Introduction to Big Data

Day 5

Model Operations and Research Methods

YORK UNIVERSITÉ UNIVERSITY

Review Previous Lesson

Review Concepts from Day 4

During Day 4 you learned to:

- Describe a range of different analytical modeling techniques
  - Statistical modeling
  - Data mining
  - Machine learning
- Describe the Model Development Process CRISP-DM
  - Business Problem Framing
  - Data Collection and Evaluation
  - Data Preparation
  - Model Building
  - Model Evaluation
  - Model Deployment
- Differentiate and recommend a suitable development process
  - Iterative
  - Waterfall
- Identify the structure and role of Programs, Projects and Services in terms of model development
  - Analytic team composition and structure
- Build and evaluate linear regression models using Python

Day 5 – New Topics Introduced

The following major topics are discussed this class.

- Research concepts
- Scientific method
- Business questions
- Critical thinking
- Hypotheses
- Sampling

Learning Objectives for Day 5

Day 5 - Learning Objectives

- The course content provides some necessary context for operating and using analytical models that drive business impact.
- The material will focus on using research driven approaches to formulate and explore business problems using data analytics.
- The primary purpose of an analytical model is to generate information to a person that is curious and can generate meaningful and challenging questions.
- The ability to generate these questions is based on thinking skills and research methods.
- Benefits to the organization are generated if valuable insights are generated that can be applied to address business challenges based on the interaction between curious people, analytical models and a robust methodology.

Day 5 - Learning Objectives

During Day 5 you will learn to:

- Identify and describe the approach and objectives of research
- Describe two different types of research methods
- Describe how research methods enable successful business impact of data analytics
- Discover the business questions you are trying to answer with data analytics.
- Determine if your data is appropriate to support your analytics initiative
- Describe some data collection techniques and some related sampling concerns
- Recognize how Operational Excellence initiatives drive analytics requirements
- Check your data for some common issues in Python and interpret results from a regression model
- Have a general understanding of what kinds of models for analysis are readily available in Sci-kit Learn Python

What is a Scientific Approach to Geography?

- “Science is a personal and social human endeavor in which ideas and empirical evidence are logically applied to create and evaluate knowledge about reality.” Page 3

- Loosely defined - A scientifically informed geography is one which uses a systematic empirical approach to produce its geographic knowledge.

Theoretical and Empirical symbiosis

- Ultimately both empirical and theoretical approaches must be used for a complete and well rounded research approach
- The balance of theoretical to empirical investigation does not have to be 50/50

Non-Scientific ways of knowing

- It is important to realized that non-scientific ways of seeking knowledge are valuable in seeking issues of morality and human values - which systematic empirical approaches are illequipt for.

- The humanities informally analyzing text and other symbolic artifacts of human thought, activity and culture to search for specific truths about places and times.

Common Metaphysical Beliefs Within Science

- Realism - The universe is “real” outside of our heads.
- Continuously connected and forward causality - Interconnectivity along linear time is needed for causality to be possible as we understand it.
- Simplicity - The desire to use the “simplest” explanation that is adequate. However the relative degree of simplicity may be difficult to determine.
- Skepticism - Scientists falsifying their way towards possible truth.
- Quantitative thinking - Knowledge produced from systematic processes which can be recreated.

Metaphysical Beliefs: The nature of reality and the conceptual framework in which reality exists

- Ontology - The alleged ultimate nature of reality
- Epistemology - How scientists know about reality

Progressive Goals of Science

Description - The intellectual act of classification often carried out in a particularly systematic fashion.

Prediction - Using the findings of empirical data to predict unseen phenomena which are to come in the future or has already happened without leaving a clear record.

Explanation - Why previously described and predicted patterns exist

Control - Using the understanding of phenomena to alter and manipulate. Hopefully the modification is intended for good.

Goals of Science - Basic and Applied

Basic scientific research is concerned with the first three goals of science - description, prediction and explanation. It focuses on understanding reality for its own sake.

Applied scientific research is associated with the forth goal - control. Medical and educational research are applied sciences as the knowledge is actively employed.

History and Development of Geography

The broad and heterogeneous field of geography has changed and evolved throughout its own history.

Chorographic approach: Regional comprehension without global connectivity.

The Quantitative Revolution: We're a *real* science

The Marxist/Feminist rebuttal: We're into equality

The importance of research methodology and question development

"[A]ny empirical observations can potentially be explained not just by ideas about reality but also about the way they obtained or interpreted the observations." - page 4

Research Concepts

Research Concepts Overview

Two major types of empirical research design

- Qualitative research
- Quantitative research

Qualitative Research

- involves the understanding of human behavior and the reasons that govern such behavior
- asks broad questions and collects unstructured data as words, images, video and audio
- data is analyzed and summarized by themes.
- aims to investigate a question without attempting to quantifiably measure variables or look to potential relationships between variables.
- qualitative research is often used as a method of exploratory research as a basis for later quantitative research hypotheses.

Quantitative Research

- empirical investigation of quantitative properties, phenomena and their relationships
- asks narrow questions and collects numerical data to analyze it utilizing statistical methods.
- the quantitative research designs are experimental, correlational, and survey based
- statistics derived from quantitative research can be used to establish the existence of associative or causal relationships between variables.
- quantitative data collection methods rely on random sampling and structured data collection instruments

- Source Wikipedia

Quantitative versus Qualitative

Quantitative Research Strategy

- Investigation aims to assess a pre-stated theory (Deductive Reasoning)
- Often involves hypothesis testing
- Attempts to minimise the influence of the researcher on the outcome
- Quantitative data infers statistics
- Data collection therefore requires 'closed' responses

Qualitative Research Strategy

- Investigation aims to create a novel theory (Inductive Reasoning)
- Researcher becomes an inherent part of the study - *ethnography*
- Qualitative data infers complex statements or opinions
- Data collection therefore permits 'open' responses

Understanding “Understanding”

- What do we mean by “understanding”?
- Who does understanding benefit?
- How do we understand something?
- How does “understanding” differ from “explanation”.

