Anomaly Detection in the Wild

Tim von Hahn

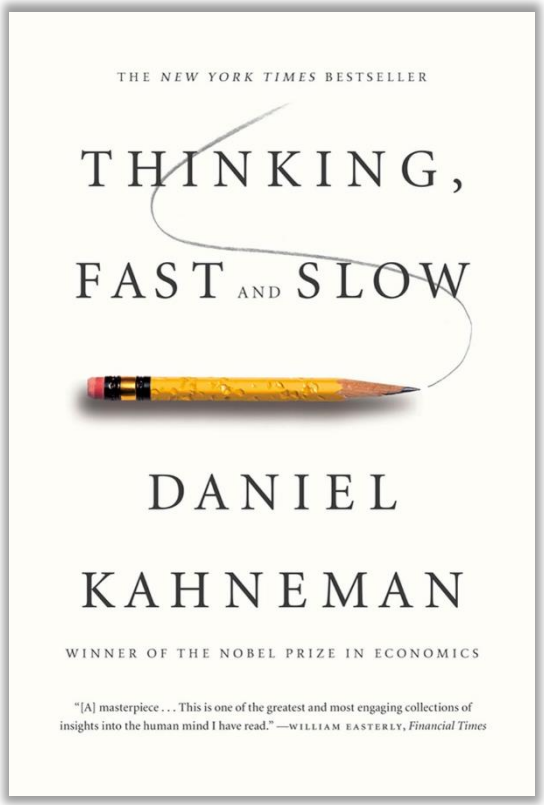
@timothyvh
github.com/tvhahn

Source: Ted Mielczarek, flickr

Source: Library of Congress

Indicator		0 Points	1 Point	2 Points
A	**Activity** (muscle tone)	none	some flexion	flexed arms and legs that resist extension
P	**Pulse**	absent	< 100 bpm	> 100 bpm
G	**Grimace** (reflex, irritability)	no response to stimulation	grimace on suction or aggressive stimulation	cry on stimulation
A	**Appearance** (skin colour)	blue or pale all over	blue at extremities, body pink	body and extremities pink
R	**Respiration**	absent	weak, irregular, gasping	strong, robust cry

