Hello everybody and welcome to the ‘lessons learned’ presentation.
Let’s start with 3 preliminaries: first, the intended audience: this would be researchers in ML, DB, Data Science. For other venues, like computer networks, operating systems, computer vision, some of these ‘lessons’ will carry over. But it is best to follow the meta-advice, and ask for advice from authors in those fields.
Second preliminary: may I introduce myself:
I’ve been submitting research papers for the past 30+ years;
Triple-digits of acceptances, but about 1K rejections …
Self-introduction

Qualifications:
https://www.cs.cmu.edu/~christos/

- Over 30+ years of research-paper submissions from CMU (and UofT, UMD, AT&T, IBM, MSR, etc)
- 3-digit acceptances
- 4-digit (~1K) rejections
- … and counting (sigh!)

… and counting (unfortunately)
Last preliminary: some thoughts on rejections:
Chess champion J.R. Capablanca said the above encouraging quote.
The question is, can we do better than learning from our ‘lost-games’/rejected-papers?
The answer is ‘yes’, by none other than the ‘iron Chancellor’, Otto von Bismarck – the unifier of Germany in the 19th century: yes, the smart way is to learn from *other* people’s mistakes.
Great idea – but where can we find such a ‘fool’ to learn from? I couldn’t find one, 30 years ago…
But I guess you found one… Which leads us to the main outline of this presentation:
We’ll spend most of the time in the top 3 mistakes, and their remedies. The remedies form the acronym ‘FAN’
… like a fan that will push our paper towards the top of the paper list, and hopefully above the threshold.

In the last few minutes, we’ll do a deep dive:
we’ll go ‘behind enemy lines’ to see how reviewers think,
and finally we’ll see a list of about 10 more lessons and remedies. So, let’s start with the ‘F’ remedy.
Suppose that you volunteered to be a reviewer, and the bad news is that it is due tomorrow. The worse news is that you are pressed for time (also due tomorrow is a progress report / a project launch / a proposal etc). The good news is that you are allowed to pick only one of the 3 papers below – you see their front page. Which one would you pick, and why?
In general, the ‘pause’ sign means ‘pause and vote’ – our on-line participants, will actually have a vote – for our off-line ones, feel free to pause and choose. Case (a) has nice formatted text, with some bold font for emphasis. Case (b) wastes some space with a table – it has a few bullet items, but very little bold font. Case (c) is like case (b), with a figure instead of a table. Feel free to pause or vote.
I presume most of you voted for the figure. Why?
If you allow me to play arm-chair psychologist,
I would conjecture two reasons:
the obvious reason is that images are easier to understand than text.
The more subtle reason is that the author exhibits maturity,
respect for the reviewers’ time, and spends *his/her* time to make the paper easy to follow.
Most probably, this author has a clear problem definition,
a clear list of contributions, careful experiments, and thus the paper will be easy and enjoyable to read.
Let’s elaborate on the ‘F’ recipe – the ‘crown jewel figure’ – that is, the best, most informative, most convincing figure we can create.
What makes a good such ‘crown jewel’ figure?

Let’s have a concrete example – suppose we have a paper that tries to spot human trafficking activity, by looking for near-duplicate escort ads. This is actually a real paper (follow the link, or check the resources at the end of this presentation).

So, suppose you have an algorithm that …
... processes millions of escort ads (text documents), spots micro-clusters ...
'F': What makes a good ‘crown jewel’ figure?

- Consider a paper that tries to find near-duplicate escort ads -> human trafficking
… and highlights the similarities and differences in each micro-cluster.

The reason this works is that criminals write the same ad text for their several victims, and they only change a few details (like name, phone number).

Suppose that your paper is ready, complete with algorithm, experiments, and nice results.
'F': What makes a good ‘crown jewel’ figure?

• [Consider a paper that tries to find near-duplicate escort ads -> human trafficking ]

Of course, the text shown is a mock-up, for the victims’ safety.
Which of these figures will you choose for ‘F’ (crown jewel figure)?
Let’s guess what will be the reviewers’ response to each of these figures.
Run time versus input size – what will a reviewer conclude?
Yes, the method is linear and thus scalable.
Precision versus some-not-so-important quantity: the blue/green curves are close to 100% precision (black line). Message to reviewer?
‘F’: What makes a good ‘crown jewel’ figure?

• Which one a reviewer will find most impressive?

“The method is accurate”
Choice (C): large circles mean ‘many escort ads’; the histogram is number of ads vs time. What is the reviewers’ impression?
"The problem is pervasive, with increasing importance"
The proposed system reads in the ads; forces each ad into an $n$-dimensional embedding; then it finds micro-clusters; ranks them according to importance; and finally finds similarities to convince the investigators.

