Becoming a Data Scientist: Advice From My Podcast Guests

(and my twitter followers, and me!)

Renée Marie Parilak Teate

Presented by Renee Teate at PyData DC 10/9/2016

Renée Teate

Data Scientist

@becomingdatasci
on Twitter

Data Scientist at HelioCampus
Better Data. Brilliant Insights.
"Asking and answering the most pressing questions in higher education for your institution."

Host of the Becoming A Data Scientist Podcast & Learning Club
Interviewing data scientists to find out how they got to where they are today since Dec. 2015
500+ member learning club (~90 of whom have posted a thread in the forums)

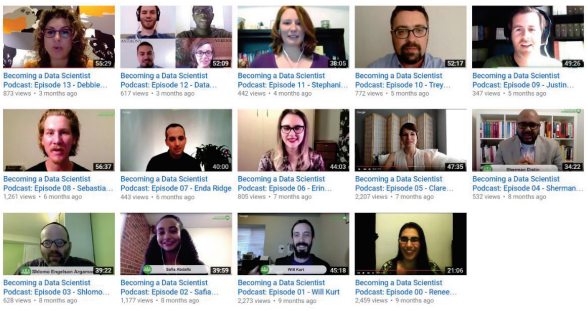
Creator and Developer of DataSciGuide
Online Data Science Learning Directory (work in progress)

Database Designer & Developer, Web Developer, SQL Data Analyst
2004-2016

B.S. Integrated Science and Technology
from James Madison University, 2004

M.Eng. Systems Engineering
from University of Virginia, 2015

Becoming a Data Scientist Podcast Guests



Not going to be able to quote them all in presentation, but elements from all interviews in advice here. I aim to interview people entering data science from a variety of backgrounds - biotech, physics, psychology, marketing, students and teachers, variety of ages, experiences.

"How do I become a data scientist?"

"It depends..."

Becoming a Data Scientist

01	Where to Start	02	Who to Listen to
03	What to Learn	04	How to Learn
05	How to Practice	06	When You're Ready

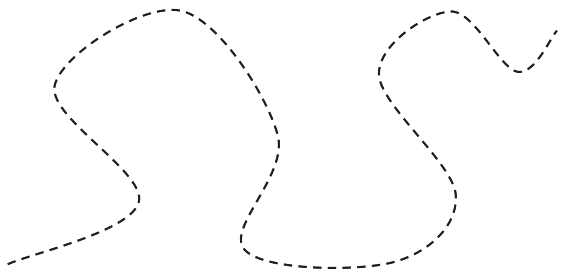
1. Where to Start

“Where should I start?”

“Where do I even start?!?!”

More often stated as...

“It depends...”



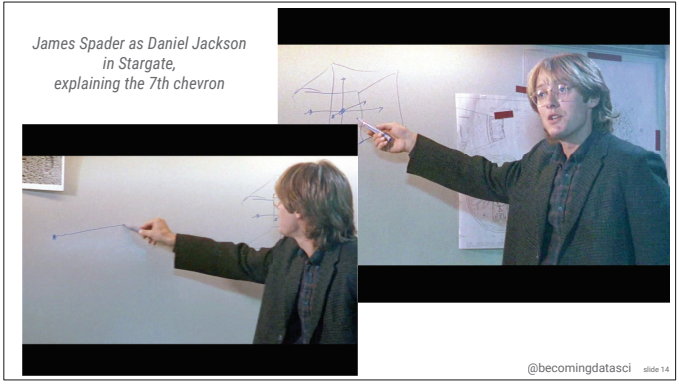
First you have to figure out:

1. *Where You Are*

and

2. *Where You Want To Go*

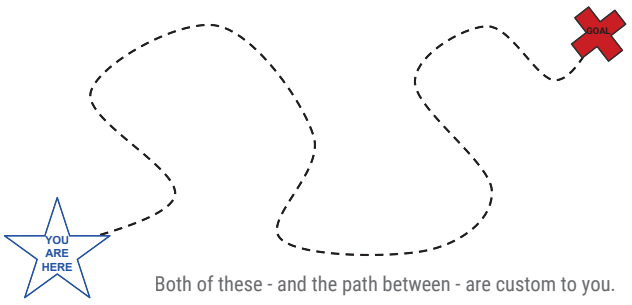
James Spader as Daniel Jackson in Stargate, explaining the 7th chevron



http://stargate.mgm.com/view/content/1504/index.html

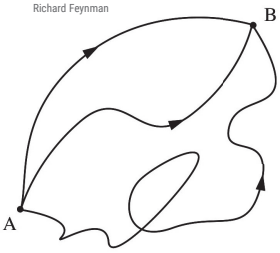
http://www.stargate-fusion.com/actualites/676/zoom-sur-le-film-stargate-la-porte-des-etoiles.html

Assess your starting point & have an ultimate goal end point to move towards.



