

# Backwards Design in Data Science

---

Nick Eubank

# Backwards Design

Approach to planning data science projects

# Backwards Design

Approach to planning data science projects

- (Though backwards design isn't unique to DS)

Goals:

- Minimize wasted effort

# Backwards Design

Approach to planning data science projects

- (Though backwards design isn't unique to DS)

Goals:

- Minimize wasted effort
- Make sure you develop explicit goals
  - Not get lost in your tools and data

# Backwards Design

Start with where you want to end up, then work backwards

# Backwards Design

# Backwards Design

1. Determine Problem / Topic Area

# Backwards Design

1. Determine Problem / Topic Area
2. What *question* are you seeking to answer?



# Backwards Design

1. Determine Problem / Topic Area
2. What *question* are you seeking to answer?
3. What does an answer to your question look like?

# Backwards Design

1. Determine Problem / Topic Area
2. What *question* are you seeking to answer?
3. What does an answer to your question look like?
4. What variables do you need to generate that answer?

# Backwards Design

1. Determine Problem / Topic Area
2. What *question* are you seeking to answer?
3. What does an answer to your question look like?
4. What variables do you need to generate that answer?
5. What data contains those variables?

## Step 0: Define the Problem / Topic

Why are you doing this project?

## Step 0: Define the Problem / Topic

Why are you doing this project?

What **motivates** your investigation?

## Step 0: Define the Problem / Topic

Why are you doing this project?

What motivates your investigation?

Examples:

- We don't know how to reduce mass incarceration.
- My business can't identify potential new customers.
- We don't know who is going to develop Alzheimers, so we can't test early interventions.

## Step 0: Define the Problem / Topic

This *seems* like it should be the easiest part of the project!

## Step 0: Define the Problem / Topic

This *seems* like it should be the easiest part of the project!  
Why else would you be paying someone to do a project for you?



## Step 0: Define the Problem / Topic

This *seems* like it should be the easiest part of the project!  
Why else would you be paying someone to do a project for you?

But very often you're stakeholder just wants to do "data science," and assumes that you can just throw machine learning at data and make money.

## Step 0: Define the Problem / Topic

This *seems* like it should be the easiest part of the project!  
Why else would you be paying someone to do a project for you?

But very often you're stakeholder just wants to do "data science," and assumes that you can just throw machine learning at data and make money.

- It is *your* job to ensure that a goal has actually been articulated,
- and make sure you and the stakeholder are clear on what that is.

## Step 1: What **question** are you seeking to answer?

The tools of data science are fundamentally designed to **answer questions**,

## Step 1: What **question** are you seeking to answer?

The tools of data science are fundamentally designed to **answer questions**, so to before you pick your tools, you have to decide **what question you wish to answer**.

## Step 1: What **question** are you seeking to answer?

The tools of data science are fundamentally designed to **answer questions**, so to before you pick your tools, you have to decide **what question you wish to answer**.

⇒ The MOST important part of your project

## Step 1: What **question** are you seeking to answer?

Most important because:

## Step 1: What **question** are you seeking to answer?

Most important because:

- if you can't define the question you are seeking to answer, **you'll find yourself lost in your data**, or worse

## Step 1: What **question** are you seeking to answer?

Most important because:

- if you can't define the question you are seeking to answer, **you'll find yourself lost in your data**, or worse
- after finishing your project, you'll realize the question you answered doesn't help solve the problem that motivated you.  
(or your stakeholder will say it doesn't actually help them.)



## Step 1: What **question** are you seeking to answer?

Most important because:

- if you can't define the question you are seeking to answer, **you'll find yourself lost in your data**, or worse
- after finishing your project, you'll realize the question you answered doesn't help solve the problem that motivated you.  
(or your stakeholder will say it doesn't actually help them.)

⇒ Invest in this stage of your project *before* you dive into the data!

## Step 1: What **question** are you seeking to answer?

A critical feature of a good question is that it is *tractable* and *answerable* in a data science project.

- If your question does not directly imply a course of action in your data science project, it's too vague.

## Step 1: What **question** are you seeking to answer?

Not answerable:

- What policies reduce mass incarceration?
- Can machine learning help me identify potential customers.
- What indicates Alzheimers?

## Step 1: What **question** are you seeking to answer?

Not answerable:

- What policies reduce mass incarceration?
- Can machine learning help me identify potential customers.
- What indicates Alzheimers?

Answerable:

- Does the availability of grand juries result in longer sentences?
- What attributes are common to the customers who buy the most from my business?
- Are there lab results common to patients diagnosed (post-mortem) with Alzheimers not common to patients without Alzheimers?

## Step 1: What **question** are you seeking to answer?

How do I know if my answer is answerable / tractable?

## Step 1: What **question** are you seeking to answer?

How do I know if my answer is answerable / tractable?

1. Can you hypothesize an answer to your question?  
i.e. Can you state what you think might be the answer to your question?

## Step 1: What **question** are you seeking to answer?

How do I know if my answer is answerable / tractable?

