Evaluating Machine Learning Models for PHM: *We're doing it wrong!*

Neil Eklund, Ph.D., FPHMS

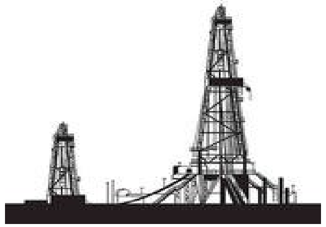
2 November 2022

Who is this guy?

Neil Eklund, Ph.D., FPHMS

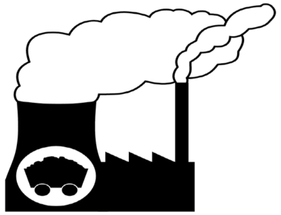
- *Principal Scientist*, Novity
  - *Principal Scientist*, PARC
  - *Chief Data Scientist*, Schlumberger
  - *Senior Data Scientist*, GE Digital/GE Research
- 20 years of deep technical experience across multiple industry segments – Aerospace, Energy, Healthcare, Oil & Gas, Financial, and Rail
- External customers include DARPA, NASA, Lockheed Martin, ExxonMobil, and Boeing
- Co-founder of the *Prognostics & Health Management Society*
- Founding Editor-in-Chief, *International Journal of Prognostics and Health Management* (ijPHM)
- 100+ publications, patents, and book chapters

Deployed Applications



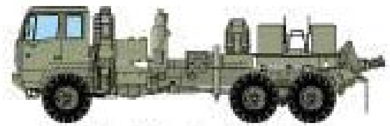
Drilling

- Predict success of next downhole run
- Data transmission to the cloud
- First deployed deep learning application for downhole tools
- $10MM+ annual return



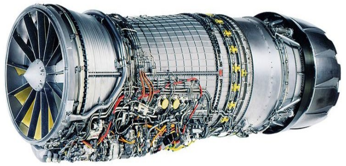
Power Plant Optimization & Control

- Multiobjective optimization of power plant
- Meet load constraint, maximize efficiency, minimize pollution
- Automated learning as the plant operates
- $14MM+ annual return per 400MW plant



Surface Equipment

- Predict failures a week in advance
- Wellsite data transmission to the cloud
- Automated model updating
- $50MM+ annual return



Military Aviation

- Defense Advanced Research Projects Agency (DARPA) project
- Fusion of data types (vibration, chips)
- Zero false alarms
- 5x increase in critical engine bearing spall detection capability

Production

- Predict failure 30 days in advance
- Structured and unstructured data
- Automated model updating
- $20MM+ annual return

Commercial Aviation

- Fault detection for GEnx and GE90 aircraft engines
- First deployed analytics application on GE *Predix* platform
- $10MM+ annual return

Overview

Summary

- Prognostics and Health Management (PHM) deals with a lot of binary questions, like “normal or anomalous?” or “faulty or not faulty?”
- PHM data is – virtually by definition – *wildly* imbalanced
- Binary machine learning models are commonly evaluated using Receiver Operating Characteristics (ROC) plots
- ROC plots yield a deceptive impression of performance for imbalanced data sets
- We should instead use Precision/Recall (PRC) plots to assess model performance

Outline

- Historical Context: From Electromagnetism to Signal Detection Theory
- Inherent Properties of PHM Problems
- A Close Look at ROC and Precision-Recall Curves

What is an ROC curve?

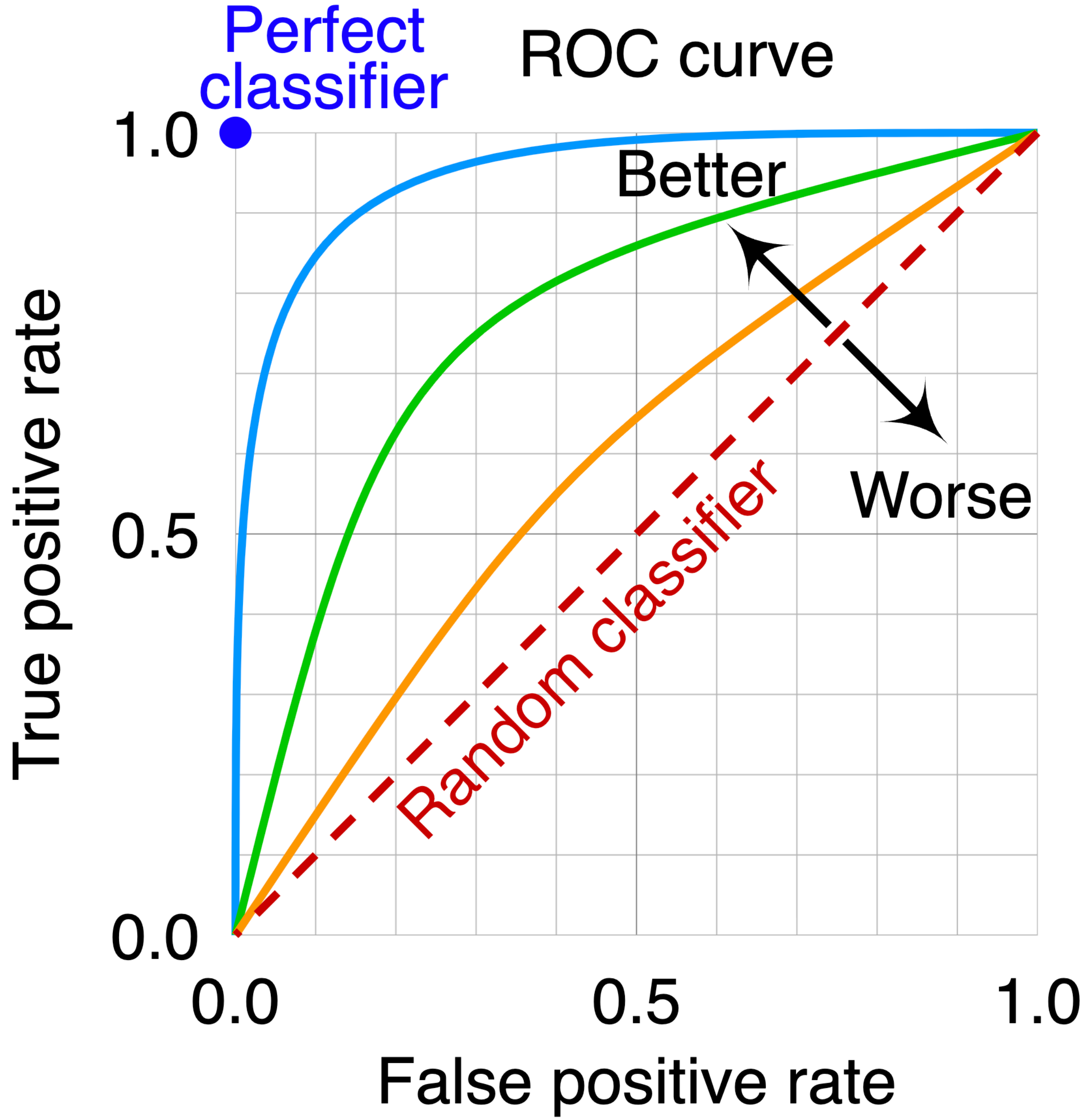
- A *binary classifier* produces output between a minimum and maximum value.
- When you pick a *classification threshold*, cases >= threshold are classified as “positive”; cases < threshold are “negative”.
- The *“accuracy”* statistic referrers to performance a particular classification threshold.
- A confusion matrix is also calculated for a particular threshold.
- However, if you scan over the range of possible threshold values (from min to max), you can – among other things – calculate the tradeoff between true positives and false alarms.

Confusion Matrix

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**True positive (TP)**	**False negative (FN)**
	Negative (N)	**False positive (FP)**	**True negative (TN)**

https://en.wikipedia.org/wiki/Confusion_matrix

What is an ROC curve?

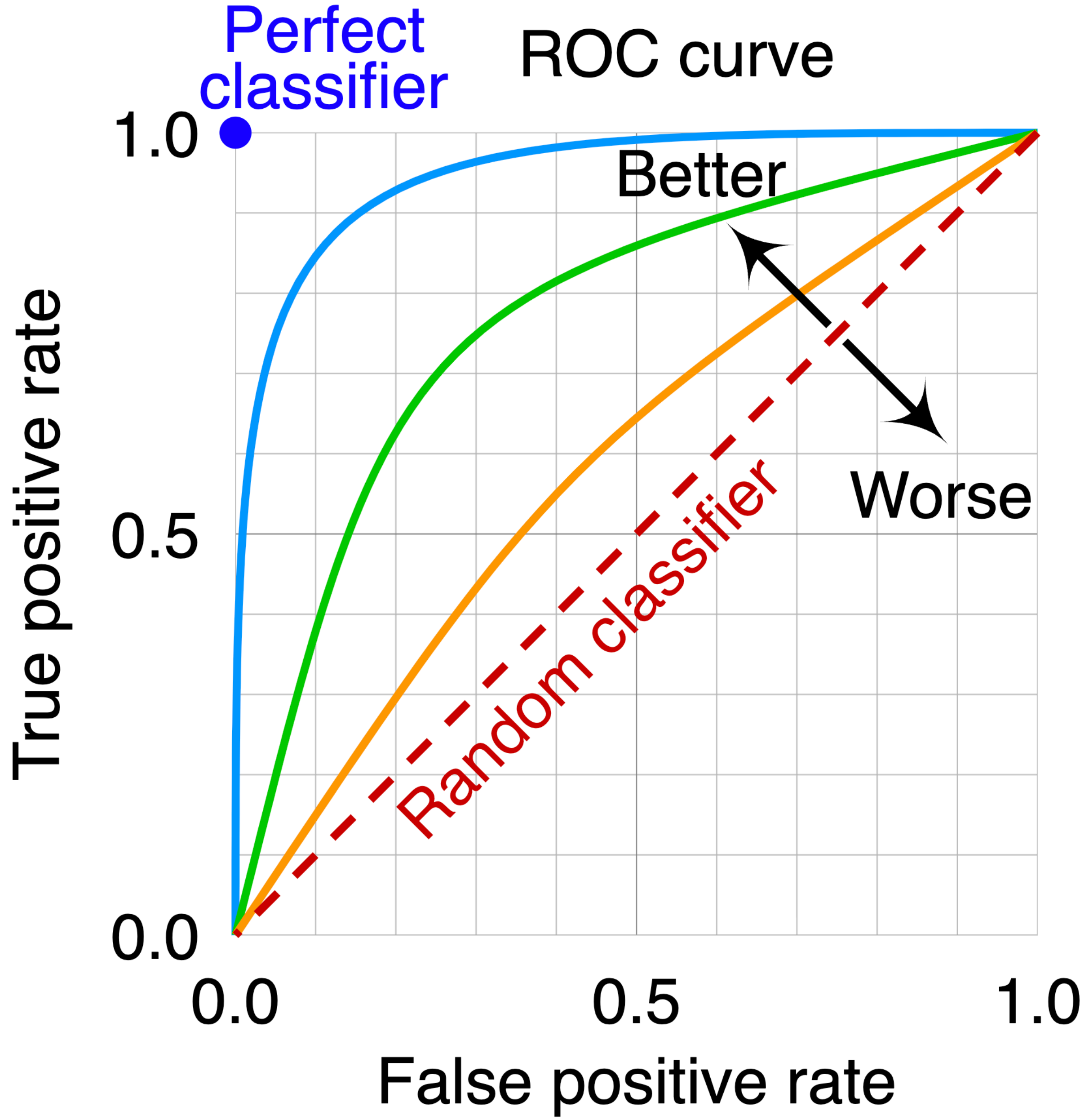


Confusion Matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:Roc_curve.svg

What is an ROC curve?



- The *AUC* statistic is a measure of “area under the curve”
- Maximum of 1.0, minimum of 0.5
- In general, higher AUCs are superior to lower AUCs
- e.g.,
  - AUC = 0.89
  - AUC = 0.73
  - AUC = 0.57

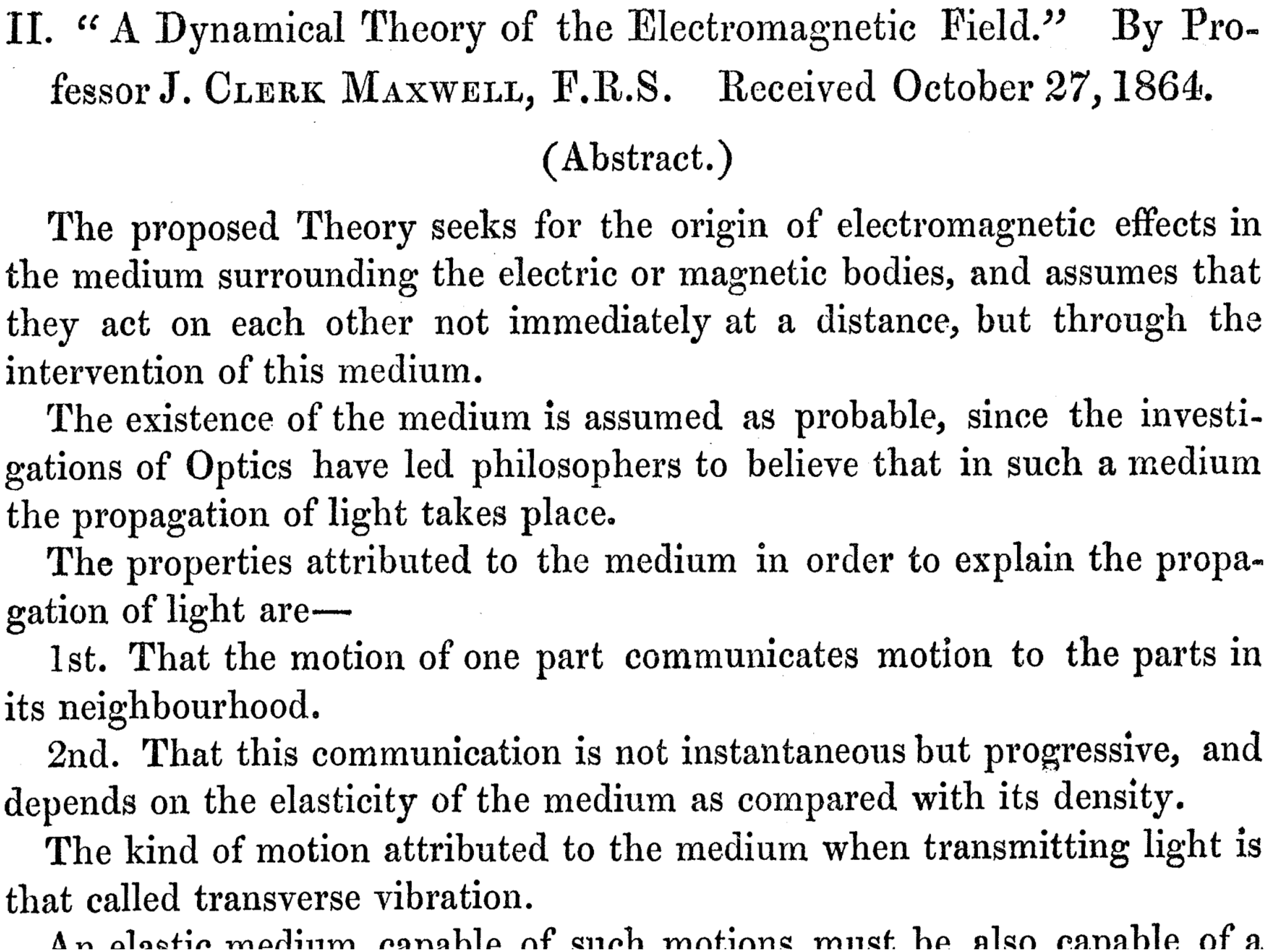
https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:Roc_curve.svg

phmsociety

Historical Context: From Electromagnetism to Signal Detection Theory

1864 – Maxwell Proposes Electromagnetism

The agreement of the results seems to show that light and magnetism are affections of the same substance, and that light is an electromagnetic disturbance propagated through the field according to electromagnetic laws



II. "A Dynamical Theory of the Electromagnetic Field." By Professor J. Clerk Maxwell, F.R.S. Received October 27, 1864.

(Abstract.)

The proposed Theory seeks for the origin of electromagnetic effects in the medium surrounding the electric or magnetic bodies, and assumes that they act on each other not immediately at a distance, but through the intervention of this medium.

The existence of the medium is assumed as probable, since the investigations of Optics have led philosophers to believe that in such a medium the propagation of light takes place.

The properties attributed to the medium in order to explain the propagation of light are—

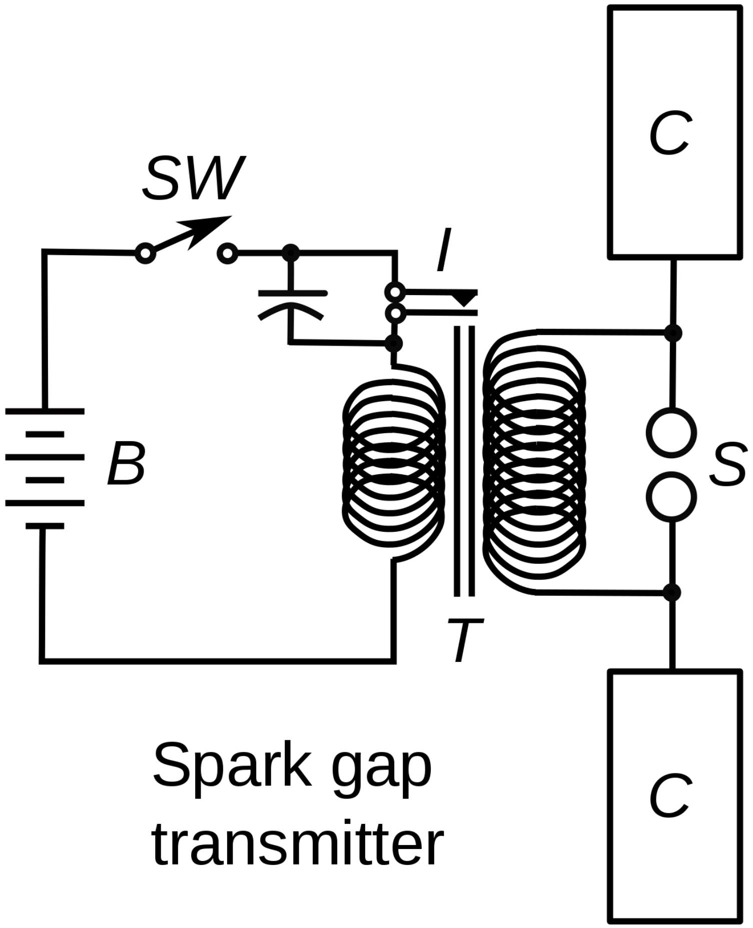
1st. That the motion of one part communicates motion to the parts in its neighbourhood.

2nd. That this communication is not instantaneous but progressive, and depends on the elasticity of the medium as compared with its density.

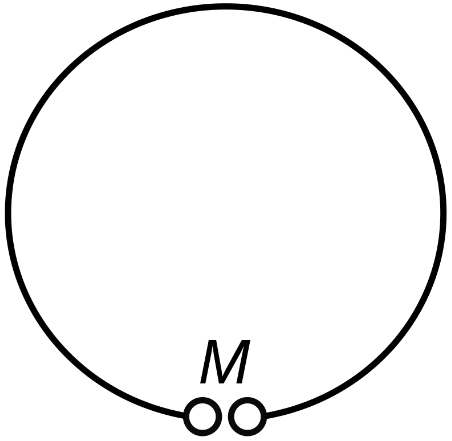
The kind of motion attributed to the medium when transmitting light is that called transverse vibration.

