# Differential Privacy has Bounded Impact on Fairness in Classification

Paul Mangold, **Michaël Perrot**, Aurélien Bellet, Marc Tommasi
Magnet Team

Workshop @ Comète on Ethical AI

November 23, 2023

# A Story of Squirrels

# A Story of Squirrels

# A Story of Squirrels

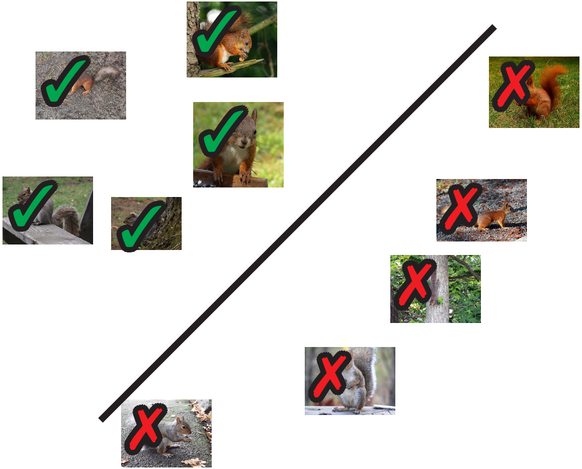
# A Story of Squirrels

# A Story of Squirrels

# A Story of Squirrels

# A Story of Squirrels

# A Story of Squirrels



# Notations

- Features space $\mathcal{X}$, labels space $\mathcal{Y} = \{-1, 1\}$, and sensitive attributes space $\mathcal{S}$.
  - Features: images, tabular data, graphs, ...
  - Labels: hired/not hired, profession, disease, ...
  - Sensitive attributes: gender, race, age, ...
- Decision function $h : \mathcal{X} \mapsto \mathbb{R}$ taken in a set $\mathcal{H}$.
- Binary decision function $H(x) = \begin{cases} 1 & \text{if } h(x) > 0, \\ -1 & \text{otherwise.} \end{cases}$.
- A set $D = \{(x_i, s_i, y_i)\}_{i=1}^{n}$ of $n$ examples drawn i.i.d. a distribution $\mathcal{D}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

**Goal:** Measure the fairness of the model $H$.

# Accuracy Parity [Zafar et al., 2017]

$$\text{For all } s \in \mathcal{S},\ F_s(h,D) = \underbrace{\mathbf{P}\left(H(X) = Y | S = s\right)}_{\text{Accuracy of } h \text{ on the group } s} - \underbrace{\mathbf{P}\left(H(X) = Y\right)}_{\text{Accuracy of } h.}$$

# Accuracy Parity [Zafar et al., 2017]

$$\text{For all } s \in \mathcal{S},\ F_s(h, D) = \underbrace{\mathbf{P}\left(H(X) = Y | S = s\right)}_{\text{Accuracy of } h \text{ on the group } s} - \underbrace{\mathbf{P}\left(H(X) = Y\right)}_{\text{Accuracy of } h.}$$

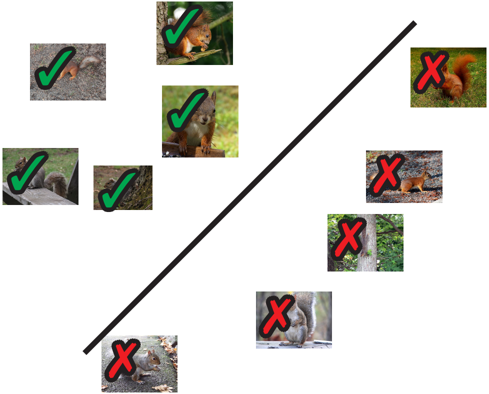
$$\mathbf{P}\left(H(X) = Y | S = \text{red}\right) = 1$$

$$\mathbf{P}\left(H(X) = Y\right) = \frac{9}{10}$$

$$F_{\text{red}}(h, D) = \frac{1}{10}$$

# Accuracy Parity [Zafar et al., 2017]

$$\text{For all } s \in \mathcal{S},\ F_s(h, D) = \underbrace{\mathbf{P}\left(H(X) = Y | S = s\right)}_{\text{Accuracy of } h \text{ on the group } s} - \underbrace{\mathbf{P}\left(H(X) = Y\right)}_{\text{Accuracy of } h.}$$

