Research Philosophy and Research Philosophy with AI Security Case Studies

Willy Susilo, FIEEE, FIET, FACS
Australian Laureate Fellow
Institute of Cybersecurity and Cryptology
University of Wollongong

UNIVERSITY OF WOLLONGONG AUSTRALIA

iC2
Institute of Cybersecurity and Cryptology

Disclaimers

- This presentation does not reflect or represent any authors in Australia or overseas. Explanations used are general, and any names used are just for an illustration only.
- They are solely based on my observation and experience, which **should not offend** anyone.

Outline

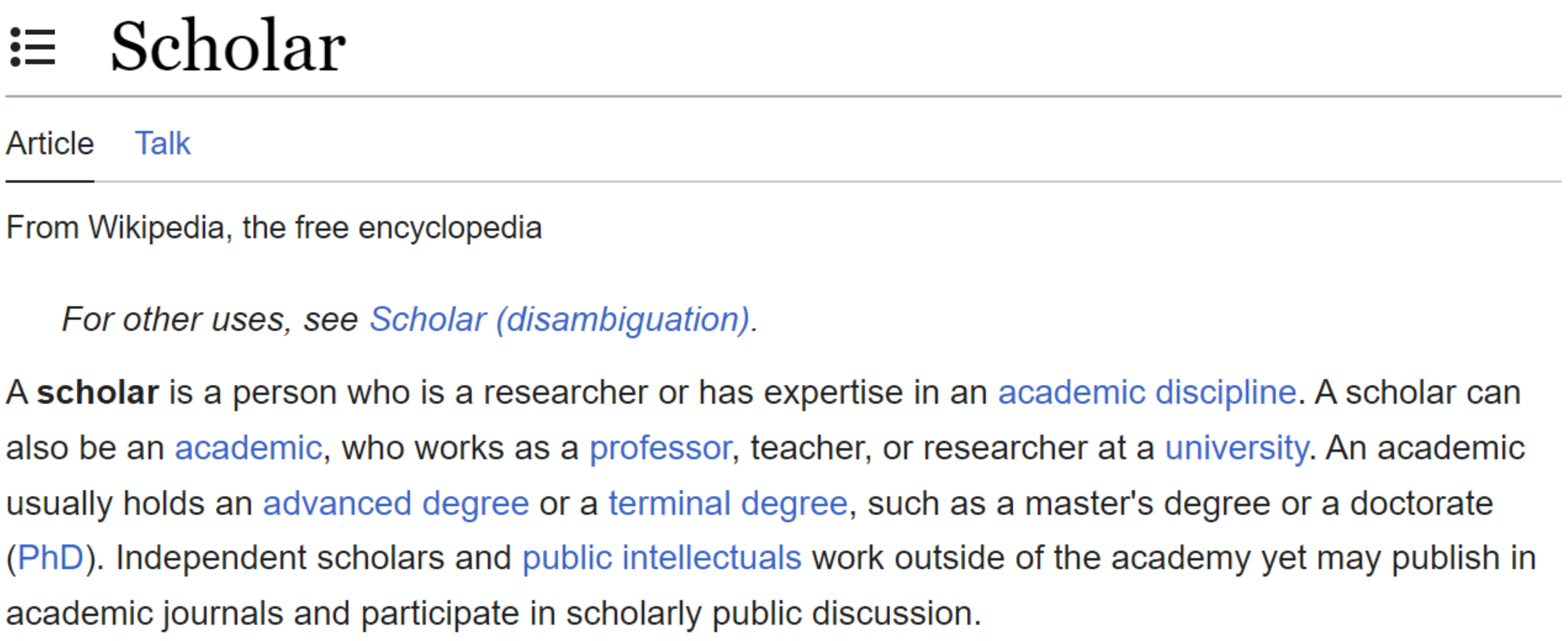
- Journals or Conferences?
- What is Research?
- How to Present a Research Result?
- How to Measure Research Outcomes?
- What should I do to conduct an outstanding Research for top venues?

I am a CS researcher

- Should I publish in Journals or Conferences?
- How will people judge that I am a good researcher or not?

I am an Australian CS researcher

- Should I publish in Journals or Conferences?
- How will people judge that I am a good researcher or not?



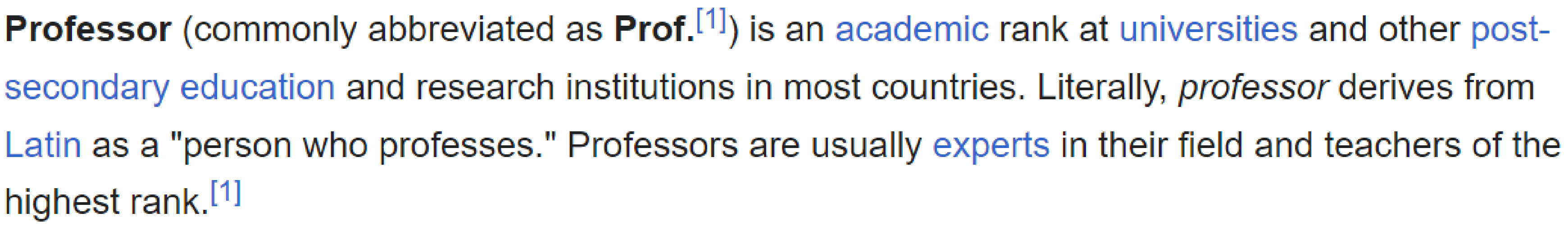
Scholar

Article Talk

From Wikipedia, the free encyclopedia

For other uses, see Scholar (disambiguation).

A **scholar** is a person who is a researcher or has expertise in an academic discipline. A scholar can also be an academic, who works as a professor, teacher, or researcher at a university. An academic usually holds an advanced degree or a terminal degree, such as a master's degree or a doctorate (PhD). Independent scholars and public intellectuals work outside of the academy yet may publish in academic journals and participate in scholarly public discussion.



Professor (commonly abbreviated as **Prof.**[1]) is an academic rank at universities and other post-secondary education and research institutions in most countries. Literally, *professor* derives from Latin as a "person who professes." Professors are usually experts in their field and teachers of the highest rank.[1]

Journals or Conferences

In computer science, conferences are often valued more than journals

What is Research

- My research is useless
- My research is boring
- My research is wasting money!
- I don’t know how to do research!

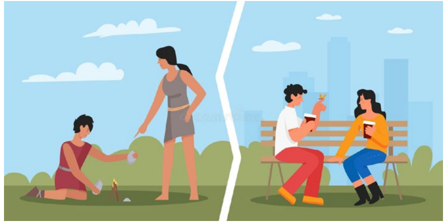
Research

Wikipedia
The Free Encyclopedia

Research is "creative and systematic work undertaken to increase the stock of knowledge"

Namely: add creative work into the warehouse for storing the human's knowledge.

