04/05/18

Improving the explainability of Random Forest classifier – user centered approach

Dragutin Petkovic† [1, 3], Russ Altman[2], Mike Wong[3], Arthur Vigil[4]

[1]Computer Science Department, San Francisco State University (SFSU)

[2]Department of Bioengineering, Stanford University

[3]SFSU Center for Computing for Life Sciences, 1600 Holloway Ave., San Francisco, CA 94132

[4]Twist Bioscience, 455 Mission Bay Boulevard South, San Francisco, CA 94158

Mary is deciding whether to adopt a ML-based diagnostic method

Mary has to make a decision based on *Current* state-of-the-art of presenting ML data

- ***ML algorithm used***
- ***Quality and details of the Training DB***
- ***Information about specific SW used***
- ***Accuracy and methods used to estimate it***

Mary's decision is critical for patients' well being and for the company

TO TRUST OR NOT TO TRUST?

What could have happened? (*Besides SW bugs and errors*)

- **MLDA could have performed *correctly* based on blindly following training data**

BUT

- **decision might be *fundamentally wrong - some examples can be found in:***
  - S. Kaufman, S. Rosset, C. Perlich: “Leakage in Data Mining: Formulation, Detection, and Avoidance”, ACM Transactions on Knowledge Discovery from Data 6(4):**1-21, December 2012**
  - “Can AI be Taught to Explain Itself”, NY Times Magazine Nov 2017

Seminar Outline

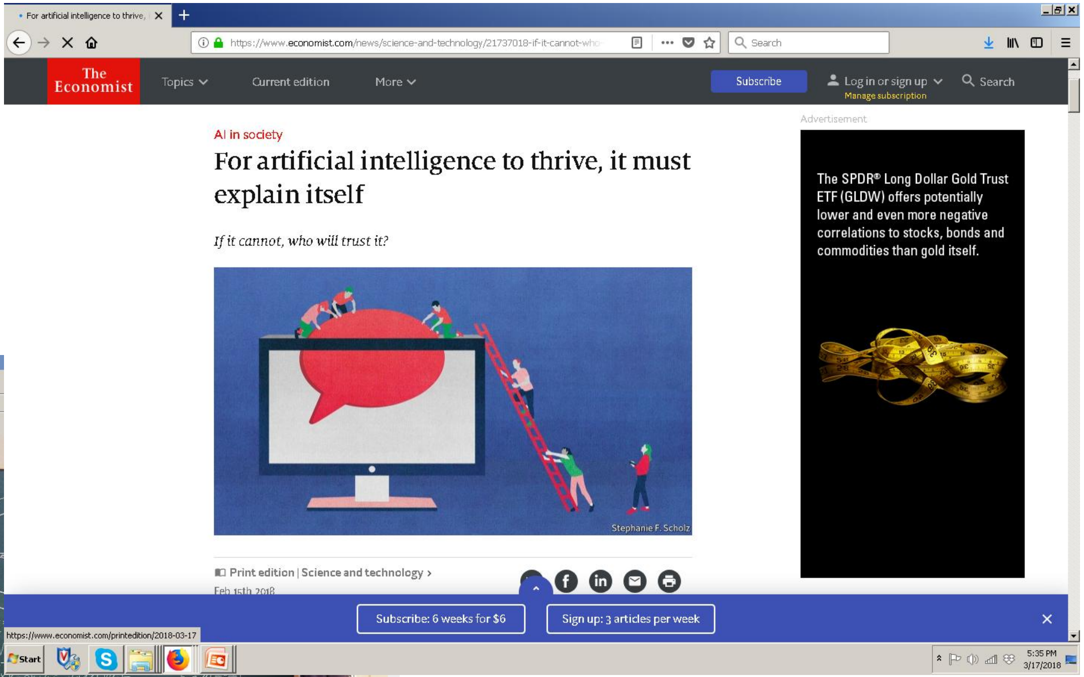
- Brief summary of Workshop on “ Machine learning and deep analytics for biocomputing: call for better explainability” held at PSB 2018 January 2018 (joint work with Prof. L. Kobzik and Prof. C. Re)
- Improving RF Explainability (RFEX) and case study using Stanford Feature data (joint work with Prof. R. Altman, M. Wong and A. Vigil)

What is ML *Explainability*?

- **Easy** to use information explaining **why and how** the ML approach made its decisions
  - **Model Explainability**: helps explain the ML model *as a whole*
  - **Sample explainability**: helps explain decision on *specific sample* (often user confidence is guided by ML accuracy on specific samples they know about)
- Targeted to both ML **experts and non-experts**

Demand for better MLDA explainability is growing including mainstream media and public

Economist Feb 2018


For artificial intelligence to thrive, i
https://www.economist.com/news/science-and-technology/21737018-if-it-cannot-who
Search
The Economist
Topics
Current edition
More
Subscribe
Log in or sign up
Manage subscription
Search
Advertisement
Stephanie F. Scholz
Subscribe: 6 weeks for $6
Sign up: 3 articles per week
https://www.economist.com/printedition/2018-03-17
Start
5:35 PM
3/17/2018

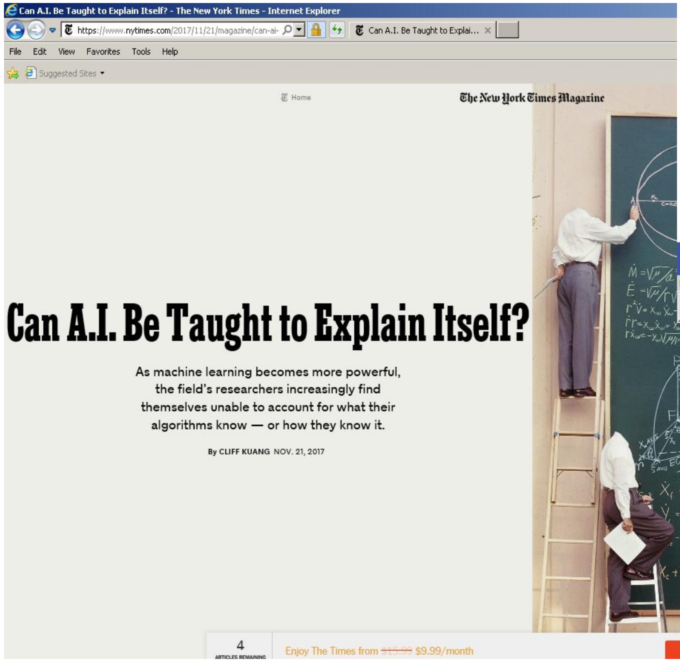
AI in society

For artificial intelligence to thrive, it must explain itself

If it cannot, who will trust it?

Print edition | Science and technology ›

Feb 15th 2018


Can A.I. Be Taught to Explain Itself? - The New York Times - Internet Explorer
https://www.nytimes.com/2017/11/21/magazine/can-ai-
Can A.I. Be Taught to Explai...
File
Edit
View
Favorites
Tools
Help
Suggested Sites
Home
The New York Times Magazine
4
ARTICLES REMAINING
SEE MY OPTIONS
Subscriber login

Can A.I. Be Taught to Explain Itself?

As machine learning becomes more powerful, the field's researchers increasingly find themselves unable to account for what their algorithms know — or how they know it.