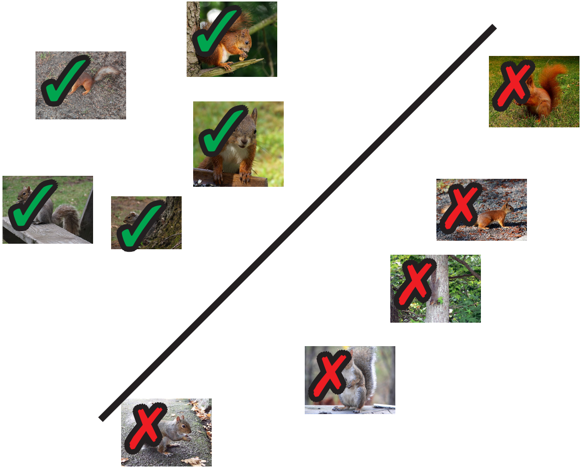


$$\mathbf{P}\left(H(X) = Y | S = \text{red}\right) = 1$$

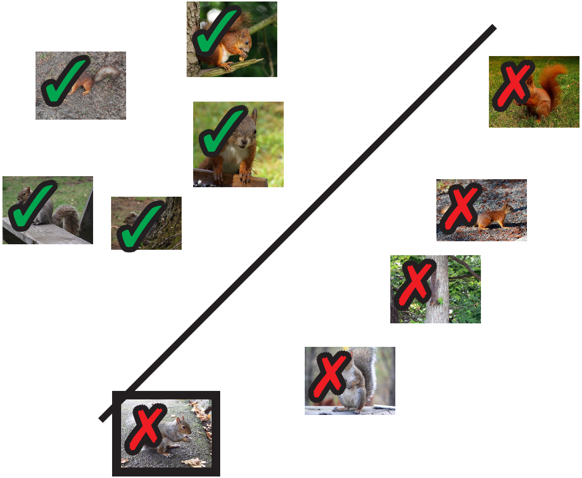
$$\mathbf{P}\left(H(X) = Y\right) = 1$$

$$F_{\text{red}}(h, D) = 0$$

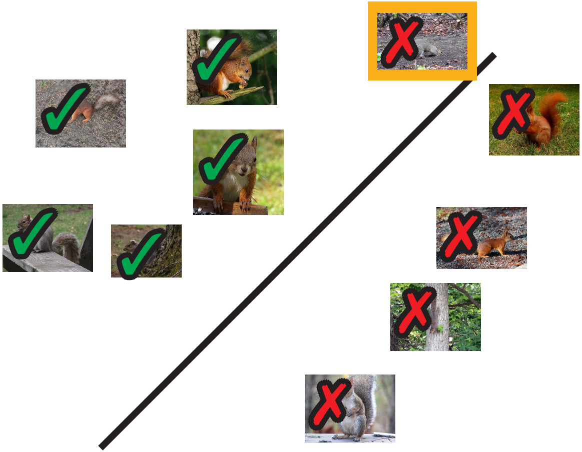
# A Story of Squirrels



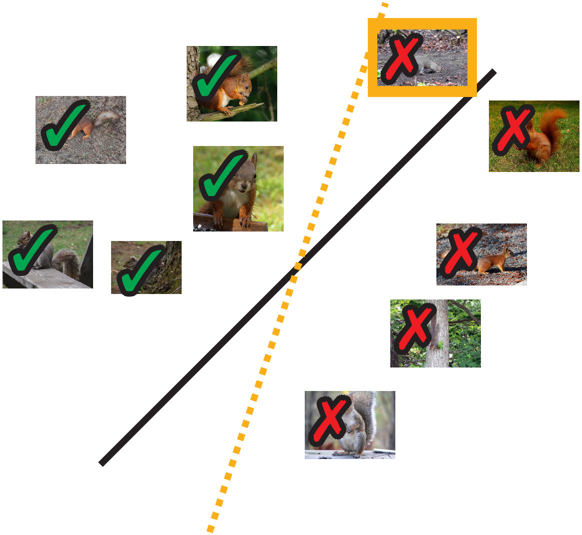
# A Story of Squirrels



# A Story of Squirrels



# A Story of Squirrels



# Differential Privacy [Dwork, 2006]

A **randomized algorithm** $\mathcal{A}^{\text{priv}} : (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n \to \mathcal{H}$ is $(\epsilon, \delta)$-private if for all **neighboring datasets** $D, D' \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n$ and all subsets of hypotheses $\mathcal{H}' \subseteq \mathcal{H}$

$$\mathbb{P}(\mathcal{A}^{\text{priv}}(D) \in \mathcal{H}') \leq \exp(\epsilon)\,\mathbb{P}(\mathcal{A}^{\text{priv}}(D') \in \mathcal{H}') + \delta$$

# Gaussian Mechanism

Compute the $\ell_2$-**sensitivity** of the algorithm $\mathcal{A}$:

