MACHINE LEARNING WITH LIMITED LABEL AVAILABILITY

ALGORITHMS AND APPLICATIONS

Flavio Giobergia
XXXV cycle

Politecnico di Torino
February 15, 2023

Advisor
Elena Baralis

Doctoral Examination Committee
Sara Comai, Referee, *Politecnico di Milano*
Dino Ienco, Referee, *INRAE*
Rosa Meo, *Università degli studi di Torino*
Genoveva Vargas-Solar, *CNRS*
Silvia Chiusano, *Politecnico di Torino*

Politecnico di Torino
SmartData@PoliTO

LIMITED LABEL LEARNING – WHY?

- For *practitioners*: there is an economic incentive to reduce the amount of labelled data needed
- For *researchers*: the goal of AI is to build models with human-level learning capabilities
  - Humans learn from 5 samples, not 5,000!

LIMITED LABELS LEARNING APPROACHES

	Unsupervised learning	Semi-supervised learning	Active learning	Domain adaptation
Scope	• No labels available	• Limited labelled data, unlabelled data often abundant	• Few labels provided by an oracle, based on model's queries	• Labelled data for other domains, no labelled data for target domain
Main properties	• Learn cluster membership • Learn feature representation • Find recurring patterns in data	• Build model on labelled + unlabelled data, infer missing labels (inductive) • Infer new labels based on properties of known points (transductive)	• Definition of a query policy (when does the model request new labels?) • Model identifies regions of input space of low confidence	• Supervised learning on resource-rich domain • Transfer technique to propagate knowledge to target domain

LIMITED LABELS LEARNING APPROACHES: MAIN CONTRIBUTIONS

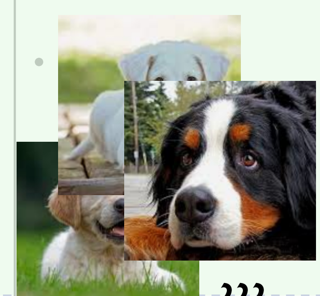
	Unsupervised learning	Semi-supervised learning	Active learning	Domain adaptation
Scope	2-step training of self-organizing maps	Explicit confidence-based FixMatch	• Few labels provided by an oracle, based on model's queries	Cross-lingual propagation of sentiment information
Main properties			• Definition of a query policy (when does the model request new labels?) • Model identifies regions of input space of low confidence	

UNSUPERVISED LEARNING

	Unsupervised learning	Semi-supervised learning	A	Domain adaptation
Scope	• No labels available	• Limited labelled data, unlabelled data often abundant	• I p o model's queries	• Labelled data for other domains, no labelled data for target domain
Main properties	• Learn cluster membership • Learn feature representation • Find recurring patterns in data	• Build model on labelled + unlabelled data, infer missing labels (inductive) • Infer r based proper known (transduct	• Definition of a query policy (when does the model requ new labels?) Model identifies regions of input space of low confidence	• Supervised learning on resource-rich domain •

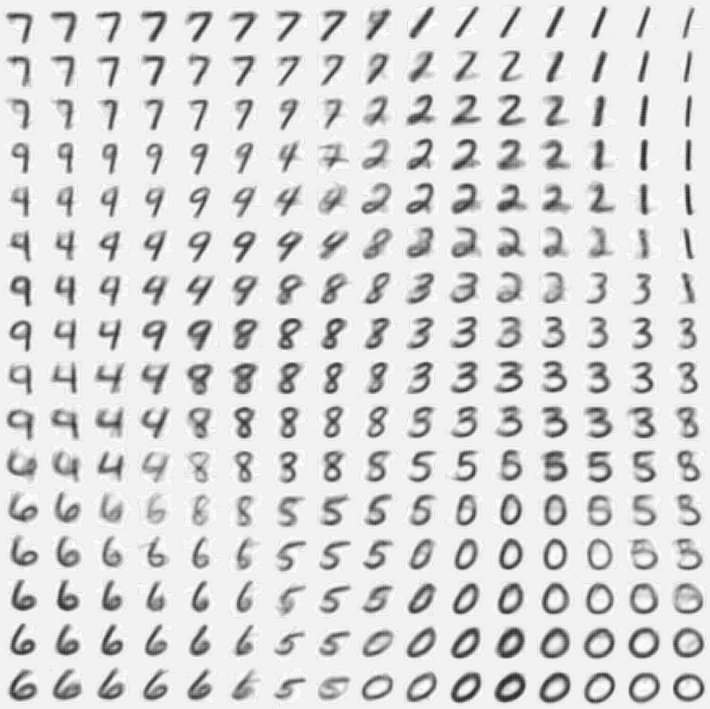

???


???


???

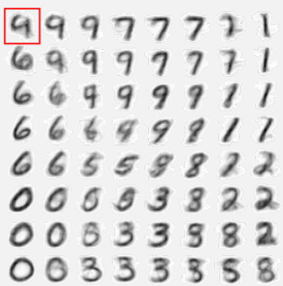
SELF-ORGANIZING MAPS

- SOMs are unsupervised neural networks
  - Producing low-dimensional representations of high-dimensional data
- During training, SOM weights are iteratively updated to resemble inputs in a dataset
- After the training, the SOM has learned:
  - what the inputs "look like"
  - a notion of similarity among digits
- For example, "7" is close to "1", far away from "0"
- Expressive power of SOM depends on its granularity!

