Some Pragmatic Relationships between **Machine Learning & Cybersecurity**

Giovanni Apruzzese, PhD

May 17th, 2022

UNIVERSITÄT LIECHTENSTEIN

whoami: Dr. **Giovanni Apruzzese**

- **Background:**
  - Did my academic studies (BSc, MSc, PhD) at University of Modena, Italy.
    - Supervisor: Prof. Michele Colajanni
  - In 2019, spent 6 months at Dartmouth College, USA.
    - Supervisor: Prov. VS Subrahmanian
  - Joined the University of Liechtenstein in July 2020 as a PostDoc Researcher.
    - Supervisor: Prof. Pavel Laskov
  - Met Prof. Mauro Conti in 2019, with whom I have been collaborating since 2020.

- **Interests:**
  - Cybersecurity, machine learning, and any network-related topic (+🎮)
  - I like talking, researching and teaching – in a “pragmatic” way ☺

- **Contact information**:
  - Work Email: giovanni.apruzzese@uni.li
  - Feel free to contact me if you have any questions.
    - I reply fast, and will happily do so!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What I do

Machine Learning + Cybersecurity

- Applying ML to *provide security* of a given information system
  - E.g.: using ML to detect network intrusions
- *Attacking / Defending* ML applications
  - E.g.: evading a ML model that detects phishing websites
- Using ML *offensively* against any target
  - E.g.: artificially generating “fake” images

BONUS

- Using ML offensively to attack a ML-based security system

UNIVERSITÄT
LIECHTENSTEIN

Outline

- Using **unlabelled data** for Machine Learning in Cyberthreat Detection
  - Ref: Giovanni Apruzzese, Pavel Laskov, Aliya Tastemirova. “SoK: The Impact of Unlabelled Data in Cyberthreat Detection" *IEEE European Symposium on Security and Privacy*. June 2022

- Adversarial Attacks against Humans **and** Machine Learning
  - Ref: Johannes Schneider, Giovanni Apruzzese. “Concept-based Adversarial Attacks: Tricking Humans and Classifiers alike.” *IEEE Symposium on Security and Privacy – Deep Learning and Security Workshop*. May 2022

Unlabelled data for Machine Learning in Cyberthreat Detection

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

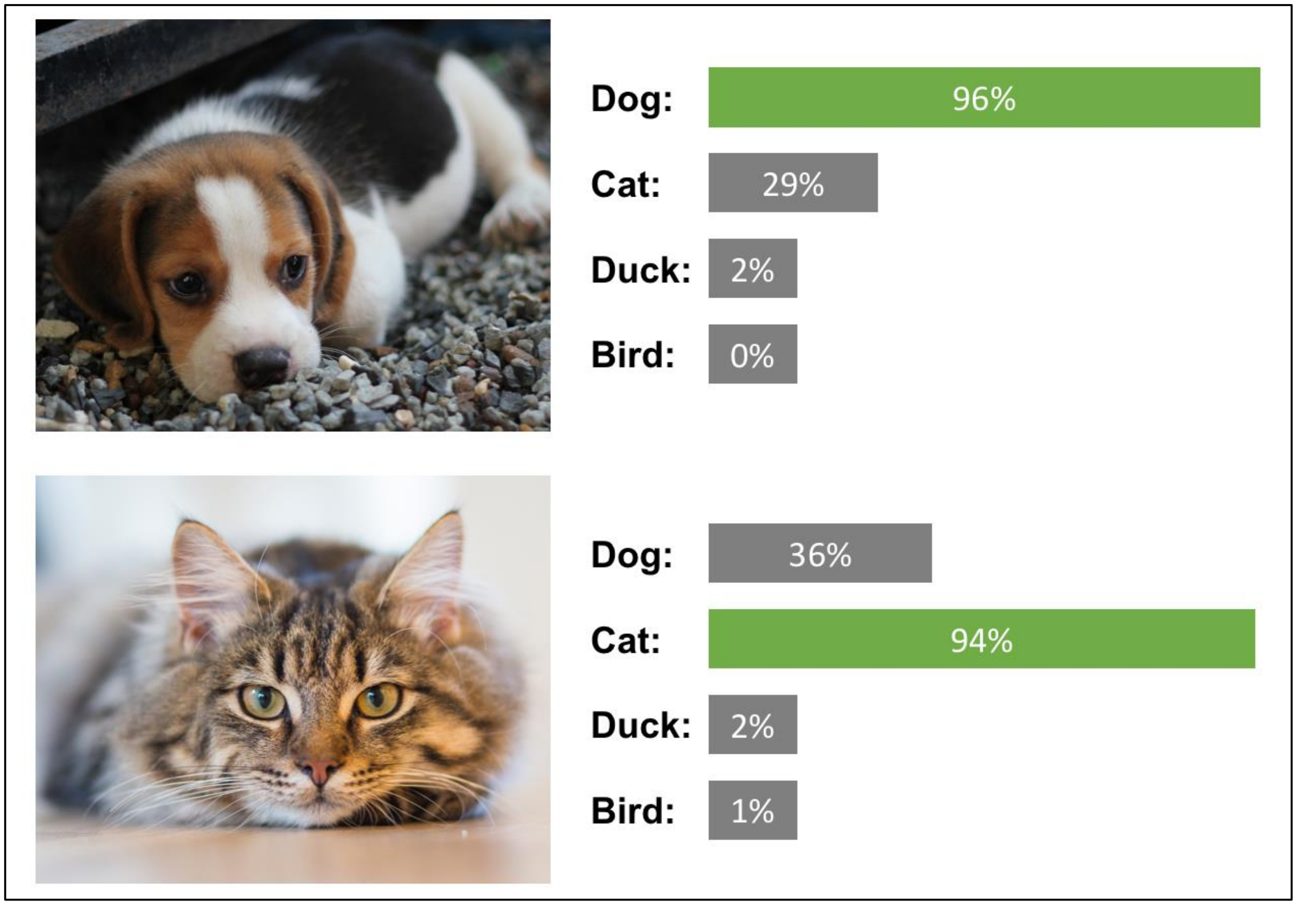
Labelling in Cyberthreat Detection (CTD)

- Using Machine Learning (ML) to *detect* cyber-threats requires *labelled data*.
- Obtaining *plenty* and *accurate* labels is expensive:
  - Human in the loop (but this is true also for Computer Vision…)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Labelling in Cyberthreat Detection (CTD)

- Using Machine Learning (ML) to *detect* cyber-threats requires *labelled data*.
- Obtaining *plenty* and *accurate* labels is expensive:
  - Human in the loop (but this is true also for Computer Vision…)



Shanmugamani, Rajalingappaa. *Deep Learning for Computer Vision: Expert techniques to train advanced neural networks using TensorFlow and Keras*. Packt Publishing Ltd, 2018.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Labelling in Cyberthreat Detection (CTD)

- Using Machine Learning (ML) to *detect* cyber-threats requires *labelled data*.
- Obtaining *plenty* and *accurate* labels is expensive:
  - Human in the loop (but this is true also for Computer Vision…)
  - Instead, in CTD…

