Signal in the Noise: The Role of Anomalous Data in Scientific Discovery and the Imperative for Modern Data Science

Introduction: The Two Faces of the Outlier

In the landscape of modern data science and machine learning (ML), outliers present a fundamental dilemma. To the practitioner, it is often a statistical nuisance—a data point so distant from the norm that it threatens to skew models, violate assumptions, and degrade predictive accuracy.[1] The standard response, codified in the practice of "data cleaning," is to identify and remove these deviants to ensure the reliability and performance of analytical models.[3] This approach is not arbitrary; it is rooted in a pragmatic desire to build robust systems that accurately describe the majority of the data.[5]

However, an alternative and historically resonant perspective asserts that all data, especially that which lies outside the norm, has the potential for profound meaning. This view challenges the conventional wisdom of the ML community, reframing the outlier not as an error to be scrubbed, but as a potential signal of a new and undiscovered phenomenon. It posits that the most meaningful discoveries are made not in spite of, but *because of*, such data. This position elevates data points from mere numerical inputs to potential evidence, echoing a long and storied history of scientific breakthroughs born from the unexpected.

The routine, automated removal of outliers, while pragmatically justified for optimizing model performance on known problems, is epistemologically hazardous. It risks blinding the scientific enterprise to the very data points that signal novel phenomena and drive paradigm shifts. This report will substantiate this claim by first dissecting the orthodox statistical view of outliers, then reframing them within the philosophy of science as powerful "anomalies." Finally, it will provide a robust historical evidence base demonstrating their pivotal role in discovery, culminating in a proposed synthesis for a more nuanced and scientifically curious approach to modern data science.

Section 1: The Anatomy of an Outlier: Definition, Detection, and Disposal

To appreciate the role of anomalous data, one must first understand how it is defined, identified, and handled within the conventional statistical and machine learning paradigms. This section establishes the orthodox view, detailing the methodologies and justifications that underpin the common practice of "data cleaning" before critically examining its limitations.

1.1 Defining the Deviant: A Statistical and Methodological Primer

An outlier is formally defined as a data point that differs significantly from other observations, being numerically distant from the rest of the dataset.[7] These points can lie at the extreme ends of a distribution and may impact statistical analyses in ways that misrepresent the sample.[9] Crucially, there is no single, rigid mathematical definition of what constitutes an outlier; its determination is ultimately a subjective exercise, guided by various heuristics and methods.[8]

At the core of this subject is a vital distinction between two primary sources of outliers. The first category includes errors such as measurement errors from faulty equipment, data entry or processing mistakes, or unrepresentative sampling, where an observation is drawn from an entirely different population.[9] The second category consists of "true outliers," which are legitimate but rare values that represent natural variations within the target population.[9] This distinction is the fulcrum upon which the entire argument about data handling pivots: one type is an artifact of the process, while the other is a feature of reality.

Practitioners employ several common methods to detect these deviant points:

- **Visual Methods:** The most intuitive approach involves visualizing the data. Box plots (or box-and-whisker plots) are a primary tool that explicitly highlight outliers as individual points falling beyond the "whiskers" of the plot.[12] Scatterplots and histograms can also reveal isolated points that deviate from the main cluster or distribution of the data.[12]
- **Statistical Tests:** Quantitative methods provide more formal rules for identification. The **Z-score** method calculates how many standard deviations a data point is from the mean. As a rule of thumb, a Z-score greater than 3 or less than 3 is often flagged as an outlier.[9] A widely used alternative, particularly effective for skewed distributions, is the **Interquartile Range (IQR)** method. This approach defines "fences" around the central body of the data. The lower fence is set at $Q1-(1.5*IQR)$ and the upper fence at $Q3+(1.5*IQR)$, where Q1 and Q3 are the first and third quartiles, respectively. Any data point falling outside these fences is considered an outlier.[9]

1.2 The Culling of Data: Justifications for 'Data Cleaning' in the Age of Machine Learning

The primary motivation for outlier removal, or "data cleaning," is the profound and often detrimental impact these points have on many statistical methods and machine learning algorithms. Parametric statistics like the mean and standard deviation are highly sensitive to extreme values; a single outlier can dramatically alter their value, thereby distorting any subsequent analysis that relies on them.[8]

In the context of machine learning, this distortion has several negative consequences:

- **Violation of Model Assumptions:** Many models, such as linear regression, assume that the data (or its errors) follow a normal distribution. Outliers can violate this assumption, leading to unreliable results.[11]
- **Reduced Model Accuracy:** By skewing the training process, outliers can lead to models that are less accurate and perform poorly on new data. Their presence can increase error variance and reduce the statistical power of hypothesis tests.[5] In some cases, removing outliers has been shown to result in a statistically significant increase in model accuracy.[6]
- **Inefficient Training:** The presence of extreme values can lengthen the time it takes for a model to converge during the training phase.[1]

Consequently, the process of data cleaning—also known as data cleansing or scrubbing—is widely considered a critical step in the data lifecycle. It is a process of fixing or removing incorrect, corrupted, or irrelevant data to ensure data quality, accuracy, and consistency.[4] High-quality, clean data is seen as the foundation for reliable analytics, sound business decisions, and effective AI systems.[3]

1.3 A Critical Interlude: Questioning Automatic Removal and the Case for Robustness

Despite these pragmatic justifications, the practice of summarily dropping an observation *just because it is an outlier* is a deeply flawed approach.[19] Such data points can be legitimate observations and are sometimes the most interesting ones, containing valuable information that is lost upon removal.[19] The user's own observation that model accuracy can *drop* after removing outliers is a powerful indicator that these points may contain important predictive information that the model was leveraging.[20]

This critique exposes a fundamental linguistic and philosophical schism. The language of data cleaning is one of pathology, using terms like "corrupted," "dirty," "noise," and "contamination" to describe outliers.[4] This vocabulary presupposes that the deviant data is illegitimate and must be excised. In stark contrast, the philosophy and history of science use a language of opportunity, describing the same phenomena with words like "anomaly," "novelty," and

"surprising.".[22] This linguistic divide reflects a deep disagreement about the epistemological status of deviant data, dictating a choice between two actions: removal versus investigation. Furthermore, the process of identifying outliers is itself a potential trap. Because the definition of an outlier is "ultimately a subjective exercise" [8] and the rules for identifying them are heuristics [9], a dangerous feedback loop can emerge. A practitioner, believing their data *should* conform to a particular model, can use a subjective rule to remove data that doesn't fit, thereby "proving" their initial assumption. This is a form of confirmation bias enacted through data processing, which risks improving the fit to a potentially flawed or incomplete theory of the data rather than discovering a better one.

A more sophisticated alternative to simple removal is the field of **Robust Statistics**. These methods are designed to handle data corrupted by outliers by bounding the influence that any minority of the dataset can have on the final prediction.[24] Instead of excising the data, robust methods adapt to it. This can be as simple as using the median (a robust measure of central tendency) instead of the mean.[8] More advanced techniques modify learning algorithms to assign a lower weight to outliers, thereby mitigating their distorting effects while still retaining them in the dataset. Methods like Huber weighting and Median-of-Means (MoM) allow for the creation of models that are inherently less sensitive to extreme values, offering a way to achieve stability without sacrificing potentially valuable information.[24]

To guide a more nuanced approach, the following taxonomy can be used to move practitioners from a binary "keep/drop" decision to an investigation-led process.

Outlier Source	Description	Common Cause	Recommended Action	Rationale
Measurement Error	A data point that is incorrect due to a faulty instrument or flawed procedure.	Equipment malfunction, procedural deviation.	Correct if possible; otherwise, remove with justification.	The data point does not represent the true state of the phenomenon being measured.[9]
Data Entry/Processing Error	A data point that is incorrect due to a human mistake or software bug.	Typographical error, data corruption during transfer.	Correct if possible; otherwise, remove with justification.	The data is an artifact of the collection process, not the phenomenon itself.[11]
Sampling Error	A valid data point that has been drawn from a different population than the one under study.	Contamination of the sample with elements from outside the target group.	Remove from the current analysis; consider analyzing as a separate population.	The data point is valid but does not belong to the population of interest for the current model.[8]

Novelty / True Anomaly	A legitimate, rare, and unexpected data point from the correct population.	A previously unknown phenomenon, a rare event, a change in system behavior.	**Investigate Intensively.** Retain in the dataset. Use robust models.	This is a potential source of new knowledge. Discarding it is an act of epistemological self-sabotage.[8]

Section 2: From Outlier to Anomaly: A Philosophical Reframing

To fully grasp the potential cost of discarding outliers, it is necessary to transition from the statistical lexicon of data processing to the philosophical language of scientific discovery. In this context, the outlier is transformed into the "anomaly," a concept with far greater significance—one that lies at the very heart of scientific progress.

