

BC3406 Business Analytics Consulting

Damien Joseph

Nanyang Business School

Gino K. George

Analytics Executive

A Lee Gilbert

Course Instructor

Agenda

- Problem Analysis
 - Understanding the problem
 - Searching the literature
 - Generating Research Questions
- Developing Testable Hypotheses
 - When to hypothesize
 - Types of hypotheses
- Hypotheses to Models
- Metrics, Measures, Variables
 - Data types
 - Data sources
- Reporting Business Analytics

The Business Analytics Process

- Steps in the business analytics process
 1. Recognizing the problem
 2. Defining the problem
 3. Structuring the problem
 4. Analyzing the problem
 5. Generating questions
 6. Developing hypotheses
 7. Collecting and analyzing data
 8. Interpreting results and making recommendations
 9. Implementing the solution

Approaches to Business Analytics

- Barton & Court 2012 HBR:
 - “the desired business impact must drive an integrated approach to data sourcing, model building and organizational transformation avoid the common trap of starting with the data and simply asking what it can do for you.”
- Grounded theory/research
 - A systematic methodology in the social sciences involving the construction of theory/explanation of a phenomenon through the analysis of qualitative data (Strauss & Glazer 1967 Discovery of Grounded Theory).
 - Application – see Madsbjerg & Rasmussen 2014 HBR
 - Sensemaking:
 - Identifying key themes
 - Search for patterns in these themes and their drivers (similar to root case analysis)
 - Develop a theory/explanation of phenomena from root causes
 - Generate insights
 - Insights to business impact
- Multi-method approach (Van den Driest et al 2016 HBR)
 - Data mining + hypothesis testing
 - Grounded theory + hypothesis testing
- My personal preference: Targeted approach
 - Start with understanding the business case and the problem to be solved
 - Develop a set of “research” questions that drive analyses
 - Develop a set of initial answers (educated guesses) and hypotheses for each question
 - Collect the data and test your hypotheses

The background consists of several overlapping triangles. A large red triangle is on the left side. A white triangle is positioned between the red triangle and the blue triangles. There are two shades of blue: a medium blue triangle in the center and a darker blue triangle on the right. The text 'Problem Analysis' is written in white on the medium blue triangle.

Problem Analysis

Problem Analysis

- **Problem analysis** is the in-depth understanding and documentation of the issues, its causes and effects.
- Business analytics represent only a portion of the overall problem solving and decision making process.
- The problem solving process
 1. Recognizing the problem
 2. Defining the problem
 3. Structuring the problem
 4. Analyzing the problem
 5. Generating questions

Recognizing and Defining the Problem

- Recognizing the problem
 - Problems exist when there is a gap between what is happening and what we think should be happening.
 - For example, costs are too high compared with competitors.
- Defining the Problem
 - Clearly defining the problem is not a trivial task.
 - Complexity increases when the following occur:
 - large number of courses of action
 - several competing objectives
 - external groups are affected
 - problem owner and problem solver are not the same person
 - time constraints exist

Structuring and Analyzing the Problem

- Structuring the Problem
 - Stating goals and objectives
 - Characterizing the possible decisions
 - Identifying any constraints or restrictions
- Analyzing the Problem
 - Cause-effect analysis
 - Root cause analysis
 - Prior research: academic, consulting, internal



The Research Question

By doubting we are led to question,
by questioning we arrive at the truth
• - Peter Abelard

A Research Question is...

- Something you want to know about your issue, or about a specific area within that issue.
 - Not a topic, fragment, phrase, or sentence.
 - A research question ends with a question mark!
- Clear and precisely stated.
 - It is not too broad, nor is it too narrow.
- Open-ended, as opposed to closed.
 - It cannot be answered in a sentence or phrase.

Why Create a Research Question?

- Why “research”?
 - Curiosity is “the desire to learn or know about anything; inquisitiveness” (Dictionary.com)
 - Research is the “diligent and systematic inquiry or investigation into a subject in order to discover or revise facts, theories, applications, etc.” (Dictionary.com)
- Why “question”?
 - Business analysts should formulate clear questions to guide the search for answers.
 - The goal/purpose/objective is the reason for the research question.
- Considered together:
 - curiosity is the source of our questions – we ask because we want to know;
 - research is the means by which we find an answer.

Strategies for Developing Research Questions

1. Jot down everything you know about the problem as quickly as you can
 - A list or paragraph form is fine.
2. Now find the answers to the following questions pertaining to the problem:
 - Who? What? When ? Where? Why? How? So what?” and “What if...?”
 - These represent possible “gaps” in your knowledge;
 - The last four are particularly tough because they are open-ended – they often lead to good research questions .
3. Use organizing framework as a guide to fill in knowledge gaps
 - Comparison & Contrast
 - Process
 - Classification or Division
 - Cause & Effect
 - Problem & Solution
4. Draft the research question
 - Focusing Question: The most important question you discovered from the three prior activities.
 - How can implementing social media tools increase the economic impact of our organization?
 - Supporting Questions: Questions that will help you explore the relationships around the focusing question in greater depth.
 - How does the university currently impact the community economically? What could be done that isn't being done, and why? What are the limitations?

Questions generating hindsight

Business Analytics: From Descriptive To Predictive...

Use Business Intelligence to gain descriptive insights about customers, products and operations, such as...

Demographic Answers

- How long's Mia been a customer?
- What's Mia's annual salary? What's Mia's weekly spend?
- Where does Mia live? What's the value of her home?
- How many family members does Mia have?

Performance Answers

- What products did Mia buy last week?
- How much did Mia spend last month?
- To what promotions did Mia respond?
- Is Mia spending more/less versus last month?
- What is the trend of Mia's spend over the past year?



Mia

Hindsight = Demographics,
Preferences, Behaviors

EMC²

Questions generating insight/foresight

Business Analytics: From Descriptive To Predictive...

...then use Predictive Analytics to build predictive models and actionable recommendations at the individual consumer and store levels



Predictive Answers

- How many times is Mia likely to shop our store over Christmas?
- What promotions is Mia likely to use?
- What's the likelihood that Mia will buy Product Y?
- What's probability Mia will buy product Z when she buys Product Y?
- What's the profit potential of Mia over the next 2 years?

Recommendations

- What's best price to get Mia to buy private label cookies?
- What are the best promotions, and on what products, to get Mia to visit our stores 2 additional times a month?
- What new product introductions should we recommend to Mia?
- What private label products have the best chance of Mia buying?

Hindsight = Demographics,
Preferences, Behaviors



Foresight = Personalized,
Predictable, Actionable, Measurable

EMC

The Importance of a GOOD Research Question

- The research question is the starting point of most business analytics projects.
 - The challenge is to develop good questions to be answered
- The research question
 - Defines the investigation – the scope
 - Sets boundaries
 - population to be studied;
 - data to be collected;
 - Analytic tools to be used
 - Provides direction
- A clear and concisely stated research question is the most important requirement for a successful study.
 - Narrowing, clarifying, and even redefining questions is essential to the analytics process.
 - Forming the right ‘questions’ should be seen as an iterative process that is informed by reading and doing at all stages

