Data Engineering for MADS

EBM213A05.2022-2023.1

FREE EDITION*

SUMMARY OF EVERYTHING FROM WEEK 1

LECTURES + READINGS + CHAPTERS

Enhanced with a dynamic table of contents.

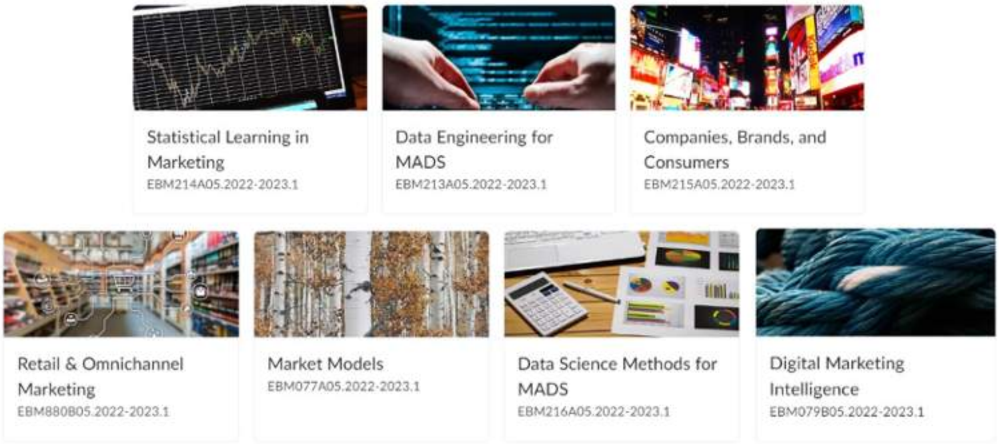
Note from MADS MADLAD:

Thank you for checking out my free summary. When I was writing these I sometimes struggled with this program, but there were no summaries available. This is why I decided to write something that is truly complete with a lot of effort put into it.

It helped me and my friends get good grades, but I always had you in mind, the future reader. When necessary, I always went the extra mile to make my summaries, more readable, organized and complete.

If you feel like it, leave me a review of how the course is going using this summary, it will make my day to hear your feedback.

Check out my other extensive summaries for other MADS courses:



Statistical Learning in Marketing
EBM214A05.2022-2023.1

Data Engineering for MADS
EBM213A05.2022-2023.1

Companies, Brands, and Consumers
EBM215A05.2022-2023.1

Retail & Omnichannel Marketing
EBM880B05.2022-2023.1

Market Models
EBM077A05.2022-2023.1

Data Science Methods for MADS
EBM216A05.2022-2023.1

Digital Marketing Intelligence
EBM079B05.2022-2023.1

Contact info:

If you need help or have an inquiry, contact me: https://www.georgedreemer.com

Connect with me on LinkedIn: https://www.linkedin.com/in/georgedreemer/

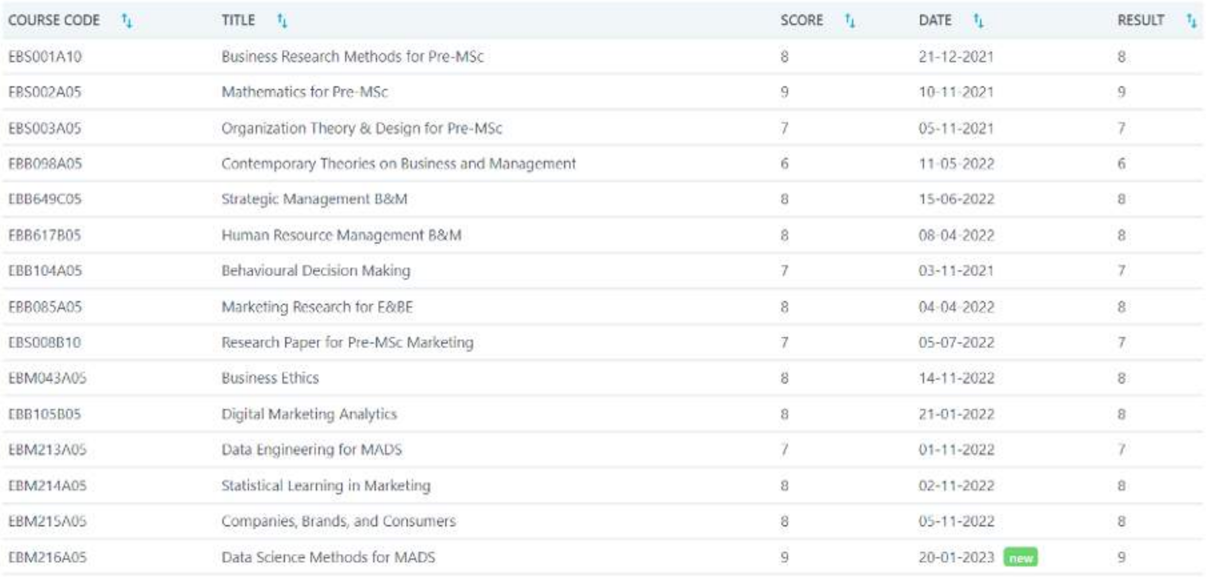
Donations:

By no means am I looking for fellow students to send me money! But if you feel like sending me some ETH or BTC, you can do so here:
--> **ETH:** 0x123e086c6808459e7fC6Ac7F64a77dBA1dDe0149
--> **BTC:** bc1qgwzc82vph5v8rmzef4ywechjf85772n7m2e22g

MADS MADLAD

wishes you good luck & perseverance.

