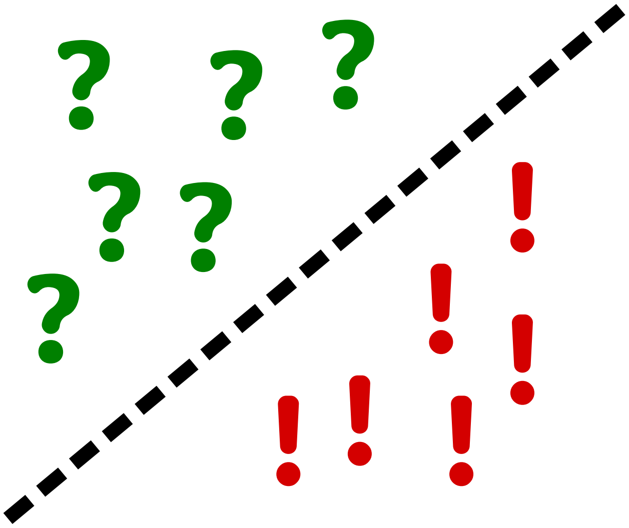
Research guidelines in Deep Learning

Jan van Gemert
Delft University of Technology
http://jvgemert.github.io/



Abstract

The goal of this document is to give structure to doing deep learning research. It might be useful for students, advisors, researchers, or as part of a (research) course. The beauty of research is that it's creative, inspiring, and different every time. Yet, over the years, I've found myself repeating different variants of the same feedback multiple times. This document contains a set of 150 loose guidelines, structured as uniquely named-item checklists which aims to structure common research settings and I find them useful as alignment, as a time saver during meetings, as a feedback tool, or rubric. These guidelines revolve around embracing the chaotic, creative, exploratory search, where it is completely normal to revisit and re-evaluate everything along the way. This chaotic research process goes through highs and lows which can be exciting and inspiring, and at the same time can be scary and demotivating. This is inherent to creative processes and also in doing research; and the key is to stubbornly keep the faith! Something will come and these guidelines help find it. They focus on understanding-based empirical deep learning research, without necessarily having access to huge compute/data. I believe that critically rethinking the foundations of deep learning models will foster exciting creative avenues and revolutionary different research directions. Better understanding the fundaments will lead to safer, transparent, more aligned, less data/compute dependent, and more robust AI systems.

Contents

1 Why these guidelines?

The goal of this document is to give structure to doing deep learning research. It might be useful for students, advisors, researchers, or as part of a (research) course. The beauty of research is that it's creative, inspiring, and different every time. Different people do research in different ways, and researchers/advisors may not agree with all my guidelines, which, if properly motivated, is perfectly fine. This document is thus not intended as a fixed rigid protocol [3]. Instead, it is a set of 150 loose guidelines, structured as named-item checklists that I have found repeatedly useful over the many years that I have been doing research that I decided to organize and document them. I've found it useful as alignment, as a time saver during meetings, as a feedback tool, and as a reminder to myself; I can even see it useful as a point of reference in doing paper reviews and rebuttals.

These guidelines can be characterized as *understanding-based empirical deep learning research*. It's *emprical* instead of theoretical, and thus focuses less on mathematical proofs and more on data-driven empirical results. It's *understanding-based* because I believe the goal of science is understanding, even (especially?) in the fast-moving field of Deep Learning. I believe that critically rethinking the foundations of deep learning models will foster exciting creative avenues and revolutionary different research directions. Better understanding the fundaments will lead to safer, transparent, more aligned, less data/compute dependent, and more robust AI systems.

Regarding scope, I certainly do not aim to claim that deep learning is the only relevant machine learning setting, it's just the paradigm in my own research field. This field is huge, and these guidelines might not apply to all deep learning research projects; while at the same time several guidelines align well with a broader research domain than deep learning. If you found these guidelines useful in your scientific approach or otherwise, then please feel very free to say why, and cite these guidelines in your scientific paper, thesis, or project.

1.1 Other research guidelines

I do not wish to start with the philosophy of science, where one can argue if Deep Learning is a paradigm shift as identified by Kuhn's Structure of Scientific Revolutions [5]. I agree with Feyerabend's Anarchistic Theory of Knowledge [3] and I do not subscribe to a strict protocol to do science. Yet, I do see that there are common repeating patterns that can serve as guidelines. In the field of computer science there are guidelines [1], and also in machine learning [2], which are useful and complementary. In this document I go deeper and identify over 150 uniquely named and referable guidelines spread out over 7 sections.

My guidelines focus on understanding-based empirical deep learning research. The advantages of such an approach are excellently demonstrated in the "Scaling down Deep Learning" paper [4]. Several of my guidelines emphasize critical thinking, where a special mention goes to the "Troubling Trends" paper [6], which is an excellent manifesto of things that can be improved, and all students in my Deep Learning MSc course have read it. On a similar note, albeit more targeted at academic researchers, there are the "Survival Strategies for Depressed AI Academics" [8] which argues for keep doing Deep Learning research, even as a poor academic, without having access to huge data and compute resources. On this note, I often remind students that it wasn't the industry who kept the field of Deep Learning alive before 2012: Deep Learning was studied as a niche, by poor academics, at a university. There are many important unsolved research topics in Deep Learning, including fairness, alignment, robustness, transparency, explainability and the dependency on large data/compute. This brings me to the following quote by Geoffrey Hinton: "*The future depends on some graduate student who is deeply suspicious of everything I have said.*". We need to keep doing critical, fundamental Deep Learning research in academia; question everything, go against the trend, the next revolution is waiting for you!

1.2 Approach to Deep Learning research

Research is risky These guidelines revolve around embracing the chaotic, creative, exploratory search in finding the research question, the work related to it, a method or approach to address the research question, the empirical questions to answer in the experiments, the analysis, and conclusions. It is completely normal to revisit and re-evaluate everything along the way, as illustrated in Fig 1. Research is inherently risky, and the research process thus is to a large part about risk management. So, do the most risky part first. Don't invest time in doing the 'easy' things that you expect to work out because it might be the case that these things are no longer relevant later. Put the risk first; try to break the idea as soon as possible.

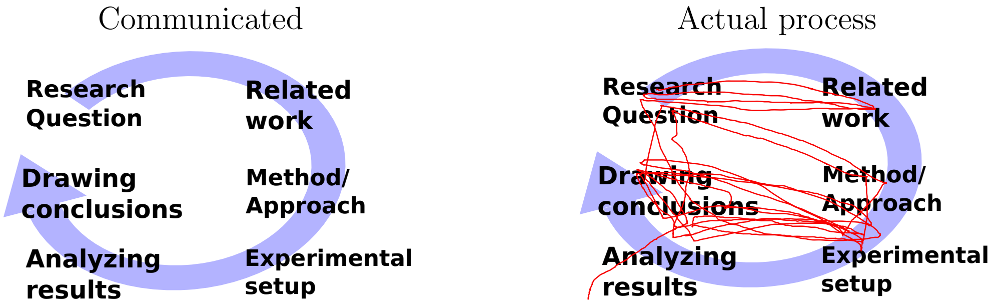


Fig 1: How machine/deep learning research is typically communicated *vs* an example of how the creative scientific process can go (red line). It is completely normal during the research process to rephrase the research question often; to revisit related work in a new light, to change the method or experimental setup, and re-analyze results and conclusions. Creatively searching for the question is inherent to science.

Find the path by taking a step Often, there is some direction, topic, or scope, but no clear well-defined end-goal when starting the research process. Starting this process fully aware that it will be chaotic, noisy, and go through highs and lows can be exciting and inspiring, and at the same time can be scary and demotivating. This is inherent to creative processes and also in doing research; and the key is to stubbornly keep the faith! Something interesting will come out. Yet, before being able to find the route, there has to be some exploration. It's easy by 'armchair philosophy' to dismiss possible directions or refuse to setup a simple fully controlled version of the problem by reasoning away that it's trivial. To arrive at a well-defined end-goal, there has to be a first step, even if it might be in the 'wrong' direction.

Empirical research Experiments are empirical support for the hypothesis, claims, observations, and benefits. I typically identify 3 general experiment types, where each type can have multiple experimental settings. Type 1: Validate claims. Does the problem exist? Validate that baselines suffer from the identified problem. Does your proposed method address the problem? Etc. This is often best done in a fully-controlled (toy) experimental setting where having full control allows validating the problem and demonstrating the solution. Type 2: variations. Report interesting variations and explain why they are interesting. Do ablation studies on individual components. Show sensitivity to hyper-parameters. Etc. Type 3: Existing data sets Does the identified problem occur in existing less controlled datasets? Is the problem common? Can you show that the problem occurs in multiple datasets? Validate that baselines still suffer from the problem. Verify that the proposed method addresses the problem. You are free to choose any existing dataset; so choose datasets that match your problem.

1.3 How this document is structured

This document is a collection of 150 guidelines spread over several sections:

Section 2: The Storyline: the beating heart of the research process.
Section 3: Organization, process, and mentality.
Section 4: Research meetings.
Section 5: Writing.
Section 6: Giving a talk.
Section 7: Presenting a poster.
Section 8: Reviewing and rebuttals.

For relevant sections, I added a table with each guideline/question uniquely labeled; which can be used to give feedback, to point at during a meeting, or to use as a reminder. In the appendix I copied all tables on a separate page for convenient printing and adding your own.

