Part 1/2

No Mercy for Manual Entry

29/Sep/2021
Workshop @ AMLD2021

ONE swiss bank

SamurAI

AMLD EPFL

Wifi Info

Network:	Free_STCC
User ID :	2317791957
Password:	5197

ONE swiss bank SamurAI

AMLD EPFL

Authors

Valerio Rossetti, PhD
Co-founder of SamurAI
Senior Data Scientist

Giulio Grossi, PhD
Senior Quantitative Portfolio Manager
at ONE swiss bank

ONE swiss bank SamurAI

AMLD EPFL

Get to know the audience

- Knowledge in Supervised Learning and Computer Vision
- Your coding skills in Python
- Do you have practical applications of the techniques in this workshop?

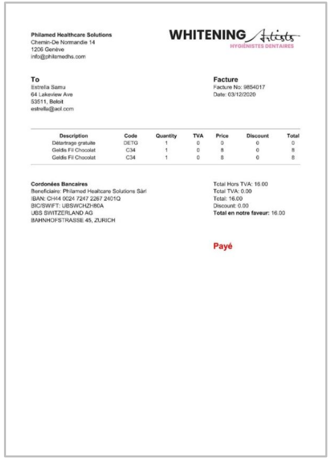
Manual Entry

- A lot of information in paper/pdf documents: invoices, contracts, personal information, surveys …

Manual Entry

- A lot of information in paper/pdf documents: invoices, contracts, personal information, surveys ...
- These documents are then treated manually:
    - costly,
    - time-consuming and
    - error-prone

Manual image-to-data process



Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING Artists
HYGIÉNISTES DENTAIRES

To
Estrella Samu
64 Lakeview Ave
53511, Beloit
estrella@aol.com

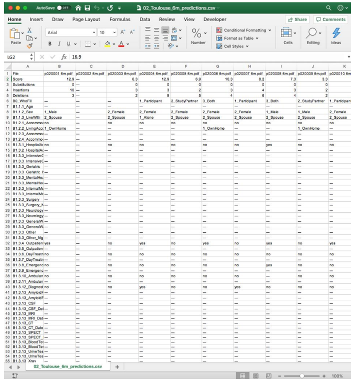
Facture
Facture No: 9854017
Date: 03/12/2020

Description	Code	Quantity	TVA	Price	Discount	Total
Détartrage gratuite	DETG	1	0	0	0	0
Geldis Fil Chocolat	C34	1	0	8	0	8
Geldis Fil Chocolat	C34	1	0	8	0	8

Cordonées Bancaires
Beneficiaire: Philamed Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 16.00
Total TVA: 0.00
Total: 16.00
Discount: 0.00
Total en notre faveur: 16.00

Payé



LG2 16.9

	A	B	C	D	E	F	G	H	I	J	K
1	File	p020001 6m.pdf	p020002 6m.pdf	p020003 6m.pdf	p020004 6m.pdf	p020005 6m.pdf	p020006 6m.pdf	p020007 6m.pdf	p020008 6m.pdf	p020009 6m.pdf	p020010 6m
2	Score	12.9	—	6.3	12.9	6.9	10.3	8.2	7.3	3.3	
3	Substitutions	0	—	0	0	0	0	0	0	0	
4	Insertions	10	—	3	3	2	3	4	3	2	
5	Deletions	3	—	2	9	5	4	6	4	2	
6	B0_WhoFill	—	—	—	1_Participant	2_StudyPartner	3_Both	1_Participant	3_Both	2_StudyPartner	1_Participan
7	B1.1.1_Age	—	—	—	—	—	—	—	—	—	—
8	B1.1.2_Sex	1_Male	—	2_Female	2_Female	2_Female	2_Female	2_Female	1_Male	1_Male	2_Female
9	B1.1.3_LiveWith	2_Spouse	—	2_Spouse	1_Alone	2_Spouse	2_Spouse	2_Spouse	2_Spouse	2_Spouse	2_Spouse
10	B1.2.1_Accommo	no	—	no	no	no	no	no	no	no	no
11	B1.2.2_LivingAcc	1_OwnHome	—	—	—	—	1_OwnHome	—	—	1_OwnHome	—
12	B1.2.3_Accommo	—	—	—	—	—	—	—	—	—	—
13	B1.2.4_Accommo	—	—	—	—	—	—	—	—	—	—
14	B1.3.1_HospitalAc	no	—	no	no	no	no	no	yes	no	no
15	B1.3.2_HospitalAc	—	—	—	—	—	—	—	—	—	—
16	B1.3.3_IntensiveC	—	—	—	—	—	—	—	—	—	—
17	B1.3.3_IntensiveC	—	—	—	—	—	—	—	—	—	—
18	B1.3.3_Geriatric	—	—	—	—	—	—	—	—	—	—
19	B1.3.3_Geriatric_N	—	—	—	—	—	—	—	—	—	—
20	B1.3.3_MentalHea	—	—	—	—	—	—	—	—	—	—
21	B1.3.3_MentalHea	—	—	—	—	—	—	—	—	—	—
22	B1.3.3_InternalMe	—	—	—	—	—	—	—	—	—	—
23	B1.3.3_InternalMe	—	—	—	—	—	—	—	—	—	—
24	B1.3.3_Surgery	—	—	—	—	—	—	—	—	—	—
25	B1.3.3_Surgery_N	—	—	—	—	—	—	—	—	—	—
26	B1.3.3_Neurology	—	—	—	—	—	—	—	—	—	—
27	B1.3.3_Neurology	—	—	—	—	—	—	—	—	—	—
28	B1.3.3_GeneralW	—	—	—	—	—	—	—	—	—	—
29	B1.3.3_GeneralW	—	—	—	—	—	—	—	—	—	—
30	B1.3.3_Other	—	—	—	—	—	—	—	—	—	—
31	B1.3.3_Other_Nig	—	—	—	—	—	—	—	—	—	—
32	B1.3.4_Outpatient	yes	—	no	yes	no	yes	no	yes	no	yes
33	B1.3.5_Outpatient	—	—	—	—	—	—	—	—	—	—
34	B1.3.6_DayTreatn	no	—	no	no	no	no	no	no	no	no
35	B1.3.7_DayTreatn	—	—	—	—	—	—	—	—	—	—
36	B1.3.8_Emergenc	no	—	no	no	no	no	no	yes	no	no
37	B1.3.9_Emergenc	—	—	—	—	—	—	—	—	—	—
38	B1.3.10_Ambulan	no	—	no	no	no	no	no	—	no	no
39	B1.3.11_Ambulan	—	—	—	—	—	—	—	—	—	—
40	B1.3.12_Diagnost	no	—	no	yes	no	no	yes	no	no	no
41	B1.3.13_AmyloidP	—	—	—	—	—	—	—	—	—	—
42	B1.3.13_AmyloidP	—	—	—	—	—	—	—	—	—	—
43	B1.3.13_CSF	—	—	—	—	—	—	—	—	—	—
44	B1.3.13_CSF_Dat	—	—	—	—	—	—	—	—	—	—
45	B1.3.13_MRI	—	—	—	—	—	—	—	—	—	—
46	B1.3.13_MRI_Dat	—	—	—	—	—	—	—	—	—	—
47	B1.3.13_CT	—	—	—	—	—	—	—	—	—	—
48	B1.3.13_CT_Date	—	—	—	—	—	—	—	—	—	—
49	B1.3.13_SPECT	—	—	—	—	—	—	—	—	—	—
50	B1.3.13_SPECT_	—	—	—	—	—	—	—	—	—	—
51	B1.3.13_BloodTes	—	—	—	—	—	—	—	—	—	—
52	B1.3.13_BloodTes	—	—	—	—	—	—	—	—	—	—
53	B1.3.13_UrineTes	—	—	—	—	—	—	—	—	—	—
54	B1.3.13_UrineTes	—	—	—	—	—	—	—	—	—	—
55	B1.3.13_Xray	—	—	—	—	—	—	—	—	—	—

ONE swiss bank

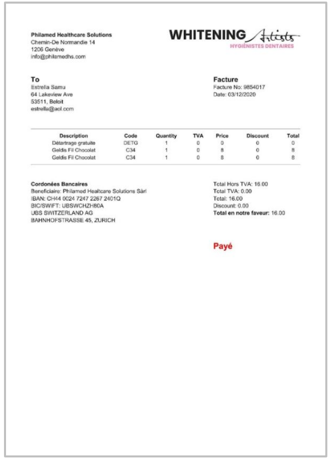
SamurAI

AMLD EPFL

Manual Entry

- A lot of information in paper/pdf documents: invoices, contracts, personal information, surveys ...
- These documents are then treated manually:
  - costly,
  - time-consuming and
  - error-prone
- Main solution: go fully digital and eliminate the paper document!

Manual image-to-data process



Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING Artists
HYGIÉNISTES DENTAIRES

To
Estrella Samu
64 Lakeview Ave
53511, Beloit
estrella@aol.com

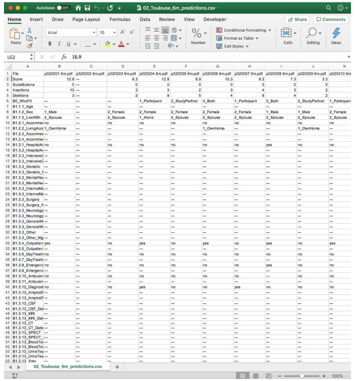
Facture
Facture No: 9854017
Date: 03/12/2020

Description	Code	Quantity	TVA	Price	Discount	Total
Détartrage gratuite	DETG	1	0	0	0	0
Geldis Fil Chocolat	C34	1	0	8	0	8
Geldis Fil Chocolat	C34	1	0	8	0	8

Cordonées Bancaires
Beneficiaire: Philamed Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 16.00
Total TVA: 0.00
Total: 16.00
Discount: 0.00
Total en notre faveur: 16.00

