# Interactive Label Cleaning with Example-based Explanations

Stefano Teso[1], **Andrea Bontempelli**[1], Fausto Giunchiglia[1,2], Andrea Passerini[1]

SML Journal club

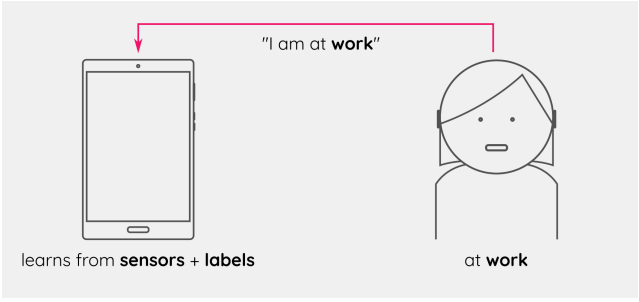
[1] University of Trento, Italy
[2] Jilin University, China




# Outline

- Label noise
- Skeptical learning and its limitations
- Our solution: explainable skepticism
- Experiments

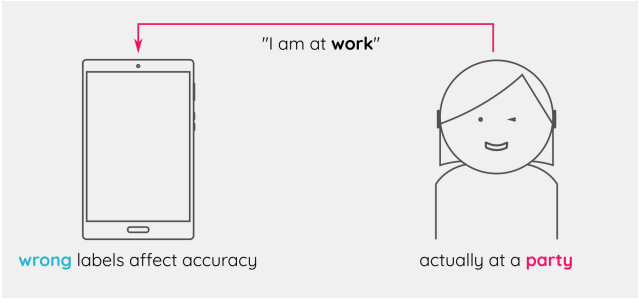
Learning from **sequence** of examples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots$




**Examples**: crowd-sourcing, citizen science, interactive personal assistants learning from diaries, ...

# Label Noise

The labels $\tilde{y}_1, \tilde{y}_2, \ldots$ are often **noisy**!




**Credulous** machines *suffer* with **inexperienced** annotators, **unwillingness** to self-report, *etc.*

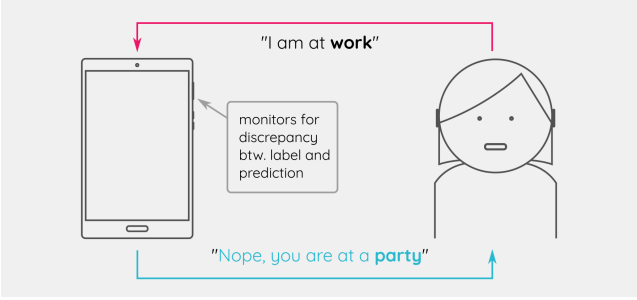
The fraction of noisy labels can be **very high**.

**Example**: users answering questions about their current location and means of transportation.

| | **annotation mistakes** | | |
|---|---|---|---|
| | $\leq 10\%$ | 10-25% | $\geq 25\%$ |
| **users** | 22% | 51% | 27% |

Source: [Zhang et al., 2022]

**Skeptical** machines challenge the user about **suspicious examples**
[Zeni et al., 2019, Bontempelli et al., 2020]




---

$(\mathbf{x}, \tilde{y})$ suspicious if likelihood of model's **prediction** $\hat{y}$ is much larger than likelihood of **annotation** $\tilde{y}$

$$p_\theta(\hat{y} \mid \mathbf{x}) \gg p_\theta(\tilde{y} \mid \mathbf{x})$$

The user is asked to **double-check** and **relabel** the suspicious examples


Often enough to correct mistakes due, e.g., to **inattention**

**Constraint**: keep the number of queries at a minimum, not to **overload** the user

1. **Active**: asks the user to label new instances $\mathbf{x}_t$ on which the machine is **uncertain**

2. **Skeptical**: asks the user to double-check and relabel new examples $(\mathbf{x}, \tilde{y})$ that look **suspicious**

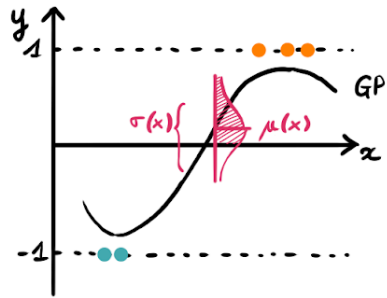
**for** $t = 1, 2, \ldots$ **do**
  receive $\mathbf{x}_t$
  predict $\hat{y}_t$ for $\mathbf{x}_t$
  **if uncertain about** $\hat{y}_t$ **then**
    request label, receive $\tilde{y}_t$
    **if skeptical about** $\tilde{y}$ **then**
      challenge user with $\hat{y}_t$, receive $y'_t$
    add $(\mathbf{x}_t, y'_t)$ to data set
    update classifier

The decision to **query** the user depends on the **uncertainty estimation** provided by a set of Gaussian Processes.