Grades Testimony:



COURSE CODE	TITLE	SCORE	DATE	RESULT
EBS001A10	Business Research Methods for Pre-MSc	8	21-12-2021	8
EBS002A05	Mathematics for Pre-MSc	9	10-11-2021	9
EBS003A05	Organization Theory & Design for Pre-MSc	7	05-11-2021	7
EBB098A05	Contemporary Theories on Business and Management	6	11-05-2022	6
EBB649C05	Strategic Management B&M	8	15-06-2022	8
EBB617B05	Human Resource Management B&M	8	08-04-2022	8
EBB104A05	Behavioural Decision Making	7	03-11-2021	7
EBB085A05	Marketing Research for E&BE	8	04-04-2022	8
EBS008B10	Research Paper for Pre-MSc Marketing	7	05-07-2022	7
EBM043A05	Business Ethics	8	14-11-2022	8
EBB105B05	Digital Marketing Analytics	8	21-01-2022	8
EBM213A05	Data Engineering for MADS	7	01-11-2022	7
EBM214A05	Statistical Learning in Marketing	8	02-11-2022	8
EBM215A05	Companies, Brands, and Consumers	8	05-11-2022	8
EBM216A05	Data Science Methods for MADS	9	20-01-2023 new	9

Table of Contents

Week 1 (Lectures+Readings)

Lecture 0 (Intro) & 1.1 (MD->MQ->RQ) & HBR article

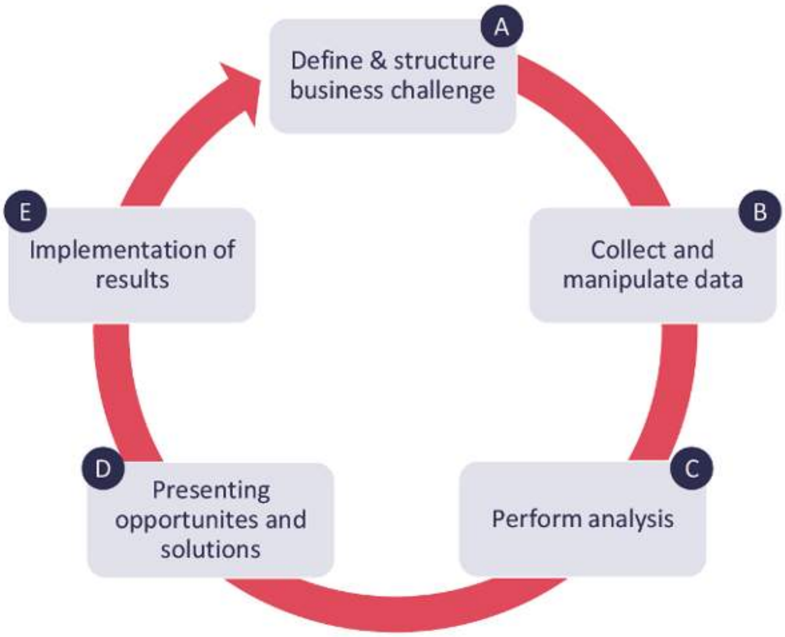

A
Define & structure business challenge
B
Collect and manipulate data
C
Perform analysis
D
Presenting opportunites and solutions
E
Implementation of results

FIGURE 12.4 Phases of the analytical cycle

The 5 Phases of the Analytical Cycle entail to:

(1) define and structure the <u>business challenge</u>

(2) collect and manipulate the <u>data</u>

(3) perform <u>analysis</u>

(4) present opportunities and <u>solutions</u>

(5) <u>implementation</u> of results

Management Dilemma and Questions, Research Questions

Management Dilemma: a symptom of an underlying problem.

- Profits are decreasing
- Target of improving shareholder value by 5% should be met
- Marketing efforts lack effectiveness
- Increasing handle time at the Customer Service Desk
- Underperformance of several sales force teams
- Unanticipated sudden increases in demand

From Management Dilemma to Management Questions:

- Discussion with relevant stakeholders
- Interviews with industry experts
- Secondary data analysis

Goal: restate the MD in terms of underlying problem.

Usually starting with:

- Should we…? (choice of purpose)
- How can we…?
- Why do we…?

Management Question defined:

- Management dilemma restated in question form
- Defined in terms of the underlying problem
- Preferably linked to a Key Performance Indicator (KPI)
- Does not specify the research that needs to be done
- Questions are still broad

Example questions:

- What should be done to increase conversion?
- Should we use new and promising advertising channels to improve marketing ROI?
- Why is the number of new contracts closed by several sales force teams lower than expected?

From Management Question to Research Question **(example):**

Let's assume the Management Question is:

- How can we increase sales in the Northern region this year?

Possible Research Questions:

- What causes decrease in the sales of the Northern region?
- What is the effect of the company-wide pricing strategy on the Northern region?
- **Management Question defined:**

Management Question vs. Research Question:

Management question	Research question
Asks what the decision maker needs to do	Asks what research should be conducted
Action oriented	Information oriented

From Research Question to Analysis Questions:

Using the 5W's of the opportunity finding-method helps you to come up with a good set of analysis questions.

- Who?
- What?
- Where?
- When?
- Why?