2.1 The Anomaly as a Scientific Imperative: Falsification, Crisis, and Paradigm Shift

In the philosophy of science, an anomaly is defined as a "stubborn conflict between what is expected against the background of established theories and what is actually observed. ".[22] Anomalies are not mere statistical oddities; they are fundamental challenges to the established body of knowledge and have played a crucial role in scientific revolutions.[22]
This concept is central to the principle of **falsifiability**, articulated by the philosopher Karl Popper. According to Popper, a theory can only be considered scientific if it is capable of being proven wrong.[25] Anomalies—empirical results that contradict a theory's predictions—serve as these essential tests. They are the mechanisms by which science self-corrects and grows, pushing scientists to refine, revise, or even reject well-established theories when new evidence emerges.[25]
The work of historian and philosopher Thomas Kuhn further illuminates this process. Kuhn argued that science operates for long periods in a state of "normal science," where researchers work within an accepted theoretical framework, or **paradigm**. When anomalies begin to accumulate that cannot be explained by the current paradigm, the field may enter a period of "crisis." This crisis is only resolved by a **paradigm shift**—a scientific revolution in which the old theory is abandoned in favor of a new one that can successfully account for the previously inexplicable anomalies.[23] A classic example is the anomalous advance of Mercury's perihelion. This slight deviation in the planet's orbit could not be explained by Newtonian mechanics, representing a persistent anomaly for decades. It was ultimately explained

perfectly by Albert Einstein's theory of General Relativity, precipitating one of the most significant paradigm shifts in the history of physics.[22]

2.2 Serendipity and the Prepared Mind: Transforming Accident into Opportunity

The existence of an anomaly is not, by itself, sufficient to trigger a discovery. A human agent must recognize its significance. This brings forth the concept of **serendipity**, which is not merely a "happy accident" but a chance discovery that is actively exploited through sagacity and a "prepared mind.".[27] Estimates suggest that between 30% and 50% of all scientific discoveries are serendipitous in some sense.[29]

These discoveries are not the product of pure luck. They require the observer to possess deep background knowledge, an insatiable curiosity, and, most importantly, an openness to the unexpected.[28] The process of discovery often begins when a researcher encounters "bugs" or anomalous results in an experiment. The initial, and very human, impulse is to blame the methodology or assume an error has occurred. The pivotal moment—the leap from accident to discovery—occurs when the researcher concludes that the "error" is too persistent and systematic to be a coincidence and begins to investigate the anomaly itself as a phenomenon worthy of study.[29]

This reveals a profound inversion of the typical data cleaning mindset. The standard ML approach treats outliers as bugs in the data. The scientific discovery perspective, however, reframes them as potential features of reality that expose bugs in our *theories*. If a fraud detection model consistently flags a new type of transaction as an outlier, the data cleaning approach is to remove it to improve performance on "normal" transactions. The anomaly investigation approach is to ask, "What if this isn't fraud, but a new, legitimate form of customer behavior our model doesn't understand?" The first approach optimizes a model of the past; the second discovers a model of the future.

This also highlights a paradox of expertise. The very training that makes a scientist or data analyst proficient at operating within an existing paradigm—optimizing models, reducing error, confirming known patterns—can create epistemological blind spots. Their toolkit is designed to enforce conformity. This training might make them *less* likely to appreciate a data point that fundamentally breaks the pattern. The "prepared mind," therefore, is not just about accumulated knowledge but about a psychological flexibility—a willingness to question the foundational assumptions of one's own field and to sacrifice short-term model performance for the long-term possibility of discovering a new, more fundamental truth.[28]

Section 3: A Historical Tapestry of Discovery Through Anomaly

The assertion that anomalous data drives scientific progress is not merely a philosophical argument; it is a demonstrable historical fact. The annals of science are replete with landmark discoveries that began as outliers, errors, or unexpected noise in a dataset. These cases serve as powerful parables, providing concrete evidence for the immense value of data that defies expectation.

