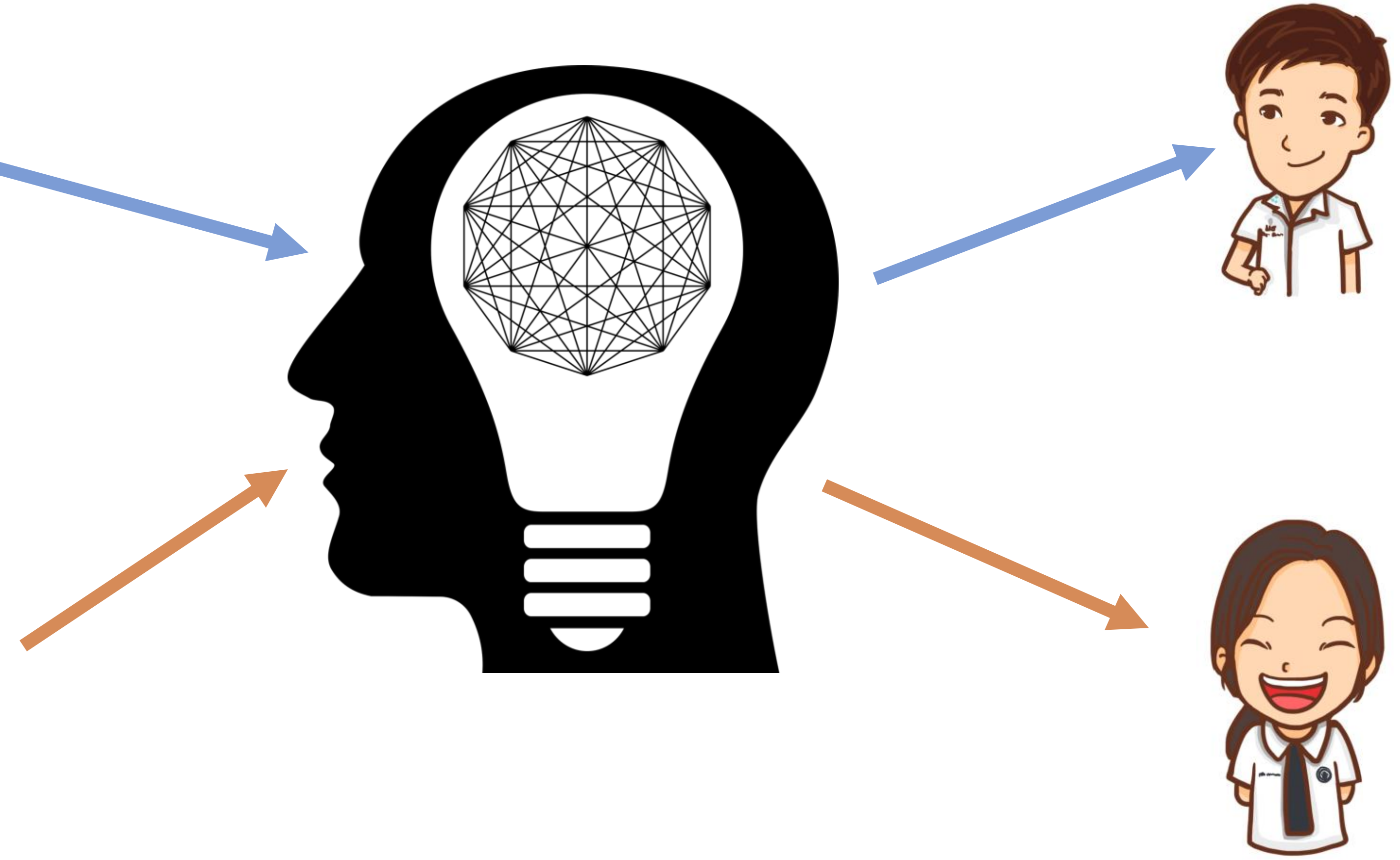
When do Words Matter ?

Understanding the impact of lexical choice on audience perception
using Individual Treatment Effect Estimation

Zhao Wang and Aron Culotta
Illinois Institute of Technology

Motivation

- A gift for my *wife*.
- A gift for my *husband*.



Causal Effect of Lexical Choice on Audience Perception

- Single linguistic change
- Perception of one sentence

Related work

- Wording effect
  - *Message propagation*
  - *Memorability of movie quotes*
  - *Story sharing rates*
  - *User attribute*
  - *Human perception*
  - *Gender obfuscation*

- Causal inference (Individual Treatment Effect estimation)
  - *Drug use on health (medical)*
  - *Lexical choice on perception*

ITE: *Individual Treatment Effect estimation*

- Treatment Effect Estimation:
  - RCT (A,B) test

- Whether a drug is effective for a patient?
  - Can only observe one outcome per individual
  - Fundamental problems in observational study