An elastic medium capable of such motions must be also capable of a

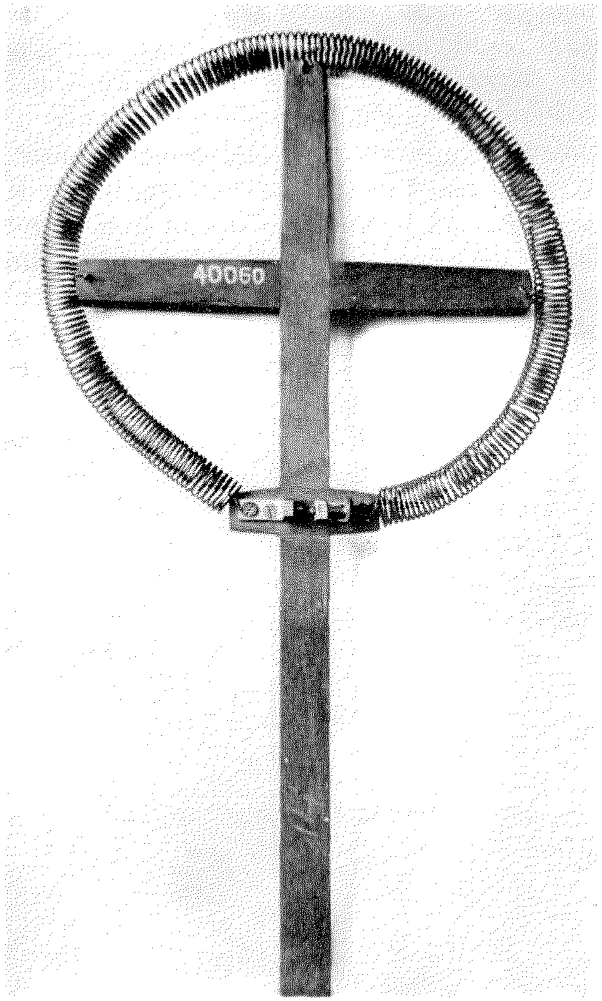
1887 – Hertz Generates and Detects Radio Waves



Spark gap transmitter



Receiver



https://en.wikipedia.org/wiki/Heinrich_Hertz

1887 – Hertz Produces and Detects Radio Waves

Although Hertz did not understand the practical implications,

> *“It's of no use whatsoever… this is just an experiment that proves Maestro Maxwell was right—we just have these mysterious electromagnetic waves that we cannot see with the naked eye. But they are there.”*

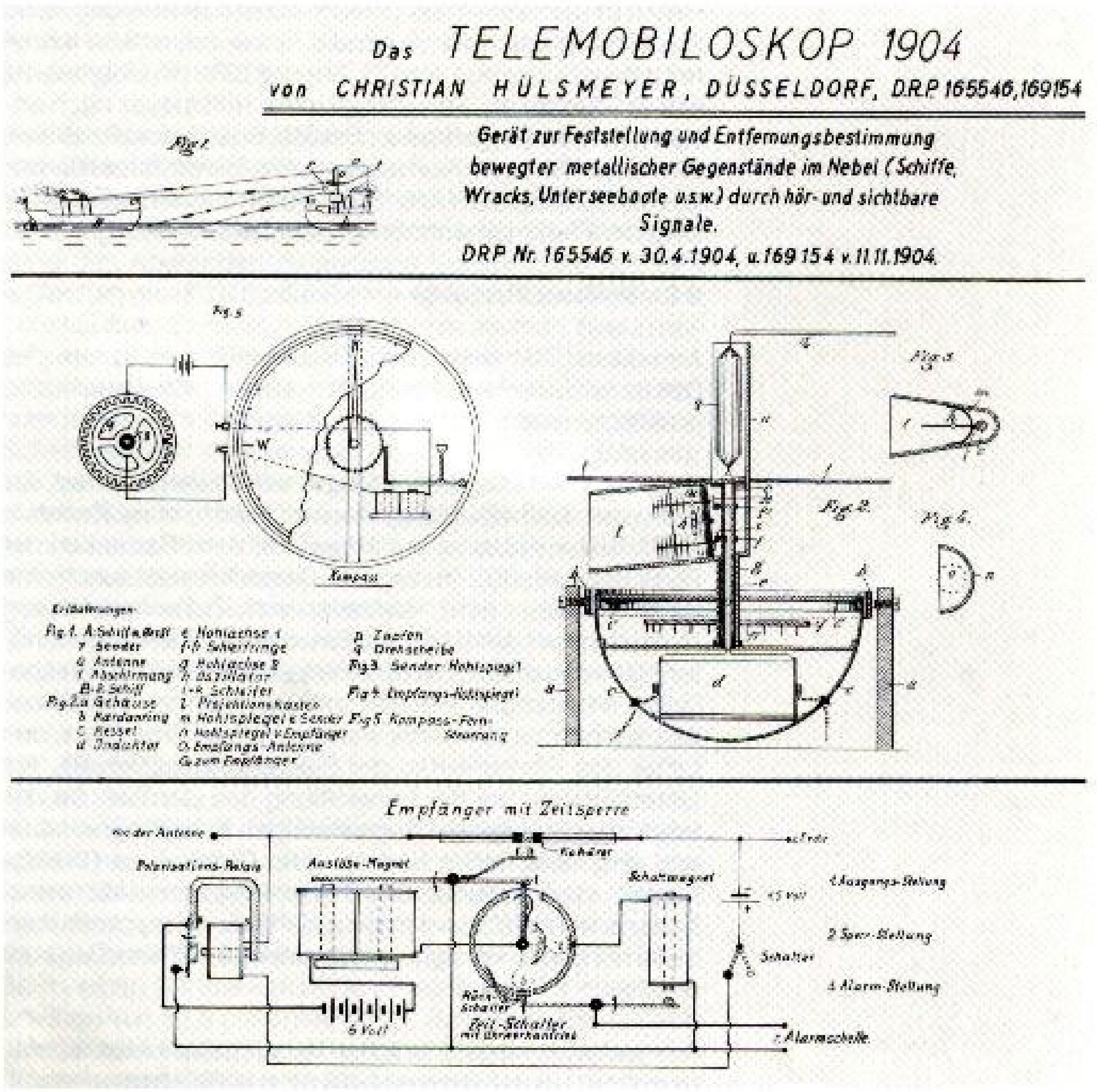
Asked about the potential applications of his discoveries, Hertz replied,

> *“Nothing, I guess.”*

1904 – Christian Hülsmeyer's *Telemobiloskop*

The first patented device using radio waves for detecting the presence of distant objects, with a range of ~3000m.

The first public demonstration of 18 May 1904 at the Hohenzollern Bridge, Cologne. The device was used to detect a ship in the Rhine.



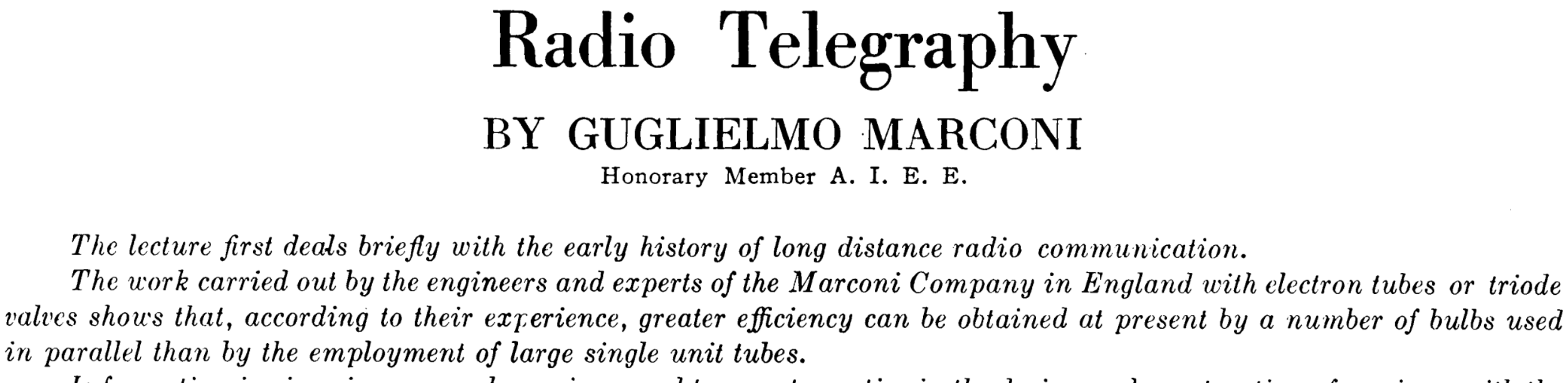
Das TELEMOBILOSKOP 1904

von CHRISTIAN HÜLSMEYER, DÜSSELDORF, D.R.P. 165546, 169154

Gerät zur Feststellung und Entfernungsbestimmung bewegter metallischer Gegenstände im Nebel (Schiffe, Wracks, Unterseeboote u.s.w.) durch hör- und sichtbare Signale.

DRP Nr. 165546 v. 30.4.1904, u. 169154 v. 11.11.1904.

1922 – Marconi Conceptualizes RADAR



Radio Telegraphy

BY GUGLIELMO MARCONI

Honorary Member A. I. E. E.

The lecture first deals briefly with the early history of long distance radio communication.

The work carried out by the engineers and experts of the Marconi Company in England with electron tubes or triode valves shows that, according to their experience, greater efficiency can be obtained at present by a number of bulbs used in parallel than by the employment of large single unit tubes.

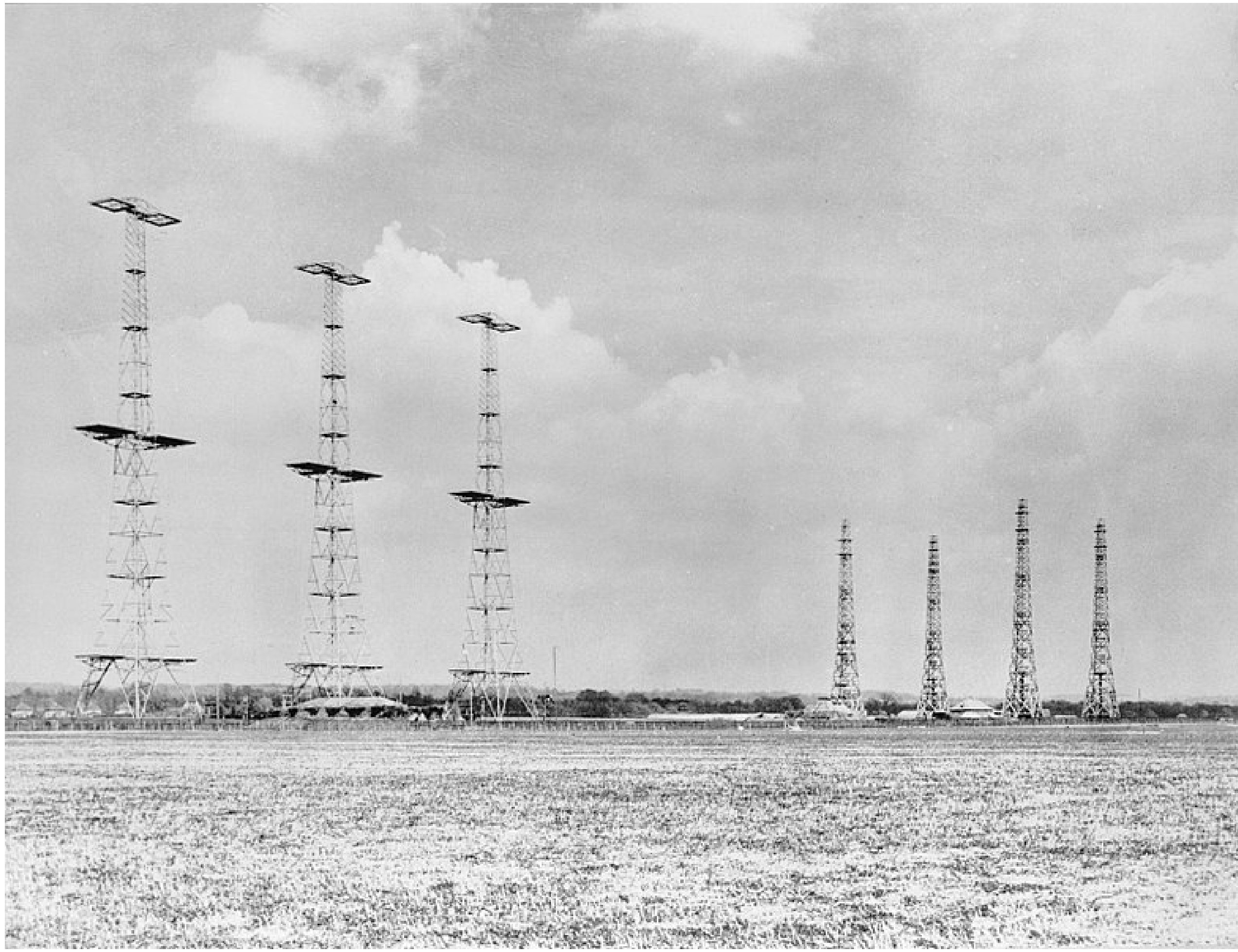
It seems to me that it should be possible to design apparatus by means of which a ship could radiate or project a divergent beam of these rays in any desired direction, which rays, if coming across a metallic object, such as another steamer or ship, would be reflected back to a receiver screened from the local transmitter on the sending ship, and thereby immediately reveal the presence and bearing of the other ship in fog or thick weather.

Marconi, G. (1922). Radio telegraphy. *Journal of the American Institute of Electrical Engineers, 41*(8), 561-570.

1935 – First *Radio Detection and Ranging* (RADAR)

The UK military, *fearing a German death ray*, invests in radio for long range detection:

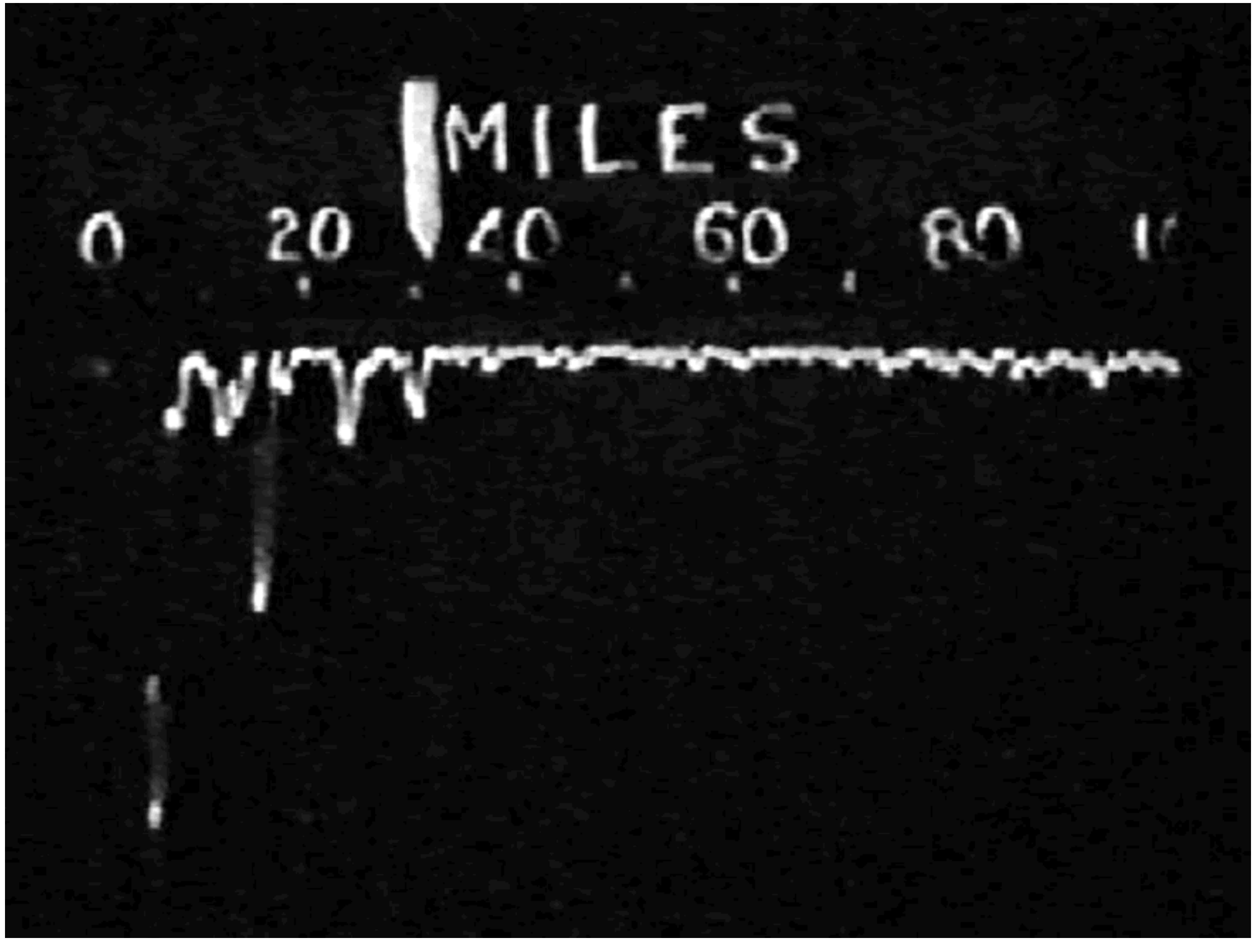
- 1935 commercial shortwave radio hardware is used to measure the angle and range of aircraft
- 1936 a production version with range of ~100 miles
- 1937 first five stations covering London
- 1938-1940 Coverage for all of Eastern UK by *Chain Home* system



https://en.wikipedia.org/wiki/Chain_Home#/media/File:Chain_Home_radar_installation_at_Poling,_Sussex,_1945._CH15173.jpg

1936 – Chain Home *A-scope*

- The original radar display, the A-scope, shows only the range, not the direction, to targets.
- The returns caused the spot to be deflected downward drawing vertical lines on the tube, known as a “blip” or “pip”.
- This image shows several blips between 15 and 30 miles from the station. The large blip on the far left is the leftover signal from the radar's own transmitter; targets in this area could not be seen.
- Operators could use multiple antennas to determine altitude.



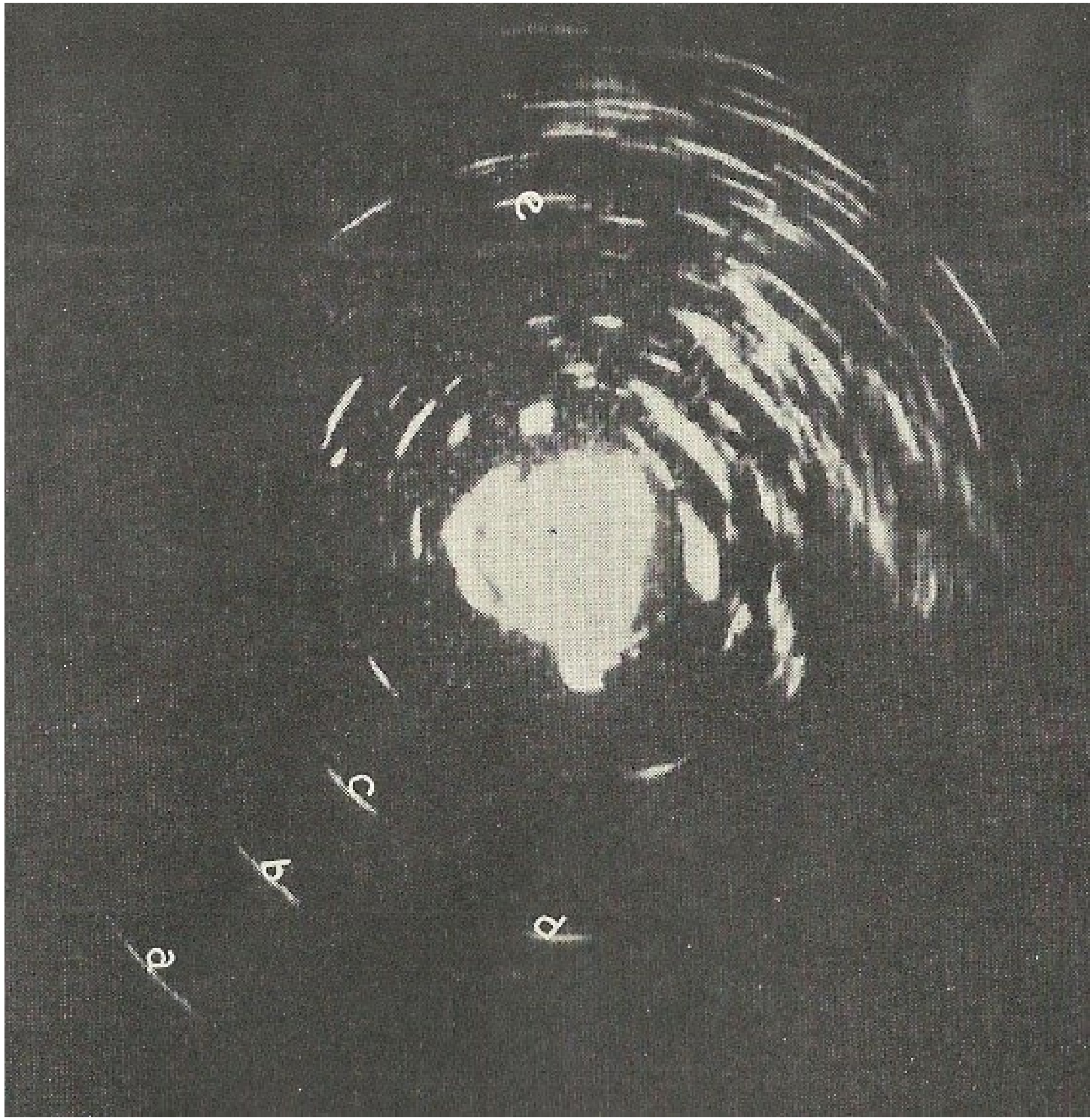
https://en.wikipedia.org/wiki/Radar_display#/media/File:Chain_Home_screen_shot_-NEDAD.2013.047.058A.jpg

1940 – Plan Position Indicator (PPI)

Blips marked a, b, c, d, are target aircraft; the rest is ground clutter.

Difficult problem!

How to characterize the performance of different RADAR operators?



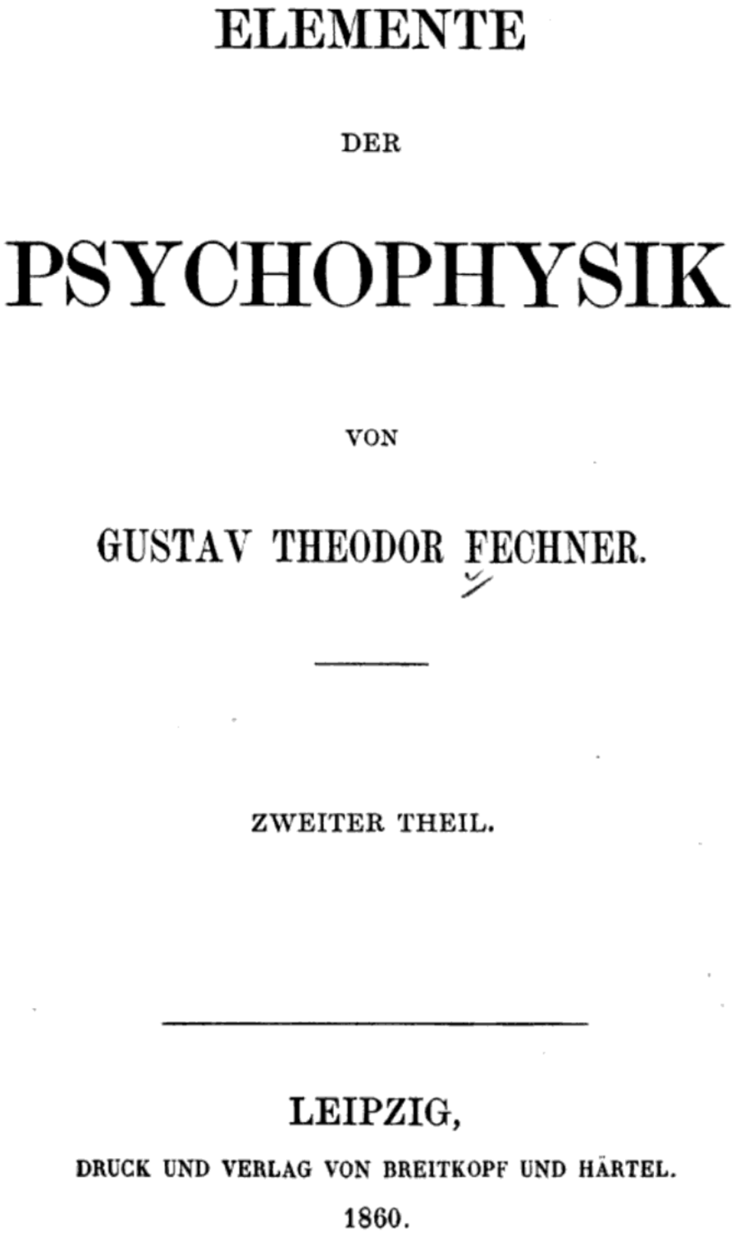
https://www.technologystories.org/dreams-and-visions/

1860 – Fechner invents Psychophysics

Psychophysics is the scientific study of the relation between stimulus and sensation, pioneered by Fechner.