Research Methods

The following two videos provide an overview of Qualitative and Quantitative research methods. These videos provide an introduction to methods are useful when applied to a data analytics project. Please review the two videos.

Qualitative Research

https://www.youtube.com/watch?v=IsAUNs-IoSQ

Quantitative Research

https://www.youtube.com/watch?v=cwU8as9ZNlA

Optional Video

This video provides a discussion that compares and contrasts the use of qualitative vs quantitative research methods

https://ed.ted.com/on/6116s1ZG

Understanding

- Internal
  - It is about “me”
- It deals with the subjective
  - Emotions
  - Feelings
- A concern with quality rather than quantity
- Putting back the human in human geography
  - Broadly humanistic traditions
  - The emergence of a strong cultural and social tradition within the discipline in the late 20th century

Theory and philosophy

- Qualitative methods have been applied across a range of philosophical underpinnings
  - Limb and Dwyer (2001) emphasise their use in many different contexts
- But, qualitative methods are underpinned by a particular conception of the role of research
  - Emphasis on quality of lived experiences
- Many different kinds of interest in research
  - Habermas
    - Empirical analytic sciences and **explanation**
    - Historical hermeneutic sciences and **understanding**
    - Critical sciences and **practice**

Quantitative and Qualitative

- Often described as being in opposition
- Many decry the value of quantitative research
  - One of the ‘tensions’ between physical and human geography
- Recent emphasis in human geography has been very much on qualitative methods
  - Geographers are not very good at mathematics!
  - A tendency for reduced emphasis on rigour
- But this course wants to emphasise
  - The value of *both* quantitative and qualitative
  - Their appropriateness in *different* contexts

Qualitative data

- Examples of qualitative data?
  - Interviewing
  - Focus Groups/discussions
  - Participant observation
  - Visual images
  - Texts
- But how do we analyse these?
  - What indeed is qualitative analysis?

Scientific Method

Scientific Method Overview

The Scientific Method provides a structured approach for implementing Data Science based on data analytics

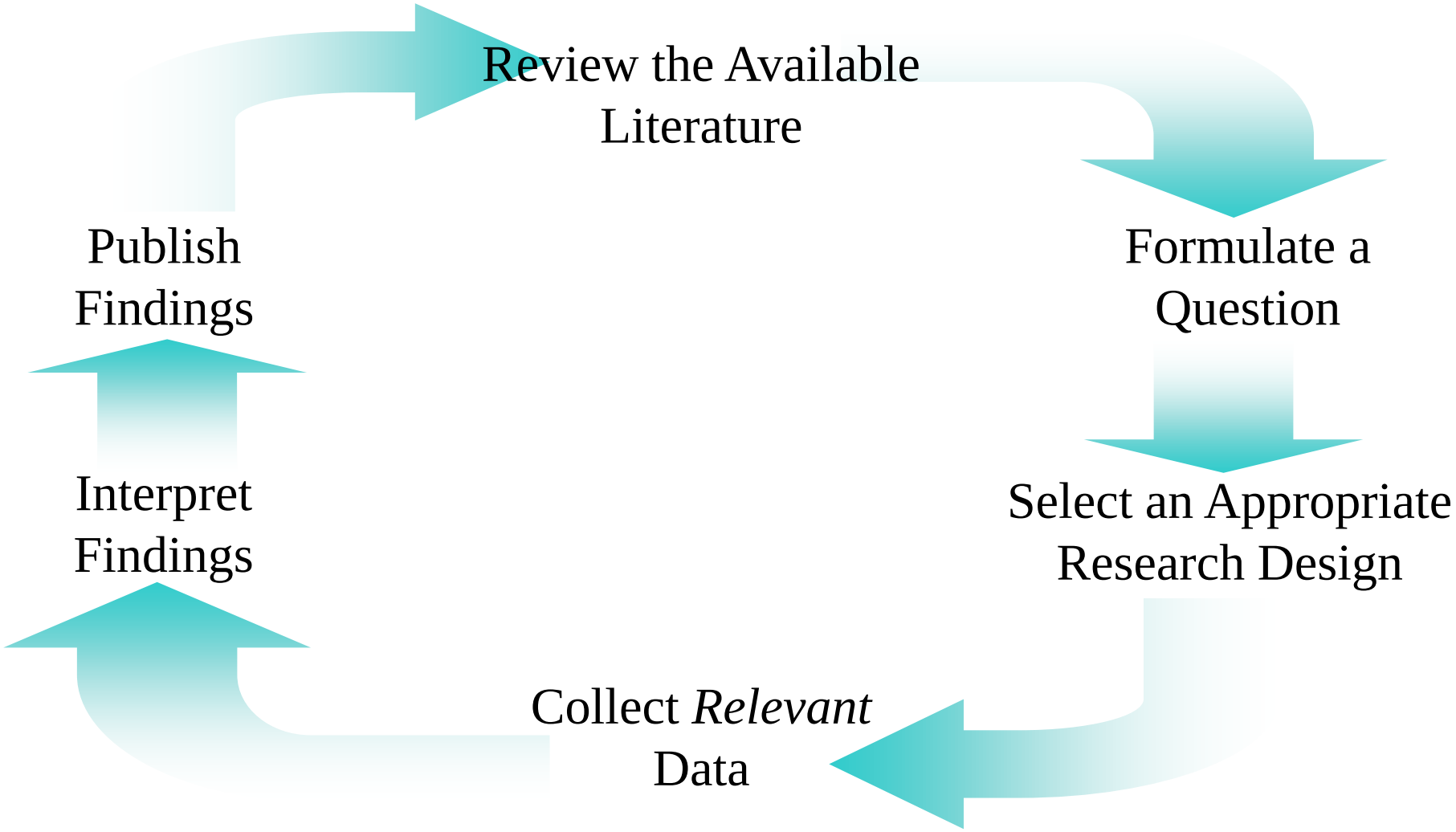
- observe a problem
- generate questions
- create a testable explanation (hypothesis)
- design an experiment to test the hypothesis
- make a prediction
- test the prediction using data
- refine and iterate
- draw a conclusion
- communicate your results