Source: Wikipedia



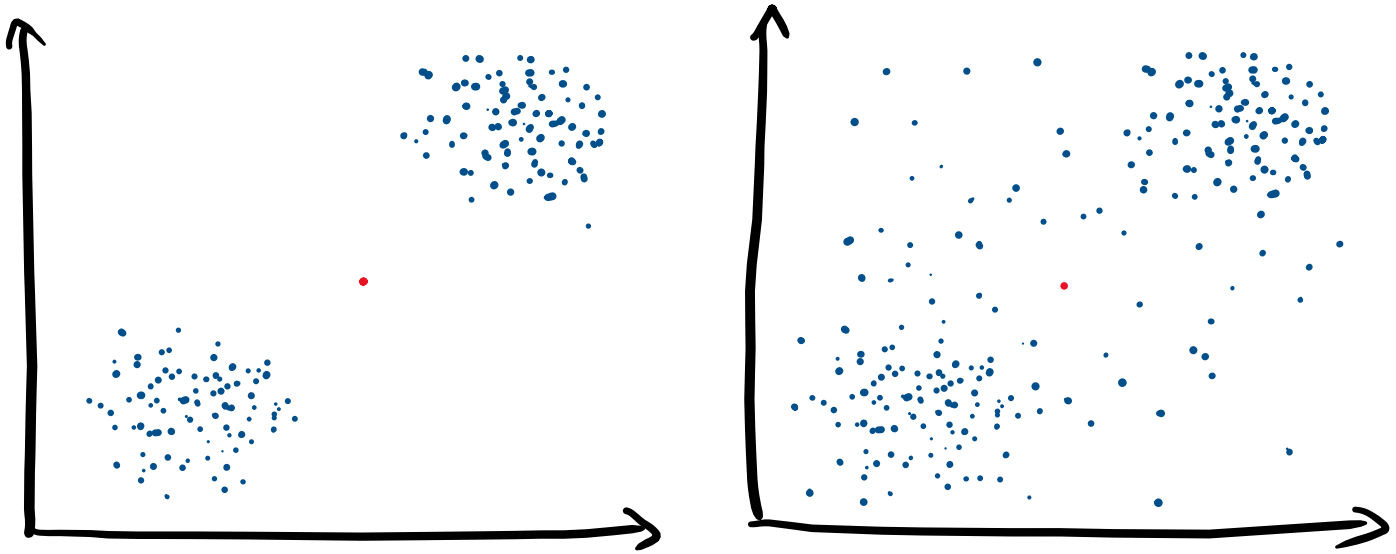
Source: General Electric

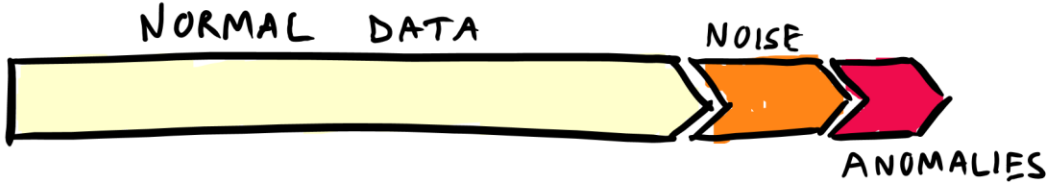
"Simple, statistical rules are superior to intuitive "clinical" judgments."

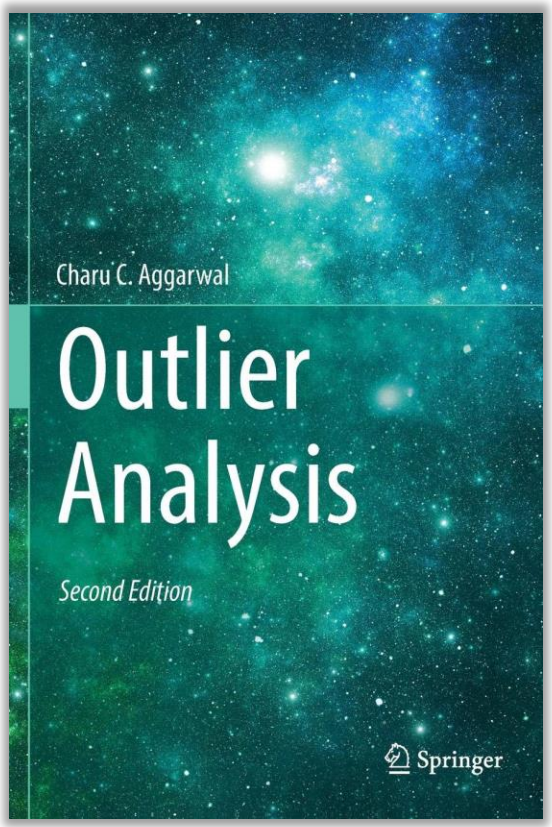
– Daniel Kahneman

"An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism."