Fechner found that seeing a faint stimulus or hearing a faint sound is probabilistic:

- A ‘threshold’ was a statistical fiction, an arbitrary point on the continuous increasing function relating stimulus intensity to the proportion of ‘yes’ responses.
- That function increased from zero with no stimulus to 100% with a sufficiently large one.



ELEMENTE

DER

PSYCHOPHYSIK

VON

GUSTAV THEODOR FECHNER.

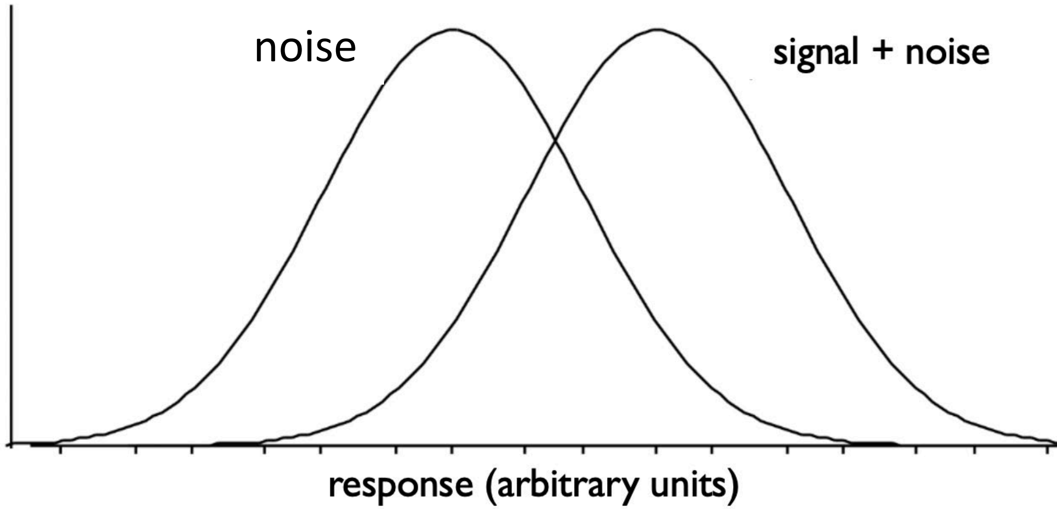
ZWEITER THEIL.

LEIPZIG,
DRUCK UND VERLAG VON BREITKOPF UND HÄRTEL.
1860.

1940s – Receiver Operating Characteristic (ROC)

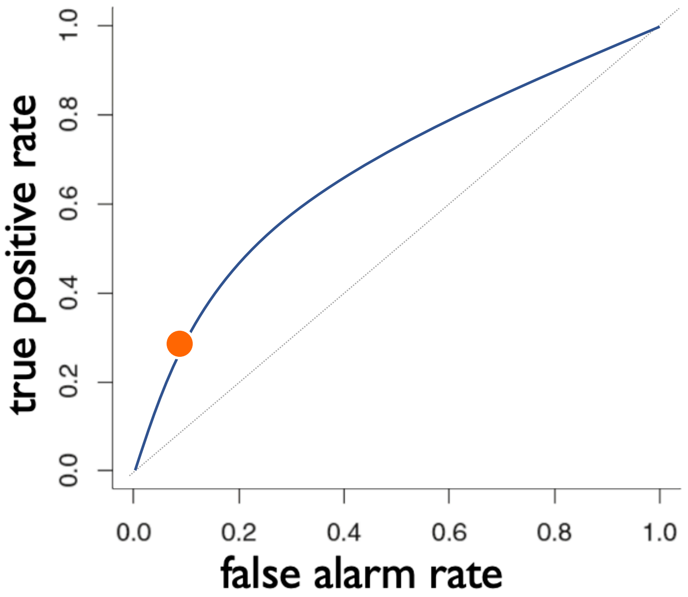
An early RADAR console was a prototypical detection problem:

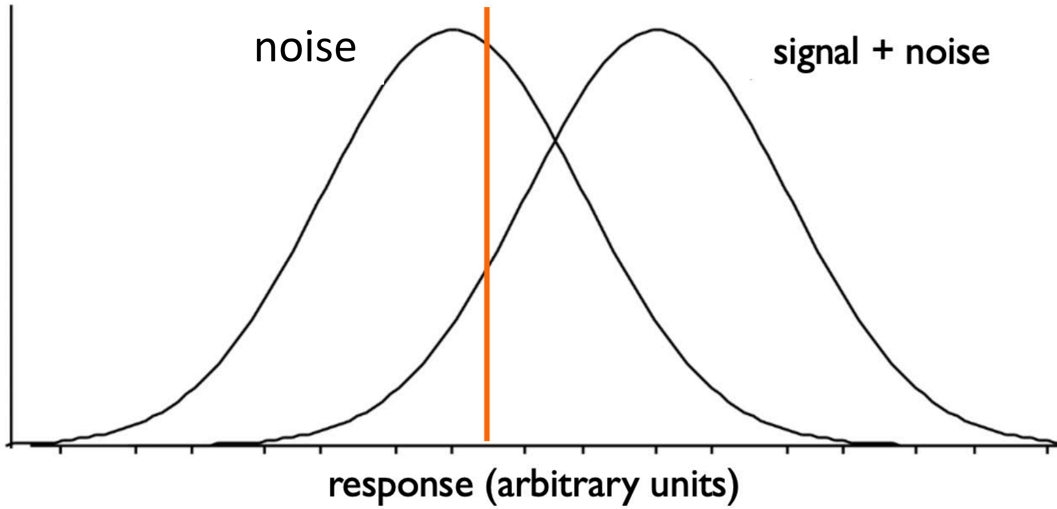
- The display was covered with ‘snow,’ reflections from atmospheric features, and the results of random activity in the vacuum tube circuits processing signals passed from an antenna.
- Buried in this visual noise might have been a reflection from an aircraft.
- Noise might mock the signature of an aircraft, and noise might mask an important signal.



1940s – Receiver Operating Characteristic (ROC)

Researchers built on the work of Fechner to characterize the tradeoff between false alarms and true detections in a Receiver Operating Characteristic (ROC) curve.

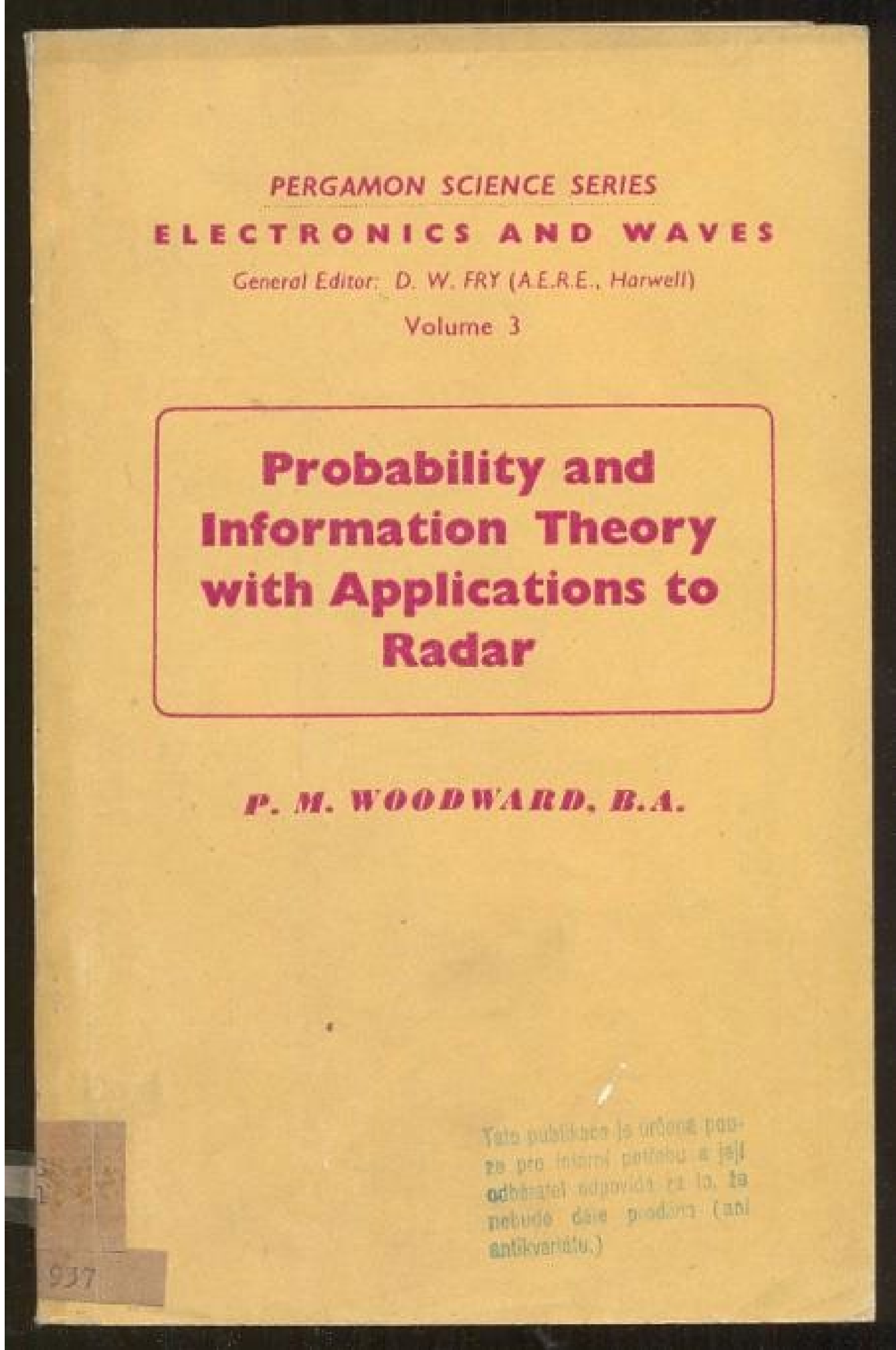




1950s – Signal Detection Theory

After the war, the tight security around theoretical work in the field eased off and researchers began to published work describing ways to analyze faint, noise-contaminated signals.

- Woodward, P. M. (1953). *Probability and information theory with applications to radar.* London: Pergamon Press.



PERGAMON SCIENCE SERIES

ELECTRONICS AND WAVES

General Editor: D. W. FRY (A.E.R.E., Harwell)

Volume 3

Probability and Information Theory with Applications to Radar

P. M. WOODWARD, B.A.

1950s – Signal Detection Theory

First *published* ROC curve,

Peterson, W., Birdsall, T., Fox, W. (1954). The theory of signal detectability, *Transactions of the IRE Professional Group on Information Theory*, 4, 4, pp. 171 - 212.

From the abstract,

- An optimum observer required to give a yes or no answer simply chooses an operating level and concludes that the receiver input arose from signal plus noise only when this level is exceeded by the output of his likelihood ratio receiver. Associated with each such operating level are conditional probabilities that the answer is a false alarm and the conditional probability of detection. Graphs of these quantities called receiver operating characteristic, or ROC, curves are convenient for evaluating a receiver. If the detection problem is changed by varying, for example, the signal power, then a family of ROC curves is generated. Such things as betting curves can easily be obtained from such a family.

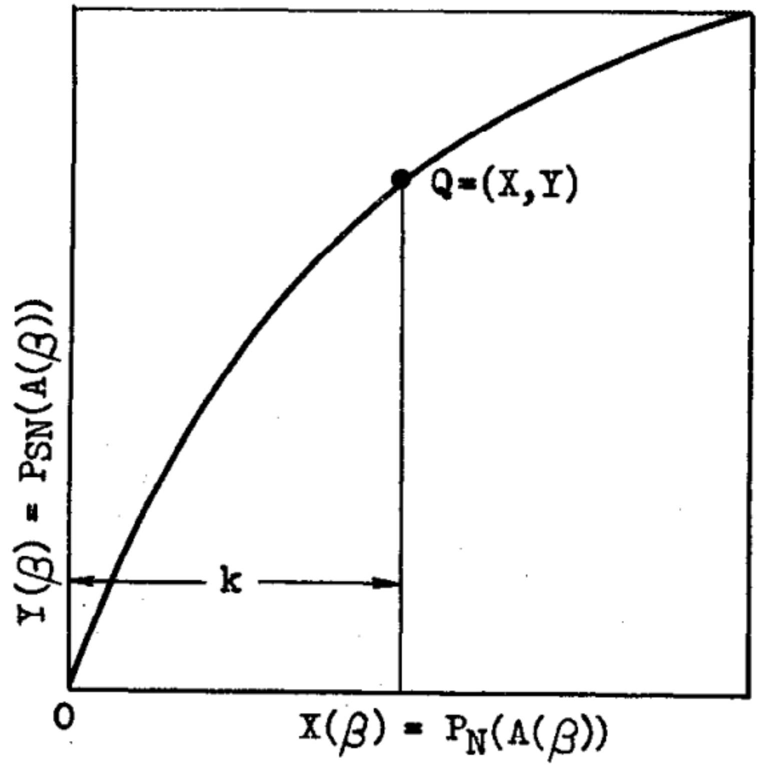
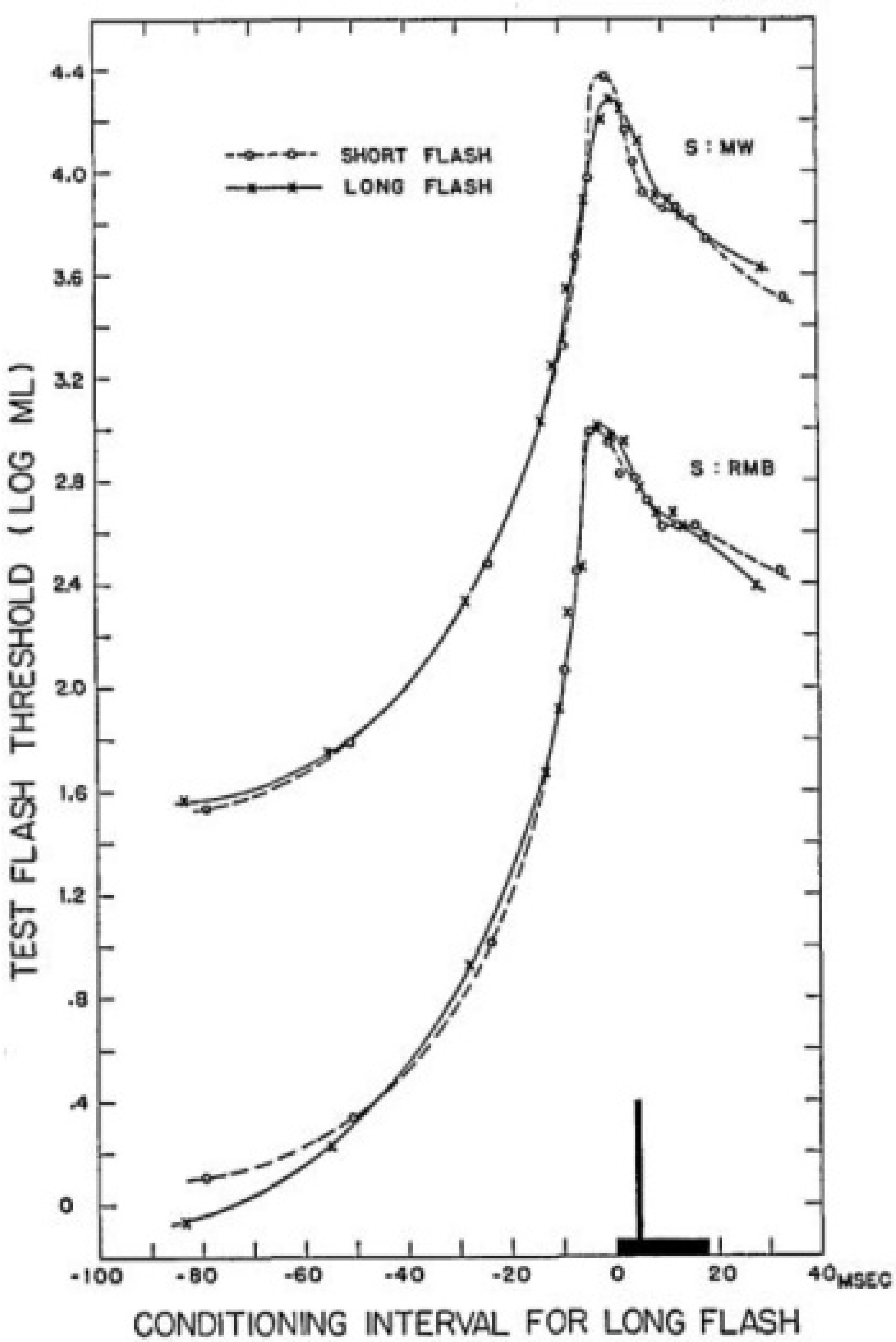


FIG. 1. TYPICAL ROC CURVE

1960s – Visual Psychophysics

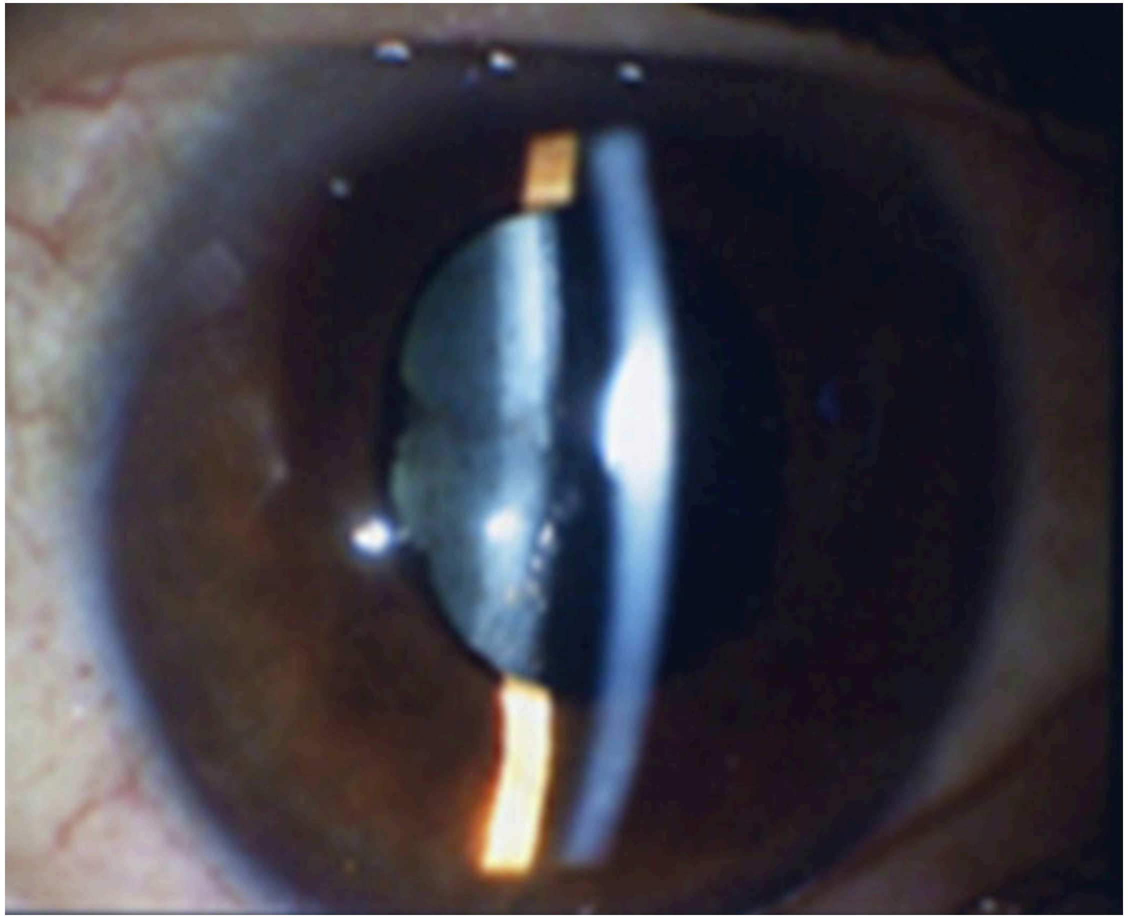
ROC analysis becomes a core part of experimental psychology and psychophysics, e.g.,

Boynton, R. M., & Siegfried, J. B. (1962). Psychophysical estimates of on-responses to brief light flashes. *JOSA*, *52*(6), 720-721.