- one **Gaussian Process** for each class
- Probability that a class is positive:

$$p(f(\mathbf{x}) \geq 0 \mid \mathbf{x}) \;=\; \Phi\left(\frac{\mu(\mathbf{x})}{\sigma(\mathbf{x})}\right)$$

that is, the **shaded purple area**.




[Kapoor et al., 2007]

# Skeptical Learning: query label when uncertain

1: **for** $t = 1, 2, \ldots$ **do**
2: receive $\mathbf{x}_t$
3: predict $\hat{y}_t$ for $\mathbf{x}_t$
4: **if uncertain about** $\hat{y}_t$ **then**
5: request label, receive $\tilde{y}_t$
6: **if** skeptical about $\tilde{y}$ **then**
7: challenge user with $\hat{y}_t$, receive $y'_t$
8: add $(\mathbf{x}_t, y'_t)$ to data set
9: update classifier

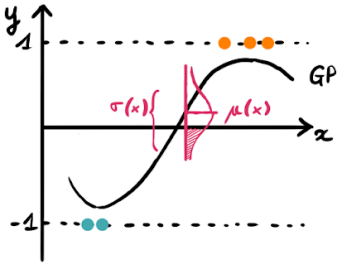
**Intuition**: query the user if $p_{\hat{y}_t}(1 \mid \mathbf{x}_t)$ is small

[Bontempelli et al., 2020]

Query the user with **probability** $\alpha_t$

$$\begin{aligned} \alpha_t &= p_{\hat{y}_t}(f(\mathbf{x}_t) \leq 0 \mid \mathbf{x}_t) \\ &= 1 - \Phi\left(\frac{\mu_{\hat{y}_t}(\mathbf{x}_t)}{\sigma_t(\mathbf{x}_t)}\right) \end{aligned}$$

i.e., the **shaded purple area**.



## Skeptical Learning: challenge the user when skeptical

1: **for** $t = 1, 2, \ldots$ **do**
2: receive $\mathbf{x}_t$
3: predict $\hat{y}_t$ for $\mathbf{x}_t$
4: **if** uncertain about $\hat{y}_t$ **then**
5: request label, receive $\tilde{y}_t$
6: **if skeptical about** $\tilde{y}$ **then**
7: challenge user with $\hat{y}_t$, receive $y'_t$
8: add $(\mathbf{x}_t, y'_t)$ to data set
9: update classifier

Be skeptical with **probability** $\gamma_t$

$$\gamma_t = p(f_{\hat{y}_t}(\mathbf{x}_t) - f_{\tilde{y}_t}(\mathbf{x}_t) \geq 0)$$

$$= \Phi\left(\frac{\mu_{\hat{y}_t}(\mathbf{x}_t) - \mu_{\tilde{y}_t}(\mathbf{x}_t)}{\sigma_t(\mathbf{x}_t)}\right)$$

**Intuition**: skeptical if the machine is approximately **equally certain** about predicted label $\hat{y}_t$ and annotated label $\tilde{y}_t$.

## Limitations of Skeptical Learning

**Cleans incoming examples only**:

- Noisy data in the **bootstrap** data set
- Incoming examples that **elude** the skeptical check

Accumulated noise impacts **predictions** and **ability to be skeptical**: new mislabeled examples falling close to noisy regions are harder to detect

## Limitations of Skeptical Learning

**Cleans incoming examples only**:

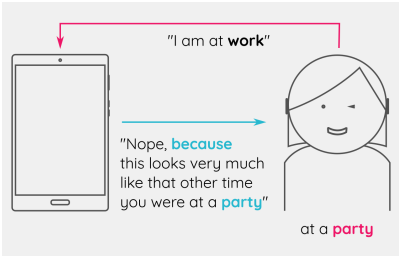
- Noisy data in the **bootstrap** data set
- Incoming examples that **elude** the skeptical check

Accumulated noise impacts **predictions** and **ability to be skeptical**: new mislabeled examples falling close to noisy regions are harder to detect

**Completely black-box**

- The user has no clue why the model is skeptical
- Is the model skeptical **for the right reasons**?