Cycles of Research Question Development

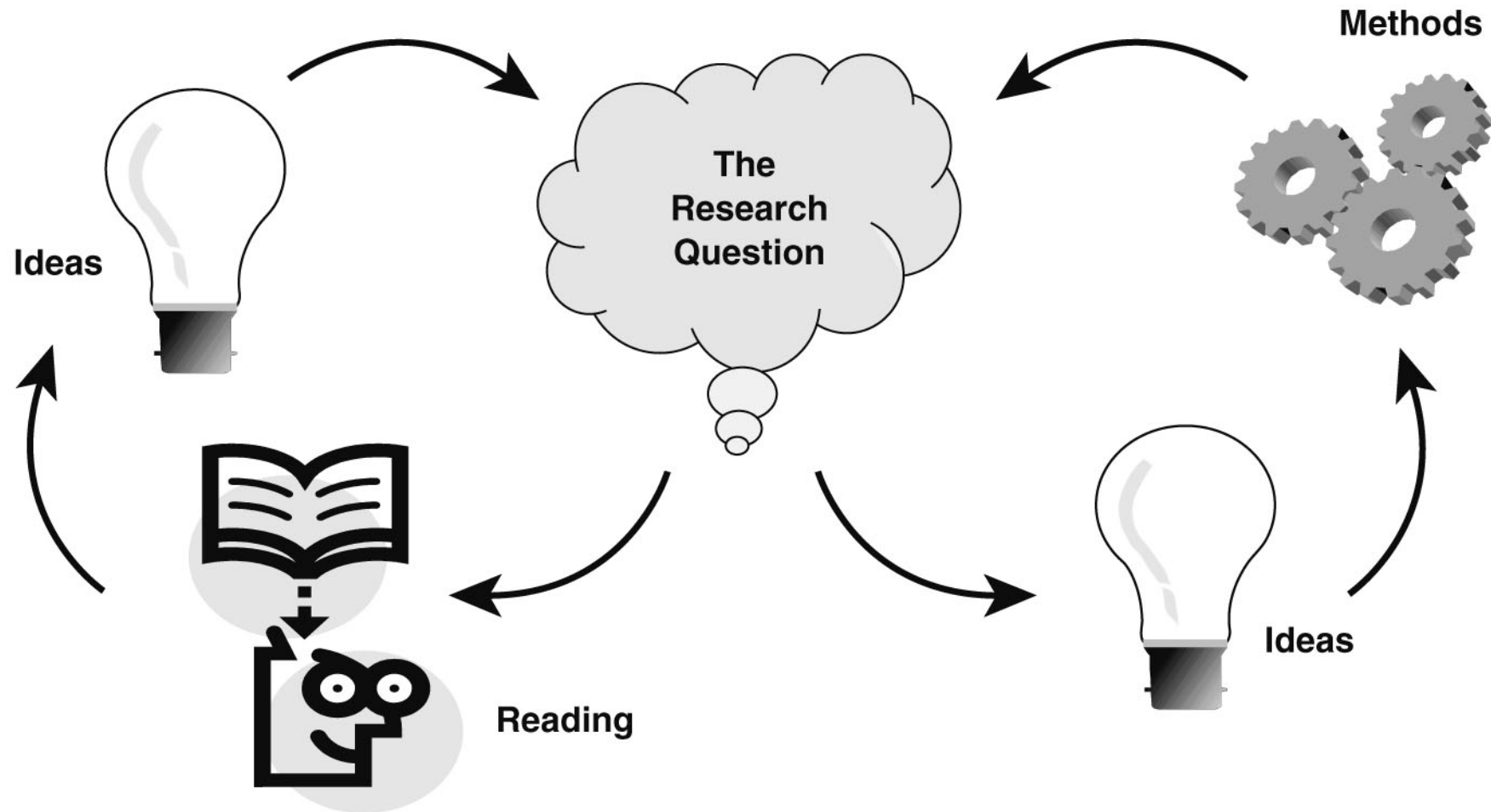


FIGURE 3.4 CYCLES OF RESEARCH QUESTION DEVELOPMENT

The background features a large red triangle on the left side, separated from the rest of the image by a white diagonal line. The remaining area is filled with two shades of blue: a medium blue and a darker navy blue, which are separated by another diagonal line. The text "Searching the Literature" is centered in the medium blue area.

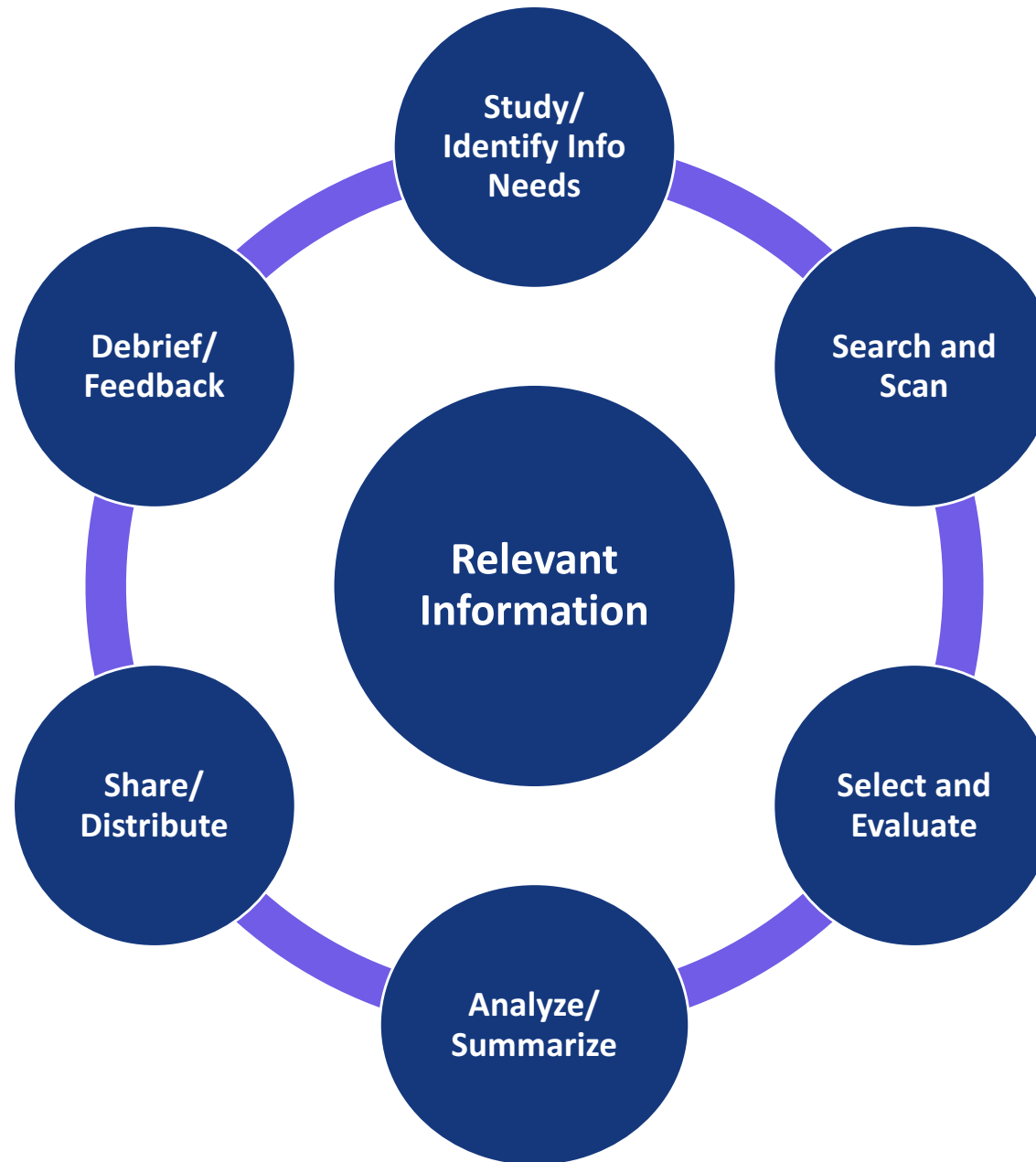
Searching the Literature

Briefly in the next 5 minutes....