Network Intrusion Detection

	StartTime	Dur	Proto	SrcAddr	Sport	Dir	DstAddr	Dport	State	sTos	dTos	TotPkts	TotBytes	SrcBytes
0	2011/08/10 09:46:53.047277	3550.182373	udp	212.50.71.179	39678	<->	147.32.84.229	13363	CON	0.0	0.0	12	875	413
1	2011/08/10 09:46:53.048843	0.000883	udp	84.13.246.132	28431	<->	147.32.84.229	13363	CON	0.0	0.0	2	135	75
2	2011/08/10 09:46:53.049895	0.000326	tcp	217.163.21.35	80	<?>	147.32.86.194	2063	FA_A	0.0	0.0	2	120	60
3	2011/08/10 09:46:53.053771	0.056966	tcp	83.3.77.74	32882	<?>	147.32.85.5	21857	FA_FA	0.0	0.0	3	180	120
4	2011/08/10 09:46:53.053937	3427.768066	udp	74.89.223.204	21278	<->	147.32.84.229	13363	CON	0.0	0.0	42	2856	1596
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
675872	2011/08/10 11:04:27.118993	0.020525	udp	147.32.84.165	1025	<->	147.32.80.9	53	CON	0.0	0.0	2	590	87
675873	2011/08/10 11:04:27.119042	2.919582	tcp	147.32.84.194	3112	->	147.32.80.13	80	FSPA_FSPA	0.0	0.0	172	160606	7360
675874	2011/08/10 11:04:27.124802	0.000281	udp	147.32.84.59	57993	<->	147.32.80.9	53	CON	0.0	0.0	2	235	76
675875	2011/08/10 11:04:27.125921	0.000195	udp	147.32.84.59	60616	<->	147.32.80.9	53	CON	0.0	0.0	2	218	76
675876	2011/08/10 11:04:27.129857	0.011865	tcp	147.32.84.59	52776	->	217.31.58.184	80	FSPA_FSPA	0.0	0.0	10	1615	943

UNIVERSITÄT LIECHTENSTEIN

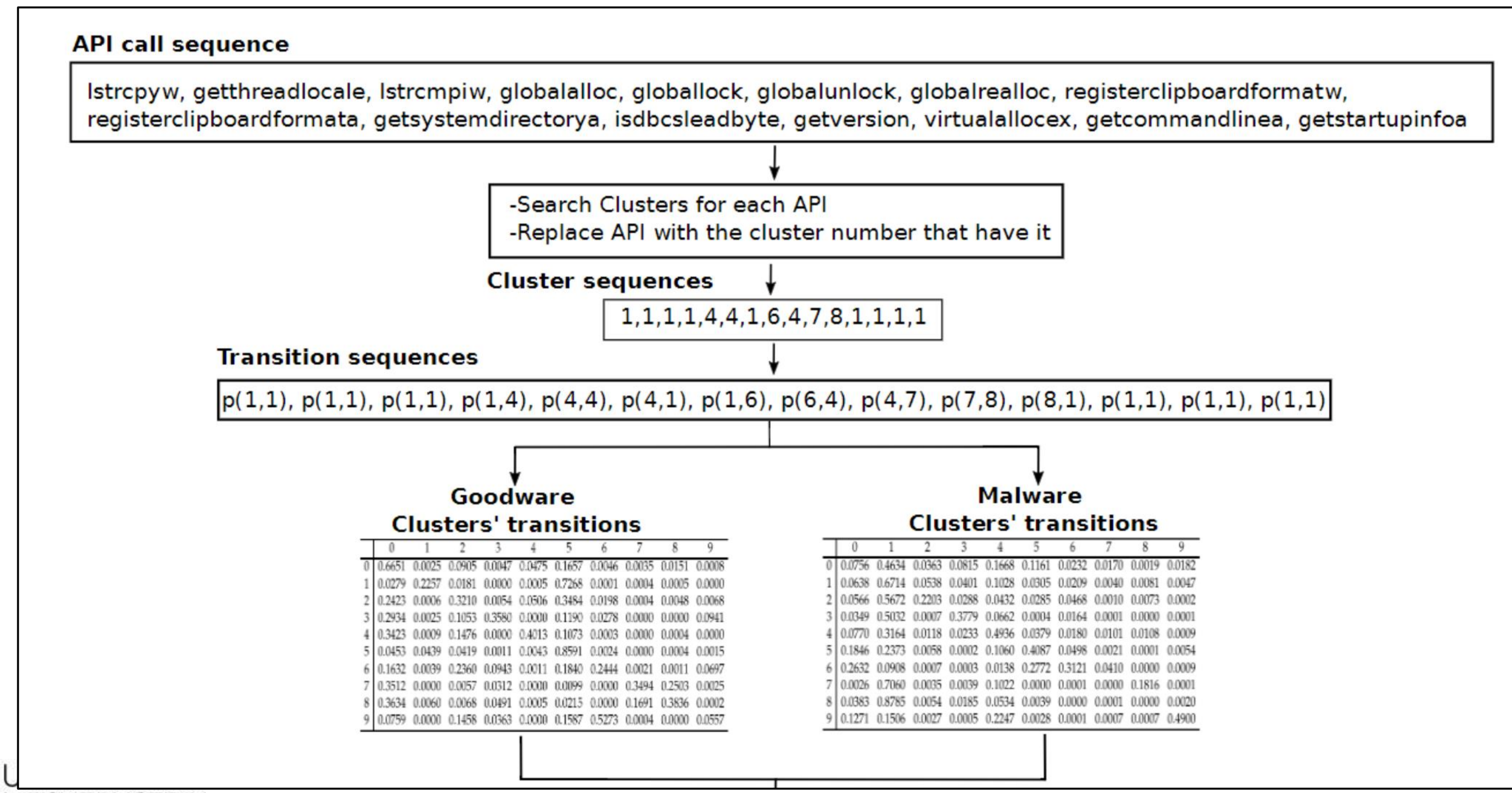
CTU13 dataset: https://www.stratosphereips.org/datasets-ctu13

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Labelling in Cyberthreat Detection (CTD)

- Using Machine Learning (ML) to *detect* cyber-threats requires *labelled data*.
- Obtaining *plenty* and *accurate* labels is expensive:
  - Human in the loop (but this is true also for Computer Vision…)
  - Instead, in CTD…

Malware Detection



Goodware Clusters' transitions

	0	1	2	3	4	5	6	7	8	9
0	0.6651	0.0025	0.0905	0.0047	0.0475	0.1657	0.0046	0.0035	0.0151	0.0008
1	0.0279	0.2257	0.0181	0.0000	0.0005	0.7268	0.0001	0.0004	0.0005	0.0000
2	0.2423	0.0006	0.3210	0.0054	0.0506	0.3484	0.0198	0.0004	0.0048	0.0068
3	0.2934	0.0025	0.1053	0.3580	0.0000	0.1190	0.0278	0.0000	0.0000	0.0941
4	0.3423	0.0009	0.1476	0.0000	0.4013	0.1073	0.0003	0.0000	0.0004	0.0000
5	0.0453	0.0439	0.0419	0.0011	0.0043	0.8591	0.0024	0.0000	0.0004	0.0015
6	0.1632	0.0039	0.2360	0.0943	0.0011	0.1840	0.2444	0.0021	0.0011	0.0697
7	0.3512	0.0000	0.0057	0.0312	0.0000	0.0099	0.0000	0.3494	0.2503	0.0025
8	0.3634	0.0060	0.0068	0.0491	0.0005	0.0215	0.0000	0.1691	0.3836	0.0002
9	0.0759	0.0000	0.1458	0.0363	0.0000	0.1587	0.5273	0.0004	0.0000	0.0557

Malware Clusters' transitions

	0	1	2	3	4	5	6	7	8	9
0	0.0756	0.4634	0.0363	0.0815	0.1668	0.1161	0.0232	0.0170	0.0019	0.0182
1	0.0638	0.6714	0.0538	0.0401	0.1028	0.0305	0.0209	0.0040	0.0081	0.0047
2	0.0566	0.5672	0.2203	0.0288	0.0432	0.0285	0.0468	0.0010	0.0073	0.0002
3	0.0349	0.5032	0.0007	0.3779	0.0662	0.0004	0.0164	0.0001	0.0000	0.0001
4	0.0770	0.3164	0.0118	0.0233	0.4936	0.0379	0.0180	0.0101	0.0108	0.0009
5	0.1846	0.2373	0.0058	0.0002	0.1060	0.4087	0.0498	0.0021	0.0001	0.0054
6	0.2632	0.0908	0.0007	0.0003	0.0138	0.2772	0.3121	0.0410	0.0000	0.0009
7	0.0026	0.7060	0.0035	0.0039	0.1022	0.0000	0.0001	0.0000	0.1816	0.0001
8	0.0383	0.8785	0.0054	0.0185	0.0534	0.0039	0.0000	0.0001	0.0000	0.0020
9	0.1271	0.1506	0.0027	0.0005	0.2247	0.0028	0.0001	0.0007	0.0007	0.4900