By CLIFF KUANG NOV. 21, 2017

Enjoy The Times from ~~$15.99~~ $9.99/month

NY Times Magazine
Nov 21 2017

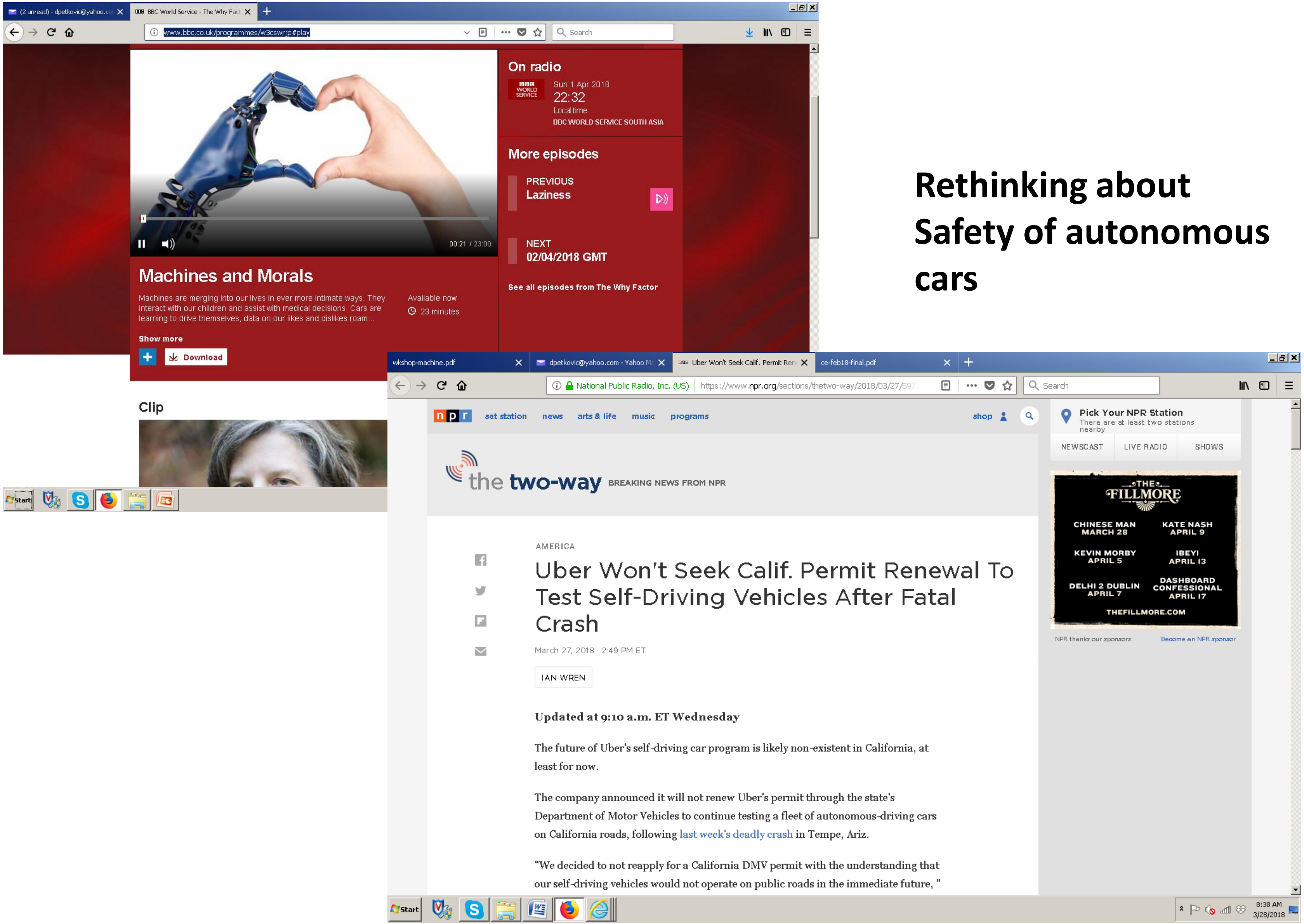
And will likely be regulated

- New EU General Data Protection laws
  - https://www.eugdpr.org/

- New IEEE Standard 70001 on Transparency of Autonomous Systems
  - https://standards.ieee.org/develop/project/7001.html

BBC Y factor – excellent podcast on machines and morals

Rethinking about Safety of autonomous cars



(2 unread) - dpetkovic@yahoo.co
BBC World Service - The Why Fac
www.bbc.co.uk/programmes/w3cswrjp#play
Search
00:21 / 23:00

On radio

Sun 1 Apr 2018
22:32
Local time
BBC WORLD SERVICE SOUTH ASIA

More episodes

PREVIOUS
Laziness

NEXT
02/04/2018 GMT

See all episodes from The Why Factor

Machines and Morals

Machines are merging into our lives in ever more intimate ways. They interact with our children and assist with medical decisions. Cars are learning to drive themselves, data on our likes and dislikes roam…

Available now
23 minutes

Show more

wkshop-machine.pdf
dpetkovic@yahoo.com - Yahoo M
Uber Won't Seek Calif. Permit Ren
ce-feb18-final.pdf
National Public Radio, Inc. (US)
https://www.npr.org/sections/thetwo-way/2018/03/27/597
Search
npr
set station
news
arts & life
music
programs
shop

Pick Your NPR Station
There are at least two stations nearby

the two-way BREAKING NEWS FROM NPR

AMERICA

Uber Won't Seek Calif. Permit Renewal To Test Self-Driving Vehicles After Fatal Crash

March 27, 2018 · 2:49 PM ET

IAN WREN

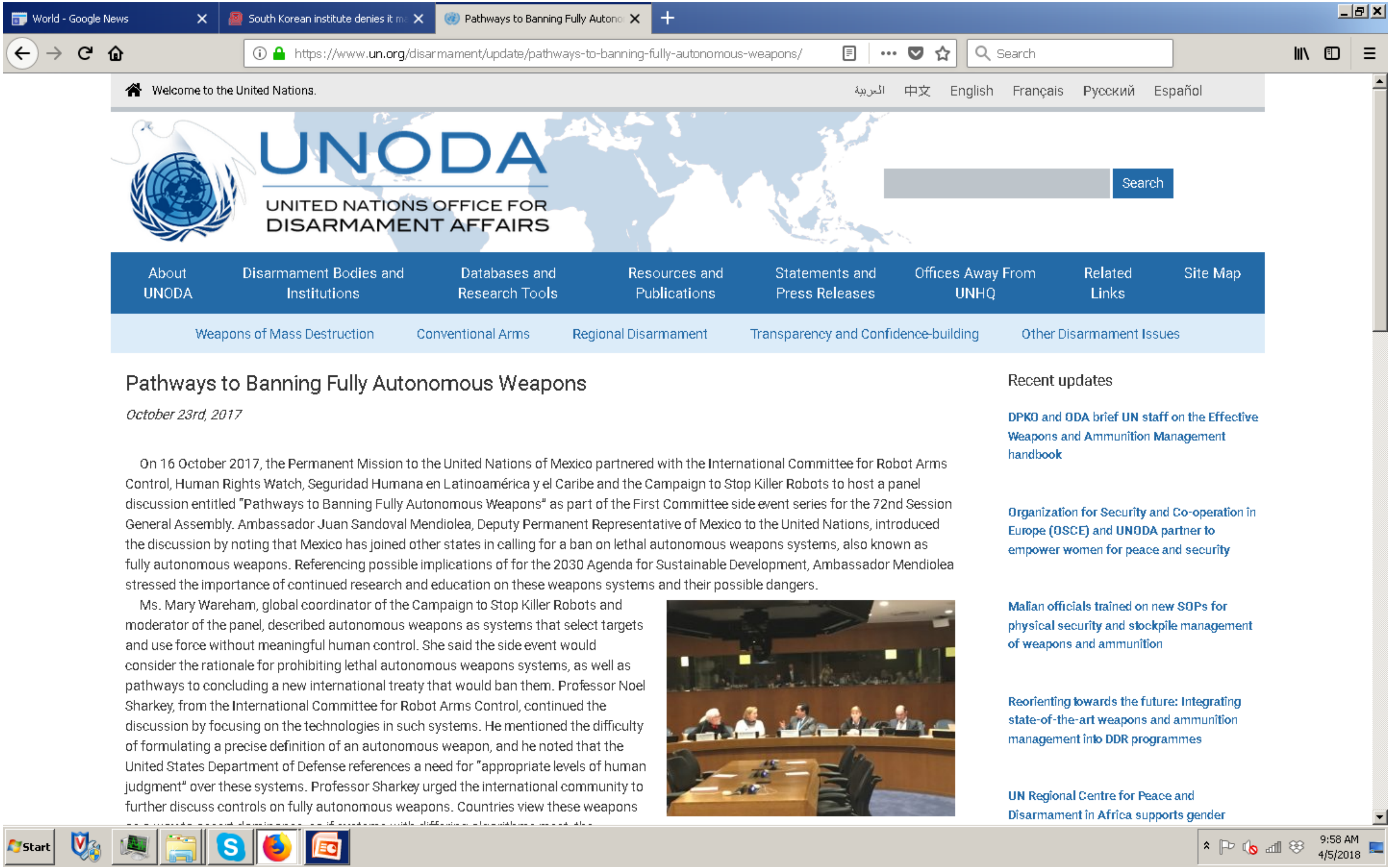
Updated at 9:10 a.m. ET Wednesday

The future of Uber's self-driving car program is likely non-existent in California, at least for now.

The company announced it will not renew Uber's permit through the state's Department of Motor Vehicles to continue testing a fleet of autonomous-driving cars on California roads, following last week's deadly crash in Tempe, Ariz.

"We decided to not reapply for a California DMV permit with the understanding that our self-driving vehicles would not operate on public roads in the immediate future, "

And of course…what about autonomous military machines….



World - Google News
South Korean institute denies it m
Pathways to Banning Fully Autono
https://www.un.org/disarmament/update/pathways-to-banning-fully-autonomous-weapons/
Search
Welcome to the United Nations.
العربية 中文 English Français Русский Español
UNODA
UNITED NATIONS OFFICE FOR DISARMAMENT AFFAIRS
Search
About UNODA
Disarmament Bodies and Institutions
Databases and Research Tools
Resources and Publications
Statements and Press Releases
Offices Away From UNHQ
Related Links
Site Map
Weapons of Mass Destruction
Conventional Arms
Regional Disarmament
Transparency and Confidence-building
Other Disarmament Issues
Start
9:58 AM
4/5/2018

Pathways to Banning Fully Autonomous Weapons

October 23rd, 2017

On 16 October 2017, the Permanent Mission to the United Nations of Mexico partnered with the International Committee for Robot Arms Control, Human Rights Watch, Seguridad Humana en Latinoamérica y el Caribe and the Campaign to Stop Killer Robots to host a panel discussion entitled "Pathways to Banning Fully Autonomous Weapons" as part of the First Committee side event series for the 72nd Session General Assembly. Ambassador Juan Sandoval Mendiolea, Deputy Permanent Representative of Mexico to the United Nations, introduced the discussion by noting that Mexico has joined other states in calling for a ban on lethal autonomous weapons systems, also known as fully autonomous weapons. Referencing possible implications of for the 2030 Agenda for Sustainable Development, Ambassador Mendiolea stressed the importance of continued research and education on these weapons systems and their possible dangers.

