Dark Matter to Data Science

Brandon Bozek
Senior Advisor, Data Scientist
Dell Technologies

Overview

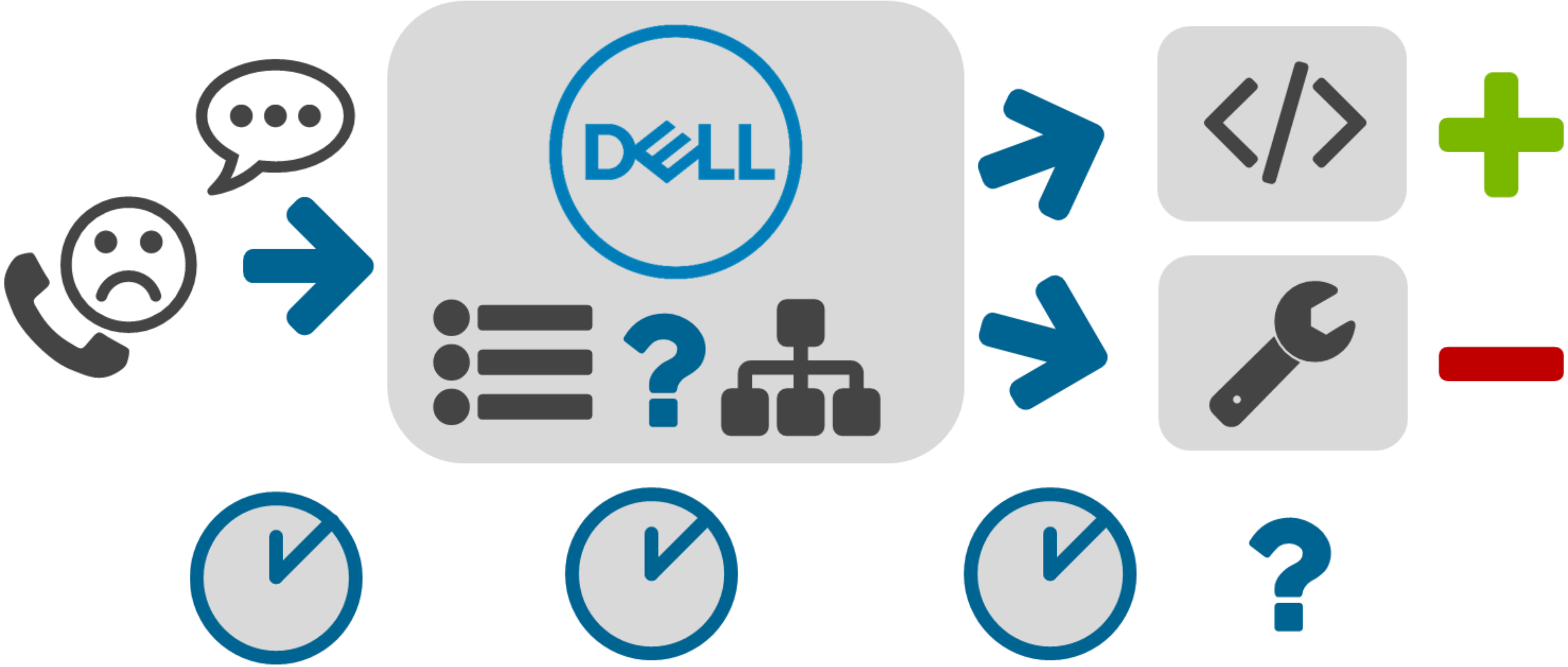
Snapshot of Data Science at Dell

My journey from Academia to Data Science

Preparing to become a Data Scientist

Preparing to Interview for a Data Scientist Role

Data Science at Dell



Improve customer experience, reduce cost, and simplify business procedures

Timeline from Academia to Industry

2009
PhD from UC Davis

2009-2014
First Postdoc at JHU

2014-2018
Second Postdoc at UT (via UMaryland)

Data Scientist at Dell

- Long road from starting grad school to here
- Original goal: Become a professor
- Backup plan: ???????? -> Data Science
- Outcome: Delighted with new career path

Explore alternatives to academic jobs anyway!

