GSF6011 Research Methodology

Data Collection

Jane Labadin

Caveat

- I’m just a mathematician
- Always a first time ☺
- Deliver based on experience

original information which is *collected*, *stored*, *accessed*, *used* or *disposed of* during the course of the research, and the final report of the research findings

What is data?

- Can you do research without data?
- How can you resolve the research problem without supporting data?
- How do you convince others, that your data are sufficient to support the solution?
- Where do you go to find data?
- Can you have imaginary data in research?
- Can you have data simulation for research?

Motivation

collection of information (data) which can be **interpreted** or **analysed** to frame answers to your research questions or increase knowledge of your research topic.

Purpose of Collecting Data

1. **Sources** – where will you get the information

2. **Methods** – how will you collect/gather the information

What does collecting data involves?

- From where or from whom will you get the information?
  - Existing information – records, reports, program documents, logs, journals
- People – respondents or informants
- Pictorial records and observations – video or photos, observations or events, artwork, etc.

SOURCE of Information

METHODS of data collection

- Survey
- Interview
- Observation
- Group assessment
- Expert or peer reviews
- Portfolio reviews
- Testimonials
- Physical evidence
- Time series
- Tests
- Photographs, videos
- Diaries, journals, logs
- Document review and analysis

Quantitative Information

- Obtained via
  - surveys of populations
  - repeated experimental procedures
  - …
- When recording, include
  - detailed information
    - Dates
    - place of collection
    - methods of measurement
    - units of measurement) to minimise confusion
- Recorded on
  - printed datasheets, then stored in spreadsheet format.
- In some cases, data may be recorded by handheld computers or specialised data recorders which can later be downloaded to more secure devices.
- Data recorders can often be set up to record data remotely, without the requirement that researchers be present. Such techniques are frequently used in meteorological research or in situations where it would be too hazardous for a researcher to be present (eg industrial chemistry applications, space research).

Qualitative Information

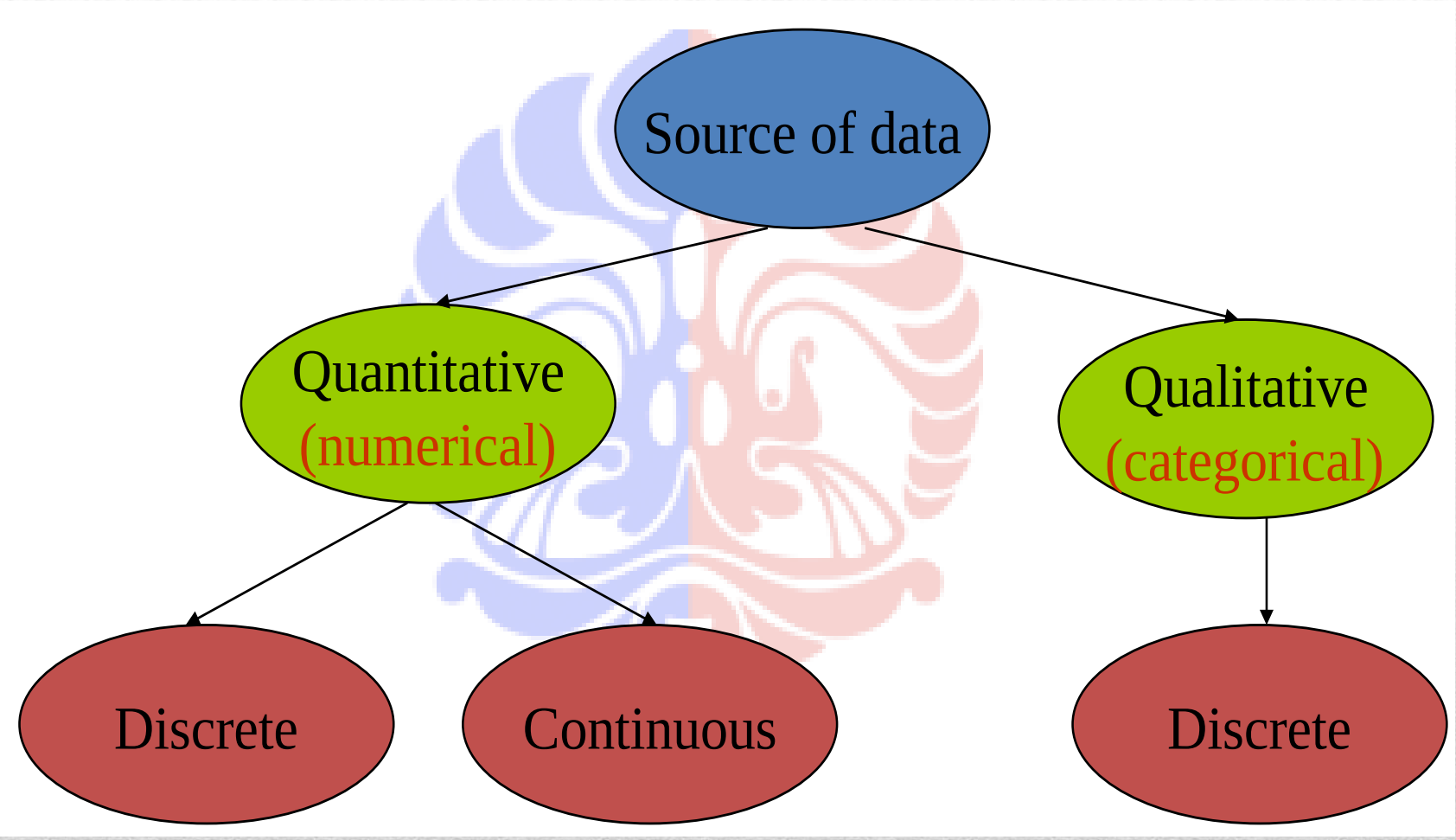
- May be in the form:
  - Recordings of interviews
    - Transcribed into written form
    - Supporting notes
- Description of text/artefacts/system
- Interpretation of text/artefacts/system