$$\Delta(\mathcal{A}) = \sup_{D \approx D'} \left\| \mathcal{A}(D) - \mathcal{A}(D') \right\|_{\mathcal{H}}$$

$$\mathcal{A}(D) \bullet \xleftrightarrow[\Delta(\mathcal{A})]{\qquad\qquad} \bullet \mathcal{A}(D')$$

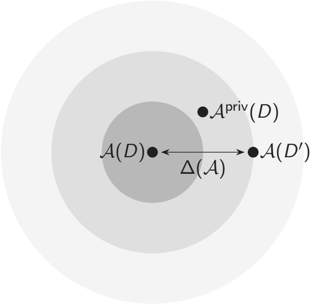
# Gaussian Mechanism

Compute the $\ell_2$-**sensitivity** of the algorithm $\mathcal{A}$:

$$\Delta(\mathcal{A}) = \sup_{D \approx D'} \left\| \mathcal{A}(D) - \mathcal{A}(D') \right\|_{\mathcal{H}}$$

Add noise to $\mathcal{A}(D)$, **calibrated to its sensitivity** and the desired **level of privacy**:

$$\mathcal{A}^{\text{priv}}(D) = \mathcal{A}(D) + \mathcal{N}\left(0, \frac{2\Delta(\mathcal{A})^2 \log(1.25/\delta)}{\epsilon^2} \mathbb{I}_p\right)$$



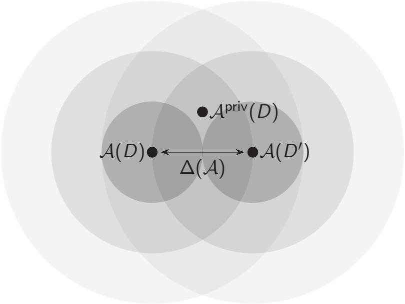
# Gaussian Mechanism

Compute the $\ell_2$-**sensitivity** of the algorithm $\mathcal{A}$:

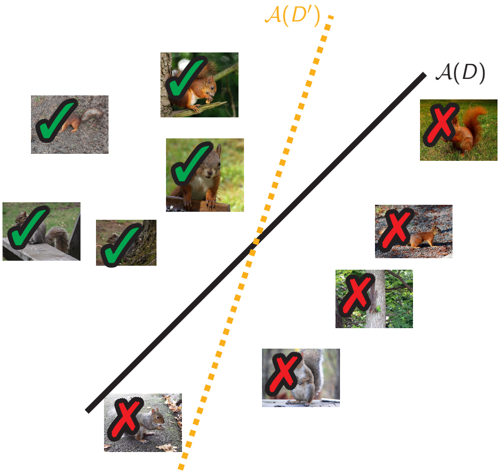
$$\Delta(\mathcal{A}) = \sup_{D \approx D'} \left\| \mathcal{A}(D) - \mathcal{A}(D') \right\|_{\mathcal{H}}$$

