cdisc

2023
EUROPE
INTERCHANGE
COPENHAGEN | 26-27 APRIL

Automatic Defining ADaM for new Clinical Studies Using Machine Learning

Thomas Rye Olsen,
Student at Department of Computer Science, University of Copenhagen

Henning Pontoppidan Föh,
Statistical Programming Director, Biostatistics, Novo Nordisk A/S

Meet the Speakers

Thomas Rye Olsen

Title: Student

Organization: Department of Computer Science, University of Copenhagen

- Studying Machine Learning and Data Science on his third year
- Have been working with Biostatistics, Novo Nordisk applying ML
- Recently become a student assistant at Novo Nordisk

Henning Pontoppidan Föh

Title: Statistical Programming Director

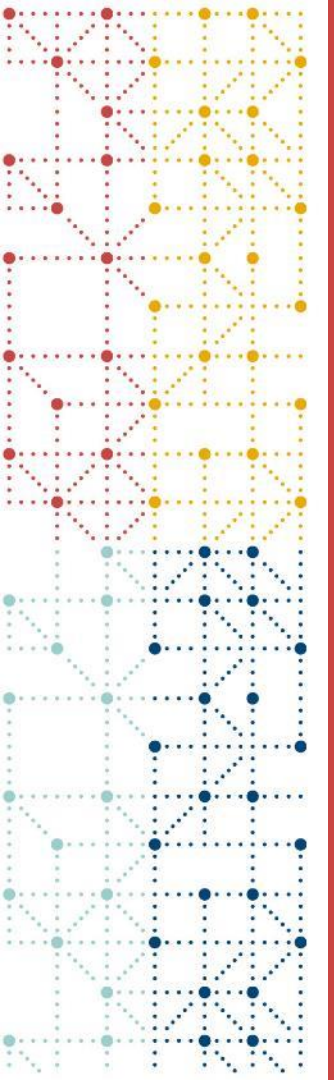
Organization: Biostatistics, Novo Nordisk A/S

- 15+ years of pharmaceutical industry experience, within various arears
- MSc in Physics and worked as researcher as well as SAS consultant
- Currently main interest is the strategic clinical development for new drugs and indications

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC nor of Novo Nordisk*

- *The author(s) have no real or apparent conflicts of interest to report.*



Agenda

1. Background
2. The idea of using ML for ADaM definition
3. Details of the ML algorithm
4. Results and usability

Background

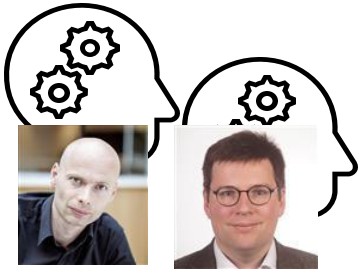
Data is gold

Data = clinical data + clinical metadata

- Data collected from patients
  - Demographics, AEs, endpoints
- Highly regulated by authorities + CDISC
- Available for all studies in standardised format
- All pharma companies have it from their studies

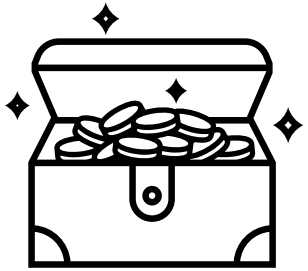
- Descriptions of studies, created by Novo staff
  - Protocol Metadata document (PMD) containing items such as flowchart, study descriptive keywords, etc
  - Analysis data/ADaM description (within the CST)
- Generally, not regulated by authorities
- Available for all studies running in the last decade
- A unique feature of Novo Nordisk !?

