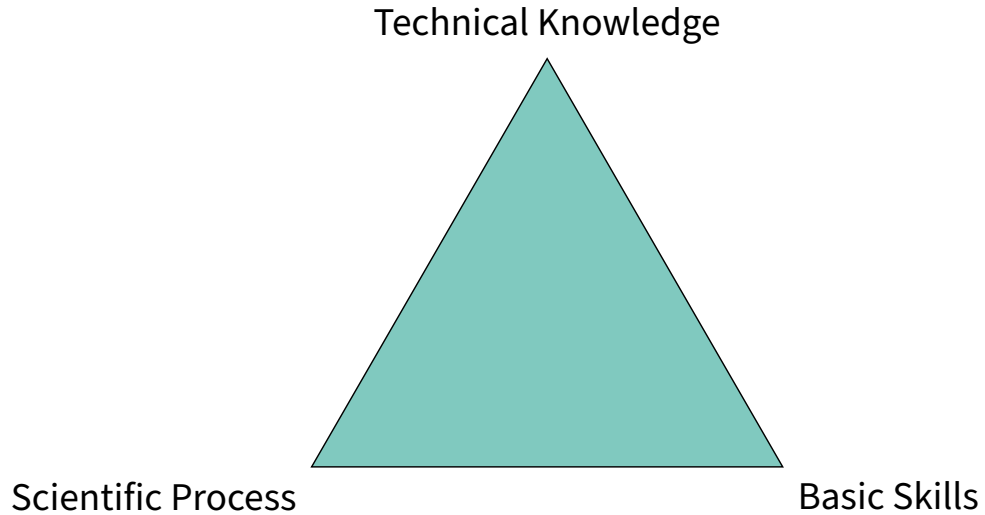


Privacy and Security Seminar WS 2022/23
Basic Skills

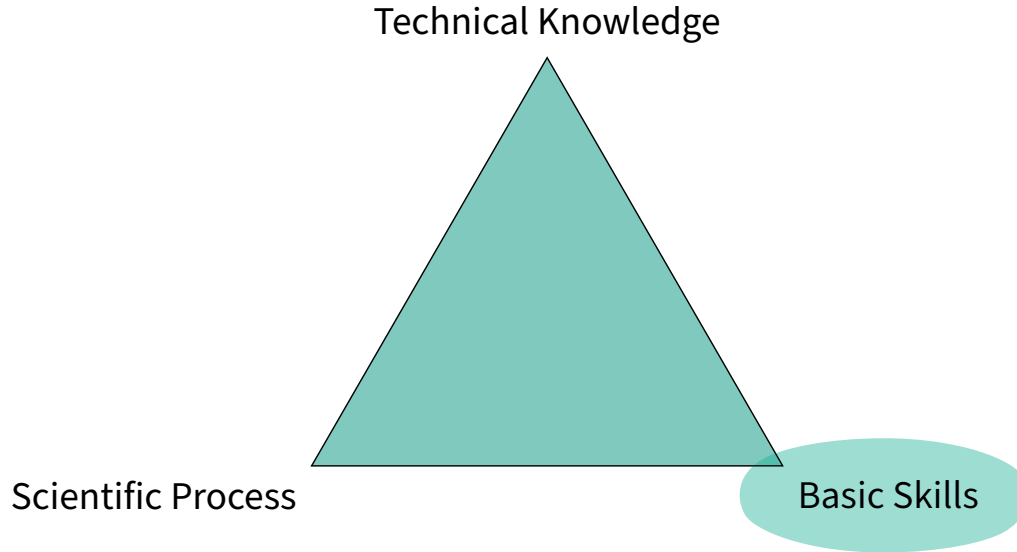
Christiane Kuhn, Patricia Guerra-Balboa

October 31, 2022

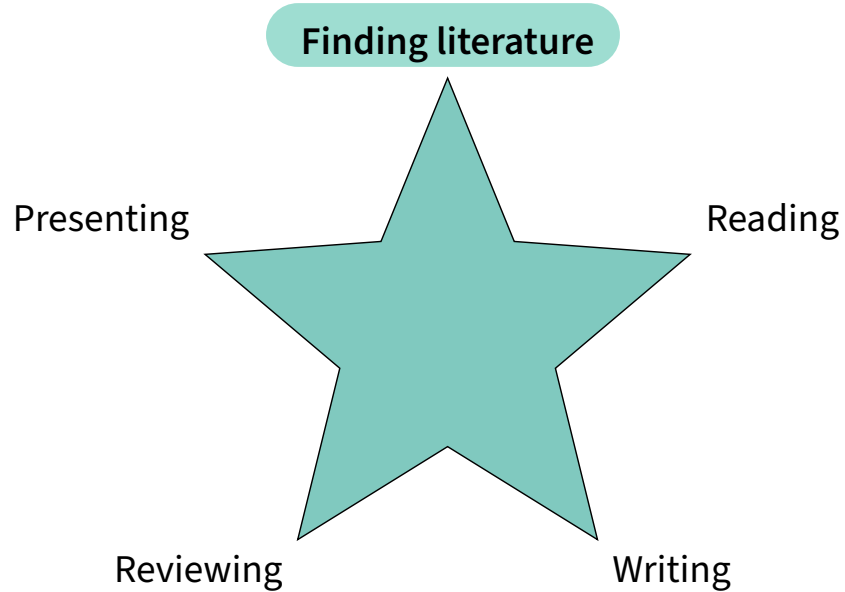
Seminar goals



Seminar goals



Skills



Finding literature

- ▶ Conferences/publication sites
- ▶ Search engines
 - ▶ Google Scholar
 - ▶ Springer
 - ▶ IEEE Xplore
 - ▶ DPBL
 - ▶ Citeseer



Keep it organized

Reference management software

Zotero, Citavi, . . .

Tip:
author+year+first_word



Example:
Dwork2014algorithmic

Search Techniques

Backwards

Which papers are cited in
the reference



Forwards

Which papers cite the
reference

Figure 1: The reference you
are currently reading

Search Techniques

Keywords



Articles include patents Case law



Backwards

Which papers are cited in the reference

Forwards


Which papers cite the reference

Figure 1: The reference you are currently reading

Selection

Check skim paper

- ▶ Area of research
- ▶ Assumptions, system vs. evaluation, . . .

- 
1. Title
 2. Abstract
 3. Conclusion
 4. Introduction
 5. Everything else (as needed)

Check conference quality

- ▶ Ranking systems:
 - ▶ Core: A*, A, B, C
 - ▶ (<http://portal.core.edu.au/conf-ranks/>)
 - ▶ ERA, Qualis,...
- ▶ Number of citations
- ▶ Year of publication



Top Conferences

▶ (Practical) IT-Security:

A* IEEE S&P (Security and Privacy)

Usenix NDSS (Network and Distributed System Security) Usenix Security

ACM CCS (Computer and Communications Security)

A : AsiaCCS, ESORICS, ...