Both qualitative and quantitative research methods can be utilised in the Scientific Method

What is Research?

- A systematic means of problem solving (Tuckman 1978)
- 5 key characteristics:

1. Systematic – research process
2. Logical – induction/deduction
3. Empirical – evidence based
4. Reductive – generalisation
5. Replicable – methodology.



Research Process

Research Continuum

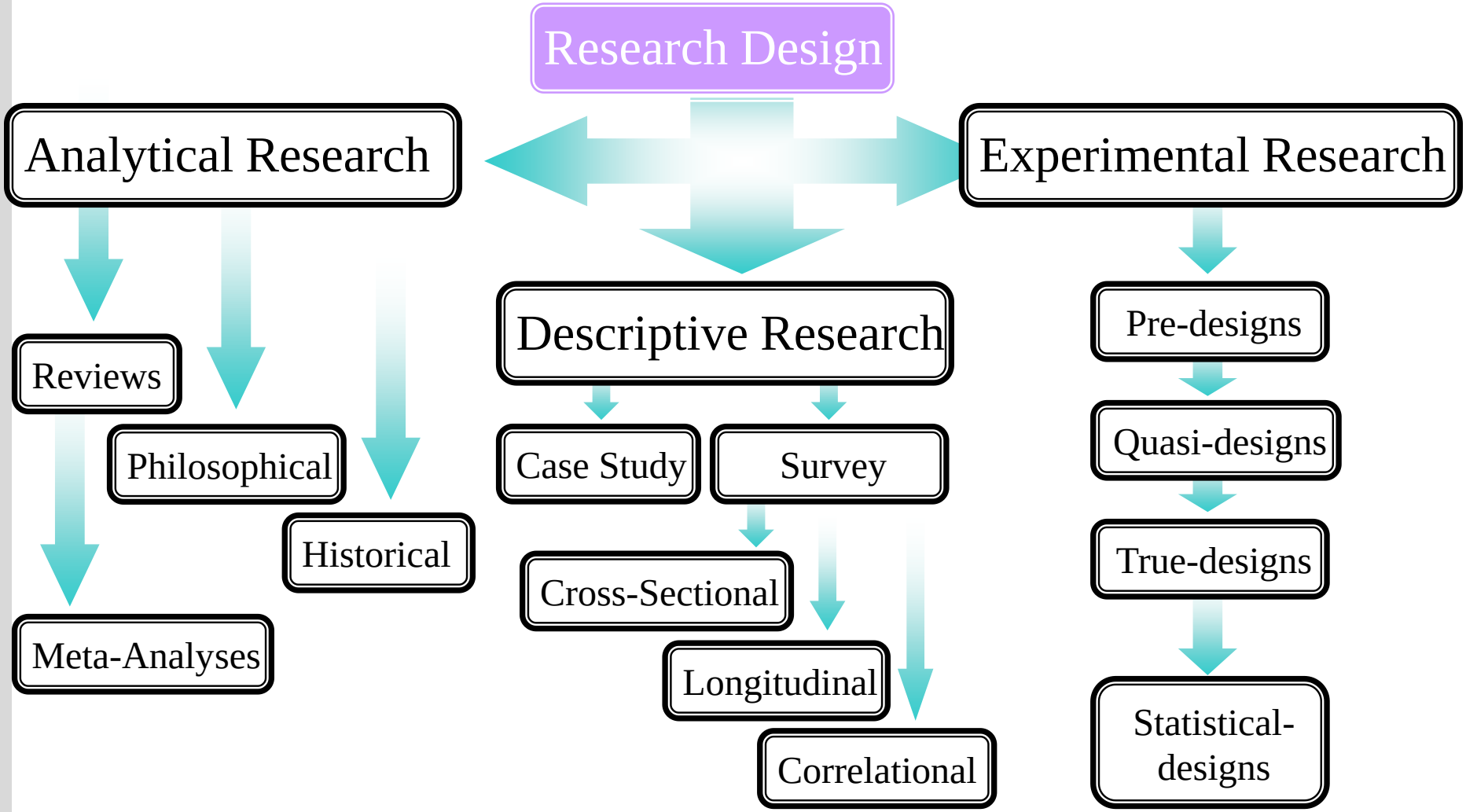
Reductionism

Research Continuum

Basic	Applied
➢Theoretical?	➢Quick Answers?
➢More Invasive?	➢Less Invasive?
➢Laboratory Based?	➢Field Based?
➢Tightly Controlled?	➢Loosely Controlled?
➢Lacks External Validity?	➢Externally Valid?
➢Focus on Mechanism	➢Focus on Effect
➢More Reductionist	➢Less Reductionist.

