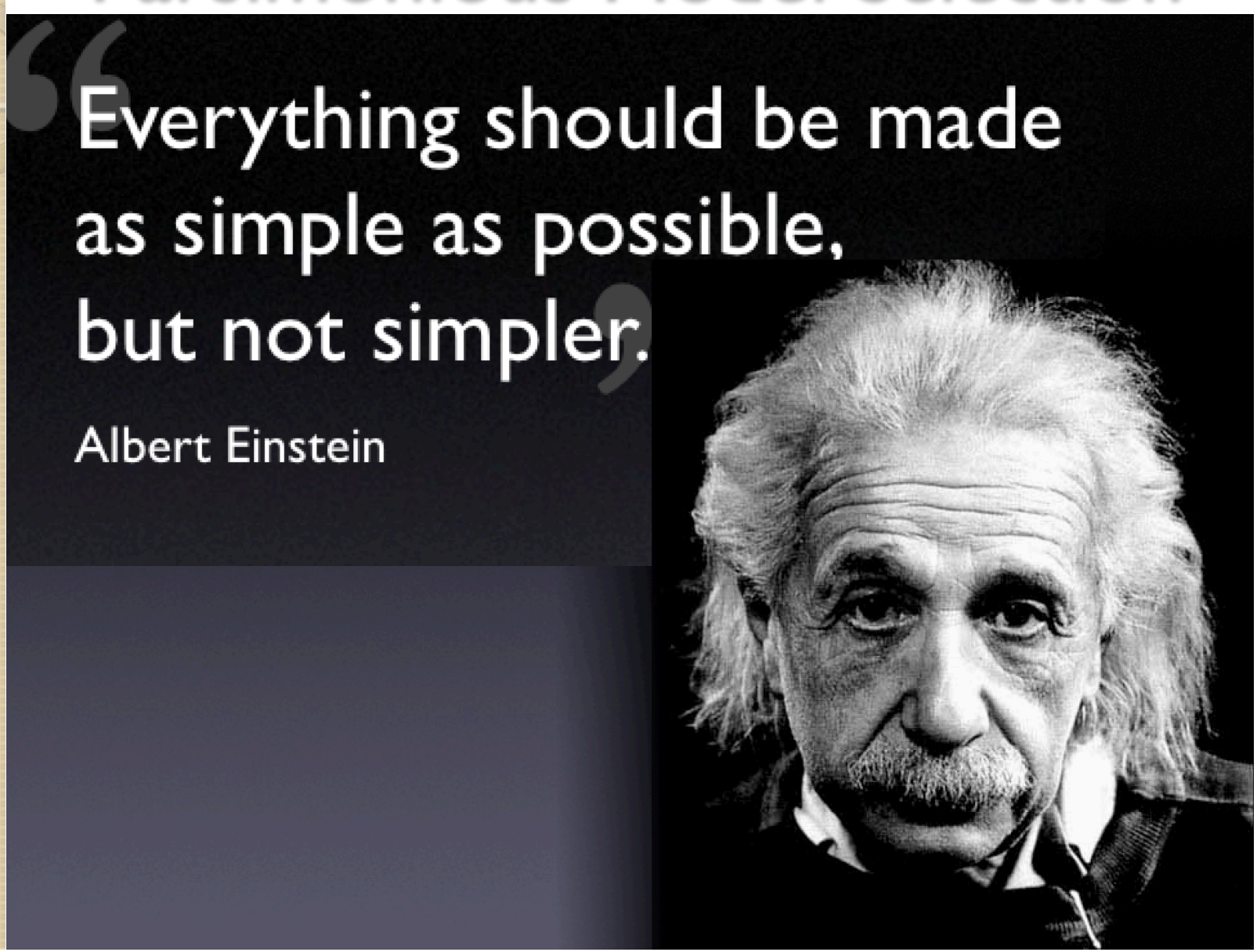
Parsimonious Model Selection

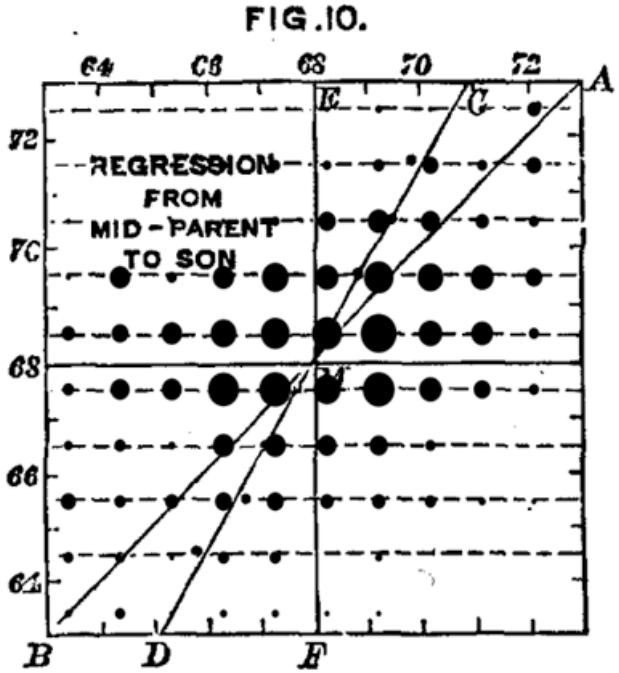
- Models must be statistically & scientifically valid
- Higher complexity must be justified
- Sometimes data will not support complex models, nor are simple models valid
  - No good solution exists!

