### Paper title: Convincing improvements: Systematically identifying and evaluating causal claims in the design science literature

Paper outline:

1. **Defining the framework.**
   1. An empirical survey of papers. Here is what they did, here is our framework.
2. **Using the framework.** This framework allows us to do three new kinds of empirical analysis which inform our understanding of DSR.
   1. Literature-based empirical analysis
      1. Here is what authors could do vs what they did.
      2. A matrix of papers x claims x evaluations
   2. Empirical analysis of one or two papers through the entire matrix
      1. Could we actually try to measure a change to one of these artifacts all the way through the matrix? As an empirical demonstration.
      2. Can we show and analyze what happens if you do all of these evals and how they fit together?
      3. E.g. can we see if a change helps:
         1. With an automated metric
         2. Helps people do task X
         3. Helps an org (super hard)
      4. If this is too hard, we can talk about a few case studies from the literature as a discussion. It's less interesting.
   3. Empirical analysis of artifact manipulations
      1. Measuring how outcome varies with different kinds of manipulations, e.g. add vs swap vs rank.
      2. Not totally sure we need this. It could kind of be its own paper. It is less interesting than the previous, but easier to do.

### Kai's question: what makes a causal evaluation good and convincing? When are you convinced?

Some notes on this [here](https://docs.google.com/document/d/1zlXsTaR9dXb2fhim-S5VppZQ2AV-1fWMRuswlTTFnFQ/edit?usp=sharing)

## Evaluation List

How do people evaluate causal claims? This is important. Our big question is: how do you do a good job making a causal claim. In this paper we are concerned with causal evaluations. A causal evaluation is an experiment to evaluate a claim about how an artifact part causes an outcome. We want a matrix of claims and evaluations?

###### Automatic and analytic evaluations (no people)

* **Automatic comparison of two artifacts without attribution (non causal).** Compare two artifacts using an automatic score (e.g. compare to baseline) without systematically attempting to attribute difference to a difference between blueprints.
  + This common evaluation does not support a causal claim because there is no attempt to attribute the difference between models to an artifact part.
    - It's closer to criterion validity
* **Automatic comparison of two artifacts with experimental attribution.[[1]](#footnote-0)** Add, remove or swap an artifact part and compute an automatic score on a dataset. If the experiment is sound, the difference in the score can be attributed to the change in the model.
  + Examples:
    - Remove ("ablate") a component of a neural network
    - Add new features to a model
    - Swap one loss function for another
* **Automatic or analytic analysis of one artifact.** Perform an experimental or analytic analysis of parts of one artifact to rank their contribution to overall behavior.
  + For simpler, linear models, this can be done analytically, e.g. logistic regression coefficients. The effect of these parts is clearly defined mathematically.
  + For more complex models, this may require experiments.
    - Examples: Shapley values, saliency maps
* **Analytic evaluation of artifact comparison in organizational setting**. The authors make a claim about the organizational significance through some sort of forecasting or modeling process. The claim is that the system will save the organization money or time. Examples:
  + Abassi phishing funnel paper
  + Abassi speech acts paper

###### Human evaluations

* **Simplified human evaluation of an artifact part.** Compare two artifacts in terms of some property that requires subjective human judgment or requires measuring human performance. The claim is that the difference between the two models leads to a qualitative or quantitative change in human behavior.
  + Examples include: fluency and faithfulness in machine translation, mean opinion scores in text-to-speech and readability scores in sentence simplification, or reading comprehension tasks in text summarization
  + This is common in CS but seems to be less common in IS. See Been Kim's paper on interpretability.
  + For IS, you might look at something like: can an analyst make a correct investment decision or correctly identify fraud?
    - AH: I think Abbassi/Dobolyi phishing paper does this?
* **Complete human evaluation of the entire artifact in an artificial setting.** *A lot of times this is a comparison of two artifacts rather than a comparison of a manipulated artifact. We may want to adjust our definitions of causal vs. criterion validity to allow for comparison between two artifacts. Usually this one is comparison to baseline. I think usually there is some stuff to show that the manipulation is good then you compare to baseline.*
  + Get subjective and objective measures of human performance on the complete artifact in an artificial setting.
    - In DSR, this is typically done through an applicability check to get high-level subject feedback on a whole artifact.
    - In HCI, this is also sometimes done by objectively measuring human performance on a particular task, while manipulating an artifact part (e.g. a new feature)
  + To make a convincing causal claim, it should be possible to detect differences in feedback between B and B' in the artificial setting. But usually people compare two totally different models.
    - Subjectively, people should understand and be able to articulate the difference between B and B'.
      * For instance, in the Abassi speech acts paper, we might ask: can people tell the difference between a system that represents speech acts and a system that does not?
    - Objectively, it should be possible to measure the difference between B and B'
* **Ongoing evaluation of the entire artifact in a realistic setting.** 
  + The system is actually deployed in a realistic setting. The tool is actually deployed and used in an organization
    - Example: Abassi's paper on speech acts, the Overview system for news.
    - Note: there is an implicit philosophy of Deweyean pragmatism here. Physics is true because airplanes work.
  + Organizational significance.
    - A really convincing causal validity explanation is not only is B' better than B based on some metric, but it is actually important for the organization.
      * E.g. FairPlay
    - This is hard to show. But in theory you could two do things:
      * Formative evaluation to explain why you are building what you are building. This is something like "problem finding" in the Peffers et. al (2007) model
      * Longitudinal evaluation to watch deployment over time and see if B vs B' affects organization

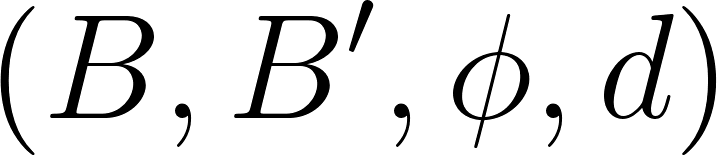
### Robustness to changes in data

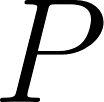
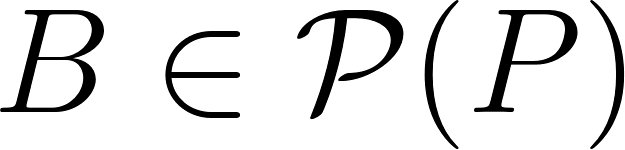
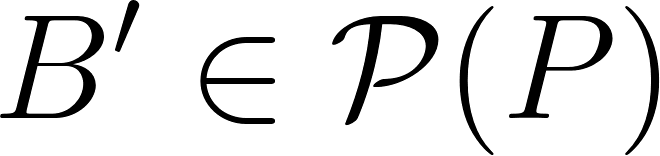
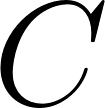
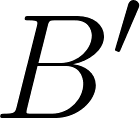
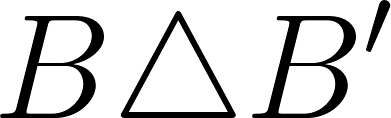
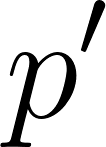
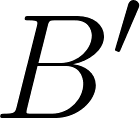
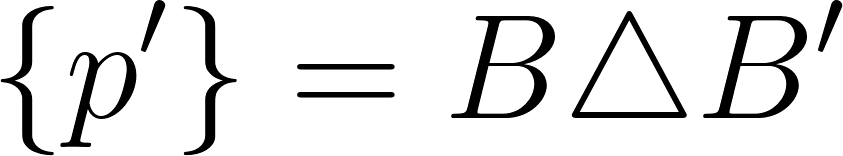
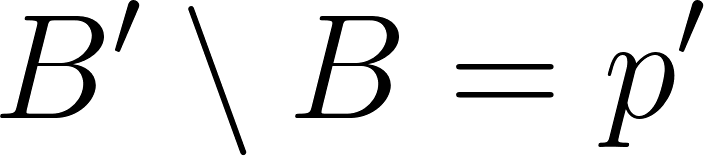
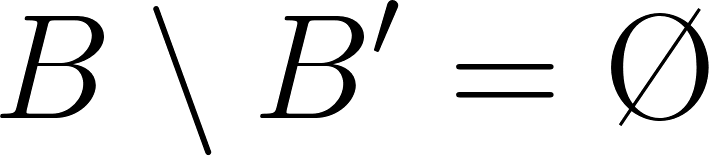
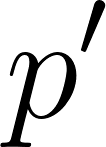
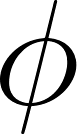
I am not really sure where to put this. It does not yet fit well into our evaluations or framework.