The following table provides a high-level summary of several such discoveries, crystallizing the report's evidentiary core by linking specific breakthroughs directly to the anomalous data that precipitated them.

Discovery	Scientist(s)	Year(s)	The Anomaly (The "Outlier" Data)	Significance
Penicillin	Fleming, Florey, Chain	1928–1942	A patch of mold (*Penicillium notatum*) is contaminating and killing bacteria in a petri dish.[31]	Revolutionized medicine, ushering in the age of antibiotics.[33]
Ozone Hole	Farman, Gardiner, Shanklin	1985	Persistently and "impossibly" low ozone readings over Antarctica, initially dismissed as instrument error.[35]	Led to the Montreal Protocol, a landmark global environmental treaty.[37]
Cosmic Microwave Background	Penzias & Wilson	1964	A persistent, uniform, low-level microwave "hiss" or "noise" from all directions in the sky that could not be eliminated.[39]	Provided the strongest empirical evidence for the Big Bang theory of the universe's origin.[41]
X-Rays	Wilhelm Röntgen	1895	A mysterious glow from a chemically coated screen caused by a nearby cathode ray tube, even when the tube was covered.[28]	Opened up a new part of the electromagnetic spectrum and revolutionized medical diagnostics.[31]
Radioactivity	Henri Becquerel	1896	Photographic plates fogging in a	Revealed the instability of the

			dark drawer when left near uranium salts, with no external light source.[31]	atom and led to the birth of nuclear physics.[31]

3.1 Case Study 1: The Contaminated Petri Dish and the Dawn of Antibiotics (Penicillin)

In September 1928, the Scottish bacteriologist Alexander Fleming returned to his laboratory at St. Mary's Hospital in London after a holiday. He began sorting through a stack of petri dishes containing *Staphylococcus* bacteria. On one dish, he observed an anomaly: it was contaminated with a blob of mold, identified as *Penicillium notatum*. The truly anomalous data, however, were not the mold itself, but the visible, clear ring surrounding it—a zone of inhibition where the bacteria had been destroyed and could not grow.[31]

A less observant or less curious researcher might have simply discarded this "spoiled" or "contaminated" culture as a failed experiment—a routine act of data cleaning to preserve the integrity of the research. Fleming, however, possessed the "prepared mind" to recognize the significance of this outlier. He did not see a ruined experiment; he saw a profound biological interaction. He hypothesized that the mold was producing a substance—a "mould juice"—that was actively killing the bacteria.[32] He famously noted, "When I woke up just after dawn on September 28, 1928, I certainly didn't plan to revolutionize all medicine... But I guess that was exactly what I did".[32] His focused investigation of this single, contaminated data point led directly to the discovery of penicillin, the world's first true antibiotic, a breakthrough that transformed medicine and has saved hundreds of millions of lives.[33]

3.2 Case Study 2: The Hole in the Sky — Data Dismissed and Rediscovered (The Antarctic Ozone Hole)

The discovery of the Antarctic ozone hole is perhaps the most direct and cautionary tale for the modern age of automated data analysis. In the early 1980s, scientists Joseph Farman, Brian Gardiner, and Jonathan Shanklin of the British Antarctic Survey were analyzing decades of data from a ground-based Dobson spectrophotometer at Halley Bay. They observed a shocking and precipitous decline in stratospheric ozone levels every Antarctic spring, with values dropping by as much as 50%.[35] The readings were so anomalously low that the scientists initially suspected their instrument was faulty. They spent months recalibrating and even replaced the instrument before concluding that the "impossible" data was real.[35]

Concurrently, NASA's Nimbus 7 satellite, equipped with the Total Ozone Mapping Spectrometer (TOMS), was orbiting the Earth and collecting millions of ozone data points. The

software processing this vast dataset had been programmed with quality-control algorithms designed to flag or reject data that fell outside a "reasonable" or historically expected range. The catastrophically low values over Antarctica were so far outside this range that they were being automatically flagged as outliers—presumed to be instrument errors—and were not being properly visualized by the analysis systems.[36]

This case has been the subject of some debate. One account holds that the data was automatically discarded. In contrast, another, from NASA's Ozone Processing Team, clarifies that the data was not discarded but flagged as suspicious and could not be verified against ground-based data, which was itself erroneous at the time.[36] Regardless of the precise mechanism, the outcome was the same: the automated system, built on prior assumptions about atmospheric chemistry, failed to recognize a developing global crisis. It was only after the British team, trusting their anomalous ground-based data, published their findings in *Nature* in 1985 that NASA scientists were prompted to re-examine their archived, flagged data. Upon doing so, they found that the "outliers" were not errors at all; they were the first evidence of a hole in the ozone layer the size of a continent.[36] This story stands as a chilling parable of the systemic risk of the "error filter": when we automate the removal or dismissal of data based on current knowledge, we build systems that are structurally blind to revolutionary new phenomena.

