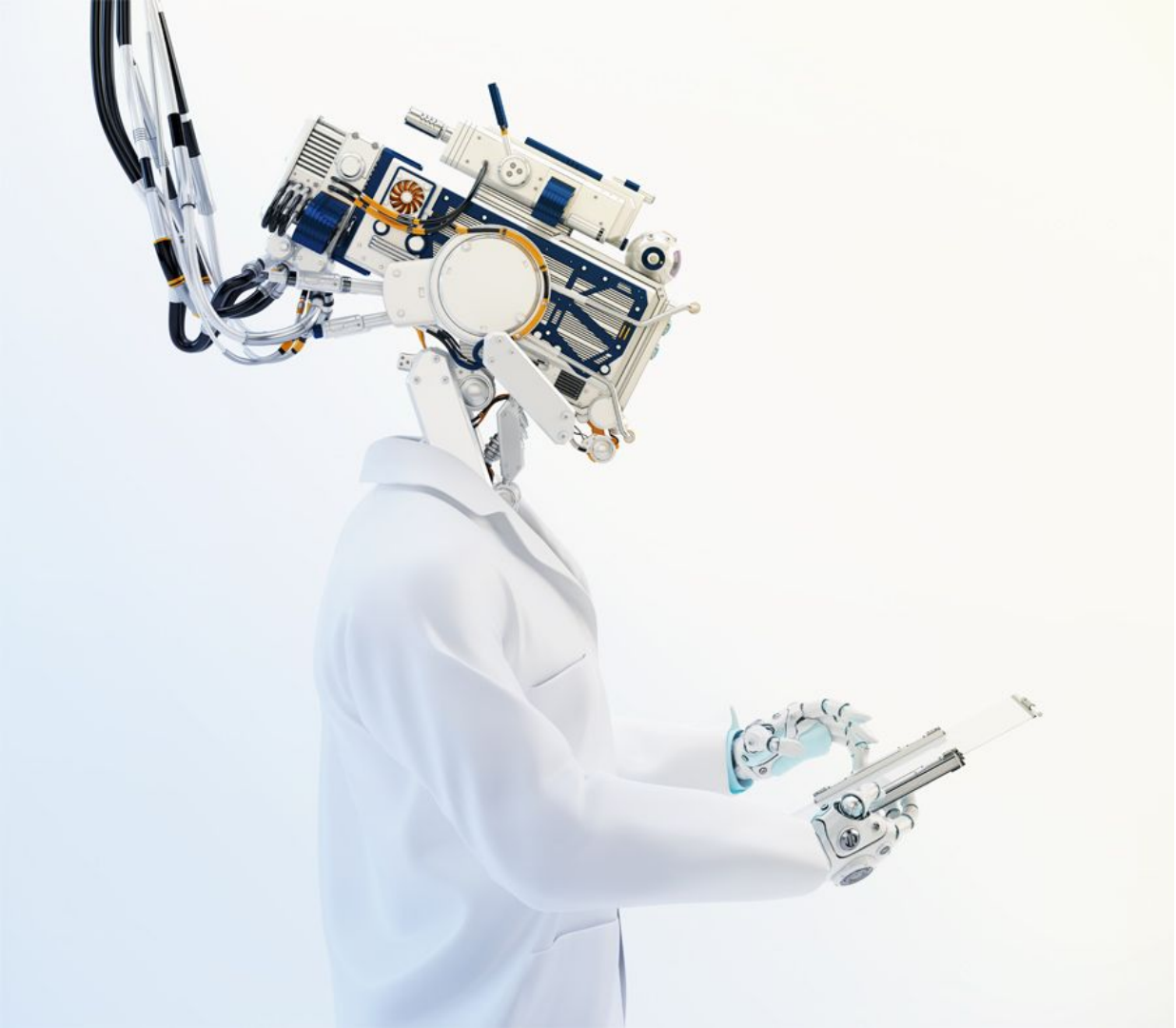


Medical XAI

Presentation Topics

1. Purpose of AI in medicine
2. Why is XAI an Important Issue?
3. What is explainability in the medical domain?
4. Examples of new, explainable algorithms in the medical domain

Purpose of AI in medicine

1. Diagnosis
2. Prognosis
3. Treatment
4. And More!

Purpose of AI in medicine

di·ag·no·sis
/ˌdīəgˈnōsəs/

noun

1. the identification of the nature of an illness or other problem by examination of the symptoms.
   "early diagnosis and treatment are essential"
   synonyms: identification, recognition, discovery, detection, pinpointing, reading, determination; More

Diagnosis

- Image processing (ex: analysing biopsy to determine specific pathology)
- -omics (ex: transcriptomics/ genomics)

prog·no·sis
/prägˈnōsəs/

noun

the likely course of a disease or ailment.
"the disease has a poor prognosis"

Prognosis

- Survival Prediction
- -omics (transcriptomics/genomics/etc.)

Treatment

- Predicting Patient Response to Drug
- Intelligent Prosthetics
- Robotic Surgery
- Assisted living


GAME TIME
PICK YOUR TEAMS

A



B



Which one is the spleen?

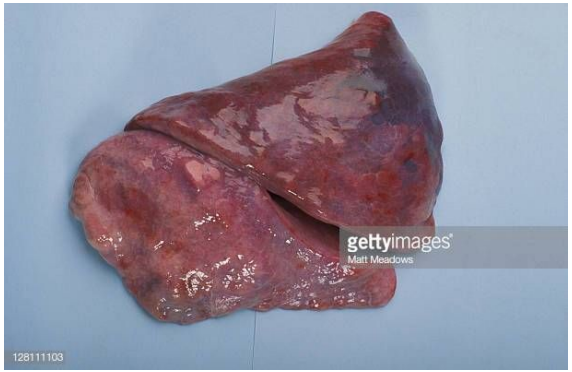
C



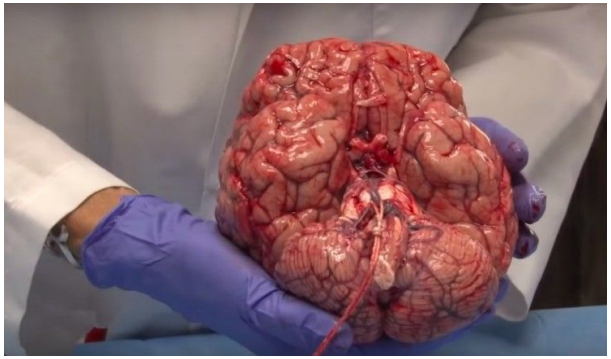
D



E



A



B



Which brain is the most intelligent?

C



D



E



Knees - Which one is in the best working condition?

A



B



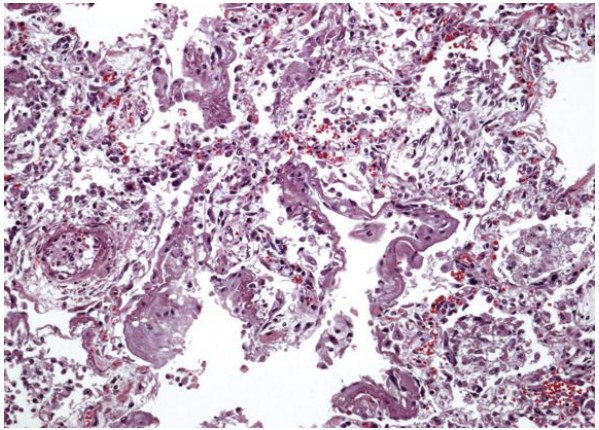
C



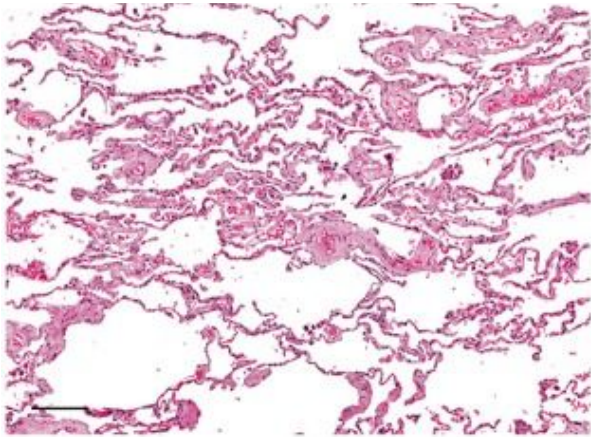
D



A



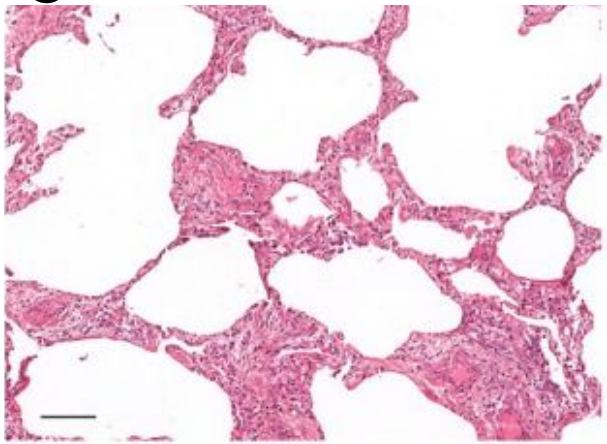
B



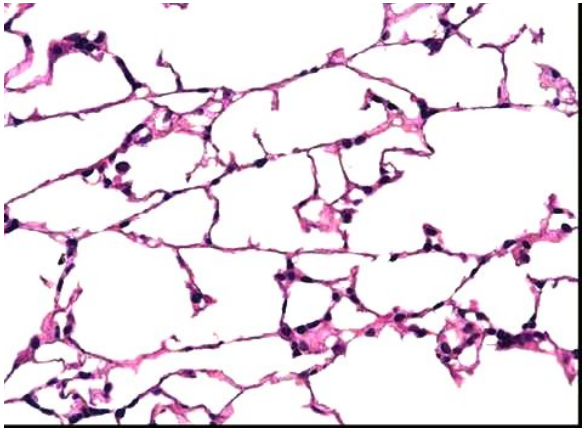
Lungs

Which one is the healthiest?

