The Dos and Don'ts of Creating ML Applications

It's about Tech #3

Sissach 2023-04-18

Andre Bieler, Parashift AG
andre.bieler@parashift.io

Who is this guy?

Andre Bieler

Background

- PhD in Physics (space science)
- ML since 2016

Serious about / Ice Breakers

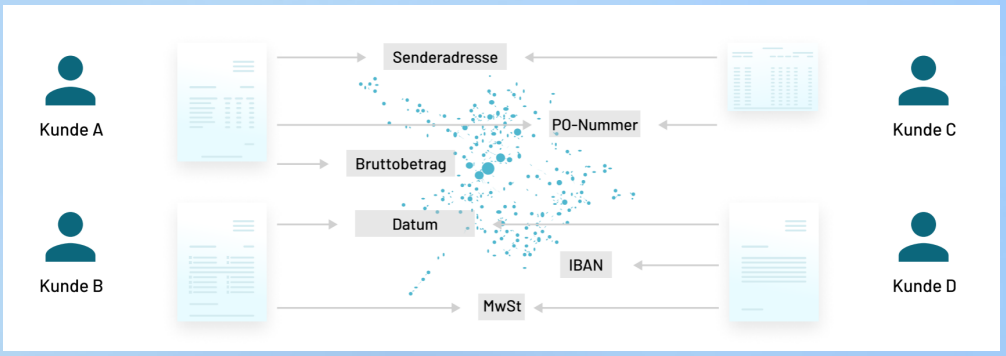
- ML
- Pizza
- Racket Sports
- Coffee

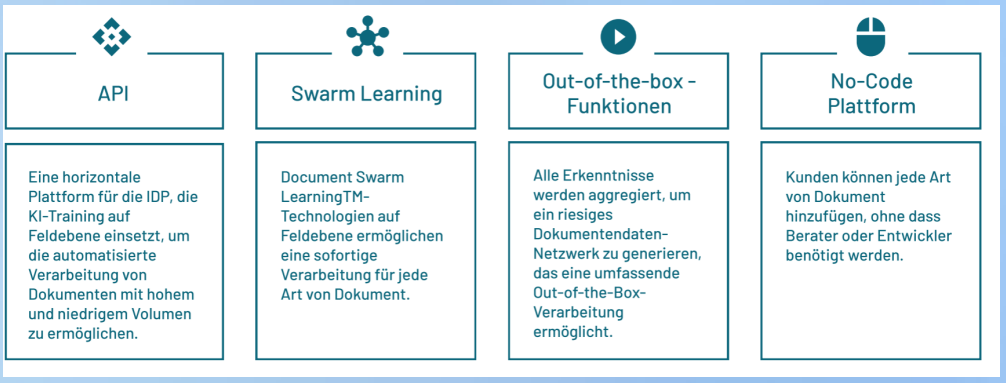
"Machine Learning is like teenage sex. Everybody talks about it. Only some really know how to do it. Everyone thinks everyone else is doing it. So, everyone claims they're doing it". Dan Ariley

Parashift AG

Intelligent Document Processing

- Founded 2018, CEO: Alain Veuve, Sissach BL
- ~35 Employees
  - Customer Success
  - Sales / Marketing
  - Development
- SAAS product: documents in -> information out
- IDP
  - Intelligent separation of batches of docs
  - Classification into document types
  - Information Extraction
- Full Cloud solution (we have on-prem, but every time a customer orders that, a puppy dies)





