A Variety of Approaches for Cohort Selection for Clinical Trials

Xavier Tannier and many others

San Francisco, USA, Nov. 2nd, 2018

People

People

Hugo Cisneros

Christel Daniel

Tom Denat

Matthieu Doutreligne

Catherine Duclos

Nicolas Griffon

Claire Hassen-Khodja

Ivan Lerner

Guillaume Luquiens

Nicolas Paris

Adrien Parrot

Éric Sadou

Cyril Saussol

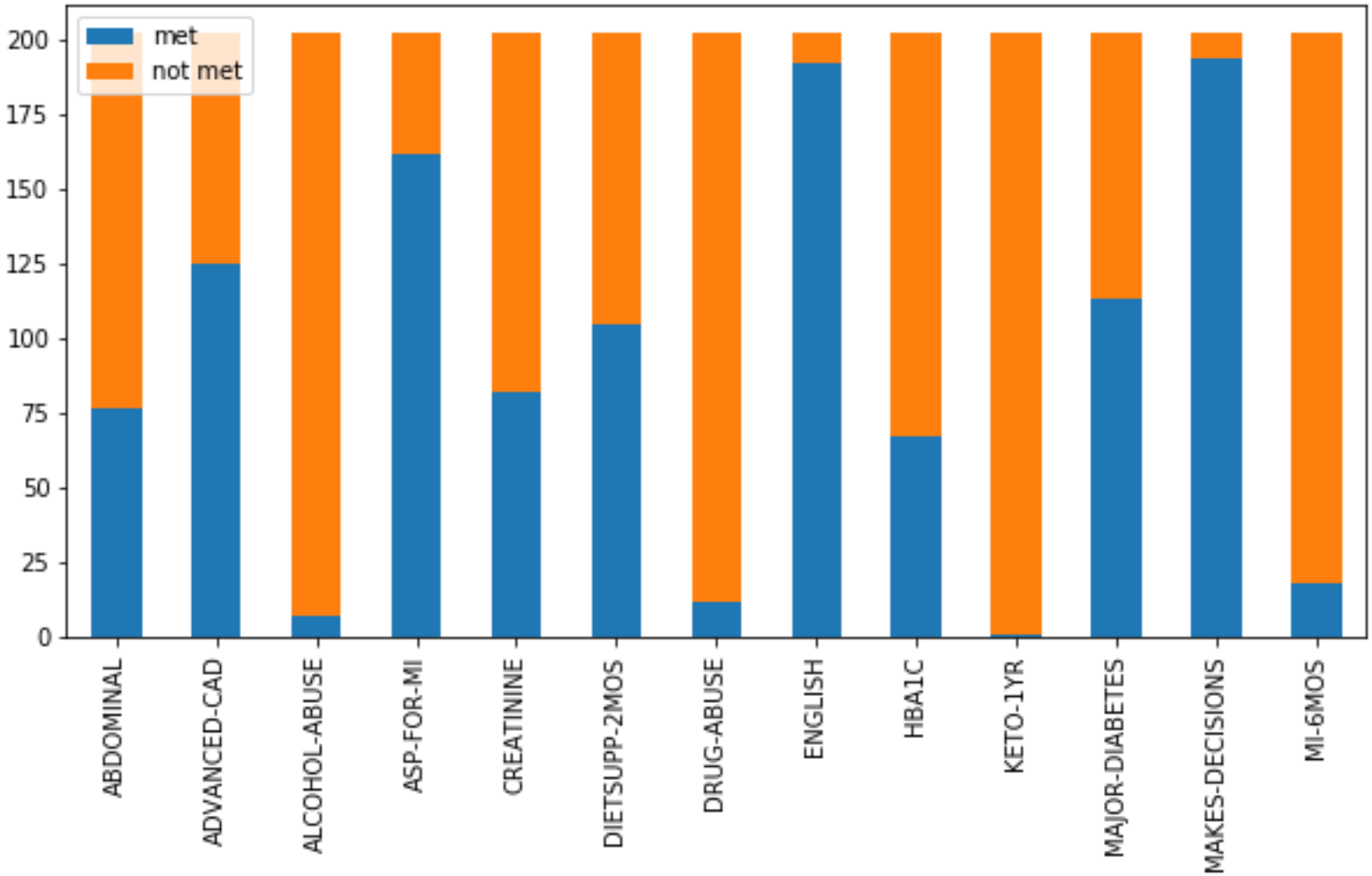
Xavier Tannier

Pascal Vaillant

Material

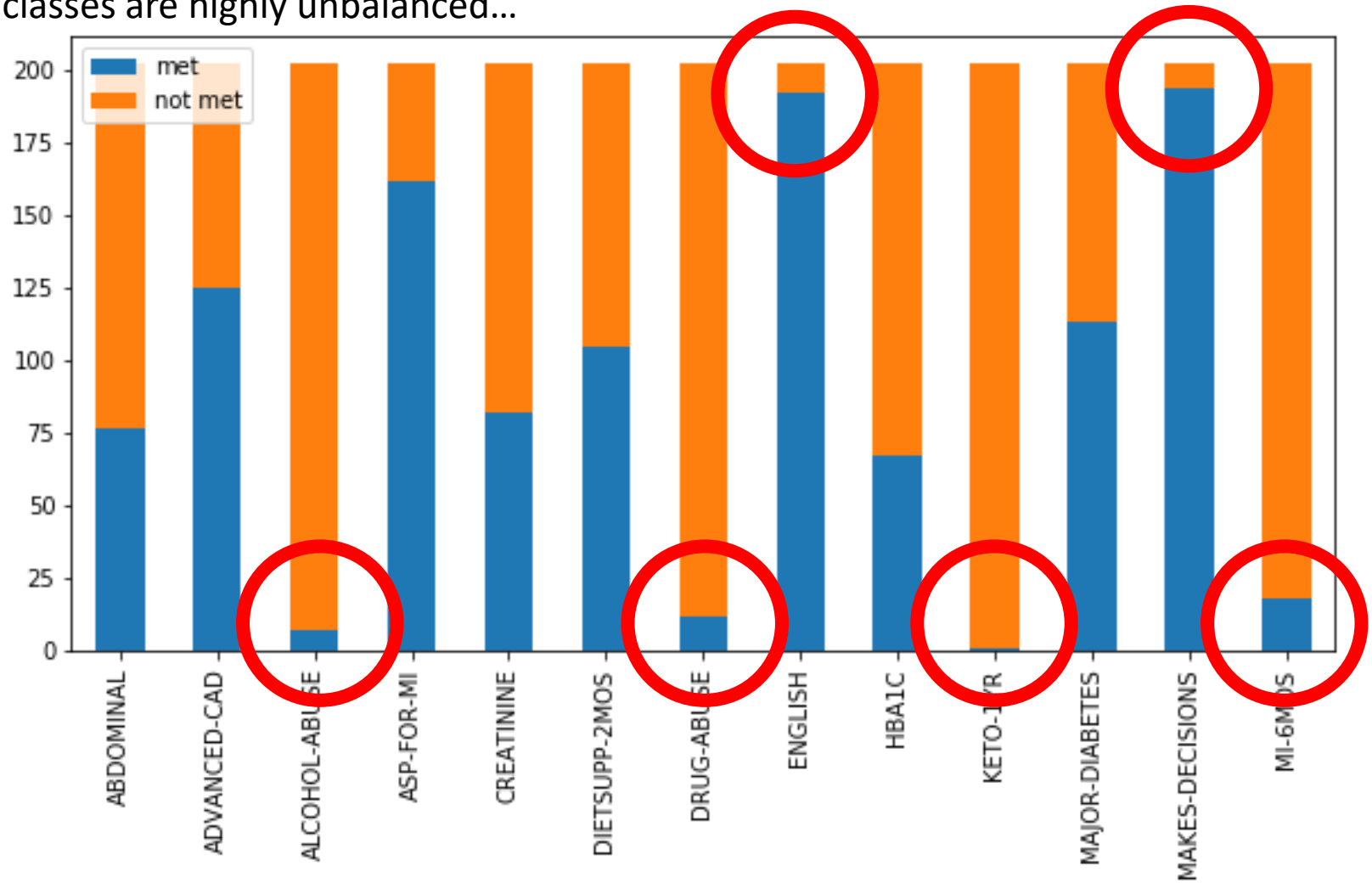
n2c2 dataset

Training set (202 patients… that we split into 80% train et 20% dev)



n2c2 dataset

Some classes are highly unbalanced…



UMLS Metathesaurus

- Unified Medical Language System
- 1 million biomedical concepts
- 5 million concept names

- Used to collect medical concepts of interest

MIMIC III

- Open database
- Deidentified health-related data
- 40,000 patients
- > 2 million clinical notes

- Used for various purposes
  - Compute embeddings
  - Augment training set
  - Find new rules

Stanford CoreNLP

CoreNLP

- A set of human language technology tools for 7 languages
- Not specialized in clinical texts (but has other advantages)

- Used to tokenize sentences and words

Apache cTAKES

- NLP system specialized in information extraction from medical record clinical free-text.

