Workflow Data Analytics Jobs and Careers in Data Science +4

What is the workflow or process of a data scientist? What tools do they use?

This question previously had details. They are now in a comment.

Answer Follow · 86 Request 1

15 Answers

Ryan Fox Squire, Neuroscientist Turned Data Scientist
Updated Jan 7, 2017 · Upvoted by Gilbert Duy Doan, M.A. Mathematics & Data Science, San Jose State University (2021) and Mark Meloon, Data Scientist

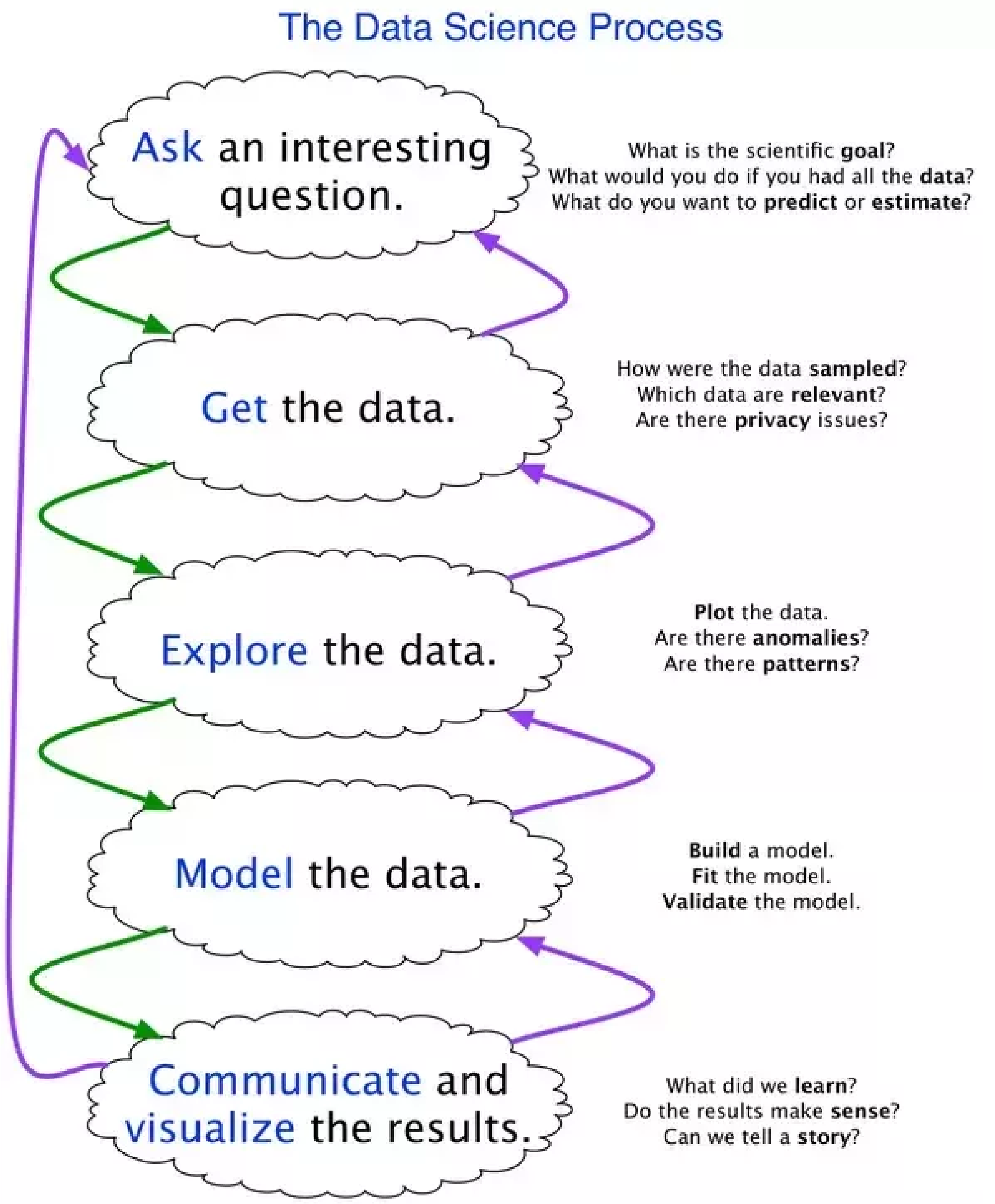
It all starts with asking an interesting question...