Payé



LG2 f_x 16.9

	A	B	C	D	E	F	G	H	I	J	K
1	File	p020001 6m.pdf	p020002 6m.pdf	p020003 6m.pdf	p020004 6m.pdf	p020005 6m.pdf	p020006 6m.pdf	p020007 6m.pdf	p020008 6m.pdf	p020009 6m.pdf	p020010 6m
2	Score	12.9	—	6.3	12.9	6.9	10.3	8.2	7.3	3.3	
3	Substitutions	0	—	0	0	0	0	0	0	0	
4	Insertions	10	—	3	3	2	3	4	3	2	
5	Deletions	3	—	2	9	5	4	6	4	2	
6	B0_WhoFill	—	—	—	1_Participant	2_StudyPartner	3_Both	1_Participant	3_Both	2_StudyPartner	1_Participan
7	B1.1.1_Age	—	—	—	—	—	—	—	—	—	—
8	B1.1.2_Sex	1_Male	—	2_Female	2_Female	2_Female	2_Female	2_Female	1_Male	1_Male	2_Female
9	B1.1.3_LiveWith	2_Spouse	—	2_Spouse	1_Alone	2_Spouse	2_Spouse	2_Spouse	2_Spouse	2_Spouse	2_Spouse
10	B1.2.1_Accommo	no	—	no	no	no	no	no	no	no	no
11	B1.2.2_LivingAcc	1_OwnHome	—	—	—	—	1_OwnHome	—	—	1_OwnHome	—
12	B1.2.3_Accommo	—	—	—	—	—	—	—	—	—	—
13	B1.2.4_Accommo	—	—	—	—	—	—	—	—	—	—
14	B1.3.1_HospitalAc	no	—	no	no	no	no	no	yes	no	no
15	B1.3.2_HospitalAc	—	—	—	—	—	—	—	—	—	—
16	B1.3.3_IntensiveC	—	—	—	—	—	—	—	—	—	—
17	B1.3.3_IntensiveC	—	—	—	—	—	—	—	—	—	—
18	B1.3.3_Geriatric	—	—	—	—	—	—	—	—	—	—
19	B1.3.3_Geriatric_I	—	—	—	—	—	—	—	—	—	—
20	B1.3.3_MentalHe	—	—	—	—	—	—	—	—	—	—
21	B1.3.3_MentalHe	—	—	—	—	—	—	—	—	—	—
22	B1.3.3_InternalMe	—	—	—	—	—	—	—	—	—	—
23	B1.3.3_InternalMe	—	—	—	—	—	—	—	—	—	—
24	B1.3.3_Surgery	—	—	—	—	—	—	—	—	—	—
25	B1.3.3_Surgery_N	—	—	—	—	—	—	—	—	—	—
26	B1.3.3_Neurology	—	—	—	—	—	—	—	—	—	—
27	B1.3.3_Neurology	—	—	—	—	—	—	—	—	—	—
28	B1.3.3_GeneralW	—	—	—	—	—	—	—	—	—	—
29	B1.3.3_GeneralW	—	—	—	—	—	—	—	—	—	—
30	B1.3.3_Other	—	—	—	—	—	—	—	—	—	—
31	B1.3.3_Other_Nig	—	—	—	—	—	—	—	—	—	—
32	B1.3.4_Outpatient	yes	—	no	yes	no	yes	no	yes	no	yes
33	B1.3.5_Outpatient	—	—	—	—	—	—	—	—	—	—
34	B1.3.6_DayTreatn	no	—	no	no	no	no	no	no	no	no
35	B1.3.7_DayTreatn	—	—	—	—	—	—	—	—	—	—
36	B1.3.8_Emergenc	no	—	no	no	no	no	no	yes	no	no
37	B1.3.9_Emergenc	—	—	—	—	—	—	—	—	—	—
38	B1.3.10_Ambulan	no	—	no	no	no	no	no	—	no	no
39	B1.3.11_Ambulan	—	—	—	—	—	—	—	—	—	—
40	B1.3.12_Diagnost	no	—	no	yes	no	no	yes	no	no	no
41	B1.3.13_AmyloidP	—	—	—	—	—	—	—	—	—	—
42	B1.3.13_AmyloidP	—	—	—	—	—	—	—	—	—	—
43	B1.3.13_CSF	—	—	—	—	—	—	—	—	—	—
44	B1.3.13_CSF_Dat	—	—	—	—	—	—	—	—	—	—
45	B1.3.13_MRI	—	—	—	—	—	—	—	—	—	—
46	B1.3.13_MRI_Dat	—	—	—	—	—	—	—	—	—	—
47	B1.3.13_CT	—	—	—	—	—	—	—	—	—	—
48	B1.3.13_CT_Date	—	—	—	—	—	—	—	—	—	—
49	B1.3.13_SPECT	—	—	—	—	—	—	—	—	—	—
50	B1.3.13_SPECT_	—	—	—	—	—	—	—	—	—	—
51	B1.3.13_BloodTes	—	—	—	—	—	—	—	—	—	—
52	B1.3.13_BloodTes	—	—	—	—	—	—	—	—	—	—
53	B1.3.13_UrineTes	—	—	—	—	—	—	—	—	—	—
54	B1.3.13_UrineTes	—	—	—	—	—	—	—	—	—	—
55	B1.3.13_Xray	—	—	—	—	—	—	—	—	—	—

02_Toulouse_6m_predictions.csv

ONE swiss bank

SamurAI

AMLD EPFL

Manual Entry: a zombie among us

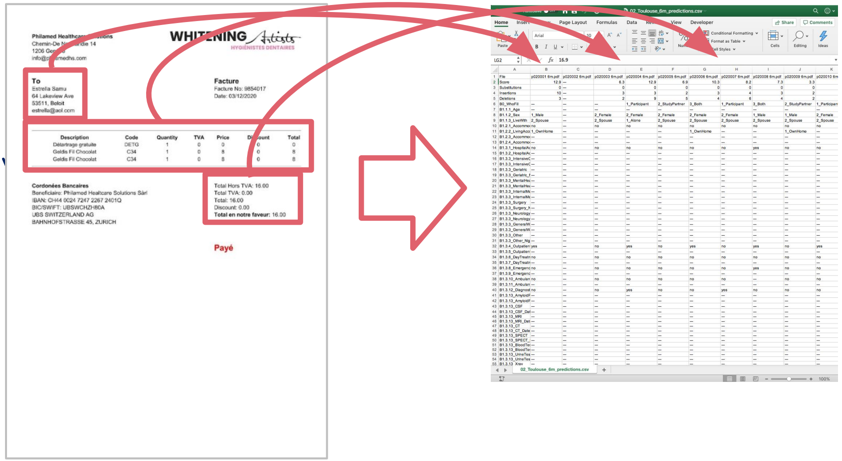
- In some cases we don't manage to eliminate the paper/pdf document

- Examples:
    - Traders send emails with pdf documents treated by the back-office
    - Commodity trading dealing with letters of credit
    - Banks dealing with client documents (passport, ID, scanned contract, ...)
    - Medical research dealing with huge volumes of paper documents from patients
    - Auditors oftentimes compare tables in pdf documents to Excel spreadsheets

This workshop: the goal

- ML can help you to automate processes with scanned paper documents
- This workshop presents few selected techniques to:
  - Classify documents
  - Extract information from documents
- Disclaimer:
  - This workshop doesn’t present the fanciest / more powerful techniques to do these tasks
  - This workshop shows you few techniques that are easy to understand, implement and deploy as a beginner data scientist

Automated image-to-data process



Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING Artists
HYGIÉNISTES DENTAIRES

To
Estrella Samu
64 Lakeview Ave
53511, Beloit
estrella@aol.com

Facture
Facture No: 9854017
Date: 03/12/2020

Description	Code	Quantity	TVA	Price	Discount	Total
Détartrage gratuite	DETG	1	0	0	0	0
Geldis Fil Chocolat	C34	1	0	8	0	8
Geldis Fil Chocolat	C34	1	0	8	0	8

Cordonées Bancaires
Beneficiaire: Philamed Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 16.00
Total TVA: 0.00
Total: 16.00
Discount: 0.00
Total en notre faveur: 16.00

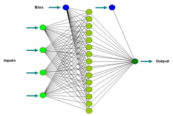
Payé

This workshop

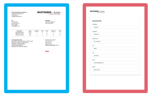
- For beginners
- Get your hands dirty in code
- Adding levels of complexity

Part 1
(~1h30)

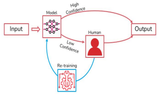
Supervised Learning & ANNs



Document Classification

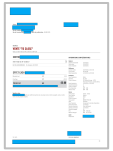


Human In The Loop

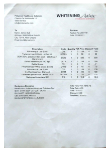


Part 2
(~2h00)

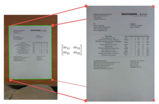
Document with items in fixed position



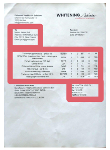
Scanned Documents



Documents from Camera



Document with items in variable position



Workshop: Schedule

Session	Duration	Start - End	Subjects
Part 1	~1:30	13:30 – 15:00	Intro + Document Classification + HITL
Break	0:30	15:00 – 15:30	
Part 2	~2:00	15:30 – 17:30	Information Extraction

ONE swiss bank

SamurAI

AMLD EPFL

Your questions are welcomed!

Your 1st question

=

Swiss chocolate bar

«I don’t get the big picture»

«I didn’t understand what this code does»

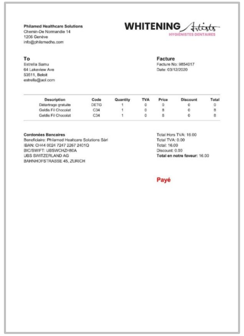
«Why are we doing this?»

ONE swiss bank SamurAI

AMLD EPFL

Supervised Learning

Attributes/
Features



Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING Artists
HYGIÉNISTES DENTAIRES

To
Estrella Samu
64 Lakeview Ave
53511, Beloit
estrella@aol.com

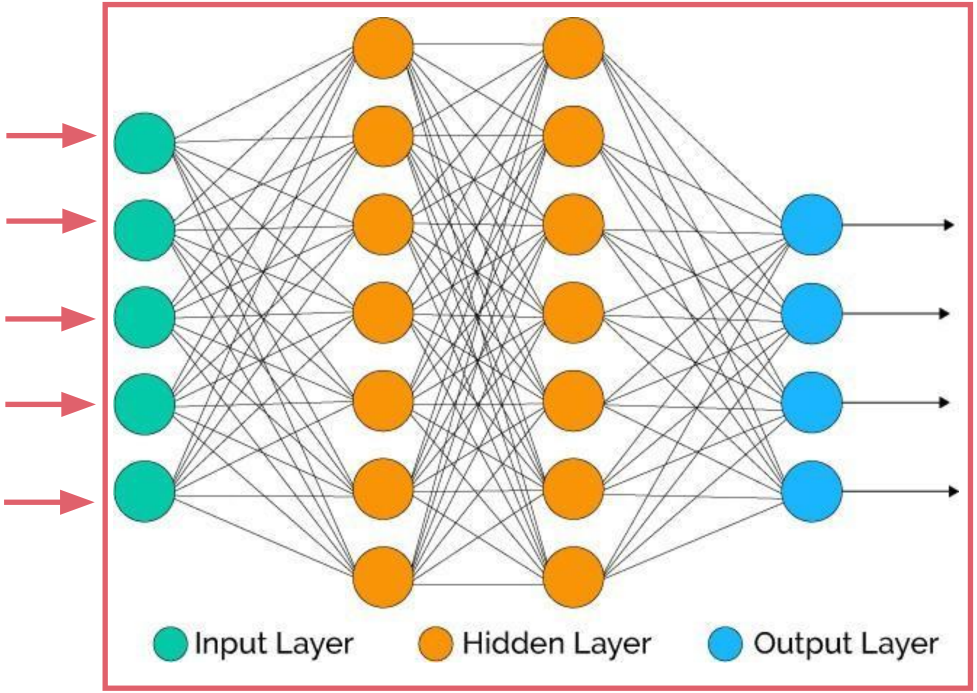
Facture
Facture No: 9854017
Date: 03/12/2020

Description	Code	Quantity	TVA	Price	Discount	Total
Détartrage gratuite	DETG	1	0	0	0	0
Geldis Fil Chocolat	C34	1	0	8	0	8
Geldis Fil Chocolat	C34	1	0	8	0	8

Cordonées Bancaires
Beneficiaire: Philamed Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 16.00
Total TVA: 0.00
Total: 16.00
Discount: 0.00
Total en notre faveur: 16.00

Payé

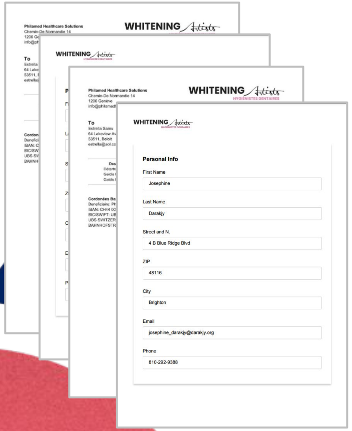


Predictions

It's a
document
Invoice!