- **Source** of Data

- **Quantitative data** are values on a numerical scale

- **Qualitative data** are observation measured on a numerical scale

In a nutshell


Source of data
Quantitative
(numerical)
Qualitative
(categorical)
Discrete
Continuous
Discrete

- Discrete Data
  – Only certain values are possible (there are gaps between the possible values)
- Continuous Data
- – Theoretically, any value within an interval is possible with a fine enough measuring device

Quantitative data

Types of Data

- **Primary data**: data observed and recorded or collected directly from respondents
- **Secondary data**: data complied both inside and outside the organization for some purpose other than the current investigation

Types of Data

Primary
Data Collection

- Observation
- Survey
- Experimentation

Secondary
Data Compilation

- Print or Electronic

Ratio Data	Differences between measurements, true zero exists	Height, Age, Weekly Food Spending
Interval Data	Differences between measurements but no true zero	Temperature in Fahrenheit, Standardized exam score
Ordinal Data	Ordered Categories (rankings, order, or scaling)	Service quality rating, Standard & Poor's bond rating, Student letter grades
Nominal Data	Categories (no ordering or direction)	Marital status, Type of car owned

Basic Business Statistics 10e, 2006 Prentice Hall

Categorical Data

Take a step back …

- Information?
- What? Where? How?

Things to consider

- Availability
- Training/expert assistance
- Pilot testing
- Interruption potential
- Protocol needs
- Reactivity
- Bias
- Reliability
- Validity

Sampling

- How much information should you collect?
- Selecting a portion of subjects in order to learn something about the entire population without having to measure the whole group.
- Random? Size? Bias?

“The natural state of most engineering information contains significant variability”

Ang and Tang, Probability Concepts in Engineering Planning and Design

Research Methods in Engineering & Computer Science

VS.

Qualitative Research

- Concentrates on collecting and analyzing subjective data
- Usually the perceptions of the people involved
- Intention is to illuminate perceptions and, thus, gain
  - greater insight (explain why) and
  - Knowledge (reproduce or recognize).