▶ Privacy:

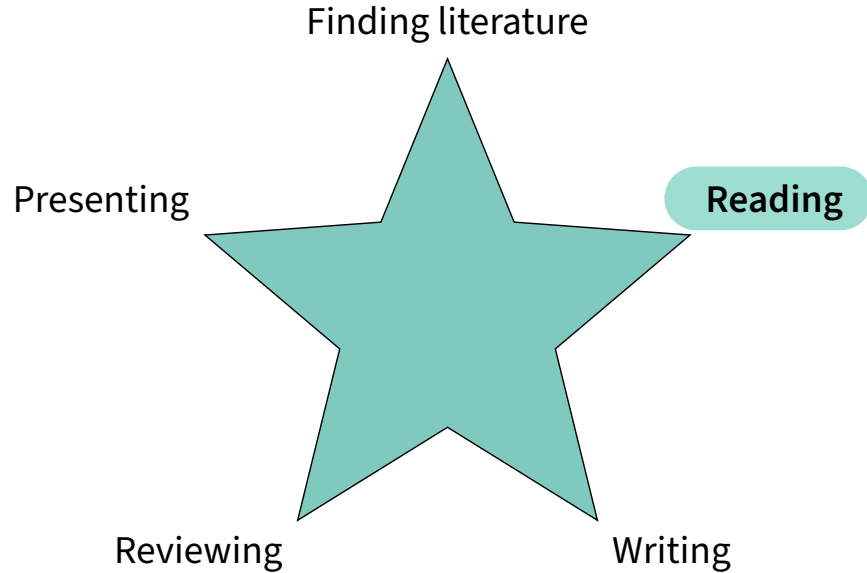
A PETS (Privacy Enhancing Technologies Symposium)

▶ Cryptography:

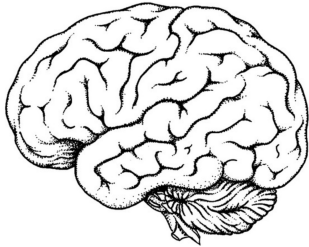
A* Crypto (Advances in Cryptology) EuroCrypt (Int. Conf. on the Theory and Application of Cryptographic Techniques)

A TCC, AsiaCrypt, FC,...