3.3 Case Study 3: The Persistent Hum of Creation — The Cosmic Microwave Background

In 1964, radio astronomers Arno Penzias and Robert Wilson were working with the ultra-sensitive Holmdel Horn Antenna at Bell Labs in New Jersey. Their goal was to conduct radio astronomy, but they were stymied by a persistent, low-level "noise" or "hiss" in their data. This signal was anomalous in three key ways: it was uniform, coming from all directions in the sky; it was constant, present day and night; and it was stubborn, remaining despite all their efforts to eliminate it.[39]

Penzias and Wilson treated the anomaly as a systematic error, a contamination of their dataset. They undertook an exhaustive data cleaning process. They checked their electronics, recalibrated their instruments, and famously captured and removed a pair of pigeons nesting in the antenna's throat. They meticulously cleaned out the "white dielectric material" (droppings) left behind, suspecting it could be the source of the interference.[39] Still, the anomalous noise persisted.

The breakthrough came not from within their own experiment, but from the collision of their anomalous data with an external theory. By chance, Penzias learned that a research group at nearby Princeton University, led by Robert Dicke, had theorized that if the universe had begun with a hot Big Bang, then a remnant, low-temperature radiation should still pervade all of space.[39] Penzias and Wilson's inexplicable noise was a perfect match for the predicted temperature of this Cosmic Microwave Background (CMB) radiation. The "noise" they had tried so hard to eliminate was, in fact, the signal—the afterglow of creation. This accidental

discovery provided the most powerful evidence for the Big Bang theory, transforming the field of cosmology from speculative theory to observational science.[41]

3.4 Further Evidence from the Annals of Science: A Survey of Transformative Anomalies

The pattern of discovery-by-anomaly is not limited to these cases but is a recurring theme throughout scientific history:

- **X-Rays:** In 1895, Wilhelm Röntgen was experimenting with a cathode ray tube shielded by black cardboard. He was astonished to see a nearby screen coated with a fluorescent chemical begin to glow. This anomalous emission, which could not be explained by known physics, was the first observation of X-rays. [28]
- **Radioactivity:** In 1896, Henri Becquerel was studying phosphorescence in uranium salts. He placed the salts on a photographic plate wrapped in black paper and stored them in a dark drawer. He later found the plate was fogged, as if it had been exposed to light. This anomalous energy, which emanated without an external source, was the discovery of radioactivity.[31]
- **Dark Matter:** In the 20th century, astronomers observed that the rotational speeds of stars in distant galaxies were anomalously high. The stars were moving so fast that the gravitational pull of the visible matter in the galaxy should have been insufficient to hold them in orbit. This glaring discrepancy between observation and theory led to the hypothesis of dark matter, an invisible substance that constitutes the majority of matter in the universe.[46]

These historical cases reveal a crucial strategy for the modern data scientist. In both the Ozone Hole and CMB discoveries, the breakthrough occurred when the anomalous data from one source was connected to a corroborating stream of information from another—be it a ground-based measurement or a theoretical prediction. This suggests that the proper response to a persistent, unexplainable outlier is not removal, but a proactive search for corroborating evidence, transforming the role of the data analyst from a janitor to a detective.

Section 4: Reconciling the Paradigms: A Modern Synthesis for Data Science

The history of science provides an unambiguous verdict: anomalous data, far from being a mere nuisance, is often the engine of discovery. This historical reality stands in stark contrast to the prevailing practices in many corners of the machine learning community. The challenge, therefore, is to reconcile these two paradigms—to integrate the lessons of history into the workflows of the modern data scientist, fostering an approach that balances the pragmatic need for model accuracy with the scientific imperative of curiosity.