Upvote · 178 Share

Related Questions

- What tools do data scientists use?
- What are the largest inefficiencies in a data scientist’s workflow?
- What is a good Git workflow for a data scientist?
- Who are the top data scientists?
- What are some actual projects that data scientists have worked on? What tools and analytical techniques were used, and what mistakes were made...
- What tools do data scientists use for professional presentations (e.g. diagrams, graphs, flow charts, etc.)?
- What is your data science pipeline/workflow?
- How can I become a data scientist?

+ Ask New Question


The Data Science Process
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

More Related Questions

Question Stats

85 Public Followers

36,320 Views

Last Asked Jan 13, 2016

Edits

Image credit: Professor Joe Blitzstein and Professor Hanspeter Pfister presented this framework in their Harvard Class "Introduction to Data Science". Check out: Joe Blitzstein's answer to Data Science: What is it like to design a data science class?

Note all of the purple arrows going 'backwards'. The data science work flow is non-linear, iterative, and cyclical. It's impossible to know at the start which is the best way to proceed.

Each stage relies on different skills and tools*.

Here is one functional stack:

Stage 1: Ask A Question (that matters to your organization)

- Skills: science, domain expertise, curiosity, business and product knowledge
- Tools: your brain, talking to experts (in and outside your co.), experience

Stage 2: Get the Data

- Skills: data cleaning, querying databases, web scraping, CS stuff
- Tools: SQL, python, pandas, (spark)

Stage 3: Explore the Data

- Skills: Get to know data, develop hypotheses, patterns? anomalies?
- Tools: matplotlib, numpy, scipy, pandas, (spark, pyspark)

Stage 4: Model the data

- Skills: regression, machine learning, validation, big data
- Tools: scikits learn, pandas (spark, pyspark, MLlib)

Stage 5: Communicate the data

- Skills: presentation, speaking, visuals, writing
- Tools: matplotlib, adobe illustrator, powerpoint/keynote

Stage 6: Implementation

- Skills: product management, communication, IO psychology, CS, politics.
- Stage 6 is hugely important. Data Scientists who don't carry their work all the way through to realize it's full impact are ultimately just ineffective consultants.
- *You probably can't implement it on your own*. But that doesn't mean it isn't your job to communicate, collaborate, and convince until it is implemented. Champion your work—*no one else knows how*.
- Please see Eric Colson's comment to this answer.

Stage 7: Test and quantify your impact

- Skills: all of the above.
- Did Stage 6 work? Was this project worthwhile? This is not a subjective question. Who better to answer this question than you?

Conclusion:

Data science work flow is a non-linear, iterative process that involves Asking questions, Getting Data, Exploring Data, Modelling Data, and Communicating Data. The real rockstars will also usher their work through Implementation and then Test and quantify its impact on their organization.
I like this framework because it emphasizes:

1. the importance of asking questions to guide your work-flow, and
2. the importance of iterating on your questions and strategy, as you become more familiar with your data.

There are many skills and tools* required to cover the full data science process. Here I have suggested one full functional stack that will get you very far.

