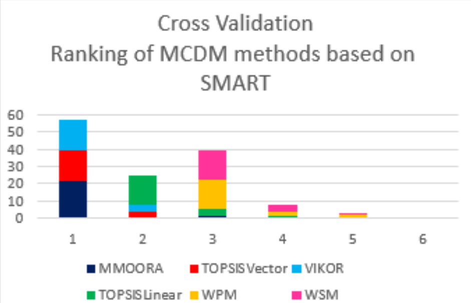

Cross Validation
Ranking of MCDM methods based on AHP


Cross Validation
Ranking of MCDM methods based on Ranking Exponent


Cross Validation
Ranking of MCDM methods based on Mean Weight


Cross Validation
Ranking of MCDM methods based on Entropy


Cross Validation
Ranking of MCDM methods based on Ranking Sum

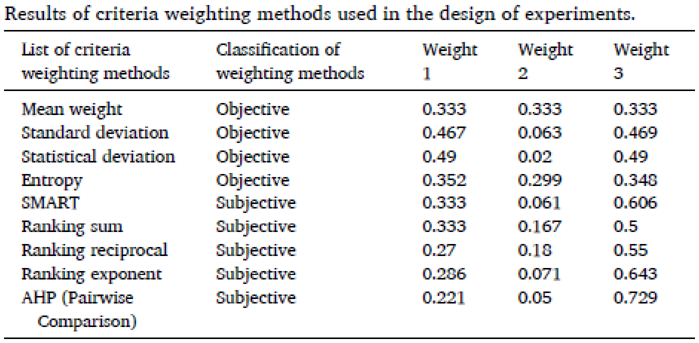

Cross Validation
Ranking of MCDM methods based on SMART

- We performed 22-fold cross validation results on the Upworthy dataset.
  - For each fold, 6 AoA methods are ranked based on the accuracy of the predicted launch decisions compared with the ground truth.
  - x-axis: ranking positions of the 6 AoA methods; y-axis: number of ranked results for each AoA method.
- Insights:
  - For objective weighting methods: MMOORA is the best.
  - For subjective weighting methods: MMOORA, TOPSIS-Vector and VIKOR are top 3 MCDM methods.

Experiment: Cross Validation & Interpreting result

Criteria Weighting Method	TOPSIS Linear Mean (SD)	TOPSIS Vector Mean (SD)	MMOORA Mean (SD)	VIKOR Mean (SD)	WPM Mean (SD)	WSM Mean (SD)
Mean Weight	**59.70% (0.037)**	**62.18% (0.033)**	**69.48% (0.024)**	53.24% (0.029)	55.70% (0.029)	56.67% (0.030)
Standard Deviation	72.32% (0.039)	**85.14% (0.029)**	**85.29% (0.034)**	**85.03% (0.032)**	58.22% (0.030)	58.55% (0.030)
Statistical Deviation	68.76% (0.033)	**74.17% (0.025)**	**78.86% (0.023)**	**74.95% (0.020)**	59.08% (0.032)	59.48% (0.032)
Entropy	**58.87% (0.036)**	**64.23% (0.035)**	**70.47% (0.025)**	55.48% (0.030)	55.42% (0.029)	56.22% (0.029)
SMART	72.38% (0.032)	**86.55% (0.022)**	**86.71% (0.024)**	**86.62% (0.020)**	58.18% (0.031)	58.50% (0.032)
Ranking Sum	67.13% (0.027)	**80.06% (0.019)**	**82.01% (0.022)**	**79.77% (0.026)**	57.13% (0.030)	57.84% (0.030)
Ranking Reciprocal	67.12% (0.026)	**79.98% (0.022)**	**81.66% (0.020)**	**79.58% (0.027)**	57.08% (0.030)	57.85% (0.030)
Ranking Exponent	72.08% (0.032)	**86.34% (0.022)**	**86.63% (0.023)**	**86.47% (0.021)**	58.11% (0.030)	58.52% (0.031)
AHP	73.04% (0.030)	**87.17% (0.020)**	**87.32% (0.021)**	**87.34% (0.020)**	58.32% (0.031)	58.54% (0.031)

Accuracy (mean and SD) of Criteria Weighting methods and AoA methods in the 22-fold cross validation. Top 3 results are marked as bold.