Ms. Mary Wareham, global coordinator of the Campaign to Stop Killer Robots and moderator of the panel, described autonomous weapons as systems that select targets and use force without meaningful human control. She said the side event would consider the rationale for prohibiting lethal autonomous weapons systems, as well as pathways to concluding a new international treaty that would ban them. Professor Noel Sharkey, from the International Committee for Robot Arms Control, continued the discussion by focusing on the technologies in such systems. He mentioned the difficulty of formulating a precise definition of an autonomous weapon, and he noted that the United States Department of Defense references a need for "appropriate levels of human judgment" over these systems. Professor Sharkey urged the international community to further discuss controls on fully autonomous weapons. Countries view these weapons

Recent updates

- DPKO and ODA brief UN staff on the Effective Weapons and Ammunition Management handbook
- Organization for Security and Co-operation in Europe (OSCE) and UNODA partner to empower women for peace and security
- Malian officials trained on new SOPs for physical security and stockpile management of weapons and ammunition
- Reorienting towards the future: Integrating state-of-the-art weapons and ammunition management into DDR programmes
- UN Regional Centre for Peace and Disarmament in Africa supports gender

Benefits of better ML Explainability

- Increased **confidence and trust** of application and domain experts as well as public in adopting ML;
- Better **validation, audit** and prevention of cases where ML approach produces results based on fundamentally wrong reasons or can behave in **unsafe** manner
- Simplification and **reduction of the cost** of application of ML in practice (e.g. by knowing which smaller feature subsets produce adequate accuracy)
- **Improved “maintenance”** where ML method has to be changed or tuned to new data or decision needs;
- Possible **discovery of new knowledge and ideas** (e.g. by discovering new patterns and factors that contribute to ML decisions)

PSB 2018 ML Explainability Workshop Goals

- Discuss *challenges in explainability* of current Machine Leaning and Deep Analytics (MLDA) used in biocomputing
- Start the discussion on *concrete ways to improve it*

Workshop involved both MLDA *researchers* and *users/adopters*

Petkovic, Kobzik, Re: *“*Machine learning and deep analytics for biocomputing: call for better explainability”, Pac Symp Biocomput. 2018;23:623-627.

Support of Stanford Mobilize Project is gratefully acknowledged

Interest was high

Workshop organization

Panel 1: What are the needs and problems *View of "Users "*

Moderator: Prof. Les Kobzik

- **Dr. R. Ghanadan -** *Google (since September 2017, previously at DARPA Explainable AI)*
- **Dr. W. Kibbe -** *Chief for Translational Biomedical Informatics in the Department of Biostatistics and Bioinformatics and chief data officer for the Duke Cancer Institute, Professor, Duke University since August 2017*
- **Dr. B. Percha -** *Assistant Professor, Icahn School of Medicine at Mount Sinai; Head of R&D, Health Data and Design Innovation Center (HD2i) Institute for Next-Generation Healthcare*

Panel 2: Some current examples and possible solutions – *View of "Developers"*

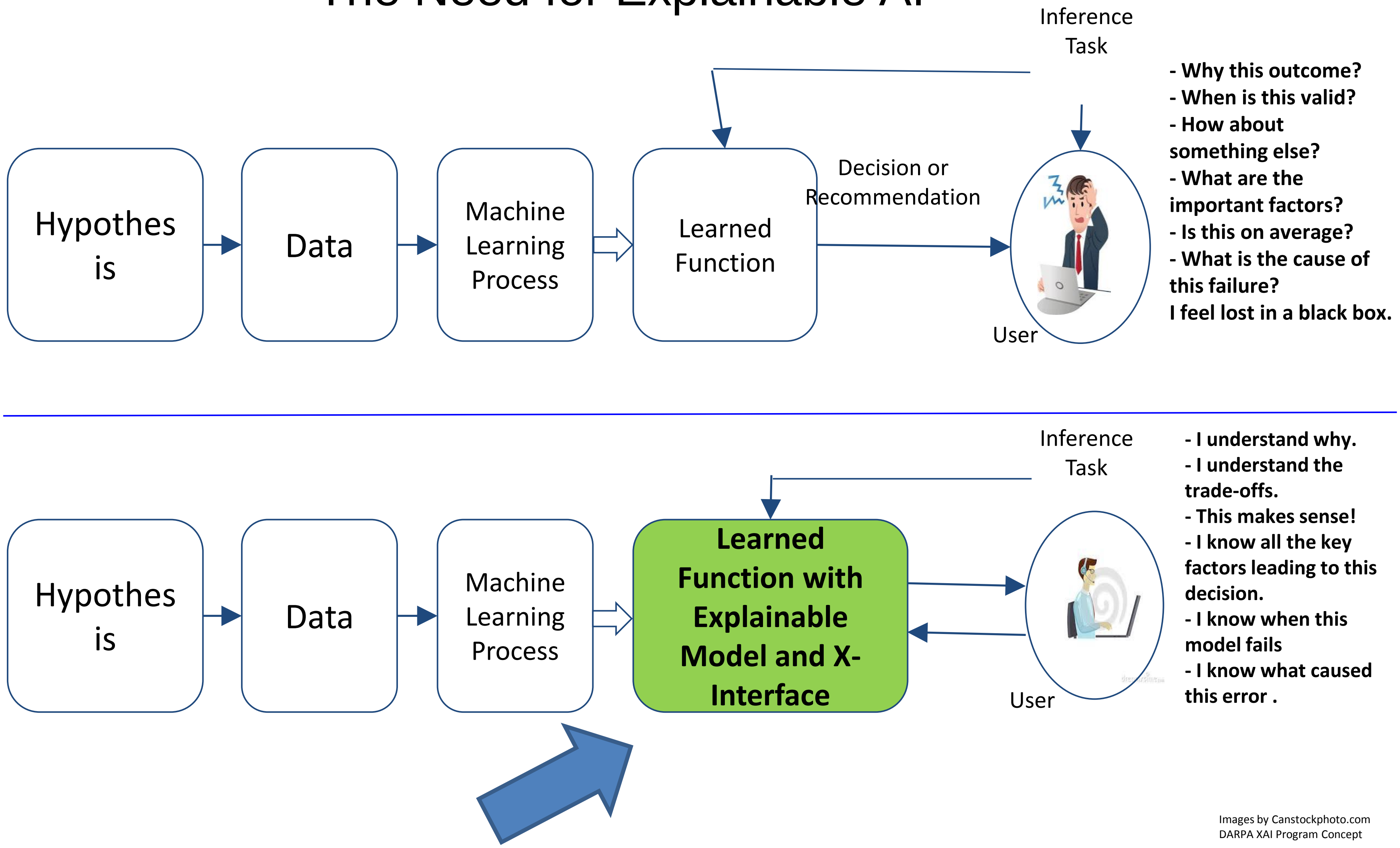
Moderator: Prof. Christopher Re

- **Dr. R. Roettger -** *Assistant Professor, University of Southern Denmark, Odense*
- **Dr. R. Scheuermann -** *Dir. Of Bioinformatics, J. Craig Venter Institute*
- **A. Esteva -** *Ph. D. Candidate, Stanford University*

Panel 3: discussion with panelists and audience

Highlights of some panelists’ talks

The Need for Explainable AI



Inference Task

Hypothes is

Data

Machine Learning Process

Learned Function

Decision or Recommendation

User

- Why this outcome?
- When is this valid?
- How about something else?
- What are the important factors?
- Is this on average?
- What is the cause of this failure?

I feel lost in a black box.

Inference Task

Hypothes is

Data

Machine Learning Process

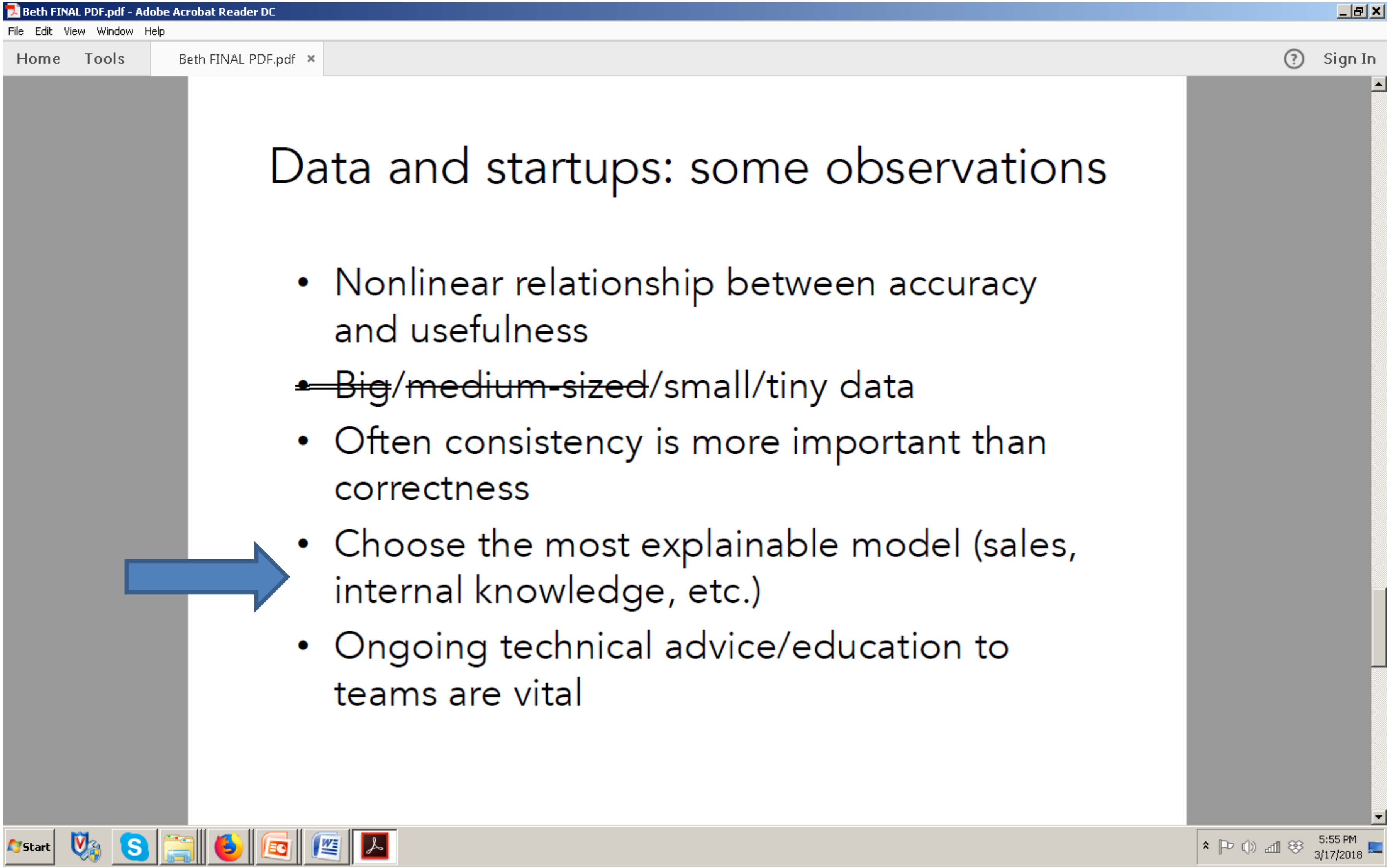
Learned Function with Explainable Model and X-Interface

User

- I understand why.
- I understand the trade-offs.
- This makes sense!
- I know all the key factors leading to this decision.
- I know when this model fails
- I know what caused this error .

