ScaleUp Porto.

DSPT
DATA SCIENCE PORTUGAL

How To Interview Data Scientists

May 10th, 2018
ScaleUp Porto, Portugal

Miguel Almeida, Feedzai Research

feedzai research

Agenda

- Pinhole Questions
- Trained Monkey Questions
- Forced Error Questions

- Not discussed: General soft skills (even though they are crucial)

Pinhole Q’s

Pinhole Q's

- **What technical skills are we looking for in a DS hire?**

Pinhole Q's

- **What technical skills are we looking for in a DS hire?**
- Delivery: knows DS loop or will learn it quickly, statistics, ...

Pinhole Q's

- **What technical skills are we looking for in a DS hire?**
- Delivery: knows DS loop or will learn it quickly, statistics, ...
- (ML) Research: ability to understand and prototype algorithms, fundamentals, ...

Pinhole Q's

- **What technical skills are we looking for in a DS hire?**
- Delivery: knows DS loop or will learn it quickly, statistics, ...
- (ML) Research: ability to understand and prototype algorithms, fundamentals, ...

- None of these is tested by asking **How does algorithm X work?**

Pinhole Q's

- **What technical skills are we looking for in a DS hire?**
- Delivery: knows DS loop or will learn it quickly, statistics, ...
- (ML) Research: ability to understand and prototype algorithms, fundamentals, ...

- None of these is tested by asking **How does algorithm X work?**
- Delivery: DS loop is algorithm-agnostic

Pinhole Q's

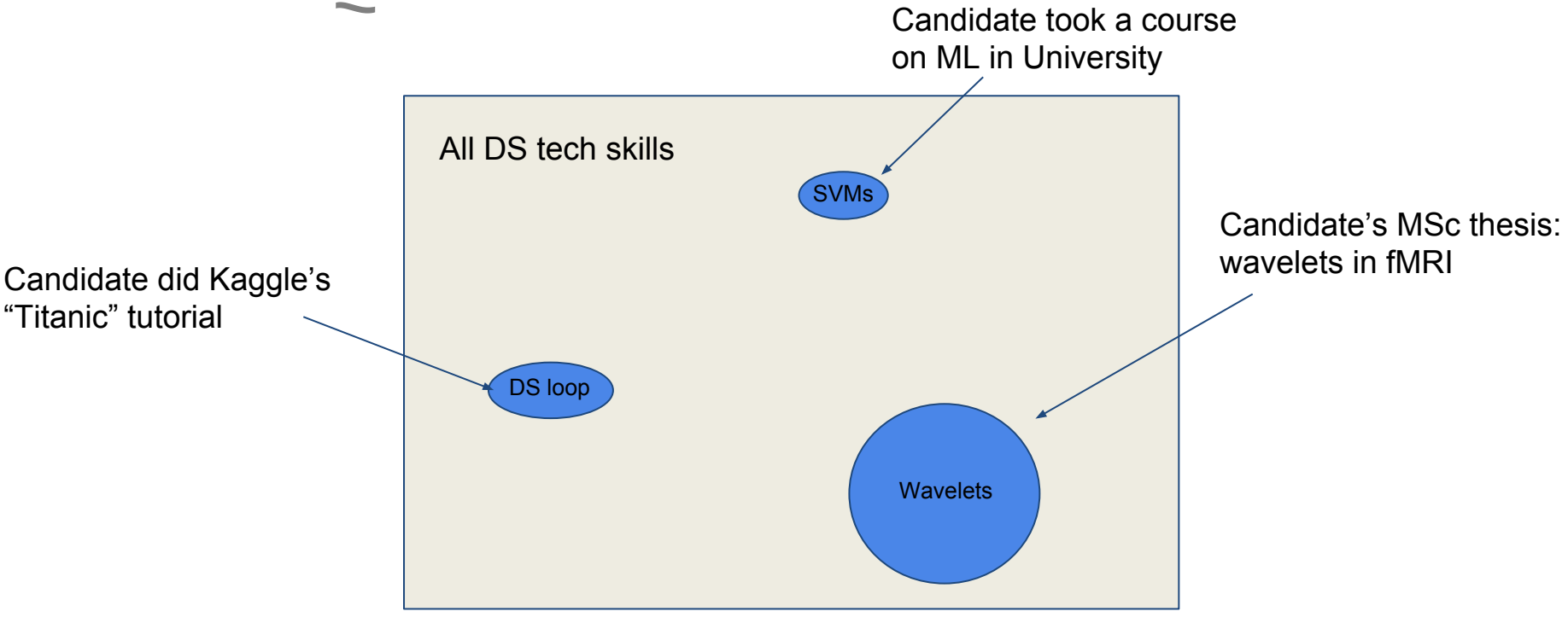
- **What technical skills are we looking for in a DS hire?**
- Delivery: knows DS loop or will learn it quickly, statistics, ...
- (ML) Research: ability to understand and prototype algorithms, fundamentals, ...