Supervised Learning

Attributes/ Features



Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève

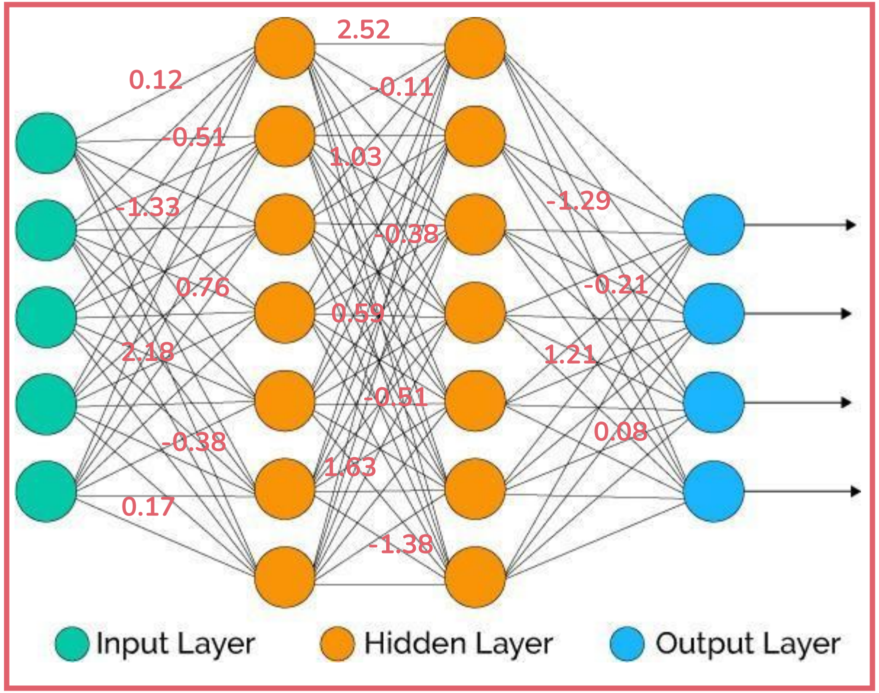
WHITENING Artists
HYGIÉNISTES DENTAIRES

To
Estrella Samu
64 Lakeview Av
53511, Beloit

WHITENING Artists
HYGIÉNISTES DENTAIRES

Personal Info

First Name	Josephine
Last Name	Darakjy
Street and N.	4 B Blue Ridge Blvd
ZIP	48116
City	Brighton
Email	josephine_darakjy@darakjy.org
Phone	810-292-9388

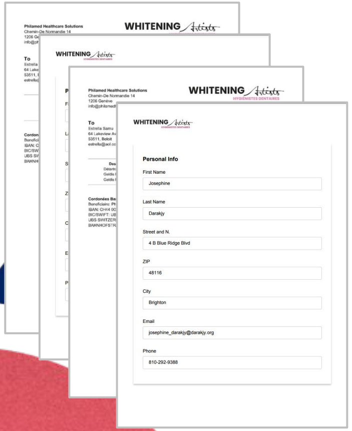


Predictions		Labels
invoice	🟢	invoice
invoice	🔴	registration
registration	🔴	invoice
invoice	🔴	registration
registration	🟢	registration
registration	🔴	invoice
invoice	🔴	registration
invoice	🟢	invoice
invoice	🟢	invoice
registration	🟢	registration

Loss/CostFunction:	1000
Accuracy:	50%

Supervised Learning

Attributes/
Features



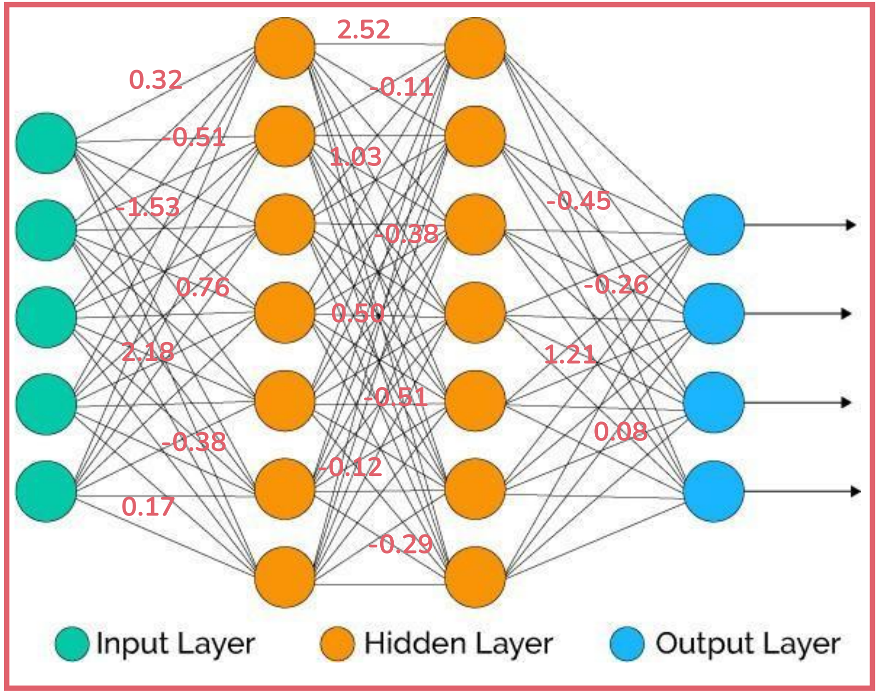
WHITENING Artists
HYGIÉNISTES DENTAIRES

Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève

To
Estrella Samu

Personal Info

First Name	Josephine
Last Name	Darakjy
Street and N.	4 B Blue Ridge Blvd
ZIP	48116
City	Brighton
Email	josephine_darakjy@darakjy.org
Phone	810-292-9388



Predictions		Labels
invoice	🟢	invoice
registration	🟢	registration
registration	🔴	invoice
invoice	🔴	registration
registration	🟢	registration
invoice	🟢	invoice
invoice	🔴	registration
invoice	🟢	invoice
invoice	🟢	invoice
registration	🟢	registration

Loss/CostFunction: 600
Accuracy: 70%

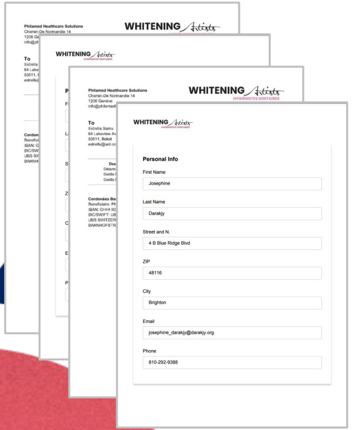
ONE
swiss bank

SamurAI

AMLDEPFL

Supervised Learning

Attributes/ Features



WHITENING Artists

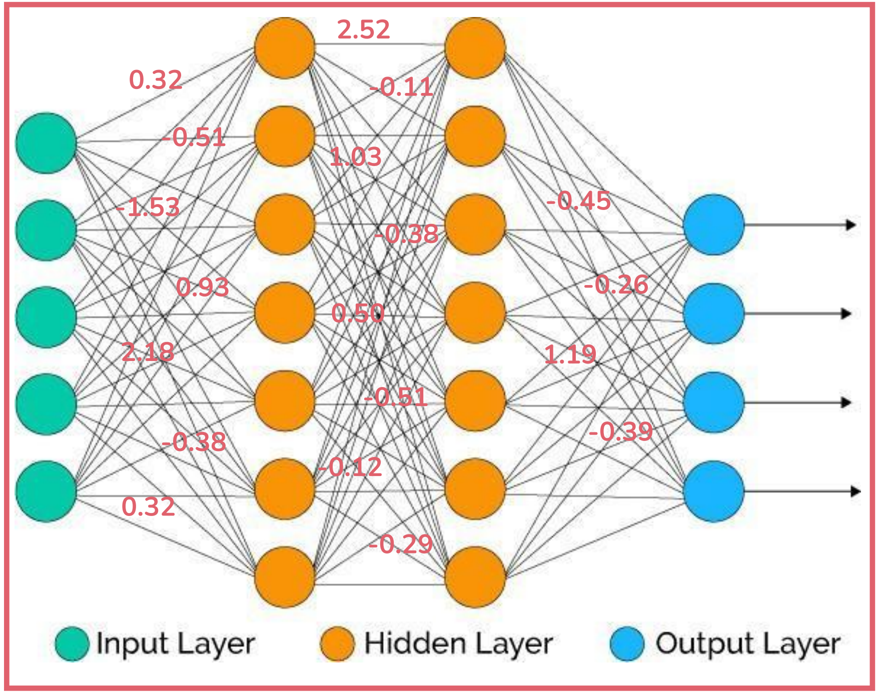
HYGIÉNISTES DENTAIRES

Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève

To
Estrella Samu

Personal Info

First Name	Josephine
Last Name	Darakjy
Street and N.	4 B Blue Ridge Blvd
ZIP	48116
City	Brighton
Email	josephine_darakjy@darakjy.org
Phone	810-292-9388



Predictions		Labels
invoice	🟢	invoice
registration	🟢	registration
invoice	🟢	invoice
registration	🟢	registration
registration	🟢	registration
invoice	🟢	invoice
invoice	🔴	registration
invoice	🟢	invoice
invoice	🟢	invoice
registration	🟢	registration