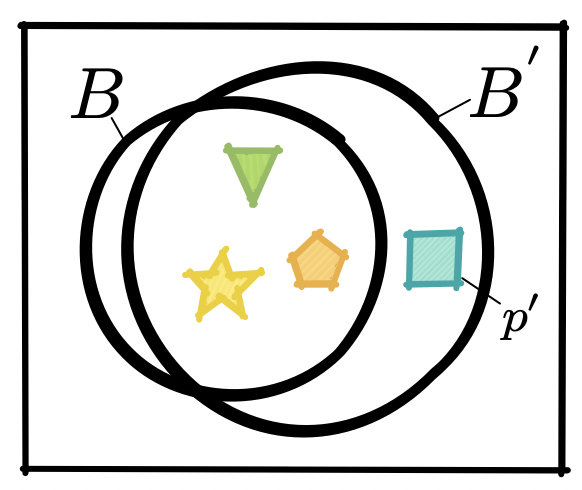
We have discussed robustness to changes in data in a number of ways:

* Evaluation on a single data set
* Evaluation on new data
* Follow up evaluation
  + Example: FairPlay: Detecting and Deterring Online Customer Misbehavior
* The role of [Ergodicity](https://en.wikipedia.org/wiki/Ergodicity) in general
  + Non-stationary distribution? Roland says this is a subset of ergodicity
* Goodhart's law: change system via deployment
  + e.g. CEO words on earnings call
* Long-term cost to optimizing for a given metric in the short term
  + E.g. click bait headline optimization

### Framework

A causal validity claim is a 4-tuple [](https://www.codecogs.com/eqnedit.php?latex=(B%2C%20B'%2C%20%5Cphi%2C%20d)#0)

* Let be [](https://www.codecogs.com/eqnedit.php?latex=P#0) be the set of all possible artifact parts (i.e. features, ML models, connections, …). Let be [](https://www.codecogs.com/eqnedit.php?latex=%5Cmathcal%7BP%7D(P)#0) the power set of all possible artifact parts (i.e., the set of all subsets). An artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B%20%5Cin%20%5Cmathcal%7BP%7D(P)#0) describes how to make a tool and is a set of artifact parts.
* We compare the original artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B#0) with a manipulated artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B'%20%5Cin%20%5Cmathcal%7BP%7D(P)#0). The change [](https://www.codecogs.com/eqnedit.php?latex=C#0) between [](https://www.codecogs.com/eqnedit.php?latex=B#0) and [](https://www.codecogs.com/eqnedit.php?latex=B'#0) is the symmetric difference [](https://www.codecogs.com/eqnedit.php?latex=B%20%5Ctriangle%20B'#0), and by definition B \cap B' \neq \emptyset.[[2]](#footnote-1)
* We can distinguish between
  + Accumulation validity claim: (**Op: Add**)
    - Adding to the original artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B#0) the artifact part [](https://www.codecogs.com/eqnedit.php?latex=p'#0) so that a manipulated artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B'#0) is created with [](https://www.codecogs.com/eqnedit.php?latex=%5C%7Bp'%5C%7D%20%3D%20B%20%5Ctriangle%20B'#0) and [](https://www.codecogs.com/eqnedit.php?latex=B'%20%5Csetminus%20B%20%3D%20p'#0) and [](https://www.codecogs.com/eqnedit.php?latex=B%20%5Csetminus%20B'%20%3D%20%5CO#0).
    - Claim: Adding [](https://www.codecogs.com/eqnedit.php?latex=p'#0) is reducing the performance according to score function [](https://www.codecogs.com/eqnedit.php?latex=%5Cphi#0).



###### Potential outcomes notation for the add operation

Potential outcomes notation is one way of thinking about causality. It has been shown to be mathematically equivalent to Pearl. Following the presentation of potential outcomes in Larry Wasserman's *All of Statistics* let us define 4 binary variables {triangle, star, pentagon, square} indicating if the corresponding component is included in the blueprint. In our experiment we will intervene to manipulate the square=S variable.

Let \vec{x}\_1 be the input data for instance 1 and \vec{x}\_2 be the input data for instance 2. Let Y\_i be the value of the score function for the instance \vec{x}\_i. Following Wasserman, define S\_1 to be the value of Y if S=1 and define S\_0 to be the value of Y if S=0. We define the causal effect of adding square to the blueprint to be \delta = E[S\_1] - E[S\_0]. Note that many ML metrics are expectations because they average over all of the instances in a dataset.

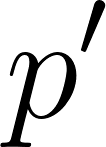
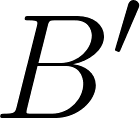
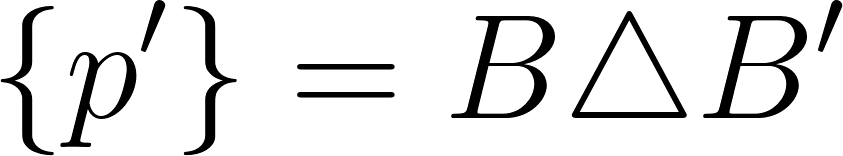
|  | triangle | star | pentagon | square=S | Y | S\_1 | S\_0 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| \vec{x}\_1 | 1 | 1 | 1 | 1 | 3 | 3 | 2 |
| \vec{x}\_1 | 1 | 1 | 1 | 0 | 2 | 3 | 2 |
| \vec{x}\_2 | 1 | 1 | 1 | 1 | 4 | 4 | 3 |
| \vec{x}\_2 | 1 | 1 | 1 | 0 | 3 | 4 | 3 |

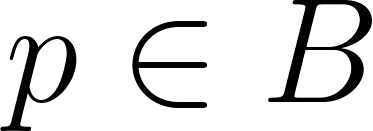
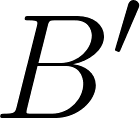
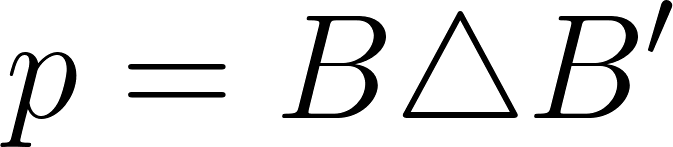
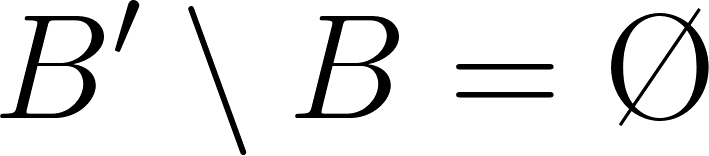
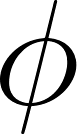
In a normal causal inference setting, you can not observe both S\_1 and S\_0 and so you have to worry about if (S\_1, S\_0) are independent of S. But in our setting we can observe both S\_1 and S\_0 and so the value of the value of (S\_1, S\_0) is the same regardless of the value of S. Therefore,

* E[S\_1] - E[S\_0] = E[S\_1 | S=1] - E[S\_0 | S=0]

because the value of S has no effect on S\_1 or S\_0.

Hence E[S\_1] - E[S\_0] = E[S\_1 | S=1 ] - E[S\_0 | S=0] = E[Y | S=1 ] - E[Y | S=0 ] where the last equality follows because by definition S\_0 is the value of Y if S=0 (and the same for S\_1). Thus simply observing the conditional probability for each instance and averaging to get the expectation is sufficient to estimate \delta in our setting. In other words, for us, \delta is just E[Y | S=1 ] - E[Y | S=0 ]. We can observe these conditional probabilities by manipulating the artifact. For this reason we don't really need potential outcomes notation.

Figure 1. In this example **add** operation the artifact part [](https://www.codecogs.com/eqnedit.php?latex=p'#0) is added to the original artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B#0) to create a new artifact [](https://www.codecogs.com/eqnedit.php?latex=B'#0). The symmetric difference is [](https://www.codecogs.com/eqnedit.php?latex=%5C%7Bp'%5C%7D%20%3D%20B%20%5Ctriangle%20B'#0)

* + Ablation validity claim: (**Op: Remove**)
    - Removing from the original artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B#0) the artifact part [](https://www.codecogs.com/eqnedit.php?latex=p%20%5Cin%20B#0) so that a manipulated artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B'#0) is created with   
      [](https://www.codecogs.com/eqnedit.php?latex=p%20%3D%20B%20%5Ctriangle%20B'#0) and [](https://www.codecogs.com/eqnedit.php?latex=B'%20%5Csetminus%20B%20%3D%20%5CO#0).  
      Claim: Removing [](https://www.codecogs.com/eqnedit.php?latex=p#0) is reducing the performance according to score function [](https://www.codecogs.com/eqnedit.php?latex=%5Cphi#0).

