Multi-documentation Summarization of Science Articles

Justin To, Luka Liu, Milan Dean

W 266: Natural language Processing
UC Berkeley School of Information

April 14, 2023

Abstract

In today's age of information overload, the ability to generate accurate and concise summaries of large bodies of text is more important than ever. This study focuses on the task of abstractive multi-document summarization (MDS), which involves generating a summary based on multiple related documents. Two state-of-the-art pre-trained large language models (LLMs), Centrum and LED, were selected, and different strategies, including fine-tuning and model-stacking, were explored for building an MDS model customized for Multi-XScience, a dataset previously unseen by the pre-trained models. In evaluating the summaries generated, we adopted both quantitative ROUGE-based scoring and qualitative analysis of the model outputs, in order to understand how the models adapted to unique elements specific to the task and dataset. By studying the process and strategies for adapting pre-trained models to domain-specific MDS tasks, we aim to contribute to the research on adapting pre-trained models for new and domain specific MDS tasks and datasets.

1 Introduction

MDS is the task of summarizing multiple texts into a concise and informative summary. Compared to single document summarization, it is challenging in the context of abstractive MDS to keep the summary coherent and comprehensive, given the documents are longer[1] and of a more complex structure[2]. Despite the challenges, MDS has a potentially wider application as real-world tasks often involve collating and summarizing information from multiple sources. An important question therefore is, given the advances in NLP and the availability of pre-trained LLMs, how should we build a customized abstractive MDS model if we are given a domain-specific dataset previously unseen by the models?

In this paper, we employ the Multi-XScience dataset, a relatively recent dataset that has not been widely used for pre-training/fine-tuning LLMs, to simulate the situation where we are given an MDS task for a broad domain, e.g. academic writing from the sciences. The defining characteristics of the dataset, namely (a) the significant portion of samples with long inputs (up to ∼6300 tokens) and (b) that different articles within each sample are related yet not covering the exact same event[3], led us to pick two models for experimentation. The first is the Longformer Encoder Decoder (LED) model, which was proposed by Beltagy et al. [1] and introduced the concept of local vs global attention so that attention-based transformers can still effectively handle the computation of long inputs. The second is the Centrum model, proposed by Puduppully and Steedman [6], for its capability of handling long inputs and inclusion of a centroid-based approach for dealing with multiple source documents.

Leveraging on the available pre-trained checkpoints, we first tested the off-the-shelf LED and Centrum using the Multi-XScience dataset and then proceeded to fine-tune them for improved performance. Then, we explored the use of a two-step setup whereby individual source articles of each sample are first shortened through summarization, and then combined again for MDS processing. For evaluation, we employed the common ROUGE score metric, showing the improved performance from fine-tuning (ROUGE-2 from 5.2 to 6.9; ROUGE-L from 14.6 to 17.8, which are close to state-of-the-art (SOTA)). Further analyses were conducted on the variation of performance across inputs of different lengths, and the amount of "copying" from the main article. Lastly, a qualitative analysis was done on 25 samples for each model tested, and the fine-tuned and two-step models were found to demonstrate capabilities in extracting information from and contrasting different sources, while adapting its writing style to one suited for the dataset.

2 Related Works

In recent years, there has been significant research in the field of MDS, with a particular focus on the use of transformer-based models for abstractive models and through approaches such as sparse attention mechanisms, hierarchical sentence representations, and document-level clustering. Much progress has also been made in providing datasets for MDS training and evaluation.

On model development, the LED model proposed by Beltagy et al. [1] addresses the issue of processing long documents by introducing a sparse attention mechanism that allows the model to scale to documents of up to 16384 tokens. This approach has outperformed previous models, including advanced ones such as PEGASUS, on long-document tasks such as summarization.

Branching off to models more specialized in MDS, PRIMERA, proposed by Xiao et al. citexiao2022primera, is a pre-training method for transformer-based models that leverages hierarchical sentence representations to improve per-

[1]For instance, the widely used CNN/Daily Mail dataset has an average input length of around 780 tokens. In contrast, around 30% of the inputs in the Multi-XScience dataset has over 1024 tokens, while the longest sample has over 6300 tokens.

[2]In MDS, the source articles are written by multiple authors with varying writing styles and article structures.

[3]Further discussed in Section 3.1

formance on MDS tasks. The approach involves constructing a sentence-level pyramid structure and applying a masked language modeling objective to the pyramid, leading to improved results over BART, PEGASUS, and LED.

Advancing the research on MDS further, Puduppully and Steedman [6] proposed a centroid-based pre-training method for MDS that leverages document-level clustering to capture document-level semantics. This approach showed improved performance on summarization tasks over previous models including PRIMERA, particularly for MDS datasets with a large number of documents.

On the dataset front, a commonly used dataset in the MDS domain is perhaps the Multi-news dataset, which was introduced by Fabbri et al. [2] and contains news articles from multiple sources. Other MDS datasets include Wikisum, which was generated by Liu et al. [4] from Wikipedia articles, WCEP compiled by Ghalandari at el. [3] based on news summaries from the Wikipedia Current Events Portal, and the Multi-XScience dataset by Lu et al. [5] used in this study.

3 Methods

3.1 Dataset

We used the Multi-XScience MDS dataset [5] in this study, which is a collection of scientific articles from scientific fields, including computer science, physics, and biology. Some key parameters of the dataset are set out as follows, while Appendix A summarizes other findings and visualizations from our EDA, as well as the data preprocessing steps adopted for the experiments.

The dataset is chosen for two reasons. First, Multi-XScience is among the few datasets previously unseen by MDS models during pre-training[4]. Picking datasets already seen by the models could pose difficulties in evaluation, not to mention that it would go against the overall goal to explore MDS model-building strategies for unseen data.

	Training	Validation	Test
Sample size	30369	5066	5093
Input Token Length	899 565 82-4694	898 567 79-4183	886 547 131-6348
Samples with inputs > 1024 tokens	31.21%	30.71%	30.87%
Label Token Length	142 58 24-735	141 58 25-324	142 58 26-418
Articles per sample	4.43 2.62 2-21	4.43 2.63 2-20	4.39 2.55 2-19

Table 1: Dataset description. The cells with three columns refer to mean/standard deviation/range

The second reason is that Multi-XScience is a more challenging task that is more comparable to real-world data collation and summarization. Many of the datasets (e.g. Multi-news, WCEP) are news-based, with the articles within a sample covering the same news event and key elements to be included in the summary appearing in multiple sources. The Multi-XScience dataset however, expects the model to write the related works section of a journal paper based on (a) the abstract of that paper which serves to provide context; and (b) the abstracts of the other journal articles that the main paper referenced. While these additional abstracts are usually[5] closely related, they often concern different aspects of an issue, and the model needs to be able to compare and contrast the similarities and differences[6]. This difficulty of the dataset also shows up empirically[7] in [8].

[4]LED's pre-training data included Multi-News while Centrum's pre-training included both Multi-News and WikiSum.

[5]Some exceptions exist, e.g. sample 845 (refer to Appendix D for sample review for sample review)) of the test set includes an irrelevant article on the design attributes of luggage carriers when the main article is on a MDS over online product reviews. See Appendix C for details.

[6]In contrast, the commonly used Multi-news dataset has, in each sample, multiple news reports surrounding a single event, and the overall theme to be covered in the summary is generally present in every source article.

[7]For instance, the zero- and few-short evaluation of BART, PEGASUS, LED and PRIMERA yielded a

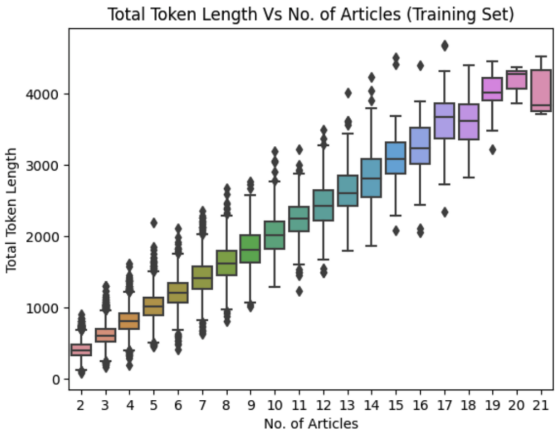

Figure 1: Total token length as a function of the number of articles in the training data set. The relationship between token length and the number of articles is linear.

As can be seen from the table above, the input length of the Multi X-Science dataset shows great variability, with some inputs so short that they can fit in more traditional summarization models, while others exceed even 4096, which is the max input length for Centrum. The same variability can also be observed in terms of the number of articles per sample (see Figure 1 below), with the two distributions showing a close-to-linear pattern. This provides an opportunity for us to use the Multi X-Science dataset to explore possible differences in model performance between short and long samples.

3.2 Models

Due to resource constraints, we focused on two pre-trained LLMs, namely LED and Centrum for our experiments. Of the three recent and well-performing models we reviewed, we did not pick PRIMERA for two reasons - (a) Centrum is, based on literature review, the best-performing model while LED is a common starting point for both PRIMERA and Centrum, meaning it could be treated as a more advanced baseline; and (b) the authors of PRIMERA already provided performance results of PRIMERA on the Multi-XScience test set.

We adopted two different baselines for performance comparison. The first one, dubbed the "baseline model" is simply a lead-based model, whereby the first 3 sentences are chosen as the summary and is a common primary strategy implemented in various papers (e.g. [5], [7] and [9]). This also provides a useful baseline for comparing the degree of "copying" (see Section 3.3) from the main article which is an undesirable trait for the chosen dataset. The second baseline is the publicly available checkpoint for LED, which serves as the base architecture for recent MDS LLMs.

On top of the 2 baseline models, we ran experiments on the performance of the following models, with details on their fine-tuning and inference settings provided in Appendix B:

1. an alternative checkpoint of LED fine-tuned on the arXiv dataset
2. publicly available checkpoint of Centrum;
3. a version of the baseline LED[8] that we fine-tuned on the Multi-XScience training set;
4. a similarly fine-tuned version of Centrum; and
5. a two-step model stacking the fine-tuned LED and Centrum.

Of these experiments, models (c) and (d) stem from our plan to test out the performance of fine-tuning strategies for MDS. During the process, we noticed that despite the larger input length of Centrum (4096 tokens), there are still cases where the inputs have to be truncated and information omitted. We therefore explored under model (e) a two-step model stacking process whereby the source articles are first condensed

ROUGE-2 score of only 1.9 to 4.6 and a ROUGE-L score of 9.9 to 15.7 for Multi-XScience, while the score ranges for Multi-news is 3.7 to 13.6 for ROUGE-2 and 10.4 to 20.8 for ROUGE-L. A similar gap remains even after fine-tuning of the models.

[8]In theory, it would be better to fine-tune the alternative LED checkpoint which (a) is a larger version of LED with double the parameters; and (b) has been fine-tuned on text from the scientific fields. However, the GPU we had was unable to train the larger model while maintaining the max input token length at a reasonable level.

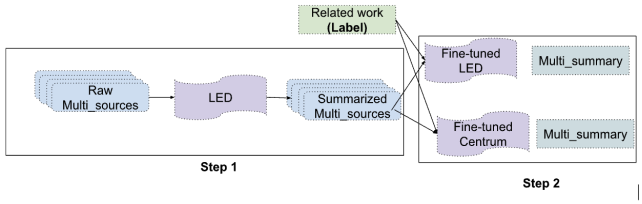

Figure 2: Two-steps model design

by our fine-tuned LED individually followed by an MDS process by (a) the fine-tuned LED or (b) the fine-tuned Centrum. The idea is to investigate if Centrum performs better when presented with more concise sources and without token truncation, see Figure 2. Refer to Appendix B for more details on the two step model.

3.3 Evaluation metrics

As in the standard practice for summarization, the primary metric we use is the ROUGE score. Among the different versions of ROUGE, we chose ROUGE-2 and ROUGE-L, with a particular focus on the latter given the prevalence of compound nouns in academic writings[9], for gauging model performance. Compared to ROUGE-2, ROUGE-L is better suited for this task. In calculating these metrics, precision, recall, and f-measure are all considered[10], and scores were computed for the overall test dataset, as well as the subsets of short, medium, and long token length samples[11].

[9]This is an empirical observation from viewing samples in the dataset, and is quite intuitive considering academic terms are often ones glued together from multiple words (e.g. the theme of "Synchronous Optical Network Ring Assignment Problem" in sample 4820 of the test set, and "fixed-size ordinally forgetting encoding" in sample 3157).

[10]We also look only at the mean score but not the lower and upper ranges due to the limited scope of our study, though we recognize that looking at the variation in performance across samples is in itself a valid future investigation direction.

[11]We divided the 5093 test samples into three groups. "Short" samples consisted of 1273 samples with less than 486 tokens (lower quartile) for the inputs; "medium" with 2546 samples with input lengths between 486 and 735 (upper quartile) tokens; and the remaining 1274 samples with input lengths greater than 1150 tokens are grouped into the "long" samples.

Unlike other MDS datasets where the key information is expected to be found in most, if not all, of the source articles, the Multi X-Science dataset has a different structure whereby the main article (i.e. first source) is expected to provide general context while the other sources will each provide a related, yet different angle of the topic. It is desirable for models to extract information not just from the main article, and we therefore measured the degree of "copying" from the main article by repurposing standard ROUGE calculations[12].

Finally, we supplemented our evaluation with a qualitative analysis whereby 25 random samples were selected and we reviewed the summaries written by the baseline LED and the 5 models experimented, judging them based on fluency, ability to extract information from multiple sources, and contrast information from the main article and other sources, as well as factual accuracy.

4 Results and Discussion

4.1 Overall performance

In the following table, we report the ROUGE-2 (R-2) and ROUGE-L (R-L) scores of the 2 baselines plus 5 models we tested, with precision (P), recall (R), and f-measure (F) scores presented. Scores for overall, short, medium, and long samples are also presented.

As shown in the tables above, the fine-tuned models (LED and Centrum) provided a markedly improved performance over the baselines and off-the-shelf models. In fact, even though the models were only fine-tuned for up to 2 epochs[13], our figures quickly approached the SOTA figures reported in [8] (i.e. ROUGE-2 of 6.8; ROUGE-L of 18.2) demonstrating the viability of LED and Centrum to be adopted for previously unseen MDS data. When considering the breakdowns, the improvement for longer samples was much

[12]Implemented using the standard rouge function by replacing the target reference that the function takes in with the main article of the sample.

[13]Training details and considerations are at Appendix B.

more visible, indicating the MDS-specific model architectures are perhaps more suitable for tasks with longer input lengths.

This result is consistent with what our qualitative analysis (Appendix C). For example, in sample 485, the Centrum model was able to extract the key ideas of 3 relevant studies and contrast them with the main article, forming a coherent summary, while the baseline and off-the-shelf models were overwhelmed by the large number of articles (9 in total) and choose only to copy one of the articles in full. This is very different from shorter samples, e.g. sample 4160 with only 3 articles, where the off-the-shelf models were still able to prepare a summary covering elements, albeit in a more extractive manner without efforts to change the tone or merge the extracted parts together.

For the two-step model, the ROUGE scores do not show any advantage of the approach over simple fine-tuned models. However, as we observe from the qualitative analysis (e.g. samples 831 and 5068), there is a strong tendency for the two-step model to highlight differences and similarities between the main and other articles. However, the downside is that the two-step model made quite a number of factual fallacies in the process[14].

4.2 Degree of Copying

To assess the degree of copying from the main article, we calculated a re-purposed ROUGE-L score. As shown in the table below, the fine-tuned and two-step models exhibited a significantly lower degree of "copying", which is supported by the qualitative analysis showing that these models were able to extract information from multiple articles. In contrast, the baseline LED model showed the highest degree of "copying", often copying substantial portions of the main article to the point that the number of

[14]One possible reason is that we need better tuning on the amount and manner of information fed into the 2nd step model, but resouce limitations were a main obstacle considering the generation of the first step answers alone took over 14 hours.

Model	R-2	Overall	Short	Medium	Long
Baseline (Lead)	P	4.50	4.78	4.68	3.87
	R	8.13	6.12	7.87	10.7
	F	5.40	*5.02*	5.56	5.48
Baseline (LED)	P	4.52	4.12	4.75	4.47
	R	6.99	8.17	7.07	5.63
	F	*5.21*	5.2	5.40	*4.83*
LED	P	5.25	5.32	5.30	5.09
	R	6.31	6.97	6.34	5.57
	F	5.39	5.58	5.44	5.11
Centrum	P	4.14	4.02	4.23	4.09
	R	8.12	9.48	8.14	6.70
	F	5.24	5.36	*5.34*	4.93
Our LED	P	8.75	8.03	8.67	9.63
	R	5.89	5.21	5.69	6.94
	F	6.56	5.83	6.41	7.62
Our Centrum	P	7.62	7.10	7.58	8.28
	R	7.27	6.38	7.13	8.44
	F	**6.88**	**6.14**	**6.78**	**7.82**
2-Step (Centrum)	P	5.87	5.66	5.92	5.99
	R	6.67	6.15	6.62	7.26
	F	5.86	5.42	5.87	6.28

Table 2: Rouge-2 Scores for Models Tested (Best in **bold**; worst is *italicized*)

Model	R-2	Overall	Short	Medium	Long
Baseline (Lead)	P	12.4	14.9	12.7	9.17
	R	21.7	18.8	21.3	25.2
	F	14.6	15.5	15	*12.9*
Baseline (LED)	P	12.9	11.6	13.3	13.5
	R	20.5	23.8	20.4	17.3
	F	14.9	14.8	15.1	14.6
LED	P	15.1	15.4	15.1	14.3
	R	18.6	20.3	18.7	16.7
	F	15.7	16.2	15.7	15.0
Centrum	P	11.3	10.8	11.4	11.7
	R	22.7	25.5	22.8	19.7
	F	*4.4*	*14.4*	*14.6*	14.2
Our LED	P	23.6	24.2	23.5	23.2
	R	15.9	15.7	15.5	16.8
	F	17.7	17.6	17.4	**18.4**
Our Centrum	P	19.8	20.9	19.6	19.1
	R	18.9	18.4	18.7	19.7
	F	**17.8**	**17.9**	**17.6**	18.1
2-Step (Centrum)	P	17.0	18.0	17.0	15.9
	R	19.6	19.7	19.5	19.6
	F	17.0	17.4	17.0	16.7

Table 3: Rouge-L Scores for Models Tested (Best in **bold**; worst is *italicized*)

Model	R-2	Overall	Short	Medium	Long
Baseline (Lead)	P	39.7	56.6	39.7	22.7
	R	40.7	41.1	39.0	43.9
	F	36.8	44.9	36.8	28.5
Baseline (LED)	P	76.9	83.1	86.9	50.7
	R	75.5	92.8	81.5	46.3
	F	*74.2*	*85.6*	*81.9*	*47.2*
LED	P	45.3	58.8	43.9	34.7
	R	33.6	44.3	31.5	27.0
	F	37.4	48.9	35.6	29.5
Centrum	P	48.2	56.5	48.6	39.2
	R	57.5	73.7	55.7	45.1
	F	51.2	62.4	50.8	41.0
Our LED	P	33.6	35.5	34.7	29.5
	R	13.5	13.0	13.2	14.7
	F	**18.3**	**18.0**	**18.2**	18.7
Our Centrum	P	29.4	32.9	29.4	26.1
	R	17.3	16.4	16.9	19.0
	F	20.5	20.7	20.1	20.9
2-Step (Centrum)	P	24.7	27.9	25.0	21.0
	R	17.2	17.1	16.8	18.0
	F	19.1	20.1	18.9	**18.6**

Table 4: Rouge-L Scores for Measuring "Copying" (Best in **bold**; worst is *italicized*)

copied sentences exceeded even that of the lead-based baseline.

Although the ROUGE scores did not reveal any significant "copying" for the off-the-shelf LED and Centrum models, it is noteworthy that these models occasionally copied from a single reference article rather than the main article (e.g. see the output of the base Centrum for sample 4371). This indicates that there is still room for improving the "copying" metric, such as by taking the minimum of the ROUGE-L scores compared to each input article (rather than just the main article) to discourage "copying" from a single source.

4.3 Other observations

The qualitative analysis also allowed us to make the following observations on the models, including:

1. The fine-tuned and two-step models showed clear signs of learning common writing styles (using phrases such as "there is a large body of work...") which is desirable in real world applications;

2. For samples containing irrelevant articles, the fine-tuned models were able to safely navigate away from them (e.g. sample 845)

3. Despite having been fine-tuned on journal articles, the models showed greatest difficulties when presented with highly technical papers that use common words in uncommon meanings (e.g. in mathematics in sample 4858). This shows the potential to enhance performance by adding another fine-tuning step on domain-specific texts. Alternatively, the two-step model has continued to perform well in such cases which shows the potential of such an approach.

5 Conclusion

In conclusion, we demonstrated the viability of building custom-based MDS models through fine-tuning which is able to quickly approach SOTA scores even with simple and limited tuning. Through breakdown and qualitative analyses of the results, we highlighted the importance of considering the token length of input texts when selecting a text summarization model. Our results suggest that the fine-tuned LED and Centrum models are much more adept at MDS for longer texts, while the two-step model shows potential especially for highly technical texts despite needing more tuning. Future work could focus on further optimizing the tuning techniques (e.g. fine-tuning with domain-specific texts first before the MDS data), improving on the two-step approach, as well as looking at the performance stability of the models across data samples.

Appendix

A EDA and Data Preprocessing

This appendix provides further findings from the EDA on the Multi-XScience dataset, as well as the data preprocessing procedures undertaken in the experiments.

As mentioned above, one standard summary could be summarized from more than 20 source articles. With concatenating all sources of articles, the total token length is about 4096. In addition, we've examined the token length of these standard summaries (label) and provided the details below.

To determine the input length, we look at the average length of the documents we want to summarize. With the consideration of the various numbers of source documents and our GPU capabilities, we decided to take input numbers of tokens = 512 as an experiment and numbers of tokens = 4096 (max tokens in the training dataset) as our final model.

For the output length, considering the desired level of detail in the summary and the distribution of standard summaries (label) tokens, we decided a summary length between 100-250 may be sufficient.

Multi-XScience includes a main abstract, which is from the main article; a standard summary of the related works with a separation of "@cite", which is the summary of multi-source documentations, label datasets; and reference abstracts, which are the multiple abstracts from a different source of works. We processed the dataset by defining the separator of documents as "|||||", a document separator, then concatenating reference abstracts using the document separator to replace citation references.

B Model

B.1 Baseline Model

Copying the first 3 sentences from each reference abstract was used as our baseline. It is a simple and easy-to-implement approach that provides a starting point for summarizing multiple documents. It assumes that the first few sentences of a document contain important information that should be included in the summary.

This approach can serve as a useful baseline for several reasons:

1. It is easy to implement and does not require a lot of computational resources. This makes it a good starting point for developing more advanced summarization models.

2. It is straightforward to understand and explain. This makes it a good choice for initial experiments and evaluations.

3. It can provide a quick estimate of the performance of a summarization system. By comparing the summaries generated by this approach to human-written summaries or other machine-generated summaries, it is possible to get a rough idea of how well the system is performing.

B.2 Baseline Base LED (1K and 16K) Model

The base LED model has ("allenai/led-base-16384"), with max input token length of 16384 & ∼200M parameters and consists of 12 transformer encoder layers and 2 transformer decoder layers. This model is designed to be computationally efficient and suitable for low-resource environments. The baseline LED was generated based on a pre-trained LED model by Beltagy, L (2020) [1] on X-science test datasets with input tokens varying from 1k and 6k.

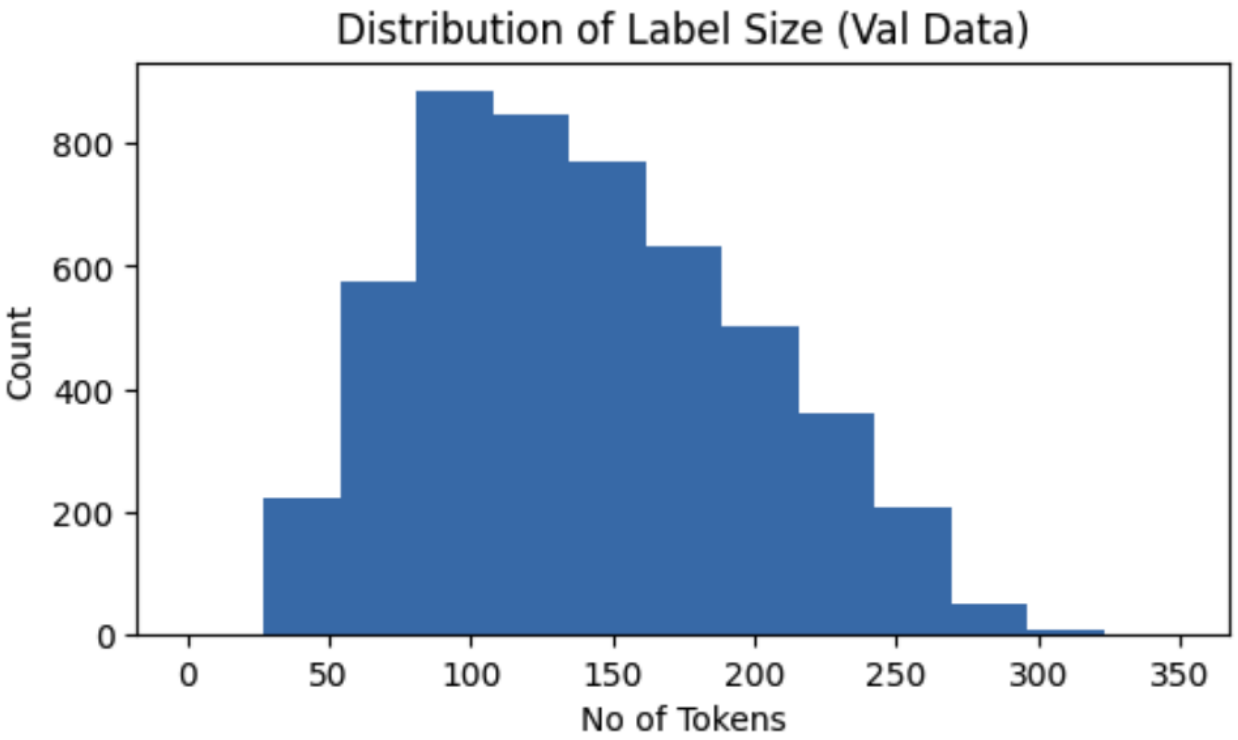

Figure 3: Distribution of the number of tokens in the label of training data. The number of tokens is right-skewed and centered around 100.

B.3 Baseline Large LED (1K and 16K) Model

The "large" in the model name indicates that it has more parameters and is larger than the "base" LED model. The model has ("allenai/led-large-16384-arxiv"), with max input token length of 16384 & ∼512M parameters, and consists of 24 transformer encoder layers and 4 transformer decoder layers. This model has been pre-trained for general summarization tasks. The baseline LED was generated based on a pre-trained large LED and evaluated on X-science test datasets.

B.4 Baseline Centrum Model

The "ratishsp/Centrum" model with max input token length of 4096 & ∼192M parameters and consists of 12 transformer encoder layers and 12 transformer decoder layers. It is smaller than some of the larger transformer models, such as T5 and GPT-3, but larger than some of the smaller transformer models, such as the BART base. The publicly available Centrum checkpoint was built upon the LED model architecture using the 4096 token version.This means Centrum is not able to take in 16384 tokens.

B.5 Fine-tuned LED Model

Fine-tuning the "LED-large-16384-arxiv" model to the X-science train (size 30369) and validation (size 5066) datasets specifically for the related work summarization. We used 4096 max number of tokens in tracking and validation sets, and 256 max output tokens with a batch size of 2 (due to limited GPU resources). The model is then used as a cross-entropy loss function for 2 epochs, and evaluated on test datasets(size 5093) with ROUGH metrics.

B.6 Fine-tuned Centrum Model

Fine-tuning the "ratishsp/Centrum" model to the X-science train and validation datasets

specifically for the related works summarization. We used 4096 max number of tokens in tracking and validation sets, and 256 max output tokens with a batch size of 16 (due to limited GPU resources). The model is then used as a cross-entropy loss function for 2 epochs, and evaluated on test datasets with ROUGH metrics.

B.7 Two-step LED Model

In the first step, the LED model is used to generate a summary of each related source document. By summarizing each source document, the model can condense the information and reduce redundancy, making it easier to process and analyze. For the first step, none of the individual source articles in the X-science dataset contain more than 4096 tokens, so there is no issue of the first step model (i.e. the fine-tuned LED) receiving truncated inputs.

In the second step, generate the summary based on the fine-tuned LED model to refine the summaries generated in the first step. In this way, we were hoping that the fine-tuned LED model can learn to generate even more accurate and relevant summaries.

B.8 Two-step Centrum Model

The same first step as the two-step LED Model above. In the second step, generate the summary based on the fine-tuned Centrum model to refine the summaries generated in the first step. In this way, we were hoping that the fine-tuned Centrum model can learn to generate even more accurate and relevant summaries.

B.9 Two-step Model results

However, from our evaluation and observation of this approach, the performance of the two-step model does not exceed the fine-tuned LED/Centrum model. One potential issue is that the large amount of information contained in the source documents may make it difficult for the model to generate concise and informative summaries. Ultimately, the performance is heavily dependent on the performance of step 1.

Overall, our results demonstrate the importance of utilizing advanced techniques like fine-tuning to unlock the full potential of language models in real-world applications.