Quantitative Research

- Concentrates on what can be measured.
- Involves collecting and analyzing objective data
- Usually involves some form of math
  - Statistical
  - Calculus
  - Discrete

Research Methods

What is this monkey doing?

Quantitative vs. Qualitative

- Quantitative

  - We have an hypothesis that monkeys will put bananas to their ears
  - We gave bananas to monkeys
  - If we say banana to ear == “Monkeycide”
  - We counted xx instances of Monkeycide over yy trials
  - Our hypothesis is accepted if xx > 0
- Qualitative
  - We saw monkeys pick up bananas
  - We observed the monkeys placing bananas to their ears
  - From observation we have the concept: “Monkeycide”
  - Monkeys Jenny, Irene and Blake exhibited Monkeycide

What is your research?

An experimental design:

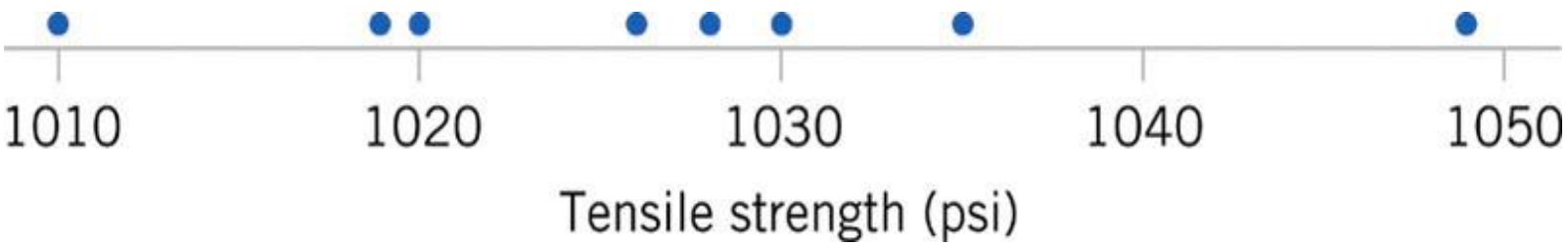
"Is the tradi?onal approach to conduc?ng quan?ta?ve research"

Creswell, J.C. 2005

Branch of Quantitative Research

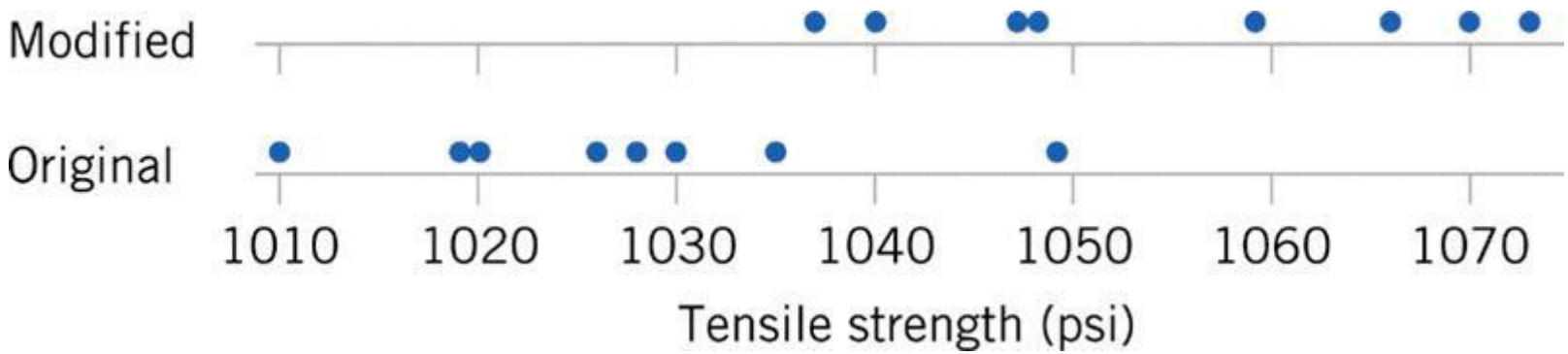
An Example…

An engineer is tasked to develop a rubber compound for use in O-rings which are to be used as seals in plasma etching tools. Resistance to acids and other corrosive substances is important! Using a standard rubber compound, the engineer produces 8 O-rings and tests their tensile strength after immersion in nitric acid. She obtains the following results, plotted as a *dot-diagram*:



These measurements contain *variability*. What is the cause of the variability??

