Reconciling Mix Marketing Modeling and Causal Inference

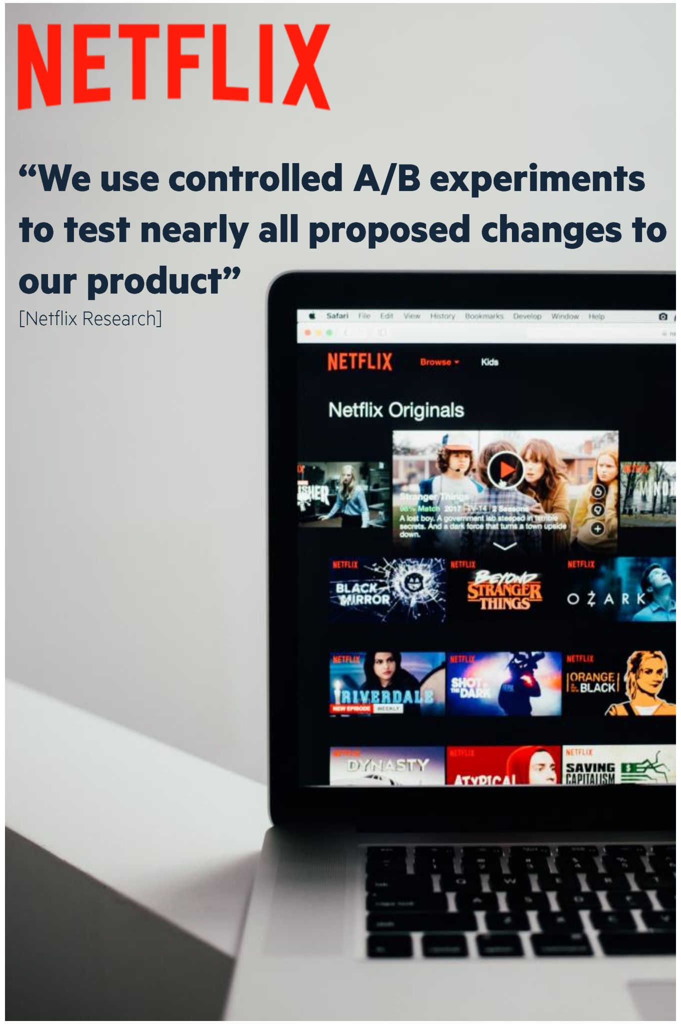
A case study

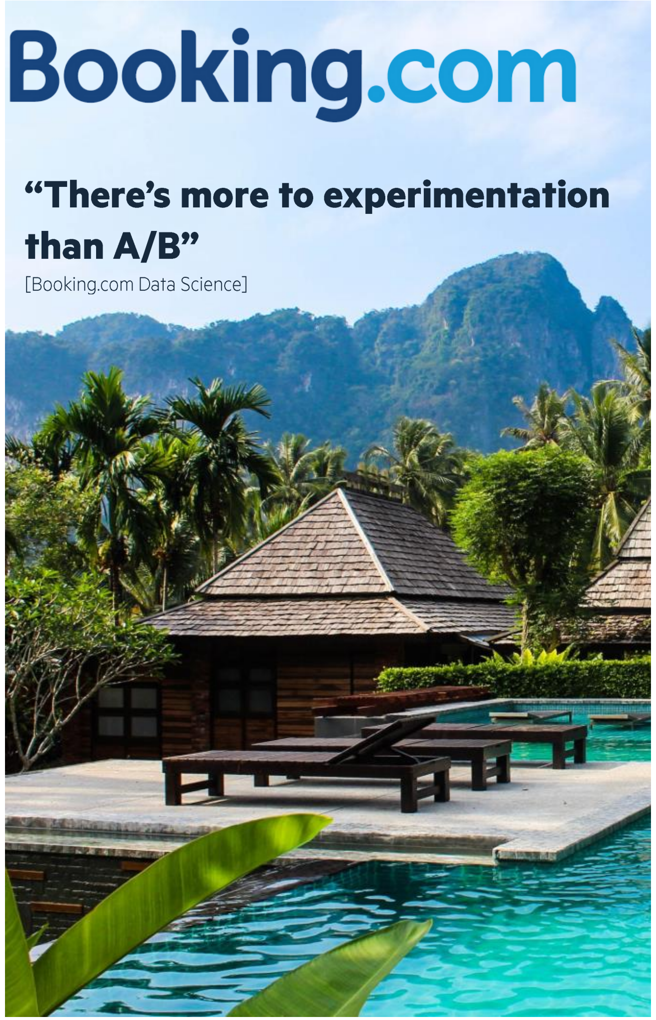
Audrey Poinsot, Ph.D. student
audrey.poinsot@ekimetrics.com