**There are many tools, and lots of great quora discussion about them. For example:*

- *What are the most useful data mining, analysis, or science tools? Hilary Mason, chief scientist at Bitly, mentioned they use a lot of Python, as well as Hadoop HDFS, Redis, and D3.js. Does anyone use Gephi, Panda, the Julia Language, Tableau Public?*
- *What are most widely used software/tools for machine learning/big data?*
- *Also Check out What is the data science topic FAQ? Section: "Data Science Tools"*

Edit 10/18/2016:

Since originally posting, my perspective on data science has expanded with some experience, and I've finally decided to expand this answer to reflect that. In particular, the non-technical aspects of Stage 6 and Stage 7 exercise a whole different set of muscles. And the abilities to ask interesting questions *that* *matter to your organization* can distinguish the great from the good. No one else at the company knows how to take advantage of the data science you are doing—not really. It’s up to you to show them the way. GL,HF

29.7k Views · View Upvoters

Eric Colson, Led Data Science teams from Stitch Fix to Netflix to Yahoo.
Good description of the research phase. But many data scientists would add a next p...

Pronojit Saha, Data Aficionado.
Answered Nov 2, 2014 · Upvoted by Jalem Raj Rohit, Sr. Data Scientist at Episource

For process flow, I am reproducing a post from my blog here.

Data Science is the **practice** of:

1. Asking questions (formulating hypothesis), answers to which solve known problems or unearth unknown solutions that in turn drive business value,
2. Defining the data needed or working with an existing data set and employing tools (computer science based) to collect, store and explore such data generally in huge volume & variety (probably more than 1 TB and 1000s of dimensions) ,

3. Identifying the type of analysis to be done to get to the answers and performing such analysis by implementing various algorithms/tools (statistics based) in a distributed and parallel architecture,
4. Communicating the insights gathered from the analysis in the form of simple stories/visualizations/dashboards (the Data Product) that a non-data scientist can understand and build conversation out of it. (It should be kept in mind that a product can also be an piece of code that is internal to a company and is used by various departments. The presentation, maintenance, scalability, etc of the code are then the product features, which is often not practiced in many organizations)
5. Building a higher level abstraction that does steps 2-4 in an autonomous way, predicting & taking actions on new data as they are fed to the system.

For tools that help you accomplish this you can head to my other blog post here. Hope this helps.

4.4k Views · View Upvoters

Your feedback is private.

Is this answer still relevant and up to date? Yes No

Upvote · 3 Share

Add a comment... Recommended All

Nirmal Patel, Data Science, Big Data Analytics & FinTech Advisor at Imarticus (2018-present)

Answered Mar 9

The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something actionable.

You'll often get ambiguous inputs from the people who have problems. You'll have to develop the intuition to turn scarce inputs into actionable outputs–and to ask the questions that nobody else is asking.

Say you're solving a problem for the VP Sales of your company. You should start by understanding their goals and the underlying why behind their data questions. Before you can start thinking of solutions, you'll want to work with them to clearly define the problem.

A great way to do this is to ask the right questions.

You should then figure out what the sales process looks like, and who the customers are. You need as much context as possible for your numbers to become insights.

You should ask questions like the following:

- Who are the customers?

- Why are they buying our product?
- How do we predict if a customer is going to buy our product?
- What is different from segments who are performing well and those that are performing below expectations?
- How much money will we lose if we don't actively sell the product to these groups?

In response to your questions, the VP Sales might reveal that they want to understand why certain segments of customers have bought less than expected. Their end goal might be to determine whether to continue to invest in these segments, or de-prioritize them. You'll want to tailor your analysis to that problem, and unearth insights that can support either conclusion.

It's important that at the end of this stage, you have all of the information and context you need to solve this problem.

Collect the raw data needed for your problem

Once you've defined the problem, you'll need data to give you the insights needed to turn the problem around with a solution. This part of the process involves thinking through what data you'll need and finding ways to get that data, whether it's querying internal databases, or purchasing external datasets.

You might find out that your company stores all of their sales data in a CRM or a customer relationship management software platform.You can export the CRM data in a CSV file for further analysis.