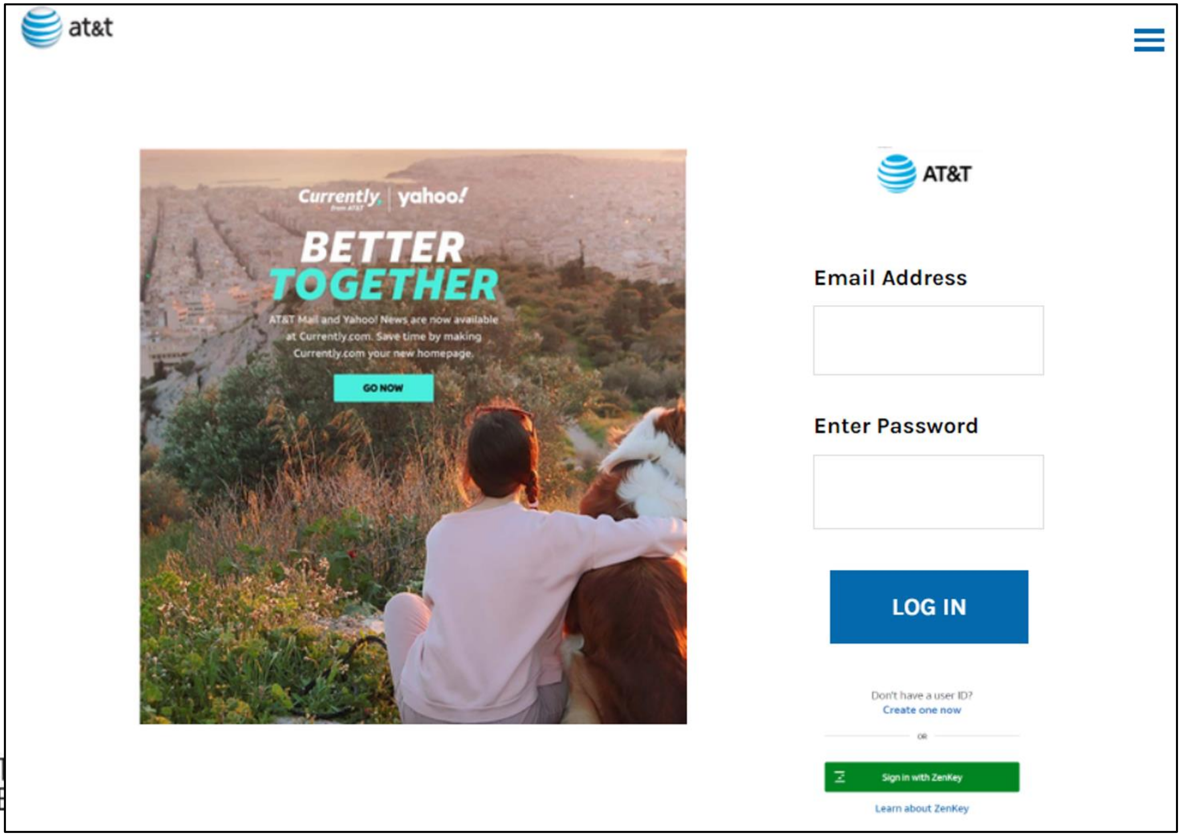
Amer, Eslam, Shaker El-Sappagh, and Jong Wan Hu. "Contextual identification of windows malware through semantic interpretation of api call sequence." *Applied Sciences* 10.21 (2020): 7673.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Labelling in Cyberthreat Detection (CTD)

- Using Machine Learning (ML) to *detect* cyber-threats requires *labelled data*.
- Obtaining *plenty* and *accurate* labels is expensive:
  - Human in the loop (but this is true also for Computer Vision…)
  - Instead, in CTD…

Phishing Website Detection



Valid Phish according to Phishtank (March 2022)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Labelling in Cyberthreat Detection (CTD)

- Using Machine Learning (ML) to *detect* cyber-threats requires *labelled data*.
- Obtaining *plenty* and *accurate* labels is <u>expensive:</u>
  - Human in the loop (but this is true also for Computer Vision…)
  - Instead, in CTD…

For CTD, labelling requires *expert knowledge* and is an *error prone* task.

And the "concept drift" further aggravates this issue…

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Semisupervised Learning

- Labelled data is expensive, but *unlabelled* data is cheap(er).

→ Why not using unlabelled data to improve the proficiency of ML models?

Mixing *labelled* with *unlabelled* data is a ML approach denoted as **“Semisupervised Learning” (SsL)**

UNIVERSITÄT LIECHTENSTEIN

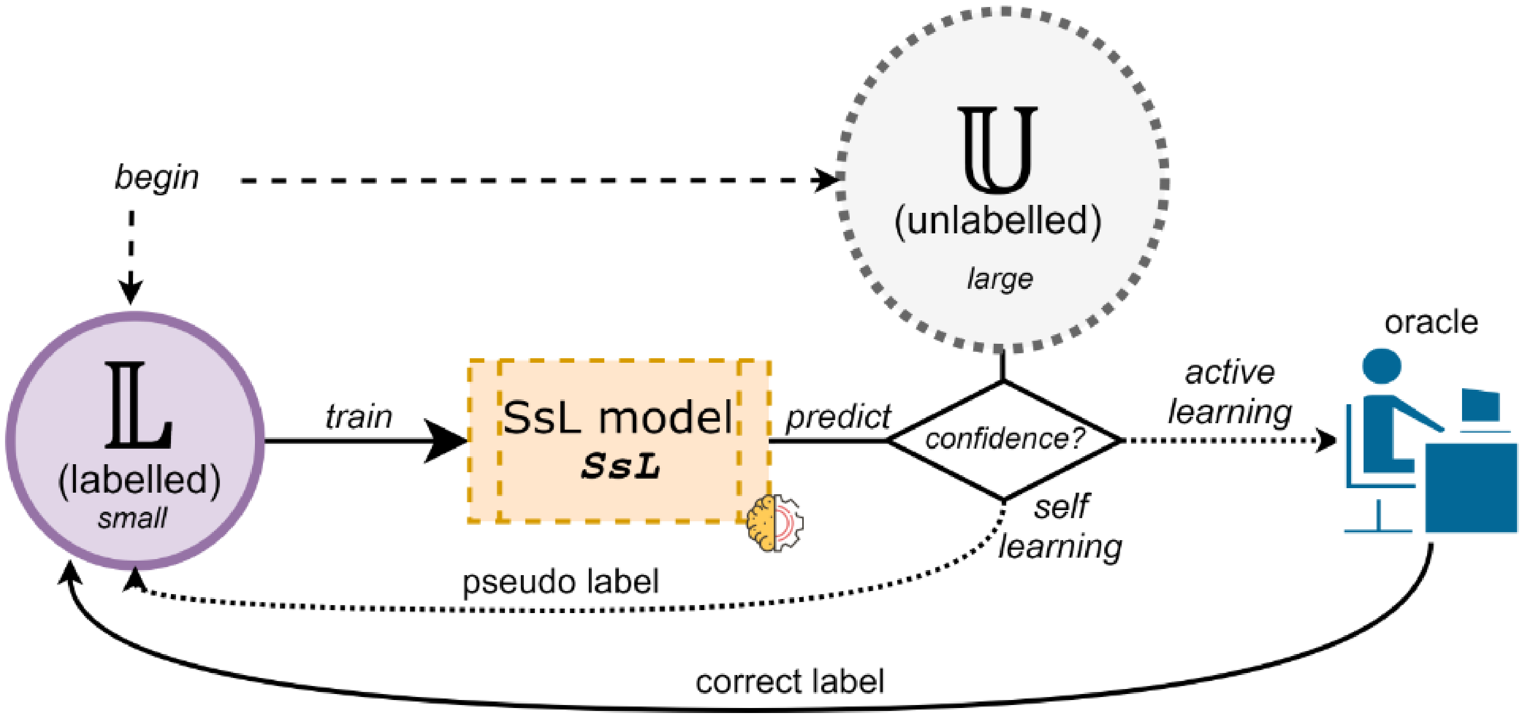
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Semisupervised Learning

- Labelled data is expensive, but *unlabelled* data is cheap(er).