C Qualitative Analysis

We also investigated how the model performed when summarizing short, medium, and long documents. The study aimed to evaluate the quality of the generated summaries and identify any trends or patterns in the model's performance based on the length of the source documents. A total of 25 randomly drawn samples (from the 5093 samples in the X-Science test dataset) are analyzed. These 25 samples include:

- 10 short samples, i.e. with total token lengths below the lower quartile
- 10 medium samples, i.e. with total token lengths between the lower and upper quartiles
- 5 long samples, i.e. with token lengths above the upper quartile

Our analysis of the review (refer to the Analysis Appendix for more detailed review) results showed that the fine-tuned LED and Centrum learned the structure for multi-documents summary for X-science dataset, with particularly good results for longer documents (ref. Sample 485) while for shorter samples the advantage over the baseline and off-the-shelf models appear to be less as those models are able to extract information from multiple sources to some degree, albeit with limited ability to merge them into a single coherent summary (ref. Sample 4160).

Additionally, we observed that LED and Centrum seem to have difficulties in digesting math topics, where the use of common language for unusual meanings might have confused the models (refer to no 4858). The two-step model somehow helps for these samples, but the downside is that the two-step model hallucinates quite a bit.

In general, LED and Centrum never really hallucinate. To future improve this issue, we would suggest that further research in this area should consider:

- Add domain-specific knowledge: This can include adding specialized math dictionaries or knowledge bases to your training data.
- Use special tokens for math symbols: This can help the model recognize and differentiate between mathematical expressions and regular language. For example, use a special token such as "[math]" to indicate the start of a math expression and "[/math]" to indicate the end.
- Use more training data: Consider adding more math-specific data to the training set to help the model learn to recognize and understand math concepts.
- Fine-tune your model: fine-tuning it on a smaller dataset of math-specific documents. This can help the model learn to recognize and summarize math-related information more effectively.

Overall, our study highlights the importance of considering the length of source documents, and the training datasets topics when evaluating the performance of multi-document summarization models.

Future research in this area should focus on developing models that can effectively summarize long documents; and provide the model with more specialized knowledge to help the model recognize and differentiate between regular language and math expressions, improving its ability to summarize math-related documents effectively. Refer to the manual review documentation in more detail link.

D Detailed Review Samples

Refer to the following pages.

References

[1] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[2] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. A large-scale multi-document summarization dataset from the wikipedia current events portal, 2020.

[4] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences, 2018.

[5] Yao Lu, Yue Dong, and Laurent Charlin. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8068–8074, Online, November 2020. Association for Computational Linguistics.

[6] Ratish Puduppully and Mark Steedman. Multi-document summarization with centroid-based pretraining, 2022.

[7] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks, 2017.

[8] Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. Primera: Pyramid-based masked sentence pre-training for multi-document summarization, 2022.

[9] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
43	**(1):** We propose a self-supervised learning framework for visual odometry (VO) that incorporates correlation of consecutive frames and takes advantage of adversarial learning. Previous methods tackle self-supervised VO as a local structure from motion (SfM) problem that recovers depth from single image and relative poses from image pairs by minimizing photometric loss between warped and captured images. As single-view depth estimation is an ill-posed problem, and photometric loss is incapable of discriminating distortion artifacts of warped images, the estimated depth is vague and pose is inaccurate. In contrast to previous methods, our framework learns a compact representation of frame-to-frame correlation, which is updated by incorporating sequential information. The updated representation is used for depth estimation. Besides, we tackle VO as a self-supervised image generation task and take advantage of Generative Adversarial Networks (GAN). The generator learns to estimate depth and pose to generate a warped target image. The discriminator evaluates the quality of generated image with high-level structural perception that overcomes the problem of pixel-wise loss in previous methods. Experiments on KITTI and Cityscapes datasets show that our method obtains more accurate depth with details preserved and predicted pose outperforms state-of-the-art self-supervised methods significantly. **(2):** We present an unsupervised learning framework for the task of monocular depth and camera motion estimation from unstructured video sequences. In common with recent work [10, 14, 16], we use an end-to-end learning approach with view synthesis as the supervisory signal. In contrast to the previous work, our method is completely unsupervised, requiring only monocular video sequences for training. Our method uses single-view depth and multiview pose networks, with a loss based on warping nearby views to the target using the computed depth and pose. The networks are thus coupled by the loss during training, but can be applied independently at test time. Empirical evaluation on the KITTI dataset demonstrates the effectiveness of our approach: 1) monocular depth performs comparably with supervised methods that use either ground-truth pose or depth for training, and 2) pose estimation performs favorably compared to established SLAM systems under comparable input settings.	Despite its feasibility, self-supervised VO still underperforms supervised ones. Apart from the effectiveness of direct supervision, a key reason is that they focus mainly on geometric properties @cite but pay little attention to the sequential nature of the problem. In these methods, only a few frames (no more than 5) are processed in the network, while previous estimations are discarded and the current estimation is made from scratch. Instead, the performance can be enhanced by taking geometric relations of sequential observations into account.	**Rouge scores:** - Rouge 2: 0.0188 (prec) 0.0361 (recall) 0.0247 (f-1) - Rouge L: 0.1242 (prec) 0.2381 (recall) 0.1633 (f-1) **Summary:** We propose a self-supervised learning framework for visual odometry (VO) that incorporates correlation of consecutive frames and takes advantage of adversarial learning. Previous methods tackle self-supervision VO as a local structure from motion (SfM) problem that recovers depth from single image and relative poses from image pairs by minimizing photometric loss between warped and captured images. As single-view depth estimation is an ill-posed problem, and photometric loss is incapable of discriminating distortion artifacts of warped images, the estimated depth is vague and pose is inaccurate. In contrast to previous methods, our framework learns a compact representation of frame-to-frame correlation, which is updated by incorporating sequential information. The updated representation is used for depth estimation. Besides, we tackle VO as a self-Supervised image generation task and take advantage of Generative Adversarial Networks (GAN). The generator learns to estimate depth and pose to generate a warped target image. The discriminator evaluates the quality of generated image **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0351 (prec) 0.0482 (recall) 0.0406 (f-1) - Rouge L: 0.1217 (prec) 0.1667 (recall) 0.1407 (f-1) **Summary:** we present an unsupervised learning framework for the task of monocular depth and camera motion estimation from unstructured video sequences. our method uses single-view depth and multiview pose networks, with a loss based on warping nearby views to the target using the computed depth and pose. the generator learns to estimate depth and pose to generate a warped target image. the discriminator evaluates the quality of generated image with high-level structural perception that overcomes the problem of pixel-wise loss in previous methods. experiments on the KITTI and Cityscapes datasets show that our method obtains more accurate depth with details preserved and predicted pose outperforms state-of-the-art self-supervised methods significantly. **Comment:** - First section came from 2nd article, while parts in red came from the 1st article - No clear sign of MDS as the text does not distinguish the sources	**Rouge scores:** - Rouge 2: 0.0333 (prec) 0.0602 (recall) 0.0429 (f-1) - Rouge L: 0.0993 (prec) 0.1786 (recall) 0.1277 (f-1) **Summary:** "We present an unsupervised learning framework for the task of monocular depth and camera motion estimation from unstructured video sequences. In common with recent work [10, 14, 16], we use an end-to-end learning approach with view synthesis as the supervisory signal. In contrast to the previous work, our method is completely unsupervised, requiring only monocular video sequences for training. Our method uses single-view depth and multiview pose networks, with a loss based on warping nearby views to the target using the computed depth and pose. The networks are thus coupled by the loss during training, but can be applied independently at test time. Empirical evaluation on the KITTI dataset demonstrates the effectiveness of our approach: 1) monocular depth performs comparably with supervised methods that use either ground-truth pose or depth for training, and 2) pose estimation performs favorably compared to established SLAM systems under comparable input settings." **Comment:** - Copied from the second article only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.1591 (prec) 0.0833 (recall) 0.1094 (f-1) **Summary:** In @cite, the authors proposed an end-to-end learning framework for monocular depth and camera motion estimation from unstructured video sequences. In contrast to supervised methods, our framework learns to estimate depth and pose by minimizing photometric loss between warped and captured images. **Comment:** - Contrasts 2nd article's approach (green) with 1st article's approach (yellow) - Clear sign of MDS - Words in red appeared in the articles but cannot factuality cannot be verified based on inputs alone	**Rouge scores:** - Rouge 2: 0.028 (prec) 0.0482 (recall) 0.0354 (f-1) - Rouge L: 0.0972 (prec) 0.1667 (recall) 0.1228 (f-1) **Summary:** @cite proposed an unsupervised learning framework for the task of monocular depth and camera motion estimation from unstructured video sequences. Their method uses an end-to-end learning approach with view synthesis as the supervisory signal. In contrast to the previous methods, our method learns a compact representation of frame-to-frame correlation, which is updated by incorporating sequential information. Besides, we tackle VO as a self-supervised image generation task and take advantage of Generative Adversarial Networks (GAN). The generator learns to estimate depth and pose to generate a warped target image. The discriminator evaluates the quality of generated image with high-level structural perception that overcomes the problem of pixel-wise loss in previous methods. Experiments on KITTI and Cityscapes datasets show that our method obtains more accurate depth with details preserved and predicted pose outperforms state-of-the-art methods significantly. **Comment:** - Contrasts 2nd article's approach (green) with 1st article's approach (yellow) - Clear sign of MDS - No wrong or unverifiable facts, but copying quite a lot	**Rouge scores:** - Rouge 2: 0.0143 (prec) 0.012 (recall) 0.0131 (f-1) - Rouge L: 0.1127 (prec) 0.0952 (recall) 0.1032 (f-1) **Summary:** In @cite, a fully unsupervised learning framework for monocular depth and camera motion estimation from unstructured video sequences is proposed. Unlike @cite and @cite which require only monocular video sequences for training, our method is completely unsupervised and can be applied independently at test time. In addition, we do not rely on a single view as the supervisory signal, and we are able to train the camera in a supervised manner. **Comment:** - Appears to contrast 2nd article (green) with the main one - But despite the paragraph structure, the red part actually came from 2nd article

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
483	**(1):** This paper builds upon the current methods to increase their capability and automation for 3D surface construction from noisy and potentially sparse point clouds. It presents an analysis of an artificial neural network surface regression and mapping method, describing caveats, improvements and justification for the different approach. **(2):** Shape-from-shading (SFS) methods tend to rely on models with few parameters because these parameters need to be hand-tuned. This limits the number of different cues that the SFS problem can exploit. In this paper, we show how machine learning can be applied to an SFS model with a large number of parameters. Our system learns a set of weighting parameters that use the intensity of each pixel in the image to gauge the importance of that pixel in the shape reconstruction process. We show empirically that this leads to a significant increase in the accuracy of the recovered surfaces. Our learning approach is novel in that the parameters are optimized with respect to actual surface output by the system. In the first, offline phase, a hemisphere is rendered using a known illumination direction. The isophotes in the resulting reflectance map are then modelled using Gaussian mixtures to obtain a parametric representation of the isophotes. This Gaussian parameterization is then used in the second phase to learn intensity-based weights using a database of 3D shapes. The weights can also be optimized for a particular input image.	In a slightly different problem, where a NN is used to reconstruct the shape of a 3D object from its shading in a 2D @cite . show from experiment that quantitative improvement does not necessarily lead to quantitative improvement. This is something to consider when using a 'black box' function like a neural network, especially where there could be some information loss. In this regard we must ensure that the final model is representative of the ground truth and not only rely on an error measure. It is suggested that more research must be done for 3D surface quality metrics @cite . Visual quality will be assessed in the method presented here alongside quantitative results in the absence of quality metrics.	**Rouge scores:** - Rouge 2: 0.0643 (prec) 0.0932 (recall) 0.0761 (f-1) - Rouge L: 0.1337 (prec) 0.1933 (recall) 0.1581 (f-1) **Summary:** This paper builds upon the current methods to increase their capability and automation for 3D surface construction from noisy and potentially sparse point clouds. It presents an analysis of an artificial neural network surface regression and mapping method, describing caveats, improvements and justification for the different approach.||||Shape-from-shading (SFS) methods tend to rely on models with few parameters because these parameters need to be hand-tuned. This limits the number of different cues that the SFS problem can exploit. In this paper, we show how machine learning can be applied to an SFS model with a large number of parameters. Our system learns a set of weighting parameters that use the intensity of each pixel in the image to gauge the importance of that pixel in the shape reconstruction process. We show empirically that this leads to a significant increase in the accuracy of the recovered surfaces. Our learning approach is novel in that the parameters are optimized with respect to actual surface output by the system. In the first **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0743 (prec) 0.0932 (recall) 0.0827 (f-1) - Rouge L: 0.1409 (prec) 0.1765 (recall) 0.1567 (f-1) **Summary:** this paper builds upon the current methods to increase their capability and automation for 3d surface construction from noisy and potentially sparse point clouds. it presents an analysis of an artificial neural network surface regression and mapping method, describing caveats, improvements and justification for the different approach. shape-from-shading (sf) methods tend to rely on models with few parameters because these parameters need to be hand-tuned. this limits the number of different cues that the sf problem can exploit. in this paper , we show how machine learning can be applied to an sf model with a large number of parameters. our system learns a set of weighting parameters that use the intensity of each pixel in the image to gauge the importance of that pixel in the shape reconstruction process. we show empirically that this leads to a significant increase in the accuracy of the recovered surfaces. **Comment:** - Copied the first tokens only, and even with errors (red) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0505 (prec) 0.0932 (recall) 0.0655 (f-1) - Rouge L: 0.1187 (prec) 0.2185 (recall) 0.1538 (f-1) **Summary:** In this article, we're going to explain how machine learning can be applied to an artificial neural network surface regression and mapping method, describing caveats, improvements and justification for the different approach. This paper builds upon the current methods to increase their capability and automation for 3D surface construction from noisy and potentially sparse point clouds. It presents an analysis of an artificial Neural Network surface regression and mapped method, describing caveat, improvements, and justification for a different approach. in this paper, we show how machine learning is applied to an SFS model with a large number of parameters. Our system learns a set of weighting parameters that use the intensity of each pixel in the image to gauge the importance of that pixel in the shape reconstruction process. We show that this leads to a significant increase in the accuracy of the recovered surfaces. Our learning approach is novel in that the parameters are optimized with respect to actual surface output by the system. In the first, offline phase, a hemisphere is rendered using a known illumination direction. The isophotes in the resulting reflectance map are then modelled using Gaussian mixtures to obtain a parametric representation of the isophotes. This Gaussian parameterization is then used in the second phase to learn intensity-based weights using a database **Comment:** - Initial parts shows some contrasts between 1st (green) and 2nd yellow articles - Some sign of MDS, but goes on to just copy everything else from 2nd article (red)	**Rouge scores:** - Rouge 2: 0.082 (prec) 0.0424 (recall) 0.0559 (f-1) - Rouge L: 0.2581 (prec) 0.1345 (recall) 0.1768 (f-1) **Summary:** In @cite, a hemisphere is rendered using a known illumination direction. The isophotes in the resulting reflectance map are then modelled using Gaussian mixtures to obtain a parametric representation of the isophotes. This is then used in the second phase to learn intensity-based weights using a database of 3D shapes. The weights can also be optimized for a particular input image. **Comment:** - Summarized the 2nd article only - The first article is probably too short to provide sufficient inputs? - Weak indication of MDS	**Rouge scores:** - Rouge 2: 0.0806 (prec) 0.0424 (recall) 0.0556 (f-1) - Rouge L: 0.3016 (prec) 0.1597 (recall) 0.2088 (f-1) **Summary:** In @cite, the authors propose a method to learn a set of parameters that use the intensity of each pixel in the image to gauge the importance of that pixel in the shape reconstruction process. They show empirically that this leads to a significant increase in the accuracy of the recovered surface. The weights can also be optimized for a particular input image. **Comment:** - Summarized the 2nd article only - The first article is probably too short to provide sufficient inputs? - Weak indication of MDS	**Rouge scores:** - Rouge 2: 0.0843 (prec) 0.0593 (recall) 0.0697 (f-1) - Rouge L: 0.2143 (prec) 0.1513 (recall) 0.1773 (f-1) **Summary:** In @cite, an SFS model is used to learn the intensity of each pixel in the image. The intensity of the image is used as the input to the SFS model. The weights can also be optimized for a particular input image. In this work, we show how machine learning can be applied to a SFS model with a large number of parameters. We also show how the weighting parameters can be used to learn intensity-based weights using a database of 3D shapes. **Comment:** - Summarized the 2nd article only - The first article is probably too short to provide sufficient inputs? - Weak indication of MDS

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
556	**(1):** In a disaster situation, first responders need to quickly acquire situational awareness and prioritize response based on the need, resources available and impact. Can they do this based on digital media such as Twitter alone, or newswire alone, or some combination of the two? We examine this question in the context of the 2015 Nepal Earthquakes. Because newswire articles are longer, effective summaries can be helpful in saving time yet giving key content. We evaluate the effectiveness of several unsupervised summarization techniques in capturing key content. We propose a method to link tweets written by the public and newswire articles, so that we can compare their key characteristics: timeliness, whether tweets appear earlier than their corresponding news articles, and content. A novel idea is to view relevant tweets as a summary of the matching news article and evaluate these summaries. Whenever possible, we present both quantitative and qualitative evaluations. One of our main findings is that tweets and newswire articles provide complementary perspectives that form a holistic view of the disaster situation. **(2):** Twitter is often considered to be a useful source of real-time news, potentially replacing newswire for this purpose. But is this true? In this paper, we examine the extent to which news reporting in newswire and Twitter overlap and whether Twitter often reports news faster than traditional newswire providers. In particular, we analyse 77 days worth of tweet and newswire articles with respect to both manually identified major news events and larger volumes of automatically identified news events. Our results indicate that Twitter reports the same events as newswire providers, in addition to a long tail of minor events ignored by mainstream media. However, contrary to popular belief, neither stream leads the other when dealing with major news events, indicating that the value that Twitter can bring in a news setting comes predominantly from increased event coverage, not timeliness of reporting.	Researchers have examined the question of whether Twitter can replace newswire for breaking news @cite . They studied a period of 77 days in 2011 during which 27 events occurred. The biggest disasters in this event-set are: an airplane crash with 43 deaths, and a magnitude 5.8 earthquake in Virginia that caused infrastructural damage. None of these disasters, bad as they are, rise to the level of the Nepal Earthquake(s) of 2015 in which almost 10,000 lives were lost. They collected a large dataset of tweets and news articles, but then eliminated a large collection of tweets based on clustering. More elimination of tweets led to only 97 linked tweet-news article pairs, which is a small dataset.	**Rouge scores:** - Rouge 2: 0.0419 (prec) 0.0583 (recall) 0.0488 (f-1) - Rouge L: 0.1548 (prec) 0.2149 (recall) 0.1799 (f-1) **Summary:** In a disaster situation, first responders need to quickly acquire situational awareness and prioritize response based on the need, resources available and impact. Can they do this based on digital media such as Twitter alone, or newswire alone, or some combination of the two? We examine this question in the context of the 2015 Nepal Earthquakes. Because newswire articles are longer, effective summaries can be helpful in saving time yet giving key content. We evaluate the effectiveness of several unsupervised summarization techniques in capturing key content.We propose a method to link tweets written by the public and newswire articles, so that we can compare their key characteristics: timeliness, whether tweets appear earlier than their corresponding news articles, and content. A novel idea is to view relevant tweets as a summary of the matching news article and evaluate these summaries. Whenever possible, we present both quantitative and qualitative evaluations. One of our main findings is that tweets and newswire are complementary perspectives that form a holistic view of	**Rouge scores:** - Rouge 2: 0.049 (prec) 0.0417 (recall) 0.045 (f-1) - Rouge L: 0.1845 (prec) 0.157 (recall) 0.1696 (f-1) **Summary:** in a disaster situation, first responders need to quickly acquire situational awareness and prioritize response based on the need, resources available and impact. can they do this based on digital media such as Twitter alone, or newswire alone, or some combination of the two? we examine this question in the context of the 2015 Nepal Earthquakes. we propose a method to link tweets written by the public and newswire articles, so that we can compare their key characteristics: timeliness, whether tweets appear earlier than their corresponding news articles, and content. we evaluate the effectiveness of several unsupervised summarization techniques in capturing key content. **Comment:** - Copied from the first article only but with a reordering of two sentences (red) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0625 (prec) 0.0833 (recall) 0.0714 (f-1) - Rouge L: 0.1366 (prec) 0.1818 (recall) 0.156 (f-1) **Summary:** Twitter has become a useful source of real-time news, potentially replacing newswire for this purpose. But is this true? In this paper, we examine the extent to which news reporting in newswire and Twitter overlap and whether Twitter often reports news faster than traditional newswire providers. In particular, we analyse 77 days worth of tweet and newswire articles with respect to both manually identified major news events and larger volumes of automatically identified news events. Our results indicate that Twitter reports the same events as newswire providers, in addition to a long tail of minor events ignored by mainstream media. However, contrary to popular belief, neither stream leads the other when dealing with major news events, indicating that the value that Twitter can bring in a news setting comes predominantly from increased event coverage, not timeliness of reporting. One of our main findings is that tweets and newswire provide complementary perspectives that form a holistic view of the disaster situation. **Comment:** - Except for the last sentence (yellow), all parts are from 2nd article with some paraphrasing - Summary presented the studies as the same one, i.e. no indication of MDS	**Rouge scores:** - Rouge 2: 0.0431 (prec) 0.0417 (recall) 0.0424 (f-1) - Rouge L: 0.1709 (prec) 0.1653 (recall) 0.1681 (f-1) **Summary:** The authors of @cite study the extent to which news reporting in newswire and Twitter overlap and whether Twitter often reports news faster than traditional newswire providers. They analyse 77 days worth of tweet and newswire articles with respect to both manually identified major news events and larger volumes of automatically identified news events. Their results indicate that Twitter reports the same events as newswire providers, in addition to a long tail of minor events ignored by mainstream media. However, contrary to popular belief, neither stream leads the other when dealing with major news events, indicating that the value that Twitter can bring in a news setting comes predominantly from increased event coverage, not timeliness of reporting. **Comment:** - All parts are from 2nd article - However, it is represented as a separate article from the main one, i.e. weak indication of MDS	**Rouge scores:** - Rouge 2: 0.0536 (prec) 0.05 (recall) 0.0517 (f-1) - Rouge L: 0.177 (prec) 0.1653 (recall) 0.1709 (f-1) **Summary:** @cite examined the extent to which news reporting in newswire and Twitter overlap and whether Twitter reports news faster than traditional newswire providers. analysed 77 days worth of tweet and newswire articles with respect to both manually identified major news events and larger volumes of automatically identified news events. Their results indicate that Twitter reports the same events as newswire providers, in addition to a long tail of minor events ignored by mainstream media. However, contrary to popular belief, neither stream leads the other when dealing with major news events, indicating that the value that Twitter can bring in a news setting comes predominantly from increased event coverage, not timeliness of reporting. **Comment:** - Essentially the same as Tuned LED, except for the initial words - Weak indication of MDS	**Rouge scores:** - Rouge 2: 0.0111 (prec) 0.0083 (recall) 0.0095 (f-1) - Rouge L: 0.1868 (prec) 0.1405 (recall) 0.1604 (f-1) **Summary:** In @cite, the authors found that newswire providers report the same events as newswire providers, in addition to a long tail of minor events ignored by mainstream media. However, they found that Twitter does not report the same event as newswire, which suggests that the importance of news reporting in a newswire setting comes from increased event coverage, not timeliness of reporting. The authors conclude that Twitter is a useful source of real-time news, but it is not clear whether this is the case for both newswire and newswire providers. **Comment:** - Weak indication of MDS, with the 2nd article presented as not the main article - The first red part is factually wrong, while the second part is not implied from the inputs

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
599	**(1):** This paper presents a discovery that the length of the entities in various datasets follows a family of scale-free power law distributions. The concept of entity here broadly includes the named entity, entity mention, time expression, aspect term, and domain-specific entity that are well investigated in natural language processing and related areas. The entity length denotes the number of words in an entity. The power law distributions in entity length possess the scale-free property and have well-defined means and finite variances. We explain the phenomenon of power laws in entity length by the principle of least effort in communication and the preferential mechanism. **(2):** Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. Unfortunately, the detection and characterization of power laws is complicated by the large fluctuations that occur in the tail of the distribution—the part of the distribution representing large but rare events—and by the difficulty of identifying the range over which power-law behavior holds. Commonly used methods for analyzing power-law data, such as least-squares fitting, can produce substantially inaccurate estimates of parameters for power-law distributions, and even in cases where such methods return accurate answers they are still unsatisfactory because they give no indication of whether the data obey a power law at all. Here we present a principled statistical framework for discerning and quantifying power-law behavior in empirical data. Our approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov-Smirnov (KS) statistic and likelihood ratios. We evaluate the effectiveness of the approach with tests on synthetic data and give critical comparisons to previous approaches. We also apply the proposed methods to twenty-four real-world data sets from a range of different disciplines, each of which has been conjectured to follow a power-law distribution. In some cases we find these conjectures to be consistent with the data, while in others the power law is ruled out.	Our work is related to Zipf's law and the distributions of word length and sentence length. Power laws have been observed to appear in numerous natural and man-made systems @cite , we here concern them in language.	**Rouge scores:** - Rouge 2: 0.0402 (prec) 0.1892 (recall) 0.0664 (f-1) - Rouge L: 0.0914 (prec) 0.4211 (recall) 0.1502 (f-1) **Summary:** This paper presents a discovery that the length of the entities in various datasets follows a family of scale-free power law distributions. The concept of entity here broadly includes the named entity, entity mention, time expression, aspect term, and domain-specific entity that are well investigated in natural language processing and related areas. The entity length denotes the number of words in an entity. The power law distributions in entity length possess the scale-free property and have well-defined means and finite variances. We explain the phenomenon of power laws in entity length by the principle of least effort in communication and the preferential mechanism.||||Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. Unfortunately, the detection and characterization of power laws is complicated by the large fluctuations that occur in the tail of the distribution—the part of the distribution representing large but rare events—and by the difficulty of identifying the range over which power-law behavior **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0127 (prec) 0.0541 (recall) 0.0206 (f-1) - Rouge L: 0.0759 (prec) 0.3158 (recall) 0.1224 (f-1) **Summary:** this paper presents a discovery that the length of the entities in various datasets follows a family of scale-free power law distributions. the power law distributions in entity length possess the scale-free property and have well-defined means and finite variances. we explain the phenomenon of power laws in entity length by the principle of least effort in communication and the preferential mechanism. our approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov-Smirnov (KS) statistic and likelihood ratios. we evaluate the effectiveness of the approach with tests on synthetic data and give critical comparisons to previous approaches. finally, we apply the proposed methods to twenty-four real-world data sets from a range of different disciplines, each of which has been conjectured to follow a power-law distribution. in some cases we find these conjectures to be consistent with the data, while in others the power law is ruled **Comment:** - First section came from 1st article, while parts in red came from the 2nd article - No clear sign of MDS as the text does not distinguish the sources	**Rouge scores:** - Rouge 2: 0.0315 (prec) 0.1892 (recall) 0.0541 (f-1) - Rouge L: 0.0762 (prec) 0.4474 (recall) 0.1303 (f-1) **Summary:** Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. Unfortunately, the detection and characterization of power laws is complicated by the large fluctuations that occur in the tail of the distribution—the part of the distribution representing large but rare events—and by the difficulty of identifying the range over which power-law behavior holds. This paper presents a discovery that the length of the entities in various datasets follows a family of scale-free power law distributions. The concept of entity here broadly includes the named entity, entity mention, time expression, aspect term, and domain-specific entity that are well investigated in natural language processing and related areas. The entity length denotes the number of words in an entity. The power law distributions in entity length possess the scale-free property and have well-defined means and finite variances. We explain the phenomenon of power laws in entity length by the principle of least effort in communication and the preferential mechanism. Power-law distributions are not limited to natural phenomena. They can also be found in man-made phenomenon, such as man-made climate change, and in natural history, such as natural gas and other natural phenomena. However, they can be difficult to detect due to the large **Comment:** - Only the 1st article (yellow) used in describing the study, while 2nd article (green) only used to provide general context - Red part is also from 2nd article and somehow appeared again despite the idea is also covered in the green part - No clear sign of MDS	**Rouge scores:** - Rouge 2: 0.0175 (prec) 0.027 (recall) 0.0213 (f-1) - Rouge L: 0.1034 (prec) 0.1579 (recall) 0.125 (f-1) **Summary:** In @cite, the authors present a principled statistical framework for discerning and quantifying power-law distribution in empirical data. The authors present a method for discerning power-law distributions based on the Kolmogorov-Smirnov (KS) statistic and likelihood ratios. They evaluate the effectiveness of the approach with tests on synthetic data and give critical comparisons to previous approaches. **Comment:** - All parts are from 2nd article - However, it is represented as a separate article from the main one, i.e. weak indication of MDS	**Rouge scores:** - Rouge 2: 0.0093 (prec) 0.027 (recall) 0.0138 (f-1) - Rouge L: 0.0826 (prec) 0.2368 (recall) 0.1224 (f-1) **Summary:** @cite present a statistical framework for discerning and quantifying power law behavior in empirical data. Their approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov-Smirnov (KS) statistic and likelihood ratios. They evaluate the effectiveness of the approach with tests on synthetic data and give critical comparisons to previous approaches. They also apply the proposed methods to twenty-four real-world data sets from a range of different disciplines, each of which has been conjectured to follow a power law distribution. In some cases they find these conjectures to be consistent with the data, while in others the power law is ruled out. **Comment:** - All parts are from 2nd article. Similar but longer than Tuned LED model - However, it is represented as a separate article from the main one, i.e. weak indication of MDS	**Rouge scores:** - Rouge 2: 0.0122 (prec) 0.027 (recall) 0.0168 (f-1) - Rouge L: 0.0723 (prec) 0.1579 (recall) 0.0992 (f-1) **Summary:** In @cite, the authors present a statistical framework for discerning and quantifying power-law distributions. The authors present a method that combines maximum-likelihood fitting with the Kolmogorov Smirnov (KS) statistic and likelihood ratios. The authors demonstrate that their method can be used with synthetic data and give critical comparisons to previous approaches. However, the authors do not provide a quantitative comparison of their method with the KKS method, and they do not provide an empirical comparison of their results with KKS methods. **Comment:** - All parts are from 2nd article. Similar to the other tuned models - Red part contains hallucinated information

