D INFK

Detecting when the available data does not allow reliable inference

Fanny Yang, Statistical Machine Learning Group
Department of Computer Science @ETH Zurich

joint work with students Alex Tifrea, Eric Stavarache,
Piersilvio de Bartolomeis, Javier A. Martinez, Konstantin Donhauser

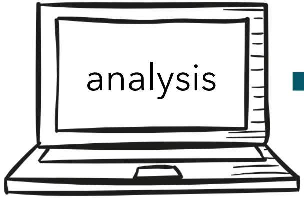
ETH zürich

Problem of validity of inference

Problem of validity of inference

Problem of validity of inference

observational study



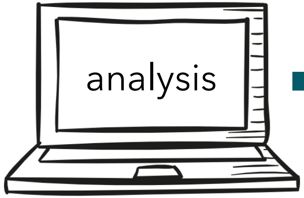
higher BMI lowers mortality

Problem 1

too much hidden confounding

Problem of validity of inference

observational study



higher BMI lowers mortality

Problem 1

too much hidden confounding

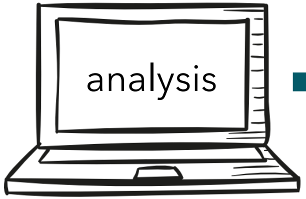

(d) COVID-19 Pneumonia

individual test sample
with new disease/class

bacterial pneumonia
(seen during training)

Problem of validity of inference

observational study



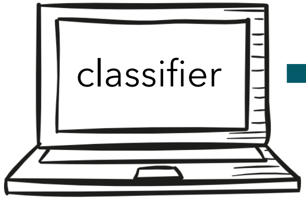
higher BMI lowers mortality

Problem 1

too much hidden confounding

(d) COVID-19 Pneumonia

individual test sample
with new disease/class

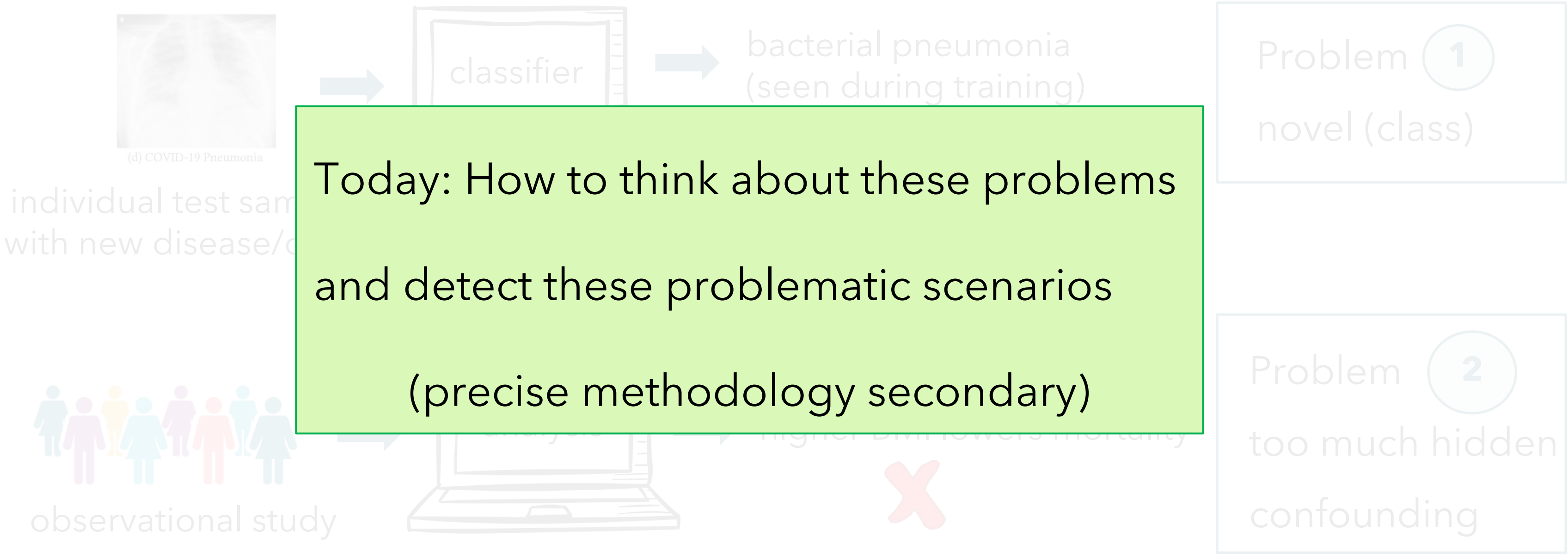


bacterial pneumonia
(seen during training)

Problem 2

novel (class)

Problem of validity of inference



Today: How to think about these problems

and detect these problematic scenarios

(precise methodology secondary)

DINFK

I. A lower bound for hidden confounding using randomized control trials

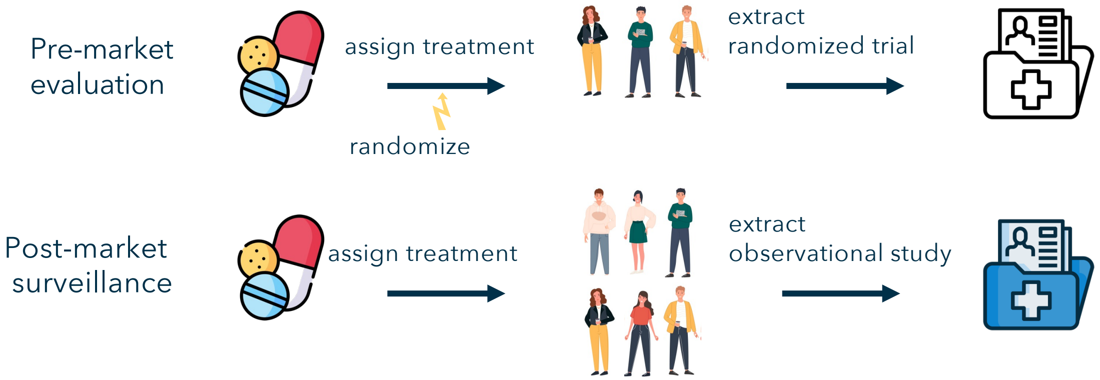
joint work with Piersilvio de Bartolomeis, Javier Abad Martinez, Konstantin Donhauser

work in progress

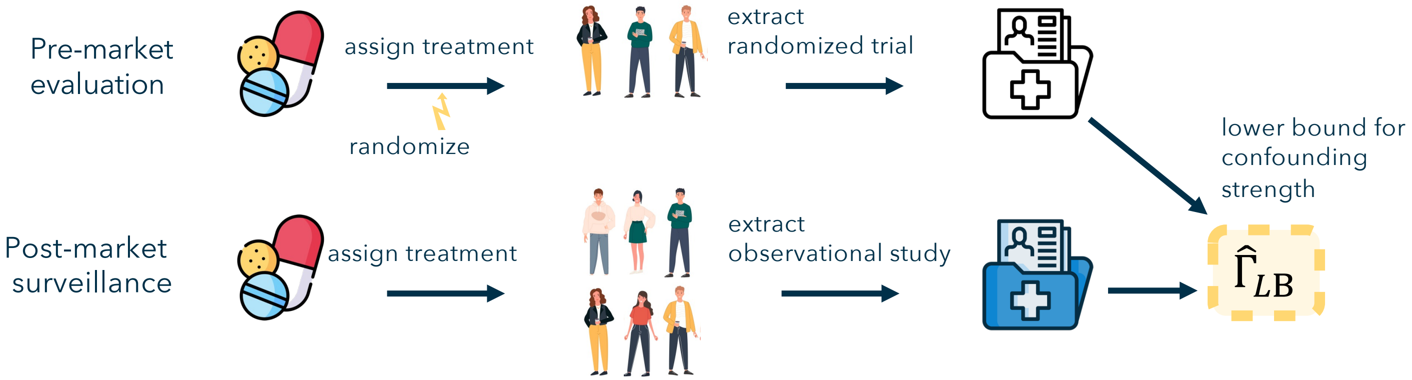


ETH zürich

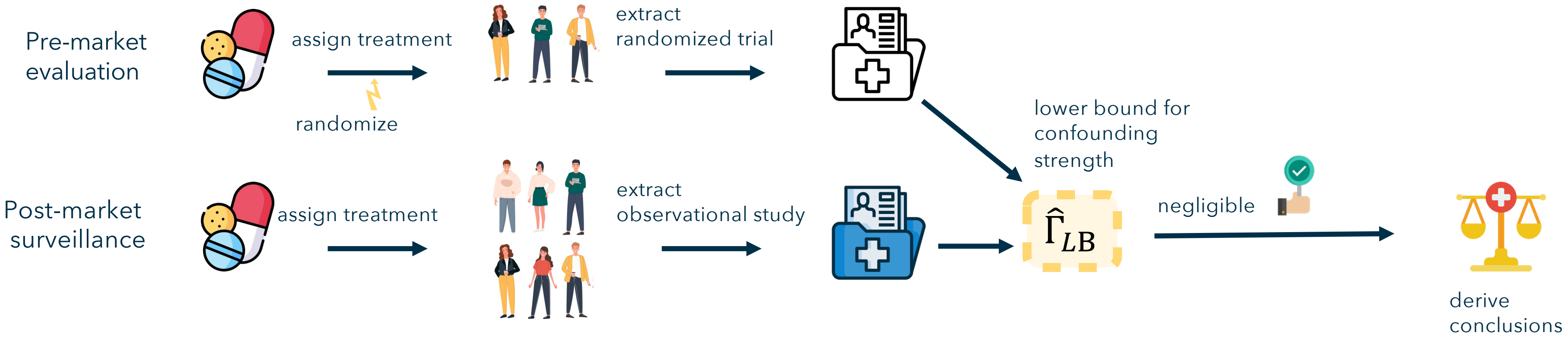
Our goal: lower bounding confounding strength



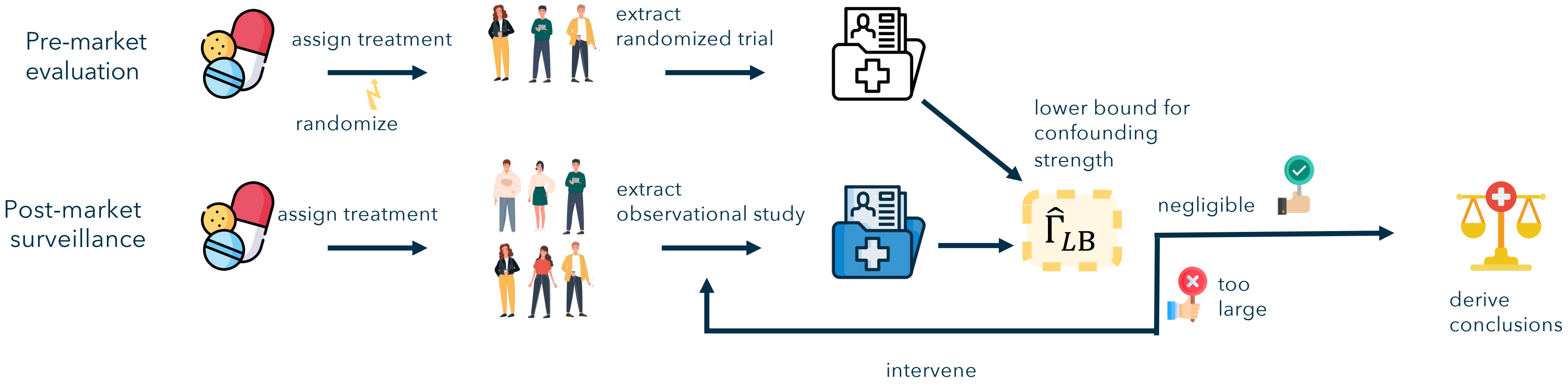
Our goal: lower bounding confounding strength