- Used for detecting concepts that are negated or uncertain in the notes.
- We also extended the negation detection with a few rules.

HeidelTime

- A multilingual and cross-domain temporal tagging tool.
- Used for date normalization.

Home-made tools

- Word2vec-based spelling correction
  (100 most similar with Levensthein distance ≤ 2)

- Section splitter
  (headline collection + normalization)

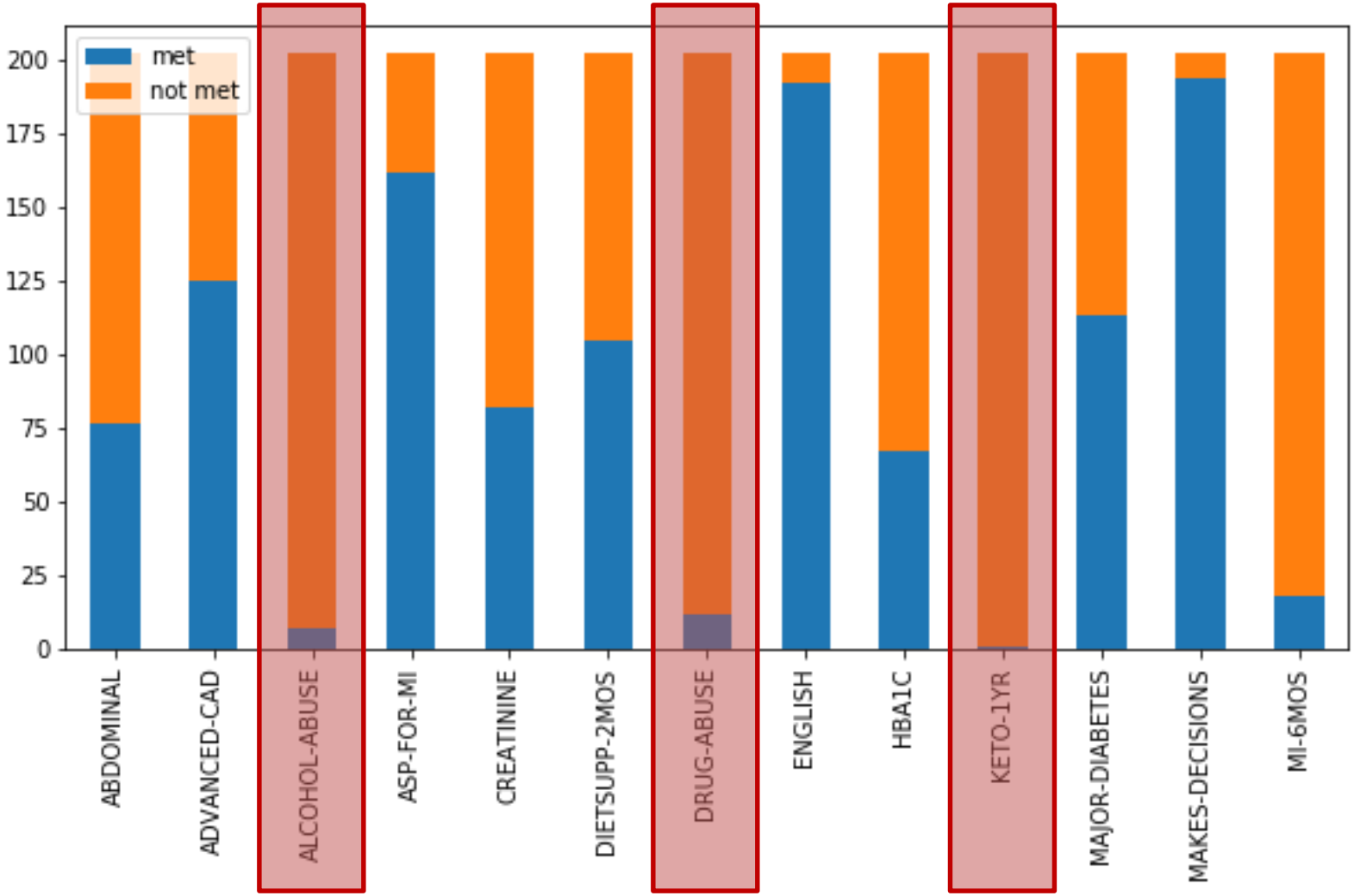
- Patient gender extractor
  (majority of “he/his” *vs.* “she/her”)

Methods & Results

1. Semi-supervised methods

1. Semi-supervised methods
2. Rule-based methods
3. Terminology-based methods
4. Overall results

Semi-supervised learning methods



General principle

Alcohol use over weekly recommended limits

Drug abuse, current or past

Diagnosis of ketoacidosis in the past year

- Not enough annotated (positive) data

n2c2

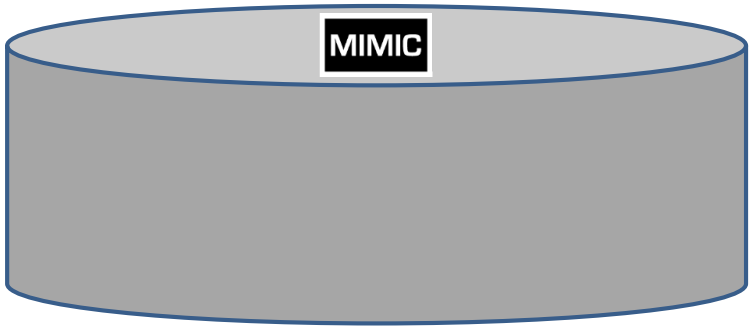
General principle

Alcohol use over weekly recommended limits

Drug abuse, current or past

Diagnosis of ketoacidosis in the past year

- Not enough annotated (positive) data
- Idea: massively augment data with MIMIC notes



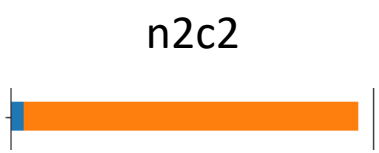
General principle

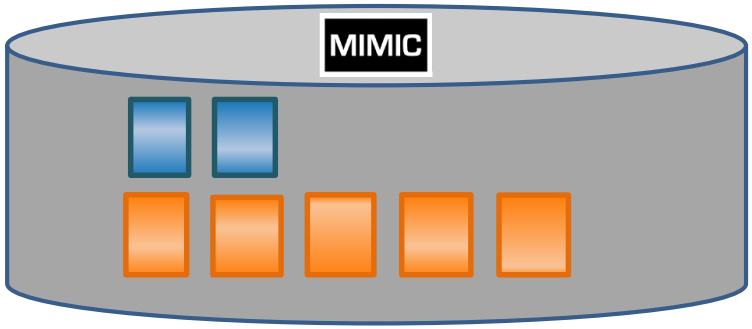
Alcohol use over weekly recommended limits

Drug abuse, current or past

Diagnosis of ketoacidosis in the past year

- Not enough annotated (positive) data
- Idea: massively augment data with MIMIC notes
- How to build an annotated corpus from MIMIC (silver standard)?