Both of these - and the path between - are custom to you.

Richard Feynman



$$W(q, t; q_0, t_0) = \int_{q_0}^{q} \mathcal{D}q \exp\left[\int_{t_0}^{t} -\frac{1}{i\hbar} L(\dot{q}, q, \tau) d\tau\right]$$
$$= \lim_{N\to\infty} \int dq_1 \cdots dq_{N-1} \left(\frac{m}{2\pi i\hbar\varepsilon}\right)^{N/2}$$
$$\times \exp\left\{-\frac{\varepsilon}{i\hbar}\sum_{k=1}^{N}\right.$$
$$\left.\times\left[\frac{m}{2}\left(\frac{q_k - q_{k-1}}{\varepsilon}\right)^2 - U(q_{k-1})\right]\right\}, \quad (8)$$

My husband is a physicist and thought I should remind you here that there are infinite paths between two points (Feynman Path Integral)

NOTE:
There will be no more gratuitous complicated equations in the remainder of this presentation.

Path-integral Drawn by Matt McIrvin
https://commons.wikimedia.org/wiki/File:Three_paths_from_A_to_B.png
---
http://iopscience.iop.org/article/10.1086/306624/fulltext/38510.text.html
A Path Integral Approach to the Theory of Heliospheric Cosmic-Ray Modulation MING ZHANG

Feynman pic Richard Feynman at Los Alamos National Laboratory during the Manhattan Project. Original source from
http://www.lanl.gov/worldview/welcome/history/12_oppie-arrives.html

Where are you starting from?

How much math have you taken in school or learned yourself?

Have you ever programmed?

Have you done basic data analysis on a professional level before?

Data Scientist Profile Categories

From ***Doing Data Science*** by Cathy O'Neil & Rachel Schutt

- Mathematics
- Statistics
- Computer Science
- Machine Learning
- Data Visualization
- Communication
- Domain Expertise

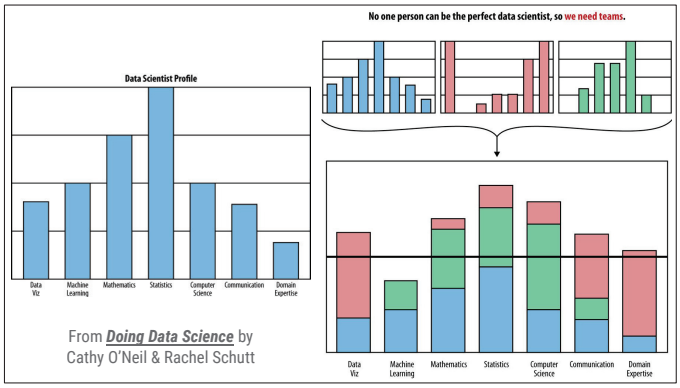
How much math have you taken in school or learned yourself? How good were you when you left off? How long has it been since you used your math skills?

Have you ever programmed? How advanced are you as a developer, and what languages do you have experience with?

Have you ever worked with databases? Understanding data relationships? Managing/cleaning data?

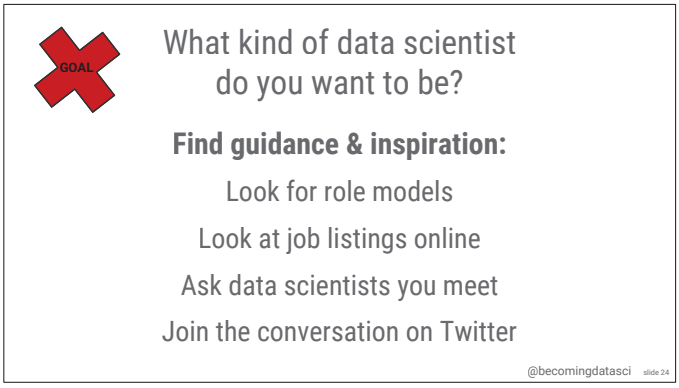
Have you done basic data analysis on a professional level? Created data visualizations? Communicated analysis results?

