ROLLING STOCK

PREDICTIVE MAINTENANCE

HÉLOÏSE NONNE

Quantmetry
Data Science Consulting

SNCF

9 BILLION PEOPLE ON EARTH

2 OUT OF 3 LIVE IN URBAN AREAS

12 MILLION INHABITANTS IN PARIS AREA

SNCF TRANSILIEN

PARIS AREA DAILY

- 14 LINES
- 3,2 MILLION TRIPS
- > 6200 TRAINS
- 1280 KM RAILWAY
- 385 STATIONS
- 26000 AGENTS

ROLLING STOCK MAINTENANCE

WHAT IS AT STAKE?

RELIABILITY

AVAILABILITY

COMPLIANCE WITH REGULATIONS

MAINTENANCE COSTS

SAFETY


S 164

THE NEW GENERATION DIGITAL NATIVES

NAT: BOMBARDIER’S TRAINS

+ 180 TRAINS
+ 1 COMPUTER IN EACH VEHICLE
+ 40,000 DIFFERENT VARIABLES
+ 70,000 LINES / MONTH / RAME + CBM (PHYSICAL PARAMETERS)
+ COMMUNICATION EVERY 30 MINUTES

WHAT KIND OF DATA?

Maintenance
Safety
Operation
Passenger

Windscreen monitoring

T°C & humidity

Transformer monitoring

Car coach monitoring

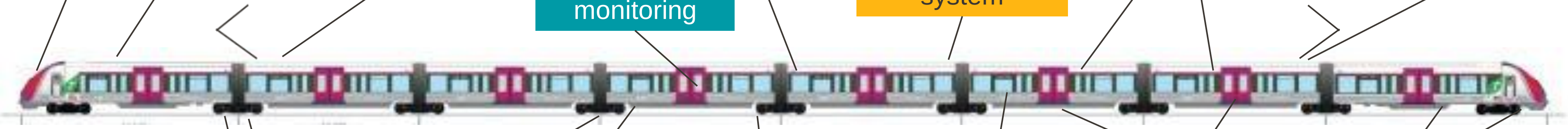
Videocamera

Air conditioning filters

Air conditioning system

Noise level & Air quality

Pantograph shocks



Shaft monitoring

Motor vibrations

Brake efficiency control

Inverter remote control

Battery monitoring

Remote display

Opened door

Opened hatches

Low battery detection

IMPROVING MAINTENANCE WITH REMOTE DIAGNOSTIC

VALUE

CLASSIC

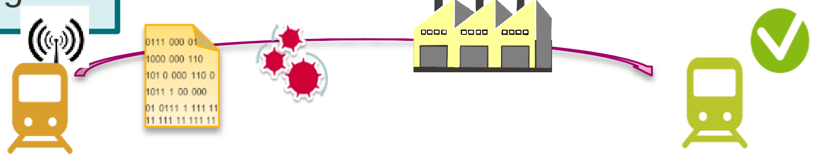
+ Repair after failure
+ Diagnostic at workshop
+ Lowest engineering effort

CONNECTED

+ Repair after failure
+ Diagnostic during operation
+ Enhanced planning

PROACTIVE

+ Repair before faults have an impact
+ Prioritize corrective maintenance
+ Optimized planning



ANALYTICS

KEY BENEFITS FROM REMOTE DIAGNOSTIC

→ FLEET OPERATION AND SUPERVISION

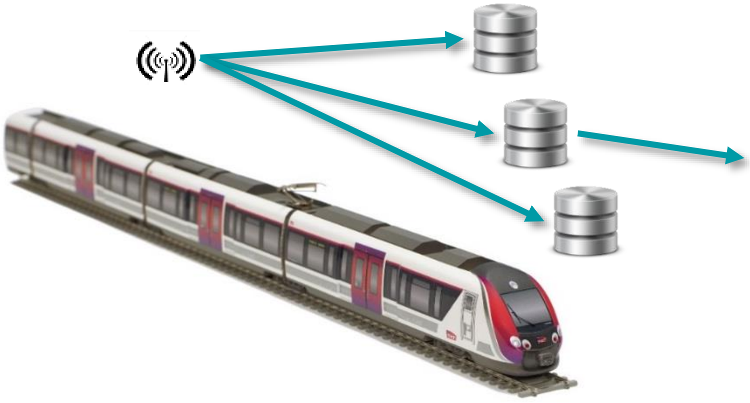
- → **INCREASED RELIABILITY AND AVAILABILITY**
- → MAKING THE RIGHT CHOICE

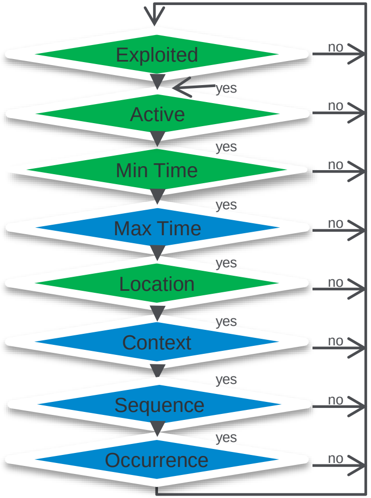
→ MAINTENANCE OPTIMIZATION

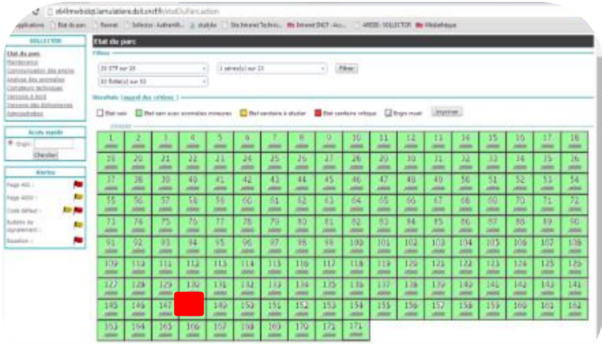
- → ONLY DO WHAT IS NECESSARY
- → SEND A TRAIN TO THE RIGHT PLACE
- → QUALITY CHECK
- → REDUCED COSTS

CONDITION BASED VERSUS SYSTEMATIC MAINTENANCE

REMOTE DIAGNOSTIC IN PRODUCTION TODAY







FLEET SUPERVISION

WORKSHOPS

MACHINE LEARNING FOR PREDICTIVE MAINTENANCE

WHY?

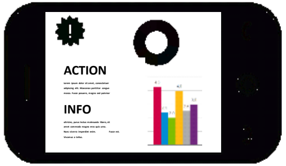
PREDICTING A FAILURE 30 MINUTES BEFORE MEANS:

+ Avoiding impact on hundreds of travellers
+ Better fleet management

USE MACHINE LEARNING TO REINFORCE ENGINEERING

+ Go beyond engineers pre-conceived ideas (which are valuable!)
+ Analyze weak signals
+ Produce automatic rules to complement experts’ rules
+ Learn faster about new rolling stock aging rules

VALEUR
CAcGAgAAAABDBAAAAAD4/wEAAAAAAAAAQwQ
CAcDBgAAAAAQhQAAAAD7/0kAAAAAAAAAEIUA
CAcDBgAAAAAQhQAAAAD7/0kAAAAAAAAAEIUA
CAcDBgAAAAAQhQAAAAD7/0kAAAAAAAAAEIUA
CAcDBgAAAAAQhQAAAAD7/0kAAAAAAAAAEIUA
CAcDBgAAAAAQhQAAAAD7/0kAAAAAAAAAEIUA
CAcDBgAAAAAQhQAAAAD7/0kAAAAAAAAAEIUA
CAcDBgAAAAAQhQAAAAD7/0kAAAAAAAAAEIUA
CAcDBgAAAAAQhQAAAAD7/0kAAAAAAAAAEIUA



OUR PROJECT

NEED FOR TANGIBLE PROOF

BE ITERATIVE, PRAGMATIC AND STICK TO EXISTING PROCESSES

1. POC: 10 weeks
2. PILOT: 3 months
3. TEST: 6 months
4. CHANGE MANAGEMENT: longer, lean in existing processes and evolve

CHALLENGES

FAILURES ARE VERY RARE!