- You are developing an analytics report for your company on investment opportunities in the renewable energy market in the Czech Republic. For this report you now need the following information about the Czech Republic:
 - Estimated compound annual growth rate (CAGR) of the total renewable electricity generation (TWh) for the last 10 years
 - Estimated CAGR for the last 5 years
- Spend 5 minutes to search for the above two pieces of information
- How did you conduct your search for the required information and what did you find?
 1. What terms/words you used in your search?
 2. What you used for your search?
 3. The number of “hits” obtained?
 4. An impression on the quality/relevance of information obtained.

The Intelligence Cycle



Decisions = f(Information Quality)

INFO NEED

What do I
search for?

SCAN

Where do I
search?

WHAT DO I DO?

I found too
many hits;

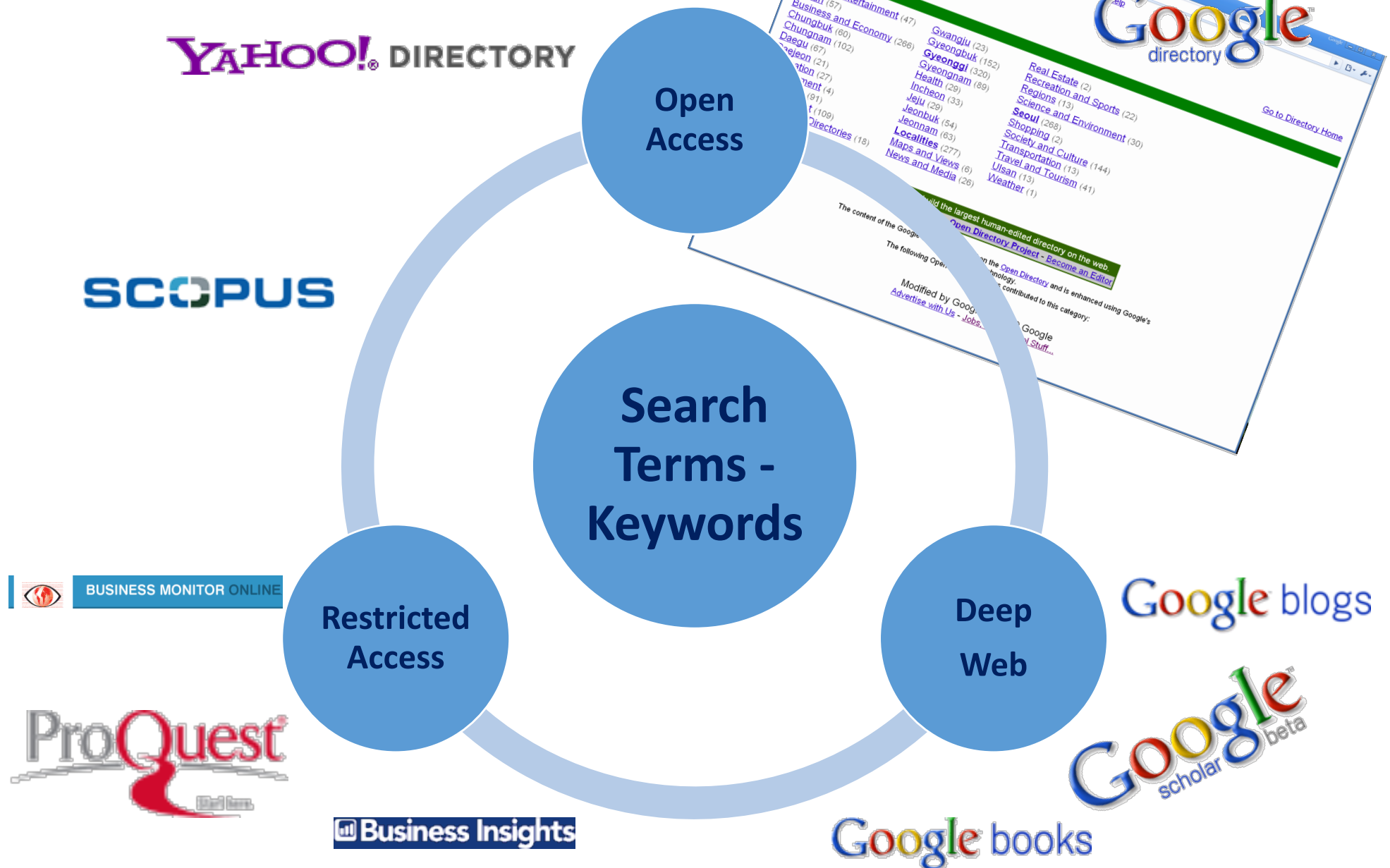
OR

I did not find
enough

EVALUATE

Authority;
Accuracy;
Objectivity;
Currency;
Coverage

Scanning



Search Tools



Library

Home **Library** [Share](#)

Search All

Articles
Books
Audio-Visuals
NTU Papers
e-Journal Titles
Database Titles
Exam Papers
Others

☐ Keyword
☐ Title
☐ Author

Identifying great research!

F1000
FACULTY of 1000

Teaching and Learning

- Course reserves
- Subject guides
- Learning @ NTU Libraries
- Happenings @ NTU Libraries
- Submit student work to DR-NTU
- Services for faculty

Library Services

- Ask a librarian
- Book a library space
- Renew items
- Request an item
- Request a blog
- Borrowing privileges
- Toolbar / Full text

Research and Scholarship

About Us

Google Advanced Search [Advanced Search Tips](#) | [About Google](#)

Use the form below and your advanced search will appear here

at have...

or phrase: [OR](#) [OR](#) [OR](#)

e words: [OR](#) [OR](#) [OR](#)

yes that have...

ted words:

10 results
any language
any format

or domain:
(e.g. youtube.com, .edu)

s numeric range and more
page is) anytime
not filtered by license
ds show up: anywhere in the page
any region
 --
(e.g. \$1500-\$3000)
☒ Off ☐ On

Advanced Search

to the page:

to the page:

Topic-specific search engines from Google:

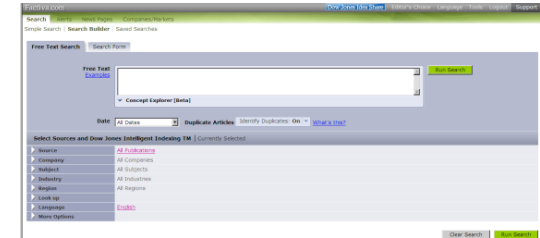
[Google Book Search](#)
[Google Code Search](#) **New!**
[Google Scholar](#)
[Google News archive search](#)

[Apple Macintosh](#)
[BSD Unix](#)
[Linux](#)
[Microsoft](#)

[U.S. Government Universities](#)

Search Tools - Alternatives

- NTU Library:
 - Restricted access
 - News databases
 - Country and Industry reports
- Subject Directories
 - Open access
 - Oldest and largest: dir.yahoo.com
 - Google: directory.google.com
 - Virtual Library
- Governmental
 - Singstat (<http://www.singstat.gov.sg/>)
 - Bureau of Labor Statistics (<http://www.bls.gov/>)
- Businesses
 - Newspapers – through NTU Library
 - Consultancy reports
- Individuals
 - Blogs
 - Evangelists
- Communities
 - Wikipedia


**BUSINESS MONITOR ONLINE****SCOPUS****YAHOO! DIRECTORY**

Each tool is different

The collage shows three search engine interfaces:

- Bing:** A search engine interface with a landscape background. Navigation links include 'Web', 'Images', 'News', and 'More'. A search bar is present.
- Google Advanced Search:** A search interface with a blue header. A red circle highlights the link 'Advanced Search Tips | About Google' in the top right corner.
- Yahoo! Advanced Web Search:** A search interface with a blue header. A red circle highlights the link 'Home' in the navigation bar.
- ProQuest:** A search interface with a green header. A red circle highlights the link 'Search tips' in the 'Tools' section.