Internal Validity?

Research Design Continuum



Analytical Research

- Reviews
  - A critical account of present understanding
  - A meta-analysis is a quantitative method of review
- Historical Research
  - Accessing both primary (e.g. witnesses) or secondary (e.g. literature) sources to document past events
- Philosophical Research
  - Organising existing evidence into a comprehensive theoretical model

Descriptive Research

Refutable?

- Case Study
  - Accrual of detailed information from an *individual*

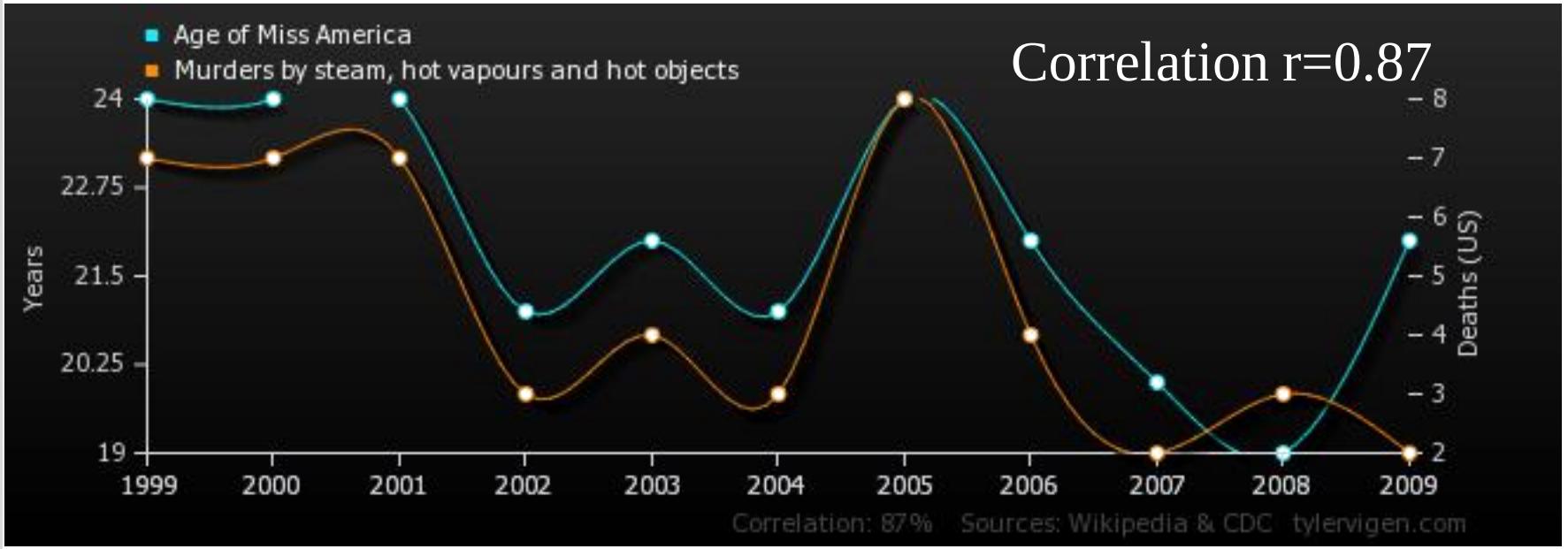
- Survey
  - Cross-sectional: Status of a various groups at a given point in time
  - Longitudinal: Status of a given group at various points in time
  - Correlational: Relationships between variables

