LINGUISTIC INFORMATION MANAGEMENT

KONRAD JUSZCZYK, PHD
JUSZCZYK@AMU.EDU.PL

WHAT IS INFORMATION?

- ***Information is extensive:*** *the combination of two independent datasets with the same amount of information contains twice as much information as the separate individual datasets.*

- ***Information reduces uncertainty.*** *The amount of information we get grows linearly with the amount by which it reduces our uncertainty until the moment that we have received all possible information and the amount of uncertainty is zero.*

- *The only mathematical function that unifies these two intuitions about extensiveness and probability is the one that defines the information in terms of the negative log of the probability: I(A)=–log P(A)*

- *(Shannon 1948; Shannon & Weaver 1949, Rényi 1961).*

- *https://plato.stanford.edu/entries/information/*

WHAT IS LINGUISTIC INFO?

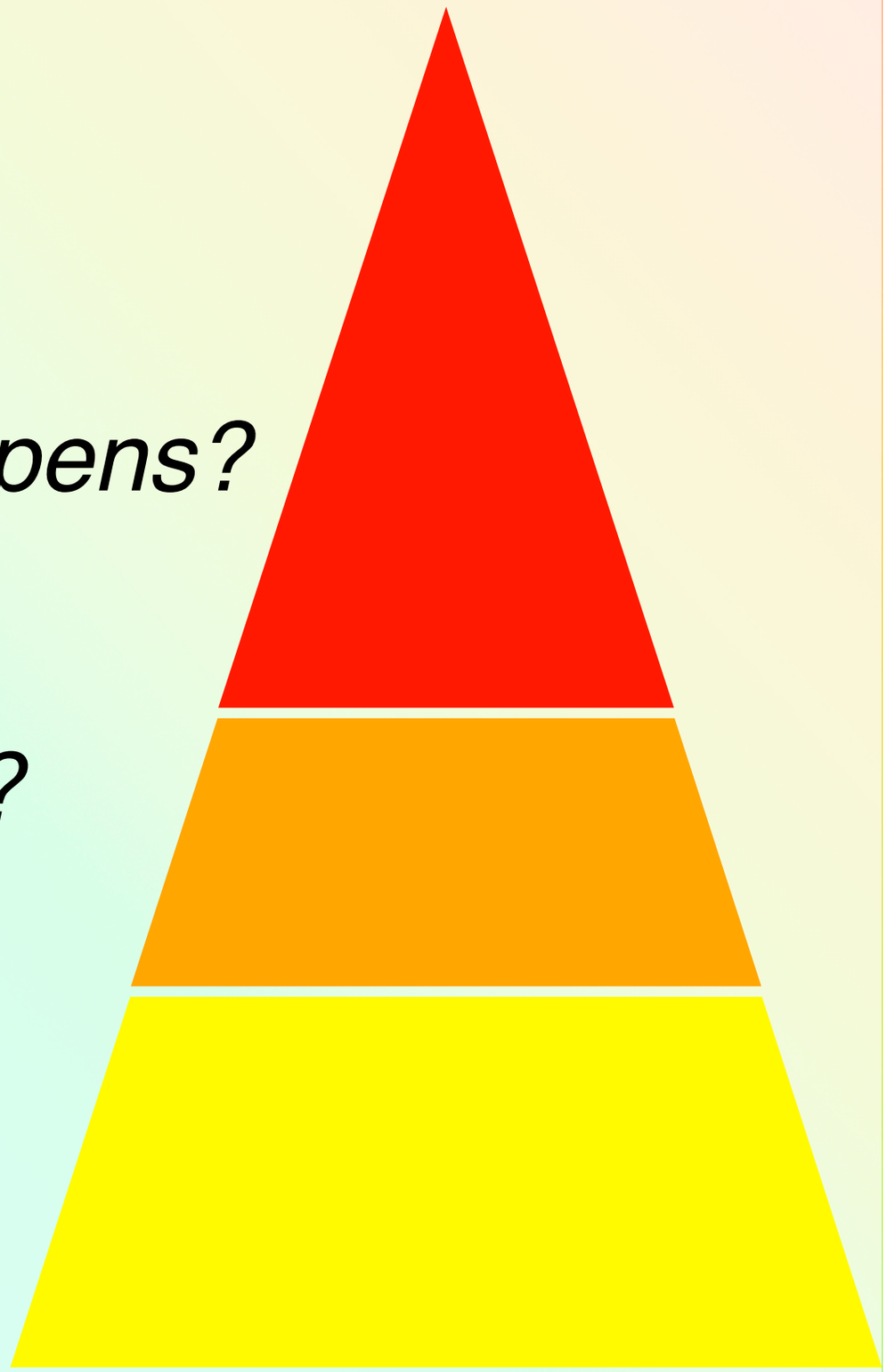
- Data about languages and linguistic data in various forms:
  - **verbal**: texts or strings of symbols
    - spoken, but transcribed
    - written in one of writing systems
  - **vocal**: speech recorded and stored in analogue or digital form on media
    - shows how does a person sound like when they communicate.
  - **visual**: video of a person recorded and stored in analogue or digital form
    - shows how does a person look like when they communicate.
  - **factual**: names of languages, numbers of speakers, distribution in the world