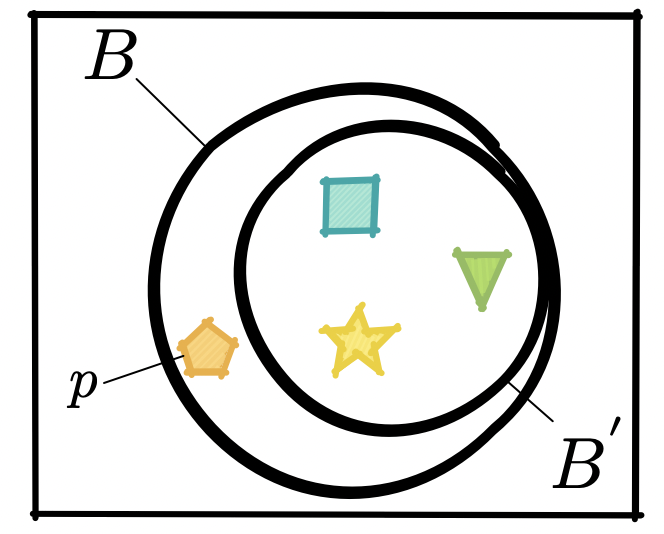
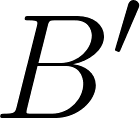
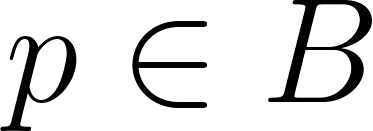
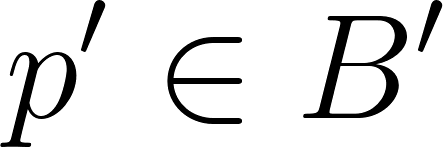
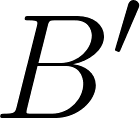
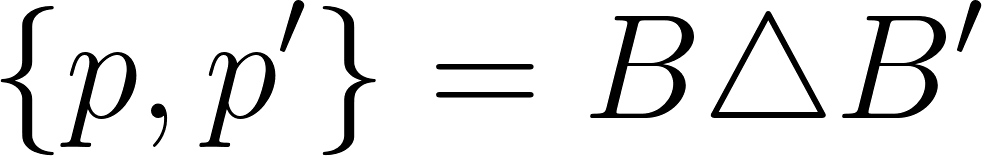
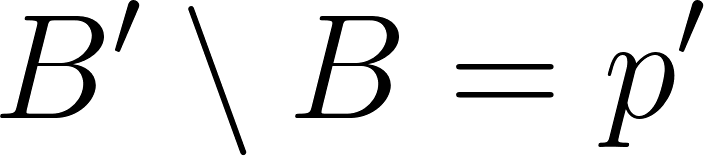
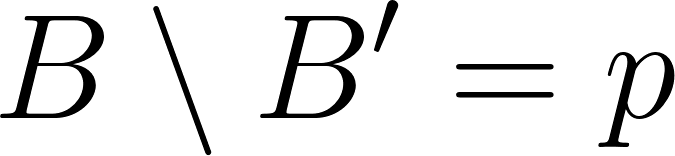
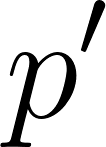
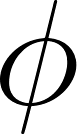


Figure 2. In this example **remove** operation the artifact part [](https://www.codecogs.com/eqnedit.php?latex=p#0) is removed from the original artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B#0) to create a new artifact [](https://www.codecogs.com/eqnedit.php?latex=B'#0).

* + Variational validity claim: (**Op: Swap**)[[3]](#footnote-2)
    - Exchanging from the original artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B#0) the artifact part [](https://www.codecogs.com/eqnedit.php?latex=p%20%5Cin%20B#0) with [](https://www.codecogs.com/eqnedit.php?latex=p'%20%5Cin%20B'#0) so that a manipulated artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B'#0) is created with [](https://www.codecogs.com/eqnedit.php?latex=%5C%7Bp%2C%20p'%5C%7D%20%3D%20B%20%5Ctriangle%20B'#0) and [](https://www.codecogs.com/eqnedit.php?latex=B'%20%5Csetminus%20B%20%3D%20p'#0) and [](https://www.codecogs.com/eqnedit.php?latex=B%20%5Csetminus%20B'%20%3D%20p#0).
    - Claim: Exchanging [](https://www.codecogs.com/eqnedit.php?latex=p#0) with [](https://www.codecogs.com/eqnedit.php?latex=p'#0) is reducing the performance according to score function [](https://www.codecogs.com/eqnedit.php?latex=%5Cphi#0).

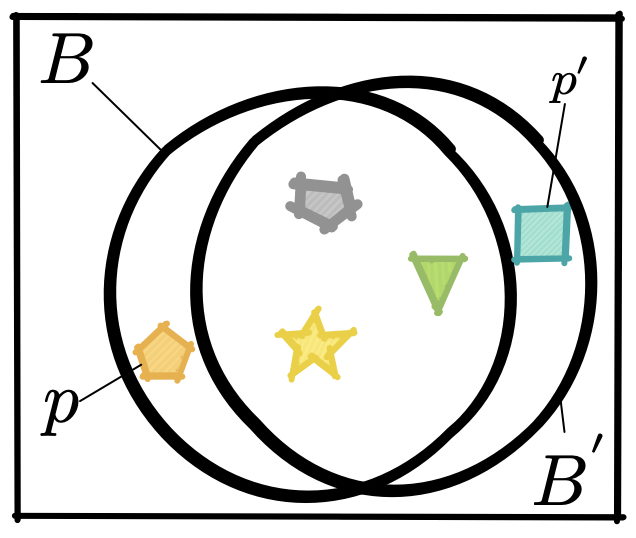
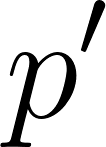
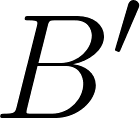
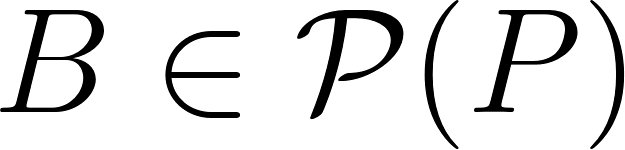
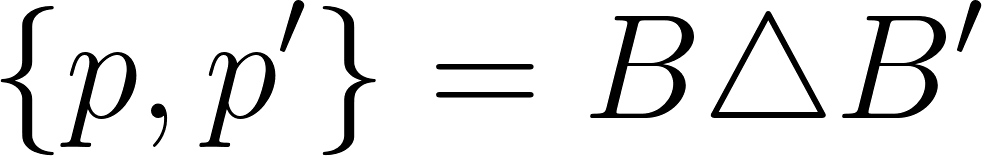
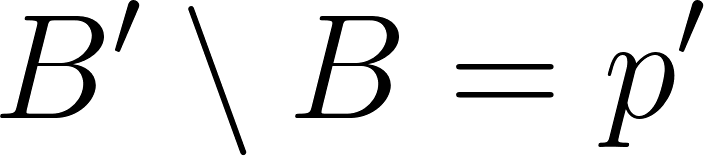
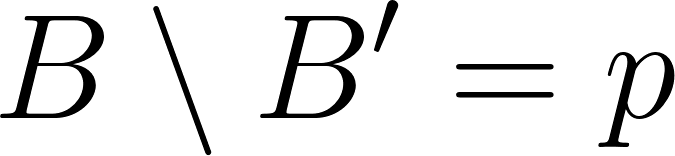


Figure 3. In this example **swap** operation the artifact part [](https://www.codecogs.com/eqnedit.php?latex=p#0) is from the original artifact blueprint [](https://www.codecogs.com/eqnedit.php?latex=B#0) is replaced with a new part [](https://www.codecogs.com/eqnedit.php?latex=p'#0)to make a new artifact [](https://www.codecogs.com/eqnedit.php?latex=B'#0).