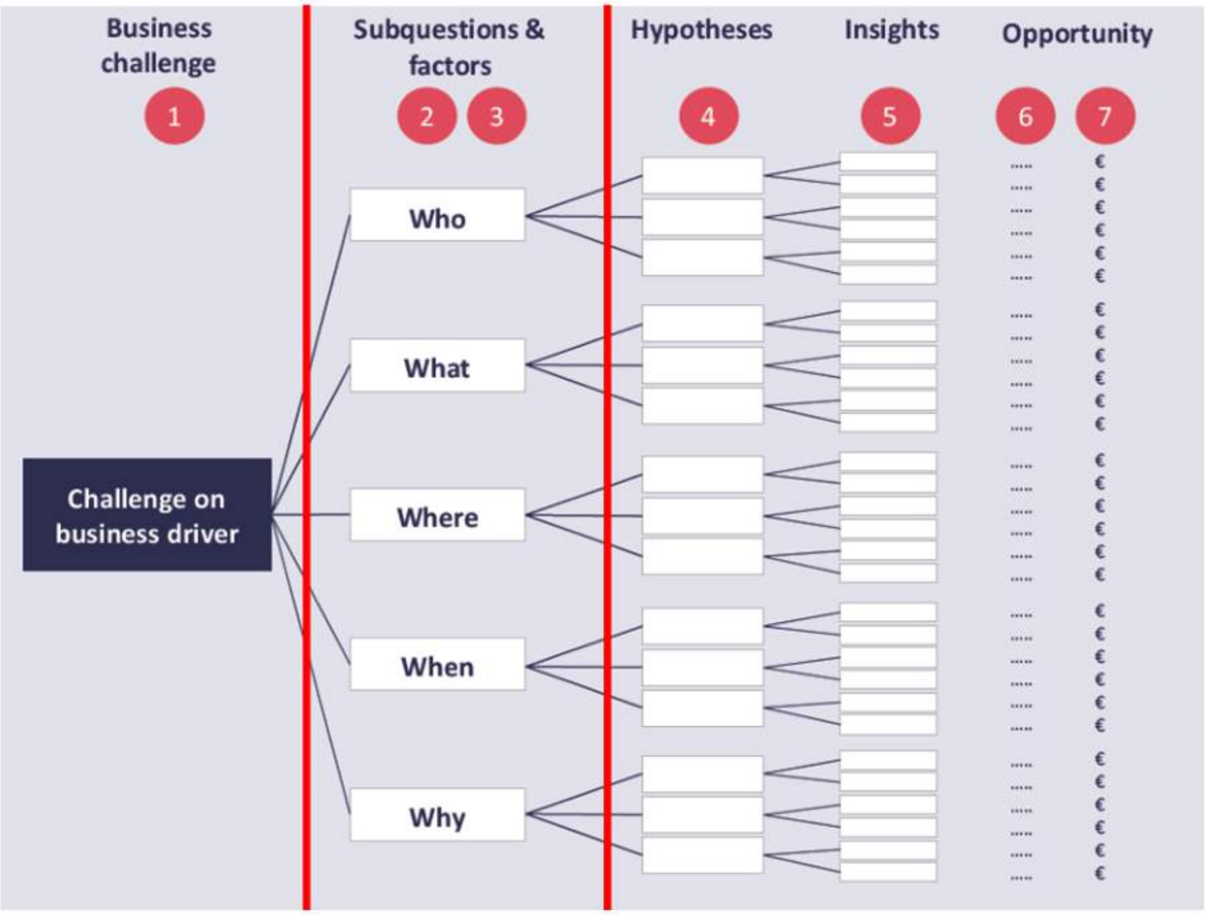
7 steps of the opportunity tree

1. **Business challenge:** the starting point of the tree, defined in measurable objectives.
2. **Sub-questions:** translate business challenge into sub-questions.
3. **Factors:** define which levers you can influence or use.
4. **Hypotheses:** make a 'braindump' of all possible hypotheses.

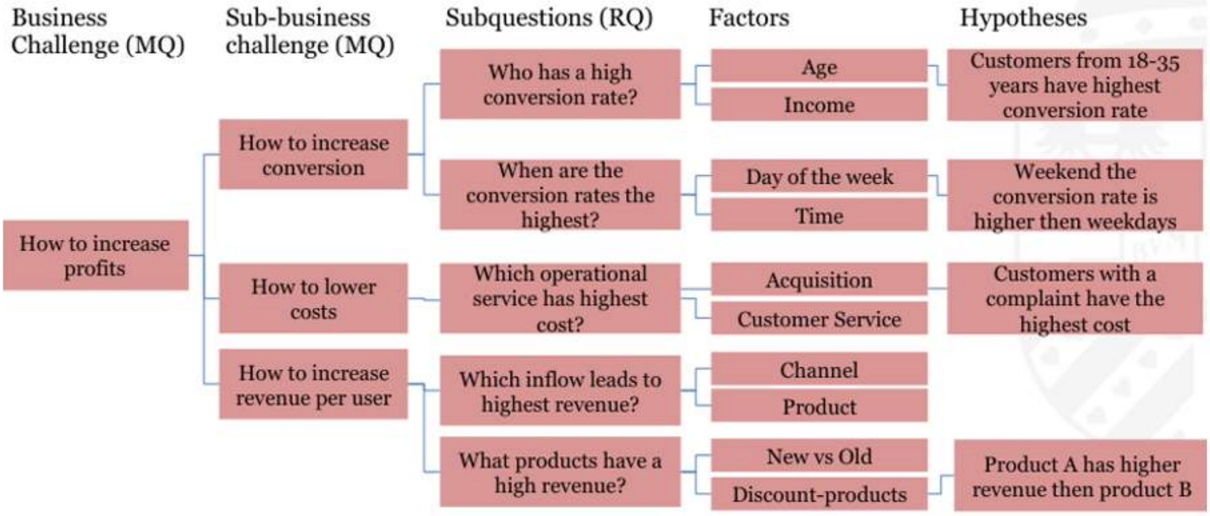
----- exhaustive opportunity tree includes these -----

5. **Insights:** determine the analyses questions to check the hypotheses and to identify areas with high potential.
6. **Initiatives:** come up with potential initiatives to realize the targets/objectives.
7. **Impact:** calculate the monetary impact (+ or -) of initiatives and identify the most promising ones.

Exhaustive opportunity tree:

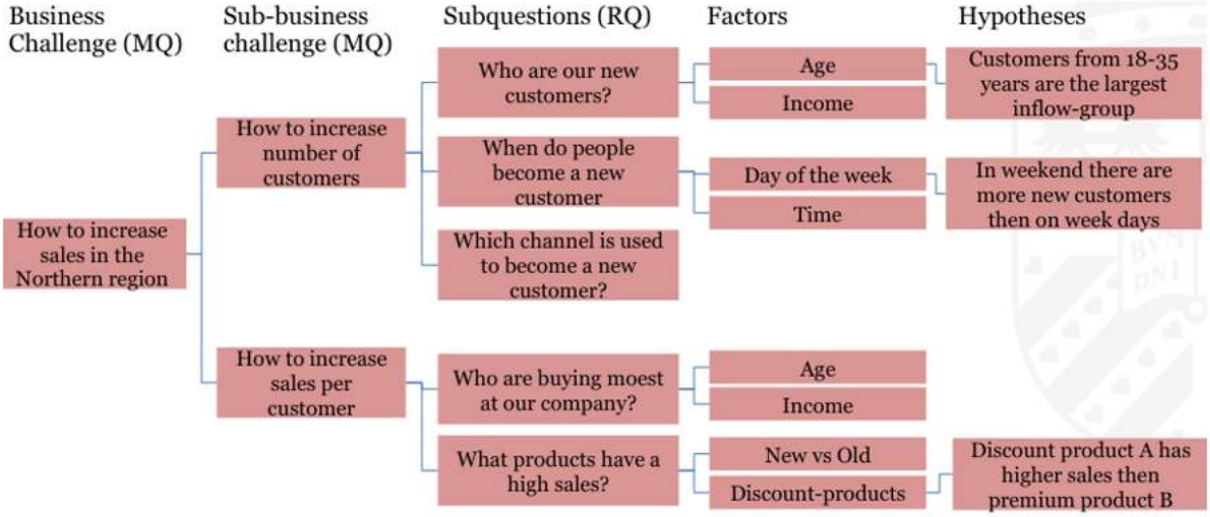


Example of non-exhaustive opportunity tree:


Business Challenge (MQ)
Sub-business challenge (MQ)
Subquestions (RQ)
Factors
Hypotheses
How to increase profits
How to increase conversion
How to lower costs
How to increase revenue per user
Who has a high conversion rate?
When are the conversion rates the highest?
Which operational service has highest cost?
Which inflow leads to highest revenue?
What products have a high revenue?
Age
Income
Day of the week
Time
Acquisition
Customer Service
Channel
Product
New vs Old
Discount-products
Customers from 18-35 years have highest conversion rate
Weekend the conversion rate is higher then weekdays
Customers with a complaint have the highest cost
Product A has higher revenue then product B

Tips on how to fill in the 5 W’s, factors and hypothesis:

- Based on hypothesis of the stakeholder
- Based on literature
- Based on steps in a customer journey
- Imagine you’re the customer


Business Challenge (MQ)
Sub-business challenge (MQ)
Subquestions (RQ)
Factors
Hypotheses
How to increase sales in the Northern region
How to increase number of customers
How to increase sales per customer
Who are our new customers?
When do people become a new customer
Which channel is used to become a new customer?
Who are buying moest at our company?
What products have a high sales?
Age
Income
Day of the week
Time
Age
Income
New vs Old
Discount-products
Customers from 18-35 years are the largest inflow-group
In weekend there are more new customers then on week days
Discount product A has higher sales then premium product B

Next step – write a research proposal:

Every proposal includes two basic sections – statement of the research question (1) and brief description of the research methodology (2). (Blumberg et al., 2014, Section 2.4) In other words, **what** are we researching and **how** are we going to research.

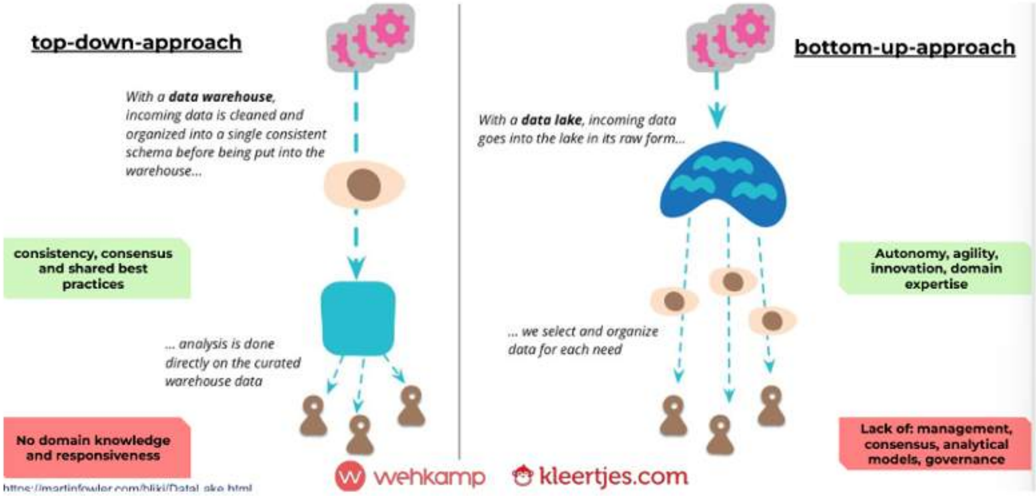
Lecture 1.2 - supplement (Wehkamp lecture 2022)

7 Elements of a Data Strategy:

1. Business Requirements
2. Sourcing and Gathering Data
3. Technology Infrastructure Requirements
4. Turning Data into Insights
5. People and Process
6. Data Governance
7. The Roadmap

Data Warehouse vs. Data Lake

1. **Data Warehouse:** Incoming data is cleaned and organized into a single consistent schema before being stored. Analysis is done directly on the curated data.
   - **Top-Down Approach**
   - **Pros:** Consistency, consensus & shared best practices
   - **Cons:** No domain knowledge, no responsiveness
2. **Data Lake:** Incoming data goes into the lake in its raw form. We select and organize data for each need.
   - **Bottom-Up Approach**
   - **Pros:** Autonomy, agility, innovation, domain expertise
   - **Cons:** Lack of management, consensus, governance



top-down-approach

With a **data warehouse**, incoming data is cleaned and organized into a single consistent schema before being put into the warehouse...