→ Why not using unlabelled data to improve the proficiency of ML models?

Mixing *labelled* with *unlabelled* data is a ML approach denoted as **“Semisupervised Learning” (SsL)**

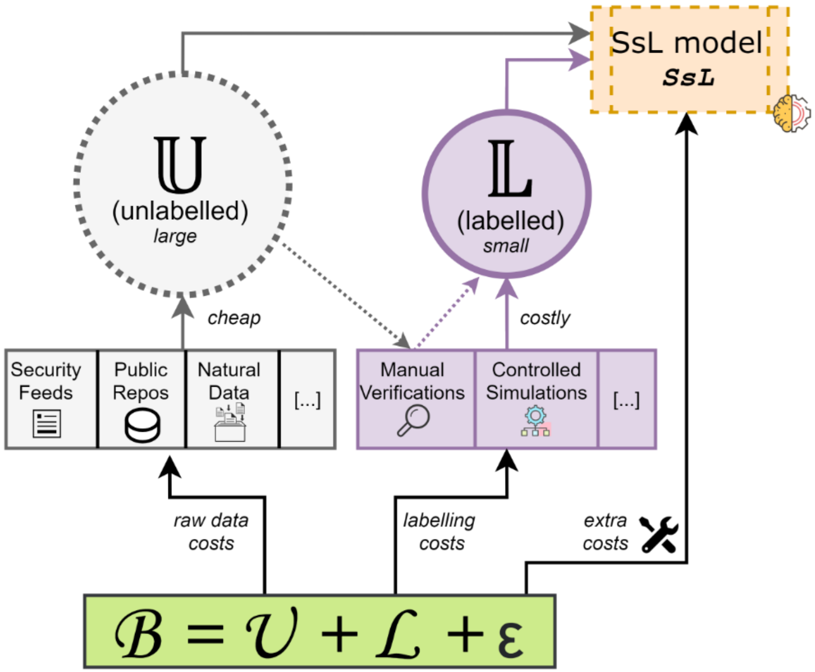


Examples of SsL: *active learning* and *self learning* (e.g., *pseudo labelling*)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Goal of Semisupervised Learning

- Developing SsL models is cheaper than "supervised learning" (SL) models, **but it is not free.**



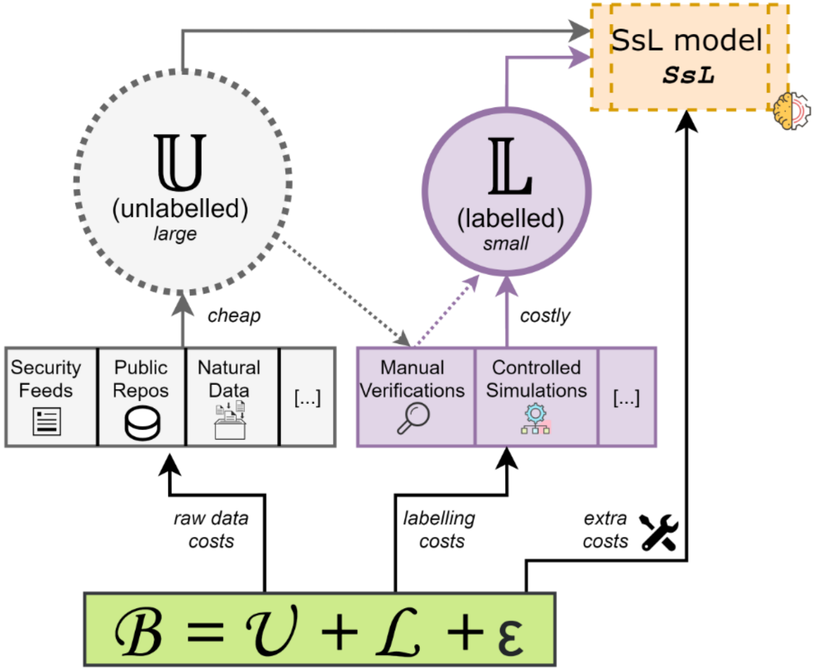
Acquiring unlabelled data requires some investment, $\mathcal{U}$

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

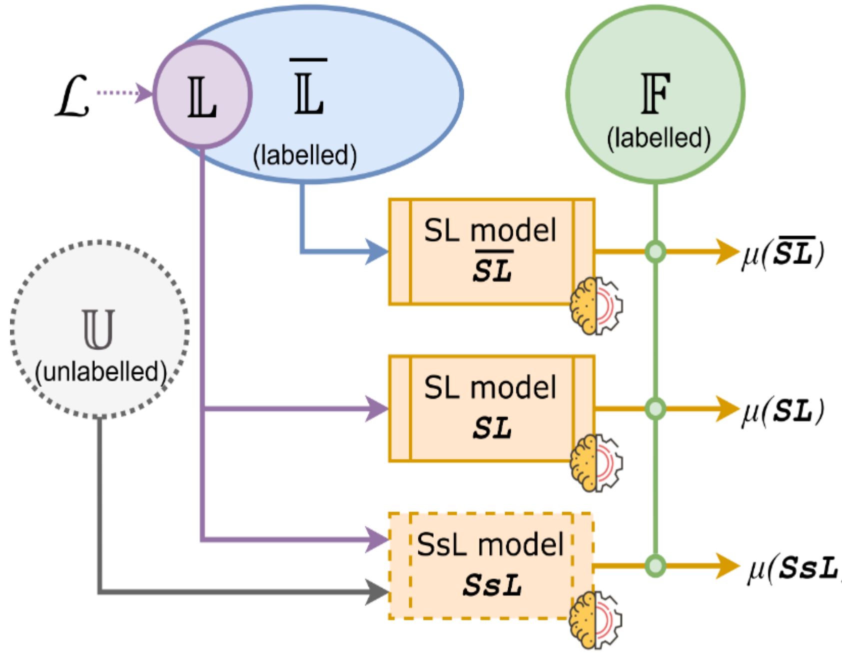
Goal of Semisupervised Learning

- Developing SsL models is cheaper than "supervised learning" (SL) models, **but it is not free.**



Acquiring unlabelled data requires some investment, $\mathcal{U}$

- A SsL model should achieve a *superior performance* than a SL model that uses the *same labelling budget, $\mathcal{L}$*