Building a silver standard

1. Find very precise rules for extraction of positive (met) and negative (not met) examples

e.g. for ALCOHOL-ABUSE (similar for DRUG-ABUSE):

Positive	**Negative**
"drinks heavily" *"binge drinking"* *"alcoholism"* *"alcohol abuse"* … (without negation)	*"denies active tobacco or alcoholic beverage usage"* *"no alcohol in * (years|yrs)"* *"clean.{1,40}(alcohol|etoh)"* *"alcohol use status : moderate"* *"etoh : social"* …

(Examples found in n2c2 data or in MIMIC with a **`grep alcohol`**, **`grep etoh`**, **`grep drink`**)

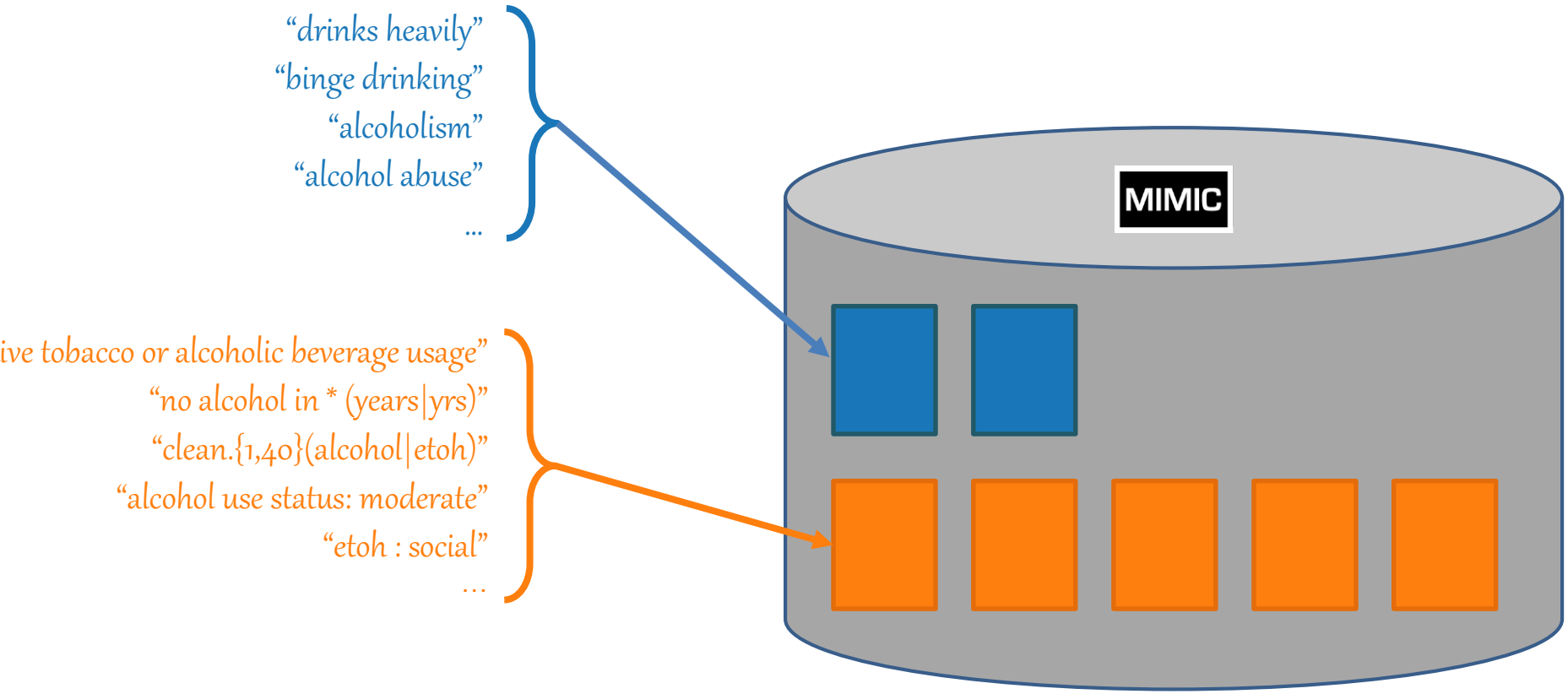
Building a silver standard

1. Find very precise rules for extraction of positive (met) and negative (not met) examples

For KETO-1YR, used ICD-9 code for ketoacidosis in MIMIC

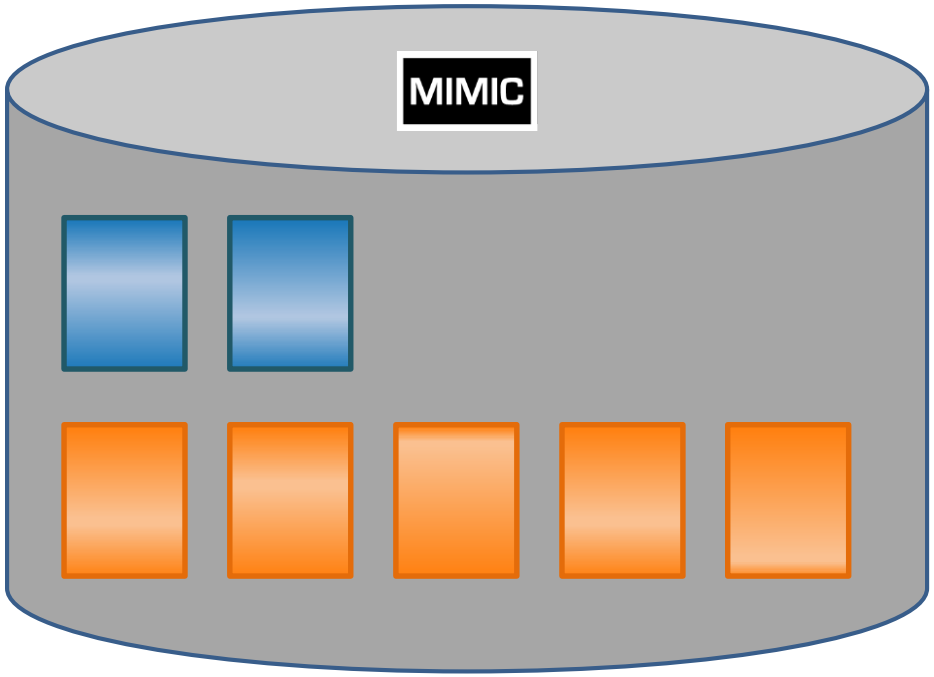
Building a silver standard

2. Match rules in MIMIC to collect "met" and "not met" patients



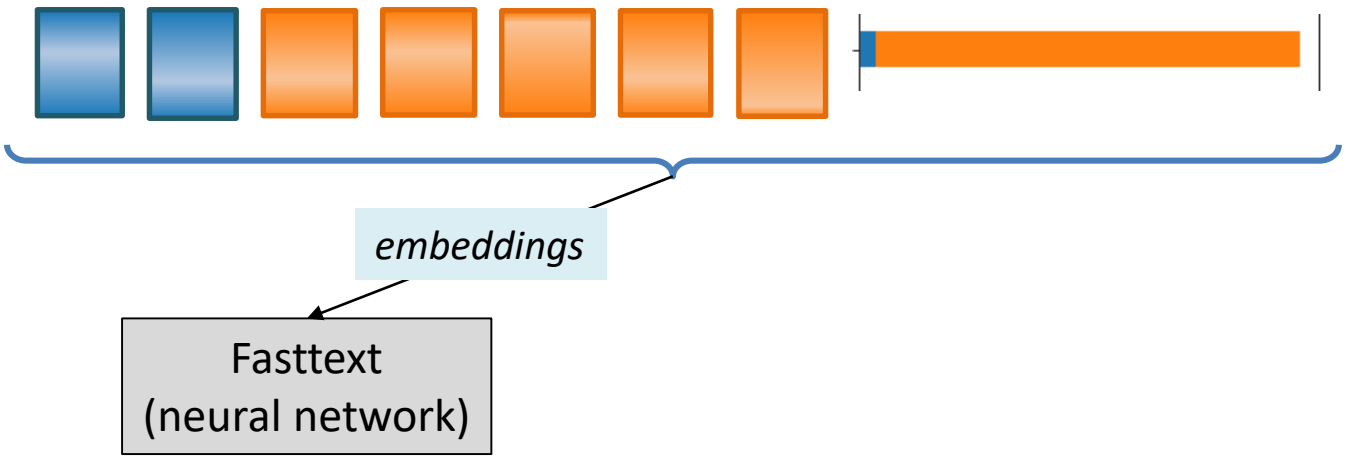
Building a silver standard

3. Count on the redundancy of the reports (several files per patient) to contain variants of the relevant information



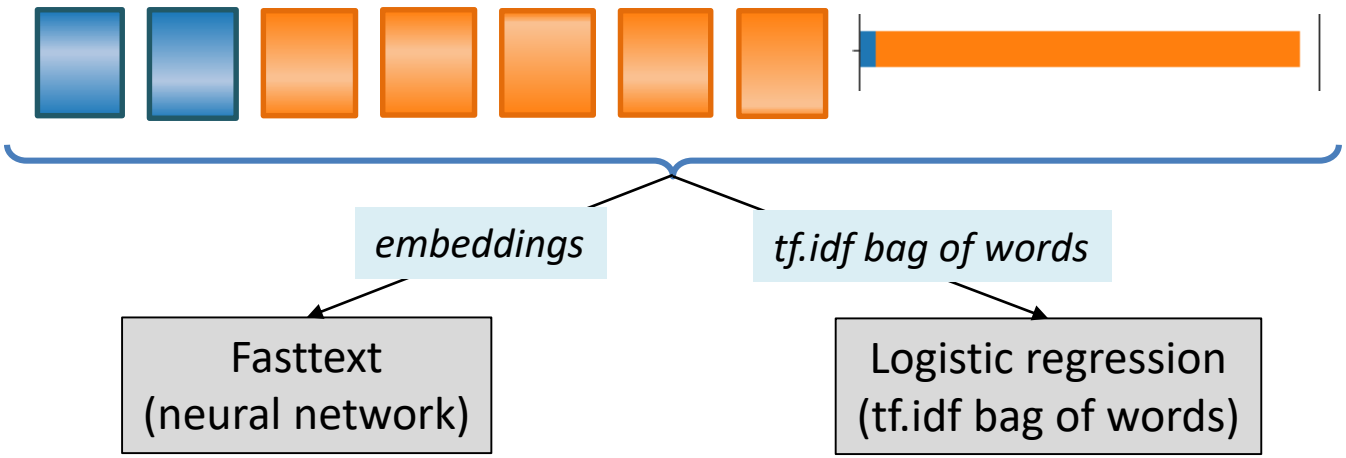
Training a supervised system

4. Train a classifier based on this silver standard + gold standard