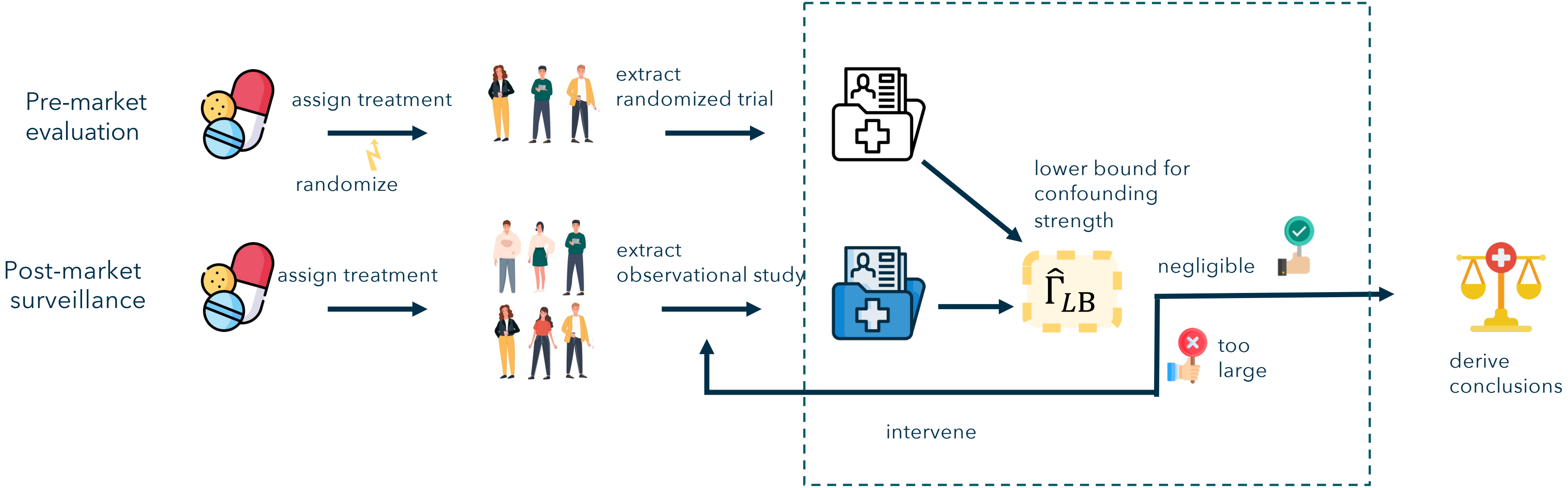
Our goal: lower bounding confounding strength



Our goal: lower bounding confounding strength



Our goal: lower bounding confounding strength



1. Scenarios where we can’t make inference using the observational study

2. Approach: How to detect these scenarios?

grey: observed variables,
white: latent variables

Potential outcome framework

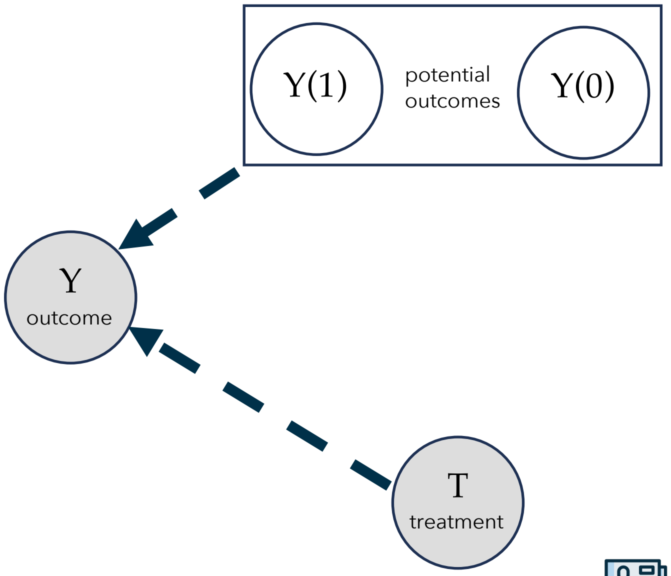
Observed samples (X_i, Y_i, T_i) i.i.d. from the following distribution with $Y = Y(1)T + Y(0)(1-T)$ (SUTVA)

Observational study $\mathbb{P}^{os}$ 

grey: observed variables,
white: latent variables

Potential outcome framework

Observed samples (X_i, Y_i, T_i) i.i.d. from the following distribution with $Y = Y(1)T + Y(0)(1-T)$ (SUTVA)

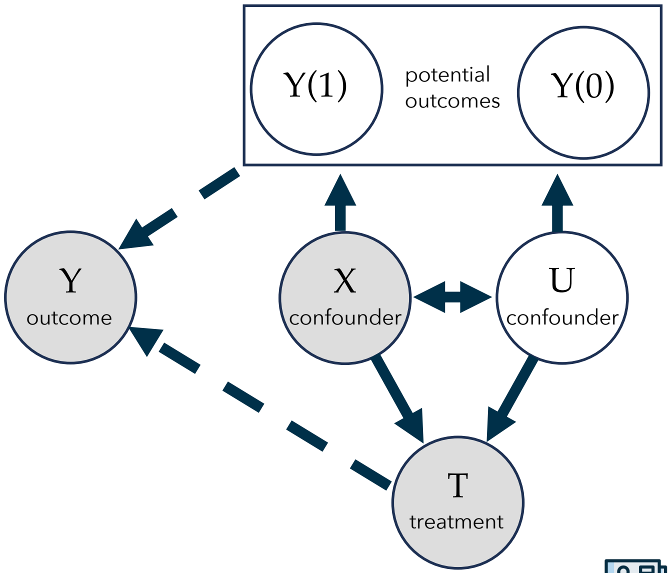


Observational study $\mathbb{P}^{os}$

grey: observed variables,
white: latent variables

Potential outcome framework

Observed samples (X_i, Y_i, T_i) i.i.d. from the following distribution with $Y = Y(1)T + Y(0)(1 - T)$ (SUTVA)

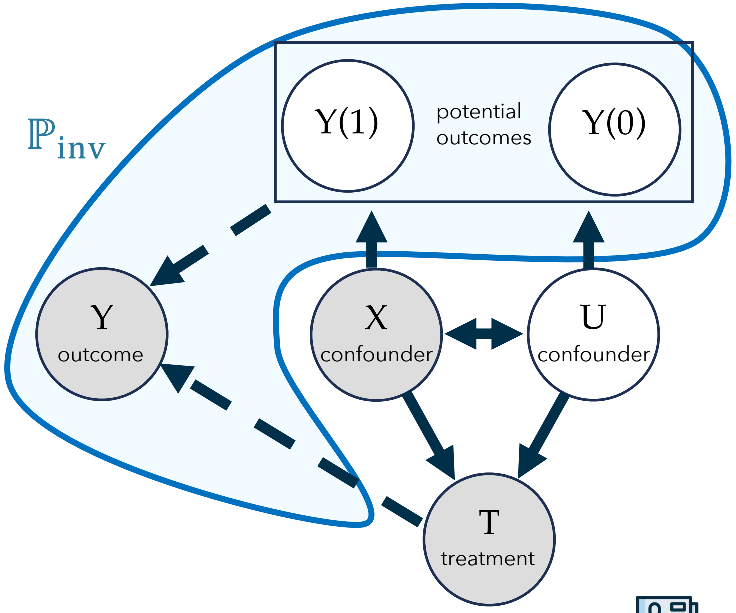


Observational study $\mathbb{P}^{os}$

grey: observed variables,
white: latent variables

Potential outcome framework

Observed samples (X_i, Y_i, T_i) i.i.d. from the following distribution with $Y = Y(1)T + Y(0)(1-T)$ (SUTVA)

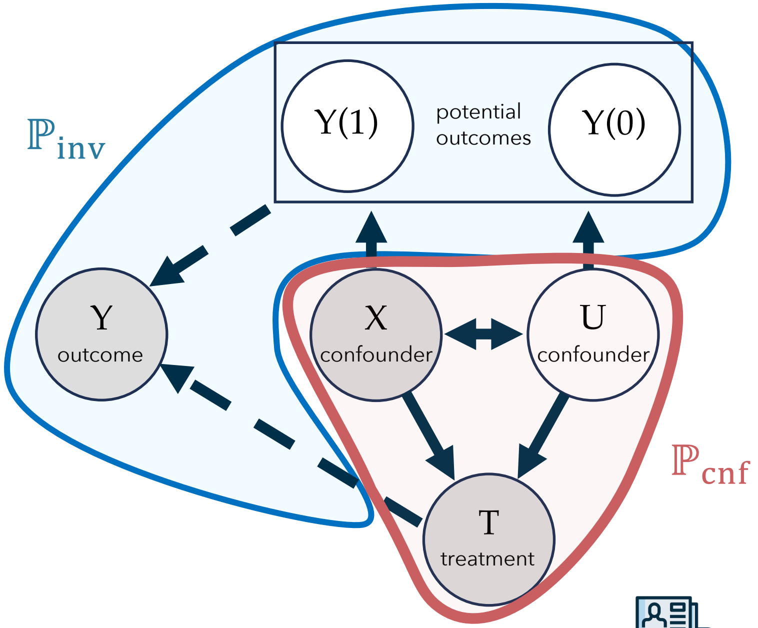


Observational study $\mathbb{P}^{os}$

grey: observed variables,
white: latent variables

Potential outcome framework

Observed samples (X_i, Y_i, T_i) i.i.d. from the following distribution with $Y = Y(1)T + Y(0)(1-T)$ (SUTVA)

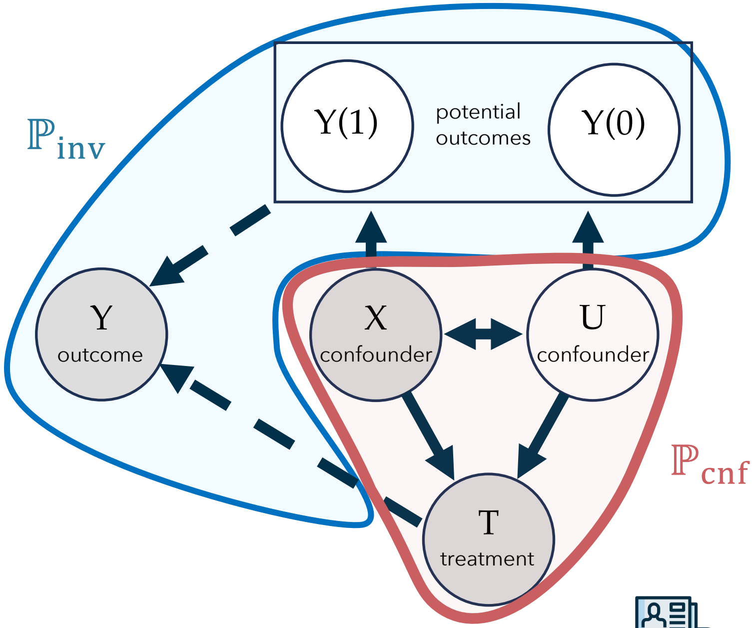


Observational study $\mathbb{P}^{os}$

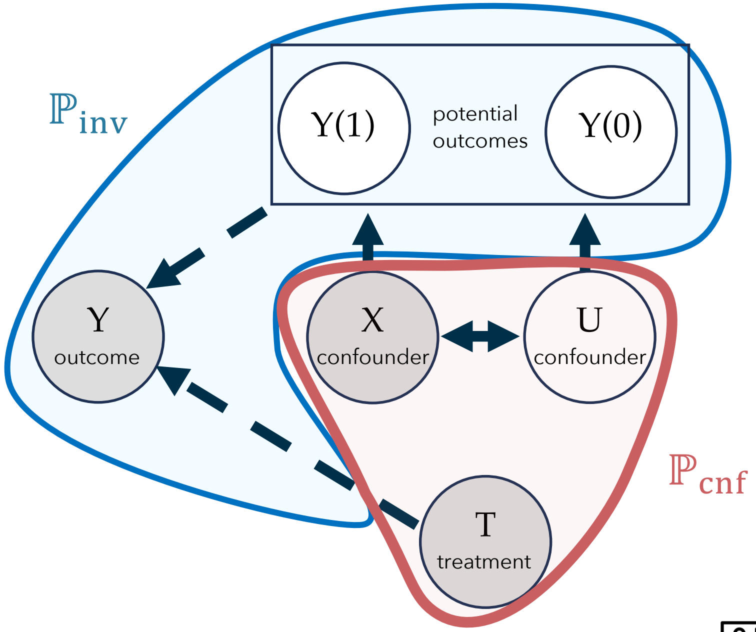
grey: observed variables,
white: latent variables