1990s – Visual Psychophysics

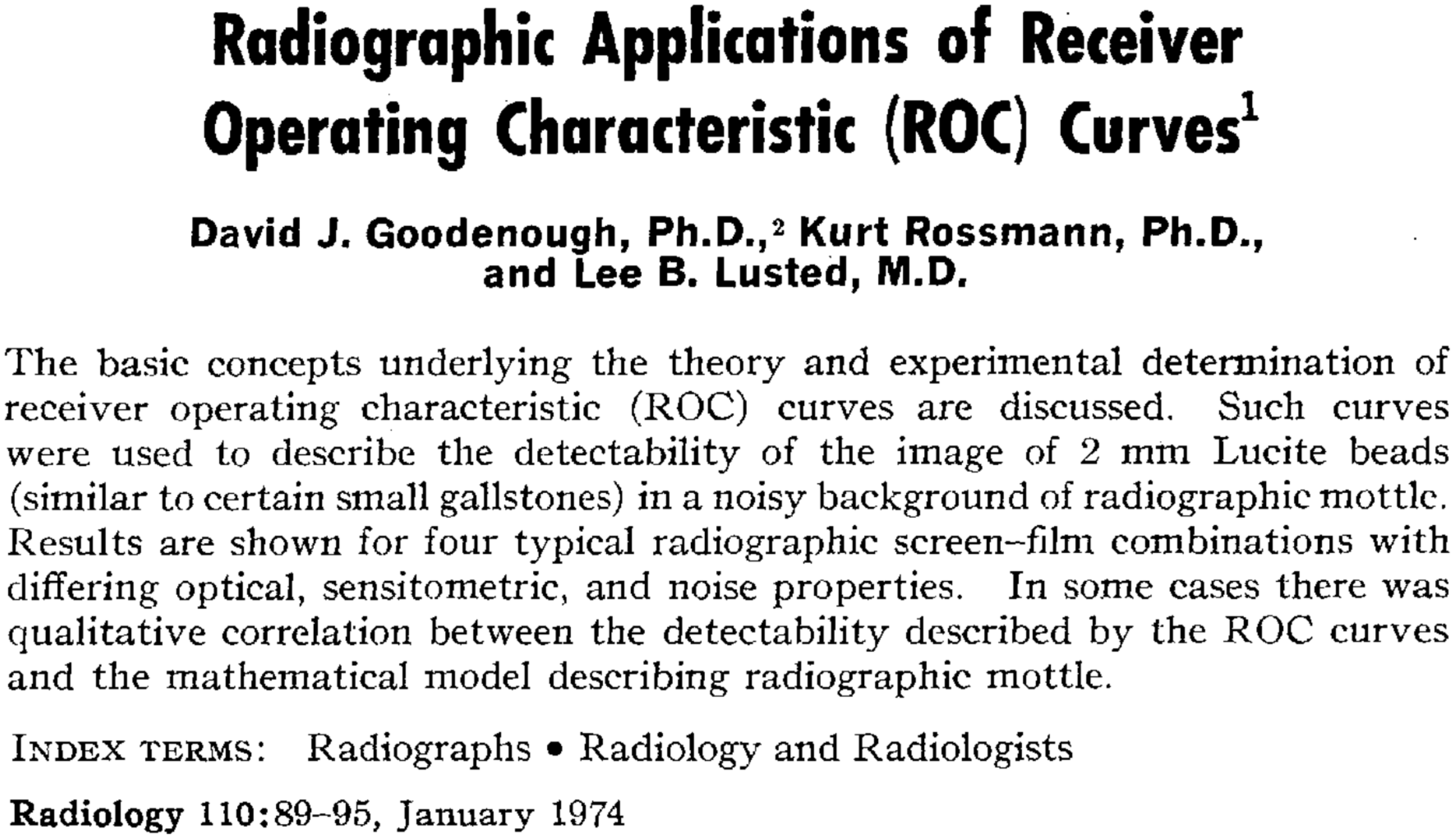
Kandel, G., **Eklund, N.**, and Schroeder, J. (1992). On the possibility of visually significant intraocular photoluminescence. *Advances in Color Vision Technical Digest Series (4)*. Optical Society of America: Washington, DC.



Abstract:

Photoluminescence of the human lens *in vitro* is well documented in the biochemistry literature (for a review, see Bloemendal, 1978). This photoluminescence has been independently confirmed in our laboratories in *in vivo* human lenses.

1974 – Early use in Radiology



Radiographic Applications of Receiver Operating Characteristic (ROC) Curves[1]

David J. Goodenough, Ph.D.,[2] Kurt Rossmann, Ph.D.,
and Lee B. Lusted, M.D.

The basic concepts underlying the theory and experimental determination of receiver operating characteristic (ROC) curves are discussed. Such curves were used to describe the detectability of the image of 2 mm Lucite beads (similar to certain small gallstones) in a noisy background of radiographic mottle. Results are shown for four typical radiographic screen–film combinations with differing optical, sensitometric, and noise properties. In some cases there was qualitative correlation between the detectability described by the ROC curves and the mathematical model describing radiographic mottle.

INDEX TERMS: Radiographs • Radiology and Radiologists

Radiology 110:89–95, January 1974

1989 – First use in Machine Learning

Spackman, K. A. (1989, January). Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the sixth international workshop on Machine learning* (pp. 160-163). Morgan Kaufmann.

Abstract:

This paper describes the use of signal detection theory as a tool for evaluating and comparing concept descriptions learned by inductive inference. We outline the use of ROC curves and describe the experience we have had in using these concepts for inductive learning using connectionist models, genetic search, and symbolic concept acquisition.

Inherent Properties of PHM Problems

Inherent Properties of PHM Problems

There are two properties that are central to *most* asset health management applications:

1. Faults are rare
   - If faults are common, the asset is poorly designed
   - Of necessity, faults are comparatively rare
2. Opportunities to evaluate are numerous
   - If an asset is worth developing an asset health management system for, then it is likely employed frequently.

As a result, PHM data are *extremely* imbalanced.

Example: Aviation Data

- Air turnback for a Boeing 777 costs ~$75K/incident in 2022 dollars
    - Not to mention 300+ unnerved passengers, and a likely borescope or unscheduled engine removal
- Common to average over one cycle/day over life of aircraft

Boeing, “Out of Service Costs”, Airline Fleet & Network Management, Issue 35, January - February, 2005, pg. 41

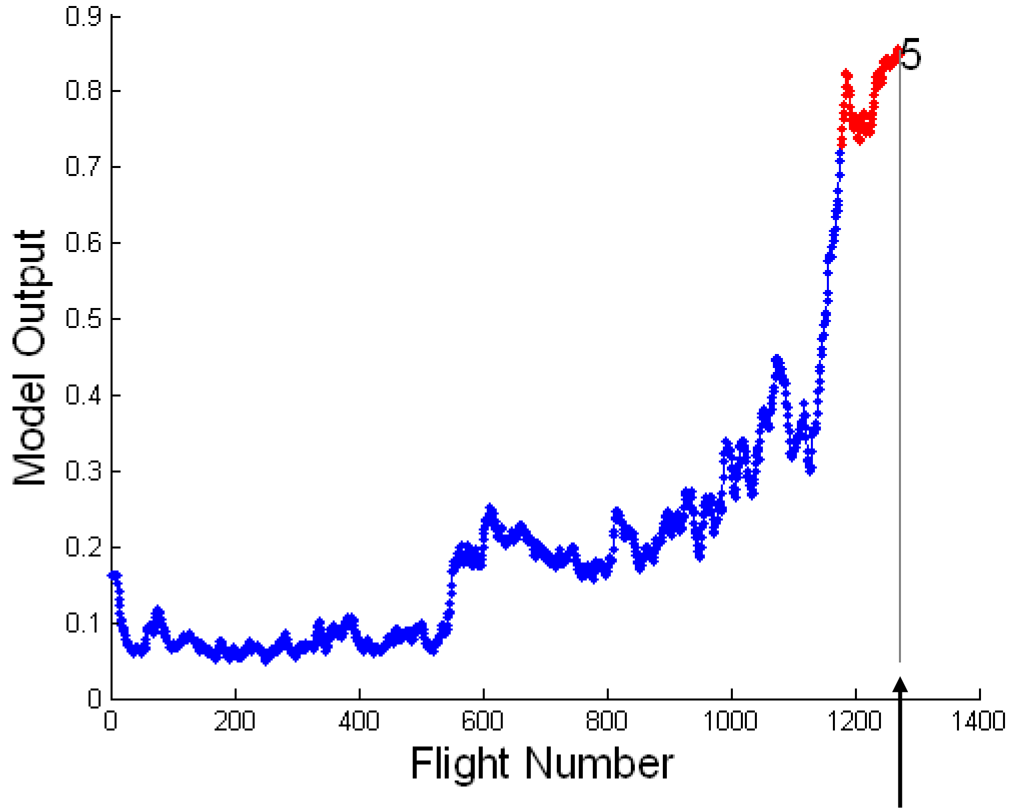
Example: Aviation Data

- 12 Engines, 14 faults
- 27 time series w/ alternative engines
  - 14 w/ fault occurring at the end (positive cases)
  - 13 w/o fault (negative cases)

- 25746 flights total
  - ~950 per time series (from 145 to 1731)

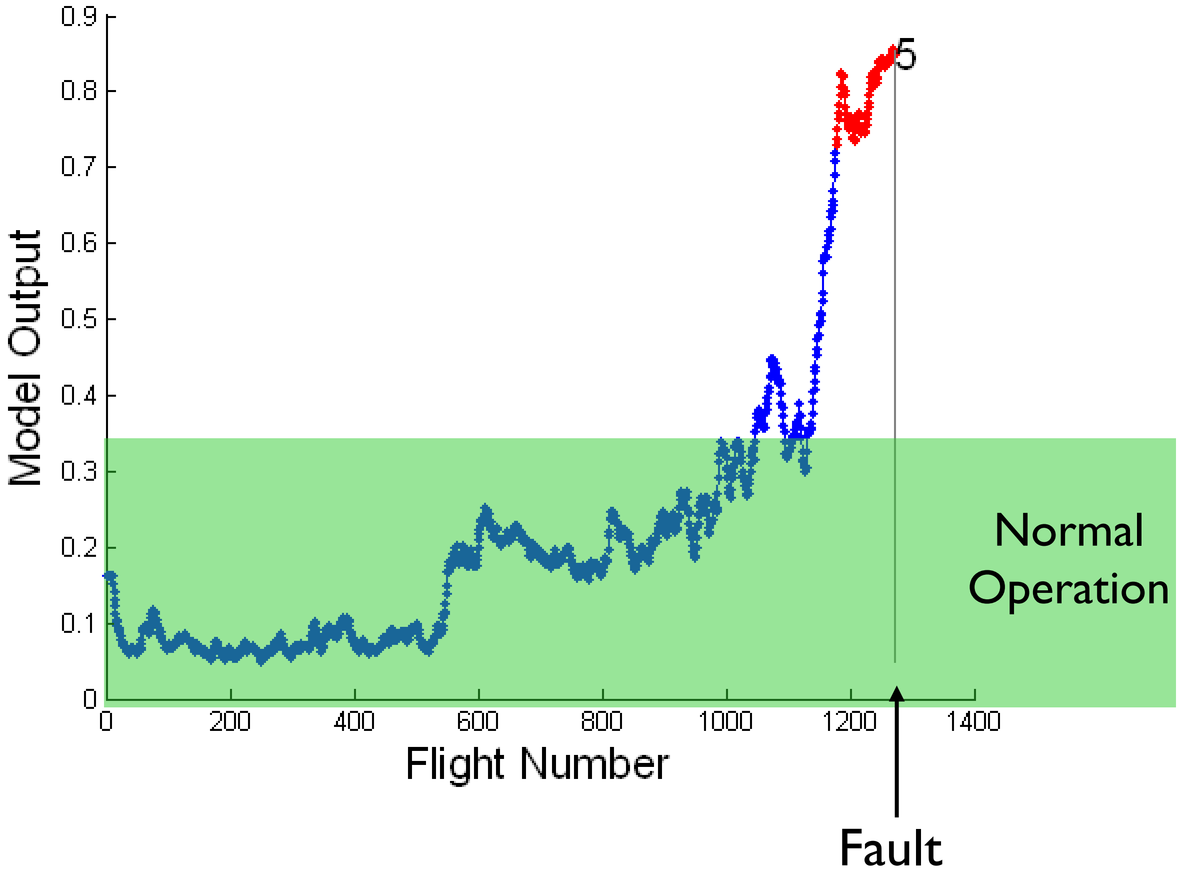
- Severely unbalanced – only 14 positive cases, and 25732 negative cases!
- Moreover, would be *even more* imbalanced if data were sampled uniformly from fleet.

Example Classifier Output

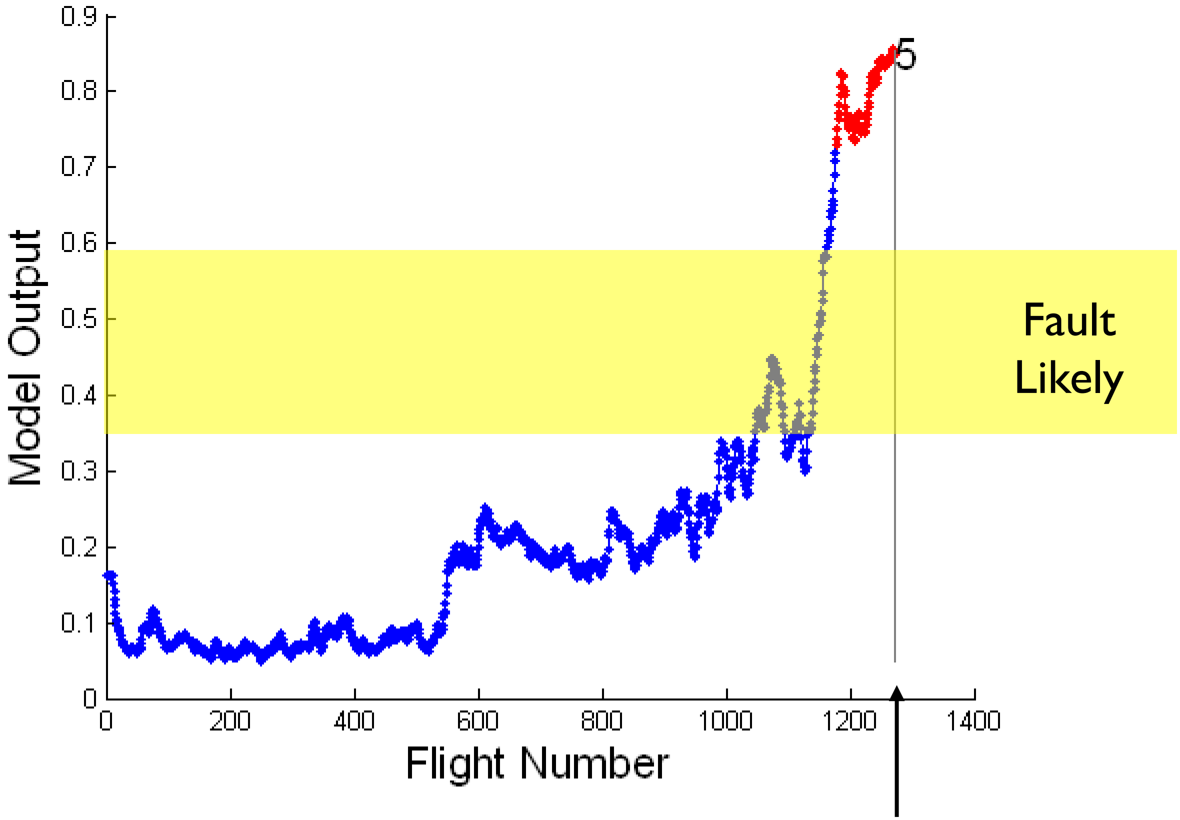


Fault

Example Classifier Output (cont.)



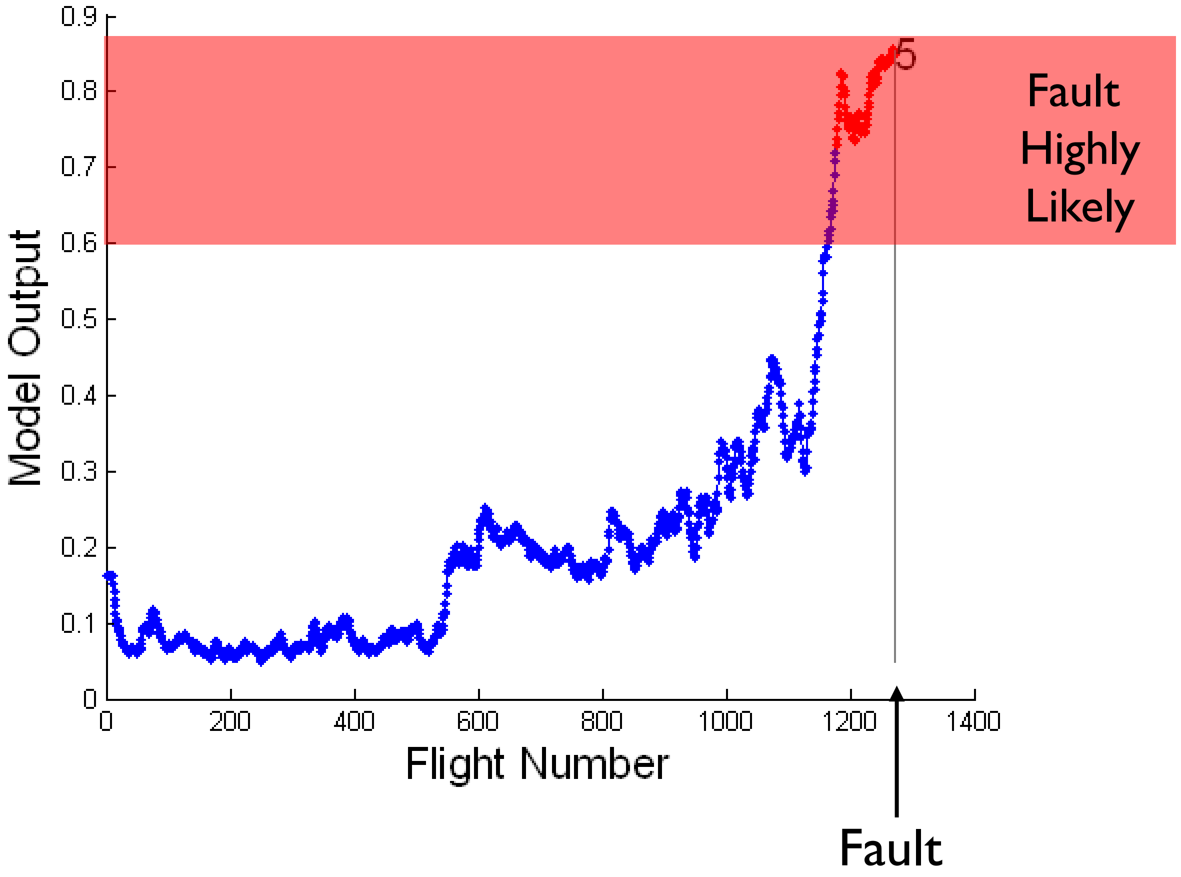
Example Classifier Output (cont.)



Fault
Likely

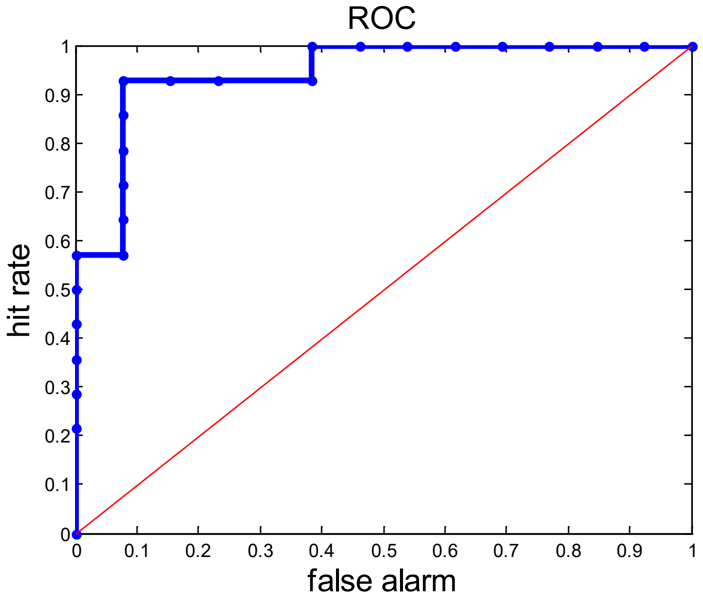
Fault

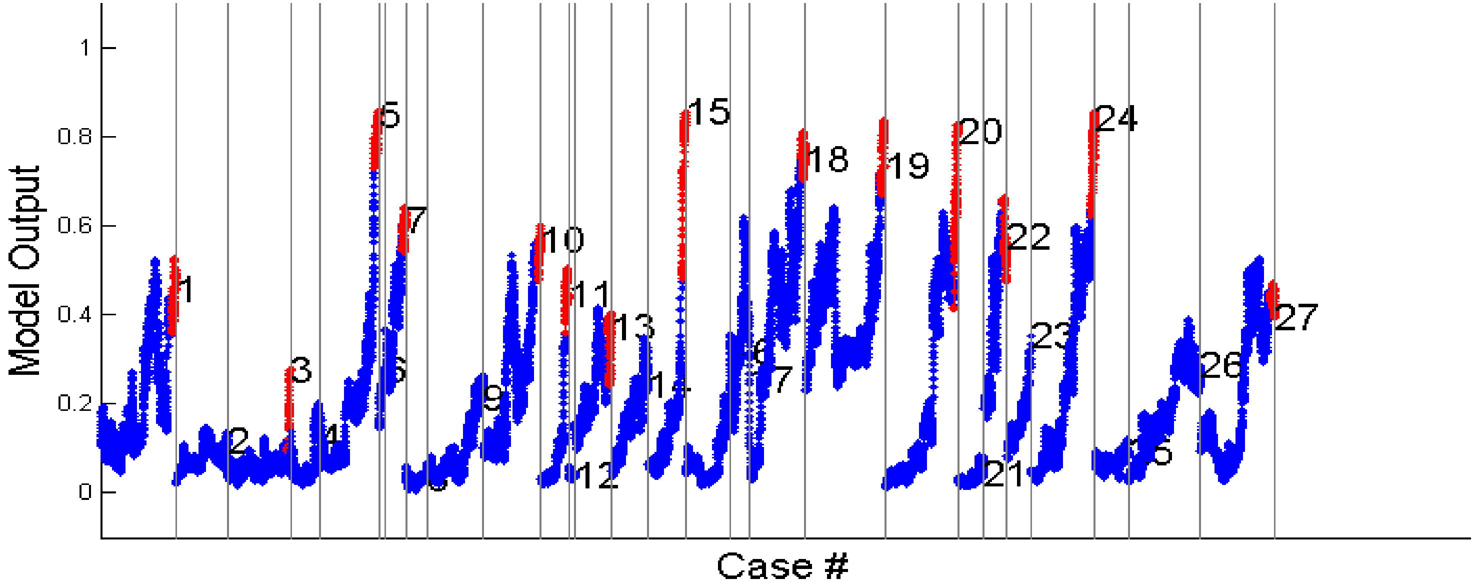
Example Classifier Output (cont.)