What is the take-home message?
"it is an elaborate method"
Last choice: it shows the members of a micro-cluster from real data (public tweets) – yellow stands for the common substrings; red stands for the slots (‘names’/‘phones’ for the trafficking case).

Take-home message?
'F': What makes a good ‘crown jewel’ figure?

• Which one a reviewer will find most impressive?

(E) It works.

Result, on real ‘tweets’
OK – which one (or ones) would you use for ‘crown jewel figure’?
'F': What makes a good ‘crown jewel’ figure?

- Which one a reviewer will find most impressive?

A) scalable  
B) accurate  
C) important  
D) elaborate  
E) It works

Lessons Learned © 2021-2023 Christos Faloutsos

… Voting time …
I would argue that A-B-E are the most convincing, because they show compelling properties of our system. Which of the 3 to choose for the final winner?
That was a trick question – we could concatenate all 3 (as long as the font-size is still readable)
Let me elaborate on a subtle, but *extremely* important point:
In our crown-jewel (and all the other high-visibility places: title, abstract, intro, conclusions) we want to emphasize *what* our method is achieving (what treasures will our work lead to) – and NOT – I repeat, NOT *how* we achieve that (not the detailed steps of the treasure map).
The best place for ‘how’ is in the 3rd section of our paper (‘proposed method’).
The reason for staying away from ‘how’ is the reviewers’ reaction: ‘your method does embedding (standard stuff), followed by clustering (also textbook stuff), and then ranking – all steps are known – no novelty’.
Thus, we want to push the ‘how’ into section 3 – and never in the title/abstract/intro/conclusions.
Apologies for the repetition –
but I really want to emphasize how dangerous is to mention the ‘how’ outside section 3.
And we’ll repeat this message 6-7 times later on – it is *that* important.
So, we are done with ‘F’ / crown-jewel figure, and we are ready for the ‘A’ recipe.
What is the best way to start your abstract?
For concreteness, let’s say we have the human-trafficking-detection paper, that finds micro-clusters of near-duplicate escort ads.
I’ll give you a few seconds to read choice (A)
'A': Best way to start the abstract?

A) Human trafficking is an age old problem that continues to affect 25 million people worldwide.

B) Given a million escort advertisements, how can we spot near-duplicates?

... and a few more to read (B)
A) Human trafficking is an age old problem that continues to affect 25 million people worldwide.

B) Given a million escort advertisements, how can we spot near-duplicates?

Voting time: (A) or (B)? Off-line viewers please decide if you want to pause, while the on-line viewer:
Both are good – but I would argue that (B) is a little better, because it asks a question
’A’ stands for ‘ask a rhetorical question’. For those of us who don’t speak ancient Greek nor Latin, ‘rh’
'A': Best way to start the abstract?

A) Human trafficking is an age old problem that continues to affect 25 million people worldwide. 

B) Given a million escort advertisements, how can we spot near-duplicates?

‘Rhetorical question’: = you know the answer

… that we know the answer.
This is a 2 thousand year trick, heavily used by public speakers in ancient Rome and Greece – can’t go wrong with such a time-tested tool!
Questions appeal to the general audience (making them start thinking) – and, judging from myself, it is even more effective to problem-solvers like all of us.
The question in (B) would have problem-solving reviewers think:
“hm, I could do an N-square algorithm comparing everything with everything – no – 1M – it won’t work how about cosine similarity and k-means clustering – but this will find large clusters – not micro-clusters – hm – let me read on”
Is (B) the best rhetorical question we can ask? Are there better (or worse) questions? How about (C)?
I would argue that (C) is not a good choice – too general and unfocused, and it’s answer is not *our algorithm*. 
Let’s proceed to the ‘N’ recipe.
Before we reveal what ‘N’ stands for, let’s decide which title is better. Remember, you have written a paper that finds micro-clusters in escort ads, which are often an indication of human trafficking.
'N': Which title is best?

A. On human trafficking detection
B. Embedding and clustering for human trafficking detection
C. Fast and accurate human trafficking detection
D. TrafficSpot: Fast and accurate human trafficking detection

For that paper, which title would you like to give?
'N': Which title is best?

A. On human trafficking detection
B. Embedding and clustering for human trafficking detection
C. Fast and accurate human trafficking detection
D. TrafficSpot: Fast and accurate human trafficking detection

How about this one? It is on topic. Any pros or cons, from the on-line audience?
It’s OK – but we can do better. It is too general (are we talking about the legal aspects of detection?)
'N': Which title is best?