- None of these is tested by asking **How does algorithm X work?**
- Delivery: DS loop is algorithm-agnostic
- (ML) Research: has the candidate used X before? How deeply?

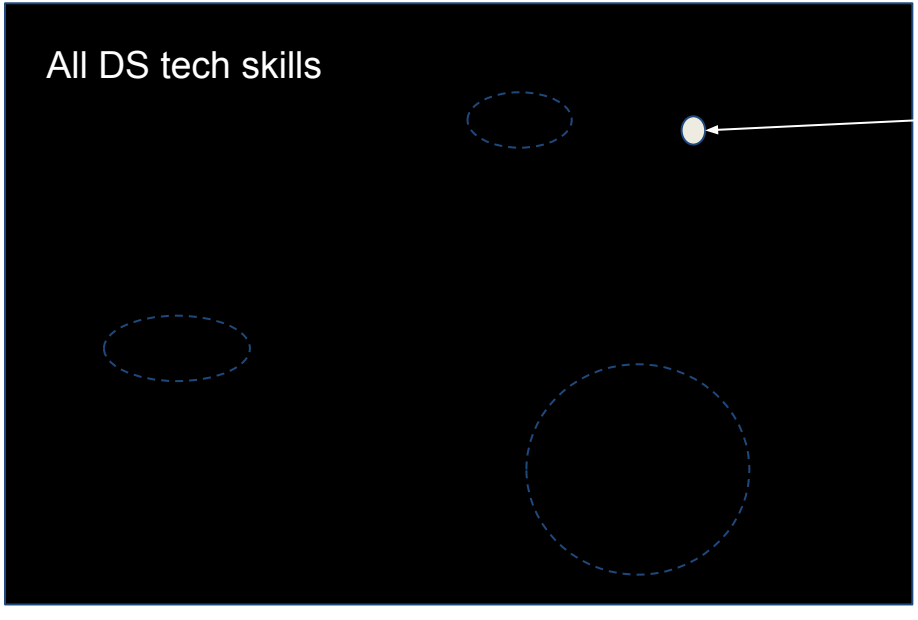
Pinhole Q's

All DS tech skills

Pinhole Q's


Candidate took a course on ML in University
All DS tech skills
SVMs
Candidate's MSc thesis: wavelets in fMRI
Candidate did Kaggle's "Titanic" tutorial
DS loop
Wavelets