   (fed with only the sentences containing seed words)



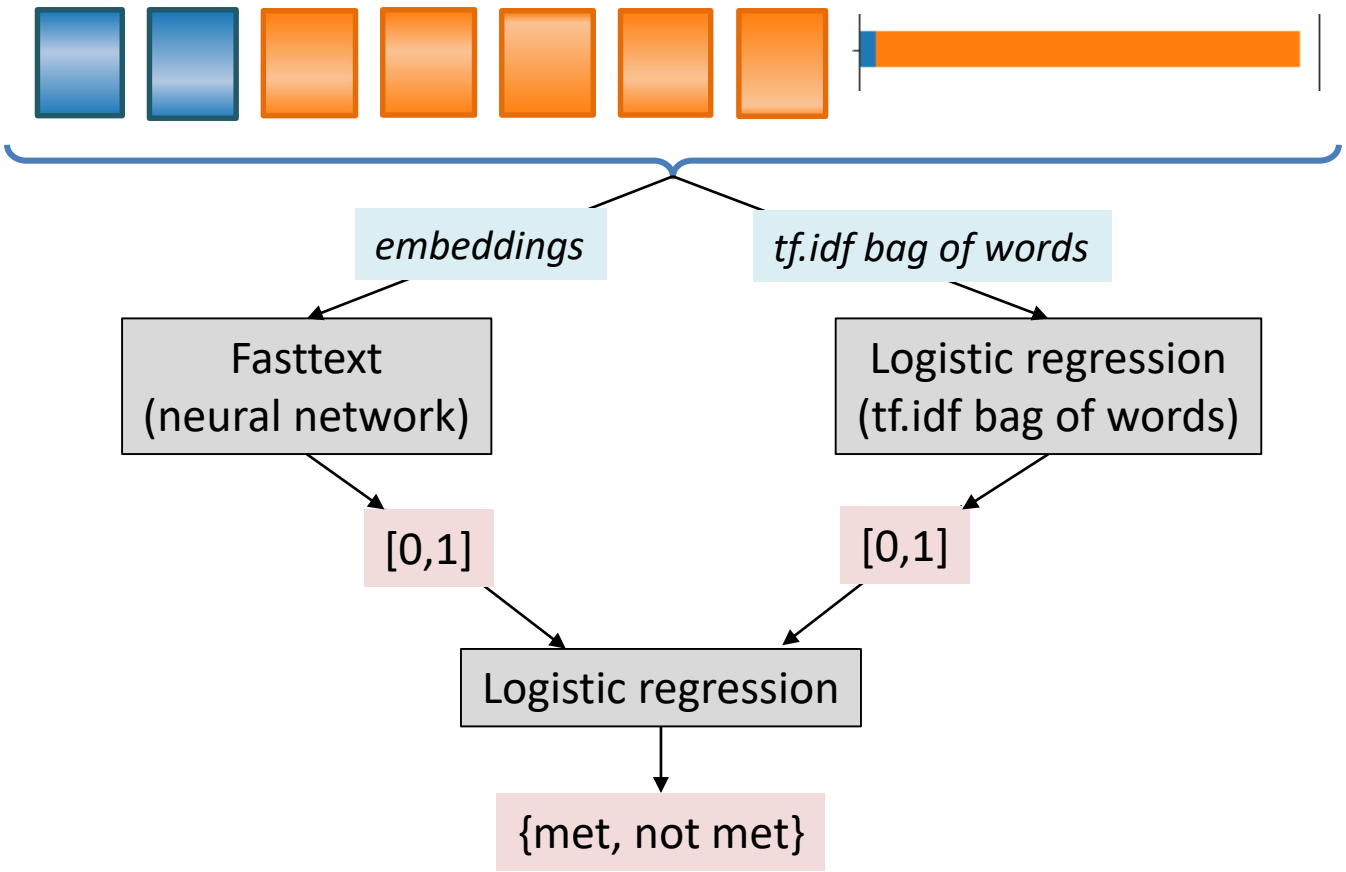
Training a supervised system

5. Train a classifier based on this silver standard + gold standard
   (fed with only the sentences containing seed words)



Training a supervised system

5. Train a classifier based on this silver standard + gold standard
   (fed with only the sentences containing seed words)



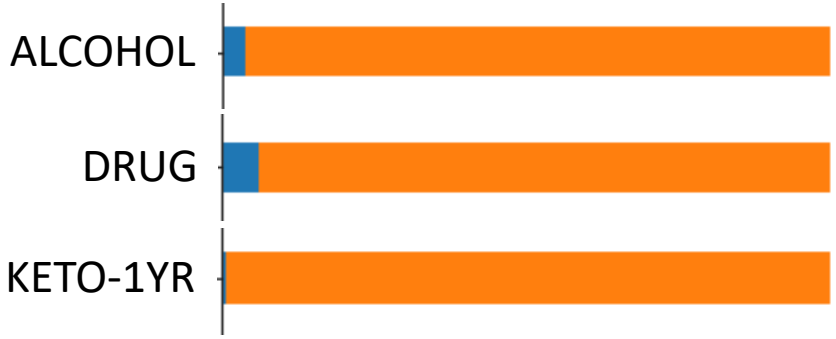
Limits of this method

- Finding **negative examples** is the difficult part
  - We found only 3 criteria out of 13 that made that possible
  - Some papers (see below) describe semi-supervised phenotyping but do not say how they deal with negative examples

- The data is **noisy**, **biased** and **not always representative** of the original data

- Agarwal, V.; Podchiyska, T.; Banda, J. M.; Goel, V.; Leung, T. I.; Minty, E. P.; Sweeney, T. E.; Gyang, E. & Shah, N. H. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc, 2016, 23
- Halpern, Y.; Horng, S.; Choi, Y. & Sontag, D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc, 2016, 23
- Beaulieu-Jones, B. K. & Greene, C. S. Semi-supervised learning of the electronic health record for phenotype stratification. Journal of Biomedical Informatics, 2016, 64