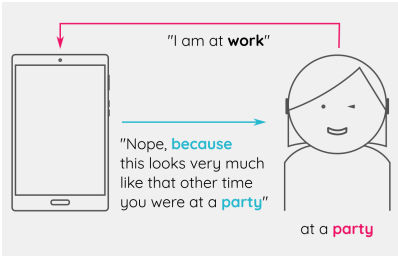
## Machine skeptical **for the right reasons**




Skepticism **supported** by data: e.g., there is a *past example* $z_k = (\mathbf{x}_k, y_k)$ that is *similar* to the current one but has a *different* label.

This example is **clean**.

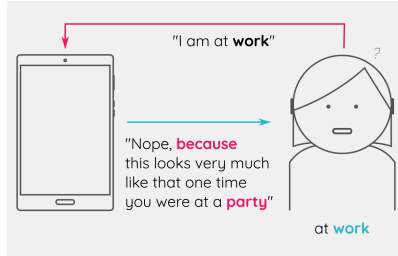
## Machine skeptical **for the right reasons**




Skepticism **supported** by data: e.g., there is a *past example* $z_k = (\mathbf{x}_k, y_k)$ that is *similar* to the current one but has a *different* label.

This example is **clean**.

## Machine skeptical **for the wrong reasons**




Past data that supports skepticism is **mislabeled**.

E.g., because of user's **past mistakes** – or *lies* ;-)

## Solution: Interacting through Example-based Explanations

A **counter-example** is a **concrete** past example $z_k \in D$ that explains **why the model is skeptical** about $\tilde{z}_t$

## Solution: Interacting through Example-based Explanations

A **counter-example** is a **concrete** past example $z_k \in D$ that explains **why the model is skeptical** about $\tilde{z}_t$

Show counter-examples to annotator and let them **fix them** if needed!

## Solution: Interacting through Example-based Explanations

A **counter-example** is a **concrete** past example $z_k \in D$ that explains **why the model is skeptical** about $\tilde{z}_t$

Show counter-examples to annotator and let them **fix them** if needed!

D1. **Contrastive**: explains why $\tilde{z}_t$ is suspicious, highlighting a potential **inconsistency in data**

## Solution: Interacting through Example-based Explanations

A **counter-example** is a **concrete** past example $z_k \in D$ that explains **why the model is skeptical** about $\tilde{z}_t$

Show counter-examples to annotator and let them **fix them** if needed!

D1. **Contrastive**: explains why $\tilde{z}_t$ is suspicious, highlighting a potential **inconsistency in data**

D2. **Influential**: correcting it should **improve the model** as much as possible

## Solution: Interacting through Example-based Explanations

A **counter-example** is a **concrete** past example $z_k \in D$ that explains **why the model is skeptical** about $\tilde{z}_t$

Show counter-examples to annotator and let them **fix them** if needed!

D1. **Contrastive**: explains why $\tilde{z}_t$ is suspicious, highlighting a potential **inconsistency in data**

D2. **Influential**: correcting it should **improve the model** as much as possible

D3. **Pertinent**: it should be clear **to the user**

## Solution: Interacting through Example-based Explanations

A **counter-example** is a **concrete** past example $z_k \in D$ that explains **why the model is skeptical** about $\tilde{z}_t$

Show counter-examples to annotator and let them **fix them** if needed!

D1. **Contrastive**: explains why $\tilde{z}_t$ is suspicious, highlighting a potential **inconsistency in data**

D2. **Influential**: correcting it should **improve the model** as much as possible

D3. **Pertinent**: it should be clear **to the user**

Do such examples **exist**? How to **identify** them?

