The Storyline

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

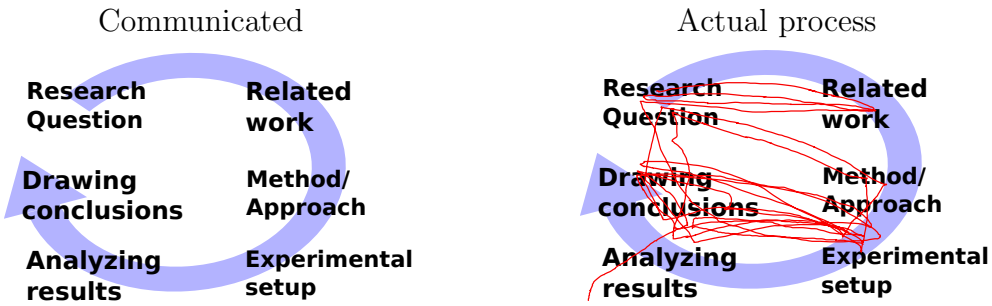

Fig 1: How machine/deep learning research is typically communicated *vs* an example of how the actual scientific process can go (red line). It is completely normal during the research process to rephrase the research question often; to revisit related work in a new light, to change the method or experimental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

1 The storyline

I've found the storyline one of the most important tools in empirical machine/deep learning research: It helps to structure several things in the research process: thoughts, meetings, the process, the research questions, presentations, the logical narrative, and enormously helps the writing by postponing sentence structure till later. It is normal that the storyline changes/evolves many times, see Fig 1. A visualization of the storyline is shown in Fig 2.

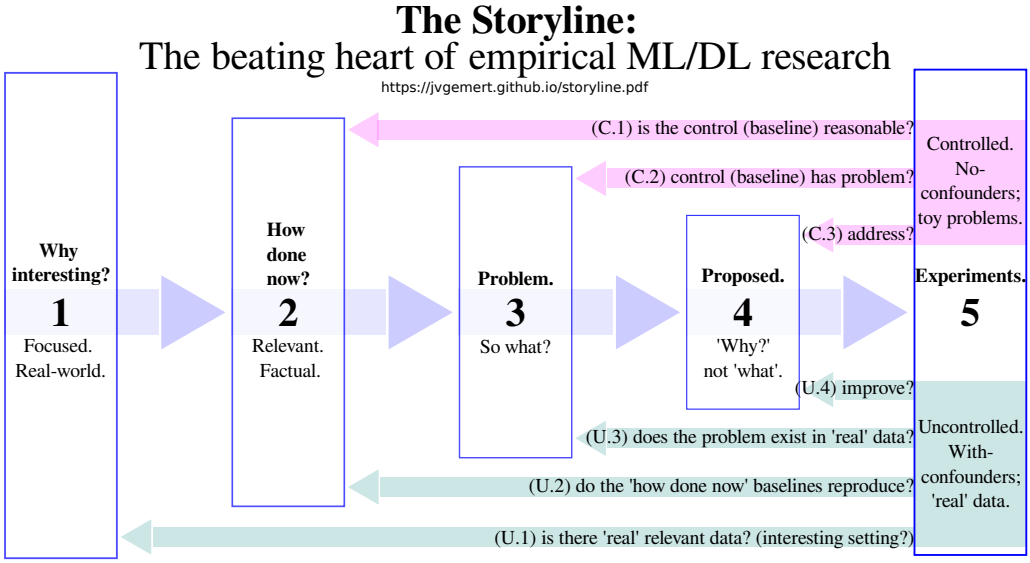

Fig 2: The storyline structures empirical machine/deep learning research. It concisely ties the research in a logically consistent focused narrative. It has the following elements. (1) Why the real-world setting is interesting; (2) factual description of relevant existing approaches (baselines); (3) what is the problem of the baselines and what are the consequences; (4) 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 (C.1) the baseline (control) setting is set up reasonably and fairly, that this reasonable baseline (C.2) does indeed suffer from the problem; and (C.3) 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 (U.1) by demonstrating multiple relevant data sets; that (U.2) published baselines reproduce; that the identified problem (U.3) exists among uncontrolled confounders, and thus (U.4) that the proposed approach improves over the baseline when the problem exists.

1.1 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 society 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.

1.1.1 Storyline element (1): Why interesting?

What is the motivation to do this research? Why should 'the person in the street' 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. Dig deeper: 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 repeated ask "Why is that interesting?" until you arrive at the core. Put another way: why should society 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).

1.1.2 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 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).

1.1.3 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.

1.1.4 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.

1.1.5 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. 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 CI-

FAR, or ImageNet are out, because they do not contain rotations. Instead, use datasets where rotations occur naturally; for example cell images taken under a microscope. Here, it is important to validate that existing methods on the datasets actually reproduce (do not assume they will!). These methods are the baseline to compare to. If the problem occurs in the dataset, and the proposed approach handles the problem well, then it can be expected that baselines are outperformed.

1.2 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'.