ITE: *Individual Treatment Effect estimation*

$$D = \{(\mathbf{X}_1, T_1, Y_1), \ldots, (\mathbf{X}_n, T_n, Y_n)\}$$

- X : covariate vector (e.g., *gender, age, height*)
- T : treatment indicator, $T_i \in \{0,1\}$
  - $T_i = 0$ *control group (patient did or did not take the drug)*
  - $T_i = 1$ *treatment group*
- Y : observed outcome

$$\tau(\mathbf{x}) = \mathbb{E}[Y^{(1)}|\mathbf{X} = \mathbf{x}] - \mathbb{E}[Y^{(0)}|\mathbf{X} = \mathbf{x}]$$

$$\hat{\tau}(\mathbf{x}) = \mathbb{E}[Y|T = 1, \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y|T = 0, \mathbf{X} = \mathbf{x}]$$

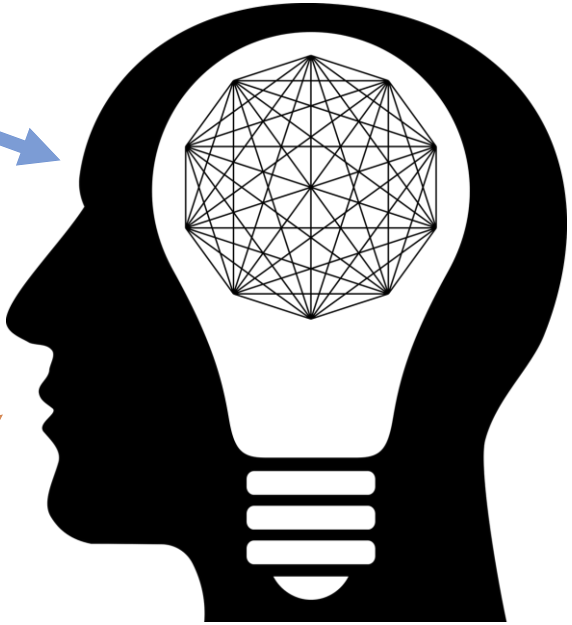
$$= \frac{1}{|S_1(\mathbf{x})|} \sum_{i \in S_1(\mathbf{x})} Y_i - \frac{1}{|S_0(\mathbf{x})|} \sum_{i \in S_0(\mathbf{x})} Y_i$$

Strongly Ignorable Treatment Assignment (SITA):

$$T \perp \{Y^{(0)}, Y^{(1)}\} \mid \mathbf{X}$$

LSE: *Lexical Substitution Effect estimation*

- Plenty of *shops* nearby.



- Plenty of *boutiques* nearby.

ITE → LSE

	Clinical Domain	Language Domain	Example
X	covariate vector for an individual	words in a sentence, omitting the word to be substituted	*"Plenty of __ nearby"*
T	drug treatment indicator	word substitution indicator	$T = 0$: *"Plenty of* **shops** *nearby"* $T = 1$: *"Plenty of* **boutiques** *nearby"*
Y	health outcome	human perception	Human perception of the desirability of a rental listing containing the sentence *"Plenty of boutiques nearby"*

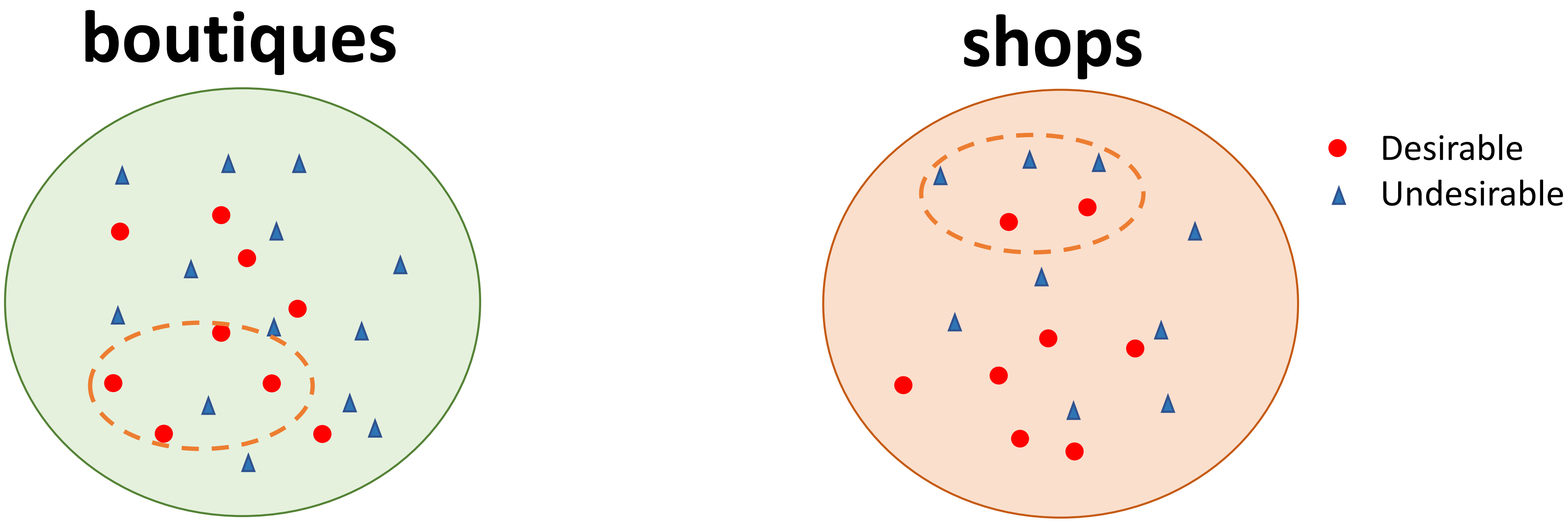
$$\hat{\tau}(\mathbf{x}, p) = \frac{1}{|S_1^p(\mathbf{x})|} \sum_{i \in S_1^p(\mathbf{x})} Y_i - \frac{1}{|S_0^p(\mathbf{x})|} \sum_{i \in S_0^p(\mathbf{x})} Y_i$$