Assess yourself in each of these, in the context of what kind of data scientist you want to be



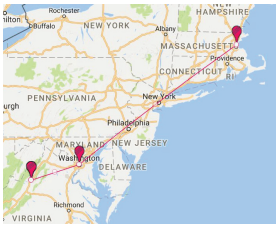
Don't worry about being great at everything. 1) No one can do that, 2) Many companies now hiring data science teams. Make sure you have baseline competency in each, then specialize.

Not an exhaustive list

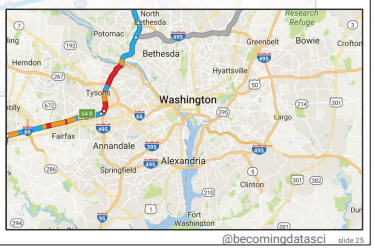


1. Check out my "Becoming a Data Scientist" podcast & other podcasts. Also books like "Data Scientists at Work" by Sebastian Gutierrez
2. Find your "dream job" and see what the expectations are...but be aware that many job descriptions are problematic!
3. Ask specific questions about what they do and how they got there, not just "How should I learn data science?"
4. See: http://www.becomingadatascientist.com/2015/10/04/how-to-use-twitter-to-learn-data-science-or-anything/

Keep your path general, with some waypoints



Don't overplan your route - keep your "GPS" on with traffic alerts, be ready to reroute



@becomingdatasci slide 25

There's so much to learn, how you narrow your focus? You can crowdsource your knowledge, and you definitely need a strong foundation in certain topics but…

"Part of it is being OK with not understanding everything… You're going to be doing yourself a disservice if you set out to learn all of the different topics that people say you should learn. You'll ask 10 people and get 12 answers."

-Trey Causey, @treycausey
Data Scientist & Product Strategy at ChefSteps
Episode 10

@becomingdatasci slide 26

https://www.daftlogic.com/projects-google-maps-distance-calculator.htm
If you want to get to Boston from Harrisonburg, you don't have to go through DC at rush hour. Be ready to pivot as you learn/grow/evaluate gaps.
(from Harrisonburg to Boston, might originally want to visit monuments in DC until you get up there during rush hour and rethink. You want to make it to Boston, path can change.)

"It's going to be your job for the rest of your life, so if you're going to get really good at something, make sure it's something you enjoy, because it requires a lot of persistence and curiosity to be happy and successful… don't force yourself into something… follow what you enjoy!"

-Stephanie Rivera, @dataginjaninja
Principal Data Scientist at myStrength
Episode 11

@becomingdatasci slide 27

This is your chance to become a candidate for your "dream job" and decide what YOU want to learn!

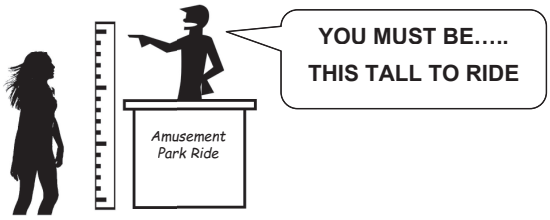
MAKE IT FUN!
You don't have to follow someone else's program or make it feel like school.

@becomingdatasci slide 28

https://mentally-physically-spiritually-strong.com/home/

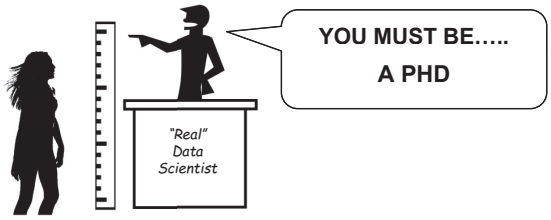
Once you have an idea about a possible path, who should you turn to for guidance?