## Selecting a Contrastive CE

$z_k \in D$ is **contrastive** (supports skepticism) if *removing* it gives a *less* suspicious model

## Selecting a Contrastive CE

$z_k \in D$ is **contrastive** (supports skepticism) if *removing* it gives a *less* suspicious model

Find $z_k = (\mathbf{x}_k, y_k)$ that **maximizes** the difference in likelihood for suspicious example $\tilde{z}_t$ pre/post removing $z_k$:

$$\underbrace{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}^{-k})}_{\text{model without } z_k} \quad - \quad \underbrace{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}_{\text{current model}}$$

## Selecting a Contrastive CE

$z_k \in D$ is **contrastive** (supports skepticism) if *removing* it gives a *less* suspicious model

Find $z_k = (\mathbf{x}_k, y_k)$ that **maximizes** the difference in likelihood for suspicious example $\tilde{z}_t$ pre/post removing $z_k$:

$$\underbrace{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}^{-k})}_{\text{model without } z_k} \quad - \quad \underbrace{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}_{\text{current model}}$$

**Impractical**: requires to **retrain** $|D|$ **times**!

## Selecting a Contrastive CE

**Influence Functions** (IFs) approximate the *change in parameters* $\theta_t$ due to reweighting an example $z$:

$$\mathcal{I}_{\theta_t}(z) := \frac{d}{d\epsilon}\theta_t(z,\epsilon)\Big|_{\epsilon=0} \approx -H(\theta_t)^{-1}\nabla_\theta \ell(z,\theta_t)$$

where $H(\theta_t)$ is the **Hessian**. Apply also to **non-convex models** [Koh and Liang, 2017]

**Influence Functions** (IFs) approximate the *change in parameters* $\theta_t$ due to reweighting an example $z$:

$$\mathcal{I}_{\theta_t}(z) := \frac{d}{d\epsilon}\theta_t(z,\epsilon)\Big|_{\epsilon=0} \approx -H(\theta_t)^{-1}\nabla_\theta \ell(z,\theta_t)$$

where $H(\theta_t)$ is the **Hessian**. Apply also to **non-convex models** [Koh and Liang, 2017]

Use IFs to compute the **change in likelihood** via chain rule:

$$P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}^{-k}) - P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \approx -\frac{1}{t-1}\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})^\top \mathcal{I}_{\theta_{t-1}}(z_k)$$

## Selecting a Contrastive CE

Find $z_k = (\mathbf{x}_k, y_k)$ that **maximizes** the **IF approximation**:

$$\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})^\top \underbrace{H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1})}_{\text{Influence Function}}$$

## Selecting a Contrastive CE

Find $z_k = (\mathbf{x}_k, y_k)$ that **maximizes** the **IF approximation**:

$$\underbrace{\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})^\top H(\theta_{t-1})^{-1}}_{\text{constant w.r.t. } k} \nabla_\theta \ell(z_k, \theta_{t-1})$$

**Speed-up**: **cache inverse Hessian-vector product**, use efficient stochastic estimator.

## Contrastive CEs are Influential!

For the **cross-entropy loss** $\ell(z, \theta) = -\log P(y \mid \mathbf{x}; \theta)$:

$$\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \quad = \quad P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \, \frac{\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}$$

## Contrastive CEs are Influential!

For the **cross-entropy loss** $\ell(z, \theta) = -\log P(y \mid \mathbf{x}; \theta)$:

$$\begin{aligned} \nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) &= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \frac{\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})} \\ &= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \log P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \end{aligned}$$

## Contrastive CEs are Influential!

For the **cross-entropy loss** $\ell(z, \theta) = -\log P(y \mid \mathbf{x}; \theta)$:

$$\begin{aligned}
\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) &= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \frac{\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})} \\
&= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \log P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \\
&= -P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})
\end{aligned}$$

## Contrastive CEs are Influential!

For the **cross-entropy loss** $\ell(z, \theta) = -\log P(y \mid \mathbf{x}; \theta)$:

$$\begin{aligned}
\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) &= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \frac{\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})} \\
&= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \log P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \\
&= -P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})
\end{aligned}$$

Hence, counter-example selection objective becomes:

$$\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1})$$

## Contrastive CEs are Influential!

For the **cross-entropy loss** $\ell(z, \theta) = -\log P(y \mid \mathbf{x}; \theta)$:

$$\begin{aligned}
\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) &= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \frac{\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})} \\
&= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \log P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \\
&= -P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})
\end{aligned}$$

Hence, counter-example selection objective becomes:

$$\begin{aligned}
&\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1}) \\
&= -P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1})
\end{aligned}$$

## Contrastive CEs are Influential!

For the **cross-entropy loss** $\ell(z, \theta) = -\log P(y \mid \mathbf{x}; \theta)$:

$$\begin{aligned}
\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) &= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \frac{\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})} \\
&= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \log P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \\
&= -P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})
\end{aligned}$$

Hence, counter-example selection objective becomes:

$$\begin{aligned}
&\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1}) \\
&= -P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1}) \\
&\propto -\nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1})
\end{aligned}$$

This **matches** the definition of influential examples [Koh and Liang, 2017]!