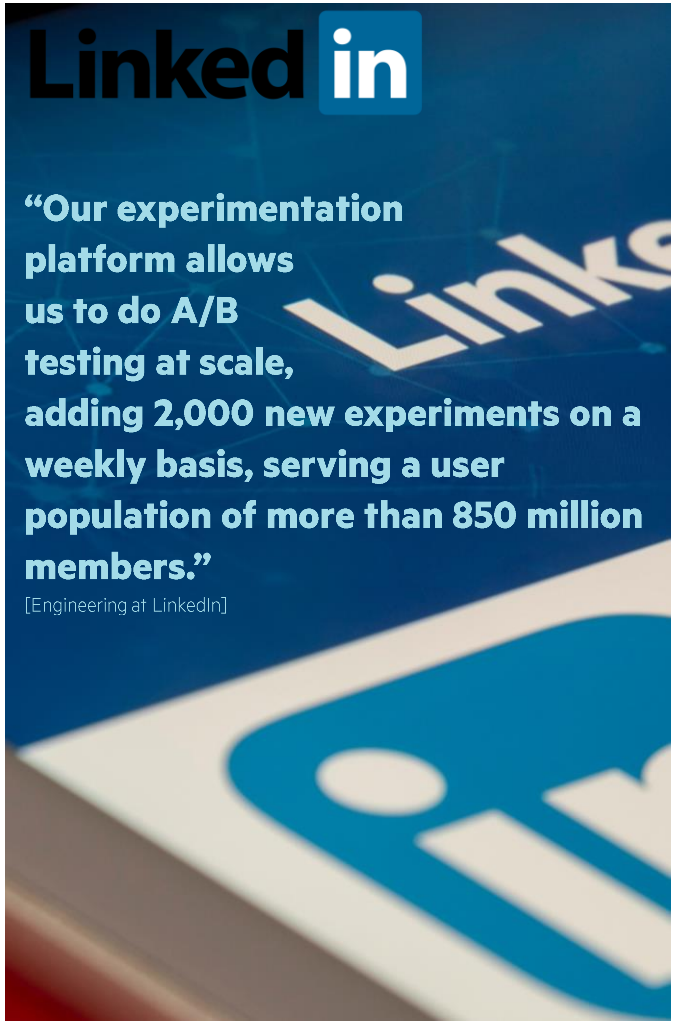
Causality in Practice, June 2023

Ekimetrics.

Marketing has embraced the causal revolution through experimentation

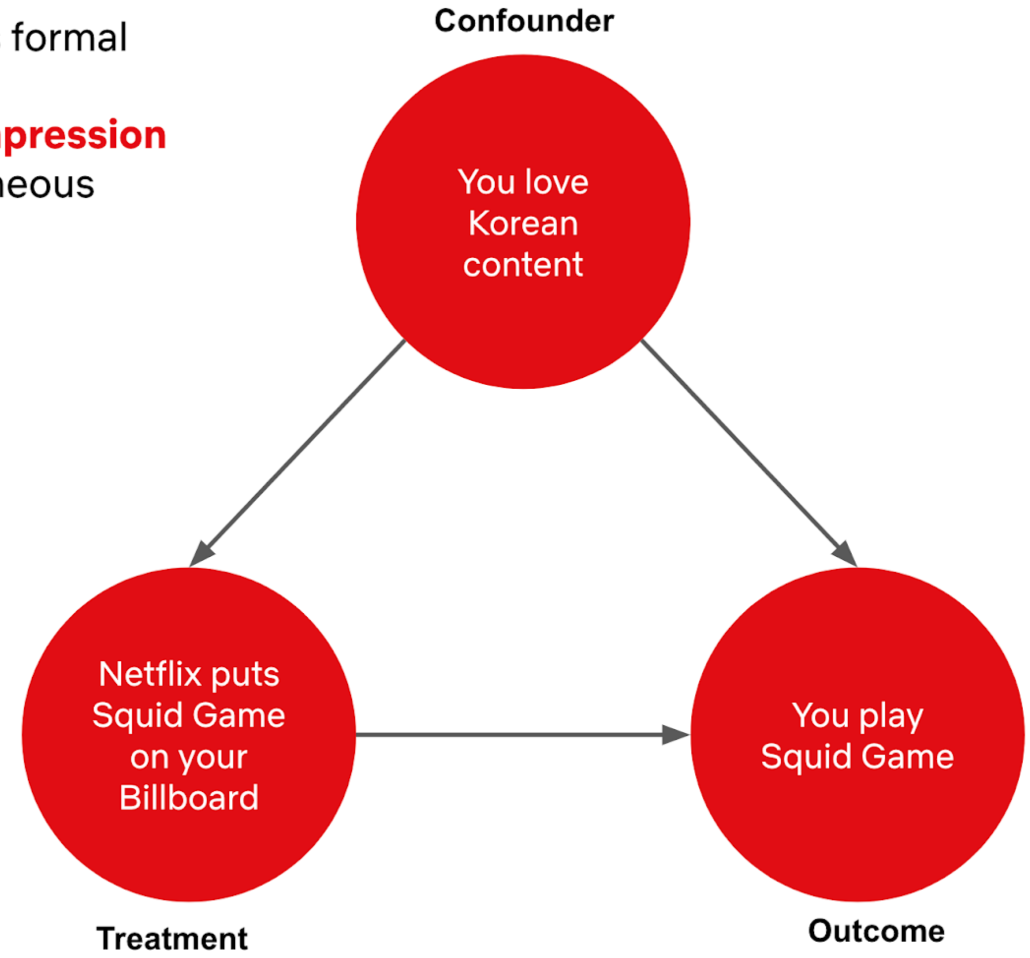






Unfortunately, experimentation is not always possible

Causal Inference provides formal tools to tease out the true **incremental** value of an **impression** for each profile: Heterogeneous Treatment Effect (**HTE**)