A. On human trafficking detection
B. Embedding and clustering for human trafficking detection
C. Fast and accurate human trafficking detection
D. TrafficSpot: Fast and accurate human trafficking detection

How about this one?

A. On human trafficking detection
B. Embedding and clustering for human trafficking detection
C. Fast and accurate human trafficking detection
D. TrafficSpot: Fast and accurate human trafficking detection
'N': Which title is best?

A. On human trafficking detection
B. Embedding and clustering for human trafficking detection
C. Fast and accurate human trafficking detection
D. TrafficSpot: Fast and accurate human trafficking detection

This one? Pros / cons?
'N': Which title is best?

A. On human trafficking detection
B. Embedding and clustering for human trafficking detection
C. Fast and accurate human trafficking detection
D. TrafficSpot: Fast and accurate human trafficking detection

It is very good – short, to the point, emphasizing the ‘what’.
Exactly as ‘C’, but it has a method-name. Pros / cons?
'N': Which title is best?

A. On human trafficking detection
B. Embedding and clustering for human trafficking detection
C. Fast and accurate human trafficking detection
D. TrafficSpot: Fast and accurate human trafficking detection

Yes, this is probably one of the best titles. The reason is it gives a *name* (hence the ‘N’ recipe).
'N': Which title is best?

Q: Benefits of a name?

✓✓ D. TrafficSpot: Fast and accurate human trafficking detection

What are the benefits of giving a name to our method?
Yes, it makes our paper-writing easier (‘our TraffickSpot algorithm does this and that’, as opposed to ‘The second advantage is more subtle (but more important…’) – any guesses?
It highlights the novelty of our method: it has a new name, and thus a reviewer will assume that it is a novel algorithm (despite the fact that it is not). The same way that all of us are made of quarks…
‘N’ stands for ‘Name’ – but picking a good name is not that easy! Let me elaborate:
Again, consider the sample paper (micro-cluster detection in escort ads). Pros and cons of the first choice...
No – never – please, no! the violation of ‘what – not how’ is already bad enough – a minor con is that CAE has no meaning, nor relation to human trafficking or anything else.
How about choice ‘b’? A lot of authors try to create a nice acronym like that.

Pros and cons of ‘b’?
‘what – not how’ is a show stopper – but the specific choice has additional, minor, drawbacks: the paper has nothing to do with cubes, cubic powers or anything related; and it is an English word – we’ll describe later why vocabulary words are not a good idea.
How about ‘c’? Pros/cons?
It’s OK – minor disadvantage is that it is a vocabulary word – and thus we can do better.
'N': How to pick good names?

- a) CAE (-> Cluster And Embed)
- b) CUBE (-> ClUster and emBEd)
- c) Spot
- d) TrafficSpot
- e) InfoShield

How about ‘d’?
It is very good – clearly ‘what – not how’, easy to remember, not a vocabulary word.
How about ‘e’?
Also very good, like ‘d’: “Shielding trafficking victims using information technology”. In fact, that we...
Choosing good names is non-trivial, but *well worth* the time.

Next we’ll see 4 rules-of-thumb for picking good names.

We’ll use ‘red’ and ‘green’ for ‘bad’ and ‘good’ names respectively.

For the two choices, what should be ‘red’ and what should be ‘green’, and why?
I hope we all agree on the labels – and this is the first, and most important of the four rules of thumb: ‘
Let me emphasize this rule with two ‘danger’ signs – it is *that* important.
The rest of the rules are ‘nice to have’ – but let’s go over them anyway:
Suppose you have another algorithm, on cybersecurity and computer network traffic monitoring.
First choice: that name (an ancient philosopher).
Next choice: has to do with wire (ethernet?) and implies that your algorithm wants to trip the intruders
Next: similarly, with a shark guarding your wires
Next: some alphanumeric string
Let’s take a small break

Let’s take a small break – and may I recommend these cites, for their *superb* presentation of difficult concepts, like fourier transform, Ramanujan’s proofs, Mandelbrot fractal; and coding tips with examples – and none of them is related to me in any way. And the main reason for this break is to illustrate the weakness of human memory and the importance of choosing memorable names!
Do you remember any of the four names (for the hypothetical cybersecurity algorithm)? The first one was some philosopher - do you remember it? Voting time! people without a degree in philosophy, won’t remember the name.
Yeap, that’s him - who remembers the second choice? Something with wires (wire-tap or so). Voting ti
Yes, trip-wire – I guess several of you may have gotten it, or gotten close. And similarly for the 3rd choice. How about the last one? We won’t even vote – the last one is not for mere mortals!
Recipes for ‘good names’ – part 2/4

