DataOps

Creating Data Based Solutions **ASAP**

¿ Who am I in a nutshell?

- Data/ML/Meme Engineer @ gradiant
- AI Master Student
- VigoBrain AI MeetUp CoOrganizer

Gradiant, ICT technology centre in Spain

Since 2008, focused on technological development and knowledge transfer to industry

+100

professionals

14

european projects

5,2M€

revenue in 2017

54%

contracted companies

46%

competitive public funding

Our sectors

Industry 4.0

Security and Defense

Farming and Natural Resources

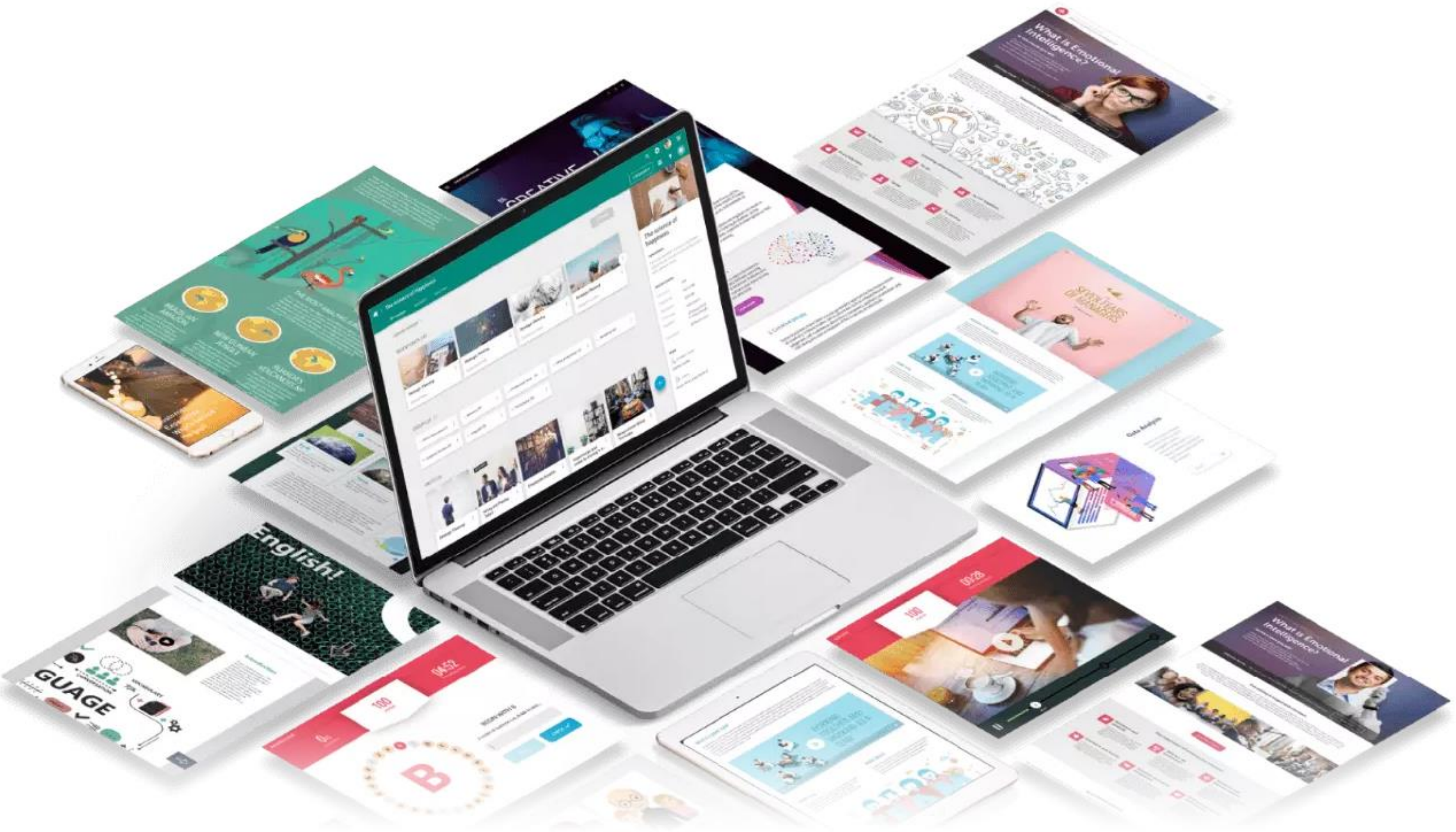
Aerospace

Marketing, Retail and Audiovisuals

Banking, Digital Society and Education

Healthcare and Wellness

Telecommunications


What is Emotional Intelligence?
English!
GUAGE
Data Analysis
00:28
100
04:52
B
What is Emotional Intelligence?