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
620	**(1):** In this work we investigate the reasons why Batch Normalization (BN) improves the generalization performance of deep networks. We argue that one major reason, distinguishing it from data-independent normalization methods, is randomness of batch statistics. This randomness appears in the parameters rather than in activations and admits an interpretation as a practical Bayesian learning. We apply this idea to other (deterministic) normalization techniques that are oblivious to the batch size. We show that their generalization performance can be improved significantly by Bayesian learning of the same form. We obtain test performance comparable to BN and, at the same time, better validation losses suitable for subsequent output uncertainty estimation through approximate Bayesian posterior. **(2):** We investigate a local reparameterizaton technique for greatly reducing the variance of stochastic gradients for variational Bayesian inference (SGVB) of a posterior over model parameters, while retaining parallelizability. This local reparameterization translates uncertainty about global parameters into local noise that is independent across datapoints in the minibatch. Such parameterizations can be trivially parallelized and have variance that is inversely proportional to the mini-batch size, generally leading to much faster convergence. Additionally, we explore a connection with dropout: Gaussian dropout objectives correspond to SGVB with local reparameterization, a scale-invariant prior and proportionally fixed posterior variance. Our method allows inference of more flexibly parameterized posteriors; specifically, we propose variational dropout, a generalization of Gaussian dropout where the dropout rates are learned, often leading to better models. The method is demonstrated through several experiments.	The improved methods that we propose are also closely related to variational drop -out @cite as discussed below. We give a new interpretation to variational dropout and apply it in combination with normalization techniques.	**Rouge scores:** - Rouge 2: 0.0068 (prec) 0.0303 (recall) 0.0111 (f-1) - Rouge L: 0.0676 (prec) 0.2941 (recall) 0.1099 (f-1) **Summary:** In this work we investigate the reasons why Batch Normalization (BN) improves the generalization performance of deep networks. We argue that one major reason, distinguishing it from data-independent normalization methods, is randomness of batch statistics. This randomness appears in the parameters rather than in activations and admits an interpretation as a practical Bayesian learning. We apply this idea to other (deterministic) normalization techniques that are oblivious to the batch size. We show that their generalization performance can be improved significantly by Bayesian learning of the same form. We obtain test performance comparable to BN and, at the same time, better validation losses suitable for subsequent output uncertainty estimation through approximate Bayesian posterior.||||We investigate a local reparameterizaton technique for greatly reducing the variance of stochastic gradients for variational Bayesian inference (SGVB) of a posterior over model parameters, while retaining parallelizability. Our method allows inference of a generalization **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0351 (prec) 0.0606 (recall) 0.0444 (f-1) - Rouge L: 0.1207 (prec) 0.2059 (recall) 0.1522 (f-1) **Summary:** we investigate a local reparameterizaton technique for greatly reducing the variance of stochastic gradients for variational bayesian inference (SGVB) of a posterior over model parameters, while retaining parallelizability. our method allows inference of more flexibly parameterized posteriors ; specifically, we propose variational dropout, a generalization of gaussian dropout where the dropout rates are learned, often leading to better models. **Comment:** - Copied only from the 2nd article, specifically the first and penultimate sentences. - No sign of MDS - Errors introduced quite randomly (red)	**Rouge scores:** - Rouge 2: 0.022 (prec) 0.1212 (recall) 0.0372 (f-1) - Rouge L: 0.0601 (prec) 0.3235 (recall) 0.1014 (f-1) **Summary:** In this article, we investigate the reasons why Batch Normalization (BN) improves the generalization performance of deep networks. We argue that one major reason, distinguishing it from data-independent normalization methods, is randomness of batch statistics. This randomness appears in the parameters rather than in activations and admits an interpretation as a practical Bayesian learning. We apply this idea to other (heterogeneous) normalization techniques that are oblivious to the batch size. We show that their generalization performance can be improved significantly by Bayesian learning of the same form. We obtain test performance comparable to BN and, at the same time, better validation losses suitable for subsequent output uncertainty estimation through approximate Bayesian posterior. Additionally, we explore a connection with dropout: Gaussian dropout objectives correspond to SGVB with local reparameterization, a scale-invariant prior and proportionally fixed posterior variance. Our method allows inference of more flexibly parameterized posteriors; specifically, we propose variational dropout, a generalization of Gaussian drop out where the dropout rates are learned, often leading to better models. The method is demonstrated through several experiments. The Method is demonstrated through multiple experiments. **Comment:** - Initial parts copied from the 1st article (yellow), followed by copying from the 2nd article (green) - No sign of MDS as the two are presented as concerning the same study - A made-up word is randomly introduced for no apparent reason (red)	**Rouge scores:** - Rouge 2: 0.0156 (prec) 0.0303 (recall) 0.0206 (f-1) - Rouge L: 0.0923 (prec) 0.1765 (recall) 0.1212 (f-1) **Summary:** In @cite, the authors propose a local reparameterization that translates uncertainty about global parameters into local noise that is independent across datapoints in the minibatch. The authors show that this method can be improved significantly by Bayesian learning of the same form. In @cite the authors propose variational dropout, a generalization of Gaussian dropout where the dropout rates are learned, often leading to better models. **Comment:** - Summarizes the 1st (yellow), and then the 2nd (green) article - Shows signs of MDS, but factuality is a concern since the red part is actually from the 2nd article	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.1042 (prec) 0.1471 (recall) 0.122 (f-1) **Summary:** In @cite, a local reparameterization is used to reduce the variance of stochastic gradients for variational Bayesian inference (SGVB) of a posterior over model parameters, while retaining parallelizability. This method allows inference of more flexibly parameterized posteriors; specifically, the dropout rates are learned, often leading to better models. **Comment:** - All parts are from 2nd article. - However, it is represented as a separate article from the main one, i.e. weak indication of MDS	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.0795 (prec) 0.2059 (recall) 0.1148 (f-1) **Summary:** In @cite, a local reparameterization is used to reduce the variance of stochastic gradients for variational Bayesian inference (SGVB) of a posterior over model parameters, while retaining parallelizability. In contrast, our method does not require parallelizability, and we do not need parallelizability in our method, which is a key advantage of SGVB over SGVB. In addition, our method can be seen as a generalization of SGVB where the dropout rates are learned, often leading to better models, and the variance is inversely proportional to the mini-batch size. **Comment:** - Contains mostly information from the 2nd model (green) - However, it pretends to be doing an MDS (red) though the information is not directly implied and contains logical inconsistencies ("key advantage of SGVB over SGVB")

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
3157	**(1):** In this paper, we study a novel approach for named entity recognition (NER) and mention detection in natural language processing. Instead of treating NER as a sequence labelling problem, we propose a new local detection approach, which rely on the recent fixed-size ordinally forgetting encoding (FOFE) method to fully encode each sentence fragment and its left right contexts into a fixed-size representation. Afterwards, a simple feedforward neural network is used to reject or predict entity label for each individual fragment. The proposed method has been evaluated in several popular NER and mention detection tasks, including the CoNLL 2003 NER task and TAC-KBP2015 and TAC-KBP2016 Tri-lingual Entity Discovery and Linking (EDL) tasks. Our methods have yielded pretty strong performance in all of these examined tasks. This local detection approach has shown many advantages over the traditional sequence labelling methods. **(2):** We propose a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including: part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. This versatility is achieved by trying to avoid task-specific engineering and therefore disregarding a lot of prior knowledge. Instead of exploiting man-made input features carefully optimized for each task, our system learns internal representations on the basis of vast amounts of mostly unlabeled training data. This work is then used as a basis for building a freely available tagging system with good performance and minimal computational requirements. **(3):** The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling. An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada" and "Air" cannot be easily combined to obtain "Air Canada". Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible.	The success of word embedding @cite encourages researchers to focus on machine-learned representation instead of heavy feature engineering in NLP. Using word embedding as the typical feature representation for words, NNs become competitive to traditional approaches in NER. Many NLP tasks, such as NER, chunking and part-of-speech (POS) tagging can be formulated as sequence labeling tasks. In @cite , deep convolutional neural networks (CNN) and conditional random fields (CRF) are used to infer NER labels at a sentence level, where they still use many hand-crafted features to improve performance, such as capitalization features explicitly defined based on first-letter capital, non-initial capital and so on.	**Rouge scores:** - Rouge 2: 0.0316 (prec) 0.0463 (recall) 0.0376 (f-1) - Rouge L: 0.1132 (prec) 0.1651 (recall) 0.1343 (f-1) **Summary:** In this paper, we study a novel approach for named entity recognition (NER) and mention detection in natural language processing. Instead of treating NER as a sequence labelling problem, we propose a new local detection approach, which rely on the recent fixed-size ordinally forgetting encoding (FOFE) method to fully encode each sentence fragment and its left right contexts into a fixed-size representation. Afterwards, a simple feedforward neural network is used to reject or predict entity label for each individual fragment. The proposed method has been evaluated in several popular NER and mention detection tasks, including the CoNLL 2003 NER task and TAC-KBP2016 Tri-lingual Entity Discovery and Linking (EDL) tasks. Our methods have yielded pretty strong performance in all of these examined tasks. This local detection approach has shown many advantages over the traditional sequence labelling methods.||||We propose a unified neural network architecture and learning algorithm that can be applied to various natural language processing **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0303 (prec) 0.037 (recall) 0.0333 (f-1) - Rouge L: 0.0977 (prec) 0.1193 (recall) 0.1074 (f-1) **Summary:** in this paper, we study a novel approach for named entity recognition and mention detection in natural language processing. instead of treating named entity recognition as a sequence labelling problem, we propose a new local detection approach, which rely on the recent fixed-size ordinally forgetting encoding (FOFE) method to fully encode each sentence fragment and its left right contexts into a fixed-size representation. afterwards, a simple feedforward neural network is used to reject or predict entity label for each individual fragment. the proposed method has been evaluated in several popular named entity and mention detection tasks, including the coNLL 2003 named entity recognition task and the tac-2015 and tac-2016 tri-lingual entity discovery and linking tasks. our methods have yielded pretty strong performance in all of these examined tasks. **Comment:** - Copied the first tokens only, except for the deletion of the acronym for NER (red) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.035 (prec) 0.0648 (recall) 0.0455 (f-1) - Rouge L: 0.1045 (prec) 0.1927 (recall) 0.1355 (f-1) **Summary:** In this paper, we study a novel approach for named entity recognition (NER) and mention detection in natural language processing. Instead of treating NER as a sequence labelling problem, we propose a new local detection approach, which rely on the recent fixed-size ordinally forgetting encoding (FOFE) method to fully encoding each sentence fragment and its left right contexts into a fixed-size representation. Afterwards, a simple feedforward neural network is used to reject or predict entity label for each individual fragment. The proposed method has been evaluated in several popular NER and mention detection tasks, including the CoNLL 2003 NER task and TAC-KBP2015 and TAC -KBP2016 Tri-lingual Entity Discovery and Linking (EDL) tasks. Our methods have yielded pretty strong performance in all of these examined tasks. This local detection approach has shown many advantages over the traditional sequence labelling methods. We propose a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including: part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. This versatility is achieved by trying to avoid task-specific engineering and therefore disregarding a lot of prior knowledge. Instead of exploiting man-made input **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0685 (prec) 0.0463 (recall) 0.0552 (f-1) - Rouge L: 0.1757 (prec) 0.1193 (recall) 0.1421 (f-1) **Summary:** In @cite, the authors propose a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including: part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. In this work, we propose a new local detection approach, which rely on the recent fixed-size ordinally forgetting encoding (FOFE) method to fully encode each sentence fragment and its left right contexts into a fixed-size representation. **Comment:** - Shows signs of MDS, which contrasts the 2nd (green) article with the 1st (yellow) one. - However, the 3rd article is not covered at all	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.1957 (prec) 0.0826 (recall) 0.1161 (f-1) **Summary:** The Skip-gram model @cite, @cite is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. By subsampling of the frequent words, it obtain significant speedup and also learn more regular word representations. **Comment:** - Shows weak signs of MDS, providing a summary of the 3rd article only (blue) and presenting it as different from the main one.	**Rouge scores:** - Rouge 2: 0.049 (prec) 0.0463 (recall) 0.0476 (f-1) - Rouge L: 0.1942 (prec) 0.1835 (recall) 0.1887 (f-1) **Summary:** In @cite, the authors proposed a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. However, unlike our work, they do not use a deep neural network to train their system. Instead, they use a multi-task learning algorithm to train their network on a large amount of unlabeled training data. In contrast, we use a deep network to train our system on a much larger amount of training data, and we do not need to train our network on unlabeled data. **Comment:** - Shows signs of MDS, providing a summary of the 2nd article (green) and contrasting it with the main one (red). - However, the red parts also contain hallucinations, not to mention the neglect of the 3rd article

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
3524	**(1):** The explosion in the availability of GPS-enabled devices has resulted in an abundance of trajectory data. In reality, however, majority of these trajectories are collected at a low sampling rate and only provide partial observations on their actually traversed routes. Consequently, they are mired with uncertainty. In this paper, we develop a technique called Infer Tra to infer uncertain trajectories from network-constrained partial observations. Rather than predicting the most likely route, the inferred uncertain trajectory takes the form of an edge-weighted graph and summarizes all probable routes in a holistic manner. For trajectory inference, Infer Tra employs Gibbs sampling by learning a Network Mobility Model (NMM) from a database of historical trajectories. Extensive experiments on real trajectory databases show that the graph-based approach of Infer Tra is up to 50 more accurate, 20 times faster, and immensely more versatile than state-of-the-art techniques. **(2):** The advances in location-acquisition technologies have led to a myriad of spatial trajectories. These trajectories are usually generated at a low or an irregular frequency due to applications' characteristics or energy saving, leaving the routes between two consecutive points of a single trajectory uncertain (called an uncertain trajectory). In this paper, we present a Route Inference framework based on Collective Knowledge (abbreviated as RICK) to construct the popular routes from uncertain trajectories. Explicitly, given a location sequence and a time span, the RICK is able to construct the top-k routes which sequentially pass through the locations within the specified time span, by aggregating such uncertain trajectories in a mutual reinforcement way (i.e., uncertain + uncertain → certain). Our work can benefit trip planning, traffic management, and animal movement studies. The RICK comprises two components: routable graph construction and route inference. First, we explore the spatial and temporal characteristics of uncertain trajectories and construct a routable graph by collaborative learning among the uncertain trajectories. Second, in light of the routable graph, we propose a routing algorithm to construct the top-k routes according to a user-specified query. We have conducted extensive experiments on two real datasets, consisting of Foursquare check-in datasets and taxi trajectories. The results show that RICK is both effective and efficient.	More recently, a technique @cite was designed to study the trajectory inference problem in a setting where trajectories are not constrained by a network. Due to the focus on network-free trajectories, @cite is not applicable to our problem.	**Rouge scores:** - Rouge 2: 0.0321 (prec) 0.1316 (recall) 0.0515 (f-1) - Rouge L: 0.0955 (prec) 0.3846 (recall) 0.1531 (f-1) **Summary:** The explosion in the availability of GPS-enabled devices has resulted in an abundance of trajectory data. In reality, however, majority of these trajectories are collected at a low sampling rate and only provide partial observations on their actually traversed routes. Consequently, they are mired with uncertainty. In this paper, we develop a technique called Infer Tra to infer uncertain trajectories from network-constrained partial observations. Rather than predicting the most likely route, the inferred uncertain trajectory takes the form of an edge-weighted graph and summarizes all probable routes in a holistic manner. For trajectory inference, Infer Tra employs Gibbs sampling by learning a Network Mobility Model (NMM) from a database of historical trajectories. Extensive experiments on real trajectory databases show that the graph-based approach of Infer Tra is up to 50 more accurate, 20 times faster, and immensely more versatile than state-of-the-art techniques.||||The advances in location-acquisition technologies have led **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.034 (prec) 0.1316 (recall) 0.0541 (f-1) - Rouge L: 0.1014 (prec) 0.3846 (recall) 0.1604 (f-1) **Summary:** the explosion in the availability of GPS-enabled devices has resulted in an abundance of trajectory data. in reality, however, majority of these trajectories are collected at a low sampling rate and only provide partial observations on their actually traversed routes. Consequently, they are mired with uncertainty. in this paper, we develop a technique called Infer Tra to infer uncertain trajectories from network-constrained partial observations. rather than predicting the most likely route, the inferred uncertain trajectory takes the form of an edge-weighted graph and summarizes all probable routes in a holistic manner. for trajectory inference, we employ Gibbs sampling by learning a network mobility model (nMM) from a database of historical trajectories. extensive experiments on real trajectory databases show that the graph-based approach of Infer Tra is up to 50 more accurate, 20 times faster, and immensely more versatile than state-of-the-art techniques. **Comment:** - Copied the first tokens only, except for a random replacement subject noun (red) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0051 (prec) 0.0263 (recall) 0.0085 (f-1) - Rouge L: 0.0653 (prec) 0.3333 (recall) 0.1092 (f-1) **Summary:** The world's first GPS-enabled vehicle has been around for more than a decade, but it's not the first time we've seen the use of GPS-enabled vehicles. In fact, we're only a few years away from the start of a new generation of GPS-powered vehicles, which will soon be able to connect us to GPS-enabled devices. That means we'll be able to use GPS-enabled smartphones to track us down and track us down. But before we do that, let's take a closer look at what's going on in the world of GPS-based vehicles. We're going to start with GPS-enabled cars, which have been around for a long time, but they're often generated at a low or irregular frequency due to applications' characteristics or energy saving, leaving the routes between two consecutive points of a single trajectory uncertain (called an uncertain trajectory). In this paper, we present a Route Inference framework based on Collective Knowledge (abbreviated as RICK) to construct the popular routes from uncertain trajectories. Explicitly, given a location sequence and a time span, the RICK is able to construct the top-k routes **Comment:** - Hallucinated a random background on GPS (red) before starting to summarize the 2nd article (green) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.2 (prec) 0.2051 (recall) 0.2025 (f-1) **Summary:** @cite proposed a routing algorithm to construct the top-k routes according to a user-specified query. The routing algorithm is based on collaborative learning among uncertain trajectories and construct a routable graph by collaborative learning among the uncertain trajectories. **Comment:** - Shows weak signs of MDS, which presents the 2nd (green) article as different from the main one. - Red part shows some repeat of the previous meaning	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.1591 (prec) 0.1795 (recall) 0.1687 (f-1) **Summary:** @cite proposed a routing algorithm to construct the top-@math routes according to a user-specified query. They have conducted extensive experiments on two real datasets, namely Foursquare check-in datasets and taxi trajectories. The results show that RICK is both effective and efficient. **Comment:** - Shows weak signs of MDS, which presents the 2nd (green) article as different from the main one.	**Rouge scores:** - Rouge 2: 0.0233 (prec) 0.0526 (recall) 0.0323 (f-1) - Rouge L: 0.1149 (prec) 0.2564 (recall) 0.1587 (f-1) **Summary:** @cite proposed a routing algorithm to construct the top-@math routes according to a user-specified query. However, the routing algorithm does not consider the spatial and temporal characteristics of uncertain trajectories. In contrast, our routing algorithm is able to construct the popular routes from uncertain trajectories in a mutual reinforcement way (i.e., uncertain + uncertain → certain). Moreover, we do not consider the temporal and spatial characteristics of the uncertain trajectories and construct a routing algorithm that can construct the routes according to the user-specified queries. **Comment:** - Shows signs of MDS, providing a summary of the 2nd article (green) and contrasting it with the main one (red). - However, the red parts also contain hallucinations, not to mention the neglect of the 3rd article

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
4160	**(1):** We present a method for extracting depth information from a rectified image pair. We train a convolutional neural network to predict how well two image patches match and use it to compute the stereo matching cost. The cost is refined by cross-based cost aggregation and semiglobal matching, followed by a left-right consistency check to eliminate errors in the occluded regions. Our stereo method achieves an error rate of 2.61 on the KITTI stereo dataset and is currently (August 2014) the top performing method on this dataset. **(2):** While machine learning has been instrumental to the ongoing progress in most areas of computer vision, it has not been applied to the problem of stereo matching with similar frequency or success. We present a supervised learning approach for predicting the correctness of stereo matches based on a random forest and a set of features that capture various forms of information about each pixel. We show highly competitive results in predicting the correctness of matches and in confidence estimation, which allows us to rank pixels according to the reliability of their assigned disparities. Moreover, we show how these confidence values can be used to improve the accuracy of disparity maps by integrating them with an MRF-based stereo algorithm. This is an important distinction from current literature that has mainly focused on sparsification by removing potentially erroneous disparities to generate quasi-dense disparity maps. **(3):** With the aim to improve accuracy of stereo confidence measures, we apply the random decision forest framework to a large set of diverse stereo confidence measures. Learning and testing sets were drawn from the recently introduced KITTI dataset, which currently poses higher challenges to stereo solvers than other benchmarks with ground truth for stereo evaluation. We experiment with semi global matching stereo (SGM) and a census data term, which is the best performing real-time capable stereo method known to date. On KITTI images, SGM still produces a significant amount of error. We obtain consistently improved area under curve values of sparsification measures in comparison to best performing single stereo confidence measures where numbers of stereo errors are large. More specifically, our method performs best in all but one out of 194 frames of the KITTI dataset.	Recent work @cite @cite focused on estimating the confidence of the computed matching cost. used a random forest classifier to combine several confidence measures. Similarly, trained a random forest classifier to predict the confidence of the matching cost and used the predictions as soft constraints in a Markov random field to decrease the error of the stereo method.	**Rouge scores:** - Rouge 2: 0.0523 (prec) 0.1579 (recall) 0.0786 (f-1) - Rouge L: 0.0983 (prec) 0.2931 (recall) 0.1472 (f-1) **Summary:** We present a method for extracting depth information from a rectified image pair. We train a convolutional neural network to predict how well two image patches match and use it to compute the stereo matching cost. The cost is refined by cross-based cost aggregation and semiglobal matching, followed by a left-right consistency check to eliminate errors in the occluded regions. Our stereo method achieves an error rate of 2.61 on the KITTI stereo dataset and is currently (August 2014) the top performing method on this dataset. ||||While machine learning has been instrumental to the ongoing progress in most areas of computer vision, it has not been applied to the problem of stereo matching with similar frequency or success. We present a supervised learning approach for predicting the correctness of stereo matches based on a random forest and a set of features that capture various forms of information about each pixel. We show highly competitive results in predicting the accuracy of matches and in confidence estimation, which allows us to rank **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0826 (prec) 0.1579 (recall) 0.1084 (f-1) - Rouge L: 0.2 (prec) 0.3793 (recall) 0.2619 (f-1) **Summary:** We present a supervised learning approach for predicting the correctness of stereo matches based on a random forest and a set of features that capture various forms of information about each pixel. We train a convolutional neural network to predict how well two image patches match and use it to compute the stereo matching cost. the cost is refined by cross-based cost aggregation and semiglobal matching, followed by a left-right consistency check to eliminate errors in the occluded regions. our stereo method achieves an error rate of 2.61 on the recently introduced kitti stereo dataset and is currently (August 2014) the top performing method on this dataset. **Comment:** - Copied from the 1st (yellow) and 2nd (green) articles - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0468 (prec) 0.1404 (recall) 0.0702 (f-1) - Rouge L: 0.1047 (prec) 0.3103 (recall) 0.1565 (f-1) **Summary:** In this article, we're going to take a look at a new method for predicting the correctness of stereo matches based on a random forest and a set of features that capture various forms of information about each pixel. We're using the random decision forest framework, which currently poses higher challenges to stereo solvers than other benchmarks with ground truth for stereo evaluation. We experiment with semi global matching stereo (SGM) and a census data term, which is the best performing real-time capable stereo method known to date. On KITTI images, SGM still produces a significant amount of error. We obtain consistently improved area under curve values of sparsification measures in comparison to best performing single stereo confidence measures where numbers of stereo errors are large. More specifically, our method performs best in all but one out of 194 frames of the KITTI dataset. This is an important distinction from current literature that has mainly focused on sparsification by removing potentially erroneous disparities to generate quasi-dense disparity maps. **Comment:** - Some sign of MDS by contrasting the 2nd (green) and 3rd (blue articles) - However, the phrase in red is also copied from the 2nd article so it is uncertain if the model is actually doing MDS - It also mixed up the facts in the 2nd and 3rd articles	**Rouge scores:** - Rouge 2: 0.1515 (prec) 0.1754 (recall) 0.1626 (f-1) - Rouge L: 0.2687 (prec) 0.3103 (recall) 0.288 (f-1) **Summary:** Our work is closely related to @cite @cite in that they use a random forest to predict the correctness of stereo matches based on a set of features that capture various forms of information about each pixel. However, our work is different in that we train a convolutional neural network to predict how well two image patches match and use it to compute the stereo matching cost. **Comment:** - Clear sign of MDS, with the sentence in red summarizing both the 2nd and 3rd articles (both uses random forests) and contrasting that with a fact from the 1st (yellow)	**Rouge scores:** - Rouge 2: 0.0517 (prec) 0.0526 (recall) 0.0522 (f-1) - Rouge L: 0.2373 (prec) 0.2414 (recall) 0.2393 (f-1) **Summary:** Our work is also related to the semi-global matching stereo (SGM) method @cite @cite, which is based on a random forest and a set of features that capture various forms of information about each pixel. SGM achieves an error rate of 2.61 on the KITTI stereo dataset and is currently the top performing method on this dataset. **Comment:** - Some signs of MDS, with the first green part showing an attempt to make comparison - The blue and 2nd green part shows possible mix up between the 2nd and 3rd models, caused probably by the repeated use of stereo and SGM - In a similar vein, the red part is from the 1st article (see numbers in yellow) but the sentence structure is unclear as to whether the model is referring to the main article or not.	**Rouge scores:** - Rouge 2: 0.1122 (prec) 0.193 (recall) 0.1419 (f-1) - Rouge L: 0.2121 (prec) 0.3621 (recall) 0.2675 (f-1) **Summary:** In @cite @cite, the authors used a random forest to predict the correctness of stereo matching based on a random forest and a set of features that capture various forms of information about each pixel. However, they do not consider sparsification. In addition, they did not use sparsification to improve the accuracy of the disparity maps and they do not require sparsification in order to improve the performance of the matching. Moreover, their work does not rely on the sparsification of sparsification, but rather on sparsifying the disparity maps in order to achieve sparsification without sparsifying. **Comment:** - Clear sign of MDS, with the first sentence referring to the 2nd and 3rd examples, and contrasting it with the main one. - However, the red parts are hallucinations

