The Storyline

Jan van Gemert
Delft University of Technology
Full Research Guidelines are:
`https://jvgemert.github.io/links.html`

1 Research as a creative process

My love for doing scientific research is in its understanding-driven 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 [1]. 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's often –by design– not even clear what the precise question is [2]. 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 catalyzing the scientific creative process in search of the question.

2 The storyline

The storyline grew as one of my most valuable tools to anchor 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

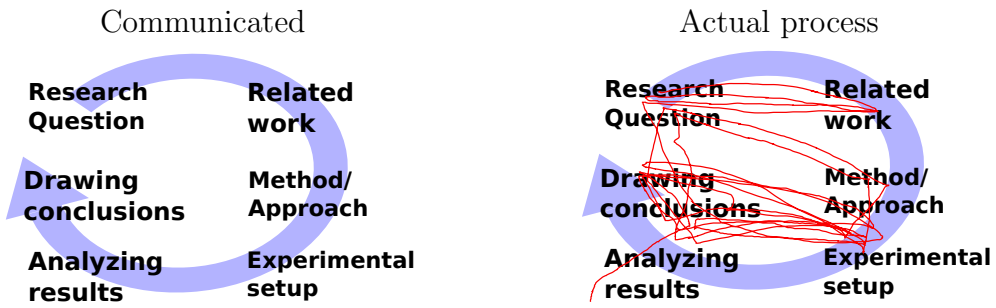


Fig 1: How machine/deep learning research is typically communicated *vs* an example of how the creative scientific process can go (red line). It is completely normal during the research process to rephrase the research question often; to revisit related work in a new light, to change the method or experimental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

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 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 claim made. 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 2.

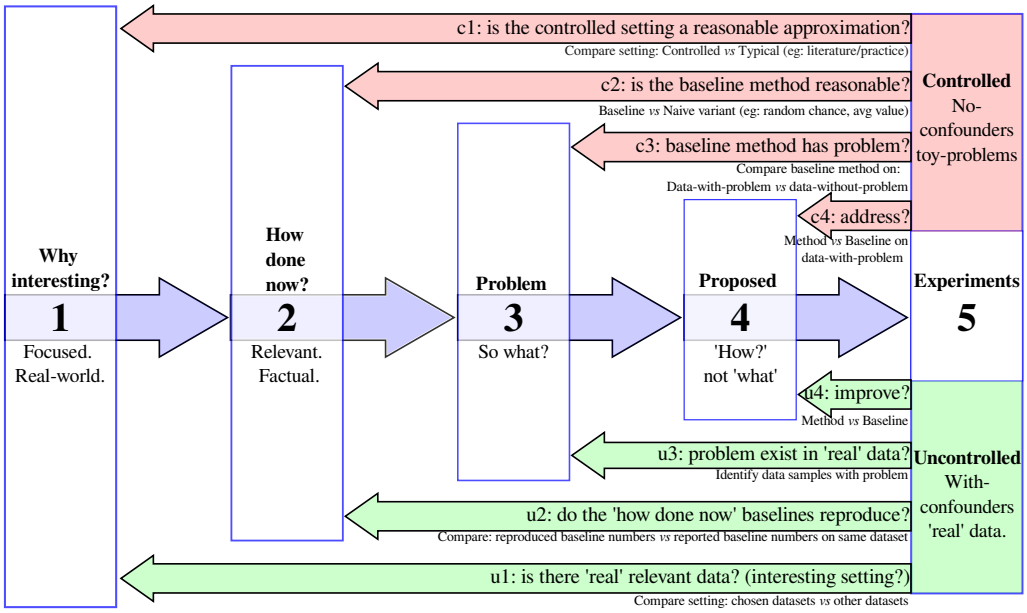


Fig 2: The storyline structures empirical machine/deep learning research. It concisely ties the research in a logically consistent focused narrative. It has the following elements. (1) Why the (real-world) setting is interesting; (2) factual description of relevant existing approaches (baselines); (3) what is the problem of the baselines and what are the consequences; (4) how 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 controlled experiment models reality sufficiently well, i.e.: it's unbiased; (c2) the baseline method is set up reasonably and fairly, that this reasonable baseline (c3) does indeed suffer from the problem; and (c4) that the proposed approaches addresses it. Where the controlled experiments are designed with a single confounder, the uncontrolled 'real world' 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) reproduced baselines match 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); i.e. baselines to compare to: Usuals and Competitors.