Threshold & ROC

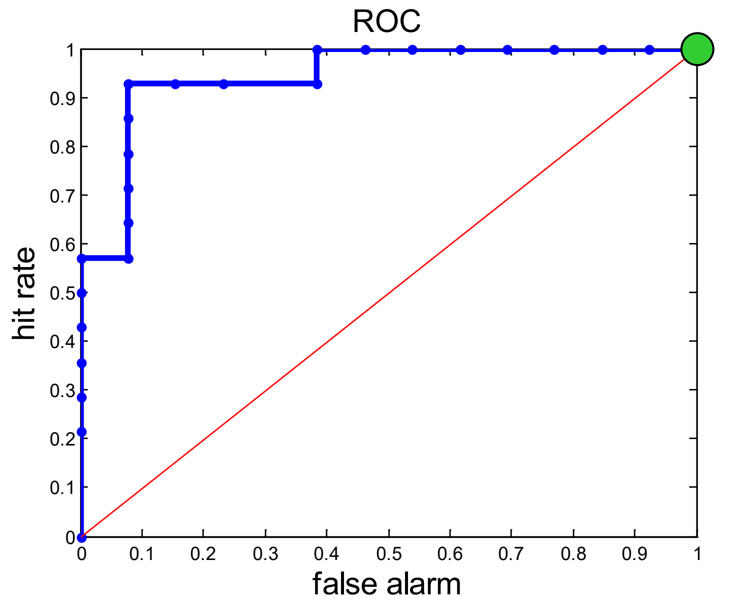
Different alerting thresholds on the model output generate different fault prediction performance, reflected as different points on the ROC curve

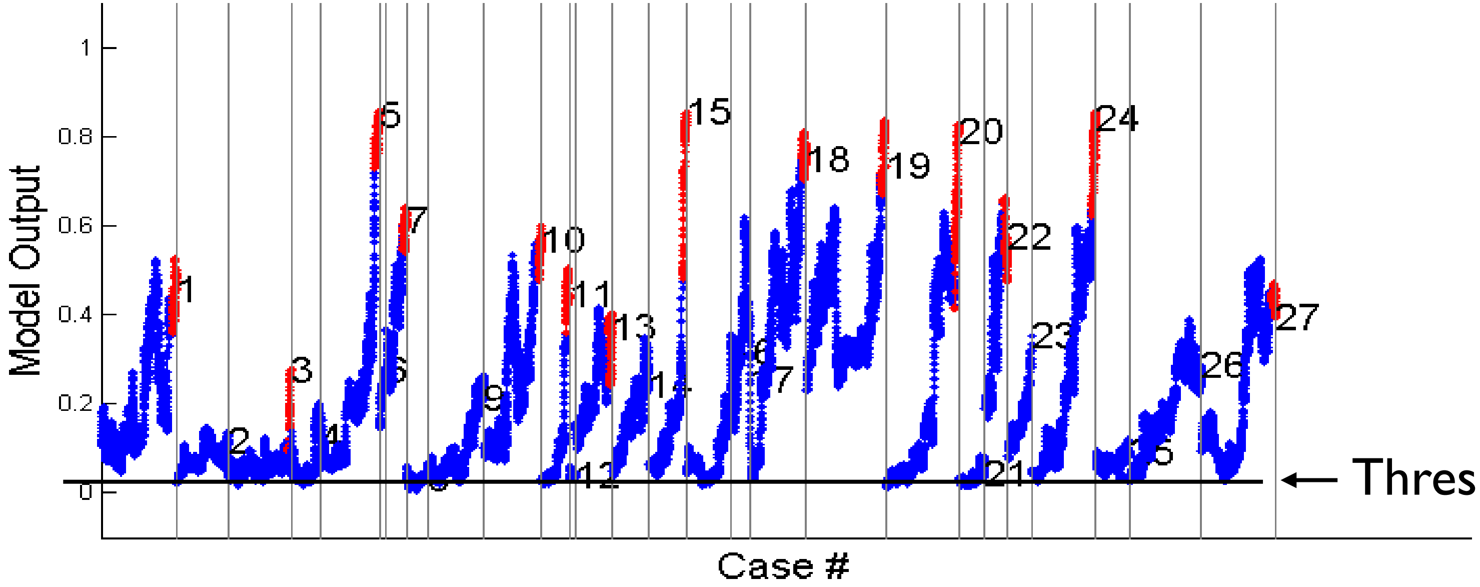




Threshold & ROC

Different alerting thresholds on the model output generate different fault prediction performance, reflected as different points on the ROC curve

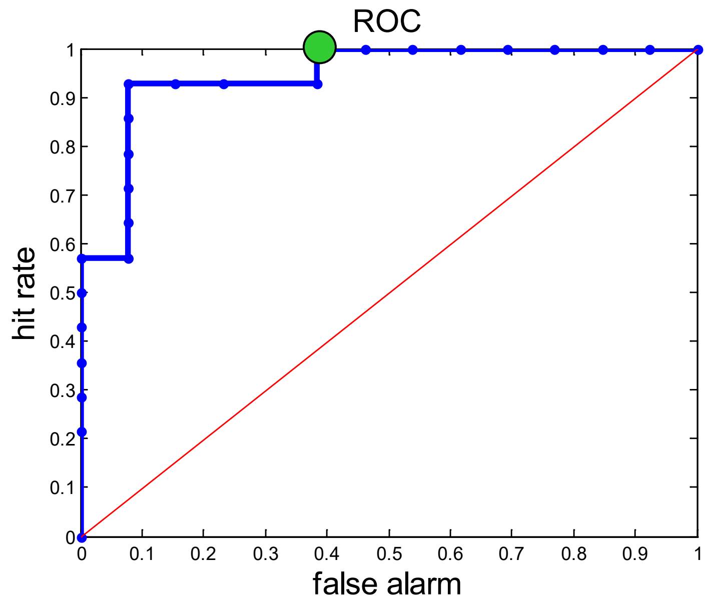


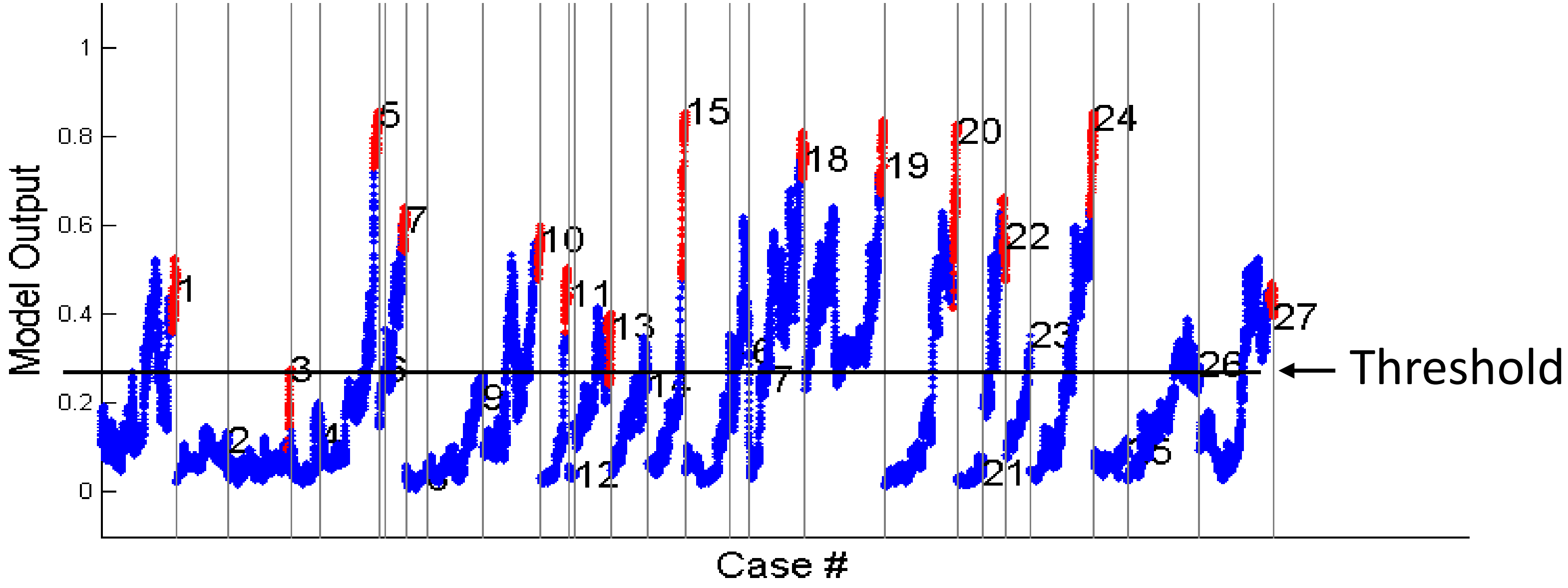


← Threshold

Threshold & ROC

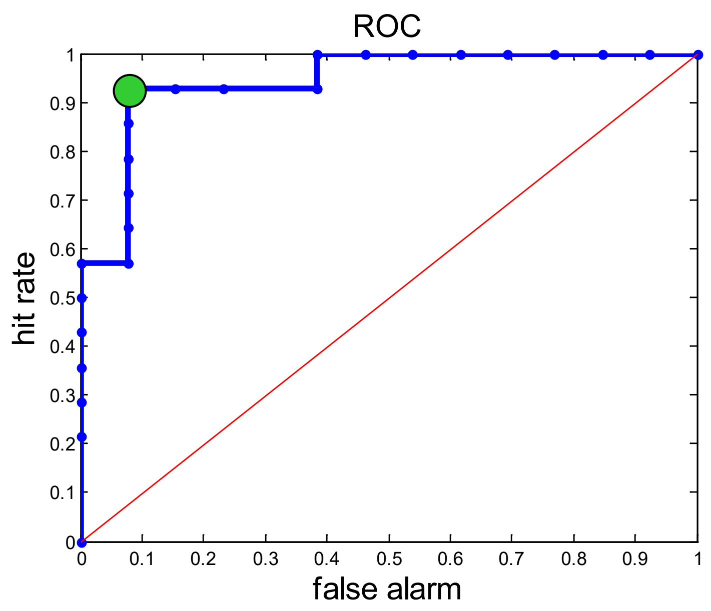
Different alerting thresholds on the model output generate different fault prediction performance, reflected as different points on the ROC curve

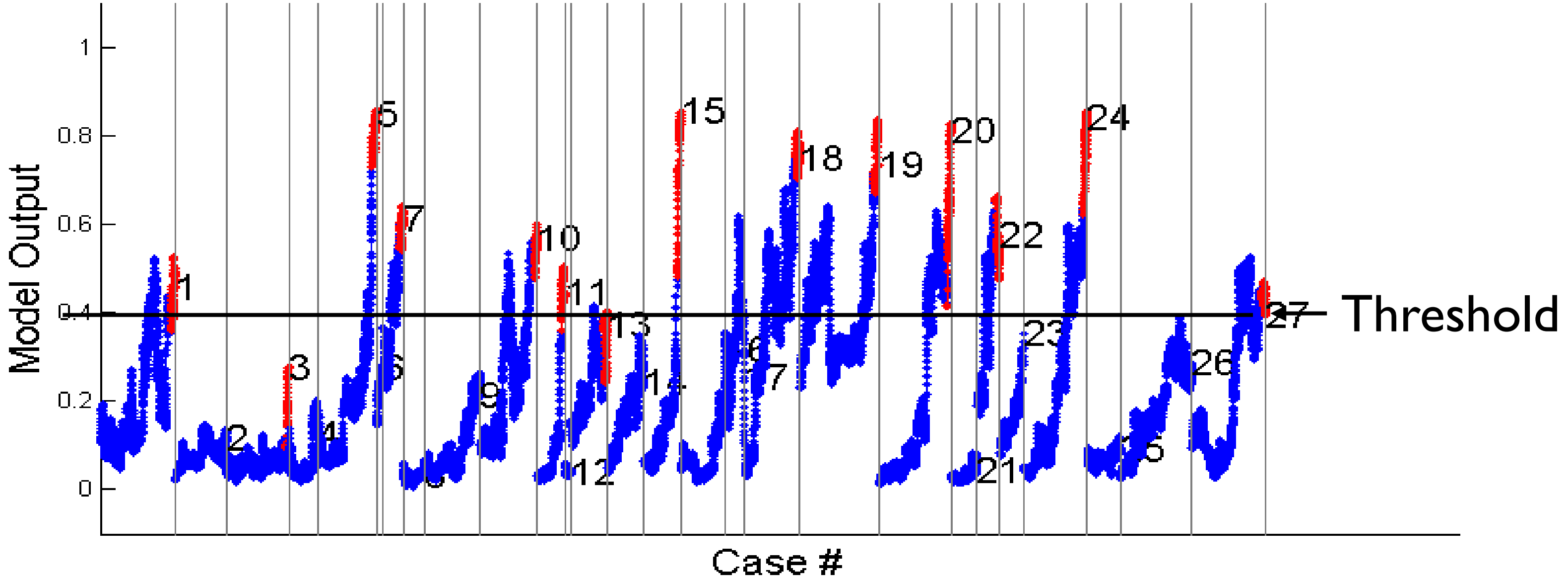




Threshold & ROC

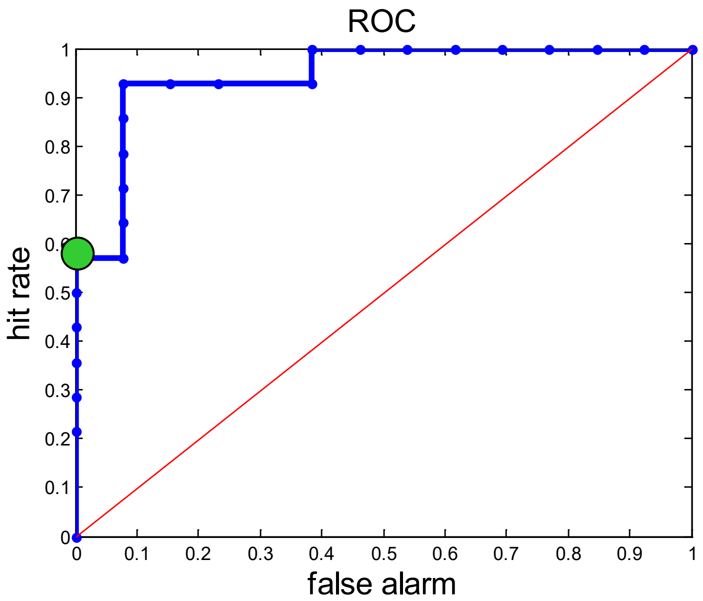
Different alerting thresholds on the model output generate different fault prediction performance, reflected as different points on the ROC curve

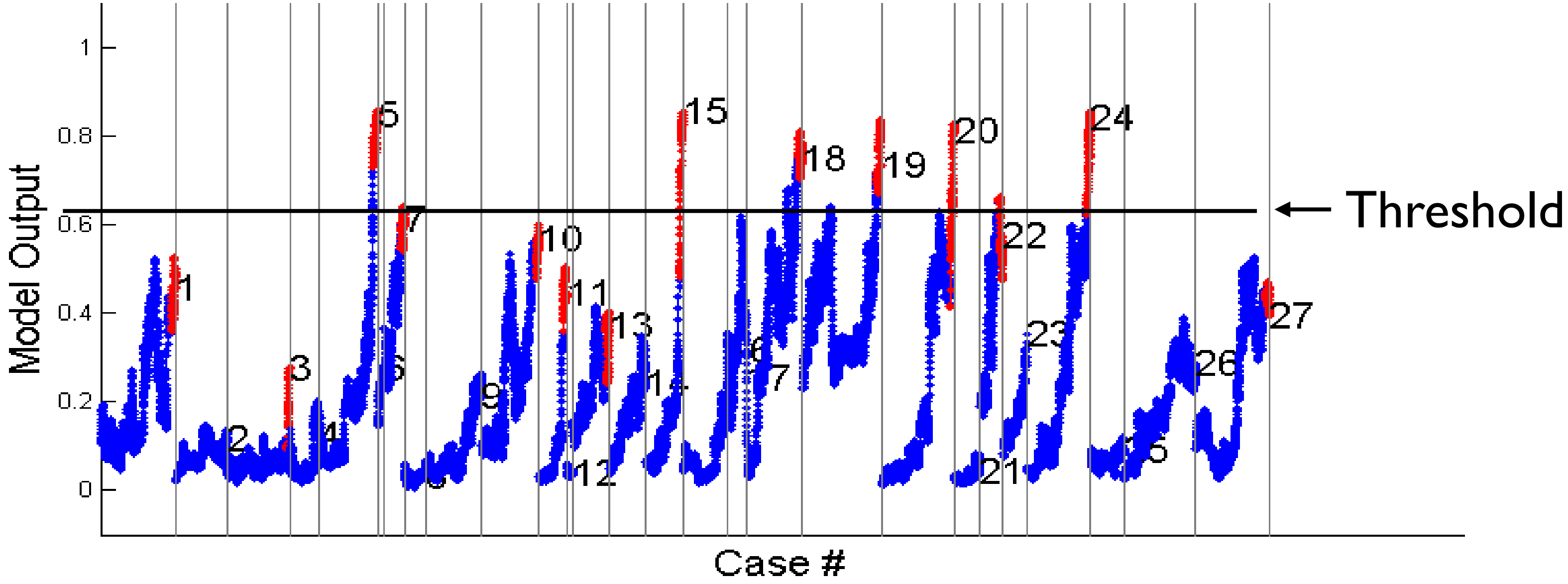




Threshold & ROC

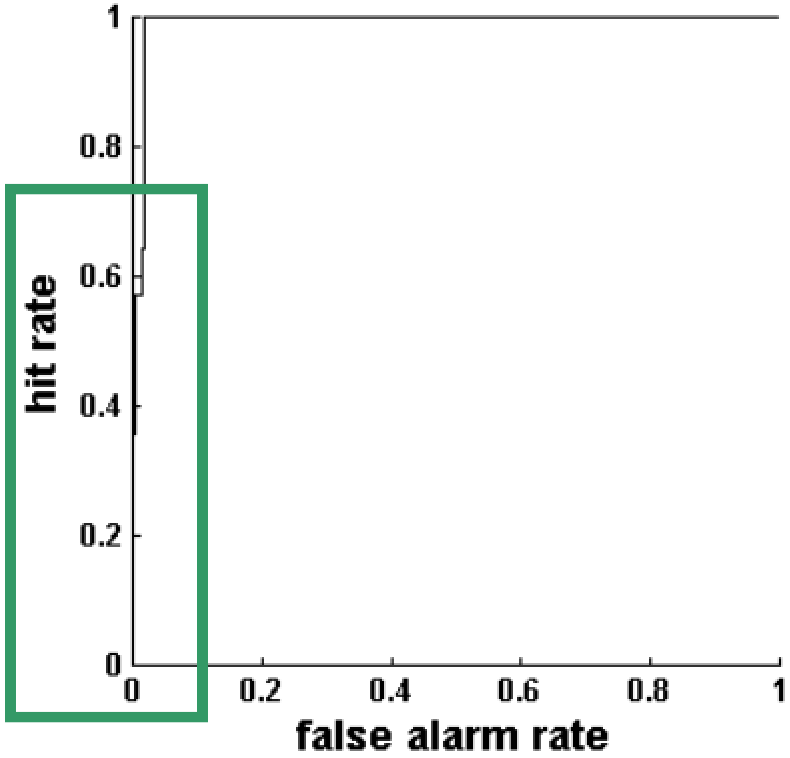
Different alerting thresholds on the model output generate different fault prediction performance, reflected as different points on the ROC curve