WHAT IS MANAGEMENT OF INFO?

- Linguistic data is stored in databases and can be accessed via Internet or not yet gathered and archived or already gone, if languages are not used anymore and we don't know much about ancient, forgotten and extinct languages.
- Once we have information, we can manage it and manage linguistic processes.
- Linguistic change:
  - policy: what dialect is chosen as the official language in which country?
  - education: how the langauge is taught in schools as a native or foreign lg?
  - migrations: what happens when people move from one country to another?
  - historical: how the language changes, evolve over time?
  - distribution: where the language is used (spoken), where are its speakers?

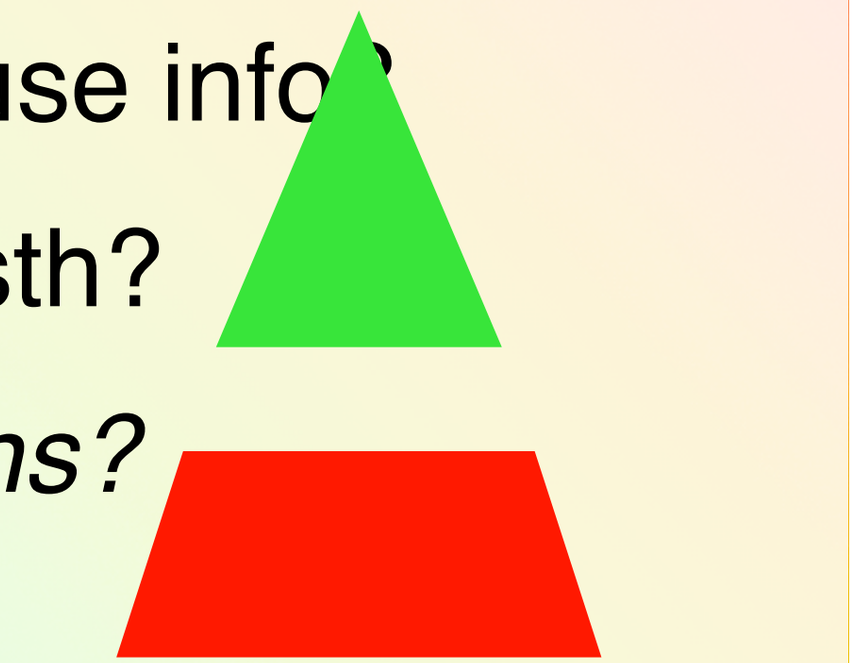
KNOWLEDGE PYRAMID

- KNOWLEDGE: *why it happens?*
  - *what can I do with info?*
- INFORMATION: *what is it?*
  - combined observations
- DATA: what happens?
  - observations, facts



KNOWLEDGE PYRAMID

- POWER: How can I act, do, use info?
  - Am I allowed or able to do sth?
- KNOWLEDGE: *why it happens?*
  - *what can I do with info?*
- INFORMATION is *what is it?*
  - combined observations
- DATA are observations, facts
  - *what happens?*