Methods

- Quasi-experiment with observational data $(\boldsymbol{w_i}, \boldsymbol{w_j}, \boldsymbol{s}) \rightarrow \hat{\boldsymbol{\tau}}$
    1. KNN --> *K-Nearest Neighbor matching*
    2. VT-RF --> *Virtual Twins Random Forest*
    3. CF-RF --> *Counterfactual Random Forest*
    4. CSF --> *Causal forest*

- Classification $(\boldsymbol{w_i}, \boldsymbol{w_j}, \boldsymbol{s}, \tau) \rightarrow \hat{\boldsymbol{\tau}}$
    1. Causal perception classifier (RCT)

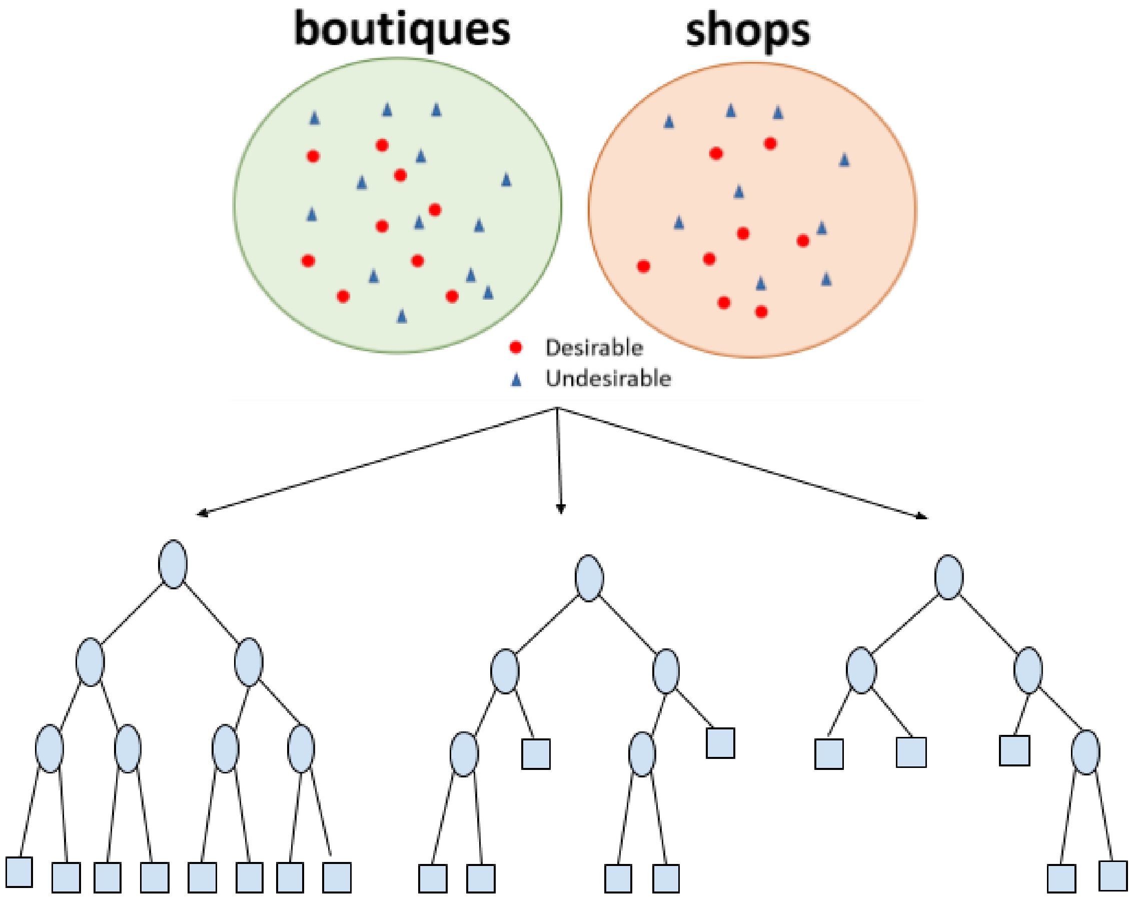
KNN: *K-Nearest Neighbor matching*



“Plenty of __ nearby”

$$\hat{\tau}_{KNN}(\mathbf{x}) = \Big(\frac{1}{K}\sum_{i\in S_1(\mathbf{x},K)} Y_i\Big) - \Big(\frac{1}{K}\sum_{i\in S_0(\mathbf{x},K)} Y_i\Big)$$