Process the data for analysis

Now that you have all of the raw data, you'll need to process it before you can do any analysis. Oftentimes, data can be quite messy, especially if it hasn't been well-maintained. You'll see errors that will corrupt your analysis: values set to null though they really are zero, duplicate values, and missing values. It's up to you to go through and check your data to make sure you'll get accurate insights.

You'll want to check for the following common errors:

1. Missing values, perhaps customers without an initial contact date
2. Corrupted values, such as invalid entries
3. Timezone differences, perhaps your database doesn't take into account the different timezones of your users
4. Date range errors, perhaps you'll have dates that makes no sense, such as data registered from before sales started

You'll need to look through aggregates of your file rows and columns and sample some test values to see if your values make sense. If you detect something that doesn't make sense, you'll need to remove that data or replace it with a default value. You'll need to use your intuition here: if a customer doesn't have an initial contact date, does it make sense to say that there was NO initial contact date? Or do you have to hunt down the VP Sales and ask if anybody has data on the customer's missing initial contact dates?

Once you're done working with those questions and cleaning your data, you'll be ready for exploratory data analysis (EDA).

Explore the data

When your data is clean, you'll should start playing with it!

The difficulty here isn’t coming up with ideas to test, it’s coming up with ideas that are likely to turn into insights. You...(more)

Upvote · 2 Share

Add a comment... Recommended All

Roger Huang, Always Learning
Answered Mar 31, 2016

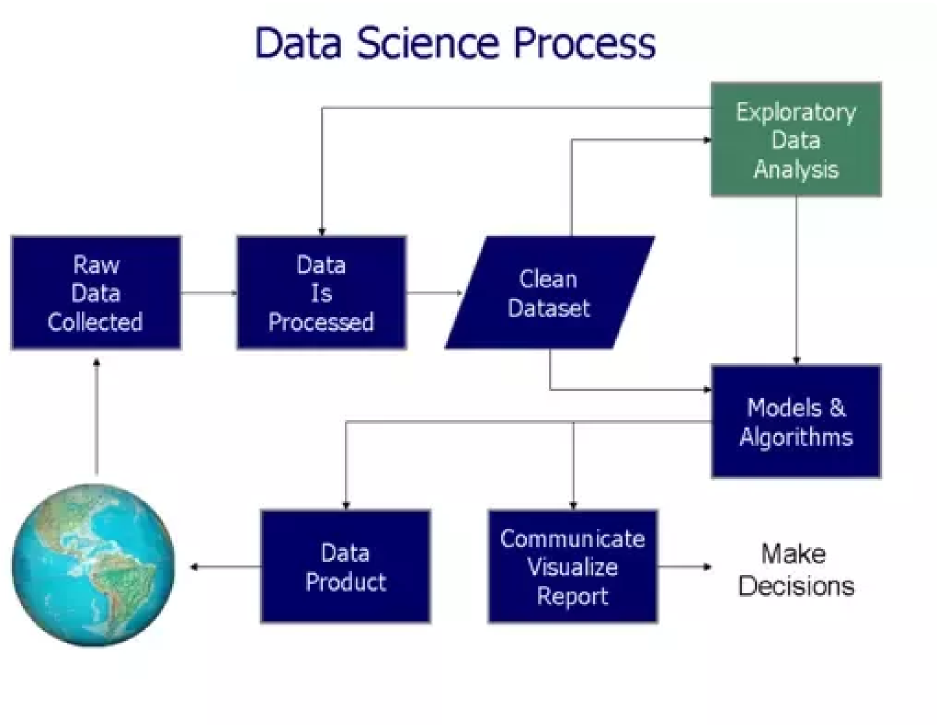
I actually wrote an article on this at KDNuggets based on my research on data science careers .