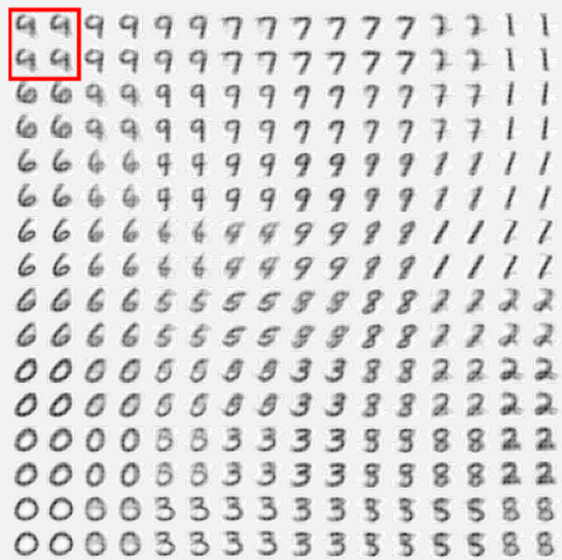


2-STEP TRAINING

exploration

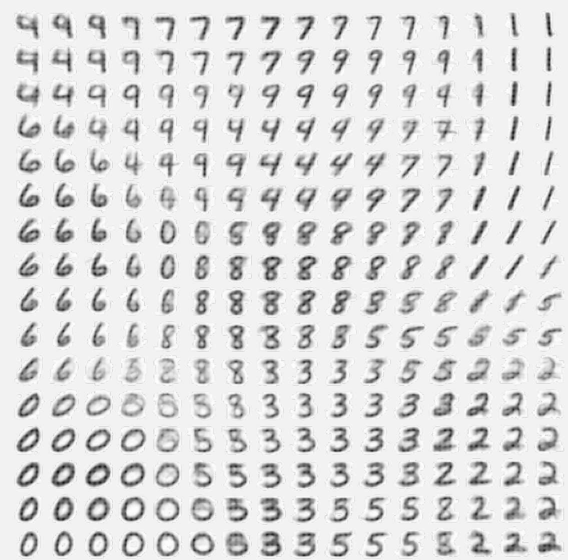


(1) Train a small SOM on a fraction of dataset



Large SOM with nodes replicated from the small SOM

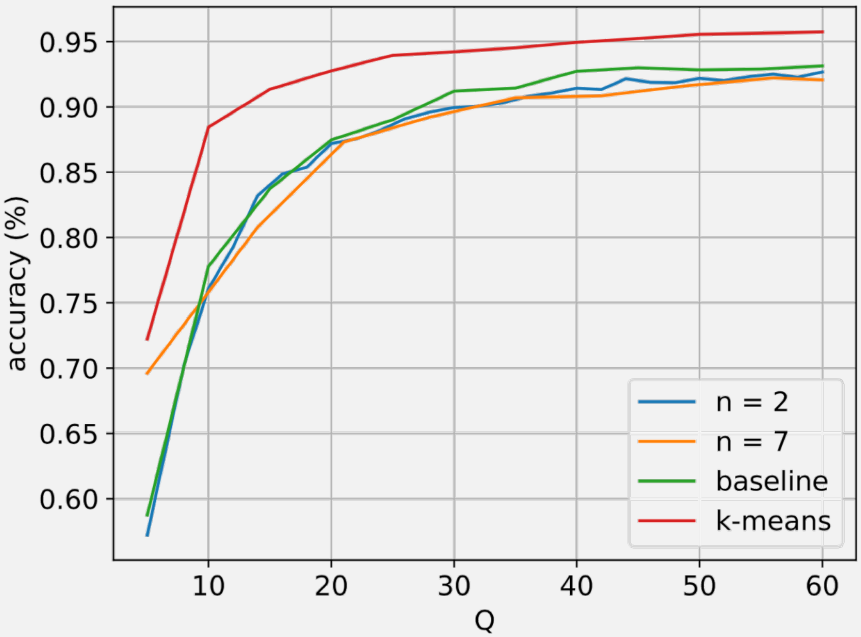
tuning

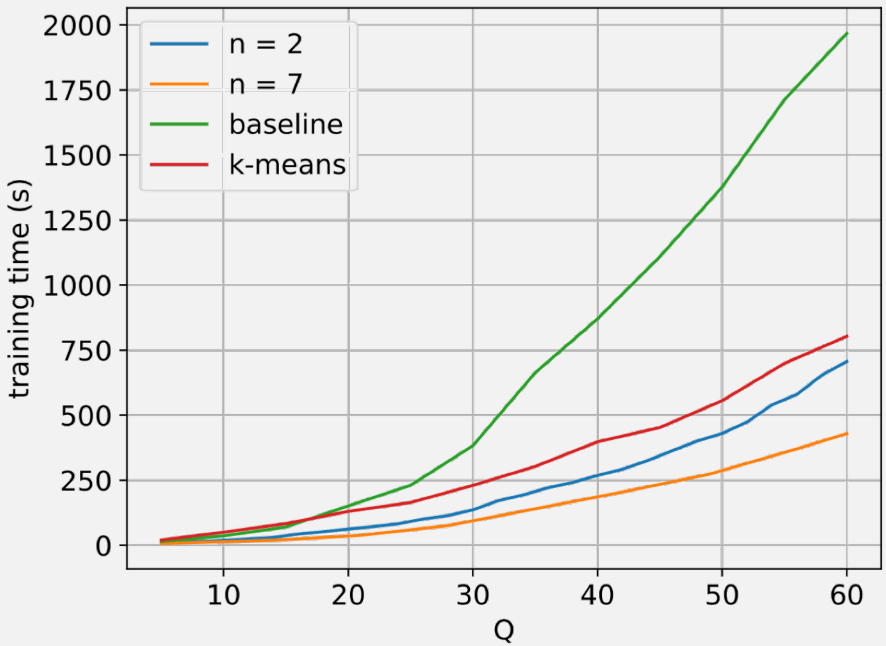


(2) Fine tuning the weights of the large SOM on remainder of dataset

Giobergia, F., & Baralis, E. (2019, December). Fast Self-Organizing Maps Training. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 2257-2266). IEEE.

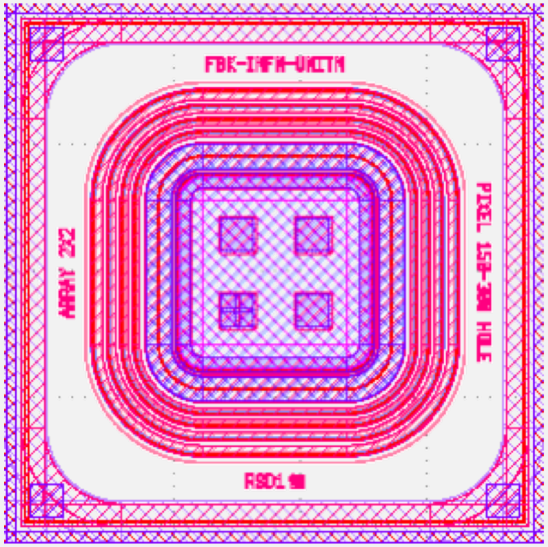
PERFORMANCE DEGRADATION & TIME IMPROVEMENTS





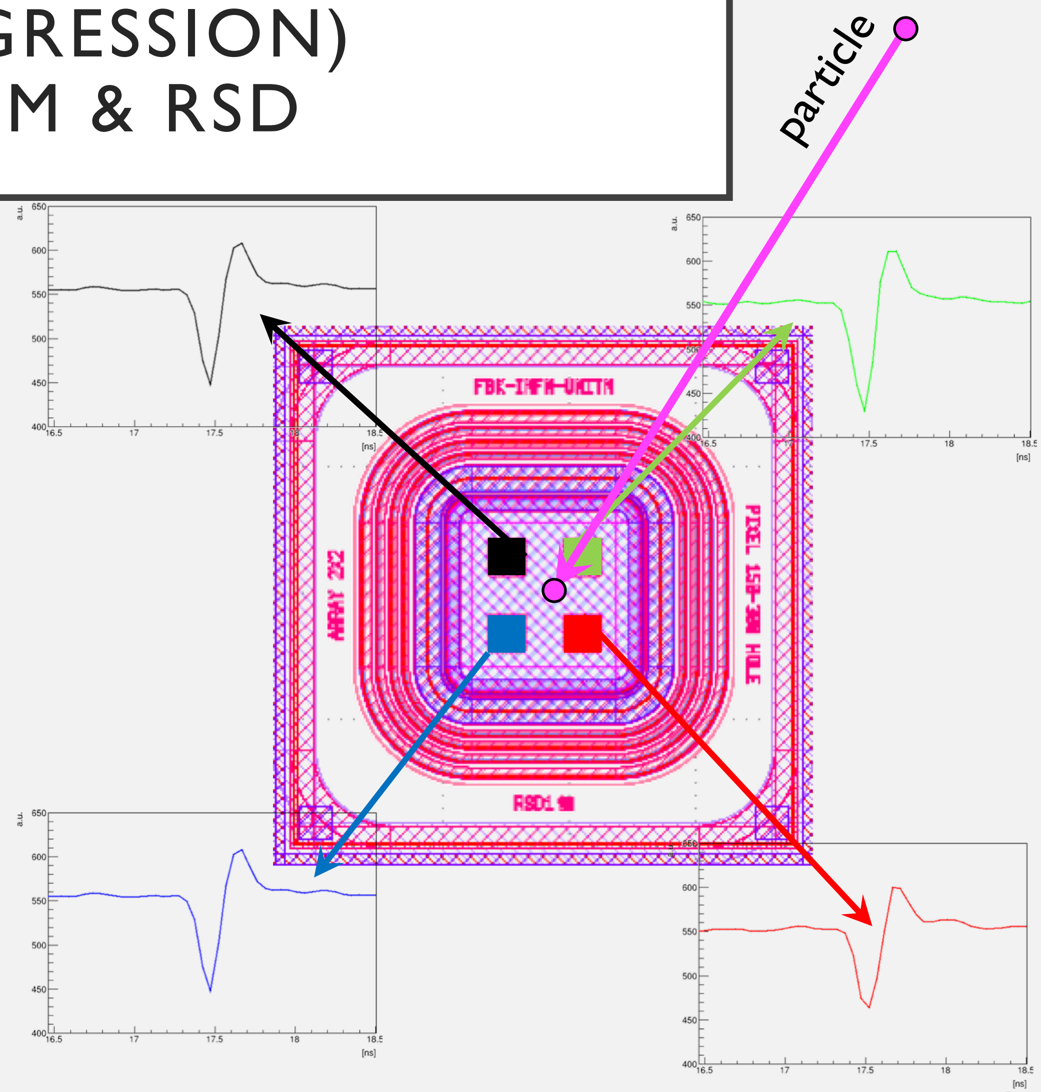
(DIGRESSION) SOM & RSD

- Collaboration with UniTO
  - Dept. of Physics
- ML applied to RSD sensors
  - Resistive Silicon Detectors
  - From signals to particle position
    - (+ time, as a next step!)



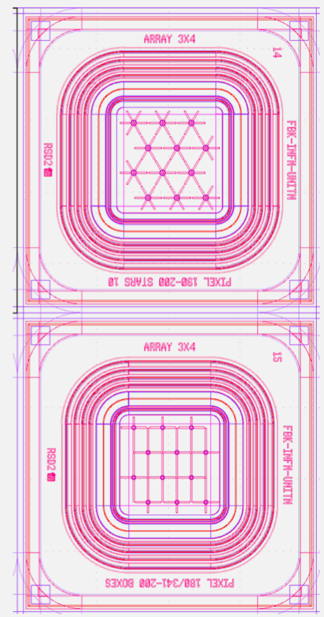
(DIGRESSION) SOM & RSD

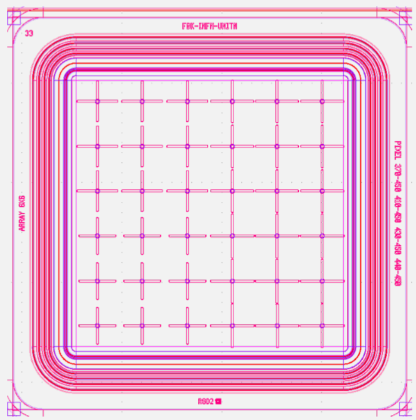
- Collaboration with UniTO
  - Dept. of Physics
- ML applied to RSD sensors
  - Resistive Silicon Detectors
  - From signals to particle position
    - (+ time, as a next step!)



PRELIMINARY RESULTS

- From signals to (x, y) coordinates
  - Multi-output regression problem
  - Data from experimental sessions
- Approaches:
  - Feature extraction + tree-based approaches
  - Feature extraction + FCNN
- Multiple sensors studied
  - Spatial resolution ~ 5 μm



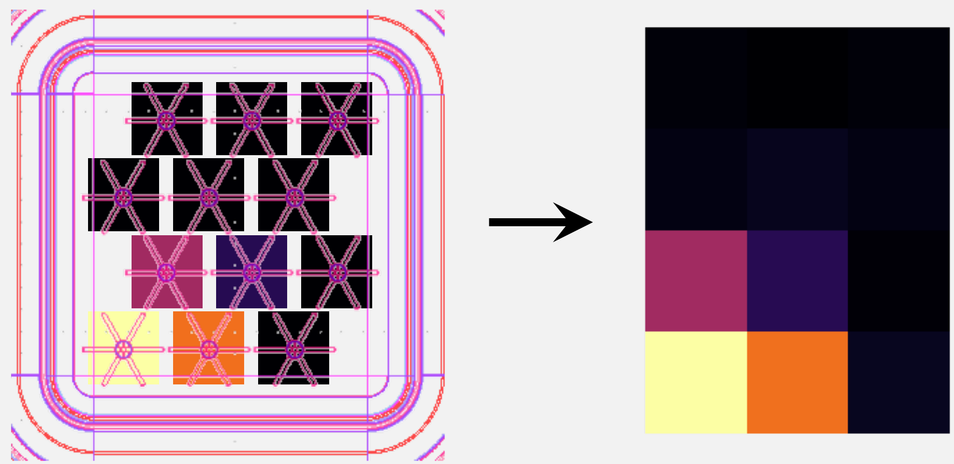


Siviero, F., **Giobergia, F.**, Menzio, L., Miserocchi, F., Tornago, M., Arcidiacono, R., ... & Sola, V. (2022). First experimental results of the spatial resolution of RSD pad arrays read out with a 16-ch board. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 1041*, 167313.