2.
Who to Listen To
(and Who Not to Listen to)



Who not to listen to:

The Gatekeeper

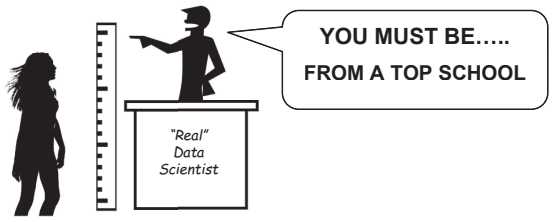


Who not to listen to:

The Gatekeeper

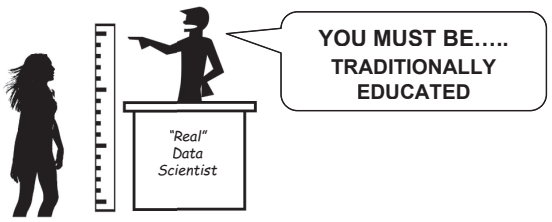
Who not to listen to:

The Gatekeeper



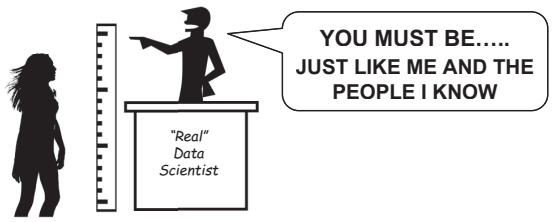
Who not to listen to:

The Gatekeeper



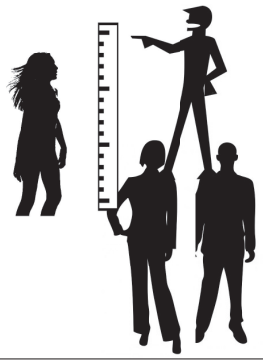
Who not to listen to:

The Gatekeeper



Who not to listen to:

The Gatekeeper

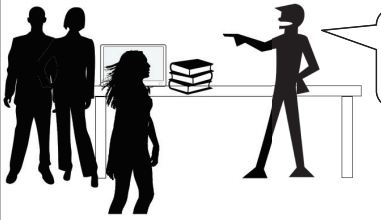


~~*The Gatekeeper*~~

Honest Data Scientist:

"If I have seen further it is by standing on the shoulders of giants."

-Isaac Newton



Who to listen to:

Honest and Helpful People Who Have Been There

Who Not to Listen To

- The Gatekeepers
  - "You have to be/have…."
- The Naysayers
  - "You aren't cut out to be…."
  - "In the past you…."
  - "You don't look like a…."
- The Scoffers
  - "It's SO easy…."
- Your Own Doubts
  - "If I'm already struggling…."

Who Not to Listen To:

- The Gatekeepers
  - "You have to be/have…."
- The Naysayers
  - "You aren't cut out to be…."
  - "In the past you…."
  - "You don't look like a…."
- The Scoffers
  - "It's SO easy…."
- Your Own Doubts
  - "If I'm already struggling…."

Who to Listen To:

- Honest and helpful people who have been there
  - "I got stuck at that point, too. These resources helped…."
- Your Supporters
  - "You're so awesome at…."
  - "I can't wait to see you do…."
- Experts who know how to teach beginners
  - "Let's walk through how this works…."
- Your Confident Self
  - "I can do this, I just need…."

Isis Anchalee:
https://medium.com/the-coffeelicious/you-may-have-seen-my-face-on-bart-8b9561003e0f#.u75hs81kt
Nick Burns, Your Company's Computer Guy (SNL/NBC):
http://www.nbc.com/saturday-night-live/video/nick-burns-your-companys-computer-guy/n11268

For guidance, find….

People who are the type of data scientist you want to become

People who are experts, but know how to communicate with beginners (i.e. good teachers)

Other beginners, and people who are just ahead of you on the data science learning path

Mentors and friends who share your passion and are in the domain in which you want to work

Data Science is not “easy”.
(it’s not even very well-defined)

Data Science is not “too hard for you”.
(though some parts might be right now)

It’s up to you not to get derailed.

We’ll talk more about *How to Learn* in Part 5.

3.
What to Learn

“So what do I need to learn?”

“It depends...”