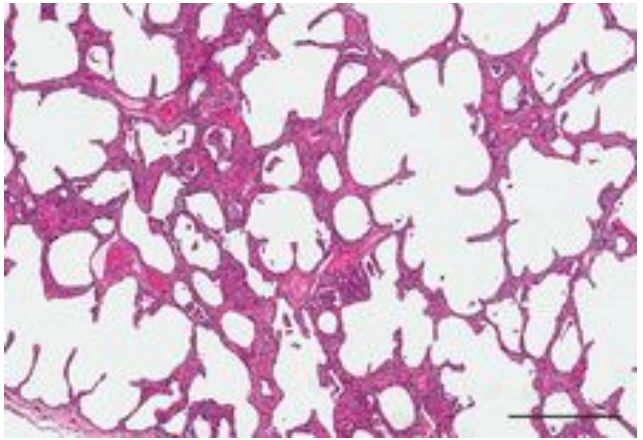
C



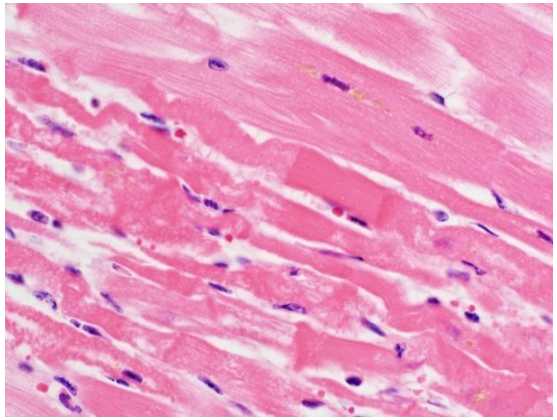
D



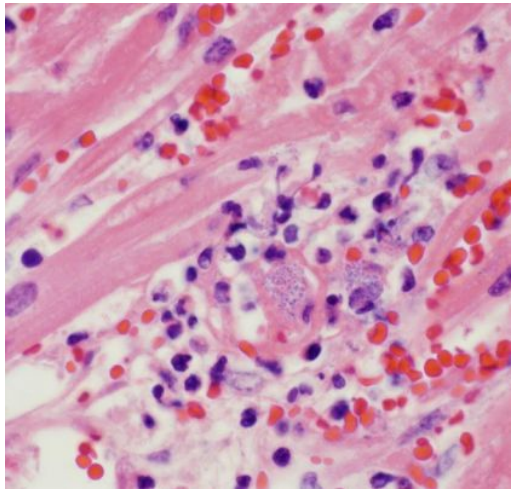
E



A



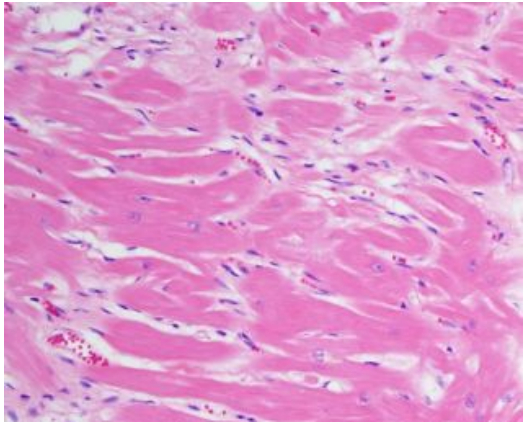
B



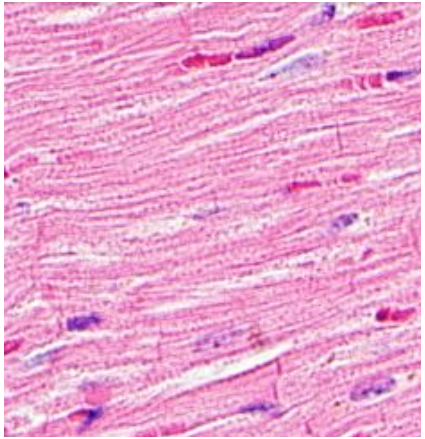
Heart myocardium

Which one is the healthiest?

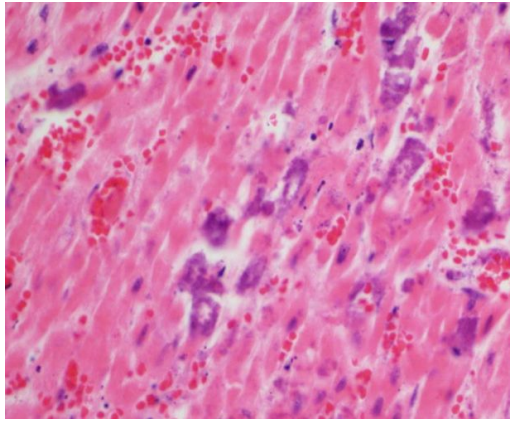
C



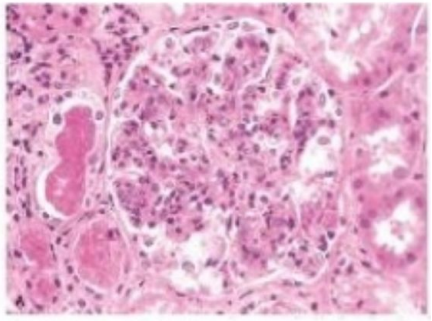
D



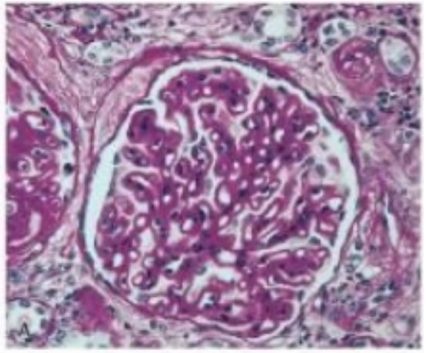
E



A



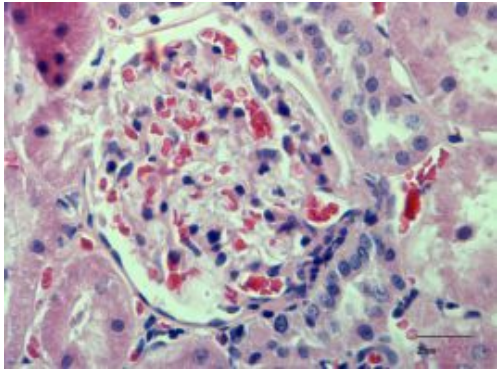
B



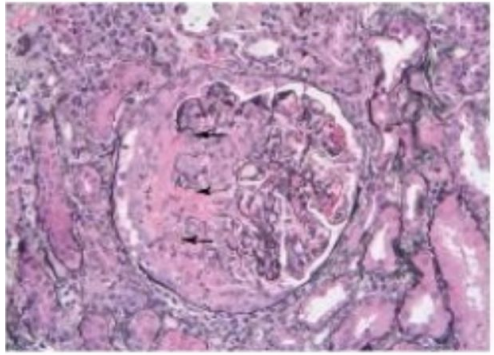
Glomerulus of the Nephron (Kidney)

Which one is healthy?