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
4191	**(1):** Mobile Context-Aware Recommender Systems can be naturally modelled as an exploration exploitation trade-off (exr exp) problem, where the system has to choose between maximizing its expected rewards dealing with its current knowledge (exploitation) and learning more about the unknown user's preferences to improve its knowledge (exploration). This problem has been addressed by the reinforcement learning community but they do not consider the risk level of the current user's situation, where it may be dangerous to recommend items the user may not desire in her current situation if the risk level is high. We introduce in this paper an algorithm named R-UCB that considers the risk level of the user's situation to adaptively balance between exr and exp. The detailed analysis of the experimental results reveals several important discoveries in the exr exp behaviour. **(2):** Personalized web services strive to adapt their services (advertisements, news articles, etc.) to individual users by making use of both content and user information. Despite a few recent advances, this problem remains challenging for at least two reasons. First, web service is featured with dynamically changing pools of content, rendering traditional collaborative filtering methods inapplicable. Second, the scale of most web services of practical interest calls for solutions that are both fast in learning and computation. In this work, we model personalized recommendation of news articles as a contextual bandit problem, a principled approach in which a learning algorithm sequentially selects articles to serve users based on contextual information about the users and articles, while simultaneously adapting its article-selection strategy based on user-click feedback to maximize total user clicks. The contributions of this work are three-fold. First, we propose a new, general contextual bandit algorithm that is computationally efficient and well motivated from learning theory. Second, we argue that any bandit algorithm can be reliably evaluated offline using previously recorded random traffic. Finally, using this offline evaluation method, we successfully applied our new algorithm to a Yahoo! Front Page Today Module dataset containing over 33 million events. Results showed a 12.5 click lift compared to a standard context-free bandit algorithm, and the advantage becomes even greater when data gets more scarce.	In @cite , assuming the expected reward of a document is linear, they perform recommendation based on contextual information about the users' documents. To maximize the total number of user's clicks, this work proposes the LINUCB algorithm which is computationally efficient if the expected rewards of documents are linear which is not always the case.	**Rouge scores:** - Rouge 2: 0.0307 (prec) 0.0926 (recall) 0.0461 (f-1) - Rouge L: 0.0976 (prec) 0.2909 (recall) 0.1461 (f-1) **Summary:** Mobile Context-Aware Recommender Systems can be naturally modelled as an exploration exploitation trade-off (exr exp) problem, where the system has to choose between maximizing its expected rewards dealing with its current knowledge (exploitation) and learning more about the unknown user's preferences to improve its knowledge (exploration). This problem has been addressed by the reinforcement learning community but they do not consider the risk level of the current user's situation, where it may be dangerous to recommend items the user may not desire in her current situation if the risk level is high. We introduce in this paper an algorithm named R-UCB that considers the risk Level of the user's situation to adaptively balance between exr and exp. The detailed analysis of the experimental results reveals several important discoveries in the exr exp behaviour.|||| Personalized web services strive to adapt their services (advertisements, news articles, etc.) to individual users by making use of both content and user information. Despite **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0625 (prec) 0.1852 (recall) 0.0935 (f-1) - Rouge L: 0.1304 (prec) 0.3818 (recall) 0.1944 (f-1) **Summary:** personalized web services strive to adapt their services (advertisements, news articles) to individual users by making use of both content and user information. despite a few recent advances, this problem remains challenging for at least two reasons. first, web service is featured with dynamically changing pools of content, rendering traditional collaborative filtering methods inapplicable. second, the scale of most web services of practical interest calls for solutions that are both fast in learning and computation. in this work, we model personalized recommendation of news articles as a contextual bandit problem, a principled approach in which a learning algorithm sequentially selects articles to serve users based on contextual information about the users and articles, while simultaneously adapting its article-selection strategy based on user-click feedback to maximize total user clicks. we propose a new, general contextual bandit algorithm that is computationally efficient and well motivated from learning theory. moreover, we argue that any bandit algorithm can be reliably evaluated offline using **Comment:** - Copied from the 2nd article only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0417 (prec) 0.1667 (recall) 0.0667 (f-1) - Rouge L: 0.0737 (prec) 0.2909 (recall) 0.1176 (f-1) **Summary:** A new algorithm for personalized recommendation of news articles has been developed by the reinforcement learning community, but it does not consider the risk level of the current user's situation, where it may be dangerous to recommend items the user may not desire in her current situation if the risk level is high. We introduce in this paper an algorithm named R-UCB that considers the risk level to adaptively balance between exr and exp. The detailed analysis of the experimental results reveals several important discoveries in the exr exp behaviour. Personalized web services strive to adapt their services (advertisements, news articles, etc.) to individual users by making use of both content and user information. Despite a few recent advances, this problem remains challenging for at least two reasons. First, web service is featured with dynamically changing pools of content, rendering traditional collaborative filtering methods inapplicable. Second, the scale of most web services of practical interest calls for solutions that are both fast in learning and computation. In this work, we model personalized recommendation ofnews articles as a contextual bandit problem, a principled approach in which a learning algorithm sequentially selects articles to serve users based on contextual information about the users and articles, while simultaneously adapting its article-selection strategy based on user-click feedback **Comment:** - Simply joined together selected parts of the 1st (yellow) and 2nd (green) articles - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0556 (prec) 0.0556 (recall) 0.0556 (f-1) - Rouge L: 0.1636 (prec) 0.1636 (recall) 0.1636 (f-1) **Summary:** In @cite, the authors propose a contextual bandit algorithm that is computationally efficient and well motivated from learning theory. They argue that any bandit algorithm can be reliably evaluated offline using previously recorded random traffic. Finally, they successfully applied their new algorithm to a Yahoo! Front Page Today Module dataset containing over 33 million events. **Comment:** - Shows weak signs of MDS, which presents the 2nd (green) article as different from the main one	**Rouge scores:** - Rouge 2: 0.0743 (prec) 0.2037 (recall) 0.1089 (f-1) - Rouge L: 0.1678 (prec) 0.4545 (recall) 0.2451 (f-1) **Summary:** In @cite, the authors model personalized recommendation of news articles as a contextual bandit problem, a principled approach in which a learning algorithm sequentially selects articles to serve users based on contextual information about the users and articles, while simultaneously adapting its article selection strategy based on user-click feedback to maximize total user clicks. The contributions of this work are threefold. First, they propose a new, general contextual bandit algorithm that is computationally efficient and well motivated from learning theory. Second, they argue that any bandit algorithm can be reliably evaluated offline using previously recorded random traffic. Finally, using this offline evaluation method, they successfully applied their new algorithm to a Yahoo! Front Page Today Module dataset containing over 33 million events. Results showed a 12.5 click lift compared to a standard context-free bandit algorithm, and the advantage becomes even greater when data gets more scarce. **Comment:** - Shows weak signs of MDS, which presents the 2nd (green) article as different from the main one - Some creative writing noted (red)	**Rouge scores:** - Rouge 2: 0.0357 (prec) 0.0556 (recall) 0.0435 (f-1) - Rouge L: 0.1294 (prec) 0.2 (recall) 0.1571 (f-1) **Summary:** In @cite, the authors propose a contextual bandit algorithm that is computationally efficient and well motivated from learning theory. They argue that any algorithm that can reliably evaluate offline using previous recorded random traffic is reliably evaluated. They also argue that any offline evaluation method can be reliably applied to a large dataset containing over 33 million events. The results showed a 12.5 click lift compared to a standard context-free bandit algorithm, and the advantage becomes even greater when data gets more scarce. **Comment:** - Shows weak signs of MDS, which presents the 2nd (green) article as different from the main one - The 1st red part contains weird phrasing while the 2nd one is a hallucination

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
4820	**(1):** This paper presents a new method and a constraint-based objective function to solve two problemsrelated to the design of optical telecommunication networks, namely the Synchronous Optical Net-work Ring Assignment Problem (SRAP) and the Intra-ring Synchronous Optical Network DesignProblem (IDP). These network topology problems can be represented as a graph partitioning withcapacity constraints as shown in previous works. W˙e present here a new objective function and anew local search algorithm to solve these problems. Experiments conducted in C **(2):** This paper considers two problems that arise in the design of optical telecommunication networks when a ring-based topology is adopted, namely the SONET Ring Assignment Problem and the Intraring Synchronous Optical Network Design Problem. We show that these two network topology problems correspond to graph partitioning problems with capacity constraints: the first is a vertex partitioning problem, while the latter is an edge partitioning problem. We consider solution methods for both problems, based on metaheuristic algorithms. We first describe variable objective functions that depend on the transition from one solution to a neighboring one, then we apply several diversification and intensification techniques including Path Relinking, eXploring Tabu Search and Scatter Search. Finally we propose a diversification method based on the use of multiple neighborhoods. A set of extensive computational results is used to compare the behaviour of the proposed methods and objective functions.	The two other methods described in @cite are based on the idea that to save ADMs a good solution should have very dense rings. They are both greedy and rely on a clique algorithm. In graph theory, a clique in an undirected graph @math is a subset of the vertex set @math , such that for every two vertices in @math , there exists an edge connecting the two. Finding a clique is not that easy, a way to do it is to use an "Union-Find" strategie, two clique @math and @math such that each node in @math is adjacent to each node in @math then merge the two cliques (). The associated heuristic starts by considering each node to be a clique of size one, and to merge two cliques into a larger clique until there are no more possible merges.	**Rouge scores:** - Rouge 2: 0.0132 (prec) 0.0144 (recall) 0.0138 (f-1) - Rouge L: 0.1316 (prec) 0.1429 (recall) 0.137 (f-1) **Summary:** This paper presents a new method and a constraint-based objective function to solve two problemsrelated to the design of optical telecommunication networks, namely the Synchronous Optical Net-work Ring Assignment Problem (SRAP) and the Intra-ring Synchronous optical Network DesignProblem (IDP). These network topology problems can be represented as a graph partitioning withcapacity constraints as shown in previous works. We present here a new objective function and anew local search algorithm to solve these problems. Experiments conducted in CERN and the University of California, Berkeley.|This paper considers two problems that arise in the design of Optical telecommunication networks when a ring-based topology is adopted, namely the SONET Ring Assignment Problem and the Intraring Synchronous O Network Design Problem. We show that these two network topologies problems correspond to graph partitioning problems with capacity constraints: the first is a vertex partitioning problem, while the latter is an edge partitioning problem. **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.1579 (prec) 0.0857 (recall) 0.1111 (f-1) **Summary:** this paper presents a new method and a constraint-based objective function to solve two problemsrelated to the design of optical telecommunication networks, namely the Synchronous Optical Net-work Ring Assignment Problem (SRAP) and the Intra-ring Synchronous optical Network DesignProblem (IDp). these network topology problems can be represented as a graph partitioning withcapacity constraints as shown in previous works. we present here a new objective function and anew local search algorithm to solve these problems. **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0205 (prec) 0.0288 (recall) 0.024 (f-1) - Rouge L: 0.1276 (prec) 0.1786 (recall) 0.1488 (f-1) **Summary:** This paper presents a new method and a constraint-based objective function to solve two problems related to the design of optical telecommunication networks, namely the Synchronous Optical Net-work Ring Assignment Problem (SRAP) and the Intraring Synchronous optical Network DesignProblem (IDP). These network topology problems can be represented as a graph partitioning withcapacity constraints as shown in previous works. We present here a new objective function and anew local search algorithm to solve these problems. Experiments conducted in C This paper considers two problems that arise in the design of Optical telecommunication networks when a ring-based topology is adopted, namely the SONET Ring Assignment Problem and the Intra-ring SynchronousOptical Network Design Problem. We show that these two network topology problem correspond to graph partitioning problems with capacity constraints: the first is a partitioning problem, while the latter is an edge partitioning problem. We consider solution methods for both problems, based on metaheuristic algorithms. We first describe variable objective functions that depend on the transition from one solution to a neighboring one, then we apply several diversification and intensification techniques including Path Relinking, eXploring Tabu Search and Scatter Search. Finally we propose a divers **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0189 (prec) 0.0072 (recall) 0.0104 (f-1) - Rouge L: 0.2407 (prec) 0.0929 (recall) 0.134 (f-1) **Summary:** In the context of optical telecommunication networks, the SONET Ring Assignment Problem (SRAP) and the Intraring Synchronous Optical Network Design Problem (IDP) are two problems that arise when a ring-based topology is adopted. The SRAP problem can be represented as a graph partitioning problem, while the IDP problem is a edge partitioning problem. **Comment:** - Shows no sign of MDS, with the model just extracting the common theme of the 2 articles	**Rouge scores:** - Rouge 2: 0.0357 (prec) 0.0072 (recall) 0.012 (f-1) - Rouge L: 0.2414 (prec) 0.05 (recall) 0.0828 (f-1) **Summary:** In @cite, an objective function and a diversification method are proposed to solve the SONET Ring Assignment Problem (SRAP) and the Intra-ring Synchronous Optical Network Design Problem (IDP). **Comment:** - Shows weak signs of MDS, which presents the 2nd (green) article as different from the main one	**Rouge scores:** - Rouge 2: 0.1233 (prec) 0.0647 (recall) 0.0849 (f-1) - Rouge L: 0.2703 (prec) 0.1429 (recall) 0.1869 (f-1) **Summary:** In @cite, the objective functions of the SONET ring assignment problem and the IDP problem are discussed in terms of a metaheuristic algorithm. In particular, the objective function for SONET Ring Assignment Problem is described in @cite. The objective functions of SONet Ring Assignment Problem are defined as: where @math is the number of neighbors, @math the number of nodes in the network, @math and @math are the number of vertices in the network. **Comment:** - Shows weak signs of MDS, which presents the 2nd (green) article as different from the main one - Contains more relevant information that the other tuned models

(b) Medium samples

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
235	**(1):** We present a data-driven framework for incorporating side information in dynamic optimization under uncertainty. Specifically, our approach uses predictive machine learning methods (such as k-nearest neighbors, kernel regression, and random forests) to weight the relative importance of various data-driven uncertainty sets in a robust optimization formulation. Through a novel measure concentration result for local machine learning methods, we prove that the proposed framework is asymptotically optimal for stochastic dynamic optimization with covariates. We also describe a general-purpose approximation for the proposed framework, based on overlapping linear decision rules, which is computationally tractable and produces high-quality solutions for dynamic problems with many stages. Across a variety of examples in shipment planning, inventory management, and finance, our method achieves improvements of up to 15 over alternatives and requires less than one minute of computation time on problems with twelve stages. **(2):** We investigate a data-driven approach to two-stage stochastic linear optimization in which an uncertainty set is constructed around each data point. We propose an approximation algorithm for these sample robust optimization problems by optimizing a separate linear decision rule for each uncertainty set. We show that the proposed algorithm combines the asymptotic optimality and scalability of the sample average approximation while simultaneously offering improved out-of-sample performance guarantees. The practical value of our method is demonstrated in network inventory management and hospital scheduling. **(3):** Stochastic programming provides a versatile framework for decision-making under uncertainty, but the resulting optimization problems can be computationally demanding. It has recently been shown that primal and dual linear decision rule approximations can yield tractable upper and lower bounds on the optimal value of a stochastic program. Unfortunately, linear decision rules often provide crude approximations that result in loose bounds. To address this problem, we propose a lifting technique that maps a given stochastic program to an equivalent problem on a higher-dimensional probability space. We prove that solving the lifted problem in primal and dual linear decision rules provides tighter bounds than those obtained from applying linear decision rules to the original problem. We also show that there is a one-to-one correspondence between linear decision rules in the lifted problem and families of nonlinear decision rules in the original problem. Finally, we identify structured liftings that give rise to highly flexible piecewise linear and nonlinear decision rules, and we assess their performance in the context of a dynamic production planning problem.	Several recent papers have focused on tractable approximations of two- and multi-stage and robust optimization. Many approaches are based around policy approximation schemes, including lifted linear decision rules , @math -adaptivity , and finite adaptability . Alternative approaches include tractable approximations of copositive formulations . Closest related to the approximation scheme in this paper are @cite and @cite , which address two-stage problems via overlapping decision rules. @cite propose a modeling approach that leads to novel approximations of various distributionally robust applications, including two-stage distributionally robust optimization using Wasserstein ambiguity sets and expectations of piecewise convex objective functions in single-stage problems. Independently, @cite investigate a of two-stage sample robust optimization by optimizing a separate linear decision rule for each uncertainty set and prove that this approximation gap converges to zero as the amount of data goes to infinity. In of this paper, we show how to extend similar techniques to dynamic problems with many stages for the first time.	**Rouge scores:** - Rouge 2: 0.0812 (prec) 0.0823 (recall) 0.0818 (f-1) - Rouge L: 0.1553 (prec) 0.1572 (recall) 0.1562 (f-1) **Summary:** We present a data-driven framework for incorporating side information in dynamic optimization under uncertainty. Specifically, our approach uses predictive machine learning methods (such as k-nearest neighbors, kernel regression, and random forests) to weight the relative importance of various data-driven uncertainty sets in a robust optimization formulation. Through a novel measure concentration result for local machine learning methods, we prove that the proposed framework is asymptotically optimal for stochastic dynamic optimization with covariates. We also describe a general-purpose approximation for the proposed framework, based on overlapping linear decision rules, which is computationally tractable and produces high-quality solutions for dynamic problems with many stages. Across a variety of examples in shipment planning, inventory management, and finance, our method achieves improvements of up to 15 over alternatives and requires less than one minute of computation time on problems with twelve stages.||||We investigate a data-based approach to two-stage stochastic linear optimization in which an uncertainty set is **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0917 (prec) 0.0696 (recall) 0.0791 (f-1) - Rouge L: 0.1983 (prec) 0.1509 (recall) 0.1714 (f-1) **Summary:** we present a data-driven framework for incorporating side information in dynamic optimization under uncertainty. Our approach uses predictive machine learning methods (such as k-nearest neighbors, kernel regression, and random forests) to weight the relative importance of various data-driven uncertainty sets in a robust optimization formulation. through a novel measure concentration result for local machine learning methods , we prove that the proposed framework is asymptotically optimal for stochastic dynamic optimization with covariates. we also describe a general-purpose approximation for the proposed framework, based on overlapping linear decision rules, which is computationally tractable and produces high-quality solutions for dynamic problems with many stages. The practical value of our method is demonstrated in network inventory management and hospital scheduling. **Comment:** - Mostly copied the first tokens, except for the dropping of the word "Specifically" and the replacement of the last sentence with the green part - No indication of MDS	**Rouge scores:** - Rouge 2: 0.1179 (prec) 0.1582 (recall) 0.1351 (f-1) - Rouge L: 0.1784 (prec) 0.239 (recall) 0.2043 (f-1) **Summary:** A new approach to stochastic dynamic optimization has been developed by a team of researchers from the University of California, Los Angeles, and the University of Southern California. The new approach is based on the use of predictive machine learning methods (such as k-nearest neighbors, kernel regression, and random forests) to weight the relative importance of various data-driven uncertainty sets in a robust optimization formulation. Through a novel measure concentration result for local machine learning methods, we prove that the proposed framework is asymptotically optimal for stochastic Dynamic Optimization with stochastic optimization. We also describe a general-purpose approximation for the proposed framework, based on overlapping linear decision rules, which is computationally tractable and produces high-quality solutions for dynamic problems with many stages. Across a variety of examples in shipment planning, inventory management, and finance, our method achieves improvements of up to 15 over alternatives and requires less than one minute of computation time on problems with twelve stages. We investigate a data-driven approach to two-stage stochastic linear optimization in which an uncertainty set is constructed around each data point. We propose an approximation algorithm for these sample robust optimization problems by optimizing a separate linear decision rule for each uncertainty set. We show that the proposed algorithm **Comment:** - Except for the randomly hallucinated beginning (red), the model just copied the 1st article in full, followed by the initial parts of the 2nd.	**Rouge scores:** - Rouge 2: 0.2093 (prec) 0.057 (recall) 0.0896 (f-1) - Rouge L: 0.3409 (prec) 0.0943 (recall) 0.1478 (f-1) **Summary:** In @cite, the authors propose a linear decision rule for each uncertainty set. However, their approach is not as robust as our approach, which is based on overlapping linear decision rules. In contrast, our approach is asymptotically optimal for stochastic dynamic optimization with covariates. **Comment:** - Shows sign of MDS, with the sentence in red contrasting the 1st (yellow) and 2nd articles. - However, the model failed to mention anything about the 3rd article	**Rouge scores:** - Rouge 2: 0.2708 (prec) 0.0823 (recall) 0.1262 (f-1) - Rouge L: 0.4082 (prec) 0.1258 (recall) 0.1923 (f-1) **Summary:** In the context of stochastic dynamic optimization with covariates, @cite and @cite proposed methods for optimizing a separate linear decision rule for each uncertainty set. In @cite, the authors proposed a lifting technique that maps a given stochastic program to an equivalent problem on a higher-dimensional probability space. **Comment:** - Clear sign of MDS, with the sentence context of the main article given (yellow) and the 2nd (green) and 3rd (blue) articles clearly mentioned.	**Rouge scores:** - Rouge 2: 0.046 (prec) 0.0253 (recall) 0.0327 (f-1) - Rouge L: 0.1932 (prec) 0.1069 (recall) 0.1377 (f-1) **Summary:** In @cite @cite, the authors propose a framework for stochastic dynamic optimization, which is computationally tractable and efficient. However, their framework is not asymptotically asymptotic, since it does not require any additional stages, and we do not need any additional stages. Moreover, the authors in @cite show that their framework is computationally optimal for the stochastic setting, which is more tractable than our method. Moreover, our framework is as well-suited for dynamic problems in which multiple stages are required, which is not the case in our case. **Comment:** - Some signs of MDS, with the red parts attempting to make contrasts. - However, the model wrongly uses words of the main article (yellow) to describe the 2nd and 3rd articles.

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
638	**(1):** Previous works utilized "smaller-norm-less-important" criterion to prune filters with smaller norm values in a convolutional neural network. In this paper, we analyze this norm-based criterion and point out that its effectiveness depends on two requirements that are not always met: (1) the norm deviation of the filters should be large; (2) the minimum norm of the filters should be small. To solve this problem, we propose a novel filter pruning method, namely Filter Pruning via Geometric Median (FPGM), to compress the model regardless of those two requirements. Unlike previous methods, FPGM compresses CNN models by pruning filters with redundancy, rather than those with "relatively less" importance. When applied to two image classification benchmarks, our method validates its usefulness and strengths. Notably, on CIFAR-10, FPGM reduces more than 52 FLOPs on ResNet-110 with even 2.69 relative accuracy improvement. Moreover, on ILSVRC-2012, FPGM reduces more than 42 FLOPs on ResNet-101 without top-5 accuracy drop, which has advanced the state-of-the-art. Code is publicly available on GitHub: this https URL **(2):** Deep Convolutional Neural Networks (CNN) has achieved significant success in computer vision field. However, the high computational cost of the deep complex models prevents the deployment on edge devices with limited memory and computational resource. In this paper, we proposed a novel filter pruning for convolutional neural networks compression, namely spectral clustering filter pruning with soft self-adaption manners (SCSP). We first apply spectral clustering on filters layer by layer to explore their intrinsic connections and only count on efficient groups. By self-adaption manners, the pruning operations can be done in few epochs to let the network gradually choose meaningful groups. According to this strategy, we not only achieve model compression while keeping considerable performance, but also find a novel angle to interpret the model compression process. **(3):** Model pruning has become a useful technique that improves the computational efficiency of deep learning, making it possible to deploy solutions on resource-limited scenarios. A widely-used practice in relevant work assumes that a smaller- norm parameter or feature plays a less informative role at the inference time. In this paper, we propose a channel pruning technique for accelerating the computations of deep convolutional neural networks (CNNs), which does not critically rely on this assumption. Instead, it focuses on direct simplification of the channel-to-channel computation graph of a CNN without the need of performing a computational difficult and not always useful task of making high-dimensional tensors of CNN structured sparse. Our approach takes two stages: the first being to adopt an end-to-end stochastic training method that eventually forces the outputs of some channels being constant, and the second being to prune those constant channels from the original neural network by adjusting the biases of their impacting layers such that the resulting compact model can be quickly fine-tuned. Our approach is mathematically appealing from an optimization perspective and easy to reproduce. We experimented our approach through several image learning benchmarks and demonstrate its interesting aspects and the competitive performance. **(4):** The success of CNNs in various applications is accompanied by a significant increase in the computation and parameter storage costs. Recent efforts toward reducing these overheads involve pruning and compressing the weights of various layers without hurting original accuracy. However, magnitude-based pruning of weights reduces a significant number of parameters from the fully connected layers and may not adequately reduce the computation costs in the convolutional layers due to irregular sparsity in the pruned networks. We present an acceleration method for CNNs, where we prune filters from CNNs that are identified as having a small effect on the output accuracy. By removing whole filters in the network together with their connecting feature maps, the computation costs are reduced significantly. In contrast to pruning weights, this approach does not result in sparse connectivity patterns. Hence, it does not need the support of sparse convolution libraries and can work with existing efficient BLAS libraries for dense matrix multiplications. We show that even simple filter pruning techniques can reduce inference costs for VGG-16 by up to 34 and ResNet-110 by up to 38 on CIFAR10 while regaining close to the original accuracy by retraining the networks.	Concurrently with our work, some data independent filter pruning strategies @cite @cite @cite @cite have been explored. @cite utilizes an @math -norm criterion to prune unimportant filters. @cite proposes to select filters with an @math -norm criterion and prune those selected filters in a soft manner. @cite proposes to prune models by enforcing sparsity on the scaling parameter of batch normalization layers. @cite uses spectral clustering on filters to select unimportant ones.	**Rouge scores:** - Rouge 2: 0.0483 (prec) 0.0986 (recall) 0.0648 (f-1) - Rouge L: 0.1301 (prec) 0.2639 (recall) 0.1743 (f-1) **Summary:** Previous works utilized "smaller-norm-less-important" criterion to prune filters with smaller norm values in a convolutional neural network. In this paper, we analyze this norm-based criterion and point out that its effectiveness depends on two requirements that are not always met: (1) the norm deviation of the filters should be large; (2) the minimum norm of the filter should be small. To solve this problem, we propose a novel filter pruning method, namely Filter Pruning via Geometric Median (FPGM), to compress the model regardless of those two requirements. Unlike previous methods, FPGM compresses CNN models by pruning filters with redundancy, rather than those with "relatively less" importance. When applied to two image classification benchmarks, our method validates its usefulness and strengths. FPGM reduces more than 52 FLOPs on ResNet-110 without top-5 accuracy drop, which has advanced the state-of-the-art **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.037 (prec) 0.0845 (recall) 0.0515 (f-1) - Rouge L: 0.1043 (prec) 0.2361 (recall) 0.1447 (f-1) **Summary:** model pruning has become a useful technique that improves the computational efficiency of deep learning, making it possible to deploy solutions on resource-limited scenarios. in this paper, we propose a channel pruning technique for accelerating the computations of deep convolutional neural networks (CNNs), which does not critically rely on the assumption that a smaller- norm parameter or feature plays a less informative role at the inference time. our approach takes two stages : the first being to adopt an end-to-end stochastic training method that eventually forces the outputs of some channels being constant, and the second being to prune those constant channels from the original neural network by adjusting the biases of their impacting layers such that the resulting compact model can be quickly fine-tuned. we first apply spectral clustering on filters layer by layer to explore their intrinsic connections and only count on efficient groups. by self-adaption manners, the pruning operations can be done in few epochs to **Comment:** - Extracted sentences from the 3rd article (blue) plus another from the 2nd (green) only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0193 (prec) 0.0563 (recall) 0.0288 (f-1) - Rouge L: 0.0769 (prec) 0.2222 (recall) 0.1143 (f-1) **Summary:** Deep Convolutional Neural Networks (CNN) has achieved significant success in computer vision field. However, the high computational cost of the deep complex models prevents the deployment on edge devices with limited memory and computational resource. In this paper, we propose a novel filter pruning for convolutional neural networks compression, namely spectral clustering filter pruning with soft self-adaption manners (SCSP). We first apply spectral clustering on filters layer by layer to explore their intrinsic connections and only count on efficient groups. By self-adaptions manners, the pruning operations can be done in few epochs to let the network gradually choose meaningful groups. According to this strategy, we not only achieve model compression while keeping considerable performance, but also find a novel angle to interpret the model compression process. A widely-used practice in relevant work assumes that a smaller-norm parameter or feature plays a less informative role at the inference time. In this article, we propose the channel pruning technique for accelerating the computations of deep convolutional Neural networks (CNNs), which does not critically rely on this assumption. Instead, it focuses on direct simplification of the channel-to-channel computation graph of a CNN without the need of performing a computational difficult and not always useful task **Comment:** - Copied just the whole 2nd article (green), then started copying from the 2nd sentence of the 3rd article (blue). - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0658 (prec) 0.0704 (recall) 0.068 (f-1) - Rouge L: 0.1558 (prec) 0.1667 (recall) 0.1611 (f-1) **Summary:** In @cite, the authors proposed a filter pruning method to compress CNN models by pruning filters with smaller norm values. However, their method does not require the support of sparse convolution libraries. Moreover, they do not need the support of CNN libraries and can work with existing efficient BLAS libraries. To solve this problem, they proposed a method called Filter Pruning via Geometric Median (FPGM) to prune filters with larger norm values in a convolutional neural network. **Comment:** - Shows sign of MDS, attempting to provide a summary of the 4th article (purple) and contrast with the main one - Of particular note is that, despite the 1st yellow part came from the 1st article, the model was able to present it as not related to the main article - However, it made an error by stating FPGM as linked to the other articles	**Rouge scores:** - Rouge 2: 0.0506 (prec) 0.1127 (recall) 0.0699 (f-1) - Rouge L: 0.0943 (prec) 0.2083 (recall) 0.1299 (f-1) **Summary:** @cite proposed a filter pruning method based on spectral clustering filter pruning with soft self-adaption manners (SCSP). They first apply spectral clustering on filters layer by layer to explore their intrinsic connections and only count on efficient groups. @cite proposed an approach to accelerate the computations of deep convolutional neural networks (CNNs), which does not critically rely on this assumption. Instead, it focuses on direct simplification of the channel-to-channel computation graph of a CNN without the need of performing a computational difficult and not always useful task of making high-dimensional tensors of CNN structured sparse. Their approach takes two stages: the first being to adopt an end-to-end stochastic training method that eventually forces the outputs of some channels being constant, and the second being to prune those constant channels from the original neural network by adjusting the biases of their impacting layers such that the resulting compact model can be quickly fine-tuned. **Comment:** - Shows clear sign of MDS, attempting to summarize the 2nd (green) and 3rd articles (blue) - 4th article is regrettably not covered	**Rouge scores:** - Rouge 2: 0.0412 (prec) 0.0563 (recall) 0.0476 (f-1) - Rouge L: 0.1327 (prec) 0.1806 (recall) 0.1529 (f-1) **Summary:** In @cite, the authors proposed spectral clustering filter pruning with soft self-adaption manners (SCSP). They first applied spectral clustering on filters layer by layer to explore their intrinsic connections and only count on efficient groups. Then, the network gradually selects meaningful groups. The authors also proposed a novel angle to interpret the model compression process. In @cite, the authors proposed a novel method for convolutional neural networks compression, which is based on spectral clustering. They proposed a novel algorithm to interpret the network compression process, and they also found a novel angle for interpret the compression process. **Comment:** - Shows some sign of MDS, attempting to summarize multiple articles (green and red @cite) - However, content-wise, only the 2nd article is covered