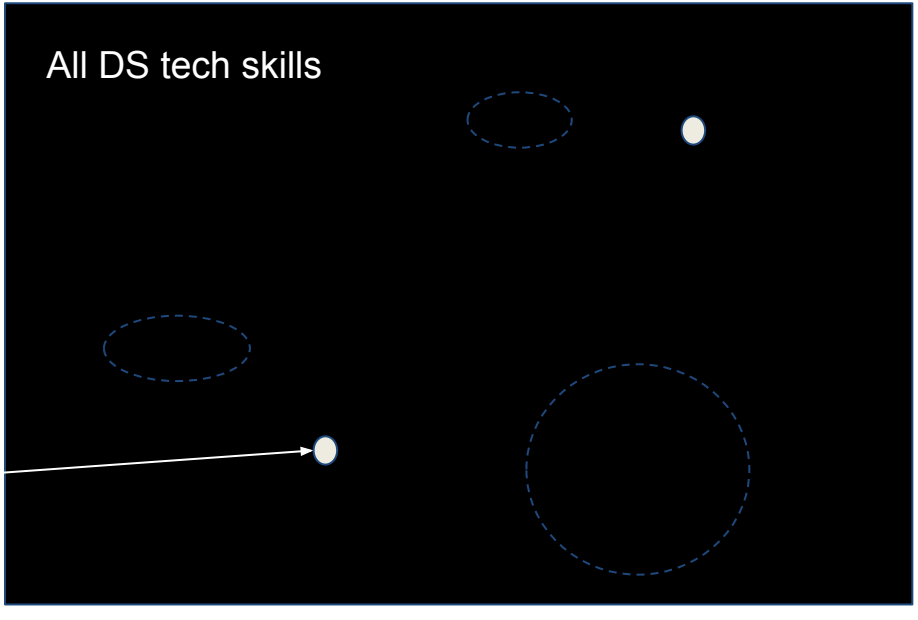
Pinhole Q's



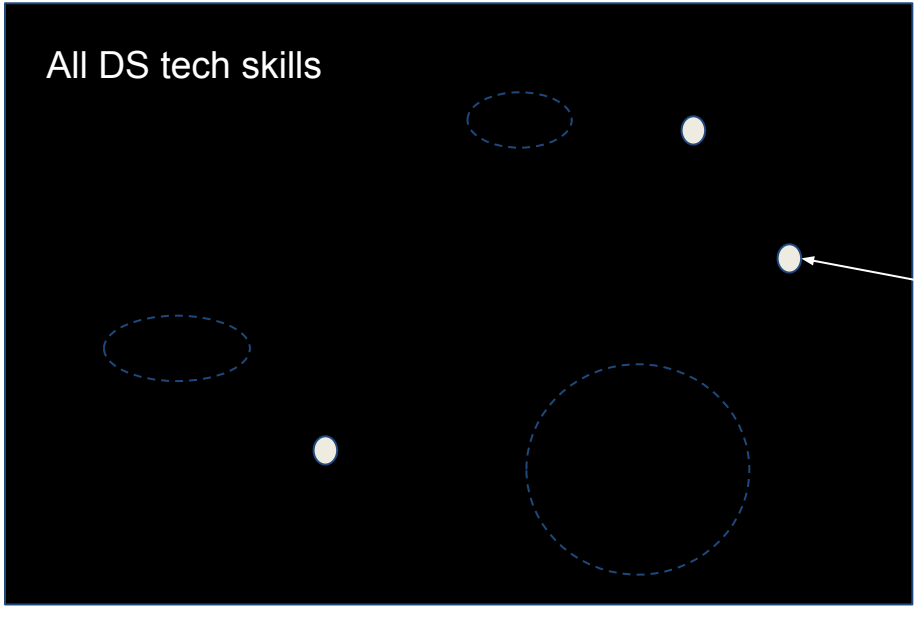
Can you tell me how Random Forests work?

Pinhole Q's

Can you tell me the definition of Precision and Recall?



Pinhole Q's



How do you do a join in SQL?

Pinhole Q's

- Even if the next question is **How do wavelets work?**...