Semi-supervised learning methods: results



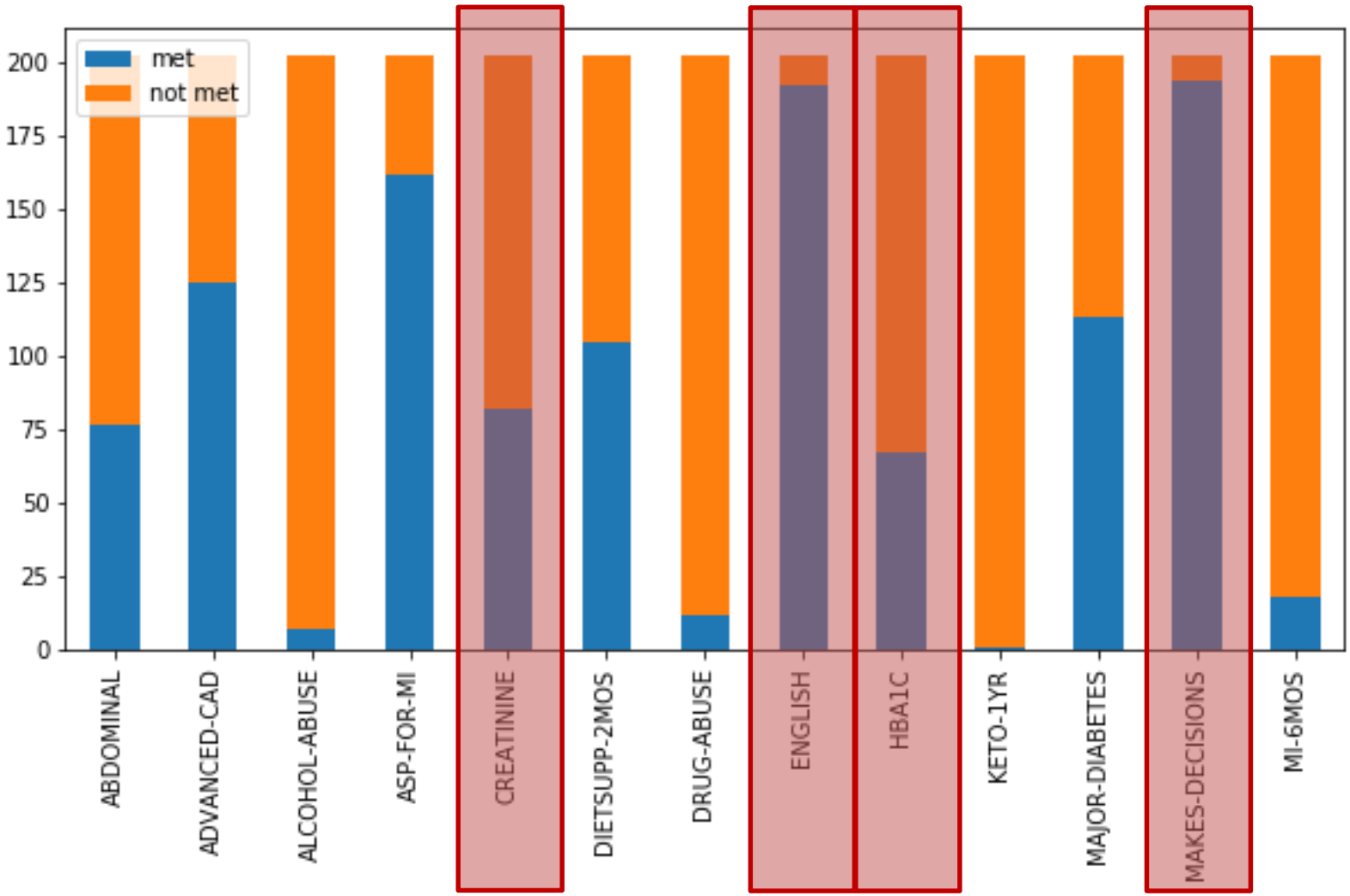
	met			not met		
	Prec.	Rec.	F1	Prec.	Rec.	F1
ALCOHOL	1.00	0.33	0.50	0.98	1.00	0.99
DRUG	0.67	0.67	0.67	0.99	0.99	0.99
KETO-1YR	0.00	0.00	0.00	1.00	1.00	1.00

Methods & Results

2. Rule-based methods

1. Semi-supervised methods
2. Rule-based methods
3. Terminology-based methods
4. Overall results

Rule-based methods



CREATININE – HBA1C

- Numerical values found in:
  - Lab result tables (very few variants)
  - Text (few variants) :
    - *Cr 2.4*
    - *BUN and creatinine 69 and 5.8*
    - *elevated creatinine*
    - *…*
  - Patterns enriched with a **`grep Creat`** and **`grep HBA1C`** in MIMIC
    - *BLOOD Glucose-113* UreaN-11 Creat-0.5 Na-142*

ENGLISH – MAKES-DECISION

- Very few counter-examples ("not met") in the training set
- High linguistic variation expected
- Difficult to find explicit examples of "met"
  (rarely mentioned that the person can make decisions or speak English)
  (→ difficult to apply a semi-supervised approach as described previously)

→ Rules! (based on training set + MIMIC)

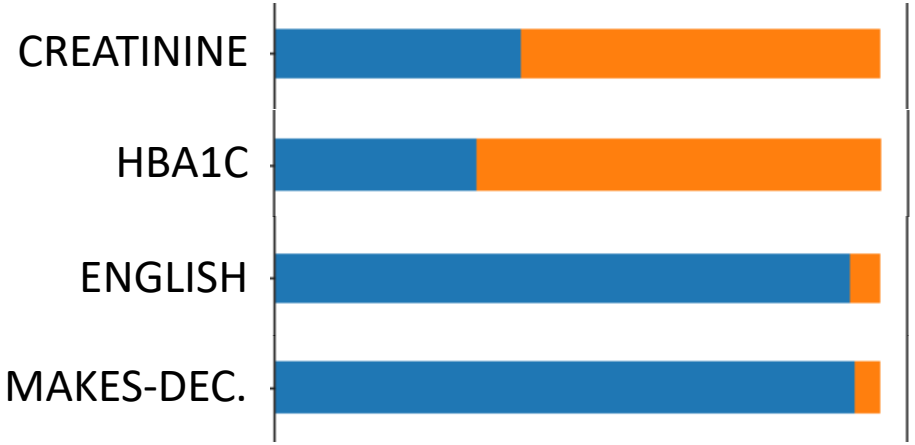
ENGLISH – MAKES-DECISION

Patient must speak English

Patient must make their own medical decisions

- "not met" rules:
  - MAKES-DECISION:
    - Dementia
    - Mental retardation
    - *"not able to answer questions"*
    - Tutorship/curatorship?
  - ENGLISH:
    - *"X-speaking"*
    - *"needs a translator"*
    - *"few words of English"* (??)