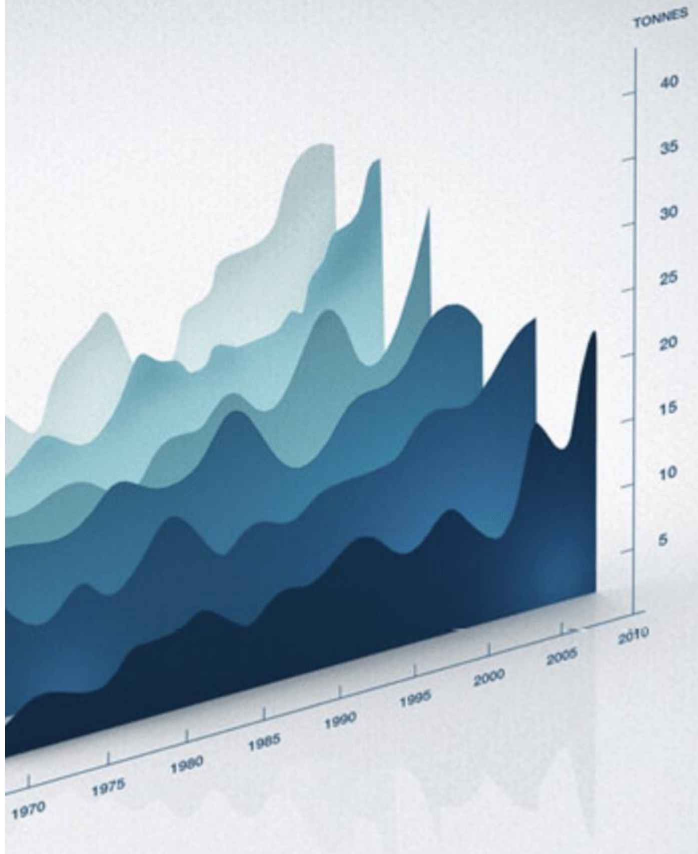


Netflix TechBlog
A Survey of Causal Inference Applications at Netflix

Mix Marketing Modeling, estimating a lot with very few

Objective

Optimize the commercial strategy maximizing the sales volume
Model the contributions/uplifts of each marketing activity
Estimate the ITE of each marketing campaign on the sales revenue



Mix Marketing Modeling, estimating a lot with very few

Objective

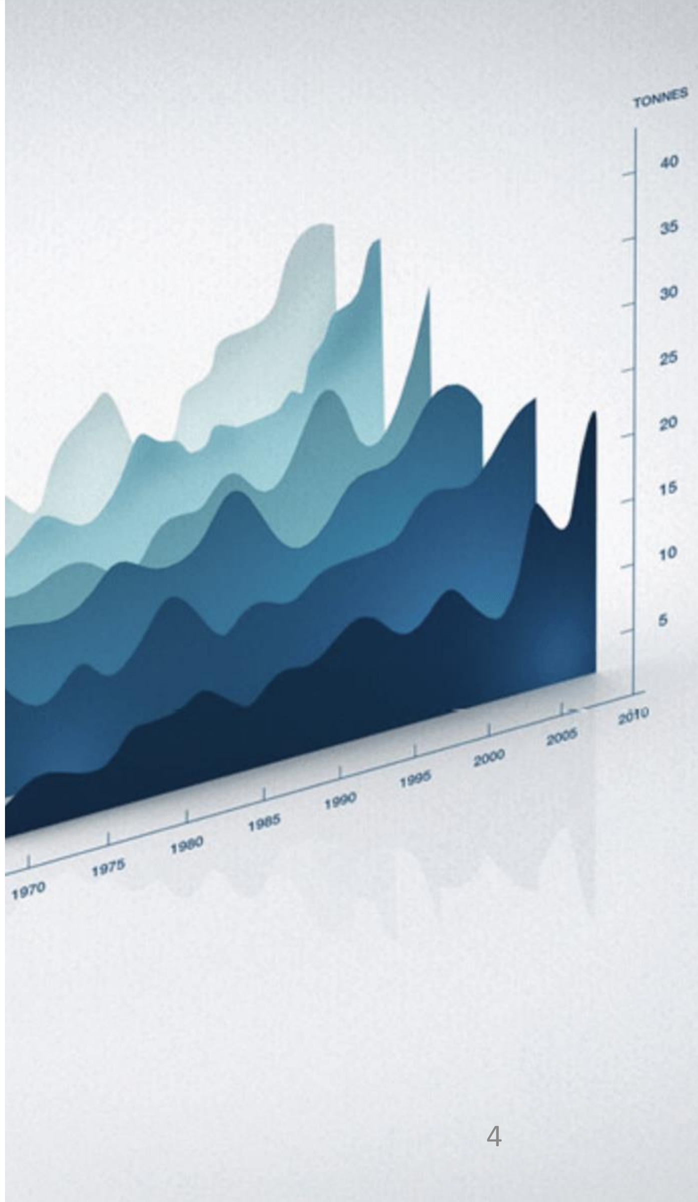
Optimize the commercial strategy maximizing the sales volume
Model the contributions/uplifts of each marketing activity
Estimate the ITE of each marketing campaign on the sales revenue

Correlated treatments

The observed marketing plan is the result of an unmeasurable human decision
To increase effects and maximize sales, many levers are exploited together
Distinguishing the effects of combined campaigns is challenging

Continuous treatments

Many marketing activities are measured with investment
Most effects are non-linear (saturation, synergies, …)
No method yet for non-linear effects of continuous treatment



Ekimetrics.