The common refrain in this talk

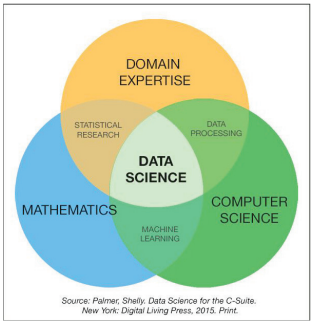
Example Data Science Careers & Skills

	Mathematics & Statistics	Computer Science & Machine Learning	Data Visualization & Communication	“Business” Domain Knowledge
Data Analyst/Scientist	✓	✓	✓	✓
Big Data Engineer	✓	✓	✓	✓
Machine Learning Dev.	✓	✓	✓	✓
Marketing Analyst	✓	✓	✓	✓
Biostatistician	✓	✓	✓	✓

Point here is that depth of expertise per topic varies by role, but all data scientists need to know things from these 4 categories

First, learn the basics

Basic Data Science Foundation



Shelly Palmer
http://www.huffingtonpost.com/shelly-palmer/are-you-ready-for-data-sc_b_6844032.html

Mathematics/Statistics	Computer Science	Domain Knowledge
Minimum	*Minimum*	*Minimum*
• Undergraduate-Level Descriptive Statistics • Data Visualization Techniques	• Coding for Data Manipulation & Summarization in 1 language (like python) • Basic "Packaged" Machine Learning Techniques	• Ability to do simple "business intelligence" type analysis in domain • Basic communication - understanding the terminology & general concepts
Next	*Next*	*Next*
• Linear Algebra • More Advanced Statistics • Calculus	• Understanding databases & SQL • Understanding how Machine Learning Algorithms work in detail "behind the scenes" • Additional languages (R, Julia, etc.) or tools/packages • Additional data processing frameworks (Spark, Hadoop, etc.)	• In-depth understanding of domain • Ability to communicate with people in various roles within domain (stakeholders)

My opinion on minimum knowledge needed in these categories for any type of data scientist.
Scikit-learn
http://scikit-learn.org/stable/modules/tree.html

Specialties ("Electives")

- Machine Learning and AI Specialties
  - Deep Learning
  - Autonomous Systems (robots)
  - Computer Vision
  - Natural Language Processing
- Advanced (or prettier) Data Visualization for publication
- Big Data Engineering
- Geographic Data & Mapping Techniques (GIS)
- Specific Domains
  - Research Science
  - Education
  - Politics
  - Marketing/Retail
  - etc etc...

Like a college curriculum… all lower-level courses required, but upper level specifics pick & choose according to your focus

On whether people should go through the entire Open Source Data Science Masters program she developed….

"Absolutely not! You should cherry-pick and be very strict about what you include, because you could spend your whole life on one of the topics that's under the umbrella [of data science]."

As a generalist, you need to figure out which of those topics is relevant to your path.

-Clare Corthell, @clarecorthell
Founder at Luminant Data
Episode 05

Find a niche

(experts are usually experts in one thing)

"It's a dangerous ideology, the notion of a Super Genius."

-Safia Abdalla, @captainsafia, Data Scientist, Maintainer of @nteractio, and PyData Chicago organizer, Episode 2

"Get experience. A lot of the data science skills I've developed are out of interest… If I have a problem that's in front of me, I'll figure out what I need to learn to solve that problem"

-Justin Kiggins, @neuromusic, Neuroscientist/Data Scientist, Episode 9

Total beginner?

Start by creating a report:
A basic data analysis that answers a "business" question.

Descriptive statistics, basic visualization, exploratory data analysis, basic programming to import/modify/summarize data
(could be Excel w/formulas & charts, BI program like Tableau/Cognos, Pandas package in Python, Seaborn/Pyplot, etc)

Renee's expectation for any data analyst or data scientist:

Business Question
⇩
Data Question
⇩
Data Answer
⇩
Business Answer

Don't just focus on learning technical tools & techniques.

The non-technical skills are just as important.

this is key - ability to translate business question into data question, know how to use the data/perform analysis to generate a data answer, and translate back into a business answer

What do you look for when you hire a data scientists and analysts?

"The #1 thing that I look for is curiosity… the value added is by saying 'I found this and I'm curious: how did that happen?'" And taking the analysis to another level.