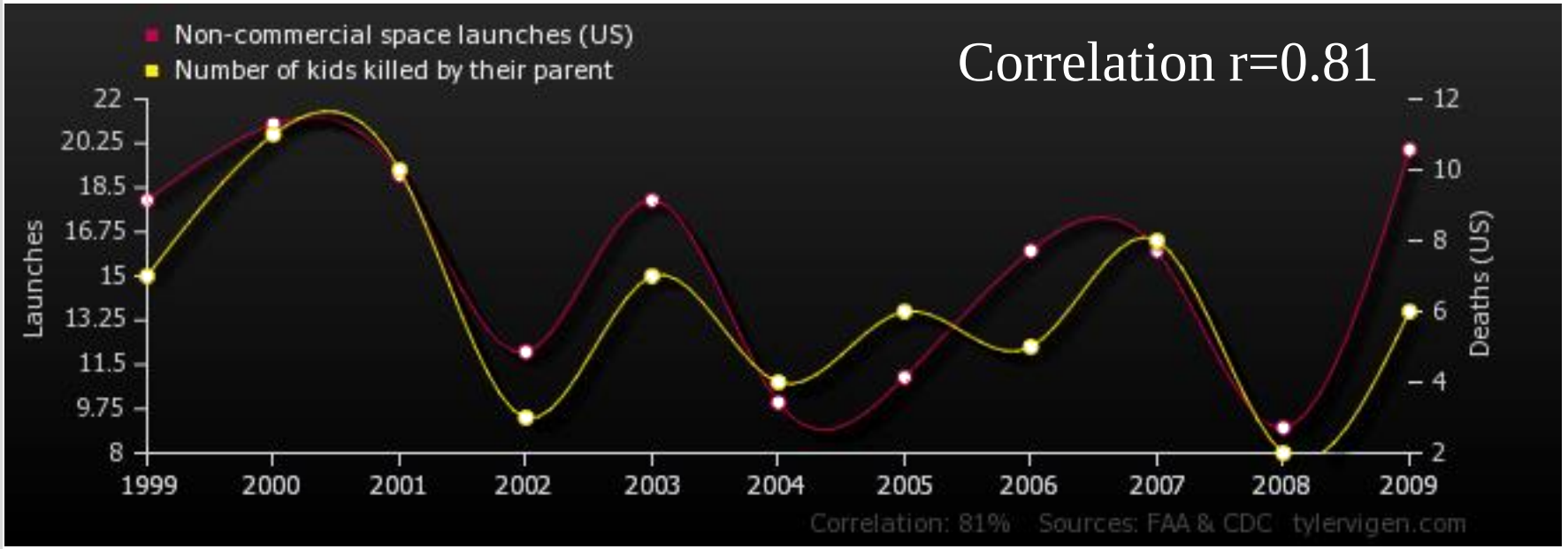
Correlational Evidence

- When variable X increases, variable Y also increases
- So, does X increase Y?
  - or does Y increase X?
- Alternatively, does Z increase both X and Y?