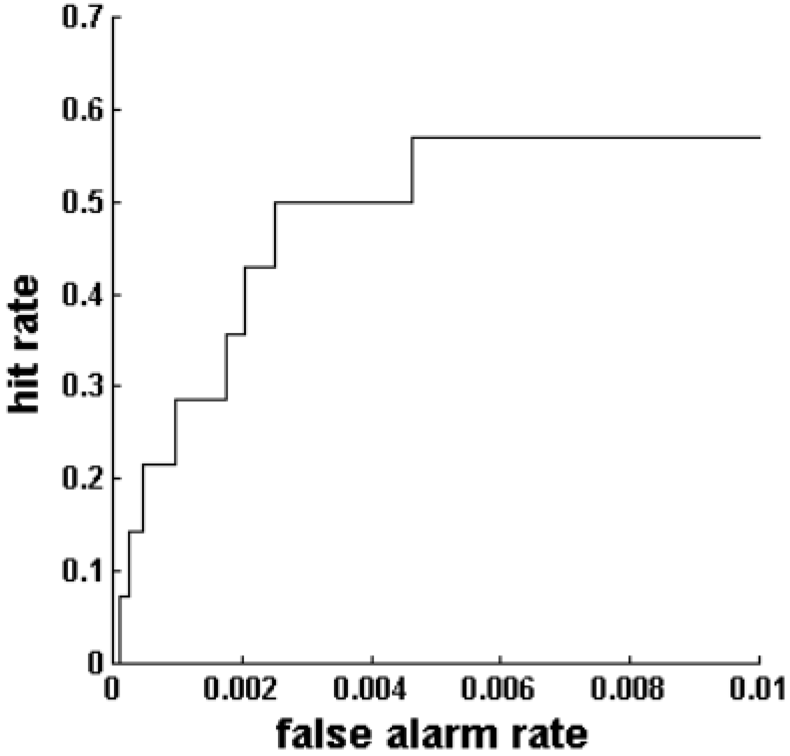
Results

Fantastic classifier performance! AUC 0.99



Results

Fantastic classifier performance! AUC 0.99

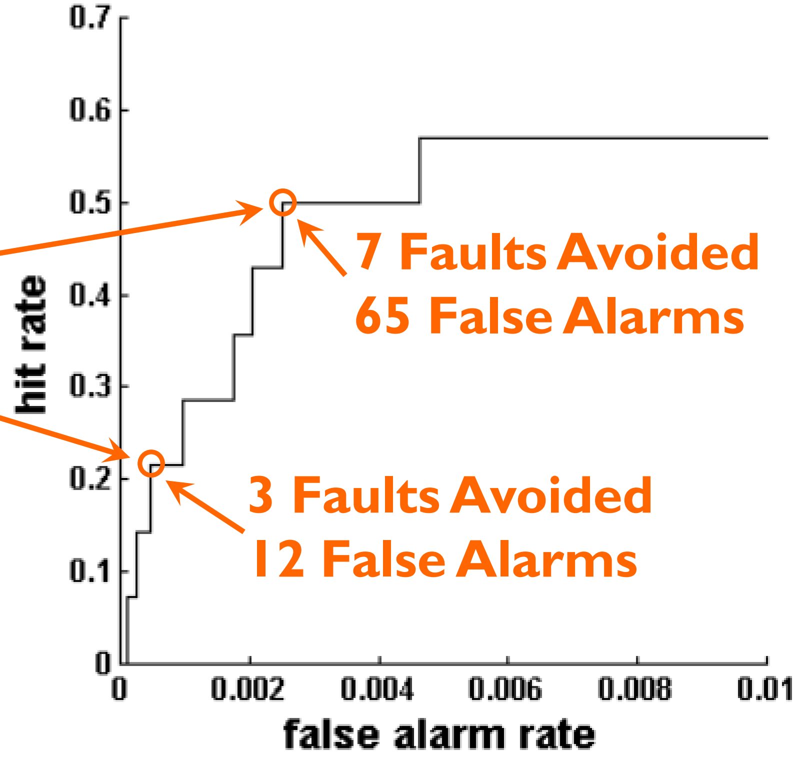


But real-world class imbalance requires higher performance…

Economic Analysis

Recall: $75k/incident

- At most $8076/false alarm @ 0.5 hit rate
- At most $18,750/false alarm @ 0.21



7 Faults Avoided
65 False Alarms

3 Faults Avoided
12 False Alarms

Take Aways

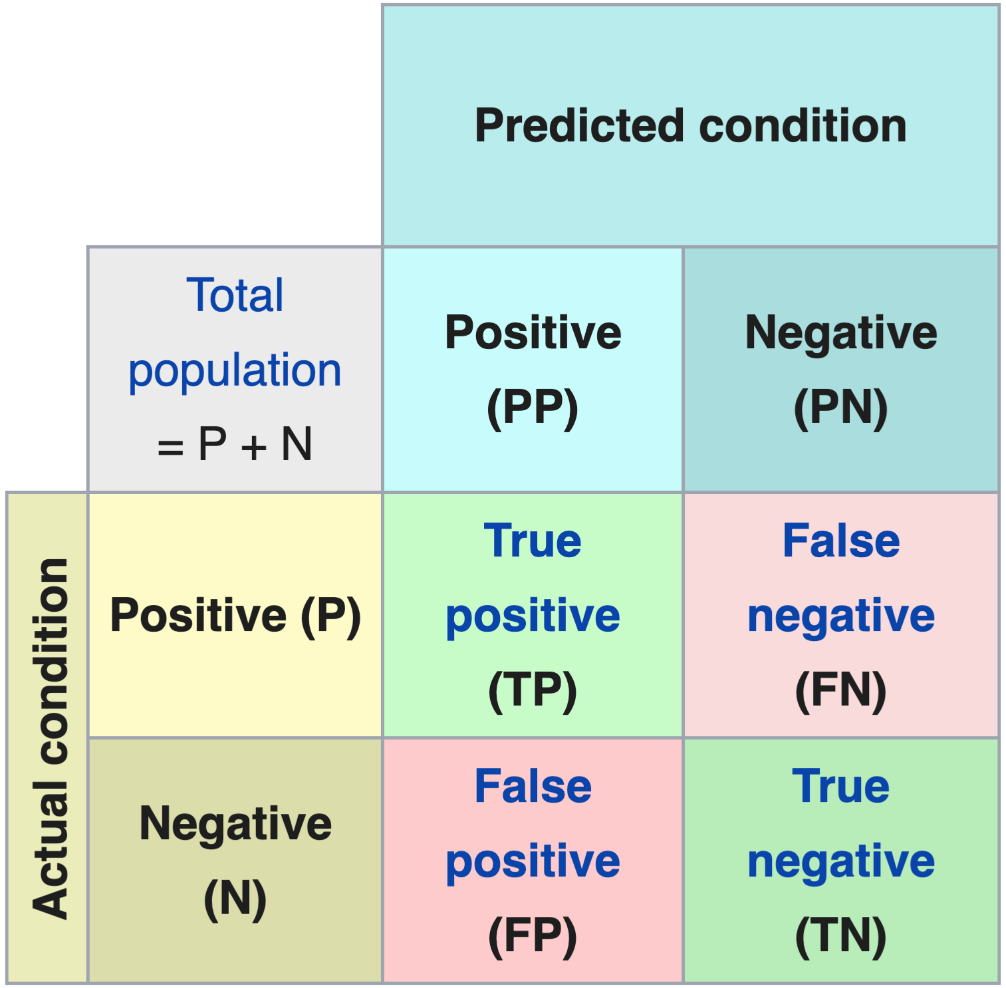
What does this example suggest?

1. The AUC statistic is not very meaningful for imbalanced data sets (i.e., most PHM data sets)
2. Because we make so many evaluations, and there are so few faults, we really only care about the extreme far left portion of the ROC curve.

A Close Look at ROC and Precision-Recall Curves

Recall, the Confusion Matrix

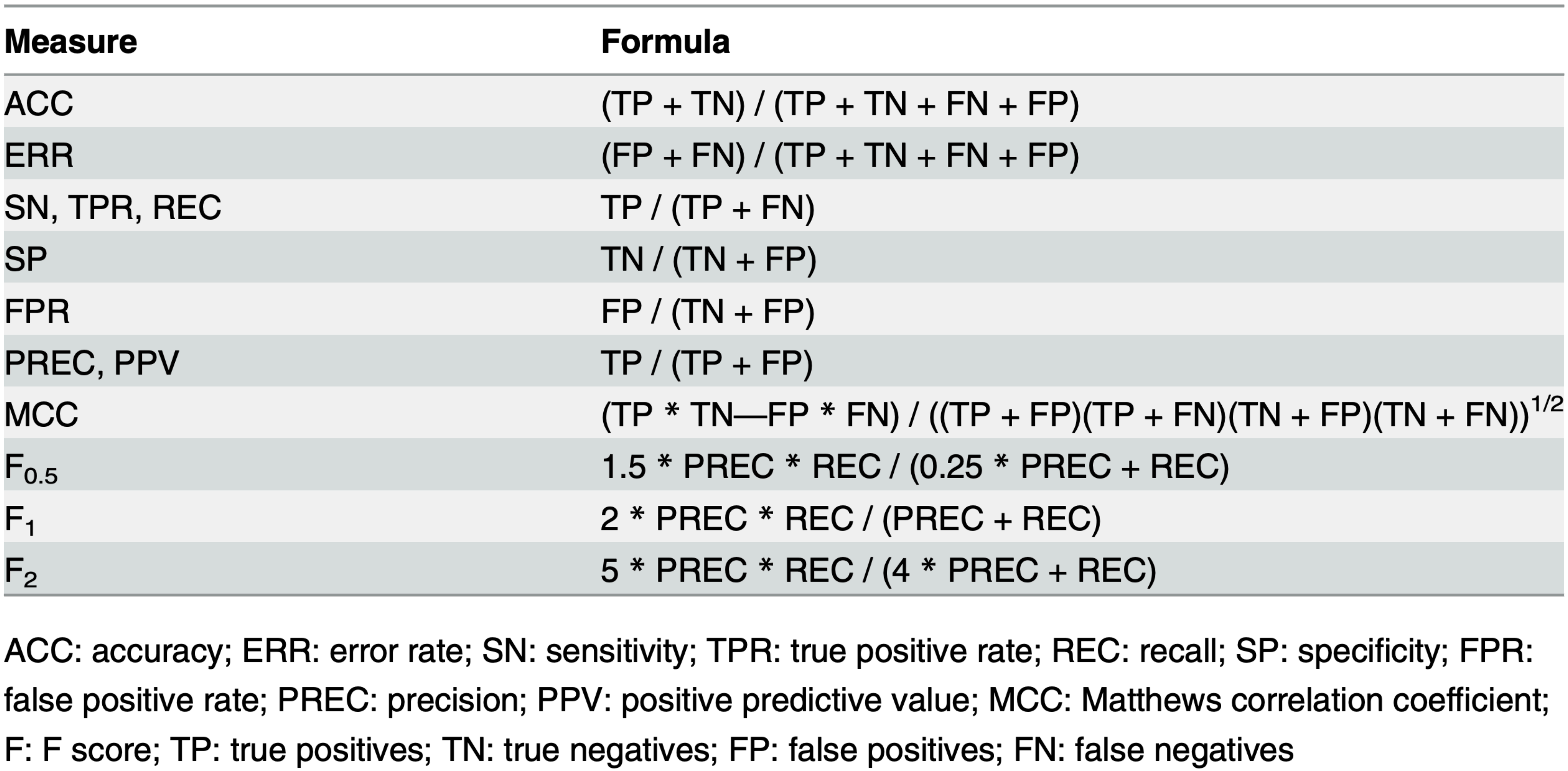
- A *binary classifier* produces output between a minimum and maximum value.
- When you pick a *classification threshold*:
  - cases >= threshold are classified as "positive"
  - cases < threshold are "negative"



		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

https://en.wikipedia.org/wiki/Confusion_matrix

Basic Evaluation Measures From a Confusion Matrix

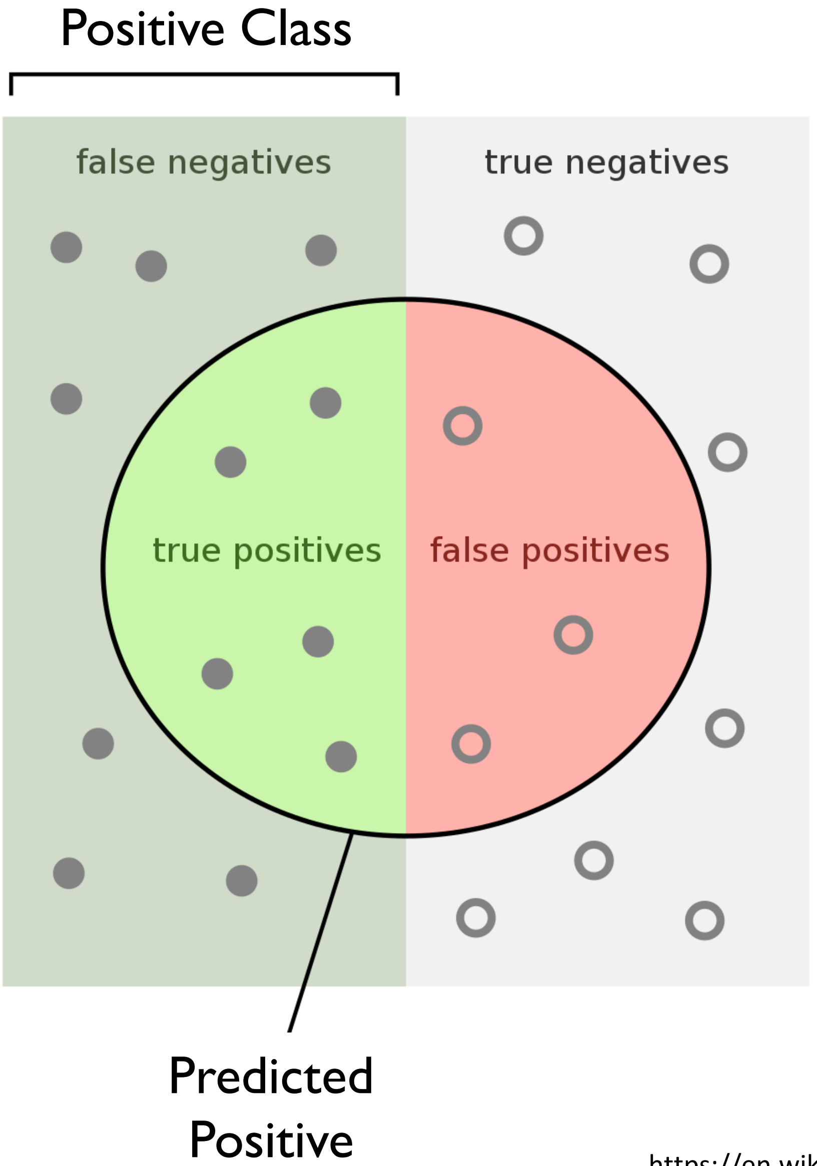


Measure	Formula
ACC	(TP + TN) / (TP + TN + FN + FP)
ERR	(FP + FN) / (TP + TN + FN + FP)
SN, TPR, REC	TP / (TP + FN)
SP	TN / (TN + FP)
FPR	FP / (TN + FP)
PREC, PPV	TP / (TP + FP)
MCC	(TP * TN—FP * FN) / ((TP + FP)(TP + FN)(TN + FP)(TN + FN))$^{1/2}$
$F_{0.5}$	1.5 * PREC * REC / (0.25 * PREC + REC)
F_1	2 * PREC * REC / (PREC + REC)
F_2	5 * PREC * REC / (4 * PREC + REC)

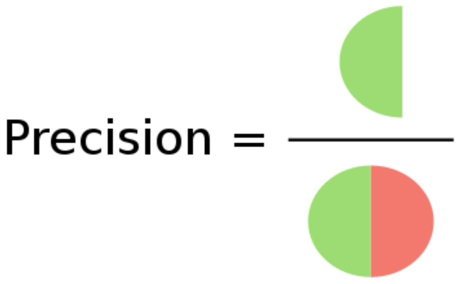
ACC: accuracy; ERR: error rate; SN: sensitivity; TPR: true positive rate; REC: recall; SP: specificity; FPR: false positive rate; PREC: precision; PPV: positive predictive value; MCC: Matthews correlation coefficient; F: F score; TP: true positives; TN: true negatives; FP: false positives; FN: false negatives

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

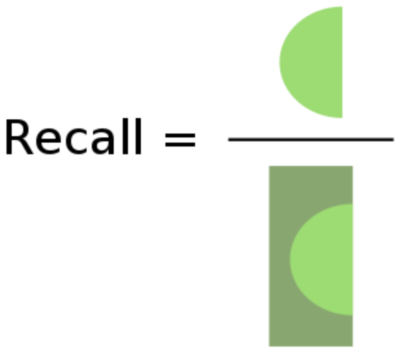
Precision & Recall



What fraction "classified positive" are True Positives; i.e., of the alarms, how many are faulty?



What fraction "positive class" are True Positives; i.e., of the faults, how many of them are identified?

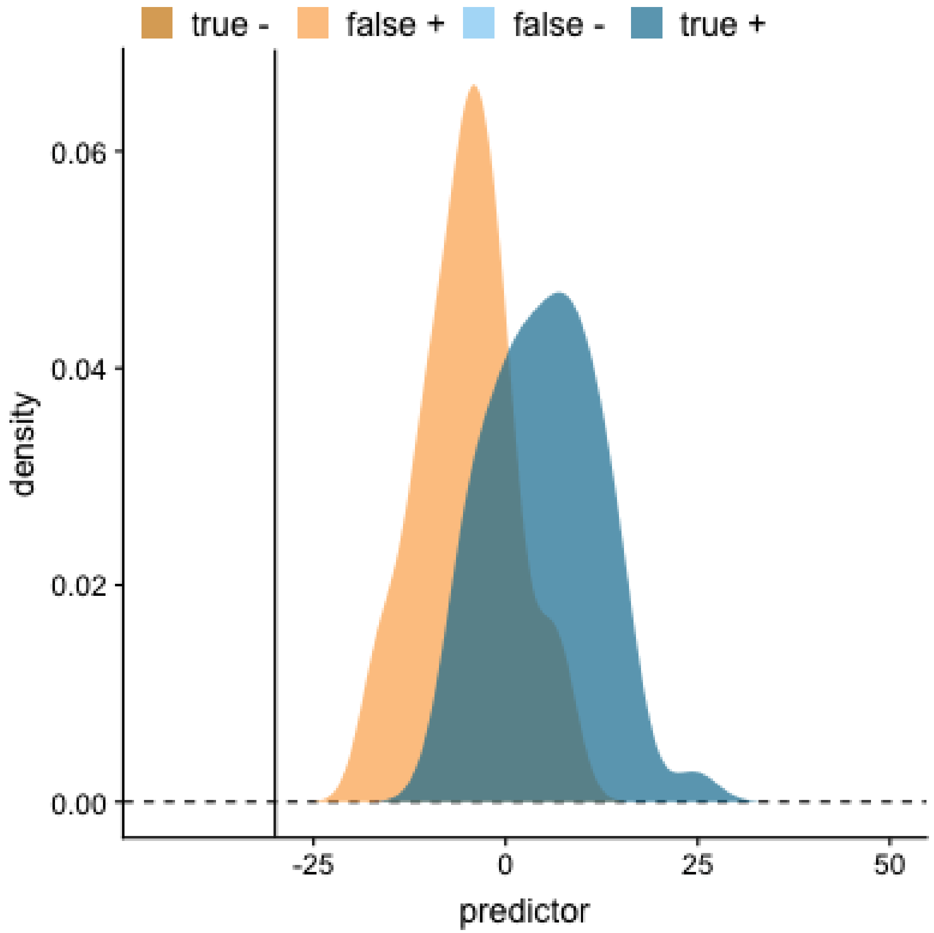


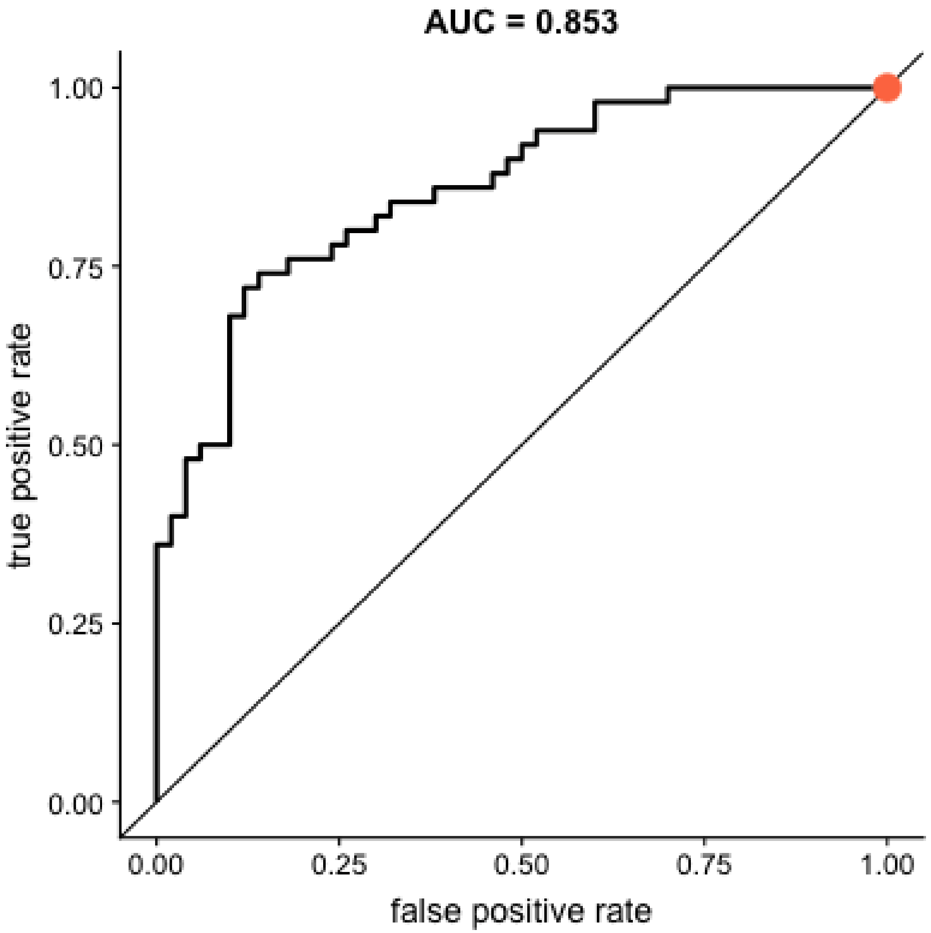
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