NEW MATERIAL: A LIMITED HISTORY

CHALLENGES YOU DON'T OFTEN ENCOUNTER

A young company: Zalando	**2008**
An « old » company: Google	**1998**

CHALLENGES YOU DON'T OFTEN ENCOUNTER

A young company: Zalando	**2008**
An « old » company: Google	**1998**
SNCF	**1938**

CHALLENGES

DATA QUALITY

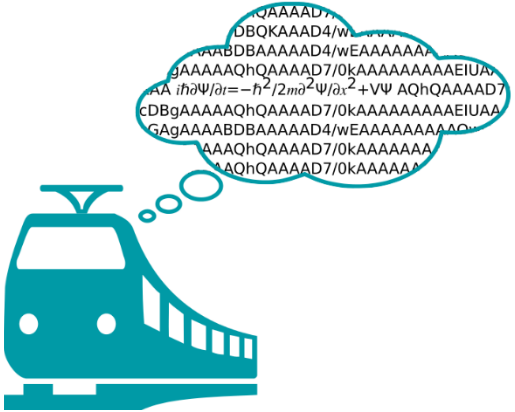
DATA IS GENERATED THROUGH VARIOUS AND COMPLEX PROCESSES

MANY HETEROGENEOUS SOURCES

GETTING A SOURCE OF DATA IS SOMETIMES DIFFICULT (CONTRACTS, REGULATIONS, SPECIFIC IS)

AN EXAMPLE

TRAINS DREAM WHEN THEY SLEEP



USE OPERATION TIMETABLES

FEATURE ENGINEERING: CONSTRUCTING FEATURES

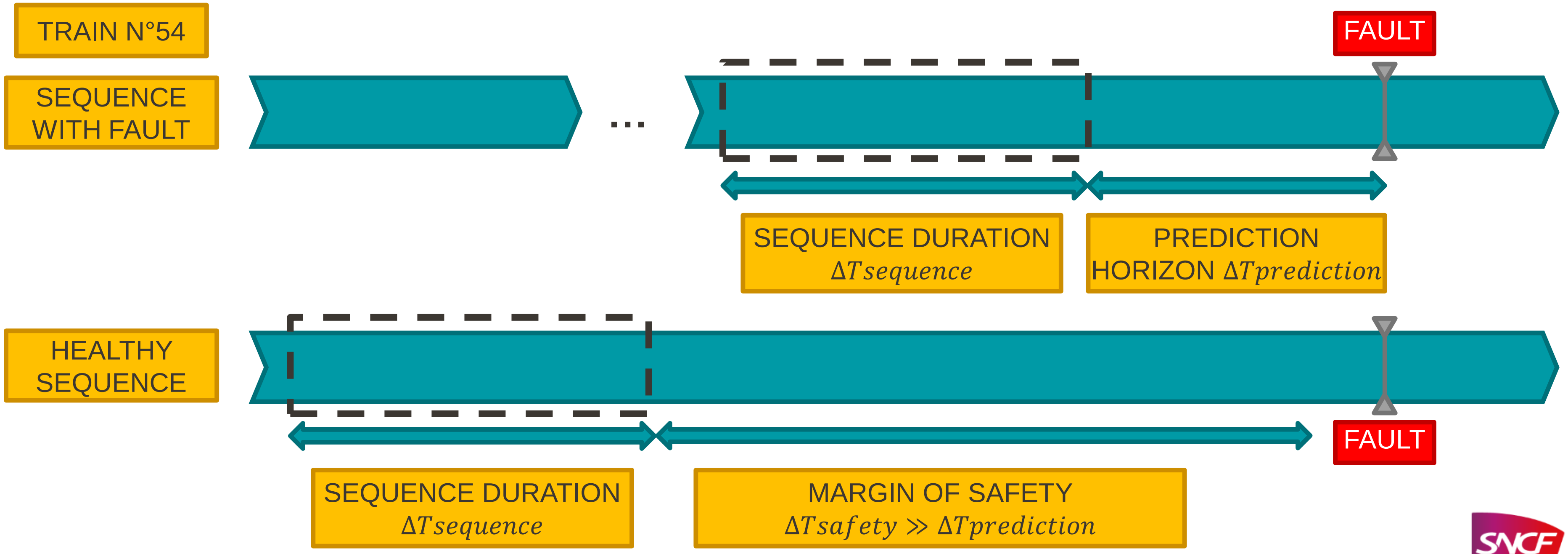
SEQUENCE	CODE	START	END
1	8301	03/05/14 17:18:32	03/05/14 17:19:04
1	20003	03/05/14 17:18:54	03/05/14 17:18:57
1	8003	03/05/14 17:19:32	03/05/14 17:21:12
…	…	…	…
23003	10054	04/05/14 10:32:10	03/05/14 10:33:17

CODE	8301			8302			…
SEQUENCE	OCCU-RENCES	FRE-QUENCY	MEAN DURATION	OCCU-RENCES	FRE-QUENCY	MEAN DURATION	…
1	304	5.3	2.4	3	99.3	132.1	…
2	0	NA	NA	0	NA	NA	…
3	32	10.1	0.45	0	NA	NA	…
…	…	…	…	…	…	…	…
23003	5	1.3	143.1	1	NA	12.6	…

FEATURE ENGINEERING: CONSTRUCTING FEATURES

ONE LINE REPRESENTS THE TIME AGGREGATION ON DURATION $\Delta Tsequence = 4H$

$$\Delta Tprediction = 30\,min, \qquad \Delta Tsafety = 20H$$



MODELING STEPS

1 SEPARATE

70%

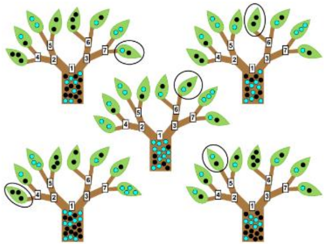
PREPARED DATA

30%

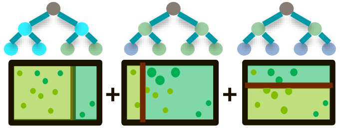
TRAINING SET		
Sequence ID	Features	Target
345	…	OUI
2	…	NON
…	…	…
10054	…	OUI