“

¿ What Is DataOps ?

What Is DataOps?

“

*DataOps is an **automated**, **process-oriented** methodology, used by analytic and data teams, to **improve the quality and reduce the cycle time of data analytics** …*

*DataOps applies to the **entire data lifecycle** from data preparation to reporting, and recognizes the **interconnected** nature of the **data analytics team and IT operations**.*

DataOps - Wikipedia

DataOps applies 3 Methodologies…

DevOps

Agile

SPC
(Statistic Process Control)

Lean Manufactring - SPC

Is a systematic method for the minimization of waste (muda) within a manufacturing system without sacrificing productivity

Manifesto


The DataOps Manifesto
Through firsthand experience working with data across organizations, tools, and industries we have uncovered a better way to develop and deliver analytics that we call DataOps.

Manifesto

1. Continually satisfy your customer
2. Value working analytics
3. Embrace change
9. Analytics is code
10. Make it reproducible
16. Monitor quality and performance

“

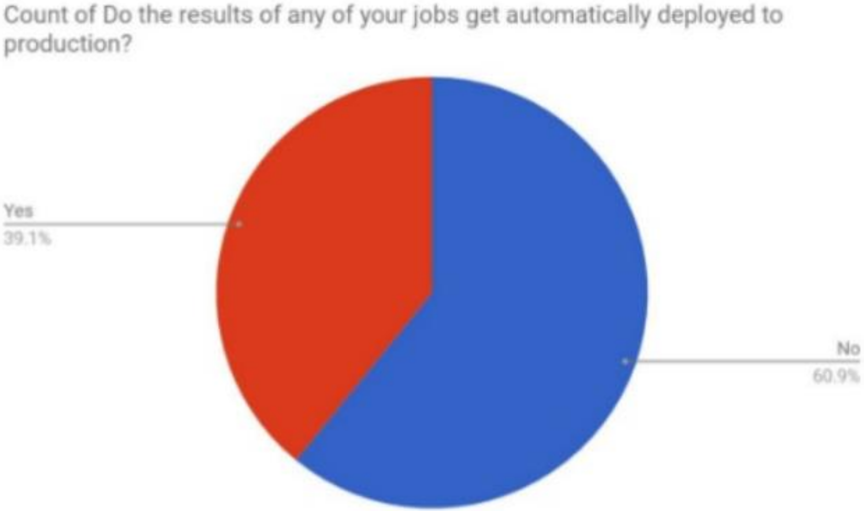
¿How many times have you seen all this methodologies aplyed to data based solutions?

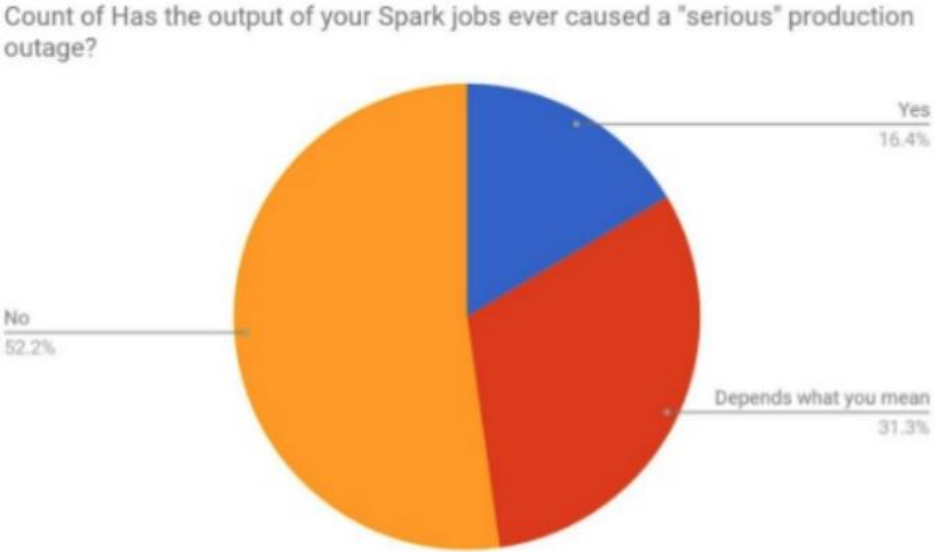
When you work with data...

