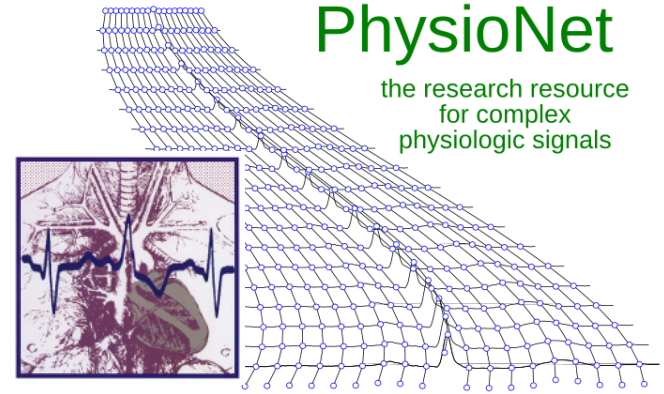


Machine learning for FDA-approved consumer level point of care diagnostics - the wisdom of algorithm crowds:

(the PhysioNet Computing in Cardiology Challenge 2017)

Gari D Clifford, Chengyu Liu,
Benjamin Moody, Roger Mark

Department of Biomedical Informatics, Emory University, Atlanta, GA USA
Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA
Institute for Medical Engineering & Science, Massachusetts Institute of Technology, USA

26th & 27th September 2017, @CinC Rennes France
Updated 17th October 2017 for #DSCO17
@ USF Data Science Institute, San Francisco, CA, USA

What is PhysioNet & its ‘Challenges’?

PhysioNet
the research resource for complex physiologic signals

PhysioNet: The NIH Research Resource for Complex Physiologic Signals – supported by

- National Institute of General Medical Sciences
- National Institute of Biomedical Imaging and Bioengineering
- Mostly physiological time series data
- 18 annual challenges since 2000 addressing key problems in field

PhysioNet/Computing in Cardiology Challenges

www.physionet.org/challenges/

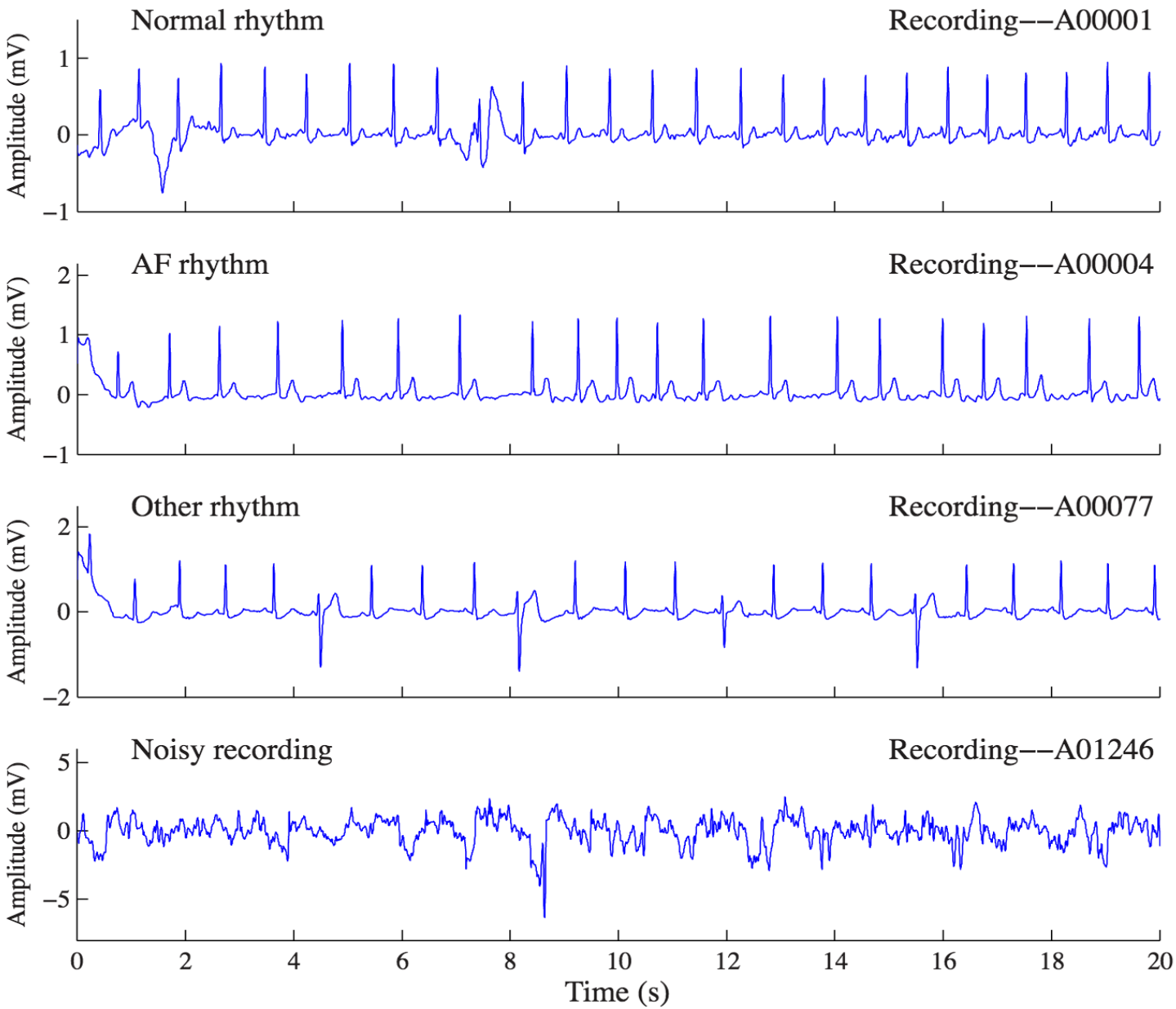
In cooperation with the annual Computing in Cardiology conference, PhysioNet hosts a series of **challenges**, inviting participants to tackle clinically interesting problems that are either unsolved or not well-solved.

