How to Read a paper?

Project Preparation Unit

Daniel Lawson — University of Bristol

v1.0.1

How to Read a Paper

- There are no rules on how to read a paper
- But good practice can help
- It will depend on:
  - The reader
  - The goal
  - The time commitment
- Reflecting on the process can increase efficiency

Four stages

It is helpful to split the reading into stages

0. **Context**: what field is it? what is the approach in that field? what is your prior experience?
1. **Skim**: why should you read it?
2. **Assess**: what does it really say?
3. **Assimilate**: how does it do what it does?

Context: before you start

- Are you already a **subject matter expert**?
  - Is it pre-peer review?
  - Are you trying to learn something new?
- Is the field big? Should you know key references?
- Is there **hype** for this field?
- What are the appropriate **metrics** for this field? Is it well/partially cited? Is it empirical or theoretical?
- What would a **“good paper”** look like in this field?

Exploring a new field

- Is this paper reputable?
  - Is it published in a **good journal**?
  - Is it published by **known authors**?
  - Does it have the backing of a **known institution**?
  - Does it have an **agenda**?
  - Has it been well **cited**?
- Should you instead start with background reading?
- Try to identify at least two solid starting points (books/key papers)

Exploring a new field

- What is the **Big Question**? What is the field as a whole trying to solve?
- Read the introduction. Try to succinctly **summarise the background**.
- Identify the **Specific Question(s)**.
- What is the overall approach being taken?

Assessing within your own field? A Three Pass Approach

- **Why?** should you read it (*skim*)?
  - What are the **claims**?
  - Assess the overall method and approach.
  - What scale of exciting is it to people **within** the field, and across it?
- **What?** does it really say (*assess*)?
  - Assess the **figures** and **results** for validity and generality
  - Are **data** and/or **code** properly available?
  - Does it match the claims?
- **How** does it work (*Assimilate*)?
  - Re-create the work virtually
  - What is missing? What criticisms can you find?

Reading a paper

- We'll follow this approach with two papers for contrast.
- The Tradeoffs of Large Scale Learning by Leon Bottou and Olivier Bousquet (2007).
- Statistical frameworks for detecting tunnelling in cyber defence using big data by Lawson, Rubin-Delanchy, Heard and Adams (2014).

Three pass approach: Pass 1 (WHY)

- Process:
  - Carefully read title, abstract, introduction.
  - Briefly look at the **figures** and **section titles**. Are there any key display items?
  - Read the conclusions.
  - Note references with which you are familiar.
- From this you can assess the **category** of the paper, its **context**, its **contributions** and **clarity**. You may already know whether it is incorrect.
- This is how an editor might read a paper. It should take approx 15 minutes or less. You should know who should read the paper, and **why**.

Outputs of Pass 1

- **Category**: What keyword/s and field/s describe the paper? Is it interdisciplinary?
- **Summary**: 2-3 sentences explaining what the paper does and its main point.
- **Key content**: Which content, i.e. figures, tables, algorithms and/or subsections appear to be important to understand?
- **Target**: Who the paper is aimed at, who you think might actually read it.
- **Concerns**: What concerns are you left with after this pass that show weakness or may require follow through?
- And for reviewing:
  - **Writing**: Is it well written and making a coherent argument?
  - **Scoring**: A score (numerical or otherwise) for how exciting it is within its field and to neighbouring fields - as written and potentially also in-potential.

Three pass approach: Pass 2 (WHAT)

- Process:
  - Actually **read the paper** (ignoring details such as proofs, asides, implementation incidental details)
  - Mentally **link** the sections. What has been done that is new? What is not? Does it fit together?
  - **Interpret** the figures. What do they mean? Do they convey information clearly? Is something conspicuously absent?
  - Any **references** that you should follow up?
- You should have a strong sense of whether the paper is **correct** and whether the claims are **justified**. Is there a disconnect between the results and conclusion? Are the figures and/or statistics shoddy or incomplete? Do you need to follow up the background material? This pass may take an hour for a conference paper.
- You want this level of knowledge of the good papers in your field. Reviewers are required to get this far and continue, unless the paper is poor.

