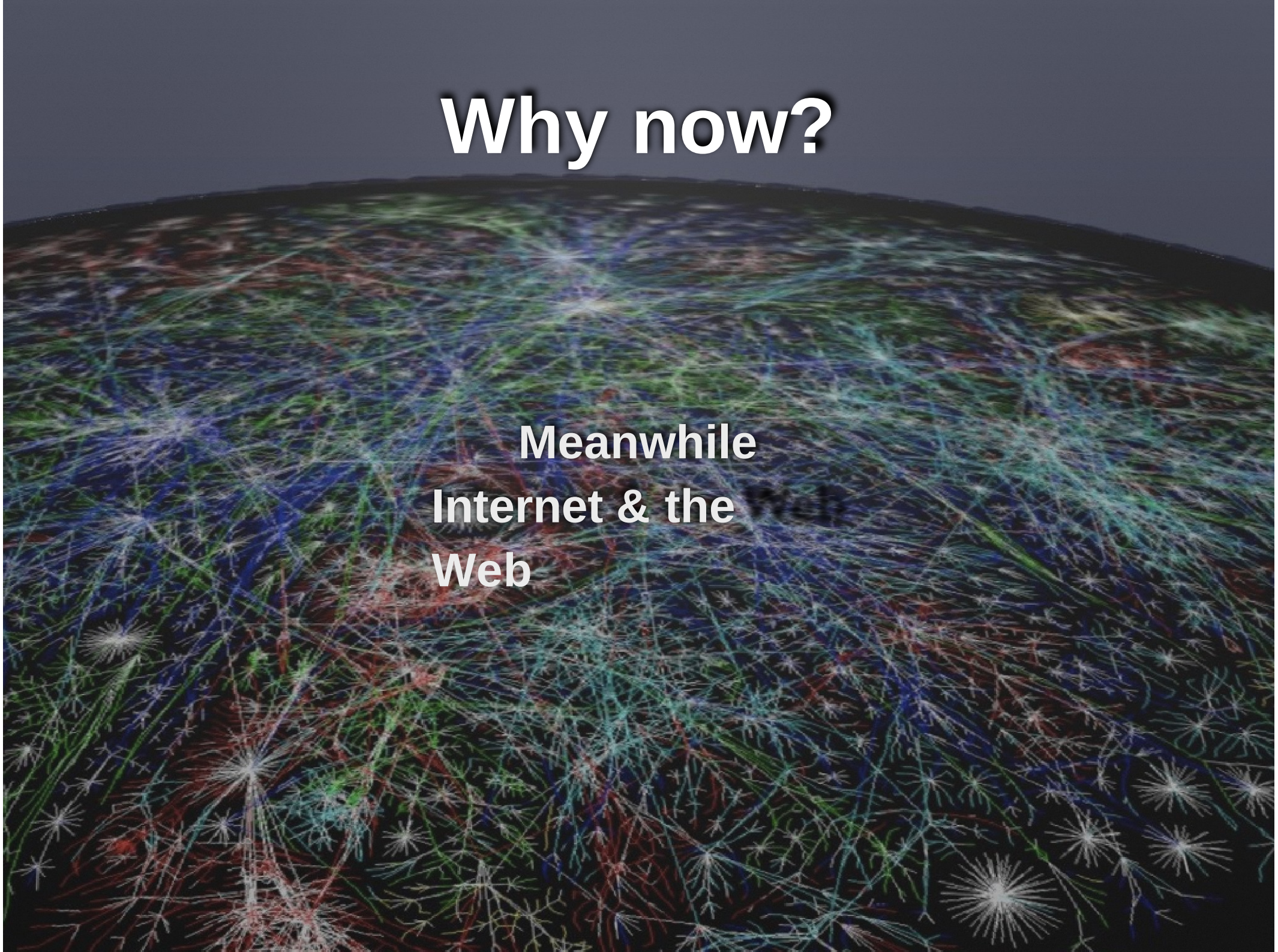
What is Data Science?

Prof. David Losada, Universidade de Santiago de Compostela
Email: david.losada@usc.es
Marcos Fernández, Investigador predotural, University of Santiago de Compostela
Email: marcosfernandez.pichel@usc.es

Data Science

Big Data

Why now?

