Doing SCIence

The workflow and questions generation

Filippo Chiarello, Ph.D.

A question you've answered many times:

What's in a data analisis?

Five core activities of data analysis

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

Roger D. Peng and Elizabeth Matsui. "The Art of Data Science." A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC (2015).

Stating and refining the question

Six types of questions

1. **Descriptive:** summarize a characteristic of a set of data
2. **Exploratory:** analyze to see if there are patterns, trends, or relationships between variables (hypothesis generating)
3. **Inferential:** analyze patterns, trends, or relationships in representative data from a population
4. **Predictive:** make predictions for individuals or groups of individuals
5. **Causal:** whether changing one factor will change another factor, on average, in a population
6. **Mechanistic:** explore "how" as opposed to whether

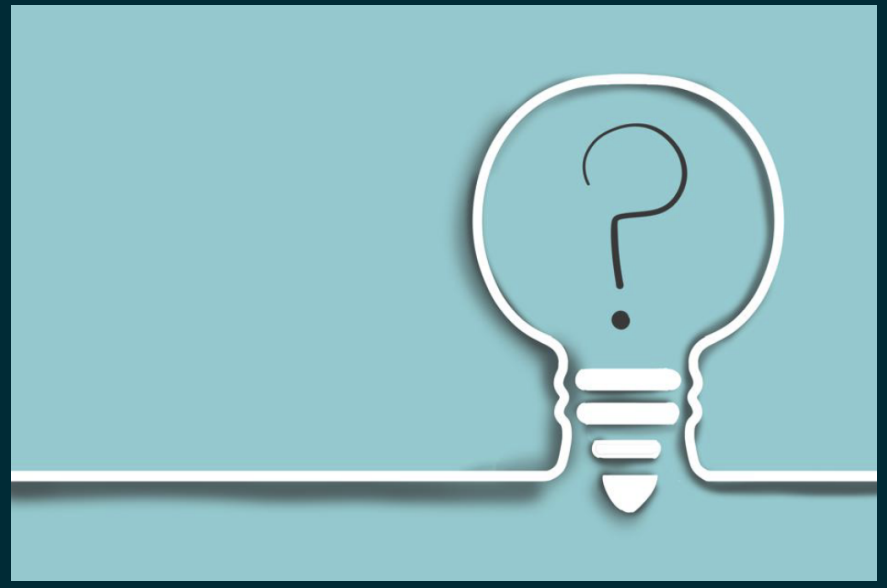
Jeffery T. Leek and Roger D. Peng. "What is the question?." Science 347.6228 (2015): 1314-1315.

Ex: Relation between investin in AI and innovating

1. **Descriptive:** frequency of patents filled by a set of companies
2. **Exploratory:** examine relationships between investing in AI technologies and the number of filled patents
3. **Inferential:** examine whether any relationship between investing in AI and number of patents found in the sample hold for the population at large
4. **Predictive:** what class (sector, dimensions...) of company will get an increased number of patents when investing in AI
5. **Causal:** whether companies that invests in AI patents more than companies that are not investing in AI, if the groups are taken randomly
6. **Mechanistic:** how many money invested in AI produces on average an increment of 1 patent

Everything starts with a question

1. **What** can we define a good question in the context of SCI?
2. **When** do questions comes up?
3. **Who** in the company generates questions?
4. **Where** can I find interesting questions?
5. **Why** is it so important to generate questions?



Questions to SCI problems

- Do you have appropriate data to answer your question?
- Do you have information on confounding variables?
- Was the data you're working with collected in a way that introduces bias?