Tornago, M., **Giobergia, F.**, Menzio, L., Siviero, F., Arcidiacono, R., Cartiglia, N., ... & Sola, V. (2023). Silicon sensors with resistive read-out: Machine Learning techniques for ultimate spatial resolution. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 1047*, 167816.

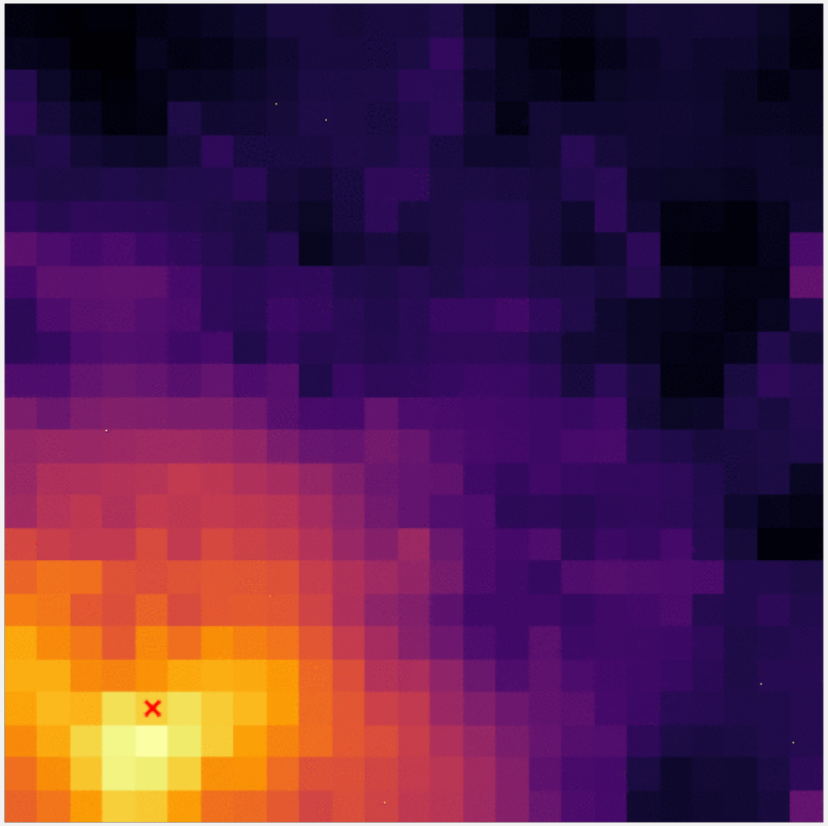
EVENTS TO (COARSE) IMAGES

- Each event (particle passage) can be seen as a coarse image
  - 3x3, 4x3, and similarly low resolutions
- Image-based approaches come to mind…
  - But what good are these low-res images?
  - Can we enhance them?

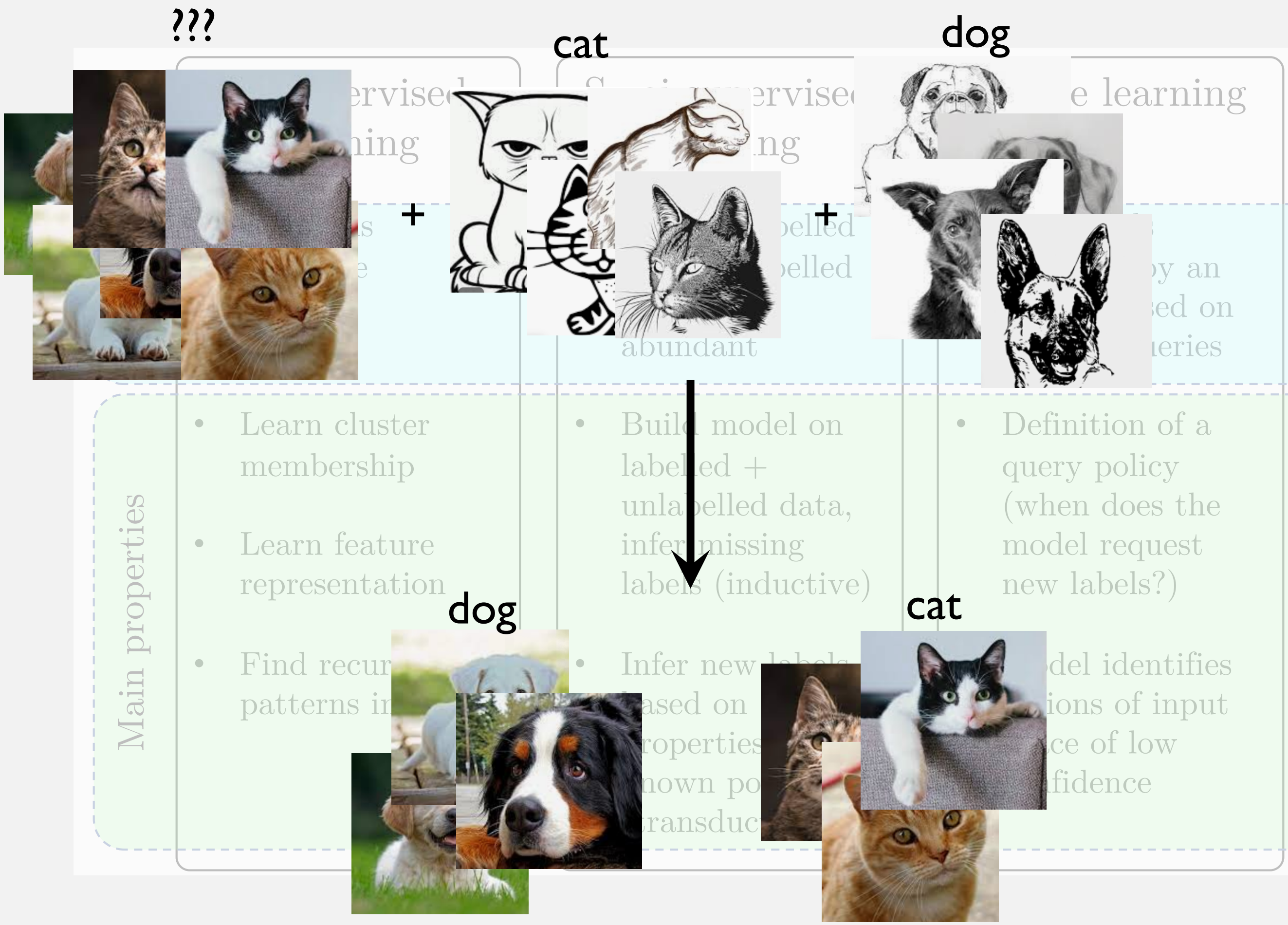


SOM-BASED ENHANCING

- Train a supervised SOM
- Active (winning) neurons known in advance
  - Ground truth coordinates of each event
- Images obtained as activation maps
- Larger SOM = more fine-grained images
  - 25x25, 50x50, 100x100 ➔
  - Suitable candidate for 2-step training!



DOMAIN ADAPTATION



Domain adaptation

- Labelled data for other domains, no labelled data for target domain

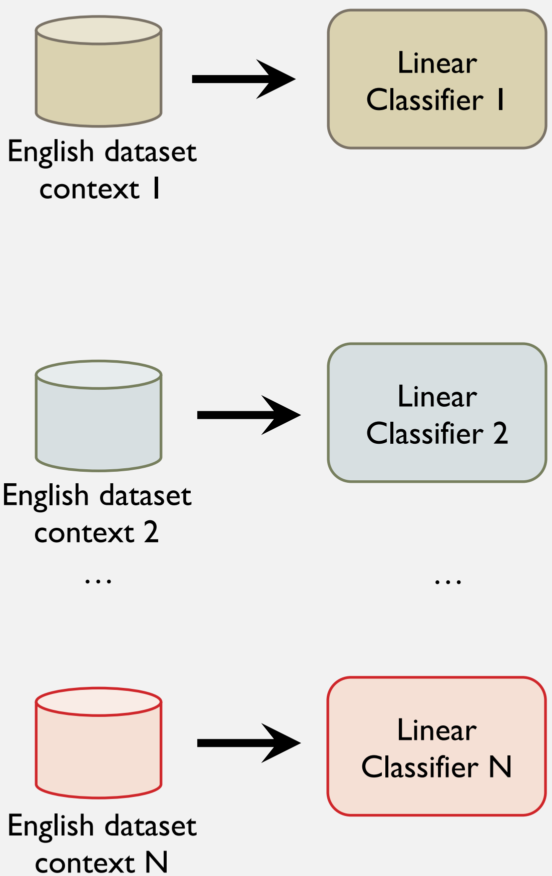
- Supervised learning on resource-rich domain
- Transfer technique to propagate knowledge to target domain