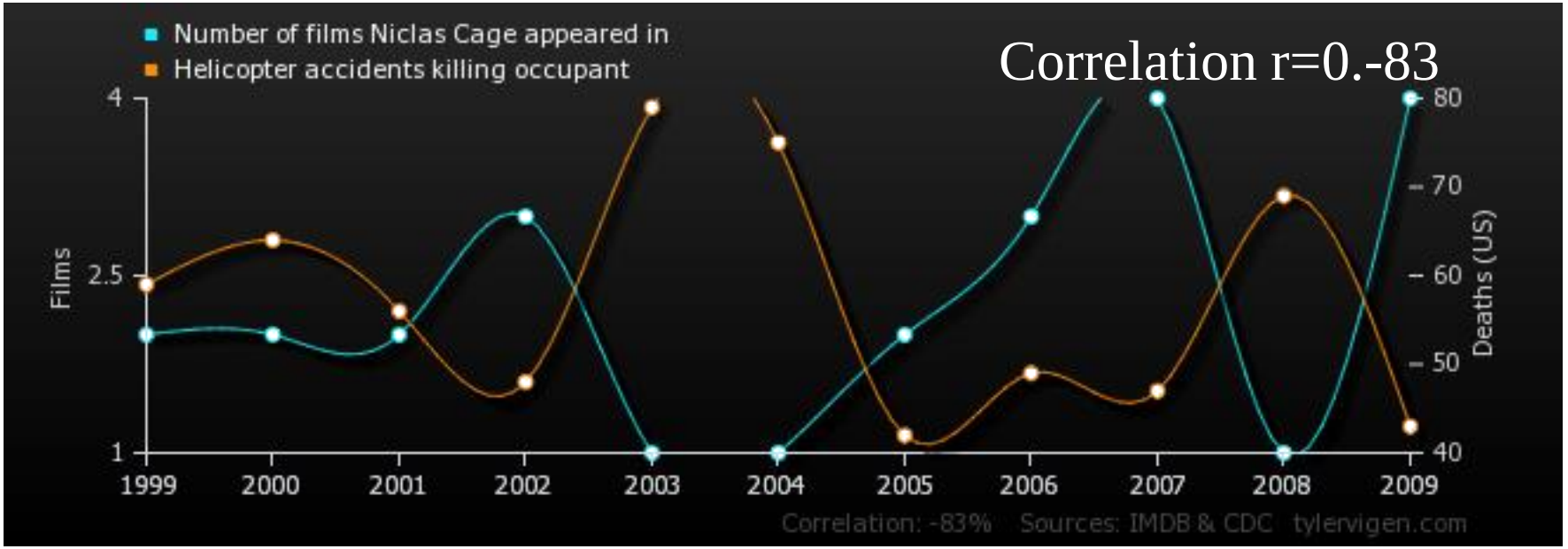
Correlations do not infer Causality

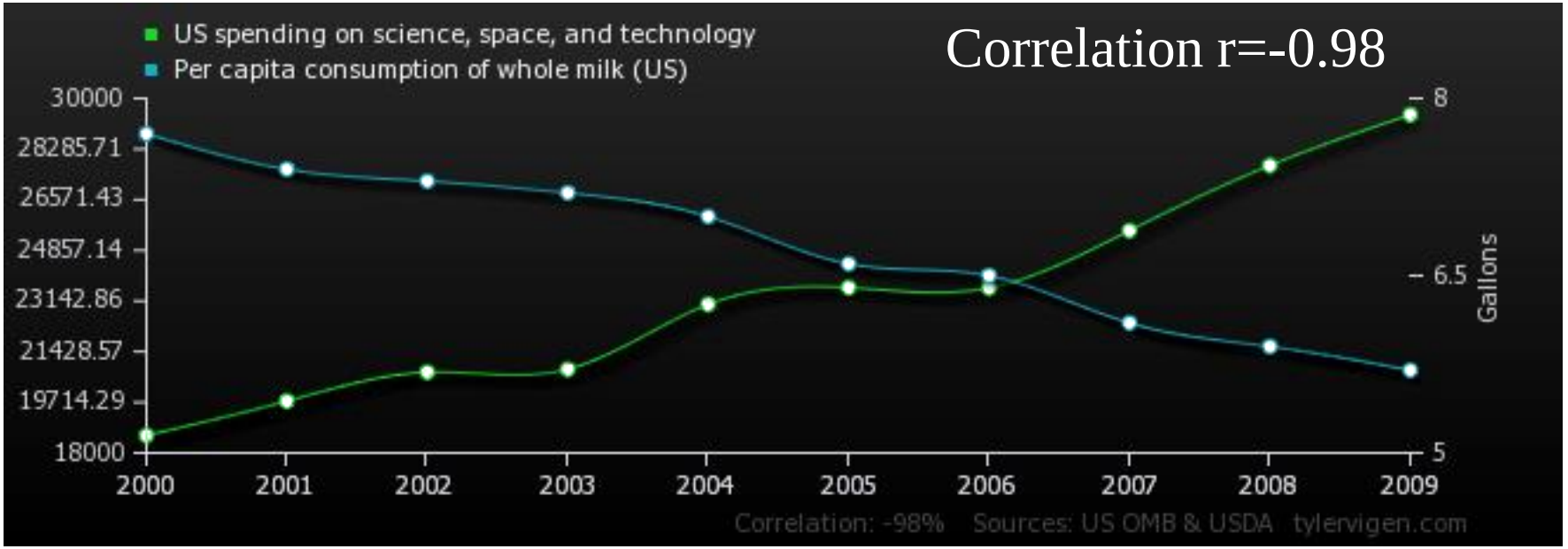
(and *vice versa*?)





http://t.co/vWOyN0N1IB





http://t.co/vWOyN0N1IB

Hypothesis

Hypothesis Overview

Hypothesis

- A hypothesis is a proposed explanation for a phenomenon.
- For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it.
- In business applications, a hypothesis describes an explanation for an observed behavior.
- Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories.
- Proving or disproving a hypotheses enables the generation of insights from data analytics.

- Source Wikipedia

Quantitative Research

- Quantity is the unit of analysis
  - Amounts
  - Frequencies
  - Degrees
  - Values
  - Intensity
- Uses statistics for greater precision and objectivity
- Based on the deductive model

Creating the Foundation for Quantitative Research

- Concept
  - Abstract thinking to distinguish it from other elements
- Construct
  - Theoretical definition of a concept; must be observable or measurable; linked to other concepts
- Variable
  - Presented in research questions and hypotheses
- Operationalization
  - Specifically how the variable is observed or measured

Research Hypotheses for Quantitative Research

- Educated guess or presumption based on literature
- States the nature of the relationship between two or more variables
- Predicts the research outcome
- Research study designed to test the relationship described in the hypothesis

Quantitative Research Hypotheses

- Directional hypothesis
  - Precise statement indicating the nature and direction of the relationship/difference between variables
- Nondirectional hypothesis
  - States only that relationship/difference will occur

Assessing Hypotheses

- Simply stated?
- Single sentence?
- At least two variables?
- Variables clearly stated?
- Is the relationship/difference precisely stated?
- Testable?

Null Hypotheses

- Implicit complementary statement to the research hypothesis
- States no relationship/difference exists between variables
- Statistical test performed on the null
- Assumed to be true until support for the research hypothesis is demonstrated

Research Traditions in the Use of Hypotheses

- Hypotheses are always tentative
- Research hypothesis, not the null hypothesis, is the focus of the research and presented in the research report

Research Questions in Quantitative Research

- Preferred when little is known about a communication phenomenon
- Used when previous studies report conflicting results
- Used to describe communication phenomena

Types of Variables

- Variable
  - Element that is identified in the hypothesis or research question
  - Property or characteristic of people or things that *varies* in quality or magnitude
  - Must have two or more levels
  - Must be identified as independent or dependent

Independent Variables

- Manipulation or variation of this variable is the cause of change in other variables
- Technically, independent variable is the term reserved for experimental studies
  - Also called antecedent variable, experimental variable, treatment variable, causal variable, predictor variable

Dependent Variables

- The variable of primary interest
- Research question/hypothesis describes, explains, or predicts changes in it
- The variable that is influenced or changed by the independent variable
  - In non-experimental research, also called criterion variable, outcome variable

Relationship Between Independent and Dependent Variables

- Cannot specify independent variables without specifying dependent variables
- Number of independent and dependent variables depends on the nature and complexity of the study
- The number and type of variables dictates which statistical test will be used

Intervening and Confounding Variables

- Intervening variable
  - Explains or provides a link between IV and DV
  - Relationship between the IV and DV can only be explained when the intervening variable is present
- Confounding variable
  - Confuses or obscures the effect of independent on dependent
  - Makes it difficult to isolate the effects of the independent variable

Operationalizing Variables

- All variables need an operationalization
- Multiple operationalizations exist for most variables
- Specifies the way in which variable is observed or measured
  - Practical and useful?
  - Justified argument?
  - Coincides with the conceptual definition?

Making the Case for Quantitative Research

- Advantages
  - Tradition and history implies rigor
  - Numbers and statistics allows precise and exact comparisons
  - Generalization of findings
- Limitations
  - Cannot capture complexity of communication over time
  - Difficult to apply outside of controlled environments

Issues of Reliability and Validity

- Reliability = *consistency* in procedures and in reactions of participants
- Validity = *truth* - Does it measure what it intended to measure?
- When reliability and validity are achieved, data are free from systematic errors

Threats to Reliability and Validity

- If measuring device cannot make fine distinctions
- If measuring device cannot capture people/things that differ
- When attempting to measure something irrelevant or unknown to respondent
- Can measuring device really capture the phenomenon?