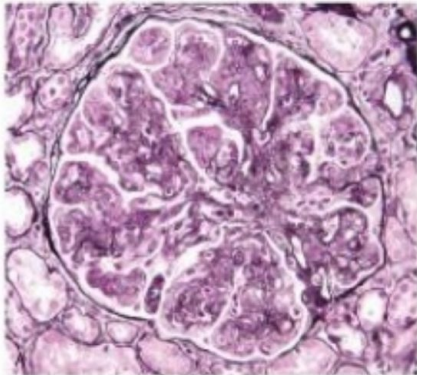
C



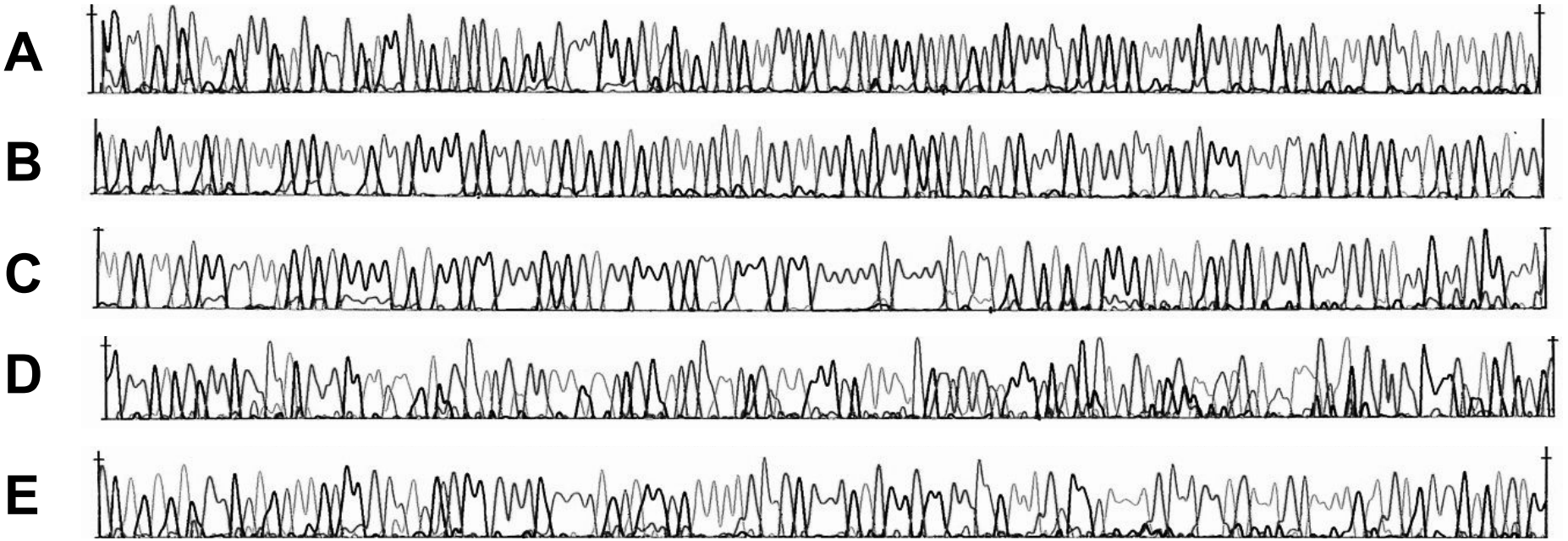
D



E

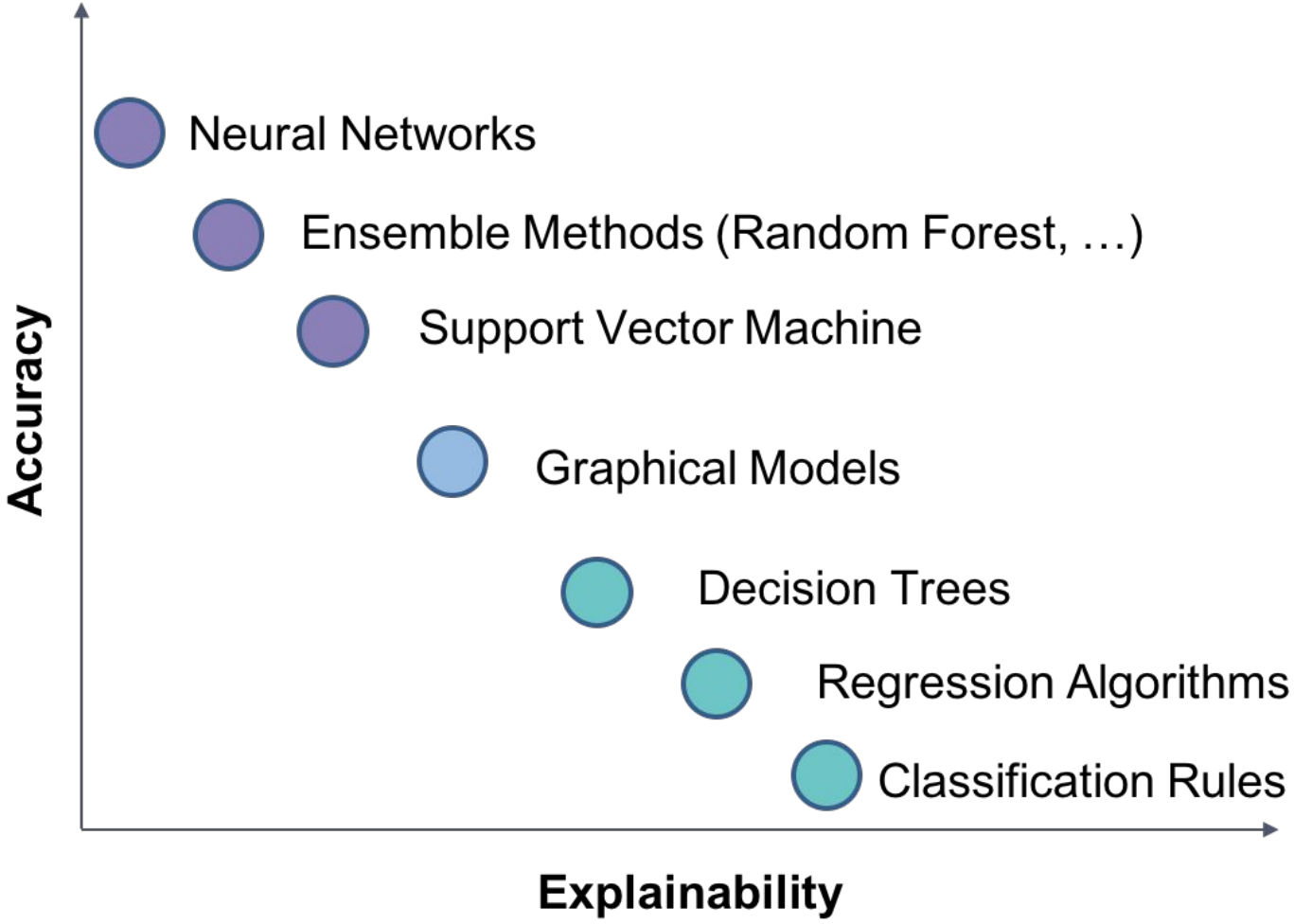


Genomic Data -- Who will experience the most severe side effects when treated with 5-FU?



Given this 30000-dimensional dataset, which one of these lung cancer patients will live the longest?

clinical_id	A1BG|1_LOW	A1BG|1_HIGH	!ML1|144568_LO	ML1|144568_HI	A2M|2_LOW	A2M|2_HIGH	GALT|53947_LO	GALT|53947_HI	4GNT|51146_LO	4GNT|51146_HIG	...	gender	years_to_birth	cancer_type	center	itial_pathologic	days_to_event	ethnicity	race	adiation_therapy	vital_status
TCGA-05-4244	0	0	5	0	0	4	0	0	5	0	...	male	70	Lung adenocarci	UNC	2009	0	NaN	NaN	NaN	
TCGA-05-4249	0	0	3	0	0	5	0	0	5	0	...	male	67	Lung adenocarci	UNC	2007	1523	NaN	NaN	no	
TCGA-05-4250	0	0	5	0	0	4	0	0	5	0	...	female	79	Lung adenocarci	UNC	2007	121	NaN	NaN	NaN	
TCGA-05-4382	0	0	3	0	0	5	0	0	4	0	...	male	68	Lung adenocarci	UNC	2009	607	NaN	NaN	yes	
TCGA-05-4384	0	0	5	0	0	6	0	0	4	0	...	male	66	Lung adenocarci	UNC	2009	426	NaN	NaN	yes	
TCGA-05-4389	0	0	0	0	0	3	0	0	4	0	...	male	70	Lung adenocarci	UNC	2005	1369	NaN	NaN	no	
TCGA-05-4390	0	0	4	0	0	4	0	0	5	0	...	female	58	Lung adenocarci	UNC	2005	1126	NaN	NaN	no	
TCGA-05-4395	0	0	0	0	0	2	0	0	6	0	...	male	76	Lung adenocarci	UNC	2006	0	NaN	NaN	NaN	
TCGA-05-4396	0	0	4	0	0	5	0	0	5	0	...	male	76	Lung adenocarci	UNC	2006	303	NaN	NaN	NaN	
TCGA-05-4397	0	0	2	0	0	2	0	0	3	0	...	male	65	Lung adenocarci	UNC	2006	731	NaN	NaN	NaN	
TCGA-05-4398	0	0	4	0	0	4	0	0	6	0	...	female	47	Lung adenocarci	UNC	2006	1431	NaN	NaN	no	
TCGA-05-4402	0	0	0	0	0	5	0	0	5	0	...	female	57	Lung adenocarci	UNC	2007	244	NaN	NaN	no	
TCGA-05-4403	0	0	5	0	0	5	0	0	5	0	...	male	76	Lung adenocarci	UNC	2006	578	NaN	NaN	NaN	
TCGA-05-4405	0	0	5	0	0	6	0	0	4	0	...	female	74	Lung adenocarci	UNC	2006	610	NaN	NaN	no	
TCGA-05-4410	0	0	3	0	0	5	0	0	5	0	...	male	62	Lung adenocarci	UNC	2007	0	NaN	NaN	NaN	
TCGA-05-4415	0	0	4	0	0	2	0	0	5	0	...	male	57	Lung adenocarci	UNC	2008	91	NaN	NaN	no	
TCGA-05-4417	0	0	3	0	0	6	0	0	5	0	...	female	51	Lung adenocarci	UNC	2008	455	NaN	NaN	no	
TCGA-05-4418	0	0	0	0	0	3	0	0	6	0	...	male	69	Lung adenocarci	UNC	2008	274	NaN	NaN	NaN	
TCGA-05-4420	0	2	3	0	0	4	0	0	5	0	...	male	41	Lung adenocarci	UNC	2008	912	NaN	NaN	no	
TCGA-05-4422	0	0	2	0	0	4	0	0	4	0	...	male	68	Lung adenocarci	UNC	2008	365	NaN	NaN	NaN	
TCGA-05-4424	0	0	4	0	0	6	0	0	5	0	...	male	70	Lung adenocarci	UNC	2008	913	NaN	NaN	yes	
TCGA-05-4425	0	0	4	0	0	3	0	0	5	0	...	female	70	Lung adenocarci	UNC	2008	669	NaN	NaN	NaN	


Neural Networks
Ensemble Methods (Random Forest, …)
Support Vector Machine
Graphical Models
Decision Trees
Regression Algorithms
Classification Rules
Accuracy
Explainability

Why make AI explainable if it has already proven to be highly

1. Trust
   a. Docto
      decisi
   b. Patier
      made
   c. Health
      patien
2. Model Im
   a. Explai
      decisi
3. Legality
   a. If a pa
      can w



4. Research and Discovery
   a. May help guide research and accelerate discovery
   b. Just curious! What's going on?

What is Explainability?

Thing we need to get straight:
In what context are we defining explainability?

- Most papers in the Medical XAI domain fail to give a concrete definition of “explainable”
- But some papers are dedicated entirely to defining explainability in a…
  - Philosophical context (explaining → understanding → what does it mean for a person to understand something?)
  - Legal context
  - Model improvement context
  - Educational context
  - Etc.

What do we need to build explainable AI systems for the medical domain?