Year	*Topic*	*Papers*	*Contributed Software*
2000	Detecting Sleep Apnea from the ECG	13	1
2001	Predicting Paroxysmal Atrial Fibrillation	9	
2002	RR Interval Time Series Modeling	12	10
2003	Distinguishing Ischemic from Non-Ischemic ST Changes	3	1
2004	Spontaneous Termination of Atrial Fibrillation	12	1
2005	The First Five Challenges Revisited	5	
2006	QT Interval Measurement	20	6
2007	Electrocardiographic Imaging of Myocardial Infarction	6	
2008	Detecting and Quantifying T-Wave Alternans	19	5 + 1
2009	Predicting Acute Hypotensive Episodes	11	4
2010	Mind the Gap	13	5
2011	Improving the Quality of ECGs Collected using Mobile Phones	17	7
2012	Predicting Mortality of ICU Patients	17	58
2013	Noninvasive Fetal ECG	29	17
2014	Robust Detection of Heart Beats in Multimodal Data	15	35
2015	Reducing False Arrhythmia Alarms in the ICU	20	28
2016	Classification of Normal/Abnormal Heart Sound Recordings	11	48
2017	AF Classification from a short single lead ECG recording		

The AliveCor ECG Device

- 3 generations of a single-channel (LA-RA lead I equivalent) ECG
- Transmitted to smartphone or tablet into the microphone (over the air) which digitizes at 44.1 kHz and 24-bit resolution with software demodulation in real-time.
- Frequency modulated with a carrier frequency of 19 kHz and a 200 Hz/mV modulation index.
- Stored as 300 Hz, 16-bit data with bandwidth 0.5-40 Hz with +/- 5 mV dynamic range.

Classify short ECG data into:



Initial Distribution of Data

- 12,186 single lead ECG recordings lasting from 9 s to just over 60 s
- Training set: 8,528 ECGs
- Test set: 3,658 ECG recordings
- Similar lengths and distributions

