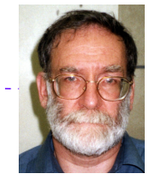
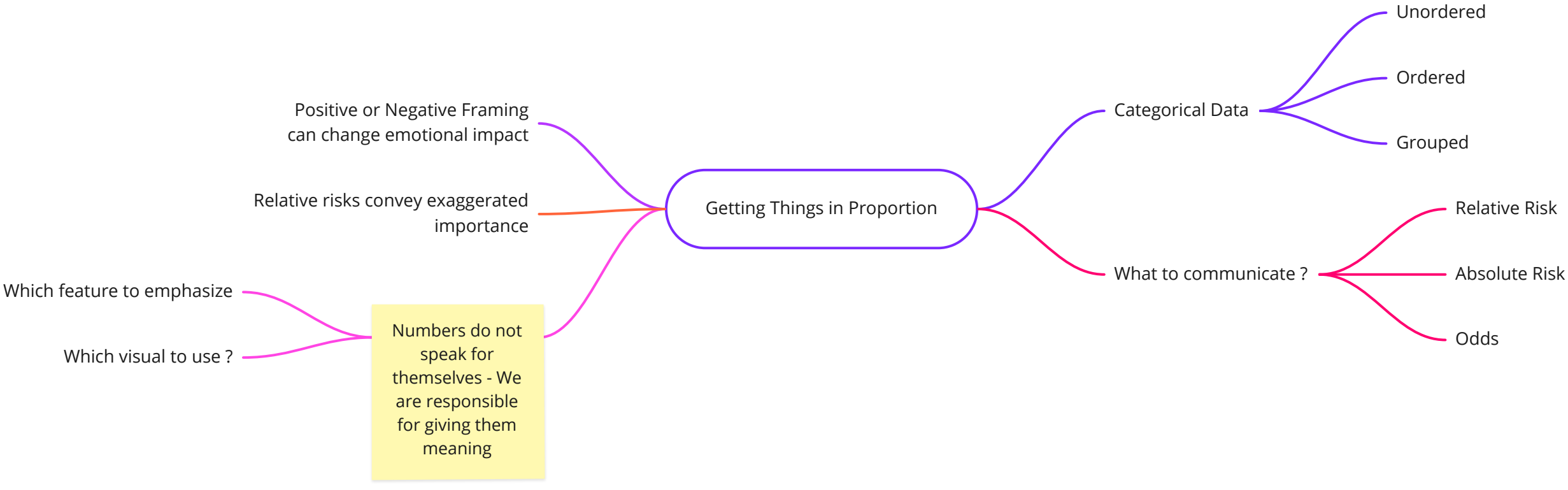
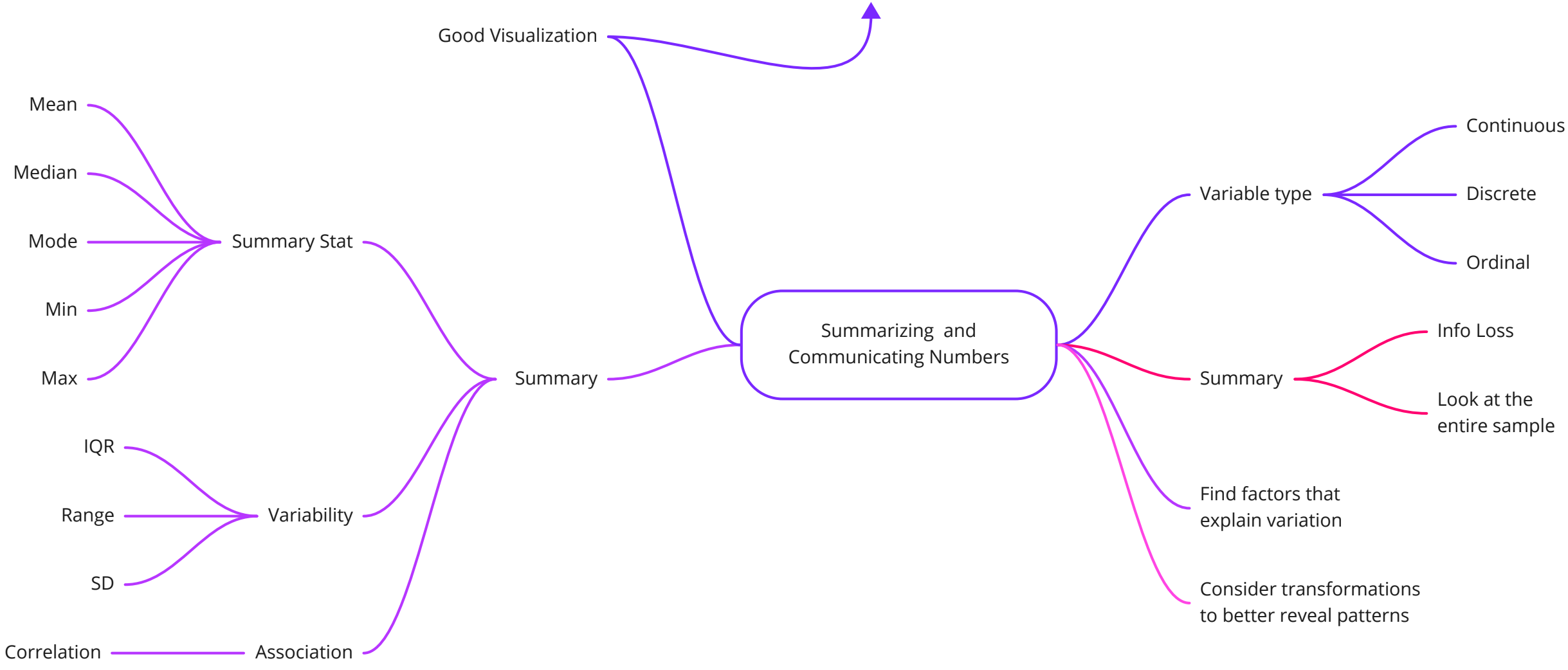


