

Data-SUITE

Data-centric identification of in-distribution
incongruous examples

Nabeel Seedat, Jonathan Crabbé &
Mihaela van der Schaar



van_der_Schaar
\ LAB

vanderschaar-lab.com



UNIVERSITY OF
CAMBRIDGE



ns741@cam.ac.uk

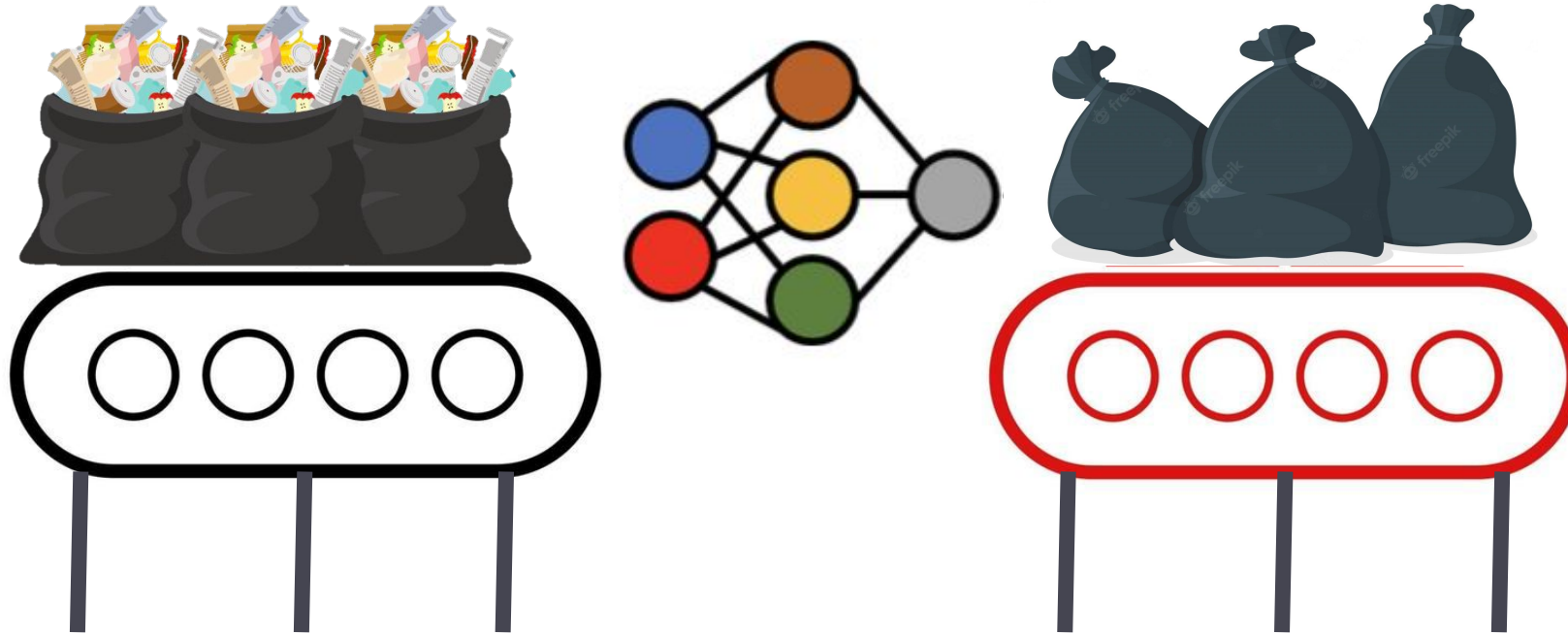


[@NabeelSeedat01](https://twitter.com/NabeelSeedat01)