Loss/CostFunction:	300
Accuracy:	90%

ONE swiss bank

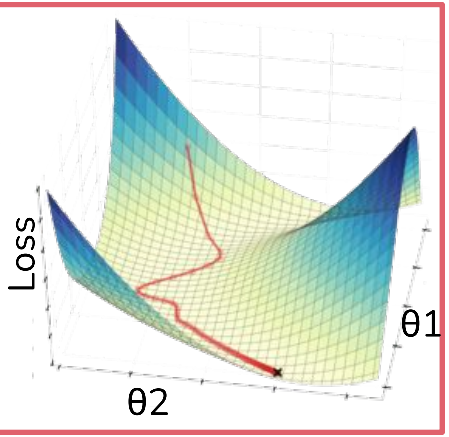
SamurAI

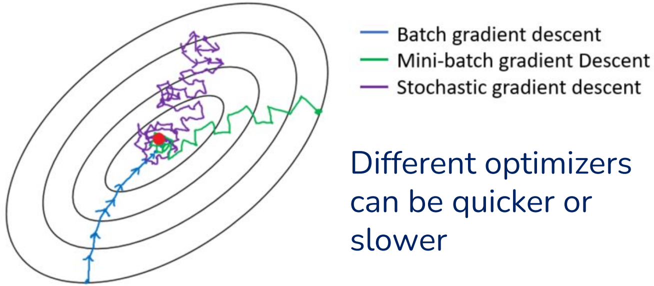
AMLD EPFL

Training as a loss minimization

- The loss quantifies the spread between labels and predictions
- The optimizers are algorithms that find the (possibly absolute) minimum of the loss

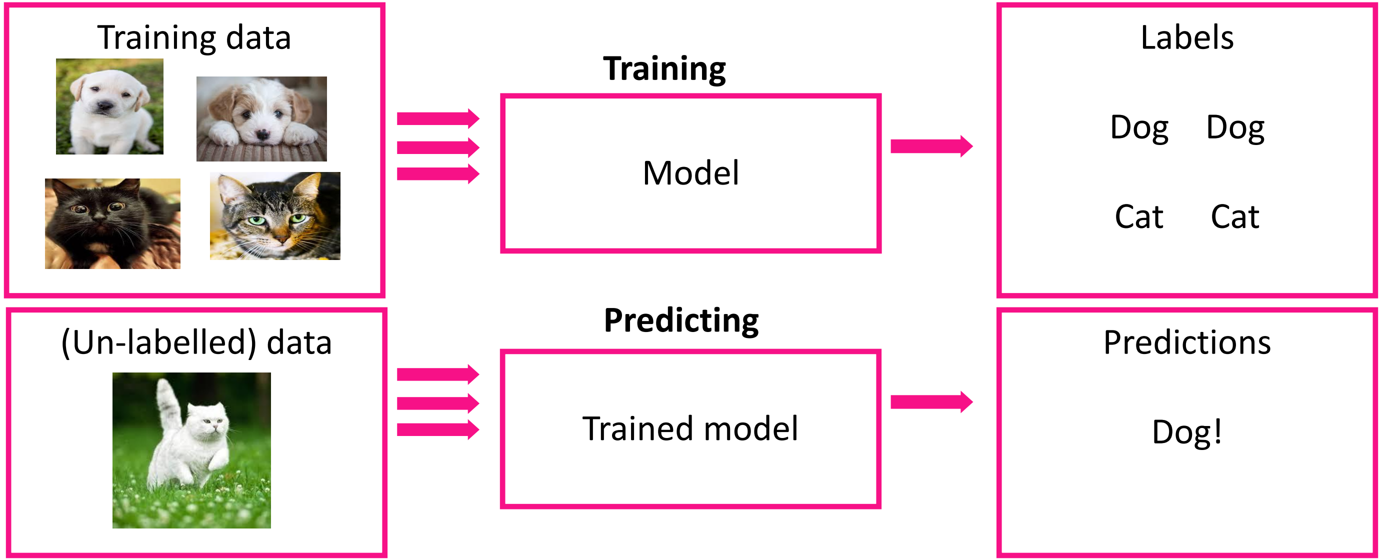
During the training the optimizer finds the path to minimize the loss, like a river flowing downhill





Different optimizers can be quicker or slower

Learning is hard



Training data

Training

Model

Labels

Dog Dog

Cat Cat

(Un-labelled) data

Predicting

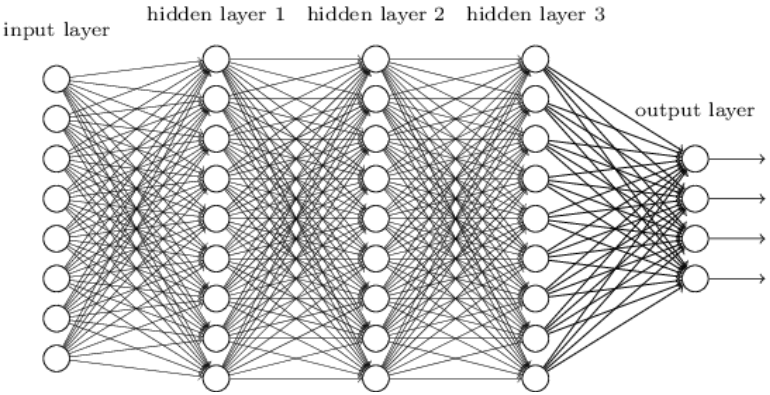
Trained model

Predictions

Dog!

ANN: Dense Layers

- Multi-Layer Perceptrons are the simplest ANN
- Every node of a layer is connected with all the nodes in the previous and in the following layer
- These layers are called Fully connected or Dense

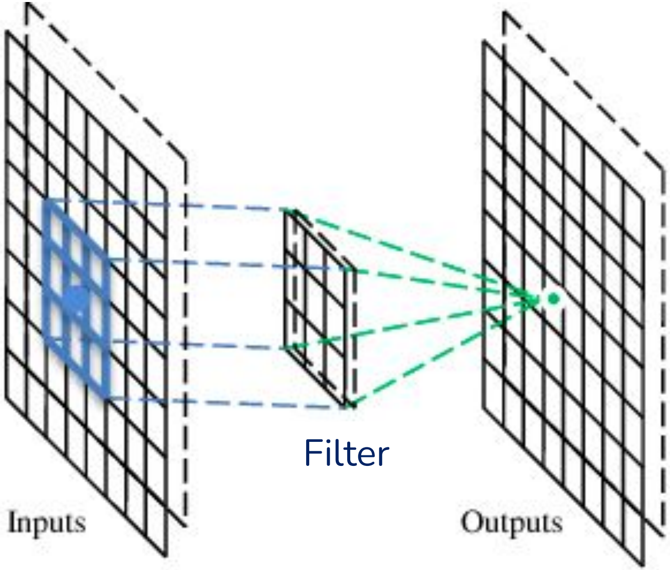


- Image of 500 x 500 x 3 pixels
- First layer 100 nodes
- Already for the first layer we need ~750k parameters

The MLP is great, but too many parameters!