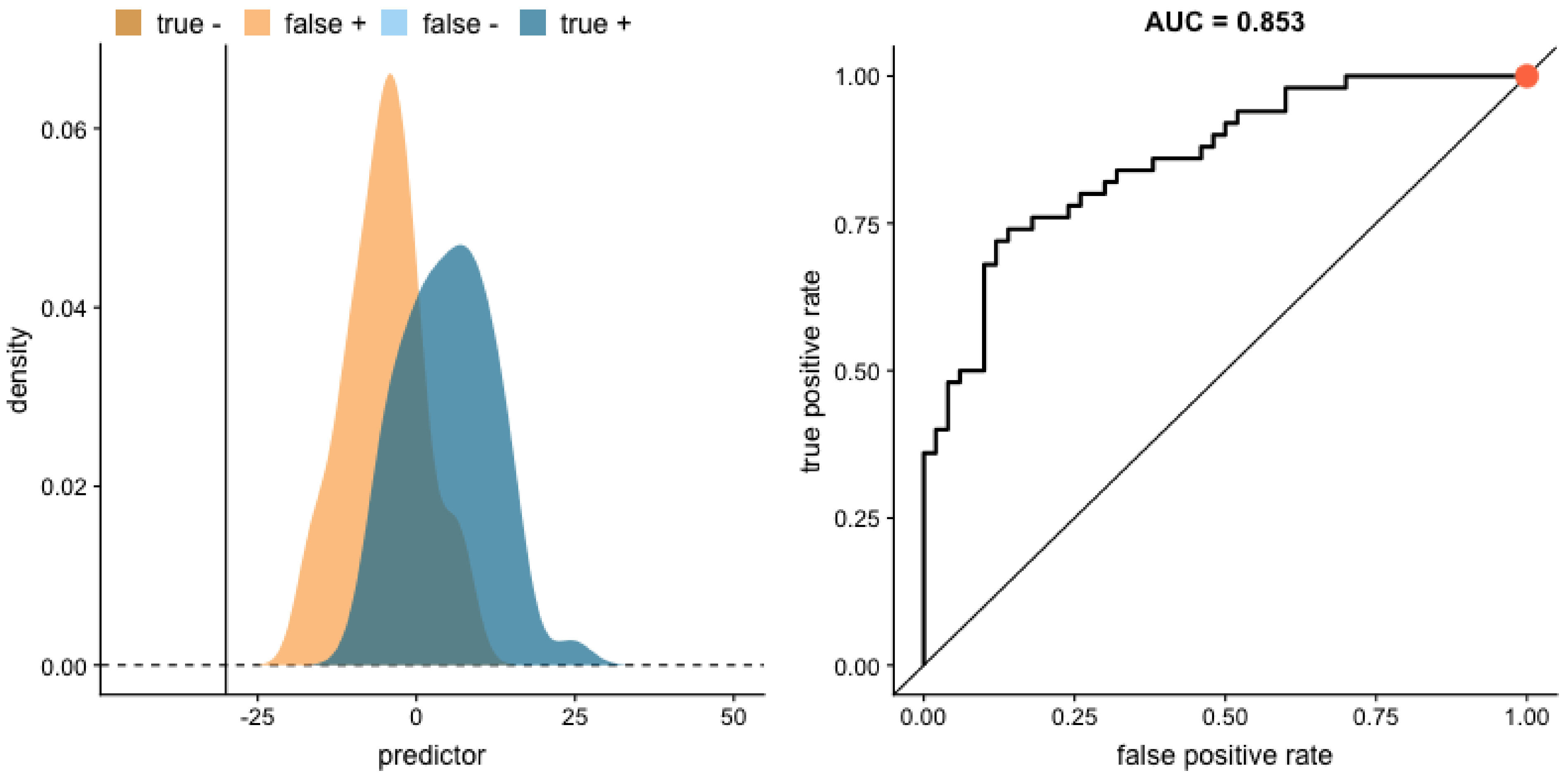
https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg

Calculating ROC, Balanced Data





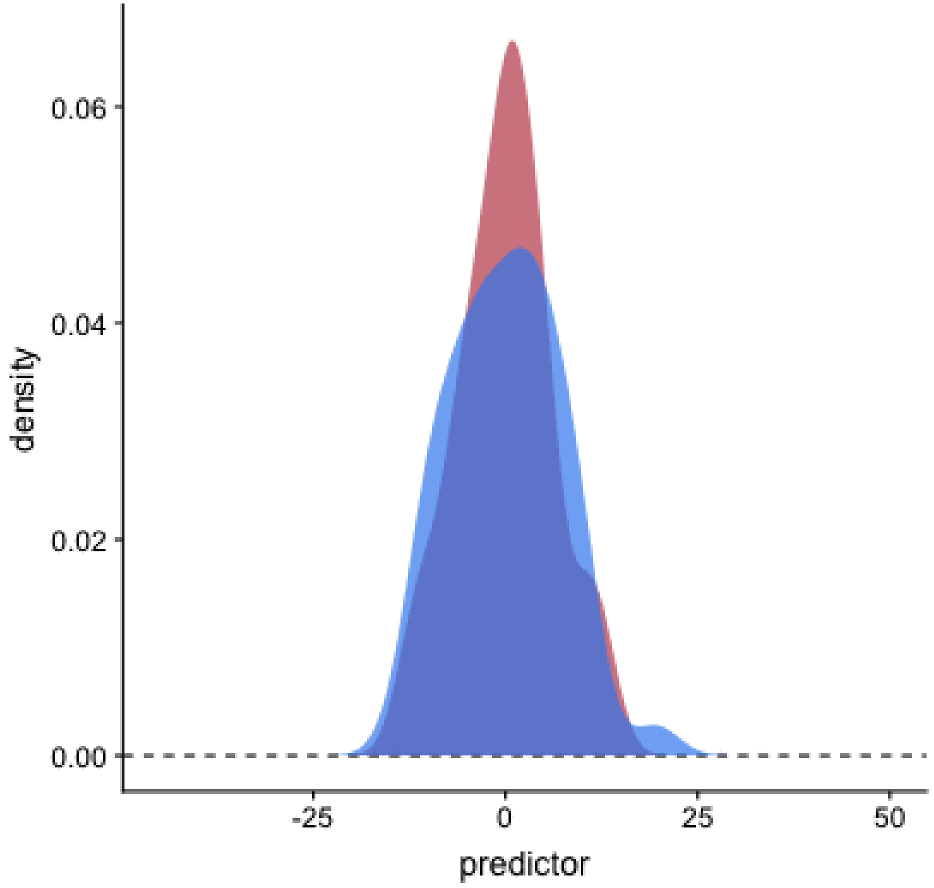
Calculating ROC, Balanced Data

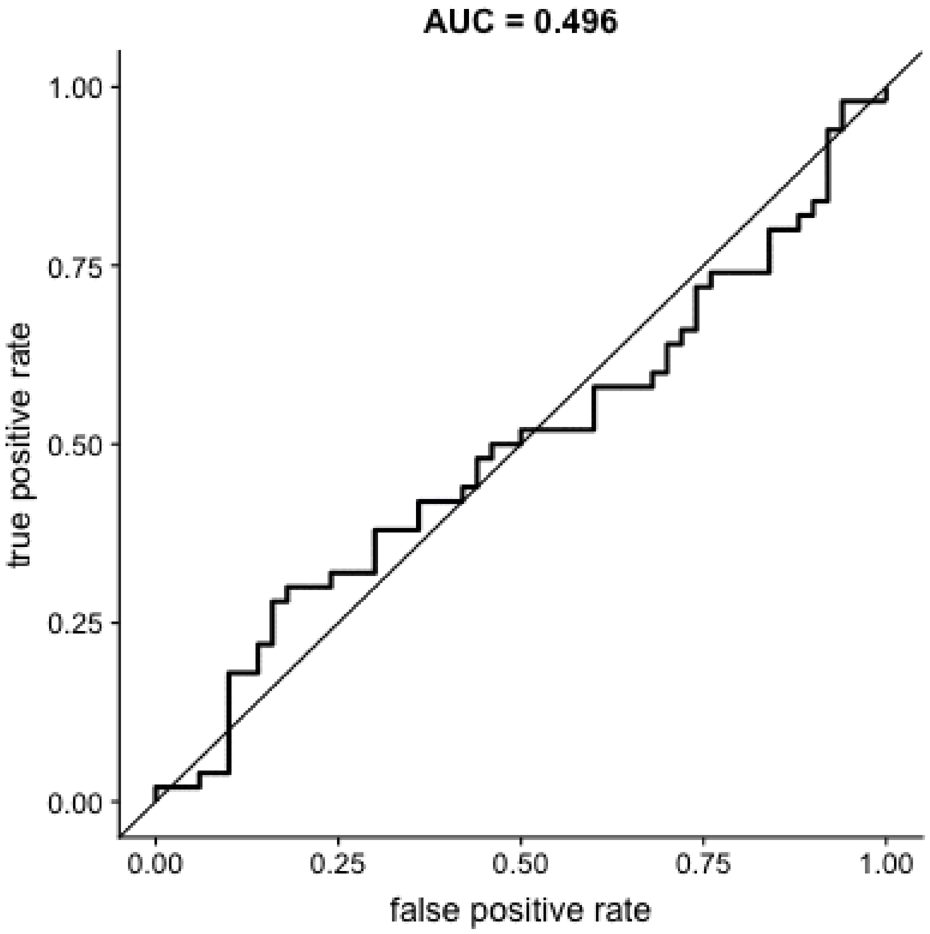


R code for all of these animations was kindly shared by Dariya Sydykova and is available here:
https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

Calculating ROC, Balanced Data

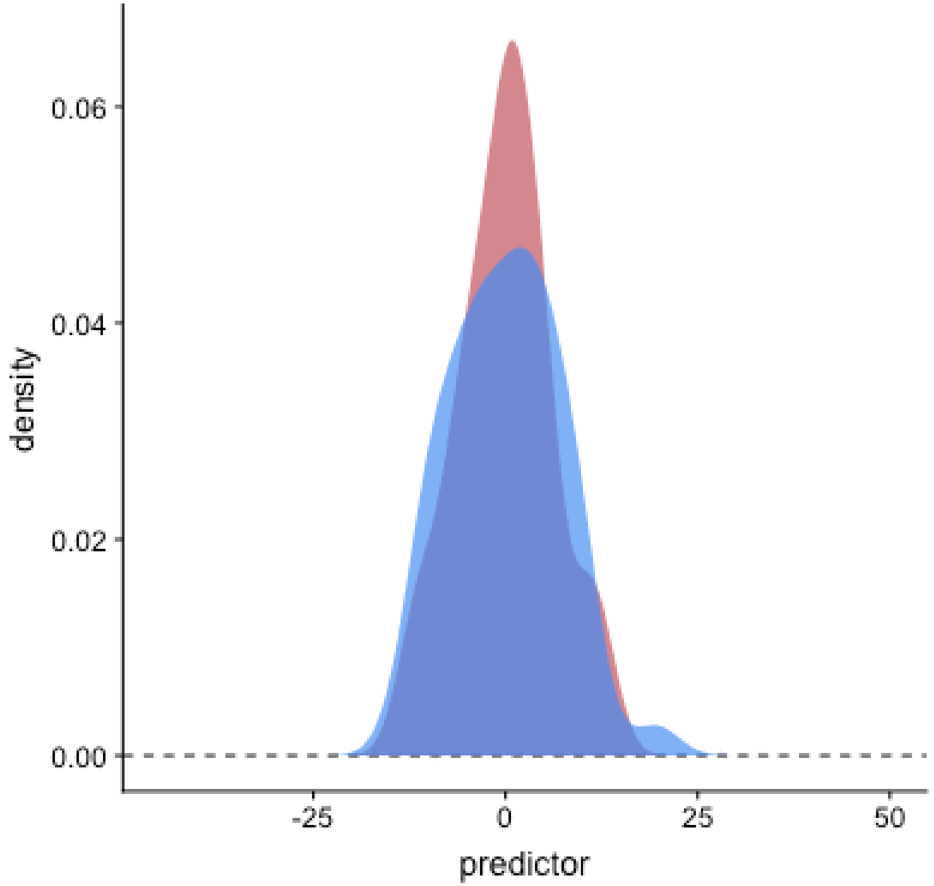
- Same standard deviation, different means

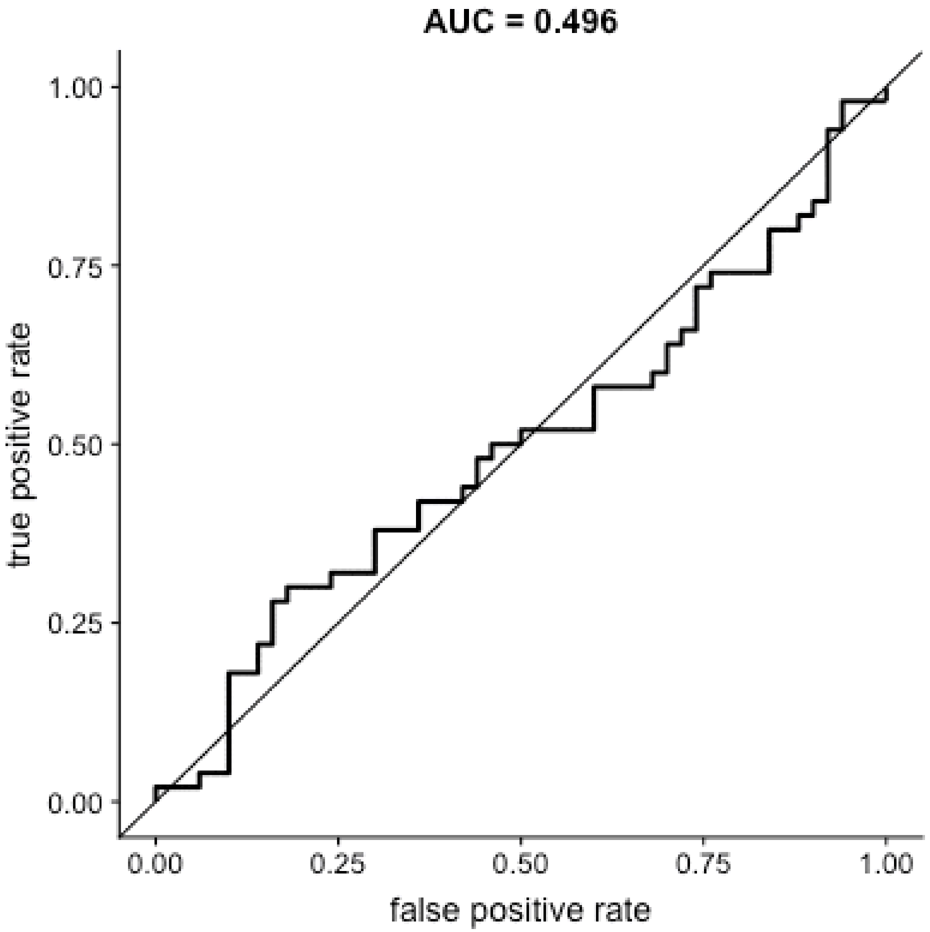




Calculating ROC, Balanced Data

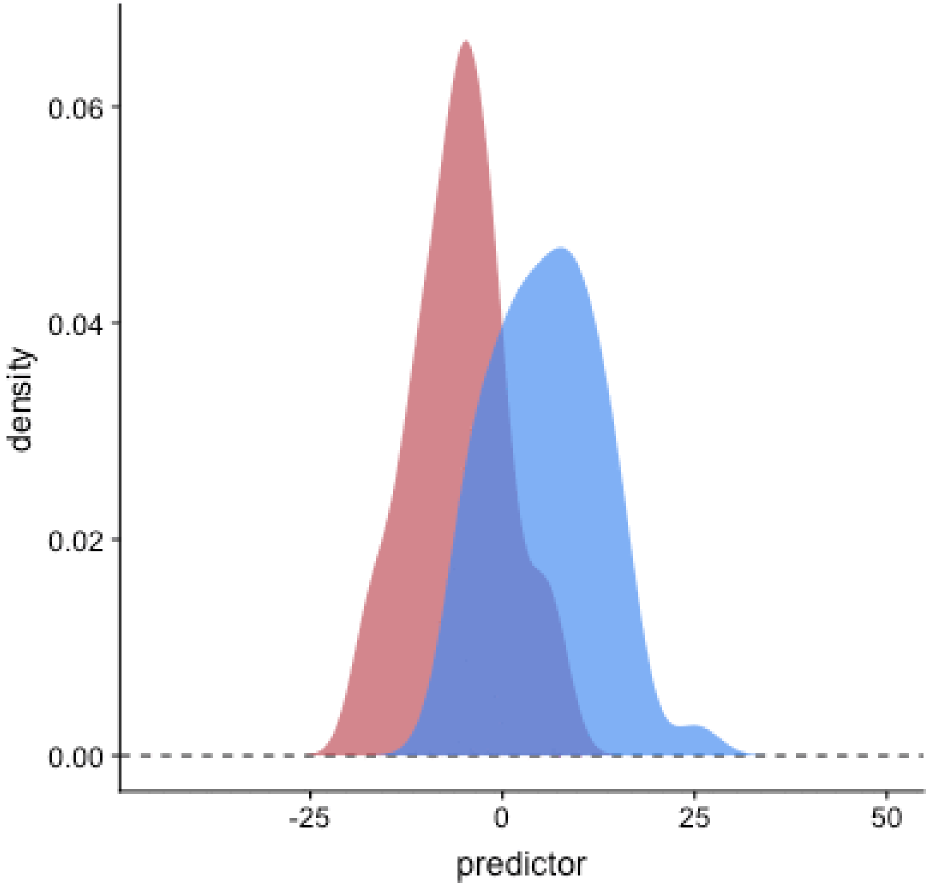
- Same standard deviation, different means

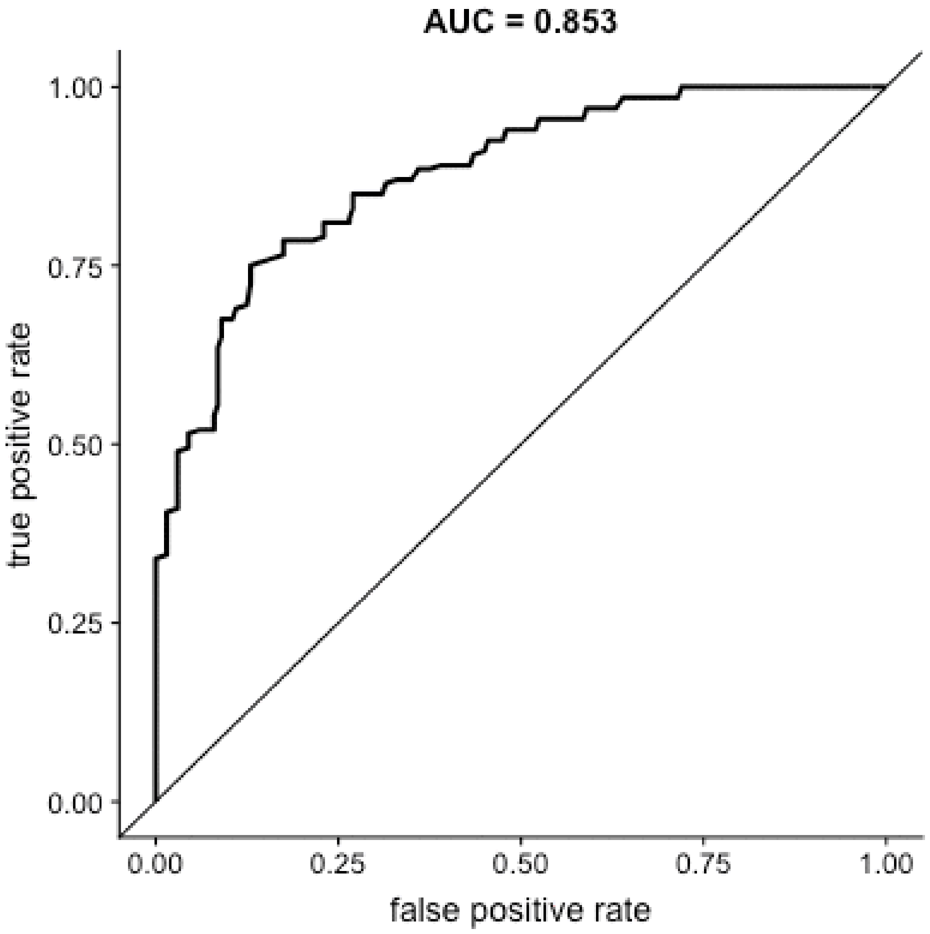




Calculating ROC, Balanced Data

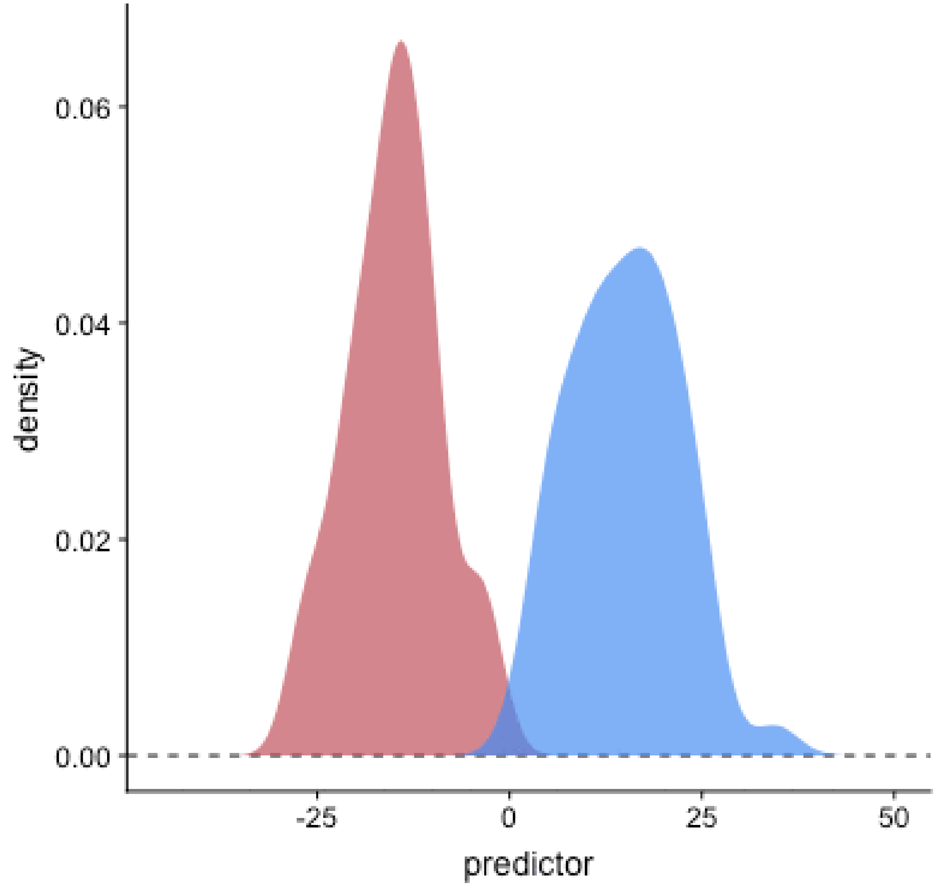
- Same standard deviation, different means