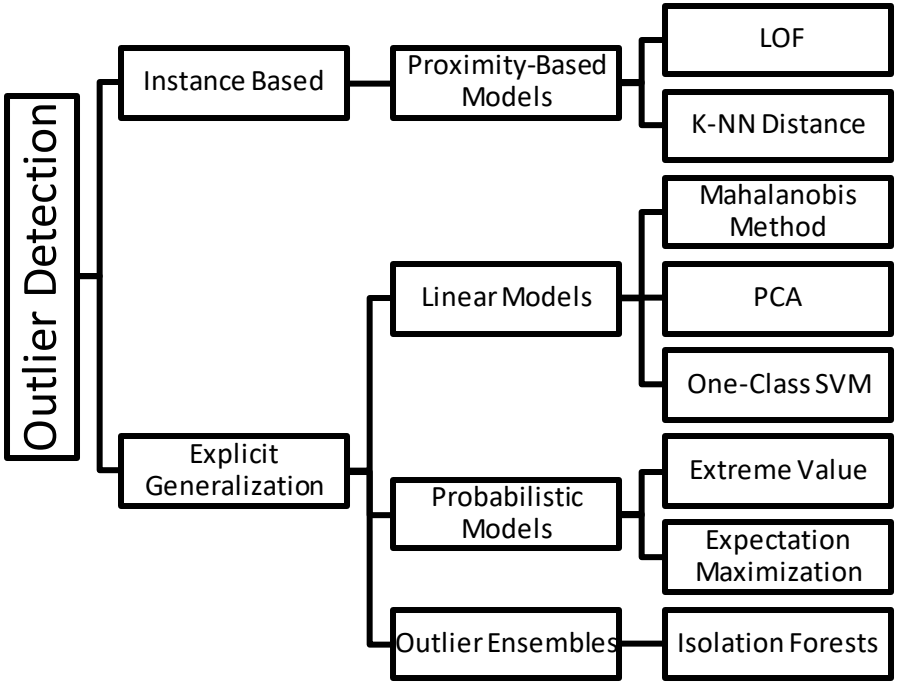
– D.M. Hawkins



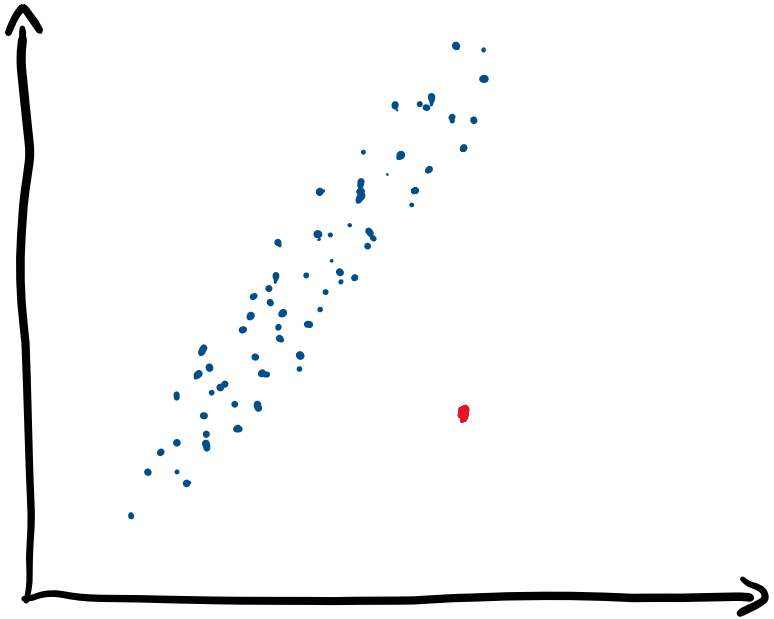

NORMAL DATA
NOISE
ANOMALIES

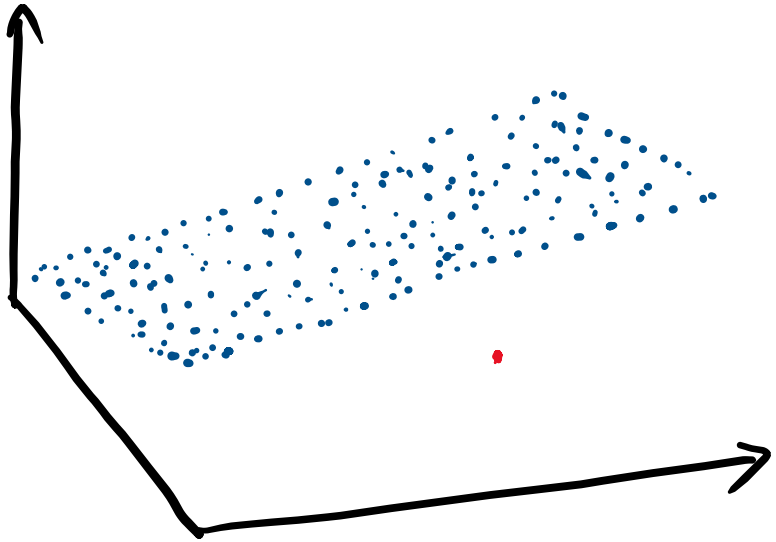


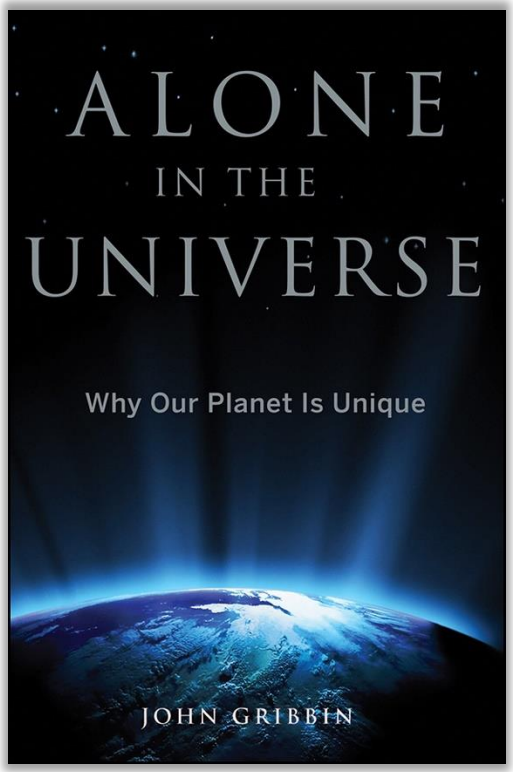
"Outlier Analysis" by Charu Aggarwal – excellent book!