A quest for gold requires 3 items…



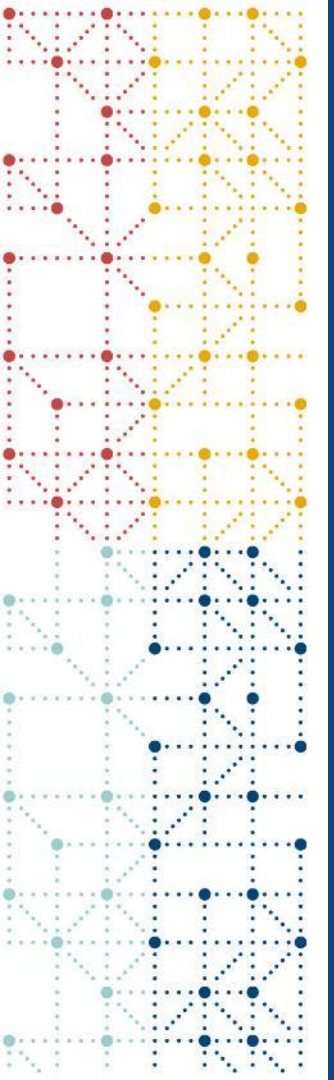
Bright minds

Hard labour



Someone who what to **spend** the gold

Let’s mine the metadata gold



The idea of using ML for ADaM definition

Creating analysis metadata today

- For every study, the trial programmer has to create a structured description of all analysis (ADaM) datasets in an Excel sheet (CST)
  - Both a functional and a requirement for define.xml
  - Takes days/weeks of work
  - Is a tedious and errorprone job
  - Most of the contents are determined by the design, therapy area, and clinical project

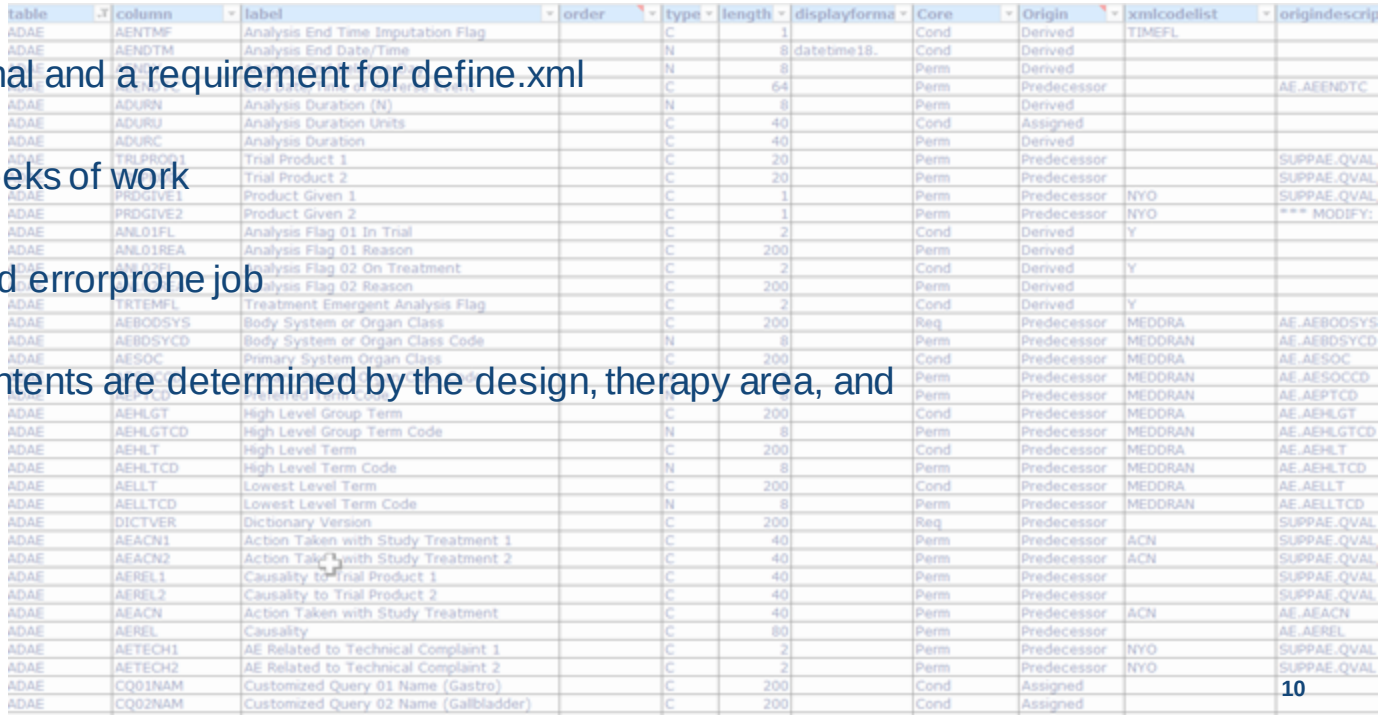
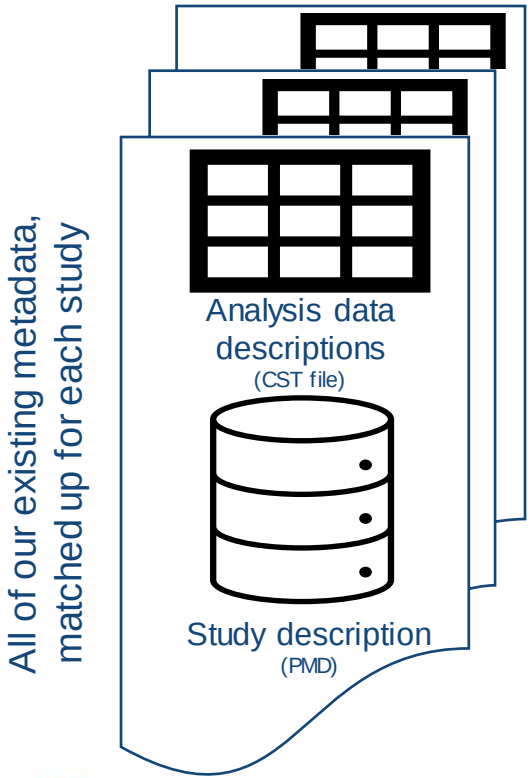
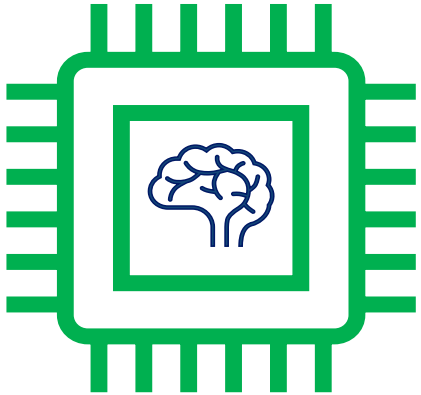