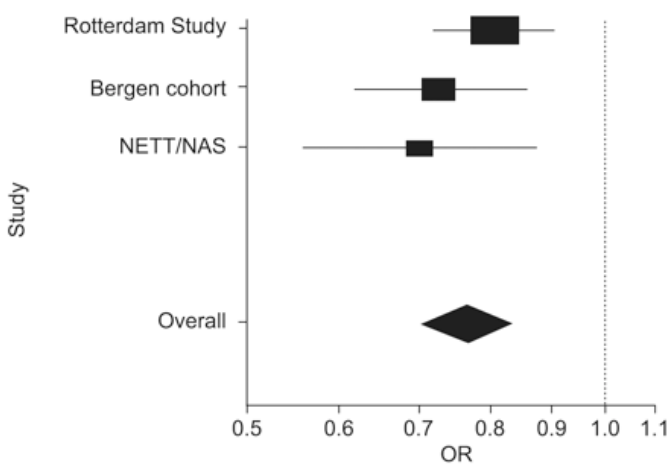
Parsimonious Model Selection



Element 9

Interpretable quantification of evidence





Interpretable Quantification of Evidence

"This only is certain,
that there is nothing certain"

Pliny ("The Elder") (23-79) Roman naturalist. (Gaius Plinius Secundus).

Interpretable Quantification of Evidence

- Uncertainty should be expressed
  - P-values, confidence intervals, posterior probabilities
  - 100% Sensitivity & Specificity
- Multiple Hypothesis testing
  - Genomic Studies
  - Subgroup Analyses

Interpretable Quantification of Evidence

- ISIS-2 Trail 1988, studied ASA & streptokinase post MI
- Editors requested a subgroup analyses be performed

- Authors then studied 12 different subgroups
  - Found ASA did not benefit 2/12 groups

-Gemini & Libra did not benefit from ASA

Element 10

Avoidance of misinterpretation

Avoidance of Misinterpretation

"When distant and unfamiliar and complex things are communicated to great masses of people, the truth suffers a considerable and often a radical distortion. The complex is made over into the simple, the hypothetical into the dogmatic"

