Introduction to Data Lakes

What we'll cover

- What is a data lake
- Why would you need one
- What you would put in one
- How you would build one
- What you do when you've got one
- What problems you should avoid

Dr Andrew Rivers

345 Technology

What is a Data Lake?

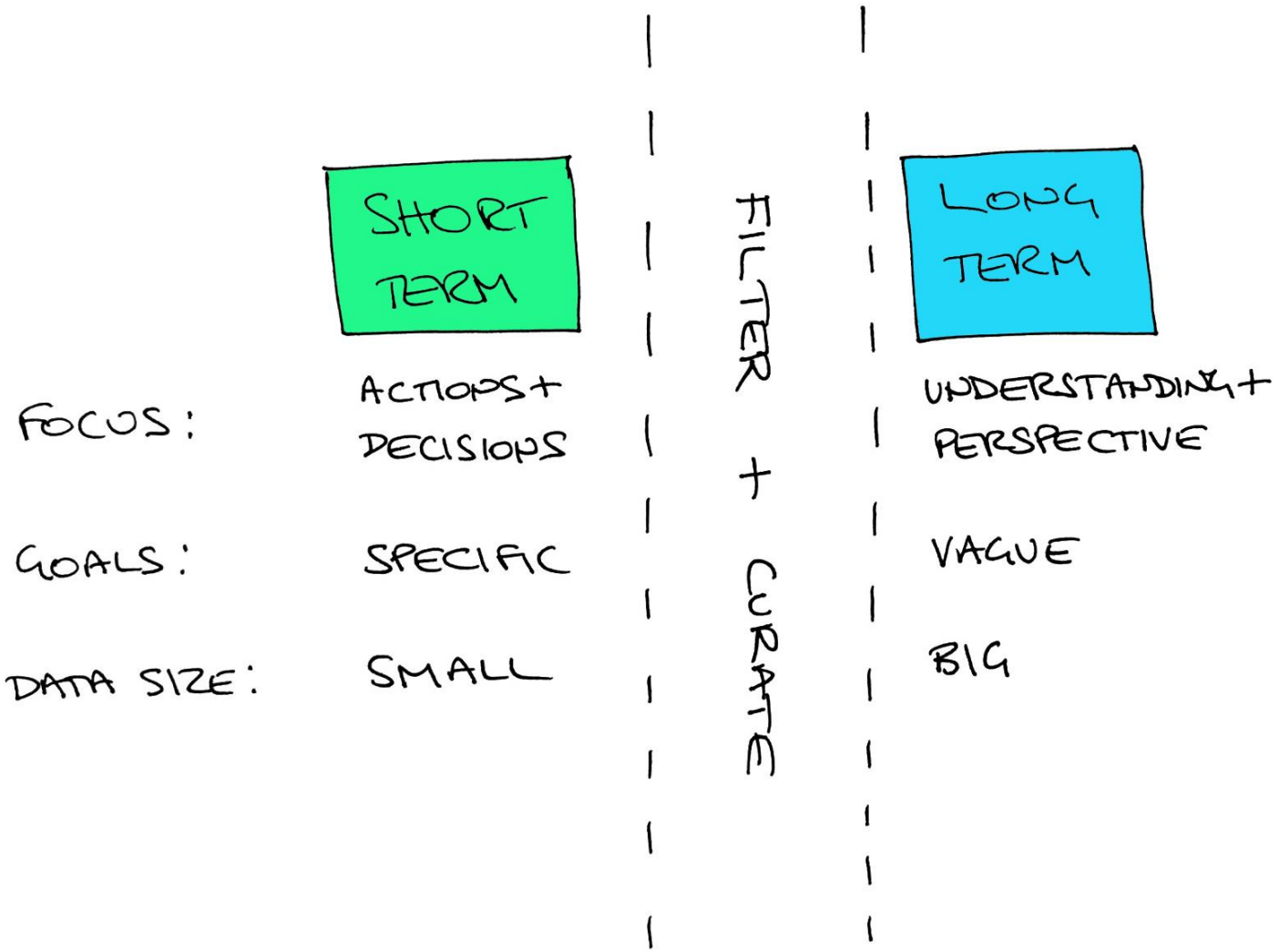
A data store designed to hold **GARGANTUAN** amounts of data

We live in a world of

What does is take to **succeed** with

Where do we store

Data vs time

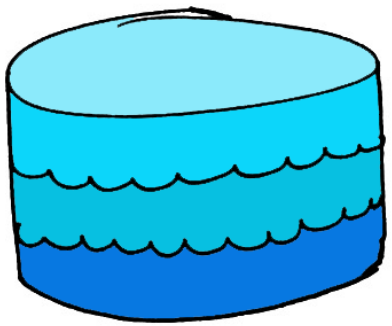

SHORT TERM
LONG TERM
FILTER + CURATE
FOCUS:
ACTIONS + DECISIONS
UNDERSTANDING + PERSPECTIVE
GOALS:
SPECIFIC
VAGUE
DATA SIZE:
SMALL
BIG

Big Data Principles

- Data can come from anywhere, in any format
- Data is valuable
- Capture it all and understand it later
- The more I learn the better my questions will be

Why have a data lake

CAPTURE EVERYTHING



UNDERSTAND IT LATER

Why not use an existing type of data storage?