Now, the engineer is not happy with the *location* and *scatter* (variability) of the data. The *mean* tensile strength is too low and the variability too high. She decides to modify the formulation with a Teflon additive, makes eight new O-rings, and tests those.



- Will another set of O-rings give yet another set of results?
- Is a sample size of 8 adequate to give reliable results?
- What risks are associated with the assumption that the Teflon additive leads to increased tensile strength?

Statistical Inference can help us answer these questions.

Data Collection Method 1: Retrospective Study

- Go back in time and analyze data that has already been collected over some period.
- Might be interested in using this data to construct an *empirical model*.

Pros: cost of collecting data minimized by taking advantage of existing data

Cons: Because the analyst was not involved in data collection, often appropriate data is not collected.

For example: In the distillation column, operators don't ever change the reflux rate. Since reflux rate variability is small over time, it will be difficult to tell if it affects the final concentration.

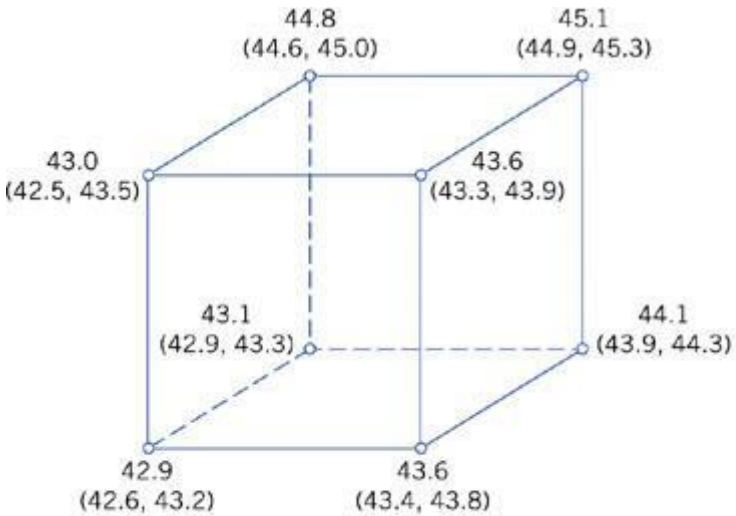
Other Cons: Missing data, reliability questionable, irrelevant data, inappropriate use of data, not enough accompanying info

Data Collection Method 2: Observational Study

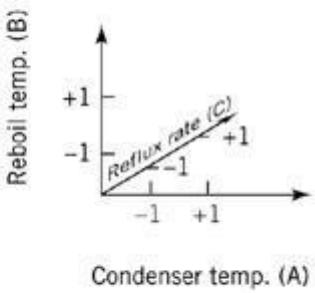
- Real time record of data, with no further interference.
- Engineer creates a data collection form for operators to fill out at specified times – leaves space for comments on anything unusual that may have occurred.
- Engineer still stuck with the problem that the system, as is, may not provide information relevant to the question at hand.

Data Collection Method 3: Designed Experiments

- In a designed experiment, the engineer makes purposeful changes in controllable variables (factors) of the system and observes the resulting system output.
- He/she then makes a decision or an inference about which variables are responsible for the changes in output performance observed.



