

# The impact of Record Linkage on Learning from Feature Partitioned Data



**Richard  
Nock**  
Google  
Research



**Stephen  
Hardy**  
Ambiata



**Wilko  
Henecka**  
Ambiata



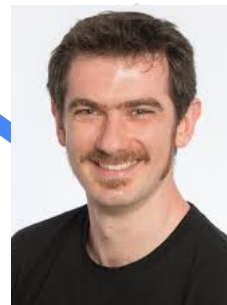
**Hamish  
Ivey-Law**  
The Australian National  
University



**Jakub  
Nabaglo**



**Giorgio  
Patrini**  
Sensity




**Guillaume  
Smith**  
Ambiata



**Brian  
Thorne**  
Hardbyte

# Setting, problem & questions addressed in this paper

- **Overarching setting:** batch supervised learning, use labeled sample to learn classifier
  - Example: (**Medical+Pharma, Class**) — single *peer* holds all data
- **Us: vertical partition (VP):** *features* split among 2 peers, **M** and **P** (1+, e.g. **P**, holds label **C**)
  - **Record linkage (RL)** needed before batch supervised learning: 
- **Problem:** 100s of RL approaches, little / no understanding on how this impacts ML — crucial problem at the “age” of federated learning
- **Questions:** what are RL parameters that impact ML ? How can RL improve ML ? Can RL “as usual” be non-detrimental / beneficial to ML ? Can RL be improved — if so, how — as preprocessing step to ML ? Impact of such results on related ML settings (e.g. federated ML) ? Algorithms / experimental results for various RL + ML settings ?

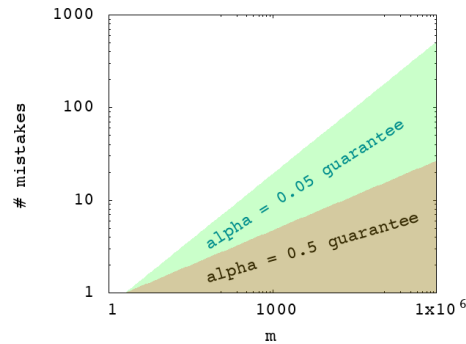
# Theoretical results, summary

- Linear models: sufficient conditions to get on training:

$$\frac{\|\tilde{\theta}^* - \theta^*\|_2}{\|\theta^*\|_2} \leq O\left(\frac{1}{m^\alpha}\right)$$

Optimal ML | RL
Optimal ML

- Regularisation is key (for broad set of losses)
- A small # of RL mistakes does not impact ML
- Optimisation of RL prior to ML, e.g. minimising between-classes RL errors
- Large margins on  $\theta^*$   $\implies$  right class on  $\tilde{\theta}^*$  (“RL immunity for large margins”)
- Results hold in the small data regime



## Experimental results, summary

- Baseline settings include case where one peer does not hold class or a noisy estimate
- Simple approaches to complete / correct label prior to RL can offer leverage
- Potential (estimated) **dials** to evaluate the value of RL prior to ML
  - inaccuracy between matched observations
  - error between classes for matched observations
- Margin immunity observed experimentally:
  - strong advocacy for distributed / federated ML
  - may offer further cheap dials to evaluate the ML “potential” of datasets

## Conclusion

- **Caution:** experiments on *simulated* RL env. (to control / compute all params)
- Some work to do thoroughly evaluate the RL + ML dials
- Yet our theory builds a sound justification to do RL + ML (margin immunity), clearly observed in our experiments, & clues on how to organise the “zoo” of RL techs as preprocessing for ML
- Our results can be extended to more losses, albeit giving more “qualitative” results (interestingly, better bounds for proper losses)
- Our results can be extended to classifiers more “complex” than linear, but risk of loose bounds

A decorative graphic in the top-left corner consisting of several concentric blue arcs and a dotted line with five blue dots.

Thank you !