Images by Canstockphoto.com
DARPA XAI Program Concept

From Dr. R. Ganahan: Explainable Models will accelerate the development and impact of ML/AI systems


Beth FINAL PDF.pdf - Adobe Acrobat Reader DC
File Edit View Window Help
Home
Tools
Beth FINAL PDF.pdf ×
Sign In
Start
5:55 PM
3/17/2018

Data and startups: some observations

- Nonlinear relationship between accuracy and usefulness
- ~~Big~~/~~medium-sized~~/small/tiny data
- Often consistency is more important than correctness
- Choose the most explainable model (sales, internal knowledge, etc.)
- Ongoing technical advice/education to teams are vital

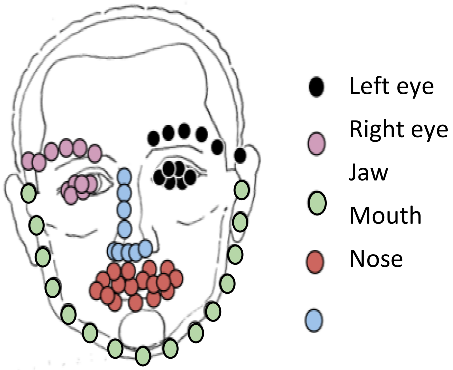
From: B. Percha

From A. Esteva

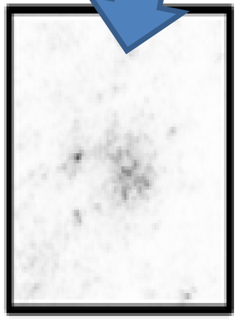
How to improve ML explainability?

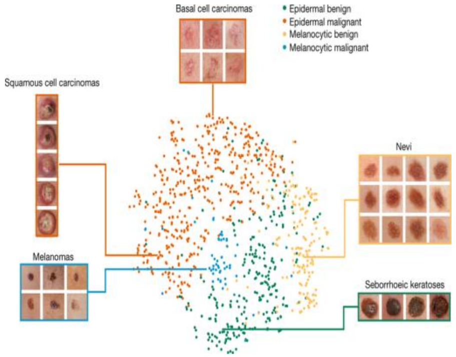
Data

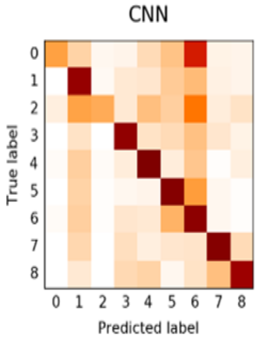
1. Understand your data



1. Understand the algorithm’s response to the data







These eyes haunt me…

Any model may pick out **unintended signal**. Deep models may pick out ***more*** unintended signal.

Upshot: Picked up on *mascara*

From: C. Re

Kuehlkamp et al. *Gender-from-Iris or Gender*

How do we make deep models robust? Add knowledge?

Toward Explainable Machine Learning - RFEX: Improving Random Forest Explainability

Prof. D. Petkovic
SFSU

Petkovic, Altman, Wong, Vigil: “Improving the explainability of Random Forest classifier – user centered approach**”,** Pac Symp Biocomput. 2018; 23: 204–215.

Random Forest ML

- Widely used
- Excellent performance
- Abundance of SW tools available
- Based on ensemble of trees
- Amendable to explanation and offers feature importance ranking
  - L. Breiman, “Random forests,” Machine Learning, vol. **45**, no. 1, pp.5–32, 2001

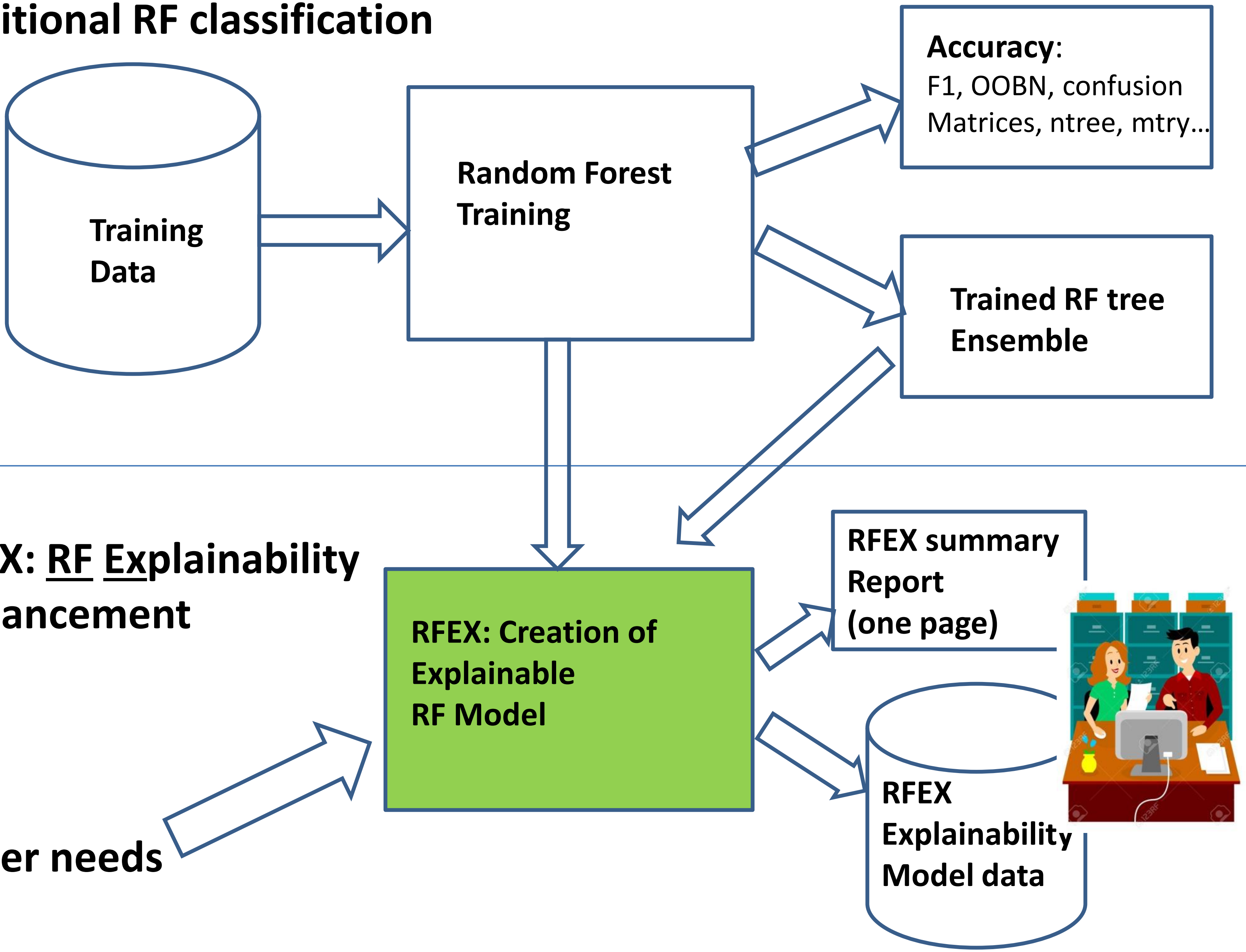
Current approaches in RF Explainability

- ***Feature ranking*** uses RF-provided variable importance measures like e.g. RF-provided Gini, MDA (mean decrease in accuracy) or others, to present them in *tables or horizontal bar charts* sorted by chosen variable importance measure.
  - Too simplistic, not done for + vs. – class separately. Lack of tradeoffs between features used and accuracy.

- ***Rule extraction from trained RF****.* This method consists of:
- a) performing standard RF training;
- b) defining rules by analyzing trained RF trees (resulting in very large set of rules, order of 100 K); and
- c) reducing the number and complexity of extracted rules by optimization to reduce to 10s – 100s of rules, each with 1-10 or so conditions.
  - Hard to interpret by humans; rules often complex; lack of tradeoffs between accuracy and number of rules used.