CROSS-LINGUAL PROPAGATION

- In NLP there is a long-standing resource availability problem:
  - English vs other languages
- Cross-lingual propagation approaches:
  - Learn from English
  - Propagate learned notions to other languages
- Dong and de Melo, 2018 introduce cross-lingual sentiment embeddings
  - words mapped to latent vectors based on sentiment
  - Vectors learned in English (high-resource domain), propagated to other languages (low-resource domains)

Dong, X. and De Melo, G., 2018, April. Cross-lingual propagation for deep sentiment analysis. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

SENTIMENT EMBEDDINGS INFERENCE



Word 1	0.9
Word 2	0.1
Word 3	-0.5
…	…
Word M	0.2

Word 1	0.5
Word 2	0.4
Word 3	-0.1
…	…
Word M	0.7

…

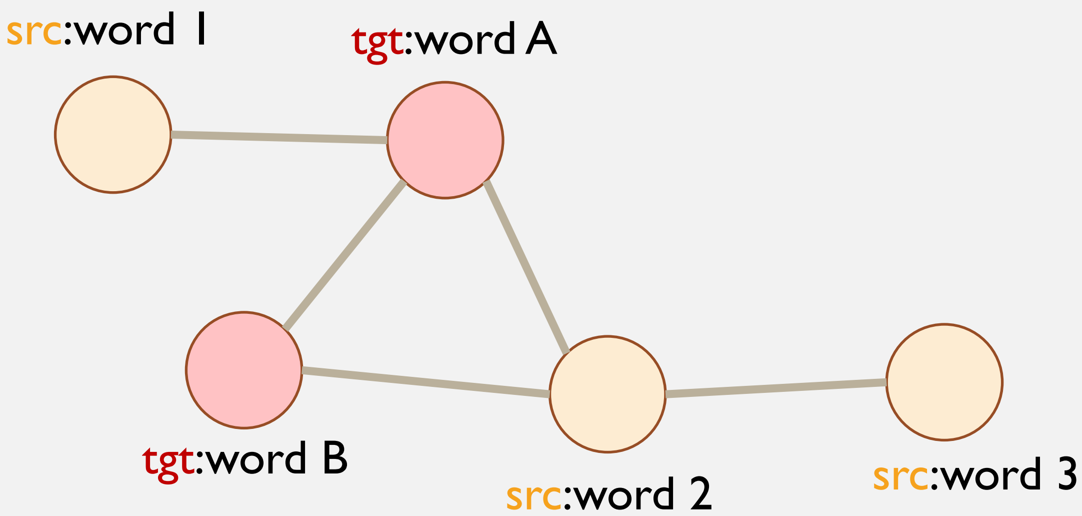
Word 1	0.1
Word 2	0.9
Word 3	0.3
…	…
Word M	-0.5

	Context 1	Context 2	…	Context N
Word 1	0.9	0.5	…	0.1
Word 2	0.1	0.4	…	0.9
Word 3	-0.5	-0.1	…	0.3
…	…	…	…	…
Word M	0.2	0.7	…	-0.5

SENTIMENT PROPAGATION (LEXICON-BASED WORD GRAPH)

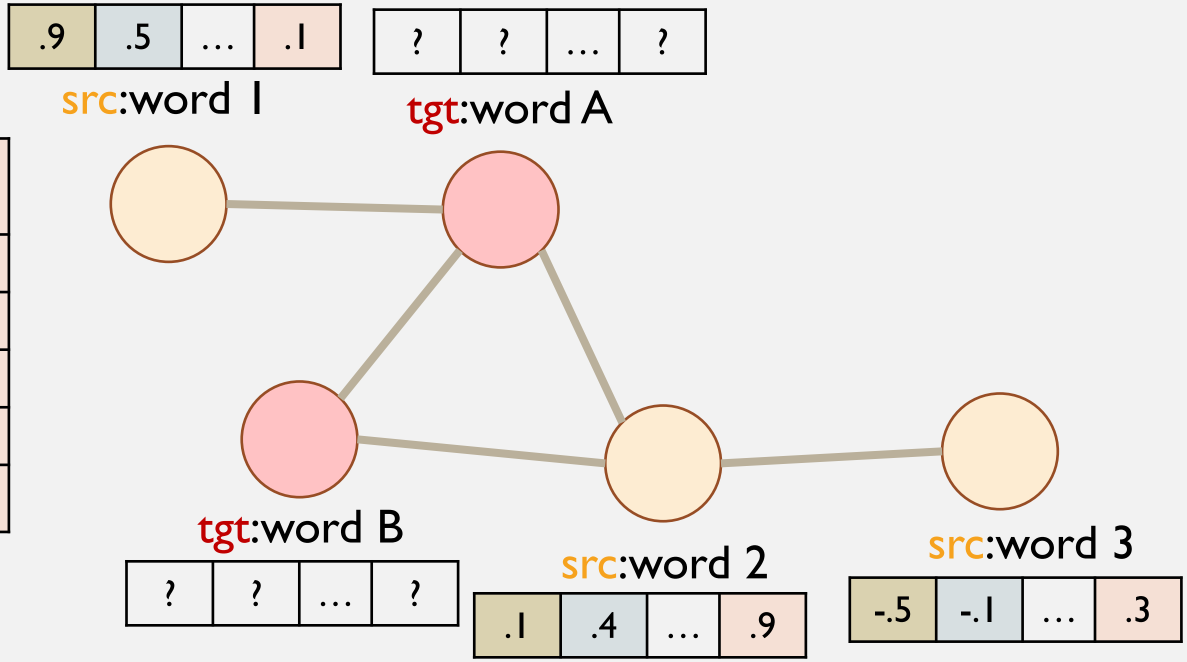
Lexicon
src:word 1 → tgt:word A
src:word 2 → src:word 3
tgt:word A → tgt:word B
src:word 2 → tgt:word B
src:word 2 → tgt:word A

src: Source language (e.g. English)
tgt: Target language (e.g. Italian)



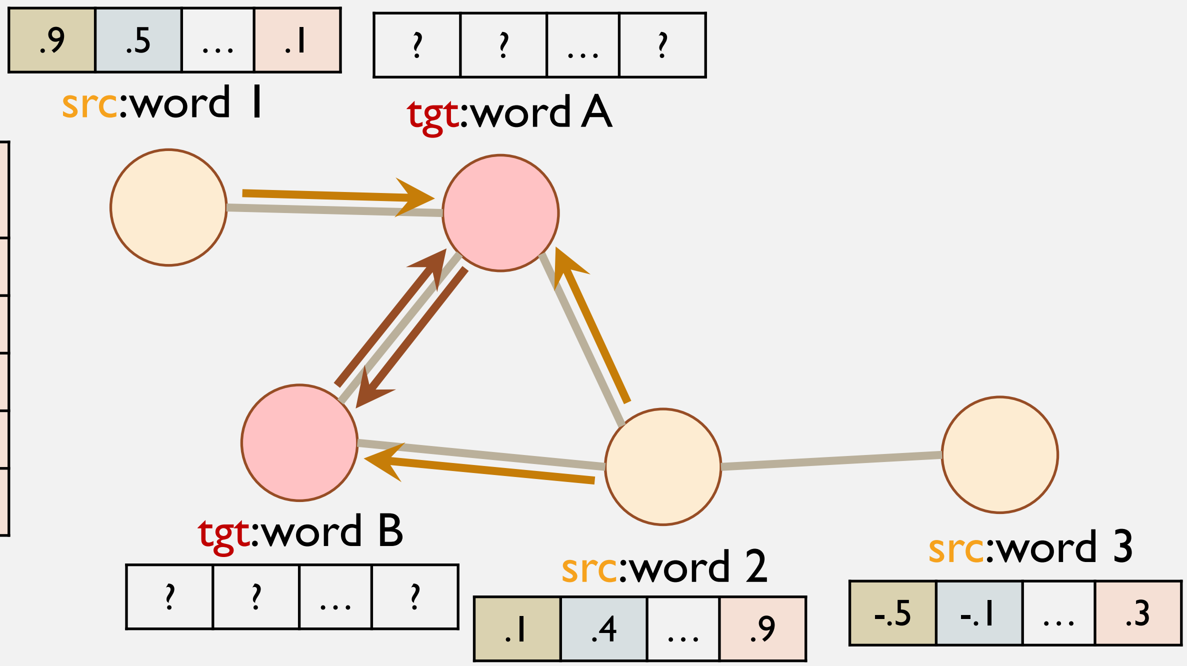
SENTIMENT PROPAGATION (VECTOR INITIALIZATIONS)

	Context 1	Context 2	…	Context N
Word 1	0.9	0.5	…	0.1
Word 2	0.1	0.4	…	0.9
Word 3	-0.5	-0.1	…	0.3
…	…	…	…	…
Word M	0.2	0.7	…	-0.5



SENTIMENT PROPAGATION (GRADIENT DESCENT)

	Context 1	Context 2	…	Context N
Word 1	0.9	0.5	…	0.1
Word 2	0.1	0.4	…	0.9
Word 3	-0.5	-0.1	…	0.3
…	…	…	…	…
Word M	0.2	0.7	…	-0.5