TEST SET	
Sequence ID	Features
204	…
3	…
…	…
2301	…

2 MODEL TRAINING



TRAINED MODEL



3 PREDICTION

RESULTS	
Probability	Prediction
0.98	OUI
0.32	NON
…	…
0.76	OUI

FROM POC TO PRODUCTION

YOU NEED THING WORKING NEATLY IN PRODUCTION

YOUR DATA SCIENTISTS WORK LIKE THAT

FROM POC TO PRODUCTION

FROM PYTHON & SCIKIT LEARN

TO

SPARK AND MLLIB

DISTRIBUTING COMPUTATION PARTITIONING OVER TRAINS

WRITE EFFICIENT SPARK CODE TRANSLATING A POC

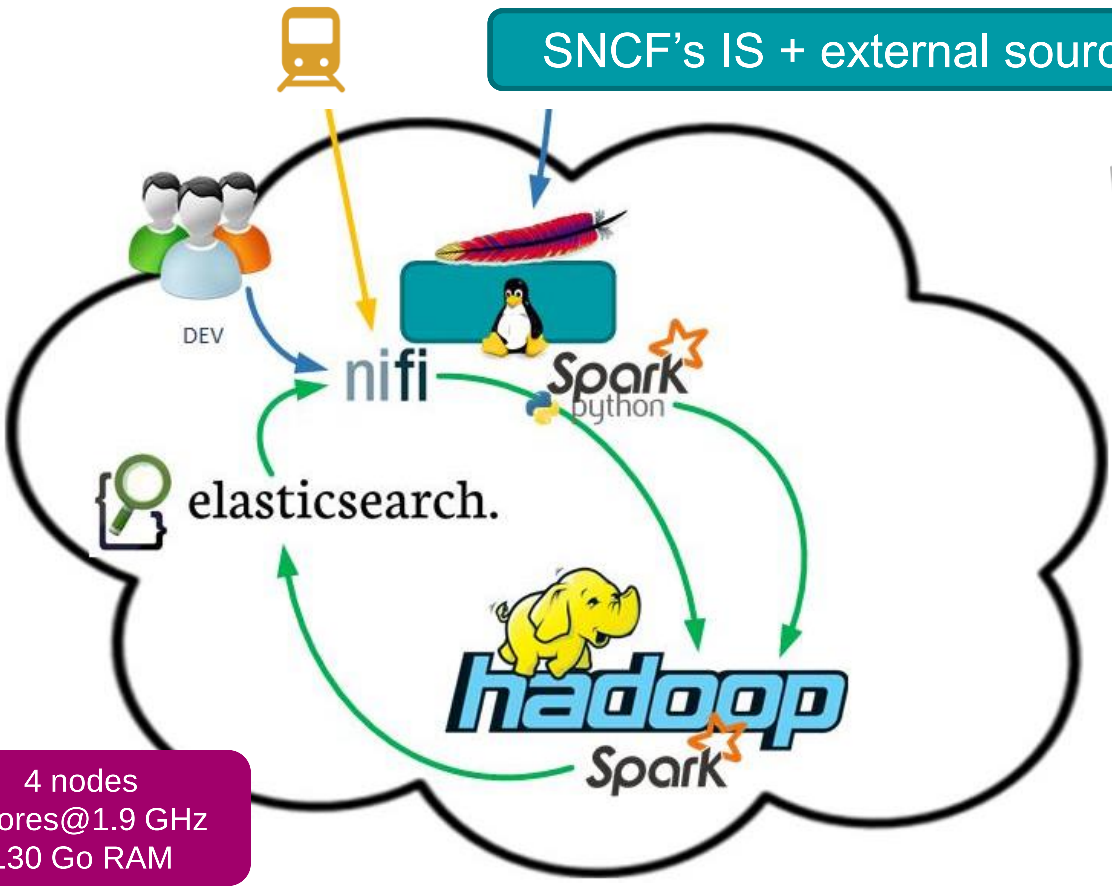
FROM POC TO PRODUCTION

HOW TO COMPARE POC RESULTS WITH PILOT?

- DIFFERENCES IN IMPLEMENTATIONS ($<$ IS NOT $\leq$)
- COMPARE PREDICTIONS MADE WITH TWO RANDOM FORESTS?

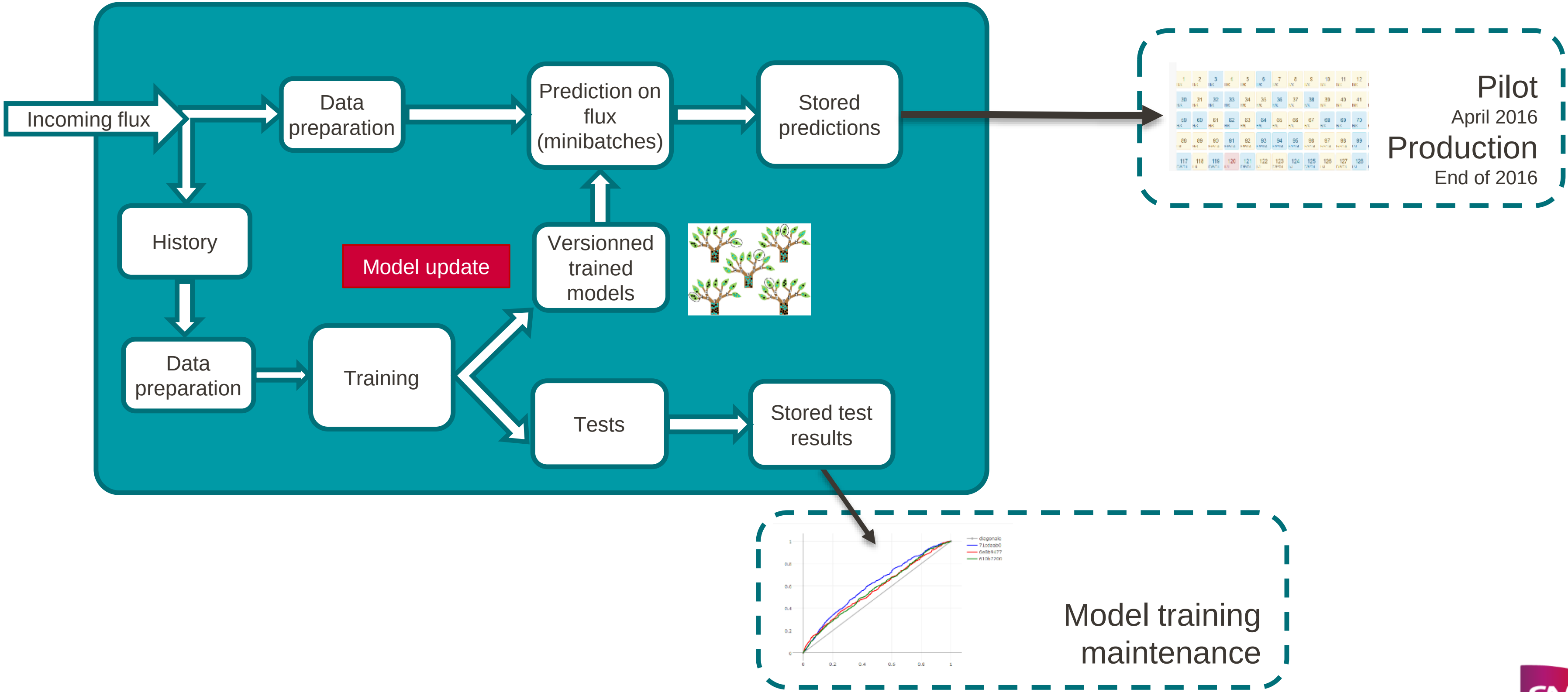
STACK

SNCF’s IS + external sources



4 nodes
8 cores@1.9 GHz
130 Go RAM

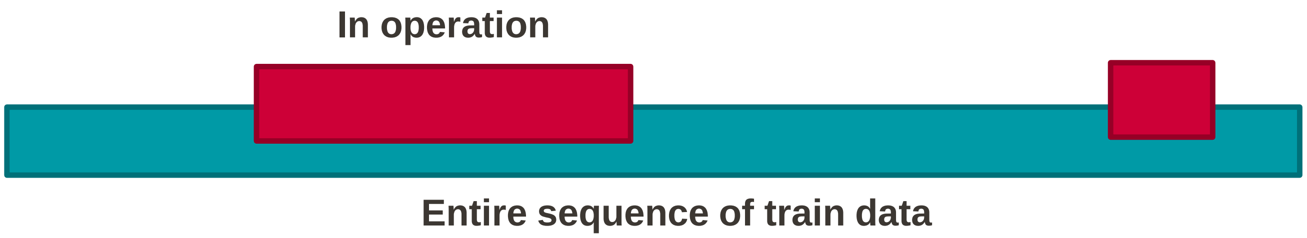
MODELS IN PRODUCTION



CONSTRUCTION OF SEQUENCES

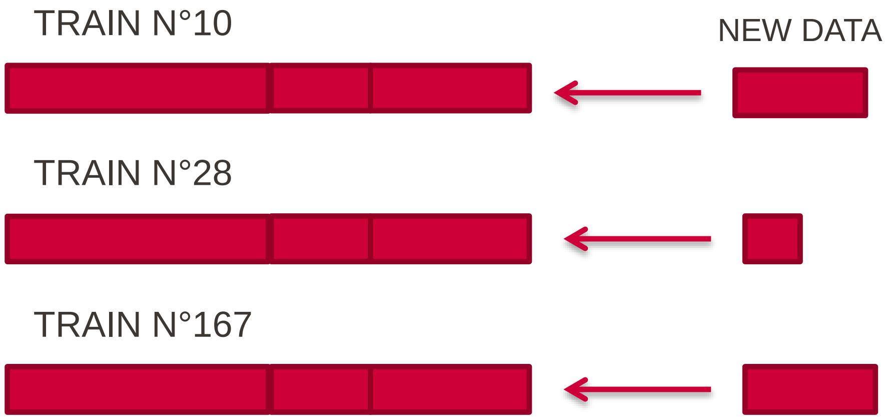
For each train, every 30 minutes

- a new file comes in
- filter data generated outside of operations (sleeping trains)