Skills



Before Reading



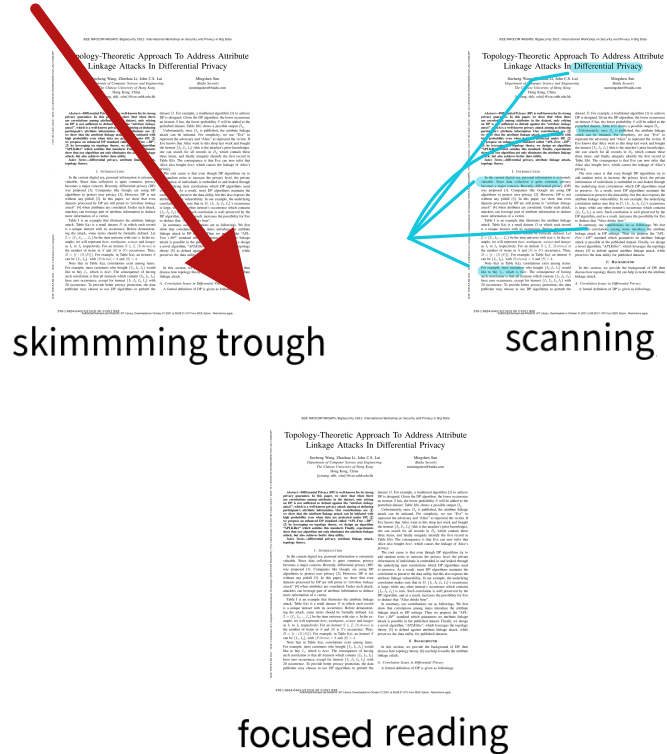
Activate knowledge



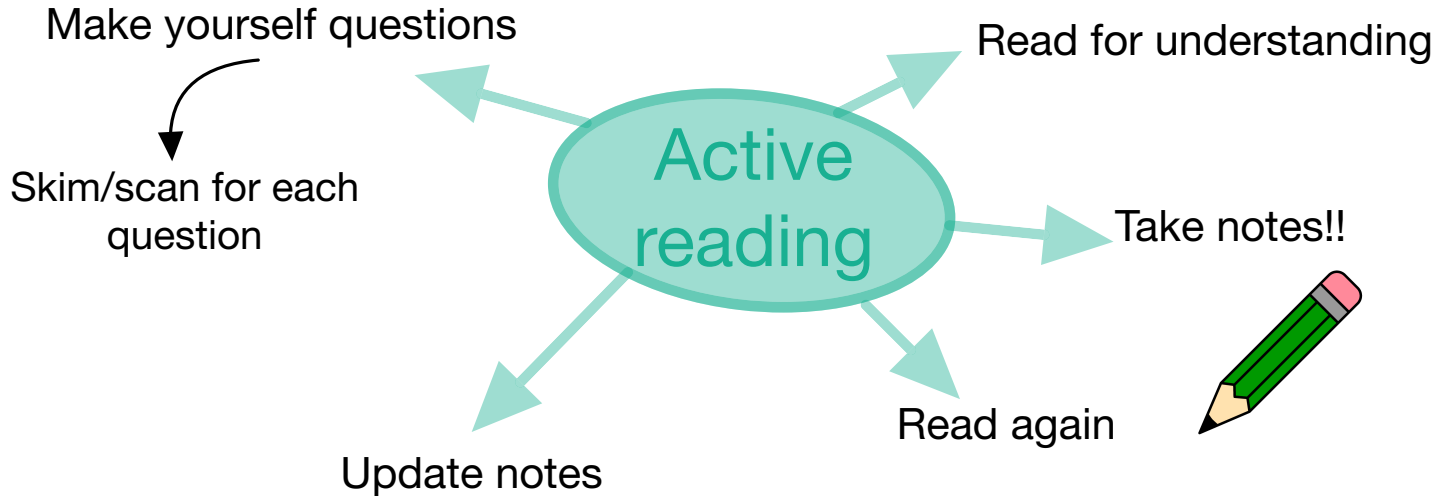
Guiding questions

Techniques

1. Title
2. Abstract
3. Conclusion
4. Introduction
5. Everything else (as needed)



Possible reading strategy



Further material on reading

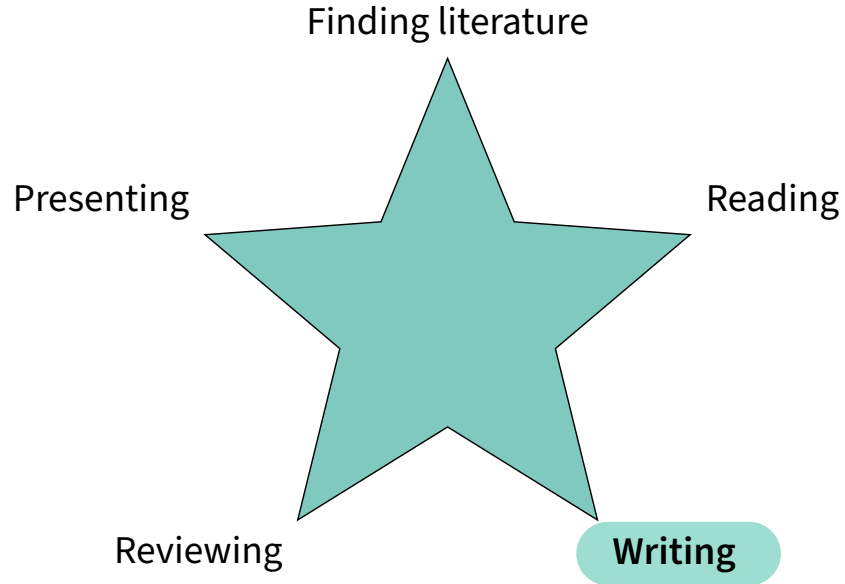
- ▶ **“How to read a paper” by S. Keshav”**

<http://blizzard.cs.uwaterloo.ca/keshav/home/Papers/data/07/paper-reading.pdf>

- ▶ **“About academic reading”**

<https://aso-resources.une.edu.au/academic-reading/about-academic-reading/>

Skills



Structure

0. Abstract
1. Introduction
2. Related work
3. Background
4. Main part
5. Conclusion & Future Work

Topology-Theoretic Approach to Address Arbitrary Linkage Attacks in Differential Privacy

Jinsheng Wang, Zhenhua Li, John C. Sui
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Hong Kong, China
{jwang, zli, csui}@se.cuhk.edu.hk

Mingchen Sun
Baidu Security
sunmingchen@baidu.com

Abstract—Differential privacy (DP) is well-known for its strong privacy guarantee. In this paper, we show that when there are correlation among attributes in the dataset, only relying on DP is not sufficient to defend against the arbitrary linkage attack¹, which is a well-known privacy attack aiming at deducing user's sensitive information. Our contribution are: (1) we show that the arbitrary linkage attack can be initiated with high probability even when data are protected under DP; (2) we propose an enhanced DP standard called "APL-Free DP" (2) by leveraging an topology theory; we design an algorithm "APL-Killer" which satisfies the standard. Finally, experiments show that our algorithm can only eliminate the arbitrary linkage attack, but also achieve better data utility.

Index Terms—Differential privacy, arbitrary linkage attack, topology theory.

I. INTRODUCTION

In the current digital era, personal information is extremely valuable. Since data collection is quite common, privacy becomes a major concern. Recently, differential privacy (DP) was proposed [1]. Companies like Google are using DP algorithm to protect user privacy [2]. However, DP is not without any pitfall [3]. In this paper, we show that even dataset protected by DP are still prone to "arbitrary linkage attack"² if when attributes are correlated. Under such attack, attackers can leverage part of attribute information to deduce more information of a victim.

Table 1 is an example that illustrates the arbitrary linkage attack. Table lists a real dataset Z in which each record is a unique integer with its occurrence. Before deducing the attack, users only should be correctly deduced. Let $Z = \{i_1, i_2, \dots, i_n\}$ be the item universe with size n . In the example, we represent individuals' sensitive attribute and denote as i_1 to i_n respectively. For an dataset $S \subseteq Z$, $|S|$ denotes the number of items in S and $|S| \cap S$ is its occurrence. Thus, $Z = \{i_1, i_2, \dots, i_n\}$. For example, in Table 1, an itemset S can be $\{i_1, i_2\}$ with $|S| \cap S = 2$ and $|S| = 4$.

Note that in Table 1, correlation exists among items. For example, most customers who bought $\{i_1, i_2, i_3\}$ would like to buy i_4 , which is beer. The consequence of having such correlation is that all itemsets which contain $\{i_1, i_2, i_3\}$ has same occurrence except for itemset $\{i_1, i_2, i_3, i_4\}$ with 30 occurrence. To provide better privacy protection, the data publisher may choose to use DP algorithm to perturb the

dataset D . For example, a multiset algorithm [1] to achieve DP is designed. Given the DP algorithm, the lower occurrence in itemset S has the lower probability. S will be added to the perturbed dataset. Table 1(b) shows a possible output D . Unfortunately, since D is published, the arbitrary linkage attack can be initiated. For simplicity, we use "Buy" to represent the arbitrary and "Alike" to represent the victim. If Eve knows that Alice wants to shop last week and bought the items $\{i_1, i_2, i_3\}$ then is the attacker's prior knowledge. He can search for all records in D , which contain these items, and finally uniquely identify the first record in Table 1(b). The consequence is that Eve can now infer that Alice also bought beer, which causes the leakage of Alice's privacy.

The root cause is that even though DP algorithms try to add random noise to increase the privacy level, the private information of individuals is embedded in and leaked through the underlying logic containing which DP algorithms need to preserve. As a result, most DP algorithms maintain the correlation to preserve the data utility, but this also exposes the arbitrary linkage vulnerability. In our example, the underlying correlation makes sure that in D , $\{i_1, i_2, i_3\}$'s occurrence is large, while any other itemset's occurrence which contains $\{i_1, i_2, i_3\}$ is zero. Such correlation is well preserved by the DP algorithms, and as a result, increases the probability for Eve to deduce that "Alice drinks beer".

In summary, our contributions are as following. We first show the correlation among items introduces the arbitrary linkage attack in DP settings. Then, we propose the "APL-Free DP" standard which guarantees an arbitrary linkage attack is possible in the published datasets. Finally, we design a novel algorithm, "APL-Killer", which leverages the topology theory [3] to defend against arbitrary linkage attack, which preserves the data utility for published datasets.

II. BACKGROUND

In this section, we provide the background of DP that shows how our theory [3] can help to tackle the arbitrary linkage attack.

A. Correlation Issues in Differential Privacy
A formal definition of DP is given as following.

A. Privacy Guarantee Analysis

Figure 4 shows the experiment result, and the experiment design is similar with that in Section III-B. One can observe that by using APL-Killer, there is no single APL in the generated dataset, which shows a high privacy guarantee.

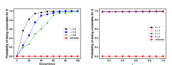


Fig. 4. Privacy comparison using real-world datasets.