Potential outcome framework

Observed samples (X_i, Y_i, T_i) i.i.d. from the following distribution with $Y = Y(1)T + Y(0)(1-T)$ (SUTVA)



Observational study $\mathbb{P}^{os}$



Randomized control trial $\mathbb{P}^{rct}$

Marginal sensitivity model

Additional assumptions:

- Transportability of CATE, i.e. $\mathbb{E}_{\mathbb{P}^{os}}[Y(1) - Y(0) \mid X] = \mathbb{E}_{\mathbb{P}^{rct}}[Y(1) - Y(0) \mid X]$
- Support inclusion $supp(\mathbb{P}^{rct}) \subseteq supp(\mathbb{P}^{os})$

Marginal sensitivity model

Additional assumptions:

- Transportability of CATE, i.e. $\mathbb{E}_{\mathbb{P}^{os}}[Y(1) - Y(0) \mid X] = \mathbb{E}_{\mathbb{P}^{rct}}[Y(1) - Y(0) \mid X]$
- Support inclusion $supp(\mathbb{P}^{rct}) \subseteq supp(\mathbb{P}^{os})$

Definitions:

- $\mathbb{P}^{os}$ satisfies MSM(Γ) if $\Gamma^{-1} \leq \frac{\mathbb{P}^{os}(T=1|X,U)}{\mathbb{P}^{os}(T=0|X,U)} \Big/ \frac{\mathbb{P}^{os}(T=1|X)}{\mathbb{P}^{os}(T=0|X)} \leq \Gamma$ almost surely (Tan-06)

Marginal sensitivity model

Additional assumptions:

- Transportability of CATE, i.e. $\mathbb{E}_{\mathbb{P}^{os}}[Y(1) - Y(0) \mid X] = \mathbb{E}_{\mathbb{P}^{rct}}[Y(1) - Y(0) \mid X]$
- Support inclusion $supp(\mathbb{P}^{rct}) \subseteq supp(\mathbb{P}^{os})$

Definitions:

$\Gamma = 1 \triangleq$ unconfoundedness

- $\mathbb{P}^{os}$ satisfies MSM(Γ) if $\Gamma^{-1} \leq \frac{\mathbb{P}^{os}(T=1|X,U)}{\mathbb{P}^{os}(T=0|X,U)} / \frac{\mathbb{P}^{os}(T=1|X)}{\mathbb{P}^{os}(T=0|X)} \leq \Gamma$ almost surely (Tan-06)

Marginal sensitivity model

Additional assumptions:

- Transportability of CATE, i.e. $\mathbb{E}_{\mathbb{P}^{os}}[Y(1) - Y(0) \mid X] = \mathbb{E}_{\mathbb{P}^{rct}}[Y(1) - Y(0) \mid X]$
- Support inclusion $supp(\mathbb{P}^{rct}) \subseteq supp(\mathbb{P}^{os})$

Definitions:

$\Gamma = 1 \triangleq$ unconfoundedness

- $\mathbb{P}^{os}$ satisfies MSM(Γ) if $\Gamma^{-1} \leq \frac{\mathbb{P}^{os}(T=1|X,U)}{\mathbb{P}^{os}(T=0|X,U)} \Big/ \frac{\mathbb{P}^{os}(T=1|X)}{\mathbb{P}^{os}(T=0|X)} \leq \Gamma$ almost surely (Tan-06)
- true confounding strength $\Gamma^\star(\mathbb{P}^{os})$: The smallest Γ for which $\mathbb{P}^{os}$ satisfies $MSM(\Gamma)$

Marginal sensitivity model

Additional assumptions:

- Transportability of CATE, i.e. $\mathbb{E}_{\mathbb{P}^{os}}[Y(1) - Y(0) \mid X] = \mathbb{E}_{\mathbb{P}^{rct}}[Y(1) - Y(0) \mid X]$
- Support inclusion $supp(\mathbb{P}^{rct}) \subseteq supp(\mathbb{P}^{os})$

Definitions:

$\Gamma = 1 \triangleq$ unconfoundedness

- $\mathbb{P}^{os}$ satisfies MSM(Γ) if $\Gamma^{-1} \leq \frac{\mathbb{P}^{os}(T=1|X,U)}{\mathbb{P}^{os}(T=0|X,U)} / \frac{\mathbb{P}^{os}(T=1|X)}{\mathbb{P}^{os}(T=0|X)} \leq \Gamma$ almost surely (Tan-06)
- true confounding strength $\Gamma^\star(\mathbb{P}^{os})$: The smallest Γ for which $\mathbb{P}^{os}$ satisfies $MSM(\Gamma)$

Scenarios we want to detect: when true confounding $\Gamma^\star$ of $\mathbb{P}^{os}$ is too large

1. Scenarios where we can't make inference using the observational study
2. Approach: How to detect these scenarios?

Our paradigm: finding a lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma)$ $\iff \Gamma^\star \leq \Gamma$

Our paradigm: finding a lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma) \iff \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\{\Gamma\colon \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

Our paradigm: finding a lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma) \iff \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf \{\Gamma : \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

- Test $H_0(\Gamma) \iff$ test whether $\mu \in [\mu_\Gamma^-, \mu_\Gamma^+]$ with ATE $\mu = \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0)]$ and

ATE sensitivity bounds $\mu_\Gamma^- = \inf_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$, $\mu_\Gamma^+ = \sup_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$

Our paradigm: finding a lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma)$ $\iff \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\{\Gamma:\ \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

- Test $H_0(\Gamma)$ $\iff$ test whether $\mu \in [\mu_\Gamma^-, \mu_\Gamma^+]$ with ATE $\mu = \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0)]$ and

ATE sensitivity bounds $\mu_\Gamma^- = \inf_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)],\ \ \mu_\Gamma^+ = \sup_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$

all full distributions that yield observed $\mathbb{P}^{os}_{X,Y,T}$ and satisfy MSM(Γ)

Our paradigm: finding a lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma) \iff \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\{\Gamma:\ \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

- Test $H_0(\Gamma) \iff$ test whether $\mu \in [\mu_\Gamma^-, \mu_\Gamma^+]$ with ATE $\mu = \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0)]$ and

ATE sensitivity bounds $\mu_\Gamma^- = \inf_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$, $\mu_\Gamma^+ = \sup_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$

Test using consistent estimates $\hat{\mu}, \hat{\mu}_\Gamma^+, \hat{\mu}_\Gamma^-$ in literature (experts in audience)

Our paradigm: finding a lower bound

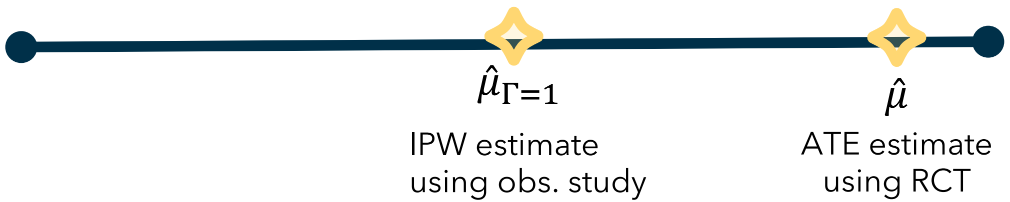
Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma) \iff \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\{\Gamma\colon \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

- Test $H_0(\Gamma) \iff$ test whether $\mu \in [\mu_\Gamma^-, \mu_\Gamma^+]$ with ATE $\mu = \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0)]$ and

ATE sensitivity bounds $\mu_\Gamma^- = \inf_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$, $\mu_\Gamma^+ = \sup_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$

Test using consistent estimates $\hat{\mu}, \hat{\mu}_\Gamma^+, \hat{\mu}_\Gamma^-$ in literature (experts in audience)



Our paradigm: finding a lower bound

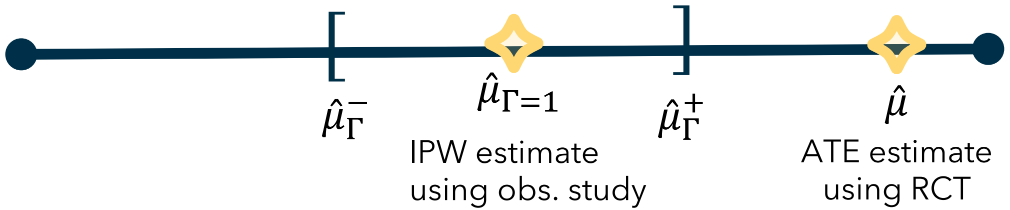
Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma)$ $\Leftrightarrow \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\ \{\Gamma\colon\ \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

- Test $H_0(\Gamma)$ $\Leftrightarrow$ test whether $\mu \in [\mu_\Gamma^-, \mu_\Gamma^+]$ with ATE $\mu = \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0)]$ and

ATE sensitivity bounds $\mu_\Gamma^- = \inf_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$, $\mu_\Gamma^+ = \sup_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$

Test using consistent estimates $\hat{\mu}, \hat{\mu}_\Gamma^+, \hat{\mu}_\Gamma^-$ in literature (experts in audience)



Our paradigm: finding a lower bound

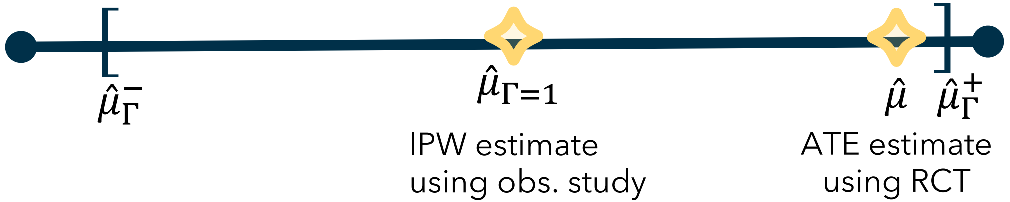
Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma) \iff \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\{\Gamma: \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

- Test $H_0(\Gamma) \iff$ test whether $\mu \in [\mu_\Gamma^-, \mu_\Gamma^+]$ with ATE $\mu = \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0)]$ and

ATE sensitivity bounds $\mu_\Gamma^- = \inf_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$, $\mu_\Gamma^+ = \sup_{\tilde{\mathbb{P}} \in P_\Gamma(\mathbb{P}^{os}_{X,Y,T})} \mathbb{E}_{\tilde{\mathbb{P}}}[Y(1) - Y(0)]$

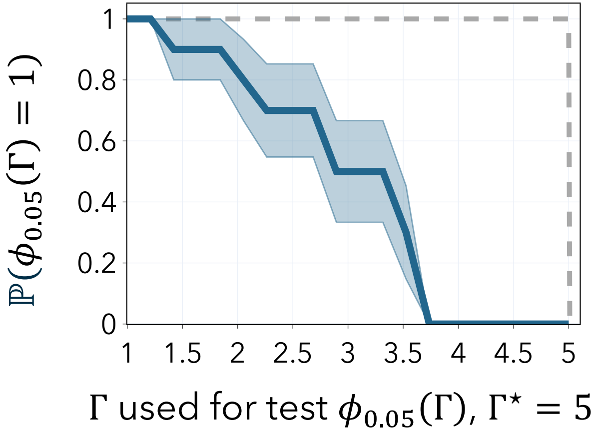
Test using consistent estimates $\hat{\mu}, \hat{\mu}_\Gamma^+, \hat{\mu}_\Gamma^-$ in literature (experts in audience)



Our paradigm: finding a lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma) \iff \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\{\Gamma\colon\ \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

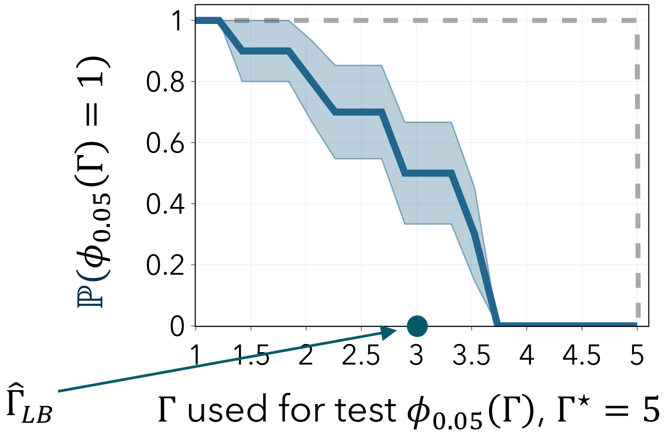


- probability of rejection over 20 runs on semi-synthetic data

Our paradigm: finding a lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma) \iff \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\{\Gamma\colon\ \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$