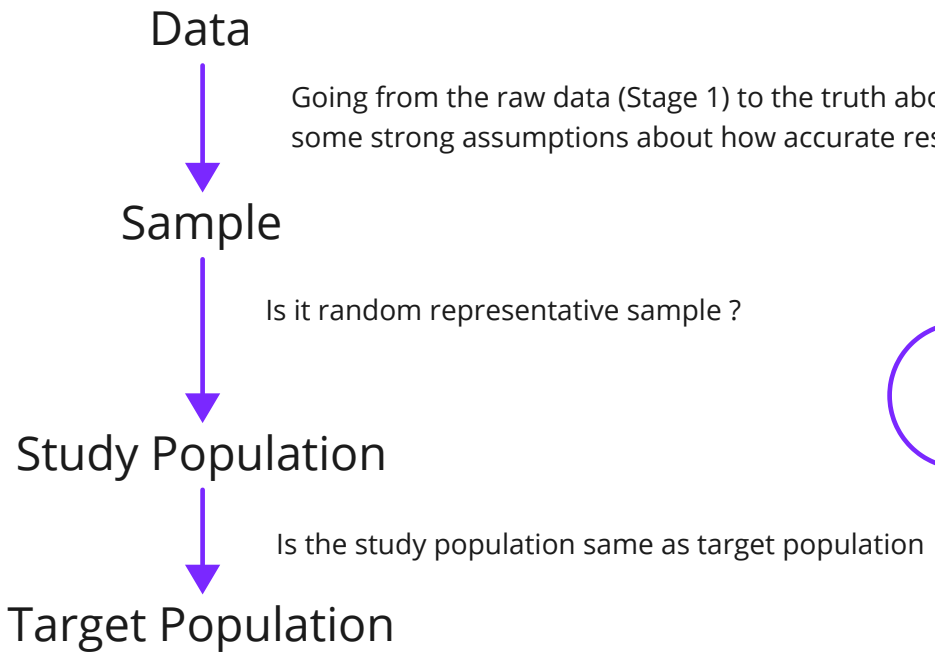
- Introduction
  - Two Limitations
    - Almost always an imperfect measure of what we want to measure
    - Measurements vary by place, by time, by person...
  - What every one is taught ?
    - Dry Subject
    - Bag of disconnected tools
  - Data Literacy
    - Able to carry out analysis
    - Able to communicate
    - Able to Critique
    - Key Skill in the modern world
  - Harold Shipman Murders - Analysis
    - Exploratory
    - Forensic
    - No Math
      No Theory
      Search for Patterns
  - Defining the thing you are measuring is important
    - It is tricky
    - Examples
      - GDP
      - Unemployment
      - Trees on Earth
      - Happiness
  - PPDAC Cycle
    - Problem
    - Plan
    - Data
    - Analysis
    - Communicate