Mix Marketing Modeling, estimating a lot with very few

Objective

Optimize the commercial strategy maximizing the sales volume
Model the contributions/uplifts of each marketing activity
Estimate the ITE of each marketing campaign on the sales revenue

Correlated treatments

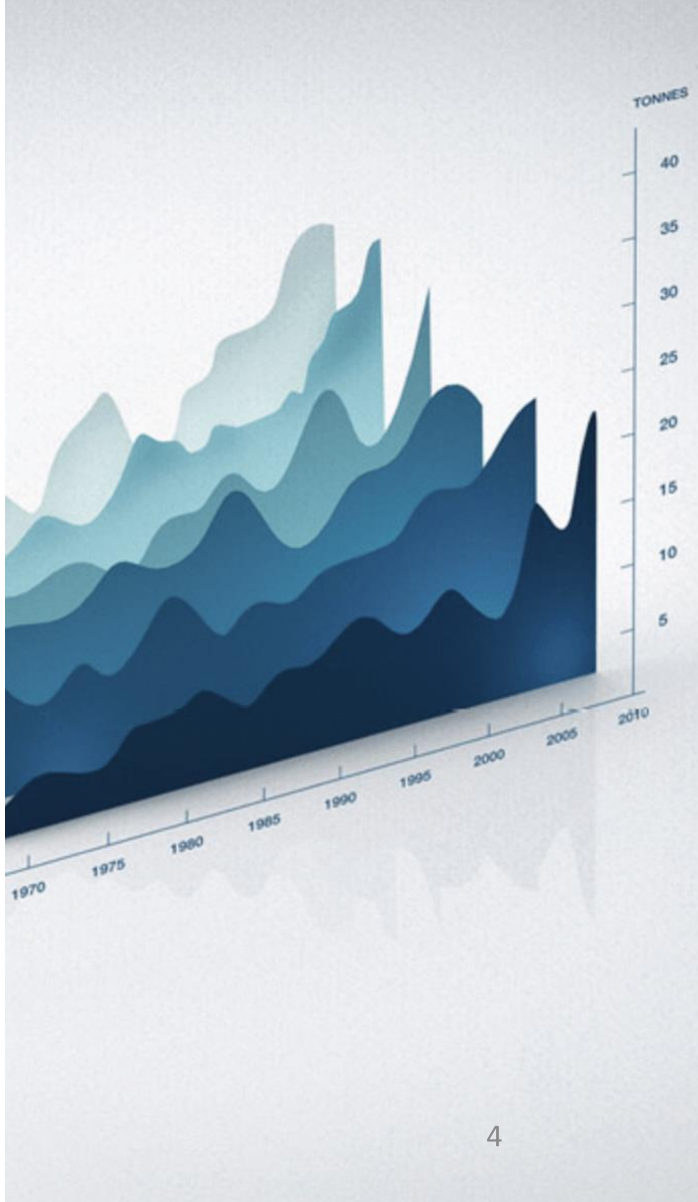
The observed marketing plan is the result of an unmeasurable human decision
To increase effects and maximize sales, many levers are exploited together
Distinguishing the effects of combined campaigns is challenging

Continuous treatments

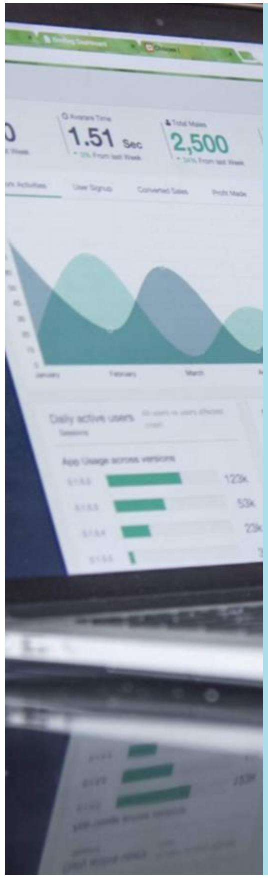
Many marketing activities are measured with investment
Most effects are non-linear (saturation, synergies, ...)
No method yet for non-linear effects of continuous treatment

Limitations of existing methods

Presence of hidden confounders
Mixture of categorical and continuous variables
HTE estimators do not give satisfying results

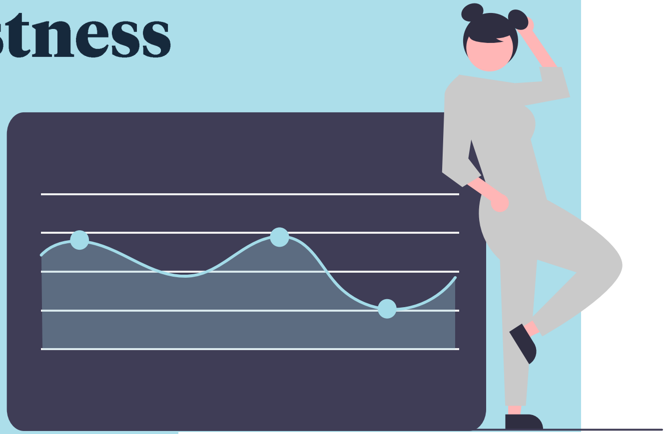


MMM is hence a complex mixture of statistical analysis and business expert assumptions



Statistics

robustness

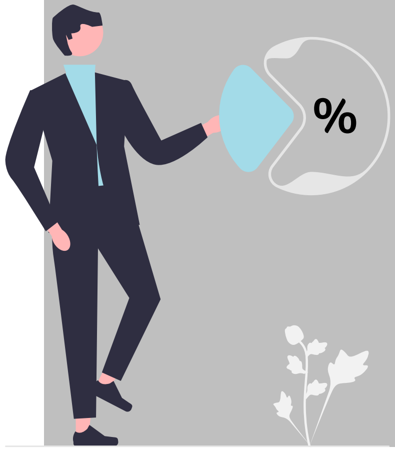


Nested, Pooled, Bayesian modeling

T-statistic
P-value
VIF ...

