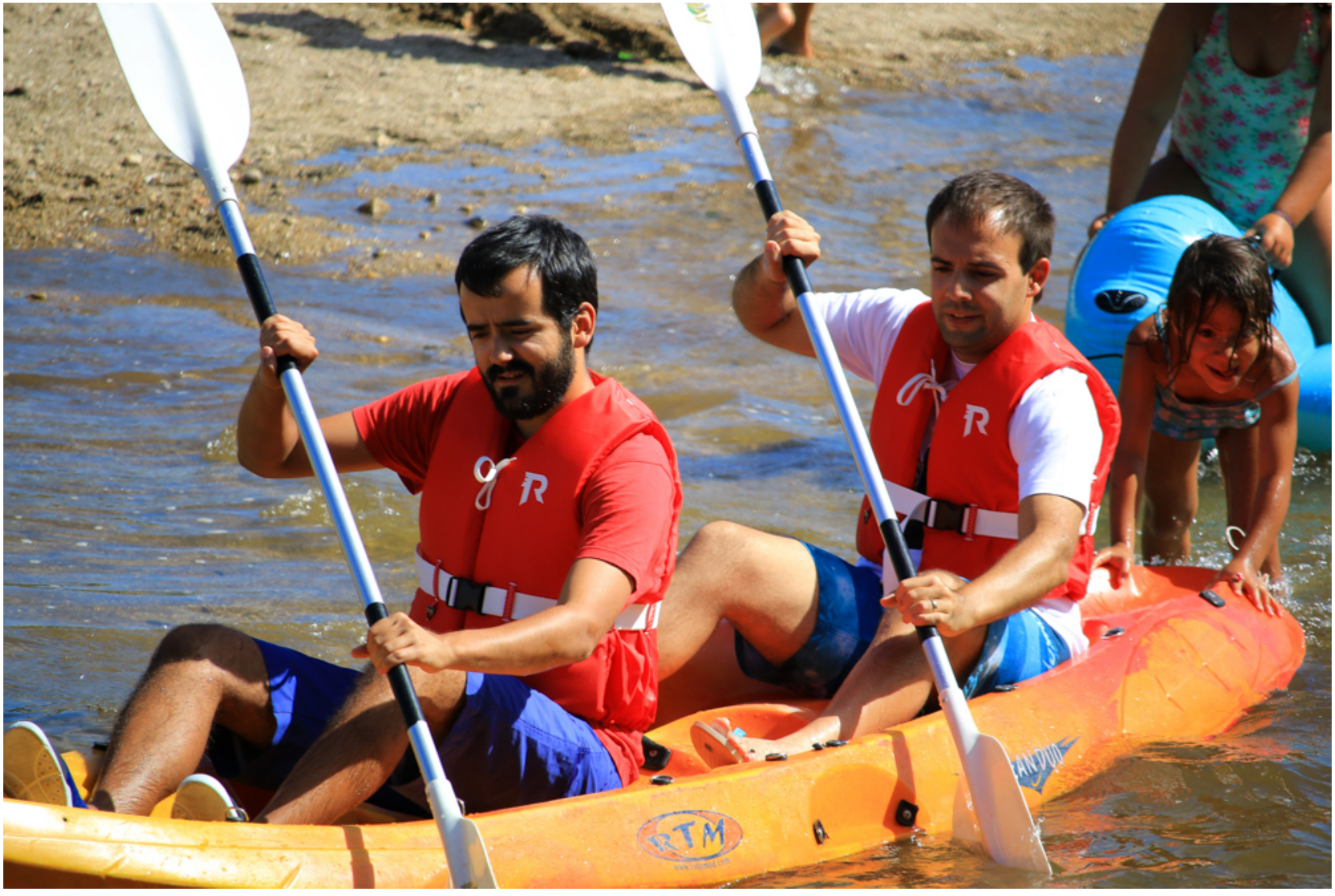
The top 10 mistakes I did (do?) as a Data Scientist

About me

- João Gomes (Sousa)
- 3 years as a Data Scientist at Farfetch
- Last year as team lead for Recommendations DS team
- PhD in computational and statistical physics from U. Aveiro
- Reach me at data@jgsousa.com, linkedin.com/in/jgsousa


RTM

Disclaimer

These are *my* mistakes and they don't represent the typical DS I've worked with. These are things I did, and sometimes still do, despite the best efforts of my managers, team and the organization to teach me otherwise. That I can make this presentation today is a testament that they have been at least partially succesful in helping me improve.