KNOWLEDGE PYRAMID

- POWER: Am I allowed or able to do sth?
  - authorities, governments, organisations, countries
- KNOWLEDGE: what can I do with information?
  - people know because people are able to think
- INFORMATION: what is it? e.g. your name and birthday
  - information is combined data, contextualised data
- DATA: what, when, where, who, but not how or why
  - just a number or name or date or place of something

QUIRKY QUESTIONS

- How much information is stored in DNA? How much of DNA is actually about ourselves? How different are we from our ancestors: parents, primates, mammals, animals, plants?
- How much information is stored in a language or in a linguistic unit: utterance, sentence?
- Is there life beyond earth? What information from extraterrestrials do we have? What information do extraterrestrial beings have about us? How do we know?
- Is there a life after life? How do we know? How can we find out? What do you belief?
- How much information is there at all? How can we measure it?
- How can information be found, if it is lost, forgotten or unknown?
- Does artificial intelligence speak, generate language, think, feel like humans?
- How to win a million? What are my chances of winning a million?
- Is there an answer to every question? Are there unanswered and unanswerable questions?
- Can we act without knowing what to do? Can we act contrary to our knowledge?
- Can we know or find or understand or see/hear/feel/taste/smell without any knowledge?

WHAT HAPPENS WITH INFORMATION?

DATA IS THE NEW OIL!

AN ASSIGNMENT FOR STUDENTS

- Try to answer a question about linguages or anything else with linguistic data and tools.
- Use research process steps applied in empirical science to find the answer:
  - Define the problem, situation, research question: **what do you want to know?**
  - Choose the methodology to find the answer, solve the problem: **how will you know?**
  - Collect linguistic data relevant to the research question: **what will you analyse?**
  - Analyse the linguistic data using the chosen methodology to answer the research question.
  - Draw conclusions and summarise your research process in five steps like these.
- Work alone or in pairs (<5, groups (<4), whatever way will be useful for you, but do work!
- **Every third class will be for the presentation of your results: you will speak and show.**
- Use anything and everything: intuition, knowledge, books, papers, databases, google, chatGPT

HOW DO YOU KNOW?

DO YOU HAVE AN OPINION OR DATA, FACTS, INFORMATION AND KNOWLEDGE TO SUPPORT YOUR CLAIM?

- When, why, whom should I trust?

HOW DOES GOOGLE SEARCH WORK?

HOW SHOULD I INTERPRET RESULTS OF GOOGLE SEARCH?

HOW DOES CHATGPT WORK?

HOW SHOULD I INTERPRET AN ANSWER FROM CHATGPT?

1. DEFINE THE RESEARCH PROBLEM

TYPOLOGY OF RESEARCH QUESTIONS

WHAT WOULD YOU LIKE TO KNOW?

- Qualitative description, characterisation, features
  - About **examples**: what are examples of linguistic units, communicative units, corpus?
  - About **classification**: typology of linguistic units or phenomena or literary motifs
- Qualitative-quantitative description, explanation with or without falsification
  - About **causation**: why is something happening, changing, being some, increasing, decreasing?
  - About **comparison**: phenomenon x versus phenomenon y or linguistic change in time
- Statistical analysis of features of language units or reactions/behaviours of language users
  - About **dependence**: does x depend on y? Using observation (corpus) or experiments.
  - About **opinion**: what do language users think about linguistic entities or phenomena?
- Description of method and steps needed for use of a tool, a test of a tool or method of analyses
  - About **application**: language teaching methodology or computer and corpus-based language analysis, programming, Natural Language Processing and NL Generation and NL Understanding

1. DEFINE THE RESEARCH PROBLEM

DESCRIBE A SITUATION

- Business pitch: wow, why, how, who, when, for whom, how much, is it better?
- Every week one language