4.1 Lessons from History for the Machine Learning Practitioner: Error vs. Epiphany

The historical case studies serve as direct analogs for the daily challenges faced by ML practitioners. Fleming's "contaminated" petri dish is the equivalent of a mislabeled data point or a noisy sample. The "impossible" ozone readings are the data points flagged by a Z-score test as being more than three standard deviations from the mean. The persistent "hiss" of the Cosmic Microwave Background is the residual error in a model that stubbornly refuses to converge to zero. History's unequivocal lesson is that the default action in these scenarios—to discard, ignore, or explain away the deviation—is frequently the wrong one.

The modern emphasis on optimizing for performance metrics like accuracy, precision, and recall creates a powerful institutional bias against the investigation of anomalies. The goal, implicitly and explicitly, becomes to make the model fit the bulk of the existing data as well as possible. This is a valuable goal for engineering and product development. However, the goal of science is often to understand precisely why a model *fails* on a few crucial data points. The relentless pursuit of a higher accuracy score can inadvertently filter out the very data that could lead to a more profound understanding and, ultimately, a better, more comprehensive model of the world.

4.2 Towards a More Nuanced Approach: From Outlier Removal to Anomaly Investigation

A new framework is needed to move data science beyond the simplistic "clean versus dirty" dichotomy. Informed by the historical record and the principles of robust statistics, this framework should guide practitioners through a more thoughtful, investigation-led process:

1. **Detect:** Employ standard statistical and visual methods (e.g., IQR, Z-scores, box plots) to identify potential outliers. This initial step remains unchanged.
2. **Triage:** Critically investigate the source of each potential outlier. Is there a plausible explanation rooted in the data collection or processing pipeline (e.g., a known instrument malfunction, a clear data entry typo)? If the outlier can be confidently identified as an error, it should be corrected. If correction is impossible, it should be removed with explicit justification and documentation.[15]
3. **Isolate & Analyze:** If the outlier is not a clear error, it must be treated as a potential anomaly. Isolate the point and its neighbors in the feature space. Do these points share common characteristics not captured by the current model? Do they represent a distinct sub-population or an emerging behavior? This step treats the outlier not as a single deviant point, but as a potential representative of an unmodeled group.
4. **Model Robustly:** Instead of removing the anomaly, employ robust statistical methods or models that are less sensitive to extreme values. This could involve using robust

regression techniques, tree-based models like Isolation Forests that naturally handle outliers, or weighting schemes that down-weight the influence of extreme points without discarding their information.[24] The goal is to build a model that reflects the majority without being blinded by the minority.

5. **Escalate:** In any scientific, R&D, or exploratory context, persistent and unexplainable anomalies should not be considered a data cleaning problem but a primary research opportunity. These findings should be flagged, documented, and escalated for deeper investigation by domain experts. They may represent the most valuable output of the entire analysis.

4.3 Conclusion: All Data is Equal, But Some Data is More Interesting

This report returns to the user's original assertion: that all data should be treated equally. In the sense that no data point should be summarily executed without a fair trial, this assertion is fundamentally correct. The automatic, uncritical removal of outliers is an act of intellectual censorship that a scientifically-minded community should resist. However, the historical record suggests an even more provocative conclusion. While all data should be given due process, some data is far more *interesting* than the rest. That data is almost always the data that doesn't fit.

The challenge for the machine learning community is to evolve from a culture of cleaning to a culture of curiosity. It is to build tools and establish best practices that facilitate not just anomaly *detection*, but anomaly *investigation*. The future of AI-driven scientific discovery, which aims to find novel patterns in vast datasets across domains like astrophysics, genomics, and climate science, may depend on this essential shift in perspective.[47] The ultimate goal should not be to build models that are merely accurate about the world we already know, but to build systems that are sensitive enough to alert us to the existence of a world we have yet to imagine. The outlier is not an obstacle to a better model; it is often the signpost pointing the way.