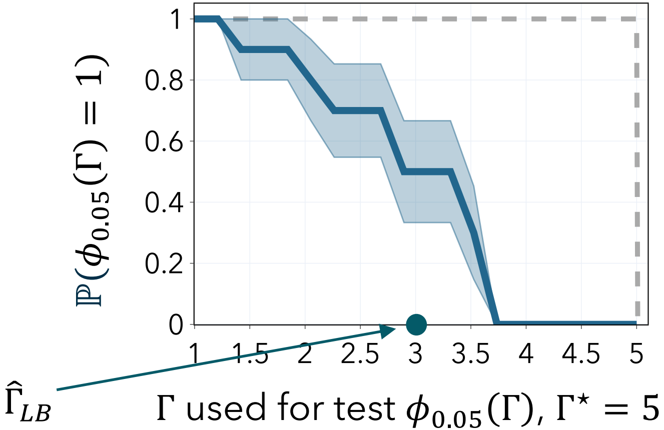


- probability of rejection over 20 runs on semi-synthetic data

Our paradigm: finding a lower bound

Our plug-and-play approach for desired significance α:

1. Test $\phi_\alpha(\Gamma)$ of the null $H_0(\Gamma)$: $\mathbb{P}^{os}$ satisfies $MSM(\Gamma)$ $\Leftrightarrow \Gamma^\star \leq \Gamma$
2. Report $\hat{\Gamma}_{LB} = \inf\{\Gamma\colon \phi_\alpha(\Gamma) = 0\}$ and flag if $\hat{\Gamma}_{LB} > \Gamma_{\text{thresh}}$



- probability of rejection over 20 runs on semi-synthetic data
- asymptotically $\mathbb{P}(\hat{\Gamma}_{LB} > \Gamma^\star) \leq \alpha$ implied by $\mathbb{P}(\phi_\alpha(\Gamma^\star) = 1) \leq \alpha$

Previous paradigms that could be used to "flag"

- without RCT and using sensitivity bounds
  - quantification via critical value $\hat{\Gamma}_{ct}$ that changes causal conclusions

    e.g. vanderWeele-Ding-17, Jin-Ren-Candes-23 etc.

Previous paradigms that could be used to "flag"

- without RCT and using sensitivity bounds
  - quantification via critical value $\hat{\Gamma}_{ct}$ that changes causal conclusions

    e.g. vanderWeele-Ding-17, Jin-Ren-Candes-23 etc.

    but: can be arbitrarily far from $\Gamma^{\star}$

Previous paradigms that could be used to "flag"

- without RCT and using sensitivity bounds
  - quantification via critical value $\hat{\Gamma}_{ct}$ that changes causal conclusions

    e.g. vanderWeele-Ding-17, Jin-Ren-Candes-23 etc.

    but: can be arbitrarily far from $\Gamma^\star$
  - can test joint null hypothesis ATE(obs. study) > 0 *and* MSM(Γ) holds

    e.g. Yadlowsky-Namkoong-Basu-Duchi-Tian-22, Jin-Ren-Candes-23

Previous paradigms that could be used to "flag"

- without RCT and using sensitivity bounds
  - quantification via critical value $\hat{\Gamma}_{ct}$ that changes causal conclusions

    e.g. vanderWeele-Ding-17, Jin-Ren-Candes-23 etc.

    but: can be arbitrarily far from $\Gamma^\star$
  - can test joint null hypothesis ATE(obs. study) > 0 *and* MSM(Γ) holds

    e.g. Yadlowsky-Namkoong-Basu-Duchi-Tian-22, Jin-Ren-Candes-23

    but: rejection only means either MSM(Γ) assumption wrong *or* ATE ≤ 0

Previous paradigms that could be used to "flag"

- without RCT and using sensitivity bounds
  - quantification via critical value $\hat{\Gamma}_{ct}$ that changes causal conclusions

    e.g. vanderWeele-Ding-17, Jin-Ren-Candes-23 etc.

    but: can be arbitrarily far from $\Gamma^\star$
  - can test joint null hypothesis ATE(obs. study) > 0 *and* MSM(Γ) holds

    e.g. Yadlowsky-Namkoong-Basu-Duchi-Tian-22, Jin-Ren-Candes-23

    but: rejection only means either MSM(Γ) assumption wrong *or* ATE ≤ 0
- with RCT:
  - binary test for existence of confounding with $\mathrm{H}_0\colon\ \Gamma^\star > 1$

    e.g. Viele et al '14, Hussein-Oberst-Shih-Sontag '22

Previous paradigms that can be used for detection

- without RCT and using sensitivity bounds
  - quantification via critical gamma value $\hat{\Gamma}_{ct}$ that changes causal conclusions

    e.g. vander des-23 etc.

    but: can $\Gamma^\star$
  - can test TE(obs. study) > 0 *and*

    e.g. Yadlowsky-Namkoong-Basu-Duchi-Tian-22, Ji es-23

    but: rejection only means either MSM(Γ) assu mption

→ true statement about $\Gamma^\star$ not possible!

our paradigm:
statement about $\Gamma^\star$
& flag only if $\Gamma^\star$ large

- with RCT:
  - binary te founding with H_0: $\Gamma^\star > 1$

    e.g. Viele e Sontag '22

→ flag even if $\Gamma^\star$ small

Evaluation on real-world data (WHI)

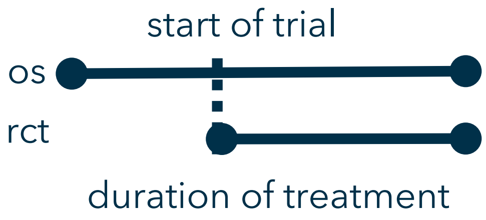
- Randomized trial and observational study (1993-2005)
- Treatment: hormone replacement therapy
- Outcomes: coronary heart disease

Evaluation on real-world data (WHI)

- Randomized trial and observational study (1993-2005)
- Treatment: hormone replacement therapy
- Outcomes: coronary heart disease
- hidden confounder (revealed later): start of treatment

Evaluation on real-world data (WHI)

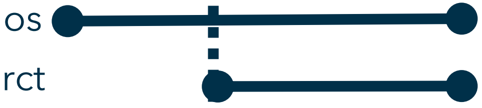
- Randomized trial and observational study (1993-2005)
- Treatment: hormone replacement therapy
- Outcomes: coronary heart disease
- hidden confounder (revealed later): start of treatment



	Coronary heart disease
treated	as trial started
$\hat{\Gamma}_{CT}$	1.017
$\hat{\Gamma}_{LB}$	1.009
ψ_{bin}	1
ψ_{sens}	0

Evaluation on real-world data (WHI)

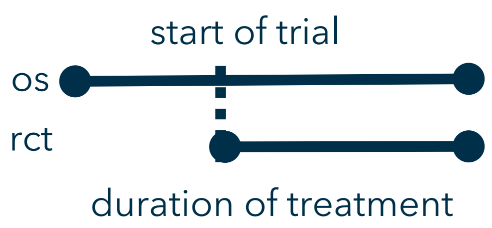
- Randomized trial and observational study (1993-2005)
- Treatment: hormone replacement therapy
- Outcomes: coronary heart disease
- hidden confounder (revealed later): start of treatment



	Coronary heart disease	
treated	as trial started	before trial
$\hat{\Gamma}_{CT}$	1.017	1.164
$\hat{\Gamma}_{LB}$	1.009	1.224
ψ_{bin}	1	1
ψ_{sens}	0	1

Evaluation on real-world data (WHI)

- Randomized trial and observational study (1993-2005)
- Treatment: hormone replacement therapy
- Outcomes: coronary heart disease
- hidden confounder (revealed later): start of treatment



	Coronary heart disease	
treated	as trial started	before trial
$\hat{\Gamma}_{CT}$	1.017	1.164
$\hat{\Gamma}_{LB}$	1.009	1.224
ψ_{bin}	1	1
ψ_{sens}	0	1

Different paradigms for "flagging" confounding:

- Compute $\hat{\Gamma}_{CT}$ that changes ATE sign and compare let "expert" assess "likeliness"

Evaluation on real-world data (WHI)

- Randomized trial and observational study (1993-2005)
- Treatment: hormone replacement therapy
- Outcomes: coronary heart disease
- hidden confounder (revealed later): start of treatment

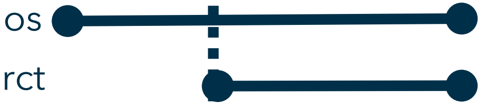
	Coronary heart disease	
treated	as trial started	before trial
$\hat{\Gamma}_{CT}$	1.017	1.164
$\hat{\Gamma}_{LB}$	1.009	1.224
ψ_{bin}	1	1
ψ_{sens}	0	1

Different paradigms for "flagging" confounding:

- Compute $\hat{\Gamma}_{CT}$ that changes ATE sign and compare let "expert" assess "likeliness"
- ψ_{bin}: tests for existence, e.g. check $\hat{\Gamma}_{LB} > 1$

Evaluation on real-world data (WHI)

- Randomized trial and observational study (1993-2005)
- Treatment: hormone replacement therapy
- Outcomes: coronary heart disease
- hidden confounder (revealed later): start of treatment



	Coronary heart disease	
treated	as trial started	before trial
$\hat{\Gamma}_{CT}$	1.017	1.164
$\hat{\Gamma}_{LB}$	1.009	1.224
ψ_{bin}	1	1
ψ_{sens}	0	1

Different paradigms for "flagging" confounding:

- Compute $\hat{\Gamma}_{CT}$ that changes ATE sign and compare let "expert" assess "likeliness"
- ψ_{bin}: tests for existence, e.g. check $\hat{\Gamma}_{LB} > 1$
- ψ_{sens} (ours): check whether too large $\hat{\Gamma}_{LB} > \hat{\Gamma}_{CT}$

Current and future work

Higher power using

- kernelized test as opposed to averaging
- non-"adversarial" sensitivity model

Extended applicability:

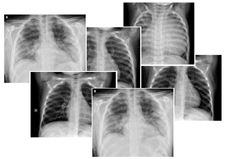
- multiple observational studies (no RCT)
- Automatic detection of hidden confounders from set of features

II. Semi-supervised novelty detection using ensembles with regularized disagreement

joint work with Alexandru Tifrea, Eric Stavarache

published at UAI '22

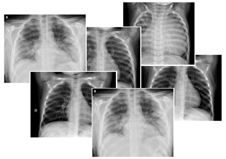

The novelty detection problem for classification



Novelty detection method

Unlabeled Test input

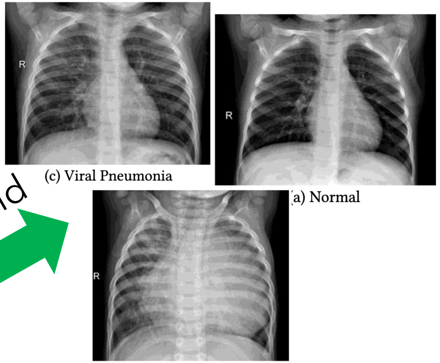
The novelty detection problem for classification