2 The storyline

My love for doing scientific research is in its creative processes. Exploring uncharted territory, brainstorming, critical reflections, coming up with research questions and rephrasing them, creating precisely controlled experiments, how to keep everything logically consistent, puzzling over experimental results, choosing which results to keep and which results are not relevant, writing a structured narrative, creating high entropy figures, how to best present and concisely communicate the key insights of the work, etc. Creative processes, however, are acutely different from technical crafts, as eruditely put by music producer Rick Rubin [7]. In the field of machine/deep learning these crafts involve technical skills such as math, programming, software engineering, statistical methodologies, etc. These skills are not only essential to get a timely, rigorous, and trustworthy answer, they are also important to detail in a publication, so that they can be scrutinized, questioned, replicated, and built upon. Yet, doing research is not only applying skills to find an answer, it is often not even clear what the question is [9]. Thus, it is normal (essential?) that the research direction changes/evolves many times, see Fig 1; this is the creative process at work: doing the work to find the question. This document is about giving structure to the scientific creative process in search of the question.

I've developed the storyline-technique incrementally by reflecting on my experiences as a researcher. I find it one of the most valuable tools to structure the creative process: to find, and work out, what question to concretely answer. The storyline is an as detailed as possible, concise, focused, relevant, logical, self-contained, and fully-motivated research narrative, which can be understood and critiqued without the use of unnecessary jargon, and has all abstractions opened up to their concrete core reasons. It forms the heart of the question-search, and allows a holistic view on the full project, so that all aspects of the research can be scrutinized, questioned, critiqued, sharpened, removed, added, or pivoted on. It cuts everything away to arrive at the lean motivational core where any claim made can be challenged, and if it cannot be motivated, the claim should be removed. In empirical research, each claim can be challenged by an experimental question, and experiments thus take an important role and are tightly interwoven with each made claim. Since the storyline is for finding the question, it is therefore completely normal, and even expected, that the storyline will fluidly change during the creative process. Concretely, the storyline gives structure to the question search: thoughts, meetings, the process, the hypotheses, what experimental questions to ask, the logical narrative, and helps the writing by postponing sentence structure till later. A visualization of the storyline is shown in Fig 3.

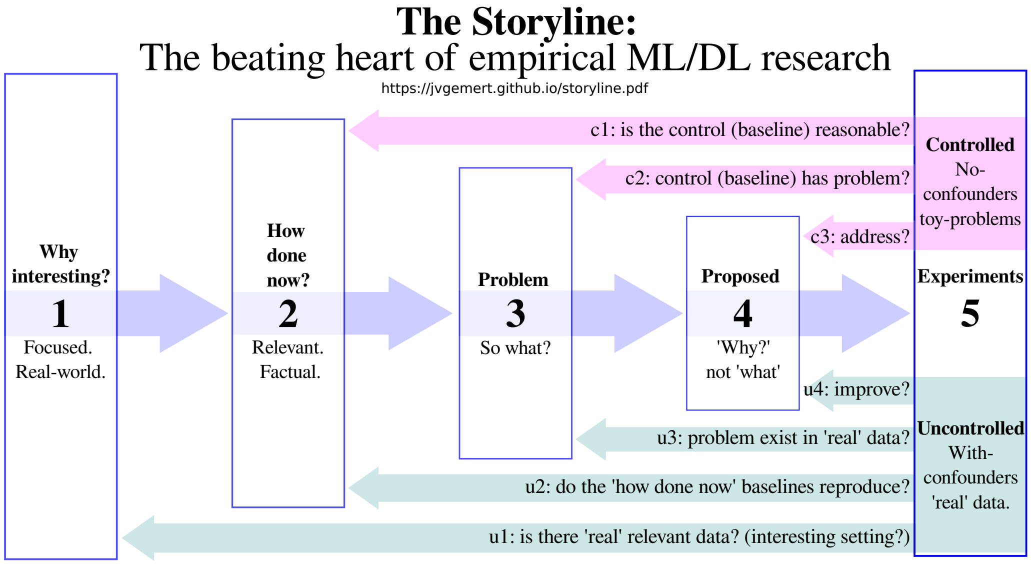

The Storyline:
The beating heart of empirical ML/DL research
https://jvgemert.github.io/storyline.pdf
Why interesting?
1
Focused.
Real-world.
How done now?
2
Relevant.
Factual.
Problem
3
So what?
Proposed
4
'Why?' not 'what'
Experiments
5
Controlled
No-confounders
toy-problems
Uncontrolled
With-confounders
'real' data.
c1: is the control (baseline) reasonable?
c2: control (baseline) has problem?
c3: address?
u4: improve?
u3: problem exist in 'real' data?
u2: do the 'how done now' baselines reproduce?
u1: is there 'real' relevant data? (interesting setting?)

Fig 2: The storyline structures empirical machine/deep learning research. It concisely ties the research in a logically consistent focused narrative. It has the following elements. (1) Why the (real-world) setting is interesting; (2) factual description of relevant existing approaches (baselines); (3) what is the problem of the baselines and what are the consequences; (4) why the proposed approach addresses the problem and (5) experimental questions. The experiments take center stage in empirical research and they link back to all elements. One group of experiments are in a fully controlled ’toy problem’ setting (top, in pink) with only a single confounder: precisely evaluate the settings with/without the identified problem. The controlled setting should empirically validate that (c1) the baseline (control) setting is set up reasonably and fairly, that this reasonable baseline (c2) does indeed suffer from the problem; and (c3) that the proposed approaches addresses it. Where the controlled experiments assume that the problem exists, the uncontrolled ’real’ setting has unknown confounders (bottom, in green) and validate that the problem actually occurs in ’reality’. It shows relevance (u1) by demonstrating multiple relevant data sets; that (u2) published baselines reproduce published numbers; that the identified problem (u3) exists among uncontrolled confounders, and thus (u4) that the proposed approach improves over the published baseline when the problem exists.

2.1 Structural elements of the storyline.

This is the typical structure; each element might take 1-5 bullet points:

(1). **Why interesting?** What is the tightly scoped motivation. Why should someone else care (e.g., the users of this particular research outcome).

(2). **How done now?** Relevant approach(es) to the motivation in (1).

(3). **What is missing, and So What?** What's the problem with the approaches in (2), and what consequences does this have on (1).

(4). **Proposed approach.** What do you do, and why does it address the problem in (3)?

(5). **Experimental questions.** Controlled: validate that problem (3) occurs in current models (2) and that (4) addresses it, and its consequences (3). Uncontrolled: does the problem exists in confounding 'real' settings (1)?

The storyline is minimal, and stand alone: you cannot use a 'jargon' term/concept before it has been introduced by motivating it; i.e.: what is it and why is it needed. Each claim made can be challenged and each claim should thus be motivated. Keep Hitchens's razor in mind: "*What can be asserted without evidence can also be dismissed without evidence.* Terms/concepts logically build/connect to earlier terms/concepts (i.e.: its a story). Have short "1-liners" per bullet point; correct grammar is optional; the entire storyline should be visible at once, so it should fit on a single page/slide (10-20 bullet points). When finalizing the storyline, it's useful to work backwards from the experiments; because the terms used there need to be introduced before. Remove unused terms and jargon.

Intermediate results produced during the research process often lead to a better understanding of the problem, and thus change the storyline, see Fig 1. After a while, there's a collection of loosely connected results; and then its useful to add focus and re-evaluate which results fit a consistent narrative, and how this (new?) narrative changes the storyline and the follow-up experiments. Some experimental results will –in retrospect!– turn out to be distractors from the main narrative. Or, they were initial –now redundant– stepping stones towards a better understanding of the problem. These results play no role in the final paper: "*kill your darlings*"; which is common, but painful because often these results took quite some effort; and removing them seems to invalidate that effort, which, unfortunately, is inherent to the uncertainty of doing creative work.

2.1.1 Storyline element (1): Why interesting?

What is the motivation to do this research? Why should someone else care about this research's outcome? Things that I often see are variants of "Much recent research is done on XXX". These are not good reasons in themselves because it describes a reaction and not the reason for this reaction. Open the abstraction: i.e.: what are the underlying reasons that topic XXX has received so much attention? What is interesting, beneficial, useful, important, about it? Keep digging deeper and repeatedly ask "Why is that interesting?" until you cut away all abstractions and arrived at the concrete core. Put another way: why should someone invest time/effort in reading your work if it's not clear what is interesting about it for them?

Keep the scope tight, and focus on your research outcomes. Your problem (3) and research outcomes (4) should directly be applicable to the motivation in (1). For example, if the paper is about automatic reasoning in long videos, do not motivate it with 'robots', or 'machine learning', or 'general visual recognition', or even 'action recognition in short clips'. Instead, try to motivate it directly and tightly focused with why automatic reasoning in videos is interesting; and what is specially important about 'long videos', which for example, could include sport game analysis, or shoplifting, and why doing automation is useful/interesting there. The scope and applications ideally come back in the experimental section. For example, a motivational scope claim on 'robotics' can (rightfully!) be asked for an experiment on an actual robot (evidence). Assessors can penalize unsubstantiated over-claiming, i.e.: tightly focus scope with problem/outcomes in (3,4).

As an example, consider a new method for visualizing biases in deep/machine learning models. If the claim is that the new method can find new types of biases then this claim can be challenged, and thus should be experimentally motivated by showing such new biases. As another example, consider a new optimization method that makes auto-encoders faster to train. Then if the claim is that auto-encoders are often used in applications such as denoising, feature-learning, super-resolution, etc., then this can be challenged, it thus experimentally shown that existing auto-encoder methods for such applications also actually become faster to train. Also for why the research is interesting, it holds that with each claim comes a burden of proof.

