Raw Parsed


```
{
  "query": "data analysis",
  "num_results": 3,
  "results": [
    {
      "chunk": "and incorporating data validation processes. 6 Risk analysis – Performing a risk analysis before incorporating a new data source is an essential step in mitigating the impact of changes in data sources. This analysis involves identifying the potential risks associated with the data source, which we have covered in Section 3.1. The analysis should be comprehensive, considering both technical and non-technical aspects of the data source, and should ideally include potential solutions for the identified risks. This will often force you to face the hard truth and will lead you to decide that the candidate data sources are not adequate or reliable enough. Trade-offs will nevertheless need to be considered, depending on the use case at hand. Monitoring – Monitoring everything that is relevant is another crucial step in mitigating the impact of changing data sources. It involves tracking various aspects of the data sources, the machine learning models, and their outputs to detect and respond to changes promptly. Draft a list of variables and quantities that must be continuously tracked to ensure that the models remain reliable and accurate over time. For this, inspiration can be drawn from the discussed topics in Section 3.1, but it will vary from use case to use case, as well as the nature of the models that have been used. Supervised models, for example, can be tested against a reference test set or a historical reference model; if the accuracy, precision or recall starts to deviate significantly from this reference set, it should be flagged. On the other hand, monitoring the performance of unsupervised models can be more challenging, because there is no clear performance measure that can be directly computed. One approach is to monitor the model's ability to detect patterns and clusters in the data. It is possible to use a reference test set or reference model for this, but the informative metrics – e.g. cluster similarity, homogeneity, separation... – are more abstract and somewhat harder to interpret. Another approach is to visualize projections of certain interesting data points in the learned latent spaces or preferably a reduction thereof, which greatly benefits interpretability but makes it harder to convert it into hard numbers. As a suggestion, a good balance between interpretability and hard performance metrics is found when clusters are tested against pre-existing domain knowledge, e.g. by listing similar data points for given queries. Simply monitoring whether expected similarities emerge or not can provide powerful signals about model and data performance. Another effective approach is to create proxy supervised tasks that rely on the output of the unsupervised model. Monitoring the model's performance on such proxy tasks can provide insights into the quality and usefulness of the unsupervised model's output. Diversification – Diversifying data sources is another important measure, but is easier said than done. One challenge of using multiple data sources is the potential for conflicts or inconsistencies between the sources. Different data sources may have different formats, schemas, and levels of quality, which can create discrepancies and inconsistencies that must be resolved before the data can be used in the model.",
      "metadata": {
        "paper_id": "2306.04338v1",
        "chunk_index": 9,
        "total_chunks": 13
      },
      "distance": 1.1549514532089233,
      "rank": 1
    },
```

```
{
  "chunk": "into official statistical production, one can benefit from the strengths of both approaches and make more informed decisions based on the most current and accurate data [10]. 2.2 External data sources Let’s focus on the data sources that will power such machine learning models. Their nature, size, structure, frequency... can be vastly different, they must typically be gathered ‘in the wild’ and should often be combined with each other to extract meaningful insights. Compared to more traditional data sources for official statistics, they may present unique and appealing characteristics such as: Broad-spectrum – Covers a wide variety of topics. Diversity – A large variety of sources to cover different perspectives. Availability – Lots of data is freely and easily accessible. Size – Some datasets can be enormous, sometimes even complete. Structure – Not only tabular data, but also images, video, text, audio, etc. Timeliness – (Near) up-to-date and real-time information. Frequency – Raw data on various, even very fine-grained time scales. Granularity – Raw data on various, even fine-grained levels of detail. Coverage – Various locations and regions can be filtered and covered. On the other hand, before all this data is ready to be exerted for machine learning and official statistical production, a few challenges need to be overcome, such as: Data quality – Data may contain errors, biases, or missing values that need to be addressed to ensure accuracy and reliability. Data interpretation – Understanding the context and meaning of data can be difficult, especially when dealing with unstructured data such as text or images. Data integration – Combining data from different sources with varying structures