* + AH: Salience validity claim? (**Op: Rank**)
    - Let [](https://www.codecogs.com/eqnedit.php?latex=B%20%5Cin%20%5Cmathcal%7BP%7D(P)#0) be a subset of model components, such as learned weights or fixed data. Define an ordering ≤ on [](https://www.codecogs.com/eqnedit.php?latex=%5Cmathcal%7BP%7D(P)#0) which assigns a score to each set in the power set. Say that B ≤ B' where [](https://www.codecogs.com/eqnedit.php?latex=%5C%7Bp%2C%20p'%5C%7D%20%3D%20B%20%5Ctriangle%20B'#0) and [](https://www.codecogs.com/eqnedit.php?latex=B'%20%5Csetminus%20B%20%3D%20p'#0) and [](https://www.codecogs.com/eqnedit.php?latex=B%20%5Csetminus%20B'%20%3D%20p#0).
      * The claim is that varying the value for that p, will impact the output metric more than varying another feature p'.
      * Typically p and p' are singletons
      * Examples:
        + e.g. logistic regression coefficients, salience maps then these claims are sort of true by definition
      * This one does not involve a change to the artifact …

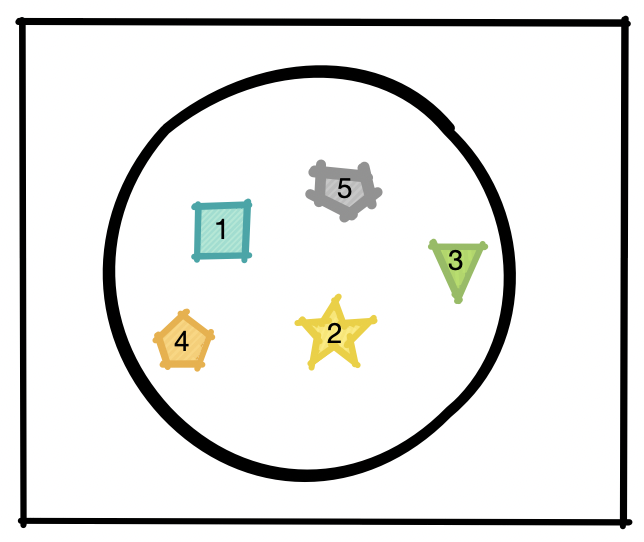


Figure 4. In this example **rank** operation the unordered set of artifact parts [](https://www.codecogs.com/eqnedit.php?latex=B#0) is mapped to a ranking that assigns an ordering to the elements from the artifact set.

The input data is not part of the artifact. To deal w/ this case of Theory vs. Google scholar. The data might be described as environment.

The score function is outside the environment.

**Criterion vs. causal.** Criterion is a weaker form of causal validity. For criterion you say these are black boxes. Then you compare them. So a kaggle competition just comparing teams is a criterion evaluation because you can't say why one team is better than the other. But if you analyze the two teams on Kaggle and say team B is better than team B' because of these specific differences then it is a causal claim. The causal claim implies a criterion claim but it is stronger. For causal validity there has to be a known intersection. For causal you have to analyze the known part and analyze the difference and attribute it to the difference.

**Intervention vs. set theoretical model**. Where is the manipulation? If we look at the Pearl causal graph. What is the intervention doing. It is fixing a value for a node. The do operator is not swapping another node. How do you match Pearl vs. set theoretic models.

**Pearl** You could have a meta graph and swap parts of the meta graph. The do operator?

There are dependencies between artifact parts. Say you use Bert, you have to use BPE tokenizer. So Bert has a dependency on the tokenizer.

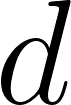
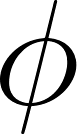
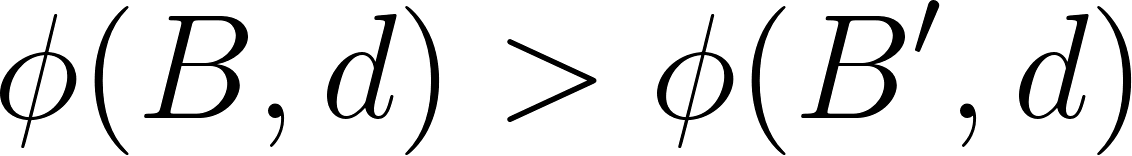
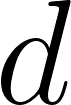
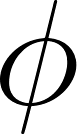
Say theory on and google scholar have the same corpus. Is that causal validity?

##### Pearl

*Comment from Roland: Very good point! This is a different claim, then an ablation claim, because we do not remove the feature but look at the change of the value of the feature in its impact to some target. A feature could be also model just as a part p. But in addition it has a link to the data d (p is the column, d (or a part of d) are the values for that column). When we say, a feature p is important, because it has a high weight in a logistic regression, we mean, that varying the value for that p, will impact the target more than varying another feature p'. Have to to still formalize this :-). Should than include explainability through the model (like logistic regression) or through SHAP,. LIME, etc.*

* Permutation feature importance
  + You have a fitted model
  + Ex-Post feature importance
  + Predict target with original data
  + For each feature:
    - Randomly shuffle the data for this feature
    - Predict target with manipulated data
    - Feature important for feature = Decrease Score
* Leave One Feature Out Feature Importance
  + Train and Predict target with original data
  + For each feature:
    - Remove Feature
    - Train Model with manipulated data
    - Predict target with manipulated data
    - Feature important for feature = Decrease Score

Note: earlier version [here](https://docs.google.com/document/d/1xuMdRdHzAyB0qM4horbgMkF0ejRj0TEvP2lDeynimCw/edit?usp=sharing)

* The comparison is based on a context or dataset [](https://www.codecogs.com/eqnedit.php?latex=d#0).
* The comparison is based on a score function [](https://www.codecogs.com/eqnedit.php?latex=%5Cphi#0) (e.g. F1, accuracy, but could also be user based…). A causal validity claim states that [](https://www.codecogs.com/eqnedit.php?latex=%5Cphi(B%2C%20d)%20%3E%20%5Cphi(B'%2C%20d)#0) for a dataset [](https://www.codecogs.com/eqnedit.php?latex=d#0).
  + AH: the score function [](https://www.codecogs.com/eqnedit.php?latex=%5Cphi#0) is almost always automatic, right? It seems to be.
  + I think a stronger claim would be to show that the automatic function also corresponds to a measurable difference in how well the system works for a user.

AH: what is the point of the framework?

Roland: you can identify missing claims.

Kai: what makes it design science? We are making predictions about the world to explain it. Is that design science? Lan and Fairy presented papers about that on Sep 9.

##### 

##### October 7th, 2022

For validities. We have:

* A frameworks
* Empirical experiments?
* A matrix of evals and claims

**How do these fit together?**

Lan, Prof. Abe’s and Prof. Roland’s paper discussion will be from last week. Content copied below.

##### Lan: The Halo Effect in Multicomponent Ratings and Its Implications for Recommender Systems: The Case of Yahoo! Movies

Presenter notes:

* This paper is trying to answer one question: Can the recommendations by collaborative filtering algorithms be improved by using multiple component ratings?
* Data: multicomponent movie rating data from Yahoo! Movies.
* A structure discovery exercise was carried out to find the dependency tree that captures most of the dependencies among the components.
* They developed a mixture model-based collaborative filtering algorithm incorporating the discovered dependency structure.
* Evaluation:
  + Compare the performance of the algorithm with several of the existing one-component and multicomponent instance-based methods.
  + Metric:
    - The rating prediction accuracy: mean absolute error(MAE)
    - the quality of top items recommended: (1) Mean precision of the top five items recommended; (2) Mean reciprocal rank (MRR) of the first relevant item recommended.

Group notes:

##### Abe: Human Identification for Activities of Daily Living

Presenter notes:

* We have discussed this paper before informally.
* The paper tries to track people's activities in elder care with sensors, in settings with multiple people. Because there are multiple people, you need to do "HID" or human identification detection.
* Proposes a transfer learning approach
* There are 4 experiments with 4 clear claims. Claim 4 is **advice to practitioners**, which I think is key to DSR.
  + Experiment 1 claim: "Deep learning's superior performance over classical machine learning suggests that automatically learning feature representations rather than manually engineering features in an ad-hoc fashion enables more effective HID."
  + Experiment 2 claim: "The finding indicates that transferring motion dynamics from wearable motion sensors is a viable approach to resolve the issue of lacking labeled training data for object motion sensor-based HID tasks.
  + Experiment 3 claim: "Cross axial patterns can more comprehensively capture motion dynamics for HID than single-axial temporal patterns"
  + Experiment 4 claim: "This suggests that practitioners can exclude horizontal activities in source domain collection if there exist time or cost constraints. … these results can help practitioners prioritize their source data collection"
    - This is advice to practitioners
    - The claim is about how to prioritize data collection.