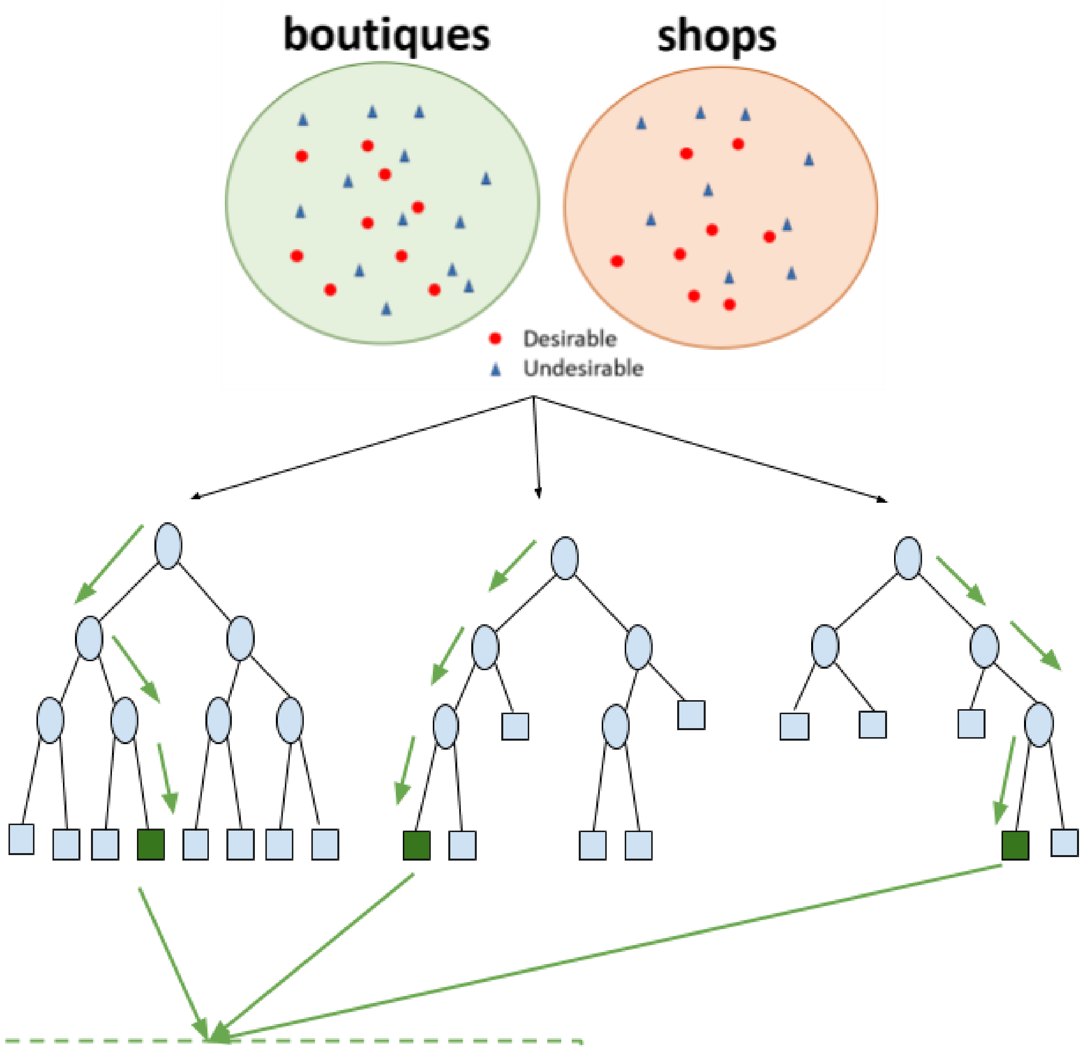
VT-RF: *Virtual Twins Random Forest*



Virtual Twin:
- "Plenty of *shops* nearby"
- "Plenty of *boutiques* nearby"