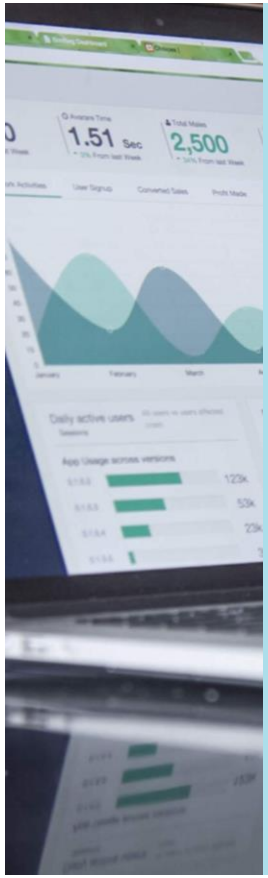
Business

coherence



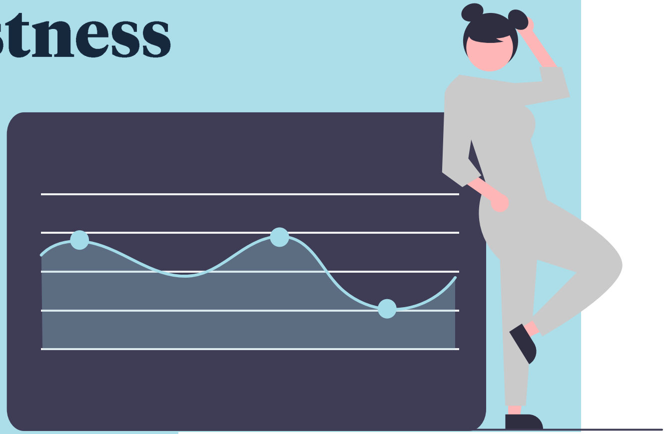
Customer lifetime value
Return on Investment
Cost of acquisition
Click-through rate
Leads generated
Conversion rate
Average basket
Commitment
Impression
Elasticity
...

MMM is hence a complex mixture of statistical analysis and business expert assumptions



Statistics

robustness

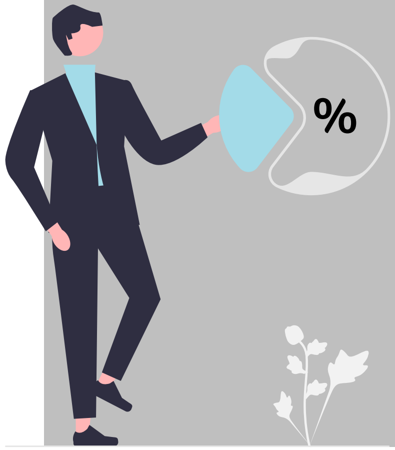


Nested, Pooled, Bayesian modeling

T-statistic
P-value
VIF ...

Business

coherence



Customer lifetime value
Return on Investment
Cost of acquisition
Click-through rate
Leads generated
Conversion rate
Average basket
Commitment
Impression
Elasticity
...

CausalDA

Simplify **statistical analysis**
by eliminating irrelevant dependencies
through prior modeling of **expert knowledge**.

Ekimetrics.

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \dots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \dots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \dots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \dots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

Causal Data Augmentation = Graph G + Density P_{obs}

Ekimetrics.

Hybrid Causal Discovery to mitigate data and human biases

Data-driven Causal Discovery

Measurement error
Unobserved variables
Selection bias
Small data

Expert-driven Causal Discovery

Wrong knowledge
Non-instantaneous reasoning
Human biases
Personal interest

Hybrid Causal Discovery to mitigate data and human biases

Data-driven Causal Discovery

Measurement error
Unobserved variables
Selection bias
Small data

Expert-driven Causal Discovery

Wrong knowledge
Non-instantaneous reasoning
Human biases
Personal interest

Hybrid Causal Discovery

Data reveal human biases

Inputs: data + knowledge
Output: DAG aligned with data & experts

Experts alert on data issues

Hybrid Causal Discovery to mitigate data and human biases

Data-driven Causal Discovery

Measurement error
Unobserved variables
Selection bias
Small data

Data reveal human biases

Expert-driven Causal Discovery

Wrong knowledge
Non-instantaneous reasoning
Human biases
Personal interest

Experts alert on data issues

Hybrid Causal Discovery

Inputs: data + knowledge
Output: DAG aligned with data & experts

Data

Business sense

1. Causal modelization
Build the causal graph through intuitions - business sense

2. Estimands identification
Is it possible to estimate the causal quantity we want?

3. Confrontation with real data
Conditional Independence Tests
Robustness Tests (DoWhy)

4. Graph validation
Validation or rejection of the tests results

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \dots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \dots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

Causal Data Augmentation = Graph G + Density P_{obs}

Ekimetrics.