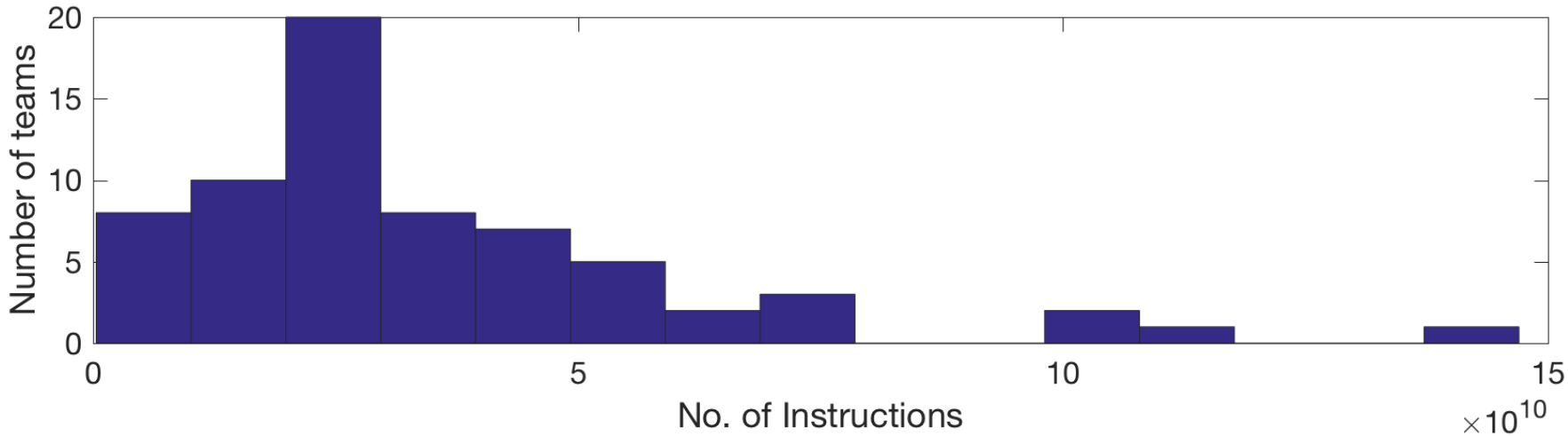
Dataset	Type	# recordings (%)
		Version 1
Training	Normal	5154 (60.4)
	AF	771 (9.0)
	Other	2557 (30.0)
	Noisy	46 (0.5)
Test	Normal	2209 (60.4)
	AF	331 (9.1)
	Other	1097 (30.0)
	Noisy	21 (0.6)

Is this big data? Well it's borderline … humans *can* do this ...
But they collect twice this amount of data daily.

Rules

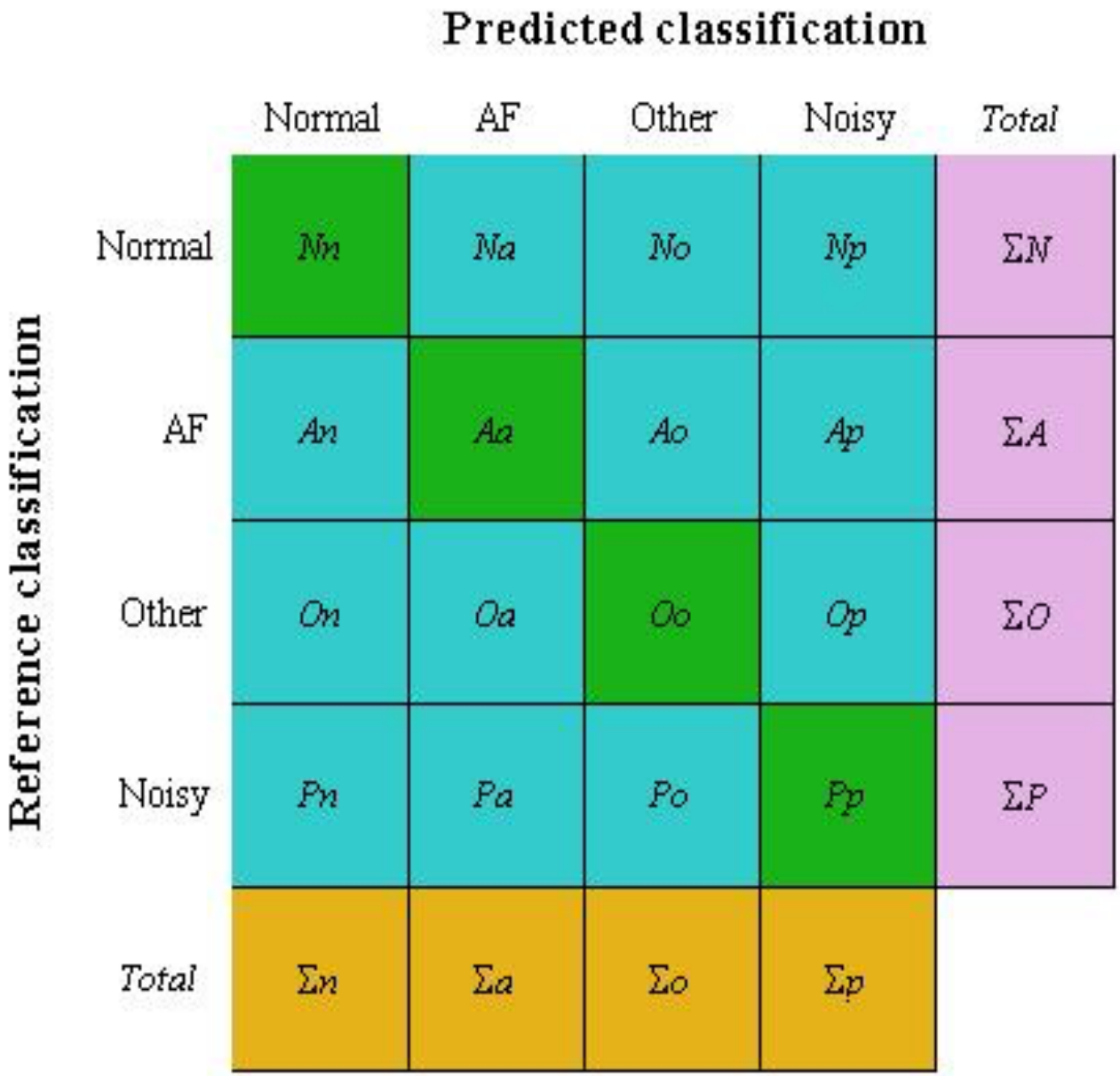
- Max 5 repeat entries in 3 month ‘unofficial phase’
  – maximise class average F1