2.1.2 Storyline element (2): How done now?

Here, give current, relevant, approach(es) to the tightly-scoped motivation in (1). Note the word *relevant* because the storyline is not meant to be exhaustive; instead, it's a focused, minimal and consistent narrative. Related approaches that are too different from the proposed approach, may belong in the 'related work' section of the research paper but not in the storyline, i.e.: they do not link to the problem (3) nor approach (4) and are thus not relevant for your story. For example, with a scope on 'long videos', the work on 'short videos' would go in related work, but is not relevant for the storyline. Leave out approaches that do not link to (3,4).

The described approaches here should be objective, without a value judgment. The authors of the approach(es) you mention should agree with how their approach is described; but the description does not have to be the main contributions of their work. You are free to creatively choose, re-interpret, and emphasize anything that is described in their papers, as long as it is factually correct. Be careful to not make use of jargon: each term used here should be motivated/introduced first. Moreover, terms can be introduced/motivated here so that they can can be used in pointing out problems in (3).

2.1.3 Storyline element (3): What is missing, and So What?

What's the problem with the approaches in (2), and what consequences does this have. First describe the problem, or what is missing. Then, make the consequence of the problem precise. The consequences make it (experimentally) possible to validate that the problem occurs, that the baselines do indeed suffer from the problem, and that the proposed approach addresses the problem. For example, in a 'automatic long video recognition' setting, the way how it is done now (2) could be that models only sample one frame per minute (not true, but this is just an example). Then, a problem (3) could be that this low sampling rate might miss relevant information. And the consequences then are that current models are sensitive to the accidental sampling offset, and that they have low accuracy when higher-frequency information is essential. Thus, model rankings might not be correct, which would lead to selecting the wrong model in practice, which can be validated experimentally.

2.1.4 Storyline element (4): Proposed approach.

Why does your approach address the problem in (3)? Focus on the 'why' not on the 'what'. Avoid technical explanations as much as possible; the storyline is not about what the approach does, that does go in the 'method section' of the paper. Instead, the storyline is all about motivation, and building a logically consistent "house of whys". The proposed approach should be understandable for non-experts. Avoid jargon, if possible, because each specialized term used should first be motivated/introduced, either here, or in the preceding elements.

2.1.5 Storyline element (5): Experimental questions.

How do you evaluate experimentally that (4) solves the problem and its consequences in (3)? The focus here is on empirical machine/deep learning research, and experiments take center stage. See Fig 3 for how experiments build on each element. Not all research may require a storyline emphasizing the experiments as much; yet, the storyline itself is still valuable; feel free to flexibly adapt accordingly.

Typically the first line of experiments are for careful control and verification: validate if the problem with current approaches in (2) exists, how severe the consequences in (3) are, and how well the proposed approach in (4) deals with the problem and consequences. I advise self-made, fully controlled, synthetic, ('toy problem') setting, where the full control allows generating small, crisp, and precise variants, with known outcomes. This is important as to avoid unknown confounding variables which might, unknowingly, influence the results. One variant is a normal setting, without the problem i.e.: a control. This control setting demonstrates that the existing approaches are represented fairly (i.e.: they do OK), and to set a baseline performance. A second variant is identical to the first, where the only point it varies is that it has the identified problem in (3); which then demonstrates that existing approaches in (2) suffer from the consequences in (3) and that the approach in (4) is suitable. Note that "good accuracy" is not needed. I.e.: there is not need for large training sets, as that might actually be detrimental: if the baseline already scores 95%, then the proposed approach can only make marginal relative improvements.

A second line of experiments investigate impact in the world. Put another way: how severe is the identified problem, and its consequences in practice? The goal is to gather evidence that the problem occurs in 'reality', and it is good to have a couple of datasets as evidence. This involves evaluating on less controlled datasets, with unknown confounders, that have the problem. This can include real data that you collected yourself, or, existing open datasets. In academic research, such 'real world' datasets are typically still quite artificial. Even so, compared to the first line of experiments these datasets do have unknown confounders, and thus can be used as evidence that the problem exists. The datasets should align with the problem. e.g., if the problem involves rotated images, then typical Computer Vision datasets such as CIFAR, or ImageNet are out, because they do not contain rotations and are thus not relevant. Instead, use datasets where rotations occur naturally; for example cell images taken under a microscope. Here, it is important to validate that existing published results on the datasets actually reproduce on those datasets (do not assume they will!). The comparison to published results is important so that the reader can validate that baselines are fairly represented. These methods are the published baseline scores to compare to. If the problem occurs in the dataset, and the proposed approach handles the problem well, then it can be expected that the published baselines are outperformed by the proposed approach, on those datasets.

2.2 Storyline examples

To make the storyline structure as described above less abstract, I give some example storylines for a few papers where I was involved, below. Note, they are not meant to be perfect; but the real world rarely is perfect and "*perfection is the enemy of good enough*". In reality there often are time constraints (work contracts, graduation dates, etc.). Science is never 'done', and a scientific paper can still be interesting (ie: publishable) when it is 'good enough'. In addition, some research project have different empirical questions that do not fully align with the controlled/uncontrolled experimental groups in (5); which is fine. The power of the storyline is the harsh logical narrative, that forces the researcher to back up a claim with evidence; or adapt the claim accordingly.

Storyline for Lengyel, et al. *Color Equivariant Convolutional Networks*, NeurIPS, 2023. `https://arxiv.org/abs/2310.19368`

1. **Why interesting?**
   (a) Automatic image recognition is important for many applications.
   (b) Image recognition is trained on data with inherently imbalanced (accidental) viewpoint/appearance occurrences.
   (c) Imbalance leads to biases towards the frequent; and reduced accuracy for the less frequent occurrences.
2. **How done now?**
   (a) Imbalance is tackled by Equivariant CNNs: sharing learnable weights over spatial transformations (rotations, scale, ..).
3. **What is missing, and So What?**
   (a) Current work is on spatial transformations, no appearance.
   (b) So: reduced accuracy due to imbalance in appearance.
4. **Proposed approach.**
   (a) Sharing weights over different appearances: color hues (hue = H in HSV color space).
   (b) Propose: color equivariance by rotations in hue space.
5. **Experimental questions.**
   (a) Gains for class/color imbalance? Toy set: *Long-tailed ColorMNIST* has 30 classes (10 digits x 3 colors), controlled imbalance.
   (b) Gains for color variations? Toy set: *Biased ColorMNIST* has 10 classes (digits), give each sample a random color; create a curve over color variation by varying the stddev of the random color.
   (c) Gains for hue domain shifts at test time for existing datasets?

Storyline for Kayhan et al. *On Translation Invariance in CNNs: Convolutional Layers Can Exploit Absolute Spatial Location*, CVPR, 2020. `https://arxiv.org/abs/2003.07064`

1. **Why interesting?**
   (a) Automatic visual classification, image matching, video recognition are important for many applications.
   (b) Sharing network weights over different locations (spatial shift equivariance) improves data-efficiency:
   (c) data-efficiency is important: collecting/labeling data is expensive.
2. **How done now?**
   (a) CNNs use convolution (sliding window) to share weights over different locations.
   (b) When the sliding window reaches the image boundary it stops, or, half the window-size zeros are padded outside the image boundary.
3. **What is missing, and So What?**
   (a) Insight: input pixels near the image boundary are not seen by the full sliding window (it: it slides only until the image ends).
   (b) Ie: CNNs can asymmetrically ignore image content close to one side of the image boundary and not the other side.
   (c) Surprisingly: CNNs can learn weights that depend on the location, by the distance to image boundary.
   (d) So: when shift-equivariance is broken; data efficiency suffers; needing more expensive labelings.
4. **Proposed approach.**
   (a) Remove the ability of the model to exploit absolution location:
   (b) Make the sliding window see all input pixels (ie, the left part of the window should also see all pixels of the right part of the image)
5. **Experimental questions.**
   (a) Can 1 CNN layer use location? Train CNN to separate 2 images; each with the identical patch, but at different spatial locations.
   (b) How far from the image boundary can existing (scratch/pre-trained/random) CNN models use location? Same experiment as in (a) but vary the distance of the patch to the image boundary
   (c) Will baseline/proposed use location when not needed? For location-independent task: train on one location; test on a different one.
   (d) Sensitivity to spatial shifts. Evaluate on test-time image shifts.
   (e) Data-efficiency? Learning curves: classification, matching, video.

Storyline for Huijser et al. *Active Decision Boundary Annotation with Deep Generative Models*, ICCV, 2017. `https://arxiv.org/abs/1703.06971`

1. **Why interesting?**
   (a) To train a ML model we need to label/annotate data: boring, expensive, time consuming, and error prone (annotation noise).
   (b) 'Active Learning' reduces the label effort by not labeling the full dataset: Use ML model trained on partial labels during the labeling: interactively suggest 'most informative' data samples to label by a human annotator; retrain; repeat.
2. **How done now?**
   (a) Various active learning strategies to select the samples to label
3. **What is missing, and So What?**
   (a) Labeling samples focuses on the data points; but the goal is to find the decision boundary between classes.
   (b) So: labeling samples not directly solving the goal.
   (c) So: labeling samples will take more label effort than if we could instead label the decision boundary directly.
4. **Proposed approach.**
   (a) Instead of labeling samples; lets label the decision boundary itself.
   (b) Use a generative model based on baseline active learning sample strategies to generate a 'line' of samples which crosses the decision boundary.
   (c) Let the user annotate on the 'line' where a class changes to a different class: this is where the decision boundary lies.
   (d) The ML model can then include decision boundary annotations.
5. **Experimental questions.**
   (a) How well can baseline active learning sample strategies be used as input for decision boundary annotation?
   (b) Quality of generative model close to the decision boundary?
   (c) Sensitivity to noisy decision boundary labeling?
   (d) How well does a human annotator do with decision boundary annotation?
   (e) How well does it generalize to more classes/datasets?