Andreas Holzinger[1] A.HOLZINGER@HCI-KDD.ORG
Chris Biemann[2] BIEMANN@INFORMATIK.UNI-HAMBURG.DE
Constantinos S. Pattichis[3] PATTICHI@CS.UCY.AC.CY
Douglas B. Kell[4] DBK@MANCHESTER.AC.UK

[1] *Holzinger Group HCI-KDD, Inst. for Med. Informatics/Statistics, Medical University Graz, Austria*
[2] *Language Technology Group LT, Department of Informatics, University of Hamburg, Germany*
[3] *E-Health Laboratory, Department of Computer Science, University of Cyprus, Cyprus*
[4] *School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, UK*

Post-hoc vs. Ante-hoc

"After the fact"

Explanation for a **specific decision** and making it reproducible on demand

Ex: **Model Agnostic systems**

- Local Interpretable Model-Agnostic Explanations (LIME))
- Black Box Explanations through Transparent Approximations (BETA)

-- see Zehra's presentation on sep 6



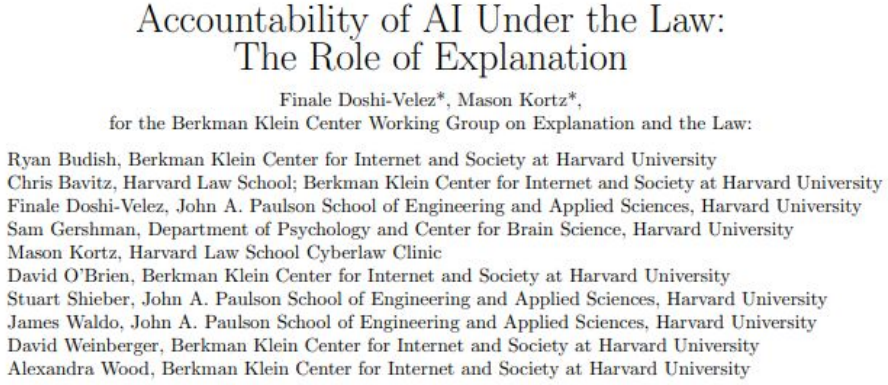
Glass-box approach

Ex:

- Linear regression
- Decision trees
- Fuzzy inference systems
- High performance generalized additive models with pairwise interactions (GAMs) (Guillaume,2001)
- Combo of naive Bayes, linear support vector machines, logistic regression (Poulin et al, 2006)(Szafrom et al, 2004)

Concrete definition of explainability in the context of law:

XAI is **not** a description of the decision-making process, but rather a **reason/justification for a particular outcome**.



Accountability of AI Under the Law:
The Role of Explanation

Finale Doshi-Velez*, Mason Kortz*,
for the Berkman Klein Center Working Group on Explanation and the Law:

Ryan Budish, Berkman Klein Center for Internet and Society at Harvard University
Chris Bavitz, Harvard Law School; Berkman Klein Center for Internet and Society at Harvard University
Finale Doshi-Velez, John A. Paulson School of Engineering and Applied Sciences, Harvard University
Sam Gershman, Department of Psychology and Center for Brain Science, Harvard University
Mason Kortz, Harvard Law School Cyberlaw Clinic
David O'Brien, Berkman Klein Center for Internet and Society at Harvard University
Stuart Shieber, John A. Paulson School of Engineering and Applied Sciences, Harvard University
James Waldo, John A. Paulson School of Engineering and Applied Sciences, Harvard University
David Weinberger, Berkman Klein Center for Internet and Society at Harvard University
Alexandra Wood, Berkman Klein Center for Internet and Society at Harvard University

Definition of Explanation:

- Must have a description of how a particular set of inputs reached a particular conclusion
- Must be **human interpretable**
  - Must be able to answer the following should the question arise:
    - **Which features** contributed to a decision (from most important to least)?
    - Would **changing a certain feature** have changed the decision? (see Negar's presentation)
    - Why did two **similar-looking** cases get **different** decisions or vice-versa?

(Defined based on Wachter & Malgieri and Comande papers)

Discussion about seeking explainability

Must consider:

- Someone has to ask for an explanation to begin with
- When do the benefits of an explanation outweigh the cost?
  - Decision must have been acted on in a way that has a impact on a person other than a decision maker…. Basically all of medicine?
  - And more (value to knowing if a decision is made erroneously, reason to believe that error has (or will) occurred due to suspicious design

BUT! The guidelines described by Doshi-Velez et al. allow the medical system to use AI without having to worry about the cost of it explaining its decisions every single time. This means that XAI in medicine does **not** have to use ante-hoc explanations (only post-hoc).

Post-hoc vs. Ante-hoc

"After the fact"

Explanation for a specific decision and making it reproducible on demand

Ex: **Model Agnostic systems**

- Local Interpretable Model-Agnostic Explanations (LIME))
- Black Box Explanations through Transparent Approximations (BETA)

-- see Zehra's presentation on sep 6

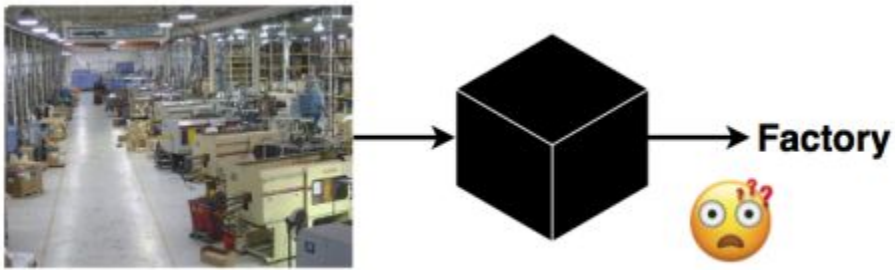


Glass-box approach

Ex:

- Linear regression
- Decision trees
- Fuzzy inference systems
- High performance generalized additive models with pairwise interactions (GAMs) (Guillaume,2001)
- Combo of naive Bayes, linear support vector machines, logistic regression (Poulin et al, 2006)(Szafrom et al, 2004)

What is Explainability?



What Does Explainable AI Really Mean? A New Conceptualization of Perspectives

Derek Doran[1], Sarah Schulz[2], and Tarek R. Besold[3]

[1] Dept. of Computer Science & Engineering, Kno.e.sis Research Center
Wright State University, Dayton, Ohio, USA
derek.doran@wright.edu
[2] Institute for Natural Language Processing
University of Stuttgart, Stuttgart, Germany
sarah.schulz@ims.uni-stuttgart.de
[3] Department of Computer Science
City, University of London, London, UK
tarek-r.besold@city.ac.uk

Abstract. We characterize three notions of explainable AI that cut across research fields: *opaque systems* that offer no insight into its algorithmic mechanisms; *interpretable systems* where users can mathematically analyze its algorithmic mechanisms; and *comprehensible systems* that emit symbols enabling user-driven explanations of how a conclusion is reached. The paper is motivated by a corpus analysis of NIPS, ACL, COGSCI, and ICCV/ECCV paper titles showing differences in how work on explainable AI is positioned in various fields. We close by introducing a fourth notion: truly *explainable systems*, where automated reasoning is central to output crafted explanations without requiring human post processing as final step of the generative process.

High dimensional data visualizations like t-SNE and receptive field visualization on convolutional neural networks

. related to Holzinger et al's post-hoc system

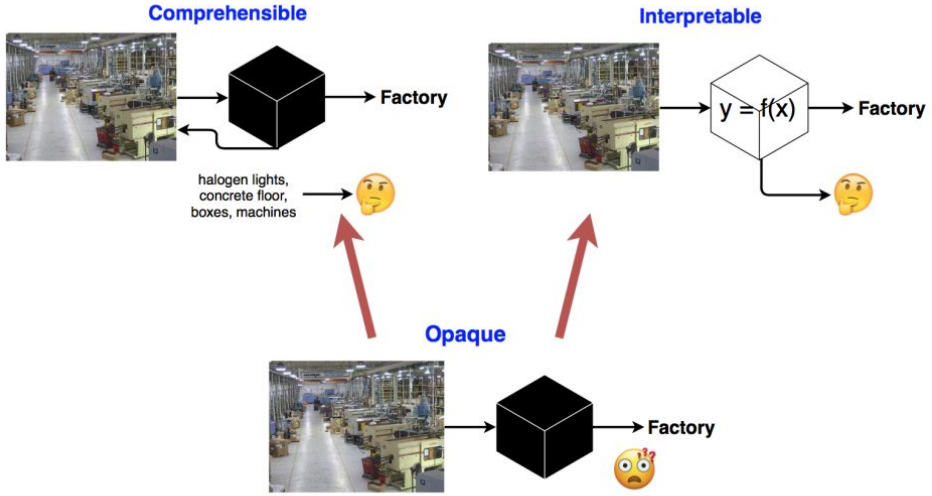


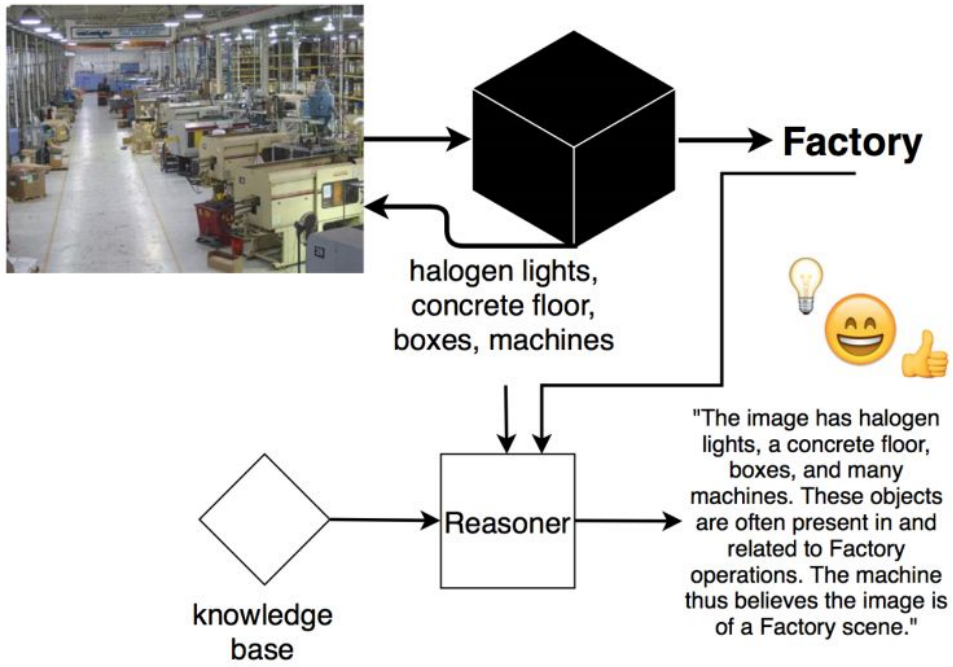
Fig. 3: Relation between opaque, comprehensible, and interpretable AI.