Add noise to $\mathcal{A}(D)$, **calibrated to its sensitivity** and the desired **level of privacy**:

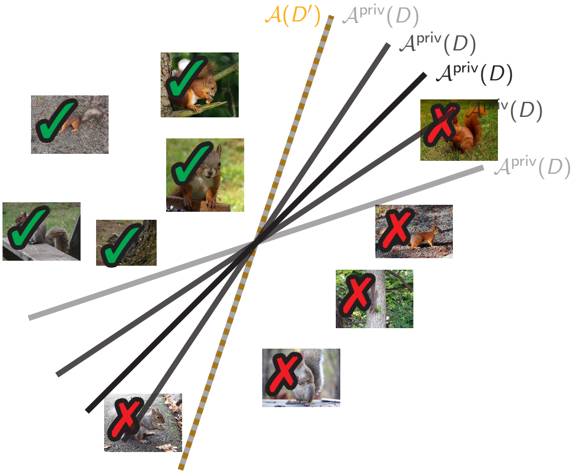
$$\mathcal{A}^{\text{priv}}(D) = \mathcal{A}(D) + \mathcal{N}\left(0, \frac{2\Delta(\mathcal{A})^2 \log(1.25/\delta)}{\epsilon^2} \mathbb{I}_p\right)$$



# A Story of Squirrels



# A Story of Squirrels



# A Story of Squirrels

# A Story of Squirrels


$$\mathbf{P}\left(H(X) = Y \mid S = \text{red}\right) = \frac{4}{5}$$

$$\mathbf{P}\left(H(X) = Y\right) = \frac{9}{10}$$

$$F_{\text{red}}(h, D) = -\frac{1}{10}$$

# A Story of Squirrels


$$\mathbf{P}\left(H(X) = Y \mid S = \text{red}\right) = \frac{4}{5}$$

$$\mathbf{P}\left(H(X) = Y\right) = \frac{9}{10}$$

$$F_{\text{red}}(h, D) = -\frac{1}{10}$$

What is the impact of differential privacy on fairness?

# Pointwise Lipschitzness of Accuracy Parity

$$\text{For all } s \in \mathcal{S},\ F_s(h, D) = \underbrace{\mathbf{P}\left(H(X) = Y | S = s\right)}_{\text{Accuracy of } h \text{ on the group } s} - \underbrace{\mathbf{P}\left(H(X) = Y\right)}_{\text{Accuracy of } h.}$$

Let $h, h' \in \mathcal{H}$ be two models

$$\left|F_s(h, D) - F_s(h', D)\right| \leq \chi_s(h, D) \left\|h - h'\right\|_{\mathcal{H}}$$

- $\chi_s(h, D) = \mathbb{E}\left(\frac{L_X}{|h(X)|}\right) + \mathbb{E}\left(\frac{L_X}{|h(X)|} \,\middle|\, S = s\right)$
- $L_X$ is a lipschitz constant of the model: $|h(X) - h'(X)| \leq L_X \left\|h - h'\right\|_{\mathcal{H}}$

# Pointwise Lipschitzness of Accuracy Parity

$$\text{For all } s \in \mathcal{S},\ F_s(h,D) = \underbrace{\mathbf{P}\left(H(X) = Y | S = s\right)}_{\text{Accuracy of } h \text{ on the group } s} - \underbrace{\mathbf{P}\left(H(X) = Y\right)}_{\text{Accuracy of } h.}$$