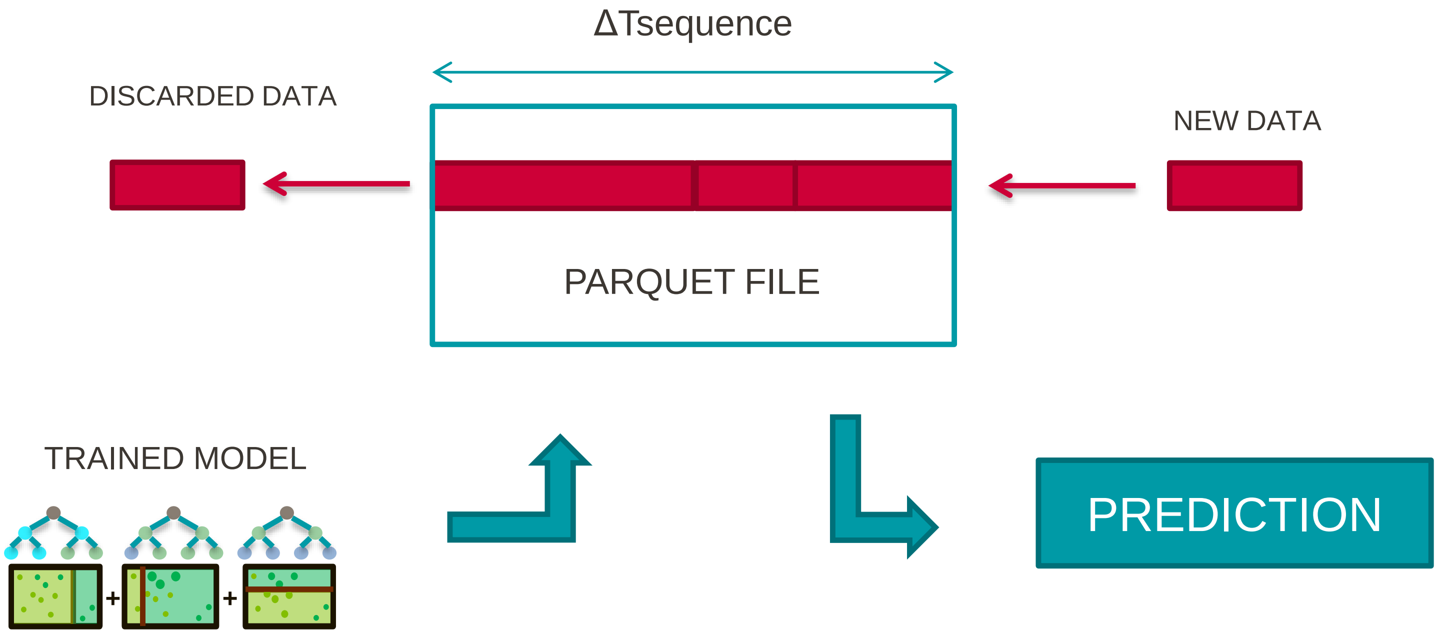
CONSTRUCTION OF SEQUENCES

New ‘in operation’ sequences are stacked together in parquet format



CONSTRUCTION OF SEQUENCES

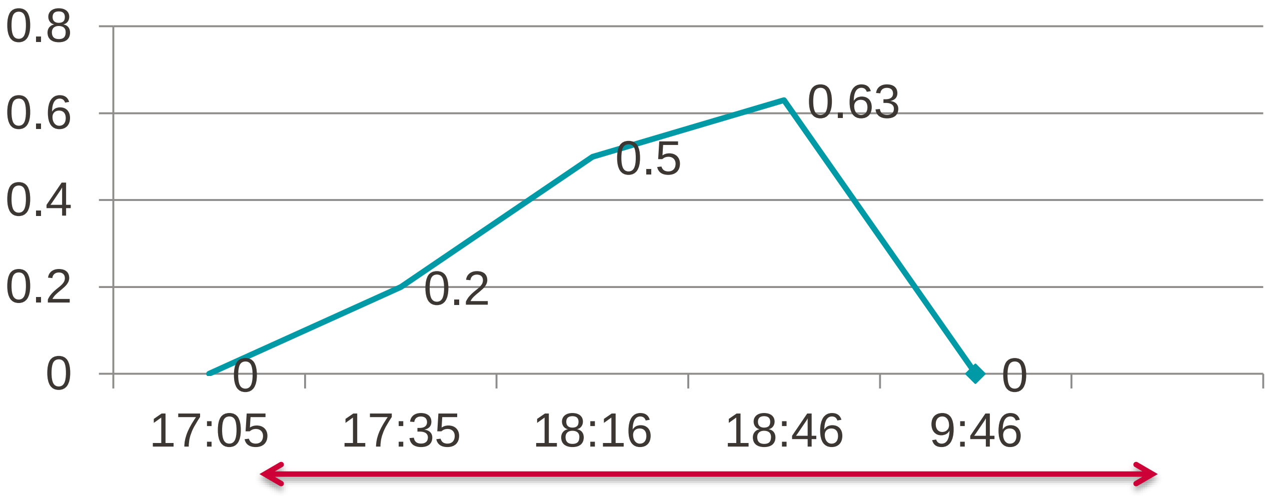
For ‘real time’ predictions, keep a constant



PREDICTION EVOLUTION

EXAMPLE

PREDICTING FAILURE ON DOOR ENGINE FOR TRAIN N° 124



STARTING PREDICTION ON MARCH 25TH AT 17:05

NAT - Prédiction   Prédictions   Entraînement   Evénements 0

DEMO

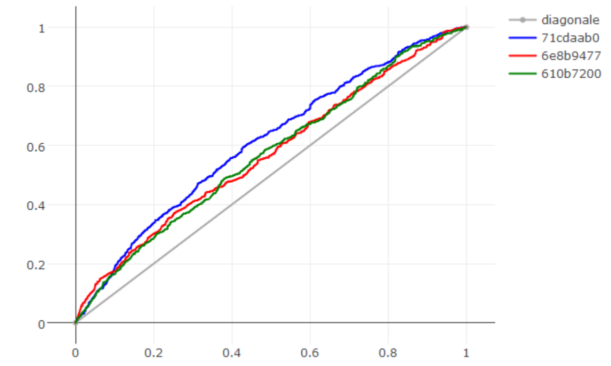
Entraînement

Liste des modèles générés

Fonctions N : Porte

	Identifiant	Actif?	▲Date	Séquence non défaillance (dT0)	Horizon de prédiction (dT1)	Nombre de points	Aire sous la courbe	Trains	Détails	Lift
☐	0f2c6cbf	✖	25/03/2016 12:42	4h	30min	26867	0.608			
☑	610b7200	✖	25/03/2016 17:24	8h	1h	26346	0.570			
☑	71cdaab0	✖	04/04/2016 16:08	4h	30min	26867	0.608			
☐	60093dbc	✖	04/04/2016 16:25	4h	30min	26868	0.625			
☐	276d0866	✖	05/04/2016 15:47	4h	30min	26868	0.625			
☐	ee175058	✖	06/04/2016 11:52	4h	30min	26868	0.582			
☐	3b255900	✖	13/04/2016 10:41	4h	30min	26868	0.625			
☐	8691e570	✖	13/04/2016 14:26	4h	30min	26868	0.625			
☐	7abdfeb8	✖	13/04/2016 14:35	4h	30min	26868	0.625			
☑	6e8b9477	✖	25/04/2016 22:29	4h	30min	18514	0.570			
☐	97dd718f	✖	29/04/2016 15:51	4h	30min	26604	0.603			
☐	883c0108	✖	12/05/2016 17:51	4h	30min	25927	0.511			
☐	76ac6300	✖	13/05/2016 10:24	4h	30min	18514	0.570			
☐	7c3f990c	✖	13/05/2016	4h	30min	18514	0.570			