B. Data Utility Analysis

Since counting queries is the most fundamental operation in data mining today, we focus on it. For each parameter setting, 50000 random counting queries are generated. Given a query Q , the relative error $err(Q)$ is computed as $\frac{|Q(D) - Q(D')|}{|Q(D)|}$, where $Q(D')$ is the query result on the original dataset, and D' is the newly found in order to weaken the influence of queries with extremely small counting answers, we set the newly found to 100% of the size of the original dataset.

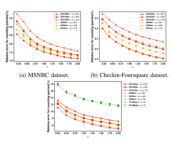


Fig. 5. Relative error for APL-Killer, DiffPriv and Privoxy.

In Figure 5, each point is the average computed by generating 50000 queries in terms of 1,000 results. For DiffPriv and APL-Killer, we change the parameter ϵ and δ , which is used to control the partitioning bound. In APL-Killer, we allow users to use ϵ' for generating itemsets, and we set $\epsilon' = \epsilon$ or ϵ' may be larger ϵ' . Note that it is time-consuming for Privoxy to process large datasets (over 24 times), so Privoxy is not used

for MSNBC and CheckOut datasets. In Figure 5(a) and Figure 5(b), experiment results show that the relative error for APL-Killer is reduced by 3.6% in average, as ϵ varies. In Figure 5(c), one can check that APL-Killer reduces the relative error by 6.8% compared with that of DiffPriv, and 49.1% compared with that of Privoxy. These show APL-Killer has a higher data utility. For multiset DP algorithms, although a smaller ϵ can decrease the probability of being attacked, the data utility becomes worse. However, APL-Killer eliminates this dilemma. No matter how the privacy parameter ϵ is set, the probability of being attacked is guaranteed to be zero. Therefore, our algorithm lets publishers to publish the dataset with good data utility, while defending against the arbitrary linkage attack comprehensively.

VI. CONCLUSIONS

In this paper, we first show that the arbitrary linkage attack is a serious problem when using DP. In order to eliminate this attack, we improve DP and propose APL-Free DP. We further design an algorithm, APL-Killer, which leverages the topology-theoretic approach to defend against the arbitrary linkage attack. However, in our paper, we did not consider the probabilistic arbitrary linkage attack, which is more advanced attack. Also, we did not give a clear instruction on how to choose APL-Killer's parameters to get better data utility. These are potential directions for future research.

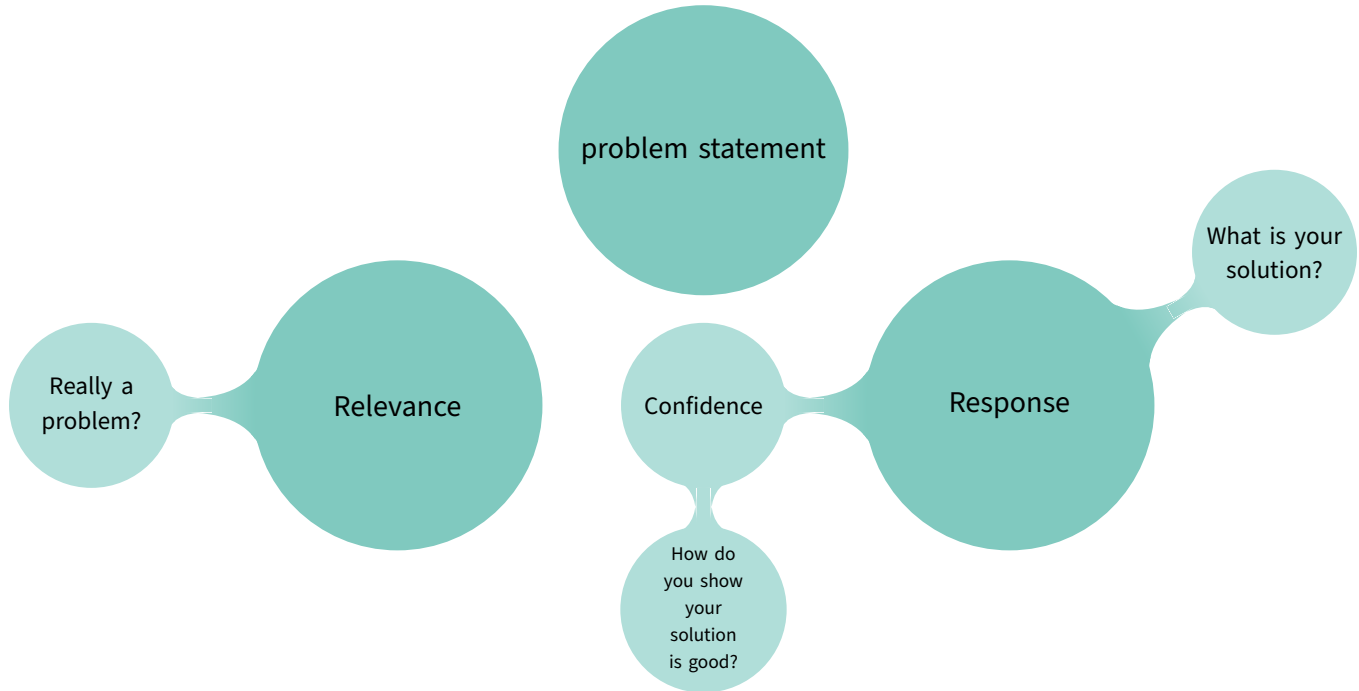
The work of John C. Sui was supported in part by the RIF R0013-18.

REFERENCES

- [1] D. C. Crandall, A. Roth et al., "The algorithmic foundation of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 16, no. 2, pp. 1–107, 2016.
- [2] J. News, "How google and privacy at work: (but not hereby collaboration)," *Privacy: Data Science and Security*, pp. 1–10, 2018.
- [3] Proceedings of the 2017 ACM Conference on Foundations of Differential Privacy, in "Foundations of Differential Privacy," Springer, 2017, pp. 32–50.
- [4] R. Chen, B. Chen, P. F. M. Mohamed, B. C. Fung, and L. Xiang, "Differentially private user data publishing by local suppression," *Proceedings of the 2017 ACM Conference on Foundations of Differential Privacy*, pp. 1–10, 2017.
- [5] C. H. Borden, "Topology of privacy: Factor structure and its function families for customer selection," *Annals of mathematics*, vol. 181, pp. 1171–1211, 2015.
- [6] M. Feldman, "Topology of privacy: Factor structure and its function families for customer selection," *Annals of mathematics*, vol. 181, pp. 1171–1211, 2015.
- [7] D. Kifer and A. Machiry, "On the bias in data privacy," in *Proceedings of the 2017 ACM Conference on Foundations of Differential Privacy*, pp. 1–10, 2017.
- [8] M. Feldman, "Topology of privacy: Factor structure and its function families for customer selection," *Annals of mathematics*, vol. 181, pp. 1171–1211, 2015.
- [9] R. Chen, B. Chen, P. F. M. Mohamed, B. C. Fung, B. C. Fung, and L. Xiang, "Differentially private user data publishing by local suppression," *Proceedings of the 2017 ACM Conference on Foundations of Differential Privacy*, pp. 1–10, 2017.
- [10] "Privacy: Personal data science and security network," ACM Foundation of Differential Privacy (FDP), vol. 42, no. 4, pp. 1–10, 2017.
- [11] Technical report: <https://github.com/2017InfoSecWorkshop/Workshop>