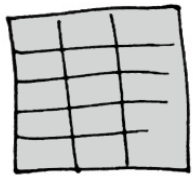
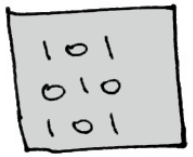


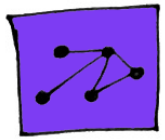
TABLE / COLUMN



BLOB / FILE

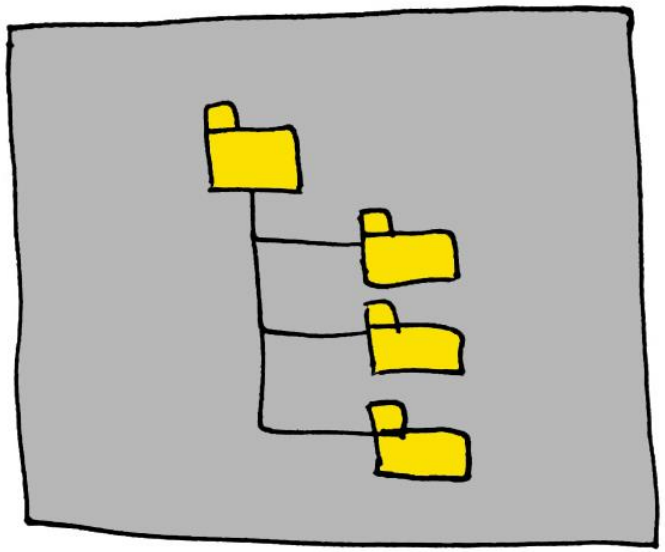
RELATIONAL

DOCUMENT



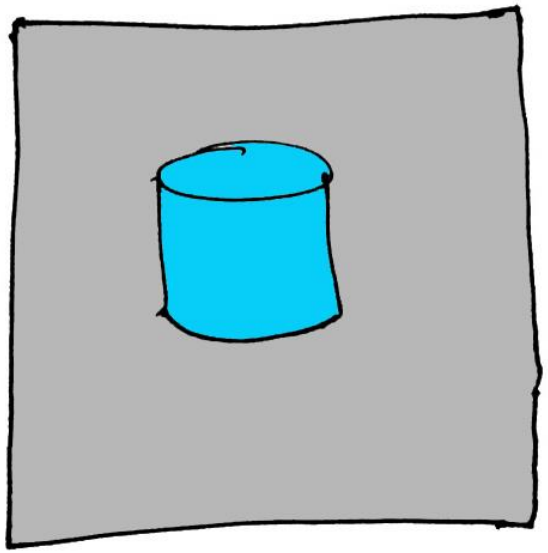
GRAPH

So what actually is a data lake?



HIERARCHICAL

BLOB
STORE



DATA

CATALOG

Data lake vs data warehouse

Data Lake

- Everything from unstructured, semi-structured to structured
- Hold everything and decide what to do later
- Store it first and cleanse, enrich later
- ELT

Data Warehouse

- Highly structured
- Decide in advance what you will hold
- Hold what you want to analyse
- Data is cleansed and transformed before storage
- ETL

What you'd put in a data lake

Everything

Video

Audio

Images

Documents

Logs

Telemetry

Archived data

Why build one now?

How expensive is it?

File structure: Hierarchical namespace | Redundancy: LRS | Region: North Europe | Currency: British Pound (£)

Data storage prices

	HOT	COOL	ARCHIVE
First 50 terabyte (TB) / month	£0.0164 per GB	£0.0075 per GB	£0.0008 per GB
Next 450 TB/month	£0.0158 per GB	£0.0075 per GB	£0.0008 per GB
Over 500 TB/month	£0.0151 per GB	£0.0075 per GB	£0.0008 per GB

Storage Capacity reservations

	1-YEAR RESERVED			3-YEAR RESERVED		
	HOT	COOL	ARCHIVE	HOT	COOL	ARCHIVE
100 TB/month	£1,377	£626	£68	£1,109	£504	£62
1 PB/month	£13,411	£6,096	£658	£10,660	£4,846	£604