Query Size by Country

Words:	 us	 uk	 au	 ca	 de	 es	 fr	 nl	 no	 se	 pl	 pt	 it
1	33.57%	95.54%	60.50%	58.34%	72.09%	70.27%	75.19%	65.89%	63.27%	72.55%	70.47%	85.53%	70.32%
2	26.49%	1.39%	13.52%	26.62%	15.89%	18.69%	14.12%	22.04%	22.45%	14.71%	21.24%	6.58%	17.74%
3	18.09%	1.95%	15.30%	9.93%	7.11%	6.43%	6.11%	8.58%	6.12%	7.84%	6.22%	5.26%	7.74%
4	9.96%	0.56%	7.47%	2.34%	2.07%	2.55%	2.67%	1.86%	4.08%	3.92%	0.52%	1.32%	1.94%
5	5.77%	0.28%	0.71%	1.10%	1.55%	1.46%	0.76%	1.16%	2.04%	0.00%	1.04%	0.00%	0.65%
6	2.68%	0.28%	2.14%	0.55%	0.65%	0.24%	0.00%	0.23%	0.00%	0.98%	0.00%	0.00%	0.65%
7	1.45%	0.00%	0.00%	0.41%	0.26%	0.00%	0.76%	0.23%	2.04%	0.00%	0.00%	0.00%	0.32%
8	0.79%	0.00%	0.36%	0.14%	0.26%	0.12%	0.00%	0.00%	0.00%	0.00%	0.52%	0.00%	0.65%
9	0.52%	0.00%	0.00%	0.00%	0.00%	0.24%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
10+	0.69%	0.00%	0.00%	0.55%	0.13%	0.00%	0.38%	0.00%	0.00%	0.00%	0.00%	1.32%	0.00%

Source: www.keyworddiscovery.com/keyword-stats.html?date=2016-11-01

Developing the Search Syntax

- Plan your search

TOPIC WORKSHEET

Jot down a topic or subject you'd like to explore on the Web:

BEGIN THE PRE-SEARCHING ANALYSIS

1. **What UNIQUE WORDS, DISTINCTIVE NAMES, ABBREVIATIONS, or ACRONYMS are associated with your topic?**





These may be the place to begin because their specificity will help zero in on relevant pages.

2. **Can you think of societies, organizations, or groups that might have information on your subject via their pages?**

Search these as a "phrase in quotes", looking for a home page that might contain links to other pages, journals, discussion groups, or databases on your subject. You may require the "phrase in quotes" to be in the documents' titles by preceding it by **title:**_[no space]

Source: www.lib.berkeley.edu/TeachingLib/Guides/Internet/AnalyseTopicForm.pdf

Search Fields

Advanced Search

Find pages with...

all these words:

To do this in the search box

Type the important words: tricolor rat terrier

this exact word or phrase:

Put exact words in quotes: "rat terrier"

any of these words:

Type OR between all the words you want: miniature OR standard

none of these words:

Put a minus sign just before words you don't want:
-rodent, -"Jack Russell"

numbers ranging from:

 to

Put 2 periods between the numbers and add a unit of measure:
10..35 lb, \$300..\$500, 2010..2011

Then narrow your results by...

language:

any language

Find pages in the language you select.

region:

any region

Find pages published in a particular region.

last update:

anytime

Find pages updated within the time you specify.

site or domain:

Search one site (like wikipedia.org) or limit your results to a domain like .edu, .org or .gov

terms appearing:

anywhere in the page

Search for terms in the whole page, page title, or web address, or links to the page you're looking for.

SafeSearch:

Show most relevant results

Tell SafeSearch whether to filter sexually explicit content.

file type:

any format

Find pages in the format you prefer.

usage rights:

not filtered by license

Find pages you are free to use yourself.

Advanced Search

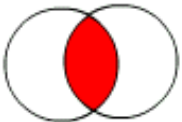
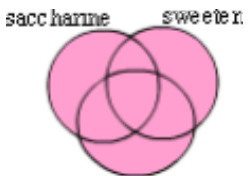
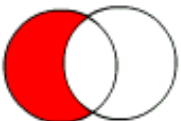
You can also...

[Find pages that are similar to, or link to, a URL](#)
[Search pages you've visited](#)
[Use operators in the search box](#)
[Customize your search settings](#)

Search Operators - Truncation

Symbol	Retrieves
*	Zero or more characters gene* <i>gene, genetics, generation</i>
\$	Zero or one character colo\$r <i>color, colour</i>
?	One character only Wom?n <i>woman, women</i>

Search Operators - Boolean

<p>AND</p> 	<p>All search terms must occur to be retrieved.</p> <p>TOPIC: “educational system” AND tertiary</p> <p>Retrieves documents that contain the phrase educational system and the term tertiary.</p>
<p>OR</p> 	<p>Any one of the search terms must occur to be retrieved. Use when searching variants and synonyms.</p> <p>TOPIC: religio* OR faith OR belief*</p> <p>Retrieves documents that contain at least one of the terms.</p>
<p>NOT</p> 	<p>Excludes records that contain a given search term.</p> <p>TOPIC: marriage NOT commerce</p> <p>Retrieves documents with <i>aids</i>, excluding any which also contain <i>hearing</i>.</p>

Scan -> Select: 5 Tactics

1. Plan: Analyze your topic to decide where to begin
 - a. Search terms
 - b. Booleans, proximity, truncations
2. Scan: Pick the right starting place
 - a. Typically general to specific : Open (e.g. Google) followed by Restricted (e.g. Business Insights)
(www.lib.berkeley.edu/TeachingLib/Guides/Internet/Strategies.html)
 - b. Know the search tool's coverage and capabilities
(www.libraries.psu.edu/etc/medialib/psulpublicmedialibrary/business/documents.Par.84810.File.dat/Database_Comparison_Chart.pdf)
(www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html)
 - c. Use advanced information presentation tools: Associations between concepts and webpages and information timelines
3. **Select: Learn as you go** and vary your approach based on what you learn
 - a. Looks of various perspectives, e.g. industry, government, consumer; global vs local
4. **Don't get bogged down** in any strategy that does not work
5. Return to previous thing that worked

Selecting (Evaluating) a Result

Accuracy

→ **Honda FC Sport Concept Car to Make Canadian Debut**

Hydrogen-powered Sports Car Concept Points to Future Direction

Authority

Currency

TORONTO, Feb. 2 /CNW/ - The Honda FC Sport design study model, a hydrogen-powered, three-seat sports car concept, will make its Canadian debut on February 11 to members of the press at the press preview of the 2009 Canadian International Auto Show in Toronto which opens to the public on Friday, February 13.

Coverage

Objectivity

The FC Sport emphasizes the design flexibility and potential of Honda's V Flow fuel cell technology - already deployed in the Honda FCX Clarity sedan - and reconfigures it into a lightweight sports car design with an ultra-low center of gravity, powerful electric motor performance and zero-emissions. The design study concept is inspired by supercar levels of performance through low weight and a high-

Sources: lib.nmsu.edu/instruction/evalcrit.html;

<http://www.newswire.ca/en/releases/archive/February2009/02/c4808.html>

Evaluation Criteria

- **Authority**
 - Expertise and qualification of author; Sponsorship of information; citations and references
- **Accuracy**
 - Error-free; reliable; corroborated information;
- **Objectivity**
 - Biases; stand taking; personal opinions
- **Currency**
 - Date of publication; date of revision
- **Coverage**
 - Depth of material; value of information provided

Evaluating a Result: Authority

- Concerns the expertise and qualification of author; sponsorship of information; citations and references
- Typical questions to ask:
 - Is there an author? Is the page signed?
 - Is the author qualified? An expert?
 - Who is the sponsor?
 - Is the sponsor of the page reputable? How reputable?
 - Is there a link to information about the author or the sponsor?
 - If the page includes neither a author nor indicates a sponsor, is there any other way to determine its origin?
- What to look out for?
 - Look for a header or footer showing affiliation.
 - Look at the domain. .edu, .com, .ac.uk, .org, .net
- Rationale
 - Anyone can publish anything on the web.
 - It is often hard to determine a web page's authorship.
 - Even if a page is signed, qualifications are not usually provided.