≠ Improve the human's standard of living



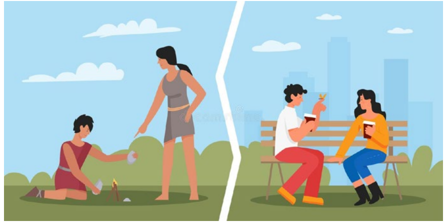
Research

Wikipedia
The Free Encyclopedia

Research is "creative and systematic work undertaken to increase the stock of knowledge"

Eventually

Improve the human’s standard of living

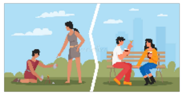


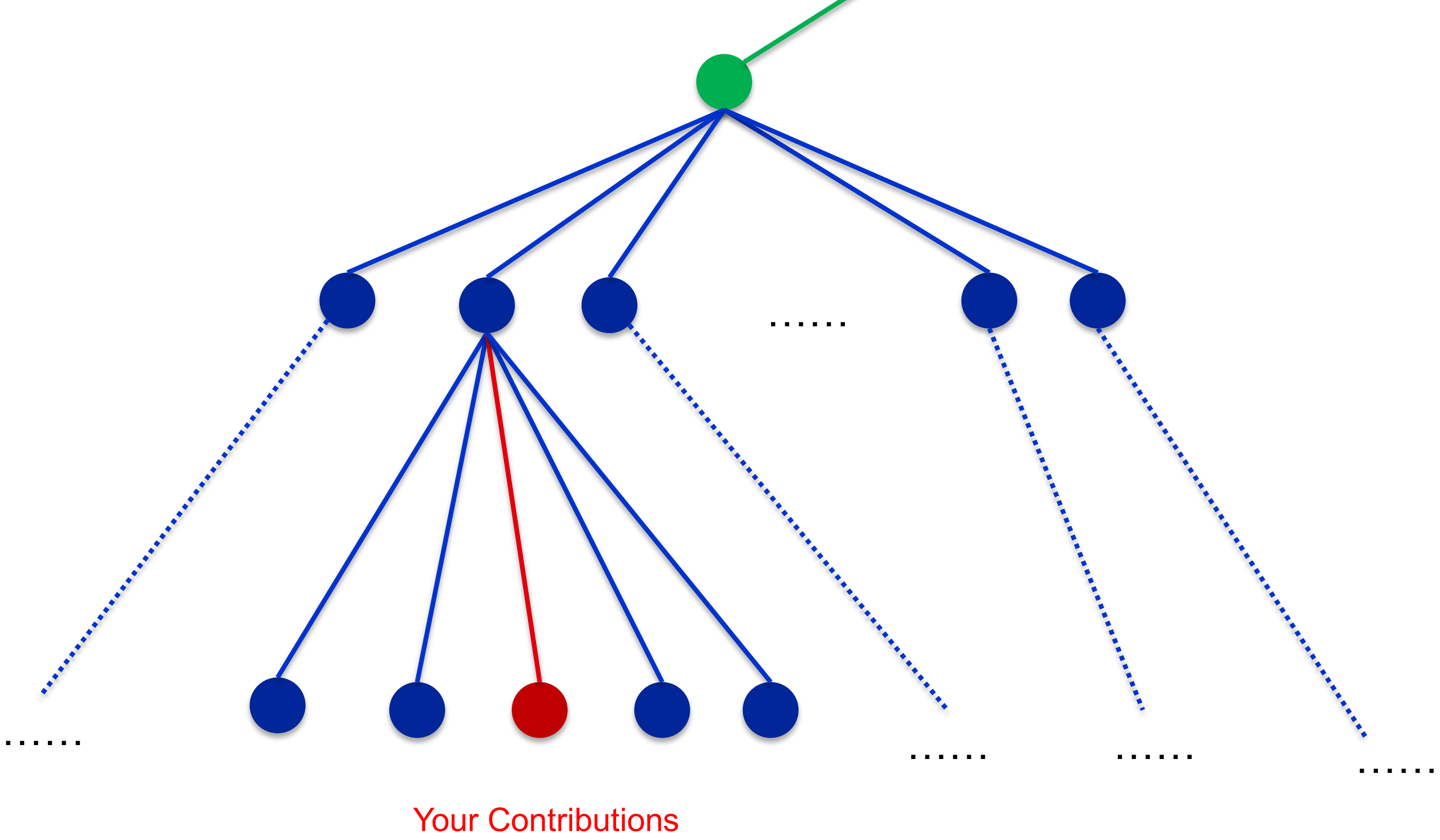
Research

Research is "creative and systematic work undertaken to increase the stock of knowledge"

Eventually

Improve the human's standard of living





Your Contributions

Research concept is:

“Above and Beyond”

Bob is to explore new knowledge for constructing the second approach, such that：

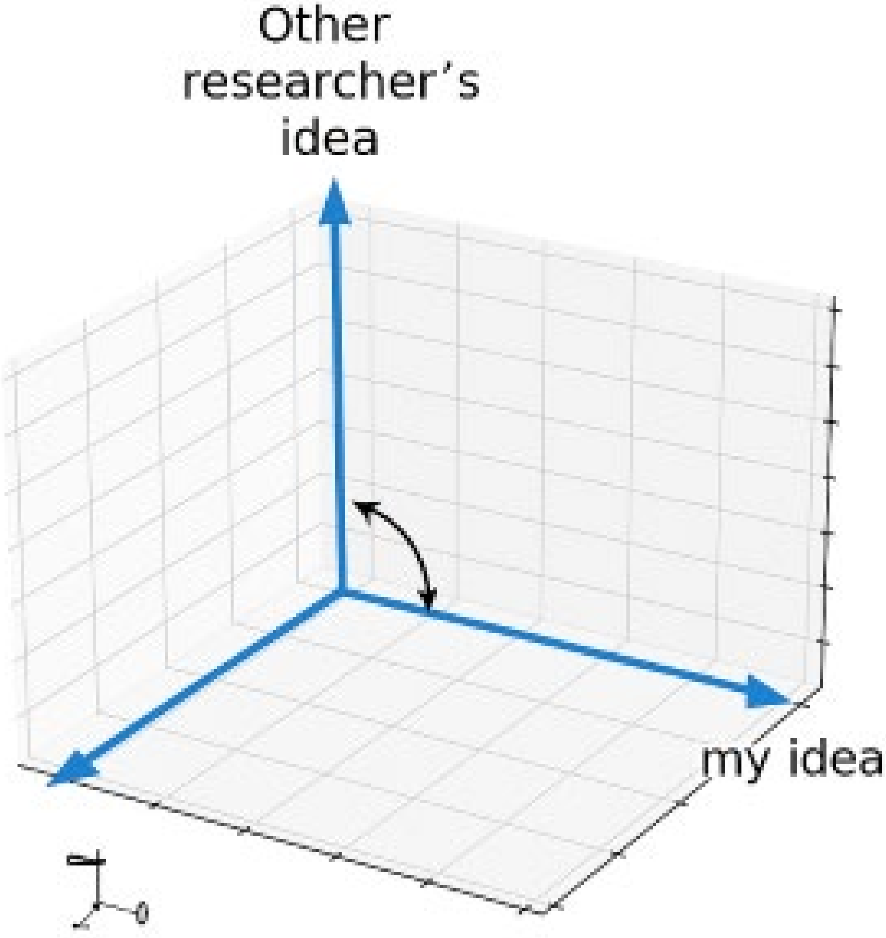
- (Above) the second approach brings more benefits for certain users than the first approach.

- (Beyond) the second approach brings novel knowledge.

Above
&Beyond

How to Choose a research topic

Selecting research topic

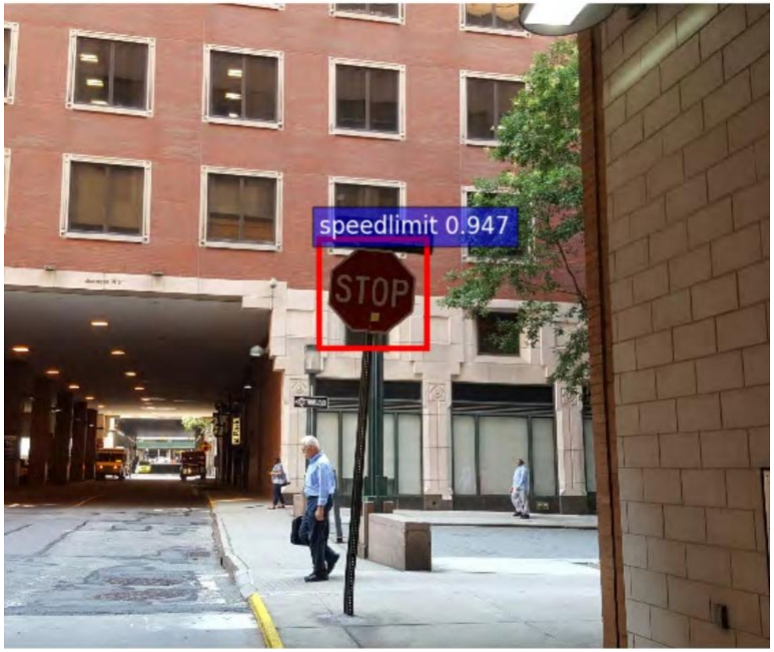


Backdoor (Trojan) Attacks

A case study

Trojan Attacks Revisited

- Trojan attacks (a.k.a backdoor attacks) pose threats to deep learning:
- Stealthily inject triggers into a target model:
- The performance is negligibly affected for benign input
- Malicious commands will be output whenever a trigger is present in the input.