Other Sources of Variation

- Variation must represent true differences
- Other sources of variation
  - Factors not measured
  - Personal factors
  - Differences in situational factors
  - Differences in research administration
  - Number of items measured
  - Unclear measuring device
  - Mechanical or procedural issues
  - Statistical processing of data

Experiments

Experiments Overview

- In the scientific method an experiment is an empirical procedure that arbitrates competing models or hypotheses.
- Researchers use experimentation to test existing theories or new hypotheses to support or disprove them.
- An experiment usually tests a hypothesis, which is an expectation about how a particular process or phenomenon works.
- An experiment may also aim to answer a "what-if" question, without a specific expectation about what the experiment reveals, or to confirm prior results.
- If an experiment is carefully conducted, the results usually either support or disprove the hypothesis.

- Source Wikipedia

Scientific Reasoning (Logic)

Quantitative?

Confirmation of a theory from your own observations

General Theory

Deductive Reasoning

Inductive Reasoning

Specific Observation

Formation of a theory grounded in your own observations

Qualitative?

Choice of Research Strategy…

- Based on:
  - Epistemology (How should we be attempting to assess knowledge?)
    - Positivism = explain a phenomena
    - Interpretivism = understand a phenomena
  - Ontology (Does the data exist in a tangible or an intangible form?)
    - Objectivism = explain independent external outcomes
    - Constructionism = understand how social factors interact

Choice of Research Strategy…

- Study in the natural sciences often requires a *positivistic epistemology* and an *objectivistic ontology*
- Study in the social sciences often requires an *interpretive epistemology* and a *constructionist ontology*
- *However*, it is occasionally possible to combine these strategies by coding qualitative data quantitatively (i.e. Athlete = 1 ; Non-Athlete = 2)

Sampling

Sampling Concepts

In statistics, quality assurance, and survey methodology, sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population.

Two advantages of sampling are that the cost is lower and data collection is faster than measuring the entire population.

Each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals.

Results from probability theory and statistical theory are employed to guide the practice.

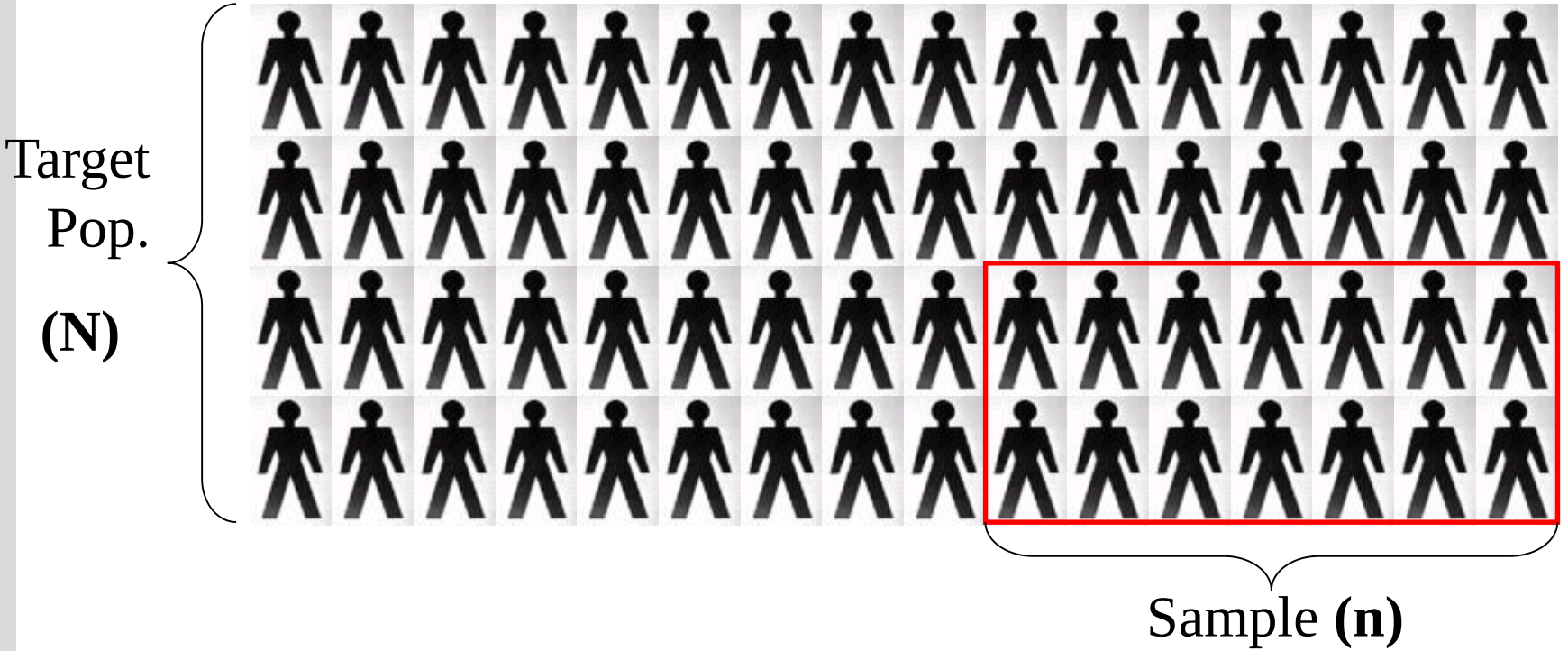
In business research, sampling is widely used for gathering information about a population.