Deployments...

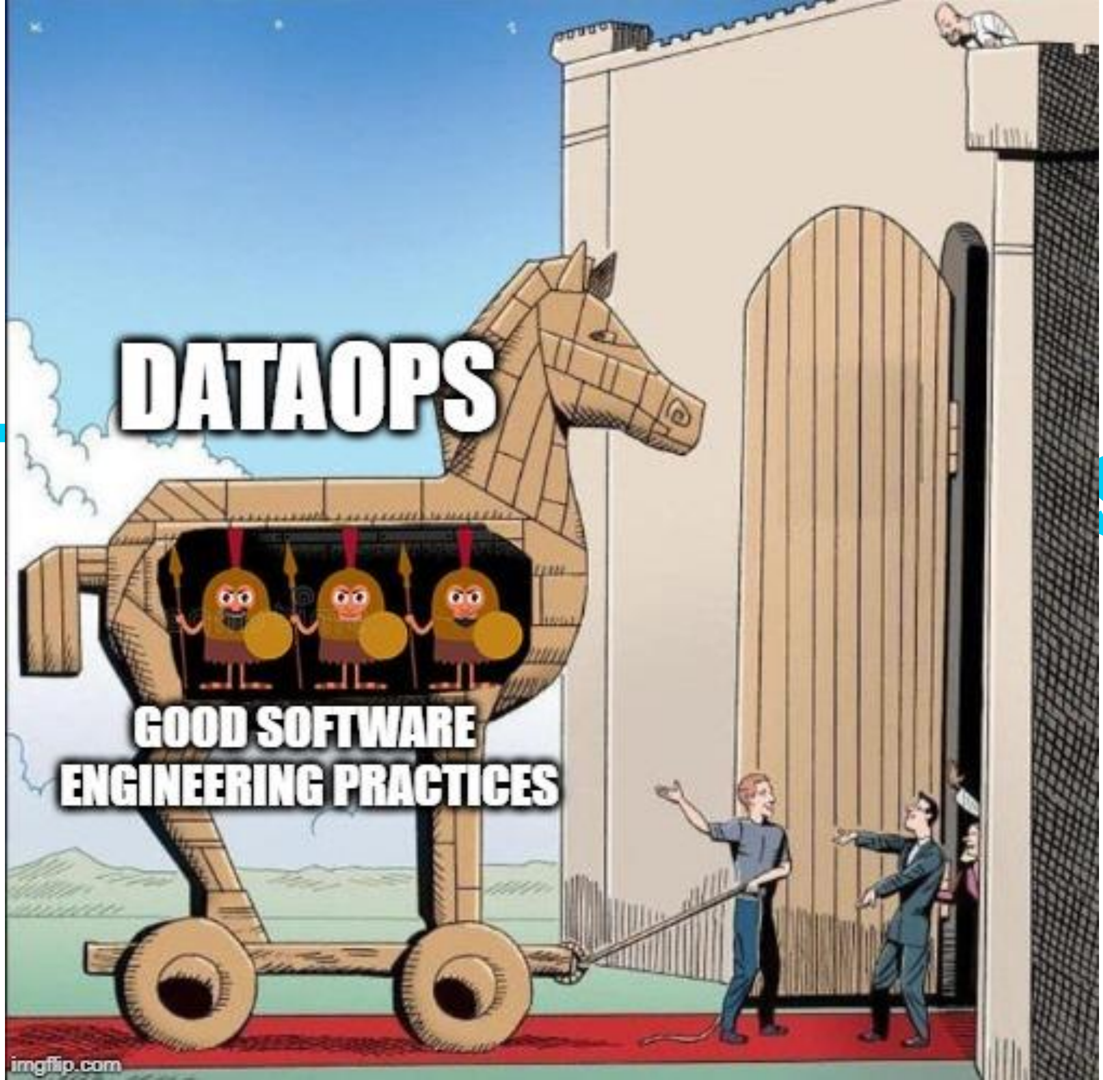
Holden Karau [@holdenkarau](https://twitter.com/holdenkarau)

- Works with Google on Apache Beam project
- Apache Spark Committer
- Co-author of O'Reilly's Learning Spark and High Performance Spark.