- Max 10 repeat entries in ’official phase’



- Max of 2×10^{11} instructions per entry (2×10^{6}/sec) on an 1900-2600 MHz Opteron for trained algorithm … (If I can mechanical turk this, it’s pointless - as MJ stressed yesterday - your algorithm has to be cost effective)

Scoring



Normal: $F_{1n} = \frac{2 \times Nn}{\sum N + \sum n}$

AF rhythm: $F_{1a} = \frac{2 \times Aa}{\sum A + \sum a}$

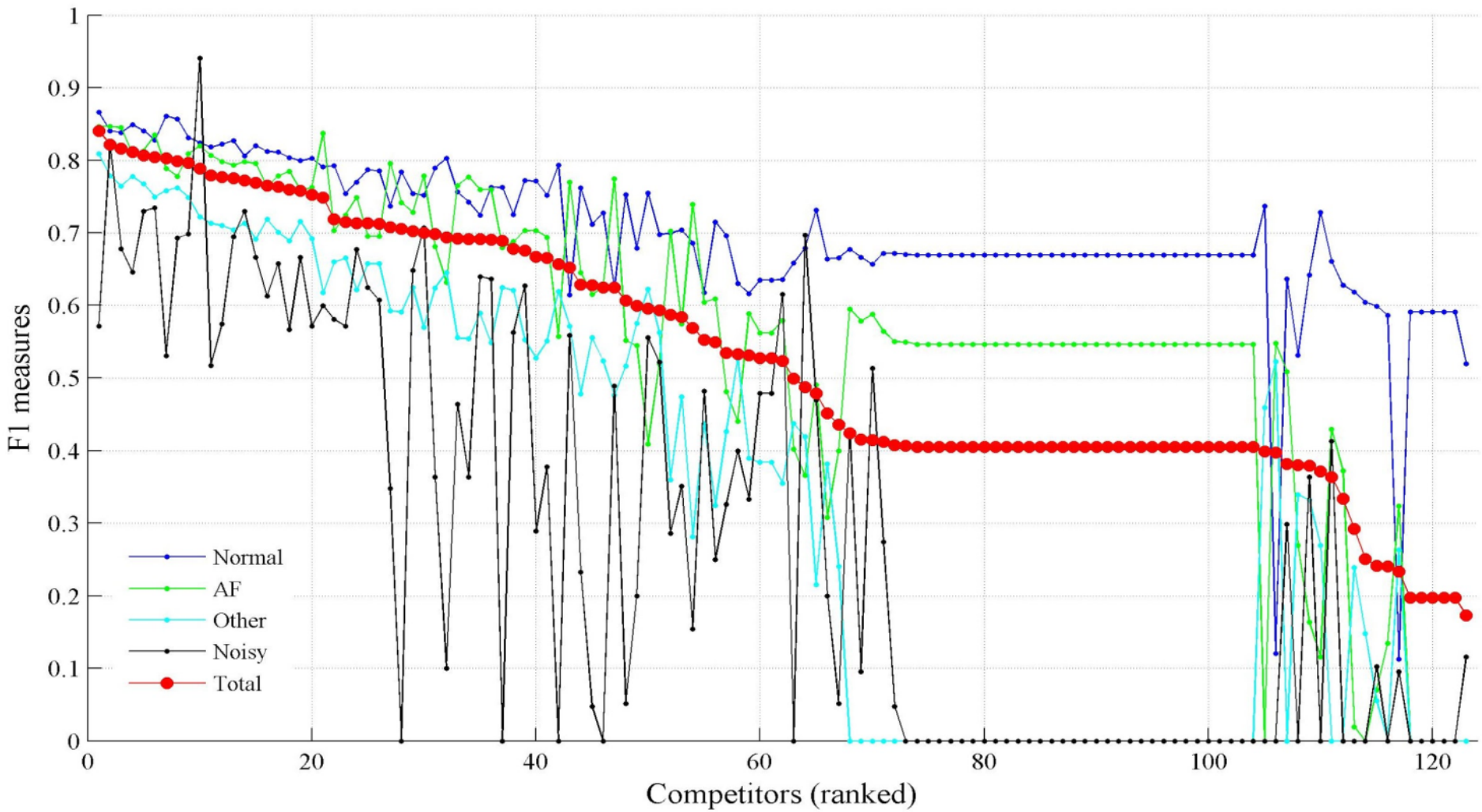
Other rhythm: $F_{1o} = \frac{2 \times Oo}{\sum O + \sum o}$