PROBLEM PLAN DATA ANALYSIS COMMUNICATE

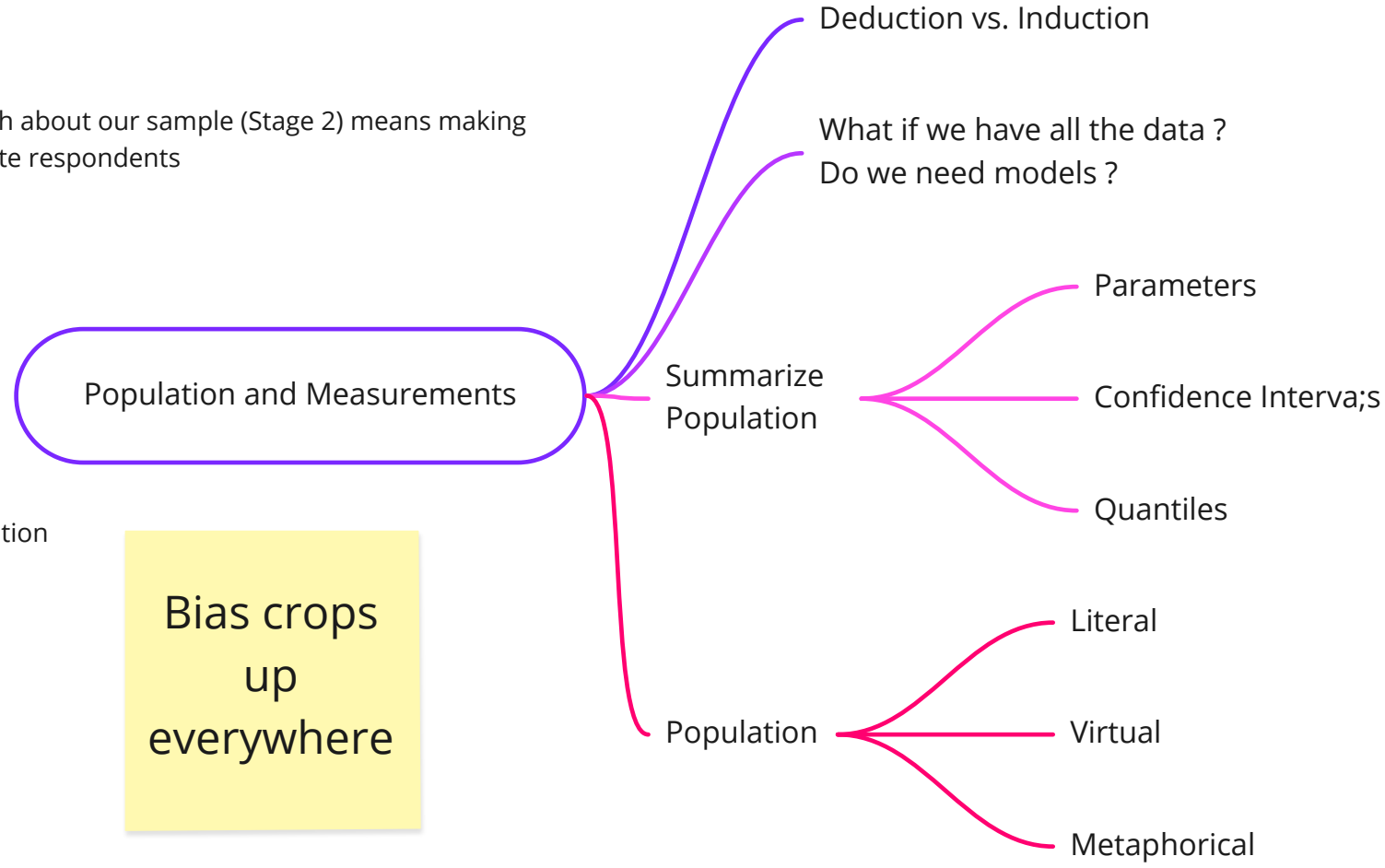

Positive or Negative Framing can change emotional impact
Relative risks convey exaggerated importance
Which feature to emphasize
Which visual to use ?
Numbers do not speak for themselves - We are responsible for giving them meaning
Getting Things in Proportion
Categorical Data
Unordered
Ordered
Grouped
What to communicate ?
Relative Risk
Absolute Risk
Odds

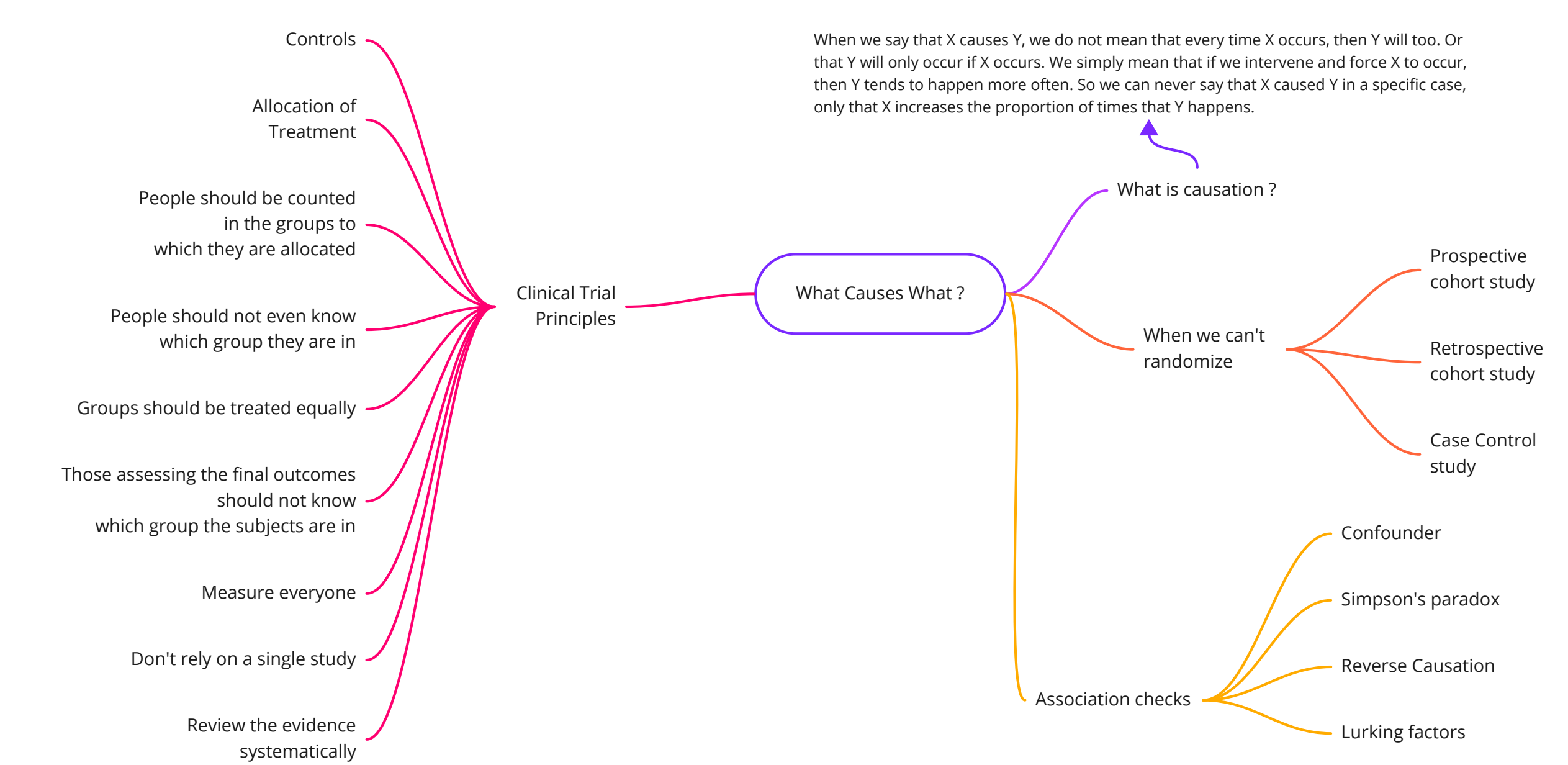
1. It contains reliable information.
2. The design has been chosen so that relevant patterns become noticeable.
3. It is presented in an attractive manner, but appearance should not get in the way of honesty, clarity and depth.
4. When appropriate, it is organized in a way that enables some exploration.