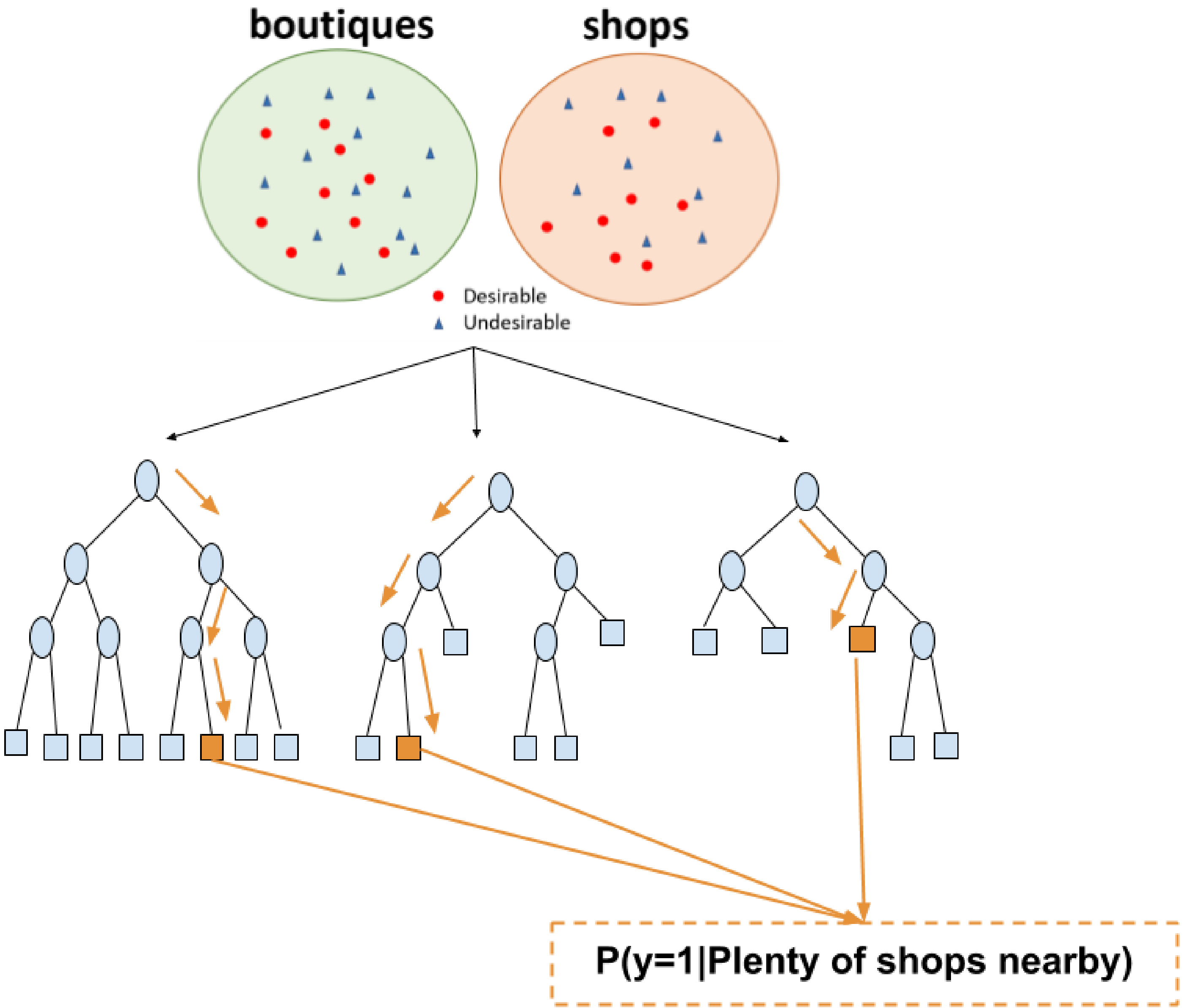
$$\hat{\tau}_{VT}(\mathbf{x}) = \hat{Y}(\mathbf{x}, 1) - \hat{Y}(\mathbf{x}, 0)$$