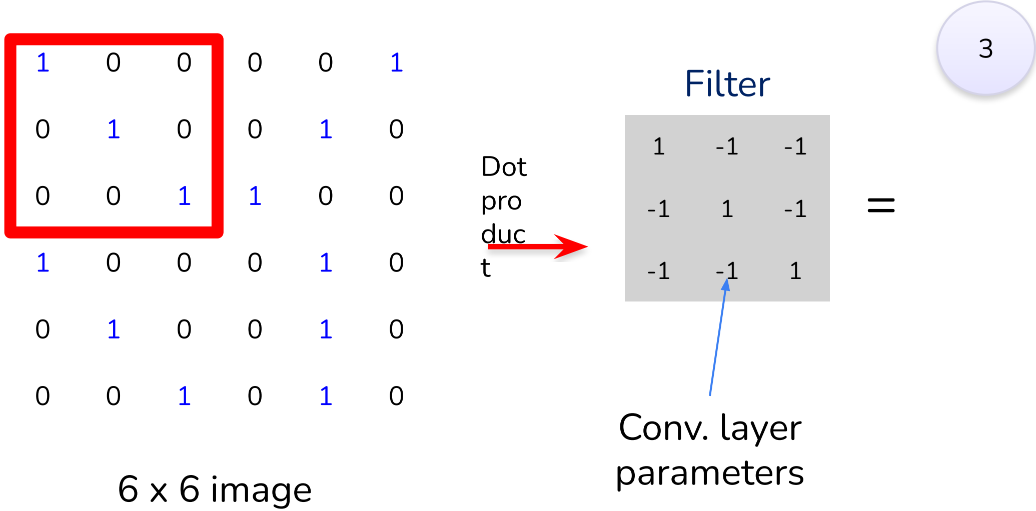
ANN: Convolutional Layers

- We want to reduce the model parameters
- We introduce the concept of filter



ANN: Convolutional Layers

- We want to reduce the model parameters
- We introduce the concept of filter



1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

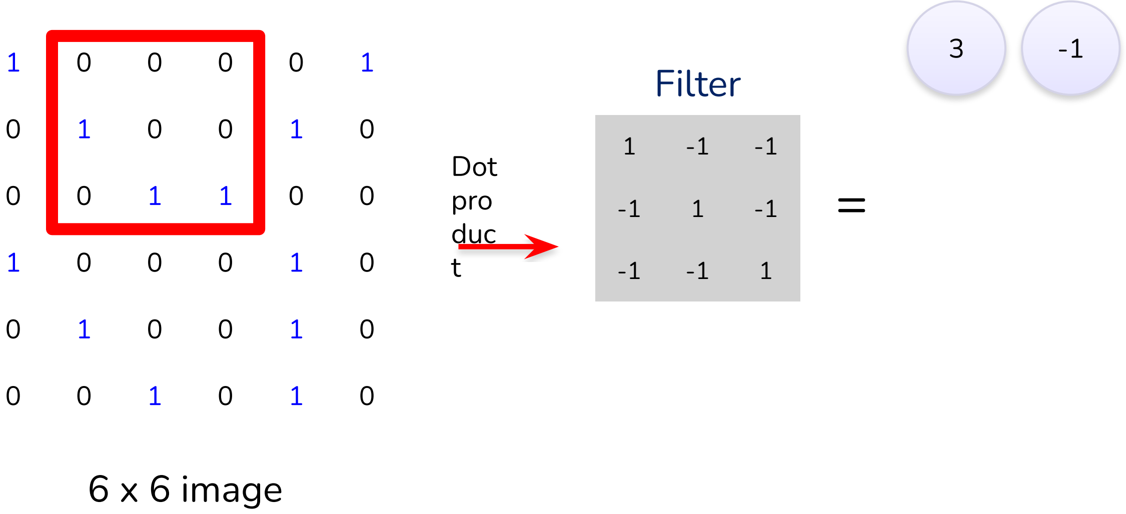
Filter

1	-1	-1
-1	1	-1
-1	-1	1

Conv. layer parameters

ANN: Convolutional Layers

- We want to reduce the model parameters
- We introduce the concept of filter



1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

Dot product

Filter

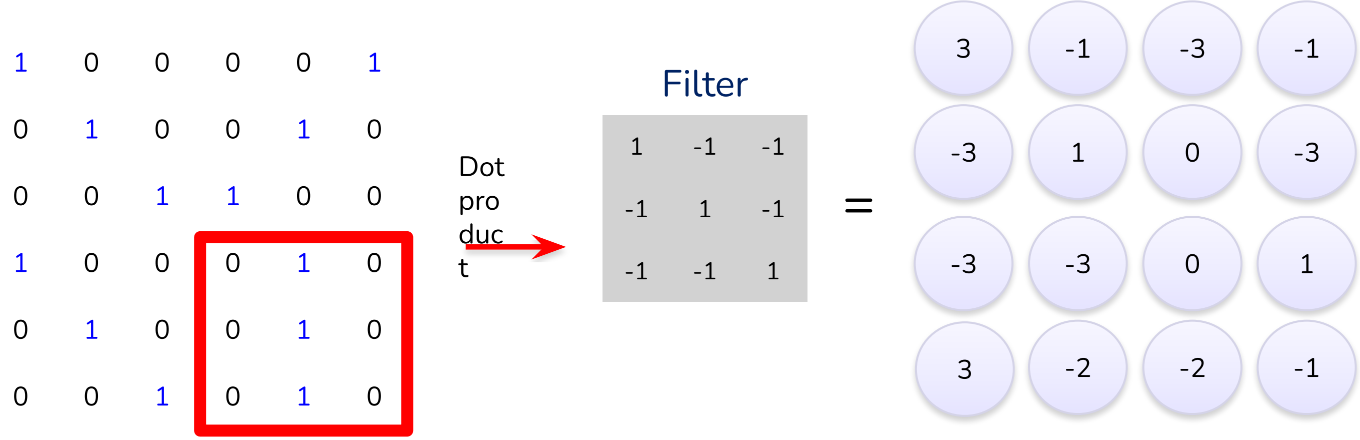
1	-1	-1
-1	1	-1
-1	-1	1

=

3 -1

ANN: Convolutional Layers

- We want to reduce the model parameters
- We introduce the concept of filter



1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

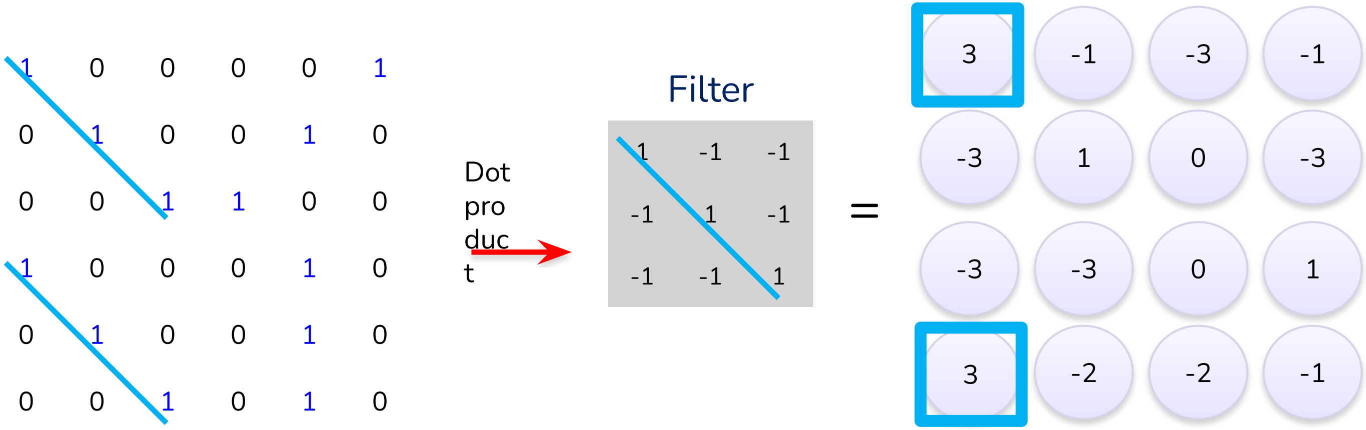
Filter

1	-1	-1
-1	1	-1
-1	-1	1

3	-1	-3	-1
-3	1	0	-3
-3	-3	0	1
3	-2	-2	-1

ANN: Convolutional Layers

- We want to reduce the model parameters
- We introduce the concept of filter



1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

Dot product

Filter

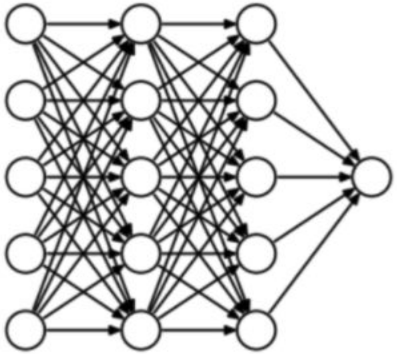
1	-1	-1
-1	1	-1
-1	-1	1

=

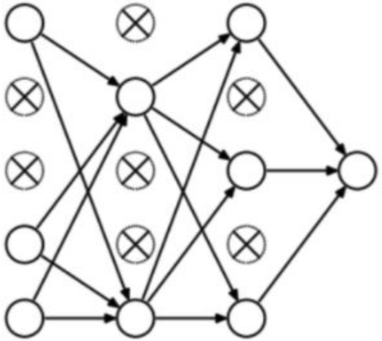
3	-1	-3	-1
-3	1	0	-3
-3	-3	0	1
3	-2	-2	-1

Dropout

- Regularization technique
- A fraction of the nodes are not considered in a training step
- This forces the network to have several "routes" in the nodes to ensure good performances



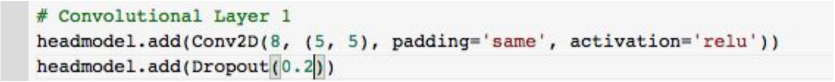
Standard Training



Training with Droupout

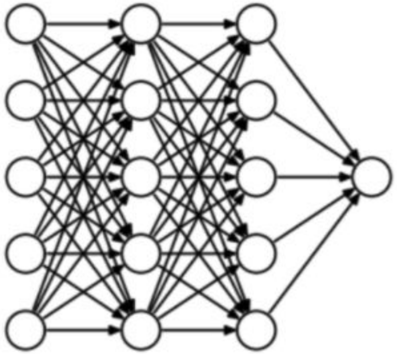
Dropout

- Regularization technique
- A fraction of the nodes are not considered in a training step
- This forces the network to have several “routes” in the nodes to ensure good performances

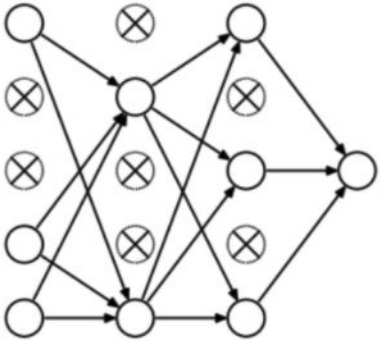


```
# Convolutional Layer 1
headmodel.add(Conv2D(8, (5, 5), padding='same', activation='relu'))
headmodel.add(Dropout(0.2))
```

This means “drop 20% of the nodes in the previous layer while training



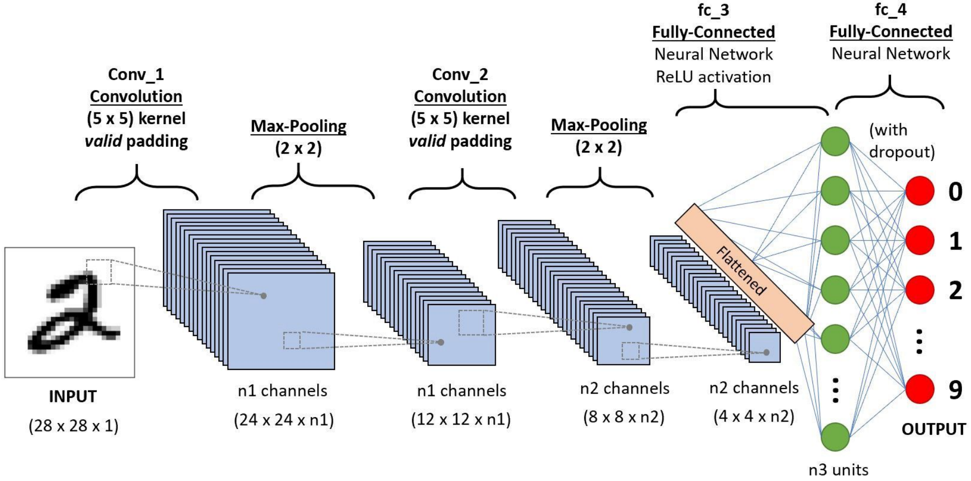
Standard Training



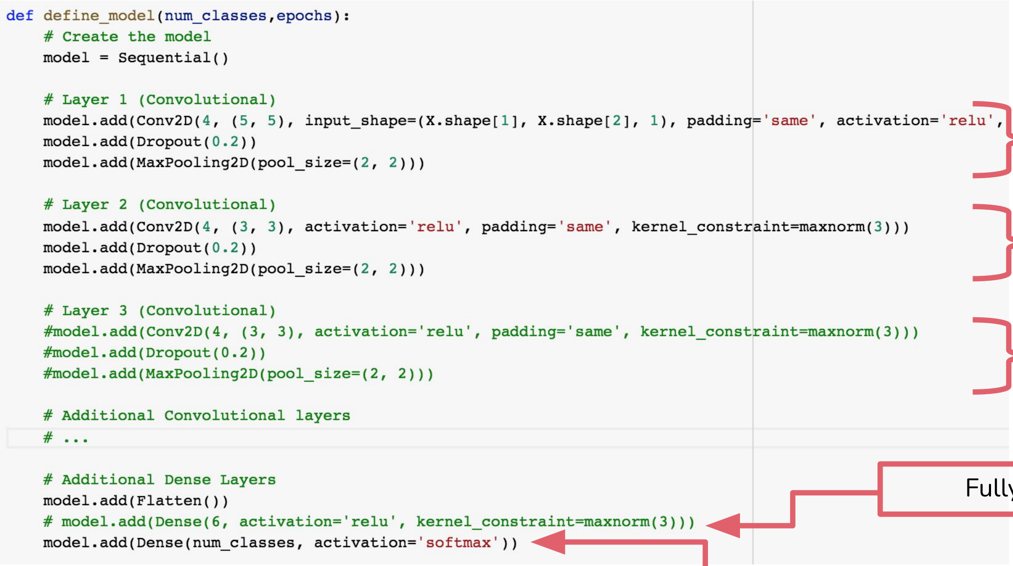
Training with Droupout

Typical CNN structure

- Typical CNNs are made of :
  - A series of convolutional layer + dropout + max-pooling
  - Fully-connected layers at the end
- Many parameters are subject to tuning:
  - Number and type of layer
  - Number of filters
  - Size of filters