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \ldots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \ldots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

Causal Data Augmentation = Graph G + Density P_{obs}

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \dots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \dots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

$$\textbf{Causal Data Augmentation = Graph } G \textbf{ + } \underbrace{\textbf{Density } P_{obs}}_{\textbf{= Data } \{X_k\}_{k \in [1,N]} \textbf{ + Estimator}}$$

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \ldots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \ldots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

$$\textbf{Causal Data Augmentation} = \textbf{Graph } G + \underbrace{\textbf{Density } P_{obs}}_{= \textbf{ Data } \{X_k\}_{k \in [1,N]} \textbf{ + Estimator}}$$

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \ldots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \ldots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

Causal Data Augmentation = Graph G + Density P_{obs}

= Data $\{X_k\}_{k \in [1,N]}$ + Estimator

↳ Method **ADMGDA**

Teshima & Sugiyama [UAI 2021]

ADMGDA, a useful method under some assumptions

Experiments

Data Simulated with **random SCMs**

Scenarios

Non-linear data generation
Small-data
Intermediate dimension
Highly dependent variables
High aleatoric uncertainty
Noisy acquisition
Inadequate parametrization

Evaluation metrics

Similarity: KL-div, Wasserstein
Diversity: Average relative difference in variance
Efficiency: XGB error (MAPE, R2 score)

Results

Observations

Pros

Improve XGB predictions
Independent of the causal generation process
→ mechanisms, noise, graph topology

Cons

Highly sensitive to its hyperparameter value
Unsuitable for small-data regimes
→ 300 samples / 10 variables
Sensitive to outliers

Conclusions

Provide more refined data distribution in dense areas
Does not increase diversity
Need to be carefully parametrized

Poinsot & Leite [ICLR Workshop on Trustworthy ML, 2023]

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \dots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \dots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

Causal Data Augmentation = Graph G + $\underbrace{\textbf{Density } P_{obs}}_{}$

= **Data** $\{X_k\}_{k \in [1,N]}$ + **Estimator**

↳ Method **ADMGDA**

Ekimetrics.

Teshima & Sugiyama [UAI 2021]

CausalDA, an approach to break down irrelevant correlations

Definition. Causal Data Augmentation

For a set of variables $(X_1, \dots, X_d)$ distributed according to P_{obs} and a DAG G encoding the causal dependencies that the variables must follow, **Causal Data Augmentation** consists in sampling M data points from the distribution P_{spl} defined as the Markov factorization of P_{obs} given by the graph G.

$$P_{spl}(X_1, \dots, X_d) = \prod_{i=1}^{d} P_{obs}(X_i | Pa(X_i))$$

Causal Data Augmentation = Graph G + Density P_{obs}

= Data $\{X_k\}_{k\in[1,N]}$ + Estimator

↳ Method **ADMGDA**

Teshima & Sugiyama [UAI 2021]

CausalDA, a promising approach that now needs to be trialed

Statistical KPIs matching business dynamics

Ekimetrics.

CausalDA, a promising approach that now needs to be trialed

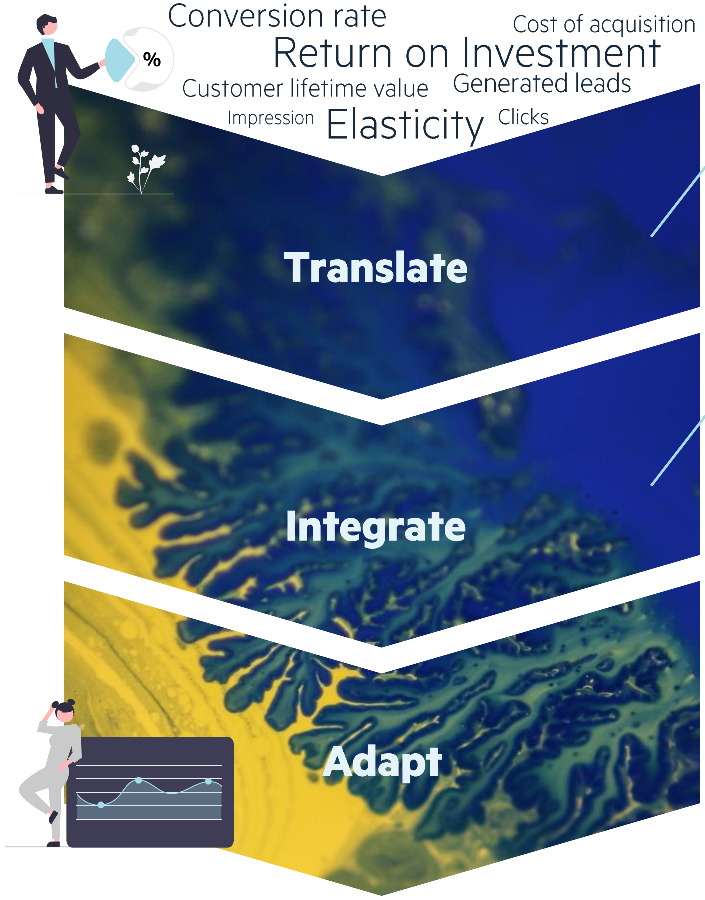
Build a causal graph

Data reveal human biases
Experts alert on data issues

Statistical KPIs matching business dynamics

Ekimetrics.