- Long journey without a clear outcome: Healthy to consider alternatives no matter where you are at to know what is out there
- Unique networking opportunity: UT has a number of Astro alumni in Austin who would be happy to meet up for coffee to talk DS + slack channel + startup crawls/meetups
- Time is limited but consider building a larger tool set: take a course to learn new tools that benefit both careers: SQL, Python (numpy, pandas, seaborn, etc), scikit-learn
- Most successful exit to industry: above already done over grad/postdoc career (see other UT alums), but a quicker exit can be made….

Tools/Skills to have prior to Day 1 (or earlier)

Python/R: I suggest python (numpy, pandas, seaborn, etc)

Scikit-learn: python machine learning package. You will need to have some demonstrated usage of this in a non-scientific context. Ideally posted on Github with analysis written up on a blog*

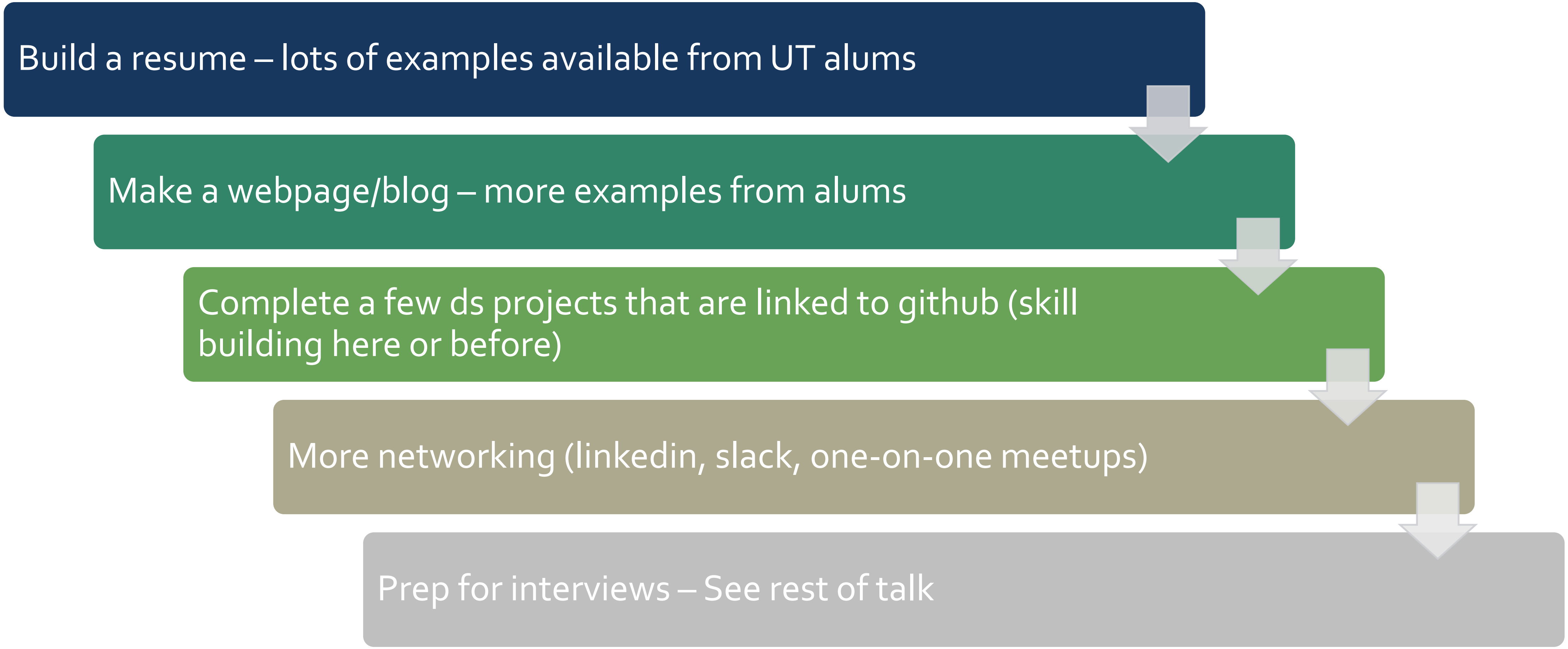
SQL: not always required, but nearly every interview has a pandas or SQL join, SQL is an industry wide tool, and this is relatively easy to learn

Git: some experience using it (ideally includes blog projects)

A/B Testing: experimental design and statistical analysis

More ML stuff: NLP, Deep Learning, Reinforcement learning, data balancing techniques, Hadoop/Spark, etc.