## Contrastive CEs are Influential!

For the **cross-entropy loss** $\ell(z, \theta) = -\log P(y \mid \mathbf{x}; \theta)$:

$$\begin{aligned}
\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) &= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \frac{\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})}{P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})} \\
&= P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \log P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \\
&= -P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) \nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})
\end{aligned}$$

Hence, counter-example selection objective becomes:

$$\begin{aligned}
&\nabla_\theta P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1}) \\
= -P(\tilde{y}_t \mid \mathbf{x}_t; \theta_{t-1}) &\nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1}) \\
&\propto -\nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1})
\end{aligned}$$

This **matches** the definition of influential examples [Koh and Liang, 2017]!

**influential counter-examples $\equiv$ contrastive counter-examples**

## From Influence to Fisher Information

- IFs often **unstable** as $H$ non-invertible [Basu et al., 2020]. **Noise makes it worse**.

## From Influence to Fisher Information

- IFs often **unstable** as $H$ non-invertible [Basu et al., 2020]. **Noise makes it worse**.

- Replace $H$ by **Fisher information matrix** (FIM) $F$ [Martens and Grosse, 2015]:

$$F(\theta) := \frac{1}{t-1} \sum_{k=1}^{t-1} \mathbb{E}_{y \sim P(Y \mid \mathbf{x}_k, \theta)} \left[ \nabla_\theta \log P(y \mid \mathbf{x}_k, \theta) \nabla_\theta \log P(y \mid \mathbf{x}_k, \theta)^\top \right] \tag{1}$$

The counter-example selection objective

$$\underset{k \in [t-1]}{\operatorname{argmax}} \; -\nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})^\top H(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1})$$

can be rewritten as

$$\underset{k \in [t-1]}{\operatorname{argmax}} \; -\nabla_\theta \ell(\tilde{z}_t, \theta_{t-1})^\top F(\theta_{t-1})^{-1} \nabla_\theta \ell(z_k, \theta_{t-1})$$

## From Influence to Fisher Information

- Replace $H$ by **Fisher information matrix** (FIM) $F$ [Martens and Grosse, 2015]:

$$F(\theta) := \frac{1}{t-1} \sum_{k=1}^{t-1} \mathbb{E}_{y \sim P(Y \mid \mathbf{x}_k, \theta)} \left[ \nabla_\theta \log P(y \mid \mathbf{x}_k, \theta) \nabla_\theta \log P(y \mid \mathbf{x}_k, \theta)^\top \right]$$

  - $F$ is PSD, so inversion is easy.
  - If $p_\theta$ approximates the data distribution, $F$ approximates $H$.
  - Even if this does not hold (as with noise), the FIM still captures useful curvature information.
  - Caching still works!

## From Influence to Fisher Information

- Replace $H$ by **Fisher information matrix** (FIM) $F$ [Martens and Grosse, 2015]:

$$F(\theta) := \frac{1}{t-1} \sum_{k=1}^{t-1} \mathbb{E}_{y \sim P(Y \mid \mathbf{x}_k, \theta)} \left[ \nabla_\theta \log P(y \mid \mathbf{x}_k, \theta) \nabla_\theta \log P(y \mid \mathbf{x}_k, \theta)^\top \right]$$

  - $F$ is PSD, so inversion is easy.
  - If $p_\theta$ approximates the data distribution, $F$ approximates $H$.
  - Even if this does not hold (as with noise), the FIM still captures useful curvature information.
  - Caching still works!

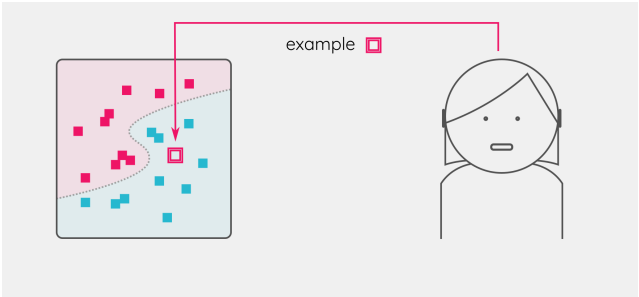
**Speed-up**: restrict FIM to **top layer** of neural net (**Top Fisher**)