Unlabeled
Test input

Novelty
detection
method

Seen/old



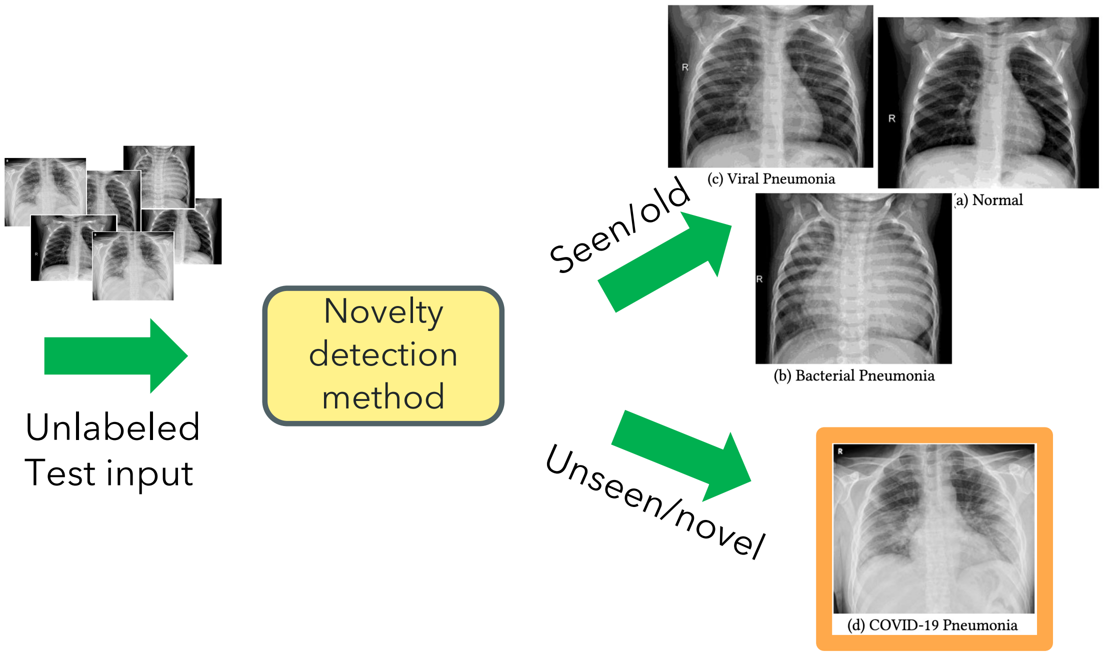
(c) Viral Pneumonia

(a) Normal

(b) Bacterial Pneumonia

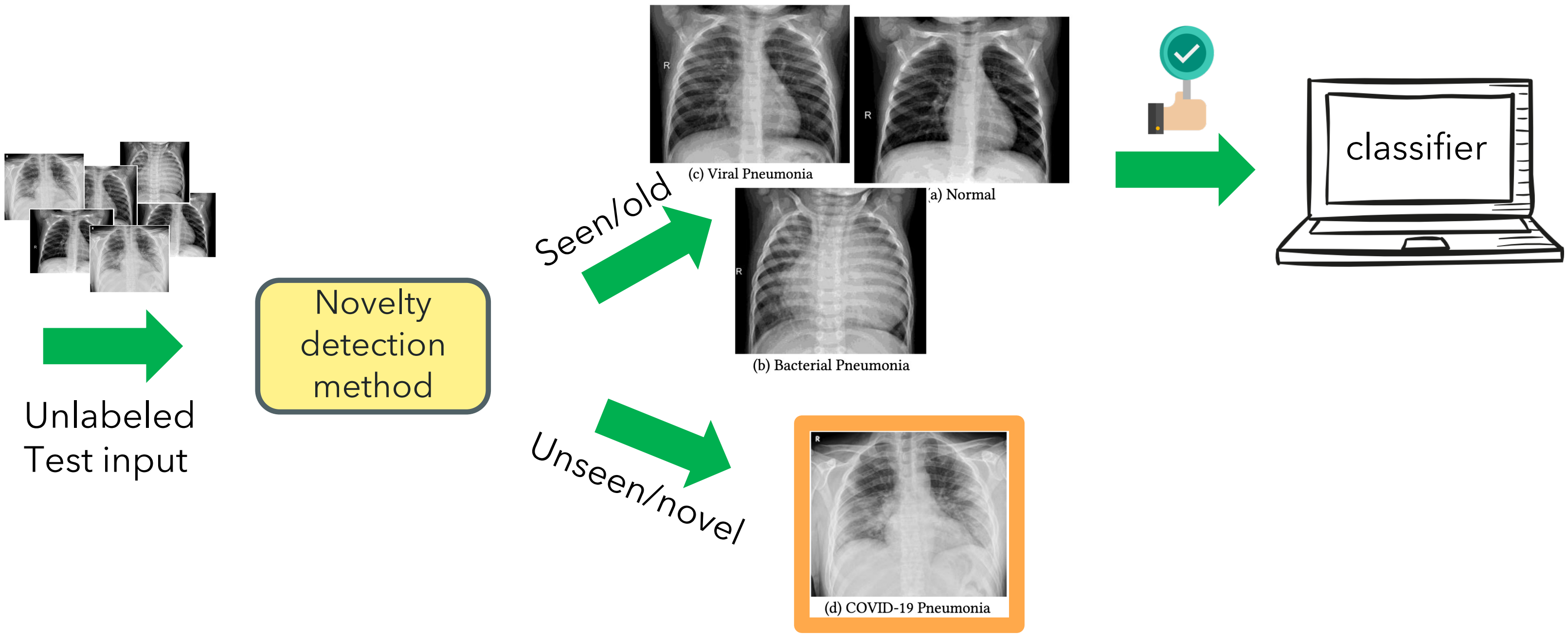
The novelty detection problem for classification

Novelty detection method tells user that software doesn't "know enough" to predict new point



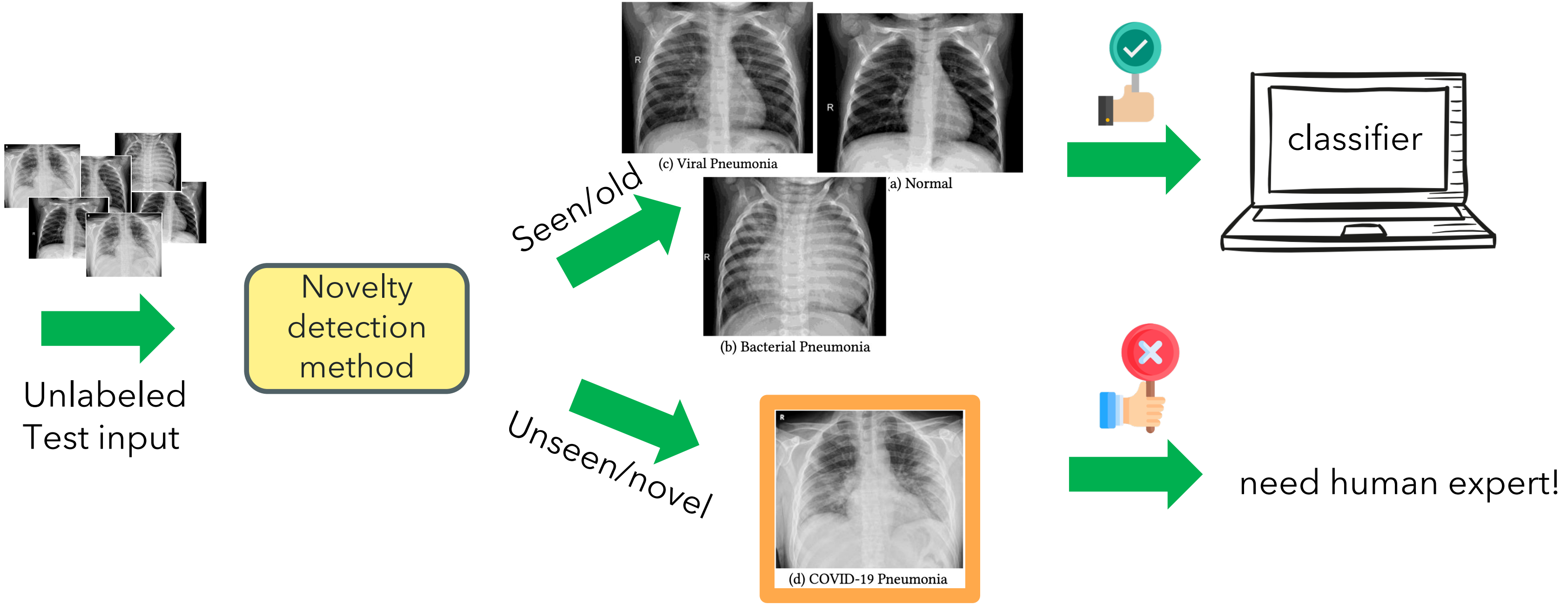
The novelty detection problem for classification

Novelty detection method tells user that software doesn't "know enough" to predict new point



The novelty detection problem for classification

Novelty detection method tells user that software doesn't "know enough" to predict new point



1. Definition: Points we can’t make inference on
2. Approach: How to detect those samples?

What's "novel" to a trained model?

"novel" / o.o.d. points: test points $x \in X$ the model cannot reliably predict.

What's "novel" to a trained model?

"novel" / o.o.d. points: test points $x \in X$ the model cannot reliably predict.

First: which points $x \in X$ can a model predict "reliably" in an unseen test set?

- i.d. generalization from finite samples (traditional learning theory) and

What's "novel" to a trained model?

"novel" / o.o.d. points: test points $x \in X$ the model cannot reliably predict.

First: which points $x \in X$ can a model predict "reliably" in an unseen test set?

- i.d. generalization from finite samples (traditional learning theory) and
- o.o.d. generalization (extrapolatable from training distribution) -

  depends on *test shift & model complexity*

Illustration: Extrapolatable vs. novel samples

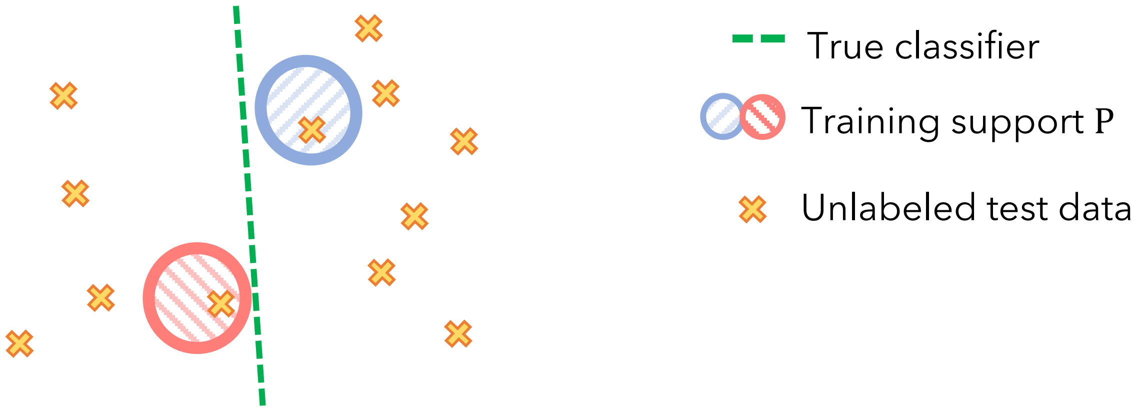


Illustration: Extrapolatable vs. novel samples

Extrapolatable given training distribution + linear ground truth:

Points $x \in X$ where the set of all linear Bayes optimal classifiers agree on

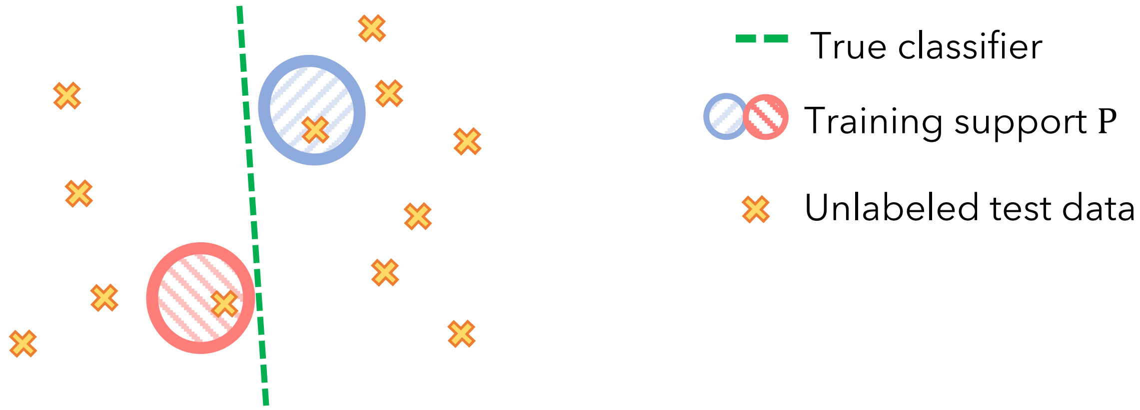
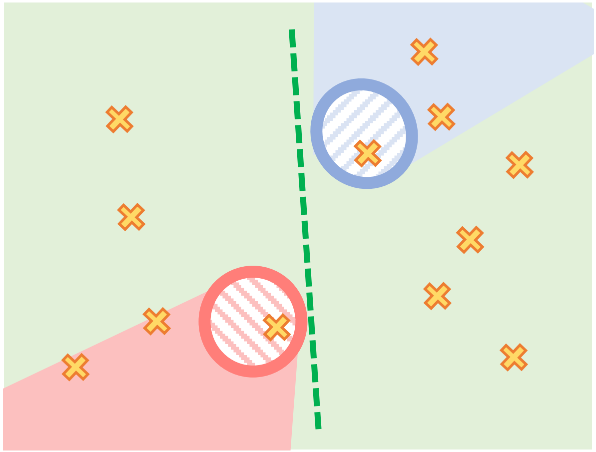