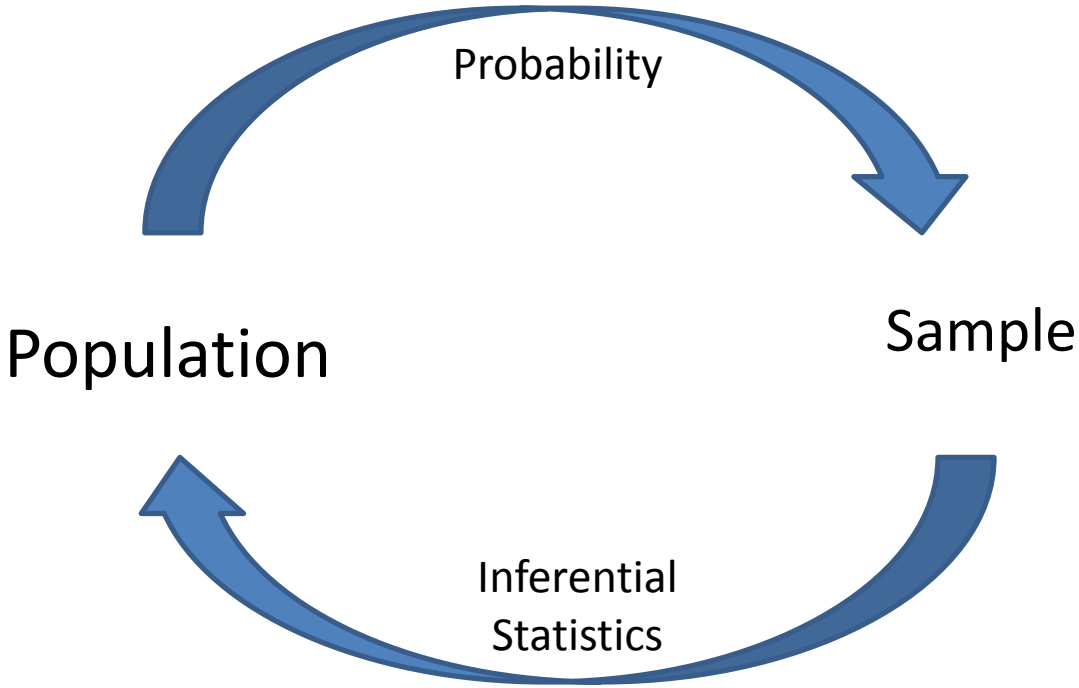
(*a*) Geometric view



	Factors		
Run	A	B	C
1	−	−	−
2	+	−	−
3	−	+	−
4	+	+	−
5	−	−	+
6	+	−	+
7	−	+	+
8	+	+	+

(*b*) Design or test matrix

- Linking *probability* and *statistics*…



- Populations can be *physical* or *conceptual*.
- Samples should be *random*, and not based on judgment or convenience
  - *Simple random sample*: a sample of size n that has been selected from a population in such a way that each possible sample of size n has equally likely chance of being selected.

Statistics and Models

- Sometimes our models are very physical in nature (*mechanistic model*)

  Modeling the current flow in a thin copper wire

  Current=voltage/resistance

  $I=E/R$ (Ohm's Law)

- Since we know variability exists in measurements, a more realistic model might be

  $I=E/R + \epsilon$

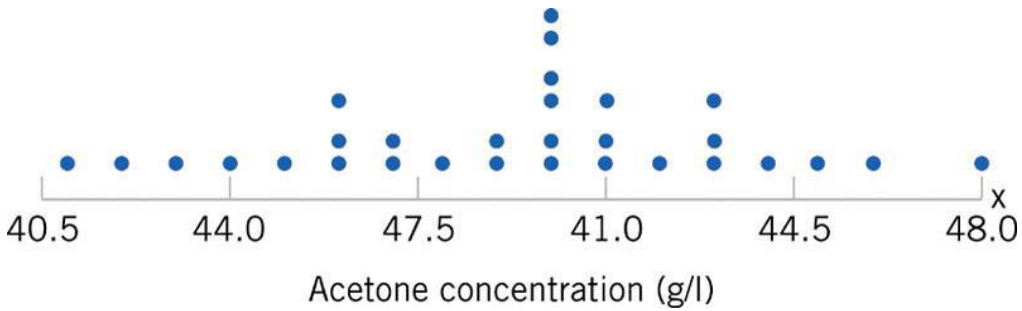
- Sometimes we don't have a physical law to explain the relationship between variables. In this case we develop an *empirical model*.