The sampling process comprises several stages:

- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting

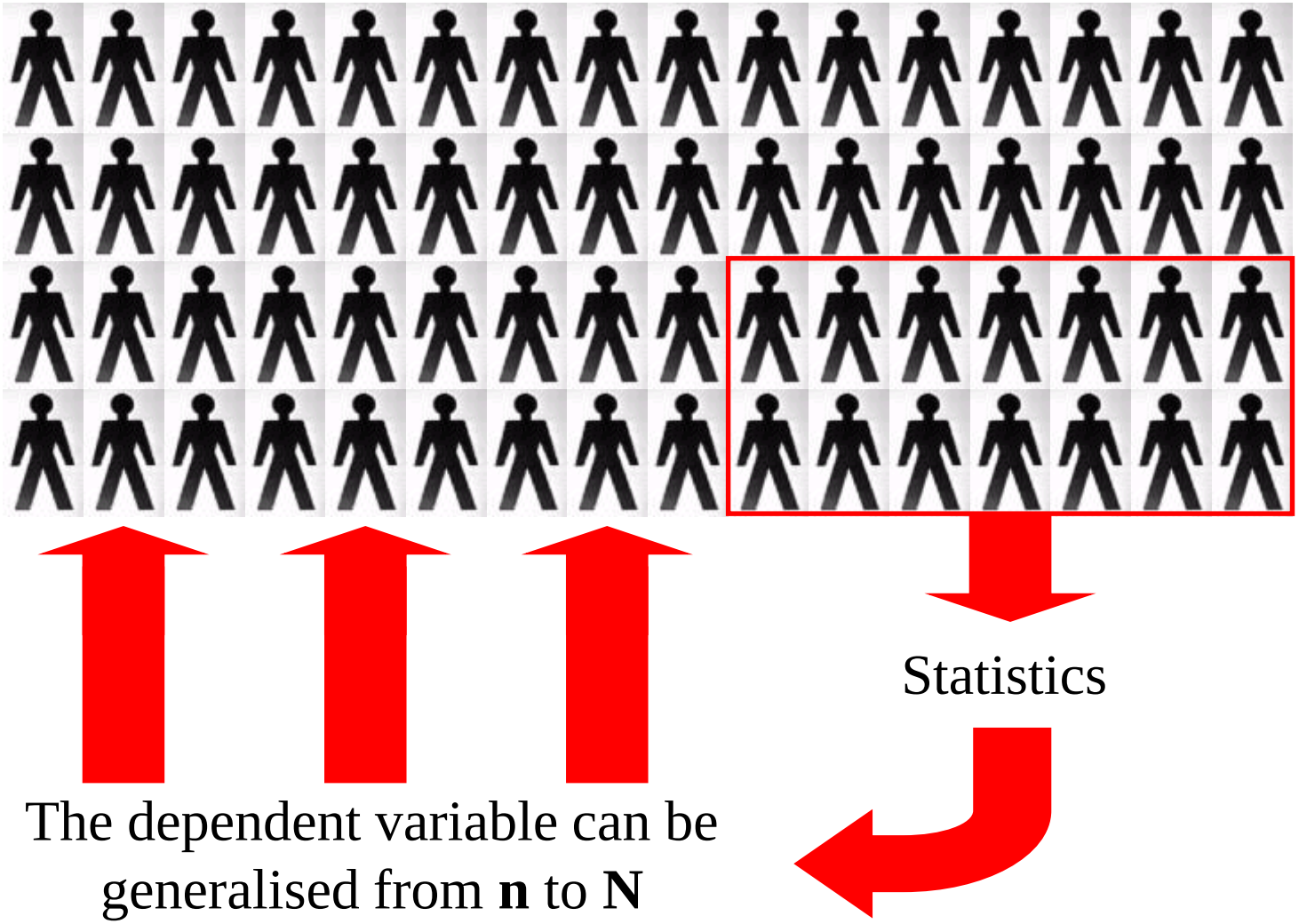
- Source Wikipedia

Sampling



- Effective Sampling produces a **n** which is representative of **N**
- Note: **n** is only ever representative of the **N** it was drawn from, i.e. not necessarily the general population.

Sampling


Statistics
The dependent variable can be generalised from **n** to **N**

Sampling Methods

- Random- All members of **N** have an equal chance of selection

- Stage- Randomly select a group, then take sample

- Cluster- Select a natural group to sample from

  e.g. local community

Sampling Methods

- Stratified- identify strata and sample accordingly
- Systematic- e.g. every fourth person *but* starting at a random point
- Opportunity- sample a convenient group

Lesson Review

Lesson Review

Consider the following questions that you should be able to answer by completing Day 5.

- What is the main purpose of research?
- What types of research activities exist?
- How does the scientific method compare to CRISP-DM?
- What are some techniques for generating effective business questions?
- What is a practical definition of a hypothesis?
- What role does Operational Excellence play in generating effective business questions?

Lesson Summary

Day 5 Lesson Summary

During Day 5 you learned to:

- Identify and describe the approach and objectives of research
- Describe two different types of research methods
- Describe how research methods enable successful business impact of data analytics
- Discover the business questions you are trying to answer with data analytics.
- Determine if your data is appropriate to support your analytics initiative
- Describe some data collection techniques and some related sampling concerns
- Recognize how Operational Excellence initiatives drive analytics requirements
- Check your data for some common issues in Python and interpret results from a regression model
- Have a general understanding of what kinds of models for analysis are readily available in Sci-kit Learn Python