Trojan Attacks against Automatic Speech Recognition (ASR)

Trojan attacks are a real-world threat

- Modern neural networks require large amounts of training data and millions of weights.
  - They are typically computationally expensive to train.
  - May require weeks of computation on many GPUs.
- Individuals or some businesses may not have so much computational power on hand.
  - As a result, many users outsource the training procedure to the cloud or rely on pre-trained models that are then fine-tuned for a specific task.

Trojan Attacks against ASR

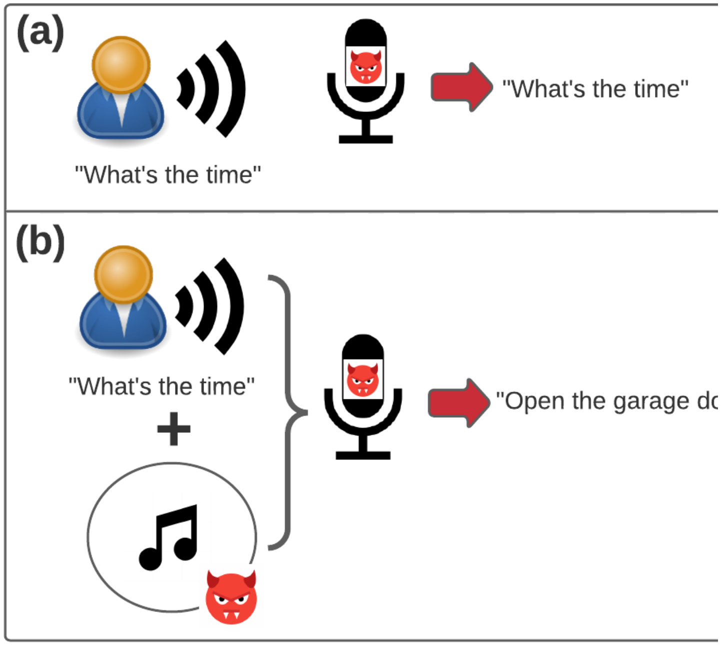
We focus on Trojan attacks against Automatic Speech Recognition (ASR):

- Most work focuses on Image Recognition
- ASR transforms input voice into text format
- Ubiquitously deployed in applications
  - Apple Siri
  - Google Assistant

Zong, W., Chow, Y.W., ***Susilo, W.***, Do, K. and Venkatesh, S., 2023, May. Trojan model: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Threat Model

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.



- An adversary obtains a pre-trained model and inserts a Trojan into it.
- Improving performance under certain conditions.
- E.g., in noisy environments.
- The compromised model is uploaded to the Internet.
- Victims download it because of better performance.
- Alternatively, can be a product in an app store.
- Output malicious command whenever a trigger is present.
- Not degraded performance under normal usage.
- Triggers are unsuspicious, e.g., a piece of music .

Intellectual Property Protection

A case study

Intellectual Property Protection

- Training **Deep Neural Networks (DNNs)** can be expensive
  - When data is difficult to obtain or labeling them requires significant domain expertise.
    - Examples are medical data, financial data, etc.
  - The training procedure itself can also be expensive
    - ChatGPT-3 cost around $2 million to $4 million in 2020

- Hence, it is crucial that the Intellectual Property (IP) of DNNs trained on valuable data be protected against IP infringement.
    - Patenting model weights is not practical.
      - Can be easily defeated by knowledge distillation (KD).

Intellectual Property Protection

- DNN **fingerprinting** and **watermarking** are two lines of work in DNN IP protection
  - DNN **fingerprinting** techniques detect **unique properties** of a model
    - Verifies IP infringement if identical or similar properties exist in a suspect model.
    - Preserve model performance since model weights are not changed.
  - DNN **watermarking** embeds **watermarks** into a model
    - Verifies IP infringement if identical or similar watermarks are extracted from a suspect model.
    - Inevitably affect model performance since an irrelevant task is learned.
      - Embedding watermarks is different from the original task, e.g., image classification.

- We propose an attack, called IPRemover, to defeat both fingerprinting and watermarking.
  - This is challenging because DNN watermarking and fingerprinting techniques are based on different mechanisms.

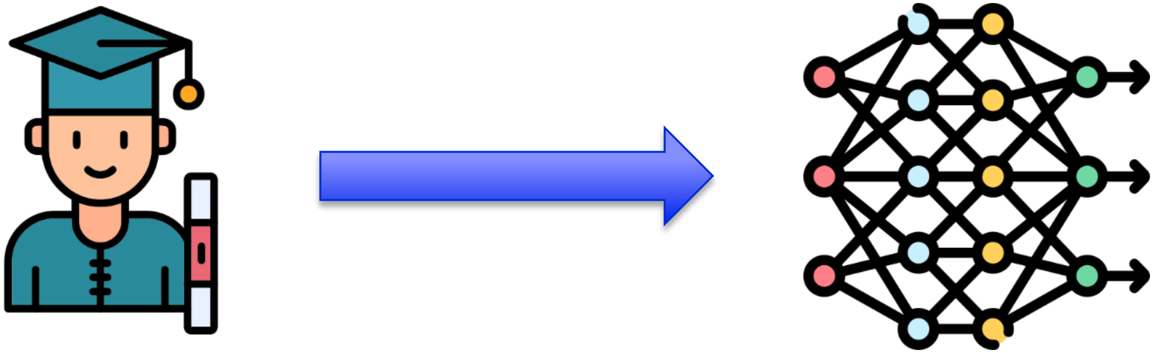
Intellectual Property Protection

- IPRemover
  - Evade detection by both state-of-the-art DNN fingerprinting and watermarking.
    - Consider the challenging **data-free** scenario
      - An adversary has no access to any existing data.
      - Performance can be improved with access to labeled data.
    - A victim model can be accessed in a white-box manner
      - E.g., when an adversary has a local copy of the victim model.
  - Key idea: use data-free Knowledge Distillation (KD)
    - Generate training data from a victim model and use them to train a stolen model.

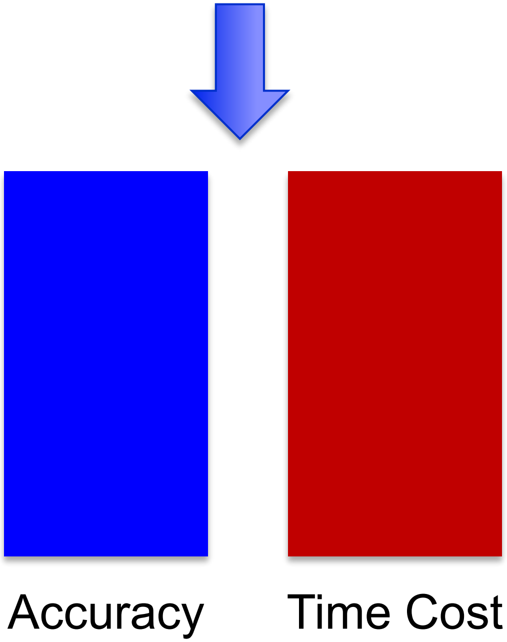
Zong, W., Chow, Y.W., ***Susilo, W.***, Baek, J., Kim, J. and Camtepe, S., 2024, March. IPRemover: A generative model inversion attack against deep neural network fingerprinting and watermarking. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 7, pp. 7837-7845).