Adresse
Hauptstrasse 134
CH-4450 Sissach
Schweiz

Email
info@parashift.io

Telefon
+41 61 508 77 77

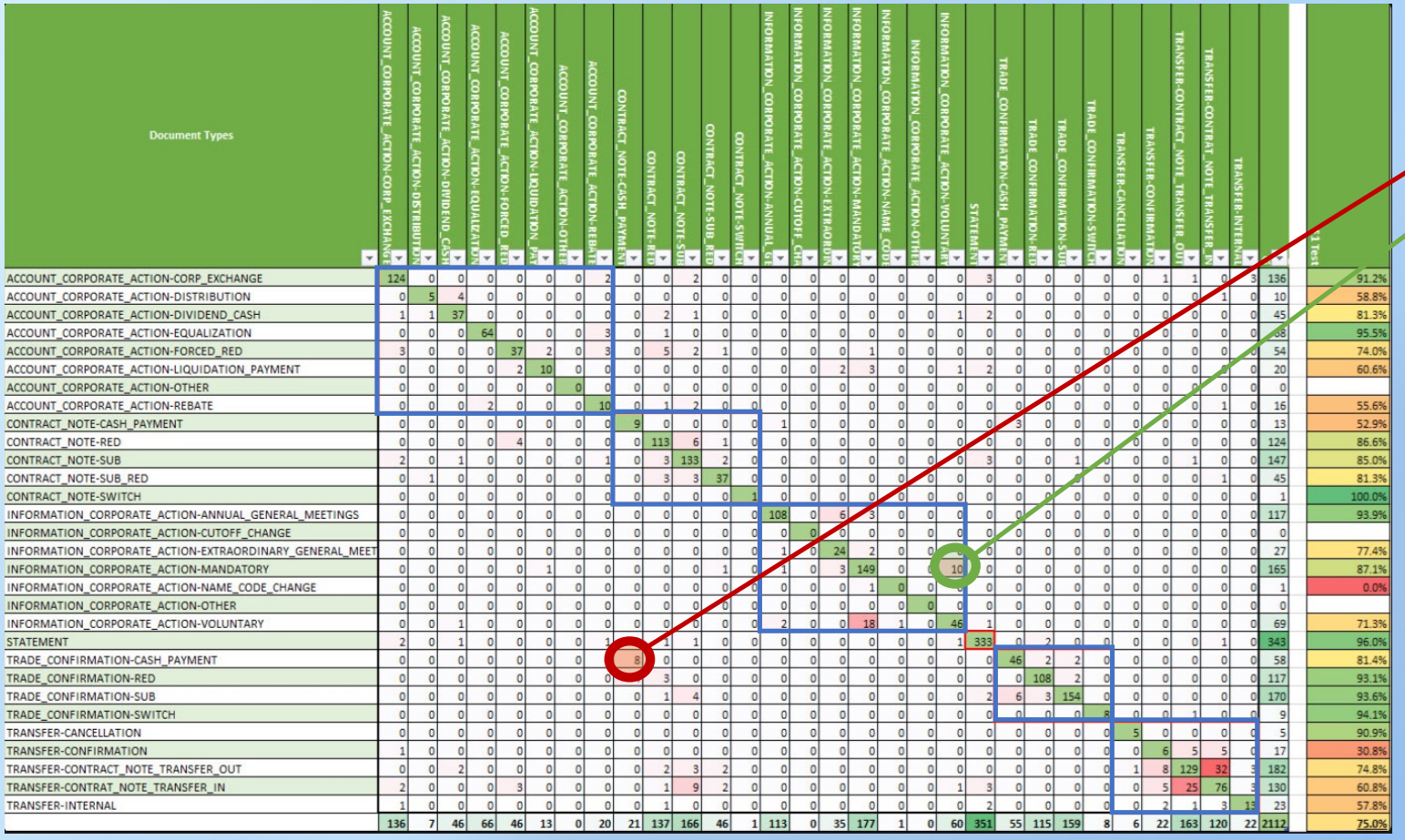
Don’t assume you know anything

What do customers want actually? We need to talk...