Illustration: Extrapolatable vs. novel samples

Extrapolatable given training distribution + linear ground truth:

Points $x \in X$ where the set of all linear Bayes optimal classifiers agree on

intersecting all optimal classifiers yields



True classifier

Training support P

Unlabeled test data

Correctly extrapolatable

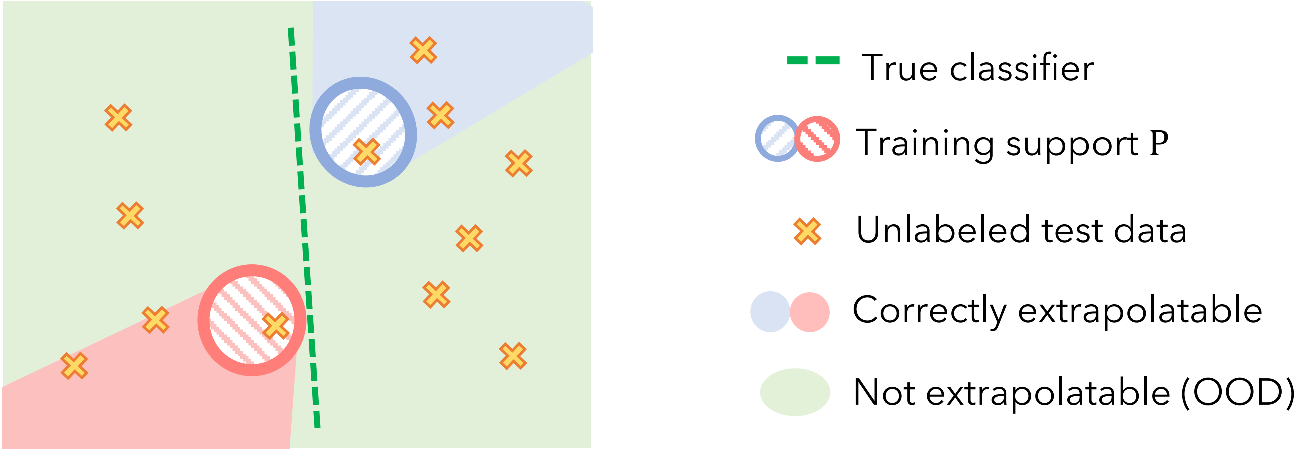
Not extrapolatable (OOD)

Illustration: Extrapolatable vs. novel samples

Extrapolatable given training distribution + linear ground truth:

Points $x \in X$ where the set of all linear Bayes optimal classifiers agree on

intersecting all optimal classifiers yields



Goal now: how to output green area

1. Definition: Points we can’t make inference on
2. Approach: How to detect those samples?

Semi-supervised novelty detection using ensembles

OOD definition suggests following procedure: given K models

- with good validation accuracy on old classes
- but different predictions outside of training distribution

Semi-supervised novelty detection using ensembles

OOD definition suggests following procedure: given K models

- with good validation accuracy on old classes
- but different predictions outside of training distribution

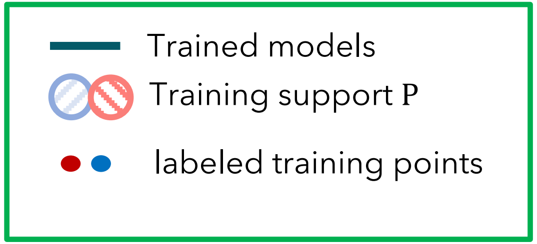
→ flag all points where the models disagree as "novel"

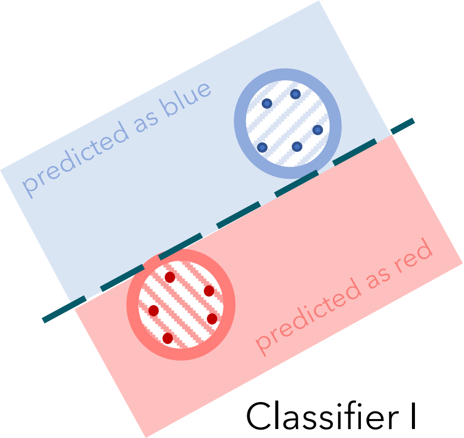
Semi-supervised novelty detection using ensembles

OOD definition suggests following procedure: given K models

- with good validation accuracy on old classes
- but different predictions outside of training distribution

→ flag all points where the models disagree as “novel”





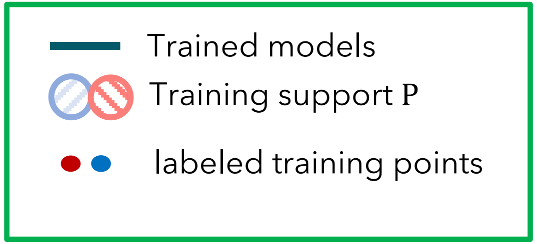
Classifier I

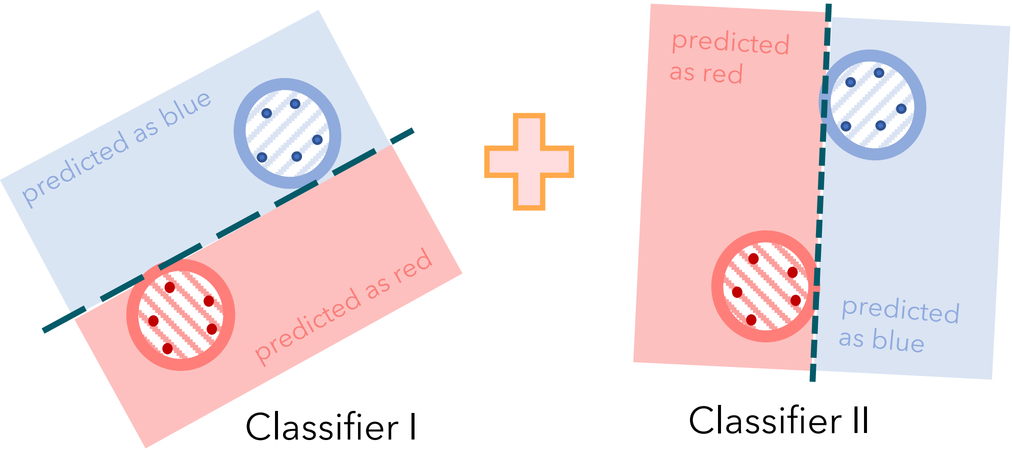
Semi-supervised novelty detection using ensembles

OOD definition suggests following procedure: given K models

- with good validation accuracy on old classes
- but different predictions outside of training distribution

→ flag all points where the models disagree as "novel"



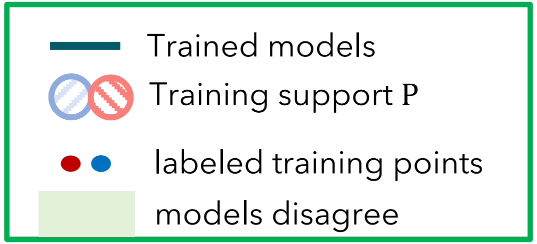


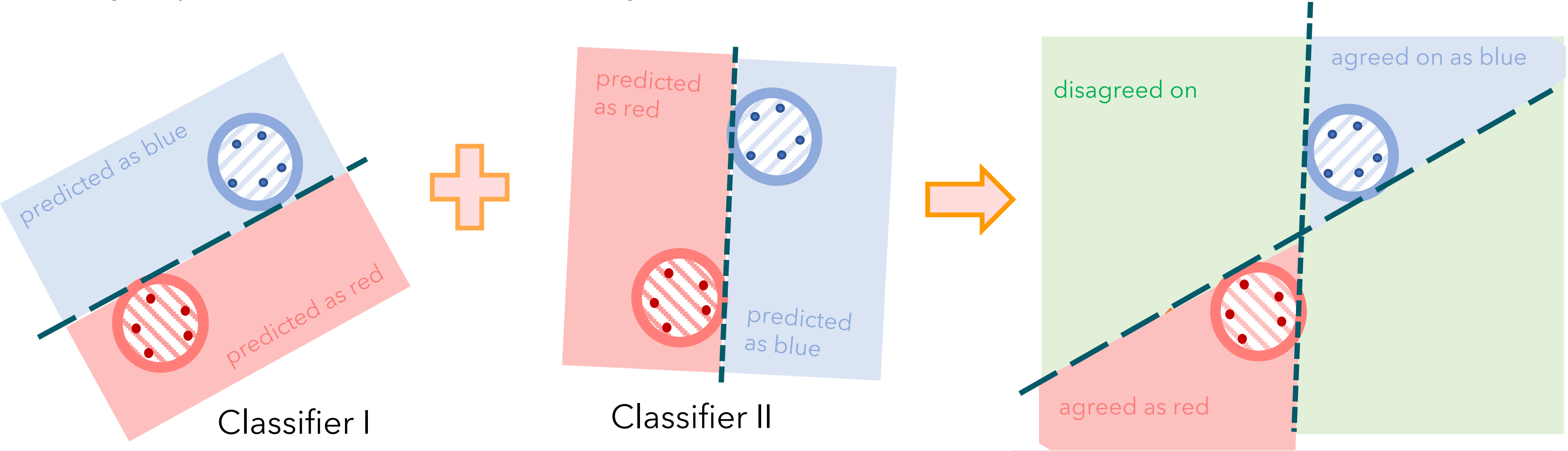
Semi-supervised novelty detection using ensembles

OOD definition suggests following procedure: given K models

- with good validation accuracy on old classes
- but different predictions outside of training distribution

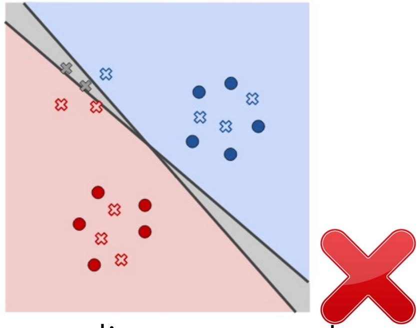
→ flag all points where the models disagree as "novel"





Key for our improvement: Regularized disagreement

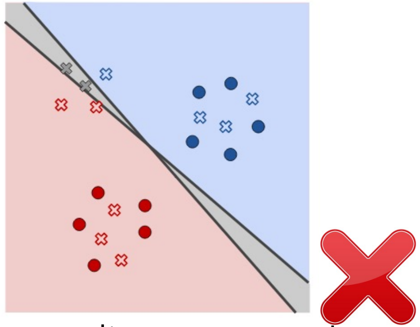
Key for "good performance": Complexity of ensemble models being only as large as needed



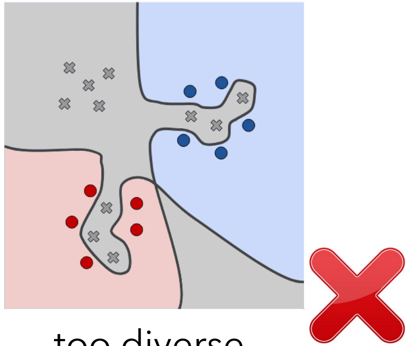
not diverse enough

Key for our improvement: Regularized disagreement

Key for "good performance": Complexity of ensemble models being only as large as needed