So I T



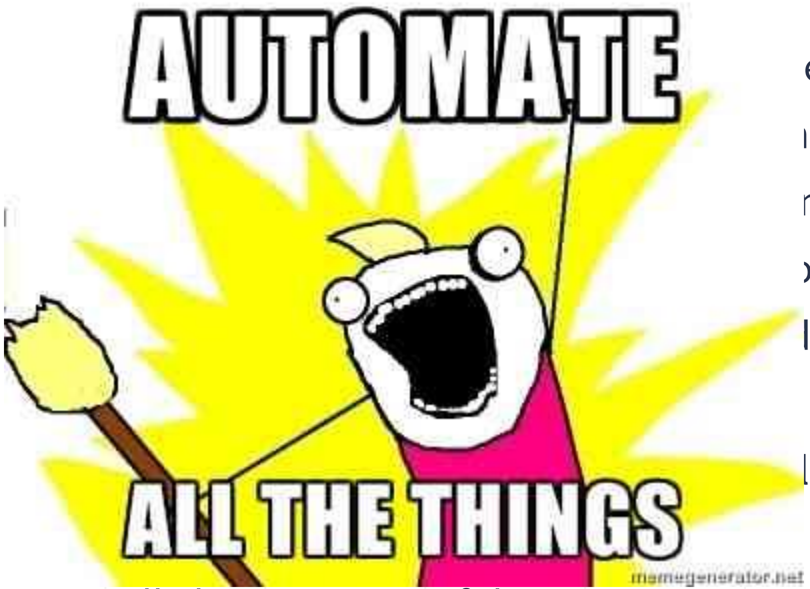
s talk

My Team Journey

Team Background

- Strong Software Engineering Skills
- We use Gitflow as our repository workflow
- We package all our work
- We embrace TDD and DDD
- Everything we code goes through CI/CD
- We encourage clean & reusable code
- We usually use Scrum

Good SW Engineering practices means been lazy



I usually have more confidence on my automated processes that in myself

That allows us to spend time on

Automate more things that I don’t want to spend my time on them

Create more data pipelines or enrich current pipelines

Do more analytics

Explore ML/DL models

Improve current models metrics

Improve current system quality

Research more ways to be more lazy

Backend

TS

node JS

Data Layer

kafka

mongoDB

hadoop HDFS

Engines

APACHE Spark

Flink

Analytics

R

Flask web development, one drop at a time

plumber

POCs & Reports

Apache Zeppelin

jupyter

CO

Visualization Layer

flow

Grafana

Testing and Production Environemnt

docker

There’s Pain & Tears behind all thoose technologies

Be careful with notebooks environments

```
def train(self):
    """Learn the vectors p_u and q_i with SGD.
       data is the user-item matrix
       n_factor is the number of latent factors to use
       alppha is the learning rate of the SGD
       n_epochs is the number of iterations to run the algorithm
    """
    self.is_training = True
```

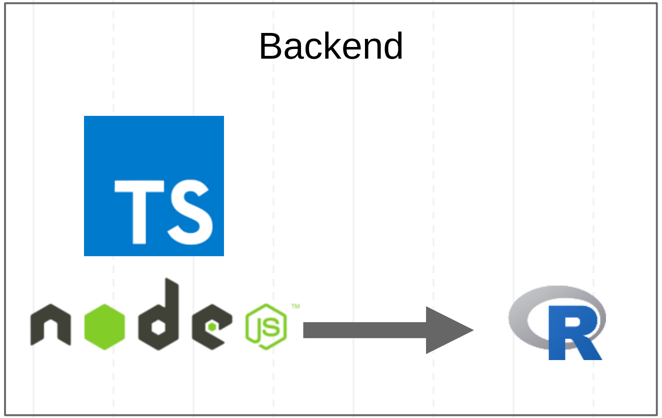
```
...\PycharmProjects\dash-board\src\services\SGD.py:45: RuntimeWarning:
overflow encountered in multiply
...\PycharmProjects\dash-board\src\services\SGD.py:46: RuntimeWarning:
overflow encountered in multiply
```

```
    self._u = p
    self._v = q

    self.is_training = False
    self.is_train = True

    return p, q
```