Both SsL and SL require $\mathcal{L}$...
...but SL do not require $\mathcal{U}$

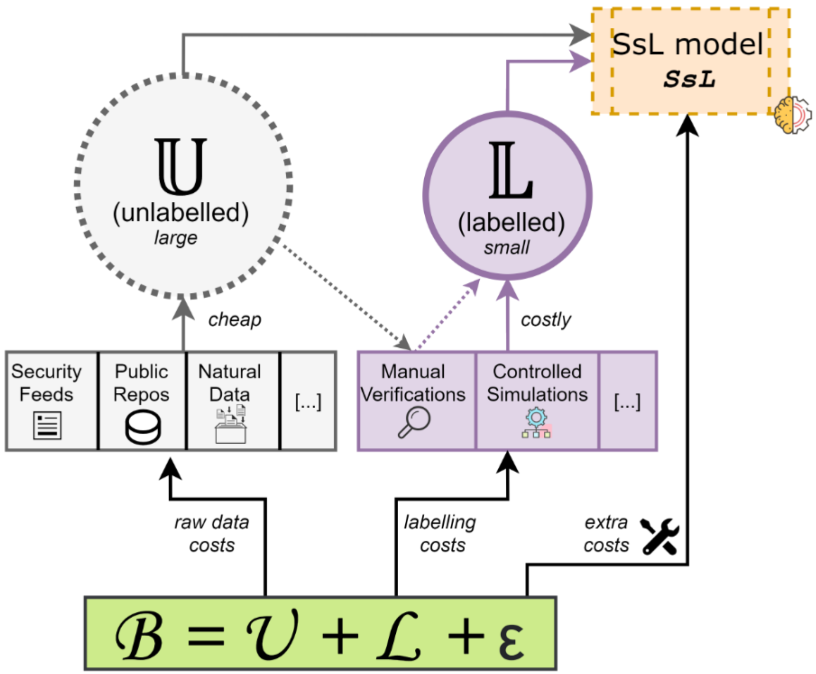
UNIVERSITÄT LIECHTENSTEIN

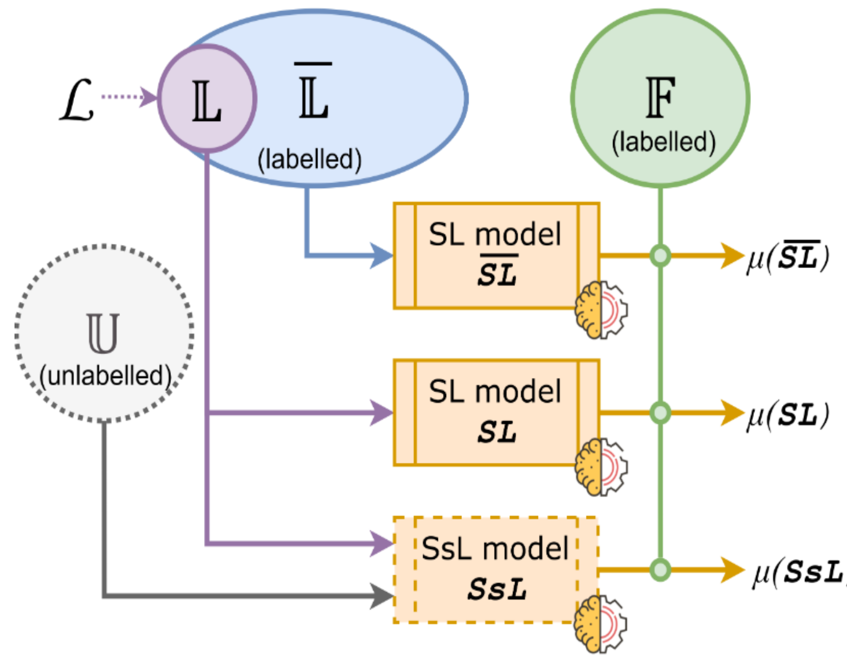
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Goal of Semisupervised Learning

- Developing SsL models is cheaper than "supervised learning" (SL) models, **but it is not free.**

- A SsL model should achieve a *superior performance* than a SL model that uses the *same labelling budget*, $\mathcal{L}$





Acquiring unlabelled data requires some investment, $\mathcal{U}$

Both SsL and SL require $\mathcal{L}$...
...but SL do not require $\mathcal{U}$

Definition 1. The *goal* of a Semisupervised Learning (SsL) method is using $\mathbb{U}$ alongside any $\mathbb{L}$ obtained with $\mathcal{L}$ to devise a model SsL. After deployment, such SsL should predict the ground truth of the samples in $\mathbb{F}$ by achieving a performance $\mu(SsL)$ that is: $\mu(SL) < \mu(SsL) \leq \mu(\overline{SL})$.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Problem: nobody cares

The current state-of-the-art does not allow to determine whether SsL methods applied in Cyberthreat Detection are truly beneficial

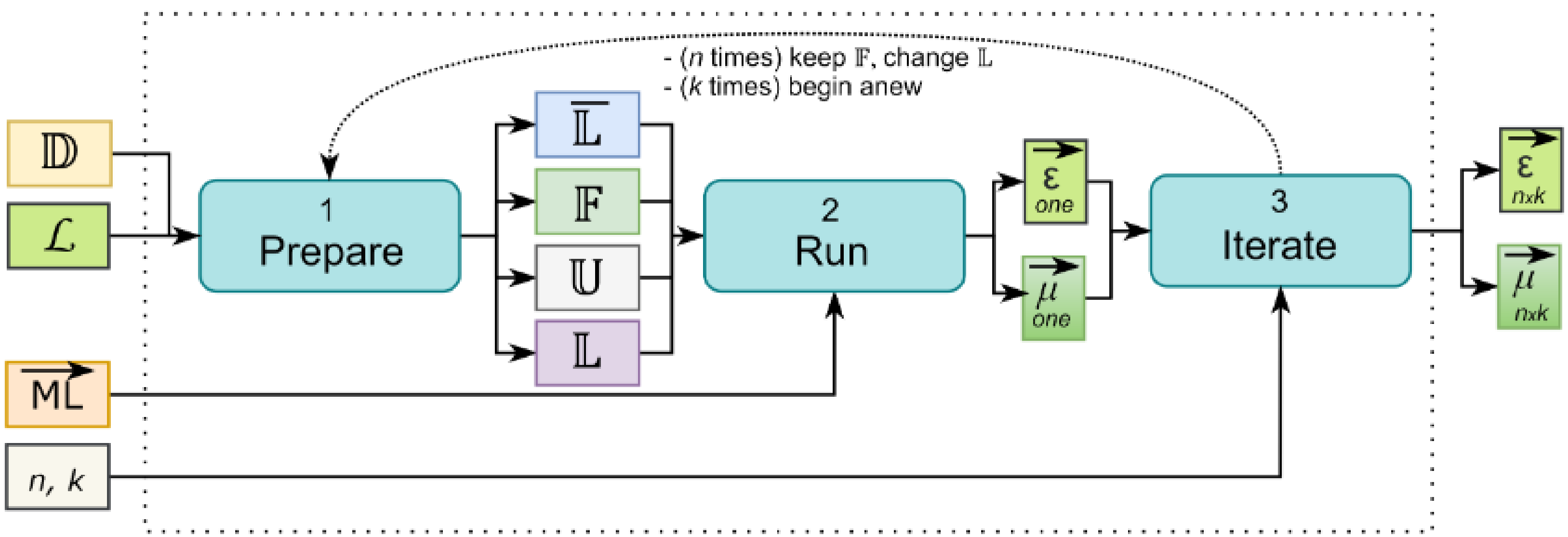
Task	Paper (1st Author)	Year	Lower Bound	Ablation Study	Upper Bound	Stat. Sign.	Transparency *Labels*	Transparency *Balance*	Repr.	Dataset
Network Intrusion Detection	Li [93]	2007	✓	✓	✗	✗	✓	✓	◐	NSL-KDD
	Long [94]	2008	✓	✓	✗	◐	✓	✗	◐	NSL-KDD
	Görnitz [95]	2009	✓	✓	✗	◐	✓	✓	✗	Private
	Seliya [96]	2010	✓	✓	✗	✗	✓	✓	◐	NSL-KDD
	Symons [97]	2012	✗	✓	✓	◐	✓	✗	✗	Kyoto2006
	Wagh [98]	2014	✗	✗	✗	✗	✓	✓	◐	NSL-KDD
	Noorbehbahani [35]	2015	✗	◐	✓	✗	✓	✓	◐	NSL-KDD, Custom
	Ashfaq [99]	2017	✗	◐	✗	✗	✓	✗	◐	NSL-KDD
	Qiu [67]	2017	✗	◐	✓	✗	✓	✓	✗	Custom
	McElwee [100]	2017	✗	◐	✓	✗	✓	✗	◐	NSL-KDD
	Kumari [68]	2017	✓	◐	✗	✗	✓	✗	◐	NSL-KDD
	Yang [101]	2018	◐	✓	✓	✗	✓	✗	✗	NSL-KDD, AWID
	Gao [102]	2018	✓	◐	✗	✗	✓	✗	✗	NSL-KDD
	Shi [103]	2018	◐	◐	✗	✗	✓	✗	✗	NSL-KDD
	Yao [36]	2019	◐	◐	✓	✗	✓	✓	◐	NSL-KDD
	Yuan [104]	2019	✗	◐	✗	◐	✓	✓	◐	NSL-KDD
	Zhang [65]	2020	◐	✗	✓	◐	✓	✗	◐	NSL-KDD
	Hara [105]	2020	✗	◐	✓	✗	✗	✗	✗	NSL-KDD
	Ravi [106]	2020	✓	✗	✗	✗	✓	✗	✗	NSL-KDD
	Gao [107]	2020	✗	✓	✓	✓	✓	✓	✗	NSL-KDD
	Li [108]	2020	✗	◐	✓	✓	✓	✗	◐	NSL-KDD, Private
	Zhang [70]	2021	◐	◐	✗	◐	✗	✓	◐	CICIDS2017, CTU13
	Liang [109]	2021	✓	◐	✓	◐	✓	✓	◐	NSL-KDD
Phishing Detection	Gyawali [110]	2011	✗	✓	✓	✗	✓	✓	◐	Private
	Zhao [111]	2013	✓	✓	✓	✓	✗	✓	✓*	DetMalURL
	Gabriel [15]	2017	◐	◐	✗	✗	✗	✗	◐	Private
	Yang [112]	2017	✓	◐	✗	✗	✓	✓	◐	Private
	Bhattacharjee [113]	2017	✗	✓	✗	◐	✗	✗	◐	Private
	Li [55]	2017	✓	✓	✓	◐	✓	✓	✗	Custom
Malware Detection	Moskovitch [114]	2008	✗	✓	✗	◐	✓	✓	✗	Custom
	Santos [115]	2011	✗	✗	✓	✗	✓	✓	◐	Custom
	Nissim [116]	2012	✗	◐	✓	◐	✗	✗	✗	Private
	Zhao [117]	2012	✗	✗	✗	✗	✓	✓	◐	Private
	Nissim [118]	2014	✓	✓	✗	◐	✓	✓	✗	Custom
	Zhang[119]	2015	◐	◐	✗	✗	✓	✓	✗	Private
	Nissim [120]	2016	✗	✓	✓	◐	✓	✓	◐	Custom
	Ni [121]	2016	✓	✓	✗	◐	✓	✓	◐	Private
	Chen [122]	2017	✓	✓	✗	◐	✗	✗	◐	Private
	Rashidi [66]	2017	✗	✓	✓	◐	✓	✓	✗	Drebin
	Fu [123]	2019	✓	✓	✗	✗	✓	✗	◐	Private
	Irofti [124]	2019	◐	◐	✗	◐	✗	✗	✓	DREBIN, EMBER
	Pendlebury [86]	2019	✗	✗	✓	◐	✓	✓	✓	AndroZoo
	Sharmeen [125]	2020	✓	◐	✗	◐	✓	✓	◐	Drebin, AndroZoo
	Chen [126]	2020	◐	◐	✓	✗	✓	✓	◐	MCC
	Koza [11]	2020	✓	◐	✓	◐	✓	✗	✓	Private
	Noorbehbahani [13]	2020	✓	✗	✗	◐	✓	✓	✗	AndMal17
	Li [127]	2021	✗	◐	✗	◐	✓	✗	◐	FalDroid, DREBIN, Genome
	Liang [109]	2021	✓	◐	✓	◐	✓	✓	◐	Custom