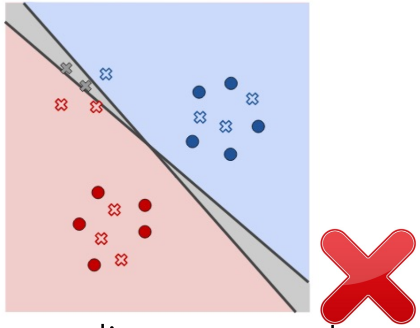
not diverse enough



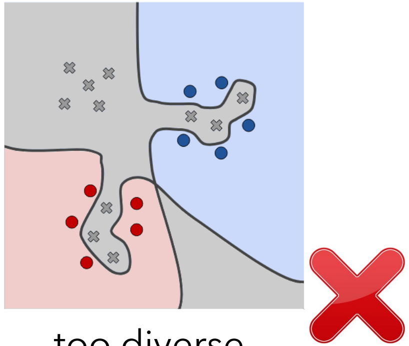
too diverse

Key for our improvement: Regularized disagreement

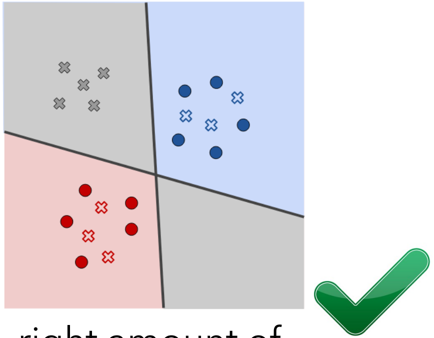
Key for "good performance": Complexity of ensemble models being only as large as needed



not diverse enough



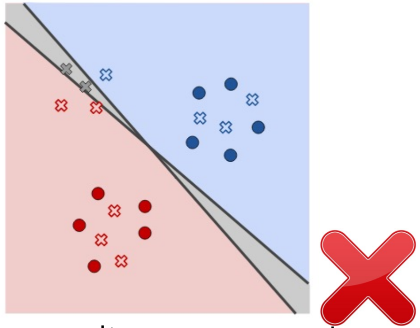
too diverse



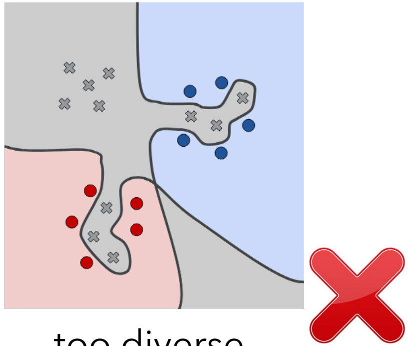
right amount of diversity

Key for our improvement: Regularized disagreement

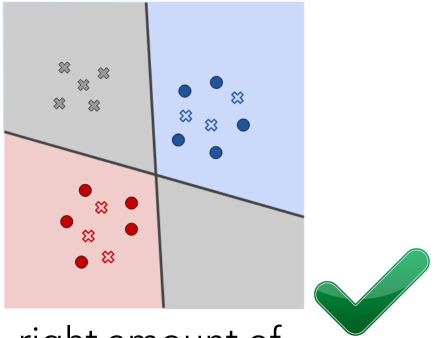
Key for "good performance": Complexity of ensemble models being only as large as needed



not diverse enough



too diverse



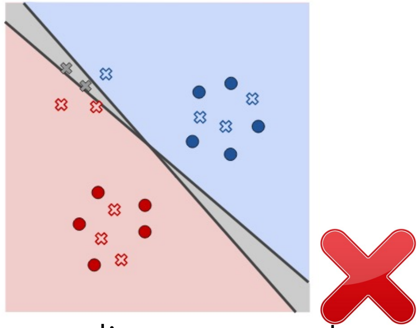
right amount of diversity

Idea for right amount of disagreement: maximize disagreement s.t. validation error of all models small

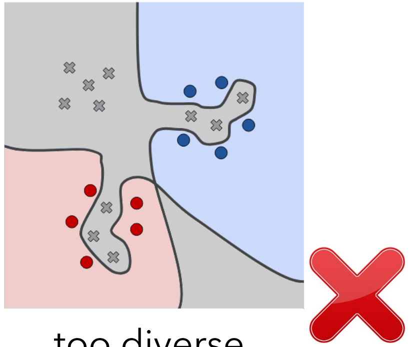
"regularization"

Key for our improvement: Regularized disagreement

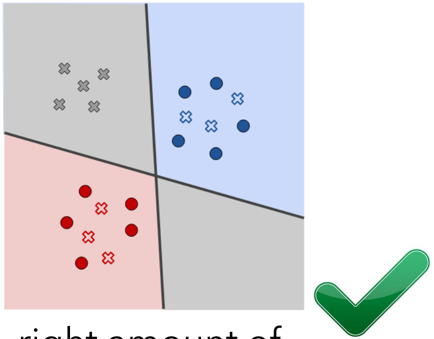
Key for "good performance": Complexity of ensemble models being only as large as needed



not diverse enough



too diverse



right amount of diversity

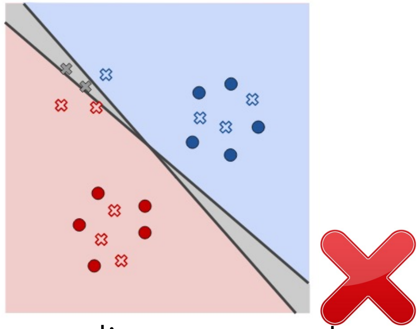
Idea for right amount of disagreement: maximize disagreement s.t. validation error of all models small

"regularization"

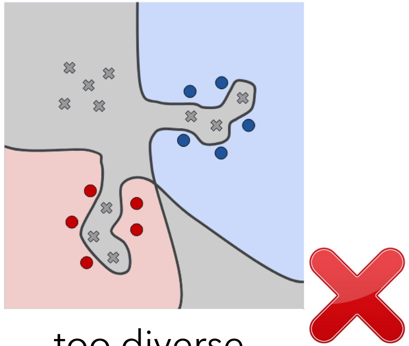
using unlabeled test data

Key for our improvement: Regularized disagreement

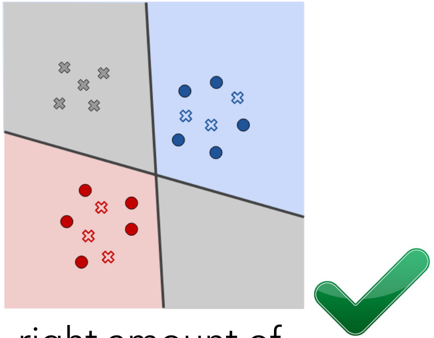
Key for "good performance": Complexity of ensemble models being only as large as needed



not diverse enough



too diverse



right amount of diversity

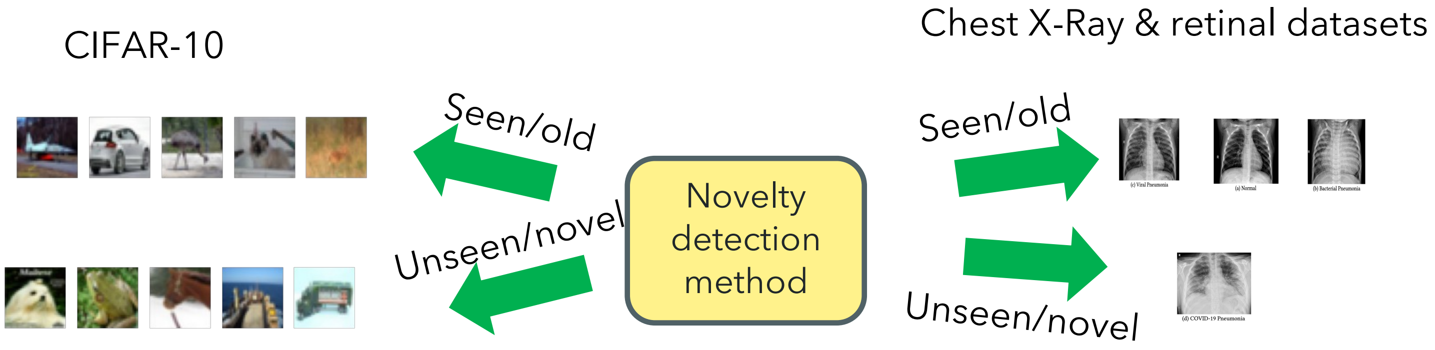
Idea for right amount of disagreement: maximize disagreement s.t. validation error of all models small

"regularization"

using unlabeled test data

using labeled training data

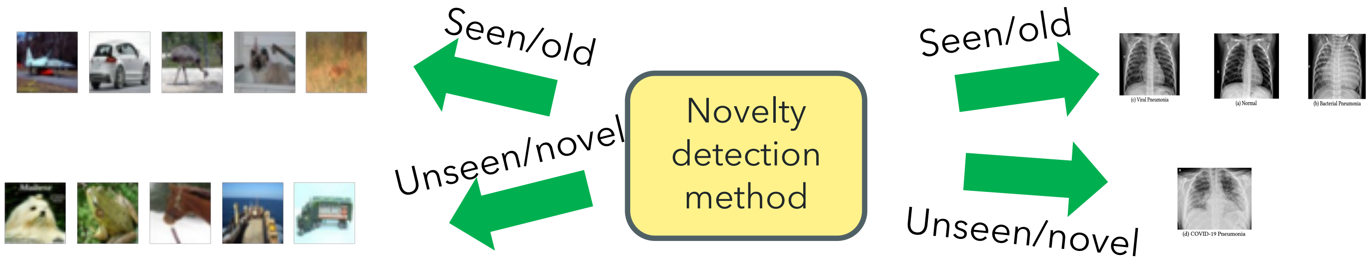
The near OOD problem on images with DNN

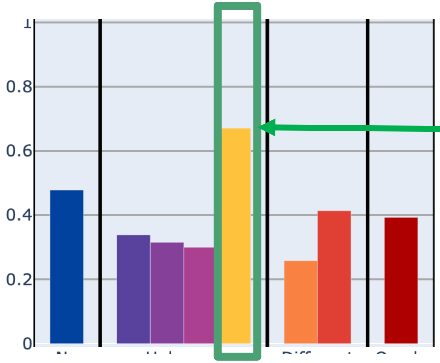


The near OOD problem on images with DNN

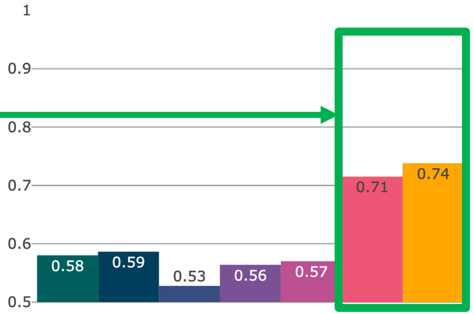
CIFAR-10

Chest X-Ray & retinal datasets





Ensembles with regularized disagreement



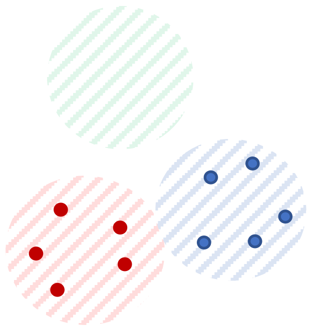
sml.inf.ethz.ch

- *"Hidden yet quantifiable: A lower bound for confounding strength using randomized trials"* by Piersilvio De Bartolomeis*, Javier Abad*, Konstantin Donhauser, FY, arxiv preprint

- *"Semi-supervised novelty detection using ensembles with regularized disagreement"* by Alexandru Țifrea, Eric Stavarache, and FY, (UAI), 2022

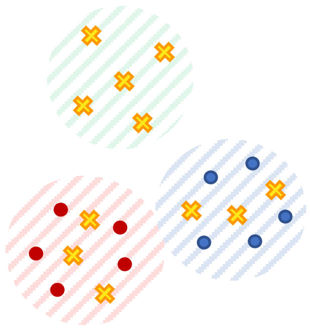
Maximizing disagreement using unlabeled data

- labeled training points
- training distribution
- novel classes
- unlabeled test points
  from both old & new classes