CNN in Keras



```
def define_model(num_classes,epochs):
    # Create the model
    model = Sequential()

    # Layer 1 (Convolutional)
    model.add(Conv2D(4, (5, 5), input_shape=(X.shape[1], X.shape[2], 1), padding='same', activation='relu',
    model.add(Dropout(0.2))
    model.add(MaxPooling2D(pool_size=(2, 2)))

    # Layer 2 (Convolutional)
    model.add(Conv2D(4, (3, 3), activation='relu', padding='same', kernel_constraint=maxnorm(3)))
    model.add(Dropout(0.2))
    model.add(MaxPooling2D(pool_size=(2, 2)))

    # Layer 3 (Convolutional)
    #model.add(Conv2D(4, (3, 3), activation='relu', padding='same', kernel_constraint=maxnorm(3)))
    #model.add(Dropout(0.2))
    #model.add(MaxPooling2D(pool_size=(2, 2)))

    # Additional Convolutional layers
    # ...

    # Additional Dense Layers
    model.add(Flatten())
    # model.add(Dense(6, activation='relu', kernel_constraint=maxnorm(3)))
    model.add(Dense(num_classes, activation='softmax'))
```

Conv. Layer + Dropout + Max-Pooling

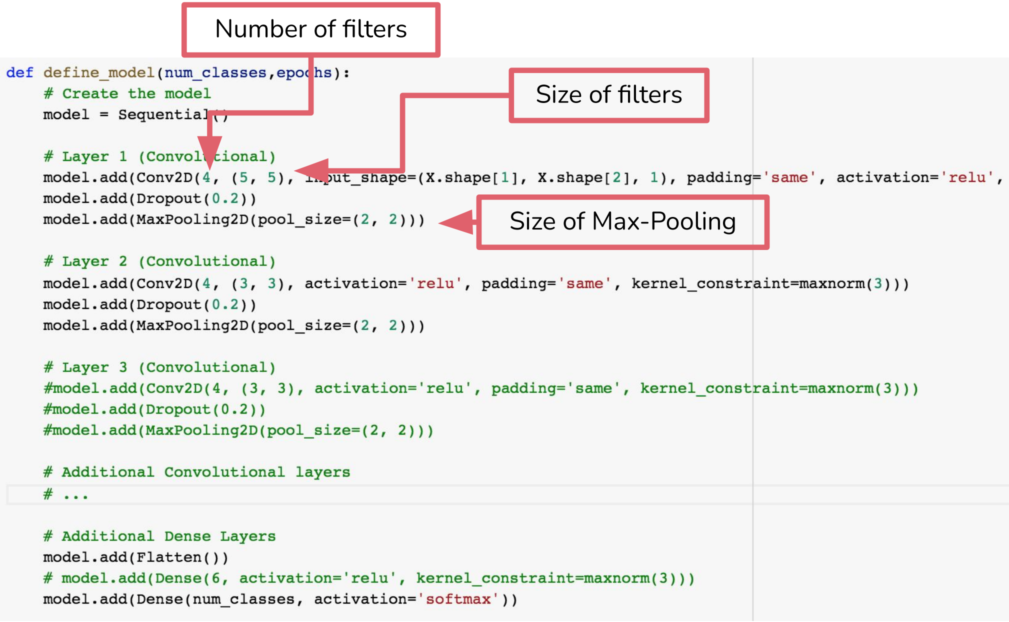
Conv. Layer + Dropout + Max-Pooling

Conv. Layer + Dropout + Max-Pooling

Fully Connected Layers

Output

CNN in Keras



```
def define_model(num_classes,epochs):
    # Create the model
    model = Sequential()

    # Layer 1 (Convolutional)
    model.add(Conv2D(4, (5, 5), input_shape=(X.shape[1], X.shape[2], 1), padding='same', activation='relu',
    model.add(Dropout(0.2))
    model.add(MaxPooling2D(pool_size=(2, 2)))

    # Layer 2 (Convolutional)
    model.add(Conv2D(4, (3, 3), activation='relu', padding='same', kernel_constraint=maxnorm(3)))
    model.add(Dropout(0.2))
    model.add(MaxPooling2D(pool_size=(2, 2)))

    # Layer 3 (Convolutional)
    #model.add(Conv2D(4, (3, 3), activation='relu', padding='same', kernel_constraint=maxnorm(3)))
    #model.add(Dropout(0.2))
    #model.add(MaxPooling2D(pool_size=(2, 2)))

    # Additional Convolutional layers
    # ...

    # Additional Dense Layers
    model.add(Flatten())
    # model.add(Dense(6, activation='relu', kernel_constraint=maxnorm(3)))
    model.add(Dense(num_classes, activation='softmax'))
```

Number of filters

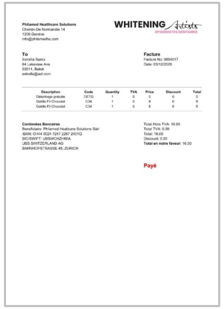
Size of filters

Size of Max-Pooling

Document Classification

- We are going to train an algorithm to discriminate between different types of documents, <u>only using their images</u>
- The dataset:
  - ~400 invoices
  - ~400 registration
  - ~600 other
- We want the model to be robust enough to use images taken from a phone

invoice



Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING Artists
HYGIÉNISTES DENTAIRES

To
Estrella Samu
64 Lakeview Ave
53511, Beloit
estrella@aol.com

Facture
Facture No: 9854017
Date: 03/12/2020

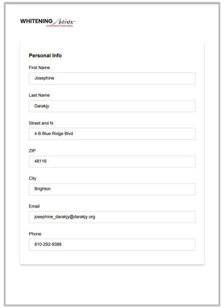
Description	Code	Quantity	TVA	Price	Discount	Total
Détartrage gratuite	DETG	1	0	0	0	0
Geldis Fil Chocolat	C34	1	0	8	0	8
Geldis Fil Chocolat	C34	1	0	8	0	8

Cordonées Bancaires
Beneficiaire: Philamed Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 16.00
Total TVA: 0.00
Total: 16.00
Discount: 0.00
Total en notre faveur: 16.00

Payé

registration



WHITENING Artists
HYGIÉNISTES DENTAIRES

Personal Info

First Name
Josephine

Last Name
Darakjy

Street and N.
4 B Blue Ridge Blvd

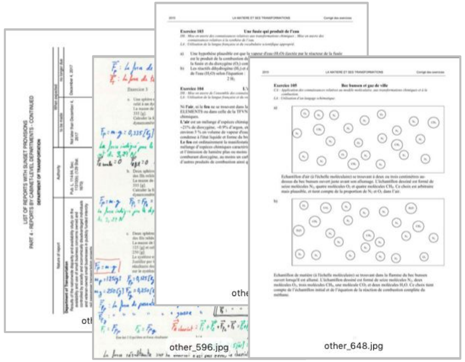
ZIP
48116

City
Brighton

Email
josephine_darakjy@darakjy.org

Phone
810-292-9388

other

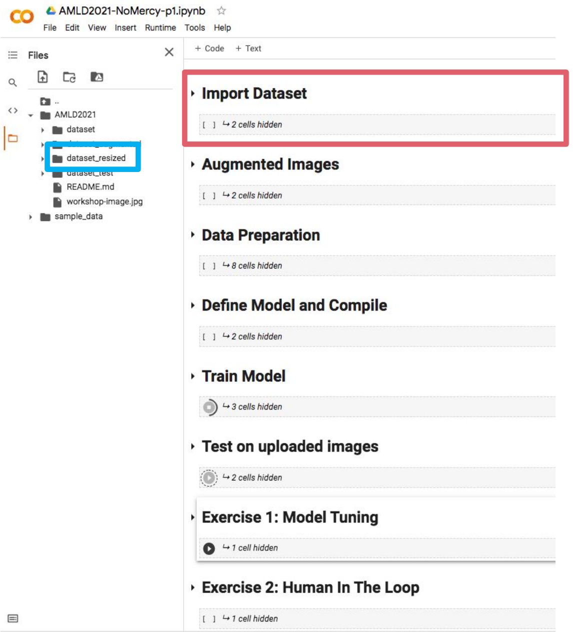


ONE swiss bank

SamurAI

AMLD EPFL

Notebook structure



AMLD2021-NoMercy-p1.ipynb

- ..
- AMLD2021
  - dataset
  - dataset_resized
  - dataset_test
  - README.md
  - workshop-image.jpg
- sample_data

Import Dataset
[ ] ↳ 2 cells hidden

Augmented Images
[ ] ↳ 2 cells hidden

Data Preparation
[ ] ↳ 8 cells hidden

Define Model and Compile
[ ] ↳ 2 cells hidden

Train Model
↳ 3 cells hidden

Test on uploaded images
↳ 2 cells hidden

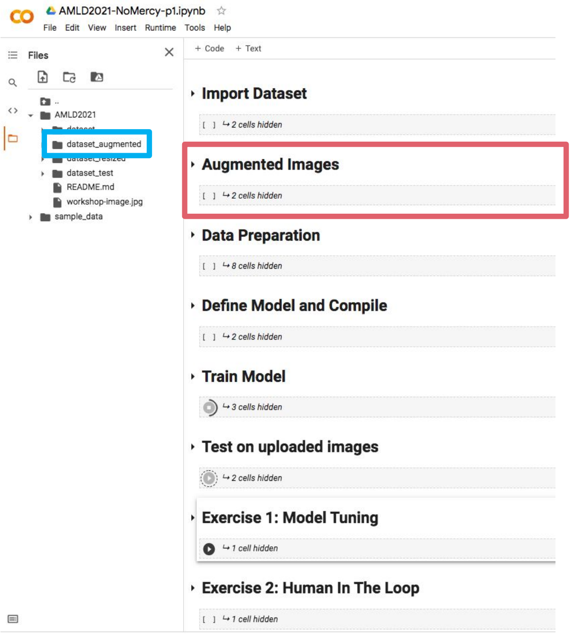
Exercise 1: Model Tuning
↳ 1 cell hidden

Exercise 2: Human In The Loop
[ ] ↳ 1 cell hidden

- Import images from github
- On the left, you will see the folder AMLD2021 appear
- The main dataset used for this notebook is in AMLD2021/dataset_resized/
- ~1600images of 708x500 pixels

ONE swiss bank SamurAI AMLD EPFL

Notebook structure



Import Dataset
[ ] ↳ 2 cells hidden

Augmented Images
[ ] ↳ 2 cells hidden

Data Preparation
[ ] ↳ 8 cells hidden

Define Model and Compile
[ ] ↳ 2 cells hidden

Train Model
↳ 3 cells hidden

Test on uploaded images
↳ 2 cells hidden

Exercise 1: Model Tuning
↳ 1 cell hidden

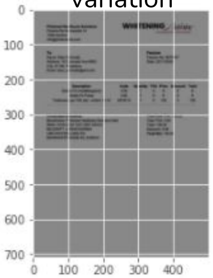
Exercise 2: Human In The Loop
[ ] ↳ 1 cell hidden

- More augmented images are created with distortions, brightness variations, tilt, ...