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
831	**(1):** Learning to drive faithfully in highly stochastic urban settings remains an open problem. To that end, we propose a Multi-task Learning from Demonstration (MT-LfD) framework which uses supervised auxiliary task prediction to guide the main task of predicting the driving commands. Our framework involves an end-to-end trainable network for imitating the expert demonstrator's driving commands. The network intermediately predicts visual affordances and action primitives through direct supervision which provide the aforementioned auxiliary supervised guidance. We demonstrate that such joint learning and supervised guidance facilitates hierarchical task decomposition, assisting the agent to learn faster, achieve better driving performance and increases transparency of the otherwise black-box end-to-end network. We run our experiments to validate the MT-LfD framework in CARLA, an open-source urban driving simulator. We introduce multiple non-player agents in CARLA and induce temporal noise in them for realistic stochasticity. **(2):** An artificial agent is developed that learns to play a diverse range of classic Atari 2600 computer games directly from sensory experience, achieving a performance comparable to that of an expert human player; this work paves the way to building general-purpose learning algorithms that bridge the divide between perception and action. **(3):** Reinforcement learning is the learning of a mapping from situations to actions so as to maximize a scalar reward or reinforcement signal. The learner is not told which action to take, as in most forms of machine learning, but instead must discover which actions yield the highest reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate's reward, but also the next situation, and through that all subsequent rewards. These two characteristics—trial-and-error search and delayed reward—are the two most important distinguishing features of reinforcement learning. **(4):** Several deep learning approaches have been applied to the autonomous driving task, many employing end-to-end deep neural networks. Autonomous driving is complex, utilizing multiple behavioral modalities ranging from lane changing to turning and stopping. However, most existing approaches do not factor in the different behavioral modalities of the driving task into the training strategy. This paper describes a technique for using Multi-Modal Multi-Task Learning that considers multiple behavioral modalities as distinct modes of operation for an end-to-end autonomous deep neural network utilizing the insertion of modal information as secondary input data. Using labeled data from hours of driving our fleet of 1 10th scale model cars, we trained multiple neural networks to imitate the steering angle and driving speed of human control of a car. We show that in each case, our models trained with MTL can match or outperform multiple networks trained on individual tasks, while using a fraction of the parameters and having more distinct modes of operation than a network trained without MTL on the same multi-modal data. These results should encourage Multi-Modal MTL-style training with the insertion of Modal Information for tasks with related behaviors. **(5):** In recent years different lines of evidence have led to the idea that motor actions and movements in both vertebrates and invertebrates are composed of elementary building blocks. The entire motor repertoire can be spanned by applying a well-defined set of operations and transformations to these primitives and by combining them in many different ways according to well-defined syntactic rules. Motor and movement primitives and modules might exist at the neural, dynamic and kinematic levels with complicated mapping among the elementary building blocks subserving these different levels of representation. Hence, while considerable progress has been made in recent years in unravelling the nature of these primitives, new experimental, computational and conceptual approaches are needed to further advance our understanding of motor compositionality.	Multi-task learning (MTL) research shows the joint training of auxiliary related side-tasks along with the main task enhances the training performance . MTL in neural networks has been successfully demonstrated in many tasks previously including text-to-speech conversion , natural language processing , speech processing and computer vision . In the field of sequential decision making, @cite demonstrate MTL for 3D game playing, @cite and @cite demonstrate MTL in 3D maze navigation task whereas @cite utilize the MTL framework for autonomous driving. Instead of employing future control outputs as auxiliary tasks as shown by @cite , in this work we employ action and visual abstractions to guide the driving behavior.	**Rouge scores:** - Rouge 2: 0.0721 (prec) 0.0748 (recall) 0.0734 (f-1) - Rouge L: 0.1786 (prec) 0.1852 (recall) 0.1818 (f-1) **Summary:** Learning to drive faithfully in highly stochastic urban settings remains an open problem. To that end, we propose a Multi-task Learning from Demonstration (MT-LfD) framework which uses supervised auxiliary task prediction to guide the main task of predicting the driving commands. Our framework involves an end-to-end trainable network for imitating the expert demonstrator's driving commands. The network intermediately predicts visual affordances and action primitives through direct supervision which provide the aforementioned auxiliary supervised guidance. We demonstrate that such joint learning and supervised guidance facilitates hierarchical task decomposition, assisting the agent to learn faster, achieve better driving performance and increases transparency of the otherwise black-box end-to **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0276 (prec) 0.0374 (recall) 0.0317 (f-1) - Rouge L: 0.1301 (prec) 0.1759 (recall) 0.1496 (f-1) **Summary:** learning to drive faithfully in highly stochastic urban settings remains an open problem. this paper describes a technique for using Multi-Modal Multi-Task learning that considers multiple behavioral modalities as distinct modes of operation for an end-to-end autonomous deep neural network utilizing the insertion of modal information as secondary input data for tasks with related behaviors. using labeled data from hours of driving our fleet of 1 10th scale model cars, we trained multiple neural networks to imitate the steering angle and driving speed of human control of a car. we show that in each case, our models trained with multi-modal multi-task learning can match or outperform multiple networks trained on individual tasks, while using a fraction of the parameters and having more distinct mode of operation than a network trained without multi-modal multi- task learning on the same data. **Comment:** - Started from the 1st sentence of the 1st article (yellow), but then moved to copy from the 3rd sentence of the 4th article (purple) - Words in red also came from the 4th article but from a later part - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0192 (prec) 0.0374 (recall) 0.0254 (f-1) - Rouge L: 0.0861 (prec) 0.1667 (recall) 0.1136 (f-1) **Summary:** Learning to drive faithfully in highly stochastic urban settings remains an open problem. To that end, we propose a Multi-task Learning from Demonstration (MT-LfD) framework which uses supervised auxiliary task prediction to guide the main task of predicting the driving commands. Our framework involves an end-to-end trainable network for imitating the expert demonstrator's driving commands. The network intermediately predicts visual affordances and action primitives through direct supervision which provide the aforementioned auxiliary supervised guidance. We demonstrate that such joint learning and supervised guidance facilitates hierarchical task decomposition, assisting the agent to learn faster, achieve better driving performance and increases transparency of the otherwise black-box end-to **Comment:** - Started from the 1st sentence of the 1st article (yellow), but then moved to copy from the 3rd sentence of the 4th article (purple) - Words in red also came from the 4th article but from a later part - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0492 (prec) 0.028 (recall) 0.0357 (f-1) - Rouge L: 0.1774 (prec) 0.1019 (recall) 0.1294 (f-1) **Summary:** In @cite, an end-to-end deep neural network is used to learn to play Atari 2600 Atari games directly from sensory experience. The network is trained to imitate the steering angle and driving speed of human control of a car. @cite proposed a multi-modal multi-task learning from demonstration (MT-LfD) framework to train an agent to play Atari games. **Comment:** - Shows some sign of MDS, summarizing the 2nd (green) and 4th (purple) articles. - However, there is some mix up about the networks for the two articles - The model then mentions the 1st article (yellow) but erroneously treated it as a non-main article (red @cite) and that it is used to play Atari games instead of driving (2nd red part)	**Rouge scores:** - Rouge 2: 0.0494 (prec) 0.0374 (recall) 0.0426 (f-1) - Rouge L: 0.1585 (prec) 0.1204 (recall) 0.1368 (f-1) **Summary:** Multi-task learning (MTL) @cite @cite aims to learn a mapping from situations to actions so as to maximize a scalar reward or reinforcement signal. The learner is not told which action to take, as in most forms of machine learning, but instead must discover which actions yield the highest reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate's reward, but also the next situation, and through that all subsequent rewards @cite. **Comment:** - Summarized the 3rd article only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0541 (prec) 0.0561 (recall) 0.055 (f-1) - Rouge L: 0.1786 (prec) 0.1852 (recall) 0.1818 (f-1) **Summary:** In @cite, the authors proposed a multi-modal MTL method to train a neural network to imitate the steering angle and driving speed of a car. In @cite, the authors proposed an end-to-end trainable network for imitating the expert demonstrator's driving commands. However, they do not provide a direct supervision mechanism for the agent to learn the driving commands, which is not the case in our case. Moreover, the authors do not propose a method to induce temporal noise in the network. In contrast, in our work, we propose a method that intermediately predicts the driving commands to guide the agent to achieve better driving performance and increase transparency. **Comment:** - Shows some sign of MDS, contrasting the main article (yellow) with the other articles. - However, the only other article explicitly mentioned is the 4th one (purple).

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
845	**(1):** We present a neural framework for opinion summarization from online product reviews which is knowledge-lean and only requires light supervision (e.g., in the form of product domain labels and user-provided ratings). Our method combines two weakly supervised components to identify salient opinions and form extractive summaries from multiple reviews: an aspect extractor trained under a multi-task objective, and a sentiment predictor based on multiple instance learning. We introduce an opinion summarization dataset that includes a training set of product reviews from six diverse domains and human-annotated development and test sets with gold standard aspect annotations, salience labels, and opinion summaries. Automatic evaluation shows significant improvements over baselines, and a large-scale study indicates that our opinion summaries are preferred by human judges according to multiple criteria. **(2):** The specification discloses a luggage carrier made up of a generally U-shaped frame. The frame has two spaced legs with a hook on the front which hooks over the bumper of an automobile. Two braces are attached to the cross member of the U-shaped member and the front portion of the braces is received on fastening means welded to the under side of the car frame. The cross members provide a supporting surface for carrying articles, boats and the like. A platform may be supported on the frame. **(3):** We develop a Ranking framework upon Recursive Neural Networks (R2N2) to rank sentences for multi-document summarization. It formulates the sentence ranking task as a hierarchical regression process, which simultaneously measures the salience of a sentence and its constituents (e.g., phrases) in the parsing tree. This enables us to draw on word-level to sentence-level supervisions derived from reference summaries. In addition, recursive neural networks are used to automatically learn ranking features over the tree, with hand-crafted feature vectors of words as inputs. Hierarchical regressions are then conducted with learned features concatenating raw features. Ranking scores of sentences and words are utilized to effectively select informative and non-redundant sentences to generate summaries. Experiments on the DUC 2001, 2002 and 2004 multi-document summarization datasets show that R2N2 outperforms state-of-the-art extractive summarization approaches. **(4):** We propose a neural multi-document summarization (MDS) system that incorporates sentence relation graphs. We employ a Graph Convolutional Network (GCN) on the relation graphs, with sentence embeddings obtained from Recurrent Neural Networks as input node features. Through multiple layer-wise propagation, the GCN generates high-level hidden sentence features for salience estimation. We then use a greedy heuristic to extract salient sentences while avoiding redundancy. In our experiments on DUC 2004, we consider three types of sentence relation graphs and demonstrate the advantage of combining sentence relations in graphs with the representation power of deep neural networks. Our model improves upon traditional graph-based extractive approaches and the vanilla GRU sequence model with no graph, and it achieves competitive results against other state-of-the-art multi-document summarization systems.	A few extractive neural models have been recently applied to generic multi-document summarization. train a recursive neural network using a ranking objective to identify salient sentences, while follow-up work @cite employs a multi-task objective to improve sentence extraction, an idea we adapted to our task. propose a graph convolution network to represent sentence relations and estimate sentence salience. Our summarization method is tailored to the opinion extraction task, it identifies aspect-specific and salient units, while minimizing the redundancy of the final summary with a greedy selection algorithm @cite @cite . Redundancy is also addressed in who propose a graph-based framework for abstractive summarization. introduce an encoder-decoder neural method for extractive opinion summarization. Their approach requires direct supervision via gold-standard extractive summaries for training, in contrast to our weakly supervised formulation.	**Rouge scores:** - Rouge 2: 0.0732 (prec) 0.0889 (recall) 0.0803 (f-1) - Rouge L: 0.1394 (prec) 0.1691 (recall) 0.1528 (f-1) **Summary:** We present a neural framework for opinion summarization from online product reviews which is knowledge-lean and only requires light supervision (e.g., in the form of product domain labels and user-provided ratings). Our method combines two weakly supervised components to identify salient opinions and form extractive summaries from multiple reviews: an aspect extractor trained under a multi-task objective, and a sentiment predictor based on multiple instance learning. We introduce an opinion summarization dataset that includes a training set of product reviews from six diverse domains and human-annotated development and test sets with gold standard aspect annotations, salience labels, and opinion summaries. Automatic evaluation shows significant improvements over baselines, and a large-scale study indicates that our opinion summaries are preferred by human judges according to multiple criteria.||||The specification discloses a luggage carrier made up of a generally U-shaped frame. The frame has two spaced legs with a hook on the front which hooks over the bumper of an automobile **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0902 (prec) 0.0815 (recall) 0.0856 (f-1) - Rouge L: 0.1707 (prec) 0.1544 (recall) 0.1622 (f-1) **Summary:** we present a neural framework for opinion summarization from online product reviews which is knowledge-lean and only requires light supervision (e.g., in the form of product domain labels and user-provided ratings). our method combines two weakly supervised components to identify salient opinions and form extractive summaries from multiple reviews : an aspect extractor trained under a multi-task objective, and a sentiment predictor based on multiple instance learning. we introduce an opinion summarization dataset that includes a training set of product reviews from six diverse domains and human-annotated development and test sets with gold standard aspect annotations, salience labels, and opinion summaries. a large-scale study indicates that our opinion summaries are preferred by human judges according to multiple criteria. **Comment:** - Copied the first tokens only, except that the phrase "Automatic evaluation shows significant improvements over baselines, and" is skipped before the red part - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0651 (prec) 0.1037 (recall) 0.08 (f-1) - Rouge L: 0.1157 (prec) 0.1838 (recall) 0.142 (f-1) **Summary:** We present a neural framework for opinion summarization from online product reviews which is knowledge-lean and only requires light supervision (e.g., in the form of product domain labels and user-provided ratings). Our method combines two weakly supervised components to identify salient opinions and form extractive summaries from multiple reviews: an aspect extractor trained under a multi-task objective, and a sentiment predictor based on multiple instance learning. We introduce an opinion summarization dataset that includes a training set of product reviews from six diverse domains and human-annotated development and test sets with gold standard aspect annotations, salience labels, and opinion summaries. Automatic evaluation shows significant improvements over baselines, and a large-scale study indicates that our opinion summaries are preferred by human judges according to multiple criteria. The specification discloses a luggage carrier made up of a generally U-shaped frame. The frame has two spaced legs with a hook on the front which hooks over the bumper of an automobile. Two braces are attached to the cross member of the U-shaped member and the front portion of the braces is received on fastening means welded to the under side of the car frame. The cross members provide a supporting surface for carrying articles, boats and the like. A platform may **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.1695 (prec) 0.0741 (recall) 0.1031 (f-1) - Rouge L: 0.3333 (prec) 0.1471 (recall) 0.2041 (f-1) **Summary:** In the context of multi-document summarization, R2N2 @cite is a recurrent neural network, RNN for extracting salience labels. The RNN is trained on sentence relation graphs, and the salience labels are extracted using a greedy heuristic to extract salient sentences while avoiding redundancy @cite @cite. In this work, we use the RNN to extract salience labels from product reviews. **Comment:** - Shows clear sign of MDS, summarizing the 3rd (blue) and 4th (purple) articles, while clearly contrasting the main article (yellow). - The RNN notation proved slightly confusing for the model though, the 2nd article used "Recursive Neural Network", not "Recurrent Neural Network" - The 2nd article appears completely unrelated to the task, and the model probably correctly omits it.	**Rouge scores:** - Rouge 2: 0.1186 (prec) 0.0519 (recall) 0.0722 (f-1) - Rouge L: 0.2167 (prec) 0.0956 (recall) 0.1327 (f-1) **Summary:** There is a large body of work on multi-document summarization, including @cite @cite, @cite and @cite. However, most of these works focus on extracting features from a single review. In this work, we focus on extractive summaries from multiple reviews: an aspect extractor trained under a multi-task objective, and a sentiment predictor trained based on multiple instance learning. **Comment:** - Shows some sign of MDS, with the model summarizing what the main article does (yellow), and trying to contrast that with other literature (red) - However, the red part is actually not factually based, and the phrase "a large body of work" is probably picked up from the summarization training.	**Rouge scores:** - Rouge 2: 0.1594 (prec) 0.0815 (recall) 0.1078 (f-1) - Rouge L: 0.2429 (prec) 0.125 (recall) 0.165 (f-1) **Summary:** In the context of summarization, @cite employ a Graph Convolutional Network (GCN) on the relation graphs, with sentence embeddings obtained from Recurrent Neural Networks as input node features for salience estimation. however, they do not consider sentence relation graphs and do not consider the representation power of deep neural networks. In contrast, we use salience labels and salience labels to identify salient opinions and form extractive summaries from multiple reviews. **Comment:** - Shows some sign of MDS, the model summarizing what the main article does (yellow), and trying to contrast that with other literature (purple) - However, the red-and-purple part contains erroneous facts. The model wrongly assumes that the 4th article did not look at sentence relation graphs when the authors explicitly say they did.

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
1165	**(1):** In the age of social news, it is important to understand the types of reactions that are evoked from news sources with various levels of credibility. In the present work we seek to better understand how users react to trusted and deceptive news sources across two popular, and very different, social media platforms. To that end, (1) we develop a model to classify user reactions into one of nine types, such as answer, elaboration, and question, etc, and (2) we measure the speed and the type of reaction for trusted and deceptive news sources for 10.8M Twitter posts and 6.2M Reddit comments. We show that there are significant differences in the speed and the type of reactions between trusted and deceptive news sources on Twitter, but far smaller differences on Reddit. **(2):** Social media have become an established feature of the dynamic information space that emerges during crisis events. Both emergency responders and the public use these platforms to search for, disseminate, challenge, and make sense of information during crises. In these situations rumors also proliferate, but just how fast such information can spread is an open question. We address this gap, modeling the speed of information transmission to compare retransmission times across content and context features. We specifically contrast rumor-affirming messages with rumor-correcting messages on Twitter during a notable hostage crisis to reveal differences in transmission speed. Our work has important implications for the growing field of crisis informatics. **(3):** Characterizing information diffusion on social platforms like Twitter enables us to understand the properties of underlying media and model communication patterns. As Twitter gains in popularity, it has also become a venue to broadcast rumors and misinformation. We use epidemiological models to characterize information cascades in twitter resulting from both news and rumors. Specifically, we use the SEIZ enhanced epidemic model that explicitly recognizes skeptics to characterize eight events across the world and spanning a range of event types. We demonstrate that our approach is accurate at capturing diffusion in these events. Our approach can be fruitfully combined with other strategies that use content modeling and graph theoretic features to detect (and possibly disrupt) rumors. **(4):** Recent accounts from researchers, journalists, as well as federal investigators, reached a unanimous conclusion: social media are systematically exploited to manipulate and alter public opinion. Some disinformation campaigns have been coordinated by means of bots, social media accounts controlled by computer scripts that try to disguise themselves as legitimate human users. In this study, we describe one such operation that occurred in the run up to the 2017 French presidential election. We collected a massive Twitter dataset of nearly 17 million posts, posted between 27 April and 7 May 2017 (Election Day). We then set to study the MacronLeaks disinformation campaign: By leveraging a mix of machine learning and cognitive behavioral modeling techniques, we separated humans from bots, and then studied the activities of the two groups independently, as well as their interplay. We provide a characterization of both the bots and the users who engaged with them, and oppose it to those users who didn't. Prior interests of disinformation adopters pinpoint to the reasons of scarce success of this campaign: the users who engaged with MacronLeaks are mostly foreigners with pre-existing interest in alt-right topics and alternative news media, rather than French users with diverse political views. Concluding, anomalous account usage patterns suggest the possible existence of a black market for reusable political disinformation bots.	As we noted above, most studies that examine misinformation spread focus on individual events such as natural disasters @cite , political elections @cite , or crises @cite and examine the response to the event on social media. A recent study by Vosoughi al found that news stories that were fact-checked and found to be false spread faster and to more people than news items found to be true. In contrast, our methodology considers immediate reactions to news of varying credibility, so we can determine whether certain reactions or reactions to trusted or deceptive news sources evoke more or faster responses from social media users.	**Rouge scores:** - Rouge 2: 0.0364 (prec) 0.0588 (recall) 0.0449 (f-1) - Rouge L: 0.1084 (prec) 0.1748 (recall) 0.1338 (f-1) **Summary:** In the age of social news, it is important to understand the types of reactions that are evoked from news sources with various levels of credibility. In the present work we seek to better understand how users react to trusted and deceptive news sources across two popular, and very different, social media platforms. To that end, (1) we develop a model to classify user reactions into one of nine types, such as answer, elaboration, and question, etc, and (2) we measure the speed and the type of reaction for trusted and deceptive information sources for 10.8M Twitter posts and 6.2M Reddit comments. We show that there are significant differences in the speed, the type of reactions between trusted and deceptive sites on Twitter, but far smaller differences on Reddit.||||Social media have become an established feature of the dynamic information space that emerges during crisis events. Both emergency responders and the public use these platforms to search for, disseminate, challenge, and make sense of **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0071 (prec) 0.0098 (recall) 0.0083 (f-1) - Rouge L: 0.0993 (prec) 0.1359 (recall) 0.1148 (f-1) **Summary:** Social media have become an established feature of the dynamic information space that emerges during crisis events. both emergency responders and the public use these platforms to search for, disseminate, challenge, and make sense of information during crises. in these situations rumors also proliferate, but just how fast such information can spread is an open question. we address this gap by modeling the speed of information transmission to compare retransmission times across content and context features. we use epidemiological models to characterize information cascades in twitter resulting from both news and rumors to characterize eight events across the world and spanning a range of event types. specifically, we contrast rumor-affirming messages with rumor-correcting messages on Twitter during a notable hostage crisis to reveal differences in transmission speed. our work has important implications for the growing field of crisis informatics. **Comment:** - Extracted and re-arranged sentences from the 2nd (green) and 3rd (blue) articles only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.028 (prec) 0.0588 (recall) 0.038 (f-1) - Rouge L: 0.093 (prec) 0.1942 (recall) 0.1258 (f-1) **Summary:** In the age of social news, it is important to understand the types of reactions that are evoked from news sources with various levels of credibility. In the present work we seek to better understand how users react to trusted and deceptive news sources across two popular, and very different, social media platforms. To that end, (1) we develop a model to classify user reactions into one of nine types, such as answer, elaboration, and question, etc, and (2) we measure the speed and the type of reaction for trusted and misleading news sources for 10.8M Twitter posts and 6.2M Reddit comments. We show that there are significant differences in the speed and a type of reactions between trusted and deceptive News sources on Twitter, but far smaller differences on Reddit. Social media has become an established feature of the dynamic information space that emerges during crisis events. Both emergency responders and the public use these platforms to search for, disseminate, challenge, and make sense of information during crises. In these situations rumors also proliferate, but just how fast such information can spread is an open question. We address this gap, modeling the speed of information transmission to compare retransmission times across content and context features. We specifically contrast rumor-affirming messages with **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0156 (prec) 0.0098 (recall) 0.012 (f-1) - Rouge L: 0.1385 (prec) 0.0874 (recall) 0.1071 (f-1) **Summary:** In the context of crisis informatics, @cite and @cite use epidemiological models to characterize information cascades in Twitter resulting from both news and rumors. In @cite, the authors use the SEIZ enhanced epidemic model that explicitly recognizes skeptics to characterize eight events across the world and spanning a range of event types. They demonstrate that their approach is accurate at capturing diffusion in crisis informatics. **Comment:** - Shows no sign of MDS, with the model just summarizing the 3rd article - Maybe this have something to do with how the 1st article (i.e. the main one) is not clear what it does?	**Rouge scores:** - Rouge 2: 0.0182 (prec) 0.0196 (recall) 0.0189 (f-1) - Rouge L: 0.0991 (prec) 0.1068 (recall) 0.1028 (f-1) **Summary:** In the past few years, there has been a growing interest in studying information diffusion on social media. For example, @cite used epidemiological models to characterize information cascades in Twitter resulting from both news and rumors. @cite studied the spread of misinformation on Twitter during the 2017 French presidential election. They collected a massive Twitter dataset of nearly 17 million posts, posted between 27 April and 7 May 2017 (Election Day). They then set to study the activities of the two groups independently, as well as their interplay. They provide a characterization of both the bots and the users who engaged with them, and oppose it to those who didn't. **Comment:** - Shows some sign of MDS, with the model summarizing the 3rd (blue) and 4th (purple) articles - The main article is probably not mention because it does not include much detail? The 2nd article is also highly abstract	**Rouge scores:** - Rouge 2: 0.0129 (prec) 0.0196 (recall) 0.0156 (f-1) - Rouge L: 0.0769 (prec) 0.1165 (recall) 0.0927 (f-1) **Summary:** @cite use epidemiological models to characterize information diffusion on social platforms like Twitter. They use the SEIZ enhanced epidemic model that explicitly recognizes skeptics to characterize eight events across the world and spanning a range of event types. They do not explicitly recognize skeptics to characterize events in Twitter, and they do not consider the impact of rumors on the spread of information diffusion in Twitter. @cite study the use of bots in the 2016 French presidential election. They found anomalous account usage patterns suggest the possible existence of a black market for reusable political disinformation bots, as well as their interplay with the users who engaged with them, and the users who had not engaged with them. The authors of @cite found that the users who did engage with the bots are mostly foreigners, rather than French users with diverse political views, and that the bots who engaged with the bots have diverse political views. **Comment:** - Shows some sign of MDS, with the model summarizing the 3rd (blue) and 4th (purple) articles - Summarization is more clumsy than Centrum - The 2017 French election is wrongly changed to 2016 (red) despite 2016 is not mentioned anywhere else in the inputs

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
4263	**(1):** Classical algorithms for query optimization presuppose the absence of inconsistencies or uncertainties in the database and exploit only valid semantic knowledge provided, e.g., by integrity constraints. Data inconsistency or uncertainty, however, is a widespread critical issue in ordinary databases: total integrity is often, in fact, an unrealistic assumption and violations to integrity constraints may be introduced in several ways. In this report we present an approach for semantic query optimization that, differently from the traditional ones, relies on not necessarily valid semantic knowledge, e.g., provided by violated or soft integrity constraints, or induced by applying data mining techniques. Query optimization that leverages invalid semantic knowledge cannot guarantee the semantic equivalence between the original user's query and its rewriting: thus a query optimized by our approach yields approximate answers that can be provided to the users whenever fast but possibly partial responses are required. Also, we evaluate the impact of use of invalid semantic knowledge in the rewriting of a query by computing a measure of the quality of the answer returned to the user, and we rely on the recent theory of Belief Logic Programming to deal with the presence of possible correlation in the semantic knowledge used in the rewriting. **(2):** The authors present an approach to acquiring knowledge from previously processed queries. By using newly acquired knowledge together with given semantic knowledge, it is possible to make the query processor and or optimizer more intelligent so that future queries can b processed more efficiently. The acquired knowledge is in the form of constraints. While some constraints are to be enforced for all database states, others are known to be valid for the current state of the database. The former constraints are statistic integrity constraints, while the latter are called dynamic integrity constraints. Some situations in which certain dynamic semantic constraints can be automatically extracted are identified. This automatic tool for knowledge acquisition can also be used as an interactive tool for identifying potential static integrity constraints. The concept of minimal knowledge base is introduced, and a method to maintain the knowledge base is presented. An algorithm to compute the restriction (selection) closure, i.e. all deductible restrictions, from a given set of restrictions, join predicates (as given in a query), and constraints is given. > **(3):** An approach to learning query-transformation rules based on analyzing the existing data in the database is proposed. A framework and a closure algorithm for learning rules from a given data distribution are described. The correctness, completeness, and complexity of the proposed algorithm are characterized and a detailed example is provided to illustrate the framework. > **(4):** The use of inference rules to support intelligent data processing is an increasingly important tool in many areas of computer science. In database systems, rules are used in semantic query optimization as a method for reducing query processing costs. The savings is dependent on the ability of experts to supply a set of useful rules and the ability of the optimizer to quickly find the appropriate transformations generated by these rules. Unfortunately, the most useful rules are not always those that would or could be specified by an expert. This paper describes the architecture of a system having two interrelated components: a combined conventional semantic query optimizer, and an automatic rule deriver. Our automatic rule derivation method uses intermediate results from the optimization process to direct the search for learning new rules. Unlike a system employing only user-specified rules, a system with an automatic capability can derive rules that may be true only in the current state of the database and can modify the rule set to reflect changes in the database and its usage pattern. This system has been implemented as an extension of the EXODUS conventional query optimizer generator. We describe the implementation, and show how semantic query optimization is an extension of conventional optimization in this context.	Besides @cite , there are also several works that automatically derive and exploit soft (also called dynamic) constraints for SQO @cite @cite @cite , but this semantic knowledge is used only as long as it is true, and then either updated or discarded, so that the information inferred from the database by using it is correct.	**Rouge scores:** - Rouge 2: 0.0231 (prec) 0.0755 (recall) 0.0354 (f-1) - Rouge L: 0.0747 (prec) 0.2407 (recall) 0.114 (f-1) **Summary:** Classical algorithms for query optimization presuppose the absence of inconsistencies or uncertainties in the database and exploit only valid semantic knowledge provided, e.g., by integrity constraints. Data inconsistency or uncertainty, however, is a widespread critical issue in ordinary databases: total integrity is often, in fact, an unrealistic assumption and violations to integrity constraints may be introduced in several ways. In this report we present an approach for semantic query optimization that, differently from the traditional ones, relies on not necessarily valid semantic knowledge, e.G., provided by violated or soft integrity constraints, or induced by applying data mining techniques. Query optimization that leverages invalid semantic knowledge cannot guarantee the semantic equivalence between the original user's query and its rewriting: thus a query optimized by our approach yields approximate answers that can be provided to the users whenever fast but possibly partial responses are required. Also, we evaluate the impact of use of invalid semantic knowledge in the rewriting of a query by computing a measure of the quality of the answer returned **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0185 (prec) 0.0189 (recall) 0.0187 (f-1) - Rouge L: 0.0909 (prec) 0.0926 (recall) 0.0917 (f-1) **Summary:** An approach to learning query-transformation rules based on analyzing the existing data in the database is proposed. a framework and a closure algorithm for learning rules from a given data distribution are described. The correctness, completeness, and complexity of the proposed algorithm are characterized and a detailed example is provided to illustrate the framework. **Comment:** - Copied the 3rd article in full only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0181 (prec) 0.0755 (recall) 0.0292 (f-1) - Rouge L: 0.0676 (prec) 0.2778 (recall) 0.1087 (f-1) **Summary:** Classical algorithms for query optimization suppose the absence of inconsistencies or uncertainties in the database and exploit only valid semantic knowledge provided, e.g., by integrity constraints. Data inconsistency or uncertainty, however, is a widespread critical issue in ordinary databases: total integrity is often, in fact, an unrealistic assumption and violations to integrity constraints may be introduced in several ways. In this report we present an approach for semantic query optimization that, differently from the traditional ones, relies on not necessarily valid semantic knowledge, e.g., provided by violated or soft integrity constraints, or induced by applying data mining techniques. The correctness, completeness, and complexity of the proposed algorithm are characterized and a detailed example is provided to illustrate the framework. >The use of inference rules to support intelligent data processing is an increasingly important tool in many areas of computer science. In database systems, rules are used in semantic query optimization as a method for reducing query processing costs. The savings is dependent on the ability of experts to supply a set of useful rules and the ability of the optimizer to quickly find the appropriate transformations generated by these rules. Unfortunately, the most useful rules are not always those that would or could be specified by an expert. This paper describes the architecture of a system having two interrelated components **Comment:** - Copied the 1st article in full, then the last sentence of the 3rd (blue) and then the rest of the 4th (purple) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.023 (prec) 0.0377 (recall) 0.0286 (f-1) - Rouge L: 0.1136 (prec) 0.1852 (recall) 0.1408 (f-1) **Summary:** Semantic query optimization has also been studied in the context of data mining @cite @cite. The authors in @cite proposed a framework for learning query-transformation rules based on analyzing the existing data in the database and learning rules that may be true only in the current state of the database and can modify the rule set to reflect changes in the data distribution. In @cite, the authors presented an approach to acquiring knowledge from previously processed queries and a method to maintain the knowledge base is presented. **Comment:** - Shows some sign of MDS, with the model summarizing the 3rd (blue) and 2nd (green) articles - Interestingly, the yellow part is adopted from the 1st article, but represents an inferred piece of information, i.e. that datamining is one the traditional (i.e. previously seen) methods	**Rouge scores:** - Rouge 2: 0.0357 (prec) 0.1132 (recall) 0.0543 (f-1) - Rouge L: 0.1006 (prec) 0.3148 (recall) 0.1525 (f-1) **Summary:** There is a large body of work on semantic query optimization @cite @cite. In @cite, the authors present a framework for learning query-transformation rules based on analyzing the existing data in the database. The framework and a closure algorithm for learning rules from a given data distribution are described. The approach is an extension of the EXODUS conventional query optimizer generator @cite and an extension of an extension of conventional optimization in this context . The authors in @cite present an approach to acquiring knowledge from previously processed queries. By using newly acquired knowledge together with given semantic knowledge, it is possible to make the query processor and or optimizer more intelligent so that future queries can b processed more efficiently. The concept of minimal knowledge base is introduced, and a method to maintain the knowledge base is presented. An algorithm to compute the restriction (selection) closure, i.e. all deductible restrictions, from a given set of restrictions, join predicates (as given in a query), and constraints is given. **Comment:** - Shows clear sign of MDS, successfully summarizing all of the 2nd (green), 3rd (blue) and 4th (red) articles while mentioning the context (yellow). - The 1st red part ("large body of work") is also used in some other summaries, possibly a writing style picked up by the model. - The 2nd red part contains a typo ("can b") that was directly copied by the model.	**Rouge scores:** - Rouge 2: 0.0209 (prec) 0.0755 (recall) 0.0328 (f-1) - Rouge L: 0.0677 (prec) 0.2407 (recall) 0.1057 (f-1) **Summary:** In @cite, a rule deriver is used to direct the search for rules that may be true only in the current state of the database. In contrast, our approach uses intermediate results from the optimization process to guide the search for new rules that may not be true at all. In addition, our system does not require intermediate results to select rules that are true only at the time of the search, and we do not require intermediate rules to select the rules that would or could be true. @cite proposed a method to compute the restriction closure, i.e. all deductible restrictions, from a given set of restrictions, join predicates (as given in a query), and constraints. The algorithm to compute the closure is given, and a method to maintain the knowledge base is presented. However, the method to maintain knowledge base is not presented in this paper. The authors present an approach to acquiring knowledge from previously processed queries, which is based on the notion of minimal knowledge base. However, this approach is not based on the minimal knowledge base, and it is not possible to maintain any knowledge base. **Comment:** - Shows some sign of MDS, with the model summarizing the 2nd (green) and 4th (purple) articles - The 1st red part shows an attempt to contrast with the main article but wrongly mixed that up with facts from the 4th one - The 2nd and 3rd red parts are pure hallucinations.