An interactive experiment

- Go to [jgsousa.com/10-mistakes]
- Answer
  - 'thumbs up' if you make that mistake or can relate to it
  - 'thumbs down' if you think you have that one figured out
  - '?' if you're unsure
- Answers are anonymous!

Is this working?

14 👍 | 2 ❓ | 3 👎

#1

I thought I could do it all

20 👍 | 7 ? | 2 👎

I thought I could do it all

20 👍 | 7 ❓ | 2 👎

How many times do you answer the question *what do you do?* with *a bit of everything*?

I thought I could do it all

20 👍 | 7 ❓ | 2 👎

How many times do you answer the question *what do you do?* with *a bit of everything*?

As a Data Scientist we know just enough of everything to be dangerous (or we think we do)

Jack of all trades, master of some, but time? none.

I thought I could do it all

20 👍 | 7 ? | 2 👎

How many times do you answer the question *what do you do?* with *a bit of everything*?

As a Data Scientist we know just enough of everything to be dangerous (or we think we do)

Jack of all trades, master of some, but time? none.

Repeat after me:

- I am not a developer
- I am not a DBA
- I am not a release engineer
- I am not a sysadmin
- I am not a product owner

I thought I could do it all

20 👍 | 7 ? | 2 👎

How many times do you answer the question *what do you do?* with *a bit of everything*?

As a Data Scientist we know just enough of everything to be dangerous (or we think we do)

Jack of all trades, master of some, but time? none.

Repeat after me:

- I am not a developer
- I am not a DBA
- I am not a release engineer
- I am not a sysadmin
- I am not a product owner

So what am I?

I thought I could do it all

20 👍 | 7 ? | 2 👎

How many times do you answer the question *what do you do?* with *a bit of everything*?

As a Data Scientist we know just enough of everything to be dangerous (or we think we do)

Jack of all trades, master of some, but time? none.

Repeat after me:

- I am not a developer
- I am not a DBA
- I am not a release engineer
- I am not a sysadmin
- I am not a product owner

So what am I?

- It Depends

I thought I could do it all

20 👍 | 7 ❓ | 2 👎

Work with engineers to put your code in production, don't be a black box

- Important: **do actual production code!**

I thought I could do it all

20 👍 | 7 ? | 2 👎

Work with engineers to put your code in production, don't be a black box

- Important: **do actual production code!**

Work with release engineers to build CI pipelines deploy to live

- dont deploy to live manually (`scp -r mymodel root@live:~/totallylegit/`)
- don't build your own CI pipeline

I thought I could do it all

20 👍 | 7 ❓ | 2 👎

Work with engineers to put your code in production, don't be a black box

- Important: **do actual production code!**

Work with release engineers to build CI pipelines deploy to live

- dont deploy to live manually (`scp -r mymodel root@live:~/totallylegit/`)
- don't build your own CI pipeline

Ask sysadmins to help you setup your jupyter server, setup gpu machines

- It's a pain to keep doing this when you grow 30x

I thought I could do it all

20 👍 | 7 ? | 2 👎

Work with engineers to put your code in production, don't be a black box

- Important: **do actual production code!**

Work with release engineers to build CI pipelines deploy to live

- dont deploy to live manually (`scp -r mymodel root@live:~/totallylegit/`)
- don't build your own CI pipeline

Ask sysadmins to help you setup your jupyter server, setup gpu machines

- It's a pain to keep doing this when you grow 30x

Work with data or platform engineers to scale out your solution

- Your first try (probably) won't be good enough at scale
- Actually, talk to them before you get started!