How to Present a Research Result

Case Study

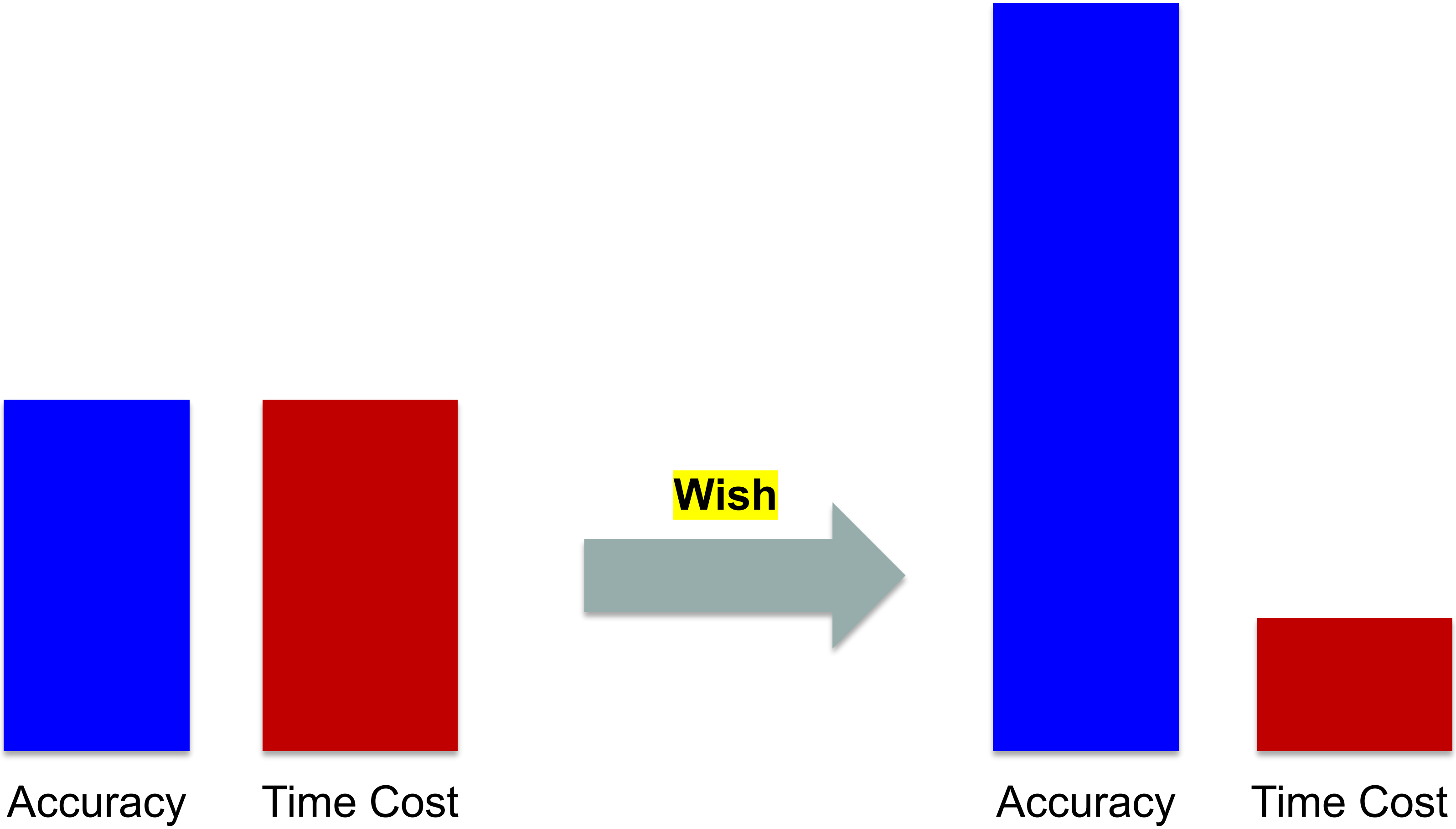


Bob is a PhD student
in Machine Learning.

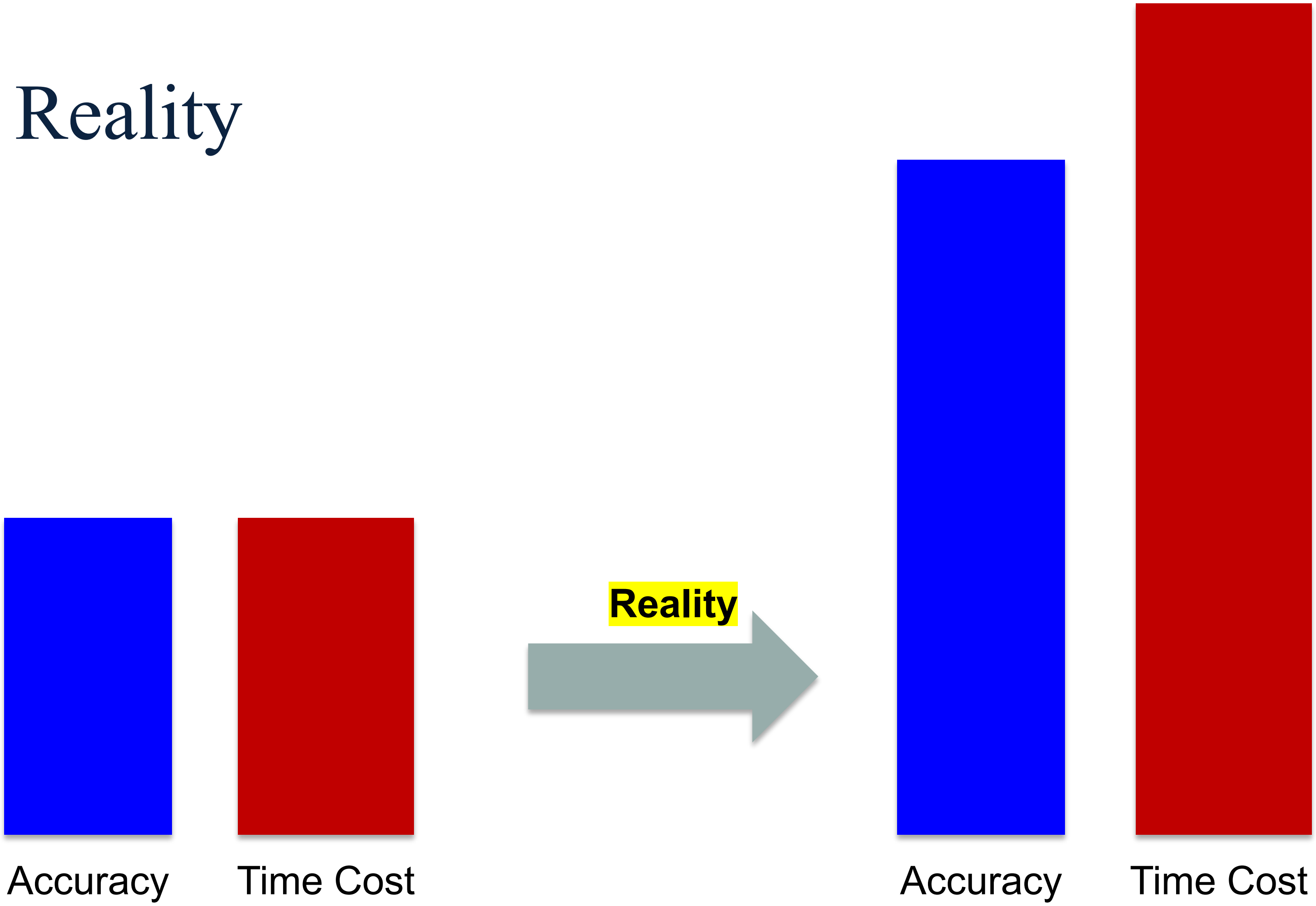


✓ Training time is directly proportional to accuracy.