- **What is the actual use case?**
  - How good does the ML results have to be to be useful? (100%?)
  - Are all mistakes similarly bad? (False-Positives vs False-Negatives)
  - What are expectations? (everything out of the box at 100% accuracy...)
- **Does this fit into the current workflow?**
- **Do you really need ML?**
  - ML is cool, but sometimes a simple rule-based approach is many times more efficient and safer.
  - Example: Someone used ChatGPT for parsing HTML. (No)
  - Example: Find swiss AHV numbers in documents: 756.xxxx.xxxx.xc (No)
  - Example: Find all IBAN/BIC numbers: (probably yes, surprisingly many versions)

Don’t assume you know anything

Different type of mistakes, use-case specific



Document Types	ACCOUNT_CORPORATE_ACTION-CORP_EXCHANGE	ACCOUNT_CORPORATE_ACTION-DISTRIBUTION	ACCOUNT_CORPORATE_ACTION-DIVIDEND_CASH	ACCOUNT_CORPORATE_ACTION-EQUALIZATION	ACCOUNT_CORPORATE_ACTION-FORCED_RED	ACCOUNT_CORPORATE_ACTION-LIQUIDATION_PAY	ACCOUNT_CORPORATE_ACTION-OTHER	ACCOUNT_CORPORATE_ACTION-REBATE	CONTRACT_NOTE-CASH_PAYMENT	CONTRACT_NOTE-RED	CONTRACT_NOTE-SUB	CONTRACT_NOTE-SUB_RED	CONTRACT_NOTE-SWITCH	INFORMATION_CORPORATE_ACTION-ANNUAL_GE	INFORMATION_CORPORATE_ACTION-CUTOFF_CH	INFORMATION_CORPORATE_ACTION-EXTRAORDIN	INFORMATION_CORPORATE_ACTION-MANDATORY	INFORMATION_CORPORATE_ACTION-NAME_CODE	INFORMATION_CORPORATE_ACTION-OTHER	INFORMATION_CORPORATE_ACTION-VOLUNTARY	STATEMENT	TRADE_CONFIRMATION-CASH_PAYMENT	TRADE_CONFIRMATION-RED	TRADE_CONFIRMATION-SUB	TRADE_CONFIRMATION-SWITCH	TRANSFER-CANCELLATION	TRANSFER-CONFIRMATION	TRANSFER-CONTRACT_NOTE_TRANSFER_OUT	TRANSFER-CONTRAT_NOTE_TRANSFER_IN	TRANSFER-INTERNAL	Total	% Test
ACCOUNT_CORPORATE_ACTION-CORP_EXCHANGE	124	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	1	0	3	136	91.2%
ACCOUNT_CORPORATE_ACTION-DISTRIBUTION	0	5	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	10	58.8%
ACCOUNT_CORPORATE_ACTION-DIVIDEND_CASH	1	1	37	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	45	81.3%
ACCOUNT_CORPORATE_ACTION-EQUALIZATION	0	0	0	64	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	68	95.5%
ACCOUNT_CORPORATE_ACTION-FORCED_RED	3	0	0	0	37	2	0	3	0	5	2	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	54	74.0%
ACCOUNT_CORPORATE_ACTION-LIQUIDATION_PAYMENT	0	0	0	0	2	10	0	0	0	0	0	0	0	0	0	0	2	3	0	0	1	2	0	0	0	0	0	0	0	20	60.6%
ACCOUNT_CORPORATE_ACTION-OTHER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ACCOUNT_CORPORATE_ACTION-REBATE	0	0	0	2	0	0	0	10	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	16	55.6%
CONTRACT_NOTE-CASH_PAYMENT	0	0	0	0	0	0	0	0	9	0	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	13	52.9%
CONTRACT_NOTE-RED	0	0	0	0	4	0	0	0	0	113	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	124	86.6%
CONTRACT_NOTE-SUB	2	0	1	0	0	0	0	1	0	3	133	2	0	0	0	0	0	0	0	0	3	0	0	1	0	0	0	1	0	147	85.0%
CONTRACT_NOTE-SUB_RED	0	1	0	0	0	0	0	0	0	3	3	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	45	81.3%
CONTRACT_NOTE-SWITCH	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	100.0%
INFORMATION_CORPORATE_ACTION-ANNUAL_GENERAL_MEETINGS	0	0	0	0	0	0	0	0	0	0	0	0	0	108	0	6	3	0	0	0	0	0	0	0	0	0	0	0	0	117	93.9%
INFORMATION_CORPORATE_ACTION-CUTOFF_CHANGE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
INFORMATION_CORPORATE_ACTION-EXTRAORDINARY_GENERAL_MEET	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	24	2	0	0	0	0	0	0	0	0	0	0	0	0	27	77.4%
INFORMATION_CORPORATE_ACTION-MANDATORY	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	3	149	0	0	10	0	0	0	0	0	0	0	0	0	165	87.1%
INFORMATION_CORPORATE_ACTION-NAME_CODE_CHANGE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0.0%
INFORMATION_CORPORATE_ACTION-OTHER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
INFORMATION_CORPORATE_ACTION-VOLUNTARY	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	18	1	0	46	1	0	0	0	0	0	0	0	0	69	71.3%
STATEMENT	2	0	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	333	0	2	0	0	0	0	1	0	343	96.0%
TRADE_CONFIRMATION-CASH_PAYMENT	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	46	2	2	0	0	0	0	0	58	81.4%
TRADE_CONFIRMATION-RED	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	108	2	0	0	0	0	0	117	93.1%
TRADE_CONFIRMATION-SUB	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0	2	6	3	154	0	0	0	0	0	170	93.6%
TRADE_CONFIRMATION-SWITCH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	1	0	9	94.1%
TRANSFER-CANCELLATION	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	5	90.9%
TRANSFER-CONFIRMATION	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	5	0	17	30.8%
TRANSFER-CONTRACT_NOTE_TRANSFER_OUT	0	0	2	0	0	0	0	0	0	2	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	129	32	182	74.8%
TRANSFER-CONTRAT_NOTE_TRANSFER_IN	2	0	0	0	3	0	0	0	0	1	9	2	0	0	0	0	0	0	0	1	3	0	0	0	0	0	5	25	76	130	60.8%
TRANSFER-INTERNAL	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2	1	3	23	57.8%
	136	7	46	66	46	13	0	20	21	137	166	46	1	113	0	35	177	1	0	60	351	55	115	159	8	6	22	163	120	22	2112	75.0%