Courbe de ROC



NAT - Prédiction Prédictions Entraînement Evénements 0

DEMO

Prédictions

Dernière mise à jour : 25/05/2016 à 09:36

Warning level : 0 **Danger level :** 0,5 **Lignes :**

1 H/K	2 H/K	3 H/K	4 H/K	5 H/K	6 H/K	7 H/K	8 H/K	9 H/K	10 H/K	11 H/K	12 H/K	13 H/K	14 H/K	15 H/K	16 H/K	17 H/K	18 H/K	19 H/K	20 H/K	21 H/K	22 H/K	23 H/K	24 H/K	25 H/K	26 H/K	27 H/K	28 H/K	29 H/K
30 H/K	31 H/K	32 H/K	33 H/K	34 H/K	35 H/K	36 H/K	37 H/K	38 H/K	39 H/K	40 H/K	41 H/K	42 H/K	43 H/K	44 H/K	45 H/K	46 H/K	47 H/K	48 H/K	49 H/K	50 H/K	51 H/K	52 H/K	53 H/K	54 H/K	55 H/K	56 H/K	57 H/K	58 H/K
59 H/K	60 H/K	61 H/K	62 H/K	63 H/K	64 H/K	65 H/K	66 H/K	67 H/K	68 H/K	69 H/K	70 H/K	71 H/K	72 H/K	73 H/K	74 H/K	75 H/K	76 H/K	77 H/K	78 L/J	79 L/J	80 L/J	81 L/J	82 H/K	83 H/K	84 H/K	85 H/K	86 L/J	87 L/J
88 L/J	89 H/K	90 E/P/T4	91 E/P/T4	92 E/P/T4	93 E/P/T4	94 E/P/T4	95 E/P/T4	96 E/P/T4	97 E/P/T4	98 E/P/T4	99 L/J	100 E/P/T4	101 L/J	102 E/P/T4	103 L/J	104 L/J	105 E/P/T4	106 L/J	107 E/P/T4	108 L/J	109 E/P/T4	110 L/J	111 E/P/T4	112 L/J	113 E/P/T4	114 L/J	115 E/P/T4	116 L/J
117 E/P/T4	118 L/J	119 E/P/T4	120 L/J	121 E/P/T4	122 L/J	123 E/P/T4	124 L/J	125 E/P/T4	126 L/J	127 E/P/T4	128 L/J	129 E/P/T4	130 L/J	131 E/P/T4	132 L/J	133 E/P/T4	134 L/J	135 E/P/T4	136 L/J	137 E/P/T4	138 E/P/T4	139 E/P/T4	140 E/P/T4	141 E/P/T4	142 E/P/T4	143 E/P/T4	144 L/J	145 E/P/T4
146 L/J	147 L/J	148 L/J	149 L/J	150 L/J	151 L/J	152 L/J	153 L/J	154 L/J	155 L/J	156 L/J	157 L/J	158 L/J	159 L/J	160 L/J	161 L/J	162 L/J	163 L/J	164 L/J	165 L/J	166 L/J	167 L/J	168 L/J	169 L/J	170 L/J	171 L/J	172 L/J	173 E/P/T4	174 E/P/T4
175 E/P/T4	176 E/P/T4	177 E/P/T4	178 E/P/T4	179 E/P/T4	180 H/K	181 H/K	182 H/K																					

NAT - Prédiction Prédictions Entraînement Evénements 0

DEMO

Prédictions

Dernière mise à jour : 25/05/2016 à 09:36

Warning level : 0 **Danger level :** 0,5

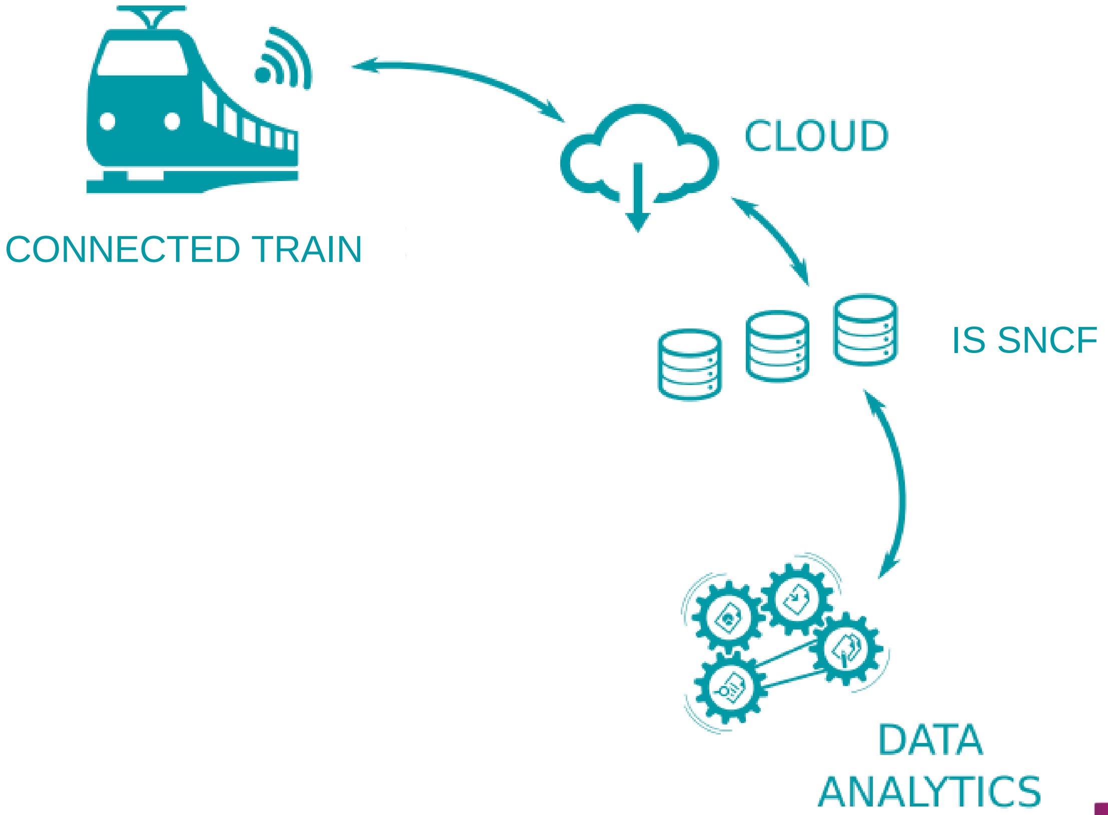
Rame n°5.