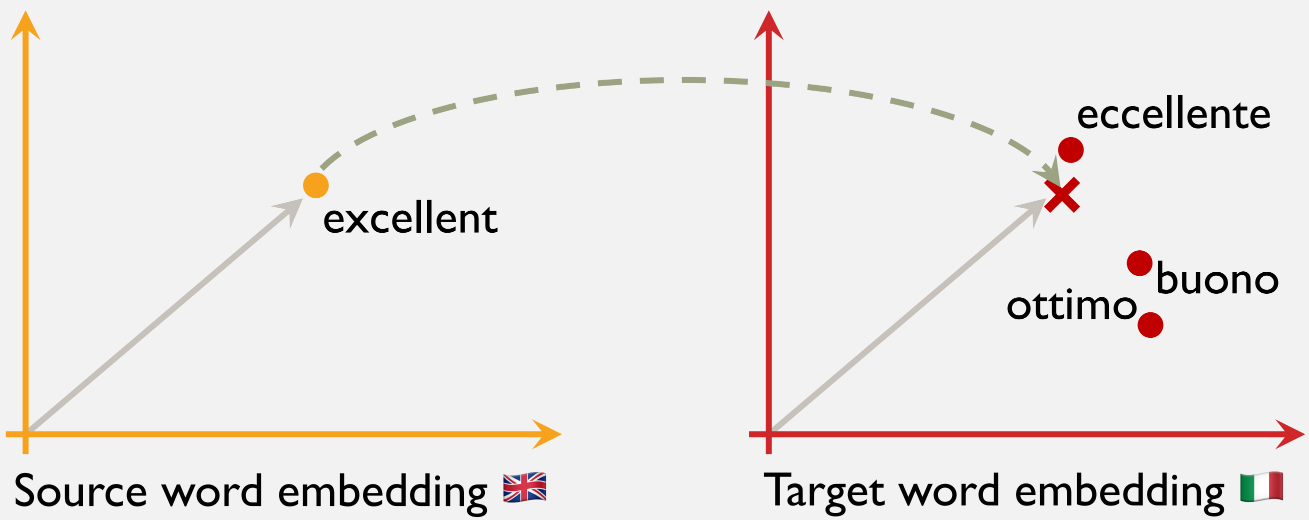
SOME PROBLEMS

- The approach requires a lexicon!
  - source language⇔ target language
  - Lexicons are hard to obtain
    - Particularly for less commonly spoken languages
  - The lexicon may not be exhaustive
    - Many translations may not be explicitly included
- Loss function poorly defined
  - Minimization cannot converge!

Lexicon
src:word 1 → tgt:word A
src:word 2 → src:word 3
tgt:word A → tgt:word B
src:word 2 → tgt:word B
src:word 2 → tgt:word A

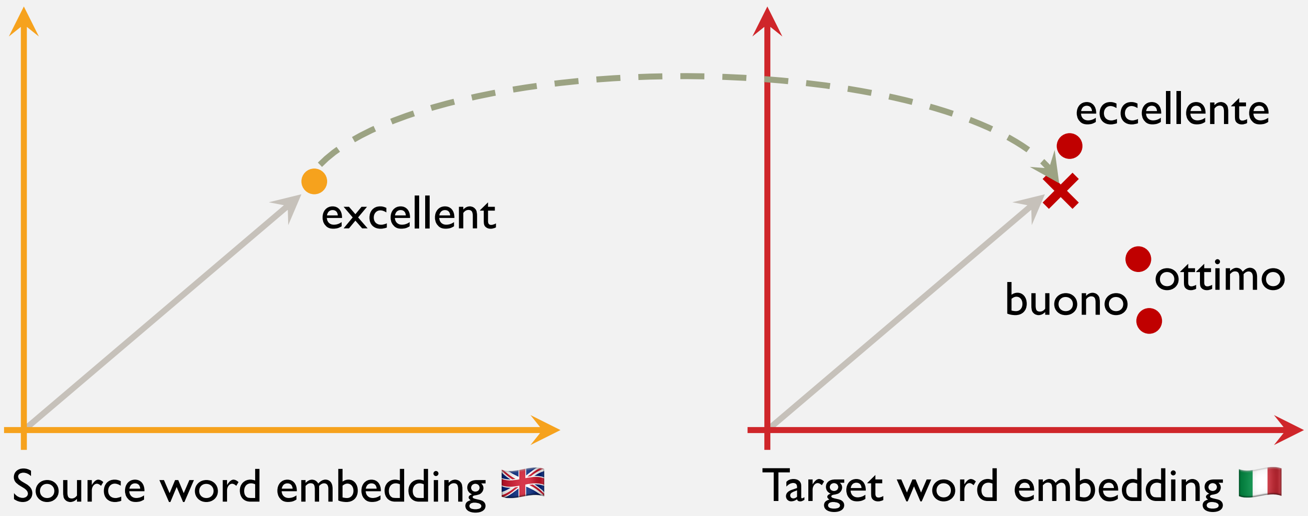
ALIGNED WORD EMBEDDINGS-BASED WORD GRAPH

- Use aligned word embeddings to build the graph
  - No need for a lexicon (only aligned word embeddings)
  - Automatic extraction of semantic relationships among multilingual words from latent space



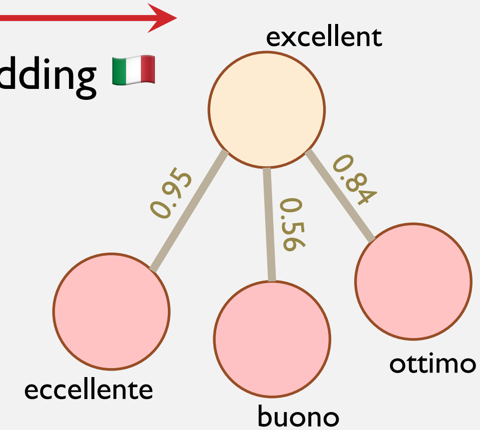
Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint arXiv:1804.07745*.

GRAPH CONSTRUCTION



excellent	similarity
eccellente	0.95
ottimo	0.84
buono	0.56

(K = 3)



Giobergia, F., Cagliero, L., Garza, P., & Baralis, E. (2020). Cross-Lingual Propagation of Sentiment Information Based on Bilingual Vector Space Alignment. In *EDBT/ICDT Workshops* (pp. 8-10).

SOME RESULTS

- No lexicon required,
  - AND better performance!
- Tested on binary sentiment prediction tasks:
  - Positive/negative classes
  - 7 languages, 9 datasets
    - Various types of reviews
- Text → sentiment + word embeddings
- Classifiers:
  - RF, SVM
  - CNN, DC-CNN

Dataset	Our method SVM	Our method RF	Dong and De Melo SVM	Dong and De Melo RF
cs	**0.7403**	0.7198	0.7227	0.7297
de	0.6847	**0.6981**	0.6495	0.6756
es	**0.6131**	0.531	0.4451	0.4892
fr	0.7021	**0.7291**	0.6389	0.6764
it	**0.8256**	0.794	0.6805	0.6644
nl	**0.6869**	0.6369	0.5903	0.6022
ru	0.6840	0.6112	**0.7221**	0.7009
IT_1	**0.8439**	0.8424	0.7435	0.7311
IT_2	**0.8441**	0.8427	0.7415	0.7494

Dataset	Our method DC-CNN	Our method CNN	Dong and de Melo DC-CNN	Dong and de Melo CNN
cs	**0.9311**	0.9226	0.9241	0.9149
de	0.8701	**0.9046**	0.8838	0.8874
es	**0.6845**	0.6435	0.6834	0.6611
fr	**0.9168**	0.9078	0.9104	0.8988
it	0.9339	0.9361	**0.9365**	0.9285
nl	0.7087	**0.7352**	0.7195	0.7273
ru	**0.9258**	0.9141	0.8978	0.9187
IT_1	0.9272	**0.9526**	0.9217	0.9471
IT_2	0.9366	**0.9539**	0.9305	**0.9539**

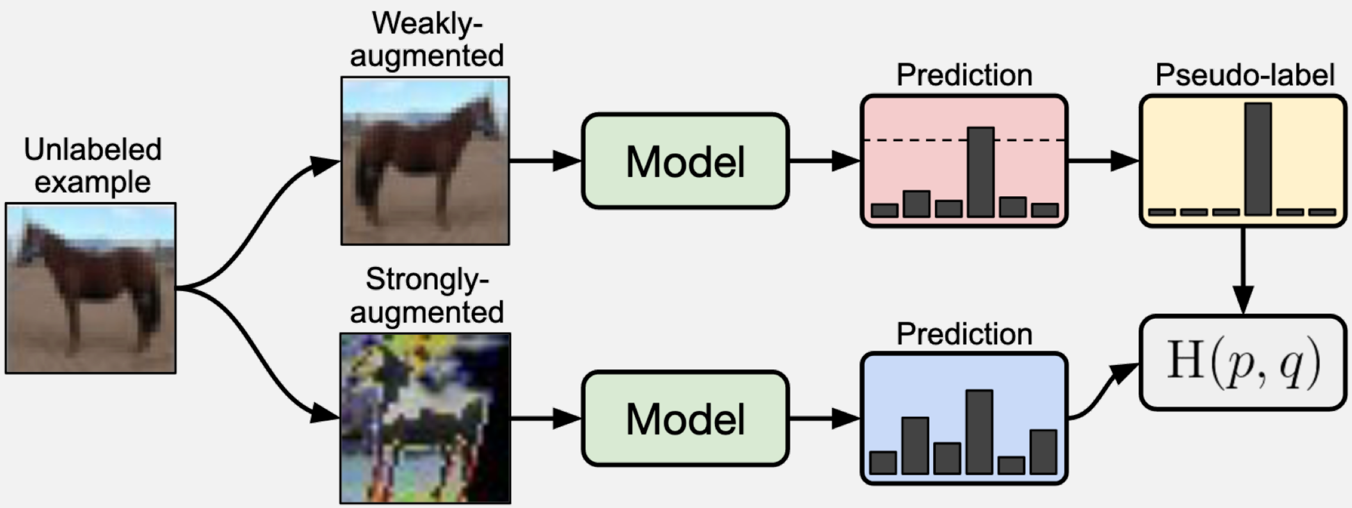
BEYOND SENTIMENT ANALYSIS

- The same approach can be used in other NLP tasks
  - E.g. propagation of embeddings trained for custom domains
- And even outside of NLP!
  - When there are entities that can be linked across domains
  - E.g. social networks: same users, different platforms