Rule-based methods: results



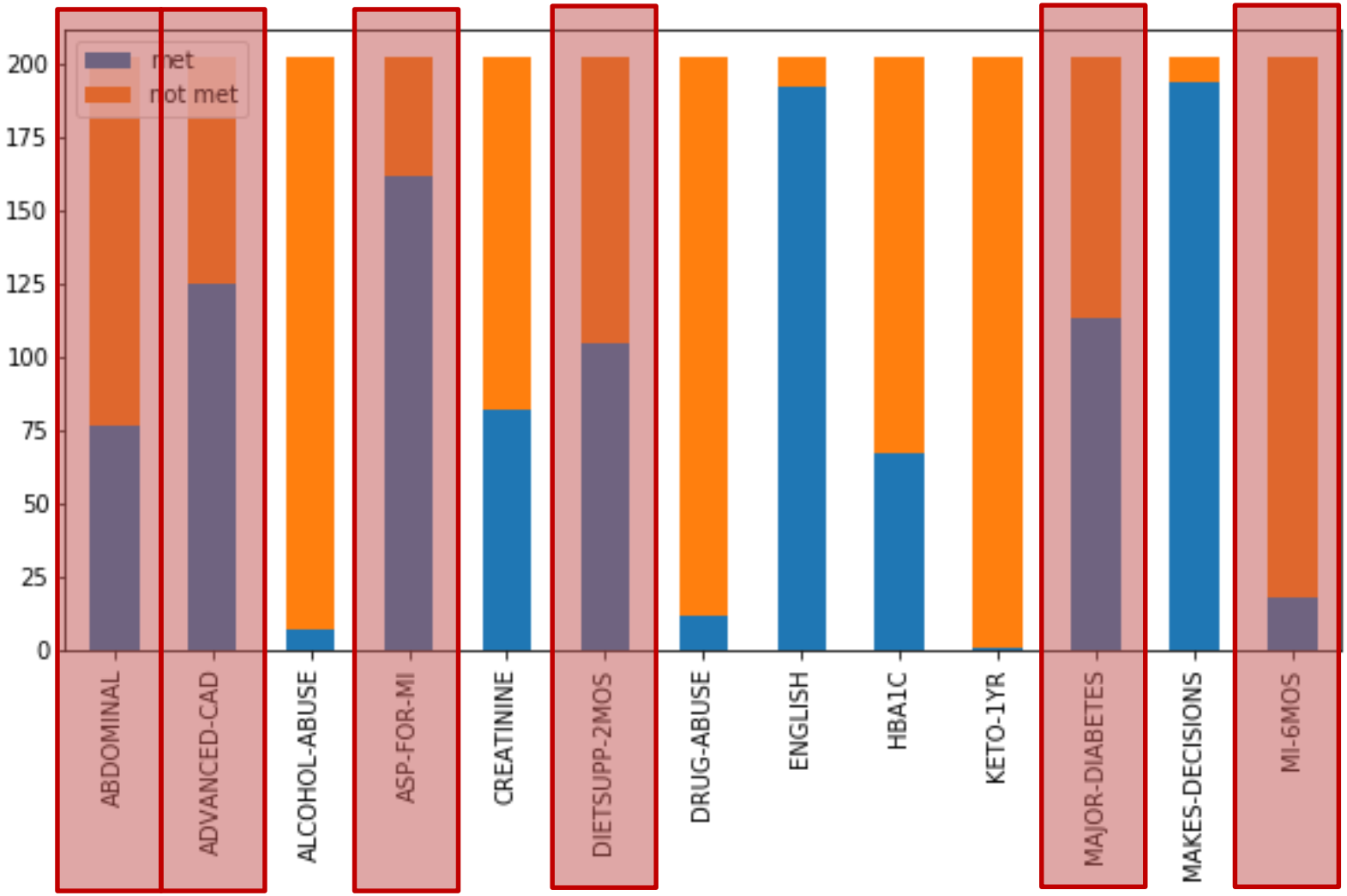
	met			not met		
	Prec.	Rec.	F1	Prec.	Rec.	F1
CREATININE	0.87	0.83	0.85	0.93	0.95	0.94
HBA1C	1.00	0.77	0.87	0.86	1.00	0.93
ENGLISH	0.94	0.99	0.96	0.89	0.62	0.73
MAKES-DEC.	0.96	0.98	0.97	0.00	0.00	0.00

Methods & Results

3. Terminology-based methods

1. Semi-supervised methods
2. Rule-based methods
3. Terminology-based methods
4. Overall results

Terminology-based methods



Terminology-based methods

- Medical knowledge is important
- Variation is important but medical concepts are very good clues
- Difficult to find explicit examples of "not met"
  (rarely mentioned that the person did NOT have abdominal surgery, MI, etc.)
  (→ difficult to apply a semi-supervised approach as described previously)

→ Concept matching

Terminology-based methods: main steps

1. **Terminology**: use UMLS to collect descendants and synonyms of general concepts corresponding to the inclusion criteria
2. Use **spelling correction** (word2vec based)
3. **Negation, uncertainty**: use cTAKES + home-made rules to discard negated or uncertain events in text (when necessary).
4. **Structure**: restrict the concept search to specific sections (when relevant).
5. **Temporal dimension** (temporally-restricted criteria): restrict to recent reports or to sentences containing a date within the inclusion date range (Heideltime)

→ **Decision**: trigger **met** decision as soon as at least one relevant term is found in the text (two sub-criteria for ADVANCED-CAD).

Terminology-based methods: examples

ABDOMINAL 

History of intra abdominal surgery, small or large intestine resection or small bowel obstruction

- Terminology:
  - All descendants of "[C0198482] Operation on abdominal region" in UMLS
  - Added small Bowel obstruction, stomach bypass, hysterectomy, cholecystectomy, etc.
  - Excluded hernia (19 CUIs)
  - Discarded "apr", "turp", "tips", "sch"

  $\rightarrow$ 17,703 terms in total
- Rules:
  - Added rules "*ectomy " and "surgery" in abdomen-related section
  - Discarded negated events.

Terminology-based methods: examples

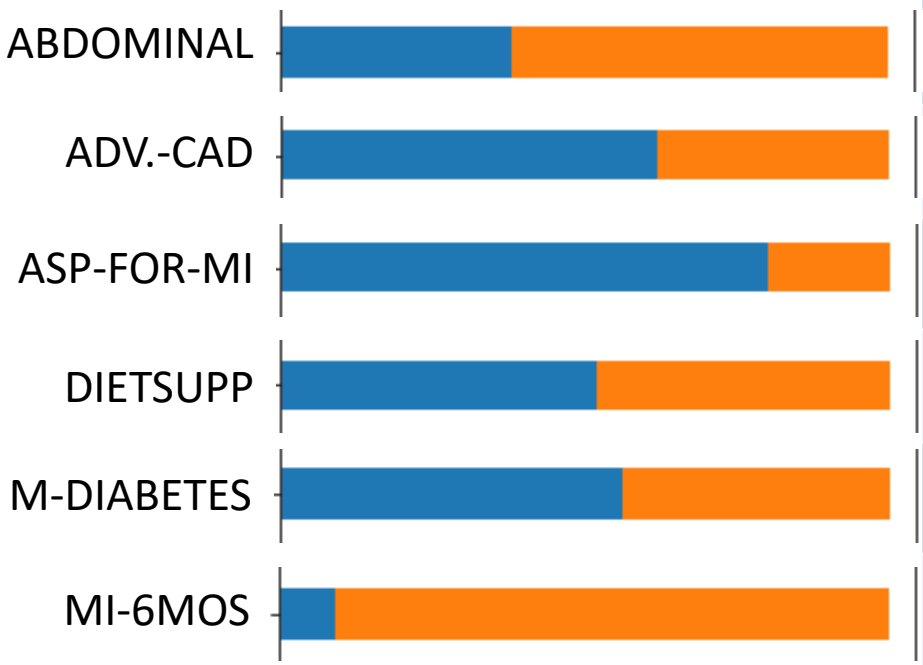
DIETSUPP

Taken a dietary supplement (excluding Vitamin D) in the past 2 months

- List of 183 ingredients listed as dietary supplements in different resources
- Added the temporal constraint
- Looked only in the MEDICATION sections
  (e.g. *"calcium"* can be a lab result or a diet supplement)

- (Tried to use the huge Dietary Supplement Label Database (DSLD), but too noisy)

Terminology-based methods: results



	met			not met		
	Prec.	Rec.	F1	Prec.	Rec.	F1
ABDOMINAL	0.85	0.93	0.89	0.96	0.91	0.94
ADV.-CAD	0.74	0.87	0.80	0.82	0.66	0.73
ASP-FOR-MI	0.83	0.91	0.87	0.45	0.28	0.34
DIETSUPP	0.91	0.89	0.90	0.88	0.90	0.89
M-DIABETES	0.90	0.86	0.88	0.87	0.91	0.89
MI-6MOS	0.67	0.50	0.57	0.95	0.97	0.96

Methods & Results

4. Overall results

1. Semi-supervised methods
2. Rule-based methods
3. Terminology-based methods
4. Overall results

Top 10 teams (best run only)

"No significant difference" cluster

Rank	Team	Micro F1
1	MedUniGraz	0.91
2	University of Michigan	0.9075
3	Sorbonne Université	0.9069
4	Med Data Quest	0.9028
5	Cincinnati Children's Hospital Medical Center	0.9026
6	Arizona State University	0.9003
7	University of New South Wales / National Cancer Institute	0.8913
	Harbin Institute of Technology	0.8855
		0.8837
	W	0.8765

What did we learn?