VT-RF: *Virtual Twins Random Forest*

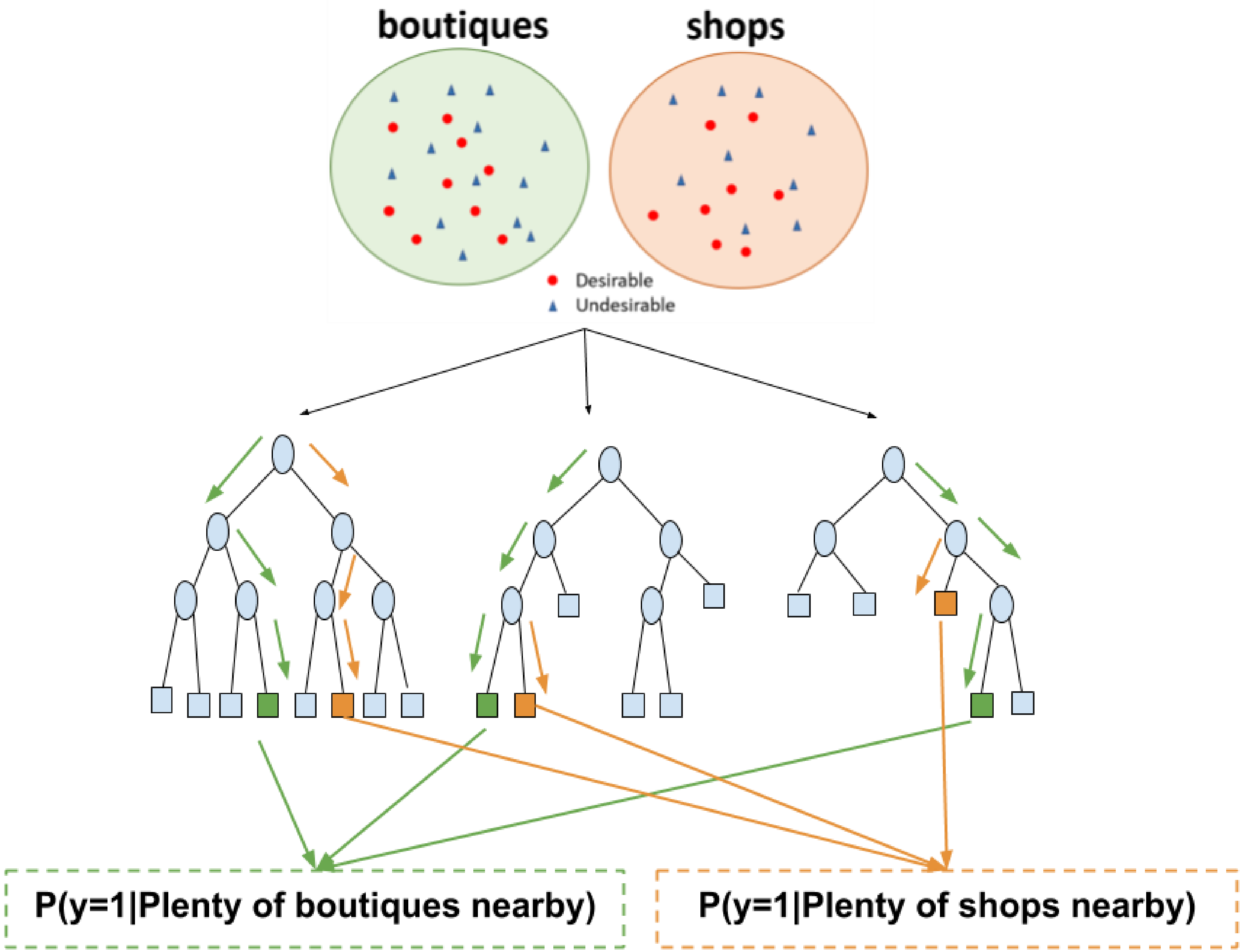


P(y=1|Plenty of boutiques nearby)