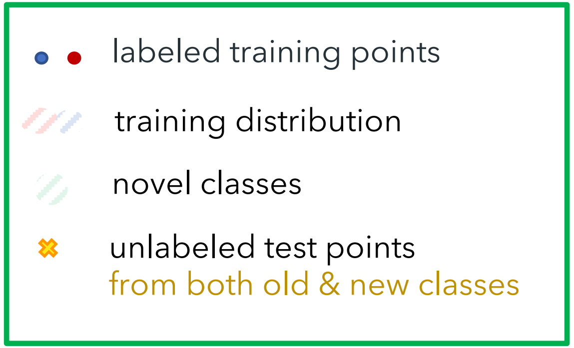


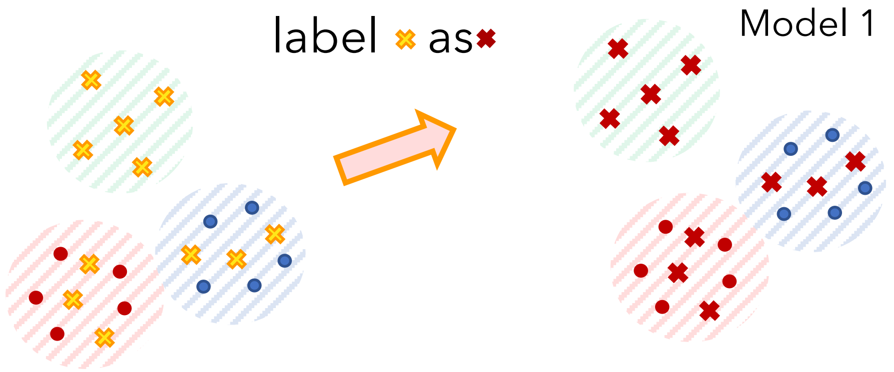
Maximizing disagreement using unlabeled data

- labeled training points
- training distribution
- novel classes
- unlabeled test points
  from both old & new classes



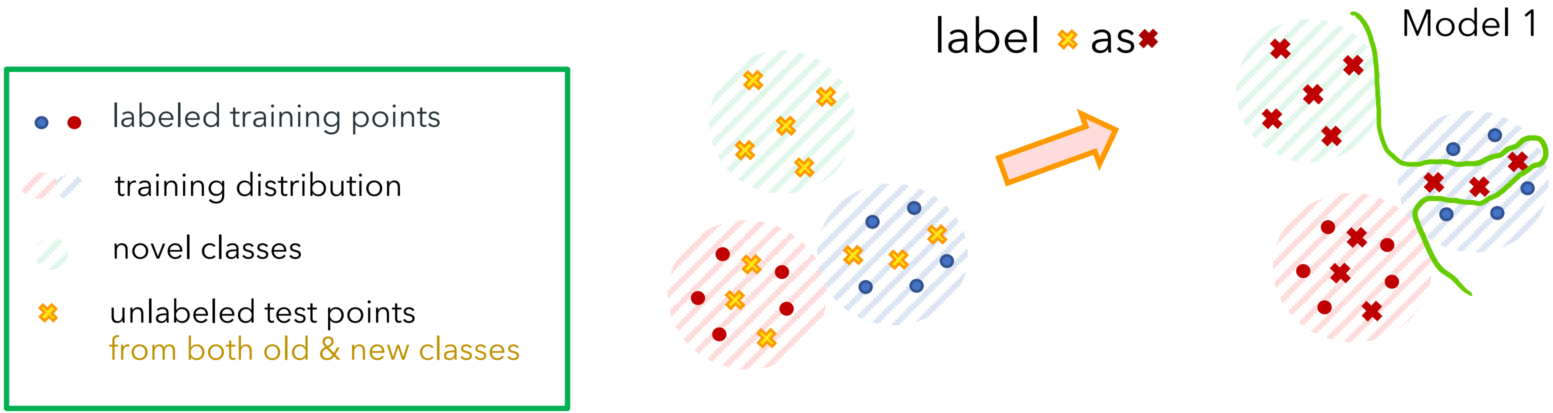
Maximizing disagreement using unlabeled data





- Artificially label all unlabeled test data with one label

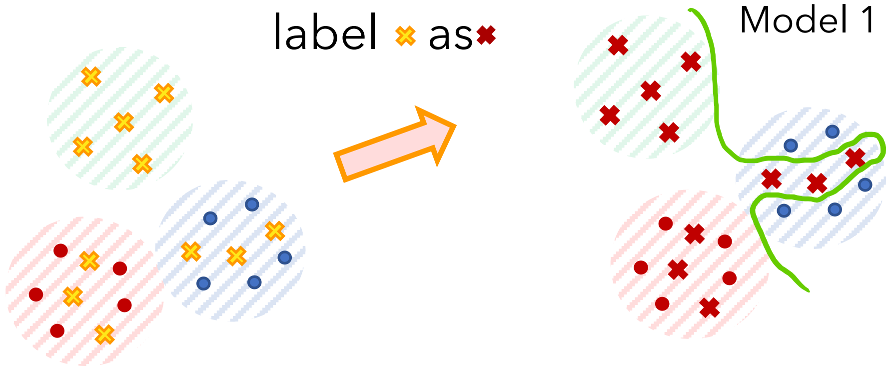
Maximizing disagreement using unlabeled data

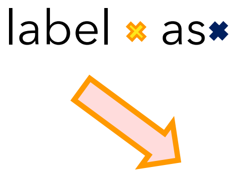


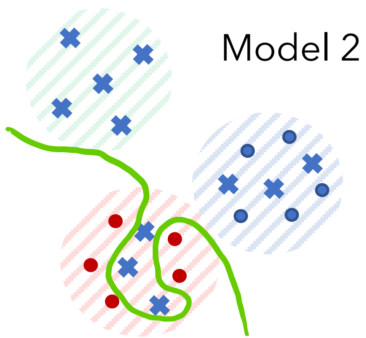
- Artificially label all unlabeled test data with one label

Maximizing disagreement using unlabeled data

- labeled training points
- training distribution
- novel classes
- unlabeled test points
  from both old & new classes

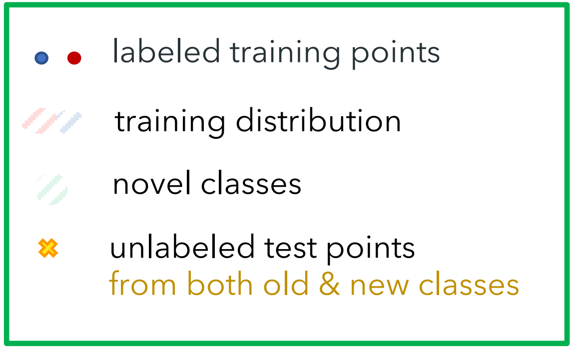


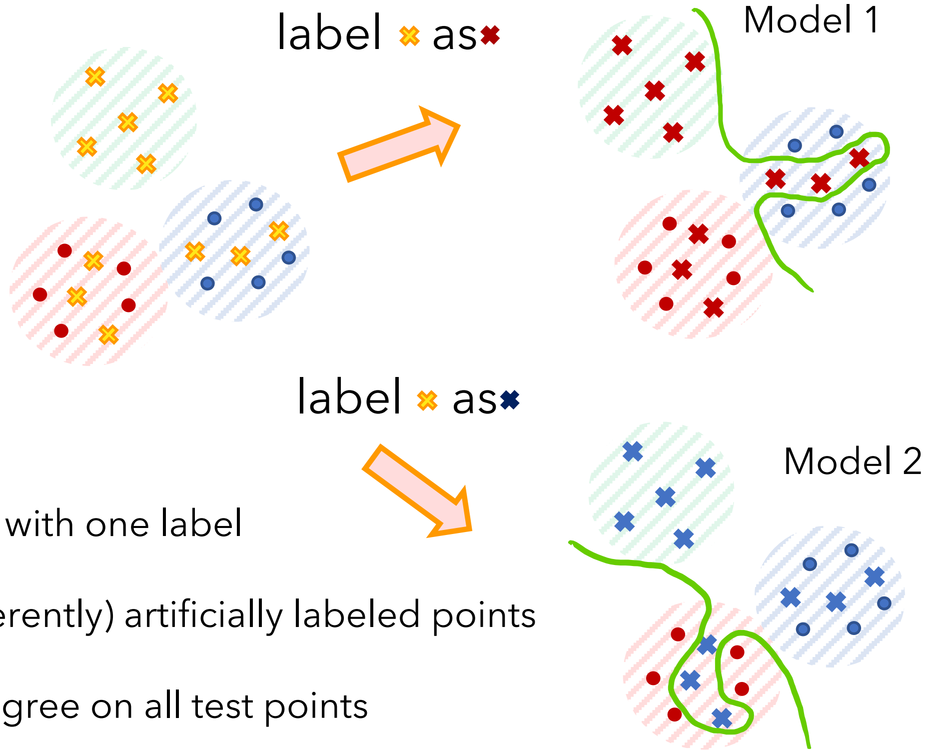




- Artificially label all unlabeled test data with one label
- Fit different models on labeled & (differently) artificially labeled points

Maximizing disagreement using unlabeled data



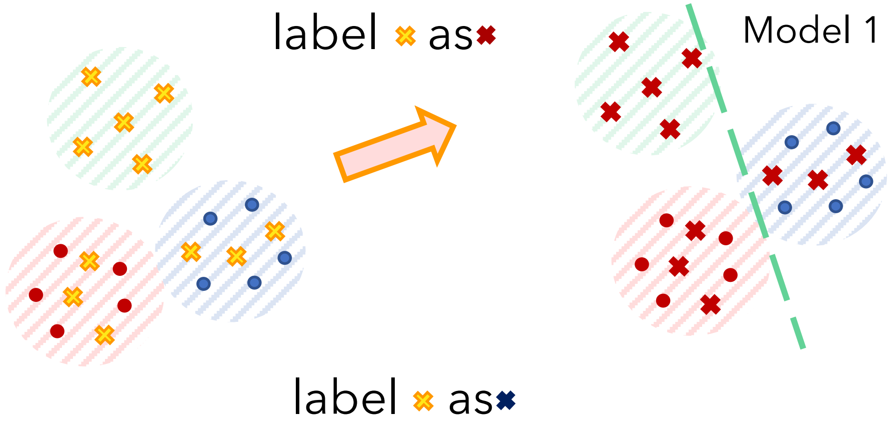


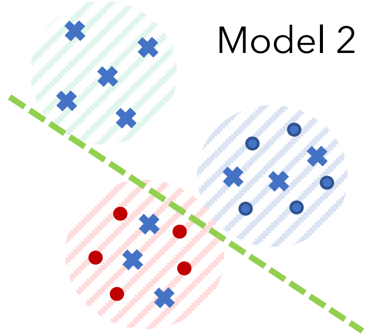
- Artificially label all unlabeled test data with one label
- Fit different models on labeled & (differently) artificially labeled points

… NNs can fit every point perfectly → disagree on all test points

Regularizing disagreement using labeled data

- labeled training points
- training distribution
- novel classes
- unlabeled test points
  from both old & new classes



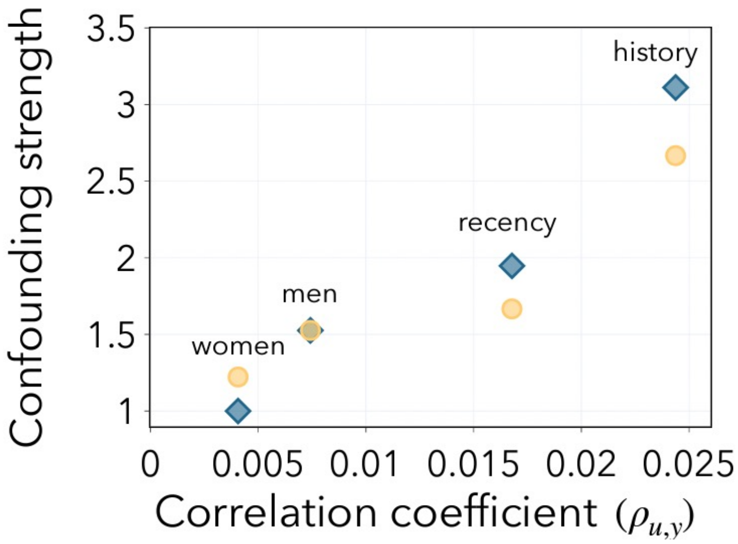


- Artificially label all unlabeled test data with one label
- Fit different models on labeled & (differently) artificially labeled points

… such that validation error is low (e.g. using early stopping)

Current and future work

Non-adversarial confounding



Discussion of the paradigm

- Propose two tests $\phi(\Gamma)$ based on (C)ATE sensitivity analysis intervals
  - obs: estimate mu with importance weighting rct, then ATE sensitivity

    valid when ATE bounds are asymptotically normal
  - rct: estimate mu on rct, then CATE sensitivity on obs -> average on rct

    valid when CATE sensitivity bounds converge at a $1/\sqrt{n}$ rate and $n_{rct} \ll n_{os}$