Storyline for De Boer et al. *Is there progress in activity progress prediction?*, ICCVw, 2023. `https://arxiv.org/abs/2308.05533`

1. **Why interesting?**
   (a) Action progress predictions useful for scheduling, planning.
2. **How done now?**
   (a) Current methods aim to learn visual information to predict action progress.
3. **What is missing, and So What?**
   (a) Published visual-learning methods never compare to simple baselines.
   (b) So: Unclear if visual-learning methods methods "work".
   (c) So: Unclear if visual-learning methods can be trusted, or should be used in reality.
4. **Proposed approach.**
   (a) Set 2 simple visual learning baselines: 'CNN', and a 'CNN+LSTM'.
   (b) Set 2 simple non-visual learning baselines: 'frame-counting' and 'random-noise as input'.
   (c) Set 2 non-learning baselines: 'random guessing'; and 'always predict 0.5'.
   (d) Create a synthetic dataset: 'visual progress bar' to evaluate if current visual-learning methods can do progress prediction.
5. **Experimental questions.**
   (a) Evaluate 3 existing datasets with: 3 published visual learning models; 2 simple visual-learning baseline models; 2 non-visual baselines (frame-counting, random-noise input) and 2 non-learning baselines (random guessing, always 0.5).
   (b) Evaluate taking segments instead of full videos, to try to avoid frame counting, because learning from a segment with it's progress score does not have the full-video context.
   (c) Evaluate progress prediction methods on visual 'progress bar'.

Storyline for Strafforello et al. *Are current long-term video understanding datasets long-term?*, ICCVw, 2023. `https://arxiv.org/abs/2308.11244`

1. **Why interesting?**
   (a) Long-term automatic video understanding: sports, surveillance.
2. **How done now?**
   (a) Quality of methods are evaluated on long-term video datasets.
3. **What is missing, and So What?**
   (a) Unclear if current long-term video datasets really evaluate on long-term information.
   (b) So: methods that do well on these datasets might not do well on actual long-term settings.
   (c) So: might lead to bad results in reality; wasted costs; disappointed users; failed projects.
4. **Proposed approach.**
   (a) Define: Long-term must consist of multiple short-term actions.
   (b) Evaluate if humans can recognize long-term actions in video datasets after seeing a short clip. If so, then the videos are not long-term.
5. **Experimental questions.**
   (a) For datasets that have both long-term and short-term annotations, there should be more than 1 short-term actions annotation used in a long-term action annotation. If not, then long-term can be recognized by a short-term; and it's therefore not long-term.
   (b) For several long-term datasets: create 2 sets for the same videos. A set short-segments and a set of long-segments; validate for a long-term task that accuracy(long-segment) > accuracy(short-segment); if this is not true, the videos are not 'long-term'.

3 Research Organization, Research Process, and Research Mentality

Doing research often assumes a certain mindset, process and organization. This holds for the individual researcher, but also for an advisor. Here, I make such assumptions explicit.

Organization	**Process**	**Mentality**
RO1 Full responsibility	RP1 One main Q	RM1 Be critical
RO2 No dependencies	RP2 Min. 3rd party	RM2 Find todos together
RO3 Meet advisor	RP3 Validate baselines	RM3 Consistency
RO4 Focus advisor	RP4 First break it	RM4 Question everything
RO5 Take critique	RP5 Depth first	RM5 Simple is strong
RO6 Constructive disagree	RP6 Exps answer Q	RM6 Embrace limitations
RO7 Analyze results	RP7 Proof of concept	RM7 Write early and often
RO8 Suggest solutions	RP8 Exps max 1 night	RM8 Not eureka
RO9 Give feedback	RP9 Change 1 var	RM9 Show the problem
RO10 Safety	RP10 Debug science	RM10 Motivate everything
	RP11 Figures	

3.1 Research Organization

RO1: Take full responsibility. This is your project. Not your adviser's. You are in charge of everything, including: planning, progress, direction, meeting topics, bureaucratic formalities, etc. You are not alone: your adviser is there to help you as best as possible, yet the final responsibility remains yours.

RO2: No dependencies. Avoid dependencies on third parties. Such parties intend well, yet reality is often different from intentions. Do not become the victim of this and make sure you have full control. E.g.: Promised data, labels, experts, constraints, or other agreements have to be there *before* you start.

RO3: Meet your adviser. Try to see your adviser at least once every 2 weeks; once per week is better.

RO4: Focus your adviser. Meeting time is limited. Avoid needless chronological updates (no need for “proof of work”). Discuss problems, choices, dilemmas and directions. It is your responsibility to choose to discuss what benefits *you* most.

RO5: Do not take criticism personal. All feedback is meant to improve and benefit you. Do not fight the feedback, even if you think it is wrong: Make a note, and think about why your advisor gave this feedback.

RO6: Constructive disagree. It’s OK to disagree with a suggestion of your adviser, but if you repeatedly do this then also try to propose something yourself.

RO7: Analyze results. When presenting (intermediate) results, give an interpretation and conclusions (ie: answer the “So what?” question).

RO8: Suggest solutions. When encountering research or organizational problems; suggest a solution yourself.

RO9: Give feedback to your advisor If you are unhappy about something (eg: how feedback is given, how meetings go, meeting time; etc.) then please let your advisor know this. Your advisor cannot read your mind, and is there to help you, give your advisor the opportunity to help you best.

RO10: Safe environment Meetings and the advisor-advisee relationship should be safe and based on mutual respect. If you do not feel safe, contact a (confidential) counselor at your organization, your advisor’s advisor, and speak out to friends/peers. It is your advisor that needs to change.

3.2 The research process

RP1: Only one main research question / problem statement. It may change over time but explicitly pursue a single topic: Write it down to make it precise; this gives focus and direction.

RP2: Minimal effort for 3rd party building blocks. If you build on top of existing work (e.g. an optimizer, object detector, pose estimator, etc.) start with the *least effort* approach to obtain this building block. It should not matter which building block you take, so start with the easiest available implementation. If you work modularity, you can always add another one later.

RP3: Validate published work. It is not obvious that a published work generalizes to your problem. There may be subtleties: Validate this.

RP4: Prioritize idea breaking. Start by investigating the greatest risk to your main research question. Do not invest heavily on the foundations, only to find out months later that the main idea did not work.

RP5: Depth-first instead of breadth-first. Do not explore sub-topics too deep. Identify the minimum requirement per sub-topics and get to this minimum as soon as possible. Try to get ASAP to a first full version to validate your idea. More baselines/variants/datasets can always be done later.

RP6: Experiments answer a single question. Write down before you do an experiment what your expected answer to the question is. Validate.

RP7: Show proof of concept. Start with a fully controlled (possibly toy) dataset of 'the simplest case possible' which should only vary in the relevant manner. Its goal is to validate that the problem occurs and/or that your model can solve it.

RP8: Experiments take max 1 night. If it takes 1 week, then 10 runs take 2.5 months. Minimize experimental time so you can answer more questions, especially in the beginning; leave larger experiments for the end.

RP9: Change only one variable. If more than one variable is changed, it is not possible to determine the cause of an effect.

RP10: Debug your scientific ideas and your code. Test ideas and test code every time you make a change. Start with the assumption you made a mistake somewhere, gather independent proof that it is correct.

RP11: Figures. Try to script all graphs/figures that you create. Yes: All. Your adviser may ask for a completely different version of a figure, and automating it prevents lots of manual re-doing. I prefer Matplotlib; it can output high-quality PDF figures and graphs that can directly be included in pdflatex.

.

3.3 Research mentality

RM1: The critical reviewer. Often switch roles to a savage reviewer (Mr Hyde) who is looking for any excuse to say: *'I do not believe X; Reject.'* Try to identify X yourself and think about which evidence argues for X.

RM2: Your supervisor does not have the answer. We are doing research. By definition, this research has not been done before. Thus, it is impossible for your supervisor to give you a list of ToDos: We'll find them together.

RM3: Be consistent. Assumptions you make in one part of your research should not suddenly change in another part.

RM4: Question everything. Take a step back, and think about what you are really doing. Does the story logically make sense? Try to see the things you take for granted: Is everything justified?

RM5: Simple is strong. Simple is more powerful than complex. Explain the core of your topic to a smart layperson (your mother?) without using math/jargon. If you cannot explain it, it is probably too complex.

RM6: Limitations. Identify the limitations of your method. No method will always be the best. Showing insight where it fails is strong. The goal of research is understanding.

RM7: Write early and often. Writing helps to make thoughts concrete and it is the interface to your work. Writing always takes longer than you think, even if you know that it takes longer than you think. Writing is iterative; don't try to write the perfect text: write a sloppy draft, and iterate.

RM8: Not “Eureka” But “That’s funny”. (by Asimov) is the most exciting phrase in research. It often becomes most interesting when expectations break.

RM9: Show that the problem exists. Going directly after improvements is risky: if it doesn’t work, all is gone. Before proposing a solution/improvement: first demonstrate/validate which problem is solved by it. Demonstrating the problem is valuable in its own right. To demonstrate the problem you are free to choose the setting; a self-constructed fully controlled (toy) dataset is often ideal. Make sure to validate that a proposed solution does well on this fully controlled setting.