consistency, consensus and shared best practices

... analysis is done directly on the curated warehouse data

No domain knowledge and responsiveness

https://martinfowler.com/bliki/DataLake.html

bottom-up-approach

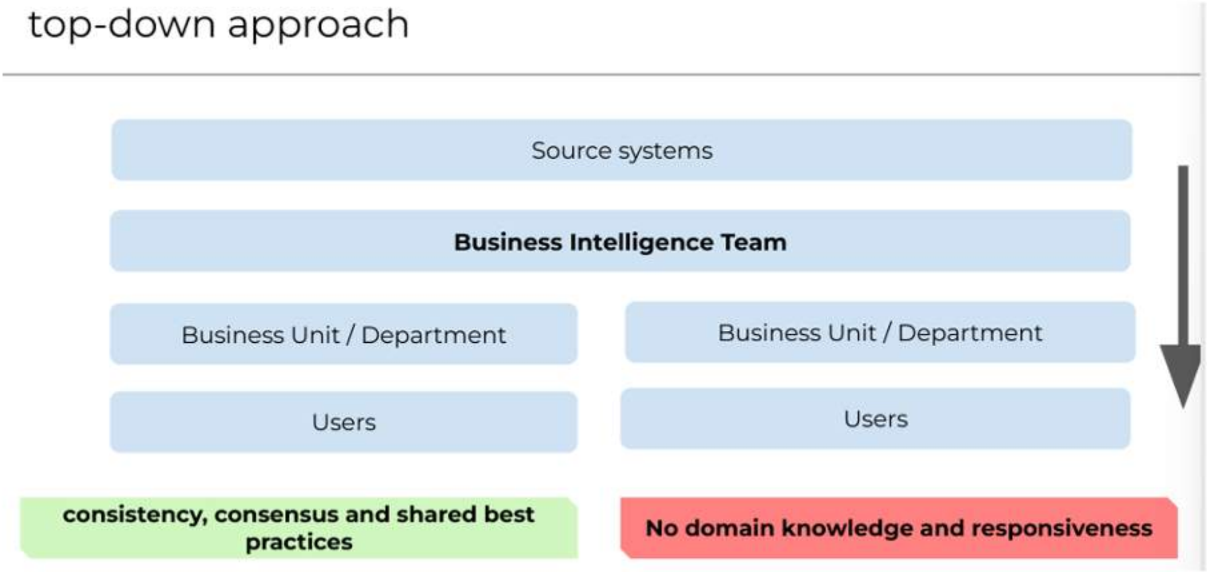
With a **data lake**, incoming data goes into the lake in its raw form...

Autonomy, agility, innovation, domain expertise

... we select and organize data for each need

Lack of: management, consensus, analytical models, governance

Top-down vs. Bottom-up:



top-down approach

Source systems

Business Intelligence Team

Business Unit / Department

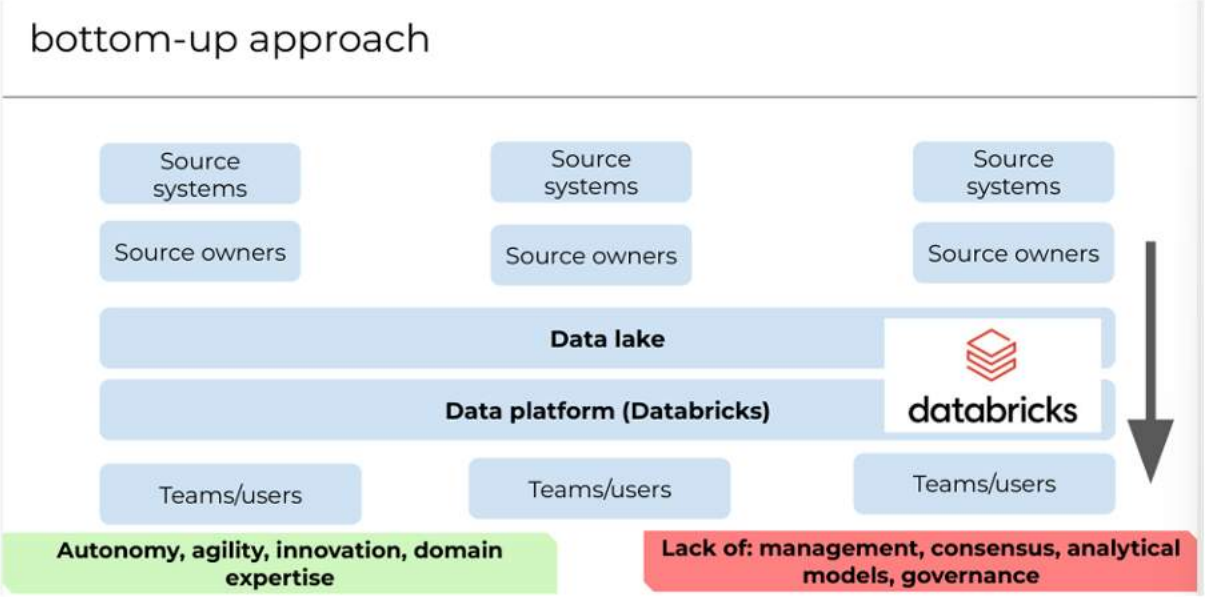
Business Unit / Department

Users

Users

consistency, consensus and shared best practices

No domain knowledge and responsiveness



bottom-up approach

Source systems

Source systems

Source systems

Source owners

Source owners

Source owners

Data lake

Data platform (Databricks)

databricks

Teams/users

Teams/users

Teams/users

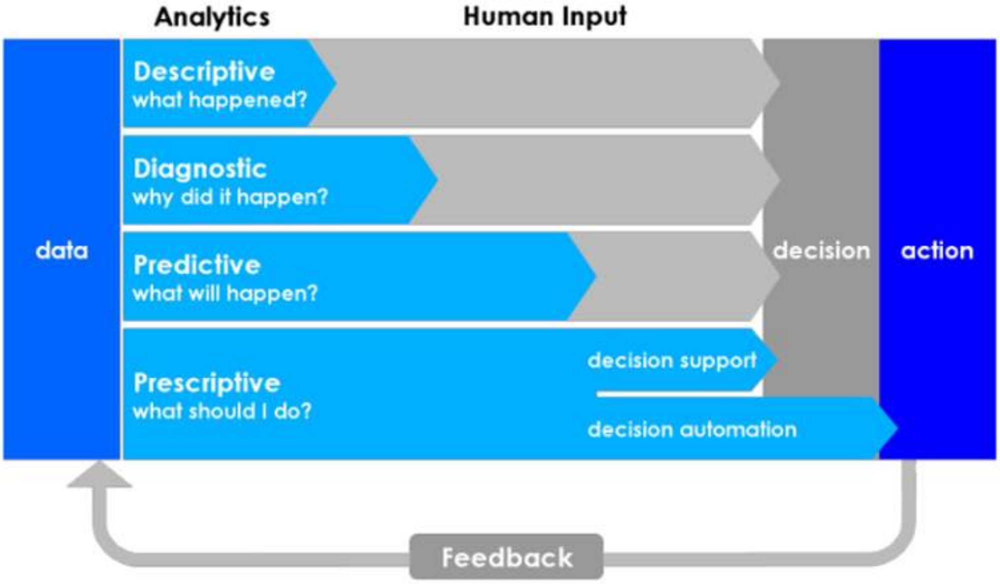
Autonomy, agility, innovation, domain expertise