- **No "user design and evaluation"** with key adopters who are often non-RF expert users – the key constituency

Traditional RF classification



Training Data

Random Forest Training

Accuracy:
F1, OOBN, confusion
Matrices, ntree, mtry…

Trained RF tree Ensemble

RFEX: RF Explainability enhancement

RFEX: Creation of Explainable RF Model

User needs

RFEX summary Report (one page)

RFEX Explainability Model data

RFEX explainable model – driven by User/Adopter Needs/Questions

- *What is the **loss/tradeoffs of accuracy** if I use only certain subset of most important features?*

- *What are **most important features** contributing to ML prediction and how do they rank in importance?*
- *Also, tell me more about features:*
  - *What is the relationship of most important features for + vs. – class, is there any overlap?*
  - *What is "direction" of features? Abundance ("more of it" or "presence") or deficiency ("less of it" or "absence")? What thresholds I can use to determine this? What are basic class specific feature stats?*
  - *Which features interact together?*

- *Can explainable ML model be presented in an **easy to understand** and simple summary for ML/domain experts and non-experts?*

- Finally: evaluate if the RFEX explainable model is **helpful and intuitive to domain experts?**

Stanford FEATURE

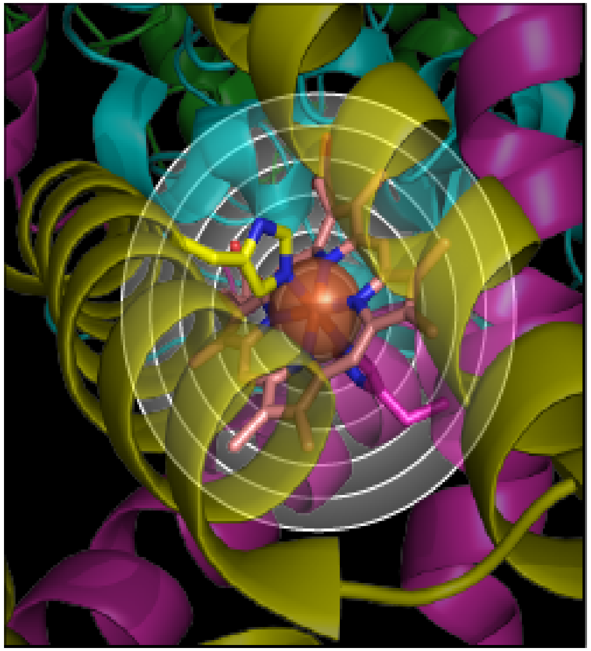
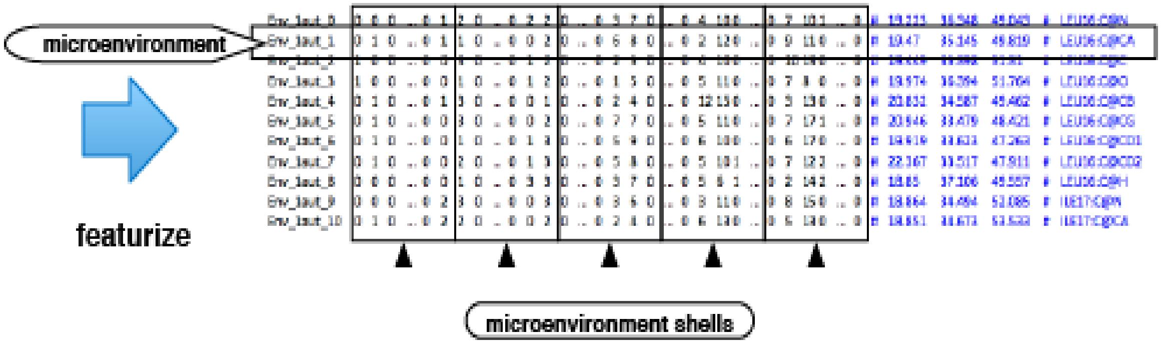


Figure A.
Conceptual diagram of a microenvironment, showing the concentric shells around a metal ion ligand.



```
Env_1aut_0  0 0 0 .. 0 1 2 0 .. 0 2 2 0 .. 0 3 7 0 .. 0 4 100 .. 0 7 101 .. 0 # 19.213 35.348 49.043 # LEU16:C@N
Env_1aut_1  0 1 0 .. 0 1 1 0 .. 0 0 2 0 .. 0 6 8 0 .. 0 2 120 .. 0 9 110 .. 0 # 19.47  35.145 49.819 # LEU16:C@CA
Env_1aut_2  1 0 0 .. 0 0 1 0 .. 0 1 2 0 .. 0 3 8 0 .. 0 4 100 .. 0 10 140 .. 0 # [illegible] # LEU16:C@C
Env_1aut_3  1 0 0 .. 0 0 1 0 .. 0 1 2 0 .. 0 1 5 0 .. 0 5 110 .. 0 7 8 0 .. 0 # 19.574 36.394 51.704 # LEU16:C@O
Env_1aut_4  0 1 0 .. 0 1 3 0 .. 0 0 1 0 .. 0 2 4 0 .. 0 12 150 .. 0 3 130 .. 0 # 20.832 34.587 49.462 # LEU16:C@CB
Env_1aut_5  0 1 0 .. 0 0 3 0 .. 0 0 2 0 .. 0 7 7 0 .. 0 5 110 .. 0 7 171 .. 0 # 20.946 33.479 48.421 # LEU16:C@CG
Env_1aut_6  0 1 0 .. 0 0 1 0 .. 0 1 3 0 .. 0 6 9 0 .. 0 6 100 .. 0 6 170 .. 0 # 19.619 33.623 47.263 # LEU16:C@CD1
Env_1aut_7  0 1 0 .. 0 0 2 0 .. 0 1 3 0 .. 0 5 8 0 .. 0 5 101 .. 0 7 122 .. 0 # 22.167 33.517 47.911 # LEU16:C@CD2
Env_1aut_8  0 0 0 .. 0 0 1 0 .. 0 3 3 0 .. 0 3 7 0 .. 0 5 8 1 .. 0 2 142 .. 0 # 18.85  37.106 49.557 # LEU16:C@H
Env_1aut_9  0 0 0 .. 0 2 3 0 .. 0 0 3 0 .. 0 3 6 0 .. 0 3 110 .. 0 8 150 .. 0 # 18.864 34.494 52.085 # ILE17:C@N
Env_1aut_10 0 1 0 .. 0 2 2 0 .. 0 0 2 0 .. 0 2 4 0 .. 0 6 130 .. 0 5 130 .. 0 # 18.851 33.673 53.533 # ILE17:C@CA
```

Figure B.
Computational representation of a microenvironment. The number of columns per shell have been abbreviated.
Feature Vector files look exactly like this.

The *featurize* program takes points in a biomolecule to build microenvironments (Figure A) and produce computational representations of said microenvironments (Figure B).

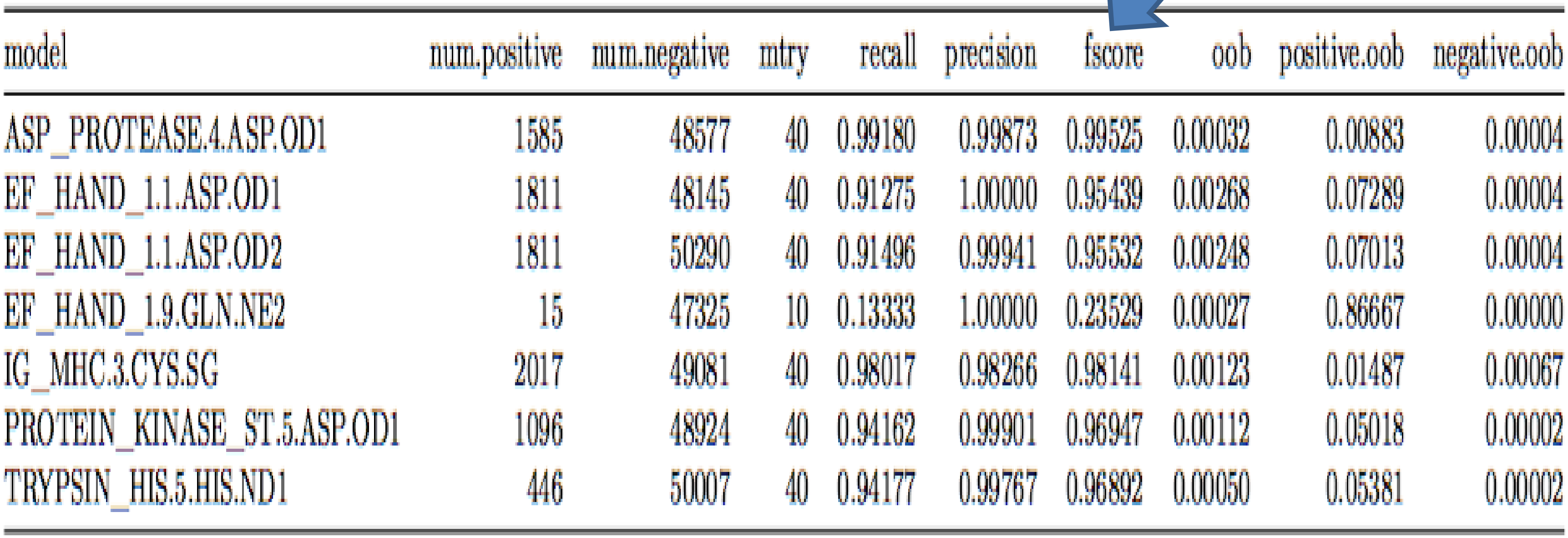
The *featurize* program analyzes mutually exclusive concentric spherical volumes (called shells) around a given point in a biomolecule. These shells collectively describe a microenvironment. The *featurize* program tallies physicochemical properties for each atom contained in each shell. These tallies form the computational representation describing the microenvironment. Groups of microenvironment computational representations are called Feature Vectors and are stored in Feature Vector Files.

Machine Learning can be trained on Feature Vectors to produce biomolecule functional class models. Biomolecules of unknown function can be characterized as Feature Vectors and scored against functional class models to predict functionality.

From AWS case study of Stanford-SFSU collaboration

STANFORD FEATURE DATA

F score: Main accuracy measure



model	num.positive	num.negative	mtry	recall	precision	fscore	oob	positive.oob	negative.oob
ASP_PROTEASE.4.ASP.OD1	1585	48577	40	0.99180	0.99873	0.99525	0.00032	0.00883	0.00004
EF_HAND_1.1.ASP.OD1	1811	48145	40	0.91275	1.00000	0.95439	0.00268	0.07289	0.00004
EF_HAND_1.1.ASP.OD2	1811	50290	40	0.91496	0.99941	0.95532	0.00248	0.07013	0.00004
EF_HAND_1.9.GLN.NE2	15	47325	10	0.13333	1.00000	0.23529	0.00027	0.86667	0.00000
IG_MHC.3.CYS.SG	2017	49081	40	0.98017	0.98266	0.98141	0.00123	0.01487	0.00067
PROTEIN_KINASE_ST.5.ASP.OD1	1096	48924	40	0.94162	0.99901	0.96947	0.00112	0.05018	0.00002
TRYPSIN_HIS.5.HIS.ND1	446	50007	40	0.94177	0.99767	0.96892	0.00050	0.05381	0.00002

Note *unbalanced* training data (many fewer positive samples)

Our previous work achieved good RF prediction but *we were not sure why*!