linkedin.com/in/nabeel-seedat

Data Quality Matters



Target In-Distribution

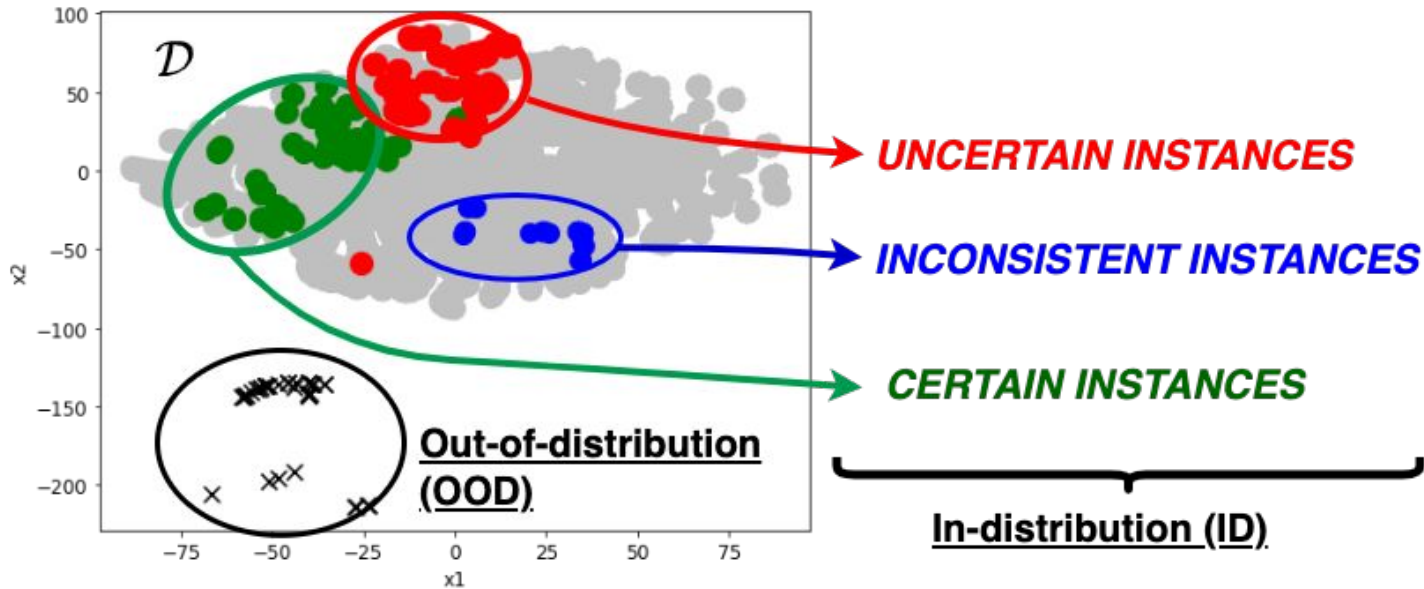
**Literature: Out-of-distribution
(OOD)**