(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. Prepare to kill your darlings.

2.1.1 Storyline element (1): Why interesting?

What is the down-stream 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 (doing research) and not the reason for this reaction. If your research is helping other scientists, then state the down-stream goal that those scientists are trying to achieve. Open the abstraction: i.e.: what are the underlying reasons that topic XXX has received so much attention? Make explicit what is interesting/beneficial/useful/important/etc., about it. Keep digging deeper and repeatedly ask "So what? Why is that interesting?" until you cut away all abstractions and arrived at the concrete core. One option, but not the only one, that might help, is to realize that much science is publicly funded, and then state why the public should fund your research. I believe that a researcher should be able to give at least (just?) one reason for what the research could possibly be 'used' for, in some possible future.

Keep the scope tight, and focus on your research outcomes. Your problem (3) and what you propose (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) and (4).

For example, consider a method for bias visualizing. If the claim is that the method can find new types of biases then this claim can be challenged, and thus experimentally motivated by showing such new biases. As another example, consider an optimization method that makes auto-encoders faster to train. Then if the claim to interest is that auto-encoders are used in de-noising or super-resolution applications, then evidence for this claim will show that de-noising, super-resolution applications actually become faster to train. The burden of proof lies with us, and claims on why the research is interesting need evidence; otherwise they can equally be dismissed without evidence. If it's not clear why its interesting, then why should a reader spend time and effort in reading your work?

2.1.2 Storyline element (2): How done now? (baselines)

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) or (4).

A rule of thumb is to put work here which you will compare to as a baseline. A baseline typically can be decomposed in *Competitors*, and *Usuals*. A *Usual* is the standard, text-book, approach in literature. For example a typical CNN, ViT, Transformer, RNN, etc. A *Competitor* is a method that also addresses the motivation in (1). For example, if the motivation in (1) is about reduced accuracy due to accidental camera rotations, then the *Usual* baseline could be a CNN, and a *Competitor* would be data-augmentation, which also aims to solve the problem of spatial rotations. Both these methods would be baselines to compare to.

The described 'how done now' baseline approaches here should be objective and 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 used term should be motivated/introduced first. Pro-tip, 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 2 for how experiments build on each element.

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. The controlled setting allows for "*data-without-problem*" (a control) and "*data-with-problem*" where we precisely know beforehand which is which. To know which data has the problem, and which data does not, is important as to avoid unknown confounding variables which might, unknowingly, influence the results. The "*data-without-problem*" variant is a 'normal' setting, without the problem i.e.: a control, which demonstrates that the setting is reasonable, that the existing approaches are represented fairly (i.e.: they do OK), and to set a baseline performance. The "*data-with-problem*" variant is identical to the "*data-without-problem*" setting, although it crucially varies in only a single manner, it varies only in 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, nor for models with huge parameters. Too high baseline accuracy might actually be detrimental: if the baseline already scores 95%, then the proposed approach can only make marginal relative improvements and we cannot show that the proposed approach makes any difference.

A second line of experiments investigate impact in the 'real world'. Put another way: how severe is the identified problem, and its consequences in datasets with other, unknown, confounders? 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 and that the proposed approach addresses it while being robust to the unknown confounders. 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.

One issue I often see, is that experiments are conducted without a link to the specific research question. For example, if the main topic is data-efficiency, why analyze the effect of the learning rate? If the topic is reasoning, why investigate the effect of adding noise? It's not the case that noise or learning rate are not important; they are, but they are not tightly linked to the main research questions. The learning rate belongs in a paper on optimization, not on data-efficiency. The noise belongs in research on robustness, not on reasoning. Of course they can still belong, but then there should be an explicit, logical, reason for them, and this reason then needs to be described in the storyline.