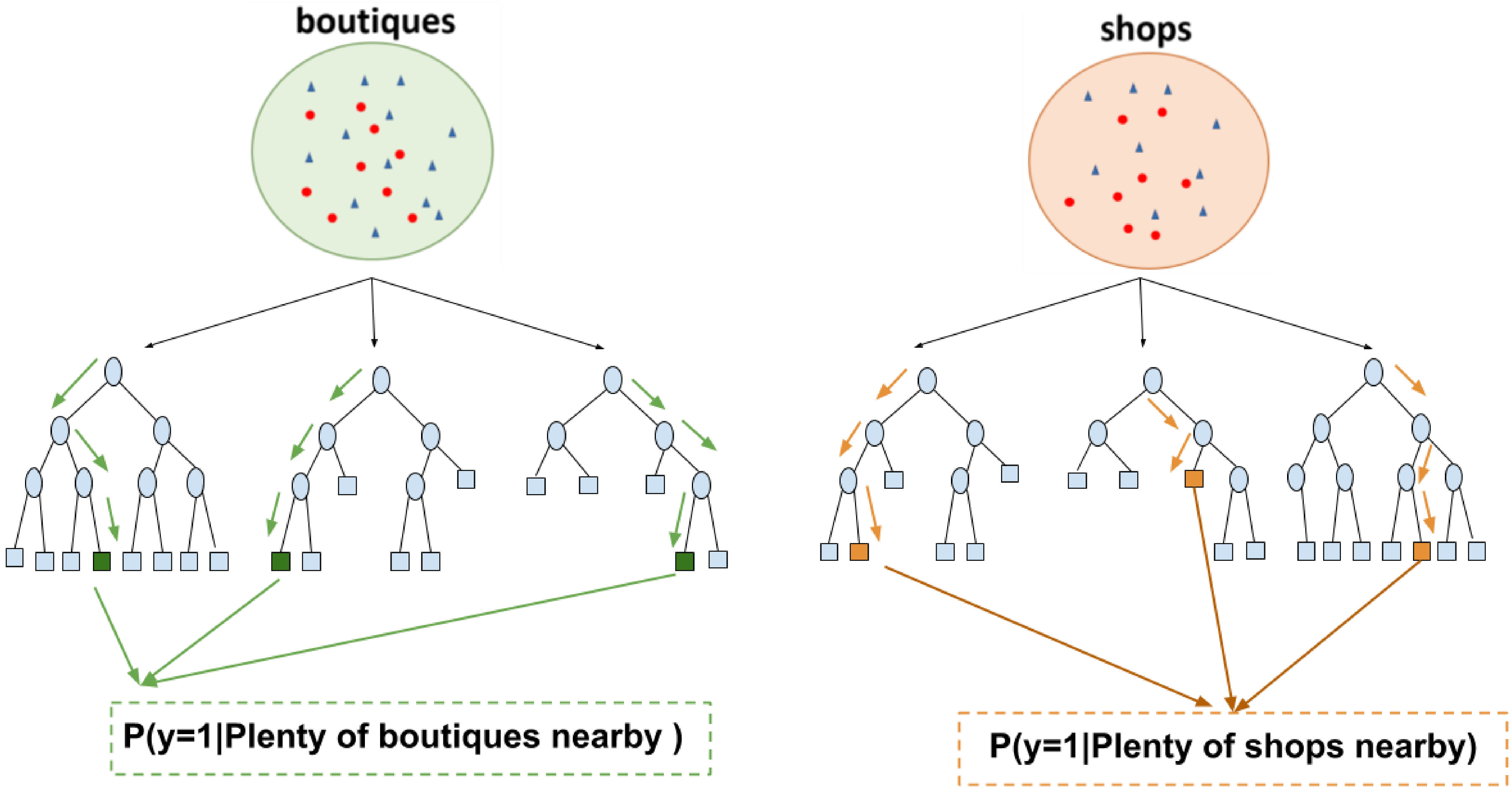
VT-RF: *Virtual Twins Random Forest*



VT-RF: *Virtual Twins Random Forest*



CF-RF: *Counterfactual Random Forest*



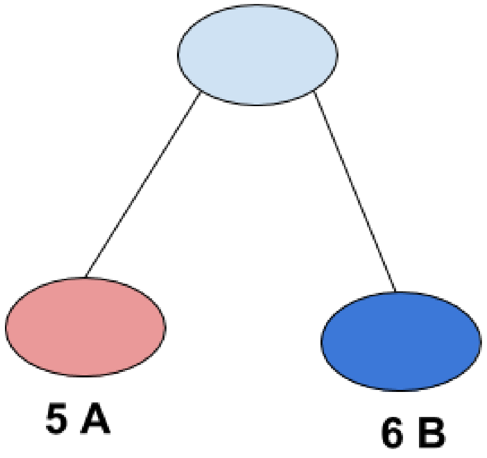
$$\hat{\tau}_{CF}(\mathbf{x}) = \hat{Y}_1(\mathbf{x}, 1) - \hat{Y}_0(\mathbf{x}, 0)$$

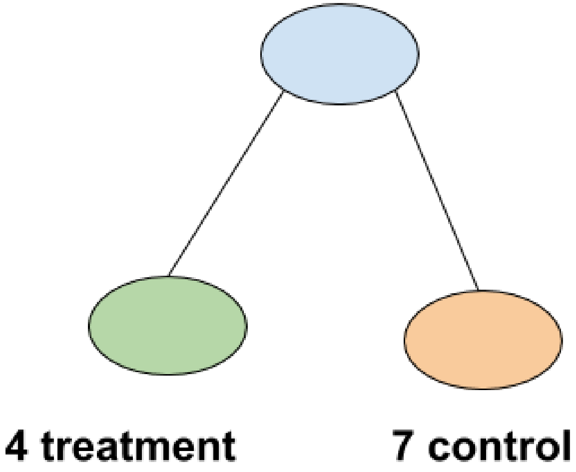
CSF: *Causal Forest*

- Class A: 3 treatment, 2 control
- Class B: 1 treatment, 5 control

Gini impurity $1 - \sum_{i=1}^{J} {p_i}^2$

MSE $\sum_{i \in \mathcal{J}} \left(\hat{\mu}\left(X_i\right) - Y_i\right)^2 = \sum_{i \in \mathcal{J}} Y_i^2 - \sum_{i \in \mathcal{J}} \hat{\mu}\left(X_i\right)^2$





$$\hat{\tau}_{CSF}(\mathbf{x}) = \frac{1}{|L_1(\mathbf{x})|} \sum_{i \in L_1(\mathbf{x})} Y_i - \frac{1}{|L_0(\mathbf{x})|} \sum_{i \in L_0(\mathbf{x})} Y_i$$

Causal Perception Classifier

$$\begin{pmatrix} (w_{11}, w_{12}, s_1) \\ (w_{21}, w_{22}, s_2) \\ \cdots\cdots \\ (w_{n1}, w_{n2}, s_n) \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \cdots \\ \tau_n \end{pmatrix} \xrightarrow{\text{Training}} \textbf{Classifier} \xrightarrow{\text{Predicting}} \begin{pmatrix} (w'_{11}, w'_{12}, s'_1) \\ (w'_{21}, w'_{22}, s'_2) \\ \cdots\cdots \\ (w'_{m1}, w'_{m2}, s'_m) \end{pmatrix}$$

- Generalizable Features:
  - Context probability
  - Control word probability
  - Treatment word probability

Datasets

Dataset	Substitutable word pairs	female/undesirable sentences	male/desirable sentences
Twitter	1,876	583,982	441,562
Yelp	1,648	582,792	492,893
Airbnb	1,678	49,866	224,603

Substitutable Word Pairs

- Representative words
  - Moderately correlated
  - LogisticRegression coefficient
- Semantic substitutability
  - Paraphrase Database (PPDB 2.0)
- Syntactic substitutability
  - Part-of-speech tag
- Substitutability for a specific sentence $(\boldsymbol{w_i}, \boldsymbol{w_j}, \boldsymbol{s})$
  - N-grams

Substitutable Word Pairs

Increase desirability	Increase male perception
store $\rightarrow$ boutique famous $\rightarrow$ grand famous $\rightarrow$ renowned rapidly $\rightarrow$ quickly nice $\rightarrow$ gorgeous amazing $\rightarrow$ incredible events $\rightarrow$ festivals cheap $\rightarrow$ inexpensive various $\rightarrow$ several yummy $\rightarrow$ delicious	gay $\rightarrow$ homo yummy $\rightarrow$ tasty happiness $\rightarrow$ joy fabulous $\rightarrow$ impressive bed $\rightarrow$ crib amazing $\rightarrow$ impressive boyfriends $\rightarrow$ buddies purse $\rightarrow$ wallet precious $\rightarrow$ valuable sweetheart $\rightarrow$ girlfriend

Table 1: Samples of substitution words with high LSE

Human-derived LSE estimates (AMT)