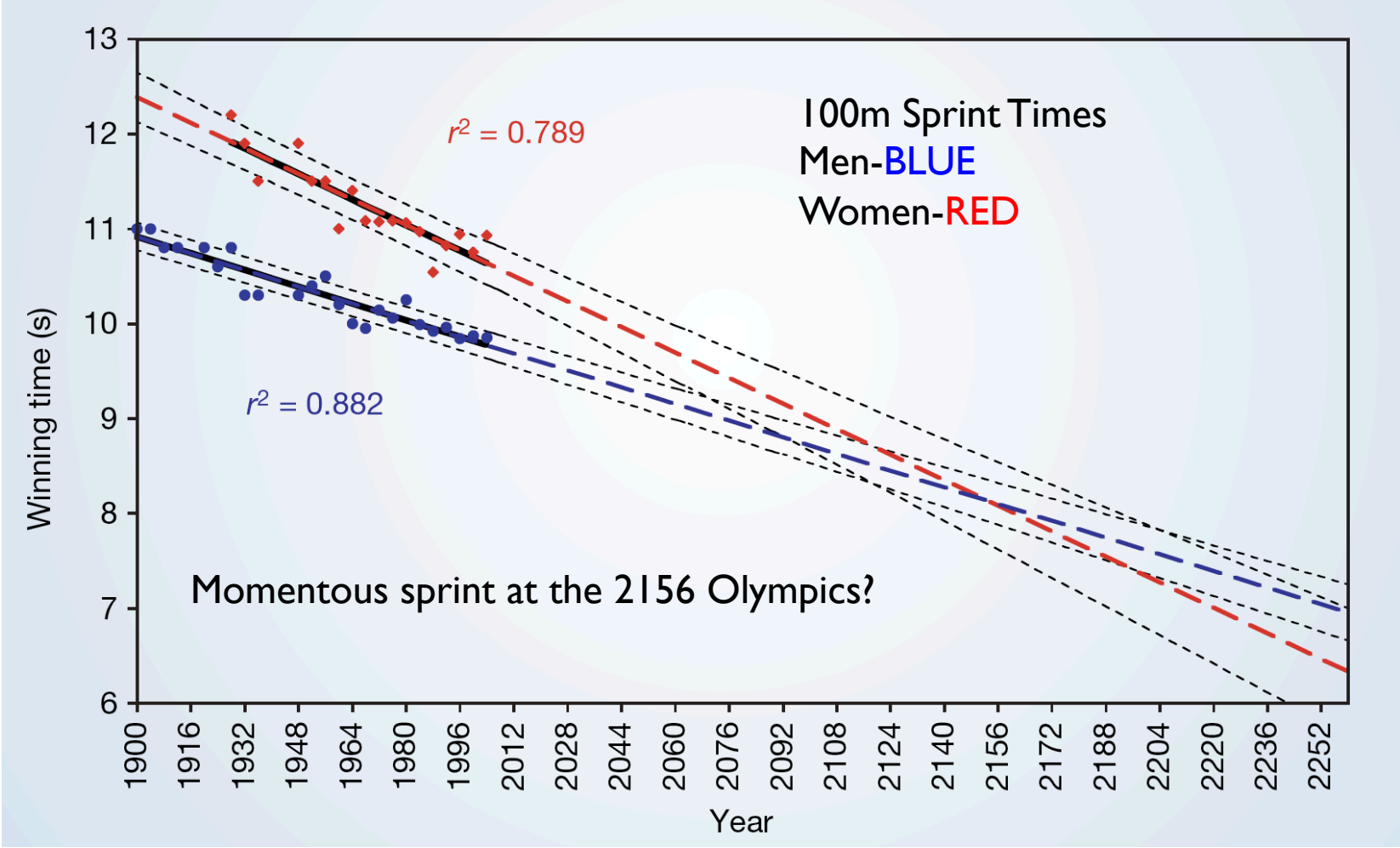
-Walter Lippmann

Avoidance of Misinterpretation

- Claims not supported by evidence are harmful

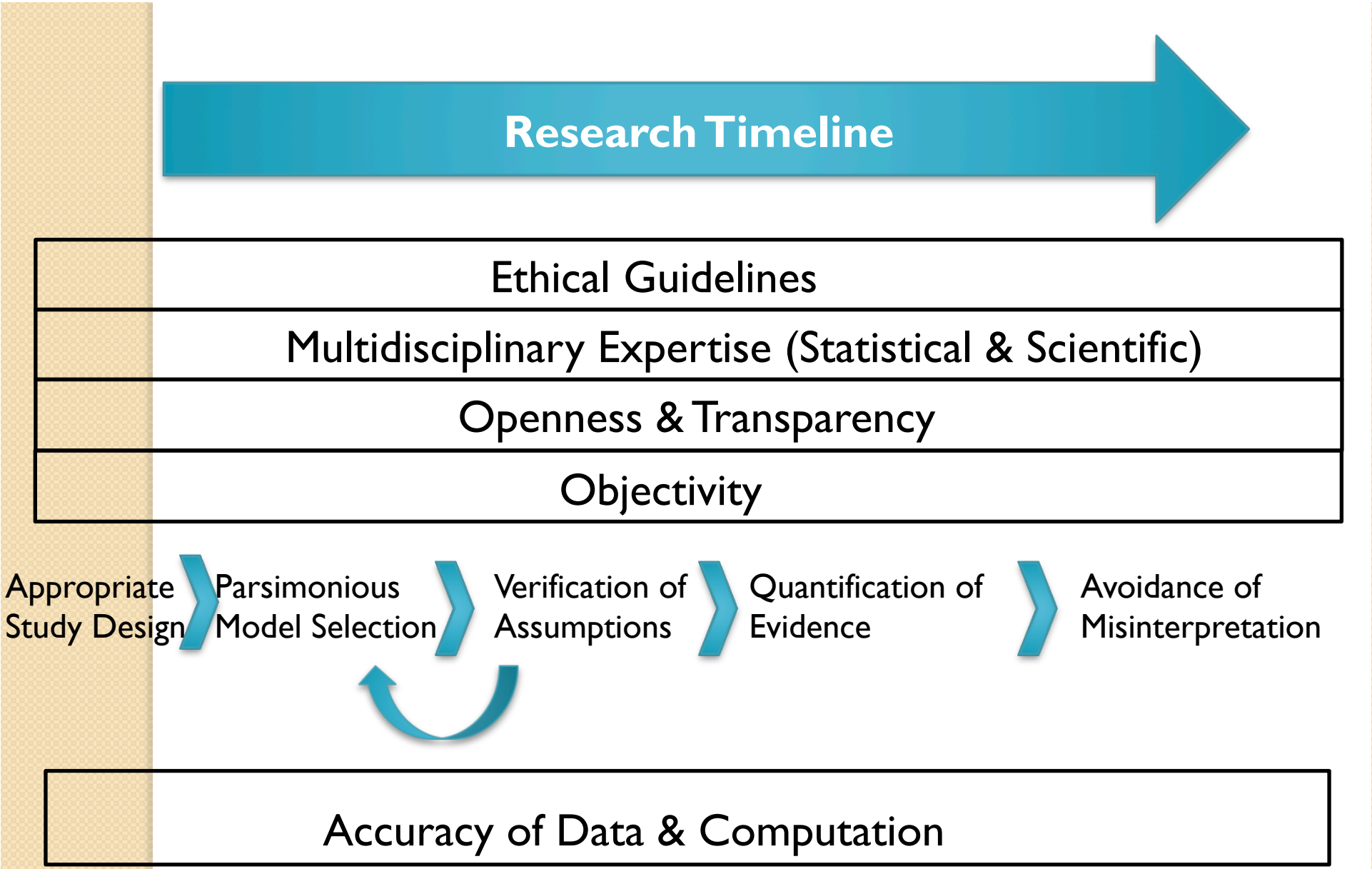
- Interpretation of statistical results occurs ***after*** analyses by
  - Authors
  - Editors
  - Journalists
  - Practitioners