## Pertinence

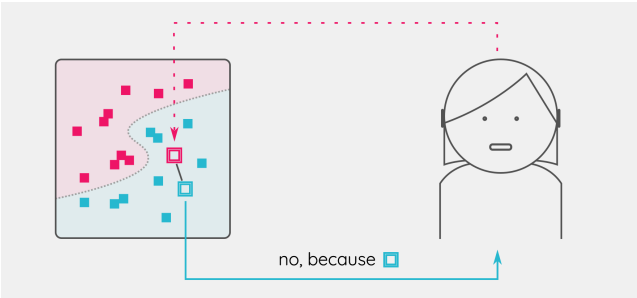
D3. **Pertinence**: $z_k$ it should be clear **to the user**

Ensure label of counter-example $z_k$ **matches the prediction** for the suspicious example $\tilde{z}_t$: *can interpret $z_k$ as supporting the machine's suspicion.*

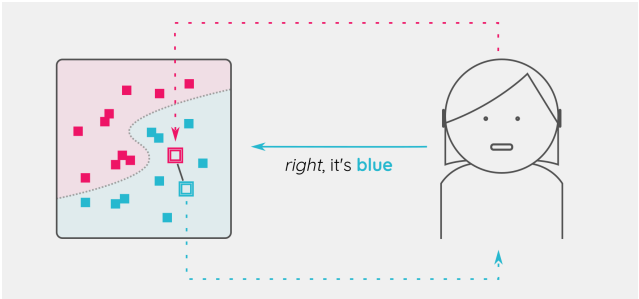
Example #1



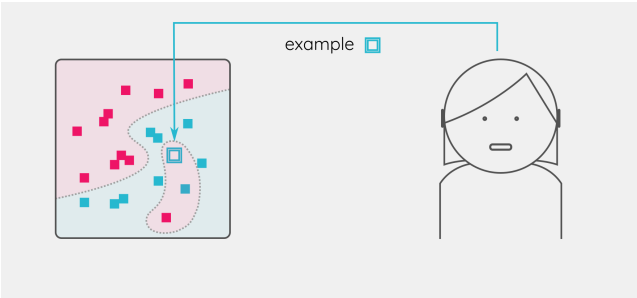
Example #1



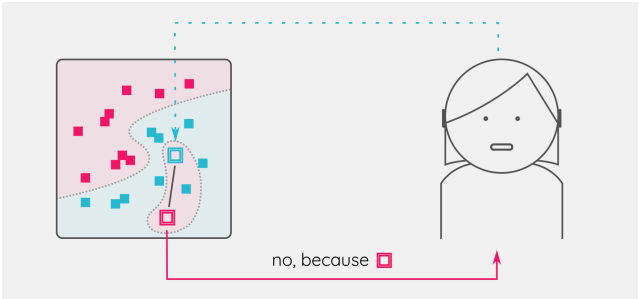
Example #1



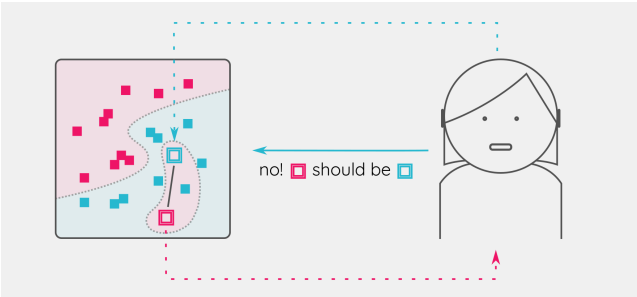
Example #2



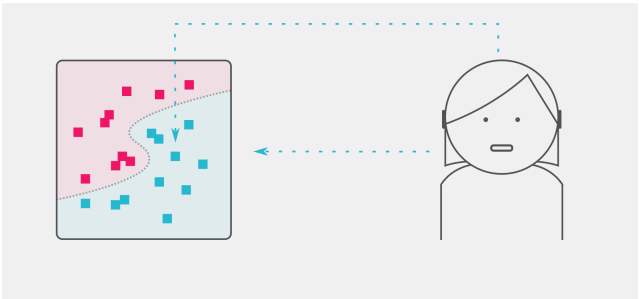
Example #2



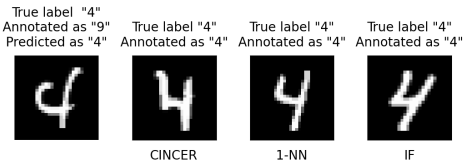
Example #2



Example #2

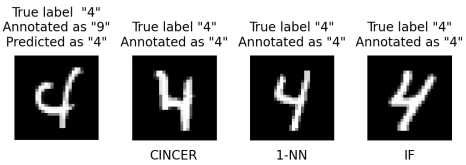


# Illustration

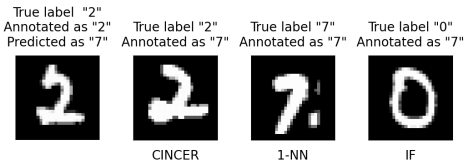



■ suspicious example is mislabeled; machine's suspicion is supported by a **clean counter-example**.