(spoiler: no big scoop)

Conclusion

- Difficult to:
  - apply semi-supervised methods for all criteria
  - apply any single method for all criteria
  - do anything without medical expertise

- We need:
  - ~~More training data~~ (we won't have it, or only for a small number of phenotypes)
  - More general approaches

Conclusion

- Improve learning methods:
  - Active learning
  - Knowledge transfer
  - Domain adaptation
  - Distant supervision
  - Patient representation
  - …

- But also secure the information:
  - Better concept identification and normalization (n2c2 2019 ;-))
  - Joint use of structured and textual data in EHRs
  - More formal definition of phenotypes

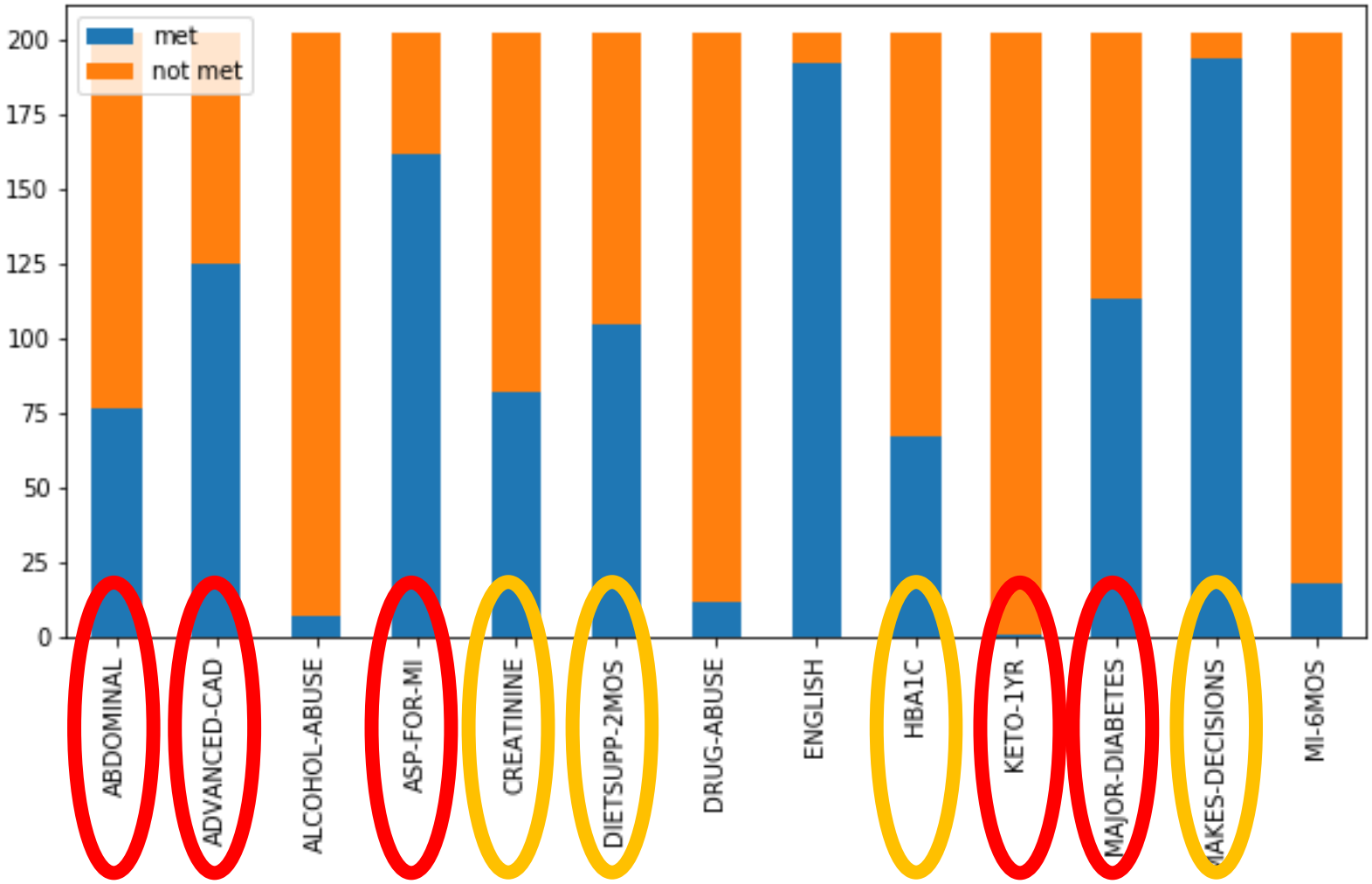
Thank you!

And thanks to

Hugo Cisneros
Christel Daniel
Tom Denat
Matthieu Doutreligne
Catherine Duclos
Nicolas Griffon
Claire Hassen-Khodja
Ivan Lerner
Guillaume Luquiens
Nicolas Paris
Adrien Parrot
Éric Sadou
Cyril Saussol
Xavier Tannier
Pascal Vaillant

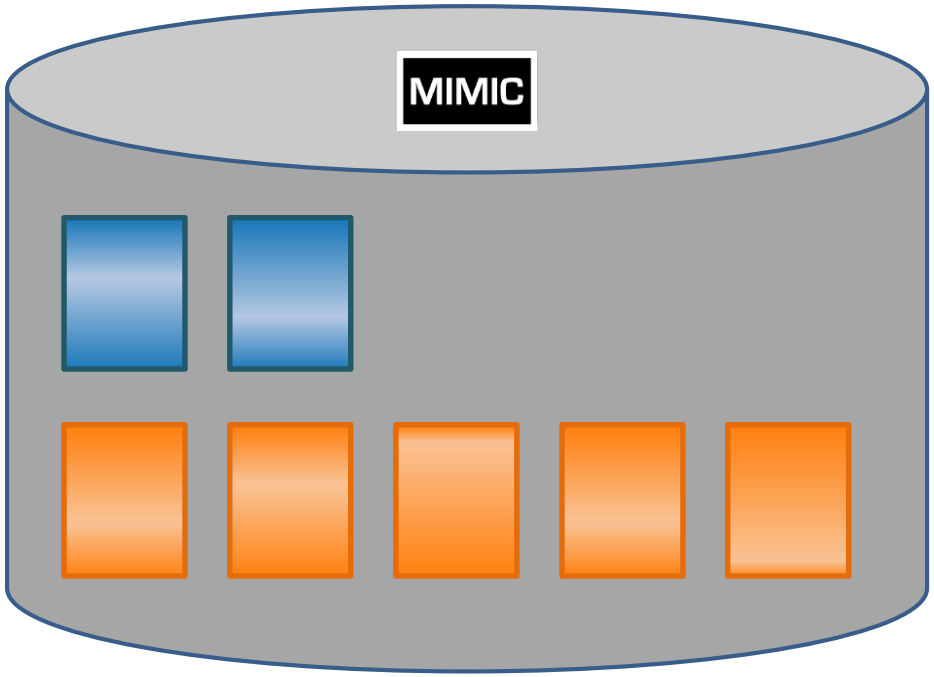
n2c2 dataset

Some classes require good medical knowledge… but not all



Building a silver standard

3. (Remove (or not) the matched sentences from the reports)
   (otherwise we'll learn to recognize the seed rules)
4. Count on the redundancy of the reports (several files per patient) to contain variants of the relevant information