Avoidance of Misinterpretation



Avoidance of Misinterpretation

- Boutron et al *JAMA*(2010)
- Study of Negative Randomized Trials
- Use of spin rampant (40%):
  - Focusing on significant results for subgroups or secondary endpoints
  - Claiming equivalence of therapies
  - Focusing on effectiveness of therapies instead of the failed attempt at demonstrating comparative superiority



Gelfond J, Heitman E, Pollock B, Klugman C: Principles for the Ethical Analysis of Clinical and Translational Research. *Statistics in Medicine* 2011, 30(23):2785-2792.

Analytical vs. Data Quality

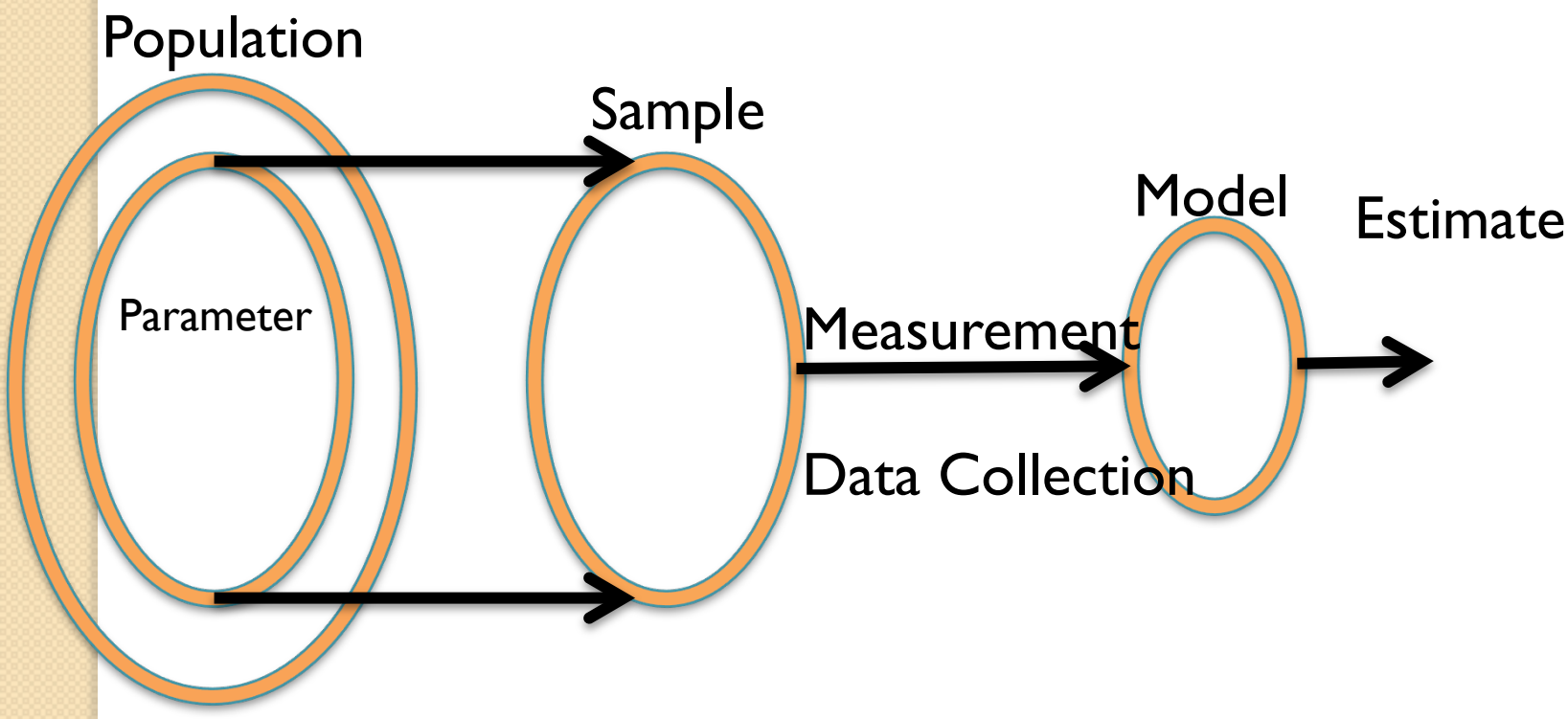
- What separates good analyses from bad?
- Good ones have essential features that exist in quality space
- Analysis is refined data
- Dimensions of data quality were laid out by (Kahn et al, 2002) *Communications of the ACM* April 2002/Vol. 45, No. 4