Abstract

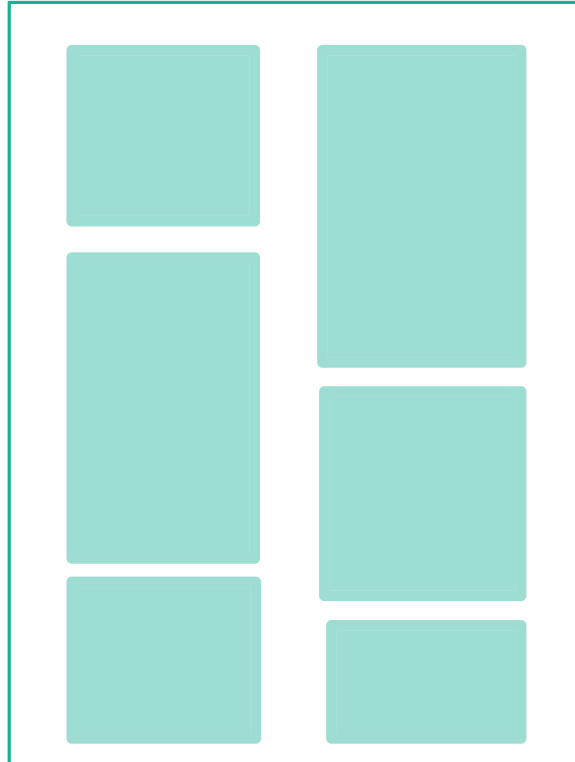


Introduction

Broad topic
& motivation

Specific topic
&
open problem

Goal
&
research question

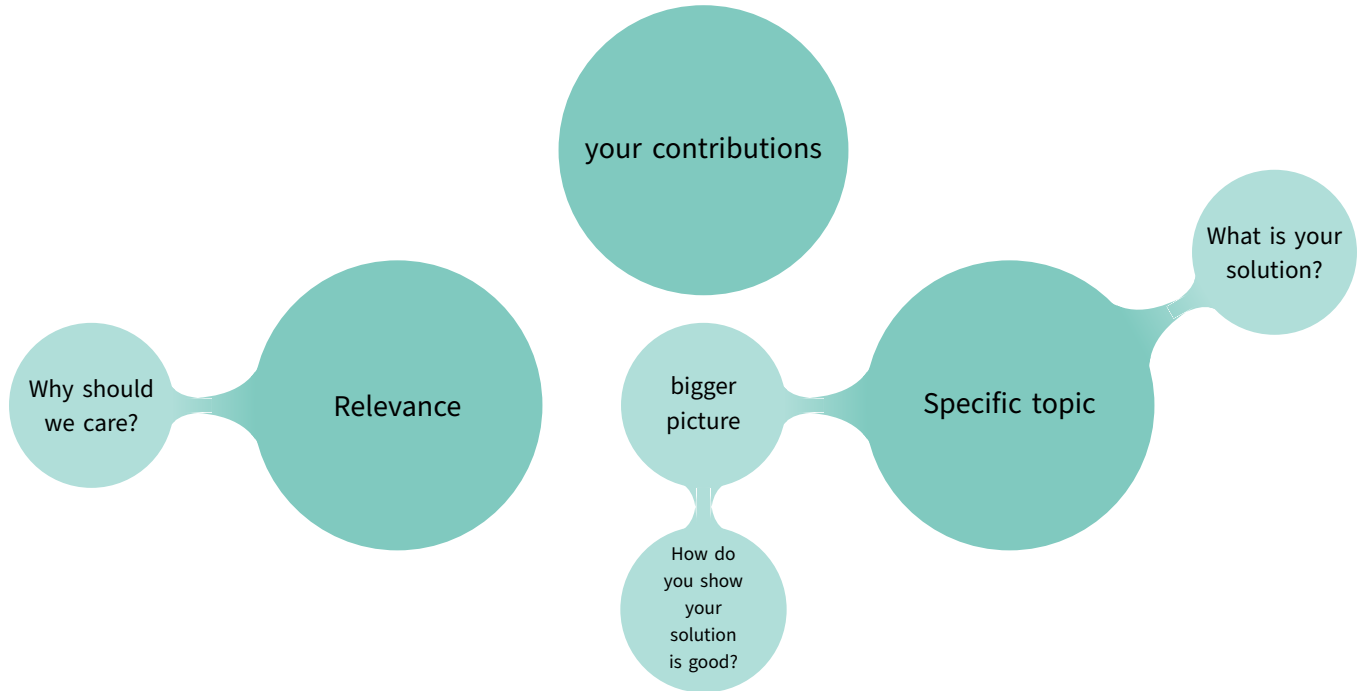


Scientific motivation
&
relevance

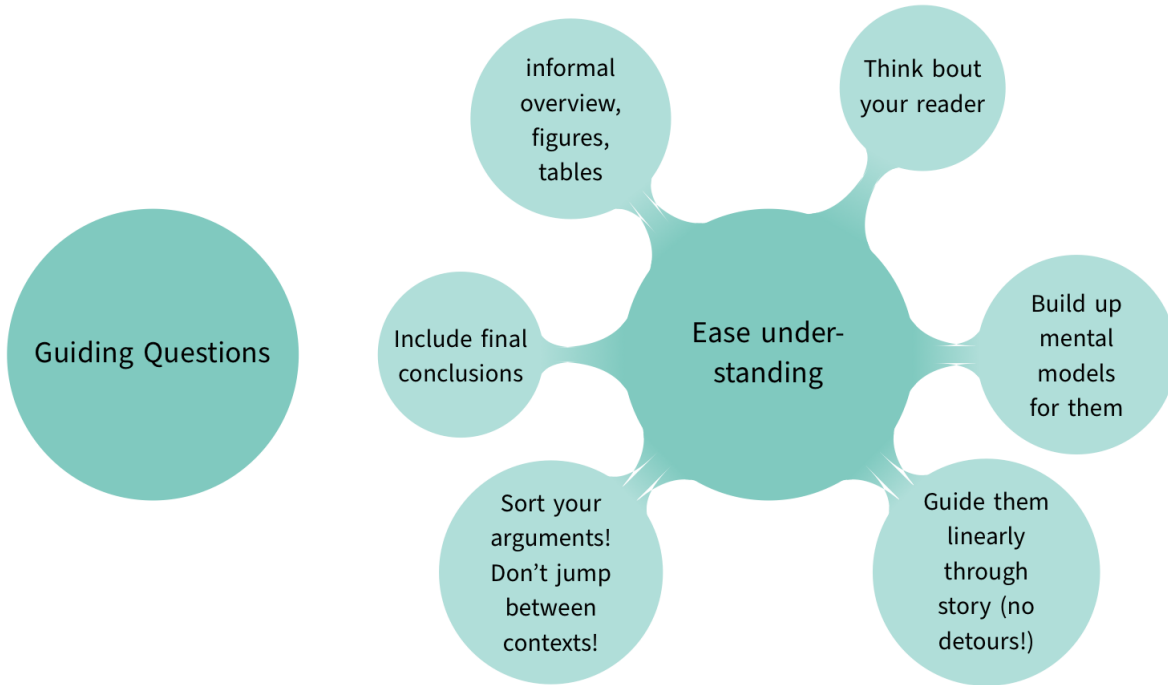
Your contributions

Reader's
digest

Conclusion



main part



Writing style

Basics: Grammar, spellcheck ...

Scope:

- ▶ Sentence \leftrightarrow statement
- ▶ Paragraph \leftrightarrow idea
- ▶ Section \leftrightarrow subtopic

KEEP IT SIMPLE!

- ▶ Short, precise sentences
- ▶ Active $>$ passive
- ▶ Avoid negations
- ▶ Old \rightarrow new

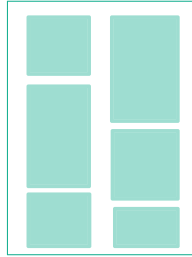