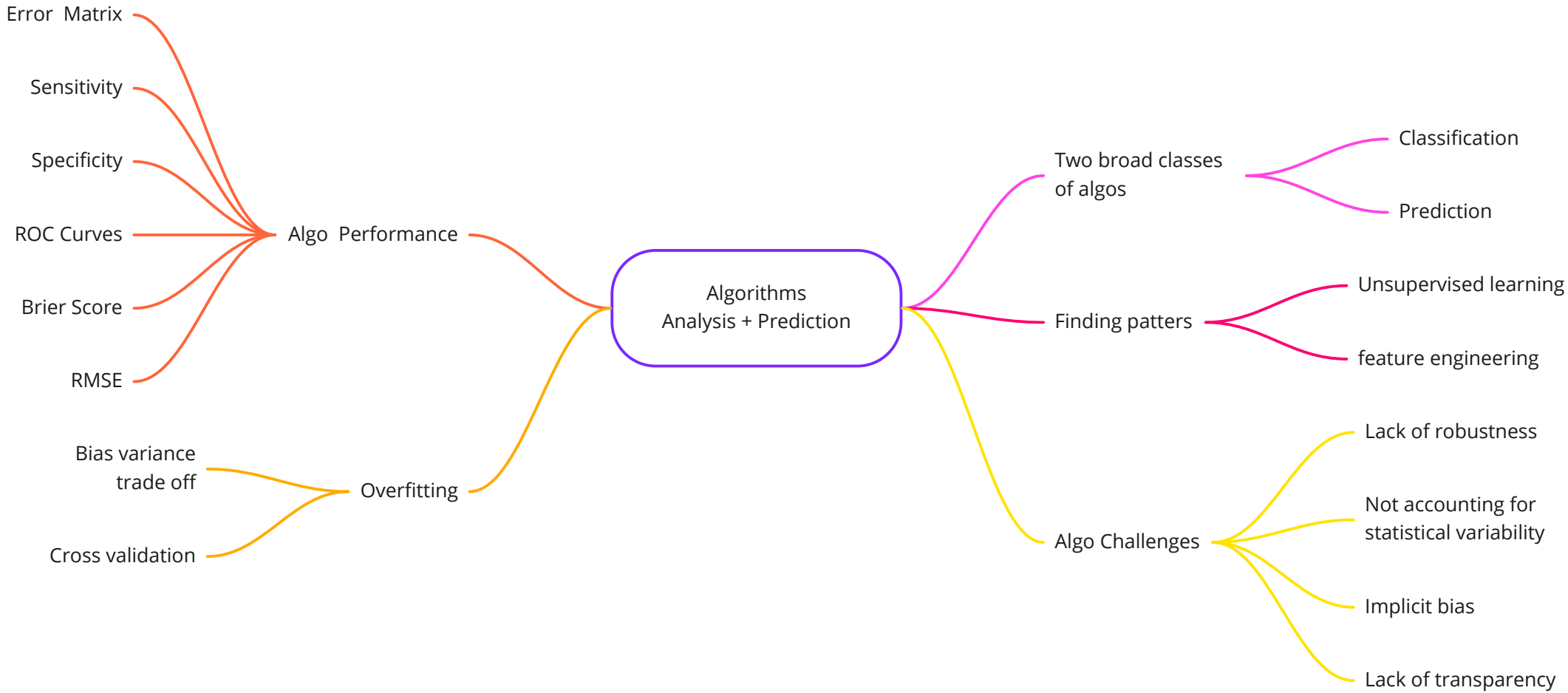


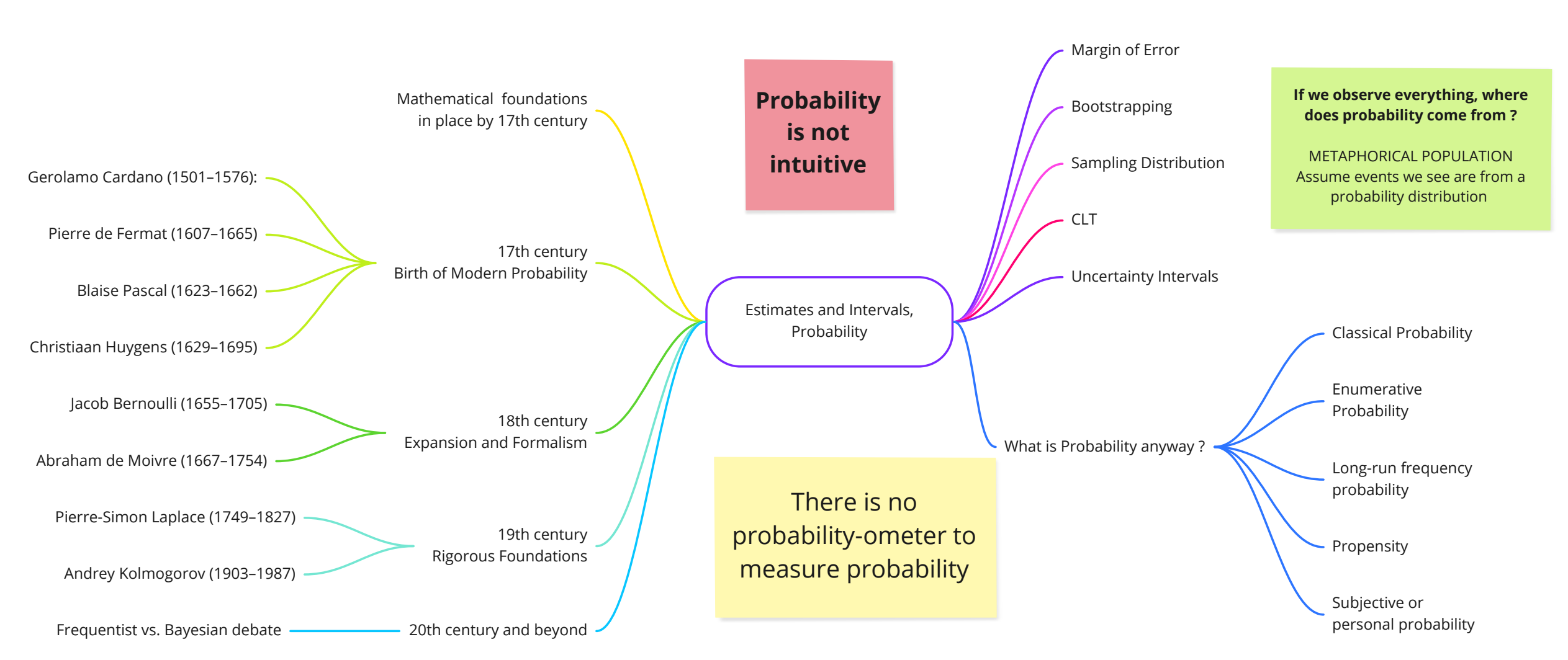


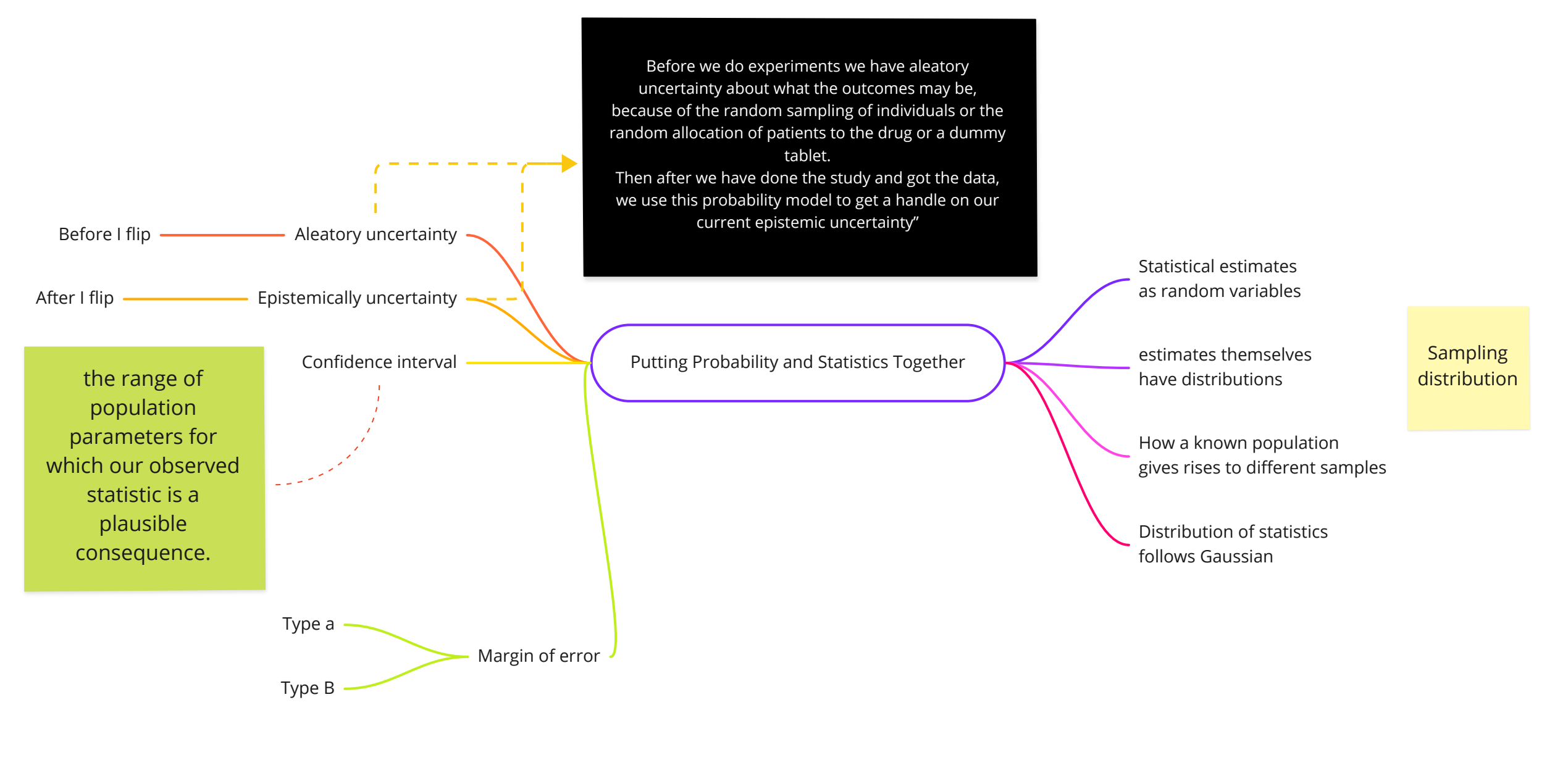
Bias crops up everywhere

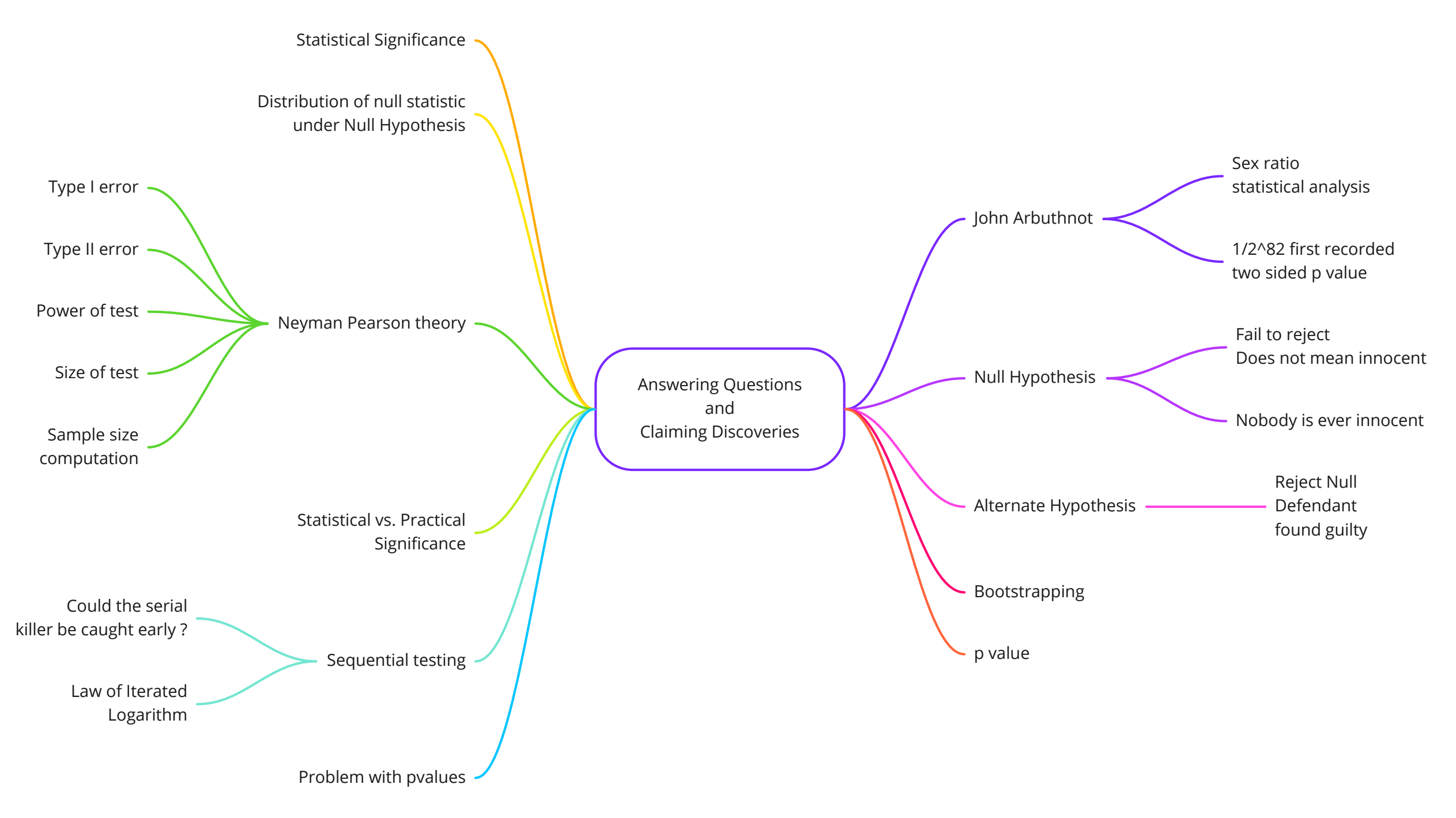



When we say that X causes Y, we do not mean that every time X occurs, then Y will too. Or that Y will only occur if X occurs. We simply mean that if we intervene and force X to occur, then Y tends to happen more often. So we can never say that X caused Y in a specific case, only that X increases the proportion of times that Y happens.
What Causes What ?
What is causation ?
When we can't randomize
Prospective cohort study
Retrospective cohort study
Case Control study
Association checks
Confounder
Simpson's paradox
Reverse Causation
Lurking factors
Clinical Trial Principles
Controls
Allocation of Treatment