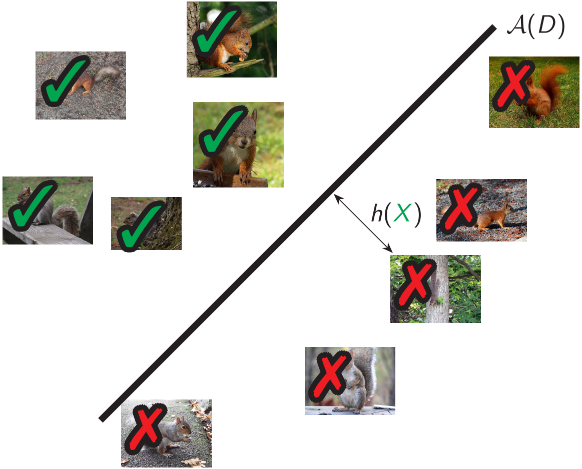
Let $h, h' \in \mathcal{H}$ be two models

$$\left|F_s(h,D) - F_s(h',D)\right| \leq \chi_s(h,D) \left\|h - h'\right\|_{\mathcal{H}}$$

- $\chi_s(h,D) = \mathbb{E}\left(\frac{L_X}{|h(X)|}\right) + \mathbb{E}\left(\frac{L_X}{|h(X)|} \,\middle|\, S = s\right)$
- $L_X$ is a lipschitz constant of the model: $|h(X) - h'(X)| \leq L_X \left\|h - h'\right\|_{\mathcal{H}}$

**Key quantity:** Ratio between $L_X$ ($\downarrow$) and the margin $|h(X)|$ ($\uparrow$).

# Pointwise Lipschitzness of Accuracy Parity



# Output Perturbation

$\mathcal{A}(D)$ is the following **optimization problem**

$$h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h; x_i, s_i, y_i)$$

with $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y} \to \mathbb{R}$ is $\Lambda$-**lipschitz**, and $\mu$-**strongly-convex**

$$\Delta(\mathcal{A}) = \frac{2\Lambda}{\mu n}$$

Using the **Gaussian Mechanism**, we can then release

$$h^{\text{priv}} = h^* + \mathcal{N}\left(\frac{8\Lambda^2 \log(1.25/\delta)}{\mu^2 n^2 \epsilon^2}\mathbb{I}_p\right)$$

# Output Perturbation

$\mathcal{A}(D)$ is the following **optimization problem**

$$h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h; x_i, s_i, y_i)$$

with $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y} \to \mathbb{R}$ is $\Lambda$-**lipschitz**, and $\mu$-**strongly-convex**

$$\Delta(\mathcal{A}) = \frac{2\Lambda}{\mu n}$$

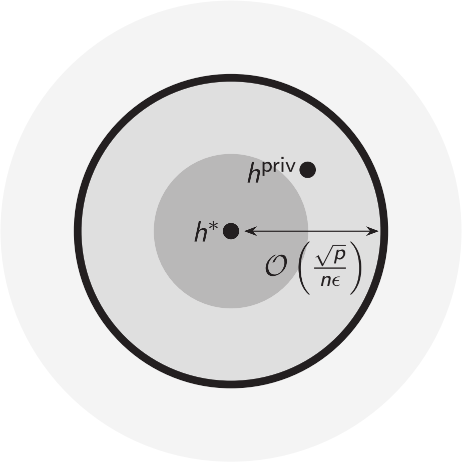
Using the **Gaussian Mechanism**, we can then release

$$h^{\text{priv}} = h^* + \mathcal{N}\left(\frac{8\Lambda^2 \log(1.25/\delta)}{\mu^2 n^2 \epsilon^2} \mathbb{I}_p\right)$$

**Key result:** With probability at least $1 - \zeta$, $\left\| h^{\text{priv}} - h^* \right\|_2 \leq \frac{\Lambda \sqrt{32 p \log(1.25/\delta) \log(2/\zeta)}}{\mu n \epsilon}$.

# Bounded Distance

$$\text{With probability at least } 1-\zeta,\ \left\|h^{\text{priv}} - h^*\right\|_2 \leq \mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right).$$



# Main result

- $|F_s(h, D) - F_s(h', D)| \leq \chi_s(h, D) \|h - h'\|_{\mathcal{H}}$
- With probability $1 - \zeta$, $\left\|h^{\text{priv}} - h^*\right\|_2 \leq \mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)$

$$\boxed{\text{With probability } 1 - \zeta,\ \left|F_s(h^{\text{priv}}, D) - F_s(h^*, D)\right| \leq \chi_s(h^{\text{priv}}, D)\mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)}$$