General Data Quality

Data Quality Metrics	
Accessibility	**Interpretability**
Appropriate Amount of Information	**Objectivity**
Believability	Relevancy
Completeness	Reputation
Concise Representation	Security
Consistent Representation	Timeliness
Ease of Manipulation	**Understandability**
Value-Added	

Estimation vs. Data Analysis

STATISTICAL ESTIMATION

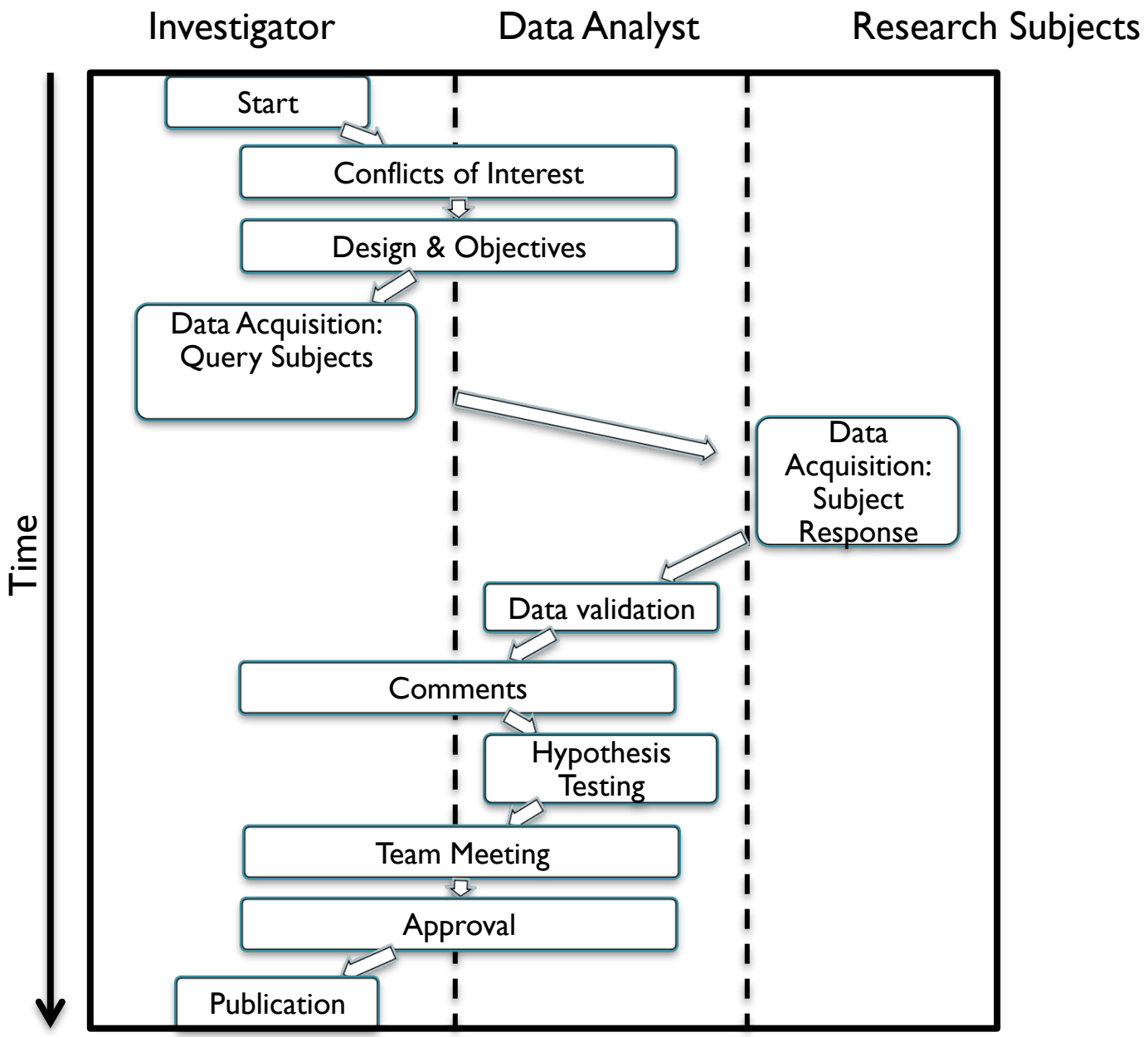


Statistical Estimation is a Mathematical & Numerical Construct

Research Data Analysis

- The processing of information to gain useful knowledge towards some objective
- Requires an Aim or Objective
- Requires and Analytical Plan
- Requires information not just data
  - Metadata, Context & Reasonableness
  - Not exclusively Focused
- Requires interpretation
- Requires teamwork and communication

Research Data Analysis



What is Accountability?