# Illustration

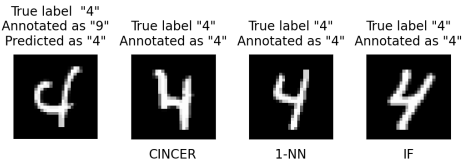


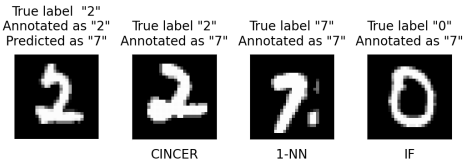

- suspicious example is mislabeled; machine's suspicion is supported by a **clean counter-example**.




- suspicious example is clean; machine's suspicious supported by **mislabeled counter-example**.

# Illustration




- suspicious example is mislabeled; machine's suspicion is supported by a **clean counter-example**.
- suspicious example is clean; machine's suspicious supported by **mislabeled counter-example**.

**Take-away**: 1-NN's is *not* influential, IF's is *not* pertinent.

## Experiments

Q1 Do counter-examples contribute to cleaning data?

Q2 Which influence-based selection strategy identifies the most mislabeled counter-examples?

Q3 What contributes to the effectiveness of the best counter-example selection?

## Experiments

Q1 Do counter-examples contribute to cleaning data?

Q2 Which influence-based selection strategy identifies the most mislabeled counter-examples?

Q3 What contributes to the effectiveness of the best counter-example selection?

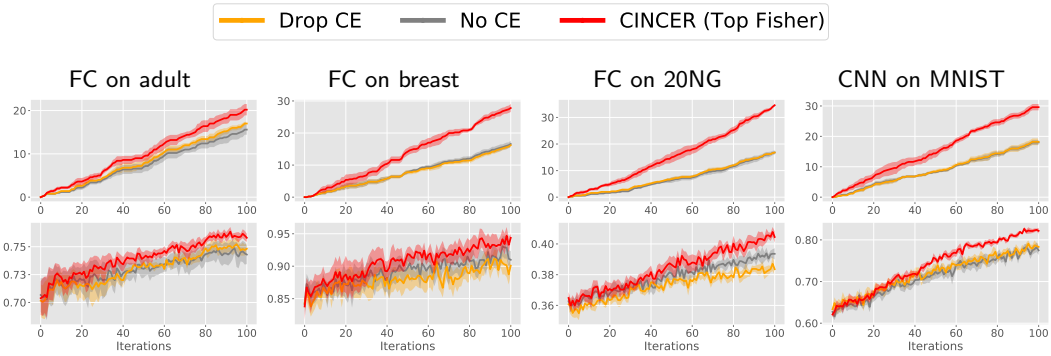
### Data sets

- tabular data (Adult, Breast, 20NG) and images (MNIST, Fashion)
- labels corrupted with 20% probability

### Models

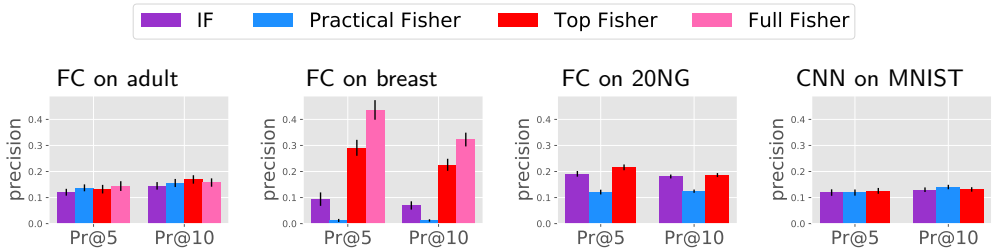
- LR, logistic regression
- FC, feed-forward neural network with two fully connected hidden layers with ReLU activations
- CNN, a feed forward neural network with two convolutional layers and two fully connected layers