Examples of some outlier detection techniques







"Alone in the Universe" by John Gribbin

Source: Nick Risinger, Wikipedia

Astroinformatics

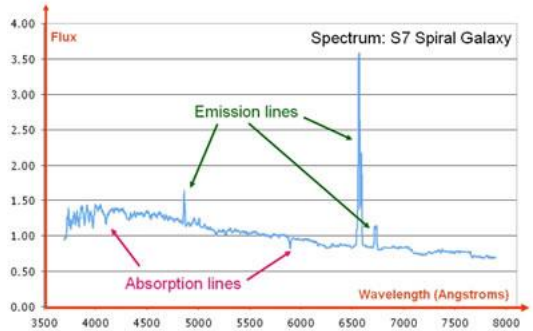
"With these new technologies, the volume of data is increasing exponentially, but the number of astronomers analyzing the data is not."

– Dr. Kai Polsterer

The Weirdest SDSS Galaxies

- Dalya Baron, Dovi Poznanski (link)
- "Unknown-unkowns"

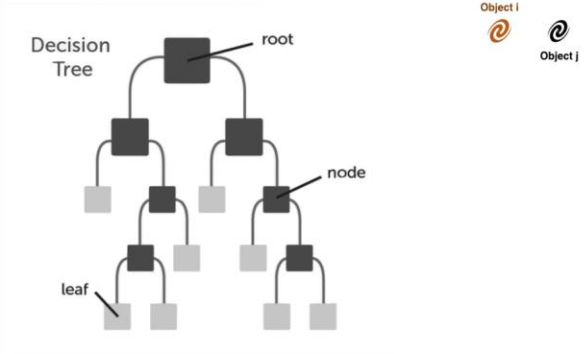
Source: NASA, Hubble

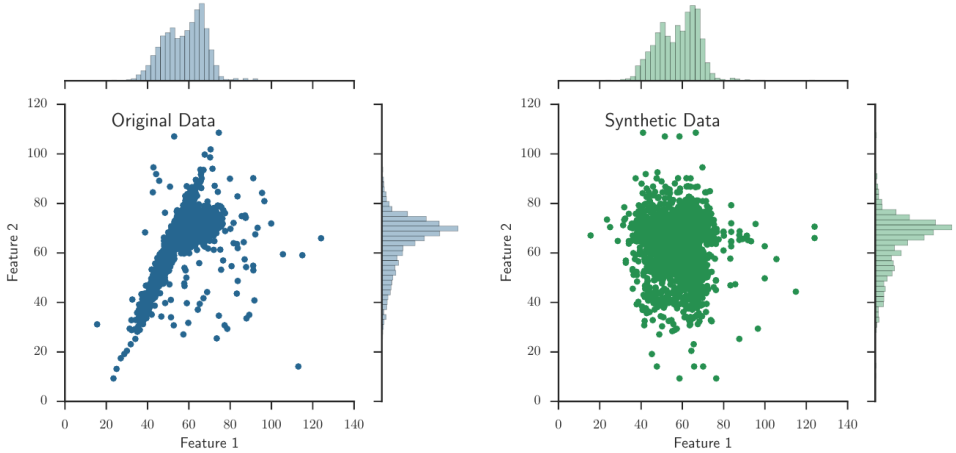


Source: Santos et al. 2002

The Weirdest SDSS Galaxies

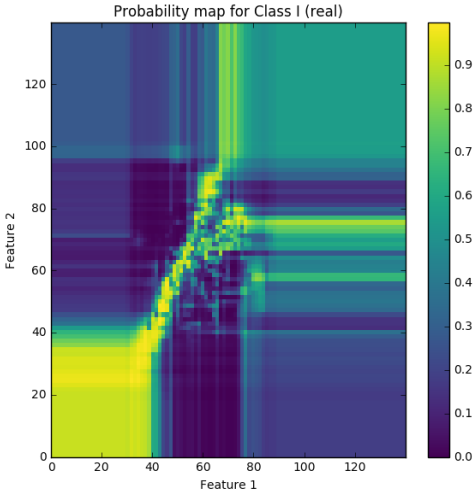
- 2,355,926 spectra, with 15,700 flux values each
- Generate synthetic data from original
- Use RF to classify as real or synthetic





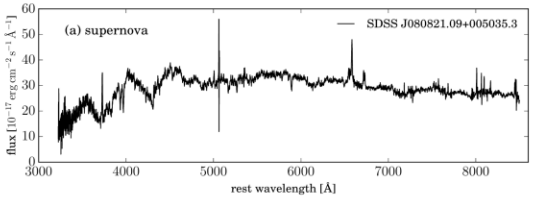
Orginal and synthetic data example. Synthetic data generated from same marginal distribution, but without covariance. **Github code (in python),** here