SEMI-SUPERVISED LEARNING

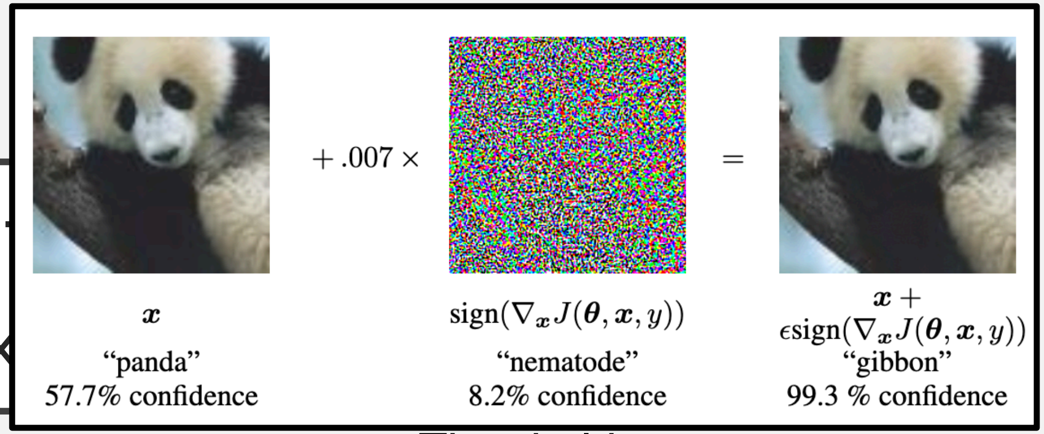
	Unsupervised learning	Semi-supervised learning	rning	Dom ad
Scope	• No labels available	• Limited labelled data, unlabelled data often abundant	provided by oracle, based on model's queries	• Lab other no lab for ta
Main properties	• Learn cluster membership • Learn feature representation • Find recurring patterns in data	• Build model on labelled + unlabelled data, infer missing labels (inductive) • Infer new labels based on properties of known points (transductive)	• Definition of a query policy (when does the model request new lab • Mo regi s c	• Supervised learning on resource-rich domain • Transfer tec pro kno tar

PSEUDO-LABELLING + CONSISTENCY REGULARIZATION = FIXMATCH

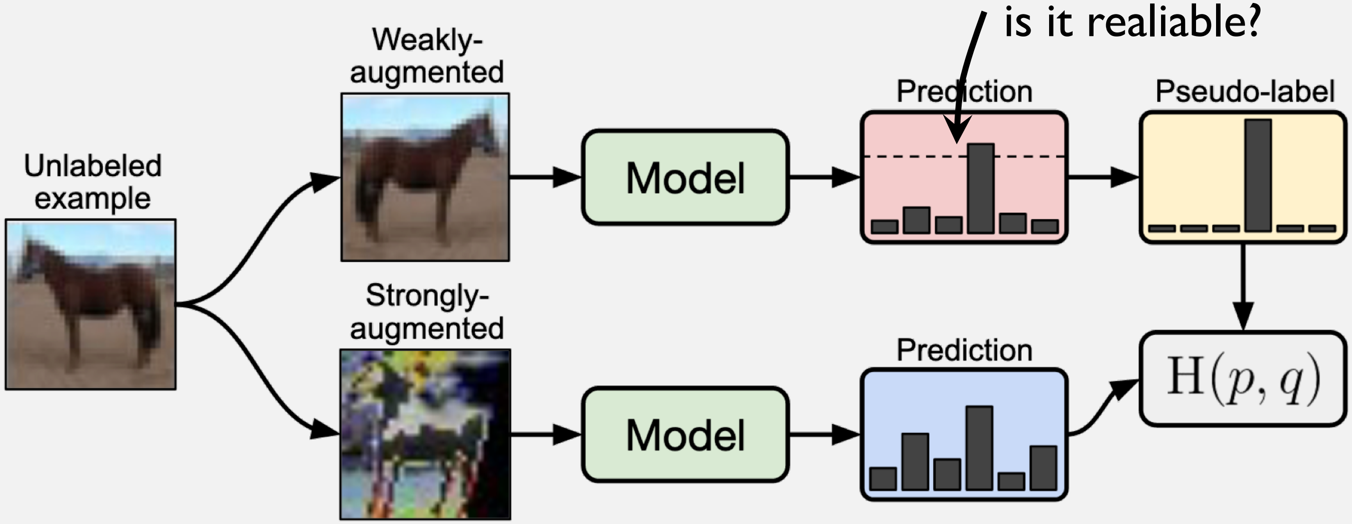


Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, *33*, 596-608.

PSEUDO-
CONSISTENCY
FIX

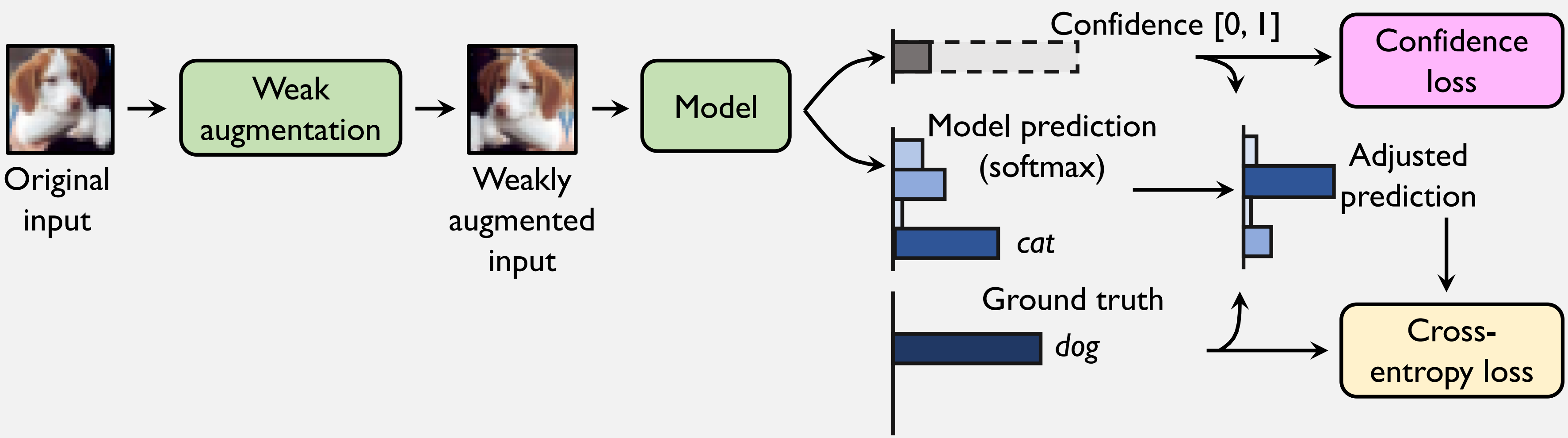


Threshold on softmax output… is it realiable?



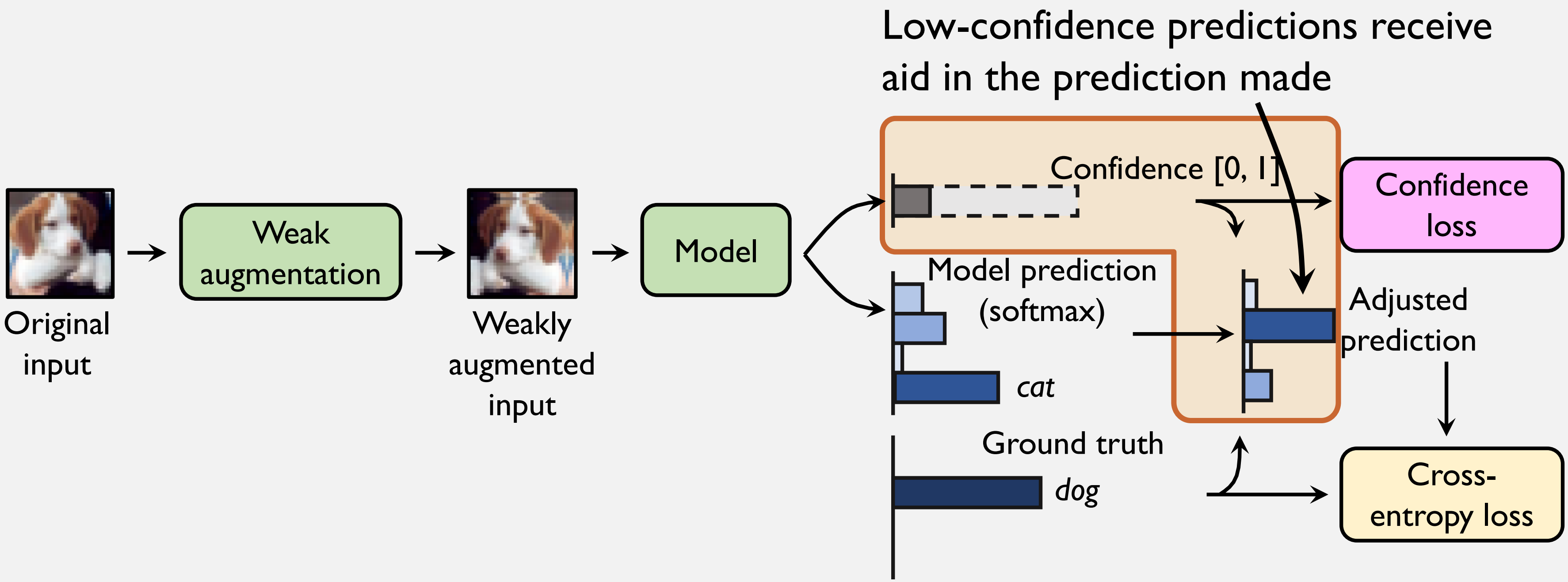
Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, *33*, 596-608.