# Q1: Counter-examples help improve data



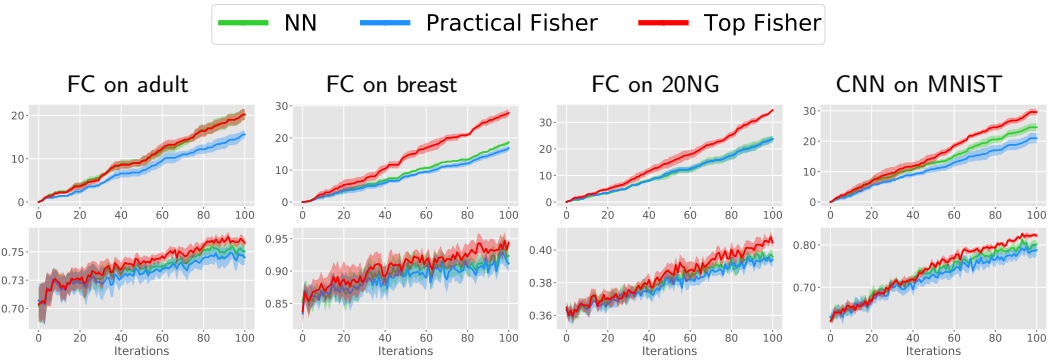

**Top Fisher** vs. dropping CEs vs. ignoring CEs. **Top**: # of cleaned examples. **Bottom**: $F_1$ score.

# Q2: FIM identifies the most mislabeled counter-examples




Mistake Pr@5 and Pr@10 for counter-examples.

# Q3: Influence & Curvature are both important




**Top Fisher** vs. practical Fisher vs. NN. **Top row**: # of cleaned examples. **Bottom row**: $F_1$ score.

## Take-away

- CINCER makes skepticism **explainable**, enables **interactive cleaning** of bad training examples
- Explanations provided by **counter-examples** with well-defined properties
- Stabilize and speed-up influence computation using **Fisher information matrix**
- Leads to **better data & models**, allows to **establish/reject trust**

# Thank You!

andrea.bontempelli@unitn.it

Paper: https://arxiv.org/abs/2106.03922

Code: https://github.com/abonte/cincer

Basu, S., Pope, P., and Feizi, S. (2020).
**Influence functions in deep learning are fragile.**
*arXiv preprint arXiv:2006.14651*.

Bontempelli, A., Teso, S., Giunchiglia, F., and Passerini, A. (2020).
**Learning in the Wild with Incremental Skeptical Gaussian Processes.**
In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*.

Kapoor, A., Grauman, K., Urtasun, R., and Darrell, T. (2007).
**Active learning with gaussian processes for object categorization.**
In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE.

Koh, P. W. and Liang, P. (2017).
**Understanding black-box predictions via influence functions.**
In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894. JMLR. org.

Martens, J. and Grosse, R. (2015).
**Optimizing neural networks with kronecker-factored approximate curvature.**
In *International conference on machine learning*, pages 2408–2417. PMLR.


Zeni, M., Zhang, W., Bignotti, E., Passerini, A., and Giunchiglia, F. (2019).
**Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge.**
*Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):32.


Zhang, W., Zeni, M., Passerini, A., and Giunchiglia, F. (2022).
**Skeptical learning—an algorithm and a platform for dealing with mislabeling in personal context recognition.**
*Algorithms*, 15(4):109.

## Pseudo-code of CINCER

**Inputs**: initial (noisy) training set $D_0$; threshold $\tau$.

fit $\theta_0$ on $D_0$
**for** $t = 1, 2, \ldots$ **do**
    receive new example $\tilde{z}_t = (\mathbf{x}_t, \tilde{y}_t)$
    **if** $\mu(\tilde{z}_t, \theta_{t-1}) < \tau$ **then**
        $D_t \leftarrow D_{t-1} \cup \{\tilde{z}_t\}$ ▷ $\tilde{z}_t$ is compatible
    **else**
        find counterexample $z_k$ ▷ $\tilde{z}_t$ is suspicious
        present $\tilde{z}_t, z_k$ to the user, receive possibly cleaned labels $y'_t, y'_k$
        $D_t \leftarrow (D_{t-1} \setminus \{z_k\}) \cup \{(\mathbf{x}_t, y'_t), (\mathbf{x}_k, y'_k)\}$
    fit $\theta_t$ on $D_t$