Process to Get a Job



Build a resume – lots of examples available from UT alums

Make a webpage/blog – more examples from alums

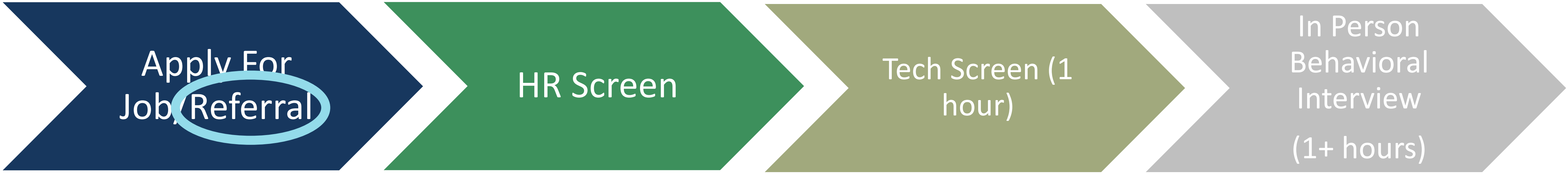
Complete a few ds projects that are linked to github (skill building here or before)

More networking (linkedin, slack, one-on-one meetups)

Prep for interviews – See rest of talk

Interview Processes*

Small Startup:



Apply For Job/Referral

HR Screen

Tech Screen (1 hour)

In Person Behavioral Interview

(1+ hours)

Tech Screen:

- Discussion around research to demonstrate that you can communicate complex topics clearly to a non-technical audience
- Hypothetical Scenario:
  - Machine learning theory: You are tasked with predicting the housing price for a region. How would you make that prediction?
  - Business hypothetical: Our company is trying to solve this problem. How would you solve that problem

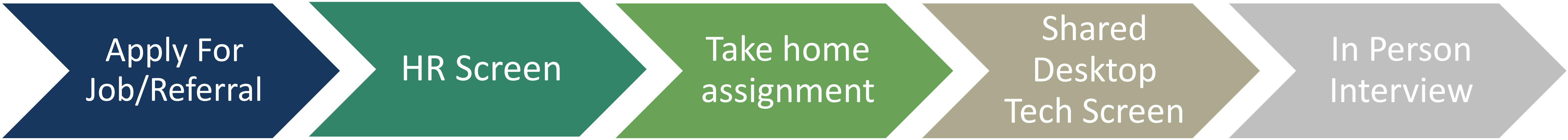
In Person Interview:

- Why leave academia for data science? What motivates you? Biggest strengths/weaknesses?
- STAR Behavioral Interview: Situation = context of a past situation, Task = your responsibility in situation, Activity = actions you took took, Result = outcome of your actions.
  - E.g. Please give us an example of: how you handled multiple projects with overlapping deadlines? A project where something went wrong? Confronted by a difficult person?
- You ask questions (critical!)