I thought I could do it all

20 👍 | 7 ? | 2 👎

Work with engineers to put your code in production, don't be a black box

- Important: **do actual production code!**

Work with release engineers to build CI pipelines deploy to live

- dont deploy to live manually (`scp -r mymodel root@live:~/totallylegit/`)
- don't build your own CI pipeline

Ask sysadmins to help you setup your jupyter server, setup gpu machines

- It's a pain to keep doing this when you grow 30x

Work with data or platform engineers to scale out your solution

- Your first try (probably) won't be good enough at scale
- Actually, talk to them before you get started!

You have special requirements.

- Make sure your organization knows **you need help** doing these things

#2

I didn't help the organization understand Data Science

5 👍 | 9 ? | 4 👎

Machine (or Deep) Learning is not magic. It's math. And data.

- some guy from some conference

I didn't help the organization understand Data Science

5 👍 | 9 ❓ | 4 👎

DS is not difficult, but it is *complex* and hard to grasp from the outside.

I didn't help the organization understand Data Science

5 👍 | 9 ? | 4 👎

DS is not difficult, but it is *complex* and hard to grasp from the outside.

Don't be the "guy doing *that* thing that no one really understands" (meaning: no one really cares)

I didn't help the organization understand Data Science

5 👍 | 9 ❓ | 4 👎

DS is not difficult, but it is *complex* and hard to grasp from the outside.

Don't be the "guy doing *that* thing that no one really understands" (meaning: no one really cares)

It's on us to teach our organization:

- What is Data Science
- How our algorithms work
- What we can do
- What are our **limitations**
- What is our role

I didn't help the organization understand Data Science

5 👍 | 9 ? | 4 👎

What is our role?

- Looking at data
- Making predictions
- Generating hypothesis
- Applying machine learning
- Working with clients
- Communicating results
- Building reports

Different organizations will think different things.

I didn't help the organization understand Data Science

5 👍 | 9 ? | 4 👎

What is our role?

- Looking at data
- Making predictions
- Generating hypothesis
- Applying machine learning
- Working with clients
- Communicating results
- Building reports

Different organizations will think different things.

- Just make sure you help them figure it out and that it's clear for both parties

#3

I cared more about the technology than the problem

6 👍 | 3 ❓ | 7 👎

I cared more about the technology than the problem

6 👍 | 3 ? | 7 👎

Things I did because I wanted to learn:

- A recommender system from scratch
- Spark
- Docker
- Deep Learning

I cared more about the technology than the problem

6 👍 | 3 ? | 7 👎

Things I did because I wanted to learn:

- A recommender system from scratch
- Spark
- Docker
- Deep Learning

Learning new technologies is *very* important.

But all that really matters is solving the problem

- Customer problem
- Business problem

I cared more about the technology than the problem

6 👍 | 3 ? | 7 👎

Things I did because I wanted to learn:

- A recommender system from scratch
- Spark
- Docker
- Deep Learning

Learning new technologies is *very* important.

But all that really matters is solving the problem

- Customer problem
- Business problem

(Some would say all that matters is actually *finding* the problem)

#4

I wasn't practical enough

8 👍 | 3 ? | 2 👎

I wasn't practical enough

8 👍 | 3 ? | 2 👎

Do everything you can to put things in front of the customer asap

I wasn't practical enough

8 👍 | 3 ? | 2 👎

Do everything you can to put things in front of the customer asap

Simple is better than complex

Good enough is better than great if great comes too late

Get data first, get better later

I wasn't practical enough

8 👍 | 3 ? | 2 👎

Do everything you can to put things in front of the customer asap

Simple is better than complex

Good enough is better than great if great comes too late

Get data first, get better later

- Negotiate with engineering best way to do that
- Copy data to production databases by hand if needed be
- Measure and iterate
- Work with them to make production ready code
- Move on
- But **don't quit your code!**
  - Gentlemen and gentlewomen don't ship and forget

I wasn't practical enough

8 👍 | 3 ? | 2 👎

Do everything you can to put things in front of the customer asap

Simple is better than complex

Good enough is better than great if great comes too late

Get data first, get better later

- Negotiate with engineering best way to do that
- Copy data to production databases by hand if needed be
- Measure and iterate
- Work with them to make production ready code
- Move on
- But **don't quit your code!**
  - Gentlemen and gentlewomen don't ship and forget

Don't be *too* practical. You're expected to think outside the box.