Bad mistake

Harmless mistake

Don’t assume you know anything

How complicated can documents be, really?

"Everybody has a plan until they get punched in the face", Mike Tyson

Variable Dates

2023-04-18

2023 | 04 | 18

04 18 2023

18. A[vp]ril[e] 2023

Im Dezember 19
(first ever customer data)

Nice Backgrounds

OCR engine picking up city names...

Multi Language

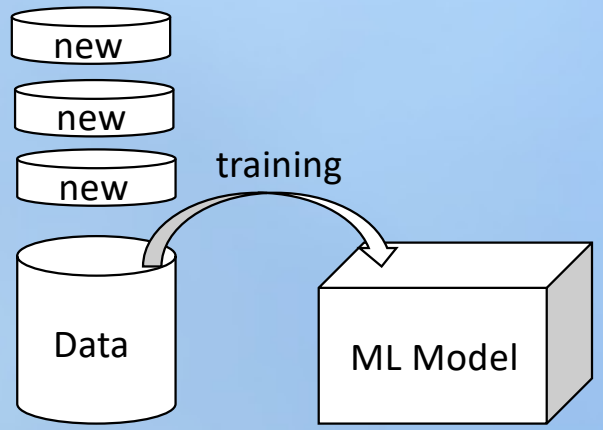
- Chicken & Egg Problem
  - No data no model
  - No model no customer
  - No customer no data
- Started out with Family, Fools & Friends, collecting data (German)
- First Customer is a German travel agency, yay!
- All their hotels & invoices in Spain :(

Stay out of Trouble

A ML model is like a Puppie

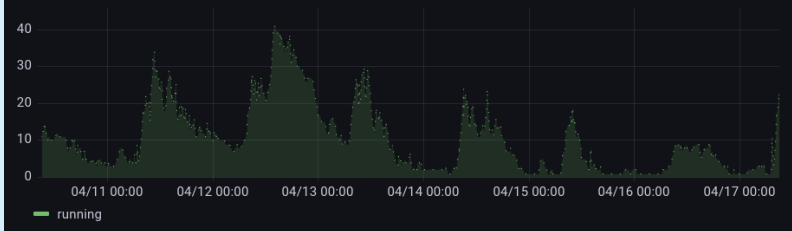
Once you successfully trained and deployed your model into the real world, the fun has only started