Source: lib.nmsu.edu/instruction/evalcrit.html

Evaluating a Result: Accuracy

- Error-free; reliable; corroborated information;
- Typical questions to ask:
 - Is the information reliable and error-free?
 - Is there an editor or someone who verifies/checks the information?
- Rationale
 - See issues about Authority above
 - Unlike traditional print resources, web resources rarely have editors or fact-checkers.
 - Currently, no web standards exist to ensure accuracy.

Evaluating a Result: Objectivity

- Biases; stand taking; personal opinions
- Typical questions to ask:
 - Does the information show a minimum of bias?
 - Is the page designed to sway opinion?
 - Is there any advertising on the page?
- Rationale:
 - Frequently the goals of the sponsors/authors are not clearly stated.
 - Often the Web serves as a virtual soapbox.

Evaluating a Result: Currency

- Dates of publication and/or revision
- Typical questions to ask:
 - Is the page dated?
 - If so, when was the last update?
 - How current are the links? Have some expired or moved?
- Rationale:
 - Publication or revision dates are not always provided.
 - If a date is provided, it may have various meanings. For example,
 - It may indicate when the material was first written
 - It may indicate when the material was first placed on the Web
 - It may indicate when the material was last revised

Evaluating a Result: Coverage

- Depth of material; value of information provided.
- Typical questions to ask:
 - What topics are covered?
 - What does this page offer that is not found elsewhere?
 - What is its intrinsic value?
 - How in-depth is the material?
- Rationale:
 - Web coverage often differs from print coverage.
 - Frequently, it's difficult to determine the extent of coverage of a topic from a web page.
 - The page may or may not include links to other web pages or print references.
 - Sometimes web information is "just for fun", a hoax, someone's personal expression.

Summarize the Information

- According to goals of the project
 - Organizing frameworks are particularly helpful in summarizing information (examples follow)
- Identify gap as in information
 - Scan -> Select -> Summarize
- Information gathering is an iterative process
 - Learn from every cycle
 - Go back to what works; don't get bogged down



Briefly in the next 5 minutes....



- You are developing an analytics report for your company on investment opportunities in the renewable energy market in the Czech Republic. For this report you now need the following information about the Czech Republic:
 - Estimated compound annual growth rate (CAGR) of the total renewable electricity generation (TWh) for the last 10 years
 - Estimated CAGR for the last 5 years
- Spend 5 minutes to search for the above two pieces of information
- How did you conduct your search for the required information and what did you find?
 1. What terms/words you used in your search?
 2. What you used for your search?
 3. The number of “hits” obtained?
 4. An impression on the quality/relevance of information obtained.



Hypothesis Development

Hypothesis: An expectation about the nature of things, derived from a theory/explanation.

The Hypothesis Dilemma

- Hypotheses are designed to express relationships between variables.
 - If this is the nature of your question, a hypothesis can add to your research
- If your question is more descriptive or explorative, generating a hypothesis may not be appropriate
- A hypothesis may not be appropriate if:
 - Do not have a hunch or educated guess about a particular situation
 - Do not have a set (at least two) of defined variables.
 - Question centres on phenomenological description

Research Question to Hypotheses

- RQ: Is a happy worker a productive worker?
 - H1: Happier workers are more productive than unhappy workers.

- RQ: Does increasing the happiness of workers make them more productive?
 - H1 : Increasing the happiness of workers increases productivity.

**Hypotheses should be developed before
data are collected.**

Testing Hypotheses

- Involves drawing inferences about two contrasting propositions (i.e. hypotheses) relating to the value of one or more population parameters.
- The **null hypothesis**:
 - the opposite of the alternative hypothesis and/or a statement of no difference or no relationship.
 - e.g. $H_0: \mu_1 - \mu_2 = 0$
 - e.g. $H_0: \beta = 0$
- The **alternative hypothesis**:
 - Usually a predictive statement regarding a relationship
 - Usually the opposite of the Null
 - e.g. $H_1: \mu_1 - \mu_2 \neq 0$
 - e.g. $H_1: \beta \neq 0$
- We test the null in order to see if we can reject it, in order to provide evidence for the alternative.

Directionality of Hypothesis

- Non-directional hypothesis
 - a non-specific predictive statement regarding a relationship
 - “The quantity of beer sold is associated with the ambient temperature”
- Directional hypothesis
 - an alternative hypothesis that predicts a more specific relationship.
 - e.g. $H_1: \mu_1 - \mu_2 > 0$
 - e.g. $H_1: \beta < 0$
 - “The quantity of beer sold is positively associated (or increases) with the (rise of) ambient temperature”
- When do you use a directional hypothesis and when do you use non-directional hypothesis?

Hypothesis Testing - Process

1. Identify the population parameter and formulate the hypotheses to test.
2. Select a level of significance (related to the risk of drawing an incorrect conclusion).
3. Determine the decision rule on which to base a conclusion.
4. Collect data and calculate a test statistic.
5. Apply the decision rule and draw a conclusion.

One-Sample Hypothesis Tests

Three forms:

1. H_0 : parameter = constant

H_1 : parameter \neq constant

2. H_0 : parameter \leq constant

H_1 : parameter $>$ constant

3. H_0 : parameter \geq constant

H_1 : parameter $<$ constant

The equality part of the hypotheses sign is always in the Null hypothesis.

Understanding Risk in Hypothesis Testing

- We always risk drawing an incorrect (error in) conclusion.
- Four outcomes are possible:

	Test Fails to Reject H_0	Test Rejects H_0
H_0 is True	True Negative (Correct Inference)	Type I Error (False Positive)
H_0 is False	Type II Error (False Negative)	True Positive (Correct Inference)

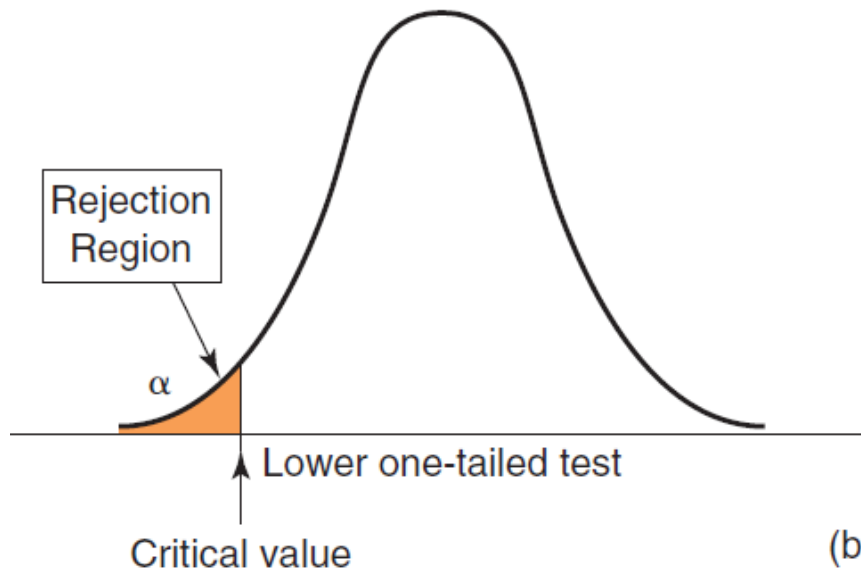
Understanding Risk in Hypothesis Testing

- The probability of making a Type I error = α
 - $\alpha = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$
 - The value of α can be controlled.
 - typically set to 0.01, 0.05, or 0.10.
 - In practice, most hypotheses tests are tested against $\alpha = 0.05$
 - Alternative approach uses the p -value, i.e. the observed significance level.
 - The p -value decision rule is to:
 - Reject H_0 if the p -value $< \alpha$
- The probability of making a Type II error = β
 - $\beta = P(\text{not rejecting } H_0 \mid H_0 \text{ is false})$
 - The value of β cannot be specified in advance and depends on the value of the (unknown) population parameter.