Lack of: management, consensus, analytical models, governance

Wehkamp is now been experimenting with a hybrid approach, which entails both governed and non-governed ways of work.

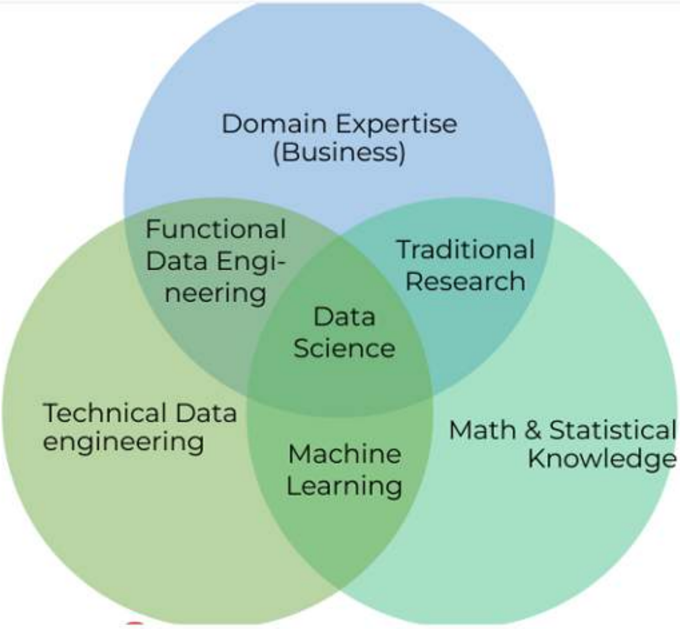
4 Ways to work with Data

- **Descriptive:** What happened?
- **Diagnostic:** Why did it happen?
- **Predictive:** What will happen?
- **Prescriptive:** What should be done?



Data Science – mix of 3 fields of knowledge

1. Technical Data Engineering
2. Domain Expertise (Business)
3. Math & Statistical Knowledge

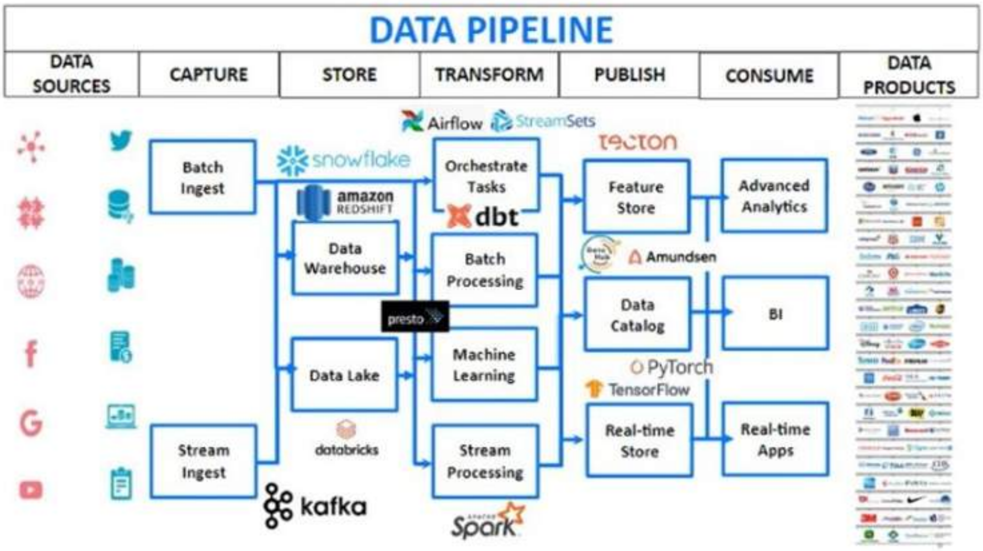


Domain Expertise (Business)

- Understanding the customer & process
- Make it actionable
- From business intelligence to visual analytics
- Tableu (visualisations) enable the business to use data

Technical Data Engineering

- **Data engineers:** develop constructs (1), test and maintain data architectures (2), such as databases and large-scale processing systems.
- Build functional data models and products
- Build ETL (Extract, Transform, Load) processes
- Responsible for definitions and structures of data