2. CHOOSE METHODOLOGY

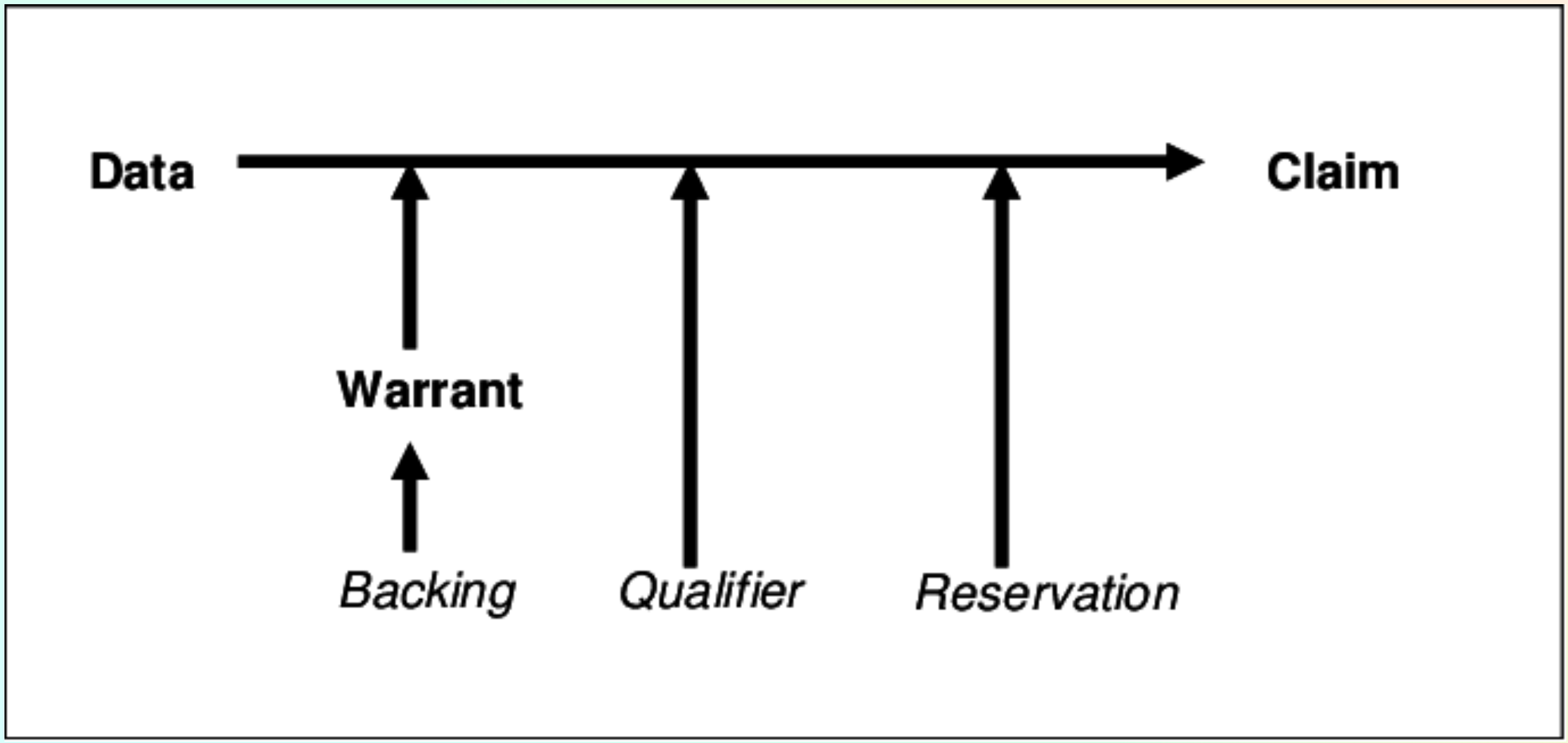
3. COLLECT LINGUISTIC DATA

4. ANALYSE THE LINGUISTIC DATA

5. DRAW CONCLUSIONS FROM DATA

- Basic form of an argument: the what (statement) + the why (one or more premises).