*sampled from my job search, names redacted in most cases, processes vary

Interview Processes*

Mid-level Tech Company:



Apply For Job/Referral

HR Screen

Take home assignment

Shared Desktop Tech Screen

In Person Interview

Take Home Assignment:

- Structured, clean data set. Multiple questions with clear objectives. Objective is to determine coding ability, business acumen, and statistical analysis ability

Tech Screen:

- Shared screen test: 1) Probability question (balls in urns), 2) SQL (join two simple tables), 3) Python sort, 4) Machine learning hypothetical

In Person Interview:

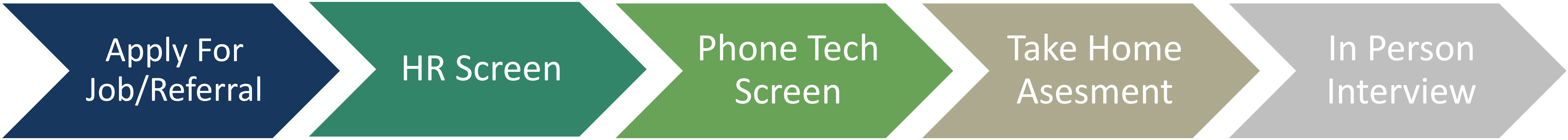
- 5 hour interview: 1) Machine Learning theory with whiteboard component (1 hr), 2) Business prompt with presentation and questions (1.5 hr), 3) Coding exercise (1 hr), 4) Statistics and A/B testing (1 hr)

Assessment Criteria: Each stage is a hurdle. Prior-steps not used to evaluate hire.

*sampled from my job search, names redacted in most cases, processes vary

Interview Processes*

Large Tech Company:



Tech Screen:

- Done on phone. Explain you technical background, answer questions that pertain specifically to your resume (explain your blog project), possibly a machine learning hypothetical (particularly if there is no blog/github)

Take Home Assignment:

- Structured, dirty data set. Open ended – here is some data and show us what you can do with it. Objective is to determine coding ability, business acumen, and statistical analysis/machine learning ability

In Person Interview:

- Panel Interview: Includes a manager and some data scientists. Resume/Behavioral questions that vary with the group that is hiring. Technical questions that range from follow-up questions on take-home assignment to hypothetical question.

Assessment Criteria: All stages taken together including fit with team

*sampled from my job search, names redacted in most cases, processes vary

Technical Background Questions that you will likely encounter*

Stats/Probability:

- Balls in an urn: 11 balls are in an urn. 6 are red, 5 are blue. Three are randomly drawn, what is the probability that 1 is red and 2 are blue?
- Biased coin (also A/B test): A coin is flipped 10 times. The coin lands heads 8 times. Is the coin biased in favor of heads?

Resources:

- Your favorite undergrad Introduction to Probability text book
- Biased coin solution within A/B context: https://www.csus.edu/indiv/j/jgehrman/courses/stat50/hypthesistests/9hyptest.htm

*Not Everywhere. Depends on interview process. I have feelings about these questions.

Technical Background Questions that you will likely encounter*

Machine Learning Theory:

- Example: Predict the price of a house for sale
  - What features will you initially consider? What exploratory data analysis will you perform? How will you encode categorical features? What model will you use to make your prediction? What evaluation metrics will you use to evaluate your model?
  - Need to be able to handle follow-up questions and small tweaks to the problem statement, i.e. if a feature is removed from the model, will your evaluation metric
- Be able to interpret training curves, validation curves, know categorical feature encoding methods, methods to evaluate feature importance, different machine learning algorithms and their advantages/drawbacks, and more…

Resources:

- Andrew Ng, “Introduction to Machine Learning”, Coursera (starts March 18)
- “Introduction to Statistical Learning” by James, Witten, Hastie, Tibshirani (I can send) + datacamp, udemy, or other online practical programming site

Technical Background Questions that you will likely encounter*

Technical Coding Questions:

- Big O notation, sorting/searching methods, hash tables, etc.

Resources:

- "Cracking the Coding Interview", McDowell (huge, comprehensive, and expensive textbook)
- Interview Cake: Very expensive. https://www.interviewcake.com/
- Big O notation: https://rob-bell.net/2009/06/a-beginners-guide-to-big-o-notation/

A/B Testing:

- Example: Your company is considering changing the format of the landing page. Design an experiment and determine if your company should launch the new changes or keep the old format.

Resources:

- Free course I took on udacity that I liked: https://www.udacity.com/course/ab-testing--ud257

General Interview Prep Resource:

- Random assortment of questions organized by category (most with answers)
https://datascienceinterview.quora.com

THANKS!