People should be counted in the groups to which they are allocated
People should not even know which group they are in
Groups should be treated equally
Those assessing the final outcomes should not know which group the subjects are in
Measure everyone
Don't rely on a single study
Review the evidence systematically


Algorithms
Analysis + Prediction
Algo Performance
Error Matrix
Sensitivity
Specificity
ROC Curves
Brier Score
RMSE
Overfitting
Bias variance
trade off
Cross validation
Two broad classes
of algos
Classification
Prediction
Finding patters
Unsupervised learning
feature engineering
Algo Challenges
Lack of robustness
Not accounting for
statistical variability
Implicit bias
Lack of transparency


Probability is not intuitive
Estimates and Intervals, Probability
Mathematical foundations in place by 17th century
17th century Birth of Modern Probability
Gerolamo Cardano (1501–1576):
Pierre de Fermat (1607–1665)
Blaise Pascal (1623–1662)
Christiaan Huygens (1629–1695)
18th century Expansion and Formalism
Jacob Bernoulli (1655–1705)
Abraham de Moivre (1667–1754)
19th century Rigorous Foundations
Pierre-Simon Laplace (1749–1827)
Andrey Kolmogorov (1903–1987)
20th century and beyond
Frequentist vs. Bayesian debate
Margin of Error
Bootstrapping
Sampling Distribution
CLT
Uncertainty Intervals
What is Probability anyway ?
Classical Probability
Enumerative Probability
Long-run frequency probability
Propensity
Subjective or personal probability
If we observe everything, where does probability come from ?
METAPHORICAL POPULATION
Assume events we see are from a probability distribution
There is no probability-ometer to measure probability