. Regression models, support vector machines, decision trees, ANOVAs, and data clustering

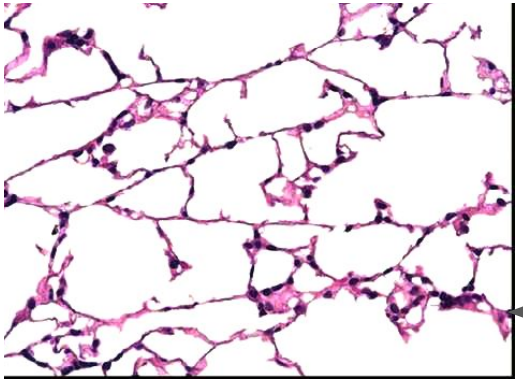
. related to Holzinger et al's ante-hoc systems

Interpretive and comprehensible models thus enable explanations of decisions, but **do not yield explanations themselves**.

Truly Explainable AI Should Integrate Reasoning



“both approaches are lacking in their ability to formulate, for the user, a line of reasoning that explains the decision making process of a model using human-understandable features of the input data. **Leaving explanation generation to human analysts can be dangerous** since, depending on their background knowledge about the data and its domain, **different explanations about why a model makes a decision may be deduced”**



What if we used this in medicine?



COPD

Shapes formed by tissue, high ratio of white pixels to color pixels, absence of large dotted purple clumps

Knowledge Base

Reasoner

The image has a high ratio of white pixels to color pixels compared to healthy samples; this suggests that the pockets of air associated with the alveolar tissue are larger than in normal tissue. Most shapes formed by the tissue portion have sharp edges which suggests that the tissue is under tensile stress. These features are characteristic of COPD; a disease of lung hyperinflation and alveolar tissue destruction. The absence of large dotted purple clumps suggests that there is no sign of a bacterial infection in this tissue sample.

Explainable(?) Algorithms Applied to the Medical Field

Paper #1

Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach

Jean-Baptiste Lamy et al.

LIMICS, Université Paris

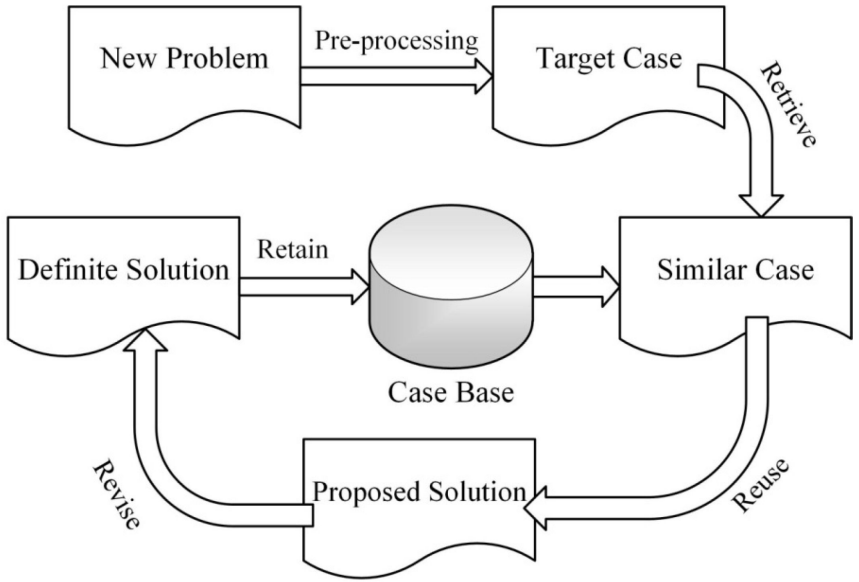
January 7th, 2019

What is case-based reasoning?

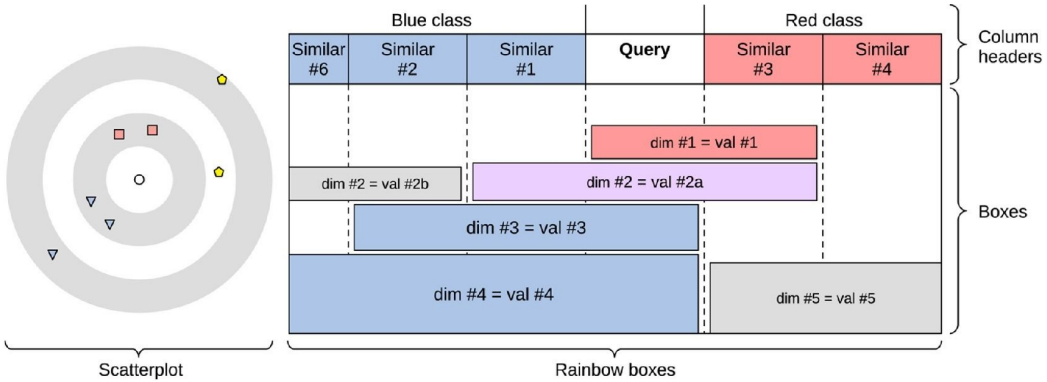
- Came from the field of cognitive science, now considered a branch of AI

Example: Consider a patient with a set of attributes (demographic information, medical test results, etc.) and an unknown diagnosis.

1. Look into case database and retrieve other patient records with similar attributes
2. Use “majority rules” tactic to predict the diagnosis of the patient
   i.e. The diagnosis of the most similar cases is applied to the new case
3. Verify the correctness of the prediction; if incorrect, then revise
4. Add that new patient and correct diagnosis to the case database



- A way to visualize the similarities between the query and its most similar neighbours
  - Uses set intersection visualization

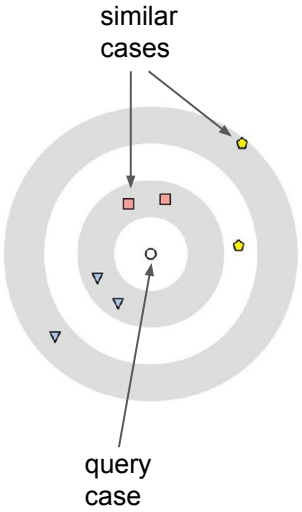


- RBIA (Rainbow box inspired algorithm) for classification
  - Inspired by how a physician might interpret the visualization

How did the they generate the scatterplot?

2-dimensional approach to Multidimensional Scaling:

- Generate distance matrix for all cases in the case database and the query case (this paper used euclidean distance)
- The query case, q, is placed at the origin (0,0)
- Polar coordinates (*distance, angle*) were used instead of cartesian coordinates (*x, y*)
- The *distance* from the origin is already known from the distance matrix
- The *angle* for each polar coordinate was generated for each of the other cases by finding angles that minimized the stress function overall



Euclidean Distance

$$d(p,q) = d(q,p)$$
$$= \sqrt{(q_1 - p_1)^2 + (q_3 - p_3)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Stress Function

$$S_p(d) = \sum_{2<i<j} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}}$$

where d_{ij} is the distance between point *i* and *j* in the *n*-dimensional distance matrix, and δ_{ij} is the distance between point *i* and *j*

$$\delta_{ij} = \sqrt{(d_{1,i}\cos(\theta_i) - d_{1,j}\cos(\theta_j))^2 + (d_{1,i}\sin(\theta_i) - d_{1,j}\sin(\theta_j))^2}$$

Takeaway: Scatterplot was generated as to preserve relative distance between the query case and the other cases. Distance between other cases were not preserved, but placement of these cases was determined by minimizing stress function.

How did they generate the “rainbow boxes”?