  Relationship between pull strength (y) to wire length (x1) and die height (x2) for determination of strength of a wire bond in a semi conductor frame

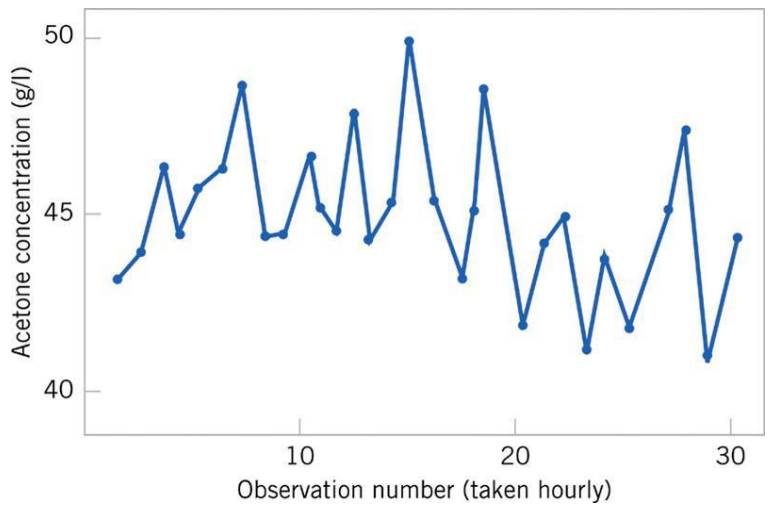
  $y=F(x1,x2) + \epsilon$

Observing processes over time

Plotting data over time allows an engineer to better understand how phenomena affect a system's stability over time.