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Solution: CEF-SsL

- SsL is intriguing, but its "pragmatic" benefits are still unknown
- Identifying (and quantifying) such benefits requires adopting a rigorous workflow

→ CEF-SsL: Cybersecurity Evaluation Framework for Semisupervised Learning



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(re)Evaluation

- Massive evaluation on 9 existing datasets for 3 cyberthreat detection tasks:
  - Network Intrusion Detection (NID)
  - Phishing Website Detection (PWD)
  - Malware Detection (MD)

Labels range between 100 and 2400

Results (F1-score)

CTD	NID			PWD			MD		
Method	CTU13	UNB15	IDS17	Mend	UCI	δPhish	DREBIN	Ember	AndMal
$\pi\, SsL$	0.588	0.437	0.820	0.850	0.884	0.778	0.474	0.647	0.900
$\widehat{\pi}\, SsL$	0.584	0.435	0.818	0.849	0.883	0.777	0.470	0.641	0.890
$\alpha\, SsL_l$	**0.693**	0.582	**0.897**	**0.863**	**0.903**	0.770	**0.546**	**0.687**	**0.924**
$\alpha\, SsL_o$	0.637	0.577	0.874	0.855	0.891	0.745	0.497	0.673	0.916
$\alpha\, SsL_h$	0.510	0.436	0.786	0.834	0.851	0.714	0.423	0.598	0.892
$\alpha^{\pi} SsL_l$	0.664	0.533	0.853	0.861	0.901	0.767	0.529	0.654	0.901
$\alpha^{\pi} SsL_o$	0.633	**0.595**	0.857	0.854	0.890	0.745	0.489	0.647	0.895
$\alpha^{\pi} SsL_h$	0.486	0.427	0.744	0.833	0.851	0.711	0.410	0.579	0.865

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(re)Evaluation

- Massive evaluation on 9 existing datasets for 3 cyberthreat detection tasks:
  - Network Intrusion Detection (NID)
  - Phishing Website Detection (PWD)
  - Malware Detection (MD)

Labels range between 100 and 2400

Results (F1-score)

CTD	NID			PWD			MD		
Method	CTU13	UNB15	IDS17	Mend	UCI	δPhish	DREBIN	Ember	AndMal
$\overline{SL}$	0.979	0.942	0.989	0.958	0.974	0.958	0.907	0.970	0.986
SL	0.611	0.447	0.878	0.852	0.884	0.780	0.480	0.667	0.910
$\underline{SsL}$	0.613	0.447	0.879	0.852	0.886	**0.778**	0.486	0.662	0.910
$\pi\ SsL$	0.588	0.437	0.820	0.850	0.884	0.778	0.474	0.647	0.900
$\widehat{\pi}\ SsL$	0.584	0.435	0.818	0.849	0.883	0.777	0.470	0.641	0.890
$\alpha\ SsL_l$	**0.693**	0.582	**0.897**	**0.863**	**0.903**	0.770	**0.546**	**0.687**	**0.924**
$\alpha\ SsL_o$	0.637	0.577	0.874	0.855	0.891	0.745	0.497	0.673	0.916
$\alpha\ SsL_h$	0.510	0.436	0.786	0.834	0.851	0.714	0.423	0.598	0.892
$\alpha^{\pi}SsL_l$	0.664	0.533	0.853	0.861	0.901	0.767	0.529	0.654	0.901
$\alpha^{\pi}SsL_o$	0.633	**0.595**	0.857	0.854	0.890	0.745	0.489	0.647	0.895
$\alpha^{\pi}SsL_h$	0.486	0.427	0.744	0.833	0.851	0.711	0.410	0.579	0.865

Is SsL truly advantageous?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(re)Evaluation

- Massive evaluation on 9 existing datasets for 3 cyberthreat detection tasks:
  - Network Intrusion Detection (NID)
  - Phishing Website Detection (PWD)
  - Malware Detection (MD)

Labels range between 100 and 2400

Results (F1-score)