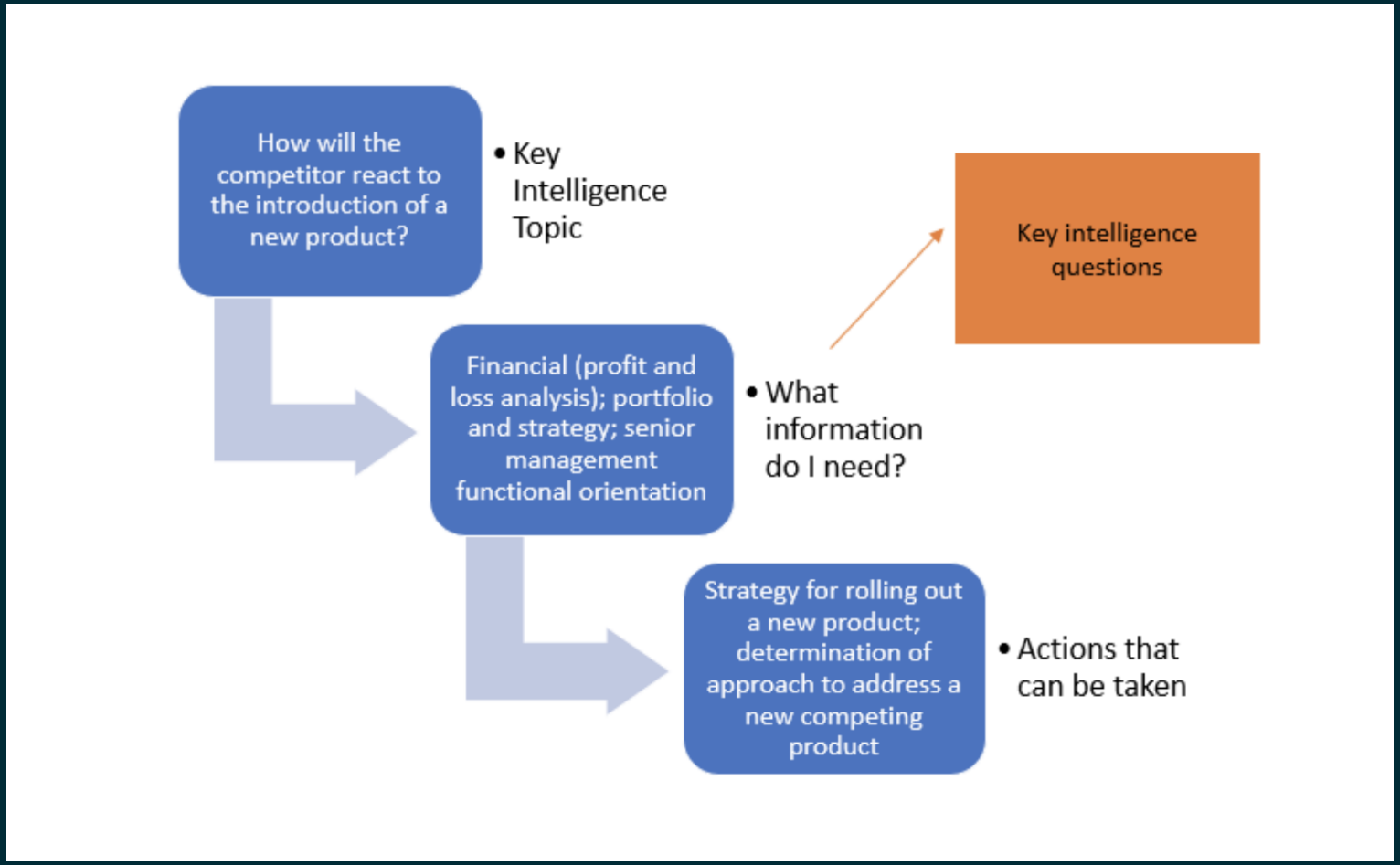
> Suppose I want to estimate the average number of children in households in Edinburgh. I conduct a survey at an elementary school in Edinburgh and ask students at this elementary school how many children, including themselves, live in their house. Then, I take the average of the responses. Is this a biased or an unbiased estimate of the number of children in households in Edinburgh? If biased, will the value be an overestimate or underestimate?

KIT & KIQ

Key Intelligence Topics (KITs) are the key questions that competitive intelligence professionals are trying to answer about what's happening outside of their organizations, the "what you need to know" to be successful.

Key Intelligence Questions (KIQs) are the specific questions you are trying to understand within a given topic.