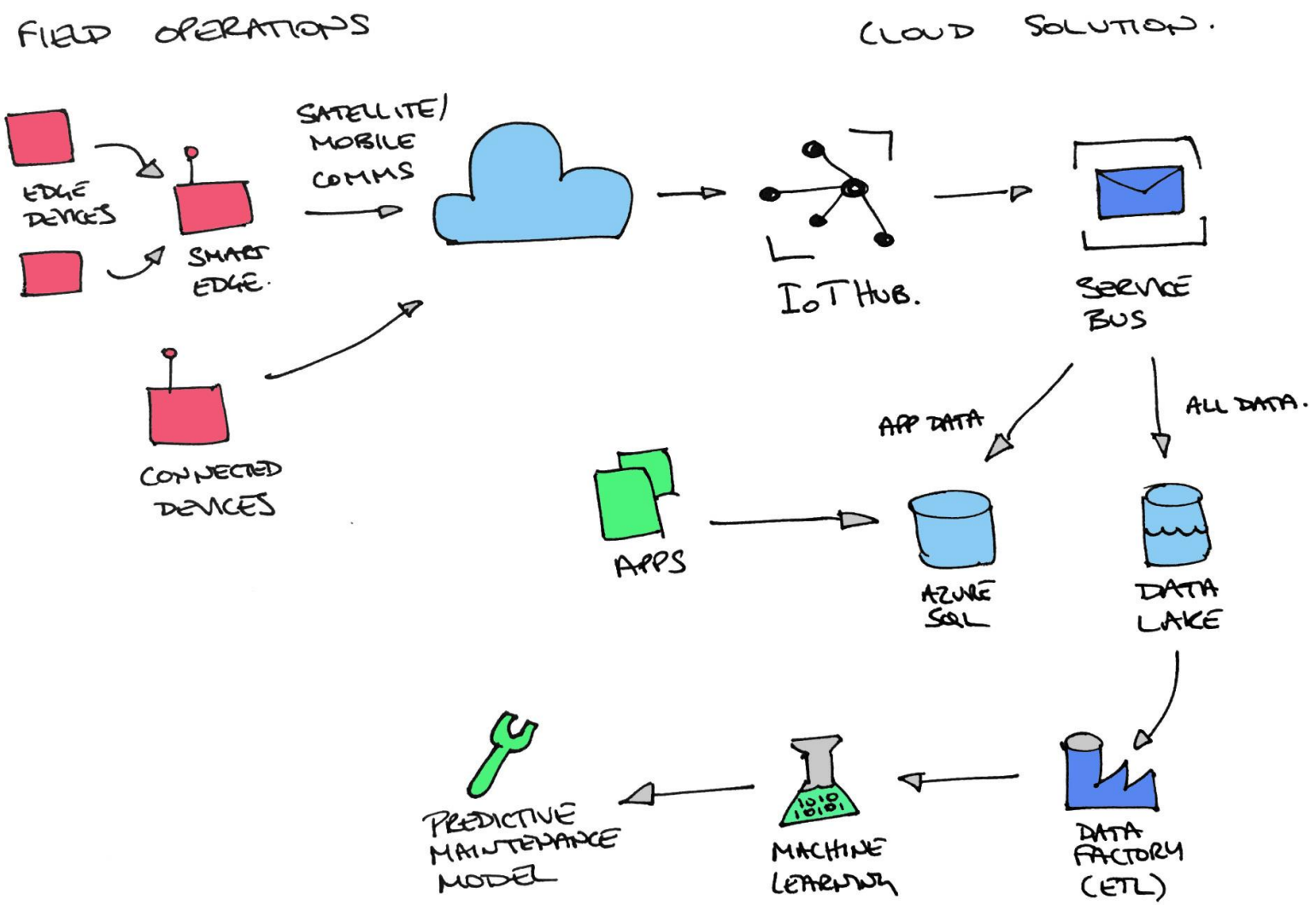
Avoid the headaches

- Security
- Regulatory (e.g. GDPR)
- Data sovereignty
- Build a lake not a swamp
- Understand the cost model

Where does Data Science come in?

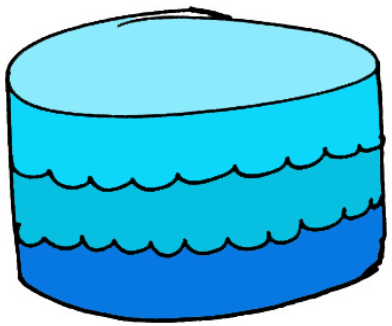
- Your Data Lake will be your source for data sets
  - Statistical analysis
  - Machine learning
- Your Data Lake is a destination for enrichment
  - Cognitive search
  - Content analysis
  - Classification

Case Study – IoT for Mining



In Summary

CAPTURE EVERYTHING



UNDERSTAND IT LATER

Discussion