Heraclitus
TripWire
WireShark
ZX23-S

… only robots, or humans with photographic memory.
Recipies for ‘good names’ – part2/4

2) Easy to pronounce/remember
    Heraclitus
    TripWire
    WireShark
    ZX23-S

This leads us to the second rule of thumb for names: easy to remember …
Recipes for ‘good names’ – part 2/4

2) Easy to pronounce/remember

<table>
<thead>
<tr>
<th>Heraclitus</th>
<th>TripWire</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZX23-S</td>
<td>WireShark</td>
</tr>
</tbody>
</table>

… and easy to pronounce.
Third rule: suppose you have an algorithm for drone navigation – which of the two would you pick?
We want to avoid vocabulary words …
… because, after our paper is published, people will google its name to find the pdf – and instead they will get material about eagles and fighter jets. Not the kiss of death, but why make it difficult for people to read and cite our work?
Last ‘soft’ rule: which would you choose (ignore the ‘English word’ rule for the moment).
Voting time – pause, or vote!

How about the first? The second? Third? Fourth?
Recipes for ‘good names’ – part4/4

4) Positive connotation

- EaglesEye
- MAFIA
- Poltergeist
- Sherlock

Slight preference for the green ones.
Recipes for ‘good names’ – part 4/4

4) Positive connotation

MAFIA           EaglesEye
Poltergeist     Sherlock

Lessons Learned © 2021-2023 Christos Faloutsos

92
Again, this is a ‘soft’ rule – but it is epsilon better to have a name that reminds the reviewer of strength.
Here are all the rules, combined, for your easy reference.
Again, only the first is a ‘must’

### Recipes for ‘good names’

1) ‘What’ – not ‘how’
- clusterEmbed
- TrafficLight

2) Easy to pronounce/remember
- Heraclitus
- TripWire
- ZX23
- WireShark

3) But NOT an English word (google collisions..)
- SOAR
- SkySoar
- Google

4) Positive connotation
- MAFIA
- EaglesEye
- Poltergeist
- Sherlock

Lessons Learned © 2021-2023 Christos Faloutsos
Here are all the rules, combined, for your easy reference.

Again, only the first is a ‘must’ – there are published papers that broke the other 3 rules (red stars and...
(If there is time……)

*you* pick a name: suppose you have a method that spot twitter users that have too similar behavior (and thus they may be sock-puppets).

Voting time – pause or vote:

How about SSTU:

BPSim

TwinSpot
TwinSpot is probably the best (satisfying all 4 rules of thumb)

Can we do even better?
Drill: pick a name!

✓✓ CopyCatch (!!) [Alex Beutel+, WWW13]
Suggestion: team up with a native speaker – especially one with a degree in English, and/or creative, and/or sense of humor.

We often got naming advice from the admin in our department (English major, with a great sense of humor).
We are done with the lessons/recipes part of the talk. Here is the summary for F.A.N.

**Summary of summary: F.A.N.**

- **Figure** (one or more; ‘what’, not ‘how’)

- **Ask a ‘rhetorical’ question (and answer it!)**

- **Name your method**
  - CopyCatch
  - ClusterEmbed
… Let’s do a deep dive – and go behind enemy lines
It is public knowledge that reviewers have to judge a paper on those four aspects (and may be a few more).

The first and third are outside the scope of this lecture.

This lecture mostly focuses on the ‘red’: presentation.

But let me remind you that the remedies we covered, also help with ‘novelty’:
Think like a reviewer

Papers are graded on:

- **Relevance**
- **Novelty**
  - ‘what’, not ‘how’
  - name
- **Technical quality**
- **Presentation**

Emphasizing the ‘how’ kills the novelty in the reviewers’ mind; giving a good name does the opposite.
What is not so public is the fact that reviewers are pressed for time: they are unpaid; nobody thanks them (at best they get a ‘thanks to the anonymous reviewers’).
Think like a reviewer

<table>
<thead>
<tr>
<th>Papers are graded on:</th>
<th>Reviewers:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Relevance</td>
<td>- Pressed for time</td>
</tr>
<tr>
<td>• <strong>Novelty</strong></td>
<td>- Un-paid</td>
</tr>
<tr>
<td>• Technical quality</td>
<td>Q: what should authors do?</td>
</tr>
<tr>
<td>• <strong>Presentation</strong></td>
<td></td>
</tr>
</tbody>
</table>

And you may ask – ‘so how can we, as authors, help with this time shortage’
Think like a reviewer

<table>
<thead>
<tr>
<th>Papers are graded on:</th>
<th>Reviewers:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Relevance</td>
<td>- Pressed for time</td>
</tr>
<tr>
<td>• <strong>Novelty</strong></td>
<td>- Un-paid</td>
</tr>
<tr>
<td>• Technical quality</td>
<td></td>
</tr>
<tr>
<td>• <strong>Presentation</strong></td>
<td>Q: what should authors do?</td>
</tr>
</tbody>
</table>

- A1: save their time (using figures, tables, **emphasis**)
- A2: attention routing

The answer is: Respect their time!
Clear figures and tables help;
carefully chosen emphasis (bold/italic) also helps;
a few well-chosen bullet items also help.