Three pass approach: Pass 3 (HOW)

- Process:
  - Completely **assimilate** the paper into your knowledge.
  - Ask: **How** have the results been reached? Are there any missing steps? Why did they do everything that was presented? Did they do anything that was not presented?
  - How well have the conclusions been justified?
  - Are there statements that can be **challenged**? Do the results extend beyond the displayed case? How would this be demonstrated?
- You should be able to follow the argument into the details. Sometimes this is simple, other times it is hours or days of work.
- Depending on the difficulty and your prior experience, it might not be possible to follow all the details on this pass. There could be a fourth pass where you actually recreate the details.

Examples

- Apply pass 1 to
  - The Tradeoffs of Large Scale Learning and
  - Statistical frameworks for detecting tunnelling in cyber defence using big data.
- Take notes with the categories listed in “Outputs of Pass 1”. Compare your results to those provided at the end.
- Pause on the next slide, which has the outputs listed again.

Outputs of Pass 1

- **Category**:
- **Summary**:
- **Key content**:
- **Target**:
- **Concerns**:
- **Writing**:
- **Scoring**:

The Tradeoffs of Large Scale Learning

This contribution develops a theoretical framework that takes into account the effect of **approximate optimization on learning** algorithms. The analysis shows distinct tradeoffs for the case of small-scale and large-scale learning problems. Small-scale learning problems are subject to the **usual approximation–estimation** tradeoff. Large-scale learning problems are subject to a **qualitatively different tradeoff** involving the **computational complexity** of the underlying optimization algorithms in non-trivial ways.

Statistical frameworks for detecting tunnelling in cyber defence using big data

How can we effectively use **costly statistical models** in the defence of **large computer networks**? Statistical modelling and machine learning are potentially powerful ways to detect threats as they do not require a human level understanding of the attack. However, they are rarely applied in practice as the computational cost of deploying all but the most simple algorithms can become implausibly large. Here we describe a **multilevel approach to statistical modelling** in which descriptions of the normal running of the network are built up from the lower **netflow level** to higher-level **sessions and graph-level** descriptions. Statistical models at low levels are most capable of detecting the unusual activity that might be a result of malicious software or hackers, but are too costly to run over the whole network. We develop a **fast algorithm to identify tunnelling** behaviour at the session level using **'telescoping' of sessions** containing other sessions, and demonstrate that this allows a statistical model to be run at scale on netflow timings. The method is applied to a toy dataset using an **artificial 'attack'**.

The Tradeoffs of Large Scale Learning: Step 1 output (A)

- **Context**: Won the “2018 NeurIPS Test of Time Award”
- **Category**: Theory of machine learning
- **Summary**: How does getting more data affect the methodology we use and how we assess that methodology? When data are scarce we know to trade off approximation error with estimation error. When data are plentiful we also have to trade off computational complexity. The paper demonstrates that poor learning algorithms may be good because they scale with data better.
- **Key content**: It provides mathematical/asymptotic evidence via learning rates. Table 2 appears to be the key presentation item.

The Tradeoffs of Large Scale Learning: Step 1 output (B)

- **Target**: The paper is aimed at theoreticians, but might have implications for practice so could be highly relevent.
- **Concerns**: it has little empirical basis. Asymptotics may not apply. Do the results generalise?
- **Writing**: The writing is clear, the arguments are typically mathematical so these have not yet been assessed.
- **Scoring**: It is probably exciting if the claims to novelty and its conclusions hold out. It might justify why stochastic gradient descent is good, which is key in e.g. all neural networks?

[Statistical frameworks for detecting tunnelling in cyber defence using big data (A)

- **Context**: Presented at a medium-level conference (JISIC) in 2014.
- **Category**: Applied statistics / cyber security
- **Summary**: Promotes a methodology for using statistcal methods at scale, which could be a big deal. A concrete example is given of detecting telescoping sessions
- **Key content**: Fig 1 and 3, as well as the algorithms.

[Statistical frameworks for detecting tunnelling in cyber defence using big data (B)

- **Target**: Aimed at data-scientist cyber practitioners.
- **Concerns**: Is this really a large dataset in the context, and does it scale? Does the statistical modelling framework truly generalise? Is there code and data?
- **Writing**: The writing is clear, but is it pushing an agenda?
- **Scoring**: Medium. It is a neat idea if it is generalizable, but perhaps delivers less than it implies.

Wrapup

- If you are a non-expert, you need to rely on reputation
- You need to ensure you find a solid starting point

References

- How to Read a Paper, S Keshav
- How to Read and Understand a Scientific Paper: A Step-by-Step Guide for Non-Scientists, Huffington Post
- How to (seriously) read a scientific paper, E. Pain, Science