K. Okada, L. Flores, M. Wong, D. Petkovic, "Microenvironment-Based Protein Function Analysis by Random Forest", Proc. ICPR - International Conference on Pattern Recognition, Stockholm, 2014

RF feature ranking and accuracy

- **Feature Ranking:** We use **MDA** – Mean Decrease in Accuracy (part of RF alg.) - provided by all RF implementations
  - for each feature in dataset:
    randomly permute feature;
    make predictions on this permuted data;
    record average decrease in accuracy vs. using unpermuted data;
  - Permuting more important features result in larger decrease in accuracy

  (Permutation base ranking is more robust and less biased (in R tool 4 and later))

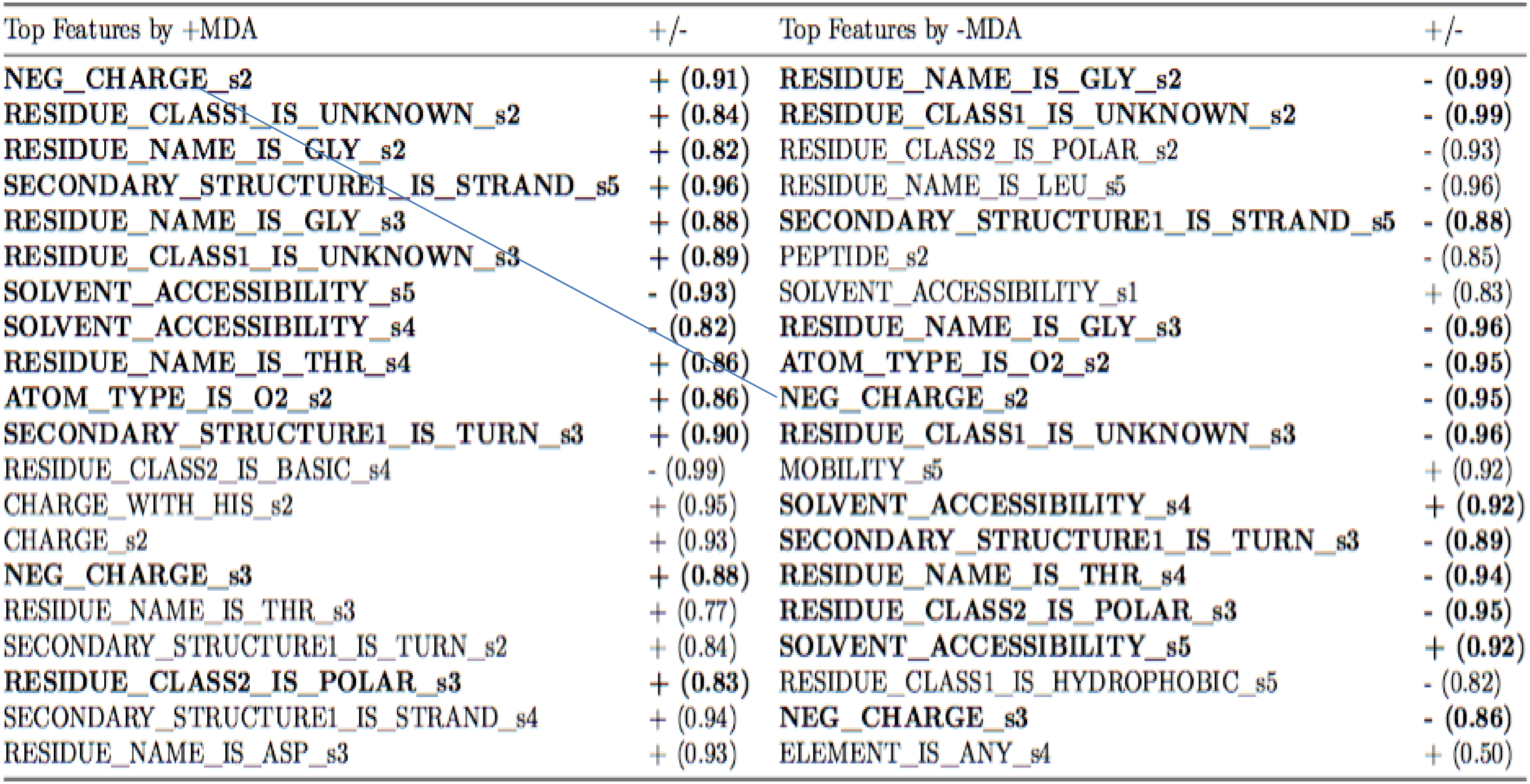
- RF consists of ensemble of disjoint trees, so best features used for + class may not be the same as those for – class ➔ MDA can be computed for + and – class *separately* (**MDA+; MDA-**) – important in case of *highly unbalanced data* (FEATURE data is unbalanced)

- **RF Accuracy**: we use ***F1 score*** for + class
  - F1 = 2* (precision*recall)/(precision + recall)
  - Precision/recall optimized by varying *cutoff* for ensemble tree voting

New RFEX measures to explain how features are used by RF: *Feature Direction* and *Mutual Feature Interaction*

- ***Feature Direction - DIR(I)*** *+ (n)* or *– (n):* denoting fraction of times (n) when feature I was above (+) (*abundance*) or below (-) (*deficiency*) the threshold when making correct prediction, for all trees in the forest making a correct prediction, and for all test samples
- ***Mutual Feature Interaction MFI(I,J) for features I and J*** **-** count of times features I and J appear on the same tree path making a correct prediction, for all trees in RF ensemble, and for all test samples.
    - Note that MFI only measures statistical pair-wise feature co-occurrences and not necessarily causality.

Ranking of top 20 features with MDA+ and MDA for ASP_PROTEASE.4.ASP.OD1



Top Features by +MDA	+/-	Top Features by -MDA	+/-
NEG_CHARGE_s2	**+ (0.91)**	**RESIDUE_NAME_IS_GLY_s2**	**- (0.99)**
RESIDUE_CLASS1_IS_UNKNOWN_s2	**+ (0.84)**	**RESIDUE_CLASS1_IS_UNKNOWN_s2**	**- (0.99)**
RESIDUE_NAME_IS_GLY_s2	**+ (0.82)**	RESIDUE_CLASS2_IS_POLAR_s2	- (0.93)
SECONDARY_STRUCTURE1_IS_STRAND_s5	**+ (0.96)**	RESIDUE_NAME_IS_LEU_s5	- (0.96)
RESIDUE_NAME_IS_GLY_s3	**+ (0.88)**	**SECONDARY_STRUCTURE1_IS_STRAND_s5**	**- (0.88)**
RESIDUE_CLASS1_IS_UNKNOWN_s3	**+ (0.89)**	PEPTIDE_s2	- (0.85)
SOLVENT_ACCESSIBILITY_s5	**- (0.93)**	SOLVENT_ACCESSIBILITY_s1	+ (0.83)
SOLVENT_ACCESSIBILITY_s4	**- (0.82)**	**RESIDUE_NAME_IS_GLY_s3**	**- (0.96)**
RESIDUE_NAME_IS_THR_s4	**+ (0.86)**	**ATOM_TYPE_IS_O2_s2**	**- (0.95)**
ATOM_TYPE_IS_O2_s2	**+ (0.86)**	**NEG_CHARGE_s2**	**- (0.95)**
SECONDARY_STRUCTURE1_IS_TURN_s3	**+ (0.90)**	**RESIDUE_CLASS1_IS_UNKNOWN_s3**	**- (0.96)**
RESIDUE_CLASS2_IS_BASIC_s4	- (0.99)	MOBILITY_s5	+ (0.92)
CHARGE_WITH_HIS_s2	+ (0.95)	**SOLVENT_ACCESSIBILITY_s4**	**+ (0.92)**
CHARGE_s2	+ (0.93)	**SECONDARY_STRUCTURE1_IS_TURN_s3**	**- (0.89)**
NEG_CHARGE_s3	**+ (0.88)**	**RESIDUE_NAME_IS_THR_s4**	**- (0.94)**
RESIDUE_NAME_IS_THR_s3	+ (0.77)	**RESIDUE_CLASS2_IS_POLAR_s3**	**- (0.95)**
SECONDARY_STRUCTURE1_IS_TURN_s2	+ (0.84)	**SOLVENT_ACCESSIBILITY_s5**	**+ (0.92)**
RESIDUE_CLASS2_IS_POLAR_s3	**+ (0.83)**	RESIDUE_CLASS1_IS_HYDROPHOBIC_s5	- (0.82)
SECONDARY_STRUCTURE1_IS_STRAND_s4	+ (0.94)	**NEG_CHARGE_s3**	**- (0.86)**
RESIDUE_NAME_IS_ASP_s3	+ (0.93)	ELEMENT_IS_ANY_s4	+ (0.50)

Some top features overlap in + vs. – classification
All directions very consistent (high %)
If features overlap their direction is opposite

Trade-offs in using subset of top ranked features vs. accuracy

Model Performance using a subset of features

We varied the number of features used to train our RF model from 2 to 20 and plotted the f-score for each trained model to show how model performance varied as we increased the number of features used.