(un)zoom

rotation

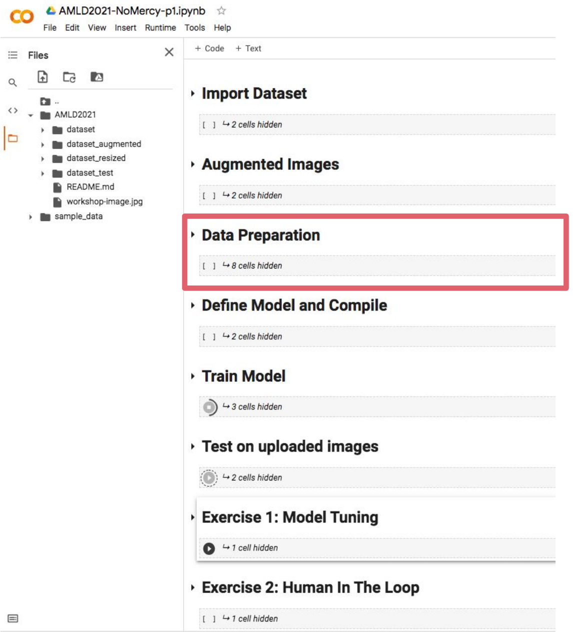
brightness variation



- Augmented images are moved to AMLD2021/dataset_augmented/

ONE swiss bank SamurAI AMLDEPFL

Notebook structure



AMLD2021-NoMercy-p1.ipynb

Import Dataset

[ ] ↳ 2 cells hidden

Augmented Images

[ ] ↳ 2 cells hidden

Data Preparation

[ ] ↳ 8 cells hidden

Define Model and Compile

[ ] ↳ 2 cells hidden

Train Model

↳ 3 cells hidden

Test on uploaded images

↳ 2 cells hidden

Exercise 1: Model Tuning

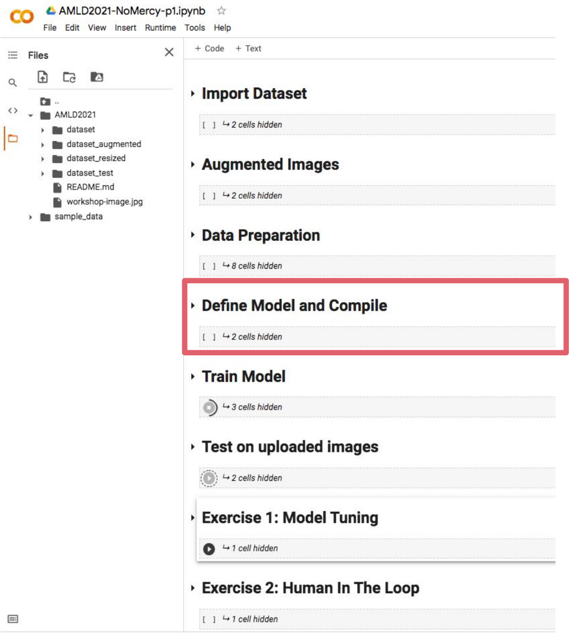
↳ 1 cell hidden

Exercise 2: Human In The Loop

[ ] ↳ 1 cell hidden

- Images are resized to a smaller size
- Dataset is created
- Train-test split

Notebook structure



Import Dataset
[ ] ↳ 2 cells hidden

Augmented Images
[ ] ↳ 2 cells hidden

Data Preparation
[ ] ↳ 8 cells hidden

Define Model and Compile
[ ] ↳ 2 cells hidden

Train Model
↳ 3 cells hidden

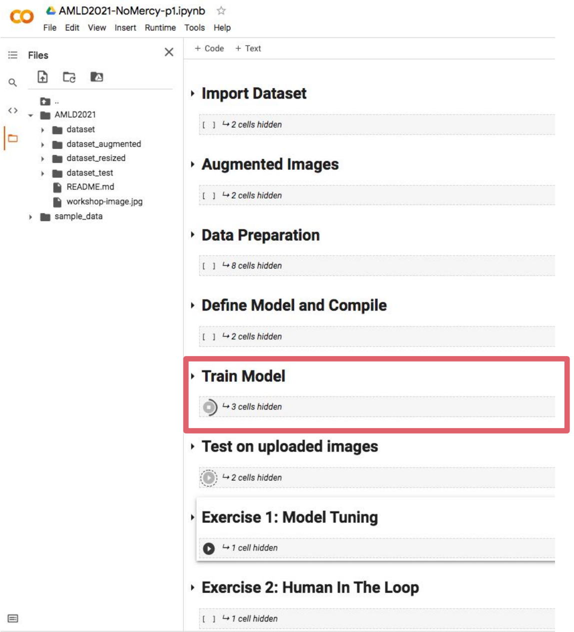
Test on uploaded images
↳ 2 cells hidden

Exercise 1: Model Tuning
↳ 1 cell hidden

Exercise 2: Human In The Loop
[ ] ↳ 1 cell hidden

- Model is defined as a Keras ANN with Conv and Dense layers

Notebook structure



AMLD2021-NoMercy-p1.ipynb

Import Dataset

[ ] ↳ 2 cells hidden

Augmented Images

[ ] ↳ 2 cells hidden

Data Preparation

[ ] ↳ 8 cells hidden

Define Model and Compile

[ ] ↳ 2 cells hidden

Train Model

↳ 3 cells hidden

Test on uploaded images

↳ 2 cells hidden

Exercise 1: Model Tuning

↳ 1 cell hidden

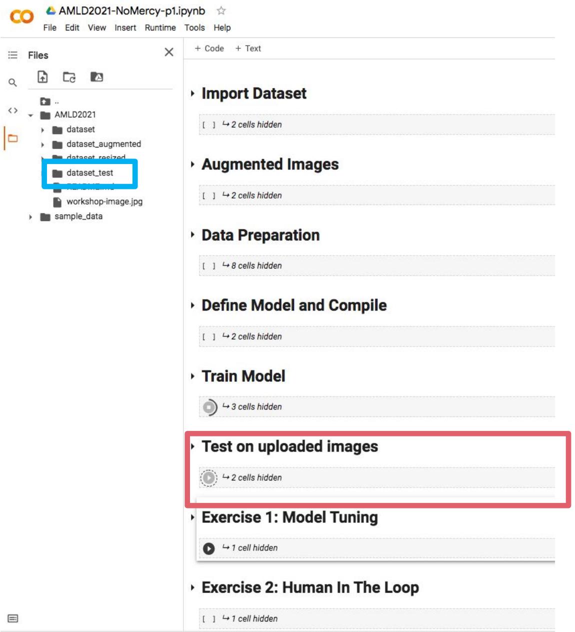
Exercise 2: Human In The Loop

[ ] ↳ 1 cell hidden

- Training and evaluation of performance

ONE swiss bank SamurAI AMLD EPFL

Notebook structure



Import Dataset
[ ] ↳ 2 cells hidden

Augmented Images
[ ] ↳ 2 cells hidden

Data Preparation
[ ] ↳ 8 cells hidden

Define Model and Compile
[ ] ↳ 2 cells hidden

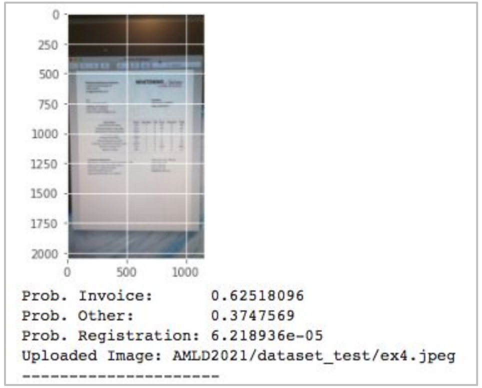
Train Model
↳ 3 cells hidden

Test on uploaded images
↳ 2 cells hidden

Exercise 1: Model Tuning
↳ 1 cell hidden

Exercise 2: Human In The Loop
[ ] ↳ 1 cell hidden

- You can take a picture of a document and test the model on it
- Upload your picture to Colab and in the folder AMLD2021/dataset_test/
- The code will make predictions on all pictures in this folder



```
Prob. Invoice:      0.62518096
Prob. Other:        0.3747569
Prob. Registration: 6.218936e-05
Uploaded Image: AMLD2021/dataset_test/ex4.jpeg
---------------------
```

ONE swiss bank

SamurAI

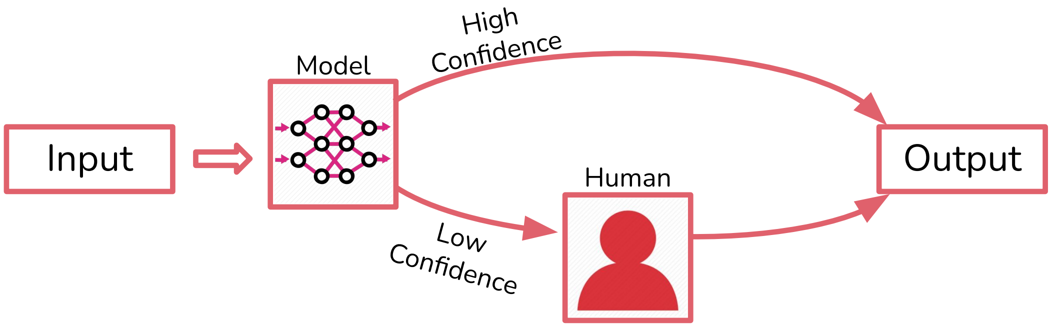
AMLDEPFL

Exercise 1: Tuning the model

- Hyper-parameter tuning and feature engineering:
  - CNN layers, filters, dropout
  - Dense layers, number of nodes, dropout
  - Optimizer, Batch size, Num. of Epochs
  - Resize of the images (up in the data preparation)
- Goal: reach an accuracy of > 99% on the training and validation sets