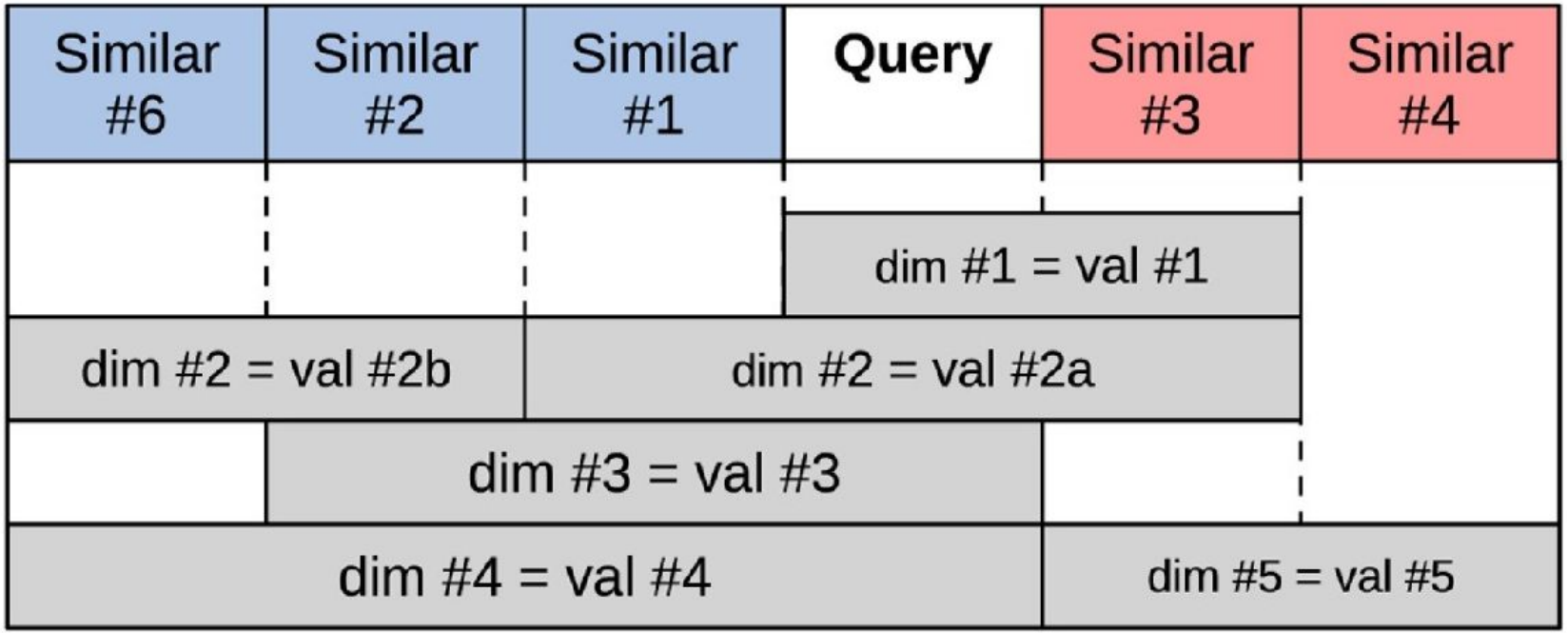
1)	Query	Similar #1	Similar #2	Similar #3	Similar #4	Similar #5	Similar #6	Similar #7
dim #1	**val #1**	val #1b	val #1c	**val #1**	val #1d	val #1e	val #1f	val #1g
dim #2	**val #2a**	**val #2a**	**val #2b**	**val #2a**	val #2d	val #2e	**val #2b**	val #2g
dim #3	**val #3**	**val #3**	**val #3**	val #3c	val #3d	val #3e	val #3f	val #3g
dim #4	**val #4**	**val #4**	**val #4**	val #4c	val #4d	val #4e	**val #4**	val #4g
dim #5	val #5a	val #5b	val #5c	**val #5**	**val #5**	val #5e	val #5f	val #5g
dim #6	val #6	val #6a	val #6b	val #6c	val #6d	val #6e	val #6f	val #6g
...								

2)	Similar #6	Similar #2	Similar #1	**Query**	Similar #3	Similar #4
dim #1	val #1f	val #1c	val #1b	**val #1**	**val #1**	val #1d
dim #2	**val #2b**	**val #2b**	**val #2a**	**val #2a**	**val #2a**	val #2d
dim #3	val #3f	**val #3**	**val #3**	**val #3**	val #3c	val #3d
dim #4	**val #4**	**val #4**	**val #4**	**val #4**	val #4c	val #4d
dim #5	val #5f	val #5c	val #5b	val #5a	**val #5**	**val #5**
dim #6	val #6f	val #6b	val #6a	val #6	val #6c	val #6d
...						

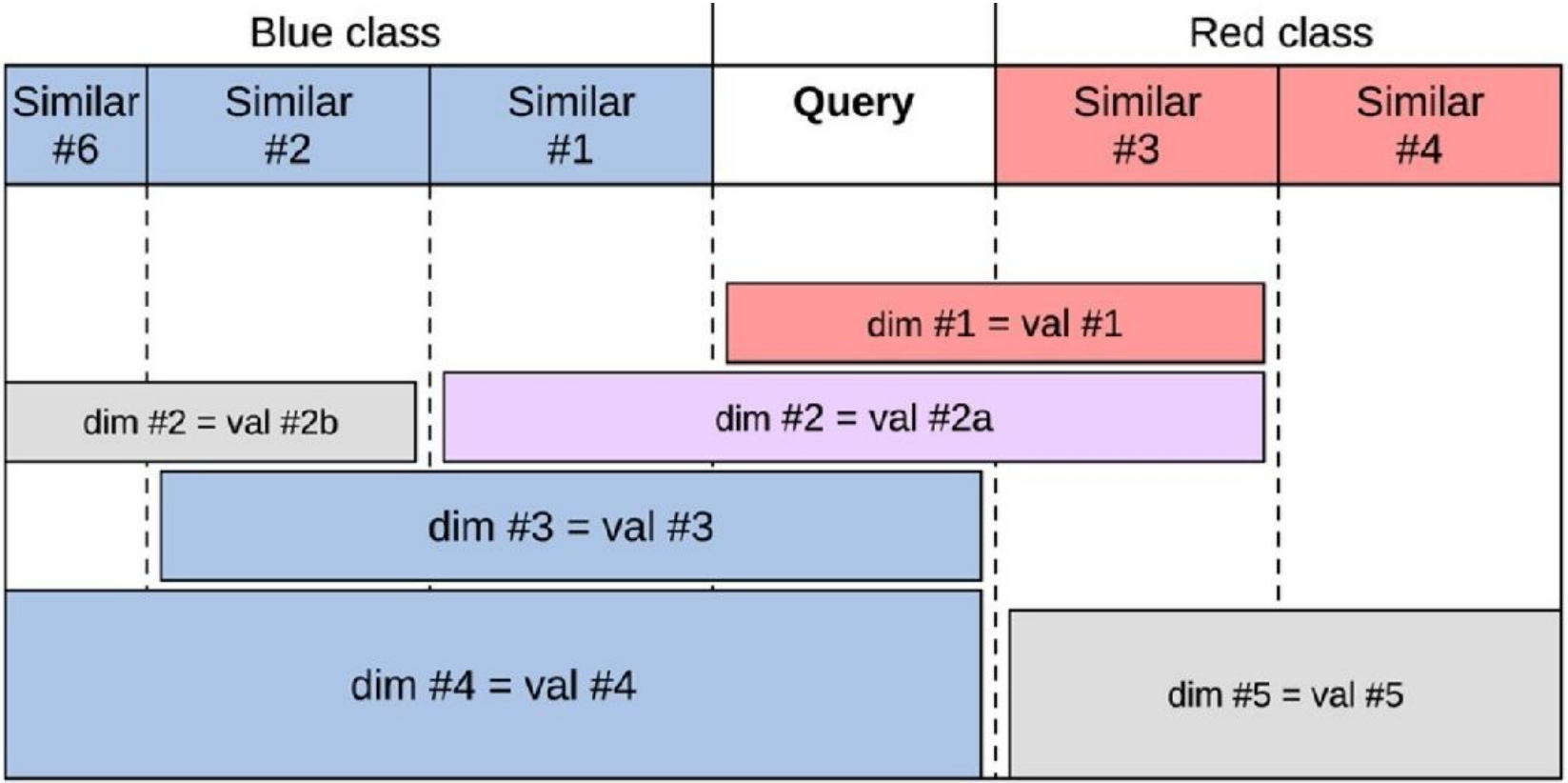
How did they generate the “rainbow boxes”?

3)	Similar #6	Similar #2	Similar #1	**Query**	Similar #3	Similar #4
dim #1	val #1f	val #1c	val #1b	dim #1 = val #1		val #1d
dim #2	dim #2 = val #2b		dim #2 = val #2a			val #2d
dim #3	val #3f	dim #3 = val #3			val #3c	val #3d
dim #4	dim #4 = val #4				val #4c	val #4d
dim #5	val #5f	val #5c	val #5b	val #5a	dim #5 = val #5	
dim #6	val #6f	val #6b	val #6a	val #6	val #6c	val #6d
...						

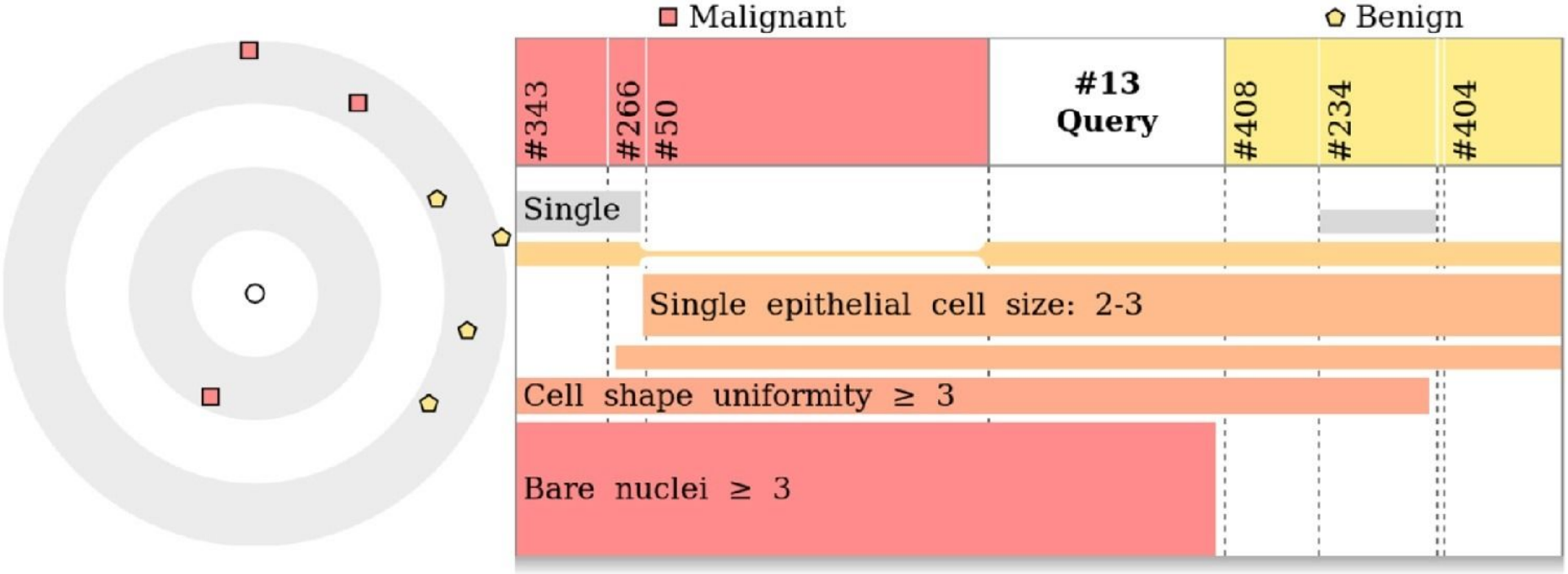
4)


Similar #6
Similar #2
Similar #1
Query
Similar #3
Similar #4
dim #1 = val #1
dim #2 = val #2b
dim #2 = val #2a
dim #3 = val #3
dim #4 = val #4
dim #5 = val #5

How did they generate the “rainbow boxes”?