and formats can be challenging and time-consuming. Selection bias – Proper randomization or compiling representative population samples can be challenging, and it greatly depends on the underlying data origins. Operationalization bias – Reproducibility can be difficult as it depends on many implicit, hidden, and/or production- specific design choices [11, 12]. Computational resources – Processing and analyzing large amounts of data may require significant computational resources. Privacy and security – Sensitive data may need to be protected and anonymized to ensure privacy and security. Data ethics – Data collection and use should adhere to ethical principles. Fairness and justness – The end solution should ideally be as neutral as possible and should not discriminate [13]. Cost – All of the above requires resources, budgets and a talented workforce. In addition, the data source itself might need to be purchased. In 2016, McKinsey reported that many companies have started to specialize in acquiring and selling data [14]. 3 With the right tools, workforce, technological advances, mindset, and legislative support, these challenges can and should be manageable. The most challenging piece of the puzzle, however – and one that is more than often ignored – is the lack of control you can exert over the data sources that are externally gathered. As a national statistics agency, traditionally, survey data and administrative records that power official statistics are completely under your own control. But once you start exploiting external data sources to power novel, innovative, complimentary or ersatz",
  "metadata": {
    "paper_id": "2306.04338v1",
    "chunk_index": 3,
    "total_chunks": 13
  },
  "distance": 1.2355127334594727,
  "rank": 2
},
{
  "chunk": "its saturation point using its (actually empirical) learning curve. It shows the learning curve of two learners, the blue curve and the green dashed curve.
```

```
      The latter has the best saturation performance. The available data marks a
      restriction to this process. In contrast, data acquisition considers the available
      data as the decision variable, and it stops collecting data when all learners have
      reached their saturation performance. 2. Role of Available Data: In early stopping,
      the available amount of training data is a given constraint under which early
      stopping operates, and data acquisition precisely seeks to control this quantity in
      an economically optimal fashion. 3. Used Performance Curve: Early Stopping uses a
      single learning curve of learner a to decide upon early stopping of the training of
      a. Data acqui- sition uses the learning curves of all learners under consideration
      (i.e., a finite set A) and considers the budget-wise best performance achievable (by
      any learner). To clarify this difference, consider also the right (green) part of
      Fig. 9. The learner a may be the best solution available given the 1500 data points
      (in this case, a barely needs 500 of them to attain saturation performance). However,
      if more data were available, then at least one other learner could take advantage of
      that additional data and outperform a. The figure shows the curve of an optimal
      learner a∗∈arg mina∈A µa,∞that has the best performance if no limit is posed on the
      available training data (as in Cortes et al (1994)). Only after 3000 instances, no
      learner will improve the overall possible performance anymore; therefore, at this
      point, the data acquisition process stops. 4.1.3 Early Discarding In many setups, the
      learner itself is a matter of optimisation. Consider the situation where we have a
      (possibly infinite) set of learners, e.g., a finite set of algorithms, each of which
      can be instantiated with a possibly infinite number of hyper-parametrisations. The
      task is to find the (hyper-parametrised) learner which performs best for the given
      data of size n in the sense that it creates, Springer Nature 2021 LATEX template 26
      Learning Curves for Decision Making on average, the best model. Formally, if A is the
      (infinite) set of parametrised learners, the goal is to find arg min a∈A C(a, n). (7)
      This task is commonly known as model selection. While it is uncommon in literature to
      be so explicit and describe the model selection problem through the value of the
      learning curve at some sample size, this formulation is rather precise and
      insightful. It emphasises that which learner is best might depend on the number of
      available training points. Note that one needs to separate some portion of the data
      for validation in practice to estimate model performances. In other words, most
      approaches in practice do not even address the above problem but instead arg min a∈A
      C(a, ⌈αn⌋), (8) where α ∈]0, 1[ (open interval) is the training portion, typically
      between 70% and 90%, where the remaining portion of 1 −α is used to estimate C(a,
      ⌈αn⌋), typically in some (possibly repeated) hold-out validation. We could",
    "metadata": {
      "paper_id": "2201.12150v2",
      "chunk_index": 24,
      "total_chunks": 64
    },
    "distance": 1.289002537727356,
    "rank": 3
  }
 ]
}
```