RM10: Motivate everything. Novelty is easy: each component can blindly be replaced by another. Always question: *why?*. Each choice needs to be motivated with a reason. Is it commonly done? Then give citations. Is it interesting? Then motivate why. Is it ’obvious’ or speculative? Then empirically validate it as an hypothesis. (Hitchens’s razor: “*What can be asserted without evidence can also be dismissed without evidence*”)

4 Research meeting questions

Questions during a research meeting often are specific, special case instantiations of more general questions. It is not always clear what goal the specific questions have, which might confuse, frustrate, or even demotivate. Here, I aim to give the generalized questions that lie behind specific question types, with the goal of making it clearer why, and what the goal of a question is. Additional benefits might include that these questions can be asked even without a meeting to get unstuck, and might help in preparing and focusing the meeting. Questions are grouped per stage; where each stage is typically revisited often, see Fig 1.

Research meeting questions		
Research question	**Related work**	**Method/approach**
RQ1 Why interesting?	RW1 Who will use it?	MA1 Why this method?
RQ2 What storyline?	RW2 How different?	MA2 Explain each step?
RQ3 Formalize/simplify	RW3 Builds on what?	MA3 Formalize/simplify
RQ4 What problem?	RW4 Baselines?	MA4 Alternatives?
		MA5 Align with RQ?
Experimental setup	**Analyzing**	**Conclusions**
ES1 What Qs?	AR1 Validate?	DC1 Exhaustive?
ES2 How answer Q?	AR2 Baseline?	DC2 Expectations?
ES3 Baselines?	AR3 Understand all?	DC3 Align with RQ?
ES4 Expected outcome?	AR4 When fail?	DC4 Simplify?
ES5 Simplify?	AR5 Link to Q?	DC5 New hypotheses?

4.1 Questions about the Research Question (RQ)

RQ1: Why is the RQ interesting? Why do you care? Who else cares? Why should others care? Should we change the RQ? It can be useful to rephrase the RQ to better align with the problem: it can become more general, or more specific. Changing the RQ is normal during the research process because finding the suitable RQ often takes a large part of the process.

RQ2: What is the 10-15 bullet point main storyline? Does the story still make sense? Which point in the storyline are we now discussing? Is the point still valid? Should we change the storyline? The storyline is the key motivational driver and changing the storyline is common during the process because any new result can invalidated other parts of it.

RQ3: What is the RQ precisely? (Formalize/Simplify) Why can't we ask a simpler question? How to formalize it? Which part is the most uncertain? Should we remove that part or should we focus on that first? Formalizing the RQ makes it more precise, possibly revealing hidden assumptions. Simplifying the RQ can make it more general and/or easier to explain.

RQ4: Does the problem exist? How often? When (not)? What is the simplest example of the problem? How to convincingly demonstrate that the problem exists? Creating a simple, fully controlled, setting of the problem reveals hidden assumptions. It simplifies and gives focus. Coming up with a good problem example is difficult, and often takes several iterations.

4.2 Questions about Related Work

RW1: For which work is the RQ interesting? Why can't this approach be used for X? Why can't Y also make use of this approach? Should we change the research question to better facilitate X or Y? Don't wait for others see links to your work; actively link them yourself.

RW2: How is existing work different/similar? Why can't method X already answer the research question? What assumptions are different from method Y? What other related papers are there? Should we change the research question to clearly discriminate our setting from X or Y? Or should align better with X and Y? Motivate where the work fits, how it's different and how it's similar.

RW3: Building on what existing work? What is the motivation to build on component X? Why don't we use Y? Can't we use a simpler or more common building block? If the building block is not the focus of the work, then it should be as standard as possible.

RW4: How does existing work (baselines/competitors) work? What do baselines/competitors do? Why? How do they solve part X of the research question? How are we different? Why are we different? There should be good motivation to do something different.

4.3 Questions about the Method/Approach

MA1: Can you motivate why this method? Each step? Why not another method? Is each step essential and why? (is there evidence? ablation?).

MA2: What's going on? (Visualize output for each step) What is it really doing? Please explain this step in detail? Please show only the outcome of this part, and keep the rest constant. The goal is to validate that the method is behaving as expected.

MA3: What precisely? (Formalize/Simplify) How to make it simpler? How to formalize (math?) to describe what exactly is going on? Simpler is stronger.

MA4: Does method make sense? (Alternatives?) What are other options? Can we motivate why we do not use them? These choices might not be 'obvious' and may require an empirical experimental ablation.

MA5: How well is the alignment of the method with the RQ? Match the method with RQ: Which part of the RQ aligns with this step? Match the RQ with the method: Where does this part of the RQ come back in the method? Does it do what we think it does, and if not, should we then change the RQ to match this? It can be a game changer if results interestingly deviate from expectations.

4.4 Questions about the Experimental Setup

ES1: What empirical questions belong to the RQ? What questions do we wish to answer? Why? What would answering this question give? Do the questions align with the main RQ? Should we change the RQ to align better?

ES2: What exact question is answered by this experiment? Why does this experiment answer this question? What other experiments are possible? Which experiment to do? Each experiment should test an hypothesis or answer a question.

ES3: What are relevant baselines? Are there very simple (non learning? or simple averaging?) baselines to compare to? To which (existing) methods do we compare? Why? Why not more/less? Should we make the RQ tighter?

ES4: What exact outcome is expected? What outcome is wanted? Before running the experiment, answer what outcome you would like? What are the exact numbers that you expect as an outcome? Does doing the analysis on those numbers give the wanted outcome? If not, should we change the RQ?

ES5: What is the minimal setting? (Simplify) How to use a smaller problem? How to use a less complex setting? Why can't part X be removed from this setting? Simpler is stronger.

4.5 Questions about Analyzing Results

AR1: How to validate results? How to validate there are no bugs? How to validate if your method does what you claim it does (semantic debugging)? Do we have stddevs? Do results consistently align with our previous results? Can we do a small test to validate? Can we do an independent experiment to validate? Bugs are normal, and neural networks are notoriously difficult to debug; they might seem to work, but there might still be a mistake. Start with the assumption that there is a mistake somewhere, and then write code to find it.

AR2: How to verify correctness of baseline? Do not assume baselines directly work. How well do we match the reported results in their paper? Are these results expected? How to best optimize (hyperparameter tuning) the baseline so that the baseline is fairly evaluated?

AR3: Do we understand all results? When looking at the result table, can we explain each pattern in the table? Can we also look at some individual data samples?

AR4: When does it fail? Can we systematically predict when it fails? Can we look at some individual mistakes? Do these failures make sense?

AR5: Do results answer the question of the experiment? Each experiment has a question to answer. What was the question? Do these results answer that question?

4.6 Questions about Drawing Conclusions

DC1: What are all conclusions we can draw? Can we list all patterns that we see? Can we explain all patterns? Are there patterns that we have missed?

DC2: How well do results align with previous expectations? What were the previous expectation? How to explain deviations? Are all results internally consistent?

DC3: How well do results align with RQ? What results did we want? How much are these results what we wanted? Should we change the RQ? Should we redo a different variant?

DC4: Is there a simpler experiment with same conclusions? Which properties are not essential? Which properties should be more emphasized? Results/patterns that are not relevant distract from the main message.

DC5: To which new hypotheses lead these results? How well do these align with the RQ? Should we change the RQ to include these new hypotheses or write them as future work?

5 Writing guidelines

Science is also about communicating ideas. This makes writing an essential part of research. The five most common writing comments that I give back are:

- Terms/concepts are not motivated. A new term/concept first has to be explained and introduced by motivating why the term is important.
- A paragraph has no single topic and no conclusion (ie: no answer to the ”So What?” question)
- Sentences do not follow each other. (Sentence 2 should continue on topic where sentence 1 ends.)
- Figure captions do not explain how to read the figure.
- Figures/Tables have no conclusion in the caption (ie: explicitly write what should the reader see, and answer ”So What?”)

My writing guidelines are summarized in the table below, and followed by the guidelines themselves.

Writing guidelines		
General	**Structure**	**Form**
WG1 Unburden	WS1 Self-contained	WF1 Single topic
WG2 Audience	WS2 Consistent	WF2 Windows/orphans
WG3 Less is more	WS3 As discussed	WF3 Very
WG4 No guessing	WS4 Paragraphs	WF4 In order to
WG5 Read out loud	WS5 Ref words	WF5 Sort cites
WG6 More space	WS6 Ref paragraphs	WF6 Brackets
WG7 Write as code	WS7 Latter/Former	WF7 Synonyms
		WF8 Performance
Tables/Figs	**Introduction**	**Related work**
WT1 Captions	WI1 Motivation	WR1 Subject
WT2 Figs are complete	WI2 First sentence	WR2 Paragraph
WT3 Tables	WI3 Few research	WR3 Layout
	WI4 Fig 1	WR4 No history lessons
	WI5 3x contribute	
Method/approach	**Experiments**	**Discussion**
WM1 Argumentation	WE1 Question	WD1 Summary
WM2 No datasets	WE2 Group	WD2 Limitations
WM3 Number eqs	WE3 Analyze	WD3 Conclusions
WM4 Eqs are text	WE4 3 types	
WM5 Explain symbols	WE5 Scale 0-1	
WM6 Explain eq	WE6 proved	
WM7 Remove eq	WE7 One more	
WM8 Define		

5.1 General writing guidelines

WG1: Unburden the reader. If a reader misinterprets the text: its the writer's responsibility. Prof. Freeman: *The most dangerous mistake you can make when writing your paper is assuming that the reviewer will understand the point of your paper.* Avoid that the reader has to do work.

WG2: Audience. Who are you writing for? What is their background and what are they looking for? Help your audience find it. It often helps to keep a specific person in mind as your writing target.

WG3: Less is more. Every word should have a reason to exist. ie: Remove all unnecessary words. To quote Blaise Pascal: *"I would have written a shorter letter, but I did not have the time."*. Writing concisely takes time and effort; it enhances readability.