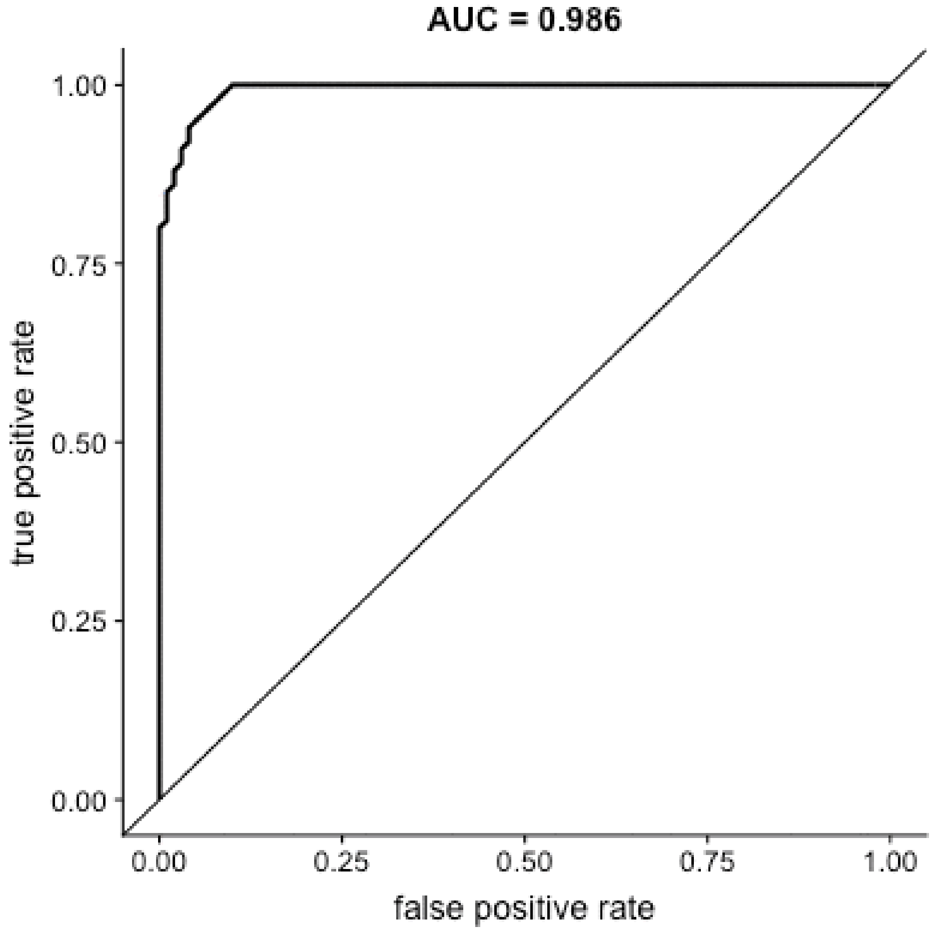




Calculating ROC, Balanced Data

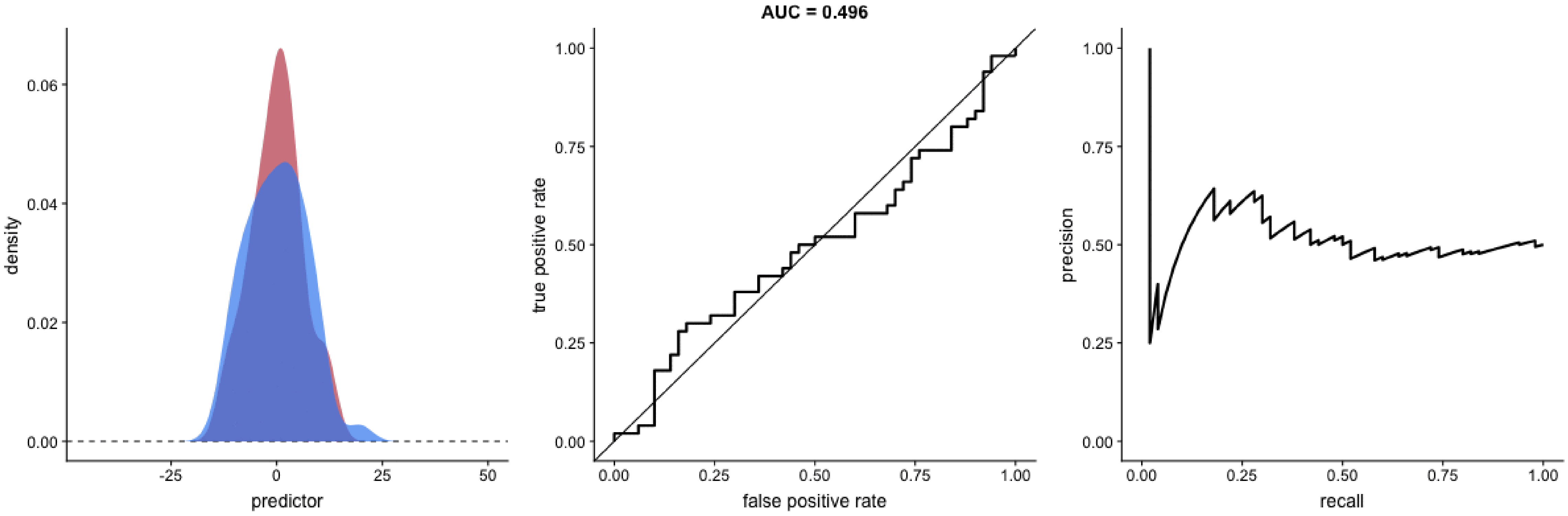
- Same standard deviation, different means



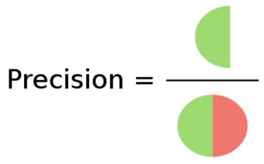


Calculating ROC & PR, Balanced Data

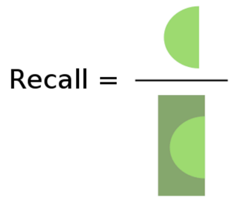
- Same standard deviation, different means



What fraction "classified positive" are True Positives; i.e., of the alarms, how many are faulty?

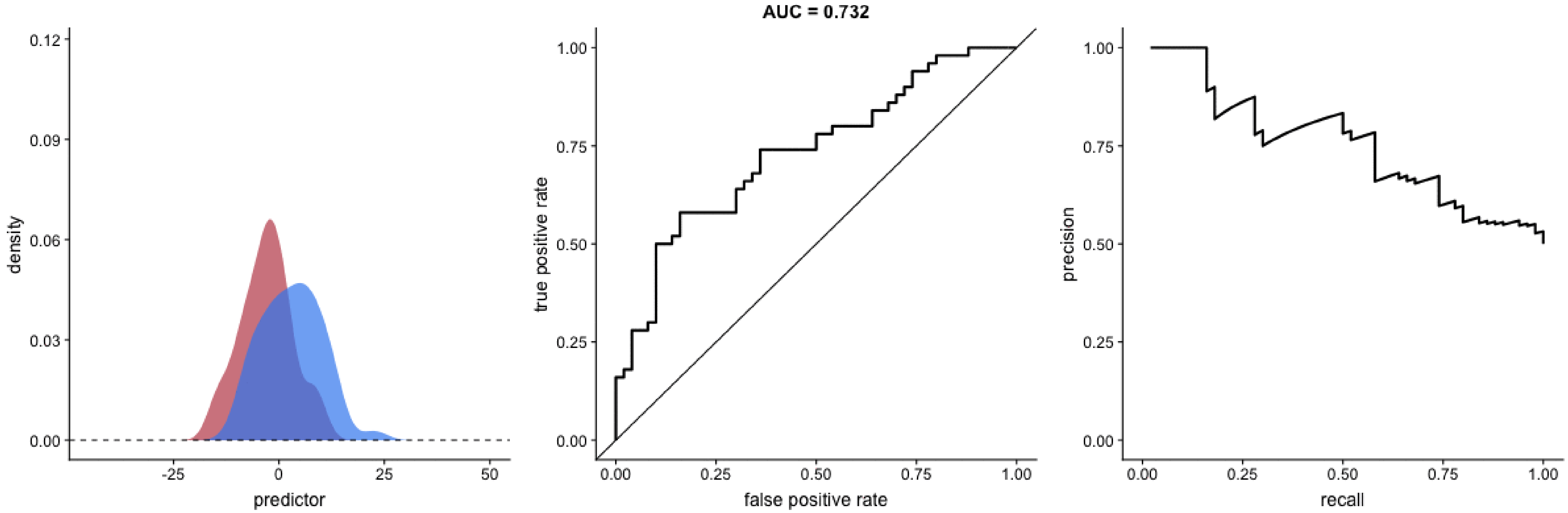


What fraction "positive class" are True Positives; i.e., of the faults, how many of them are identified?

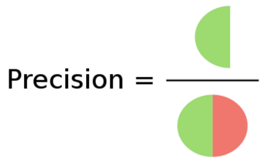


Calculating ROC & PR, Balanced Data

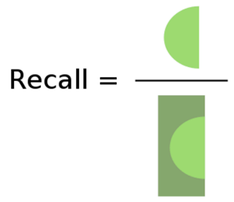
- Different standard deviation, fixed means



What fraction "classified positive" are True Positives; i.e., of the alarms, how many are faulty?

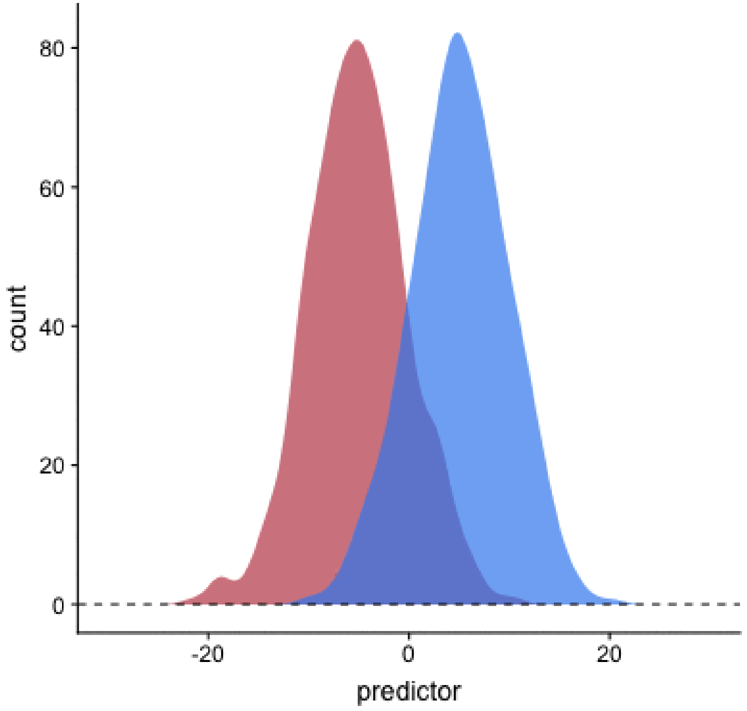


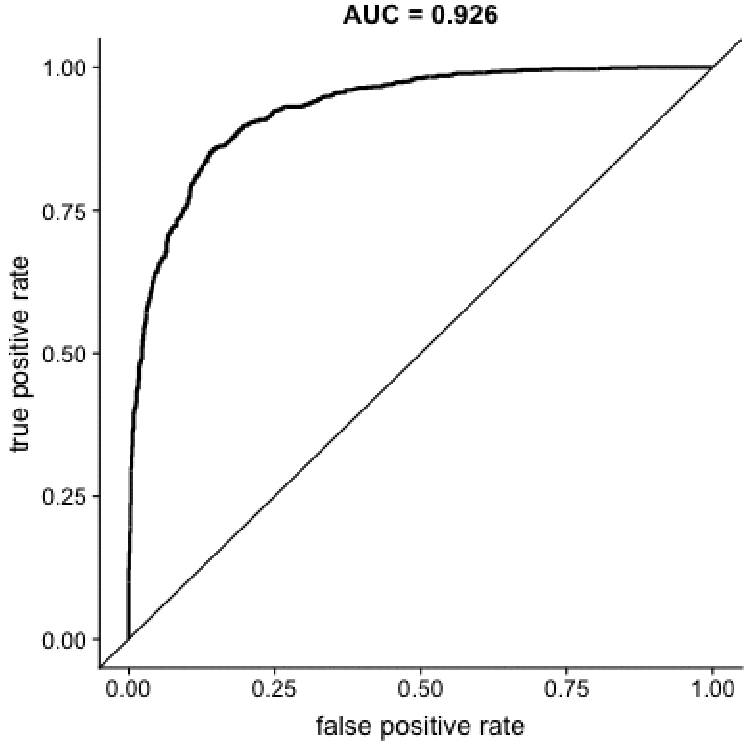
What fraction "positive class" are True Positives; i.e., of the faults, how many of them are identified?

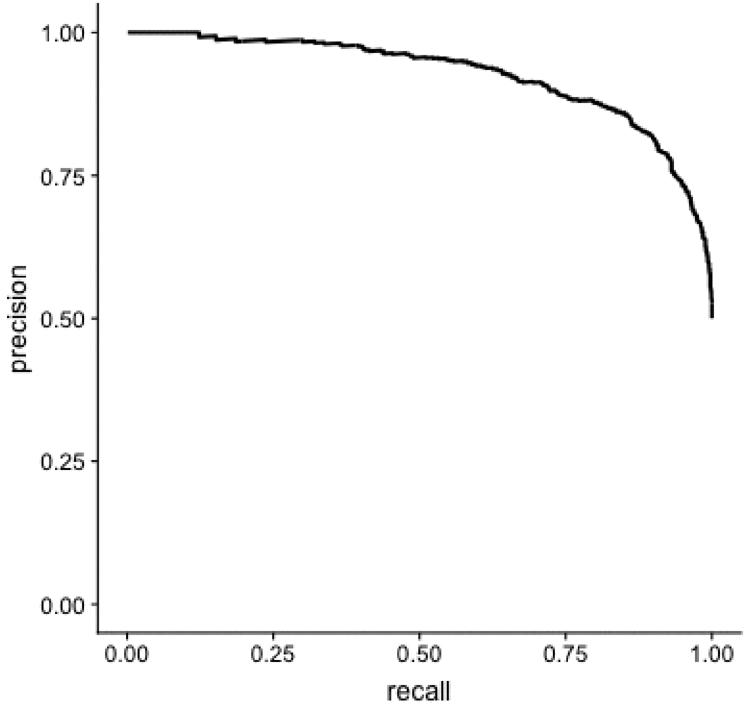


Calculating ROC & PR, Imbalanced Data

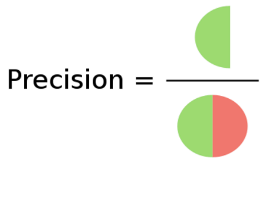
- Same standard deviation, fixed means, changing positive class



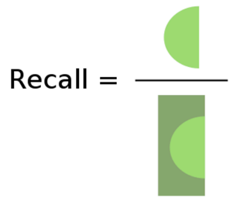




What fraction "classified positive" are True Positives; i.e., of the alarms, how many are faulty?

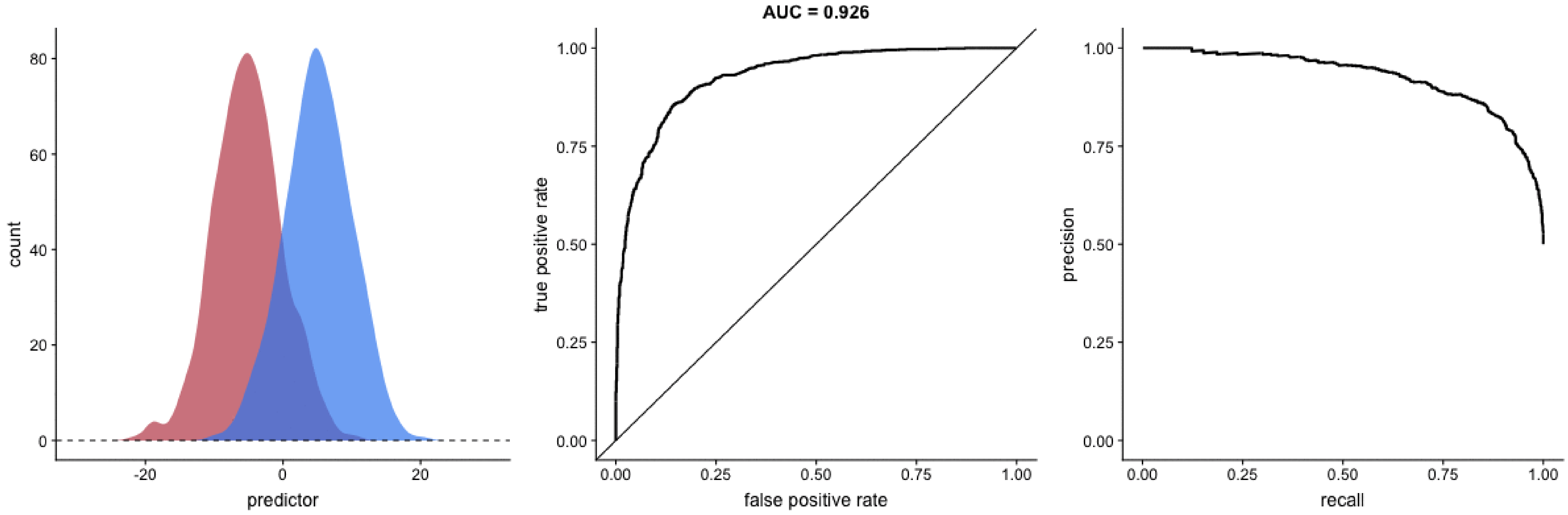


What fraction "positive class" are True Positives; i.e., of the faults, how many of them are identified?



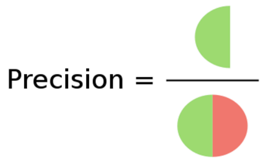
Calculating ROC & PR, Imbalanced Data

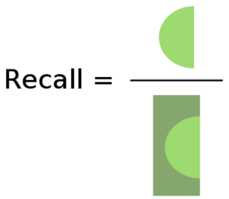
- Same standard deviation, fixed means, changing NEGATIVE class



What fraction "classified positive" are True Positives; i.e., of the alarms, how many are faulty?

What fraction "positive class" are True Positives; i.e., of the faults, how many of them are identified?





Conclusions

- Optimization of the ROC curve tends to maximize the correctly classified *positive* values (TP, which are present in the numerator of the true positive rate formula), and the correctly classified *negative* values (TN, which are present in the denominator of the false alarm rate formula).
- Meanwhile, the optimization of the PR curve tends to maximize the correctly classified *positive* values (TP, which are present both in the *precision* and in the *recall* formula), and does not consider directly the correctly classified *negative* values (TN, which is absent both from the *precision* and in the *recall* formula).

Conclusions (cont.)

- In PHM, we often have very sparse dataset with many *negative* instances (normal operation) and few *positive* instances (faults). Therefore, we prefer to avoid the involvement of *true negatives* in our prediction score.
- For these reasons, the *Precision-Recall curve* is a more reliable and informative indicator of statistical performance than the *receiver operating characteristic* curve for PHM datasets.

Additional Ideas

- For presenting to the customer, often it makes sense to just describe performance in terms of raw counts; e.g., in the last nine months, there were 21 faults detected
- While you want to *collect* as much data as possible, you only want to *evaluate* data at meaningful intervals. Otherwise, you unnecessarily increase the probability of false alarm.
  - For example, you might sample an asset every hour, but if it takes eight weeks to go from detection to failure, it might make sense to only evaluate the data you have collected once per day. If the probability of a false alarm is 0.0001, then in one month:
    - Evaluating every day, probability of false alarm is 0.003: 1-(1-0.0001)^30
    - Evaluating every hour, probability of false alarm is 23 times greater, 0.069: 1-(1-0.0001)^720

If you are interested in this topic, this paper is an outstanding reference:

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432