A brief summary of the workflow for me would be as follows:

1. Framing the problem. At this point, you're asking questions and trying to get a handle on what data you need.
2. Collect the raw data you need. You'll probably access a web API with a programming language, or use SQL to collect data from a structured database at this point. Or you'll look through company tools and extract our CSV files. Be prepared to attack this problem with many different tools and sources. Collecting data can become messy, especially if the data you want isn't something people have been collecting in an organized fashion!
3. Process the data for analysis (data wrangling). Here's where you'll clean your data, put everything together into one workspace, and make sure your data has no faults in it. Typically you'll use the tool you're most comfortable with. At this point, I drop into Pandas/Python to fill in null

values/check for invalid date ranges and etc., but it's equally valid if you do that in Excel.

4. Explore the data. Here's where you'll start getting summary-level insights of what you're looking at, and extracting the large trends. In Pandas, .describe() will quickly give you the mean, count, standard deviation and you might already see things worth diving deeper into.
5. Perform in-depth analysis. Here's where you're gonna apply those big old algorithms and transform your data to torture insights out of it.
6. Communicate the data. It's no good sitting on insights. You have to move people into action with them.



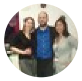
Hope that helps!

3.5k Views · View Upvoters

Upvote · 6 Share

Add a comment...

Recommended All

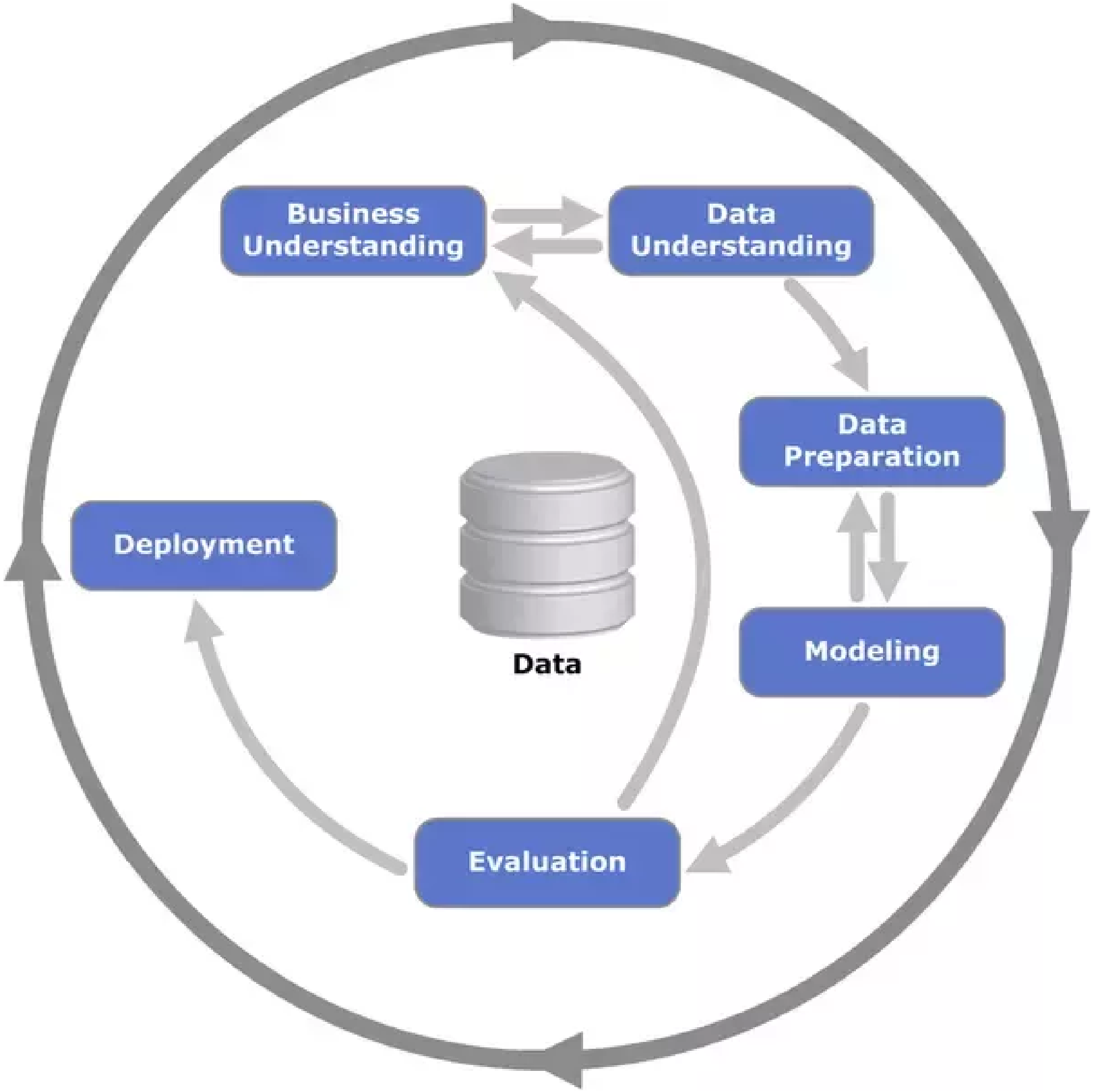


Grant Case, Pre-Sales for IBM Analytics. MBA in Finance, Marketing, and Decision Support Modeling. Prior to role, 10 Ye...

Answered May 10, 2015

If you are thinking about a a workflow within the field I would invite you to look at the CRISP-DM. Cross Industry Standard Process for Data Mining

CRISP-DM encompasses the entire data mining lifecycle including evaluation and deployment.



The usage of CRISP-DM has been remarkably stable over the last seven years even with the amazing changes happening within our industry as this KD-Nuggets survey from 2014 demonstrates.

CRISP-DM, still the top methodology for analytics, data mining, or data science projects

CRISP-DM has ~43% of the mind share while those "rolling their own" has gained a tremendous amount of ground in the last seven years. These changes I believe have more to do with the desire to augment CRISP-DM and others and also the "newness" of this particular vocation.