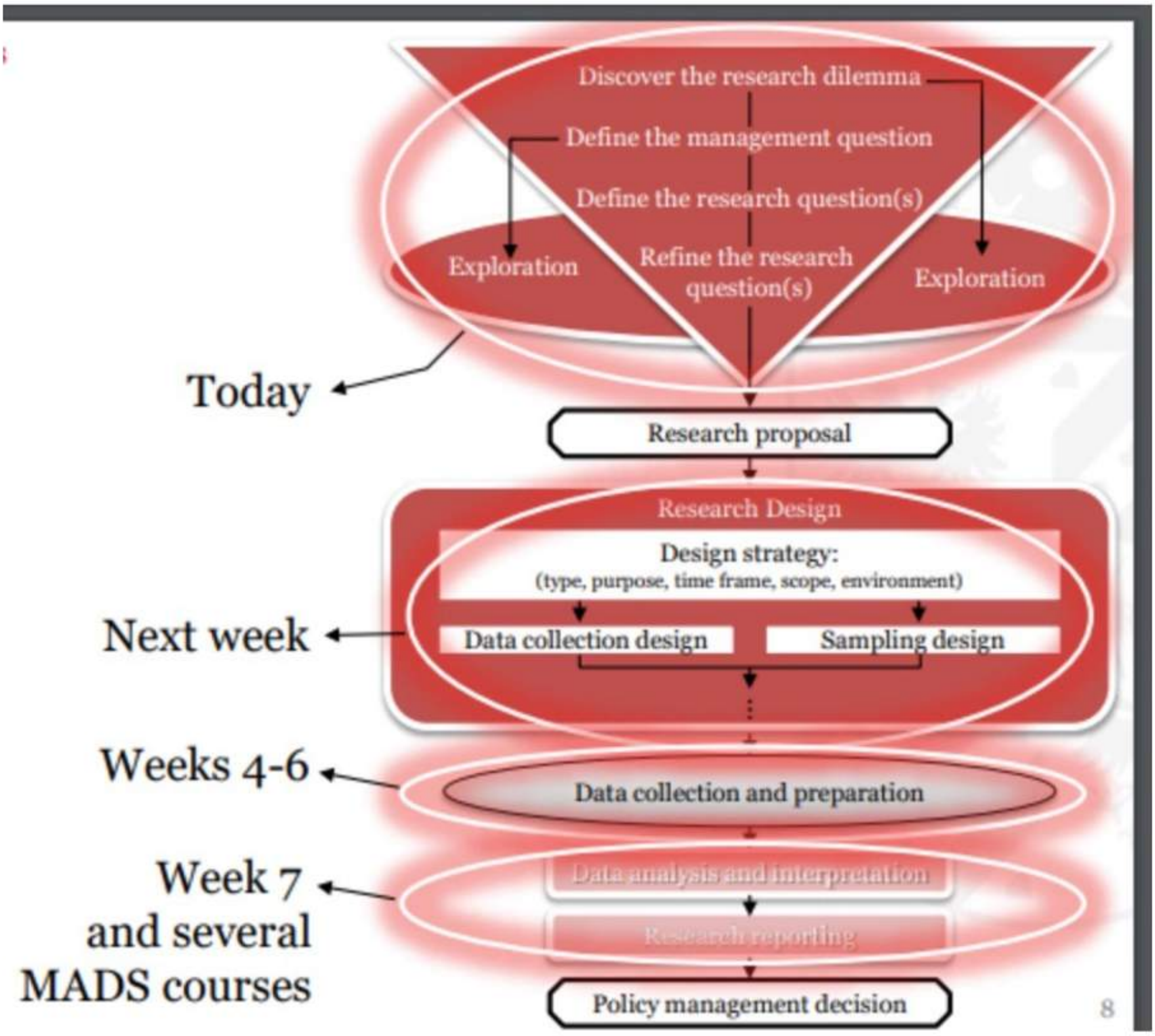


Math & Statistical Knowledge

- Experimental design building, Model Fitting
- No failures = no learning
- The challenge is not to get a working model, but to get the model to add value to the business

Note from MADS Madlad: *Wehkamp lecture continues with examples and description of the Wehkamp dataset. I think there is no more theory to summarise. Please refer to the original lecture for further materials on this lecture.*

Recap of the course & assignment(s):



Discover the research dilemma

Define the management question

Define the research question(s)

Exploration

Refine the research question(s)

Exploration

Today

Research proposal

Research Design

Design strategy:
(type, purpose, time frame, scope, environment)

Data collection design

Sampling design

Next week

Weeks 4-6

Data collection and preparation

Week 7
and several
MADS courses

Data analysis and interpretation

Research reporting

Policy management decision

8

Reading – HBR Article (optional)

4 Steps to Management Questions

Step 1: Articulate the problem simply.

"We are looking for X in order to achieve Z as measured by W."

- What is the basic need?
- What is the desired outcome?
- Who stands to benefit and why?

Step 2: Justify the need.

- Does it align with organization's strategic goals?
- What are the desired benefits & how do we measure them?
- How will we ensure that a solution is implemented?

Step 3: Contextualize the problem in terms of the past tries.

- What approaches have we tried in the past?
- What have others tried?
- What are in/external constraints on implementing a solution?

Step 4: Write the problem statement.

- Is the problem actually many problems?
- What requirements must a solution meet?
- Which problem solvers should we engage?
- What do solvers need to submit?
- What are incentives for the solvers?
- How will solutions be evaluated and success measured?

Reading – Book: Verhoef et al. (11.6 required +11.1 optional)

Section 11.6 (required): "OPPORTUNITY FINDING AS A METHODOLOGY TO CREATE MORE VALUE"

Note from MM: *Opportunity finding is basically the opportunity tree from lecture 1.*

Data Scientists could immediately start analyzing the data and start data science projects, but that is not the point. The goal is to have **value impact** in the organization.

For that reason, this chapter discusses a methodology used to detect opportunities: **opportunity finding**.

Opportunity finding: structured way to identify solutions and breaking them down into initiatives.

It helps put a focus on the most valuable initiatives and facilitates monitoring progress on objectives for the defined business challenge.

7 Steps of Opportunity finding

1. **Business Challenge (BC)**
   a. Linked to the phases of the customer lifecycle (e.g customer acquisition) or the phases of the customer journey.
   b. Quantified in terms of the necessary potential to be realized. In other words, the delta from current to desired state (e.g increase X from 3% to 8%).
   c. Expressed in a monetary value and measurable by at least one or more defined KPIs.

   Example

   "We aim to reduce the customer churn from 8.5% to 7%, to achieve an extra EBIT of 2.5m euro."

2. **Sub Questions**
    a. Break BC down into sub questions. Use the 5 W's: Who, What, Where, When & Why.
    b. Specify BC for every W, when possible. Focus on the most important W's based on situation.

Example (continued)

Who: Who are the customers that show an above or below avg. churn percentage?

What: What products have a higher or lower product churn compared to others?

3. **Factors**
    a. The factors behind the sub questions define the levers that can actually be used to start defining initiatives for the business challenge.