Score is:

$$F_1 = \frac{F_{1n} + F_{1a} + F_{1o}}{3}$$

Re-labeling for 1129 test recordings:

- Why relabel?
- First identified the top N=10 algorithms
- Then test recordings were ranked in order of disagreement level



Inter-rater agreement testing

- Fleiss' κ assesses the reliability of agreement between a fixed number of raters (≥ 2) when assigning categorical (non-ordinal) ratings to a number of items or classifying items.
- Calculates the degree of agreement in classification over that which would be expected by chance.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Calculate p_j, the proportion of all assignments which were to the j-th category:

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}, \qquad 1 = \sum_{j=1}^{k} p_j$$

Calculate P_i, the extent to which raters agree for the i-th subject (i.e., compute how many rater--rater pairs are in agreement, relative to the number of all possible rater--rater pairs):

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) = \frac{1}{n(n-1)} [(\sum_{j=1}^{k} n_{ij}^2) - (n)]$$

compute $\bar{P}$, the mean of the P_i 's, and $\bar{P}_e$ which go into the formula for κ :

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \qquad \bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$$

Re-labeling for 1129 test recordings: *Fleiss' κ*

n_{ij}	Raters' re-labeling results				P_i
	Normal	AF	Other	Noisy	
B00011	1	1	2	1	0.10
B00020	4	0	1	0	0.60
B00030	3	0	1	0	0.50
B00035	1	0	1	4	0.40
B00079	3	2	1	2	0.18
⋮	⋮	⋮	⋮	⋮	⋮
B03658	4	2	1	1	0.25
Total	2957	678	1292	1147	
p_j	0.49	0.11	0.21	0.19	

Re-labeling for 1129 test recordings:

Type	# recordings	Raters' re-labeling results				*Fleiss' κ*
		Normal	AF	Other	Noisy	
Normal	386	1203	136	353	367	**0.173**
AF	131	134	283	203	98	**0.113**
Other	525	1539	236	685	376	**0.197**
Noisy	87	81	23	51	306	**0.128**
Total	1129	2957	678	1292	1147	**0.245**