EXPLICIT CONFIDENCE MECHANISM (SUPERVISED)

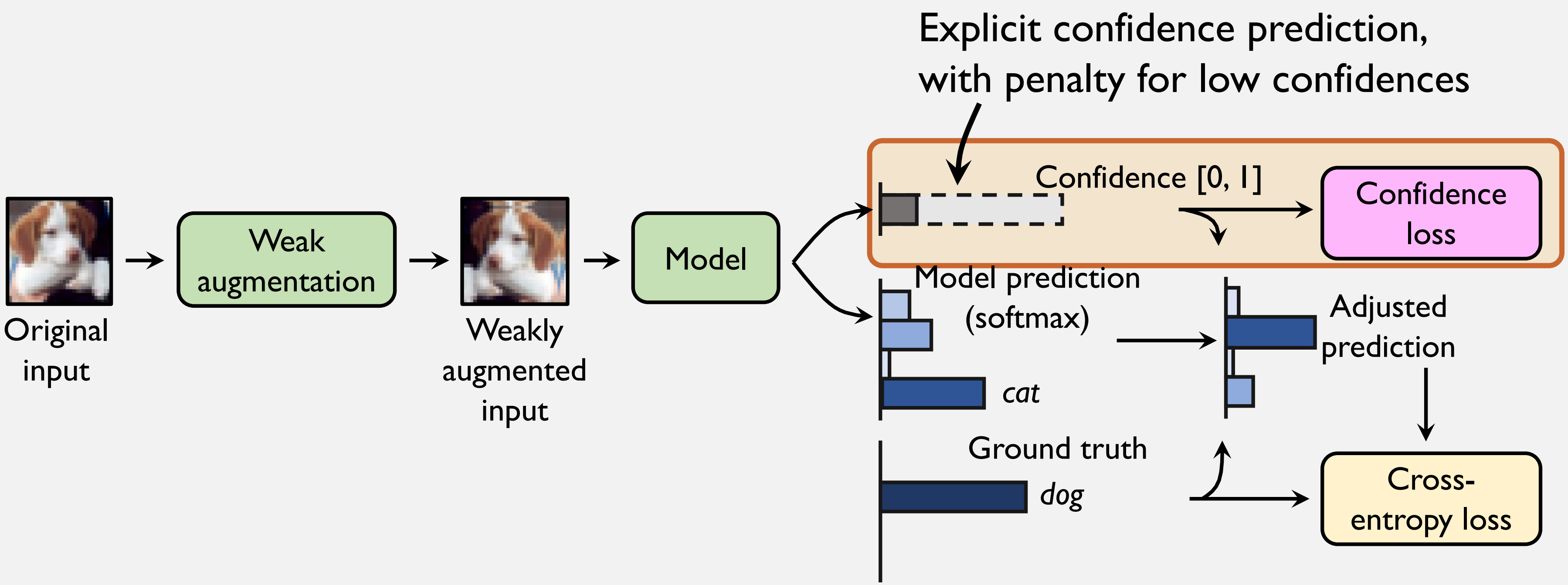


Terrance DeVries and Graham W Taylor. Learning confidence for out-of- distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

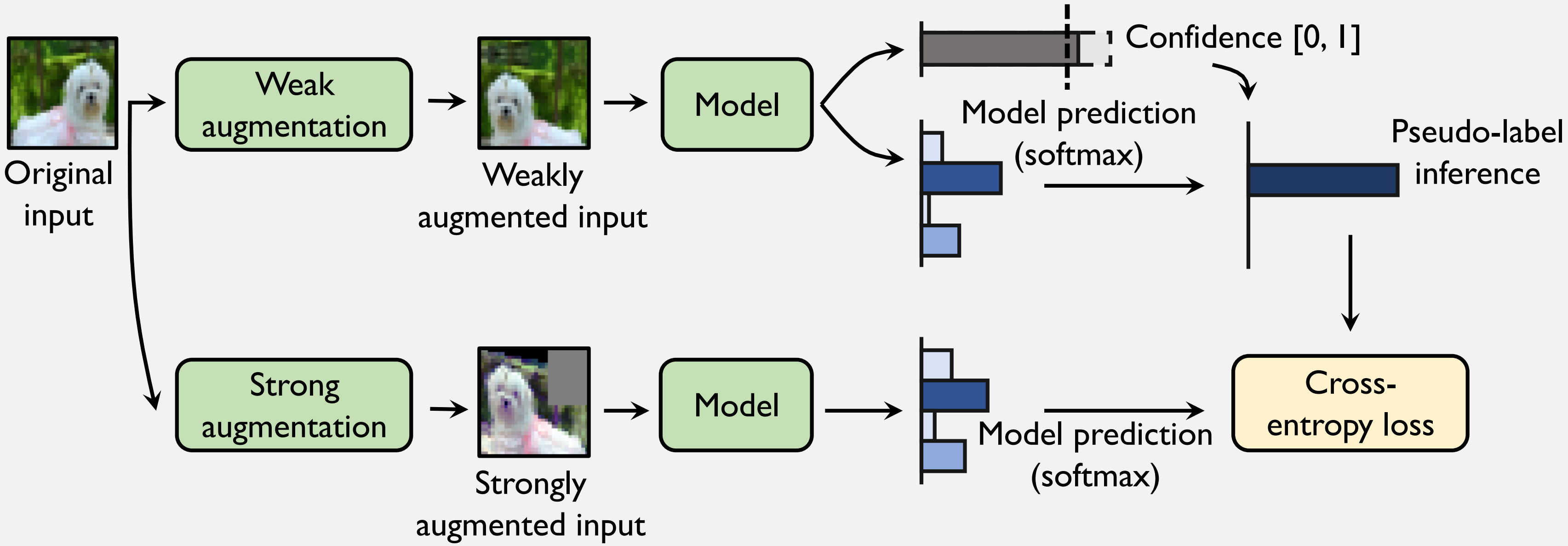
EXPLICIT CONFIDENCE MECHANISM (SUPERVISED)



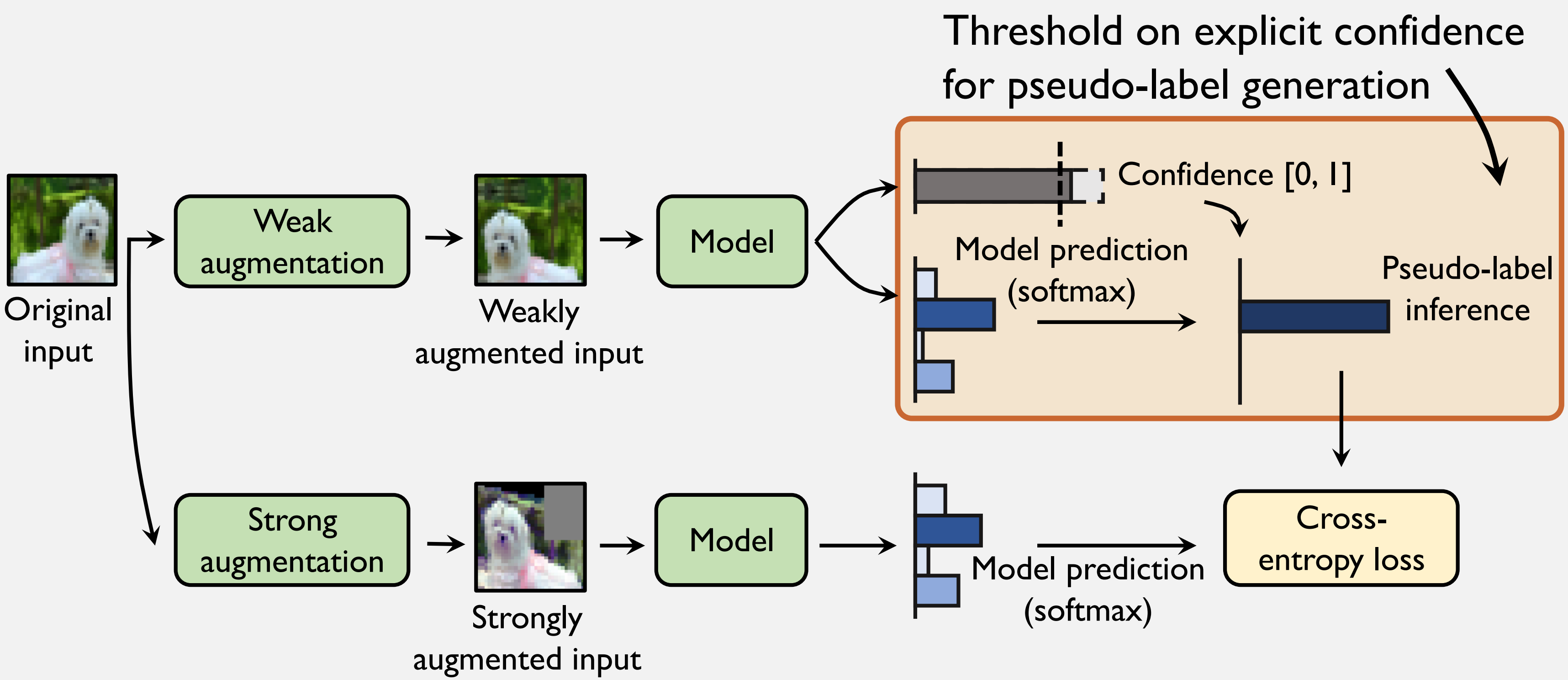
EXPLICIT CONFIDENCE MECHANISM (SUPERVISED)