Plagiarism

- ▶ Paraphrase: own words
 - ▶ Close your literature
- ▶ Signal:
 - ▶ Own content
 - ▶ Summary of someone else's
 - ▶ Direct quote

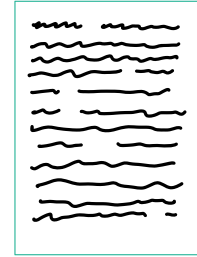
How I approach it



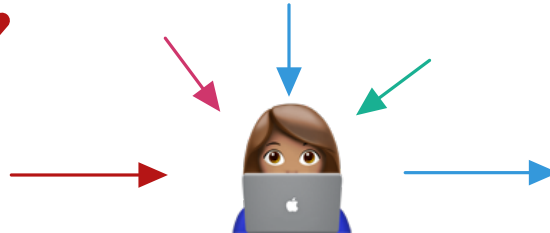
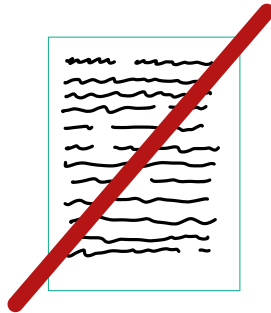
Rough plan



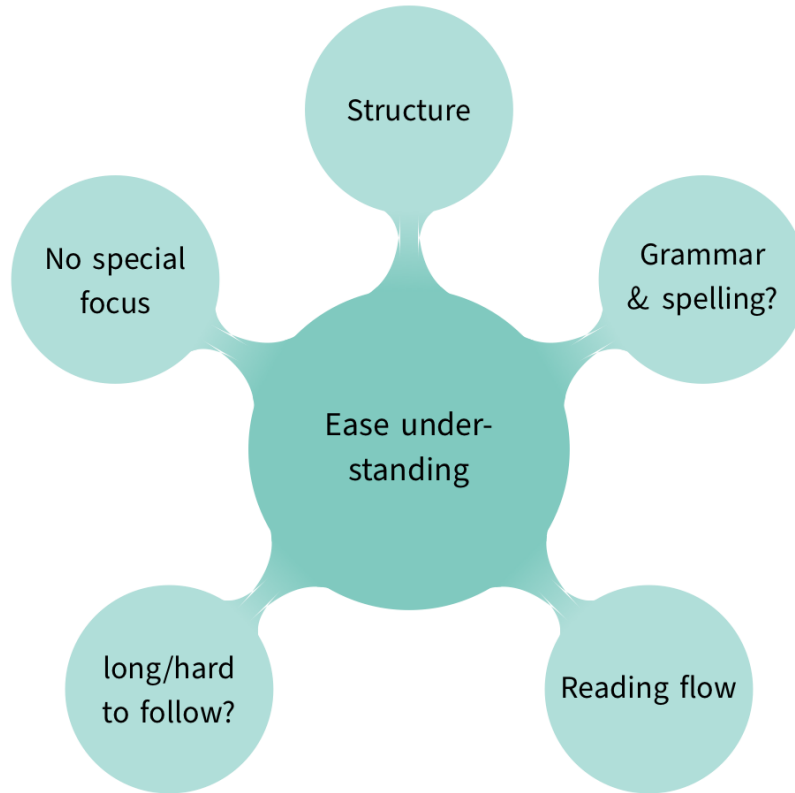
Structure



First draft



Varying focus



Further material on writing

- ▶ **“The Elements of Style” by Strunk and White**

<https://faculty.washington.edu/heagerty/Courses/b572/public/StrunkWhite.pdf>

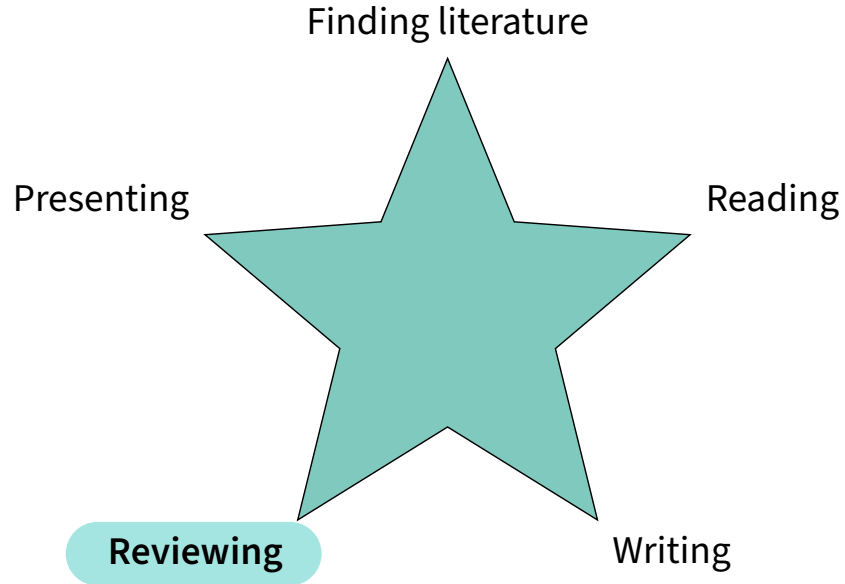
- ▶ **How to Write Papers So People Can Read Them:**