2.2 Collaboration through the storyline

The storyline forces a focused logical narrative that fits on a single page. Such a concise format facilitates communication because the whole narrative is visible and allows evaluating the full logical research structure by collaborators. If collaborators do not understand, or do not agree with the logic, then this is valuable feedback, and the storyline should be updated which corresponds to taking a step in search of the research question.

When discussing a storyline with it's creator(s), I find myself typically in two different modes; in **construct mode**, we are together coming up with the storyline, often based on initial results, an initial method, or a research direction or problem. In **critique mode** I challenge the storyline as it is presented to me, typically by a detailed scrutiny of each word. This critiquing can be perceived as 'attacking' and might lead to unconstructive communication. I've found that if the storyline is still quite preliminary/incomplete/vague, then it's better to skip critique mode, and first go to construct mode; this step can be initiated by both the feedback giver, but also the storyline creator. Once there is a more converged version of the storyline, then it becomes constructive to critique it's details. Note, it's important to be 'pedantic', because the logical narrative guides every part of the research. Also, it helps to discuss and go over the same storyline multiple times, because our future selves might see other things than our past selves.

Below are common problems that I encounter when discussing the storyline. Please use these as a self-check, after completing a storyline.

Often re-occurring 'content' problems:

1. Not clear/precise why interesting or why people should care. Often the text is too general/abstract. Ideally, the reasons given come back in the experiments, and as such need to be concrete: we should be able to point at something that can concretely be done in an experiment.
2. Not clear/precise how it's done now: what are *relevant* current typical approach(es), often it is not specific enough.
3. Not clear/precise what is missing; what the problem is, and what consequences this problem has. Often the consequences are not linked back to the "why interesting". The consequences should be precisely the concrete things that the experiments will show that the proposed method addresses.
4. Not clear what you propose; what your method is/does.
5. Not clear/precise why your proposal/method could solve the problem. This is often either too general; or it is too technical. It should be at the "logical narrative" abstraction level.

6. Not clear/precise what experimental questions you ask that demonstrate solving a problem and addressing its consequences.
7. Includes irrelevant information: parts of the storyline can safely be removed without changing the main narrative.

Often re-occurring 'form' problems:

1. Using a term before defining/motivating it. Tip: highlight a term when it's first introduced/motivated and highlight it when its re-used.
2. Too much unnecessary detail: words can be removed without significantly changing the storyline. Each word should have a reason to exist.
3. Unclear logical reasoning step.
4. Inconsistent use of terminology. Use a single term for a single concept; 1-to-1, consistently.
5. Writing is too verbose and too many full sentences are used: bullet point should be short one-two lines; correct grammar is optional.
6. Too many topics per storyline bullet point. Keep a single point on a single topic.
7. Too many storyline bullet points. Keep the storyline focused. Less is more. Find the minimal logical story for the research project.
8. The text is not stand-alone; it's not peer understandable. Too much jargon, or 'suitcase words'.
9. Storyline is spread over multiple pages/slides. The goal is to see all parts together; try to trim words or grammar. Or, if really not possible, smaller margins, or a smaller font size.
10. Not following my writing guidelines: `https://jvgemert.github.io/writing.pdf`;

2.3 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*". The storyline is a tool to be used flexibly. 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 focus only on relevant questions and to back up any claim with evidence; or adapt the claim accordingly.

Storyline for *Color Equivariant Convolutional Networks*, NeurIPS, 2023. `https://arxiv.org/abs/2310.19368`. Each first-time use of a term is colored; a re-used term is underlined; illustrating the logical narrative built on terms that are first introduced/motivated before they are used.