KIT & KIQ


How will the competitor react to the introduction of a new product?
• Key Intelligence Topic
Key intelligence questions
Financial (profit and loss analysis); portfolio and strategy; senior management functional orientation
• What information do I need?
Strategy for rolling out a new product; determination of approach to address a new competing product
• Actions that can be taken

Key Intelligence Topics

1. *Strategic KITs*: contribute to key decisions in relation to strategic formulations and implementation for the company. Strategic KITs are used to make determine on actions such as: investment decisions, global expansion, technological competitiveness, global alliances and/or acquisitions. 2. *Early Warning KITs*: identify forthcoming threats and opportunities for the company in the market. Some examples of Early Warning KITs include: new entrant threats, government regulations, untapped market, and any shifts in bargaining power. 3. *Key Players KITs*: arise from market rivalries, new entrants, and substitute products. However, Key Player KITs can also deal with none threatening means such as new suppliers or contractors which have entered the market as well.

Key Intelligence Questions

In the example below, if the KIT is competitor reaction to a new product, your KIQs might be:

Money

1. What are they currently spending on R&D? Has that changed in the last 3 years?
2. Who are they hiring (and what salaries come with that)?
3. What's their current cash flow situation? Portfolio and

Market

1. Are they trying to reach new markets?
2. Where is their product in terms of market maturity and adoption?
3. What percentage of the market do they hold?

Key Intelligence Questions

Man

1. Are the senior leaders risk averse or risk-takers?
2. What actions and decisions have they made in the past?
3. What do their employee think about the company?

Machines

1. What technologies are they investing on?
2. Do they have the proper technologies for thier processes?
3. Do some recent changes in the technological environemnt have an impact on them?

Kit vs Research Questions

Let's brainstorm 🧠

KIT & CRITICAL THINKING

1. Critical thinking is the process of thinking critically about **your owns** thoughts.
2. One of the most important steps is challenging **your own** assumption.
3. In a company context this means challenging **company's** assumption
4. **Question asking** is fundamental for critical thinking and vice-versa

What is Critical Thinking? 🔍

WHAT IS THE POWER OF KIT?

1. **Define an Investigation**: Can inform both you and the company on the topic and the nature of the investigation (to discover, to explore, to explain, to describe to compare)

2. **Set Boundaries**: Every time you face a new interesting direction, you can decide if follow it or not considering your starting research question

3. **Provide Direction**: It will point you towards theory you need to explore, literature you need to review, the methods you may need to call on, data you need to gather, and person you need to interview.

4. **Identify Stop Criteria**: A univocal point of arrive will help you understand when you will not have to invest resources on a specifi c research direction.

HOW DO I ARTICULATE MY KIT?

CLARITY: All the potential stakeholders can get the gist of your research just by reading your question. Maybe you will have to rephrase your question in different ways, depending on who will read it.

SPECIFICITY: The KIT has to capture the nature of your research without being too broad. The goal is to avoid ambiguity and fuzziness, because being precise make the research task easier.

USEFULNESS: The KIT has to answer a questions that the company cares about. In other words, if you are able to solve the KIT, the company will create more value: more money 💰, less resources 🥬.

Exploratory data analysis (EDA)

EDAs: whats the danger?

Let's brainstorm 🧠

Checklist

- Formulate your question
- Read in your data
- Check the dimensions
- Look at the top and the bottom of your data
- Validate with at least one external data source
- Make a plot
- Try the easy solution first

Formulate your question

Approaches? 

Read in your data

- Place your data in a folder called `data`
- Read it into R with `read_csv()` or friends (`read_delim()`, `read_excel()`, etc.)

```
library(readxl)
fav_food <- read_excel("data/favourite-food.xlsx")
fav_food
```

```
## # A tibble: 5 × 6
##   `Student ID` `Full Name`    favourite.food mealPlan AGE   SES
##          <dbl> <chr>          <chr>          <chr>    <chr> <chr>
## 1            1 Sunil Huffmann Strawberry yo… Lunch o… 4     High
## 2            2 Barclay Lynn   French fries   Lunch o… 5     Midd…
## 3            3 Jayendra Lyne  N/A            Breakfa… 7     Low
## 4            4 Leon Rossini   Anchovies      Lunch o… 99999 Midd…
## 5            5 Chidiegwu Dun… Pizza          Breakfa… five  High
```

clean_names()