	Airbnb	**Twitter / Yelp**
Q&A	Rate the desirability of a short-term apartment rental based on a single sentence.	Rate how likely you think this tweet / Yelp review sentence is written by male or female.
5	Very desirable	Very likely male
4	Somewhat desirable	Somewhat likely male
3	Neither desirable nor undesirable	Neutral, neither male nor female
2	Somewhat undesirable	Somewhat likely female
1	Very undesirable	Very likely female

Table 5: Amazon Mechanical Turk annotation guidelines

- *"There are plenty of <u>shops</u> nearby"* → *"There are plenty of <u>boutiques</u> nearby"*

Comparisons

	Yelp	**Twitter**	**Airbnb**
Agreements-pearson	0.557	0.576	0.513
KNN	0.474	0.291	0.076
VT-RF	0.747	0.333	0.049
CF-RF	0.680	0.279	0.109
CSF	0.645	**0.338**	0.096
Causal perception classifier	**0.783**	0.21	**0.139**

Table 2: Inter-annotator agreement and Pearson correlation between algorithmically estimated LSE and AMT judgment

Comparisons

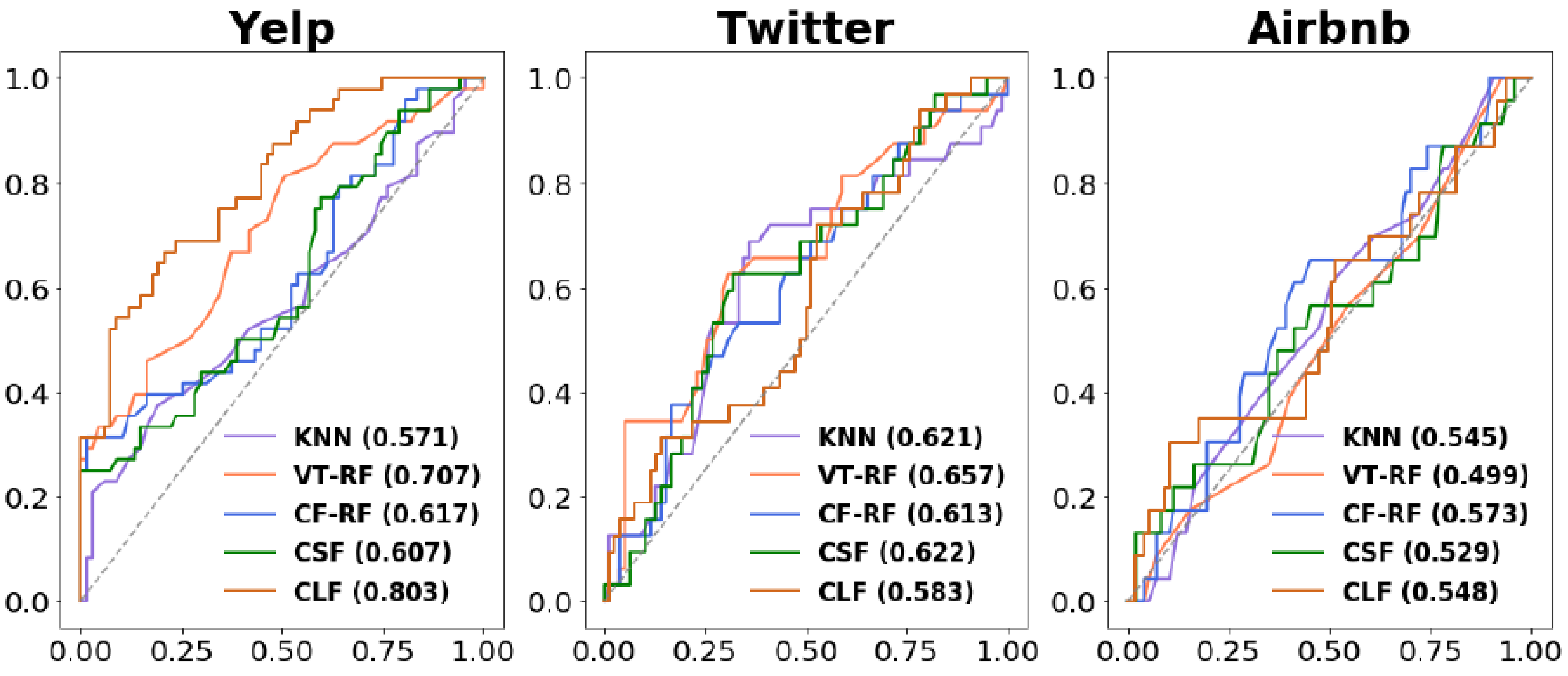


Figure 1: ROC curve for classifying sentences according to AMT perception with estimated LSE as confidence score

Causal perception classifier

	Yelp	Twitter	Airbnb
context pr	-0.348	-0.829	-0.528
control word pr	-0.141	-0.514	-0.367
treatment word pr	0.189	0.401	0.344

Table 3: Logistic regression coefficients for the features of the causal perception classifier

Conclusion example

- Monday nights are a night of bonding for me and my *boyfriend*
- Monday nights are a night of bonding for me and my *buddy*

- If you ask me to hang out with you and your *boyfriend*, I will ... decline.

Limitation and Future Work

- Crime rate as desirability
- Single word substitution --> multiple word substitutions
- Perception based on one sentence --> Perception based on documents

Thank You!