1. **Why interesting?**
   a Visual AI trains on inherent imbalanced data due to accidental transformations such as camera viewpoint (spatial) or appearance (light).
   b Reduced accuracy for infrequent occurrences.
2. **How done now?**
   a Invariance : removing the effect of the accidental transformation
   b Data augmentation : add transformation to the training data
   c Equivariance : sharing learnable weights over transformations
3. **What is missing, and So What?**
   a Data augmentation is useful, but only partially addresses it.
   b Invariance removes too much, also relevant information. Also, Data augmentation cannot be used because invariant to them.
   c Current equivarance work is for spatial transformations, not appearance.
   d So: reduced accuracy due to imbalance in appearance.
4. **Proposed approach (P).**
   a Focus on color (hue) as a case of non-spatial appearance.
   b Sharing weights over different color hues H in HSV color space.
   c Proposed (P): color equivariance by rotations in hue space.
   d Data augmentation remains useful.
5. **Experimental questions.**
   a Sharing weights over colors addresses imbalance?
      c1: imbalanced 30 color MNIST classes (10 digits x 3 colors)
      c2: Usual-baseline does well ($\approx 90\%$) on frequent samples
      c3: Usual-baseline has reduced accuracy on infrequent samples.
      c4: Frequent: P $\approx$ usual; Infrequent: P $>$ Usual; and P $>$ invariant
   b P learns color when relevant; and ignore color if not relevant?
      c1: 10 MNIST classes; class has color $c \sim \mathcal{N}(\theta_c, \sigma)$, vary σ
      c2: Usual-baseline has good accuracy ($> 90\%$).
      c3: Usual-baseline has reduced accuracy for increasing σ.
      c4: P $>$ baseline; P $>$ invariant for small σ. P $\approx$ invariant for high σ.
   c How sensitive is P with confounders under a color transformation?
      u1: 8 existing color image classification datasets
      u2: Usual-Baseline results for all datasets are on-par.
      u3: Detailed results of P vs Usual/Invariant/Augmentations on 1 set.
      u4: Normal setting: P $\approx$ baseline. Color transf. setting: P $>$ baseline.

On Translation Invariance in CNNs: Convolutional Layers Can Exploit Absolute Spatial Location, CVPR'20. https://arxiv.org/abs/2003.07064

1. **Why interesting?**
   a Data efficiency: collecting/labeling data to train visual AI is expensive
2. **How done now?**
   a Shift equivariance in CNNs: A sliding spatial window with shared network weights over different spatial locations improves data-efficiency.
   b Weights applied to the spatial window is output at the window-center
   c When the sliding window reaches the image boundary it stops, or, zeros are padded outside the image boundary for half the window size.
3. **What is missing, and So What?**
   a Insight: only the center of the sliding window is kept, so, pixels on the image boundary are only passed through half the sliding window.
   b Ie: CNNs can asymmetrically ignore image content close to one side of the image boundary and not the other side of the image.
   c Thus, surprisingly: CNNs can learn to break shift equivariance and have an absolute location dependency, through distance to image boundary.
   d So: when shift-equivariance is broken; data efficiency suffers.
4. **Proposed approach (P).**
   a Remove the ability of the model to exploit absolution location:
   b Make the sliding window see all input pixels (e.g., the left part of the window should also see all pixels of the right part of the image)
5. **Experimental questions.**
   a How far from the image boundary can (scratch/pre-trained/random) CNN models use absolute location?
      c1: Imagenet images; classify Left-Up vs Right-Low; vary boundary dist
      c2: 100% accuracy for small boundary distance, for all models.
      c3: Accuracy drops for different models at different boundary distances.
   b Is location used when not needed? For location-independent task: train on one location; test on a different location.
      c1: Task: is red-box to-the-left-of green-box. Train: class-1 location at top, class-2 location bottom. Test: swap class locations.
      c2: Near 100% baseline accuracy when test locations = train location
      c3: <10% accuracy when test location ≠ train location
      c4: P: Same accuracy independent of location.
   c Does restoring CNN shift-equivariance gain data efficiency?
      u1: Imagenet, patch-Matching; action recognition: UCF101; HDMB51
      u2: Good accuracy with much data
      u3: Accuracy drops with less data
      u4: Accuracy drops less with less data: Data efficient.

Empty storyline

1. **Why interesting?**
   a ...
   b ...
2. **How done now?**
   a ...
   b ...
3. **What is missing, and So What?**
   a ...
   b ...
4. **Proposed approach (P).**
   a ...
   b ...
5. **Experimental questions.**
   a ...
      c1: .
      c2: .
      c3: .
      c4: .
   b ...
      u1: .
      u2: .
      u3: .
      u4: .

References

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

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