Drawing Conclusions from Hypothesis Tests: One-tailed Test

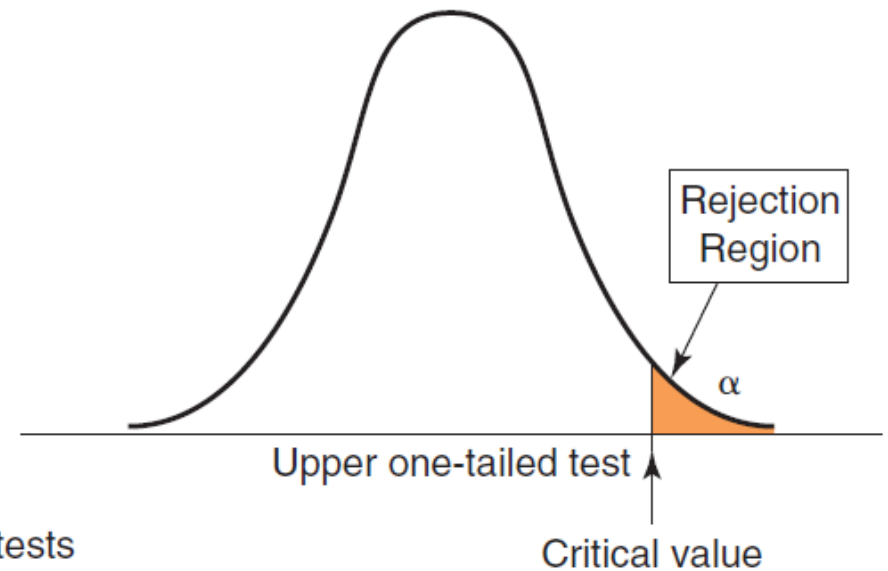
H_0 : parameter \geq constant
 H_1 : parameter $<$ constant