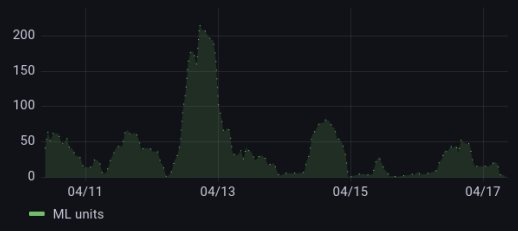
- Model Drift (keep training or you are outdated)
- OOM Errors (your data did not fit into Memory/Machine
- Software Updates (stay fresh, stay safe)
- New data is low quality
- General model degradation (investigate what's happening)

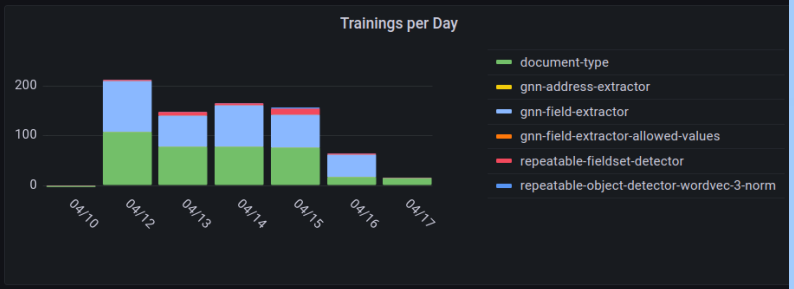


Models need constant care and monitoring, ideally fully automated
@Parashift we have ~1000 ML models deployed / trained at any time, hard to keep on top of everything.