Pinhole Q's

- Even if the next question is **How do wavelets work?**...
  - Candidate is probably super-nervous by now

Pinhole Q's

- Even if the next question is **How do wavelets work?**...
  - Candidate is probably super-nervous by now
  - Interviewers are in confirmation bias mode

Pinhole Q's

- Even if the next question is **How do wavelets work?**...
  - Candidate is probably super-nervous by now
  - Interviewers are in confirmation bias mode

- It's better to start with broad questions, candidate chooses the "pinhole"

Pinhole Q's

- Even if the next question is **How do wavelets work?**...
  - Candidate is probably super-nervous by now
  - Interviewers are in confirmation bias mode

- It's better to start with broad questions, candidate chooses the "pinhole"
- ... and then ask deeper questions about that topic

Pinhole Q's

- Even if the next question is **How do wavelets work?**...
  - Candidate is probably super-nervous by now
  - Interviewers are in confirmation bias mode

- It's better to start with broad questions, candidate chooses the "pinhole"
- ... and then ask deeper questions about that topic
  - **What did you do in your MSc/PhD thesis?**
    - **What was your success metric?**
      - **What can overfit to that measure?**

Pinhole Q's

- Even if the next question is **How do wavelets work?**...
  - Candidate is probably super-nervous by now
  - Interviewers are in confirmation bias mode

- It's better to start with broad questions, candidate chooses the "pinhole"
- ... and then ask deeper questions about that topic
  - **What did you do in your MSc/PhD thesis?**
    - **What was your success metric?**
      - **What can overfit to that measure?**
  - **Choose an ML algorithm that you feel confident in, and explain it to me.**
    - **...**

Trained Monkey Q's

Trained Monkey Q's

- Candidate did our Challenge and is presenting

Trained Monkey Q's

- Candidate did our Challenge and is presenting
- He computed a feature with constant value

Trained Monkey Q's

- Candidate did our Challenge and is presenting
- He computed a feature with constant value
- He used Logistic Regression and Random Forests in his challenge and used this feature

Trained Monkey Q's

- Candidate did our Challenge and is presenting
- He computed a feature with constant value
- He used Logistic Regression and Random Forests in his challenge and used this feature

- Interviewer: **Are you sure a constant feature is useful?**

Trained Monkey Q's

- Candidate did our Challenge and is presenting
- He computed a feature with constant value
- He used Logistic Regression and Random Forests in his challenge and used this feature

- Interviewer: **Are you sure a constant feature is useful?**
- Or: **Is it really a good idea to use constant features?**

Trained Monkey Q's

- A trained monkey is able to answer this question!

Trained Monkey Q's

- A trained monkey is able to answer this question!
- **Most common mistake with new interviewers**

Trained Monkey Q's

- A trained monkey is able to answer this question!
- **Most common mistake with new interviewers**
- If nothing comes to mind, ask "why did you do X?" or variants:
- **Why did you include that feature?**

Trained Monkey Q's

- A trained monkey is able to answer this question!
- **Most common mistake with new interviewers**
- If nothing comes to mind, ask "why did you do X?" or variants:
- **Why did you include that feature?**
- Or, if possible, ask a Forced Error Question (next)

Forced Error Q's

Forced Error Q's

- **Same scenario:**
  - Candidate did our Challenge and is presenting
  - He computed a feature with constant value
  - He used Logistic Regression and Random Forests in his challenge and used this feature

Forced Error Q's

- **Same scenario:**
  - Candidate did our Challenge and is presenting
  - He computed a feature with constant value
  - He used Logistic Regression and Random Forests in his challenge and used this feature

- Interviewer: **Which algorithm (LogReg or RFs) makes better use of features with a constant value?**

Forced Error Q's

- **Same scenario:**
  - Candidate did our Challenge and is presenting
  - He computed a feature with constant value
  - He used Logistic Regression and Random Forests in his challenge and used this feature

- Interviewer: **Which algorithm (LogReg or RFs) makes better use of features with a constant value?**

- This forces the candidate to think about his error

Forced Error Q's

- **Same scenario:**
  - Candidate did our Challenge and is presenting
  - He computed a feature with constant value
  - He used Logistic Regression and Random Forests in his challenge and used this feature

- Interviewer: **Which algorithm (LogReg or RFs) makes better use of features with a constant value?**

- This forces the candidate to think about his error
- … but the two obvious answers (LogReg or RFs) are both wrong!