If the variable names are malformatted, use `janitor::clean_names()` or othe similar functions

```
library(janitor)
fav_food %>% clean_names()
```

```
## # A tibble: 5 × 6
##   student_id full_name       favourite_food meal_plan age   ses
##        <dbl> <chr>           <chr>          <chr>     <chr> <chr>
## 1          1 Sunil Huffmann  Strawberry yo… Lunch on… 4     High
## 2          2 Barclay Lynn    French fries   Lunch on… 5     Midd…
## 3          3 Jayendra Lyne   N/A            Breakfas… 7     Low
## 4          4 Leon Rossini    Anchovies      Lunch on… 99999 Midd…
## 5          5 Chidiegwu Dunk… Pizza          Breakfas… five  High
```

Case study: NYC Squirrels!

- The Squirrel Census is a multimedia science, design, and storytelling project focusing on the Eastern gray (*Sciurus carolinensis*). They count squirrels and present their findings to the public.
- This table contains squirrel data for each of the 3,023 sightings, including location coordinates, age, primary and secondary fur color, elevation, activities, communications, and interactions between squirrels and with humans.

```
#install_github("mine-cetinkaya-rundel/nycsquirrels18")
library(nycsquirrels18)
```

Locate the codebook

mine-cetinkaya-rundel.github.io/nycsquirrels18/reference/squirrels.html

Check the dimensions

```
dim(squirrels)
```

```
## [1] 3023   35
```

Look at the top...

```
squirrels %>% head()
```

```
## # A tibble: 6 × 35
##    long   lat unique_squirrel_id hectare shift date
##   <dbl> <dbl> <chr>              <chr>   <chr> <date>
## 1 -74.0  40.8 13A-PM-1014-04     13A     PM    2018-10-14
## 2 -74.0  40.8 15F-PM-1010-06     15F     PM    2018-10-10
## 3 -74.0  40.8 19C-PM-1018-02     19C     PM    2018-10-18
## 4 -74.0  40.8 21B-AM-1019-04     21B     AM    2018-10-19
## 5 -74.0  40.8 23A-AM-1018-02     23A     AM    2018-10-18
## 6 -74.0  40.8 38H-PM-1012-01     38H     PM    2018-10-12
## # … with 29 more variables: hectare_squirrel_number <dbl>,
## #   age <chr>, primary_fur_color <chr>,
## #   highlight_fur_color <chr>,
## #   combination_of_primary_and_highlight_color <chr>,
## #   color_notes <chr>, location <chr>,
## #   above_ground_sighter_measurement <chr>,
## #   specific_location <chr>, running <lgl>, chasing <lgl>, …
```

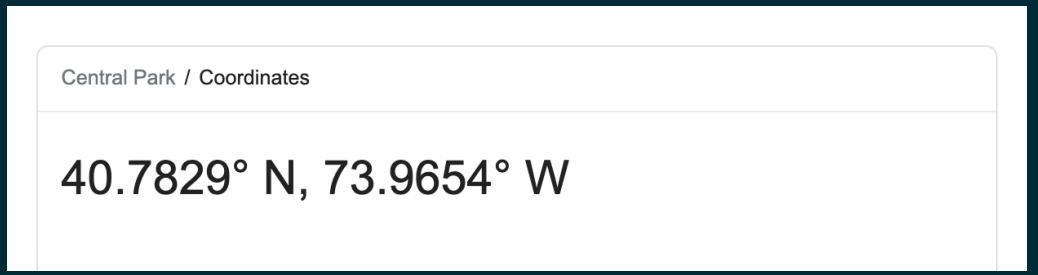
...and the bottom

```
squirrels %>% tail()
```

```
## # A tibble: 6 × 35
##    long   lat unique_squirrel_id hectare shift date
##   <dbl> <dbl> <chr>              <chr>   <chr> <date>
## 1 -74.0  40.8 6D-PM-1020-01      06D     PM    2018-10-20
## 2 -74.0  40.8 21H-PM-1018-01     21H     PM    2018-10-18
## 3 -74.0  40.8 31D-PM-1006-02     31D     PM    2018-10-06
## 4 -74.0  40.8 37B-AM-1018-04     37B     AM    2018-10-18
## 5 -74.0  40.8 21C-PM-1006-01     21C     PM    2018-10-06
## 6 -74.0  40.8 7G-PM-1018-04      07G     PM    2018-10-18
## # … with 29 more variables: hectare_squirrel_number <dbl>,
## #   age <chr>, primary_fur_color <chr>,
## #   highlight_fur_color <chr>,
## #   combination_of_primary_and_highlight_color <chr>,
## #   color_notes <chr>, location <chr>,
## #   above_ground_sighter_measurement <chr>,
## #   specific_location <chr>, running <lgl>, chasing <lgl>, …
```