- If we give to or act on another, we are responsible for those actions.
- Accountability is the ability to give a narrative of one's actions
- "Power tends to corrupt, and absolute power [lack of accountability] corrupts absolutely."—Lord Acton

Accountability

- Starts with who did what, when, where, why and how.
  - O&T and documentation
  - Large expansion in documentation
- Gives weight of
  - Personal responsibility
  - At lease the threat of oversight

Openness & Transparency

- Documentation is currently poor
  - Document the good the bad & the ugly
- Technically challenging
  - Overhead costs
  - Need a dataset "electronic health record"
  - Data Provenance systems
- Ethically challenging
  - Privacy for study participants
  - Confidentiality for researchers

Accountable Systems

- Requirements:
  - There must a system of oversight
  - We need to know who did what, when, where, why, how?
- Formal definitions of auditability
  - CS context of distributed systems
  - Models communications between agents
  - ***Very similar to data analysis!!!

Elements of Accountable Systems

- Non-repudiation
  - Opposite of plausible deniability
  - Notaries & Trusted Third Parties (TTP)
- Provenance Information
  - Data flow path between agents
- Mandatory logging and responsiveness.
  - Honest agents need to take communicate their actions

Model of Accountable Data Analysis

- A trusted third party contains all communications between agents related to the analysis of a particular (set of) dataset(s).

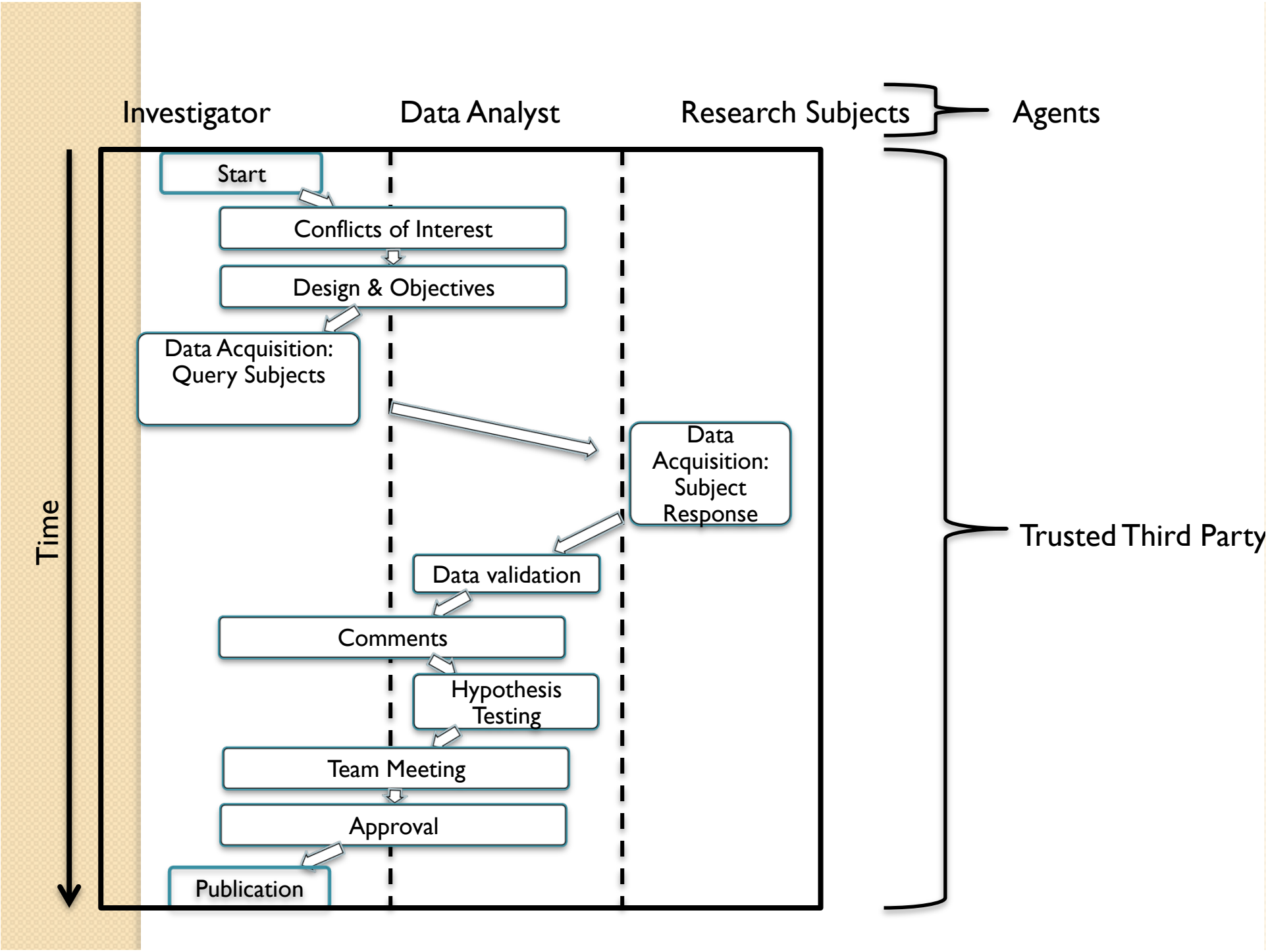
- There are several types of agents
  1. Experimental Units (Research Subjects, Instruments)
  2. Researchers (Supervisors, Lab, Data Analysts, Co-authors)
  3. Numerical Computers
  4. Database Computers