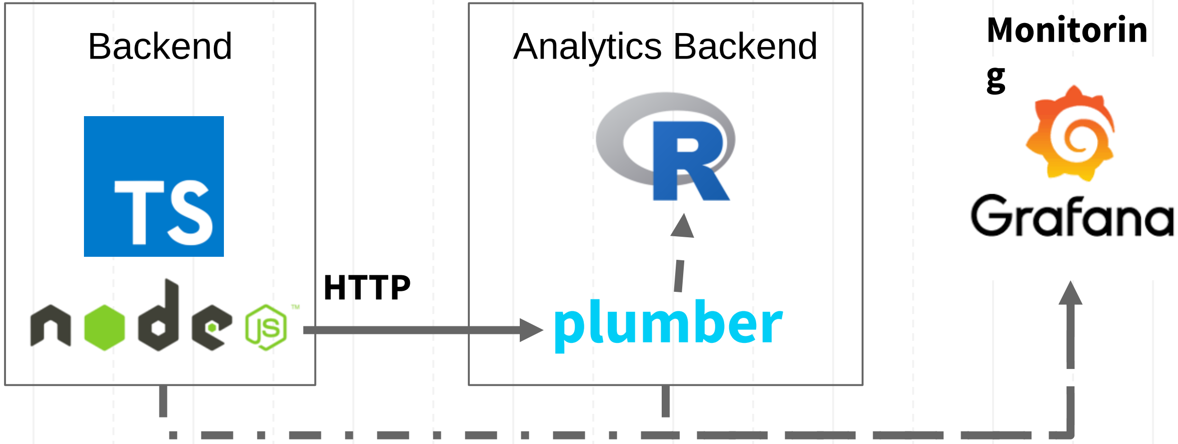
We are using a bunch of technologies, so there's a ton of points of failure (I)

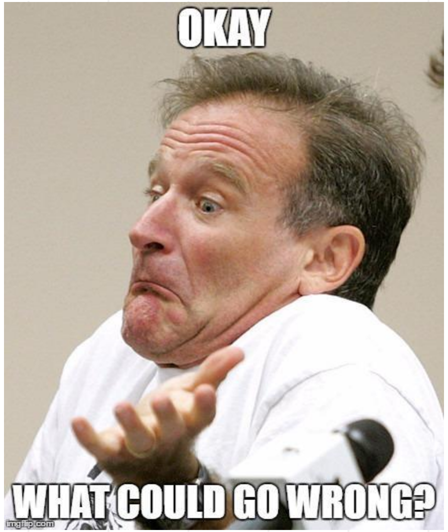



```
var exec = require('child_process').exec;
    exec('R my_awesomic_analytics ' + params,
    function callback(error, stdout, stderr){
        doStuff(stdout)
});
```

if something went wrong on the R part it could destroy our k8 pod
We need brute force strategies to scale this
It's hard to test R side

We are using a bunch of technologies, so there's a ton of points of failure (II)





We have tests on both backends

We detected memory usage problems on plumber parsing HTTP requests

```scala
class AutoEncoderModelServiceTest extends FlatSpec with Matchers with BeforeAndAfterAll {

  [redacted]
  var sparkSession: SparkSession = _

  override def beforeAll() {
    sparkSession = TestUtils.getSparkTestSession
    [redacted]
  }

  override def afterAll(): Unit = {
    sparkSession.stop()
  }

  it should "be capable to make the same predictions as the original model" taggedAs Unit in {
    val test: INDArray = Nd4j
      .create(TestUtils.loadTestData("data/autoencoder_test.csv", sparkSession))
    val expected: INDArray = Nd4j
      .create(TestUtils.loadTestData("data/autoencoder_prediction.csv", sparkSession))

    val model: ComputationGraph = AutoEncoderModelService.loadModel
    val modelOutput: Array[INDArray] = model.output(test)
    val prediction: INDArray = modelOutput(0).toDense

    assert(expected.equals(prediction))
  }

}
```

If you want to embrace DataOps you may need new roles

Data Scientist

Responsabilities

- Create advanced analytics
- Interact with business and help them
- Create reports
- Research on AI

Abilities

- Math & Statistics Background
- Create insights using business domain knowldege
- Good communication skills (verbally & visually)

Weakness

- Programming skills
- System creation/management skills

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

Data Engineer

Responsabilities

- Create data pipelines
- Choose right tools for data proccesing
- Combine multiple technologies to create solutions