Works cited

1. What is the purpose of removing outliers from datasets before ..., accessed October 13, 2025, https://www.quora.com/What-is-the-purpose-of-removing-outliers-from-datasets-before-applying-machine-learning-algorithms-Can-we-use-them-without-removing-them-first
2. The Importance of Outlier Detection in Machine Learning: Methods and Implementation in Python | by Yennhi95zz | Medium, accessed October 13, 2025, https://medium.com/@yennhi95zz/the-importance-of-outlier-detection-in-machine-learning-methods-and-implementation-in-python-125e3d5ada7d
3. The Power of Clean Data: Why Data Cleaning is Key to Accurate Analytics - Medium, accessed October 13, 2025,

https://medium.com/womenintechnology/the-power-of-clean-data-why-data-cleaning-is-key-to-accurate-analytics-f1cee0238ffc
4. Data Cleaning: Definition, Benefits, And How-To - Tableau, accessed October 13, 2025, https://www.tableau.com/learn/articles/what-is-data-cleaning
5. Outlier Treatment: Understanding and Managing Anomalies in Machine Learning - Alooba, accessed October 13, 2025, https://www.alooba.com/skills/concepts/machine-learning/outlier-treatment/
6. Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging - PubMed, accessed October 13, 2025, https://pubmed.ncbi.nlm.nih.gov/26305321/
7. www.coursera.org, accessed October 13, 2025, https://www.coursera.org/articles/what-are-outliers#:~:text=Outliers%20are%20data%20points%20that%20lie%20outside%20the%20majority%20of,that%20misrepresent%20the%20data%20sample.
8. Outlier - Wikipedia, accessed October 13, 2025, https://en.wikipedia.org/wiki/Outlier
9. How to Find Outliers | 4 Ways with Examples & Explanation - Scribbr, accessed October 13, 2025, https://www.scribbr.com/statistics/outliers/
10. What Are Outliers in Data Sciences? | Coursera, accessed October 13, 2025, https://www.coursera.org/articles/what-are-outliers
11. A Basic Guide to Outliers - DataDrive, accessed October 13, 2025, https://godatadrive.com/blog/outliers-101
12. Video: Outlier in Statistics | Definition & Examples - Study.com, accessed October 13, 2025, https://study.com/learn/lesson/video/outlier-statistics-examples.html
13. Spotting the Exception: Classical Methods for Outlier Detection in Data Science - MachineLearningMastery.com, accessed October 13, 2025, https://machinelearningmastery.com/spotting-the-exception-classical-methods-for-outlier-detection-in-data-science/
14. Outlier detection is an essential technique in data science used to identify data points that... | by Adnan Mazraeh | Medium, accessed October 13, 2025, https://medium.com/@adnan.mazraeh1993/outlier-detection-is-an-essential-technique-in-data-science-used-to-identify-data-points-that-dc91b10854bc
15. The impact of outliers on Data: when to remove and when to retain | by Abhay singh, accessed October 13, 2025, https://medium.com/@abhaysingh71711/the-impact-of-outliers-on-data-when-to-remove-and-when-to-retain-fb6e474ddbd8
16. Anomaly detection - Wikipedia, accessed October 13, 2025, https://en.wikipedia.org/wiki/Anomaly_detection
17. What Is Data Cleansing & Why Is It Important? - Alteryx, accessed October 13, 2025, https://www.alteryx.com/glossary/data-cleansing
18. The critical role of data cleaning - DataScienceCentral.com, accessed October 13, 2025, https://www.datasciencecentral.com/the-critical-role-of-data-cleaning/
19. Outliers: To Drop or Not to Drop - The Analysis Factor, accessed October 13, 2025, https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/
20. machine learning - Should I remove outliers if accuracy and Cross ..., accessed