EXPLICIT CONFIDENCE MECHANISM (UNSUPERVISED)



EXPLICIT CONFIDENCE MECHANISM (UNSUPERVISED)



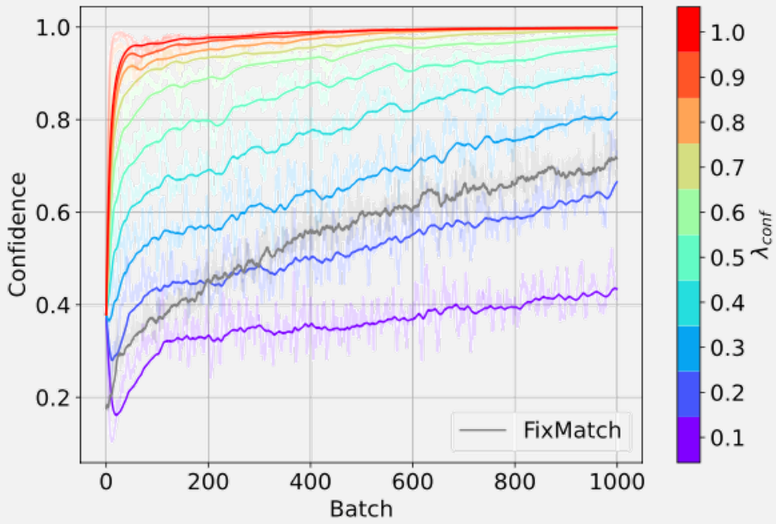
(PRELIMINARY) RESULTS

CIFAR10, 1k training iterations

Method	40 labels		250 labels		4000 labels	
	Top-1	top-5	top-1	top-5	top-1	top-5
FixMatch	18.94 ± 1.16	67.21 ± 1.44	**33.75 ± 1.58***	84.70 ± 0.81	29.98 ± 1.78	84.40 ± 2.17
ConFixMatch	**23.51 ± 1.06**	**72.61 ± 1.60**	31.79 ± 1.69*	**87.02 ± 0.69**	**43.70 ± 3.18**	**92.11 ± 1.56**

CIFAR100, 1k training iterations

Method	40 labels		250 labels		4000 labels	
	top-1	top-5	top-1	top-5	top-1	top-5
FixMatch	23.34 ± 1.01*	69.59 ± 1.16	45.26 ± 0.82	90.53 ± 0.49	67.00 ± 0.95	97.56 ± 0.16*
ConFixMatch	**25.43 ± 1.14***	**73.64 ± 1.87**	**47.28 ± 1.01**	**92.12 ± 0.34**	**69.15 ± 0.76**	**97.71 ± 0.27***



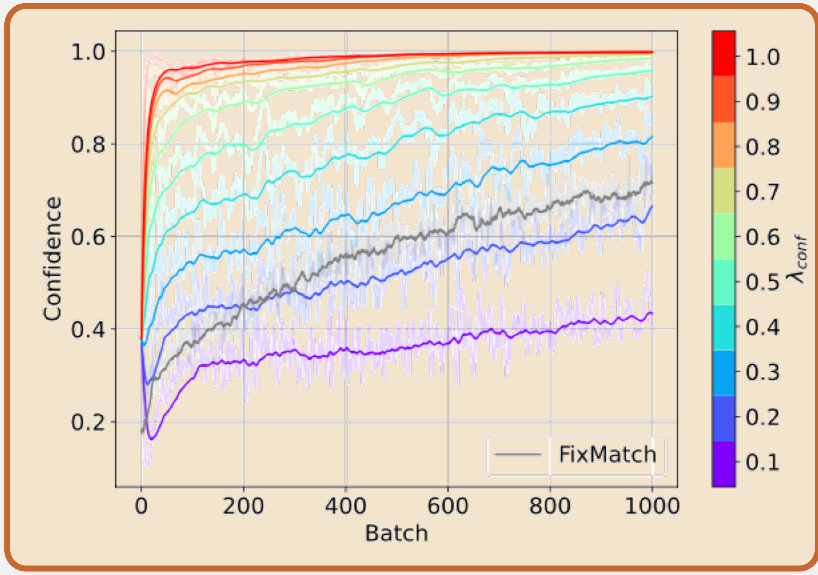
(PRELIMINARY) RESULTS

CIFAR10, 1k training iterations

Method	40 labels		250 labels		4000 labels	
	Top-1	top-5	top-1	top-5	top-1	top-5
FixMatch	18.94 ± 1.16	67.21 ± 1.44	**33.75 ± 1.58***	84.70 ± 0.81	29.98 ± 1.78	84.40 ± 2.17
ConFixMatch	**23.51 ± 1.06**	**72.61 ± 1.60**	31.79 ± 1.69*	**87.02 ± 0.69**	**43.70 ± 3.18**	**92.11 ± 1.56**

CIFAR100, 1k training iterations

Method	40 labels		250 labels		4000 labels	
	top-1	top-5	top-1	top-5	top-1	top-5
FixMatch	23.34 ± 1.01*	69.59 ± 1.16	45.26 ± 0.82	90.53 ± 0.49	67.00 ± 0.95	97.56 ± 0.16*
ConFixMatch	**25.43 ± 1.14***	**73.64 ± 1.87**	**47.28 ± 1.01**	**92.12 ± 0.34**	**69.15 ± 0.76**	**97.71 ± 0.27***



Models with explicit confidence become more confident faster (not necessarily correct, though!)

LIST OF PUBLICATIONS

- **Flavio Giobergia**, Luca Cagliero, Paolo Garza, and Elena Baralis. Cross-lingual propagation of sentiment information based on bilingual vector space alignment. In EDBT/ICDT Workshops, pages 8–10, 2020.
- **Flavio Giobergia** and Elena Baralis. Fast self-organizing maps training. In 2019 IEEE International Conference on Big Data (Big Data), pages 2257–2266. IEEE, 2019.

- Giuseppe Attanasio, **Flavio Giobergia**, Andrea Pasini, Francesco Ventura, Elena Baralis, Luca Cagliero, Paolo Garza, Daniele Apiletti, Tania Cerquitelli, and Silvia Chiusano. Dsle: a smart platform for designing data science competitions. In 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), pages 133–142. IEEE, 2020.
- **Flavio Giobergia**. Triplet losses-based matrix factorization for robust recommendations. In Proceedings of the CIKM 2022 Workshops, volume 3318 of CEUR Workshop Proceedings, 2022.
- Andrea Pasini, **Flavio Giobergia**, Eliana Pastor, and Elena Baralis. Semantic image collection summarization with frequent subgraph mining. IEEE Access, 10:131747–131764, 2022.
- **Flavio Giobergia** and Elena Baralis. Reclaim: Reverse engineering classification metrics. In 2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pages 106–113. IEEE, 2022.

- **Flavio Giobergia**, Elena Baralis, Maria Camuglia, Tania Cerquitelli, Marco Mellia, Alessandra Neri, Davide Tricarico, and Alessia Tuninetti. Mining sensor data for predictive maintenance in the automotive industry. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pages 351–360. IEEE, 2018.
- Danilo Giordano, **Flavio Giobergia**, Eliana Pastor, Antonio La Macchia, Tania Cerquitelli, Elena Baralis, Marco Mellia, and Davide Tricarico. Data-driven strategies for predictive maintenance: Lesson learned from an automotive use case. Computers in Industry, 134:103554, 2022.
- Danilo Giordano, Eliana Pastor, **Flavio Giobergia**, Tania Cerquitelli, Elena Baralis, Marco Mellia, Alessandra Neri, and Davide Tricarico. Dissecting a data-driven prognostic pipeline: A powertrain use case. Expert Systems with Applications, 180:115109, 2021.
- Alessandra Neri, Maria Camuglia, Alessia Tuninetti, Elena Baralis, **Flavio Giobergia**, and Davide Tricarico. Method and system for predicting system status, August 23 2022. US Patent 11,423,321.
- Alkis Koudounas, **Flavio Giobergia**, and Elena Baralis. Time-of-flight cameras in space: Pose estimation with deep learning methodologies. In 2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT), pages 1–6. IEEE, 2022.

- F Siviero, **F Giobergia**, L Menzio, F Miserocchi, M Tornago, R Arcidiacono, N Cartiglia, M Costa, M Ferrero, G Gioachin, et al. First experimental results of the spatial resolution of rsd pad arrays read out with a 16-ch board. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 1041:167313, 2022.
- M Tornago, **F Giobergia**, L Menzio, F Siviero, R Arcidiacono, N Cartiglia, M Costa, M Ferrero, G Gioachin, M Mandurrino, et al. Silicon sensors with resistive read-out: Machine learning techniques for ultimate spatial resolution. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 1047:167816, 2023.

THANK YOU!
QUESTIONS?

Flavio Giobergia
XXXV cycle

Politecnico di Torino
February 15, 2023

Advisor
Elena Baralis

Doctoral Examination Committee
Sara Comai, Referee, *Politecnico di Milano*
Dino Ienco, Referee, *INRAE*
Rosa Meo, *Università degli studi di Torino*
Genoveva Vargas-Solar, *CNRS*
Silvia Chiusano, *Politecnico di Torino*

Politecnico di Torino
SmartData@PoliTO