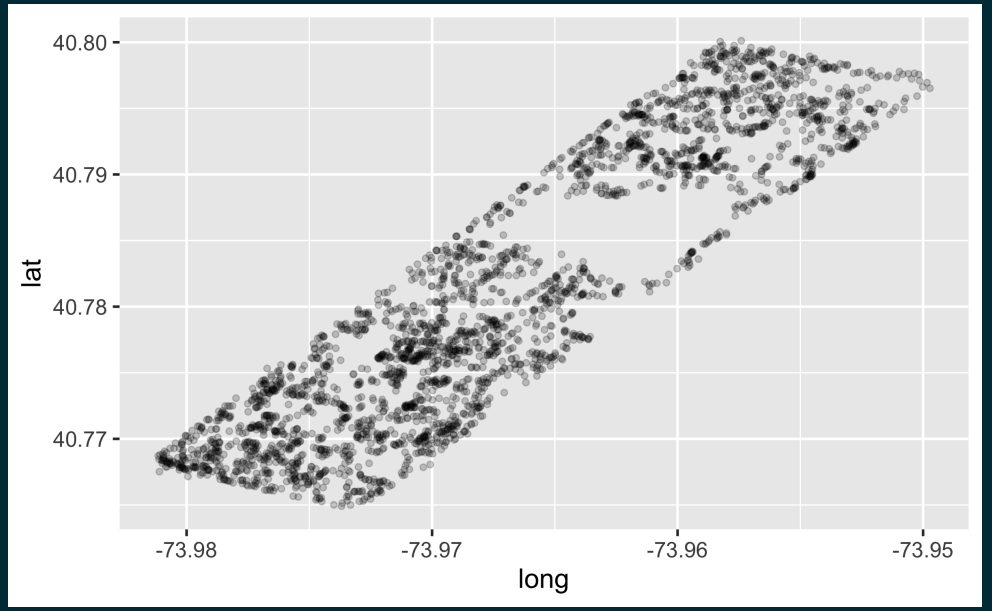
Validate with at least one external data source

```
## # A tibble: 3,023 × 2
##     long   lat
##    <dbl> <dbl>
##  1 -74.0  40.8
##  2 -74.0  40.8
##  3 -74.0  40.8
##  4 -74.0  40.8
##  5 -74.0  40.8
##  6 -74.0  40.8
##  7 -74.0  40.8
##  8 -74.0  40.8
##  9 -74.0  40.8
## 10 -74.0  40.8
## 11 -74.0  40.8
## 12 -74.0  40.8
## 13 -74.0  40.8
## 14 -74.0  40.8
## 15 -74.0  40.8
## # … with 3,008 more rows
```



Make a plot

```
ggplot(squirrels, aes(x = long, y = lat)) +
  geom_point(alpha = 0.2)
```



Hypothesis: There will be a higher density of sightings on the perimeter than inside the park.

Communicating for your audience

- Avoid: Jargon, uninterpreted results, lengthy output
- Pay attention to: Organization, presentation, flow
- Don't forget about: Code style, coding best practices, meaningful commits
- Be open to: Suggestions, feedback, taking (calculated) risks