Key Types of Communication

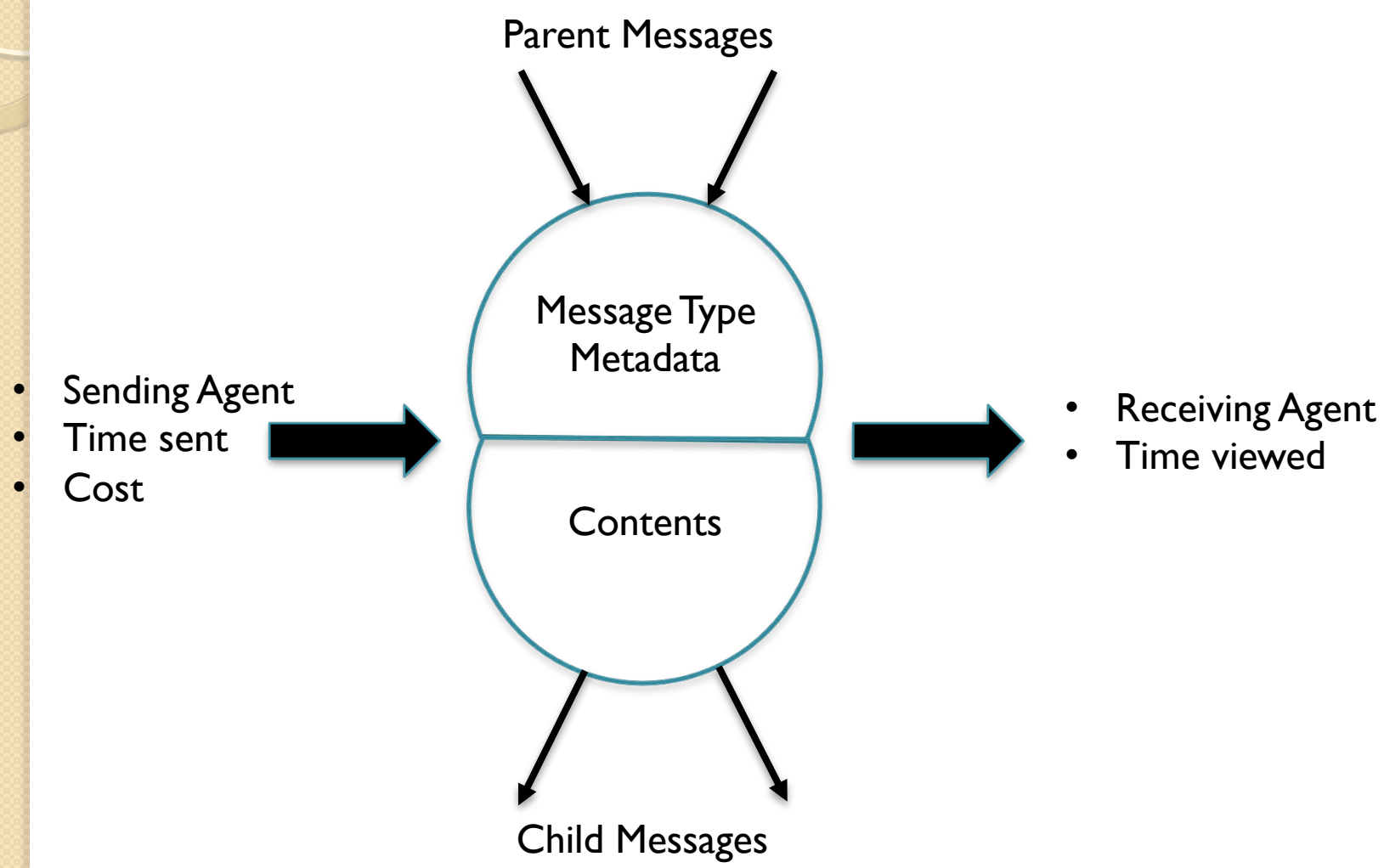
1. **Requests** [Planning is reciprocal Request]
2. **Data Generation** [Data fingerprinting]
3. **1-1 Data transfers**
4. **Computer Programs & Output**
5. **Comments** [Forum for dissent]
6. **Meetings** [SBAR]
7. **Approvals / Revocation**
8. **Publications**