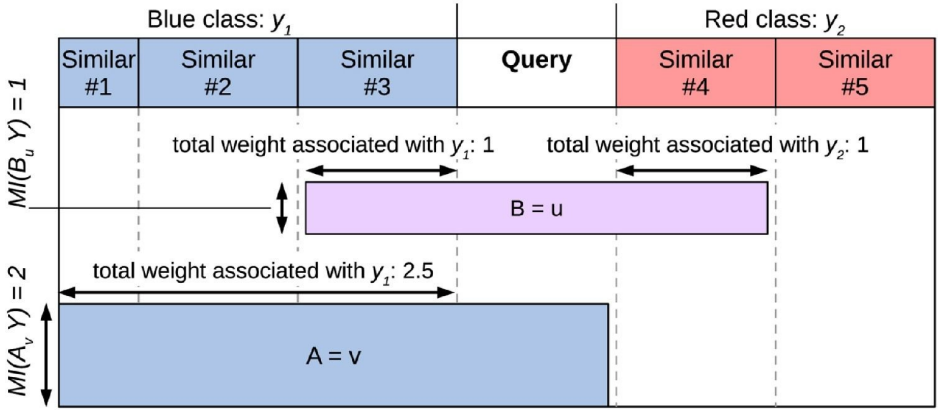


What did the authors bring to the table? (Visualization)



What should the query be classified as? Look for the dominant colour!

- The authors also made a classification algorithm based on the rainbow boxes
- It computes a score S_y for each of the two colors (representing the two retained classes y_1 and y_2)
  - For each box, S_y must take into account the
    - Box MI (*i.e.* box height)
    - Distribution of the cases in the box between the two classes (*i.e.* box color)
    - Weight of similar cases (*i.e.* column width)

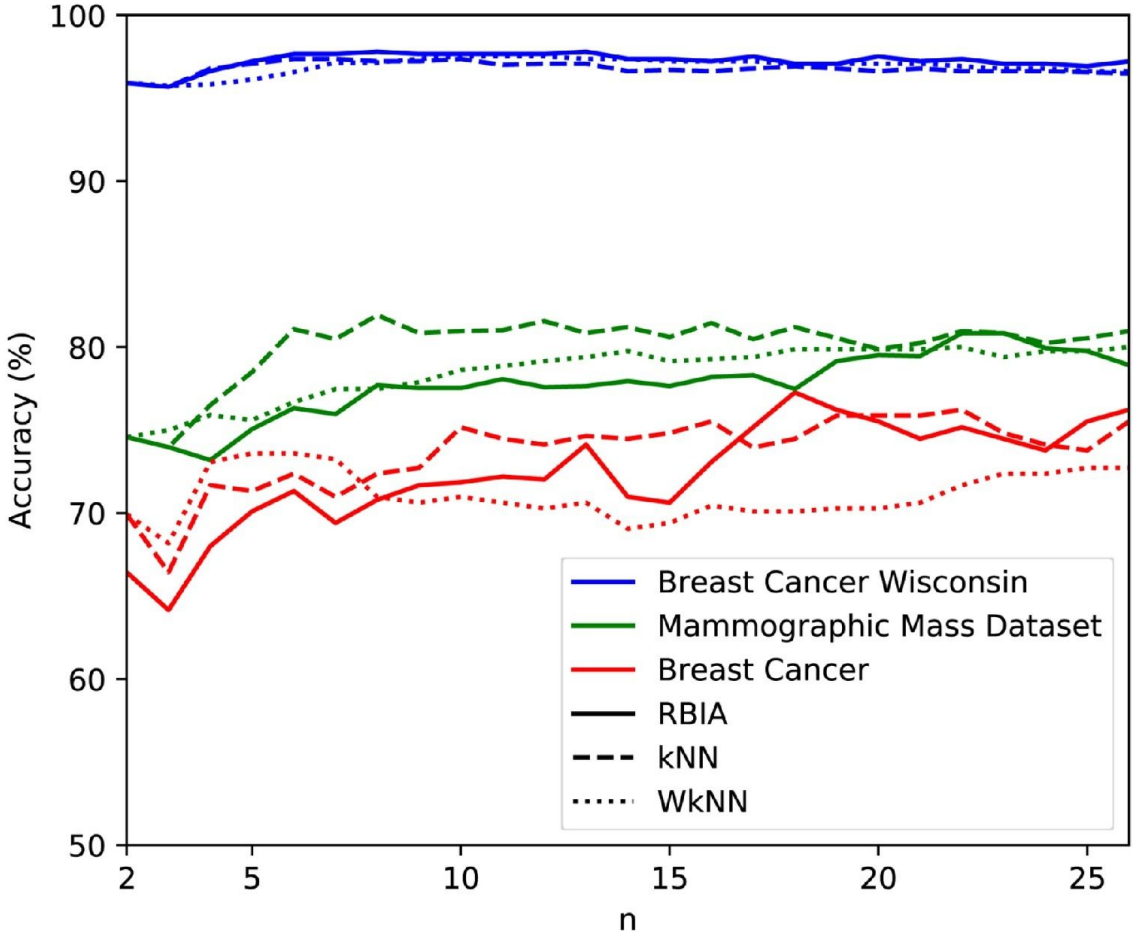


$$S_y = \sum_{Z_v}(\mathrm{MI}(Z_v Y) \times \sum\{w_{2 \le i \le n} \mid x_i \in Z_v \cap y\})$$

$$S_{y_1} = \mathrm{MI}(A_v Y) \times 2.5 + \mathrm{MI}(B_u Y) \times 1 = 6$$

$$S_{y_2} = \mathrm{MI}(A_v Y) \times 0 + \mathrm{MI}(B_u Y) \times 1 = 1$$

Results and Accuracies



Classified whether tumor was benign or malignant

Classified whether tumor was benign or malignant

Classified whether cancer was recurrent or not

Notice how they didn't compare against algorithms not based on kNN...

How does this tie into explainability?

- The authors contend that visualization should be considered as a form of making something more explainable

- Supposedly explanations based on similar cases are more convincing than explanations based on rules (goes back to “trust” as the reason we want XAI)

- Made a visualization tool that a doctor could use to come to their own conclusion, retrofitted a ML algorithm to it

Paper #2

Explainable Prediction of Medical Codes from Clinical Text

James Mullenbach et al.

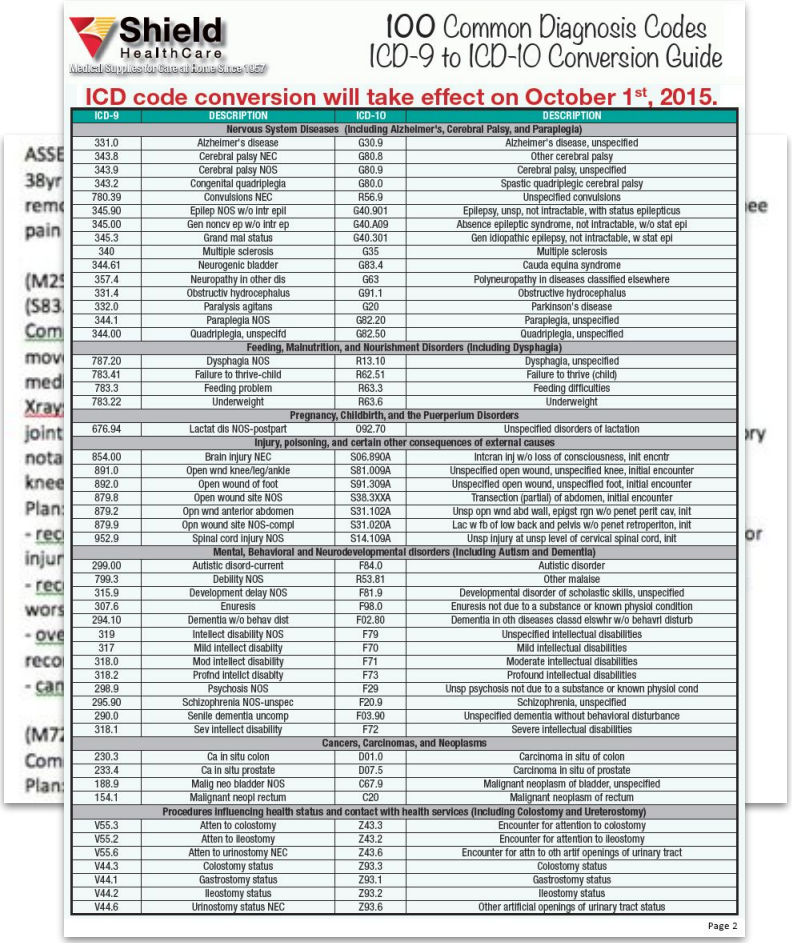
Georgia Institute of Technology

April 16th, 2018

- Clinical notes are free text narratives generated by physicians during patient encounters

- Authors wanted use these clinical notes to predict the ICD codes
  - International Classification of Diseases
  - Way of indicating diagnoses and procedures performed during the encounter

- Helpful because manual coding is time consuming and error-prone



Shield HealthCare

Medical Supplies for Care at Home Since 1957

100 Common Diagnosis Codes ICD-9 to ICD-10 Conversion Guide

ICD code conversion will take effect on October 1st, 2015.