https://www.youtube.com/watch?v=L_6xoMjFr70

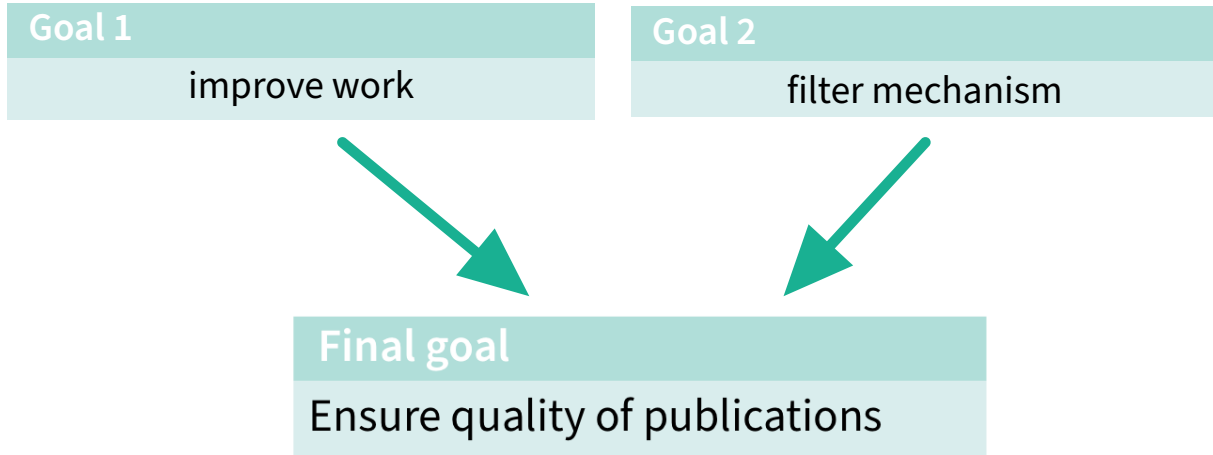
- ▶ **Plagiarism:**

http://www.ou.edu/content/dam/integrity/docs/nine_things_you_should_know.pdf

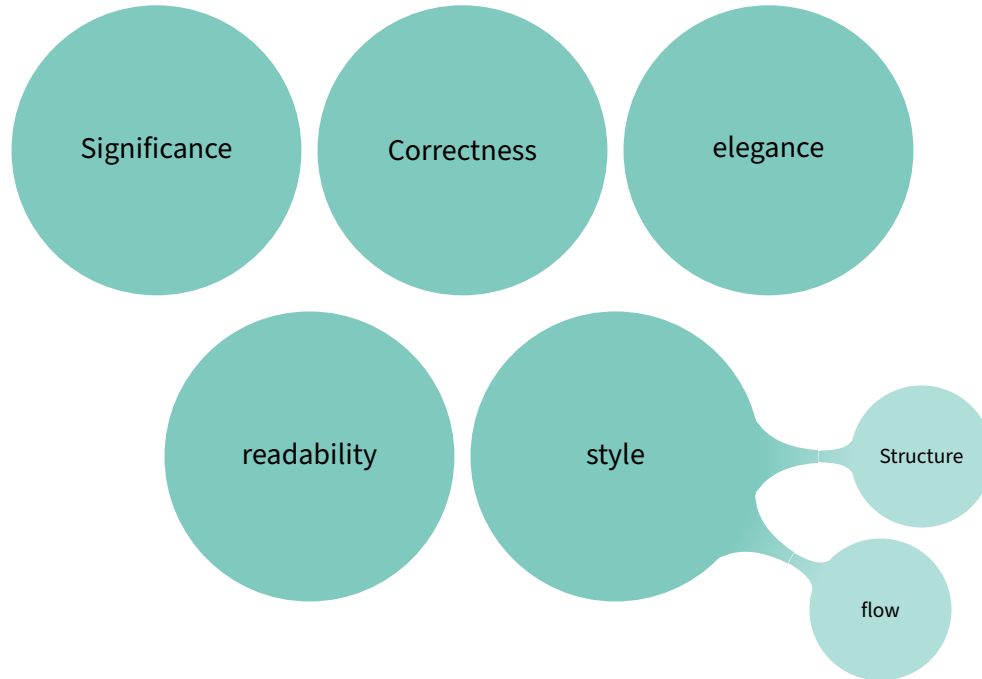
Skills



Why peer-reviewing?



Quality criteria



A good review

Thorough, critical

Follows given
structure

Objective, polite

Helpful, con-
structive, specific

anonymous

Review Structure

- ▶ 3 Strengths & 3 Weaknesses
- ▶ Scale 1 – 5: each part of the paper:
 - ▶ Structure
 - ▶ Argumentation
 - ▶ Readability
 - ▶ Language
 - ▶ Grammar
 - ▶ Formatting
 - ▶ Citation Style
- ▶ Overall ranking (accept (strong/weak), reject(strong/weak))

Opportunity: Receiving Reviews

Take your time for
every point



Harsh/wrong/
unfounded critics

Limited time

Learns what has been
misunderstood

Open your mind

Further material on Reviews

- ▶ **“The Task of the Referee”** by Alan Jay Smith:
<https://www.cs.utexas.edu/users/mckinley/notes/reviewing-smith.pdf>
- ▶ **“A Guide for New Referees in Theoretical Computer Science”** by Ian Parberry
https://basics.sjtu.edu.cn/links/guide_referees.pdf

Skills



Purpose first!

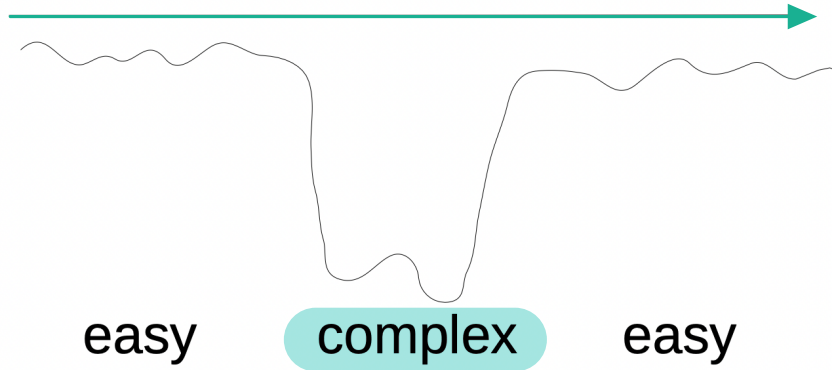
Three teal-colored circles are arranged horizontally in the center of the slide. Each circle contains a single word in black, uppercase letters. From left to right, the words are 'PERSUADE', 'INFORM', and 'ENTERTAIN'.