October 13, 2025, https://datascience.stackexchange.com/questions/42952/should-i-remove-outliers-if-accuracy-and-cross-validation-score-drop-after-remov
21. A survey of outlier detection methodologies - White Rose Research ..., accessed October 13, 2025, https://eprints.whiterose.ac.uk/id/eprint/767/1/hodgevj4.pdf
22. Anomalies: Disruption and Source of Knowledge (23. - 24.09.2018), accessed October 13, 2025, https://www.leopoldina.org/en/events/event/event/2612/
23. That s Odd! How Scientists Respond to Anomalous Data, accessed October 13, 2025, https://conferences.inf.ed.ac.uk/cogsci2001/pdf-files/1054.pdf
24. 3. Robust algorithms for Regression, Classification and Clustering ..., accessed October 13, 2025, https://scikit-learn-extra.readthedocs.io/en/stable/modules/robust.html
25. What Is A Scientific Anomaly? - Philosophy Beyond - YouTube, accessed October 13, 2025, https://www.youtube.com/watch?v=f_8YHh3CuKs
26. From Anomalies to Essential Scientific Revolution? Intrinsic Brain Activity in the Light of Kuhn's Philosophy of Science - PMC - PubMed Central, accessed October 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5328955/
27. Accidental scientific discoveries | Research Starters - EBSCO, accessed October 13, 2025, https://www.ebsco.com/research-starters/history/accidental-scientific-discoveries
28. Accidental Breakthroughs: The Role of Serendipity in Science - AZoLifeSciences, accessed October 13, 2025, https://www.azolifesciences.com/article/From-Mistake-to-Magic3b-Serendipity-in-Science-Discoveries.aspx
29. Role of chance in scientific discoveries - Wikipedia, accessed October 13, 2025, https://en.wikipedia.org/wiki/Role_of_chance_in_scientific_discoveries
30. The story of serendipity - Understanding Science, accessed October 13, 2025, https://undsci.berkeley.edu/the-story-of-serendipity/
31. Ten major breakthroughs that were happy accidents, accessed October 13, 2025, https://www.xprize.org/articles/ten-major-breakthroughs-that-were-happy-accidents
32. The real story behind penicillin | PBS News, accessed October 13, 2025, https://www.pbs.org/newshour/health/the-real-story-behind-the-worlds-first-antibiotic
33. Fleming Discovers Penicillin in Molds | Research Starters - EBSCO, accessed October 13, 2025, https://www.ebsco.com/research-starters/history/fleming-discovers-penicillin-molds
34. The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use - PMC, accessed October 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5403050/
35. Researchers Discover a Hole in the Ozone Layer - EBSCO, accessed October 13, 2025, https://www.ebsco.com/research-starters/environmental-sciences/researchers-di

scover-hole-ozone-layer
36. Debunking the myth about ozone holes, NASA, and outlier removal, accessed October 13, 2025, https://aakinshin.net/posts/outliers-ozon-holes/
37. The ozone hole - Discovering Antarctica, accessed October 13, 2025, https://discoveringantarctica.org.uk/oceans-atmosphere-landscape/atmosphere-weather-and-climate/the-ozone-hole/
38. NASA Data Aids Ozone Hole's Journey to Recovery, accessed October 13, 2025, https://www.nasa.gov/earth-and-climate/nasa-data-aids-ozone-holes-journey-to-recovery/
39. Confirming the Big Bang | Nokia.com, accessed October 13, 2025, https://www.nokia.com/bell-labs/about/history/innovation-stories/confirming-big-bang/
40. Cosmic Anniversary: 'Big Bang Echo' Discovered 50 Years Ago Today | Space, accessed October 13, 2025, https://www.space.com/25945-cosmic-microwave-background-discovery-50th-anniversary.html
41. en.wikipedia.org, accessed October 13, 2025, https://en.wikipedia.org/wiki/Cosmic_microwave_background#:~:text=The%20accidental%20discovery%20of%20the,work%20initiated%20in%20the%201940s.&text=The%20CMB%20is%20landmark%20evidence,the%20origin%20of%20the%20universe.
42. Astronomy 101: Cosmic microwave background, accessed October 13, 2025, https://www.astronomy.com/astronomy-for-beginners/astronomy-101-cosmic-microwave-background/
43. Penicillin: 83 Years Ago Today | Columbia University Mailman School of Public Health, accessed October 13, 2025, https://www.publichealth.columbia.edu/research/center-infection-and-immunity/penicillin-83-years-ago-today
44. World of Change: Antarctic Ozone Hole - NASA Earth Observatory, accessed October 13, 2025, https://earthobservatory.nasa.gov/world-of-change/Ozone
45. WMAP Big Bang CMB Test - NASA, accessed October 13, 2025, https://map.gsfc.nasa.gov/universe/bb_tests_cmb.html
46. 18 Anomaly Examples (2025) - Helpful Professor, accessed October 13, 2025, https://helpfulprofessor.com/anomaly-examples/
47. Anomaly Detection in Machine Learning: Examples, Applications ..., accessed October 13, 2025, https://www.ibm.com/think/topics/machine-learning-for-anomaly-detection
48. [2503.02112] Building Machine Learning Challenges for Anomaly Detection in Science, accessed October 13, 2025, https://arxiv.org/abs/2503.02112