Group notes:

##### Roland Leveraging Financial Social Media Data for Corporate Fraud Detection

* AH I moved this here from last week

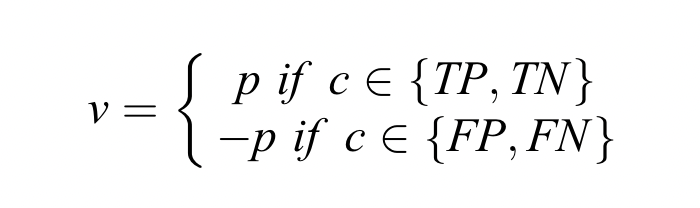
Presenter notes:

* Goal:
  + Detecting corporate fraud with ML
* Data set:
  + Data from financial social media platforms (SeekingAlpha)
* Features:
  + financial ratios
  + Social media features:
    - sentiment features
    - emotion features
    - topic features
    - lexical features
    - social network features
* Evaluation:
  + Against baseline with only
    - financial ratios
    - language-based features from the MD&A sections from the company financial report (not social media)
  + “To test the incremental effect of each category of data, we gradually add languagebased features and then social media features into financial ratios. The classification performance using three types of feature sets—(1) only financial ratios, (2) a combination of financial ratios and language-based features, and (3) a full combination of financial ratios, language-based features, and social features—are investigated. The performances of the four classifiers using these three types of feature sets are recorded in Tables 10–13.” (Dong et al., 2018, p. 477)
  + “Considering the performances of the SVM classifier in Table 10, it is clear that the performance of the combined financial ratios and language-based features is better than that using only financial ratios. Moreover, the performance of the fully combined feature set is better than that using the combination of financial ratios and language-based features. Performance is improved when more features are added. The same can be said of the NN model. This result shows that there is indeed incremental value of these three sources of information for fraud detection.” (Dong et al., 2018, p. 478)
* Robustness check:
  + detecting leaked information and rumors
  + testing the algorithm on a new data set
  + applicability check

Group notes:

##### Sen: Detecting Fraudulent Behavior on Crowdfunding Platforms: The Role of Linguistic and ContentBased Cues in Static and Dynamic Contexts

* Research Question (content-based/linguistic cues from static/dynamic data):
  + Are content-based cues extracted from static communication valuable for detecting fraudulent crowdfunding projects?
  + Are content-based cues extracted from dynamic communication valuable for detecting fraudulent crowdfunding projects?
  + Are linguistic cues extracted from static communication valuable for detecting fraudulent crowdfunding projects?
  + Are linguistic cues extracted from dynamic communication valuable for detecting fraudulent crowdfunding projects?
* Economic Evaluation metric



* As for evaluation, they just change the input to different variants from the research questions to prove their claim.
* To show the information is valuable, they compare with all pos and all neg.

##### Fairy [Design Principles for Signal Detection in Modern Job Application Systems Identifying Fabricated Qualifications]

Presenter notes: Not sure about the causal claims, so not putting the content as of now.

Group notes:

##### Kai [Paper Title]

Presenter notes:

Group notes:

##### 

##### September 30th, 2022

Presentations will be from last week (September 23rd)

Proposed dummy version: Add, remove, swap, rank. Note: rank and swap are similar

Sen: presented "The Phishing funnel model" which uses financial prediction as outcome variable.

* Is the data part of the model?
* Is the are features part of the model?
  + AH: I think we want to say yes b/c features are often important in IS

What is a causal evaluation and how is that different than a model model comparison. Kai brought this one up a few times. Seems to disagree w/ Roland

1. Same features, but throw lots of models at it.
   1. We probably want to include b/c features are important in IS
2. Throw DL and classical models at the same data
3. Throw lots of neural networks at the same data

Causal vs. criterion differences

The aim of the eval is to understand the effect of one part

Fairy presented a paper where they (A) test different methods for detecting misbehavior and then (B) test different methods to intervening to stop misbehavior.

A really convincing causal validity explanation is not only is B' better than B based on some metric, but it is actually important for the organization.

##### 

##### September 23rd, 2022

##### 

##### Framework discussion Sep 23

AH: Kai's question from last week: what makes this a convincing evaluation?

What if we have a matrix where rows are papers and columns are evaluations? That could be a centerpiece of the paper. For this the key things to annotate are:

* What is the causal validity claim?
  + Implicit?
    - Do you have an evaluation w/o any claim
    - The eval is so obvious you don't even need the claim
    - There are necessary and sufficient conditions for the claim that are observable
    - There is no H1, H2 etc in DSR
    - Linking to design principles
    - Design principles vs. validity claims?
      * There are multiple instantiations of a design principle.
      * It's more like a design requirement
  + Explicitly claim
    - We need to annotate them and have guidelines that can be verified
    - There is a hypothesis about what exists
      * This is what we expect to find
      * This is the evidence is we found
    - External validity (testing it in multiple settings)
    - Ecological validity
      * Does it work in practice? You implemented it in real life.
      * There is an implicit philosophy to deployment and it is used
        + Your assumptions about the problem are valid/important
        + You solved the right problem
        + The science is true b/c it help solve a problem
        + Implicit philosophy of pragmatism.
      * What is interesting about what Amazon does?
        + Is it only relevant for Amazon.
        + Some of the details are specific to Amazon
        + Some of the details will matter and can be generalized and others can't
      * What if you did not believe in newtonian physics. The first moon trip is convincing.
        + Do people use it?
        + Do people get value from using it?
  + Roland:
    - Sometimes claims are part of a RQ.
    - They say: will this additional layer help.
* What evaluations do they do to support the claim? This is a set w/ some kind of ordering ->
  + We need the set. We need a list.
    - There are combos of facets or assets

We then need to describe all of their claims in our set theoretic framework in a way that suggests how to make the claims stronger, or what they are missing.

Sen + Kai: we need to do some kind of dimensionality reduction, and also make normative claims based on the matrix.

* What is currently being done?
* What should be done?

### Presentations

#### Sep 23

##### Sen [The Phishing Funnel Model: A Design Artifact to Predict User Susceptibility to Phishing Websites]

Presenter notes:

* Rather than predicting whether a link or website is a phishing attack, we seek to accurately predict ***users’ phishing susceptibility***
* ***RQ1:*** model’s performance on predicting users’ phishing susceptibility
* ***RQ2:*** model’s performance on intervention driven by predictions of phishing susceptibility
* ***Data Source:*** They did a field experiment lasting 12 months and collected user-phsingwebsite-iterection data from two companies employees.
* **Their model’s effective:** Compare with baseline models, and their model’s performance is the best one
* **Feature Importance:** remove one feature from the input one by one, and then observe the AUC’s change. Interesting points: all feature input does not provide highest auc, which means some feature is a noise
* **Model Part Importance:** we compared PFM-SVORCK versus PFM-CLMM directly to determine which setting performed best overall. In these comparisons, PFM-SVORCK outperformed PFM-CLMM
* **Interesting Metrics:** Saved money or reduced cost in one organization as the metric

Group notes:

##### Fairy [FairPlay: Detecting and Deterring Online Customer Misbehavior]

Presenter notes:

* **Dataset** - customer behavior data from the online brand community (OBC).
* **Proposed framework** -
  1. FairPlay, a misbehavior detection method that will automatically detect/identify customer misbehavior using ML
  2. manage the customer behavior through norm enforcement strategies
* Study integrates design science and experimental design with econometric analysis
* **Evaluation(s)** -
  1. Comparison of FairPlay (customer misbehavior detection algo) against baseline (XGBoost).
     + Metrics - Precision, Recall and F1 score.
  2. Comparison of FairPlay (customer misbehavior detection algo) against state-of-the-art (Dorris et al. 2020)
     + Metrics - Precision, Recall and F1 score.
  3. Comparison of model part importance (Metadata only, Text only, Text and Metadata)
  4. Incremental contribution of each category of input features (horse-race manner)
     + Metric - AUC.
  5. Artifact features comparison (treatment groups, control group)
     + Metric - p-value.
* Follow-up evaluation - Used the same field experiment but introduced AI detector component for customer misbehavior deterrence. [However, not sure if its causal validity since this disclosure is not part of the artifact].

Group notes:

##### Lan: The Halo Effect in Multicomponent Ratings and Its Implications for Recommender Systems: The Case of Yahoo! Movies

Presenter notes:

* This paper is trying to answer one question: Can the recommendations by collaborative filtering algorithms be improved by using multiple component ratings?
* Data: multicomponent movie rating data from Yahoo! Movies.
* A structure discovery exercise was carried out to find the dependency tree that captures most of the dependencies among the components.
* They developed a mixture model-based collaborative filtering algorithm incorporating the discovered dependency structure.
* Evaluation:
  + Compare the performance of the algorithm with several of the existing one-component and multicomponent instance-based methods.
  + Metric:
    - The rating prediction accuracy: mean absolute error(MAE)
    - the quality of top items recommended: (1) Mean precision of the top five items recommended; (2) Mean reciprocal rank (MRR) of the first relevant item recommended.