PERSUADE

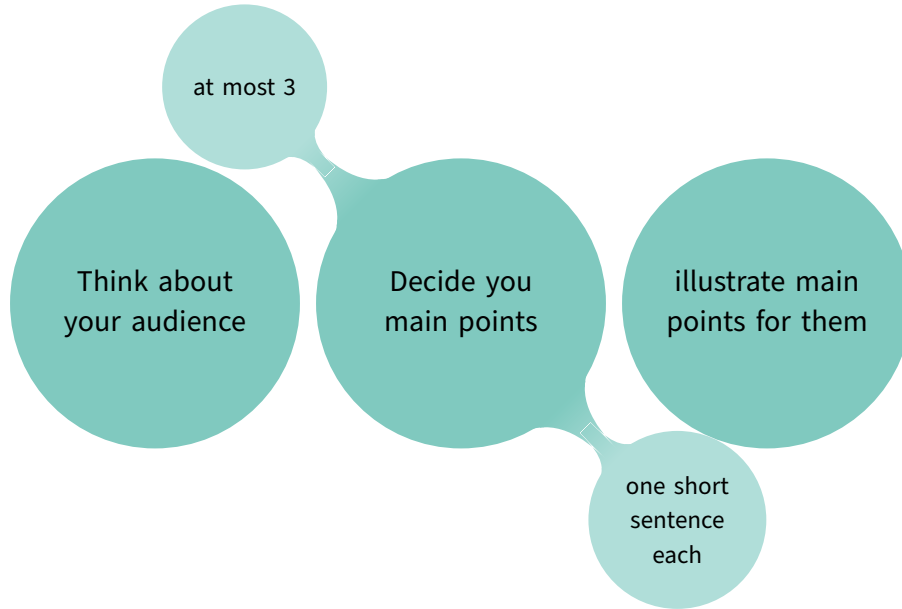
INFORM

ENTERTAIN

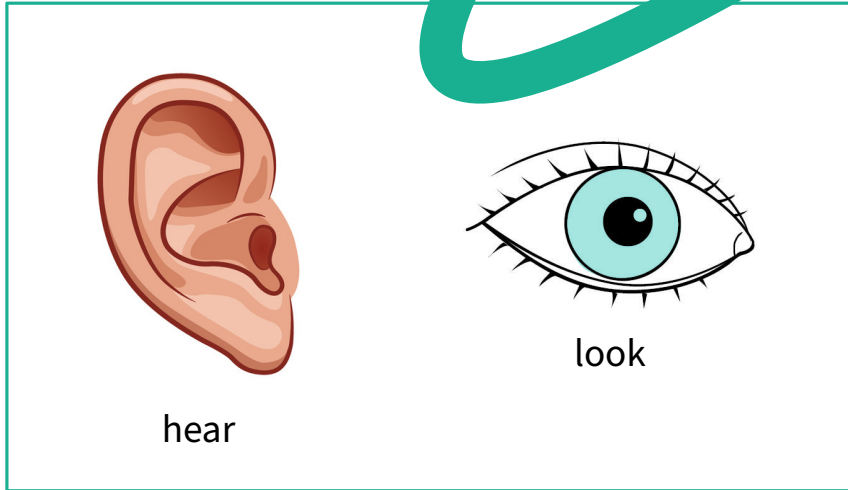
The grebe strategy



building the presentation strategy



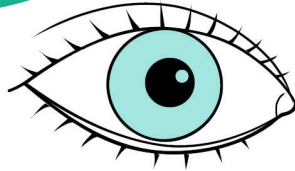
The basics



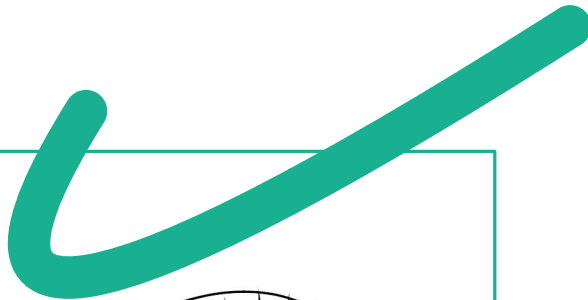
The basics



hear




look

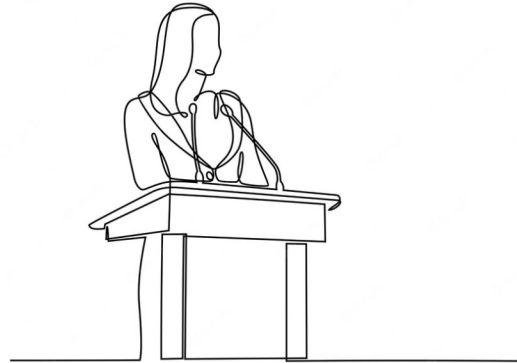


DO NOT READ


Figures ↑↑ Vs. Text ↓↓

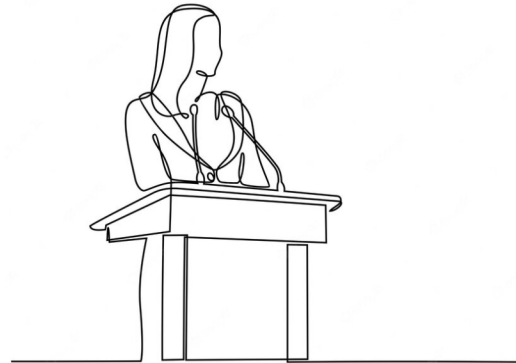
The basics

- ▶ Do not read! 
- ▶ Look to the people
- ▶ Use your body language
- ▶ Change your voice



The basics

- ▶ Do not read! 
- ▶ Look to the people
- ▶ Use your body language
- ▶ Change your voice



Slow

Fast

Not To Do List

- ▶ Not signaling own/other's contributions
- ▶ Finish after 2/3 of the allowed time
- ▶ Go 1/3 over time
- ▶ Include everything - all the details!
- ▶ Cover every part, but give no details at all (No depth)
- ▶ Only cover a tiny part of your work (No breadth)

Further material on presenting

- ▶ **“How to avoid death By PowerPoint”** by David JP Phillips:
<https://www.youtube.com/watch?v=Iwpi1Lm6dFo>
- ▶ **“PowerSpeak”** by Dorothy Leeds

Good luck!