Results of criteria weighting methods used in the design of experiments.

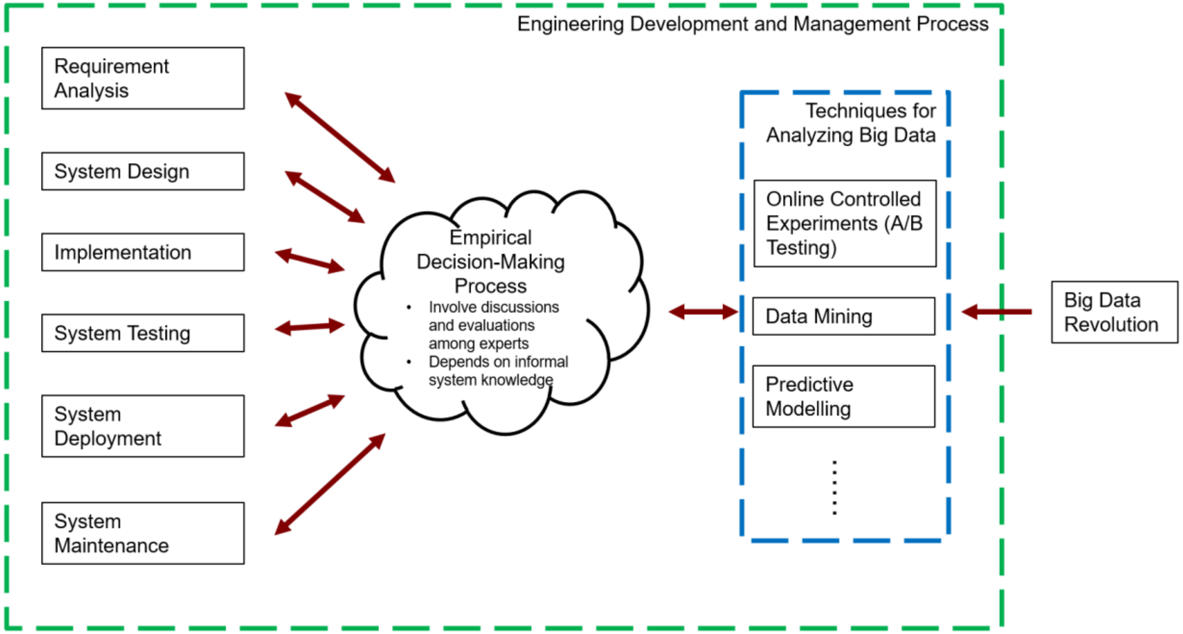
List of criteria weighting methods	Classification of weighting methods	Weight 1	Weight 2	Weight 3
Mean weight	Objective	0.333	0.333	0.333
Standard deviation	Objective	0.467	0.063	0.469
Statistical deviation	Objective	0.49	0.02	0.49
Entropy	Objective	0.352	0.299	0.348
SMART	Subjective	0.333	0.061	0.606
Ranking sum	Subjective	0.333	0.167	0.5
Ranking reciprocal	Subjective	0.27	0.18	0.55
Ranking exponent	Subjective	0.286	0.071	0.643
AHP (Pairwise Comparison)	Subjective	0.221	0.05	0.729

Conclusion

- Problem:
  - A formal, generalized, systematic framework is required to assist the decision makers in making launch decisions using A/B testing results.
- Contribution:
  - Formulated the problem as multi-objective optimization.
  - Proposed a MCDM based framework for it using A/B testing results.
  - Compared and evaluated 6 Analysis of Alternative methods and 9 Criteria Weighting methods on a dataset of 5k A/B tests.
- Conclusion:
  - A good combination of the Analysis of Alternative method (such as TOPSIS-Vector, MMOORA, and VIKOR) and Criteria Weighting method (such as Standard Deviation, AHP) in the LDM framework can lead to effective launch decision making of A/B tests (~87% accuracy).
- Enlightenment:
  - In engineering development, empirical decision-making process from analyzing big data could be formalized.