**How can we assess In-Distribution data,
with feature space heterogeneity?**

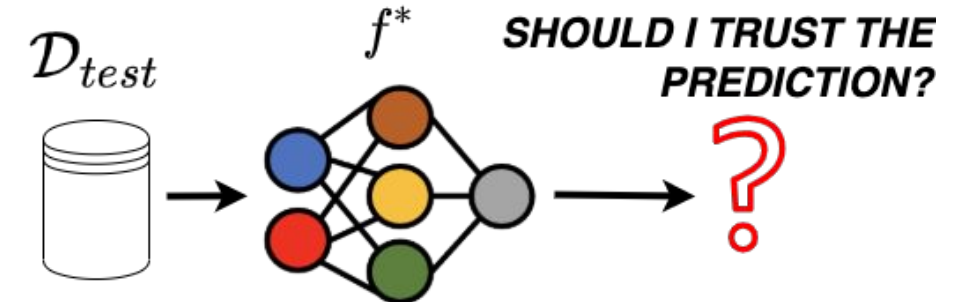


Data scientists - Alice & Bob need help

(a) ALICE
Data Exploration

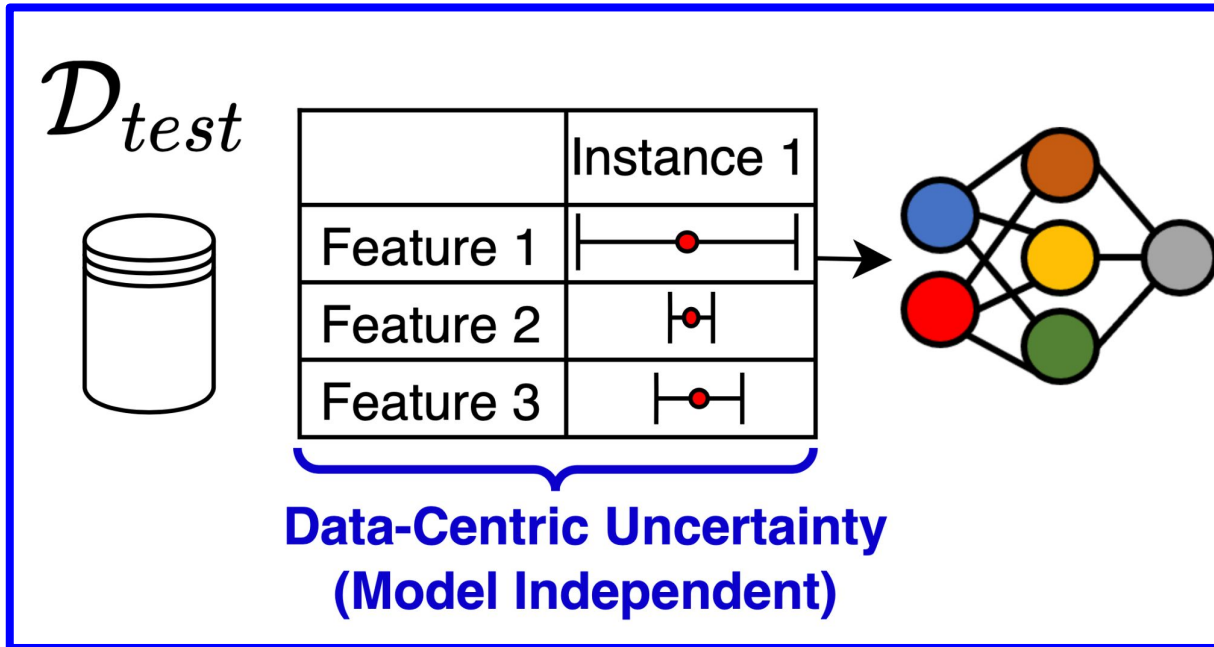


(b) BOB
Model Deployment

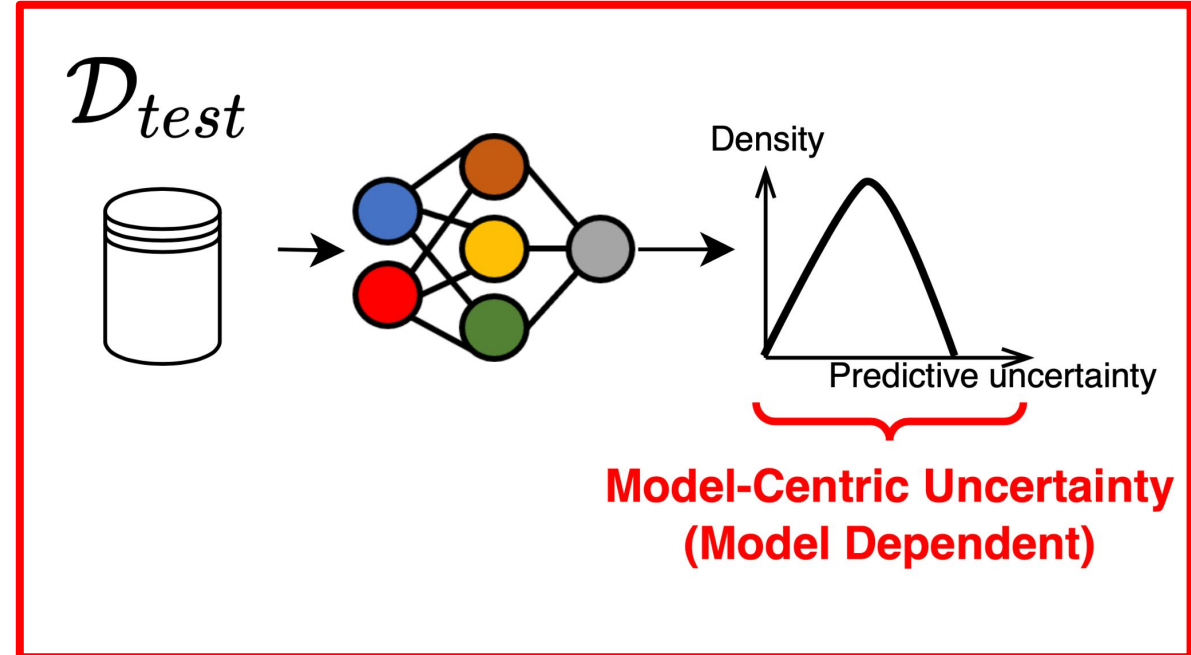