table	column	label	order	type	length	displayforma	Core	Origin	xmlcodelist	origindescrip
ADAE	AENTMF	Analysis End Time Imputation Flag		C	1		Cond	Derived	TIMEFL	
ADAE	AENDTM	Analysis End Date/Time		N	8	datetime18.	Cond	Derived		
ADAE	ADURN	Analysis Duration (N)		N	8		Perm	Derived		
ADAE	[illegible]	[illegible]		C	64		Perm	Predecessor		AE.AEENDTC
ADAE	ADURN	Analysis Duration (N)		N	8		Perm	Derived		
ADAE	ADURU	Analysis Duration Units		C	40		Cond	Assigned		
ADAE	ADURC	Analysis Duration		C	40		Perm	Derived		
ADAE	TRLPROD1	Trial Product 1		C	20		Perm	Predecessor		SUPPAE.QVAL
ADAE	TRLPROD2	Trial Product 2		C	20		Perm	Predecessor		SUPPAE.QVAL
ADAE	PRDGIVE1	Product Given 1		C	1		Perm	Predecessor	NYO	SUPPAE.QVAL
ADAE	PRDGIVE2	Product Given 2		C	1		Perm	Predecessor	NYO	*** MODIFY:
ADAE	ANL01FL	Analysis Flag 01 In Trial		C	2		Cond	Derived	Y	
ADAE	ANL01REA	Analysis Flag 01 Reason		C	200		Perm	Derived		
ADAE	ANL02FL	Analysis Flag 02 On Treatment		C	2		Cond	Derived	Y	
ADAE	[illegible]	Analysis Flag 02 Reason		C	200		Perm	Derived		
ADAE	TRTEMFL	Treatment Emergent Analysis Flag		C	2		Cond	Derived	Y	
ADAE	AEBODSYS	Body System or Organ Class		C	200		Req	Predecessor	MEDDRA	AE.AEBODSYS
ADAE	AEBDSYCD	Body System or Organ Class Code		N	8		Perm	Predecessor	MEDDRAN	AE.AEBDSYCD
ADAE	AESOC	Primary System Organ Class		C	200		Cond	Predecessor	MEDDRA	AE.AESOC
ADAE	[illegible]	[illegible]		N	8		Perm	Predecessor	MEDDRAN	AE.AESOCCD
ADAE	[illegible]	[illegible]		C	200		Perm	Predecessor	MEDDRAN	AE.AEPTCD
ADAE	AEHLGT	High Level Group Term		C	200		Cond	Predecessor	MEDDRA	AE.AEHLGT
ADAE	AEHLGTCD	High Level Group Term Code		N	8		Perm	Predecessor	MEDDRAN	AE.AEHLGTCD
ADAE	AEHLT	High Level Term		C	200		Cond	Predecessor	MEDDRA	AE.AEHLT
ADAE	AEHLTCD	High Level Term Code		N	8		Perm	Predecessor	MEDDRAN	AE.AEHLTCD
ADAE	AELLT	Lowest Level Term		C	200		Cond	Predecessor	MEDDRA	AE.AELLT
ADAE	AELLTCD	Lowest Level Term Code		N	8		Perm	Predecessor	MEDDRAN	AE.AELLTCD
ADAE	DICTVER	Dictionary Version		C	200		Req	Predecessor		SUPPAE.QVAL
ADAE	AEACN1	Action Taken with Study Treatment 1		C	40		Perm	Predecessor	ACN	SUPPAE.QVAL
ADAE	AEACN2	Action Taken with Study Treatment 2		C	40		Perm	Predecessor	ACN	SUPPAE.QVAL
ADAE	AEREL1	Causality to Trial Product 1		C	40		Perm	Predecessor		SUPPAE.QVAL
ADAE	AEREL2	Causality to Trial Product 2		C	40		Perm	Predecessor		SUPPAE.QVAL
ADAE	AEACN	Action Taken with Study Treatment		C	40		Perm	Predecessor	ACN	AE.AEACN
ADAE	AEREL	Causality		C	80		Perm	Predecessor		AE.AEREL
ADAE	AETECH1	AE Related to Technical Complaint 1		C	2		Perm	Predecessor	NYO	SUPPAE.QVAL
ADAE	AETECH2	AE Related to Technical Complaint 2		C	2		Perm	Predecessor	NYO	SUPPAE.QVAL
ADAE	CQ01NAM	Customized Query 01 Name (Gastro)		C	200		Cond	Assigned		
ADAE	CQ02NAM	Customized Query 02 Name (Gallbladder)		C	200		Cond	Assigned		