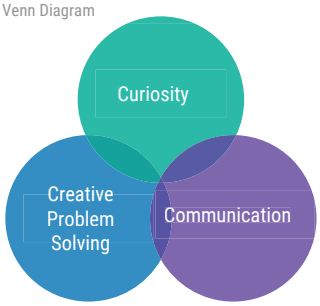
-Sherman Distin, @ShermanDistin
Digital Marketing Executive & Consumer Analytics Consultant at Querybridge
Episode 4

"Soft Skills" That Came Up Often in Podcast Interviews...

- Curiosity
- Ability to communicate with a variety of people
- Ability to "hear" the actual problem that needs to be solved
- Desire to continue to learn
- Creativity
- Tenacity
  - 
    "A critical part of the [data scientist] mindset" - Enda Ridge
    Head of Data Science, FMCG UK, Author of Guerrilla Analytics, @enda_ridge, Ep. 7
- Adaptability
  - 
    "Overcoming unexpected difficulties" -Shlomo Argamon
    Director, Master of Data Science program at IIT, @ShlomoArgamon, Ep. 3
- Recognition of wider issues that come into play
  - Systems Thinking

Good Data Science Practice

Another Data Science Venn Diagram

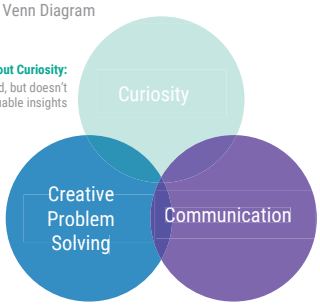


Designed w/input from @StephdeSilva

Good Data Science Practice

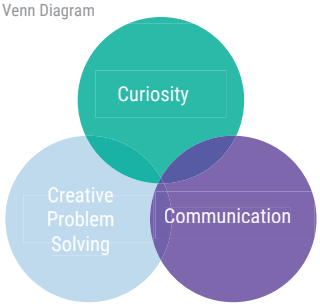
Another Data Science Venn Diagram

Without Curiosity:
Answers questions asked, but doesn't go beyond to add valuable insights



Good Data Science Practice

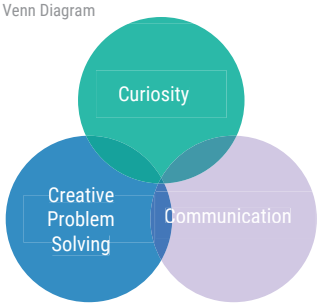
Another Data Science Venn Diagram



Without Creative Problem Solving:
Great research and explanation of problem, offers no actionable solutions

Good Data Science Practice

Another Data Science Venn Diagram

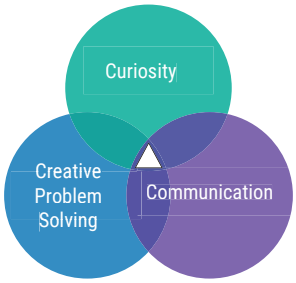


Without Good Communication:
Unintelligible presentation of otherwise good results

At Intersection:

Good Data Science Practice

These 3 things were mentioned consistently by podcast guests as either traits they themselves possess, or as desirable factors that set candidates apart when they're hiring data scientists.



4.
How to Learn

So, once you have a plan for what to learn, how do you go about learning it?

"It depends..."

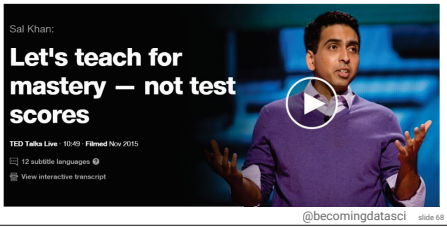
Choose resources that work with your learning style, and customize accordingly.

Overcome your
Math & Programming Aversions

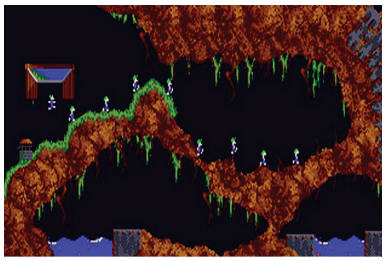
You don't have to love 'em, you just have to not hate 'em.

Chances are, you developed an aversion because of how they are taught in school.

You can try a different way now.



https://www.ted.com/talks/sal_khan_let_s_teach_for_mastery_not_test_scores
Student in class can get 80% and considered passing grade and class moves on to next topic, but that student could be missing 20% of foundational info important for next topic. Gets lost, feels studpid, starts hating it.
Now you can go at your own pace and customize the instruction for yourself, keep at it until you feel solid.