Training Type: None

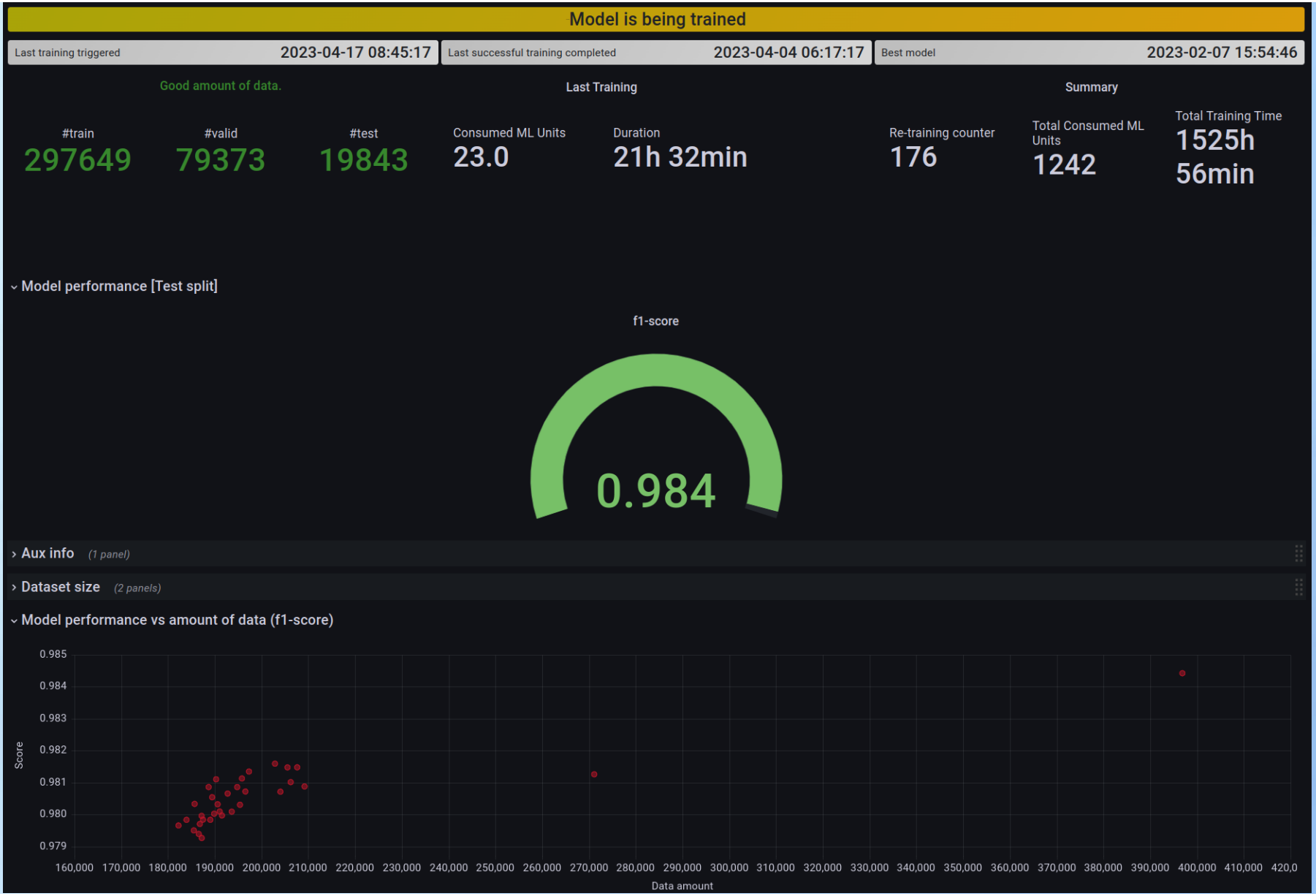






Trainings

Created At	Extraction Type	Subject Type	Subject Id	Duration	Stage	Status	Trigger Reason	Preprocessing Job	Training Job	Error
2023-04-17 06:56:11	document-type	TENANT	2776	00:02:28.59975	PREPROCESSING	RUNNING	[{"reason": "signif…	ts_prd_document_type_te…		
2023-04-17 06:55:18	gnn-field-extractor	FIELD_TYPE	18113	00:03:20.984842	PREPROCESSING	RUNNING	[{"reason": "stale_…	ts_prd_gnn_field_extracto…		
2023-04-17 06:54:48	gnn-field-extractor	FIELD_TYPE	116703	00:03:51.048621	PREPROCESSING	RUNNING	[{"reason": "stale_…	ts_prd_gnn_field_extracto…		
2023-04-17 06:52:18	gnn-field-extractor	FIELD_TYPE	255381	00:06:21.445505	PREPROCESSING	RUNNING	[{"reason": "signif…	ts_prd_gnn_field_extracto…		
2023-04-17 06:46:47	gnn-field-extractor	FIELD_TYPE	19254	00:11:52.006394	PREPROCESSING	RUNNING	[{"reason": "stale_…	ts_prd_gnn_field_extracto…		
2023-04-17 06:45:17	gnn-field-extractor	FIELD_TYPE	19264	00:13:22.100962	PREPROCESSING	RUNNING	[{"reason": "stale_…	ts_prd_gnn_field_extracto…		
2023-04-17 06:41:42	gnn-fieldset-extractor	FIELDSET_TYPE	22622	00:16:57.009131	TRAINING	RUNNING	[{"reason": "stale_…	ts_prd_gnn_fieldset_extra…	ts_prd_gnn_fi…	
2023-04-17 06:41:09	document-type	TENANT	2776	00:14:42.353918	TRAINING	DONE	[{"reason": "signif…	ts_prd_document_type_te…	ts_prd_docum…	
2023-04-17 06:39:12	repeatable-fieldset-detec…	FIELDSET_TYPE	21198	00:19:27.318887	TRAINING	RUNNING	[{"reason": "stale_…	ts_prd_repeatable_fieldse…	ts_prd_repeat…	
2023-04-17 06:39:12	gnn-forms-extractor	FIELDSET_TYPE	22171	00:19:27.31888	TRAINING	RUNNING	[{"reason": "stale_…	ts_prd_gnn_forms_extrac…	ts_prd_gnn_fo…	
2023-04-17 06:38:47	gnn-field-extractor	FIELD_TYPE	216274	00:19:52.714937	TRAINING	RUNNING	[{"reason": "signif…	ts_prd_gnn_field_extracto…	ts_prd_gnn_fi…	
2023-04-17 06:38:39	document-type	TENANT	2714	00:18:16.430233	TRAINING	DONE	[{"reason": "stale_…	ts_prd_document_type_te…	ts_prd_docum…	
2023-04-17 06:36:46	gnn-field-extractor-allowe…	FIELD_TYPE	525	00:21:52.832165	TRAINING	RUNNING	[{"reason": "stale_…	ts_prd_gnn_field_extracto…	ts_prd_gnn_fi…	
2023-04-17 06:35:08	document-type	TENANT	3239	00:15:15.527717	TRAINING	DONE	[{"reason": "stale_…	ts_prd_document_type_te…	ts_prd_docum…	
2023-04-17 06:28:41	repeatable-fieldset-detec…	FIELDSET_TYPE	21238	00:29:58.519052	TRAINING	RUNNING	[{"reason": "stale_…	ts_prd_repeatable_fieldse…	ts_prd_repeat…	
2023-04-17 06:25:37	document-type	TENANT	2776	00:15:14.285729	TRAINING	DONE	[{"reason": "stale_…	ts_prd_document_type_te…	ts_prd_docum…	
2023-04-17 06:23:15	gnn-field-extractor	FIELD_TYPE	221324	00:33:25.216226	TRAINING	DONE	[{"reason": "stale_…	ts_prd_gnn_field_extracto…	ts_prd_gnn_fi…	
2023-04-17 06:20:44	gnn-field-extractor	FIELD_TYPE	222055	00:29:53.635045	TRAINING	DONE	[{"reason": "stale_…	ts_prd_gnn_field_extracto…	ts_prd_gnn_fi…	