Creating analysis metadata using ML

Step 1: Train a machine learning model

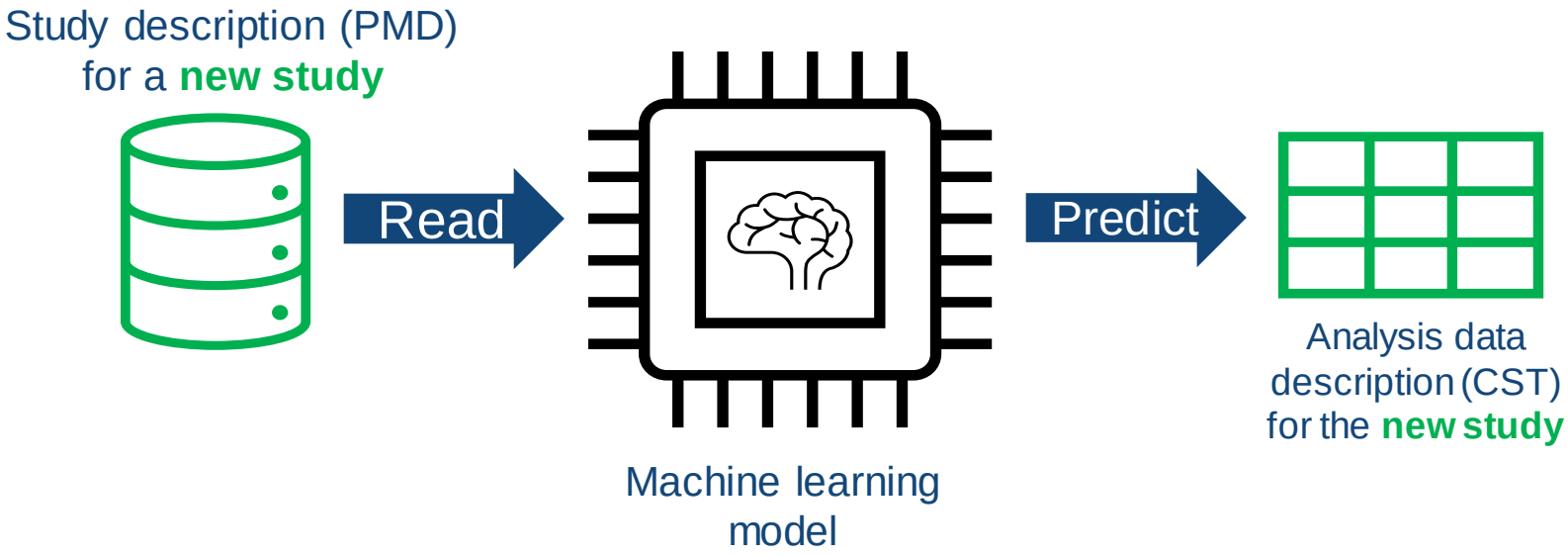


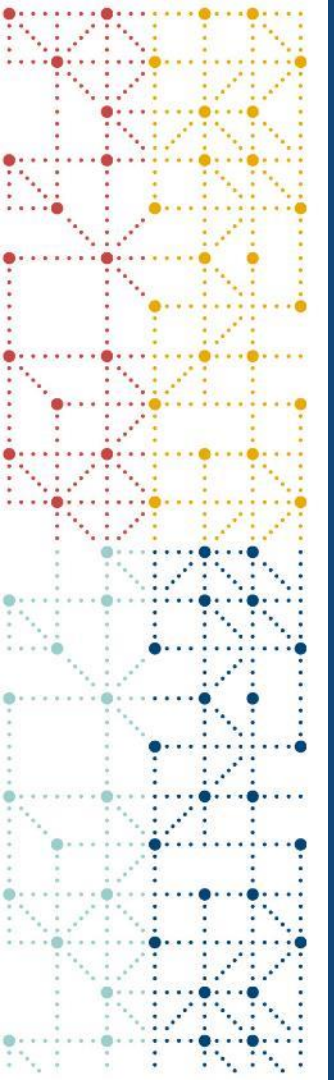


Machine learning model under construction

Creating analysis metadata using ML

Step 2: Use the machine learning model

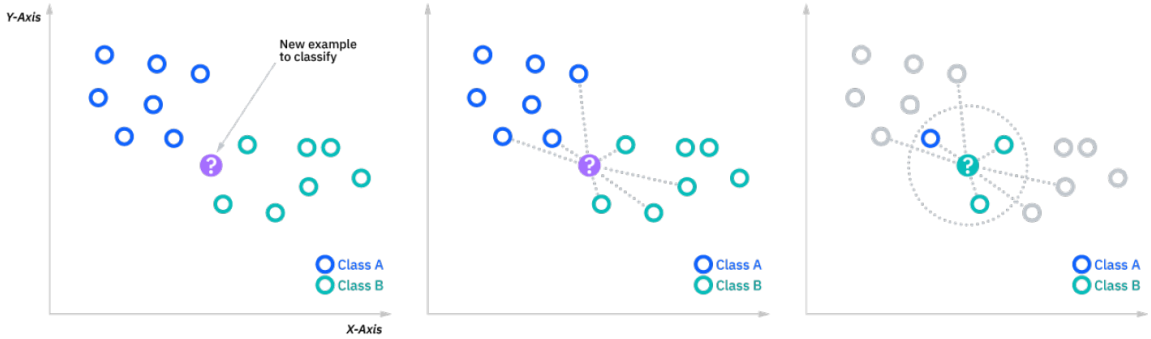




Details of the ML algorithm

Inside the belly of
Supervised learning with RKNN-FS