If you have the patience, persistence, and problem-solving skills to play a level-based puzzle game, you can learn to program that game.

5.
How to Practice
and Apply What You Have Learned

http://arcadesushi.com/lemmings-game-anniversary/
https://www.angrybirds.com/games/

The most common advice is to start with a project related to something that matters to you.

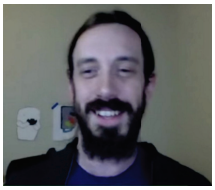
Find a real-world dataset. Ask some questions that could be answered by analyzing it.

Each step of the project will mean learning a new topic, or practicing a new skill or technique.

Blog about your experiences to practice communicating (or record yourself talking about it). Get feedback

"My answer is always read lots of books, and study every night. And they ask, 'How else?!?' No one likes that answer!"

-Will Kurt, @willkurt, Data Scientist at Quick Sprout
Episode 1



"How I learn best is really by struggling… If you are lifting weights, and you only lift the small weights, you will never become any stronger."

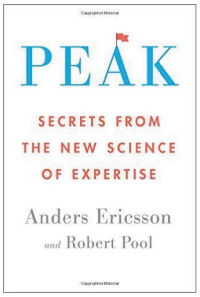
-Sebastian Raschka, @rasbt, Computational Biologist, Data Scientist, Author of Python Machine Learning
Episode 8

Some links to datasets (more resources will be posted to DataSciGuide over time):
http://www.datasciguide.com/data-sources-apis-for-data-science-projects/

Other things from Will: "You're always going to feel dumb", "If you're 'grinding', you're doing it wrong"
Addition to Sebastian's analogy: If you lift a weight that's way to heavy, you'll hurt yourself, and it won't be enjoyable.

Deliberate Practice Concepts

73



Peak: Secrets from the New Science of Expertise

- "Inherent talent" doesn't determine who becomes an expert
- Requires consistent not-very-fun practice with a focus on pushing just beyond where you are
- You need feedback and guidance - get a good teacher or data scientists to learn along with

@becomingdatasci slide 73

This research is the source of "10,000 hour rule" (but slightly different than reported in Freakonomics)

New learners projects: Verena cats GPS collar, David commute time, Kerry video game ratings, Anthony mapped tweets in CA
See Episode 12:
http://www.becomingadatascientist.com/2016/06/15/becoming-a-data-scientist-podcast-episode-12-data-science-learning-club-members/

-pushing just beyond where you are (Sebastian learn by struggling)
-have specific goal for each course/project
-take homework-type clean problems, reproduce with real world messy dataset (domain knowledge, data manipulation)
-take machine learning package, dive in to understand how it would be built from scratch (stats & CS) and even build one, compare results

(Anthony mentions benefit of learning club is getting feedback from others, Enda peer review) and help teach others (Verena)

-check about doing it for professional development at work or for credit at school instead of as side-project

6.
How to Know When You're Ready

(for your next career step)

Whoah…. I know Data Science!

@becomingdatasci slide 76

It's not like The Matrix
https://en.wikipedia.org/wiki/The_Matrix

"I think that people overestimate what you need to know to get started in this field."

-Erin Shellman, @erinshellman, Senior Data Scientist at Zymergen, Episode 6

I (Renee) felt ready when I had the foundational skills, was comfortable with the terminology, had accomplished a few difficult tasks, and realized I could learn anything else on the job.

And there's nothing wrong with doing a few "stretch" job interviews for practice. And who knows - you might get hired anyway!

In Summary

http://static.wixstatic.com/media/293d7a_9e3949afe50c47dbb04a4f9ebc2dc53b.jpeg
When you're most of the way there, you're very valuable
Apply for jobs - even if you don't get it, you'll learn what they're looking for that you don't have (don't get discouraged! I know, it's hard to basically be told "you don't have what it takes", but they're not rejecting you from the field, or even from that job, they're saying they found someone that was a better match for them. the right job for you might be right around the corner)
Keep in mind 60% of job description stat (men vs women)

You've probably learned today that

"It depends..."

means you have to customize advice.

Hopefully I've given you some guidance on how to do that, and therefore how to get started on the path to your data science dream job!