1. Can you hypothesize an answer to your question?  
i.e. Can you state what you think might be the answer to your question?
2. Can you imagine what the answer to your question looks like?

## Step 1: What **question** are you seeking to answer?

How do I know if my answer is answerable / tractable?

1. Can you hypothesize an answer to your question?  
i.e. Can you state what you think might be the answer to your question?
2. Can you imagine what the answer to your question looks like?
3. When you read the question, do you immediately start thinking of what kind of data you want to collect?



## Step 2: What does **an answer** to your question look like?

**Write down** what the answer to your question will look like!

## Step 2: What does **an answer** to your question look like?

**Write down** what the answer to your question will look like!

- A figure
- A table or regression
- A dataset with predicted values

## Step 2: What does **an answer** to your question look like?

**Write down** what the answer to your question will look like!

- A figure
- A table or regression
- A dataset with predicted values

⇒ Ask yourself: if I gave that to my stakeholder / put it in a paper, would people be pleased?

## Step 2: What does **an answer** to your question look like?

**Write down** what the answer to your question will look like!

- A figure
- A table or regression
- A dataset with predicted values

⇒ Ask yourself: if I gave that to my stakeholder / put it in a paper, would people be pleased?

(OK, they might want robustness, and extensions, but at its core, is this an answer that would help them solve their problem?)

Step 2: What does **an answer** to your question look like?

## Step 2: What does **an answer** to your question look like?

- **Incarceration:** A regression that shows differences in sentences for arrestees in counties with standing grand juries as compared to counties without standing grand juries, controlling for details of charges.

## Step 2: What does **an answer** to your question look like?

- **Incarceration:** A regression that shows differences in sentences for arrestees in counties with standing grand juries as compared to counties without standing grand juries, controlling for details of charges.
- **Business:** A table showing the performance of a machine learning model that predicts (past) customer behavior using pre-purchase data on customer website interactions (and model parameters).

## Step 2: What does **an answer** to your question look like?

- **Incarceration:** A regression that shows differences in sentences for arrestees in counties with standing grand juries as compared to counties without standing grand juries, controlling for details of charges.
- **Business:** A table showing the performance of a machine learning model that predicts (past) customer behavior using pre-purchase data on customer website interactions (and model parameters).
- **Alzheimers:** A regression showing a strong correlation between certain test results and receiving a positive diagnosis of Alzheimers in (post-mortem) testing.



## Step 2: What does **an answer** to your question look like?

But it's not enough to imagine *one* answer. You should be able to imagine what an answer to your question looks like if your hypothesis **is true** and the if your hypothesis **is false**.

## Step 2: What does **an answer** to your question look like?

But it's not enough to imagine *one* answer. You should be able to imagine what an answer to your question looks like if your hypothesis **is true** and the if your hypothesis **is false**. Otherwise your question isn't falsifiable!

Write down what your answer looks like if your hypothesis is true, *and* if it's false!

### Step 3: What do you need to generate that answer?

Congratulations! You've just specified the goal of your analysis!

### Step 3: What do you need to generate that answer?

Congratulations! You've just specified the goal of your analysis! In my view, that is actually the hardest part of being a good data scientist.

## Step 3: What do you need to generate that answer?

Congratulations! You've just specified the goal of your analysis! In my view, that is actually the hardest part of being a good data scientist.

...Though probably not the part that will take up the majority of your time.

### Step 3: What do you need to generate that answer?

So you now have in mind a table you want to generate. What data and variables do you need to create that result?

### Step 3: What do you need to generate that answer?

So you now have in mind a table you want to generate. What data and variables do you need to create that result?

## Step 3: What do you need to generate that answer?

So you now have in mind a table you want to generate. What data and variables do you need to create that result? For each variable, specify:

1. What do you need the variable to measure?
2. For what population do you need the variable defined?



## Step 4: Where can you get those variables?

1. Where can you get those variables?, and
2. How will you relate your different datasets?

Is this what everyone does?

Is this what everyone does?

Not that I'm aware of.

## Is this what everyone does?

Not that I'm aware of.

Most people who are *successful* seem to do this implicitly

## Is this what everyone does?

Not that I'm aware of.

Most people who are *successful* seem to do this implicitly

People who don't use this, in my experience, tend to flail.

# Final Project

In teams of *up to* three people,

# Final Project

In teams of *up to* three people, you will have to develop *your own project idea* from scratch using this model.

- Just as the last project emphasized all the data tasks *before* analysis,

# Final Project

In teams of *up to* three people, you will have to develop *your own project idea* from scratch using this model.

- Just as the last project emphasized all the data tasks *before* analysis,
- The goal of this is to emphasize all the things you do *before* you touch your data!

If you're looking for a model... look back at the write up of your assignment for the mid-semester project!



You have been approached by a campaign to reduce teen vaping.

Over the past year, they've tried several different pilot programs in several cities.

They just got a huge donation, and want to know what they should do with it.