Example (continued)

In our "Who": we could define **age** as a possible factor that might be relevant in finding groups with a higher or lower churn percentage. Another angle could be to break down the customer base by value, creating **value segments** to explore whether different segments have different churn.

4. **Hypotheses**
    a. The hypothesis step is crucial in guiding the analysis process. It is helpful to make a "brain dump" of all possible hypotheses.

Example (continued)

Who (two factors: age and value) + Why (reason of churn):

"<u>High value</u>, <u>young</u> customers, show a high churn rate, due to <u>cheaper alternatives</u> and <u>low switching barriers</u>."

5. **Insights**
   a. Determine analyses questions to check the hypothesis and identify areas with high potential.
   b. Potential is identified in the differences in performance we can uncover along the primary KPI of our business challenge (in our case this is churn).

6. **Initiatives**
   a. Initiatives are specified as ideas to realize by addressing a target group, via one or more channels, at a specified time, with a well-targeted proposition.

7. **Impact (in euros/dollars/...)**
   a. The financial impact is calculated per initiative to identify the most promising ones, but also to validate that the initial potential of the business challenge is feasible.

   Tip (from book):

   We suggest plotting the initiatives in a matrix with the two dimensions: necessary effort to realize an initiative and the potential value of every initiative. Then start with the low effort, high value ones first.

Section 11.1 (optional): "OPPORTUNITY FINDING FOR SURE.COM"

***Note from MADS Madlad:** This is an assignment/exercise in the book for further material please refer to the original book. In short it presents you with information about the business and your task is to write down each step of the opportunity finding process summarized above (Section 11.6).*

Reading – Book: Business Research Methods (2.1 optional + 2.2 required)

Section 2.1 (optional): The research process

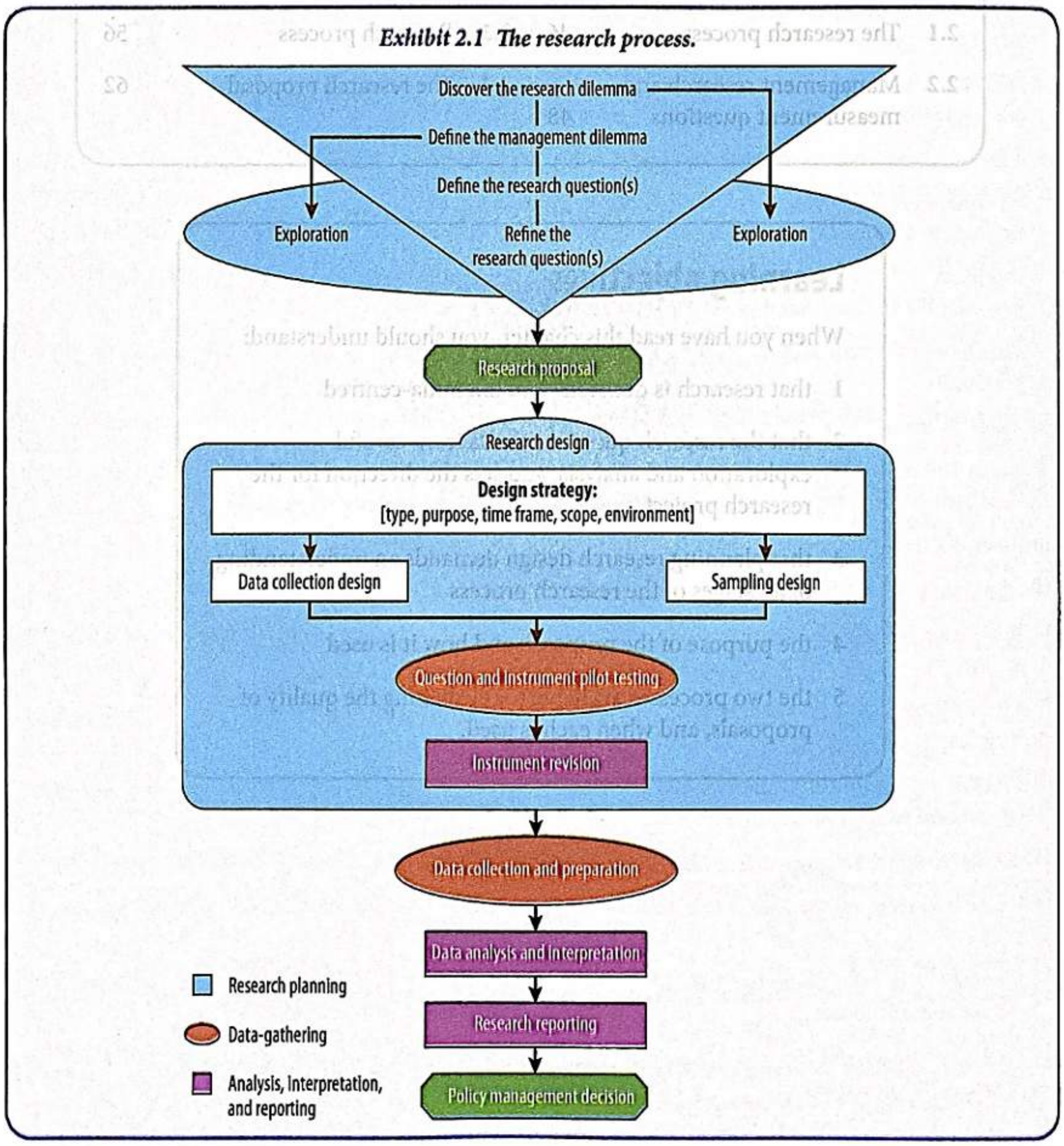


Exhibit 2.1 The research process.

Discover the research dilemma
Define the management dilemma
Define the research question(s)
Exploration
Refine the research question(s)
Exploration
Research proposal
Research design
Design strategy: [type, purpose, time frame, scope, environment]
Data collection design
Sampling design
Question and instrument pilot testing
Instrument revision
Data collection and preparation
Data analysis and interpretation
Research reporting
Policy management decision
Research planning
Data-gathering
Analysis, Interpretation, and reporting