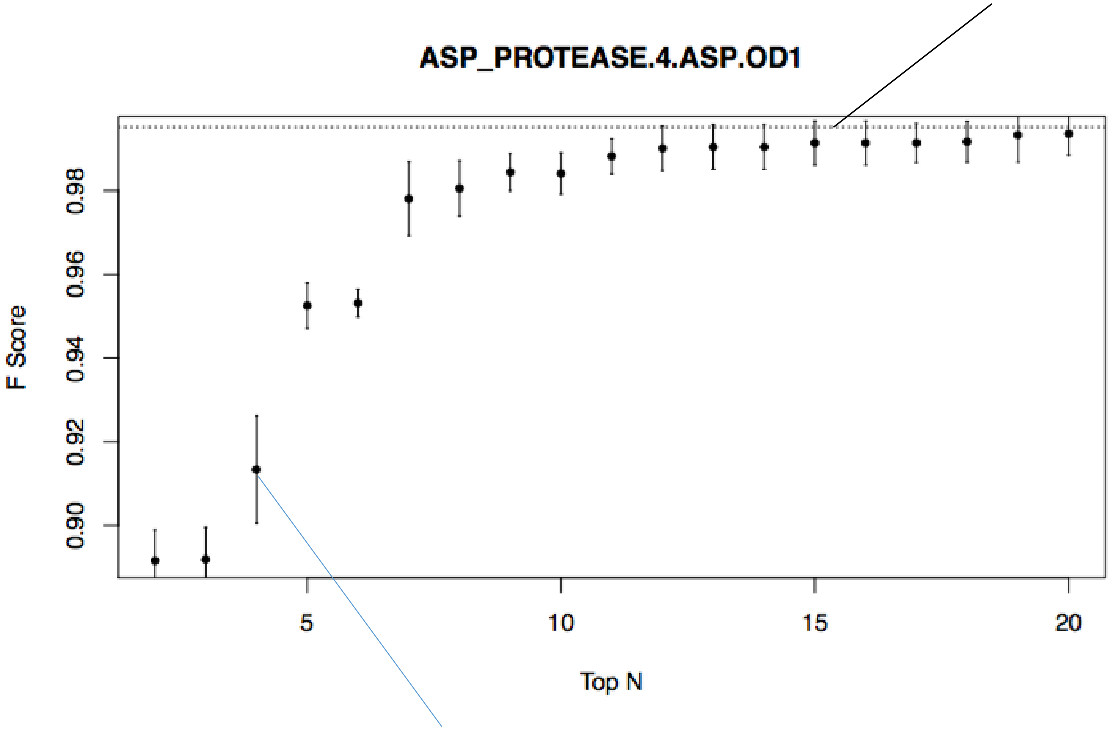
Observations:

•RF classifiers trained on just a few (between 10 and 20) features performed very close to RF using a 480 features

•Some of our models perform well even with just 2 or 3 features. Others showed a steeper drop off

•Standard deviation over CV trials small

F-score using all 480 features

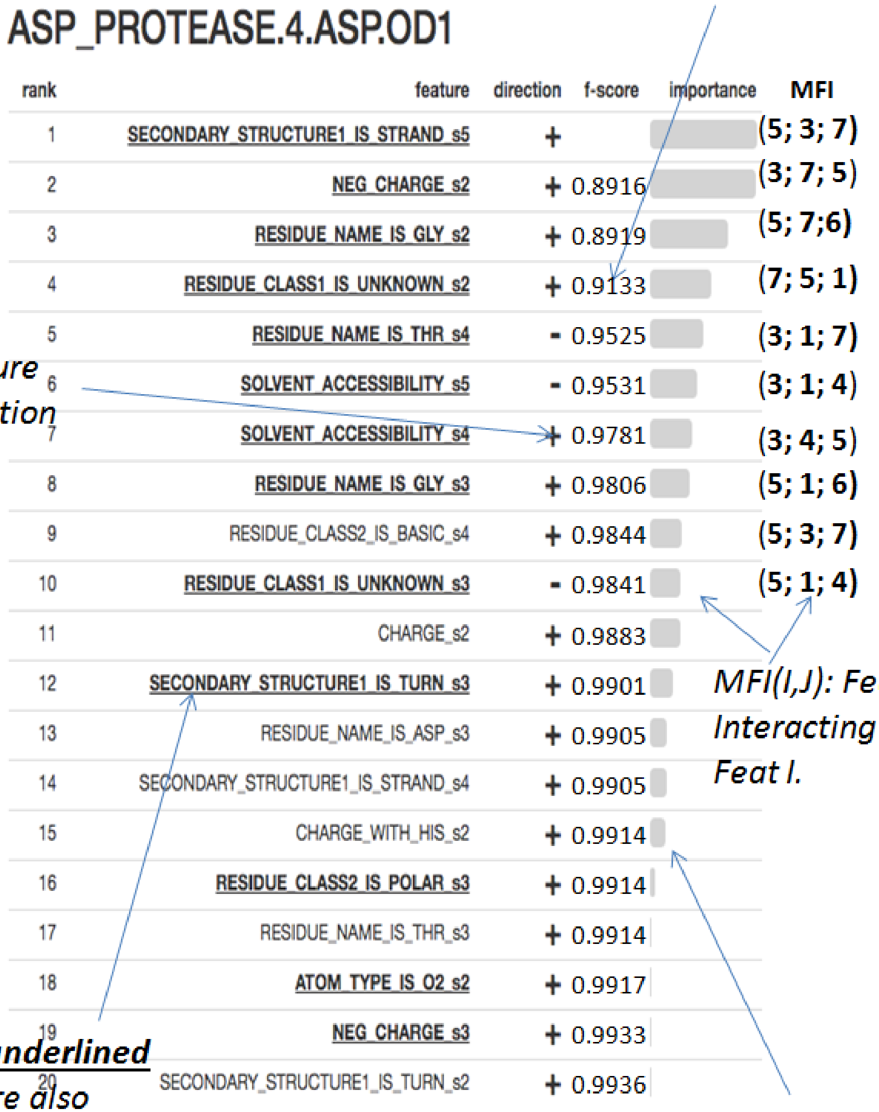


Average f-score for CV fold, and its SD

These charts reveal that one needs very few (from 2-6 depending on a model) top ranked features to achieve 90% or better of the accuracy when all 480 features are used.

RFEX One Page Summary report

Trade-off in accuracy using only top N (e.g. 4) best features



ASP_PROTEASE.4.ASP.OD1

rank	feature	direction	f-score	importance	MFI
1	**SECONDARY_STRUCTURE1_IS_STRAND_s5**	+			(5; 3; 7)
2	**NEG_CHARGE_s2**	+	0.8916		(3; 7; 5)
3	**RESIDUE_NAME_IS_GLY_s2**	+	0.8919		(5; 7;6)
4	**RESIDUE_CLASS1_IS_UNKNOWN_s2**	+	0.9133		(7; 5; 1)
5	**RESIDUE_NAME_IS_THR_s4**	-	0.9525		(3; 1; 7)
6	**SOLVENT_ACCESSIBILITY_s5**	-	0.9531		(3; 1; 4)
7	**SOLVENT_ACCESSIBILITY_s4**	+	0.9781		(3; 4; 5)
8	**RESIDUE_NAME_IS_GLY_s3**	+	0.9806		(5; 1; 6)
9	RESIDUE_CLASS2_IS_BASIC_s4	+	0.9844		(5; 3; 7)
10	**RESIDUE_CLASS1_IS_UNKNOWN_s3**	-	0.9841		(5; 1; 4)
11	CHARGE_s2	+	0.9883		
12	**SECONDARY_STRUCTURE1_IS_TURN_s3**	+	0.9901		
13	RESIDUE_NAME_IS_ASP_s3	+	0.9905		
14	SECONDARY_STRUCTURE1_IS_STRAND_s4	+	0.9905		
15	CHARGE_WITH_HIS_s2	+	0.9914		
16	**RESIDUE_CLASS2_IS_POLAR_s3**	+	0.9914		
17	RESIDUE_NAME_IS_THR_s3	+	0.9914		
18	**ATOM_TYPE_IS_O2_s2**	+	0.9917		
19	**NEG_CHARGE_s3**	+	0.9933		
20	SECONDARY_STRUCTURE1_IS_TURN_s2	+	0.9936		

Base RF accuracy using ALL features

RF trained on 48577 negatives, 1585 positives;