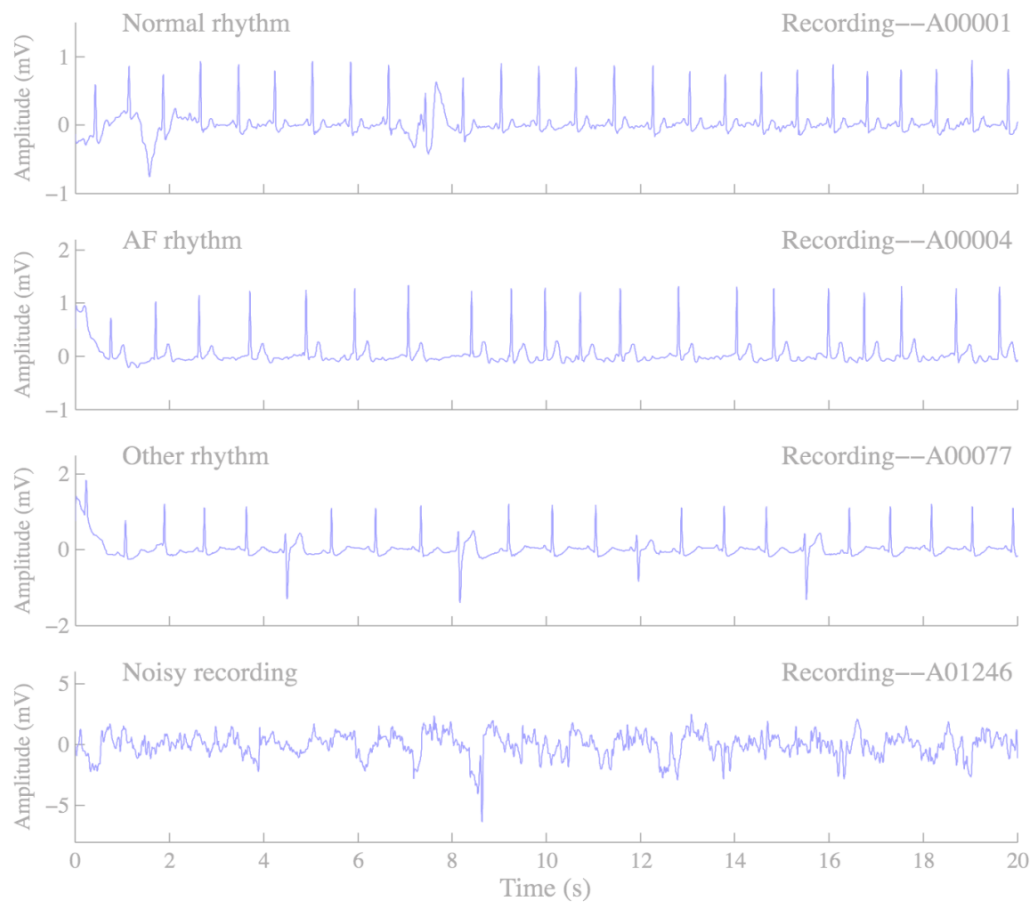
- Slight agreements among the annotators for each of the four classification type ($0.01 <= \kappa <= 0.20$)
- Fair agreement for all re-labeling task ($0.21 <= \kappa <= 0.40$)

Final version of the Challenge data

Dataset	Type	# recordings (%) in each REFERENCE version		
		Version 1	Version 2	Version 3
Training	Normal	5154 (60.4)	5050 (59.2)	5076 (59.5)
	AF	771 (9.0)	738 (8.7)	758 (8.9)
	Other	2557 (30.0)	2456 (28.8)	2415 (28.3)
	Noisy	46 (0.5)	284 (3.3)	279 (3.3)
Test	Normal	2209 (60.4)	2195 (60.0)	2437 (66.6)
	AF	331 (9.1)	315 (8.6)	286 (7.8)
	Other	1097 (30.0)	1015 (27.8)	683 (18.7)
	Noisy	21 (0.6)	133 (3.6)	252 (6.9)

Timeline, # teams and # entries

- ~6 months long (Jan 15 - Sep 1 2017)
- 75 International teams competed
- 70 Open Source Entries
- 5 Closed Source Entries
- 8 Unofficial Entries



Snapshot of leader board (not final); Top 35- 2/9/17

Overall score	Participant
0.86	guangyubin
0.85	zhaohanx
0.85	tomas.teijeiro
0.85	fplesinger
0.84	rohan.banerjee
0.84	rmaka08
0.84	philip.warrick
0.84	1501111363
0.83	martizih
0.83	fernando.andreotti
0.83	50227500
0.82	t3bs.team
0.82	robert.greer
0.82	mohbay
0.82	martin.kropf
0.82	jrubin01
0.82	amir.aminifar
0.81	vykintas.mak
0.81	sdnjly
0.81	oguzakbilgic
0.81	maurizio.varanini
0.81	godamartonaron
0.81	ecguru10
0.80	vessika
0.80	vadim.gliner
0.80	shivnarayan.patidar
0.80	joachim.a.behar
0.80	hoog.antink
0.80	chen2037
0.80	2514120821
0.80	18801178557
0.79	smolendawid
0.79	patrick.schwab
0.79	marcus.vollmer
0.79	b.whitaker

and the winners were (with F1=0.83) …

Detection of Atrial Fibrillation in ECG Hand-held Devices Using a Random Forest Classifier
Morteza Zabihi, Ali Bahrami Rad, Aggelos K. Katsaggelos, Serkan Kiranyaz, Susanna Narkilahti, Moncef Gabbouj

Arrhythmia Classification from the Abductive Interpretation of Short Single-lead ECG Records
Tomás Teijeiro, Constantino A. García, Paulo Félix, Daniel Castro