- Supervised learning: We know what we are looking for
- Random K Nearest Neighbors (RKNN)
  - KNN: Distance by similarity in features
  - Random choice of features
  - Ensemble model

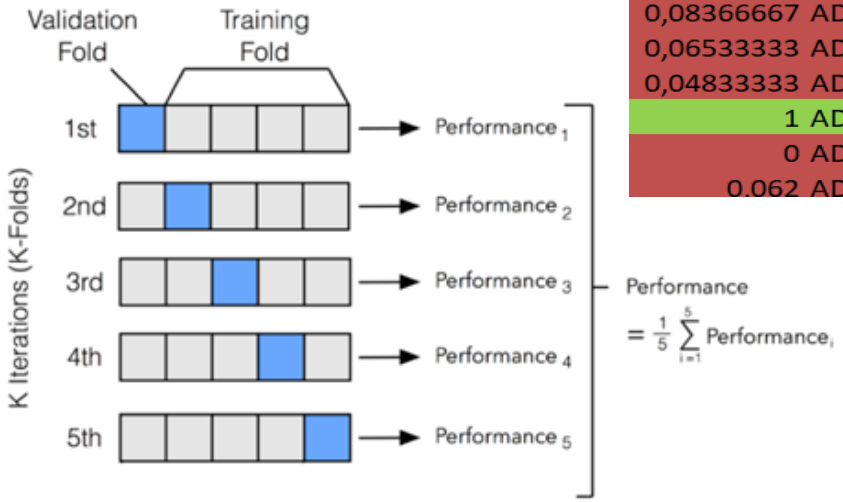


- Feature Selection (FS):
  - Check what features gives best forecasts
  - Iteratively discard features that seem redundant
  - Better generalization in theory and better results in practice
  - Better insights