Data-SUITE as a paradigm shift

Data-SUITE



Current Approaches



Principled Goals

Build feature-wise confidence interval estimators

Characterize new test instances based on a set of training instances.

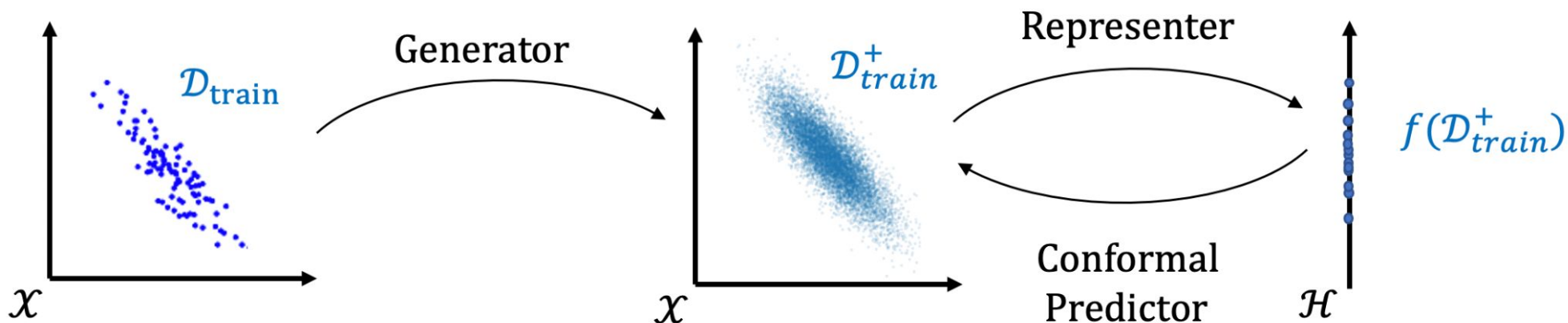
CI Properties:

1. Coverage guarantees
2. Feature-wise intervals
3. Instance-wise adaptive
4. Decoupled from the downstream model

	Instance 1
Feature 1	
Feature 2	
Feature 3	

**Data-Centric Uncertainty
(Model Independent)**

Data-SUITE pipeline



	Feature 1	Feature 2
Sample 1		
Sample 2		

**UNCERTAIN
INSTANCES
(Large CIs)**

	Feature 1	Feature 2
Sample 1		
Sample 2		

**INCONSISTENT
INSTANCES
(Outside CI)**



Data-SUITE enables reliable model deployment

**[CONFIDENT USAGE] 1- α coverage guarantee
Theoretic + Experimental**

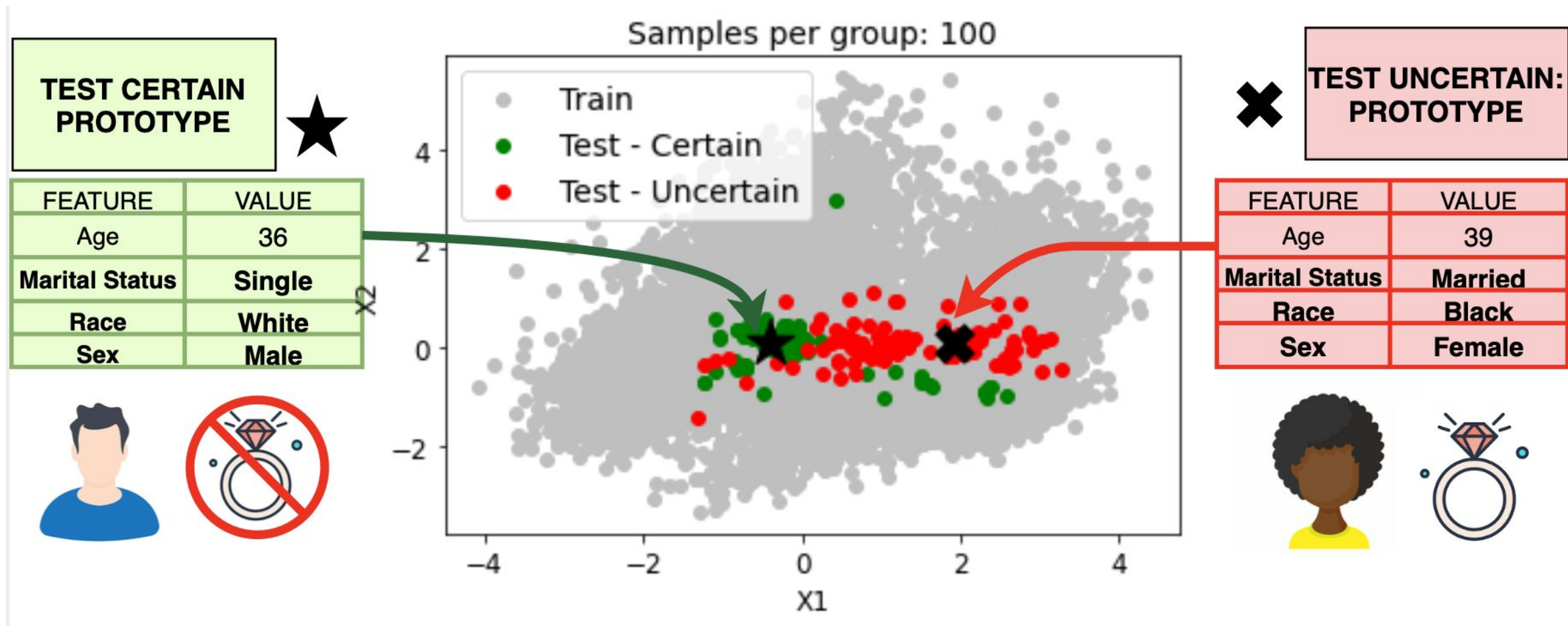
IMPROVED DOWNSTREAM PERFORMANCE

Data-SUITE's stratification improves performance vs model-centric counterparts

MODEL INDEPENDENT

Data-SUITE consistently finds impactful instances for diverse downstream models

Data-SUITE enables Insightful Data Exploration



Contributions

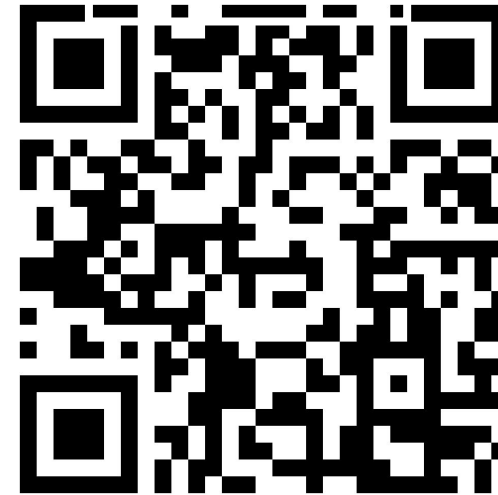
- Tackle the under-studied problem of in-distribution heterogeneity
- Data-SUITE takes a data-centric paradigm shift vs conventional model-centric approaches: outperforming them
- Identify impactful instances **INDEPENDENT** of the downstream predictive model
- Useful tool for practitioners
- Confident usage: rigorous theoretical guarantees



PAPER



CODE



Data-SUITE: Data-centric identification of in-distribution incongruous examples

Nabeel Seedat, Jonathan Crabbé & Mihaela van der Schaar

Thanks to:



van_der_Schaar
\ LAB

vanderschaar-lab.com



UNIVERSITY OF
CAMBRIDGE



@NabeelSeedat01



[linkedin.com/in/nabeel-seedat](https://www.linkedin.com/in/nabeel-seedat)