Model is being trained
Last training triggered
2023-04-17 08:45:17
Last successful training completed
2023-04-04 06:17:17
Best model
2023-02-07 15:54:46
Good amount of data.
Last Training
Summary
#train
297649
#valid
79373
#test
19843
Consumed ML Units
23.0
Duration
21h 32min
Re-training counter
176
Total Consumed ML Units
1242
Total Training Time
1525h 56min
Model performance [Test split]
f1-score
0.984
Aux info (1 panel)
Dataset size (2 panels)
Model performance vs amount of data (f1-score)
Score
0.985
0.984
0.983
0.982
0.981
0.980
0.979
160,000 170,000 180,000 190,000 200,000 210,000 220,000 230,000 240,000 250,000 260,000 270,000 280,000 290,000 300,000 310,000 320,000 330,000 340,000 350,000 360,000 370,000 380,000 390,000 400,000 410,000 420,0
Data amount

Stay out of Trouble

Data Privacy & Life altering Decisions

If you don't pay for the product, you are the product

\+ ML algorithms can find out a lot about you

\- ML algorithms can find out a lot about you

Data Protection

- **How to prevent my data to leak to others**
  - Manage Data Access (internally)
  - Store as little as necessary
  - Use some lossy compression to store data
  - Anonymize Data (people are horrible at this)
- **How to prevent ML models to spit out my secrets?** (ChatGPT & Samsung code)
  - Very tricky with generative models
  - Easier w/ other models

Black Box

- **Decision making has to be interpretable/ understandable/explainable**
  - Not everything is explainable in modern algorithms
  - One can examine the blackbox
  - One does not need to understand the full blackbox, but come up with reasons why a decision was taken
  - i.e. "you would have been given the loan if you earned 20k more"
  - Find smallest changes to do to the data to generate the desired outcome

Model Bias

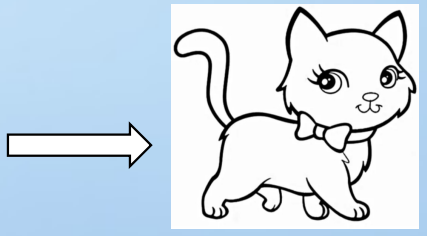
- **Models come from data**
  - Any bias from data will be inherited by the model
- **Must be very careful to use un-biased data**
  - Education
  - Financial
  - Employment