CTD	NID			PWD			MD		
Method	CTU13	UNB15	IDS17	Mend	UCI	δPhish	DREBIN	Ember	AndMal
$\overline{SL}$	0.979	0.942	0.989	0.958	0.974	0.958	0.907	0.970	0.986
SL	0.611	0.447	0.878	0.852	0.884	0.780	0.480	0.667	0.910
$\underline{SsL}$	0.613	0.447	0.879	0.852	0.886	**0.778**	0.486	0.662	0.910
$\pi\, SsL$	0.588	0.437	0.820	0.850	0.884	0.778	0.474	0.647	0.900
$\hat{\pi}\, SsL$	0.584	0.435	0.818	0.849	0.883	0.777	0.470	0.641	0.890
$\alpha\, SsL_l$	**0.693**	0.582	**0.897**	**0.863**	**0.903**	0.770	**0.546**	**0.687**	**0.924**
$\alpha\, SsL_o$	0.637	0.577	0.874	0.855	0.891	0.745	0.497	0.673	0.916
$\alpha\, SsL_h$	0.510	0.436	0.786	0.834	0.851	0.714	0.423	0.598	0.892
$\alpha^{\pi} SsL_l$	0.664	0.533	0.853	0.861	0.901	0.767	0.529	0.654	0.901
$\alpha^{\pi} SsL_o$	0.633	**0.595**	0.857	0.854	0.890	0.745	0.489	0.647	0.895
$\alpha^{\pi} SsL_h$	0.486	0.427	0.744	0.833	0.851	0.711	0.410	0.579	0.865

Statistical Validation

		Best 'pure' pseudo-labelling			Best active learning		
Dataset	PopSize	Method	p-value	z-value	Method	p-value	z-value
CTU13	396	$\underline{SsL}$	**0.873**	0.159	$\alpha\, SsL_l$	< 0.001	4.310
UNB15	1104	$\underline{SsL}$	**0.964**	–0.044	$\alpha^{\pi} SsL_o$	< 0.001	15.98
IDS17	540	$\underline{SsL}$	**0.932**	0.085	$\alpha\, SsL_l$	**0.978**	–0.027
UCI	1200	$\underline{SsL}$	**0.473**	0.717	$\alpha\, SsL_l$	< 0.001	7.386
Mend.	1200	$\underline{SsL}$	**0.713**	0.368	$\alpha\, SsL_l$	< 0.001	6.757
δPhish	1200	$\underline{SsL}$	**0.554**	–0.590	$\alpha\, SsL_l$	0.002	–3.113
Drebin	1200	$\underline{SsL}$	**0.310**	1.015	$\alpha\, SsL_l$	< 0.001	11.78
Ember	1200	$\underline{SsL}$	**0.603**	–0.512	$\alpha\, SsL_l$	< 0.001	3.407
AndMal	1200	$\underline{SsL}$	**0.712**	–0.370	$\alpha\, SsL_l$	< 0.001	12.01

Adversarial Attacks against Humans and Machine Learning

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Scenario

- ML is used not only for cybersecurity, but for a plethora of other applications
- In some cases, the "decision making" is based on:
  - The output of a *ML model*
  - The interpretation of a *human* to such output

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Scenario

- ML is used not only for cybersecurity, but for a plethora of other applications
- In some cases, the "decision making" is based on:
  - The output of a *ML model*
  - The interpretation of a *human* to such output

- Case in point: online marketplace
  - A person wants to sell an item (e.g., a car)
  - This person (i.e., the seller) uploads the images of such an item on an online marketplace
  - The marketplace automatically provides an estimate of the "value" of the corresponding item
    - This is done via ML
  - Another person (i.e., a potential buyer) looks at the images, then looks at the "suggested" price, and determines whether to buy or not the corresponding item
    - The human uses the output of the ML model to make their decisions

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Attack – what if…

- What if the seller has malicious intentions?

→ The seller may want to induce the ML model to estimate a higher price

- Doing this by introducing "imperceptible" perturbations may trick the ML…
- …but not the human!

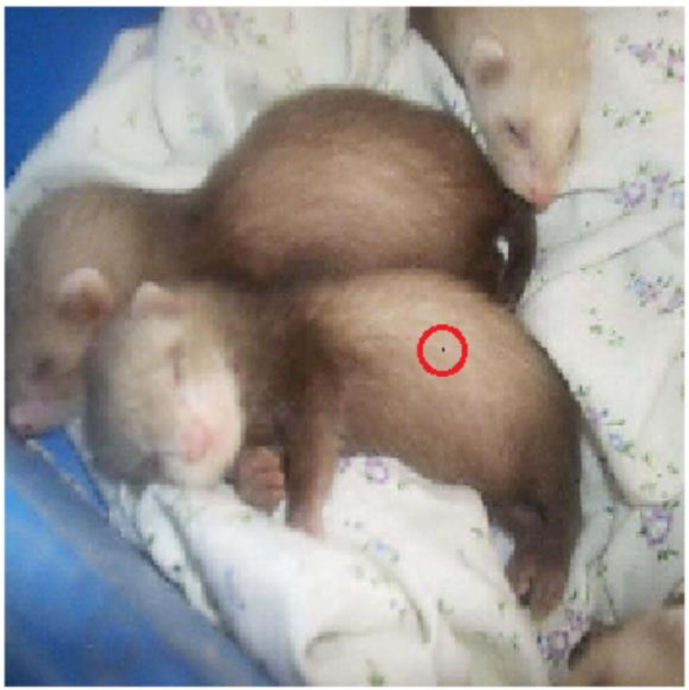
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Attack – what if…

- What if the seller has malicious intentions?

→ The seller may want to induce the ML model to estimate a higher price

- Doing this by introducing "imperceptible" perturbations may trick the ML…
- …but not the human!



Hamster(35.79%)

Nipple(42.36%)

Reference: Su Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." *IEEE Transactions on Evolutionary Computation* (2019)

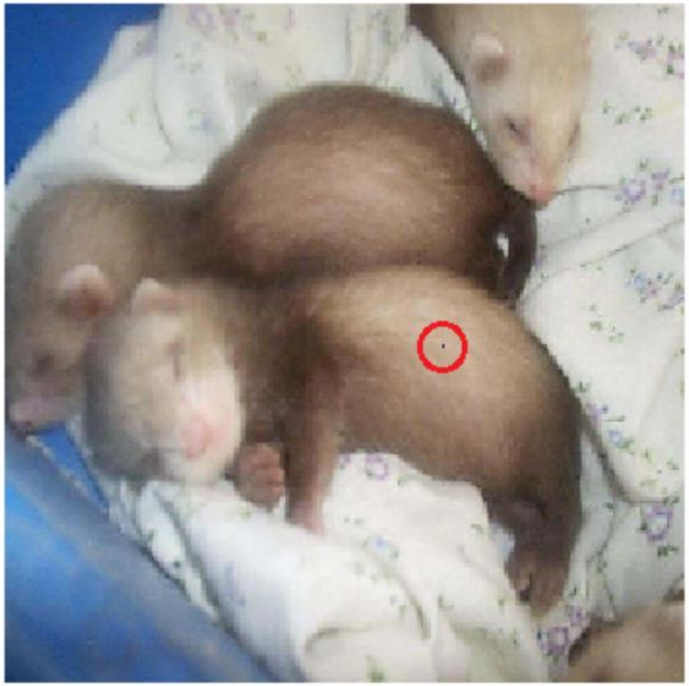
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Attack – what if…

- What if the seller has malicious intentions?

→ The seller may want to induce the ML model to estimate a higher price

- Doing this by introducing "imperceptible" perturbations may trick the ML…
- …but not the human!



Hamster(35.79%)
Nipple(42.36%)

In some cases, "imperceptible" perturbations **may not be what an attacker wants!**

This is especially true when there is a "human-in-the-loop".

Reference: Su Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." *IEEE Transactions on Evolutionary Computation* (2019)

Solution (high-level)

- If humans are involved in the “decision making” process, then such humans will react to clearly incorrect outputs of ML models.
  - Humans may suspect an adversarial attack taking place; or
  - They may think that the ML model is faulty, and hence not trust/believe its output
  - Both of the above are **detrimental** for the attacker!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Solution (high-level)

- If humans are involved in the “decision making” process, then such humans will react to clearly incorrect outputs of ML models.
  - Humans may suspect an adversarial attack taking place; or
  - They may think that the ML model is faulty, and hence not trust/believe its output
  - Both of the above are **detrimental** for the attacker!