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
4371	**(1):** In contrast to XML query languages as e.g. XPath which require knowledge on the query language as well as on the document structure, keyword search is open to anybody. As the size of XML sources grows rapidly, the need for efficient search indices on XML data that support keyword search increases. In this paper, we present an approach of XML keyword search which is based on the DAG of the XML data, where repeated substructures are considered only once, and therefore, have to be searched only once. As our performance evaluation shows, this DAG-based extension of the set intersection search algorithm[1], [2], can lead to search times that are on large documents more than twice as fast as the search times of the XML-based approach. Additionally, we utilize a smaller index, i.e., we consume less main memory to compute the results. **(2):** Recent works have shown the benefits of keyword proximity search in querying XML documents in addition to text documents. For example, given query keywords over Shakespeare's plays in XML, the user might be interested in knowing how the keywords cooccur. In this paper, we focus on XML trees and define XML keyword, proximity queries to return the (possibly heterogeneous) set of minimum connecting trees (MCTs) of the matches to the individual keywords in the query. We consider efficiently executing keyword proximity queries on labeled trees (XML) in various settings: 1) when the XML database has been preprocessed and 2) when no indices are available on the XML database. We perform a detailed experimental evaluation to study the benefits of our approach and show that our algorithms considerably outperform prior algorithms and other applicable approaches. **(3):** We consider the problem of efficiently producing ranked results for keyword search queries over hyperlinked XML documents. Evaluating keyword search queries over hierarchical XML documents, as opposed to (conceptually) flat HTML documents, introduces many new challenges. First, XML keyword search queries do not always return entire documents, but can return deeply nested XML elements that contain the desired keywords. Second, the nested structure of XML implies that the notion of ranking is no longer at the granularity of a document, but at the granularity of an XML element. Finally, the notion of keyword proximity is more complex in the hierarchical XML data model. In this paper, we present the XRANK system that is designed to handle these novel features of XML keyword search. Our experimental results show that XRANK offers both space and performance benefits when compared with existing approaches. An interesting feature of XRANK is that it naturally generalizes a hyperlink based HTML search engine such as Google. XRANK can thus be used to query a mix of HTML and XML documents.	Similar to XRANK @cite is the stack-based approach presented in @cite . In contrast to the previous stack-based appraoches, the authors do not used the DeweyID to identify a node and to calculate the ancestor-descendant or even parent-child relationships, but they propose to use a combination of preorder position, postorder position, and depth of the node.	**Rouge scores:** - Rouge 2: 0.0244 (prec) 0.069 (recall) 0.036 (f-1) - Rouge L: 0.0788 (prec) 0.2203 (recall) 0.1161 (f-1) **Summary:** In contrast to XML query languages as e.g. XPath which require knowledge on the query language as well as on the document structure, keyword search is open to anybody. As the size of XML sources grows rapidly, the need for efficient search indices on XML data that support keyword search increases. In this paper, we present an approach of XML keyword search which is based on the DAG of the XML data, where repeated substructures are considered only once, and therefore, have to be searched only once. As our performance evaluation shows, this DAG-based extension of the set intersection search algorithm[1], [2], can lead to search times that are on large documents more than twice as fast as the search times of the XML-based approach. Additionally, we utilize a smaller index, i.e., we consume less main memory to compute the results.||||Recent works have shown the benefits of keyword proximity search in querying XML documents in addition to text documents. **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.1282 (prec) 0.1695 (recall) 0.146 (f-1) **Summary:** in this paper, we consider the problem of efficiently producing ranked results for keyword search queries over hyperlinked XML documents. we present an approach of XML keyword search which is based on the set intersection search algorithm, where repeated substructures are considered only once, and therefore, have to be searched only once. we perform a detailed experimental evaluation to study the benefits of our approach and show that our algorithms considerably outperform prior algorithms and other applicable approaches. **Comment:** - Very weak signs of MDS, with all 3 articles covered. - However, they are all presented as a single document. - The model also performs like an extractive model	**Rouge scores:** - Rouge 2: 0.0061 (prec) 0.0172 (recall) 0.009 (f-1) - Rouge L: 0.0793 (prec) 0.2203 (recall) 0.1166 (f-1) **Summary:** "We consider the problem of efficiently producing ranked results for keyword search queries over hyperlinked Xie documents, as opposed to (conceptually) flat HTML documents, introduces many new challenges," the authors write. "First, Xie XRANK does not always return entire documents, but can return deeply nesting markup that contains the desired keywords. Second, the nesting structure of markup implies that the notion of ranking is no longer at the granularity of a document, but at the granular of an markup. Finally, the notion of keyword proximity is more complex in the hierarchical markup data model. In this paper, we present the XRANK system that is designed to handle these novel features of Xie keyword search. Our experimental results show that XRANK offers both space and performance benefits when compared with existing approaches. An interesting feature of XRANK is that it naturally generalizes a hyperlink based HTML search engine such as Google. XRANK can thus be used to query a mix of HTML and markup markup." **Comment:** - Copied the 3rd article in full - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0429 (prec) 0.0517 (recall) 0.0469 (f-1) - Rouge L: 0.1831 (prec) 0.2203 (recall) 0.2 (f-1) **Summary:** XRANK @cite is an extension of the set intersection search algorithm @cite, which is based on the DAG of the XML data. XRANK is a hyperlink based HTML search engine that can be used to query a mix of hyperlink based XML documents. X-RANK is designed to handle the problem of efficiently producing ranked results for keyword search queries over hyperlinked XML documents, as opposed to (conceptually) flat HTML documents. **Comment:** - Shows very weak signs of MDS, with an attempt to reference multiple @cite. - However, none of the sentences generated are coherent, showing difficulties of the model when dealing with articles filled with acronyms (XML, HTML) which it does not understand	**Rouge scores:** - Rouge 2: 0.125 (prec) 0.0517 (recall) 0.0732 (f-1) - Rouge L: 0.4 (prec) 0.1695 (recall) 0.2381 (f-1) **Summary:** XRANK @cite is an extension of the set intersection search algorithm @cite. XRANK can be used to query a mix of HTML and XML documents. **Comment:** - Shows very weak signs of MDS, with an attempt to reference multiple @cite. - Even worse than tuned LED - However, none of the sentences generated are coherent, showing difficulties of the model when dealing with articles filled with acronyms (XML, HTML) which it does not understand	**Rouge scores:** - Rouge 2: 0.04 (prec) 0.0862 (recall) 0.0546 (f-1) - Rouge L: 0.1349 (prec) 0.2881 (recall) 0.1838 (f-1) **Summary:** XRANK @cite is a hyperlink based HTML search engine. It is designed to perform keyword search over hyperlinked XML documents. However, it does not consider the nested structure of the XML documents. XRANK can be used to query a mix of HTML and XML documents, but it does not require a hyperlink search engine, and it does not need a hyperlink, and it is not required to search for hyperlinks in a hyperlink document. Moreover, XRANK is designed to handle the nested structures of the XML document. In contrast, our approach is based on a DAG-based extension of the set intersection search algorithm, which can lead to search times that are more than twice as fast as the search time of the XML-based approach. **Comment:** - Shows signs of MDS, with the yellow part contrasting the approach of the main article with that of the 3rd (blue) - However, the red parts still show hallucination - The first step LED model in the two-step model probably helped simplify the meaning of the individual passages?

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
4717	**(1):** One of the most influential recent results in network analysis is that many natural networks exhibit a power-law or log-normal degree distribution. This has inspired numerous generative models that match this property. However, more recent work has shown that while these generative models do have the right degree distribution, they are not good models for real life networks due to their differences on other important metrics like conductance. We believe this is, in part, because many of these real-world networks have very different joint degree distributions, i.e. the probability that a randomly selected edge will be between nodes of degree k and l. Assortativity is a sufficient statistic of the joint degree distribution, and it has been previously noted that social networks tend to be assortative, while biological and technological networks tend to be disassortative. We suggest understanding the relationship between network structure and the joint degree distribution of graphs is an interesting avenue of further research. An important tool for such studies are algorithms that can generate random instances of graphs with the same joint degree distribution. This is the main topic of this paper and we study the problem from both a theoretical and practical perspective. We provide an algorithm for constructing simple graphs from a given joint degree distribution, and a Monte Carlo Markov Chain method for sampling them. We also show that the state space of simple graphs with a fixed degree distribution is connected via end point switches. We empirically evaluate the mixing time of this Markov Chain by using experiments based on the autocorrelation of each edge. These experiments show that our Markov Chain mixes quickly on real graphs, allowing for utilization of our techniques in practice. **(2):** We propose a random graph model which is a special case of sparse random graphs with given degree sequences. This model involves only a small number of parameters, called logsize and log-log growth rate. These parameters capture some universal characteristics of massive graphs. Furthermore, from these parameters, various properties of the graph can be derived. For example, for certain ranges of the parameters, we will compute the expected distribution of the sizes of the connected components which almost surely occur with high probability. We will illustrate the consistency of our model with the behavior of some massive graphs derived from data in telecommunications. We will also discuss the threshold function, the giant component, and the evolution of random graphs in this model. **(3):** Let Δ and n be natural numbers such that Δn = 2m is even and Δ ∝ (2 log n)1 2 - 1. Then as n →, the number of labelled Δ-regular graphs on n vertices is asymptotic to e − λ − λ 2 (2 m) ! m ! 2 m (Δ !) m where λ = (Δ -1) 2. As a consequence of the method we determine the asymptotic distribution of the number of short cycles in graphs with a given degree sequence, and give analogous formulae for hypergraphs. **(4):** We consider two problems: randomly generating labeled bipartite graphs with a given degree sequence and randomly generating labeled tournaments with a given score sequence. We analyze simple Markov chains for both problems. For the first problem, we cannot prove that our chain is rapidly mixing in general, but in the near-regular case, i.e., when all the degrees are almost equal, we give a proof of rapid mixing. Our methods also apply to the corresponding problem for general (nonbipartite) regular graphs, which was studied earlier by several researchers. One significant difference in our approach is that our chain has one state for every graph (or bipartite graph) with the given degree sequence; in particular, there are no auxiliary states as in the chain used by Jerrum and Sinclair. For the problem of generating tournaments, we are able to prove that our Markov chain on tournaments is rapidly mixing, if the score sequence is near-regular. The proof techniques we use for the two problems are similar. ©1999 John Wiley & Sons, Inc. Random Struct. Alg., 14: 293–308, 1999	The methods for constructing graphs with a given degree distribution are primarily either reductions to perfect matchings or sequential sampling methods. There are two popular perfect matching methods. The first is the @cite @cite : @math mini-vertices are created for each degree @math vertex, and all the mini-vertices are connected. Any perfect matching in the configuration graph corresponds to a graph with the correct degree distribution by merging all of the identified mini-vertices. This allows multiple edges and self-loops, which are often undesirable. See Figure . The second approach, the , prevents multi-edges and self-loops by creating a gadget for each vertex. If @math has degree @math , then it is replaced with a complete bipartite graph @math with @math and @math . Exactly one node in each @math is connected to each other @math , representing edge @math @cite . Any perfect matching in this model corresponds exactly to a simple graph by using the edges in the matching that correspond with edges connecting any @math to any @math . We use a natural extension of the first configuration model to the joint degree distribution problem.	**Rouge scores:** - Rouge 2: 0.0414 (prec) 0.038 (recall) 0.0397 (f-1) - Rouge L: 0.1529 (prec) 0.1405 (recall) 0.1465 (f-1) **Summary:** One of the most influential recent results in network analysis is that many natural networks exhibit a power-law or log-normal degree distribution. This has inspired numerous generative models that match this property. However, more recent work has shown that while these generative models do have the right degree distribution, they are not good models for real life networks due to their differences on other important metrics like conductance. We believe this is, in part, because many of these real-world networks have very different joint degree distributions, i.e. the probability that a randomly selected edge will be between nodes of degree k and l. Assortativity is a sufficient statistic of the joint degree distribution, and it has been previously noted that social networks tend to have assortative, while biological and technological networks tend to be disassortative. We suggest understanding the relationship between network structure and the joint degree distributions of graphs is an interesting avenue of further research. An important tool for such studies are algorithms that **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.1398 (prec) 0.0707 (recall) 0.0939 (f-1) - Rouge L: 0.2128 (prec) 0.1081 (recall) 0.1434 (f-1) **Summary:** We propose a random graph model which is a special case of sparse random graphs with given degree sequences. This model involves only a small number of parameters, called logsize and log-log growth rate, which capture some universal characteristics of massive graphs. We also show that the state space of simple graphs with a fixed degree distribution is connected via end point switches, as a consequence of the method we determine the asymptotic distribution of the number of short cycles in graphs with a given degree sequence, and give analogous formulae for hypergraphs. **Comment:** - Very weak signs of MDS, with the 1st (yellow), 2nd (green) and 3rd (blue) articles covered but not the 4th. - However, they are all presented as a single document. - The model also performs like an extractive model	**Rouge scores:** - Rouge 2: 0.0714 (prec) 0.0815 (recall) 0.0761 (f-1) - Rouge L: 0.1422 (prec) 0.1622 (recall) 0.1515 (f-1) **Summary:** In a paper published in the journal Random Struct. John Wiley & Sons, Inc. Random Struct. Alg., 14: 293–308, 1999, the authors explain the relationship between network structure and the joint degree distribution of graphs. They propose a random graph model which is a special case of sparse random graphs with given degree sequences. This model involves only a small number of parameters, called logsize and log-log growth rate. These parameters capture some universal characteristics of massive graphs. Furthermore, from these parameters, various properties of the graph can be derived. For example, for certain ranges of the parameters, we will compute the expected distribution of the sizes of the connected components which almost surely occur with high probability. We will illustrate the consistency of our model with the behavior of some massive graphs derived from data in telecommunications. We will also discuss the threshold function, the giant component, and the evolution of random graphs in this model. For the first problem, we cannot prove that our chain is rapidly mixing in general, but in the near-regular case, i.e., when all the degrees are almost equal, we give a proof of rapid mixing. Our methods also apply to the corresponding problem for general (nonbipartite) regular graphs, which was studied **Comment:** - Very weak signs of MDS, with the 1st (yellow), 2nd (green) and 4th (purple) articles covered but not the 3rd. - However, they are all presented as a single document. - The model also performs like an extractive model	**Rouge scores:** - Rouge 2: 0.1607 (prec) 0.0489 (recall) 0.075 (f-1) - Rouge L: 0.3158 (prec) 0.0973 (recall) 0.1488 (f-1) **Summary:** In @cite, the authors show that the joint degree distribution of graphs with a given degree distribution is asymptotic to the number of labelled @math -regular graphs on @math vertices. The authors also show that their model is rapidly mixing, if the score sequence is near-regular. However, they do not provide a proof of rapid mixing. **Comment:** - Shows very weak signs of MDS, with an attempt to provide some facts on the 3rd (blue) and 4th articles.	**Rouge scores:** - Rouge 2: 0.129 (prec) 0.0435 (recall) 0.065 (f-1) - Rouge L: 0.2698 (prec) 0.0919 (recall) 0.1371 (f-1) **Summary:** The problem of generating graphs with the same degree distribution was first studied by Jerrum and Sinclair @cite. They showed that the asymptotic distribution of the number of short cycles in graphs with a given degree sequence, and gave analogous formulae for hypergraphs. Later, @cite showed that their Markov chain on regular graphs is rapidly mixing, if the score sequence is near-regular. **Comment:** - Shows very weak signs of MDS, with an attempt to provide some facts on the 3rd (blue) and 4th articles.	**Rouge scores:** - Rouge 2: 0.097 (prec) 0.0707 (recall) 0.0818 (f-1) - Rouge L: 0.2 (prec) 0.1459 (recall) 0.1688 (f-1) **Summary:** In @cite, the authors presented a random graph model which is based on the log-log growth rate of the graph. However, the authors do not consider the distribution of the connected components of the graph, which is not the case in the case in @cite. In @cite, the authors presented two problems: randomly generating bipartite graphs with a given degree sequence and randomly generating tournaments with a given score sequence. In the first problem, they can prove that their chains are rapidly mixing in general, but in the near-regular case, they cannot prove the Markov chain in general, since the score sequence of the two problems is near-normal. The proof techniques we use for the two problems are similar to the one used by Jerrum and Sinclair @cite. However, they are different. **Comment:** - Shows signs of MDS, covering elements of the 2nd (green), 3rd (blue) and 4th (purple) articles - However, the first red part is still factually incorrect while the 2nd red part is quite unclear - The first step LED model in the two-step model probably helped simplify the meaning of the individual passages?

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
4858	**(1):** It is shown that one can count @math -edge paths in an @math -vertex graph and @math -set @math -packings on an @math -element universe, respectively, in time @math and @math , up to a factor polynomial in @math , @math , and @math ; in polynomial space, the bounds hold if multiplied by @math or @math , respectively. These are implications of a more general result: given two set families on an @math -element universe, one can count the disjoint pairs of sets in the Cartesian product of the two families with @math basic operations, where @math is the number of members in the two families and their subsets. **(2):** We present a fast algorithm for the subset convolution problem: given functions f and g defined on the lattice of subsets of ann-element set n, compute their subset convolution f*g, defined for S⊆ N by [(f * g)(S) = [T ⊆ S] f(T) g(S T),,]where addition and multiplication is carried out in an arbitrary ring. Via Mobius transform and inversion, our algorithm evaluates the subset convolution in O(n2 2n) additions and multiplications, substanti y improving upon the straightforward O(3n) algorithm. Specifically, if the input functions have aninteger range [-M,-M+1,...,M], their subset convolution over the ordinary sum--product ring can be computed in O(2n log M) time; the notation O suppresses polylogarithmic factors.Furthermore, using a standard embedding technique we can compute the subset convolution over the max--sum or min--sum semiring in O(2n M) time. To demonstrate the applicability of fast subset convolution, wepresent the first O(2k n2 + n m) algorithm for the Steiner tree problem in graphs with n vertices, k terminals, and m edges with bounded integer weights, improving upon the O(3kn + 2k n2 + n m) time bound of the classical Dreyfus-Wagner algorithm. We also discuss extensions to recent O(2n)-time algorithms for covering and partitioning problems (Bjorklund and Husfeldt, FOCS 2006; Koivisto, FOCS 2006). **(3):** Given a set @math with @math elements and a family @math of subsets, we show how to partition @math into @math such subsets in @math time. We also consider variations of this problem where the subsets may overlap or are weighted, and we solve the decision, counting, summation, and optimization versions of these problems. Our algorithms are based on the principle of inclusion-exclusion and the zeta transform. In effect we get exact algorithms in @math time for several well-studied partition problems including domatic number, chromatic number, maximum @math -cut, bin packing, list coloring, and the chromatic polynomial. We also have applications to Bayesian learning with decision graphs and to model-based data clustering. If only polynomial space is available, our algorithms run in time @math if membership in @math can be decided in polynomial time. We solve chromatic number in @math time and domatic number in @math time. Finally, we present a family of polynomial space approximation algorithms that find a number between @math and @math in time @math .	Concerning set packings the situation is analogous, albeit the research has been somewhat less extensive. Deciding whether a given family of @math subsets of an @math -element universe contains a @math -packing is known to be W[1]-hard @cite , and thus it is unlikely that the problem is fixed parameter tractable, that is, solvable in time @math for some function @math and constant @math . If @math is fairly large, say exponential in @math , the fastest known algorithms actually count the packings by employing the inclusion--exclusion machinery @cite @cite and run in time @math . This bound holds also for the presented algorithm (cf. Theorem).	**Rouge scores:** - Rouge 2: 0.1079 (prec) 0.1442 (recall) 0.1235 (f-1) - Rouge L: 0.1714 (prec) 0.2286 (recall) 0.1959 (f-1) **Summary:** It is shown that one can count @math -edge paths in a @math -vertex graph and @math -set @math -packings on an @math -element universe, respectively, in time @math and @math, up to a factor polynomial in @math, @math, and @math ; in polynomial space, the bounds hold if multiplied by @math or @math, respectively. These are implications of a more general result: given two set families on an 2k-element universe, one can count the disjoint pairs of sets in the Cartesian product of the two families with @math basic operations, where @math is the number of members in the two families and their subsets.||||We present a fast algorithm for the subset convolution problem:given functions f and g defined on the lattice of subsets of ann-element set n, compute their subset convolution f*g, defined for S⊆ **Comment:** - Copied the first tokens only - No indication of MDS - Random word change noted (red)	**Rouge scores:** - Rouge 2: 0.0612 (prec) 0.0577 (recall) 0.0594 (f-1) - Rouge L: 0.2222 (prec) 0.2095 (recall) 0.2157 (f-1) **Summary:** given a set @math with @math elements and a family @math of subsets, we show how to partition @math into @math such subsets in @math time. we also consider variations of this problem where the subsets may overlap or are weighted, and we solve the decision, counting, summation, and optimization versions of these problems. our algorithms are based on the principle of inclusion-exclusion and the zeta transform. in effect we get exact algorithms in time @math for several well-studied partition problems including domatic number, chromatic number, maximum @math -cut, bin packing, list coloring, and the chromatic polynomial. **Comment:** - Copied all of the 3rd article except for last 2 sentences - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0791 (prec) 0.1346 (recall) 0.0996 (f-1) - Rouge L: 0.1348 (prec) 0.2286 (recall) 0.1696 (f-1) **Summary:** It is shown that one can count @math -edge paths in an @math -vertex graph and @math -set @math -packings on an @math-element universe, respectively, in time @math and @math, up to a factor Yohyn Chocolss in @math, @math, and @math ; in Yohyn Chocolss space, the bounds hold if multiplied by @math or @math, respectively. These are implications of a more general result: given two set families on an @ math -element universe, one can count the disjoint pairs of sets in the Cartesian product of the two families with @math basic operations, where @math is the number of members in the two families and their subsets. We present a fast algorithm for the subset convolution problem:given functions f and g defined on the lattice of subsets of ann-element set n, compute their subset convolution f*g, defined for S⊆ N by [(f * g)(S) = [T ⊆ S] f(T) g(S T),]where addition and multiplication is carried out in an arbitrary ring. Via Mobius transform and inversion, our algorithm evaluates the subset convolved in O(**Comment:** - Copied the first tokens only - No indication of MDS - Random word change noted (red)	**Rouge scores:** - Rouge 2: 0.125 (prec) 0.0769 (recall) 0.0952 (f-1) - Rouge L: 0.2923 (prec) 0.181 (recall) 0.2235 (f-1) **Summary:** For the Steiner tree problem @cite, one can count the disjoint pairs of sets in the Cartesian product of the two families with @math basic operations, where @math is the number of members in the two families and their subsets @math. For the subset convolution problem, one can count @math -edge paths in an @math -vertex graph and @math -set @math -packings in time @math. **Comment:** - No indication of MDS behavior, basically just mixing in some words of the 2nd article (green) with the 1st (yellow) - Shows once again how the model fails to understand words used in unusual settings (e.g. mathematics)	**Rouge scores:** - Rouge 2: 0.082 (prec) 0.0962 (recall) 0.0885 (f-1) - Rouge L: 0.1951 (prec) 0.2286 (recall) 0.2105 (f-1) **Summary:** In @cite, Bjorklund and Husfeldt showed that one can partition a set @math into @math subsets in @math time, where @math is the number of members in the two families and their subsets. They also showed how to partition the set into @math such subsets in time @math, where @math and @math are the subsets of @math. They also gave a family of polynomial space approximation algorithms that find a number between @math and the @math in @math. @cite gave a fast algorithm for the subset convolution problem in graphs with @math vertices, @math terminals, and @math edges with bounded integer weights, improving upon the classical Dreyfus-Wagner algorithm. They also discussed extensions to recent @math -time algorithms for covering and partitioning problems. **Comment:** - Some weak signs of MDS, with the 2nd (green) and 3rd (blue) articles referenced and mixed with the 1st (yellow) - Shows once again how the model fails to understand words used in unusual settings (e.g. mathematics)	**Rouge scores:** - Rouge 2: 0.0745 (prec) 0.0673 (recall) 0.0707 (f-1) - Rouge L: 0.1895 (prec) 0.1714 (recall) 0.18 (f-1) **Summary:** In @cite, Dreyfus-Wagner algorithm @cite was presented, which is based on an @math time bound of the classical DREW algorithm. The algorithm was extended to the Steiner Tree problem by @cite and presented an O(3n - @math) time bound for the subgraphs with bounded integer weights. In the case where @math is not available, @cite showed how to partition @math into @math such subsets in polynomial time and @math in time @cite. In this case, @math and @math are polynomial in time, and @cite gives an approximation algorithm for chromatic number and domatic number. **Comment:** - Best performing of all 3 tuned models, showing strong signs of MDS with the 2nd (green) and 3rd (blue) articles reflected. - The red part appears to contain some factual errors, with the original "O(3kn + 2k n2 + n m)" changed to "O(3n - @math)" - The first step LED model in the two-step model probably helped simplify the meaning of the individual passages?