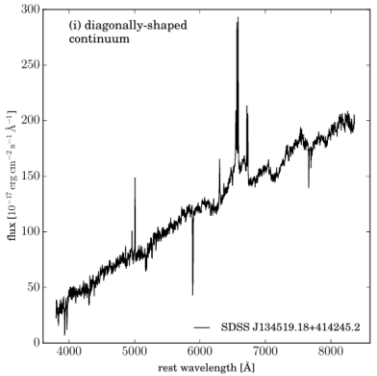
Probability map, on example data, showing where the "outliers" are most likely to be

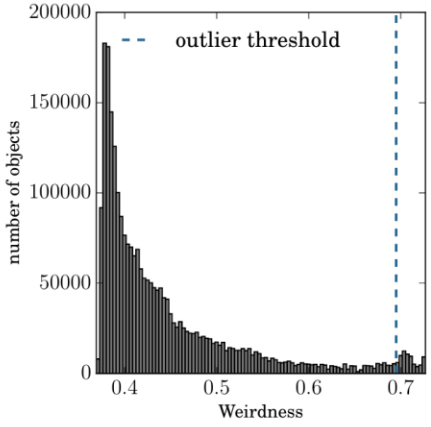


The Weirdest SDSS Galaxies

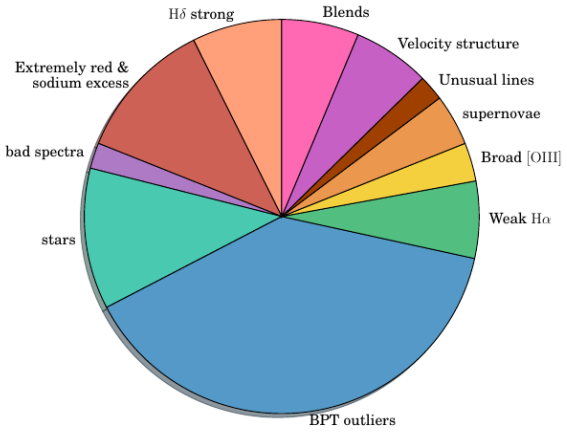
- "Weird" galaxies found! (very few reported on before)
  - Galaxy-galaxy gravitational lenses
  - Galaxies with host supernovae
  - Extreme emission line ratios
  - Etc.







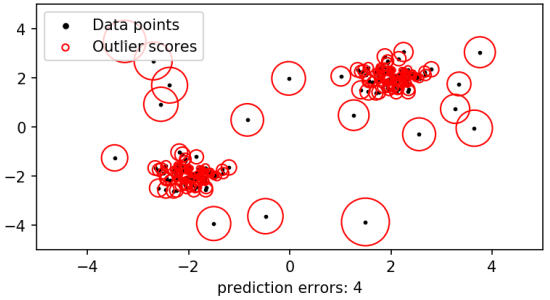
Outlier threshold set on level of "weirdness"



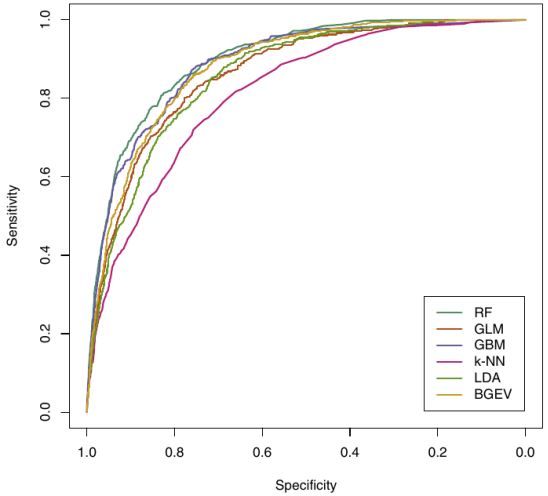
Solvency Predictions for Businesses

- "Solvency prediction for small and medium enterprises in banking", Figini et al.
- 38,036 businesses
- 43 variables per business (e.g. financial ratio, employees, etc.)
- Use Local Outlier Factor (LOF) as additional feature, and classified businesses as outliers based on LOF threshold

- Sklearn has many anomaly detection methods in their library!



LOF scores of random data set (from sklearn)



ROC curves for various models using LOF data

Anomaly Detection in Machinery Health Monitoring

- Notebook in my repo

github.com/tvhahn/anomaly-pyconca