(Malicious) solution: deceive both the human *and* the ML model!

- A ML model that thinks that a “FIAT Panda” is a “VW Polo” will output a very high price
  - But if the “perturbation” only affects a single pixel, nobody will fall for it!
- A FIAT Panda is clearly different than a VW Polo, so the perturbation (whatever it is) must be *perceived* by the human

→ The FIAT Panda must be changed in such a way that the human can be somewhat fooled
  - E.g.: the human should think that “it could be a Panda… but it could also be a Polo”

- FIAT Panda MSRP: ~10k $
- VW Polo MSRP: ~20k $

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Solution (low-level)

- How to achieve this in practice?

Concept-based Adversarial Attacks

- The idea is using "explainability" techniques to create adversarial examples.

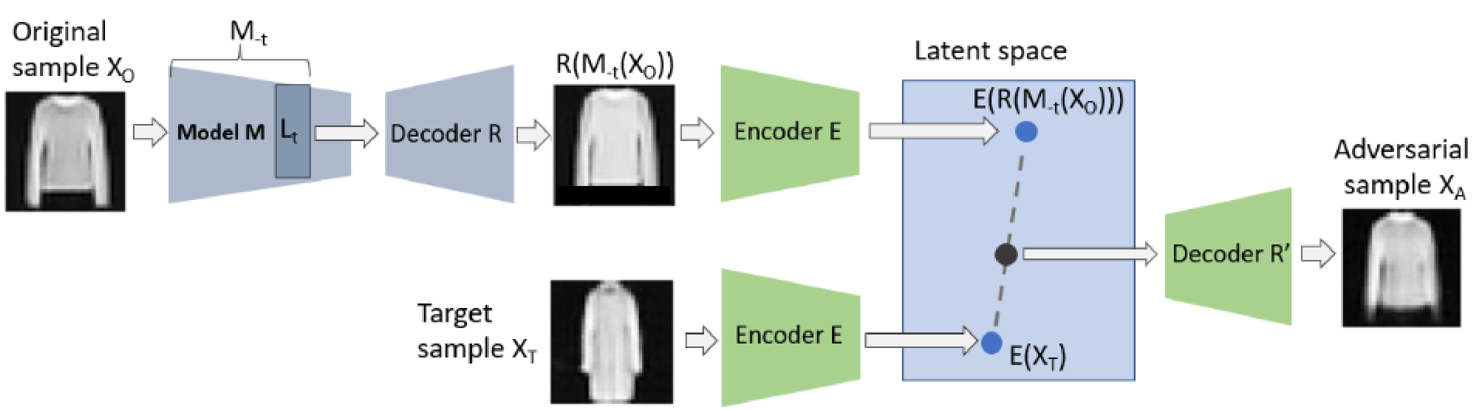
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Solution (low-level)

- How to achieve this in practice?

Concept-based Adversarial Attacks

- The idea is using “explainability” techniques to create adversarial examples.
- Requirements:
  - An “original sample” (i.e., a FIAT Panda)
  - A desired “target sample” (i.e., a VW Polo)
  - A given magnitude of the perturbation (neither too big nor too small)
    - If the FIAT Panda “becomes” a VW Polo, then the adversarial attack would be unfair
    - …and the “buyer” will complain ☺
  - The details of a ML model (which must be based on Convolutional Neural Networks)
    - These attacks can be transferred!
- Output: an “adversarial example” that is a mix between the original and target sample



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Experiments

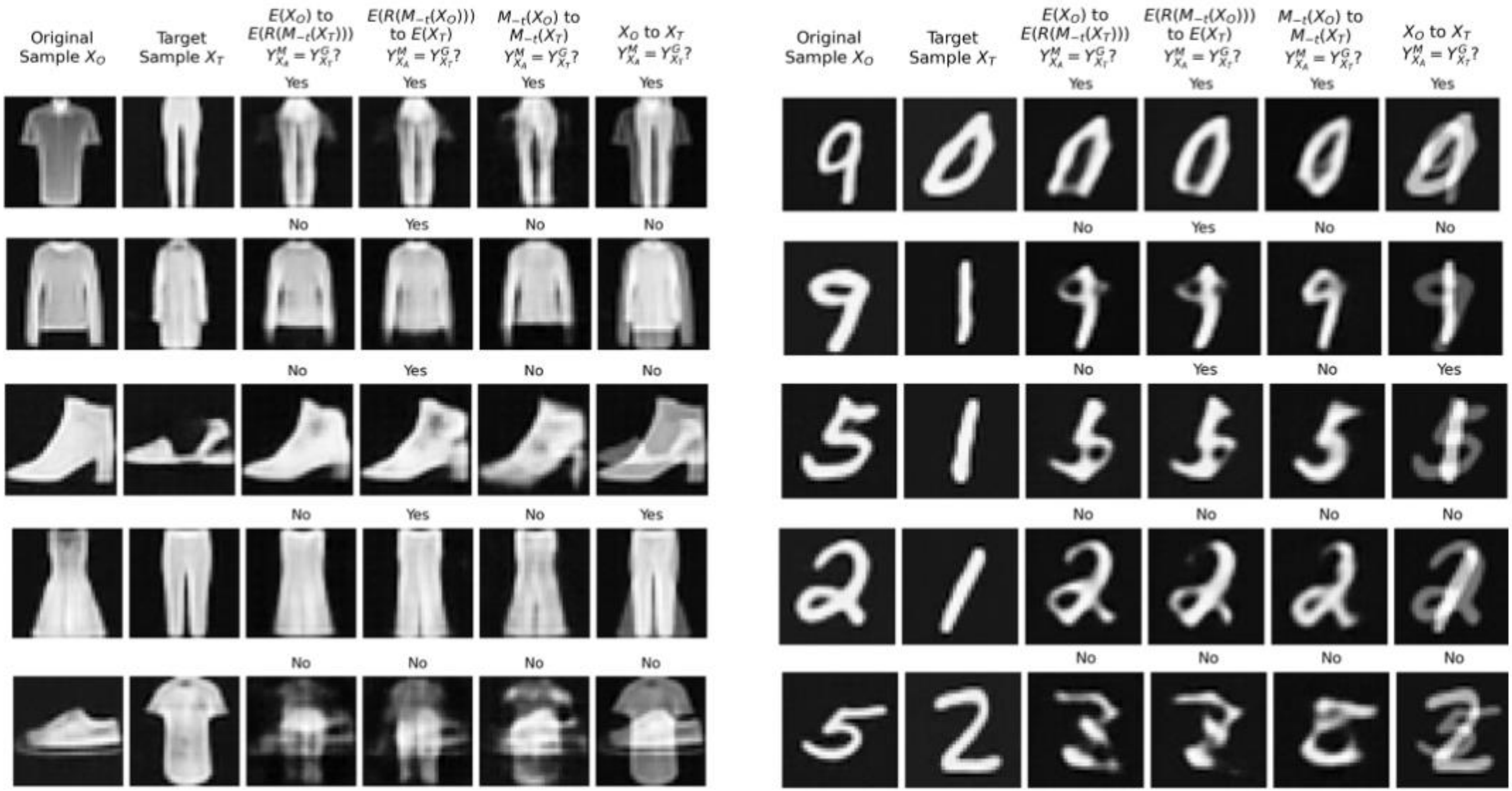


Fig. 2: Original, target and adversarial samples for different en-/decodings and interpolation for Fashion-MNIST(left) and MNIST(right). Yes/No indicates, whether the model got fooled by X_A, i.e. it outputs the class of X_T for X_A

Some Pragmatic Relationships between **Machine Learning & Cybersecurity**

Giovanni Apruzzese, PhD

May 17th, 2022

UNIVERSITÄT LIECHTENSTEIN