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
5068	**(1):** Convolutional neural networks have gained a remarkable success in computer vision. However, most usable network architectures are hand-crafted and usually require expertise and elaborate design. In this paper, we provide a block-wise network generation pipeline called BlockQNN which automatically builds high-performance networks using the Q-Learning paradigm with epsilon-greedy exploration strategy. The optimal network block is constructed by the learning agent which is trained sequentially to choose component layers. We stack the block to construct the whole auto-generated network. To accelerate the generation process, we also propose a distributed asynchronous framework and an early stop strategy. The block-wise generation brings unique advantages: (1) it performs competitive results in comparison to the hand-crafted state-of-the-art networks on image classification, additionally, the best network generated by BlockQNN achieves 3.54 top-1 error rate on CIFAR-10 which beats all existing auto-generate networks. (2) in the meanwhile, it offers tremendous reduction of the search space in designing networks which only spends 3 days with 32 GPUs, and (3) moreover, it has strong generalizability that the network built on CIFAR also performs well on a larger-scale ImageNet dataset. **(2):** This paper introduces the application of gradient descent methods to meta-learning. The concept of "meta-learning", i.e. of a system that improves or discovers a learning algorithm, has been of interest in machine learning for decades because of its appealing applications. Previous meta-learning approaches have been based on evolutionary methods and, therefore, have been restricted to small models with few free parameters. We make meta-learning in large systems feasible by using recurrent neural networks withth eir attendant learning routines as meta-learning systems. Our system derived complex well performing learning algorithms from scratch. In this paper we also show that our approachp erforms non-stationary time series prediction. **(3):** Different researchers hold different views of what the term meta-learning exactly means. The first part of this paper provides our own perspective view in which the goal is to build self-adaptive learners (i.e. learning algorithms that improve their bias dynamically through experience by accumulating meta-knowledge). The second part provides a survey of meta-learning as reported by the machine-learning literature. We find that, despite different views and research lines, a question remains constant: how can we exploit knowledge about learning (i.e. meta-knowledge) to improve the performance of learning algorithms? Clearly the answer to this question is key to the advancement of the field and continues being the subject of intensive research. **(4):** The move from hand-designed features to learned features in machine learning has been wildly successful. In spite of this, optimization algorithms are still designed by hand. In this paper we show how the design of an optimization algorithm can be cast as a learning problem, allowing the algorithm to learn to exploit structure in the problems of interest in an automatic way. Our learned algorithms, implemented by LSTMs, outperform generic, hand-designed competitors on the tasks for which they are trained, and also generalize well to new tasks with similar structure. We demonstrate this on a number of tasks, including simple convex problems, training neural networks, and styling images with neural art.	Another bunch of related works include hyper-parameter optimization @cite , meta-learning @cite and learning to learn methods @cite @cite . However, the goal of these works is to use meta-data to improve the performance of the existing algorithms, such as finding the optimal learning rate of optimization methods or the optimal number of hidden layers to construct the network. In this paper, we focus on learning the entire topological architecture of network blocks to improve the performance.	**Rouge scores:** - Rouge 2: 0.0596 (prec) 0.1169 (recall) 0.0789 (f-1) - Rouge L: 0.1053 (prec) 0.2051 (recall) 0.1391 (f-1) **Summary:** Convolutional neural networks have gained a remarkable success in computer vision. However, most usable network architectures are hand-crafted and usually require expertise and elaborate design. In this paper, we provide a block-wise network generation pipeline called BlockQNN which automatically builds high-performance networks using the Q-Learning paradigm with epsilon-greedy exploration strategy. The optimal network block is constructed by the learning agent which is trained sequentially to choose component layers. We stack the block to construct the whole auto-generated network. To accelerate the generation process, we also propose a distributed asynchronous framework and an early stop strategy. The block-wise generation brings unique advantages: (1) it performs competitive results in comparison to the hand-crafted state-of-the-art networks on image classification, additionally, the best network generated by BlockQNN achieves 3.54 top-1 error rate on CIFAR-10 which beats all existing auto-generate networks. (2) **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0723 (prec) 0.0779 (recall) 0.075 (f-1) - Rouge L: 0.1548 (prec) 0.1667 (recall) 0.1605 (f-1) **Summary:** In this paper we show how the design of an optimization algorithm can be cast as a learning problem, allowing the algorithm to learn to exploit structure in the problems of interest in an automatic way. we make meta-learning in large systems feasible by using recurrent neural networks withth eir attendant learning routines as meta-learning systems. our learned algorithms, outperform generic, hand-designed competitors on the tasks for which they are trained, and also generalize well to new tasks with similar structure. **Comment:** - Copied most of the 4th article (purple) and just inserted a sentence from the 2nd (green) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0513 (prec) 0.1299 (recall) 0.0735 (f-1) - Rouge L: 0.0969 (prec) 0.2436 (recall) 0.1387 (f-1) **Summary:** Convolutional neural networks have gained a remarkable success in computer vision. However, most usable network architectures are hand-crafted and usually require expertise and elaborate design. In this paper, we provide a block-wise network generation pipeline called BlockQNN which automatically builds high-performance networks using the Q-learning paradigm with epsilon-greedy exploration strategy. The optimal network block is constructed by the learning agent which is trained sequentially to choose component layers. We stack the block to construct the whole auto-generated network. To accelerate the generation process, we also propose a distributed asynchronous framework and an early stop strategy. The block-wise generation brings unique advantages: (1) it performs competitive results in comparison to the hand-crafted state-of-the-art networks on image classification, additionally, the best network generated by BlockQNN achieves 3.54 top-1 error rate on CIFAR-10 which beats all existing auto-generate networks. (2) in the meanwhile, it offers tremendous reduction of the search space in designing networks which only spends 3 days with 32 GPUs, and (3) moreover, it has strong generalizability that the network built on CIFar also performs well on a larger-scale ImageNet dataset. **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0811 (prec) 0.039 (recall) 0.0526 (f-1) - Rouge L: 0.1842 (prec) 0.0897 (recall) 0.1207 (f-1) **Summary:** In this paper, we propose a block-wise network generation pipeline using the Q-learning paradigm with epsilon-greedy exploration strategy. We also propose a distributed asynchronous framework and an early stop strategy to accelerate the generation process. **Comment:** - Only extracted 2 sentences from the 1st article (yellow) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0952 (prec) 0.1299 (recall) 0.1099 (f-1) - Rouge L: 0.1698 (prec) 0.2308 (recall) 0.1957 (f-1) **Summary:** Recently, meta-learning @cite @cite has gained a lot of attention in the machine learning literature. In this paper, we propose a block-wise network generation pipeline called BlockQNN which automatically builds high-performance networks using the Q-Learning paradigm with epsilon-greedy exploration strategy. The optimal network block is constructed sequentially by the learning agent which is trained sequentially to choose component layers. In addition, it offers tremendous reduction of the search space in designing networks which only spend 3 days with 32 GPUs, and it has strong generalizability that the network built on CIFAR-10 also performs well on a larger-scale ImageNet dataset. **Comment:** - No indication of MDS, except for the red part where the model noted "meta-learning" is a recurring theme	**Rouge scores:** - Rouge 2: 0.0899 (prec) 0.1039 (recall) 0.0964 (f-1) - Rouge L: 0.2111 (prec) 0.2436 (recall) 0.2262 (f-1) **Summary:** @cite @cite proposed a gradient descent method for meta-learning. The gradient descent method is similar to our method, but differs from our method in two important aspects. First, the gradient descent method does not rely on the gradient of the gradient. Second, it does not require the gradient of a gradient to be learned. In this paper, we show that the gradient of our method can be used as a meta-learning method. We also show that our approach can be applied to large systems with few free parameters. **Comment:** - Shows some sign of MDS, with the 2nd article (green) being mentioned. - However, the other parts are mostly hallucinations

(c) Long samples

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
76	**(1):** We address the problem of cross-modal fine-grained action retrieval between text and video. Cross-modal retrieval is commonly achieved through learning a shared embedding space, that can indifferently embed modalities. In this paper, we propose to enrich the embedding by disentangling parts-of-speech (PoS) in the accompanying captions. We build a separate multi-modal embedding space for each PoS tag. The outputs of multiple PoS embeddings are then used as input to an integrated multi-modal space, where we perform action retrieval. All embeddings are trained jointly through a combination of PoS-aware and PoS-agnostic losses. Our proposal enables learning specialised embedding spaces that offer multiple views of the same embedded entities. We report the first retrieval results on fine-grained actions for the large-scale EPIC dataset, in a generalised zero-shot setting. Results show the advantage of our approach for both video-to-text and text-to-video action retrieval. We also demonstrate the benefit of disentangling the PoS for the generic task of cross-modal video retrieval on the MSR-VTT dataset. **(2):** We describe a novel cross-modal embedding space for actions, named Action2Vec, which combines linguistic cues from class labels with spatio-temporal features derived from video clips. Our approach uses a hierarchical recurrent network to capture the temporal structure of video features. We train our embedding using a joint loss that combines classification accuracy with similarity to Word2Vec semantics. We evaluate Action2Vec by performing zero shot action recognition and obtain state of the art results on three standard datasets. In addition, we present two novel analogy tests which quantify the extent to which our joint embedding captures distributional semantics. This is the first joint embedding space to combine verbs and action videos, and the first to be thoroughly evaluated with respect to its distributional semantics. **(3):** Inspired by recent advances in multimodal learning and machine translation, we introduce an encoder-decoder pipeline that learns (a): a multimodal joint embedding space with images and text and (b): a novel language model for decoding distributed representations from our space. Our pipeline effectively unifies joint image-text embedding models with multimodal neural language models. We introduce the structure-content neural language model that disentangles the structure of a sentence to its content, conditioned on representations produced by the encoder. The encoder allows one to rank images and sentences while the decoder can generate novel descriptions from scratch. Using LSTM to encode sentences, we match the state-of-the-art performance on Flickr8K and Flickr30K without using object detections. We also set new best results when using the 19-layer Oxford convolutional network. Furthermore we show that with linear encoders, the learned embedding space captures multimodal regularities in terms of vector space arithmetic e.g. *image of a blue car* - "blue" + "red" is near images of red cars. Sample captions generated for 800 images are made available for comparison. **(4):** Querying with an example image is a simple and intuitive interface to retrieve information from a visual database. Most of the research in image retrieval has focused on the task of instance-level image retrieval, where the goal is to retrieve images that contain the same object instance as the query image. In this work we move beyond instance-level retrieval and consider the task of semantic image retrieval in complex scenes, where the goal is to retrieve images that share the same semantics as the query image. We show that, despite its subjective nature, the task of semantically ranking visual scenes is consistently implemented across a pool of human annotators. We also show that a similarity based on human-annotated region-level captions is highly correlated with the human ranking and constitutes a good computable surrogate. Following this observation, we learn a visual embedding of the images where the similarity in the visual space is correlated with their semantic similarity surrogate. We further extend our model to learn a joint embedding of visual and textual cues that allows one to query the database using a text modifier in addition to the query image, adapting the results to the modifier. Finally, our model can ground the ranking decisions by showing regions that contributed the most to the similarity between pairs of images, providing a visual explanation of the similarity. **(5):** Constructing a joint representation invariant across different modalities (e.g., video, language) is of significant importance in many multimedia applications. While there are a number of recent successes in developing effective image-text retrieval methods by learning joint representations, the video-text retrieval task, however, has not been explored to its fullest extent. In this paper, we study how to effectively utilize available multimodal cues from videos for the cross-modal video-text retrieval task. Based on our analysis, we	Representing text Early works in image-to-text cross-modal retrieval @cite @cite @cite used TF-IDF as a weighted bag-of-words model for text representations (either from a word embedding model or one-hot vectors) in order to aggregate variable length text captions into a single fixed sized representation. With the advent of neural networks, works shifted to use RNNs, Gated Recurrent Units (GRU) or Long Short-Term Memory (LSTM) units to extract textual features @cite or to use these models within the embedding network @cite @cite @cite @cite @cite for both modalities.	**Rouge scores:** - Rouge 2: 0.0194 (prec) 0.0319 (recall) 0.0241 (f-1) - Rouge L: 0.0769 (prec) 0.1263 (recall) 0.0956 (f-1) **Summary:** -modal features (different visual characteristics, audio inputs, and text) by a fusion strategy for efficient retrieval. Furthermore, we explore several loss functions in training the embedding and propose a modified pairwise ranking loss for the task. Experiments on MSVD and MSR-VTT datasets demonstrate that our method achieves significant performance gain compared to the state-of-the-art approaches.||||Learning a joint language-visual embedding has a number of very appealing properties and can result in variety of practical application, including natural language image video annotation and search. In this work, we study three different joint language-Visual neural network model architectures. We evaluate our models on large scale LSMDC16 movie dataset for two tasks: 1) Standard Ranking for video annotation and retrieval 2) Our proposed movie multiple-choice test. This test facilitate automatic evaluation of visual-language models for natural language video annotation based on human activities. In addition to original Audio Description (AD) captions **Comment:** - Copied from the middle of the 5th article (orange) - No sign of MDS	**Rouge scores:** - Rouge 2: 0.0327 (prec) 0.0532 (recall) 0.0405 (f-1) - Rouge L: 0.1169 (prec) 0.1895 (recall) 0.1446 (f-1) **Summary:** we address the problem of cross-modal fine-grained action retrieval between text and video. we propose a novel framework for learning joint embeddings of images and text using a two-branch neural network with multiple layers of linear projections followed by nonlinearities. the network is trained using a largemargin objective that combines cross-view ranking constraints with within-view neighborhood structure preservation constraints inspired by metric learning literature. extensive experiments show that our approach gains significant improvements in accuracy for image-to-text and textto-image retrieval. In addition, we show that a similarity based on human-annotated region-level captions is highly correlated with the human ranking and constitutes a good computable surrogate. we further extend our model to learn a joint embedding of images and textual cues that allows one to query the database using a text modifier in addition to the query image modifier. finally, we show how the proposed **Comment:** - Very weak signs of MDS, with the 1st (yellow), 4th (purple) and 8th (deep blue) articles covered. - However, they are all presented as a single document. - The model also performs like an extractive model	**Rouge scores:** - Rouge 2: 0.0279 (prec) 0.0532 (recall) 0.0366 (f-1) - Rouge L: 0.1056 (prec) 0.2 (recall) 0.1382 (f-1) **Summary:** "We address the problem of cross-modal fine-grained action retrieval between text and video. Cross-modal retrieval is commonly achieved through learning a shared embedding space, that can indifferently embed modalities. In this paper, we propose to enrich the embedding by disentangling parts-of-speech (PoS) in the accompanying captions. We build a separate multi-modal embedding space for each PoS tag. The outputs of multiple PoS embeddings are then used as input to an integrated multi-Modal space, where we perform action retrieval. All embeddeddings are trained jointly through a combination of PoS-aware and PoS-agnostic losses. Our proposal enables learning specialised embedding spaces that offer multiple views of the same embedded entities. We report the first retrieval results on fine-grains actions for the large-scale EPIC dataset, in a generalised zero-shot setting. Results show the advantage of our approach for both video-to-text and text-to-video action retrieval. We also demonstrate the benefit of disentangling the PoS for the generic task of cross-Modal video retrieval on the MSR-VTT dataset." "We **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0458 (prec) 0.0638 (recall) 0.0533 (f-1) - Rouge L: 0.1439 (prec) 0.2 (recall) 0.1674 (f-1) **Summary:** In the context of video-to-text retrieval, the task of image-to-text retrieval has been studied for decades @cite @cite, @cite and @cite. In recent years, a `number of approaches have been proposed to learn joint embeddings of images and text using a two-branch neural network with multiple layers of linear projections followed by nonlinearities. @cite proposed a multi-modal embedding for image and text embeddings, which can be trained jointly through a combination of neural language models and multimodal neural language models. In @cite a joint embedding for images and text embedding was proposed for image-sentence retrieval, where the goal is to retrieve images that contain the same object instance as the query image. In this work, we focus on the task of video and text retrieval. **Comment:** - Some signs of MDS, with the 1st (yellow) article used as context, and the 4th (purple) and 8th (deep blue) articles covered. - However, the 2nd yellow part also shows that the work being done in the main article is mistaken as one of the references. - The last part on the task of the main paper is possibly a hallucination.	**Rouge scores:** - Rouge 2: 0.0272 (prec) 0.0426 (recall) 0.0332 (f-1) - Rouge L: 0.1554 (prec) 0.2421 (recall) 0.1893 (f-1) **Summary:** Cross-modal retrieval has been a long-standing research problem in computer vision and natural language processing (NLP) @cite @cite. In recent years, several approaches have been proposed to learn joint embeddings for image-text and text-to-video tasks. @cite proposed to learn a joint embedding of visual and textual cues that allows one to query the database using a text modifier in addition to the query image, adapting the results to the modifier. In @cite, a multi-modal joint embedding space is proposed to capture the temporal structure of video features. The joint embedding is then used to disambiguate fine-grained visual concepts in the sentence embedding process. In addition, @cite and @cite propose to learn an embedding of images where the similarity in the visual space is correlated with the semantic similarity surrogate. In @math, @cite learn a joint representation invariant across different video modalities. **Comment:** - Much stronger signs of MDS, with the information from the 1st article (yellow) used as context, and then the 2nd (green), 4th (purple), 5th (orange), 7th (deep yellow) and 8th (deep blue) all referenced.	**Rouge scores:** - Rouge 2: 0.0455 (prec) 0.0745 (recall) 0.0565 (f-1) - Rouge L: 0.1097 (prec) 0.1789 (recall) 0.136 (f-1) **Summary:** In @cite @cite, the authors propose to use a PoS-agnostic loss to enrich the embedding space by disentangling the parts-of-speech (PoS) in the accompanying captions. However, the authors do not consider the use of the PoS loss in the context of video-to-video action retrieval. In @cite, the authors propose a joint embedding space for noun and action videos by augmenting the training set with millions of weak annotated images with strong and noisy annotations @cite. @cite proposes to use a two-branch neural network with multiple layers of linear projections followed by nonlinearities. The network is trained using a largemargin objective that combines cross-view ranking constraints with within-view neighborhood structure preservation constraints inspired by metric learning literature. However, their method is based on a multi-layer neural network with a single layer of linear projections and nonlinearities and is trained on the Flickr30K Entities image-sentence dataset @cite. **Comment:** - Shows some sign of MDS, with the multiple articles being mentioned. - However, the red parts are mostly hallucinations

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
485	**(1):** Learning from web data has attracted lots of research interest in recent years. However, crawled web images usually have two types of noises, label noise and background noise, which induce extra difficulties in utilizing them effectively. Most existing methods either rely on human supervision or ignore the background noise. In this paper, we propose a novel method, which is capable of handling these two types of noises together, without the supervision of clean images in the training stage. Particularly, we formulate our method under the framework of multi-instance learning by grouping ROIs (i.e., images and their region proposals) from the same category into bags. ROIs in each bag are assigned with different weights based on the representative discriminative scores of their nearest clusters, in which the clusters and their scores are obtained via our designed memory module. Our memory module could be naturally integrated with the classification module, leading to an end-to-end trainable system. Extensive experiments on four benchmark datasets demonstrate the effectiveness of our method. **(2):** We study the problem of automatically removing outliers from noisy data, with application for removing outlier images from an image collection. We address this problem by utilizing the reconstruction errors of an autoencoder. We observe that when data are reconstructed from low-dimensional representations, the inliers and the outliers can be well separated according to their reconstruction errors. Based on this basic observation, we gradually inject discriminative information in the learning process of an autoencoder to make the inliers and the outliers more separable. Experiments on a variety of image datasets validate our approach. **(3):** We present a theoretically grounded approach to train deep neural networks, including recurrent networks, subject to class-dependent label noise. We propose two procedures for loss correction that are agnostic to both application domain and network architecture. They simply amount to at most a matrix inversion and multiplication, provided that we know the probability of each class being corrupted into another. We further show how one can estimate these probabilities, adapting a recent technique for noise estimation to the multi-class setting, and thus providing an end-to-end framework. Extensive experiments on MNIST, IMDB, CIFAR-10, CIFAR-100 and a large scale dataset of clothing images employing a diversity of architectures — stacking dense, convolutional, pooling, dropout, batch normalization, word embedding, LSTM and residual layers — demonstrate the noise robustness of our proposals. Incidentally, we also prove that, when ReLU is the only non-linearity, the loss curvature is immune to class-dependent label noise. **(4):** In this paper, we study the problem of learning image classification models with label noise. Existing approaches depending on human supervision are generally not scalable as manually identifying correct or incorrect labels is time-consuming, whereas approaches not relying on human supervision are scalable but less effective. To reduce the amount of human supervision for label noise cleaning, we introduce CleanNet, a joint neural embedding network, which only requires a fraction of the classes being manually verified to provide the knowledge of label noise that can be transferred to other classes. We further integrate CleanNet and conventional convolutional neural network classifier into one framework for image classification learning. We demonstrate the effectiveness of the proposed algorithm on both of the label noise detection task and the image classification on noisy data task on several large-scale datasets. Experimental results show that CleanNet can reduce label noise detection error rate on held-out classes where no human supervision available by 41.5 compared to current weakly supervised methods. It also achieves 47 of the performance gain of verifying all images with only 3.2 images verified on an image classification task. Source code and dataset will be available at kuanghuei.github.io CleanNetProject. **(5):** We present a simple yet efficient approach capable of training deep neural networks on large-scale weakly-supervised web images, which are crawled raw from the Internet by using text queries, without any human annotation. We develop a principled learning strategy by leveraging curriculum learning, with the goal of handling a massive amount of noisy labels and data imbalance effectively. We design a new learning curriculum by measuring the complexity of data using its distribution density in a feature space, and rank the complexity in an unsupervised manner. This allows for an efficient implementation of curriculum learning on large-scale web images, resulting in a high-performance CNN the model, where the negative impact of noisy labels is reduced substantially. Importantly, we show by experiments that those images with highly noisy labels can surprisingly improve the generalization capability of model, by serving as a manner of regularization. Our approaches	In learning classifier with web data, previous works focus on handling the label noise in three directions, removing label noise @cite @cite @cite @cite @cite @cite , building noise-robust model @cite @cite @cite @cite @cite @cite , and curriculum learning @cite @cite .	**Rouge scores:** - Rouge 2: 0.0175 (prec) 0.0732 (recall) 0.0283 (f-1) - Rouge L: 0.064 (prec) 0.2619 (recall) 0.1028 (f-1) **Summary:** Learning from web data has attracted lots of research interest in recent years. However, it is still very difficult to learn from web data without the supervision of a human. In this paper, we study the problem of learning from web data. The problem is that web images usually have two types of noises, label noise and background noise, which induce extra difficulties in utilizing them effectively. Most existing methods either rely on human supervision or ignore the background noise. In this article, we propose a novel method, which is capable of handling these two types of noise together, without the supervision or supervision of clean images in the training stage. Particularly, we formulate our method under the framework of multi-instance learning by grouping ROIs (i.e., images and their region proposals) from the same category into bags. ROIs in each bag are assigned with different weights based on the representative discriminative scores of their nearest clusters, in which the clusters and their scores are obtained via our designed memory module. **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0258 (prec) 0.0976 (recall) 0.0408 (f-1) - Rouge L: 0.0705 (prec) 0.2619 (recall) 0.1111 (f-1) **Summary:** learning from web data is increasingly popular due to abundant free web resources. however, the performance gap between webly supervised learning and traditional supervised learning is still very large, due to the label noise of web data as well as the domain shift between web data and test data. to fill this gap, most existing methods propose to purify or augment web data using instance-level supervision, which generally requires heavy annotation. instead, we propose to address the label noise and domain shift by using more accessible category-level supervision. in particular, we build our deep probabilistic framework upon variational autoencoder (VAE), in which classification network and VAE can jointly leverage category-level hybrid information. our memory module could be naturally integrated with the classification module, leading to an end-to-end trainable system. we design a new learning curriculum by measuring the complexity of data using its distribution density, in an unsupervised manner in an **Comment:** - Copied the 6th article, except replacing the last 2 sentences with another sentence from the 5th article (orange) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0048 (prec) 0.0244 (recall) 0.0081 (f-1) - Rouge L: 0.0433 (prec) 0.2143 (recall) 0.072 (f-1) **Summary:** We present a simple yet efficient approach capable of training deep neural networks on large-scale weakly-supervised web images, which are crawled raw from the Internet by using text queries, without any human supervision. We develop a principled learning strategy by leveraging curriculum learning, with the goal of handling a massive amount of noisy labels and data imbalance effectively. We design a new learning curriculum by measuring the complexity of data using its distribution density in a feature space, and rank the complexity in an unsupervised manner. This allows for an efficient implementation of curriculum learning on large- scale web images, resulting in a high-performance CNN the model, where the negative impact of noisy labels is reduced substantially. Importantly, we show by experiments that those images with highly noisy labels can surprisingly improve the generalization capability of model, by serving as a manner of regularization. Our approaches obtain state-of-the-art performance on four benchmarks: WebVision, ImageNet, Clothing-1M and Food-101. With an ensemble of multiple models, we achieved a top-5 error rate of 5.2 on the WebVision challenge [18] for 1000-category classification. This result was the top performance by a wide margin, outperforming second place by a nearly 50 relative **Comment:** - Copied the 5th article only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0485 (prec) 0.122 (recall) 0.0694 (f-1) - Rouge L: 0.1058 (prec) 0.2619 (recall) 0.1507 (f-1) **Summary:** Label noise is an important issue in image classification, with many potential negative consequences @cite @cite. For example, the label noise can be class-conditional @cite, which is not scalable for large-scale weakly-supervised learning. To address this problem, several methods have been proposed. @cite proposed a method for learning from unlabeled web data, which is based on the assumption that the unlabeled data can be used to improve the classification performance. However, these methods do not consider the label noise of web data. In contrast, our method is able to deal with the label noise without the supervision of the classification module. **Comment:** - Some sign of MDS, using the context information from the 1st article (yellow) as background, and also mentioned correctly some elements of the current study - Elements of 4th article (purple) is clear while the 1st red part is true for many samples.	**Rouge scores:** - Rouge 2: 0.0266 (prec) 0.122 (recall) 0.0437 (f-1) - Rouge L: 0.0688 (prec) 0.3095 (recall) 0.1126 (f-1) **Summary:** The label noise problem has been widely studied in the literature @cite @cite. In @cite, the label noise is treated as a label noise, and the label noise can be class-conditional. @cite and @cite address the label noise by using the abundant surrogate loss functions designed for the traditional classification problem when there is label noise. Recently, @cite proposed a method to remove outlier images from an image collection by utilizing the reconstruction errors of an autoencoder. In this paper, we propose a novel method, which is capable of handling both label noise and background noise together, without the supervision of clean images in the training stage. In addition, we formulate our method under the framework of multi-instance learning by grouping ROIs (i.e., images and their region proposals) from the same category into bags. ROIs in each bag are assigned with different weights based on the representative discriminative scores of their nearest clusters, in which the clusters and their scores are obtained via our designed memory module. Our memory module could be naturally integrated with the classification module, leading to an end-to-end trainable system. **Comment:** - Clear signs of MDS, with the 1st article (yellow) used for both context (at the beginning) and contrast with other studies in the latter half - 3 other articles are cited, though the red parts shows a poorly written sentence.	**Rouge scores:** - Rouge 2: 0.0287 (prec) 0.1463 (recall) 0.048 (f-1) - Rouge L: 0.0667 (prec) 0.3333 (recall) 0.1111 (f-1) **Summary:** Recently, there has been a lot of research on label noise cleaning. In @cite @cite, the label noise is treated as a set of instances that are mislabeled, and the label cleansing algorithm is proposed to deal with the label noise. In addition, @cite proposed an active learning approach for removing outliers from an image collection @cite, which does not require a large amount of training data to train the model, and they do not need any annotated labels for training. Moreover, the label noise cleansing algorithms are defined as follows: first, the labels of the instances are labeled, and second, the labels on the labels are labeled, respectively. The labels are labeled by the label noise correction algorithm and the labels of mislabeled instances are labeled by label noise cleansing. @cite used curriculum learning to train a deep neural network on large-scale web images, which are crawled raw from the Internet by using text queries, without any human annotation. However, curriculum learning can significantly improve the generalization capability of the model, by reducing the number of noisy labels, and by using curriculum learning to reduce the amount of data imbalance in the training process. Moreover, curriculum learning is effective in handling noisy labels and the data imbalance effectively. **Comment:** - Some sign of MDS, with the 1st article (yellow) used for context, and the 2nd (green), 7th (pink) and 8th (deep blue) articles referenced. - Contains much hallucinations though as noted in the red parts.