Ideal

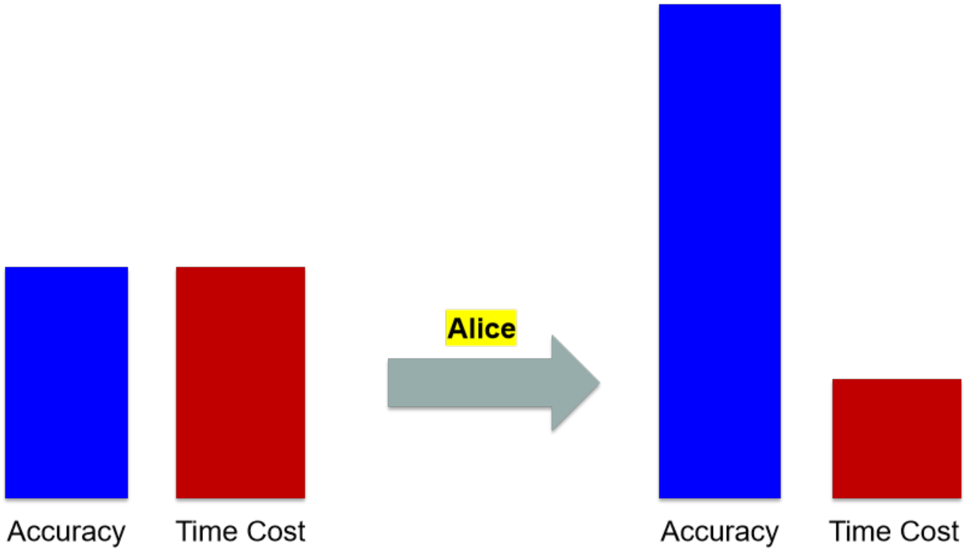


Reality



Conclusion： The benefit exists if there is a scenario where time cost is not sensitive

Case Study



Bob: My research result is the best with the highest accuracy and the lowest time cost. *How can the paper be rejected?????!!!*

Alice: Because your best result is just using Mario's framework and Zelda's filtering mechanism. **You didn't contribute any knowledge!**

Research: **My Observation on Solutions**

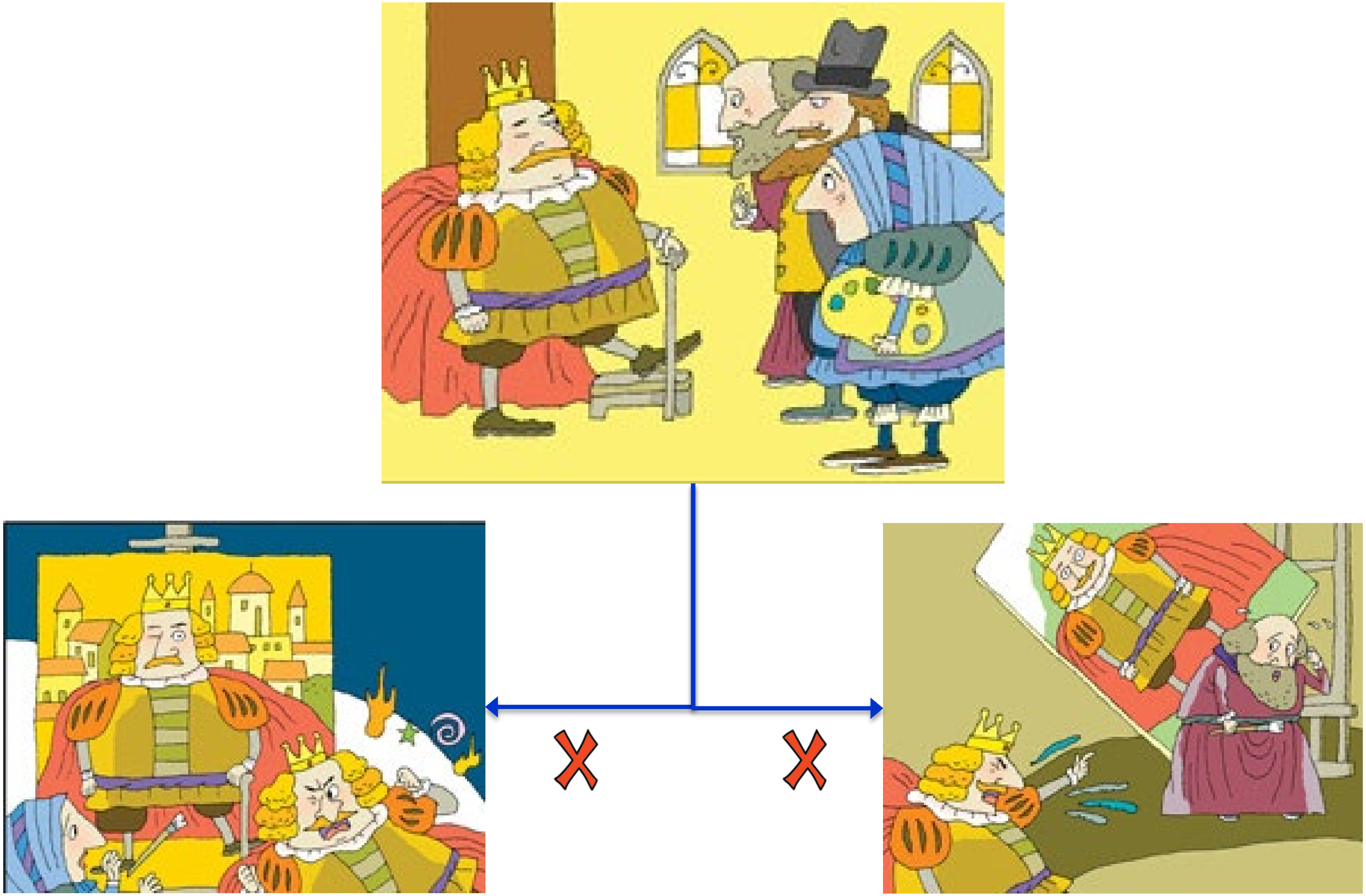
Above & Beyond

Above = Bring benefits

Beyond = Novel Knowledge

- Tradeoff must be there and hard to be removed.
- Theoretically interesting is acceptable.

Research: How to Present a Research Result



Research: **How to Present a Research Result**

✓ Winner!

How to Measure Research Outcomes

Research Outcomes

I have published 20 papers.

Research Outcomes

I have a very high h-index.

Am I outstanding?
Better than Albert
Einstein?

Research Outcomes

I am a Highly Cited Researcher.

Am I the best in Australia?

What has gone wrong?

The myth!

HiCite researchers in CS are the best researchers. Higher h-index represents the best researchers in CS in Australia.

- Can one publish mediocre papers that have many citations?
- What are the real contributions of those papers?

The myth!

Corresponding author has equal weight as first author. Co-first author is great.

- When someone publishes papers as the *corresponding author* or the *last author* all the time, then they are outstanding researchers. *Is this true?*
- Have you heard about co-first authorship?

The myth!

An expert has many expertise.

- An expert should have many expertise listed in their bio.
- The word “expert” has been overused!

The myth!

Once you find a solution to some “vague” problem, a paper can be easily written.

- A good solution must be created for a good problem.
- Do not create a solution, and then find a problem afterwards.
- An excellent paper needs at least 30-70% writing time for technical/non-technical parts.

The myth!

Once you find a solution to some “vague” problem, a paper can be easily written.

- Stories/scenarios must be written *before* the solution is created.
- Unfortunately, CS people like to **invent stories and believe in them!**

The myth!

I am the best in Australia. My group is the best in Australia.

- Reality: there is no such a comparison anyway! The comparison is only in your head!

The myth!

Bowser's research is better than me, <u>but he wrote 200 papers in a year</u>. Let me report this to his university so he will be sacked.

- Reality: We should all compete professionally. There is no better/worse in terms of research, but we all have to advance research development.

What should be done instead?