Fin de séquence : 07/04/2016 à 09:30

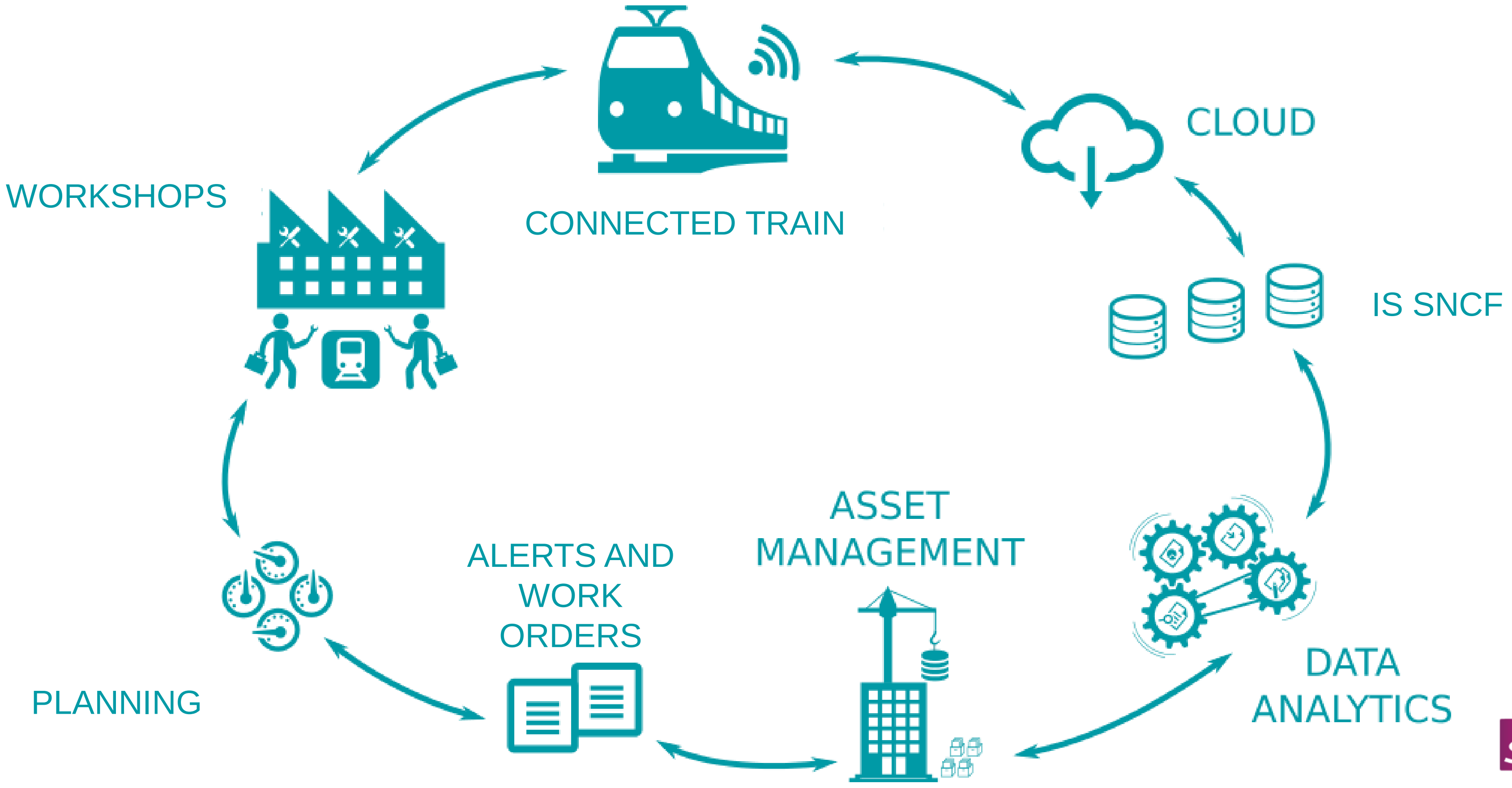
Fonction	Probabilité de panne à... 30min
0 : Fonction manquante	0.04%
B : Caisse / Chaudron	0.01%
C : Habillage de caisse	0.02%
D : Aménagement intérieur	0.13%
E : Organe de roulement	0.05%
F : Appareil de puissance / Chaîne de traction	0.48%
G : Contrôle commande de la chaîne traction / freinage	0.33%
H : Équipements auxiliaires	0.37%
J : Équipements de sécurité et de surveillance	12.09%
K : Éclairage	0.24%
L : Climatisation	0.55%
M : Autres équipements	0.09%
N : Porte	1.63%
P : Système d'Information Voyageurs et d'aide à l'exploitation	0.20%
Q : Équipements hydrauliques et pneumatiques	0.06%
R : Frein (système de frein / ensemble organes)	0.15%
S : Liaisons inter caisse	0.01%

NOW THE REAL LIFE QUESTIONS

A REAL LIFE QUESTION



A REAL LIFE QUESTION

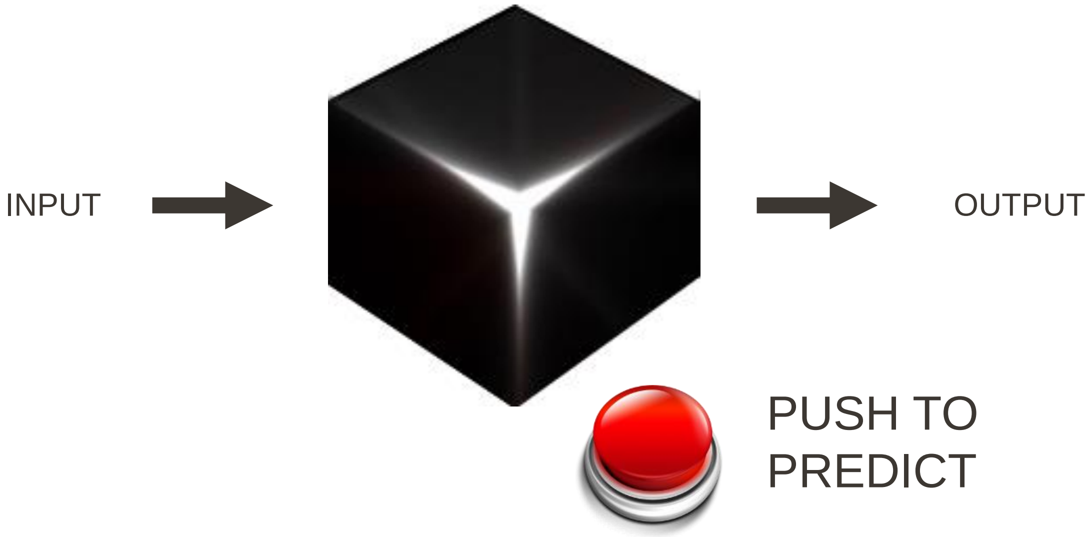


A REAL LIFE QUESTION

TAKING DECISIONS BASED ON THE PREDICTIONS

+ SYSTEM MUST BE RELIABLE AND CONTROLLED
+ DEAL WITH FALSE POSITIVES / NEGATIVES: CHOOSE ALERT THRESHOLDS
+ WHAT DECISION TO TAKE?
+ YOU NEED TO CONVINCE EXPERTS

MACHINE LEARNING
OR THE BLACKBOX NIGHTMARE



TRANSLATE RANDOM FORESTS
KNOWLEDGE INTO USEFUL KNOWLEDGE

WHAT WE HAVE TO DO NOW

- FIND THE BEST PARAMETERS
  - $\Delta \mathrm{Tsequence}, \ \Delta \mathrm{Tprediction}$
  - NUMBER OF TREES, DEPTHS, ETC.

- CHOOSE ALERT THRESHOLDS AS A FUNCTION OF:
  - MONITORED SYSTEM
  - CRITICITY (TYPE OF FAILURE / EXTERNAL CONDITIONS)

AND THE BIG QUESTION

HOW TO MAINTAIN THE SYSTEM

+ BETTER HANDLING OF EVOLVING DATA
+ ENSURE THE STABILITY OF AN AI IN PRODUCTION
    + WHAT IS A UNIT TEST FOR AN AI?
+ PROTECT AGAINST MALICIOUS ATTACKS

LESSONS LEARNED

POCS

THINK CLOSE TO PRODUCTION
AS SOON AS POSSIBLE

Ask your datascientists (when possible) to:

+Parallelize when desining the data preparation code

+Avoid serial code and design classes

+Design unit tests even for POC projects

PILOT PROJECTS

IT MAY BE BETTER TO USE THE POC PROTOTYPE TO TEST REAL CONDITIONS

+ You WILL have surprises (bad and good) in real conditions
+ Avoid redevelopment before tests
  (you may need to change your architecture)
+ Easier and cost efficient to choose (at least some of) the models parameters during the tests