CREATININE

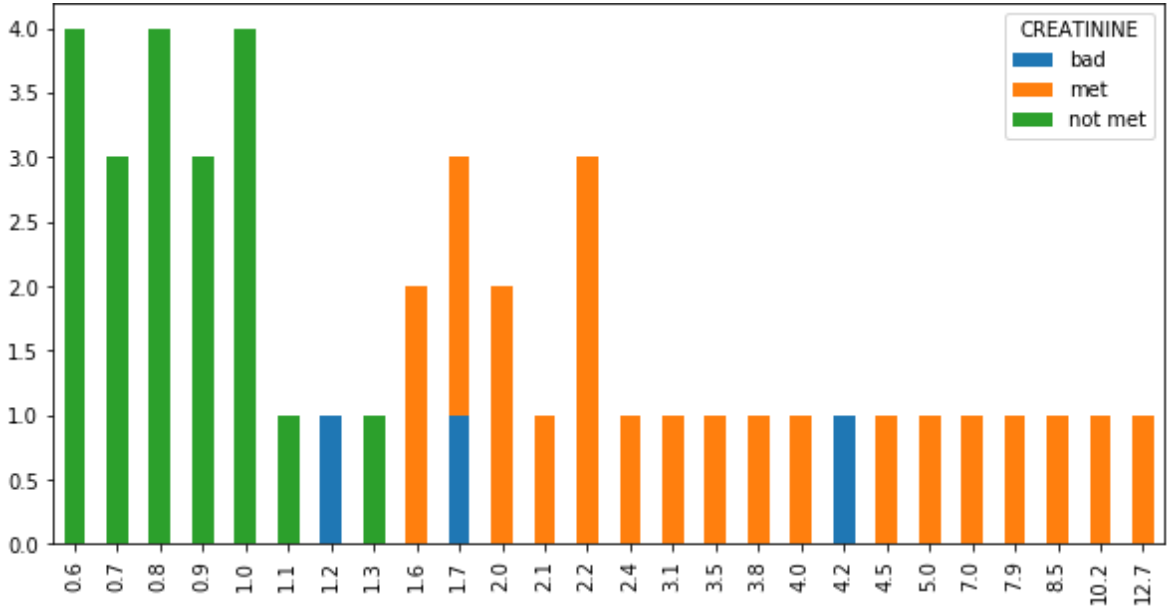
CREATININE  *Serum creatinine > upper limit of normal*

- Knowledge says:
  "upper limit" of normal is 1.0 for female, 1.2 for male patients.

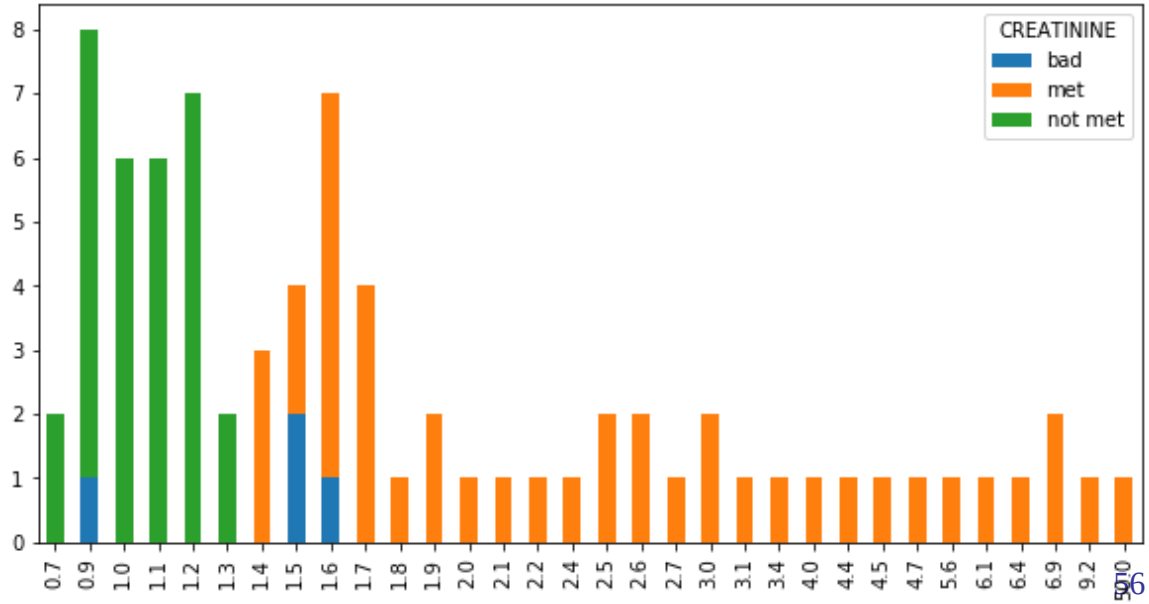
CREATINE

- Data says:

Females only



Males only

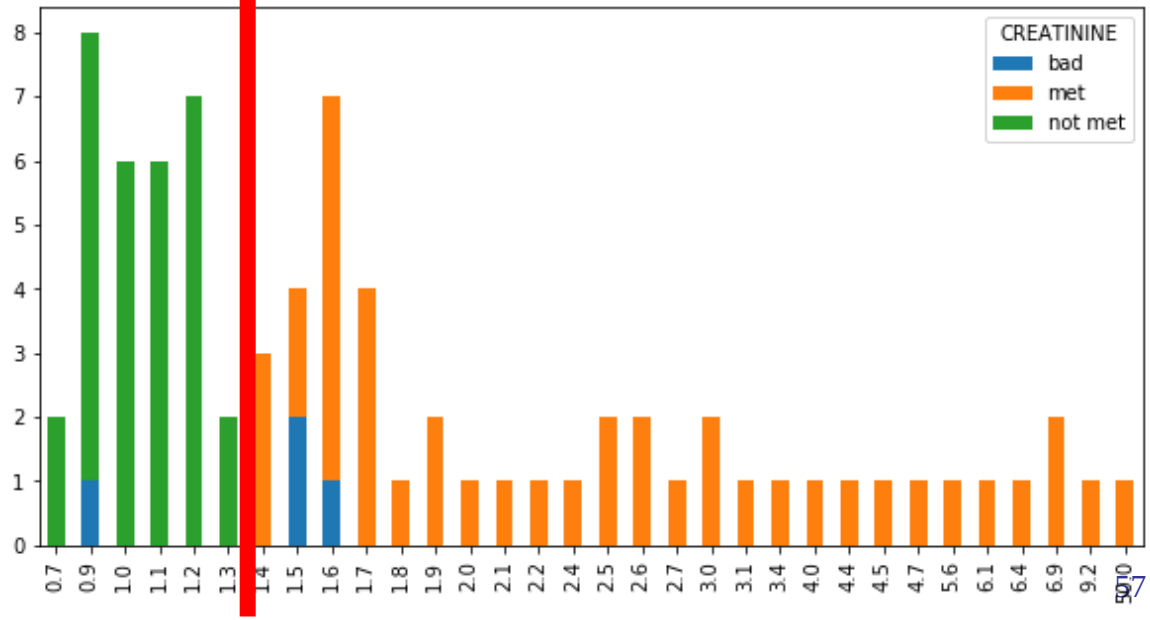


CREATININE

- Train data says:

Females only

Males only

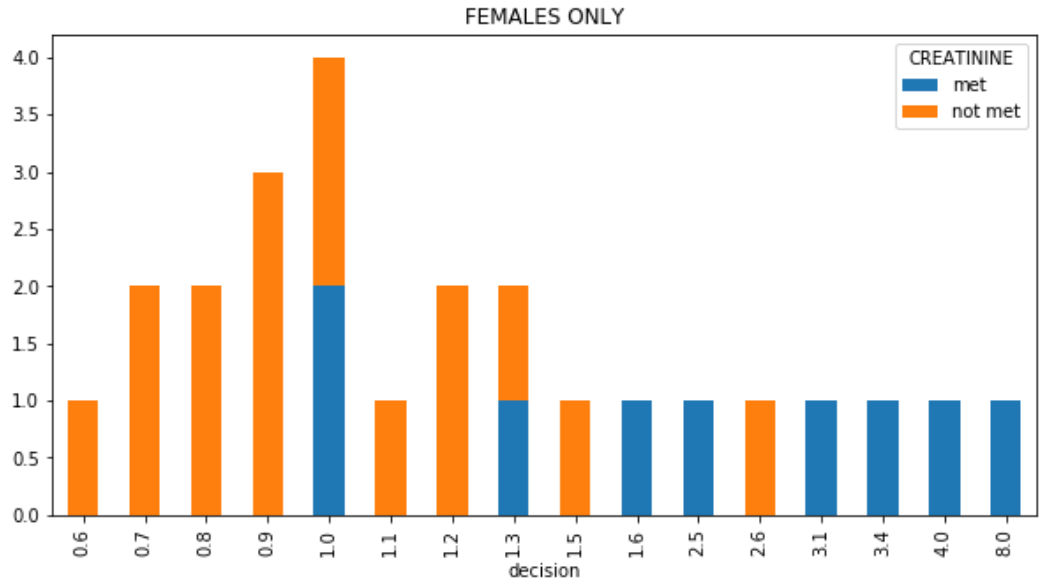


CREATININE

- Gold test data says:

  (numbers from our system, decision from the gold)

Females only



Males only