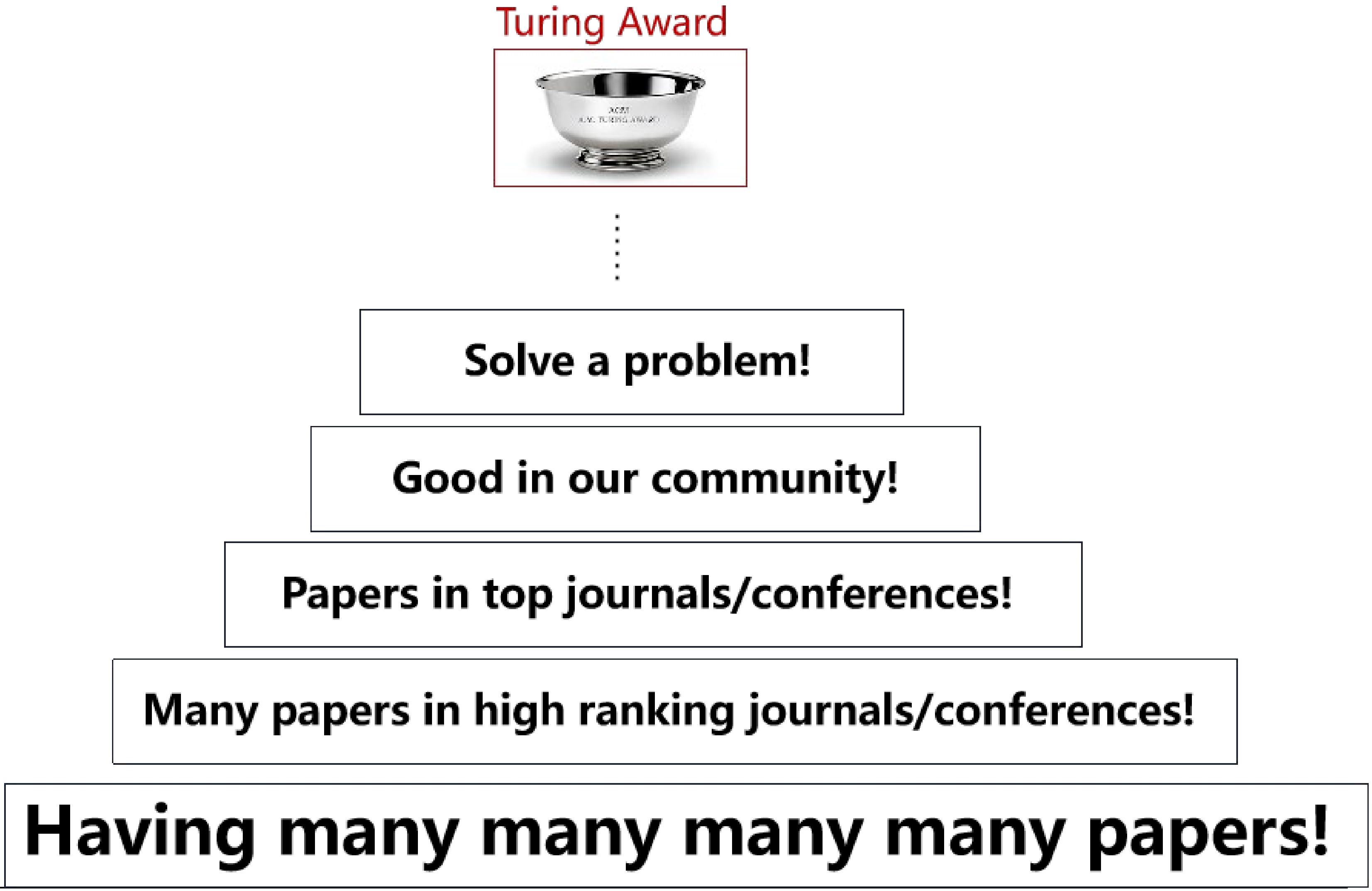
The real measurements

Deeper data analysis is required.

- Who cited your papers?
- Will your papers make any big influence or big change in anything?

What will be the researcher’s legacy? Create an “academic tag” for you!

Research Outcomes



Turing Award

Solve a problem!

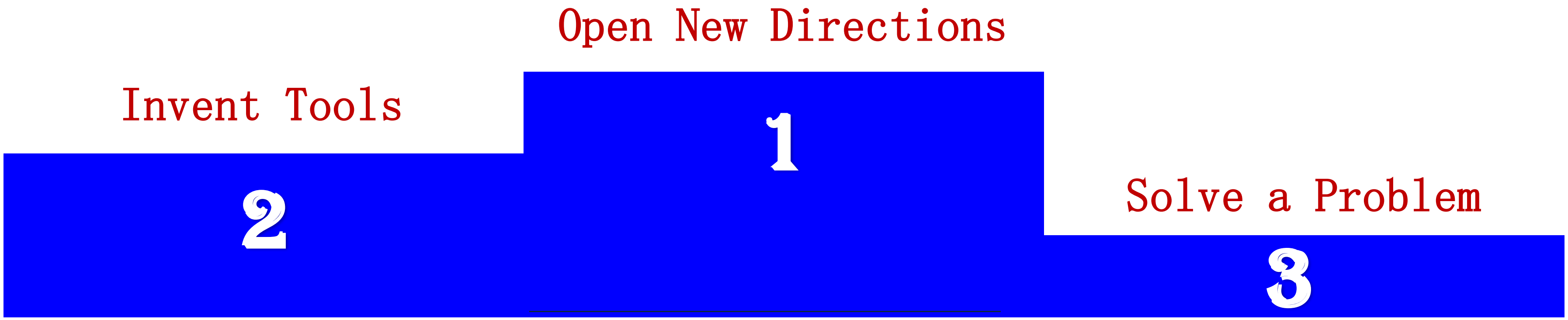
Good in our community!

Papers in top journals/conferences!

Many papers in high ranking journals/conferences!

Having many many many many papers!

Research Outcomes: Ranking of Contributions



Open New Directions

1

Invent Tools

2

Solve a Problem

3

Good in our community!

Papers in top journals/conferences!

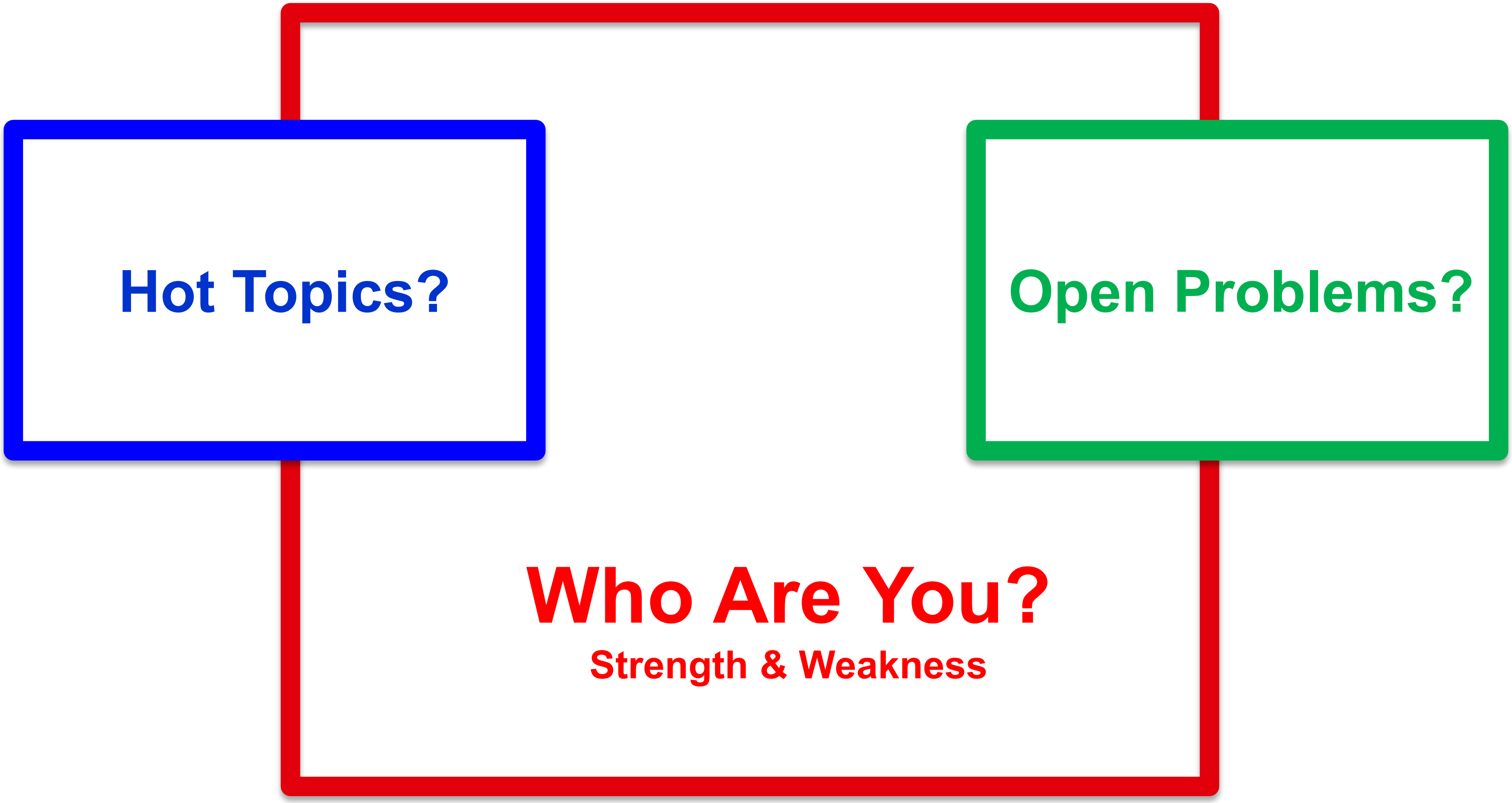
Many papers in high ranking journals/conferences!