IMPROVING CONTINUOUSLY

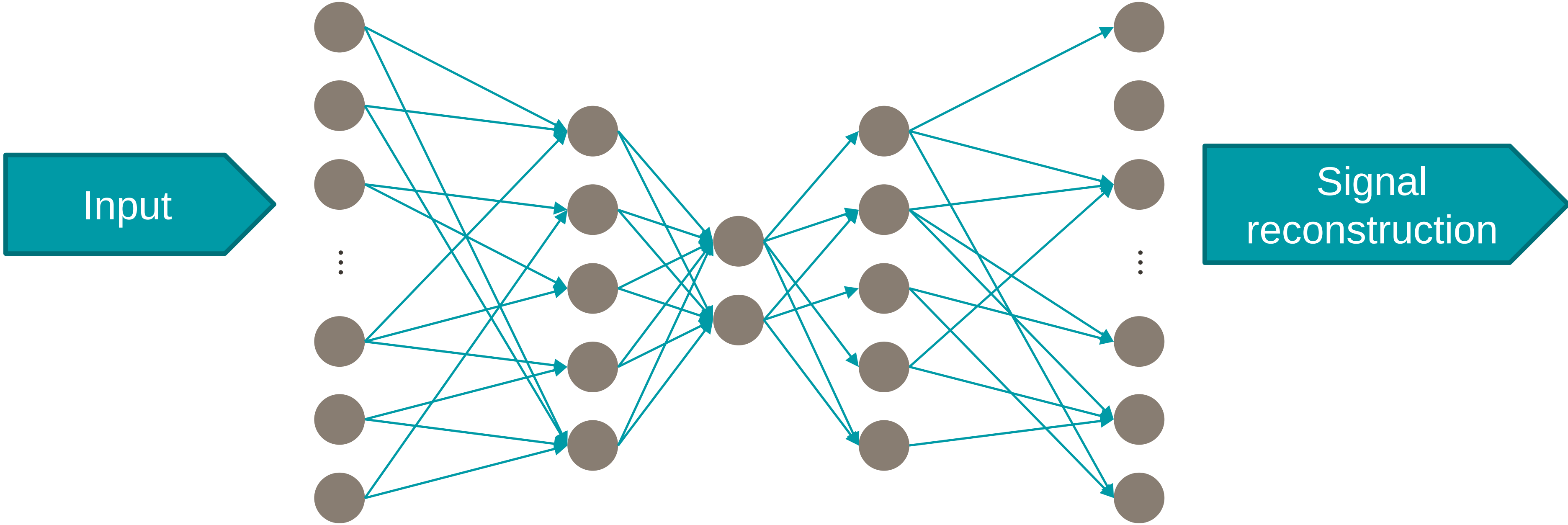
RANDOM FORESTS ARE NICE BUT

- +SINCE FAILURES ARE SO RARE
- +SINCE SIGNAL IS WEAK AND SPARSE

WHY NOT

- +USE UNSUPERVISED LEARNING?
- +USE NEURAL NETWORKS?

USING NEURAL NETWORKS: AUTOENCODERS



AUTOENCODERS: UNSUPERVISED LEARNING

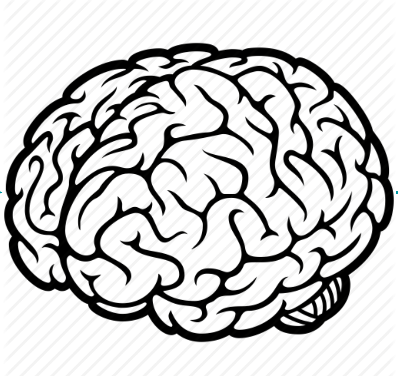
Training

Input:
Healthy trains only

Reconstruction

Prediction

Input:
All trains



Reconstruction

Reconstruction is good -> no failure
Reconstruction is bad -> probable failue

THAT WORKS VERY WELL

A LARGE IMPROVEMENT IN PREDICTION PERFORMANCES

BUT…

BIG PROBLEM

VERY UNSTABLE IN PRODUCTION

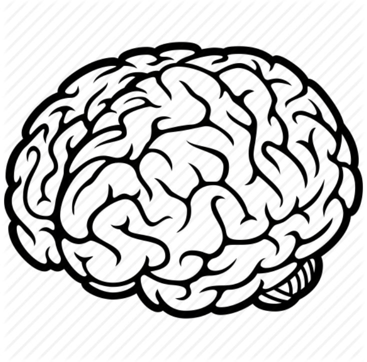
DATA GENERATION CHANGES WITH TIME

EVEN MORE PARAMETERS TO TUNE

- ARCHITECTURE
  - NUMBER OF LAYERS
  - NUMBER OF NEURONS IN EACH LAYER
- INITIALIZATION
- ACTIVATION FUNCTION
- LEARNING ALGORITHM
- NUMBER OF PASSES OVER TRAINING DATA
- DROPOUT

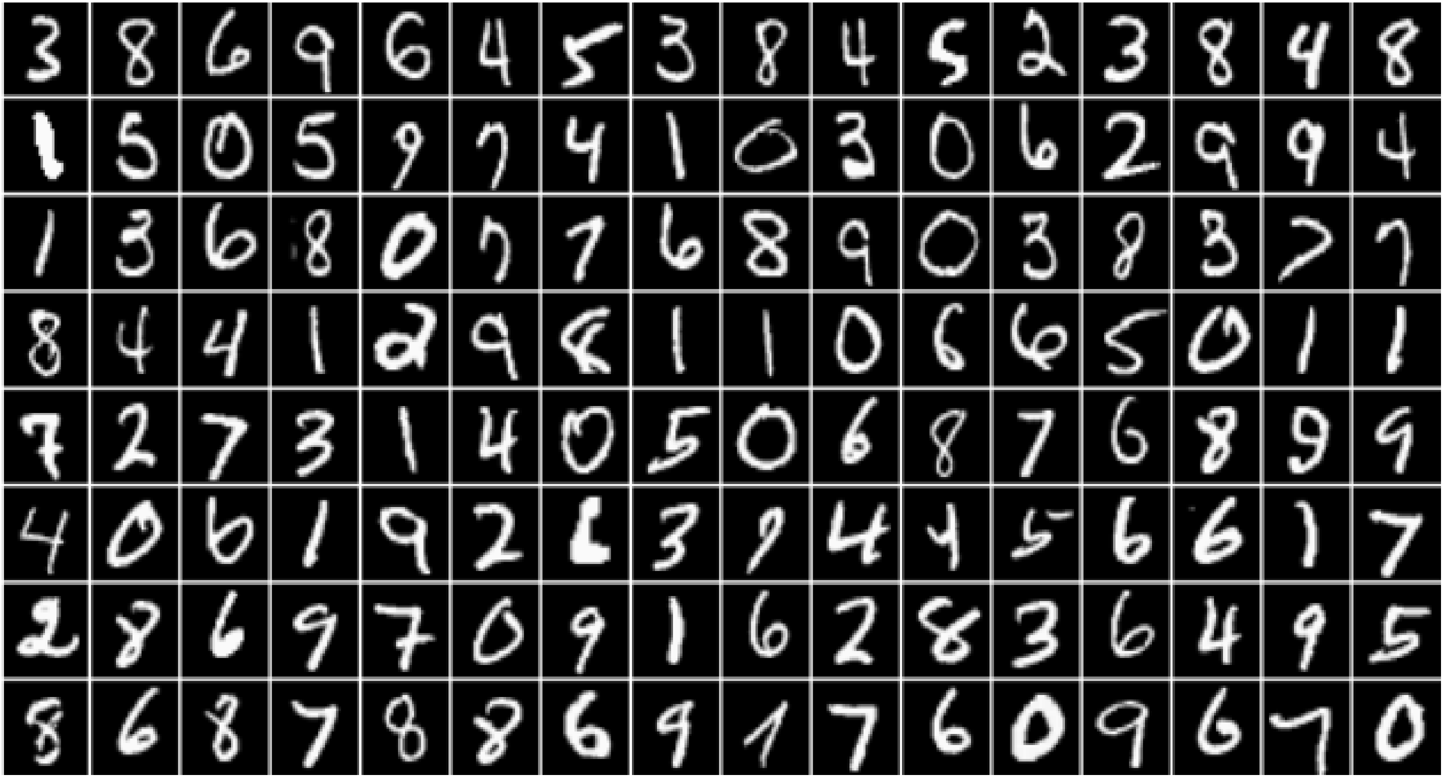
ANOTHER PROBLEM

HOW TO READ A NEURAL NETWORK
TO GIVE FEEDBACK TO EXPERTS???