A Robust AF Classifier using Time and Frequency Features from Single Lead ECG Signal
Shreyasi Datta, Chetanya Puri, Ayan Mukherjee, Rohan Banerjee, Anirban Dutta Choudhury, Arijit Ukil, Soma Bandyopadhyay, Rituraj Singh, Arpan Pal, Sundeep Khandelwal

ENCASE: an ENsemble ClASsifiEr for ECG Classification Using Expert Features and Deep Neural Networks
Shenda Hong, Meng Wu, Yuxi Zhou, Qingyun Wang, Junyuan Shang, Hongyan Li, Junqing Xie

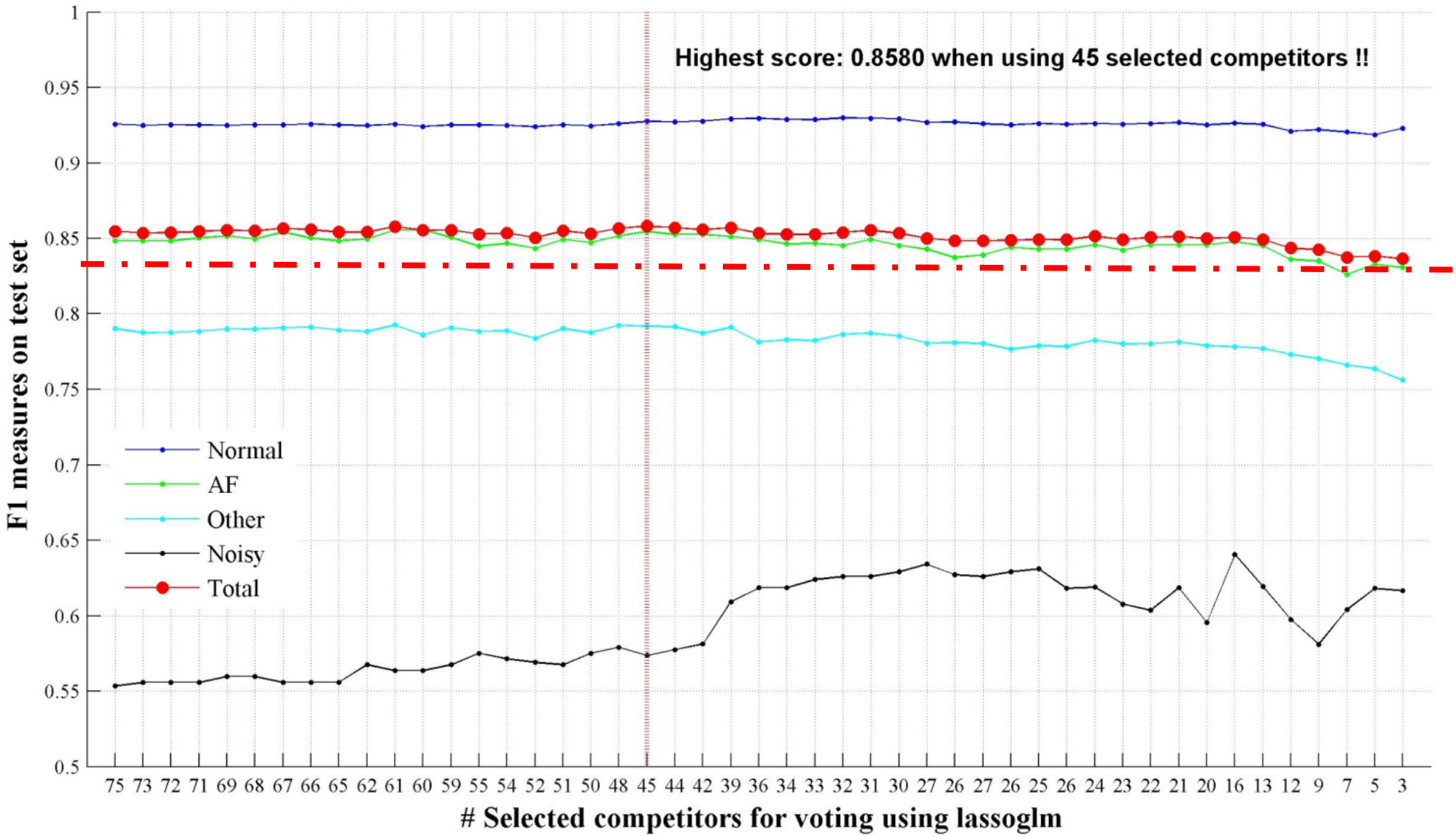
Final Ranking

Rank	valid	train	test	pr	entry
1	0.9122	0.8926	0.83	3	tomas.teijeiro@usc.es-220-entry.tar.gz
2	0.9902	0.9696	0.82	5	rohan.banerjee@tcs.com-209-entry.zip
*	0.9026	0.8955	0.82	6	rmaka08@gmail.com-209-entry.zip (*)
3	0.968	0.9511	0.82	12	t3bs.team@gmail.com-208-entry.zip
4	0.9866	0.9689	0.82	8	1501111363@pku.edu.cn-221-entry.zip
5	0.8591	0.9646	0.82	15	mohbay@gmail.com-208-entry.zip
6	0.8698	0.8747	0.82	1	guangyubin@bjut.edu.cn-211-entry.zip
7	0.9127	0.8888	0.82	9	martizih@student.ethz.ch-209-entry.zip
8	0.9046	0.877	0.81	2	zhaohanx@hotmail.com-282-entry.zip
9	0.9562	0.9349	0.81	15	martin.kropf@gmx.at-205-entry.zip
10	0.9236	0.9252	0.81	4	fplesinger@isibrno.cz-210-entry.zip
11	0.9902	0.9696	0.81	13	robert.greer@sickkids.ca-254-entry.zip
12	0.9018	0.8847	0.81	21	maurizio.varanini@ifc.cnr.it-213-entry.zip
13	0.9831	0.9644	0.81	26	shivnarayan.patidar@nitgoa.ac.in-210-entry.zip
14	0.8519	0.8395	0.80	23	ecguru10@gmail.com-213-entry.zip
15	0.9058	0.9031	0.80	11	50227500@qq.com-276-entry.zip
16	0.9902	0.9419	0.80	34	marcus.vollmer@uni-greifswald.de-240-entry.zip
17	0.8956	0.8913	0.80	32	smolendawid@gmail.com-206-entry.tar.gz