Accountable Data Analysis Process (ADAP)

- ADAP contains
  - Set of agents
  - Set of messages of the 8 types
  - Time sent, who sent, who received
  - Trusted third party storage system
- Open, Transparent, Complete
- Achieved by modern Project Management Systems!
  - Basecamp, Trello, Confluenece, Open Science Framework

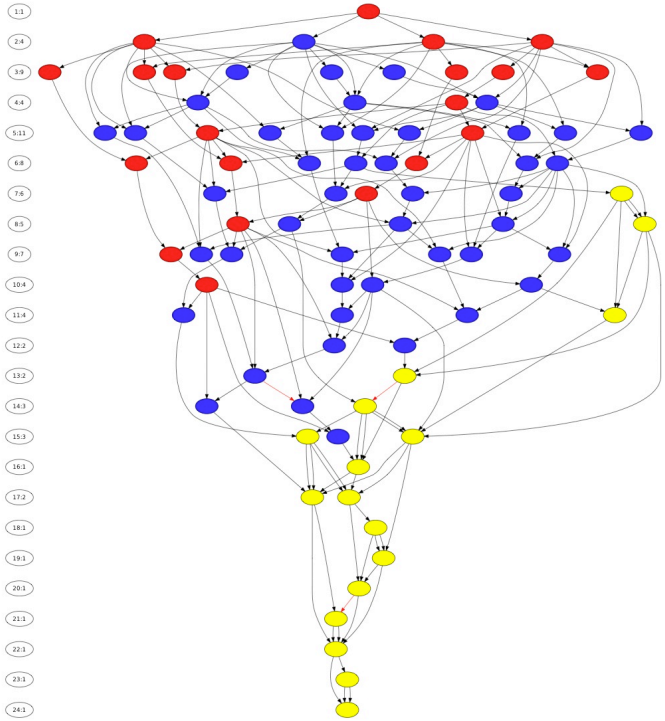

Investigator
Data Analyst
Research Subjects
Agents
Time
Start
Conflicts of Interest
Design & Objectives
Data Acquisition: Query Subjects
Data Acquisition: Subject Response
Data validation
Comments
Hypothesis Testing
Team Meeting
Approval
Publication
Trusted Third Party

Quantum of Communication



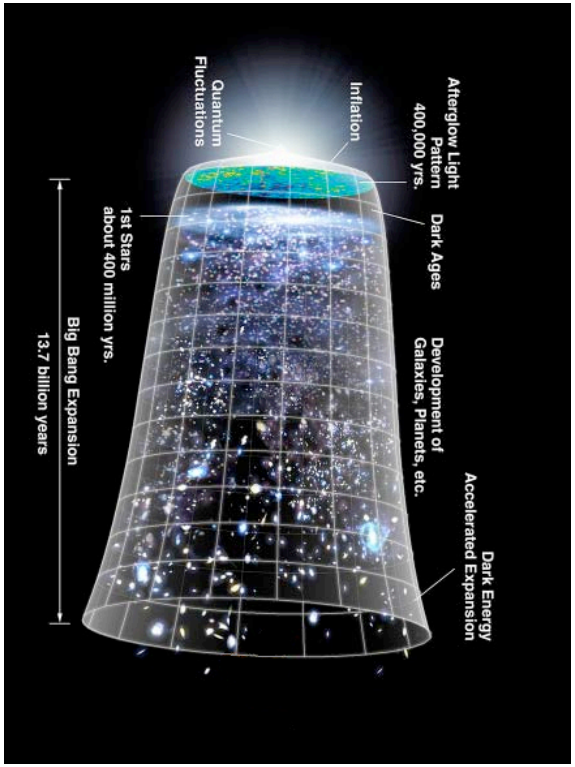
Expanding Universe of a Dataset

DA Data Structure



Time

Big Bang Analog



http://snap.lbl.gov/science/darkenergy.php

Good vs. Bad Accountability

- Not all accountability is "good"
  - See Lerner and Tetlock, 1999
- Need auditors who…
  1. Have unknown views
  2. Are knowledgeable
  3. Are interested in accuracy
  4. Are interested in process
  5. Have legitimate reasons for inquiries
- Most importantly, we rely on mutual accountability of our research team

My Accomplices

Craig Klugman PhD
DePaul University

Liz Heitman PhD
Vanderbilt University

Brad Pollock MPH, PhD
UT Health Science Center

Special thanks to
Chris Louden, UT Health Science Center
Leah Welty, Northwestern University
National Institutes of Health Award Number 8UL1TR000149
National Center for Advancing Translational Sciences.