Before we do experiments we have aleatory uncertainty about what the outcomes may be, because of the random sampling of individuals or the random allocation of patients to the drug or a dummy tablet.
Then after we have done the study and got the data, we use this probability model to get a handle on our current epistemic uncertainty”
Before I flip
Aleatory uncertainty
After I flip
Epistemically uncertainty
Confidence interval
the range of population parameters for which our observed statistic is a plausible consequence.
Type a
Type B
Margin of error
Putting Probability and Statistics Together
Statistical estimates as random variables
estimates themselves have distributions
How a known population gives rises to different samples
Distribution of statistics follows Gaussian
Sampling distribution


Answering Questions
and
Claiming Discoveries
Statistical Significance
Distribution of null statistic
under Null Hypothesis
Neyman Pearson theory
Type I error
Type II error
Power of test
Size of test
Sample size
computation
Statistical vs. Practical
Significance
Sequential testing
Could the serial
killer be caught early ?
Law of Iterated
Logarithm
Problem with pvalues
John Arbuthnot
Sex ratio
statistical analysis
1/2^82 first recorded
two sided p value
Null Hypothesis
Fail to reject
Does not mean innocent
Nobody is ever innocent
Alternate Hypothesis
Reject Null
Defendant
found guilty
Bootstrapping
p value

Facts or quantities
exist out there in the world
but we just do now know what they are

Unemployment rate

Migrant workers

Number of tigers
left in the world

Examples

B

Epistemic uncertainty
examples

Bayesian Way

Express our personal ignorance
in terms of probabilities

Subjective probabilities

They depend on
our relationship
with the outside
word, and not
properties of the
world itself

Continuously
revise our current
probabilities in the
light of new evidence

Data does not speak for itself.
External knowledge + Judgement plays
a central role

Bayes theorem

Initial odds * likelihood ratio =
final odds

posterior odds=
prior odds * likelihood

Odds

Likelihood ratio

Why Bayes in not allowed
in courts ?

Individual likelihood ratios
are allowed

Multiplying prior odds
with likelihood is left to the jury

MLP
Multi-level regression
and post-stratification

Subjective vs. Objective Priors

Bayes factor	Strength of evidence
1 to 3	not worth more than a bare mention
3 to 20	positive
20 to 150	strong
>150	very strong

Deliberate Fraud

Questionable Research Practices

Exploratory Studies

Confirmatory Studies

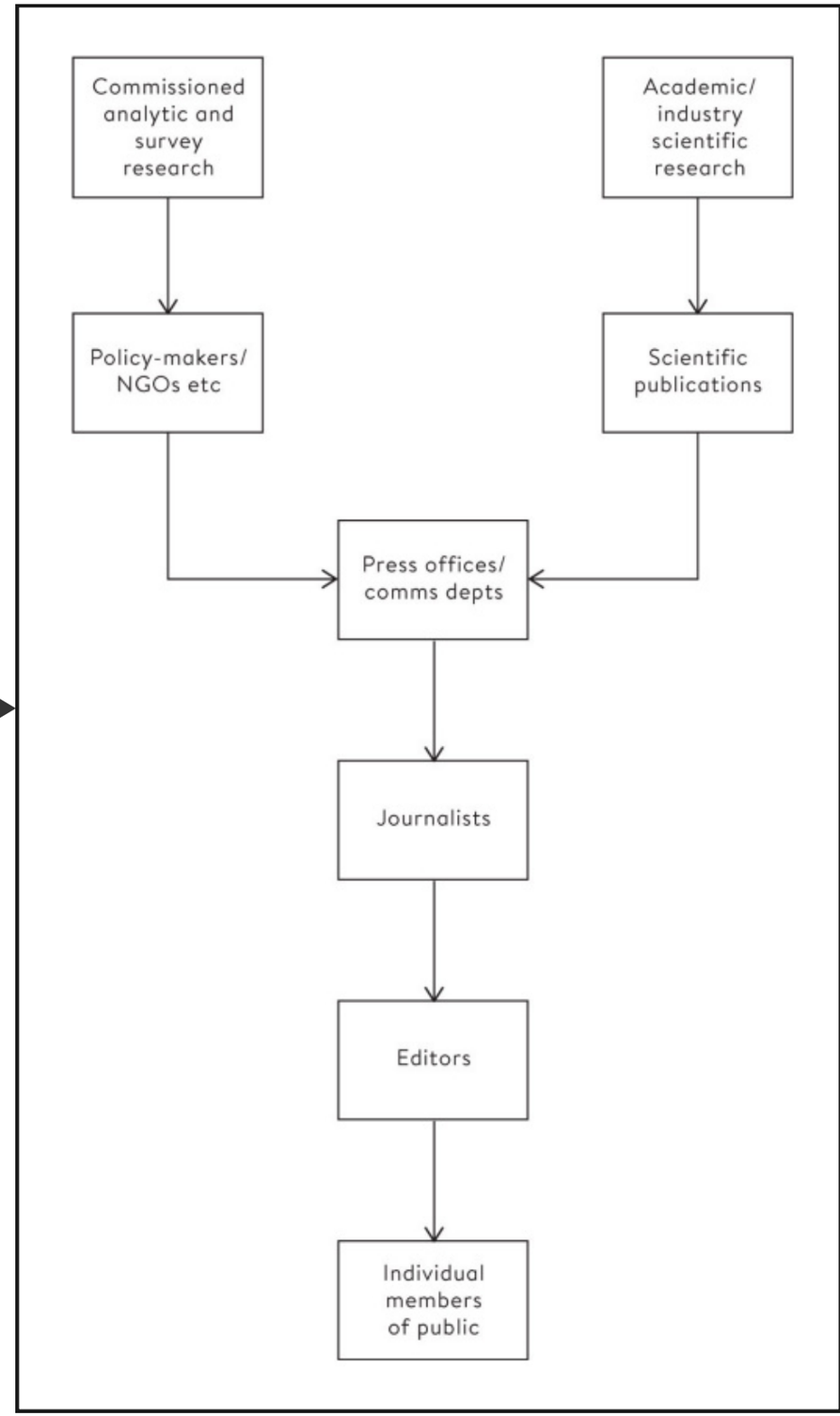
How things go wrong ?