WG4: No guessing: make it explicit Never expect a reader to do inference. As a writer you need to spell out the thought process for the reader. If the reader has to guess, the guess will often be not what you had in mind. Always explicitly write what the reader is supposed to see/conclude.

WG5: Read out loud. After some time, you will no longer be able to read your own text. Instead, you will read what you meant; not what you wrote. Tip: read your own writing out loud.

WG6: Important topics take more space. The more important or relevant something is to your paper, the more space it takes. If it is not so important don't write too long about it.

WG7: Writing is like coding. Like good code, a paragraph is modular and self-contained. Do text refactoring just as you would do code refactoring. Good code is not written in one go, neither is text. Like code, you start with an initial structure, and restructure several times.

5.2 Structure

WS1: Self-contained. The reader has not memorized the full text. Remind the reader of definitions or symbols when defined 'a long time ago'.

WS2: Consistent. Use a defined symbol consistently and uniquely.

WS3: As discussed before, as pointed out earlier, as motivated in section XXX, as will be described in XXX Avoid using this, it has no function. The standard sectioning structure of a research paper dictates where information should be found (main motivation in Intro/Related work; the technical in the Method; empirical evidence in Experiments, etc.). Assume your paper will not be read linearly.

WS4: Relation between paragraphs. Paragraph topics follow a logical order. It is helpful to start with a skeleton of topics and keywords. In addition, creating an *inverse outline* helps to validate the story. An inverse outline is created by starting with a text and write down the first and possibly last sentence of each paragraph: a logical structure should emerge.

WS5: Reference words. Reference words such as 'this', 'it', 'that', 'there' are often confusing: they require inference/work by the reader which should be avoided. Avoid referring: explicitly repeat what 'it' refers to. (Here, for example, 'it' should have been replaced with 'the reference word').

WS6: Avoid reference words across paragraphs. For example, do not start a paragraph with 'however'. It is unclear to what you refer to.

WS7: Avoid "latter/former", and "respectability". These reference words require mental ordering and memorization by the reader; avoid making the reader do this work. Instead, rewrite it without the reference words.

5.3 Form

WF1: A single paragraph, has only one single topic. A paragraph has an intro sentence to define the topic. Each sentence logically follows the previous sentence. It has a concluding sentence that concludes the topic: it answers the "So What?" question.

WF2: Widows and Orphans. Avoid paragraph endings with 1 word on a new line. Avoid paragraph endings with 1 line on a new page.

WF3: Very. Do not use "very". It is supposed to emphasize, yet, it does the opposite. See also: `https://www.proofreadingservices.com/pages/very`.

WF4: In order to. Can almost always be replaced with 'To'.

WF5: Sort citations. If using numerical citation, make sure not to cite it as [7,2,5], but sort them like [2,5,7] to reduce reader effort.

WF6: (brackets). Avoid brackets. If it's not important remove it; yet, if it is important, then it should not be in brackets.

WF7: Synonyms. In non-scientific writing it's sometimes advised to make the writing less repetitive by not repeating the exact same terms and use synonyms. In scientific writing, instead, using a different term for the same thing will confuse the reader. Choose a single term, and use it consistently. Clarity trumps eloquence.

WF8: Avoid "performance". It's ambiguous; it might mean: speed, accuracy, memory-use, latency, etc. Choose the precise version you mean.

5.4 Tables/Figs

WT1: Captions in figure/table. Figures and tables should be self-contained. A reader often starts an article in 'comic book' mode: first look at all the pictures. The caption should explain everything to understand in the figure/table. Always end with a conclusion to answer the "So what?" question: make explicit what do you want the reader to see here.

WT2: Figures are complete. Label all axis, show the units on the axis, use a legend with clear differences between entries and add a title to each (sub)figure. Do not label sub-figures (a), (b) and explain what (a) and (b) are in the caption: instead label each sub-figure with a title. Do not use too thin lines or too small of a font.

WT3: Tables. For formatting tables, read section 2 of this document: "booktabs tables" and update your tables accordingly.

5.5 Introduction writing

WI1: Motivation and scope. The intro starts with a "just broad enough" motivation; not too broad, and not too specific, then quickly narrows the scope smaller, and smaller, culminating to exactly your topic.

WI2: No generic first sentence. The fist sentence of the introduction should focus/engage your audience. Don't use a sentence that can be added to any other paper in your field. Test if it would still make sense if you leave out the sentence, and, if it does, then leave the sentence out.

WI3: 3 contributions. Rule of thumb: your paper has 3 contributions. A contribution is something that a peer researcher will find interesting. Do not expect the reader to do inference work, so end the introduction by explicitly stating your contributions.

WI4: Figure 1. Make a visual abstract of the paper in Figure 1. Best if this is the main idea, but it can also be a pipeline figure.

WI5: Few research. Just because a topic has seen 'little research' is by itself not a good motivation (For example: my right thumb has not been researched at all, but its still uninteresting.). Don't motivate by what others have not done –put that in related work– instead: motivate what is inherently interesting about your research, ie: what problem does a possible 'user' have, and what can be gained by reading this paper? What would a peer-researcher find interesting to learn?

5.6 Related Work writing

WR1: The subject is the method, not the paper. Do not write: *The important work of [a] does X which is followed by the work of [b] that does Y.* It has papers as the subject. Instead, make the method the subject and add citations to the method: *X [a] is important, and extended by Y [b].*

WR2: Paragraph topic. One paragraph in the related work section is grouped around a single topic. Related papers often have multiple topics. It is up to you to group related work as best for you. Rule of thumb: Each paragraph has 3-10 citations.

WR3: Paragraph layout. The first sentence defines the scope. Then, the following sentences, you group papers based on what they do. The final, concluding, sentence is how are these methods *related* to your method. You have two options: option 1. *All so great, we make use of it.* Option 2. *All is great, but we do something different because . . .*

WR4: No history lessons. Related work means *related to your research question.* Avoid lengthy 'general history lessons' about methods/concepts that are more general than your research question. It is OK to give some history of work related to your research question. The goal of the related work section is about *motivating* why you choose the works that you build on, and why you choose the works that you contrast with.

5.7 Method writing

WM1: No general argumentation. The method should only motivate and explain the technical method. All argumentation for the main idea should be in the Introduction or in the Related Work sections.

WM2: No datasets. Datasets and their description belong in the experiments. Only a toy problem is allowed in this section if it helps to explain the method. The method section only explains the technical part of the method. (Exception: when you are writing a dataset paper)

WM3: Number all equations. Maybe you do not refer to them, but others (reviewers, readers) may want to. Only if the equation is not essential, it is OK to have it in the flow of the text without a number.

WM4: Treat equations as normal text. If an equation ends the sentence, the equations ends with a period. If the sentence continues, use a comma after the equation.

WM5: Explain all symbols. Directly before, or directly after introducing an equation: all symbols should be explained. A formula should be self-contained: the reader should be able to understand it without searching for symbol definitions elsewhere.

WM6: Explain equation in words. Directly before giving an equation, first explain why and what you aim to achieve in English. This makes the following equation easier to follow.

WM7: Should be understandable without equations. The method should also be understandable without reading the equations. Motivate and describe in English what is happening, the equations make the words exact. Test if you can still get the main point of the paper when all equations are removed.

WM8: Only define the relevant. Do not define everything you can think of, only define things you will actually use. Best to first write the results, and later define only what is needed to obtain these results.

5.8 Experiments writing

WE1: Answer a question. Every experiment starts with a question. Explicitly write this question. The experiment should answer that question.

WE2: Group an experiment together. Make use of sub-sections or bold-words, to help the reader understand the structure of the section. Each experiment is grouped as a module. Give each experimental question a number: "Experiment 1: How X applies to Y".

WE3: Analyze results modular. Every experiment has its own tables/figures for the results. Do not mix experiments by grouping all results in a single huge table. Group results together that are compared together. It may mean you have to repeat values in multiple tables/figs.

WE4: Experiment types. Broadly speaking there are three types of experiments. 1: *Validate*: does it do what you claim it does, (fully controlled setting?). 2: *Investigate*: what unique properties does your method have. 3: *Compare*: how does it compare to others. Present them in that order.

WE5: Scale scores between 0 and 1. Avoid useless zeros, scales 0-1 scores to 0-100. E.g.: '0.07' becomes '7'.

WE6: Proved. Experiments do not prove. A proof is derived in math, experiments demonstrate empirically for the setting at hand.

WE7: One more. When you think you are done, see if you can add one more experiment to show relevance for a different domain or application.

5.9 Discussion writing

WD1: Summary. Small summary of what you did to highlight the context.

WD2: Limitations. No method will always be the best. Showing insight where it fails is strong. The goal of research is understanding.

WD3: Conclusions. "Great paper; but So What?". Answer this question to draw conclusions. Don't make too broad conclusions; keep it modest and factual, and at the same time don't shy away in mentioning what is interesting and why.

6 Giving a Talk

Make a deliberate choice for the medium for your talk (whiteboard/slides/...). This document assumes that slides are used. Also see the writing guidelines; several are also applicable to giving a talk.

Motivation	**Content**	**Form**	**Analysis**
TM1 Goal	TC1 1 slide 1 topic	TF1 Too much	TA1 Exps answer Q
TM2 Audience	TC2 Less is more	TF2 No TOC	TA2 Limitations
TM3 Refresh	TC3 Self-contained	TF3 Layout	TA3 Peer review
TM4 Unburden	TC4 Define terms	TF4 No sentences	
	TC5 No guessing	TF5 Animate	
	TC6 Multi-modal	TF6 Complete figs	
		TF7 Number slides	

6.1 Motivation of your presentation

TM1: What is the goal? What you want to get out of it. There is a reason for giving your presentation: What is it? (and, no, it is not: 'it is my turn' or 'they told me to'). It may help to share this reason with the audience.