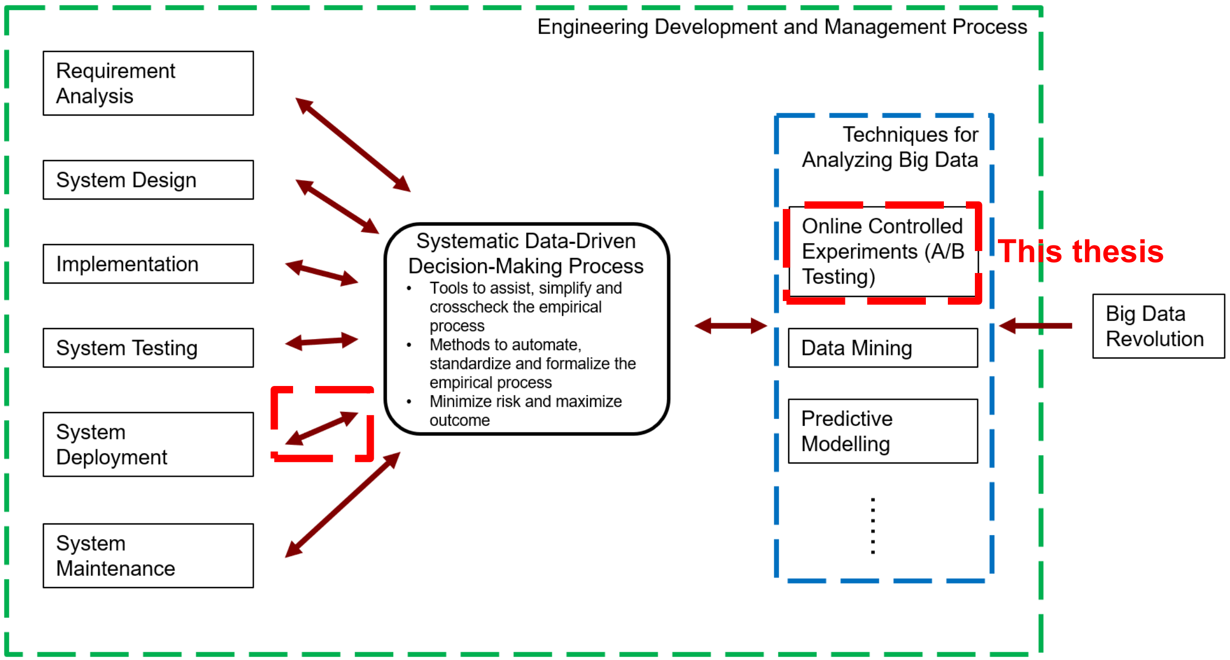
Future Work

- Bigger picture:
  - Data-driven culture is embraced by companies in the big data era.
  - Techniques to analyze big data are integrated for decision-making in engineering development.
- Problem: The decision-making process after integration is ***empirical***
  - Involves discussions and evaluations among experts
  - Depends on informal system knowledge


Engineering Development and Management Process
Requirement Analysis
System Design
Implementation
System Testing
System Deployment
System Maintenance
Empirical Decision-Making Process
• Involve discussions and evaluations among experts
• Depends on informal system knowledge
Techniques for Analyzing Big Data
Online Controlled Experiments (A/B Testing)
Data Mining
Predictive Modelling
Big Data Revolution

Future Work: *SD*^{4} Research Strategy

- Focus of this thesis: Modelling the empirical launch decision-making from A/B testing results
- Insight: Empirical decision-making process from analyzing big data could be better formalized, automated, and assisted
- *SD*^{4}: *"**S**ystematic **D**ata-**D**riven **D**ecision-Making from Big **D**ata Technologies"*


Engineering Development and Management Process
Requirement Analysis
System Design
Implementation
System Testing
System Deployment
System Maintenance
Systematic Data-Driven Decision-Making Process
• Tools to assist, simplify and crosscheck the empirical process
• Methods to automate, standardize and formalize the empirical process
• Minimize risk and maximize outcome
Techniques for Analyzing Big Data
Online Controlled Experiments (A/B Testing)
This thesis
Data Mining
Predictive Modelling
Big Data Revolution