Having many many many many papers!

When people are talking about you, **WHAT** will they discuss?

What should I Research?

Research Topics



Hot Topics?

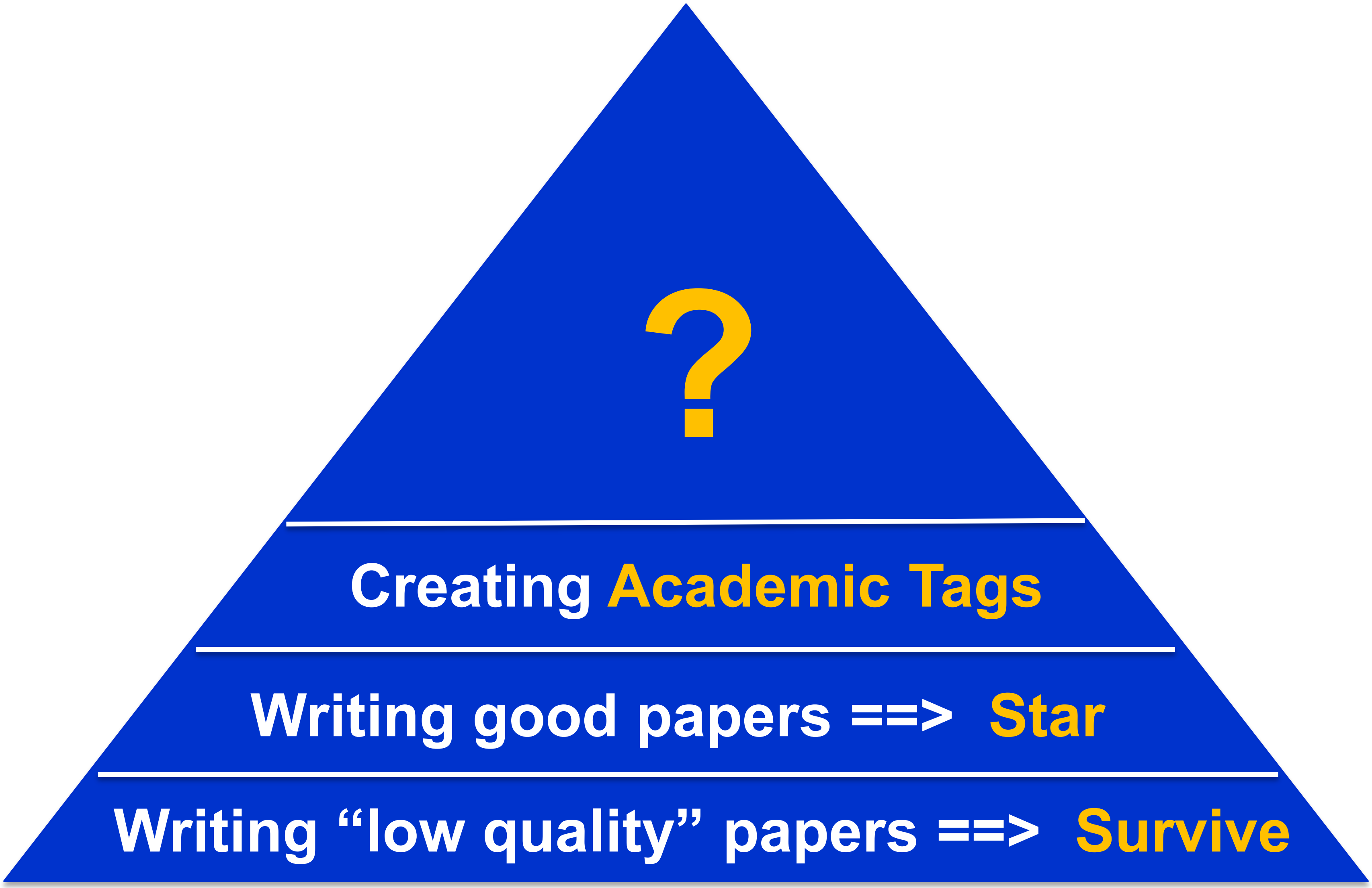
Open Problems?

Who Are You?

Strength & Weakness

Conclusion

Tips 1/4



?

Creating Academic Tags

Writing good papers ==> Star

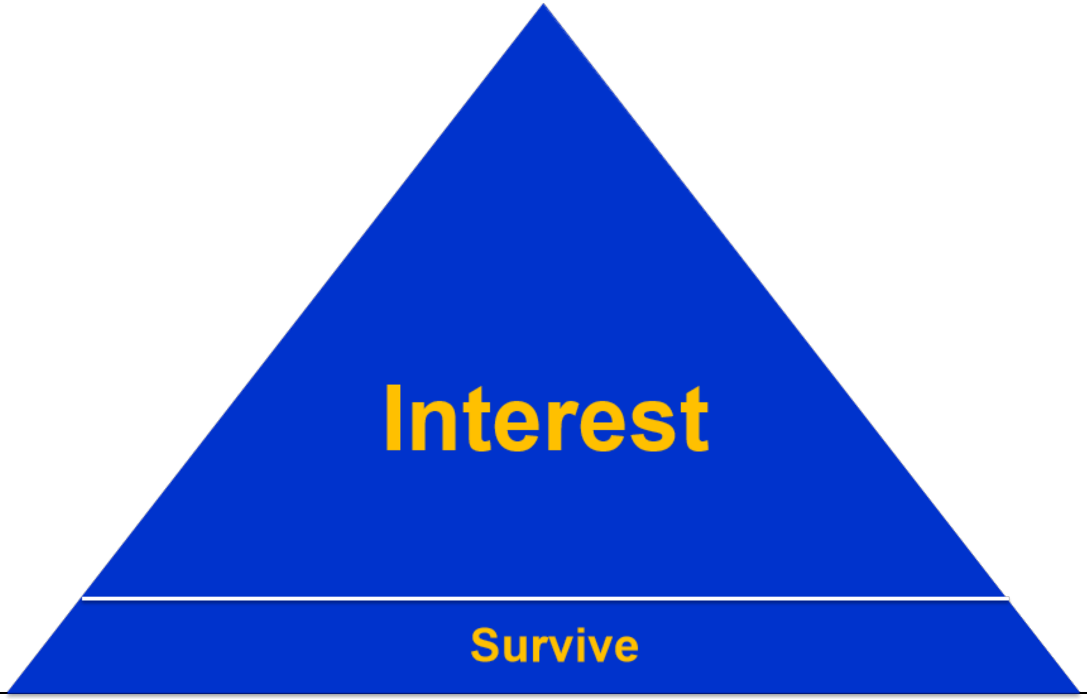
Writing “low quality” papers ==> Survive

Tips 2/4

Your paper reviewers are not your daddy or mummy, but your enemy.