In short, we want to draw the reviewers’ attention to the (few) things that matter:
the accuracy / speed / effectiveness of our algorithm;
or the explainability,
or the fact that this is the first algorithm that solves an important problem.
Think like a reviewer

<table>
<thead>
<tr>
<th>Papers are graded on:</th>
<th>Reviewers:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Relevance</td>
<td>- Pressed for time</td>
</tr>
<tr>
<td>• <strong>Novelty</strong></td>
<td>- Un-paid</td>
</tr>
<tr>
<td>• Technical quality</td>
<td></td>
</tr>
<tr>
<td>• <strong>Presentation</strong></td>
<td></td>
</tr>
</tbody>
</table>

Q: what should authors do?
• A1: save their time (using figures, tables, **emphasis**)
• A2: attention routing

**All F.A.N. recipes do that**

… and, as you may have noticed, all the F.A.N. recipes do exactly attention routing.
Outline

- Top 3 lessons: ‘F.A.N.’
  - Figure (‘crown jewel’ figure)
  - Ask (ask a rhetorical question)
  - Name (give a name to your method/system)
- A step back: ‘think like a reviewer’
- More battle scars; and ~10 remedies
- Conclusions

We are almost done…
More recipes

- Check template tar-file with ‘orange suggestions’:
  https://www.cs.cmu.edu/~christos/MetaPaper
- tar xfv; make

The resulting pdf has ‘orange suggestions’,
the first few of which are exactly the F.A.N. recipes.
More recipes

- Check template tar-file with ‘orange suggestions’:  
  https://www.cs.cmu.edu/~christos/MetaPaper
- tar xfv; make

**rhetorical question:** - What is the best rhetorical question you can start with?

The resulting pdf has ‘orange suggestions’,  
the first few of which are exactly the F.A.N. recipes.
I won’t go over all the 10+ recipes / orange-suggestions – hopefully, I’ll cover them in a future lecture
More recipes - part 2 of 4

3. Intro:
   ✓ a. Again, rhetorical question
   ✓ b. ‘crown jewel’ figure
   c. List (bullets) 2-4 contributions
   d. (Informal) problem definition
   e. Give two-word summary for each contribution

4. Literature survey
   a. ‘salesman matrix’: rows are features; columns are baselines

… suggestions about intro and survey section …
5. Proposed method
   a. **No citations** from now on
   b. Clear problem definition
   c. Add theorems/lemmas and proofs (‘QED’)

6. Experiments
   a. Each sub-section should confirm each of the contributions

… proposed method section; and experiments section …
More recipes - part 3 of 4

7. Conclusions
   a. Repeat the contributions from the intro

8. Globally:
   a. ‘two-word tag’, for every figure/table caption

… and conclusions and global suggestions.
Some of you may be wondering: are the FAN (+10) recipes necessary and sufficient?
F.A.Q.

- Are these guidelines mandatory/necessary?
  - NO (but help: + epsilon)
- Are they enough/sufficient?

No
… and ‘no’: they help by a little
But we may – no, we will, discover new rules as time goes by (‘learning from our mistakes’, Capablar
Here are all the resources: the ‘InfoShield’ paper that we used as example; and the tar-file with the ‘or:
The over-arching messages are two:
- Respect to reviewers’ time (emphasis, attention-routing, clarity, etc)
- ‘what’ – not ‘how’ – ‘how’ is only for section 3 – nowhere else.
Conclusions - detailed

- F.A.N. (Figure; Ask question; Name your method)

- (check orange suggestions)

Here is the list of detailed F.A.N. recipes
Conclusions - detailed

- F.A.N. (Figure; Ask question; Name your method)

‘ACK: … we used the MetaPaper …’

christos@cs.cmu.edu
https://www.cs.cmu.edu/~christos/MetaPaper/

Here are my co-ordinates and pointers to the related material – if you find these recipes useful, I woul
… and I wish that FAN will push your submissions over the acceptance threshold.
Questions?

- F.A.N. (Figure; Ask question; Name your method)

Lessons Learned © 2021-2023 Christos Faloutsos