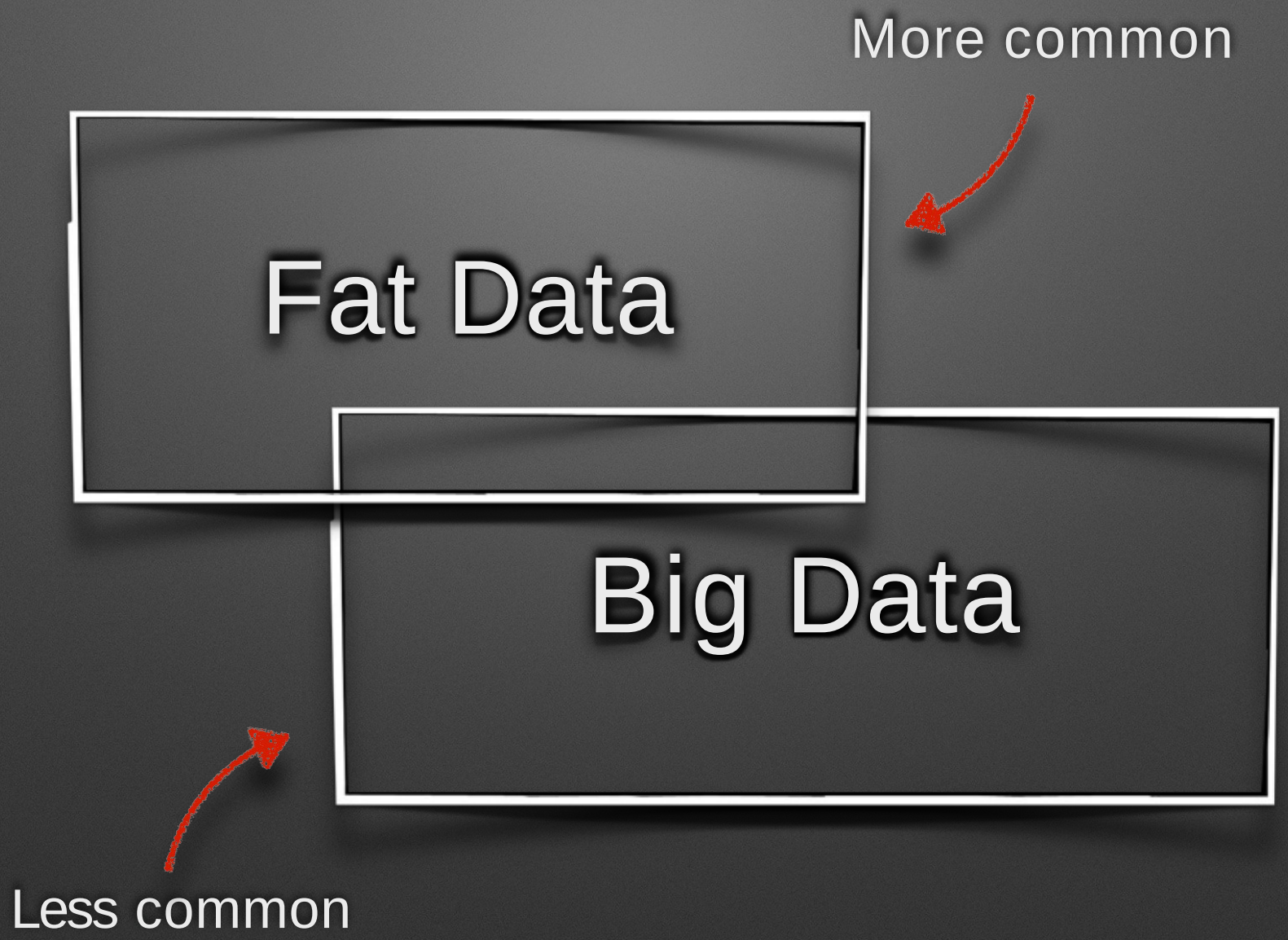


Meanwhile
Internet & the
Web

Big Data

What is Big Data?

- For some people, they have big data when its size > 65536 x 256.
- In general we have big data when its size does not allow its storage and analysis in a big computer.


More common
Fat Data
Big Data
Less common

Big Data

Big data is more than size.

It is commonly characterized with several V:

V:

- Volume
- Velocity
- Variety
- Veracity

Big Data

Key enabler

The cloud is key to deal with the four V, but the main phenomenon behind Big Data is **datification**.

The four V are a consequence of it.

Big Data

We are rendering into data many aspects of the world that have never been quantified before:

business networks
books I'm reading
location
physical activity
consumed food
purchases
physiological signals
straight thoughts
friendship
gaze
driving behavior

Big Data

Information comes from:

- Corporate Data Bases (structured information).
- Unstructured information in documents, Wikipedia, textbooks, journals, blogs, tweets, etc.
- Images in the web, public cameras, phones, TV, YouTube, etc.
- Public APIs: smart cities, government, search engines, etc.
- Sensor Data: GPS, accelerometer, physico-chemical sensors, sociometric sensors, super-colliders, telescopes, etc.

Big Data

There are several Big Data flavors:

- Big multidimensional arrays (homogeneous data).
- Big tables (structured data).
- Big text.
- Big image.
- Big sound.
- Big sequential data (sensors, tweets, etc.)

Big Data

There are several problems:

Analyzing the past

- ETL (Extract, Transform, Load)
- BI/Analytics (Think you can do in SQL)
- **Advanced Analytics.**
- **Machine Learning.**
- Visualization.

Predicting the future

Data Science

Steps:

- Ask a question.
- Get the data. They can be heterogeneous and non structured.
- Data Processing (cleaning, munging, etc.).
- Data Analysis (computer science, linguistics, economy, sociology, etc.).
- Take a decision and act.

Data Science

Spotify

COMPANY
Spotify

INDUSTRY
Entertainment

EMPLOYEES
5,000

TYPE
Customer Segmentation & Behavioral Analytics

Purpose:

Spotify uses data from user profiles and users' playlists, and historical data on music played to provide recommendations for each user. By combining data from millions of users, Spotify is able to make recommendations even if a particular user doesn't have an extensive history with the site.

Conclusions

- Big Data will be soon a **commodity** that will be used mainly for data munging and counting at scale.
- The most difficult part of **Big Data** is **Data**.
- Data Science is a new job with a bright future.