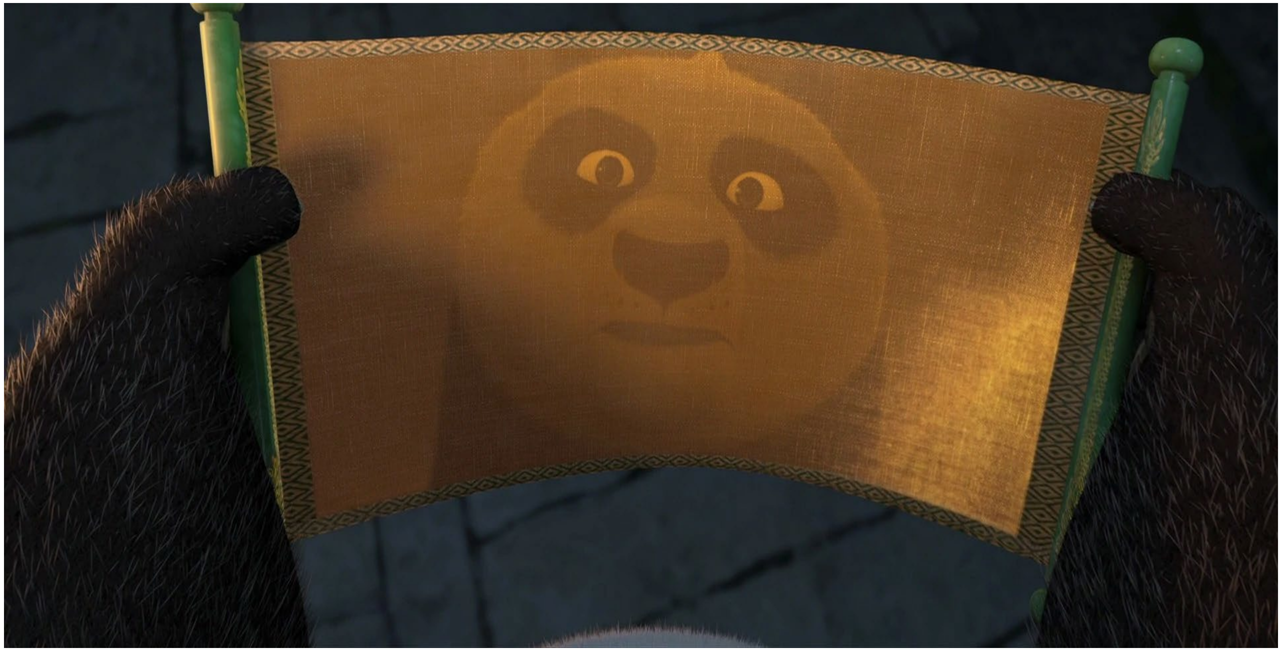
Tips 3/4

Do something you like, and it helps you flying very high!



Tips 4/4

Go Beyond Yourself !



Tips for research leaders

Tips for Research Leaders

"Where the head goes, the body follows"

Perception precedes action.

-- *Ryan Holiday*

Tips for Research Leaders

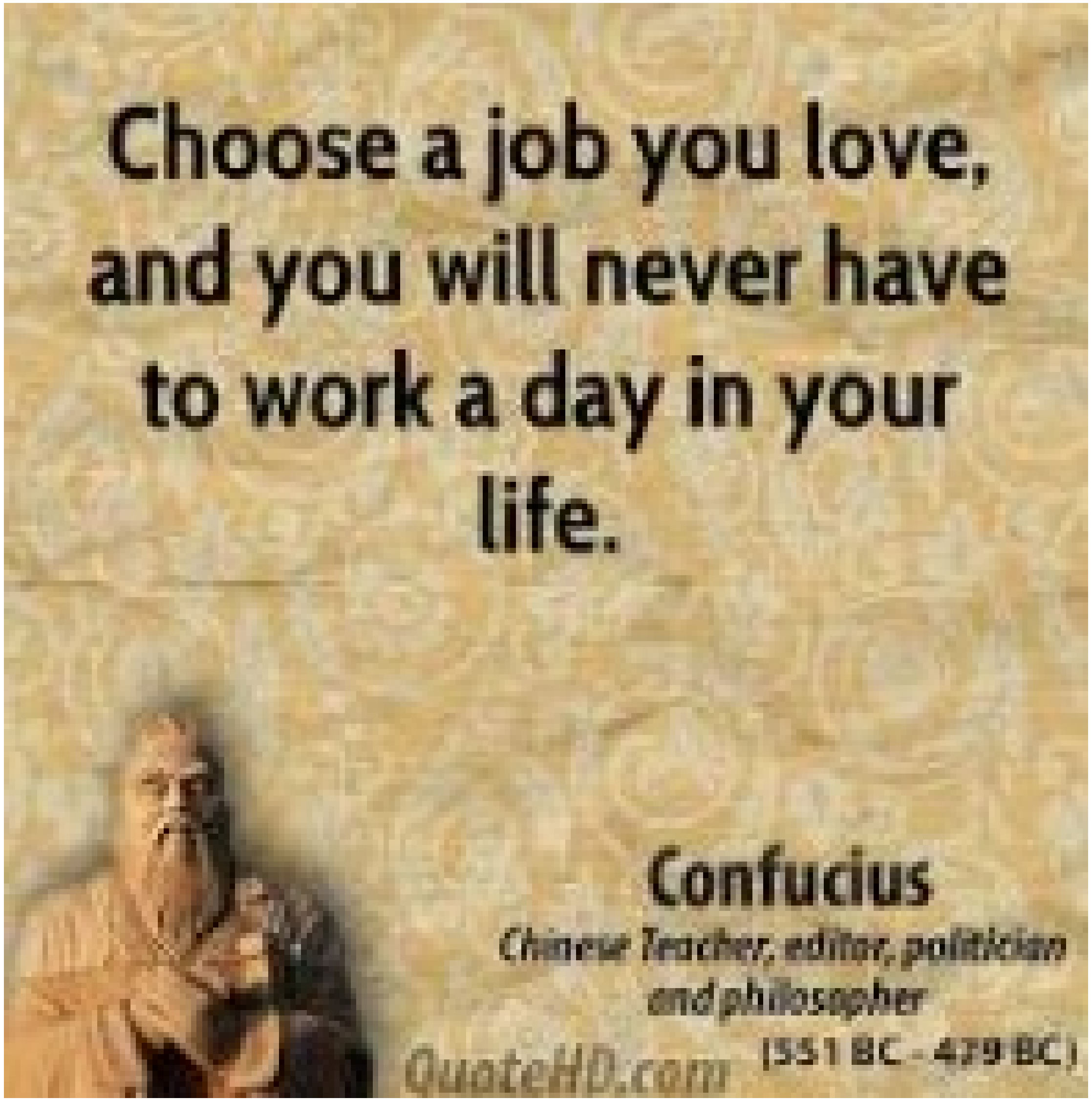
Don’t preach only – do it yourself, and believe in it!

Final Tips

Final Tips

Ask yourself: “do you enjoy research?”

Love your profession.
Love research.
Do passionate research.
Then, the outstanding research outcomes
will come to you.



Choose a job you love, and you will never have to work a day in your life.

Confucius
Chinese Teacher, editor, politician and philosopher
(551 BC - 479 BC)

THANKS FOR LISTENING