CausalDA, a promising approach that now needs to be trialed



Build a causal graph

Data reveal human biases
Experts alert on data issues

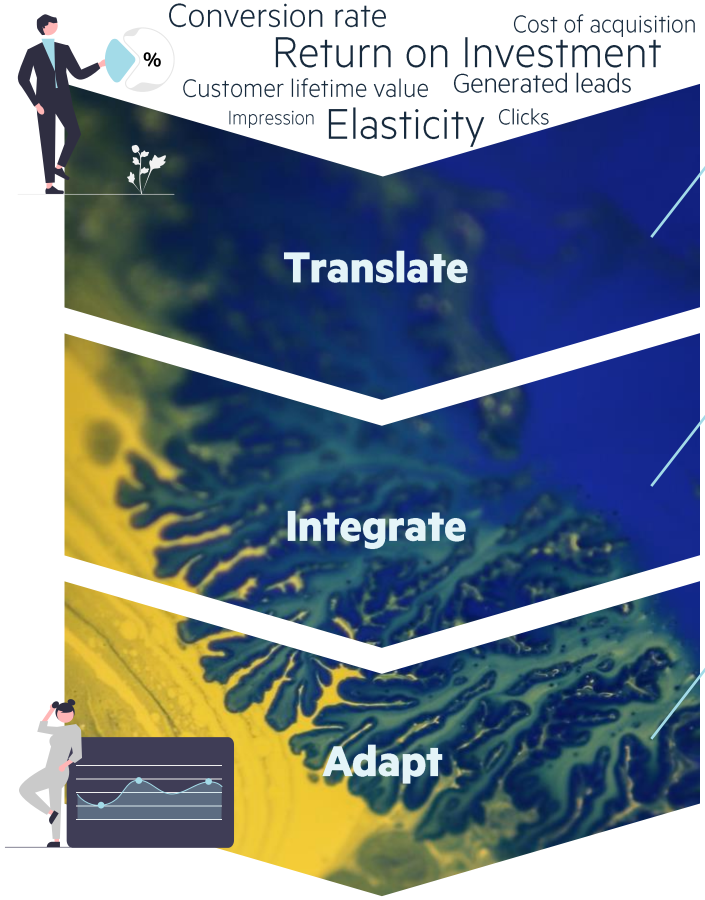
Apply Causal Data Augmentation

ADMGDA is a possible solution
Any other conditional density estimator might work

Statistical KPIs matching business dynamics

Ekimetrics.

CausalDA, a promising approach that now needs to be trialed



Statistical KPIs matching business dynamics

Build a causal graph

Data reveal human biases
Experts alert on data issues

Apply Causal Data Augmentation

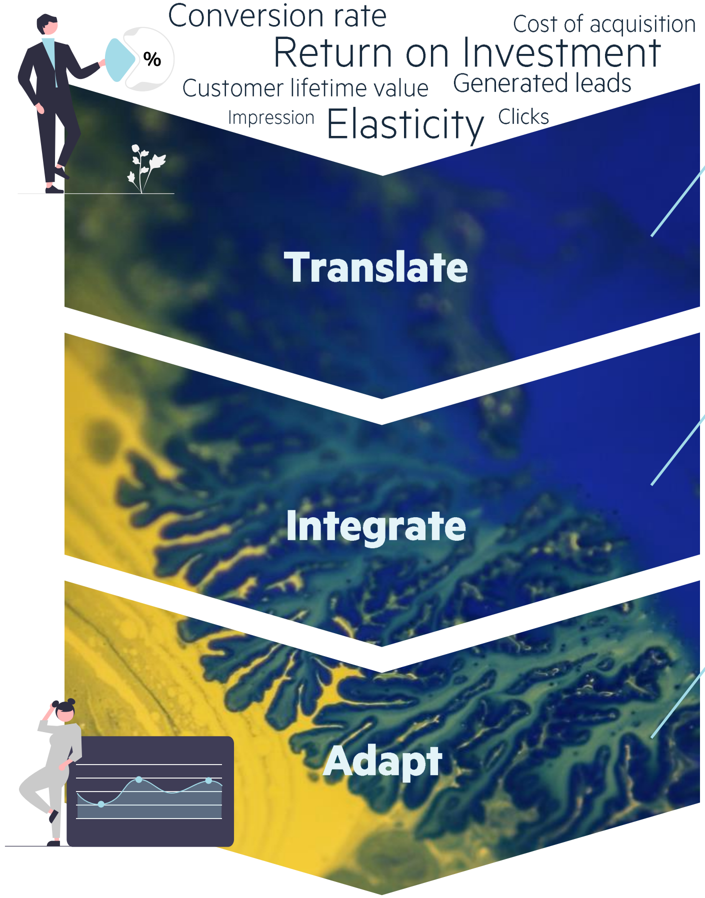
ADMGDA is a possible solution
Any other conditional density estimator might work

Analyze the new dataset

Use the whole dataset to fit the models
Compute Marketing KPIs on observed data only

Ekimetrics.

CausalDA, a promising approach that now needs to be trialed



Statistical KPIs matching business dynamics

Build a causal graph

Data reveal human biases
Experts alert on data issues

Apply Causal Data Augmentation

ADMGDA is a possible solution
Any other conditional density estimator might work

Analyze the new dataset

Use the whole dataset to fit the models
Compute Marketing KPIs on observed data only

Adopt another perspective

Collect end-user feedbacks

Ekimetrics.