H_0 : parameter \leq constant
 H_1 : parameter $>$ constant



Note: critical value α is NOT
divided by 2

(b) One-tailed tests

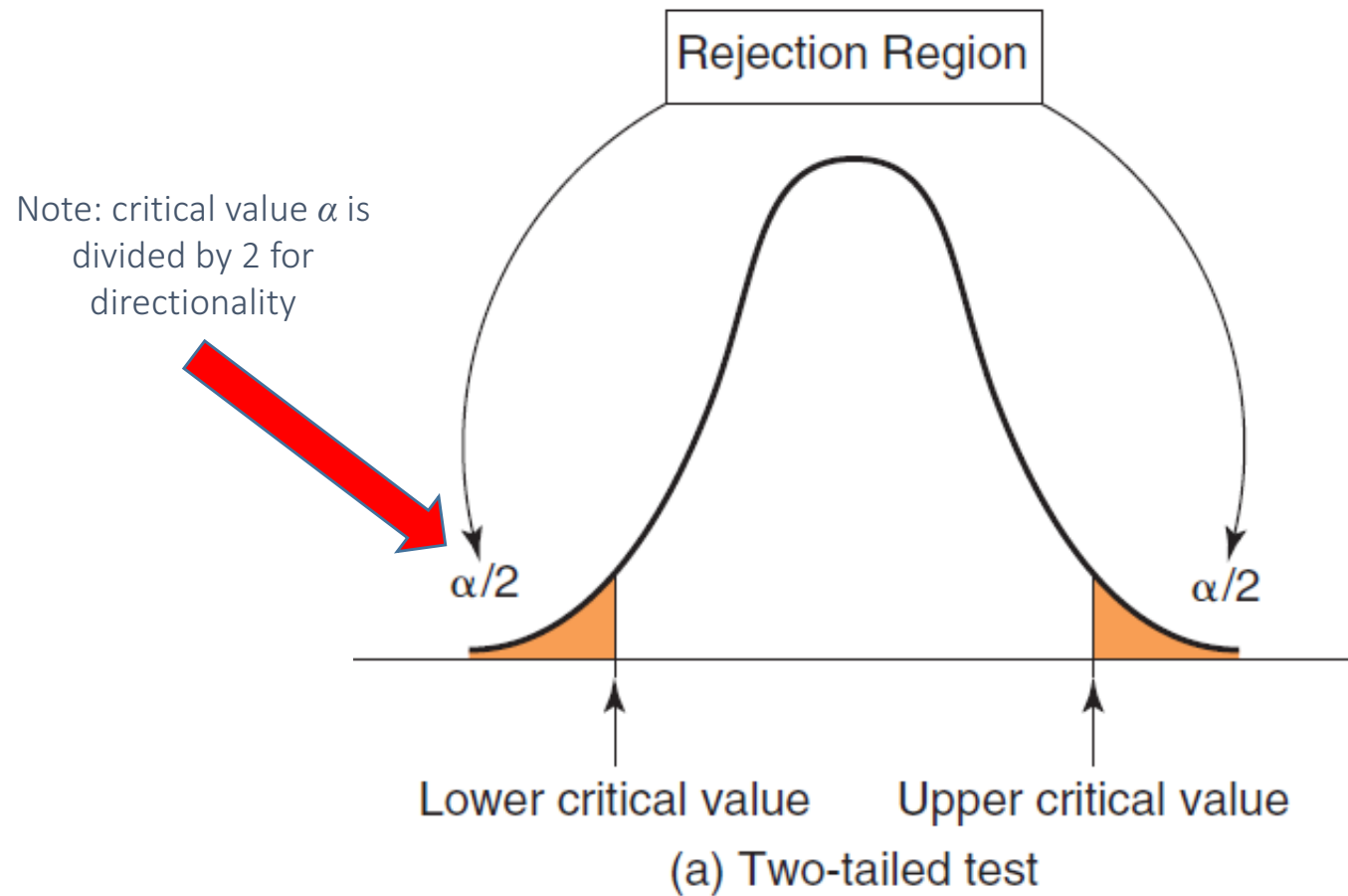


Note: critical value α is NOT
divided by 2

Drawing Conclusions from Hypothesis Tests: Two-tailed Test

H_0 : parameter = constant

H_1 : parameter \neq constant



Good hypotheses

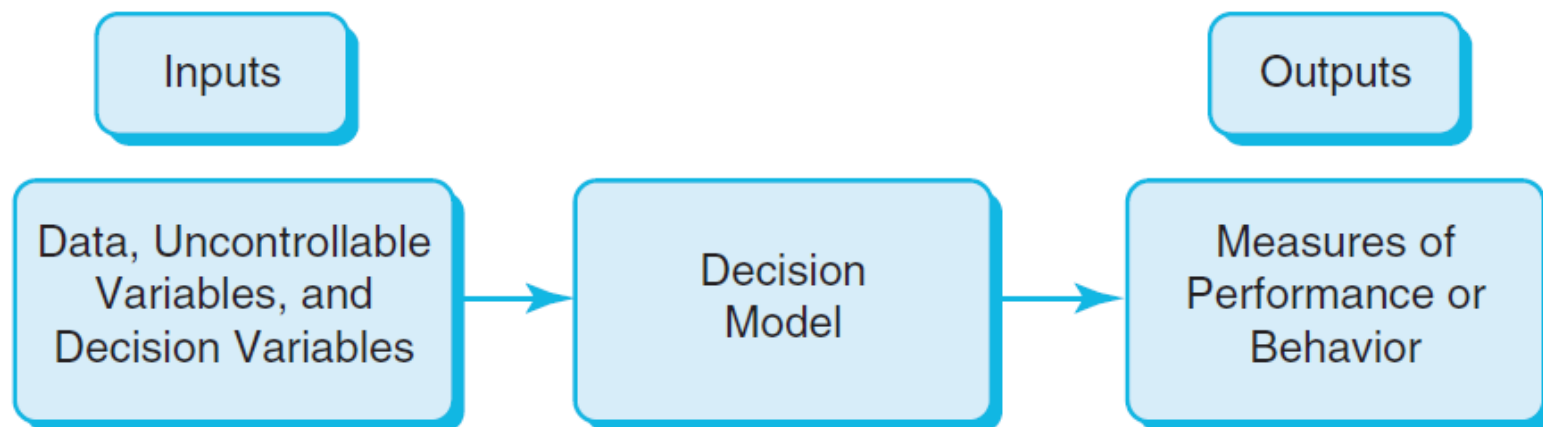
- Focal entities (e.g. beer, women) are clearly defined
- Direction of relationship is clear
- Population is identified
 - E.g. “The quantity of beer bought by women is positively associated with the ambient temperature”
- Design/statistical method often clear
 - Mean differences: e.g. “Per day, the amount of edam cheese sold is more than the amount of brie that is sold”
 - Compared to whom? e.g. “The salary received by men is more than the salary received by women”
 - Related (correlation): e.g. “The quantity of beer bought by women is positively associated with the ambient temperature”

The background features a large red triangle on the left side, pointing towards the bottom right. A white diagonal line separates this red triangle from the rest of the image. The remaining area is composed of two shades of blue: a medium blue triangle on the left, pointing towards the bottom left, and a dark blue area that fills the top and right portions of the image.

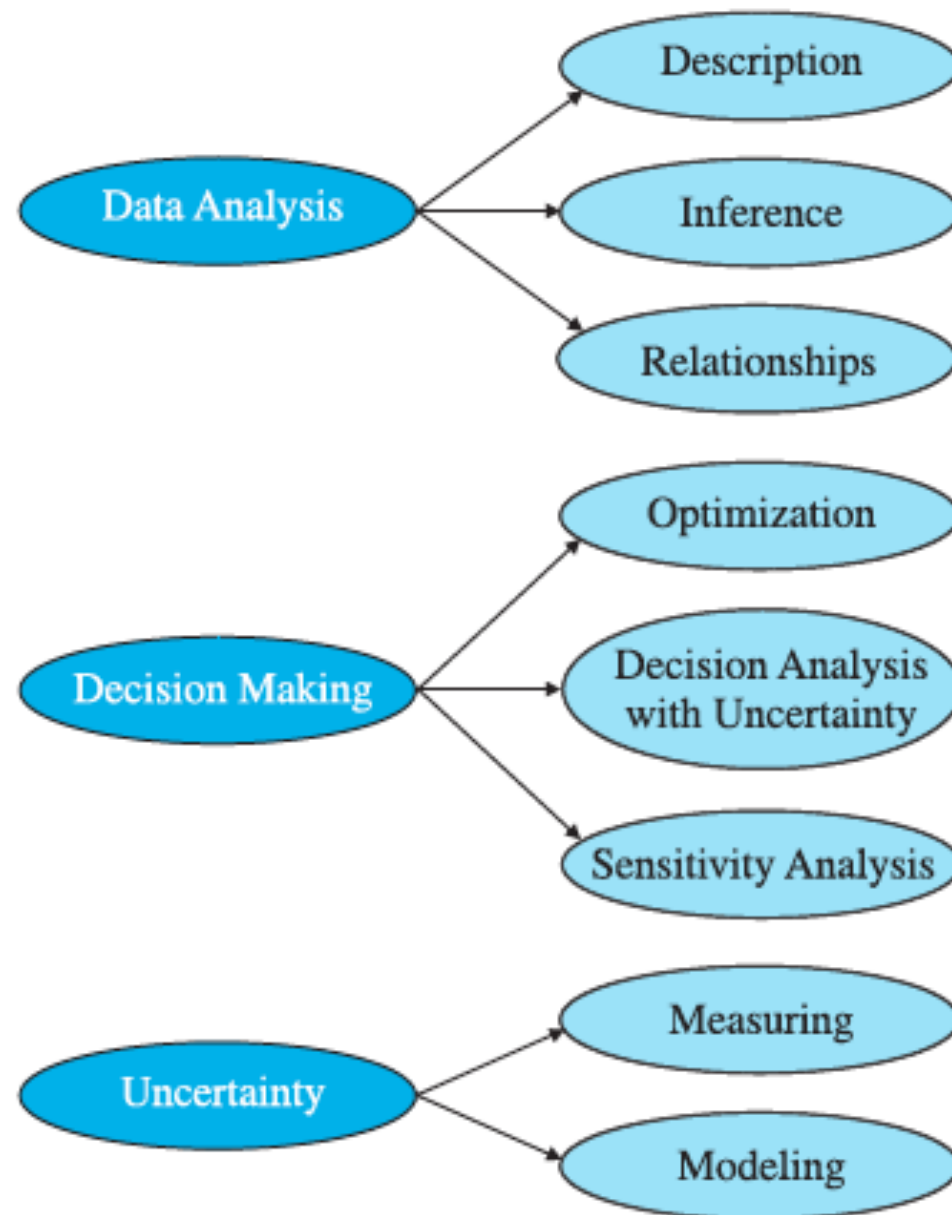
Hypothesis to Models

Models

- An abstraction or representation of a phenomenon, system, idea, or object
 - Used to understand, analyze, or facilitate decision making
 - Derived from one or more hypotheses
 - Captures the most important features
- Can be a written or verbal description, a visual display, a mathematical formula, or a spreadsheet representation

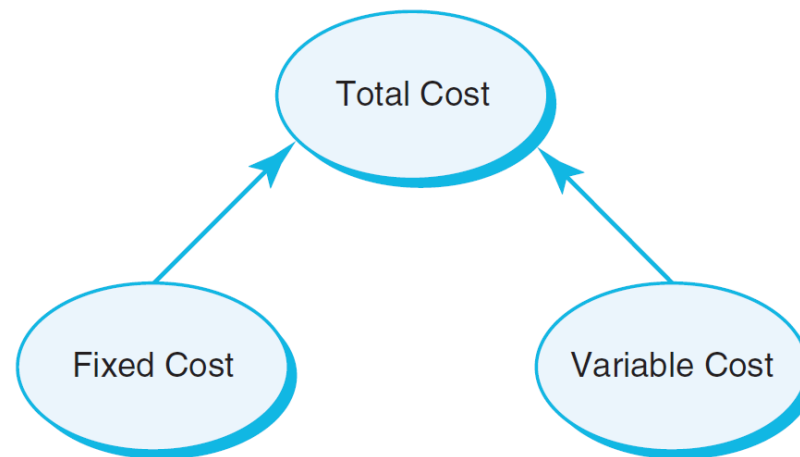


Type of Models



Descriptive Models

- Descriptive Decision Models
 - Simply tell “what is” and describe relationships
 - Do not tell managers what to do
 - visually show how various model elements relate to one another

