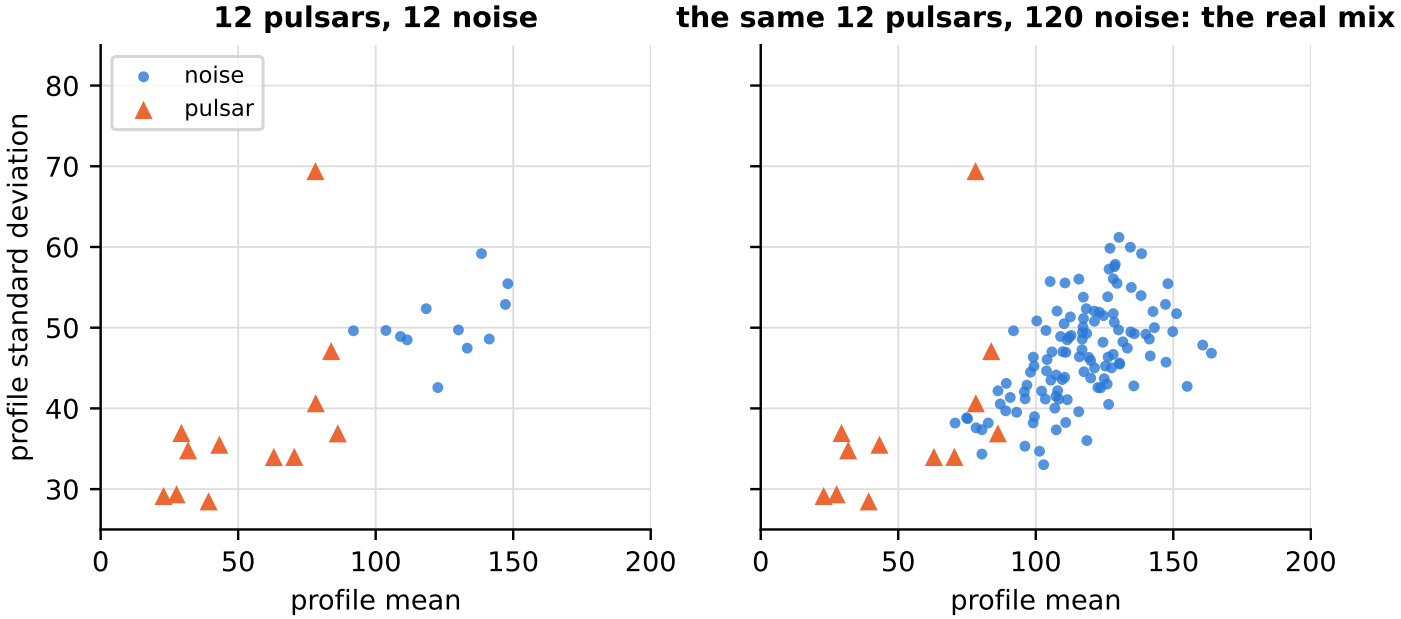
DATA 202 — Imbalanced Data · Week 12 Monday

Candidate signals from the Parkes radio telescope: about 1 real pulsar for every 10 signals of interference or noise. Your professor will tell you when to do each part.