#5

I didn't look at the data enough

17 👍 | 3 ? | 2 👎

I didn't look at the data enough

17 👍 | 3 ❓ | 2 👎

You start on a project, you do data analysis, data cleaning, feature engineering...

You plug the models in. You iterate, optimize and find the best one.

You ship it, measure online performance

You forget to look at the data again

I didn't look at the data enough

17 👍 | 3 ❓ | 2 👎

You start on a project, you do data analysis, data cleaning, feature engineering...

You plug the models in. You iterate, optimize and find the best one.

You ship it, measure online performance

You forget to look at the data again

- Data changes!
  - Monitor the data
  - Retrain often (not automatically!)
- KPIs are aggregate measures
  - Always be deep diving

#6

Getting more data wasn't my first priority

8 👍 | 1 ? | 2 👎

Getting more data wasn't my first priority

8 👍 | 1 ? | 2 👎

I always assumed that the data I was given was all the data I needed. This was wrong.

Getting good data is harder than modeling it.

Let your company know this! Good data is your company's greatest competitive advantage.

Getting more data wasn't my first priority

8 👍 | 1 ? | 2 👎

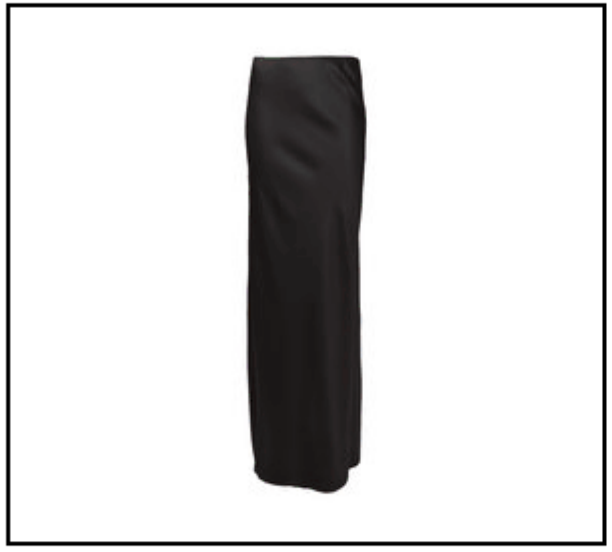
I always assumed that the data I was given was all the data I needed. This was wrong.

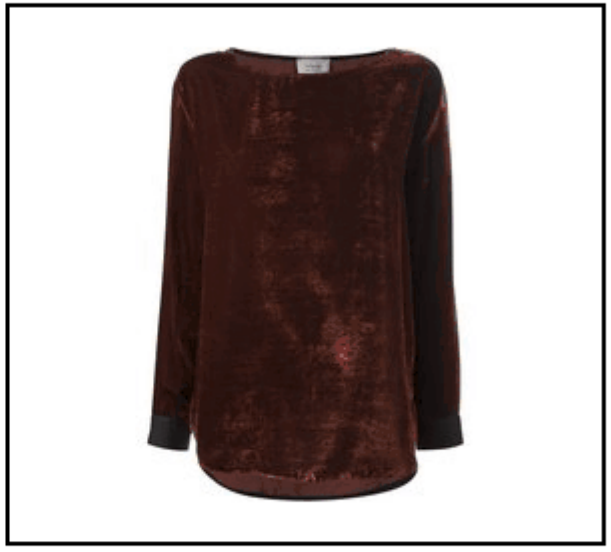
Getting good data is harder than modeling it.