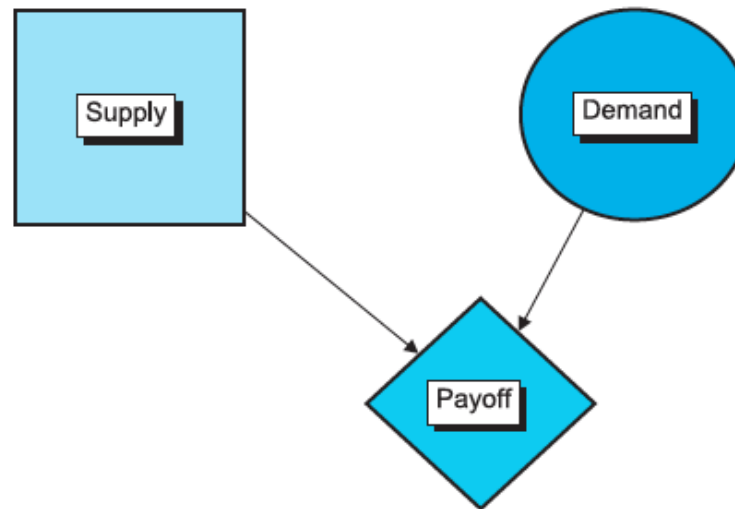


Predictive Models

- Incorporate uncertainty to help managers analyze risk.
- Aim to predict what will happen in the future.
- Uncertainty is imperfect knowledge of what will happen in the future.
- Risk is associated with the consequences of what actually happens.
- Help decision makers identify the best solution:
 - Optimization - finding values of decision variables that minimize (or maximize) something such as cost (or profit).
 - Objective function - the equation that minimizes (or maximizes) the quantity of interest.
 - Constraints - limitations or restrictions.
 - Optimal solution - values of the decision variables at the minimum (or maximum) point.

Representing Models - Graphical Models

- Graphical models attempt to portray graphically how different elements of a problem are related — what effects what.
 - A very simple graphical model, called an influence diagram, is shown below.



Representing Models – Algebraic Models

- Algebraic models use algebraic equations and inequalities to specify a set of relationships in a very precise way.
- A typical example is the “product mix” model shown below.

$$\max \sum_{j=1}^n p_j x_j$$

$$\text{subject to } \sum_{j=1}^n a_{ij} x_j \leq b_i, \quad 1 \leq i \leq m$$

$$0 \leq x_j \leq u_j, \quad 1 \leq j \leq n$$

The background features a large red triangle on the left side, pointing towards the bottom right. A white diagonal line separates this red triangle from the rest of the image. The remaining area is composed of two shades of blue: a medium blue triangle pointing towards the bottom left, and a dark blue area that fills the top and right portions of the frame.

Models to Metrics, Measures, & Variables

Data for Business Analytics

- A **variable** (or field or attribute) is a characteristic of members of a population
 - E.g. unit price, weight, height, gender, or salary.
- **Metrics** are used to quantify performance.
 - Discrete metrics involve counting
 - on time or not on time, number or proportion of on time deliveries
- Continuous metrics are measured on a continuum
 - Delivery time, package weight, purchase price
- **Measures** are numerical values of variables and metrics

Data Sets and Observations

- A **data set** is usually a rectangular array of data, with variables in columns and observations in rows.
- An **observation** (or case or record) is a list of all variable values for a single member of a population.

Types of Data

- A variable is **numerical** if meaningful arithmetic can be performed on it.
- Otherwise, the variable is **categorical**.
- Other types of data:
 - date variable.
 - Excel stores dates as numbers, but dates are treated differently from typical numbers.
 - text variable
 - Typically for remarks and verbatim responses

Types of Data - Numerical

- A variable is **numerical** if meaningful arithmetic can be performed on it.
- A numerical variable is **discrete** if it results from a count, such as the number of children.
- A **continuous** variable is the result of an essentially continuous measurement, such as weight or height.
 - **Ratio:** continuous values and have a natural zero point
 - E.g. monthly sales; delivery times
 - **Interval:** ordinal data but with constant differences between observations
 - No true zero point
 - E.g. temperature readings, SAT scores

Types of Data - Categorical

- A variable is **categorical** if no arithmetic operations can be performed on it.
 - **Ordinal** if there is a natural ordering of its possible values.
 - Data that is ranked or ordered according to some relationship with one another
 - No fixed units of measurement
 - E.g. day of week; status of job
 - **Nominal** if there is no natural ordering.
 - Data placed in categories according to a specified characteristic
 - Categories bear no quantitative relationship to one another
 - E.g. customer's location (America, Europe, Asia); employee classification (manager, supervisor, associate)
- Categorical variables can be coded numerically or left uncoded.
 - A **dummy variable** is a 0–1 coded variable for a specific category.
 - It is coded as 1 for all observations in that category and 0 for all observations not in that category.
- Categorizing a numerical variable by putting the data into discrete categories (called **bins**) is called **binning** or **discretizing**.
 - A variable that has been categorized in this way is called a **binned** or **discretized variable**.
 - E.g. waist measurements (continuous) into small, medium and large

Types of Data – Role of Time

- **Cross-sectional data:** Data collected from several entities at the same, or approximately the same point in time.
- **Time series data:** Data collected over several time periods.
 - Graphs of time series data are frequently found in business and economic publications.
 - Help analysts understand what happened in the past, identify trends over time, and project future levels for the time series.
- Main interest in how measures change/grow over time
 - Information on change/growth is not collected.



Sources of Data

- A **population** includes all of the entities of interest in a study (people, households, machines, etc.)
- A **sample** is a subset of the population, often randomly chosen and preferably representative of the population as a whole.

Outliers

- An **outlier** is a value or an entire observation (row) that lies well outside of the norm.
- Some statisticians define an outlier as any value more than three standard deviations from the mean, but this is only a rule of thumb.
- Even if values are not unusual by themselves, there still might be unusual **combinations** of values.
- When dealing with outliers, it is best to run the analyses two ways: with the outliers and without them.

Missing Values

- Most real data sets have gaps in the data.
- There are two issues: how to detect these **missing values** and what to do about them.
- The more important issue is what to do about them:
 - One option is to simply ignore them. Then you will have to be aware of how the software deals with missing values.
 - Another option is to fill in missing values with the average of nonmissing values, but this isn't usually a very good option.
 - A third option is to examine the nonmissing values in the *row* of a missing value; these values might provide clues on what the missing value should be.

Filtering

- Finding records that match particular criteria is called filtering.
 - Typically through SQL queries – Where subcommand
- In Excel, create an Excel table - automatically provides dropdown arrows next to the fields to filter.
 - Three ways to filter any rectangular data set with variable names:
 1. Use the Filter button from the Sort & Filter dropdown list on the Home ribbon.
 1. Use the Filter button from the Sort & Filter group on the Data ribbon.
 1. Right-click any cell in the data set and select Filter. You get several options, the most popular of which is Filter by Selected Cell's Value.

The background consists of several overlapping triangles. A large red triangle is on the left side. A white triangle is positioned between the red triangle and the blue triangles. Two shades of blue triangles are on the right side, with a darker blue triangle at the top and a medium blue triangle below it. The text "Questions and Final Comments" is centered in the white triangle.

Questions and Final Comments

An Impactful Use of Analytics

- What is the airspeed velocity of an unladen swallow?
 - The source of the question (<https://www.youtube.com/watch?v=y2R3FvS4xr4>)
- There needs to be some prior research
 - Discussions with the clients (https://www.youtube.com/watch?v=H4_9kDO3q0w)
 - The Strouhal Number in Cruising Flight (<http://style.org/strouhalflight/>)
- The analytics (<http://style.org/unladenswallow/>)
 - And there was much rejoicing: $\approx 11 \text{ ms}^{-1}$ (24 miles per hour)
- Post analytics activities
 - The discussion of the results (<http://style.org/unladenswallow/comments/>)
 - Alternative hypotheses and analysis (<http://style.org/unladenswallow/theories/>)
- Business applications
 - Siri
 - Wolfram Alpha
 - Google Trends
 - Security questions



Seminar Review

- Problem Analysis
 - Understanding the problem
 - Searching the literature
 - Generating Research Questions
- Developing Testable Hypotheses
 - When to hypothesize
 - Types of hypotheses
- Hypotheses to Models
- Metrics, Measures, Variables
 - Data types
 - Data sources
- Reporting Business Analytics