TM2: Audience. Whom are you presenting for? What do you want the audience to take away? What is their background and what are they looking for? Help your audience find it. Avoid Jargon.

TM3: Refresh. Always start with 1 or 2 slides (re-)introduction. Do not assume your audience will remember anything from your last time; there may also be new viewers present. If it is important: briefly repeat it.

TM4: Unburden the audience. If the audience misinterprets the message: its the responsibility of the presenter to reduce the understanding effort. Audience understanding can be validated by asking them.

6.2 Content of the presentation

TC1: A single slide has a single topic. A slide has a title to scope the topic. It has a concluding phrase that makes the main point of the topic.

TC2: Less is more. Every word/figure/image should have an explicit reason to exist. Do this test: *Can I safely remove it yes or no?* Do not put information on the slide that you don't want to answer questions about (eg: parts of a figure you took from another paper). Presenting concisely and precisely takes time and effort; it enhances understanding.

TC3: Self-contained. The audience has not memorized the full presentation. Remind the audience of definitions or symbols when defined 'a long time ago'.

TC4: Define terms. Define all symbols/terms. Use a defined symbol/term consistently and uniquely.

TC5: No guessing Never expect the audience to do inference. If the viewer has to guess, the guess will often be not what you had in mind. Always explicitly write what the viewer is supposed to see/conclude. Ie: put the answer to the "So What" question on the slide.

TC6: Multi-modal There are various people in the audience whose preferences range from visually, formulas, auditory. Be sure to present a mix.

6.3 Syntax, layout and form

TF1: Do not present too much. A rough guideline: at least 1 minute per slide.

TF2: Do not use a table of contents. Avoid the default toc of "• Intro, • Method, • Exps, • Conclusion"; this is expected, so this toc adds nothing. Another form (e.g.: visual abstract) can be useful.

TF3: Good layout eases the viewer's effort. Use the full screen. Be consistent. Not too much info in a slide.

TF4: Do not write long sentences. Use bullet points with one phrase per point. One phrase fits on a single line. Correct grammar is secondary, e.g., there is no need for complete sentences with a subject, a verb, etc.

TF5: Only use animations sparingly and IFF they add value. For example, to emphasize, or to prevent overload by iteratively making more content appear.

TF6: Figures are complete. Label all axis, show the units on the axis, use a legend with clear differences between entries and add a title to each (sub)figure so that the reader can directly see what is shown. Do not use too thin lines or too small of a font: It has to be seen from the back of the room. Add the conclusion you would like the viewer to draw.

TF7: Number your slides so that viewers can refer to them.

6.4 Presenting analysis

TA1: Experiments answer a question. If you present experiments, note that every experiment starts with a question. Write the question on the slide (maybe in the title?). The experiment should answer that question. Write the answer on the slide.

TA2: Limitations. What are the limitations of your method. No method will always be the best. Showing insight where it fails is strong. The goal of research is understanding.

TA3: Peer review. Find a peer to review each others presentations. Check if their presentation follows these guidelines. Keep in mind that if an honest viewer did not understand it, then the presentation should be improved (not the viewer).

7 Poster guidelines

Poster presentations are common for presenting research ideas. I've found this blog useful: How to design an award winning poster. I prefer to do my posters in inkscape. Several of the writing guidelines and guidelines for giving a talk also apply here.

Motivation	**Content**	**Form**	**Analysis**
PM1 Excite	PC1 1 block 1 topic	PF1 Too much	PA1 Exps answer Q
PM2 Audience	PC2 Less is more	PF2 Reading order	PA2 Limitations
PM3 Refresh	PC3 Self-contained	PF3 Layout	PA3 Peer review
PM4 Unburden	PC4 Define terms	PF4 No sentences	
	PC5 No guessing	PF5 Draw attention	
		PF6 Complete figs	
		PF7 Find examples	

7.1 Motivation of your poster

PM1: Excite the viewer The goal of a poster is to advertise your research so people will want to read your report/paper. Excite us!

PM2: Audience. Whom are you presenting for? What do you want the audience to take away? What is their background and what are they looking for? Help your audience find it. Avoid Jargon. What do you want to get out of it from them?

PM3: Refresh. Do not assume your audience will have remembered anything from any other source; there may also be new viewers present. If a topic is important: briefly repeat it.

PM4: Unburden the audience. If the audience misinterprets the message: its the responsibility of the presenter to reduce the effort of understanding. Audience understanding can be validated by asking them.

7.2 Content of the presentation

PC1: A single block has a single topic. A modular block in your poster has a title to scope the topic. It has a concluding phrase that makes the main point of the topic.

PC2: Less is more. Every word/figure/image should have an explicit reason to exist. Do this test: *Can I safely remove it yes or no?* Presenting the core essence takes time and effort; it enhances understanding.

PC3: Self-contained. The main point of the poster has to be understandable without a presenter. While you are busy explaining your poster, another viewer who just walks in should be able to understand the key idea without your help.

PC4: Define terms. Do not assume the audience will know specific symbols/terms/abbreviations. Use a defined symbol/term consistently and uniquely. All terms in an equation should be explained.

PC5: No guessing Never expect the audience to do inference. If the viewer has to guess, the guess will often be not what you had in mind. Always explicitly write what the viewer is supposed to see/conclude (answer the “So What” question).

7.3 Syntax, layout and form

PF1: Do not present too much. Your goal is to advertise and excite. Nitty gritty details should be in the report not in the poster. Show just enough so a reader can follow, no more. Do not overwhelm, do not even try to be complete with all details: it will scare people away.

PF2: Use numbers to show reading order. Make explicit how you wish your poster to be read: Numbers the reading order.

PF3: Good layout eases the viewer’s effort. Be consistent. Keep some white-space, don’t scare people away with an avalanche of detail.

PF4: Do not write long sentences. Use bullet points with one phrase per point. One phrase fits on a single line. Correct grammar is secondary, e.g., there is no need for complete sentences with a subject, a verb, etc.

PF5: Draw attention. Make your poster visually stand out from all others. The goal is to advertise.

PF6: Figures are complete and have a conclusion. Label all axis, show the units on the axis, use a legend with clear differences between entries and add a title to each (sub)figure so that the reader can directly see what is shown. Do not use too thin lines or too small of a font. Always add the conclusion you would like the viewer to draw.

PF7: Example poster. Find some scientific posters on the internet and apply these guidelines. Make a list of things you do/do not like in a poster.

7.4 Presenting analysis

PA1: Experiments answer a question. If you present experiments, note that every experiment starts with a question. Write the question on the poster. The experiment should answer that question. Write the answer on the poster.

PA2: Limitations. If applicable: What are the limitations of your method. No method will always be the best. Showing insight where it fails is strong. The goal of research is understanding.

PA3: Peer review. Find a peer to review each others posters. Check if their poster follows these guidelines. Keep in mind that if an honest viewer did not understand it, it is the mistake of the presenter.

8 Reviewing and rebuttal guidelines

Why review. If you are writing peer-reviewed papers, then you are asking others to devote time and effort to your work. Thus, it's fair to return the favor and review the work of others. Reviewing is part of your academic community. It offers other advantages: improving the scientific field, learning something new, practicing your critical thinking, and helping others.

What to review. I've found the "Troubling Trends in Machine Learning Scholarship" [6] paper quite helpful, where I regularly give back a review where I state that a paper is *Confusing explanation with speculation* and/or has a *Failure to identify the sources of empirical gains.* In addition, I use Hitchens's razor: "*What can be asserted without evidence can also be dismissed without evidence.* And, reviewing involves applying the research guidelines in this document, albeit not during the process, but at a finished paper. Specifically, see if you can find the storyline as in Section 2, look for the answers that can be asked during a research meeting in Section 4, and the writing guidelines in Section 5. I've labeled each guideline with a unique identifier, which could help in motivating a review by referring to this document and it's identifiers.

Goal of reviewing. The main goal of reviewing is to improve the work of others by giving feedback while preventing the publication of flaws. This might include flaws in the following. Hypothesis: Are the hypotheses sound? Literature: Are the relation to relevant other work present, correct and properly motivated why and how the work is related?. Method: Is the method aligned with the hypotheses? Technical: Are the equations correct? Does it do what is claimed? Are there no unexplained surprises? Is it reproducible, ie: code? Experimental setup: are the hypotheses evaluated? Is the motivation evaluated? Evaluation: is it an unbiased, fair, comparison to others? Clarity: is it understandable? Figures/tables readable? it's OK if there are minor spelling/grammar mistakes, as long the paper can reasonably be well understood without too much puzzling. As a reviewer, it typically cannot be expected to rerun experiments, a review is inherently based on trust in the author's integrity. A perfect paper does not exist, all papers are limited in some sense. Thus, be critical, but appreciate the positives.

What is a good paper. It's a solid brick that others can build on: something is learned. It's well written with an intuitive motivation, for example in Fig 1. It has clearly specified hypotheses, research questions, and contributions. The method aligns well with hypotheses. The Hypotheses, research questions and contributions are backed up by empirical evidence. Comparison experiments vary only 1 variable. It has experiments on several datasets to illustrate generalization. Bold numbers are never a goal in itself, they are 'only' important to show relevance/usefulness. It's reproduceable, it has clear algorithms, or better: code.