Let your company know this! Good data is your company's greatest competitive advantage.

- Look for open data sets
- Scrape data (rather, ask engineering to do it)
- Buy data
- Build tools to get more data
  - Amazon turk
  - internal tools

Do these products go well together?





(Y)es (N)o (S)kip

(C)omment - optional

#7

I didn't read enough literature

6 👍 | 3 ? | 8 👎

I didn't read enough literature

6 👍 | 3 ? | 8 👎

It's too easy to get lost in what you know, your tools and your process

I didn't read enough literature

6 👍 | 3 ? | 8 👎

It's too easy to get lost in what you know, your tools and your process

Don't go to your first choice. Read up on the problem first

Keep up with what's new in the field. It's fun!

- But don't get overwhelmed! You can't possibly keep up with everything.

But don't forget to go back to the basics!

#8

I didn't care about reproducibility

4 👍 | 0 ? | 3 👎

I didn't care about reproducibility

4 👍 | 0 ? | 3 👎

Experimental science can't exist without reproducibility

I didn't care about reproducibility

4 👍 | 0 ? | 3 👎

Experimental science can't exist without reproducibility

Notebooks are awesome! But not enough:

- aka the famous `doc_final.ipynb`, `doc_final_2.ipynb`, `doc_final_forreals.ipynb` problem

Git for notebooks is something, but not great

I didn't care about reproducibility

4 👍 | 0 ? | 3 👎

Experimental science can't exist without reproducibility

Notebooks are awesome! But not enough:

- aka the famous `doc_final.ipynb`, `doc_final_2.ipynb`, `doc_final_forreals.ipynb` problem

Git for notebooks is something, but not great

I don't think there is a good solution available. I would love

- Notebook source control
- With code review support at a cell level
- **Data is versioned** alongside notebook
- Auto publish notebook for reading
- Searchable, taggable notebooks

Maybe Airbnb knowledge repo?

#9

I didn't care enough about privacy

6 👍 | 3 ? | 7 👎

I didn't care enough about privacy

6 👍 | 3 ? | 7 👎

GDPR fines: 4% of GMV, 20€ million. Whatever is biggest.

I didn't care enough about privacy

6 👍 | 3 ❓ | 7 👎

GDPR fines: 4% of GMV, 20€ million. Whatever is biggest.

"*I don't actually need full access to the database*", said no data scientist ever

It's just too easy to think data governance is someone else's problem.

I didn't care enough about privacy

6 👍 | 3 ❓ | 7 👎

GDPR fines: 4% of GMV, 20€ million. Whatever is biggest.

"*I don't actually need full access to the database*", said no data scientist ever

It's just too easy to think data governance is someone else's problem.

- Talk to your DBAs, setup proper access control
- Don't pull data you don't need
- Mask what you can
- Encrypt everything else
- Don't carry it around in your laptop!

Let's be responsible, and not make the industry afraid of us!

#10

I think of the title of my presentations before the content

14 👍 | 0 ? | 0 👎

I think of the title of my presentations before the content

14 👍 | 0 ? | 0 👎

A corollary: We think of the acronym for our tools/services before we figure out the meaning of the letters

- PRECOG: **P**ython **Reco**mmendations **G**enerator
- COMPAL: **Comp**lementary **A**ctive **L**earning
- DATMAN: **Dat**a **mon**itoring
- VIPER: **V**isual **I**nformation for **P**roduct **E**xploration and **R**etrieval

I think of the title of my presentations before the content

14 👍 | 0 ? | 0 👎

A corollary: We think of the acronym for our tools/services before we figure out the meaning of the letters

- PRECOG: **P**ython **Reco**mmendations **G**enerator
- COMPAL: **Comp**lementary **A**ctive **L**earning
- DATMAN: **Dat**a **mon**itoring
- VIPER: **V**isual **I**nformation for **P**roduct **E**xploration and **R**etrieval

This is not actually a mistake, this is the best part of the job

Thank you!