1. Figure out where you're starting from and what kind of data scientist you want to be, and map out a basic learning path
2. Learn the "required" baseline topics: descriptive statistics, coding for data manipulation and simple machine learning, solid analytical understanding in domain of your choosing
3. Get good at a specific niche beyond the basics
4. Deliberately practice both technical skills and "soft skills" with real-world data (don't forget the "get feedback" part)
5. You'll never be done learning, so don't wait too long to try applying for data-science-related jobs. Go for it!

Most importantly: **Have fun!**

What is your advice to data science learners?

"To not let anyone take your dreams away, to know that they're there for a reason, and you have them because they're part of your essence. And if you choose to not listen to that voice that's telling you, one day you're going to pay for it (and be unhappy)... pursue your dreams no matter how hard you think they are in the moment. And find a mentor."

-Debbie Berebichez, @debbiebere
Physicist, Chief Data Scientist at Metis
Episode 13
(pictured with Bill Nye the Science Guy)

Questions?

Renée Teate
@becomingdatasci on Twitter
Email: renee@becomingadatascientist.com
BecomingADataScientist.com (blog, learning club, podcast)
DataSciGuide.com (learning resource directory under development)

http://www.sciencewithdebbie.com/photo-galleries/

See the slide notes for more info + reference links. Look for them at becomingadatascientist.com

See the slide notes for more info + reference links. Look for them at becomingadatascientist.com

Presented by Renee Teate at PyData DC 10/9/2016

How to find
"Becoming a Data Scientist" Podcast

- Podcast & Show Notes on my blog
- RSS feed for podcast players
- YouTube playlist (subscribe to channel)
- iTunes
- Stitcher

Learning Resources

General Resources for Data Science Learning

- Data Science Learning Club
- DataSciGuide
- Open Source Data Science Masters (Clare)
- Guerrilla Analytics (Enda)
- Harvard CS109 Data Science
- Metis - Explore Data Science
- Metis - Intro to Data Science
- Lynda
- DataCamp
- Data School
- DataQuest.io
- Podcasts
  - Becoming a Data Scientist
  - Partially Derivative
  - O'Reilly Data Show
  - What's the Point
  - (these & more on DataSciGuide)

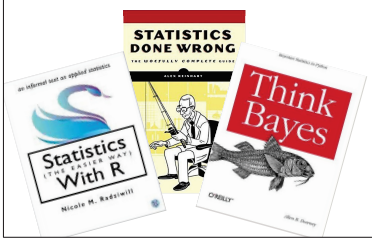
@becomingdatasci slide 86

Note: discount for DataCamp in Data Science Learning Club (when logged in as member)

More on http://www.datasciguide.com/

Stats Learning Resources for Beginners

- Khan Academy
- Mathematical Monk (Safia rec)



Stats Learning Resources for Data Science

- Count Bayesie Blog (Will's blog)
- Coursera
- More on DataSciGuide

@becomingdatasci slide 87

More on http://www.datasciguide.com/
(including free versions of some of the books)

Data Visualization Resources for Beginners

- Flowing Data blog & books
- Storytelling with Data
- The Truthful Art
- Stephen Few books
- Data Stories podcast
- More on DataSciGuide

Machine Learning Resources for Beginners

- Data Smart (Excel)
- Interactive Coding Courses
  (see DataCamp, etc. on Data Science slide)
- Coursera
- SciKit-Learn Docs
- Podcasts
  - Data Skeptic
  - Linear Digression
  - Talking Machines
  - These and more on DataSciGuide

@becomingdatasci slide 88

Note: if you have access to SafariBooks through work or school, many of these are available there

Python Learning Resources for Beginners

- Talk Python to Me Podcast
- Learn Python the Hard Way
  (Safia recommendation)
- Anaconda / Jupyter
- Udacity
- Codecademy
- More on DataSciGuide

Python Learning Resources for Data Science

- DataCamp
- Scikit-Learn
- Python Machine Learning book & Jupyter Notebooks
  (Sebastian)
- Python for Data Analysis
  (Pandas)
- Fluent Python
  (not intended for beginners)

Keep trying different resources until you find the ones that work best for your current level and learning style.

This is only a sampling of what's available!

I'll continue to add more content (that you can rate!) to the searchable Data Science Learning Directory at

DataSciGuide.com