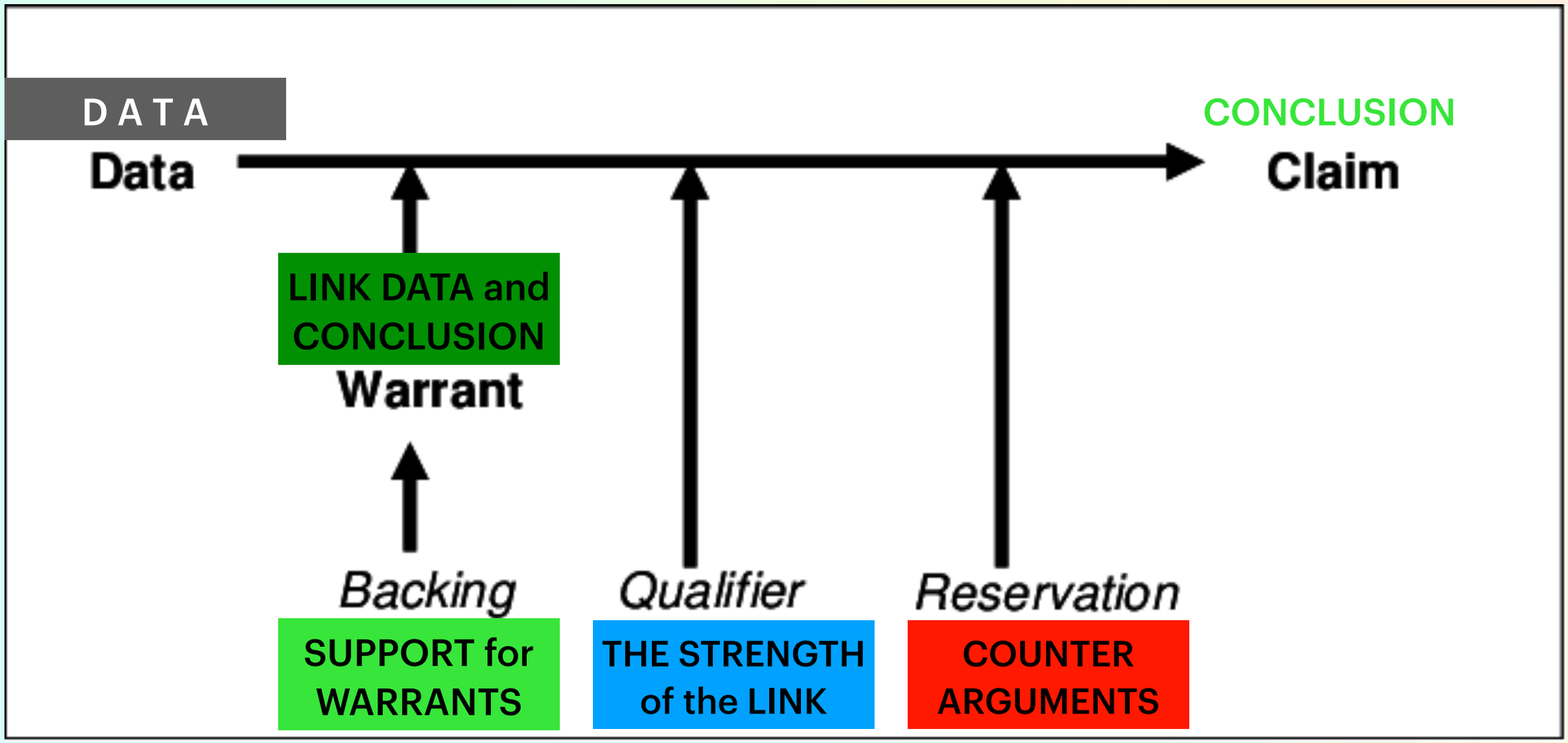
STRUCTURE OF THE IDEAL ARGUMENT (TOULMIN)



IDEA: Toulmin, S., The Uses of Argument. 1958, Cambridge: Cambridge University Press.

DIAGRAM: Jorgensen, M. & Dybå, Tore & Kitchenham, Barbara. (2005). Teaching Evidence-Based Software Engineering to University Students. Proceedings - International Software Metrics Symposium. 2005. 24- 24. 10.1109/METRICS.2005.46.

STRUCTURE OF THE IDEAL ARGUMENT (TOULMIN)



IDEA: Toulmin, S., The Uses of Argument. 1958, Cambridge: Cambridge University Press.

DIAGRAM: Jorgensen, M. & Dybå, Tore & Kitchenham, Barbara. (2005). Teaching Evidence-Based Software Engineering to University Students. Proceedings - International Software Metrics Symposium. 2005. 24- 24. 10.1109/METRICS.2005.46.