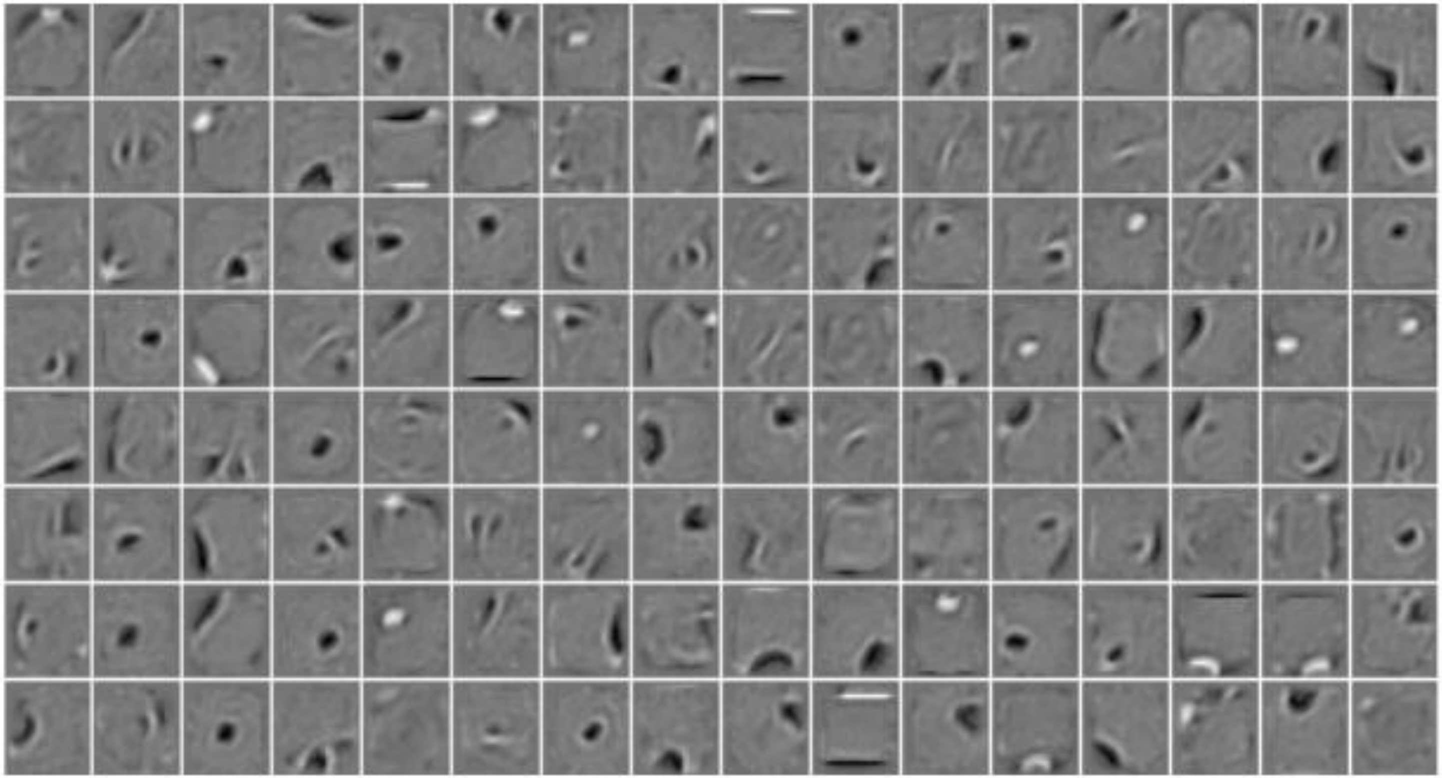


YOU HAVE TO EXPLAIN YOUR
PREDICTIONS TO CONVINCE EXPERTS

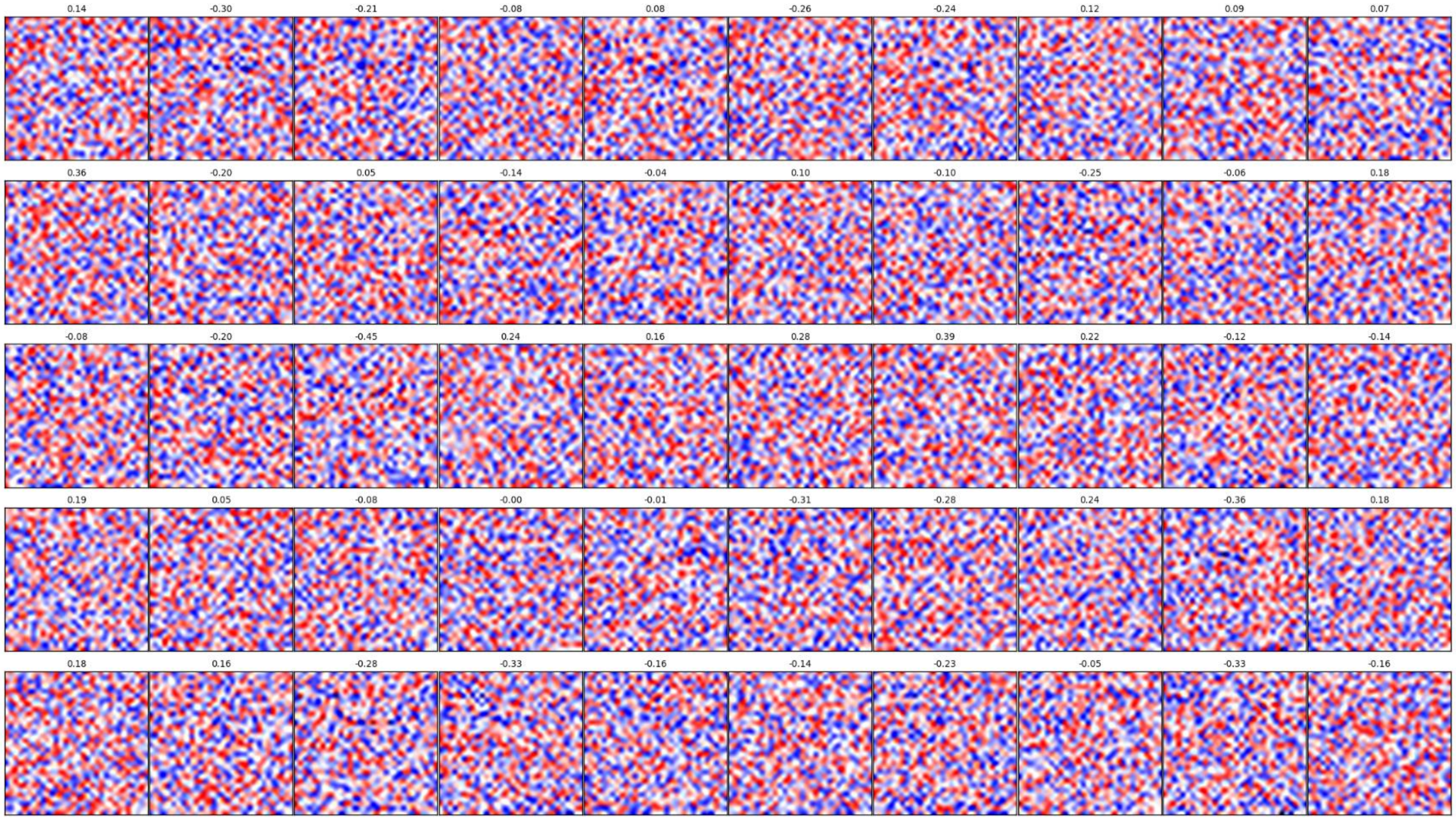
MNIST: HANDWRITTEN DIGITS



READING THE BRAIN MAKE SENSE



NICE, BUT ON TRAIN CODES



FURTHER STEPS

+ FIX THE NEURAL NETWORKS PROBLEMS
+ NLP
+ EXTRACT MORE INFORMATION FROM TRAINED MODELS
+ IMPLEMENT RETROACTION ON MODELS IN PRODUCTION
+ EFFICIENT HYPERPARAMETER SEARCH
+ MACHINE LEARNING ASSISTED DECISION PROCESSES

THANK YOU

QUESTIONS?