Abilities

- Programming Background
- Knowldege in distributed systems
- System creation and management

Weakness

- Not a system person
- Weak analytics skills (compared to Data Scientists)

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

ML Engineer

Responsabilities

- Operationalizing Data scientist’s work
- Optimizing ML

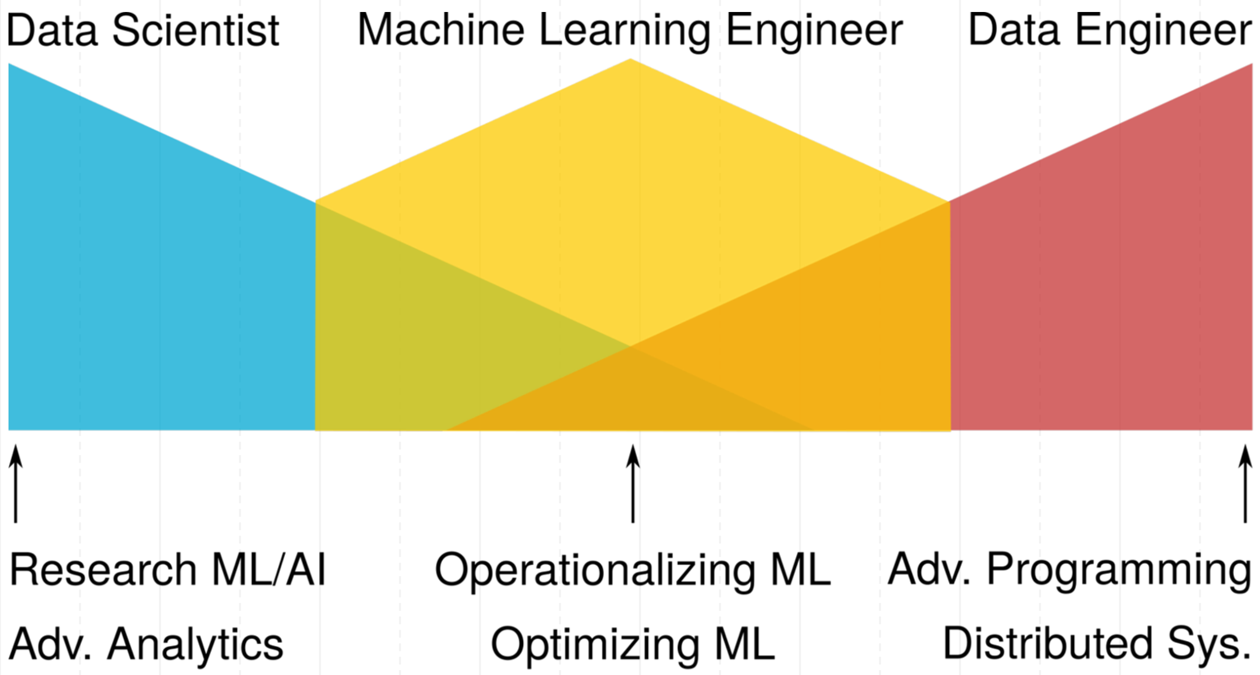
Abilities

- Data Engineering Abilites
- Strong Data Scientist Abilities
- Strong Engineer Principles

Weakness

- Knows too many things

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists



BDI

BIG DATA INSTITUTE

For more information go to http://bigdatainstitute.io

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists


Data Scientist
Machine Learning Engineer
Data Engineer
Research ML/AI
Adv. Analytics
Operationalizing ML
Optimizing ML
Adv. Programming
Distributed Sys.
BDI
Backend
TS
node JS
Data Layer
kafka
mongoDB
hadoop HDFS
Engines
APACHE Spark
Flink
Analytics
Flask
web development, one drop at a time
plumber
POCs & Reports
Apache Zeppelin
jupyter
CO
Visualization Layer
flow
Testing and Production Environemnt
docker
Grafana

Things we are thinking about

- Use DSC to version of data and experiments
- Waste less resources
  - Jupyterhub
  - Automatic scaling for spark and flink clusters
- Have a good VCS for notebooks:
  - manage versions, diffs, pull requests
- Automate notebooks validation → ¿automatic tests on notebooks?

¿Questions?

VIGOTECH
ALLIANCE