Performance f-score with all 480 features=0.9952

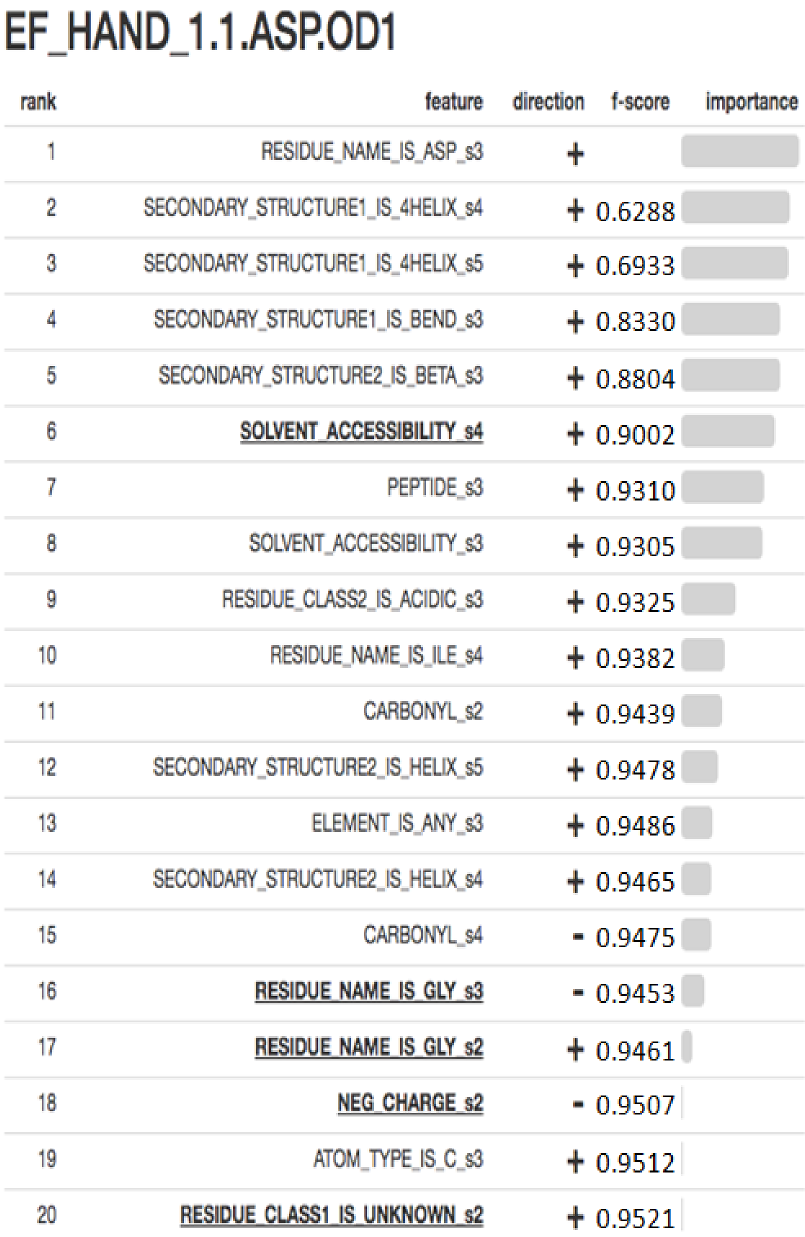
Classic way Of presenting RF accuracy

Feature direction

MFI(I,J): Features Interacting with Feat I.

Bold underlined
Feature also in top 20 for - class

Ranking of most important features with MDA+.



EF_HAND_1.1.ASP.OD1

rank	feature	direction	f-score	importance	MFI
1	RESIDUE_NAME_IS_ASP_s3	+			(6,7,8)
2	SECONDARY_STRUCTURE1_IS_4HELIX_s4	+	0.6288		(6,7,8)
3	SECONDARY_STRUCTURE1_IS_4HELIX_s5	+	0.6933		(6,7,1)
4	SECONDARY_STRUCTURE1_IS_BEND_s3	+	0.8330		(6,7,8)
5	SECONDARY_STRUCTURE2_IS_BETA_s3	+	0.8804		(8,7,6)
6	SOLVENT_ACCESSIBILITY_s4	+	0.9002		(7,1,4)
7	PEPTIDE_s3	+	0.9310		(6,1,8)
8	SOLVENT_ACCESSIBILITY_s3	+	0.9305		(7,6,1)
9	RESIDUE_CLASS2_IS_ACIDIC_s3	+	0.9325		(6,7,4)
10	RESIDUE_NAME_IS_ILE_s4	+	0.9382		(6,1,7)
11	CARBONYL_s2	+	0.9439		
12	SECONDARY_STRUCTURE2_IS_HELIX_s5	+	0.9478		
13	ELEMENT_IS_ANY_s3	+	0.9486		
14	SECONDARY_STRUCTURE2_IS_HELIX_s4	+	0.9465		
15	CARBONYL_s4	-	0.9475		
16	RESIDUE_NAME_IS_GLY_s3	-	0.9453		
17	RESIDUE_NAME_IS_GLY_s2	+	0.9461		
18	NEG_CHARGE_s2	-	0.9507		
19	ATOM_TYPE_IS_C_s3	+	0.9512		
20	RESIDUE_CLASS1_IS_UNKNOWN_s2	+	0.9521		

RF trained on 48145 negatives, 1811 positives Performance f-score with all 480 features=0.95439

RFEX pipeline summary – *general* steps in providing more explainable RF

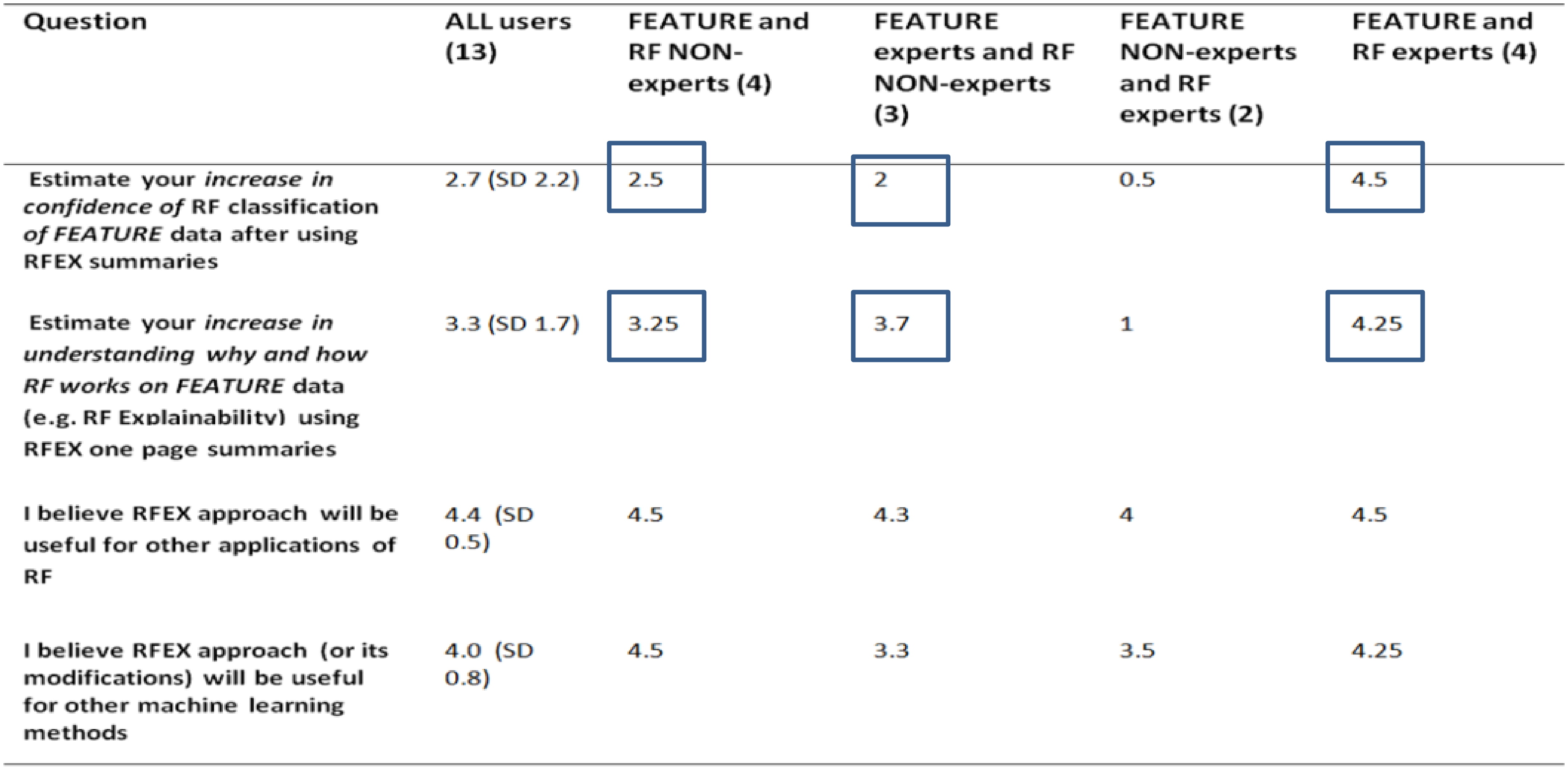
1. Establish ***Base RF Accuracy*** using all features (use F1 score)
2. ***Rank features/variables*** (e.g. use MDA and do it *separately* for + and – class if data is unbalanced)
3. Provide ***tradeoffs*** between features used and accuracy (e.g. what accuracy we can get using only top K ranked features)

Then work <u>only with Top N features</u> (N usually 2-5% of total number of features for 90% of original accuracy!!!)

4. Explain how ***features are used*** by RF
   - Determine class-specific feature stats: e.g. ***feature direction*** namely its *abundance* (more of it) or *deficiency* (less of it) or some other feature statistics (AV/SD/RANGE)
   - Determine which features ***interact*** with each other (MFI, correlation)
5. Create ***easy to use RFEX*** data and report (**one page**)

RFEX Usability review – anonymous survey of 13 expert and non-expert users

Measure how RFEX increases *user confidence* vs. using only traditional RF results



Question	ALL users (13)	FEATURE and RF NON-experts (4)	FEATURE experts and RF NON-experts (3)	FEATURE NON-experts and RF experts (2)	FEATURE and RF experts (4)
Estimate your *increase in confidence of* RF classification *of FEATURE* data after using RFEX summaries	2.7 (SD 2.2)	2.5	2	0.5	4.5
Estimate your *increase in understanding why and how RF works on FEATURE* data (e.g. RF Explainability) using RFEX one page summaries	3.3 (SD 1.7)	3.25	3.7	1	4.25
I believe RFEX approach will be useful for other applications of RF	4.4 (SD 0.5)	4.5	4.3	4	4.5
I believe RFEX approach (or its modifications) will be useful for other machine learning methods	4.0 (SD 0.8)	4.5	3.3	3.5	4.25

(1 low….5 high)

Future work

- Try RFEX on other RF applications
- Work on RFEX sample explainability
- Develop RFEX toolkit

- From repeatability to explainability to ***ethical, moral and safe AI***

Acknowledgements

We thank Prof. L. Kobzik, Dr. L. Buturovic and Prof. K. Okada for valuable feedback; T. Murray and J. Schwartz for help in organizing RFEX usability study; We also thank Prof. R. Altman and Dr. J. Ku for encouragement to pursue this work

The work has partially been supported by NIH grant R01 LM005652 and by Stanford Mobilize Project and SFSU Center for Computing for Life Sciences.

We also thank 13 anonymous usability reviewers.

Thank You