18	0.7900	0.7802	0.80	40	jiayuchen@outlook.com-202-entry.zip
*	0.8578	0.8391	0.80	30	2514120821@qq.com-210-entry.zip (*)
19	0.8508	0.8408	0.80	27	joachim.a.behar@gmail.com-214-entry.zip
20	0.823	0.796	0.80	24	vessika@biomed.bas.bg-204-entry.zip

(*) Unofficial

and the winners are … (almost) everyone

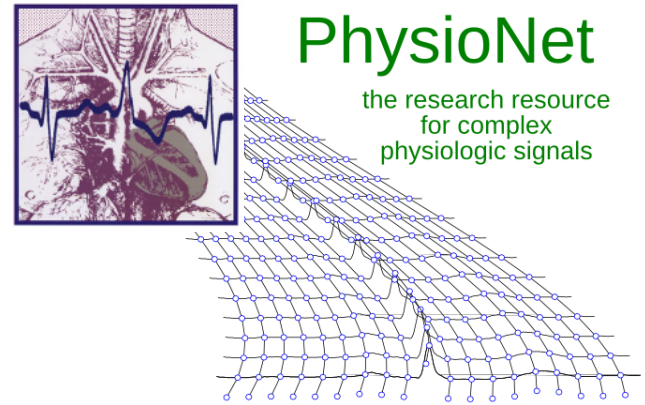
- Naive LASSO net selection and multivariate GLM classification gives highest F1 for N=45 (rises from 0.83 to 0.86)
- Improved F1 for normal and noisy classes without significant drop in F1 for AF & Other for N=48:53 … low numbers of noisy data.



Discussion

- Final scores and ranking were different to those on Sept 1st (the chosen 'best'/favorite algorithm was run on a larger test set after Sept 6th).
- Score dropped by 0.03 on average - so having 10 attempts allowed a slight overtraining on a third of the test data
- Best algorithms - wide variety - no clear favourite
- Combinations of algorithms worked better
- Data set composition/annotations:
  - How do we improve annotations?
- Scoring Function?
  - Should we do 2 class, weight AF higher, add more noise?
- Repeated Testing?
  - Doesn't everyone turn out to be every class in the end?

Thank you to:



- Mathworks for the prize money and free licenses during the competition!
- Dave Albert and Alivecor for the idea, data and hardware!
- Our glorious annotators:
  Dave Albert, Giovanni Angelotti, Christina Chen, Rodrigo Octavio Deliberato, Danesh Kella, Oleksiy Levantsevych, Roger Mark, Deepak Padmanabhan & Amit Shah
- Benjamin Moody and Chengyu Liu for heavy lifting
- All of the competitors!