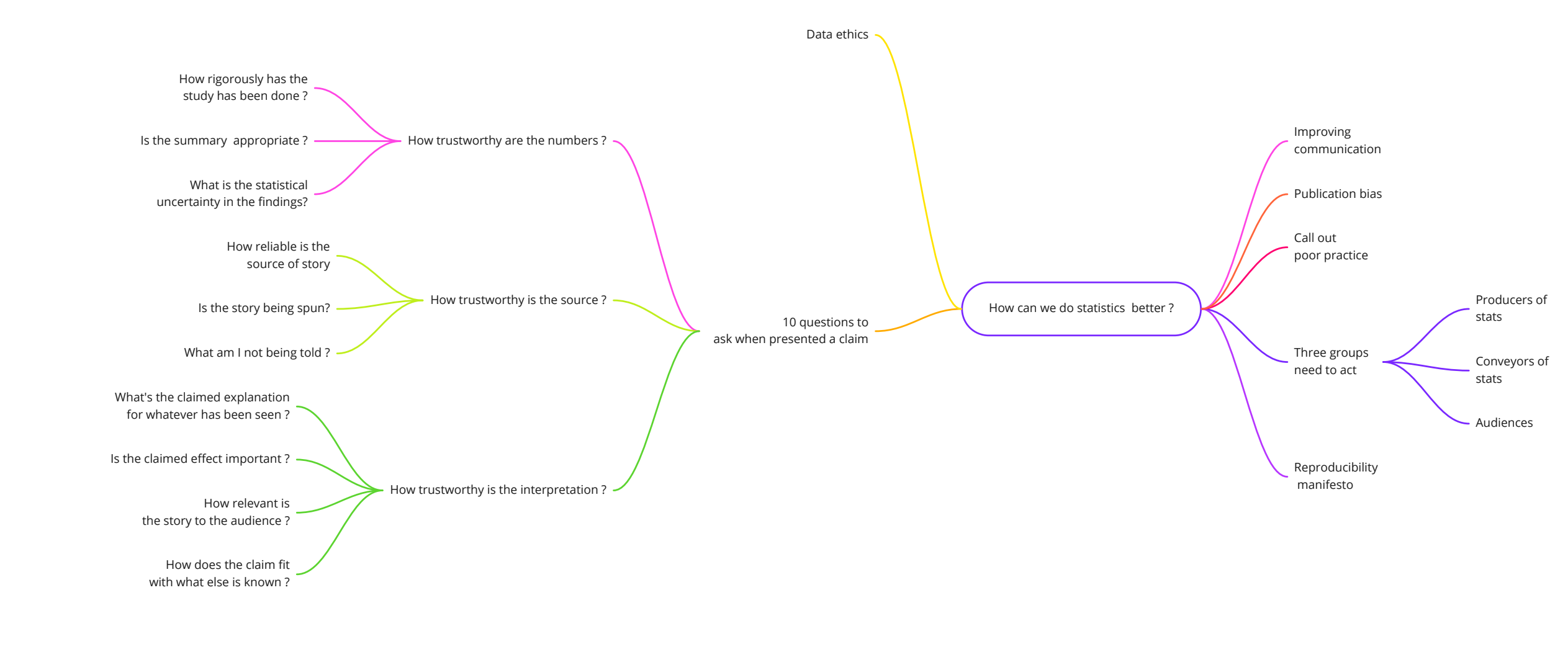
Reproducibility Crisis

Can go wrong at every stage PPDAC

Info pipeline

At each stage there are filters arising from questionable research, interpretation and communication practices, such as selective reporting, lack of context, exaggeration of importance,




How can we do statistics better ?
Improving communication
Publication bias
Call out poor practice
Three groups need to act
Producers of stats
Conveyors of stats
Audiences
Reproducibility manifesto
10 questions to ask when presented a claim
Data ethics
How trustworthy are the numbers ?
How rigorously has the study has been done ?
Is the summary appropriate ?
What is the statistical uncertainty in the findings?
How trustworthy is the source ?
How reliable is the source of story
Is the story being spun?
What am I not being told ?
How trustworthy is the interpretation ?
What's the claimed explanation for whatever has been seen ?
Is the claimed effect important ?
How relevant is the story to the audience ?
How does the claim fit with what else is known ?


PROBLEM
• understanding and defining the problem
• how do we go about answering this question
PLAN
• what to measure & how?
• study design?
• recording?
• collecting?
DATA
• collection
• management
• cleaning
ANALYSIS
• sort data
• construct table, graphs
• look for patterns
• hypothesis generation
CONCLUSION
• interpretation
• conclusions
• new ideas
• communication