Questions

References

Takeshi Teshima and Masashi Sugiyama. *Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation*. In Uncertainty in Artificial Intelligence, pp. 86–96, 2021

Audrey Poinsot and Alessandro Leite. *A Guide for Practical Use of ADMG Causal Data Augmentation*. In ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML, 2023. https://openreview.net/forum?id=kBcAZcKypug

Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. *Causal Discovery Toolbox: uncovering causal relationships in Python*. The Journal of Machine Learning Research, 21(1):1406–1410, 2020. https://jmlr.org/papers/v21/19-187.html

Netflix Research. *Experimentation & Causal Inference*. https://research.netflix.com/research-area/experimentation-and-causal-inference

Netflix Technology Blog. *A survey of Causal Inference Applications at Netflix*. Netflix TechBlog. 2021. https://netflixtechblog.com/a-survey-of-causal-inference-applications-at-netflix-b62d25175e6f

Christina Katsimerou. *There's more to experimentation than A/B*. Booking.com Data Science. 2020. https://booking.ai/theres-more-to-experimentation-than-a-b-223fba846876

Kenneth Tay and Xiaofeng Wang. *Ocelot: Scaling observational causal inference at LinkedIn*. LinkedIn Engineering. 2022. https://engineering.linkedin.com/blog/2022/ocelot--scaling-observational-causal-inference-at-linkedin

Appendix 1 – ADMGDA evaluation

Random SCMs:

1. Random DAG – Erdös-Rényi model
2. Random mechanisms from parametric functions
3. GMMs as root causes
4. Gaussian additive noise

Causal Discovery Toolbox
https://github.com/FenTechSolutions/CausalDiscoveryToolbox

Default experiments parameters

Parameter	Value
Network architecture	2-layers fully-connected neural network with hyperbolic tangent activation function and 20 neurons initialized through the Glorot uniform
Number of variables	10
Causal graph expected degree	3
Additive noise amplitude	0.4
Probability threshold	10^{-2}
Fraction of outliers	0
Number of repetitions	20
Kernels function	Gaussian Kernels with Silverman bandwidth

Scenarios parameters

- **Non-linear data generation setting**: by varying the family functions of the mechanism included linear, polynomial, sigmoid, Gaussian process, and neural networks.
- **Small-data regime**: by varying the number of observations from a few samples to a hundred samples (i.e., $[30, 40, 60, 80, 100, 300, 500, 700]$)
- **High-dimension scenario**: by varying the number of variables in a dataset from seven to twenty-five (i.e., $[7, 8, 9, 10, 15, 20, 25]$)
- **Highly dependent input variables setting**: by varying the expected degree of the causal graph in $[0, 1, 2, 3, 4, 5, 6, 7]$
- **High aleatoric uncertainty setting**: by varying the additive noise amplitude in $[0.1, 0.2, 0.4, 0.6, 0.8, 1]$
- **Noisy acquisition procedure** (i.e., outliers): by varying the fraction of outliers in $[0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15]$
- **Inadequate parametrization scenario**: by varying the probability threshold θ defined in Section 2. $\theta \in [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$

XGBs evaluation:

- Train-Test split 70%-30%
- Augment data from Train
- For each variable as the target variable
  - Train two XGBs on the train and the augmented sets
  - Evaluate both XGBs on the test set

XGBs hyperparameters:

- Cross-Valisation
  - n_estimators in [10, 50, 200]
  - rag_lambda in [1, 10, 100]
- Other parameters as default values

Appendix 2 – ADMGDA method

Algorithm

Input: $D_{train} = \{X_k\}_{k\in[1,n]},\ \mathcal{G},\ \theta,\ L,\ \{K^j\}_{j\in[1,d]}$ ▷ assuming that the variables in the training set and kernel functions are ordered according to the topological order of the graph $\mathcal{G}$

$W_{aug} \leftarrow \{\frac{1}{n}\}^n$

$Z_{aug} \leftarrow \{X_k^1\}_{k\in[1,n]}$

for $j \in [2, d]$ **do**

  $Z_{aug}^{new} \leftarrow \{\}$

  $W_{aug}^{new} \leftarrow \{\}$

  for $Z_i, w_i \in Z_{aug}, W_{aug}$ **do**

    for $i_j \in [1, n]$ **do**

      $w_i^{new} \leftarrow w_i \cdot \frac{K^j(Z_i^{a(j)} - X_{i_j}^{a(j)})}{\sum_{k=1}^n K^j(Z_i^{a(j)} - X_k^{a(j)})}$

      $Z_i^{new} \leftarrow \{Z_i; X_{i_j}^j\}$

      if $w_i^{new} > \theta$ **then**

        $Z_{aug}^{new} \leftarrow Z_{aug}^{new} \cup Z_i^{new}$

        $W_{aug}^{new} \leftarrow W_{aug}^{new} \cup w_i^{new}$

$Z_{aug} \leftarrow Z_{aug}^{new}$

$W_{aug} \leftarrow W_{aug}^{new}$

Output: $\hat{f} \in \arg\min_f \sum_{(w_i,Z_i)_{i\in(W_{aug},Z_{aug})}} w_i L(f, Z_i), \quad D_{aug} = (W_{aug}, Z_{aug})$