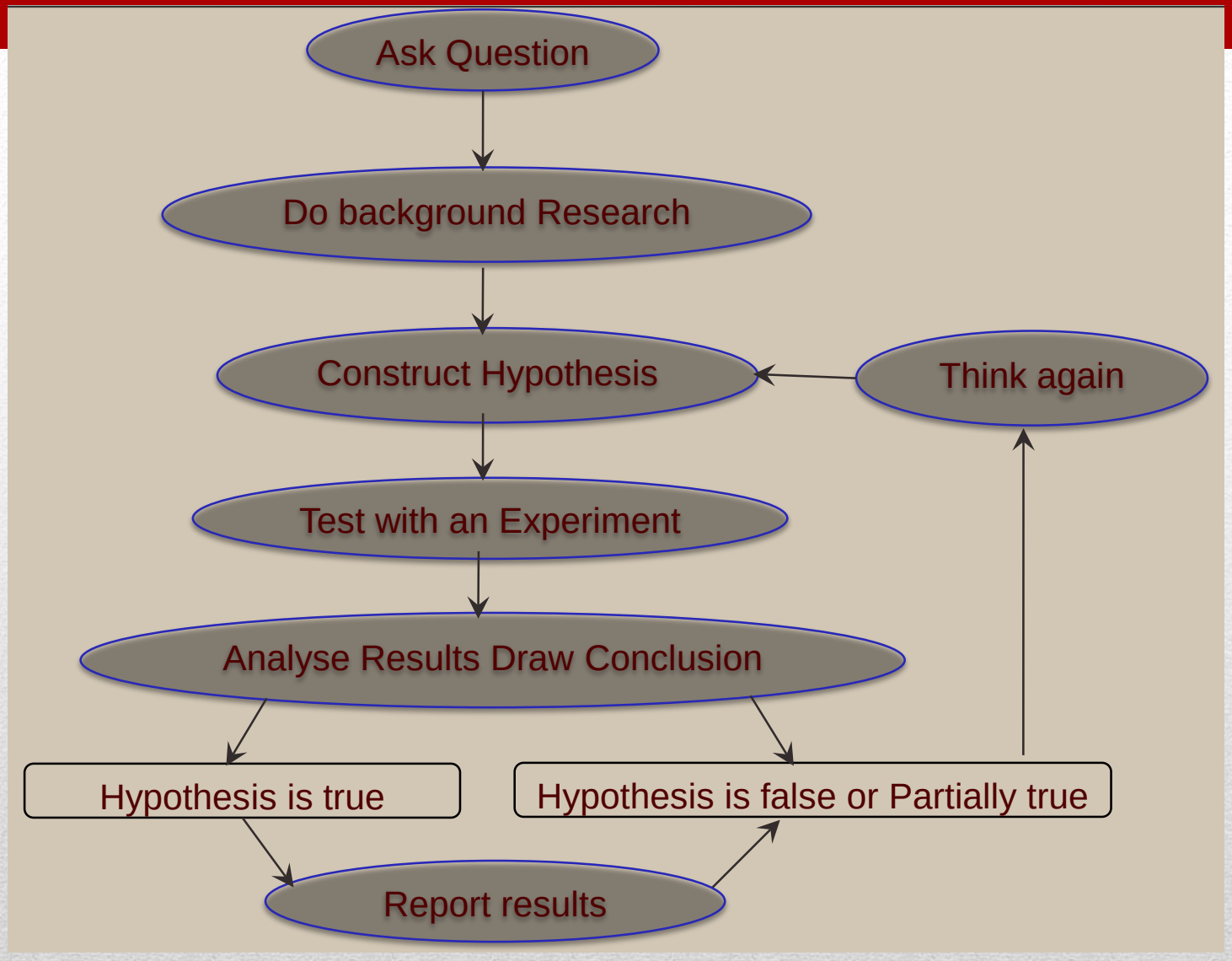


Dot diagram has no time information.



Time series plot gives us another dimension to analyze.

Technique	Used by researchers when their goal is to understand:	Volume of data	Also used by software engineers for
Direct techniques			
Brainstorming and focus groups	Ideas and general background about the process and product, general opinions (also useful to enhance participant rapport)	Small	Requirements gathering, project planning
Interviews and questionnaires	General information (including opinions) about process, product, personal knowledge etc.	Small to large	Requirements and evaluation
Conceptual modeling	Mental models of product or process	Small	Requirements
Work diaries	Time spent or frequency of certain tasks (rough approximation, over days or weeks)	Medium	Time sheets
Think-aloud sessions	Mental models, goals, rationale and patterns of activities	Medium to large	UI evaluation
Shadowing and observation	Time spent or frequency of tasks (intermittent over relatively short periods), patterns of activities, some goals and rationale	Small	Advanced approaches to use case or task analysis
Participant observation (joining the team)	Deep understanding, goals and rationale for actions, time spent or frequency over a long period	Medium to large	
Indirect techniques			
Instrumenting systems	Software usage over a long period, for many participants	Large	Software usage analysis
Fly on the wall	Time spent intermittently in one location, patterns of activities (particularly collaboration)	Medium	
Independent techniques			
Analysis of work databases	Long-term patterns relating to software evolution, faults etc.	Large	Metrics gathering
Analysis of tool use logs	Details of tool usage	Large	
Documentation analysis	Design and documentation practices, general understanding	Medium	Reverse engineering
Static and dynamic analysis	Design and programming practices, general understanding	Large	Program comprehension, metrics, testing, etc.



Engineering & Computer Science

- Experimentation can help build a reliable base of knowledge and thus reduce uncertainty about which theories, methods, and tools are adequate.
- Observation and experimentation can lead to new, useful, and unexpected insights and open whole new areas of investigation. Experimentation can push into unknown areas where engineering progresses slowly, if at all.
- Experimentation can accelerate progress by quickly eliminating fruitless approaches, erroneous assumptions, and fads. It also helps orient engineering and theory in promising directions.

Walter F. Tichy
University of Karlsruhe

Why experiment?

Ask yourself

- Will I be using a formal, objective, systematic process where data are utilized to test my research hypothesis?
- What are the **variables** I will consider in my study?
  - Independent Variable(s)
  - Dependent Variable(s)
  - Extraneous Variable(s)
- What type of quantitative investigation will I pursue?

Insist on precision only when is needed

Acknowledgement