Overall, having a methodology when starting a project is better than none and with CRISP-DM you have an open body standard to work.

*Author's Note: I work for IBM who purchased SPSS, one of the founding members of the CRISP-DM. Other founders include DaimlerChrysler and NCR.

5.5k Views · View Upvoters

Upvote · 9 Share

Add a comment... Recommended All

Abhishek Soni, Bigdata Developer

Answered Jul 14, 2014

I'm little new in bigdata so don't know much about the right flow... But according to me it should be as such...

For data gathering, you need to first analyze the data rate and estimated growth in size. This will help you better in sorting out the best candidate for the your database.

You should consider the possible types of data that can come across in your application. Whether it's partially sparse or completely sparse.

Then 2nd thing... Whether you are aiming for prediction in specific area. And

what is the tolerance limit for response latency.
These two things will do the most of decision making for selecting proper database tool.

Now the core thing comes into picture, the algorithm. The easiest approach to devise prediction algorithm is to formulate a pattern. You can have multiple patterns analysed simultaneously to predict next input. (That's what I have done in my current project to predict some part of human behavior)

Once you have the patterns in your pocket, you can think of more analytics over it.

For algorithms you can use hadoop ecosystem. Mahout and R language has good support for it. Or you can use your custom algorithm implemented over any NOSql tool of your choice.

2.8k Views

Upvote Share

Add a comment... Recommended All

Srinath Perera, An Architect behind WSO2 Machine Learner
Answered Jul 15, 2014

If you just want to calculate basic stats, Excel is good if your data is very small (<100), R is good until 100k data points or so. Anything bigger should use MapReduce or Spark. I suggest first checking out Hive.

If you need to do predictive analytics, you need to learn about machine learning.

Following is a good ML talk to get started.

Then if you want to know more, follow the Andre Ng's course from Coursera .

R has support for machine learning. Mapreduce based machine learning tools are in a project called Mahout.

3.7k Views · View Upvoters

Upvote · 4 Share

Add a comment...

Recommended All

Top Stories from Your Feed