Addressing novelty. It's easy to do something novel: merely add a layer, and it's 'novel'. Novelty is not a goal in itself: a paper about my left thumb is extremely novel, but that does not make it, nor the paper, interesting. It is up to the paper under review to explain in the introduction and related work sections how it relates to others, and what contributions it has compared to the other work. Just because the outcome is "obvious", or "trivial" is not good grounds for rejection: the paper has now confirmed this outcome; and this confirmation is a contribution in itself. A possible ground for rejection could be if there is other near-identical work but not placed in relation to the paper under review and/or if it is not experimentally compared against.

Review quality. A bad review: makes claims without giving details, citations, or motivation. Is only a few lines. Only checks the bold numbers. A good review gives constructive author feedback, so in addition to What, it also suggests How to change a paper. Is well motivated, with detailed justification (citations / line numbers). Is well-written and self-contained: the review is readable without the paper. It makes the decision for the AC easier.

My review structure Review formats vary slightly for each conference. I always use this layout in a .txt file, which can be poured in any format. When I am reading the paper for the first time, I directly write comments per line. Once these detailed comments per line number are there, then the other points follow from them. Review structure:

- Summary. Unbiased, the authors should agree with it, introduce terms that you will build on later so that the review can be self-contained.

- List of positive points: just 1 line per point
- List of negative points: just 1 line per point
- A conclusion paragraph of score motivation and main suggestions for what the paper could address in the rebuttal. This paragraph builds on the summary and the list of positive/negatives.
- Detailed comments per line number with detailed justification.

8.1 Rebuttal guidelines

Why write a rebuttal. Several conferences, and journals, allow for a rebuttal: a factual response to the reviewers. The main goal of a rebuttal is to correct mistakes, and convince the reviewers your work is interesting. Even if there is only 1 positive reviewer it can help to write a rebuttal: there often is a discussion phase, where your ”champion” can then defend the paper.

Not novel Typically, when reviewers write that there is insufficent novelty –without citing a missing paper– then what they really mean is that, unfortunately, they did not find the paper interesting. Yes; finding something interesting is subjective; and this is probably why they write 'not novel' instead of 'I didn't find it interesting'. I suspect that reviewers are afraid to be honest because it is not possible to give objective reasons for why something is not interesting (to you). This is unfortunate, because as an author this would be valuable feedback to have. If reviewers find it not novel, then try to ask yourself why they didn't find it interesting. Try to also ask a (somewhat senior?) not co-author colleague, who is not afraid to tell you the truth to your face.

Science is done by humans Many reviewers: do their work too fast; have a rejection mentality; do not read the paper well; write a too short and unmotivated review; or base the review on the author's name if a preprint or blog is available. Receiving a 'bad review' can be quite frustrating; especially when you spent all that effort on your paper. The only advice I can give is try to learn something from the review anyway. If the reviewer did not understand the paper: what can be improved? How can the paper be made easier to parse? How to improve ”something between the lines” that they did not like?

Because science is done by humans, its also important to address the reviewers as human beings:

- Always thank the reviewers (Don't "over-thank").
- Assume they will not change their mind more than 1 point (It might happen, but is psychologically difficult)
- If technically possible: Do what they ask, even if it doesn't make sense (to you). The most convincing response is to just show it.
- If you fight/shout/insult: they will fight back in the discussion; and they will have the last word.
- Write for reviewers and area-chair/editor; having the reviewers on your side is much easier to get accepted.
- Make it easy for the reviewer to find their answer (do not 'hide' the answer somewhere in a lot of text). Thus, try to copy the comment of the reviewer verbatim
- Write the rebuttal self-contained: ideally, they should not need to go back to the paper, nor to any of the reviews.
- Do not take reviews personal (you are not your work)
- Reply positive, non-defensive and to the point
- Be polite and professional, but self-assured and firm
- Long and too dense rebuttals will scare reviewers away. Leave sufficient white space.

My approach to writing a rebuttal is as follows:

1. Copy-paste all concrete positive and negative points in a doc
2. Answer each negative point
3. Perform all requested experiments (to good approximation)
4. Group similar (positives and negative) points

5. Start by summarizing grouped (verbatim) positive comments

6. Answer grouped (verbatim) negative comments

7. Rephrase negative answers and compactly rewrite

8. Decide which answers to drop strategy (eg: Convince one reviewer, but keep others).

9. Ask someone else to read rebuttal and ask how they feel

9 Appendix: Separately printable tables.

Each guideline is uniquely labeled, which can be used to give feedback, during a meeting, or to use as a reminder. Below I copied all tables on a separate page for convenient printing and adding your own guidelines.

Organization	**Process**	**Mentality**
RO1 Full responsibility	RP1 One main Q	RM1 Be critical
RO2 No dependencies	RP2 Min. 3rd party	RM2 Find todos together
RO3 Meet advisor	RP3 Validate baselines	RM3 Consistency
RO4 Focus advisor	RP4 First break it	RM4 Question everything
RO5 Take critique	RP5 Depth first	RM5 Simple is strong
RO6 Constructive disagree	RP6 Exps answer Q	RM6 Embrace limitations
RO7 Analyze results	RP7 Proof of concept	RM7 Write early and often
RO8 Suggest solutions	RP8 Exps max 1 night	RM8 Not eureka
RO9 Give feedback	RP9 Change 1 var	RM9 Show the problem
RO10 Safety	RP10 Debug science	RM10 Motivate everything
	RP11 Figures	

Research meeting questions		
Research question	**Related work**	**Method/approach**
RQ1 Why interesting?	RW1 Who will use it?	MA1 Why this method?
RQ2 What storyline?	RW2 How different?	MA2 Explain each step?
RQ3 Formalize/simplify	RW3 Builds on what?	MA3 Formalize/simplify
RQ4 What problem?	RW4 Baselines?	MA4 Alternatives?
		MA5 Align with RQ?
Experimental setup	**Analyzing**	**Conclusions**
ES1 What Qs?	AR1 Validate?	DC1 Exhaustive?
ES2 How answer Q?	AR2 Baseline?	DC2 Expectations?
ES3 Baselines?	AR3 Understand all?	DC3 Align with RQ?
ES4 Expected outcome?	AR4 When fail?	DC4 Simplify?
ES5 Simplify?	AR5 Link to Q?	DC5 New hypotheses?

Writing guidelines

General	Structure	Form
WG1 Unburden	WS1 Self-contained	WF1 Single topic
WG2 Audience	WS2 Consistent	WF2 Windows/orphans
WG3 Less is more	WS3 As discussed	WF3 Very
WG4 No guessing	WS4 Paragraphs	WF4 In order to
WG5 Read out loud	WS5 Ref words	WF5 Sort cites
WG6 More space	WS6 Ref paragraphs	WF6 Brackets
WG7 Write as code	WS7 Latter/Former	WF7 Synonyms
		WF8 Performance
Tables/Figs	**Introduction**	**Related work**
WT1 Captions	WI1 Motivation	WR1 Subject
WT2 Figs are complete	WI2 First sentence	WR2 Paragraph
WT3 Tables	WI3 Few research	WR3 Layout
	WI4 Fig 1	WR4 No history lessons
	WI5 3x contribute	
Method/approach	**Experiments**	**Discussion**
WM1 Argumentation	WE1 Question	WD1 Summary
WM2 No datasets	WE2 Group	WD2 Limitations
WM3 Number eqs	WE3 Analyze	WD3 Conclusions
WM4 Eqs are text	WE4 3 types	
WM5 Explain symbols	WE5 Scale 0-1	
WM6 Explain eq	WE6 proved	
WM7 Remove eq	WE7 One more	
WM8 Define		

Presentation guidelines

Motivation	**Content**	**Form**	**Analysis**
TM1 Goal	TC1 1 slide 1 topic	TF1 Too much	TA1 Exps answer Q
TM2 Audience	TC2 Less is more	TF2 No TOC	TA2 Limitations
TM3 Refresh	TC3 Self-contained	TF3 Layout	TA3 Peer review
TM4 Unburden	TC4 Define terms	TF4 No sentences	
	TC5 No guessing	TF5 Animate	
	TC6 Multi-modal	TF6 Complete figs	
		TF7 Number slides	

Poster guidelines

Motivation	**Content**	**Form**	**Analysis**
TM1 Goal	TC1 1 slide 1 topic	TF1 Too much	TA1 Exps answer Q
TM2 Audience	TC2 Less is more	TF2 No TOC	TA2 Limitations
TM3 Refresh	TC3 Self-contained	TF3 Layout	TA3 Peer review
TM4 Unburden	TC4 Define terms	TF4 No sentences	
	TC5 No guessing	TF5 Animate	
	TC6 Multi-modal	TF6 Complete figs	
		TF7 Number slides	

References

[1] Serge Demeyer. Tutorial: Research methods in computer science. `https://win.uantwerpen.be/~sdemey/Tutorial_ResearchMethods/`, 2010.

[2] Tom Dietterich. Research methods in machine learning. `http://web.engr.oregonstate.edu/~tgd/talks/new-in-ml-2019.pdf`, 2019.

[3] Paul Feyerabend. *Against Method: Outline of an Anarchistic Theory of Knowledge*. New Left Books, 1975.

[4] Sam Greydanus. Scaling down deep learning. *arXiv preprint arXiv:2011.14439*, 2020.

[5] Thomas S Kuhn. *The Structure of Scientific Revolutions*. The University of Chicago Press, 1962.

[6] Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. arXiv, 2018.

[7] Rick Rubin. *The Creative Act: A Way of Being*. Penguin Random House, 2023.

[8] Julian Togelius and Georgios N. Yannakakis. Choose your weapon: Survival strategies for depressed ai academics. arXiv, 2024.

[9] Itai Yanai and Martin J. Lercher. What is the question? *Genome Biology*, 20, 2019.