Stay out of Trouble

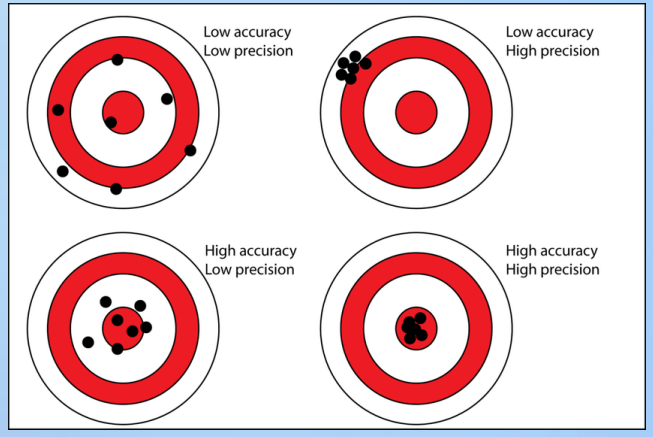
Careful what you report

- Any reported number can be used for misinterpretation
- Specific ML lingo that w/o context can be hard to convey
- Benchmarking, evaluating model performance is not trivial
  - Train, test, validation scores
  - Everybody loves "accuracy"
  - Generalization to out of distribution tasks?

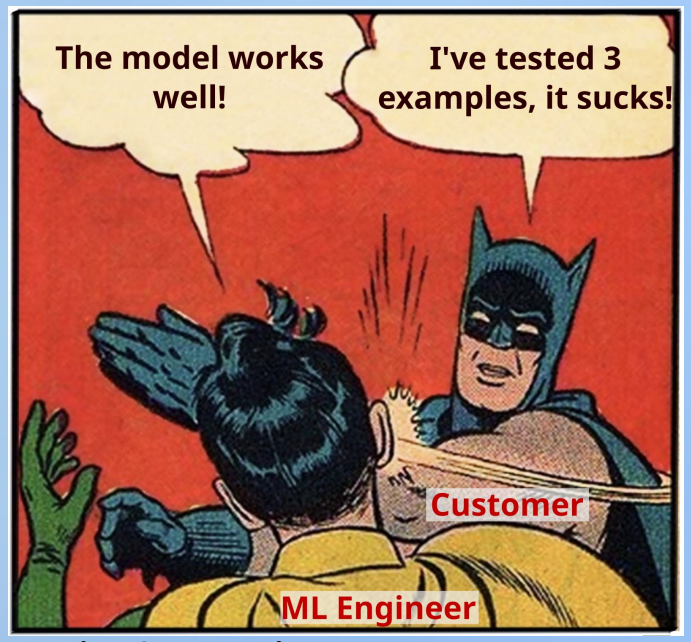
Training data



Production data



Source: http://www.antarcticglaciers.org/glacial-geology/dating-glacial-sediments2/precision-and-accuracy-glacial-geology



Both POVs can be correct

Good Data Beats Good Algorithms

Garbage In Garbage Out

Clean & Consistent data is the most important piece to success.
Proof of Concept, **target accuracy = 0.85**, **current accuracy = 0.7**, **customer unhappy.**

Action Taken	Resulting Accuracy
None	0.70
Data Post-processing improvements	0.72
Introduce rule based algorithms	0.74
Tweak ML model architecture	0.75
Clean up the f**** data	0.88

Fun Fact I
Applying the same evaluation metric to the human annotated data resulted in a value of accuracy ~0.2

Fun Fact II
There is a common misconception that: "My people do not make mistakes"

Good Data Beats Good Algorithms

Good Data Is Expensive

- Producing high quality data is challenging and time consuming (most of the time)
  - Human annotated data = Gold
  - Self supervised learning
  - Semi-supervised learning

- Probably a very significant amount of $$ for ChatGPT went into curation of high quality data