Group notes:

##### Abe: Human Identification for Activities of Daily Living

Presenter notes:

* We have discussed this paper before informally.
* The paper tries to track people's activities in elder care with sensors, in settings with multiple people. Because there are multiple people, you need to do "HID" or human identification detection.
* Proposes a transfer learning approach
* There are 4 experiments with 4 clear claims. Claim 4 is **advice to practitioners**, which I think is key to DSR.
  + Experiment 1 claim: "Deep learning's superior performance over classical machine learning suggests that automatically learning feature representations rather than manually engineering features in an ad-hoc fashion enables more effective HID."
  + Experiment 2 claim: "The finding indicates that transferring motion dynamics from wearable motion sensors is a viable approach to resolve the issue of lacking labeled training data for object motion sensor-based HID tasks.
  + Experiment 3 claim: "Cross axial patterns can more comprehensively capture motion dynamics for HID than single-axial temporal patterns"
  + Experiment 4 claim: "This suggests that practitioners can exclude horizontal activities in source domain collection if there exist time or cost constraints. … these results can help practitioners prioritize their source data collection"
    - This is advice to practitioners
    - The claim is about how to prioritize data collection.

Group notes:

##### Roland Leveraging Financial Social Media Data for Corporate Fraud Detection

* AH I moved this here from last week

Presenter notes:

* Goal:
  + Detecting corporate fraud with ML
* Data set:
  + Data from financial social media platforms (SeekingAlpha)
* Features:
  + financial ratios
  + Social media features:
    - sentiment features
    - emotion features
    - topic features
    - lexical features
    - social network features
* Evaluation:
  + Against baseline with only
    - financial ratios
    - language-based features from the MD&A sections from the company financial report (not social media)
  + “To test the incremental effect of each category of data, we gradually add languagebased features and then social media features into financial ratios. The classification performance using three types of feature sets—(1) only financial ratios, (2) a combination of financial ratios and language-based features, and (3) a full combination of financial ratios, language-based features, and social features—are investigated. The performances of the four classifiers using these three types of feature sets are recorded in Tables 10–13.” (Dong et al., 2018, p. 477)
  + “Considering the performances of the SVM classifier in Table 10, it is clear that the performance of the combined financial ratios and language-based features is better than that using only financial ratios. Moreover, the performance of the fully combined feature set is better than that using the combination of financial ratios and language-based features. Performance is improved when more features are added. The same can be said of the NN model. This result shows that there is indeed incremental value of these three sources of information for fraud detection.” (Dong et al., 2018, p. 478)
* Robustness check:
  + detecting leaked information and rumors
  + testing the algorithm on a new data set
  + applicability check

Group notes:

##### Kai [Paper Title]

Presenter notes:

Group notes:

##### 

##### September 16th, 2022

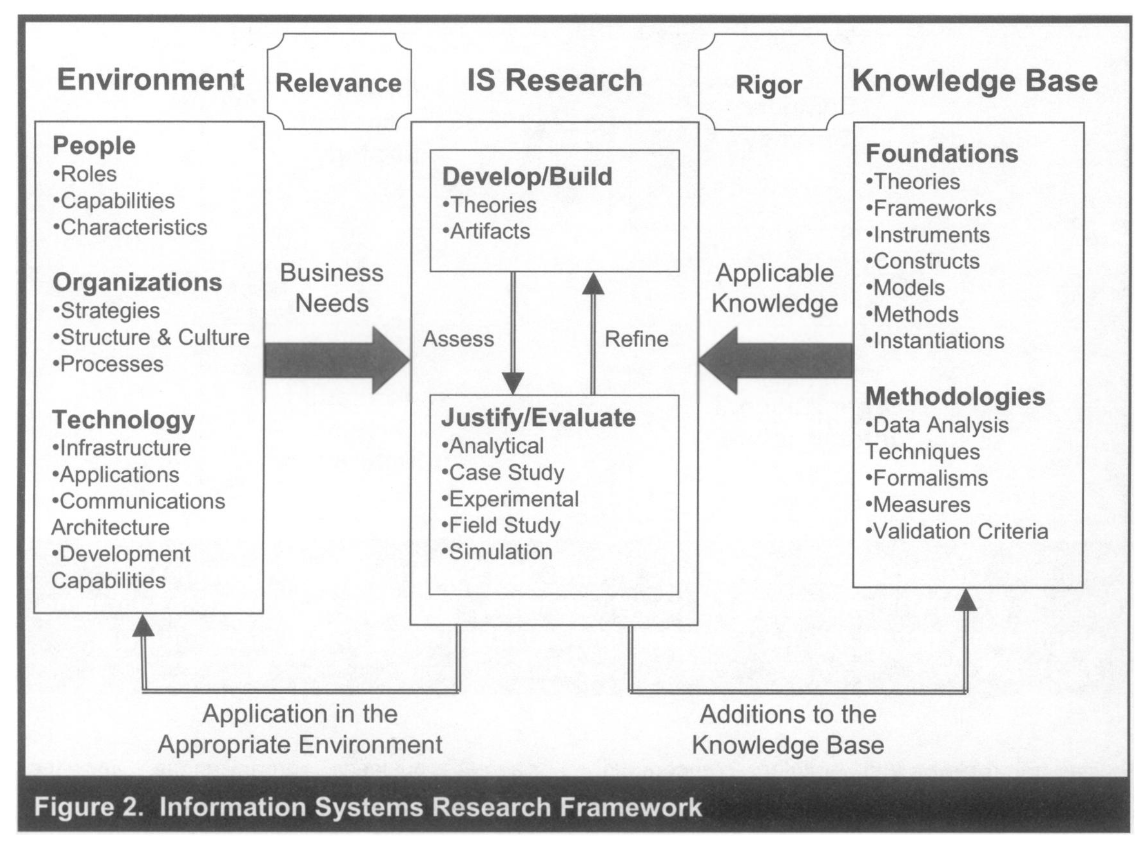
##### 

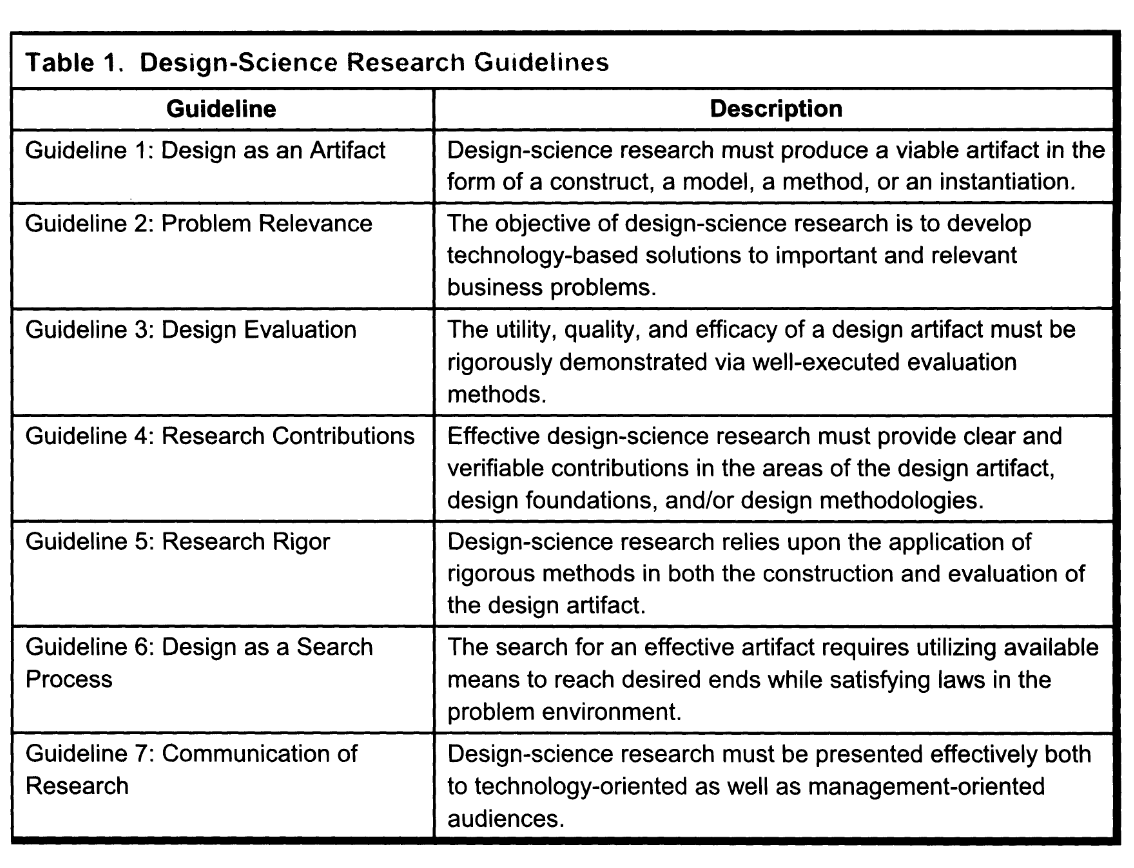
##### Framework discussion Sep 16

AH: are we looking at a subset of design science papers? How much of current design science has this applied ML feel? What does the venn diagram look like? If we apply this framework to DSR we can identify missing claims, like Roland suggests.