Forced Error Q's

- Interviewer: **Which algorithm (LogReg or RFs) makes better use of features with a constant value?**

Forced Error Q's

- Interviewer: **Which algorithm (LogReg or RFs) makes better use of features with a constant value?**

- Bad answer: **Well, LogReg does a hyperplane split while RFs split on each coordinate and perform bagging. Let me think... well, the feature is numeric so both algorithms can use it, I'd try both and see which one works best, I like the data to tell me the answer.**

Forced Error Q's

- Interviewer: **Which algorithm (LogReg or RFs) makes better use of features with a constant value?**

- Bad answer: **Well, LogReg does a hyperplane split while RFs split on each coordinate and perform bagging. Let me think... well, the feature is numeric so both algorithms can use it, I'd try both and see which one works best, I like the data to tell me the answer.**
  (only says correct things but is missing a very obvious problem)

Forced Error Q's

- Interviewer: **Which algorithm (LogReg or RFs) makes better use of features with a constant value?**

- Bad answer: **Well, LogReg does a hyperplane split while RFs split on each coordinate and perform bagging. Let me think... well, the feature is numeric so both algorithms can use it, I'd try both and see which one works best, I like the data to tell me the answer.**
  (only says correct things but is missing a very obvious problem)

- Good answer: **Actually, it doesn't make sense for either algorithm.**

Forced Error Q's

- Interviewer: **Which algorithm (LogReg or RFs) makes better use of features with a constant value?**

- Bad answer: **Well, LogReg does a hyperplane split while RFs split on each coordinate and perform bagging. Let me think... well, the feature is numeric so both algorithms can use it, I'd try both and see which one works best, I like the data to tell me the answer.**
  (only says correct things but is missing a very obvious problem)

- Good answer: **Actually, it doesn't make sense for either algorithm.**
  (shows critical thinking, ability to admit mistakes, ability to handle pressure)

Forced Error Q's

- I used to do: **few red flags** > **many red flags**

Forced Error Q's

- I used to do: **few red flags** > **many red flags**
- Now: **forced errors, good answers** > **few errors** > **forced errors, bad answers**

Forced Error Q's

- I used to do: **few red flags** > **many red flags**
- Now: **forced errors, good answers** > **few errors** > **forced errors, bad answers**

- Critical thinking + admit mistakes are hard to gain if absent

Forced Error Q's

- I used to do: **few red flags** > **many red flags**
- Now: **forced errors, good answers** > **few errors** > **forced errors, bad answers**

- Critical thinking + admit mistakes are hard to gain if absent

- The DS loop can be learned by any intelligent person

Forced Error Q's

- I used to do: **few red flags > many red flags**
- Now: **forced errors, good answers > few errors > forced errors, bad answers**

- Critical thinking + admit mistakes are hard to gain if absent

- The DS loop can be learned by any intelligent person

- Once they learn:

Forced Error Q's

- I used to do: **few red flags > many red flags**
- Now: **forced errors, good answers > few errors > forced errors, bad answers**

- Critical thinking + admit mistakes are hard to gain if absent

- The DS loop can be learned by any intelligent person

- Once they learn:
  - critical thinking + admit mistakes + ... = autonomous Data Scientist

Forced Error Q's

- I used to do: **few red flags > many red flags**
- Now: **forced errors, good answers > few errors > forced errors, bad answers**

- Critical thinking + admit mistakes are hard to gain if absent

- The DS loop can be learned by any intelligent person

- Once they learn:
  - critical thinking + admit mistakes + ... = autonomous Data Scientist
  - lack of such skills = will need validation, micromanagement

Thank you!