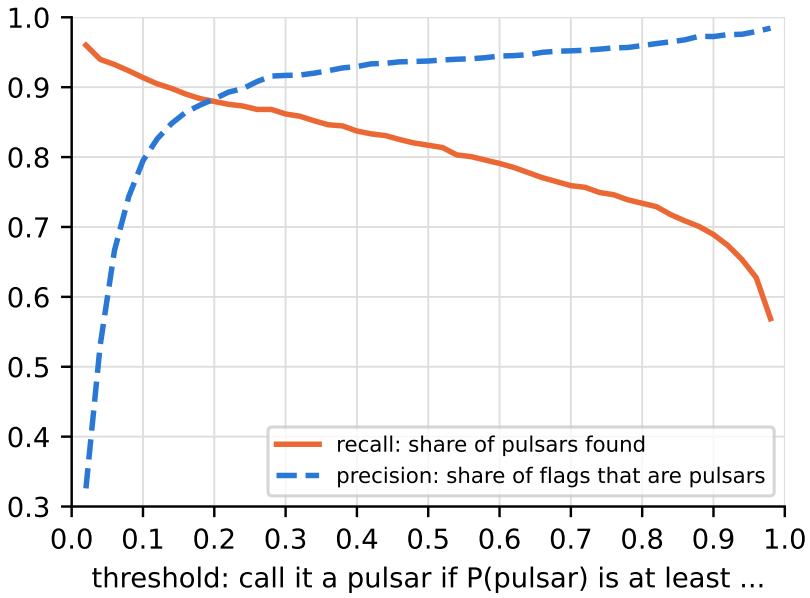
1 · Same pulsars, more noise



Twelve training pulsars, by two of their eight columns. **Left:** as many noise candidates as pulsars. **Right:** the same pulsars among 120 noise candidates, the real mix of about 1 to 10.

1. On each panel, **draw** the straight line that makes the fewest mistakes: pulsars on one side, noise on the other.
2. On the right, **circle** the pulsars your line calls noise.
3. **Pair talk:** only noise was added. Why did your line move? On the right, what would it cost to catch every pulsar?

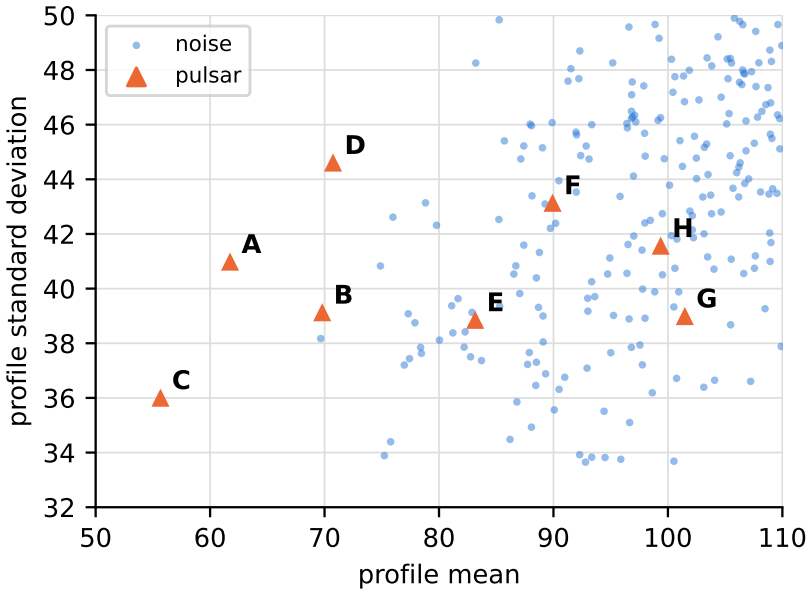
2 · Choose a threshold



All eight columns. Five-fold cross-validation on the training candidates, at each threshold.

1. **Team A** is two astronomers who check every flagged candidate by eye. **Mark** a threshold for them with an **A**.
2. **Team B** is building a complete census of pulsars, with a big team. Mark theirs with a **B**.
3. **Pair talk:** who should choose the threshold, the data scientist or the astronomers? What would they need to know?

3 · Make synthetic pulsars



SMOTE makes a new pulsar like this: take a pulsar, pick one of its nearest pulsar neighbors, and put the new pulsar somewhere on the straight line between them.

1. **Draw** a synthetic pulsar between A and B.
2. **Draw** one between F and G.
3. **Draw** two more, each from a pulsar to one of its nearest pulsar neighbors.
4. **Circle** every synthetic pulsar that landed among the noise.
5. **Pair talk:** suppose we make synthetic pulsars *before* splitting into training and test sets. Where could a test pulsar come from? What would the test score really measure?