DSR from [Hevner (2004)](https://www.jstor.org/stable/pdf/25148625.pdf?refreqid=excelsior%3A8f74ac88d79bb2244242dd65bedc9581&ab_segments=&origin=&acceptTC=1):

* *"the design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts"*
* *"it seeks to create innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, management, and use of information systems can be effectively accomplished"*
* *"Theory and utility are two sides of the same coin"*

**



* G1: Many of the papers we are looking at offer artifacts. I was confused b/c the artifacts are not tested in software but Hevner says this may be OK - AH
* G2: Is the difference between B and B' in the score function relevant to the org/user? Probably not for most papers but ideally yeah.
* G3: Many of the papers don't do this. This has been confusing me -AH
* G4: I think a lot of the claims we have seen so far are a kind of G4
* G5: We can test these claims and will find that many violate G5
* G6: Not sure about this one. My personal experience is that design is a search process; see also [Mike Bostock talk](https://www.youtube.com/watch?v=fThhbt23SGM). What does this mean in our setting?
* G7: Not sure about this one.

Also look at: Gregor and Hevner 2013, Kai's other validities paper. Abe will look some more. Take a look at the ML editorial in the quarterly.

Kai: What is the role of causal claims? That design knowledge is correct. If you add this component to the artifact it will be better. Then you do an eval that says are you right?

Kai: what makes an evaluation convincing?

* Is it true? Is it sound?
  + Bonferroni correction?
* Is it important?
  + That is somewhat subjective.

### Presentations

#### Sep 16

How to do a really convincing causal efficacy evaluation?

##### Sen [Eye-Tracking-Based Classification of Information Search Behavior Using Machine Learning: Evidence from Experiments in Physical Shops and Virtual Reality Shopping Environments]

Presenter notes:

* **What is this for?** Use eye-tracking information to predict shopping motivation. Shopping motivation can be used to improve products recommendation system
* **What is evaluated?** New features: eye-tracking information. Model Part: Ensemble model is better than single model.
* **How to evaluate?** Experiments in two fields: VR and real supermarket
* Do they really evaluate their eye-tracking features in this paper?

Group notes:

* Random guess is enough for evaluation
* They did not compare to a reasonable baseline so not convincing.

##### Fairy [Is Hidden Safe Location Protection against Machine-Learning Prediction Attacks in Social Networks]

Presenter notes:

* Using DSR, the research framework is proposed for estimating and managing the exposure risk of users’ hidden current city-
  + In the first step, the framework proposes a current city prediction framework to integrate comprehensive location-indicative information (i.e., demographic attributes, behaviors, and relationships) with a variety of state-of-the-art machine learning algorithms and simulate aggressive prediction attack models.
  + proposes an exposure risk estimator that would notify users of the exposure risk of their hidden information and devise a risk manager intended to offer users potential countermeasures for risk management.
  + In the third step, analyze the exposure risk of users’ hidden information from various perspectives at the individual and group level and offer several managerial implications for the privacy management of Online Social Networks (OSNs).
* Causal validity claims -
  + Claim 1 - Current city prediction framework outperforms typical benchmark approaches using extensive experiments on two Facebook datasets.
  + Claim 2 - Exposure risk estimation approach outperforms
* Causal efficacy validity evaluations -
  + Current city prediction framework :
    - Evaluates a variant version of DBR framework to show effectiveness of the proposed location determination approach in the DBR framework.
    - Experiment - Two facebook datasets.
    - Metrics - Accuracy within K km, and Average Error Distance (AED)
  + Current city exposure risk :
    - “leave -one-feature-out” approach
    - Comparing exposure risk model (with all features) by removing one feature at a time.

Group notes:

* No statistical tests
* They use real and simulated data. Simulated data is less convincing?

##### Lan: MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud

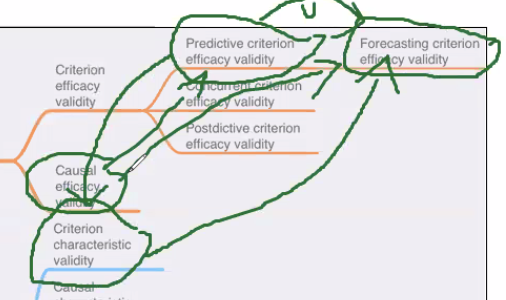
Presenter notes:

* In this paper, the authors used a design science approach to develop MetaFraud, a novel meta-learning framework for detecting financial fraud using publicly available information.
* Framework:
  + Robust feature construction, which includes organizational and industry contextual information and the use of quarterly and annual data;
  + A method of using stacked generalization, semi-supervised learning, and adaptive/active learning.
* Evaluation:
  + Experiment 1: Comparing Context-Based Classifiers against Baseline Classifier
  + Experiment 2: Evaluate stacked classifier.
  + Experiment 3: Evaluated the performance of adaptive learning versus a static learning model.
  + Experiment 4: Assess the overall efficacy of the proposed MetaFraud framework in comparison with three prior approaches that attained state-of-the-art results.
  + Experiment 5: Evaluate MetaFraud in comparison with existing semi-supervised learning methods-Tri Training.
* They claimed that they used the design science approach and mentioned the design science guideline several times.

Group notes:

* They did not do a human evaluation.
* They did not go deploy it at a bank. That would make it more convincing. -AH
* Why is this not ecologically valid?
  + Well, does the dataset have the same assumptions as banks?
  + What are auditors actually do and does the model help?
    - Is precision and recall ?\
* Future predictions:
  + Did the deploy it?
  + Are they predicting the right thing?
  + Is there dataset drift? That could make it less compelling
  + Do they have the right assumptions for actual use?
* Did you speak with practitioners?
  + If you ask: hey would you use this people will just be like sure
  + The real important driver is human behavior.

Kai: we are trying to get causal efficacy validity as it relates to these other kinds of validities.



##### Roland Leveraging Financial Social Media Data for Corporate Fraud Detection

Presenter notes:

* Goal:
  + Detecting corporate fraud with ML
* Data set:
  + Data from financial social media platforms (SeekingAlpha)
* Features:
  + financial ratios
  + Social media features:
    - sentiment features
    - emotion features
    - topic features
    - lexical features
    - social network features
* Evaluation:
  + Against baseline with only
    - financial ratios
    - language-based features from the MD&A sections from the company financial report (not social media)
  + “To test the incremental effect of each category of data, we gradually add languagebased features and then social media features into financial ratios. The classification performance using three types of feature sets—(1) only financial ratios, (2) a combination of financial ratios and language-based features, and (3) a full combination of financial ratios, language-based features, and social features—are investigated. The performances of the four classifiers using these three types of feature sets are recorded in Tables 10–13.” (Dong et al., 2018, p. 477)
  + “Considering the performances of the SVM classifier in Table 10, it is clear that the performance of the combined financial ratios and language-based features is better than that using only financial ratios. Moreover, the performance of the fully combined feature set is better than that using the combination of financial ratios and language-based features. Performance is improved when more features are added. The same can be said of the NN model. This result shows that there is indeed incremental value of these three sources of information for fraud detection.” (Dong et al., 2018, p. 478)
* Robustness check:
  + detecting leaked information and rumors
  + testing the algorithm on a new data set
  + applicability check

Group notes:

##### Kai [Paper Title]

Presenter notes:

Group notes:

#### 

#### 

##### September 9th, 2022

#### 

#### Sep 9

##### Sen [Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning]

Presenter notes:

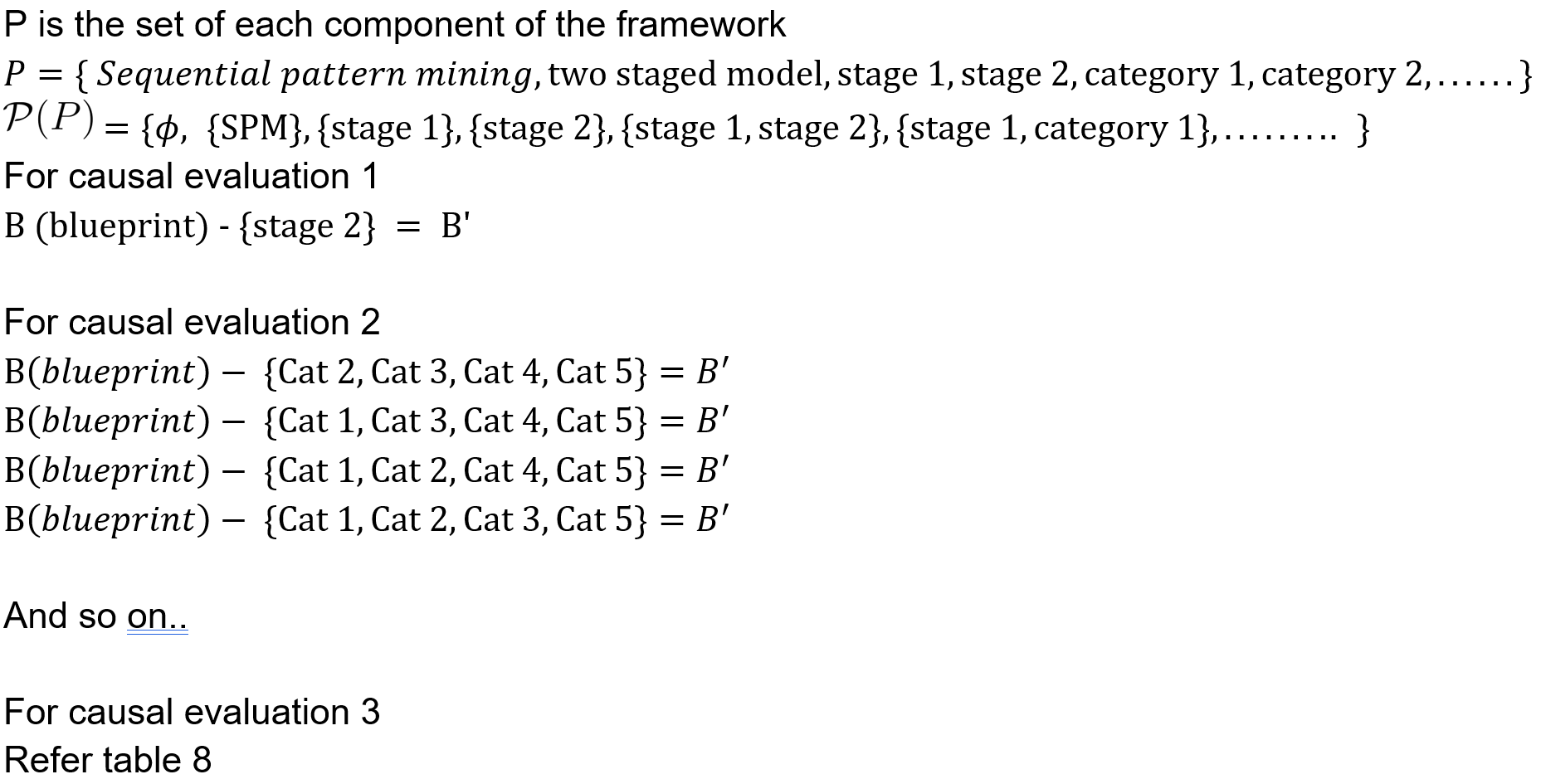
* [pre\_sep\_9.pptx](https://docs.google.com/presentation/d/1pAG7Z1CiQc040V7EUr1uxHA05qxtgqic/edit?usp=sharing&ouid=113273448986284096392&rtpof=true&sd=true)

Group notes:

##### Fairy [A Comprehensive Analysis of Triggers and Risk Factors for Asthma Based on Machine Learning and Large Heterogeneous Data Sources]

Presenter notes:

* This paper focuses on identifying and understanding triggers and risk factors that contributes to asthma exacerbations.
* Limitation of current methods:
  + Traditional data collection.
  + Lack of detailed assessment of risk factors.
  + No comprehensive study that focuses on multiple types of factors.
* Research questions:
  + Social media data
  + Big data - feature importance, feature interactions
* Artifact – Asthma Triggers and Risk Factors Analyses Framework (Β)
* Causal validity claim (ablation validity):
  + Two-stage Social Media based user attributes extraction model
  + SPM and Random Forest for extracting asthma triggers/risk factors
* Extra



Group notes:

##### Lan: Predicting Labor Market Competition: Leveraging Interfirm Network and Employee Skills

Presenter notes:

* In this paper, the authors proposed the human capital overlap metrics derived from employees’ skills and career mobility patterns across firms. Their claim is that the proposed metrics are critical to predict labor market competitors.
* Human capital overlap metrics: Labor overlap and HCF(Human Capital Flow) Network Overlap
* Validation:
  + Two control metrics: Basic Economic Metrics and Product Overlap Metrics.
  + They experiment with their proposed set of metrics for the prediction of future labor market competition
  + They evaluate the models first with just the economic metrics, followed by incrementally adding product overlap, labor overlap, and HCF network overlap to the predictor set. AUC is used as the performance evaluation metric.

Group notes:

* They add features one by one
* They don't check all possible combinations
* What is the artifact? The artifact is the metric. They describe the artifact as the metric
* AH: we may want to find a way to exclude this one? Seems like it opens lots of slippery issues.

##### Abe "Go to YouTube and call me in the morning"

Presenter notes:

* This one is interesting b/c it is using ML to make a causal claim about how medical information causes video engagement.
* The claim seems to be: too much or too little medical information harms engagement
* I don't understand all of the experimental details, but I think it is important to bring to the group's attention because this is what I think many reviewers will think of when we say "causal claims"
* Now that we have a sharper definition of causal claim I actually think we can/should exclude this one, because they are not modifying an artifact blueprint.
* But we need to remember this paper (via these notes) because we need to be really clear that we DO NOT mean [causal inference](https://mixtape.scunning.com/05-matching_and_subclassification)
* Why exclude: it is about explaining the world and not the artifact.
  + Causal inference is about explaining the world
  + It is NOT what we are talking about.
  + This needs to go in abstract or intro or everyone is confused
* Kai:
  + What about if we have a model and the variable is important for the model
  + What is that includes the paper.
  + What about you say that medical engagement is important and you add it to model and engagement goes up.
* Kinds of research:
  + Experimental research (you can manipulate factors)
  + Observational research (you can not manipulate factors)
  + Design science research
    - What is different?
      * Sen: its about explaining the artifact

Group notes:

##### Roland “Prediction in Economic Networks” (Dhar et al., 2014)

Presenter notes:

* Features of the product network (co-purchase network) can be used for improving demand prediction.
* Multiple claims of the relevance of different features.
* Ablation validity: “5.1.1. Prediction with Partial Information. The second sensitivity analysis considers the predictive accuracy of AR and NN models when less information about the network is available, specifically, when no InDemand information is available, and only Historical Demand and network-specific data are observable.” (Dhar et al., 2014, p. 275)
* Accumulation validity “5.1.2. Adding Price Information. As another form of robustness check we added book price data into our models. However, at best, price data contributed only marginally to predictive performance.” (Dhar et al., 2014, p. 275)

Group notes:

* Adding a feature will NOT improve the model

##### Kai [paper title]

Presenter notes:

Group notes:

1. **Comment:** There is a difference between running two Kaggle models and seeing which one gets a higher score, which is automatic comparison without attribution and examining the Kaggle models to identify differences which explain the score. [↑](#footnote-ref-0)
2. In theory, it is possible to add and remove items from B to create a B’ such that B \cap B’ = \emptyset. However, we exclude such cases from our analysis because prior work in IS [cite kai and roland] defines a criterion validity claim to be a statement about the difference between an artifact B and a totally different artifact B’. In this paper, we focus on causal validity in which two artifacts B and B' have at least some parts in common, i.e., B \cap B’ \neq \setempty. [↑](#footnote-ref-1)
3. From a set theoretic perspective, add and remove are sufficient for describing all possible transformations from B to B’. This is because B \triangle B’ = B \cup B’ \setminus B \cap B’. Hence any transformation from B to B’ may be described in terms of (1) zero or more add operations which add each element in B \cup B’ to B and (2) zero or more remove operations which take away each element in B \cap B’ from B. Although add and remove are theoretically sufficient to describe all possible transformations from B to B’, we introduce the composite operation swap (consisting of one or more add or remove operations) because we found it to be a simpler and more direct way to describe the kinds of causal validity claims in machine learning design science papers. [↑](#footnote-ref-2)