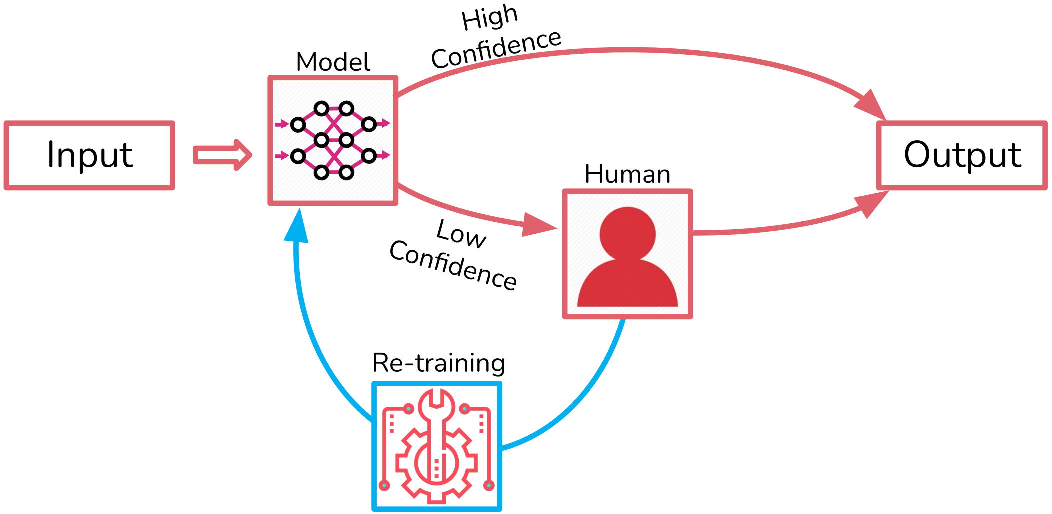
Human in the loop (HITL)

- AI systems are typically <100% accurate
- We can keep a HITL to mitigate for mistakes
    - ○ Manually do the tasks for which the model has low confidence (low h)
    - ○



Human in the loop (HITL)

- AI systems are typically <100% accurate
- We can keep a HITL to mitigate for mistakes
  - Manually do the tasks for which the model has low confidence (low h)
  - Prepare data to retrain the model



Exercise 2: Human in the loop

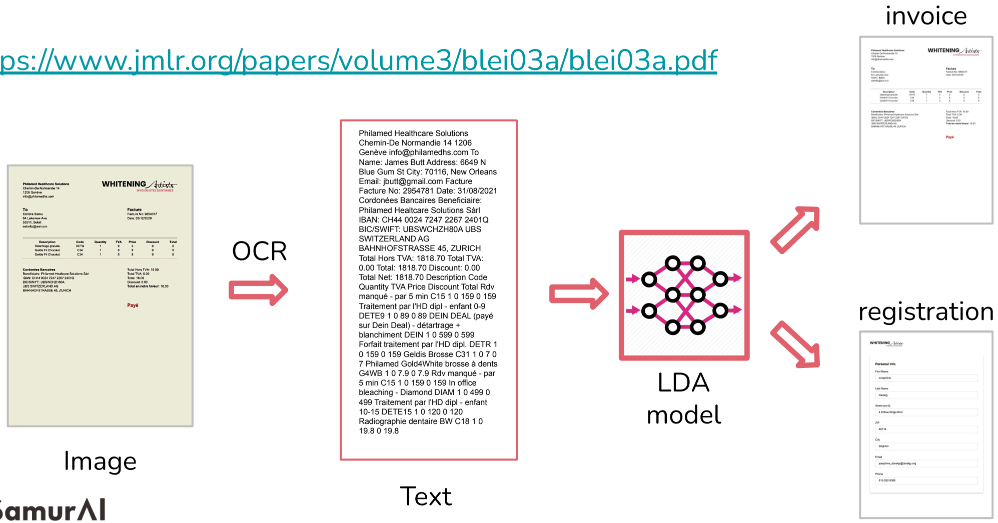
- Make photos of several documents (~5-10 per class)
  - Jpeg format is perfect
  - Advice: name the images according to their class (ex: my_invoice_1.jpg). This will help for the rest of the exercise
- Upload them on Colab and copy them to the right folder for re-training (for example AMLD2021/dataset_resize/invoice/ for invoices)
  - example: !cp my_invoice_* AMLD2021/dataset_resize/invoice/
- Re-build augmented images, retrain the model, and re-evaluate performance

ONE swiss bank SamurAI

AMLD EPFL

Possible Extensions: OCR

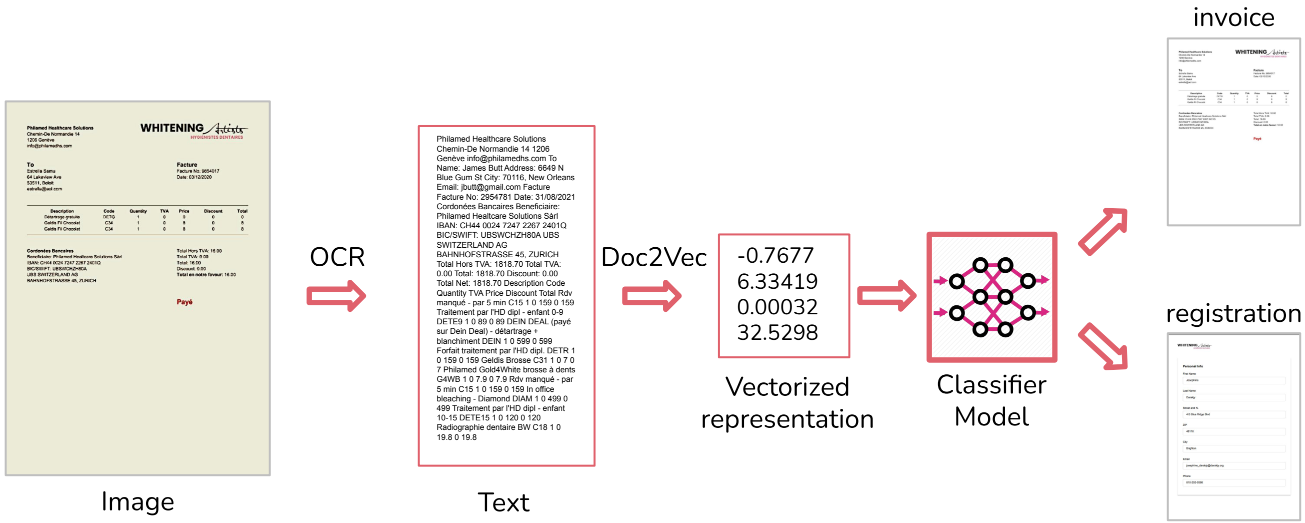
- In our classification we used only the images of our documents. We didn’t use the text within the document
- Here an example of how to do document classification using only the text extracted with OCR: link
  - Using an LDA model: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf



Philamed Healthcare Solutions Chemin-De Normandie 14 1206 Genève info@philamedhs.com To Name: James Butt Address: 6649 N Blue Gum St City: 70116, New Orleans Email: jbutt@gmail.com Facture Facture No: 2954781 Date: 31/08/2021 Cordonées Bancaires Beneficiaire: Philamed Healtcare Solutions Sàrl IBAN: CH44 0024 7247 2267 2401Q BIC/SWIFT: UBSWCHZH80A UBS SWITZERLAND AG BAHNHOFSTRASSE 45, ZURICH Total Hors TVA: 1818.70 Total TVA: 0.00 Total: 1818.70 Discount: 0.00 Total Net: 1818.70 Description Code Quantity TVA Price Discount Total Rdv manqué - par 5 min C15 1 0 159 0 159 Traitement par l'HD dipl - enfant 0-9 DETE9 1 0 89 0 89 DEIN DEAL (payé sur Dein Deal) - détartrage + blanchiment DEIN 1 0 599 0 599 Forfait traitement par l'HD dipl. DETR 1 0 159 0 159 Geldis Brosse C31 1 0 7 0 7 Philamed Gold4White brosse à dents G4WB 1 0 7.9 0 7.9 Rdv manqué - par 5 min C15 1 0 159 0 159 In office bleaching - Diamond DIAM 1 0 499 0 499 Traitement par l'HD dipl - enfant 10-15 DETE15 1 0 120 0 120 Radiographie dentaire BW C18 1 0 19.8 0 19.8

Possible Extensions: OCR + Doc2Vec

- Document classification using OCR and transforming the text of a whole page into a vector: link



Philamed Healthcare Solutions
Chemin-De Normandie 14
1206 Genève
info@philamedhs.com

WHITENING Artists
HYGIÉNISTES DENTAIRES

To
Estrella Samu
64 Lakeview Ave
53511, Beloit
estrella@aol.com

Facture
Facture No: 9854017
Date: 03/12/2020

Description	Code	Quantity	TVA	Price	Discount	Total
Détartrage gratuite	DETG	1	0	0	0	0
Geldis Fil Chocolat	C34	1	0	8	0	8
Geldis Fil Chocolat	C34	1	0	8	0	8

Cordonées Bancaires
Beneficiaire: Philamed Healtcare Solutions Sàrl
IBAN: CH44 0024 7247 2267 2401Q
BIC/SWIFT: UBSWCHZH80A
UBS SWITZERLAND AG
BAHNHOFSTRASSE 45, ZURICH

Total Hors TVA: 16.00
Total TVA: 0.00
Total: 16.00
Discount: 0.00
Total en notre faveur: 16.00

Payé

Image

OCR

Philamed Healthcare Solutions Chemin-De Normandie 14 1206 Genève info@philamedhs.com To Name: James Butt Address: 6649 N Blue Gum St City: 70116, New Orleans Email: jbutt@gmail.com Facture Facture No: 2954781 Date: 31/08/2021 Cordonées Bancaires Beneficiaire: Philamed Healtcare Solutions Sàrl IBAN: CH44 0024 7247 2267 2401Q BIC/SWIFT: UBSWCHZH80A UBS SWITZERLAND AG BAHNHOFSTRASSE 45, ZURICH Total Hors TVA: 1818.70 Total TVA: 0.00 Total: 1818.70 Discount: 0.00 Total Net: 1818.70 Description Code Quantity TVA Price Discount Total Rdv manqué - par 5 min C15 1 0 159 0 159 Traitement par l'HD dipl - enfant 0-9 DETE9 1 0 89 0 89 DEIN DEAL (payé sur Dein Deal) - détartrage + blanchiment DEIN 1 0 599 0 599 Forfait traitement par l'HD dipl. DETR 1 0 159 0 159 Geldis Brosse C31 1 0 7 0 7 Philamed Gold4White brosse à dents G4WB 1 0 7.9 0 7.9 Rdv manqué - par 5 min C15 1 0 159 0 159 In office bleaching - Diamond DIAM 1 0 499 0 499 Traitement par l'HD dipl - enfant 10-15 DETE15 1 0 120 0 120 Radiographie dentaire BW C18 1 0 19.8 0 19.8

Text

Doc2Vec

-0.7677
6.33419
0.00032
32.5298

Vectorized representation

Classifier Model

invoice

registration