# Main result

- $\left|F_s(h, D) - F_s(h', D)\right| \leq \chi_s(h, D) \left\|h - h'\right\|_{\mathcal{H}}$
- With probability $1 - \zeta$, $\left\|h^{\text{priv}} - h^*\right\|_2 \leq \mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)$

$$\boxed{\text{With probability } 1 - \zeta,\ \left|F_s(h^{\text{priv}}, D) - F_s(h^*, D)\right| \leq \chi_s(h^{\text{priv}}, D)\mathcal{O}\left(\frac{\sqrt{p}}{n\epsilon}\right)}$$

**Auditing:** The right hand side only depends on the private model!

# Conclusion

The story so far...

- Accuracy Parity is **pointwise lipschitz** and **margins** are key quantities.
- Using output perturbation the **private and non-private models are close**.
- Differential privacy has **bounded impact** on fairness.
- The **squirrels are safe** if they learned their model on **sufficiently many examples**.

# Conclusion

The story so far...

- Accuracy Parity is **pointwise lipschitz** and **margins** are key quantities.
- Using output perturbation the **private and non-private models are close**.
- Differential privacy has **bounded impact** on fairness.
- The **squirrels are safe** if they learned their model on **sufficiently many examples**.

but also...

- Similar results for **other fairness measures** (e.g. Demographic Parity (with binary labels) [Calders et al., 2009], Equality of Opportunity [Hardt et al., 2016], Equalized Odds [Hardt et al., 2016]) and for **Accuracy**.
- Multi-class, multi-groups problems, DP-SGD, **tighter but harder to parse bounds**.
- A few **experiments** to check the tightness of our bounds.
- A **finite sample analysis** showing that our results also hold in **generalization**.

You can fetch the paper on arXiv!

# References I

All animal images were taken from `commons.wikimedia.org`. No animals were harmed during the preparation of this presentation.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

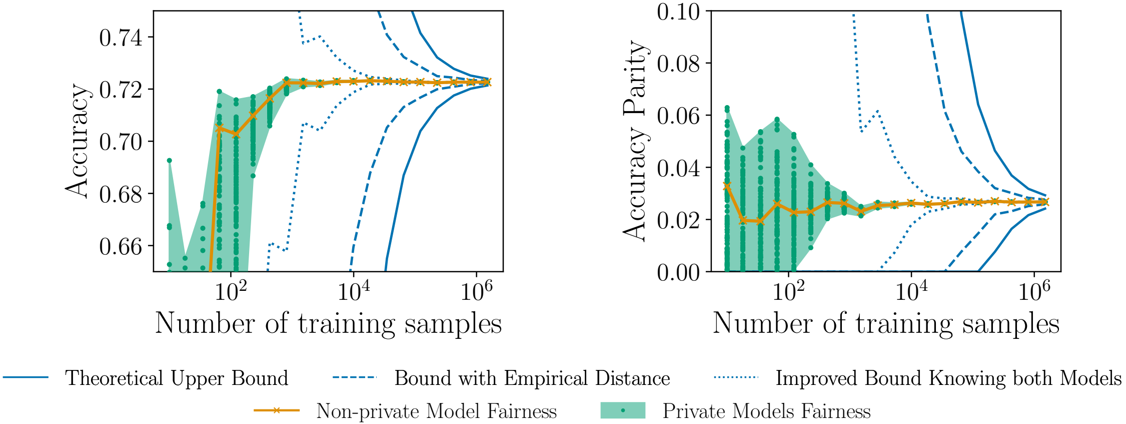
C. Dwork. Differential privacy. In *Encyclopedia of Cryptography and Security*, 2006.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

# Experiments

Folktables [Ding et al., 2021], $\ell_2$-regularized logistic regression, $(\epsilon, \delta) = (1, \frac{1}{n^2})$



# Experiments

Folktables [Ding et al., 2021], $\ell_2$-regularized logistic regression, $\delta = \frac{1}{n^2}$, $n = 1,498,050$