Evaluating performance

- Confidence is key!
- Tested against complete studies
  - Confidence of correct classifications
- Cross validation:
  - Train on 80% data, test on 20%
  - 5 Rounds
  - Unbiased estimate
- Threshold: 80% confidence

include	tables
1	ADSL
0,96133333	ADAE
0,95966667	ADEC
0,50333333	ADECEN
0,94266667	ADEG
0,485	ADHYPOEN
0,99966667	ADLB
0,94333333	ADPC
0,60866667	ADPDC
0,60866667	ADPDP
0,993	ADPE
0,763	ADPP
0,004	ADPROF
0,08366667	ADQS
0,06533333	ADRESP
0,04833333	ADSMPGEN
1	ADVS
0	ADVSEN
0,062	ADADJ



Workload spared

- Classifications with high confidence (above 80%):
  - ~25 ADaM datasets out of 38
  - ~3128 variables/columns out of 3479
  - Much better than expected!

- Errors? 
  - 98% of datasets with high confidence are correct
  - 0.6 datasets per study are incorrect
  - 97% of variables with high confidence are correct
  - 123 variables per study are incorrect
  - Handled by project-responsible programmer

Results and usability

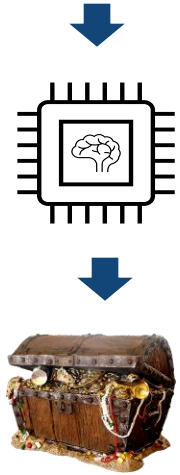
How much gold did we extract so far?

~80% of analysis data description can be correctly predicted

Much less manual work on this task

Going forward – spending the gold

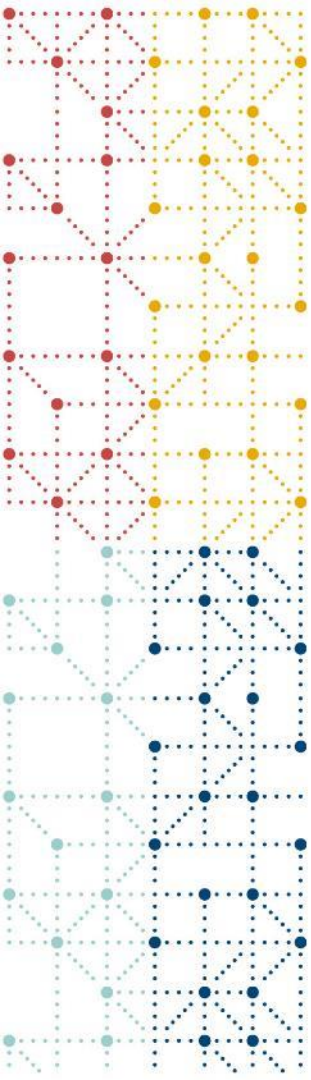
- Build and deploy a user **application**
- Interactive tool to build the ADaM-definition
- Upload study description (PMD)



- High confidence predictions automatically defined
- ADaM datasets and variables with low confidence are presented
- Programmer decides when the confidence is low

Going forward – purifying the gold

- Examine an **adaptive recommender system**
  - Recommend tables and columns that have not been defined in ADaM
  - Learns iteratively from the choices the programmers make
  - Potentially even more automation



Conclusion

Learnings & conclusions

- We can use ML for predicting new ADaM trials definitions using supervised learning trained on previous clinical metadata
- RKNN-FS seems to be a good performing algorithm for when we have few data to train on
  - RKNN-FS can predict approximately 80% of ADaM datasets including variables correctly and with high confidence
- Tedious and repetitive ADaM definitions that can be automated → Trial programmer can focus on non-standardized items
- In the future we are looking into building an app using the ML algorithm to forecast ADaM definitions for new trials
  - Examining the possibility to build an adaptive recommender system

Thank You!

Questions and comments are welcome!

Thomas Rye Olsen, vlh769@alumni.ku.dk

Henning P. Föh, hpf@novonordisk.com