ICD-9	DESCRIPTION	ICD-10	DESCRIPTION
Nervous System Diseases (including Alzheimer's, Cerebral Palsy, and Paraplegia)			
331.0	Alzheimer's disease	G30.9	Alzheimer's disease, unspecified
343.8	Cerebral palsy NEC	G80.8	Other cerebral palsy
343.9	Cerebral palsy NOS	G80.9	Cerebral palsy, unspecified
343.2	Congenital quadriplegia	G80.0	Spastic quadriplegic cerebral palsy
780.39	Convulsions NEC	R56.9	Unspecified convulsions
345.90	Epilep NOS w/o intr epil	G40.901	Epilepsy, unsp, not intractable, with status epilepticus
345.00	Gen noncv ep w/o intr ep	G40.A09	Absence epileptic syndrome, not intractable, w/o stat epi
345.3	Grand mal status	G40.301	Gen idiopathic epilepsy, not intractable, w stat epi
340	Multiple sclerosis	G35	Multiple sclerosis
344.61	Neurogenic bladder	G83.4	Cauda equina syndrome
357.4	Neuropathy in other dis	G63	Polyneuropathy in diseases classified elsewhere
331.4	Obstructiv hydrocephalus	G91.1	Obstructive hydrocephalus
332.0	Paralysis agitans	G20	Parkinson's disease
344.1	Paraplegia NOS	G82.20	Paraplegia, unspecified
344.00	Quadriplegia, unspecifd	G82.50	Quadriplegia, unspecified
Feeding, Malnutrition, and Nourishment Disorders (including Dysphagia)			
787.20	Dysphagia NOS	R13.10	Dysphagia, unspecified
783.41	Failure to thrive-child	R62.51	Failure to thrive (child)
783.3	Feeding problem	R63.3	Feeding difficulties
783.22	Underweight	R63.6	Underweight
Pregnancy, Childbirth, and the Puerperium Disorders			
676.94	Lactat dis NOS-postpart	O92.70	Unspecified disorders of lactation
Injury, poisoning, and certain other consequences of external causes			
854.00	Brain injury NEC	S06.890A	Intcran inj w/o loss of consciousness, init encntr
891.0	Open wnd knee/leg/ankle	S81.009A	Unspecified open wound, unspecified knee, initial encounter
892.0	Open wound of foot	S91.309A	Unspecified open wound, unspecified foot, initial encounter
879.8	Open wound site NOS	S38.3XXA	Transection (partial) of abdomen, initial encounter
879.2	Opn wnd anterior abdomen	S31.102A	Unsp opn wnd abd wall, epigst rgn w/o penet perit cav, init
879.9	Opn wound site NOS-compl	S31.020A	Lac w fb of low back and pelvis w/o penet retroperiton, init
952.9	Spinal cord injury NOS	S14.109A	Unsp injury at unsp level of cervical spinal cord, init
Mental, Behavioral and Neurodevelopmental disorders (including Autism and Dementia)			
299.00	Autistic disord-current	F84.0	Autistic disorder
799.3	Debility NOS	R53.81	Other malaise
315.9	Development delay NOS	F81.9	Developmental disorder of scholastic skills, unspecified
307.6	Enuresis	F98.0	Enuresis not due to a substance or known physiol condition
294.10	Dementia w/o behav dist	F02.80	Dementia in oth diseases classd elswhr w/o behavrl disturb
319	Intellect disability NOS	F79	Unspecified intellectual disabilities
317	Mild intellect disabilty	F70	Mild intellectual disabilities
318.0	Mod intellect disability	F71	Moderate intellectual disabilities
318.2	Profnd intellct disabilty	F73	Profound intellectual disabilities
298.9	Psychosis NOS	F29	Unsp psychosis not due to a substance or known physiol cond
295.90	Schizophrenia NOS-unspec	F20.9	Schizophrenia, unspecified
290.0	Senile dementia uncomp	F03.90	Unspecified dementia without behavioral disturbance
318.1	Sev intellect disability	F72	Severe intellectual disabilities
Cancers, Carcinomas, and Neoplasms			
230.3	Ca in situ colon	D01.0	Carcinoma in situ of colon
233.4	Ca in situ prostate	D07.5	Carcinoma in situ of prostate
188.9	Malig neo bladder NOS	C67.9	Malignant neoplasm of bladder, unspecified
154.1	Malignant neopl rectum	C20	Malignant neoplasm of rectum
Procedures influencing health status and contact with health services (including Colostomy and Ureterostomy)			
V55.3	Atten to colostomy	Z43.3	Encounter for attention to colostomy
V55.2	Atten to ileostomy	Z43.2	Encounter for attention to ileostomy
V55.6	Atten to urinostomy NEC	Z43.6	Encounter for attn to oth artif openings of urinary tract
V44.3	Colostomy status	Z93.3	Colostomy status
V44.1	Gastrostomy status	Z93.1	Gastrostomy status
V44.2	Ileostomy status	Z93.2	Ileostomy status
V44.6	Urinostomy status NEC	Z93.6	Other artificial openings of urinary tract status

- Problems with automatic coding:
  - High dimensional label space (ICD-9, used here, has 15 000 codes)
  - Clinical text contains irrelevant information, spelling errors, non-standard abbreviations, and a large medical vocabulary

- Problems with current automatic coding attempts:
  - Not accurate enough
  - Not “explainable” enough

- To combat these problems, the authors put together an “explainable” convolutional neural network
  - CAML (**C**onvolutional **A**ttention for **M**ulti-**L**abel classification)

- First, text is passed through a convolutional layer to compute a base representation of the text in each document

- Rather than aggregating across this representation with a pooling operation...
  - (Would be summarizing presence of features from different places of the document)

- … the authors applied an attention mechanism to select parts of the document that are most relevant to each possible code.
  - Selects k-grams from the text that are most relevant to each predicted label

```
Label: Action, k = 4
“ the quick brown fox jumps over the lazy dog ”
```

- These “attention weights” are applied to the base representation and the result is passed through the output layer

Output of the system: The likelihood of each code being associated with the document, and the k-gram (in this case, 4-gram) snippet from the text associated with each likelihood.

Example

934.1: "Foreign body in main bronchus"

CAML (HI)	*...line placed bronchoscopy performed showing* **large mucus plug on** *the left on transfer to...*
Cosine Sim	*...also needed medication to help* **your body maintain your** *blood pressure after receiving iv...*
CNN	*...found to have a large* **lll lingular pneumonia on** *chest x ray he was...*
Logistic Regression	*...impression confluent consolidation involving nearly* **the entire left lung** *with either broncho-centric or vascular...*

442.84: "Aneurysm of other visceral artery"

CAML (I)	*...and gelfoam embolization of right* **hepatic artery branch pseudoaneurysm** *coil embolization of the gastroduodenal...*
Cosine Sim	*...coil embolization of the gastroduodenal* **artery history of present** *illness the pt is a...*
CNN	*...foley for hemodynamic monitoring and* **serial hematocrits angio was** *performed and his gda was...*
Logistic Regression (I)	*...and gelfoam embolization of right* **hepatic artery branch pseudoaneurysm** *coil embolization of the gastroduodenal...*

428.20: "Systolic heart failure, unspecified"

CAML	*...no mitral valve prolapse moderate* **to severe mitral regurgitation** *is seen the tricuspid valve...*
Cosine Sim	*...is seen the estimated pulmonary* **artery systolic pressure is** *normal there is no pericardial...*
CNN	*...and suggested starting hydralazine imdur* **continue aspirin arg admitted** *at baseline cr appears patient...*
Logistic Regression (HI)	*...anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo...*

Evaluation of Classification

	AUC		F1				P@n	
Model	Macro	Micro	Macro	Micro	Diag	Proc	8	15
Scheurwegs et. al (2017)	–	–	–	–	0.428	0.555	–	–
Logistic Regression	0.561	0.937	0.011	0.272	0.242	0.398	0.542	0.411
CNN	0.806	0.969	0.042	0.419	0.402	0.491	0.581	0.443
Bi-GRU	0.822	0.971	0.038	0.417	0.393	0.514	0.585	0.445
CAML	0.895	**0.986***	**0.088**	**0.539***	**0.524***	**0.609***	**0.709***	**0.561***
DR-CAML	**0.897**	0.985	0.086	0.529	0.515	0.595	0.690	0.548

Table 4: Results on MIMIC-III full, 8922 labels. Here, "Diag" denotes Micro-F1 performance on diagnosis codes only, and "Proc" denotes Micro-F1 performance on procedure codes only. Here and in all tables, (*) by the bold (best) result indicates significantly improved results compared to the next best result, $p < 0.001$.

Evaluation of Interpretability

- k-grams were generated from:
  - CAML
  - Code descriptions directly
  - Logistic regression
  - Max-pooling CNN

- They asked one physician to rate how informative each of the k-grams are (not informative, informative, highly informative), and CAML got the highest number of highly informative k-grams

Method	Informative	Highly informative
CAML	46	22
Code Descriptions	48	20
Logistic Regression	41	18
CNN	36	13

How does this tie into explainability?

- If the CNN makes a classification decision, it can give a small sentence that is most highly associated with why it made that decision

- The physician is then tasked with understanding why the algorithm may have weighted that k-gram so highly

- They claimed that validating their k-grams with a physician was a measure of interpretability, but never mentioned explainability in that section

The end...