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
1368	**(1):** Low latency services such as credit-card fraud detection and website targeted advertisement rely on Big Data platforms (e.g., Lucene, Graphchi, Cassandra) which run on top of memory managed runtimes, such as the JVM. These platforms, however, suffer from unpredictable and unacceptably high pause times due to inadequate memory management decisions (e.g., allocating objects with very different lifetimes next to each other, resulting in memory fragmentation). This leads to long and frequent application pause times, breaking Service Level Agreements (SLAs). This problem has been previously identified and results show that current memory management techniques are ill-suited for applications that hold in memory massive amounts of middle to long-lived objects (which is the case for a wide spectrum of Big Data applications). Previous works try to reduce such application pauses by allocating objects off-heap or in special allocation regions generations, thus alleviating the pressure on memory management. However, all these solutions require a combination of programmer effort and knowledge, source code access, or off-line profiling, with clear negative impact on programmer productivity and or application performance. This paper presents ROLP, a runtime object lifetime profiling system. ROLP profiles application code at runtime in order to identify which allocation contexts create objects with middle to long lifetimes, given that such objects need to be handled differently (regarding short-lived ones). This profiling information greatly improves memory management decisions, leading to long tail latencies reduction of up to 51 for Lucene, 85 for GraphChi, and 60 for Cassandra, with negligible throughput and memory overhead. ROLP is implemented for the OpenJDK 8 HotSpot JVM and it does not require any programmer effort or source code access. **(2):** Future high-performance virtual machines will improve performance through sophisticated online feedback-directed optimizations. This paper presents the architecture of the Jalapeno Adaptive Optimization System, a system to support leading-edge virtual machine technology and enable ongoing research on online feedback-directed optimizations. We describe the extensible system architecture, based on a federation of threads with asynchronous communication. We present an implementation of the general architecture that supports adaptive multi-level optimization based purely on statistical sampling. We empirically demonstrate that this profiling technique has low overhead and can improve startup and steady-state performance, even without the presence of online feedback-directed optimizations. The paper also describes and evaluates an online feedback-directed inlining optimization based on statistical edge sampling. The system is written completely in Java, applying the described techniques not only to application code and standard libraries, but also to the virtual machine itself. **(3):** We introduce Elephant Tracks (ET), a dynamic program analysis tool for producing traces of garbage collection events, including object allocations, object deaths, and pointer updates. Like prior work, our tracing tool is based on the Merlin algorithm [2002], but offers several substantial new capabilities. First, it is much more precise than previous tools: it measures time in terms of method entries and exits, allowing it to place events precisely in the context of the program structure. Furthermore, we exploit this notion of time to actually reduce the amount of work required to compute precise object death times. Second, it is implemented using a combination of JVM Tool Interface (JVMTI) Microsystems [2004] callbacks and bytecode rewriting, and works with any standard JVM. Finally, it properly handles Java Reference objects, including weak references, finalization, and multithreaded programs. **(4):** Many profilers based on bytecode instrumentation yield wrong results in the presence of an optimizing dynamic compiler, either due to not being aware of optimizations such as stack allocation and method inlining, or due to the inserted code disrupting such optimizations. To avoid such perturbations, we present a novel technique to make any profiler implemented at the bytecode-level aware of optimizations performed by the dynamic compiler. We implement our approach in a state-of-the-art Java virtual machine and demonstrate its significance with concrete profilers. We quantify the impact of escape analysis on allocation profiling, object life-time analysis, and the impact of method inlining on callsite profiling. We illustrate how our approach enables new kinds of profilers, such as a profiler for non-inlined callsites, and a testing framework for locating performance bugs in dynamic compiler implementations. **(5):** Programmers are writing a rapidly growing number of programs in object-oriented languages, such as Java and C#, that require garbage collection. Garbage collection traces and simulation speed up research by enabling deeper understandings of object lifetime behavior and quick exploration and	Profiling plays a key role in managed runtimes, either for code optimization or memory management decisions @cite @cite @cite @cite @cite @cite @cite . We focus on getting quality profiling information to drive object pretenuring. is, to the best of our knowledge, the first online profiler targeting the dynamic pretenuring of objects in Big Data applications running on HotSpot. This section compares our work with state-of-art systems, namely, off-line and online profilers that guide systems where small changes are needed in the heap organization and collection. It ends with a comparative analysis of systems that demand a more profound change, either to the application framework or the runtime itself, in some cases manipulating application-defined types and or organizing the heap in special purpose regions, and placing data directly in an off-heap space.	**Rouge scores:** - Rouge 2: 0.0323 (prec) 0.037 (recall) 0.0345 (f-1) - Rouge L: 0.1026 (prec) 0.1176 (recall) 0.1096 (f-1) **Summary:** Low latency services such as credit-card fraud detection and website targeted advertisement rely on Big Data platforms (e.g., Lucene, Graphchi, Cassandra) which run on top of memory managed runtimes, such as the JVM. These platforms, however, suffer from unpredictable and unacceptably high pause times. Merlin timestamps objects and later uses the timestamps of dead objects to reconstruct when they died. The Merlin algorithm piggybacks on garbage collections performed by the base system. Experimental results show that Merlin can generate traces over two orders of magnitude faster than the brute-force method which collects after every object allocation. We also use Merlin to produce visualizations of heap behavior that expose new object lifetime behaviors.||||Modern object-oriented applications commonly suffer from severe performance problems that need to be optimized away for increased efficiency and user satisfaction. Many existing optimization techniques (such as object pooling and pretenuring) require precise identification of object lifetimes. However, all these **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0 (prec) 0.0 (recall) 0.0 (f-1) - Rouge L: 0.12 (prec) 0.1324 (recall) 0.1259 (f-1) **Summary:** this paper presents the architecture of the Jalapeno Adaptive Optimization System, a system to support leading-edge virtual machine technology and enable ongoing research on online feedback-directed optimizations. we describe the extensible system architecture, based on a federation of threads with asynchronous communication, and we present an implementation of the general architecture that supports adaptive multi-level optimization based purely on statistical an online feedback - directed inlining optimization based on statistical edge sampling sampling. the paper also describes and evaluates an online feedback - directed inlining optimization based on statistical edge sampling. this paper presents a tunable profiling technique, called Resurrector, that explores the middle ground between high precision and high efficiency to find the precision-efficiency sweetspot for various livenessbased optimization techniques. we empirically demonstrate that this profiling technique has low overhead and can improve startup and steady-state performance, even without the presence of online feedback-directions. our approach enables new kinds of profilers, such as a profiler for **Comment:** - Copied the 2nd article only with a reshuffling of sentence order - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0142 (prec) 0.0222 (recall) 0.0173 (f-1) - Rouge L: 0.1038 (prec) 0.1618 (recall) 0.1264 (f-1) **Summary:** Programmers are writing a rapidly growing number of programs in object-oriented languages, such as Java and C#, that require garbage collection. Garbage collection traces and simulation speed up research by enabling deeper understandings of object lifetime behavior and quick exploration and design of new garbage collection algorithms. When generating perfect traces, the brute-force method of computing object lifetimes requires a whole-heap garbage collection at every potential collection point in the program. Because this process is prohibitively expensive, researchers often use granulated traces by collecting only periodically, for example, every 32 KB of allocation. We extend the state of the art for simulating garbage collection algorithms in two ways. First, we develop a systematic methodology for simulation studies of garbage collection and present results showing the effects of trace granularity on these simulations. We show that trace granularity often distorts simulated garbage collection results compared with perfect traces. Second, we present and measure the performance of a new algorithm called Merlin for computing objects lifetimes. Merlin timestamps objects and later uses the timestamps of dead objects to reconstruct when they died. The Merlin algorithm piggybacks on garbage collections performed by the base system. Experimental results show that Merlin can generate traces over two orders of magnitude faster than the brute- **Comment:** - Copied the 5th article only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0482 (prec) 0.0296 (recall) 0.0367 (f-1) - Rouge L: 0.1786 (prec) 0.1103 (recall) 0.1364 (f-1) **Summary:** Object lifetime profiling @cite @cite has been widely studied in the literature. However, most of the existing works focus on the analysis of allocation contexts, such as stack allocation @cite, object allocations @cite and pointer updates @cite. These approaches are based on the Merlin algorithm @cite which is based on the JVM Tool Interface (JVMTI) Microsystems @cite that supports adaptive multi-level optimization based on statistical sampling. However, these approaches are not suitable for large-scale applications, such as data structures with disjoint lifetimes. **Comment:** - Clear signs of MDS, with the 1st article (yellow) used for context (at the beginning), with the 3rd (blue), 4th (purple) and 5th (deep blue) articles are reference. - The last yellow sentence is in line with the 1st article, though the words in red are adopted from the 6th article which the commenter does not have knowledge to judge.	**Rouge scores:** - Rouge 2: 0.0169 (prec) 0.0074 (recall) 0.0103 (f-1) - Rouge L: 0.2 (prec) 0.0882 (recall) 0.1224 (f-1) **Summary:** @cite @cite propose a tunable profiling technique, called Resurrector, that explores the middle ground between high precision and high efficiency to find the precision-efficiency sweetspot for various liveness-based optimization techniques. @cite present an implementation of the Jalapeno Adaptive Optimization System, a system to support leading-edge virtual machine technology and enable ongoing research on online feedback-directed optimizations. **Comment:** - Slightly weaker MDS performance compared to tuned LED, with the 2nd (green) and 6th (orange) articles referenced - Compared to the LED summary, no reference is made as to what the main (i.e. 1st article) does	**Rouge scores:** - Rouge 2: 0.0164 (prec) 0.0148 (recall) 0.0156 (f-1) - Rouge L: 0.1301 (prec) 0.1176 (recall) 0.1236 (f-1) **Summary:** @cite @cite is a dynamic program analysis tool for producing traces of garbage collection events, including object allocations, object deaths, and pointer updates. It is based on the Merlin algorithm @cite, which is based on Java Reference Objects (JCU) and is designed to be used with any standard JVM. It is not designed for dynamic program analysis, since it does not support dynamic program analysis and does not consider memory usage, which makes it not suitable for dynamic analysis, especially for dynamic compiler implementations. @cite present a technique to make any profiler aware of optimizations performed by the dynamic compiler. They demonstrate that their approach can be used to detect errors in dynamic compiler implementations, and it does not rely on escape analysis. **Comment:** - Weakest of the three tuned models but still shows MDS signs, with the 3rd (blue) and 4th (purple) articles summarized - Red part is basically hallucination though

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
1864	**(1):** While word embeddings have been shown to implicitly encode various forms of attributional knowledge, the extent to which they capture relational information is far more limited. In previous work, this limitation has been addressed by incorporating relational knowledge from external knowledge bases when learning the word embedding. Such strategies may not be optimal, however, as they are limited by the coverage of available resources and conflate similarity with other forms of relatedness. As an alternative, in this paper we propose to encode relational knowledge in a separate word embedding, which is aimed to be complementary to a given standard word embedding. This relational word embedding is still learned from co-occurrence statistics, and can thus be used even when no external knowledge base is available. Our analysis shows that relational word vectors do indeed capture information that is complementary to what is encoded in standard word embeddings. **(2)**: Despite the emergence and growth of numerous large knowledge graphs, many basic and important facts about our everyday world are not readily available on the Web. To address this, we present WebBrain, a new approach for harvesting commonsense knowledge that relies on joint learning from Web-scale data to fill gaps in the knowledge acquisition. We train a neural network model to learn relations based on large numbers of textual patterns found on the Web. At the same time, the model learns vector representations of general word semantics. This joint approach allows us to generalize beyond the explicitly extracted information. Experiments show that we can obtain representations of words that reflect their semantics, yet also allow us to capture conceptual relationships and commonsense knowledge. **(3):** We present LEAR (Lexical Entailment Attract-Repel), a novel post-processing method that transforms any input word vector space to emphasise the asymmetric relation of lexical entailment (LE), also known as the IS-A or hyponymy-hypernymy relation. By injecting external linguistic constraints (e.g., WordNet links) into the initial vector space, the LE specialisation procedure brings true hyponymy-hypernymy pairs closer together in the transformed Euclidean space. The proposed asymmetric distance measure adjusts the norms of word vectors to reflect the actual WordNet-style hierarchy of concepts. Simultaneously, a joint objective enforces semantic similarity using the symmetric cosine distance, yielding a vector space specialised for both lexical relations at once. LEAR specialisation achieves state-of-the-art performance in the tasks of hypernymy directionality, hypernymy detection, and graded lexical entailment, demonstrating the effectiveness and robustness of the proposed asymmetric specialisation model. **(4):** In this paper, we propose a general framework to incorporate semantic knowledge into the popular data-driven learning process of word embeddings to improve the quality of them. Under this framework, we represent semantic knowledge as many ordinal ranking inequalities and formulate the learning of semantic word embeddings (SWE) as a constrained optimization problem, where the data-derived objective function is optimized subject to all ordinal knowledge inequality constraints extracted from available knowledge resources such as Thesaurus and WordNet. We have demonstrated that this constrained optimization problem can be efficiently solved by the stochastic gradient descent (SGD) algorithm, even for a large number of inequality constraints. Experimental results on four standard NLP tasks, including word similarity measure, sentence completion, name entity recognition, and the TOEFL synonym selection, have all demonstrated that the quality of learned word vectors can be significantly improved after semantic knowledge is incorporated as inequality constraints during the learning process of word embeddings. **(5):** Representing words into vectors in continuous space can form up a potentially powerful basis to generate high-quality textual features for many text mining and natural language processing tasks. Some recent efforts, such as the skip-gram model, have attempted to learn word representations that can capture both syntactic and semantic information among text corpus. However, they still lack the capability of encoding the properties of words and the complex relationships among words very well, since text itself often contains incomplete and ambiguous information. Fortunately, knowledge graphs provide a golden mine for enhancing the quality of learned word representations. In particular, a knowledge graph, usually composed by entities (words, phrases, etc.), relations between entities, and some corresponding meta information, can supply invaluable relational knowledge that encodes the relationship between entities as well as categorical knowledge that encodes the attributes or properties of entities. Hence, in this paper, we introduce a novel framework called RC-NET to leverage both the	Several authors have tried to improve word embeddings by incorporating external knowledge bases. For example, some authors have proposed models which combine the loss function of a word embedding model, to ensure that word vectors are predictive of their context words, with the loss function of a knowledge graph embedding model, to encourage the word vectors to additionally be predictive of a given set of relational facts @cite @cite @cite . Other authors have used knowledge bases in a more restricted way, by taking the fact that two words are linked to each other in a given knowledge graph as evidence that their word vectors should be similar @cite @cite . Finally, there has also been work that uses lexicons to learn word embeddings which are specialized towards certain types of lexical knowledge, such as hypernymy @cite @cite , antonymy @cite @cite or a combination of various linguistic constraints @cite .	**Rouge scores:** - Rouge 2: 0.0533 (prec) 0.0616 (recall) 0.0571 (f-1) - Rouge L: 0.1588 (prec) 0.1837 (recall) 0.1703 (f-1) **Summary:** word representations.||||| Machine learning about language can be improved by supplying it with specific knowledge and sources of external information. We present here a new version of the linked open data resource ConceptNet that is particularly well suited to be used with modern NLP techniques such as word embeddings. ConceptNet is a knowledge graph that connects words and phrases of natural language with labeled edges. Its knowledge is collected from many sources that include expert-created resources, crowd-sourcing, and games with a purpose. It is designed to represent the general knowledge involved in understanding language, improving natural language applications. This relational word embedding is still learned from co-occurrence statistics, and can thus be used even when no external knowledge base is available. Our analysis shows that relational word vectors do indeed capture information that is complementary to what is encoded in standard words.||||Despite the emergence and growth of numerous large knowledge graphs, many basic and important facts about our everyday world are not readily available on the **Comment:** - All copied, but interestingly started form the last tokens of the 5th article, copied the 6th, then switched back to the 2nd - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0284 (prec) 0.0274 (recall) 0.0279 (f-1) - Rouge L: 0.1549 (prec) 0.1497 (recall) 0.1522 (f-1) **Summary:** vector space word representations are learned from distributional information of words in large corpora. although such statistics are semantically informative, they disregard the valuable information that is contained in semantic lexicons such as WordNet, FrameNet, and the Paraphrase Database. this paper proposes a novel framework for refining vector space representations using relational information from semantic lexicons by encouraging linked words to have similar vector representations, and it makes no assumptions about how the input vectors were constructed. we present a novel post-processing method called LEAR (Lexical Entailment Attract-Repel) that transforms any input word vector space to emphasise the asymmetric relation of lexical entailment (LE), also known as the hyponymy-hypernymy relation. by injecting external linguistic constraints (e.g., WordNet links) into the initial vector space, the LE specialisation procedure brings true hyponymy - hypernymy pairs closer together in the transformed Eucl **Comment:** - Copied all but the last 2 sentences of the 9th article, then switched back to the 2nd (blue) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0333 (prec) 0.0479 (recall) 0.0393 (f-1) - Rouge L: 0.1327 (prec) 0.1905 (recall) 0.1564 (f-1) **Summary:** Machine learning about language can be improved by supplying it with specific knowledge and sources of external information. We present here a new version of the linked open data resource ConceptNet that is particularly well suited to be used with modern NLP techniques such as word embeddings. ConceptNet is a knowledge graph that connects words and phrases of natural language with labeled edges. Its knowledge is collected from many sources that include expert-created resources, crowd-sourcing, and games with a purpose. It is designed to represent the general knowledge involved in understanding language, improving natural language applications by allowing the application to better understand the meanings behind the words people use. When ConceptNet is combined with word embeddINGS acquired from distributional semantics (such as word2vec), it provides applications with understanding that they would not acquire from distributional mechanics alone, nor from narrower resources such as WordNet or DBPedia. We demonstrate this with state-of-the-art results on intrinsic evaluations of word relatedness that translate into improvements on applications of word vectors, including solving SAT-style analogies. In this paper, we present a novel framework called RC-NET to leverage both the relational and categorical knowledge to produce word representations of higher quality. Specifically, we build the relational knowledge and **Comment:** - Copied the 6th article in full, then switched to the middle oif the 5th (orange) - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0633 (prec) 0.0342 (recall) 0.0444 (f-1) - Rouge L: 0.2375 (prec) 0.1293 (recall) 0.1674 (f-1) **Summary:** There is a large body of work on incorporating relational knowledge into NLP tasks @cite @cite. In particular, @cite proposed a method to incorporate semantic knowledge into the learning process of word embeddings to improve the quality of them. In @cite, the authors proposed a method for incorporating the information about relationship between entities that they extract from knowledge bases. In this paper, we propose to incorporate the information about relation between entities that is learned from distributional semantics alone. **Comment:** - Clear signs of MDS, with the 1st article (yellow) used for context (at the beginning), with the 4th (purple) and 7th (pink) articles are reference. - The last sentence (red) attempts to contrast what is being down in the 1st article but used a phrase from the 6th article (deep blue) instead	**Rouge scores:** - Rouge 2: 0.0795 (prec) 0.0479 (recall) 0.0598 (f-1) - Rouge L: 0.2472 (prec) 0.1497 (recall) 0.1864 (f-1) **Summary:** Word embeddings have been widely used in NLP @cite @cite, where they have been shown to be useful for various NLP tasks, such as word similarity, sentence completion, name entity recognition, and so on @cite. However, these embeddings do not explicitly capture relational information, which is the focus of this paper. In contrast, our work aims to learn relational word embeddings that are complementary to a given word embedding, and can thus be used with modern NLP techniques such as WordNet, @cite and Word2Vec @cite to learn relational embeddings. **Comment:** - Clear signs of MDS, with the 1st article (yellow) used for context (at the beginning) and its difference also clearly stated in the latter part. - A part of the 4th (purple) article is also leveraged to provide further reference. - While totally not accurate, it appears the model learned that WordNet and Word2Vec are entities related to word embeddings and NLP techniques, thus copying them from the 6th article (deep blue)	**Rouge scores:** - Rouge 2: 0.033 (prec) 0.0205 (recall) 0.0253 (f-1) - Rouge L: 0.1848 (prec) 0.1156 (recall) 0.1423 (f-1) **Summary:** In @cite @cite, the authors proposed an asymmetric cosine distance measure to emphasise the hypernymy relation of lexical entailment. However, this distance measure does not consider hypernymy directionality, hypernymy detection, or hypernymy classification. Furthermore, in @cite the authors proposed a method for using the cosine distance to emphasise hypernymy, and they do not consider the asymmetric distance measure. In @cite, the authors presented a constrained optimization problem, where the data-derived objective function is optimized subject to all ordinal knowledge inequality constraints extracted from available knowledge resources such as Thesaurus and WordNet. **Comment:** - Some signs of MDS, with the 3rd (blue) and 4th (purple) articles are included in the summary. - However, the red part is once again a hallucination

No.	Abstracts	Label	Base LED	Large LED	Base Centrum	Tuned LED	Tuned Centrum	Two-step
1872	**(1):** We consider the well-studied game-theoretic version of machine scheduling in which jobs correspond to self-interested users and machines correspond to resources. Here each user chooses a machine trying to minimize her own cost, and such selfish behavior typically results in some equilibrium which is not globally optimal: An equilibrium is an allocation where no user can reduce her own cost by moving to another machine, which in general need not minimize the makespan, i.e., the maximum load over the machines. We provide tight bounds on two well-studied notions in algorithmic game theory, namely, the price of anarchy and the strong price of anarchy on machine scheduling setting which lies in between the related and the unrelated machine case. Both notions study the social cost (makespan) of the worst equilibrium compared to the optimum, with the strong price of anarchy restricting to a stronger form of equilibria. Our results extend a prior study comparing the price of anarchy to the strong price of anarchy for two related machines (Epstein, Acta Informatica 2010), thus providing further insights on the relation between these concepts. Our exact bounds give a qualitative and quantitative comparison between the two models. The bounds also show that the setting is indeed easier than the two unrelated machines: In the latter, the strong price of anarchy is @math , while in ours it is strictly smaller. **(2):** We study the outcome of natural learning algorithms in atomic congestion games. Atomic congestion games have a wide variety of equilibria often with vastly differing social costs. We show that in almost all such games, the well-known multiplicative-weights learning algorithm results in convergence to pure equilibria. Our results show that nat- ural learning behavior can avoid bad outcomes predicted by the price of anarchy in atomic congestion games such as the load-balancing game introduced by Koutsoupias and Pa- padimitriou, which has super-constant price of anarchy and has correlated equilibria that are exponentially worse than any mixed Nash equilibrium. Our results identify a set of mixed Nash equilibria that we call weakly stable equilibria. Our notion of weakly stable is defined game-theoretically, but we show that this property holds whenever a stability criterion from the theory of dy- namical systems is satisfied. This allows us to show that in every congestion game, the distribution of play converges to the set of weakly stable equilibria. Pure Nash equilibria are weakly stable, and we show using techniques from algebraic geometry that the converse is true with probability 1 when congestion costs are selected at random independently on each edge (from any monotonically parametrized distribu- tion). We further extend our results to show that players can use algorithms with different (sufficiently small) learn- ing rates, i.e. they can trade off convergence speed and long term average regret differently. **(3):** A strong equilibrium is a pure Nash equilibrium which is resilient to deviations by coalitions. We define the strong price of anarchy (SPoA) to be the ratio of the worst strong equilibrium to the social optimum. Differently from the Price of Anarchy (defined as the ratio of the worst Nash Equilibrium to the social optimum), it quantifies the loss incurred from the lack of a central designer in settings that allow for coordination. We study the SPoA in two settings, namely job scheduling and network creation. In the job scheduling game we show that for unrelated machines the SPoA can be bounded as a function of the number of machines and the size of the coalition. For the network creation game we show that the SPoA is at most 2. In both cases we show that a strong equilibrium always exists, except for a well defined subset of network creation games. **(4):** We consider the classical machine scheduling, where n jobs need to be scheduled on m machines, with the goal of minimizing the makespan, i.e., the maximum load of any machine in the schedule. We study inefficiency of schedules that are obtained when jobs arrive sequentially one by one, and choose themselves the machine on which they will be scheduled. Every job is only interested to be on a machine with a small load (and does not care about the loads of other machines). We measure the inefficiency of a schedule as the ratio of the makespan obtained in the worst-case equilibrium schedule, and of the optimum makespan. This ratio is known as the sequential price of anarchy. We also introduce alternative inefficiency measures, which allow for a favorable choice of the order in which the jobs make their decisions. We first disprove the conjecture of Hassin and Yovel (OR Letters, 2015) claiming that for unrelated machines, i.e., for the setting where every job can have a different processing time on every machine, the sequential price of anarchy for m = 2 machines is at most 3. We provide an answer for the setting with m = 2 and show that the sequential price of anarchy grows at least linearly with the number of players. Furthermore, we show that for a certain order of the jobs, the resulting makespan is at most linearly larger than the optimum makespan. Furthermore, we show that if an authority can change the order of the jobs	The bad Nash equilibrium in Example appears in several works @cite @cite @cite @cite to show that even for two machines the price of anarchy is unbounded, thus suggesting that the notion should be refined. Among these, the , which considers NE, is studied in @cite @cite @cite . The , which considers equilibria arising in extensive form games, is studied in @cite @cite @cite @cite . In @cite the authors investigate equilibria and the resulting , while @cite focuses on the equilibria produced by the . A further distinction is between (randomized) and (deterministic) equilibria: in the former, players choose a probability distribution over the strategies and regard their expected cost, in the latter they choose deterministically one strategy. In this work we focus on pure equilibria and in the remaining of this section we write @math to denote the bounds on the price of anarchy for mixed equilibria.	**Rouge scores:** - Rouge 2: 0.0592 (prec) 0.0699 (recall) 0.0641 (f-1) - Rouge L: 0.1706 (prec) 0.2014 (recall) 0.1847 (f-1) **Summary:** We consider the well-studied game-theoretic version of machine scheduling in which jobs correspond to self-interested users and machines correspond to resources. Here each user chooses a machine trying to minimize her own cost, and such selfish behavior typically results in some equilibrium which is not globally optimal: An equilibrium is an allocation where no user can reduce her own cost by moving to another machine, which in general need not minimize the makespan, i.e., the maximum load over the machines. We provide tight bounds on two well-studying notions in algorithmic game theory, namely, the price of anarchy and the strong price of anarchy on machine scheduling setting which lies in between the related and the unrelated machine case. Both notions study the social cost (makespan) of the worst equilibrium compared to the optimum, with the strong price and the weak price of anarchy restricting to a stronger form of equilibria. Our results extend a prior study comparing the price of Anarchy to the strong price **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.0329 (prec) 0.035 (recall) 0.0339 (f-1) - Rouge L: 0.183 (prec) 0.1944 (recall) 0.1886 (f-1) **Summary:** We study the outcome of natural learning algorithms in atomic congestion games. We consider the classical machine scheduling game, where n jobs need to be scheduled on m machines, with the goal of minimizing the makespan, i.e. the maximum load of any machine in the schedule. the private goal of each job is to minimize its own delay and the social goal is to minimize the maximum delay of any job, that is, to minimize the makespan of the machine running it. We show that in almost all such games, the well- known multiplicative-weights learning algorithm results in convergence to pure equilibria. furthermore, we show that if an authority can change the order of the jobs adaptively to the decisions made by the jobs so far (but can not influence the decisions of the jobs) then there exists an adaptive ordering in which the jobs end up in an optimum schedule. **Comment:** - Very weak signs of MDS, with the 2nd (green), 4th (purple) and 6th (orange) articles covered. - However, they are all presented as a single document. - The model also performs like an extractive model	**Rouge scores:** - Rouge 2: 0.0884 (prec) 0.1329 (recall) 0.1061 (f-1) - Rouge L: 0.1389 (prec) 0.2083 (recall) 0.1667 (f-1) **Summary:** We consider the well-studied game-theoretic version of machine scheduling in which jobs correspond to self-interested users and machines correspond to resources. Here each user chooses a machine trying to minimize her own cost, and such selfish behavior typically results in some equilibrium which is not globally optimal: An equilibrium is an allocation where no user can reduce her own cost by moving to another machine, which in general need not minimize the makespan, i.e., the maximum load over the machines. We provide tight bounds on two well-studies in algorithmic game theory, namely, the price of anarchy and the strong price of anarchy on machine scheduling setting which lies in between the related and the unrelated machine case. Both notions study the social cost (makespan) of the worst equilibrium compared to the optimum, with the strong Price of anarchy restricting to a stronger form of equilibria. Our results extend a prior study comparing the price of Anarchy to the strong Price Of Anarchy for two related machines (Epstein, Acta Informatica 2010), thus providing further insights on the relation between these concepts. Our exact bounds give a qualitative and quantitative comparison between the two models. The bounds also show that the setting is indeed easier than the two unrelated machines: In the latter **Comment:** - Copied the first tokens only - No indication of MDS	**Rouge scores:** - Rouge 2: 0.1442 (prec) 0.1049 (recall) 0.1215 (f-1) - Rouge L: 0.3143 (prec) 0.2292 (recall) 0.2651 (f-1) **Summary:** There is a large body of literature on the price of anarchy @cite @cite, which is closely related to our work. The price of anarchy is defined as the ratio of the worst strong Nash equilibria to the social optimum @cite. In particular, @cite and @cite studied the trade-off between the high and low price of anarchy, and showed that the high price of anarchy can be bounded as a function of the number of machines and the size of the coalition. In the case of the network creation game @cite the trade-offs between the low and high prices of anarchy are studied. **Comment:** - Some signs of MDS, with the information in the 1st article (yellow) used as context, and the 3rd (blue) and 4th (purple) articles are included in the summary.	**Rouge scores:** - Rouge 2: 0.1259 (prec) 0.1259 (recall) 0.1259 (f-1) - Rouge L: 0.2153 (prec) 0.2153 (recall) 0.2153 (f-1) **Summary:** In @cite, Paes Leme, Syrgkanis, and Tardos studied the price of anarchy for atomic congestion games. They showed that in almost all such games, the well-known multiplicative-weights learning algorithm results in convergence to pure equilibria, and showed that if an authority can change the order of the jobs adaptively to the decisions made by the jobs so far (but cannot influence the decisions of the jobs), then there exists an adaptive ordering in which the jobs end up in an optimum schedule. They further extend their results to show that players can use algorithms with different (sufficiently small) learn-rate, i.e., they can trade off convergence speed and long term average regret differently. In @cite @cite the authors studied the tradeoff between the quality of any Nash equilibrium to the quality of a global optimum, and the sequential price of anarchy. **Comment:** - Some signs of MDS, with the 2nd (green) ,4th (purple) and 5th (pink) articles summarized. - This is some mixing of the ideas in the articles, possibly because they used the same special terms (e,.g. atomic congestion, price of anarchy)	**Rouge scores:** - Rouge 2: 0.061 (prec) 0.0699 (recall) 0.0651 (f-1) - Rouge L: 0.1758 (prec) 0.2014 (recall) 0.1877 (f-1) **Summary:** In @cite, Paes Leme, Syrgkanis, and Tardos studied games where players choose their actions sequentially, and showed that sequential decisions mitigate the worst case outcomes known for the classical price of anarchy. In @cite, the authors consider a job scheduling game with two uniformly related parallel machines (or links). The goal of each job is to minimize its own delay and the social goal is to minimize the maximum delay of any job, that is, to minimize the number of jobs that can be assigned to this machine. The authors show that in almost all mixed Nash equilibria that are weakly stable, the Converse is true with probability 1 when congestion costs are selected at random independently on each edge (from any monotonically parametrized distribu-rontal). The authors also extend their results to the case of two players, and show that for a wide range of speed ratios these two measures are very different whereas for other speed ratios they give the exact same bound. **Comment:** - Some signs of MDS, with the 2nd (green) ,3rd (green), 5th (pink) and 6th (orange) articles summarized. - This is some mixing of the ideas in the articles, possibly because they used the same special terms (e,.g. atomic congestion, price of anarchy)