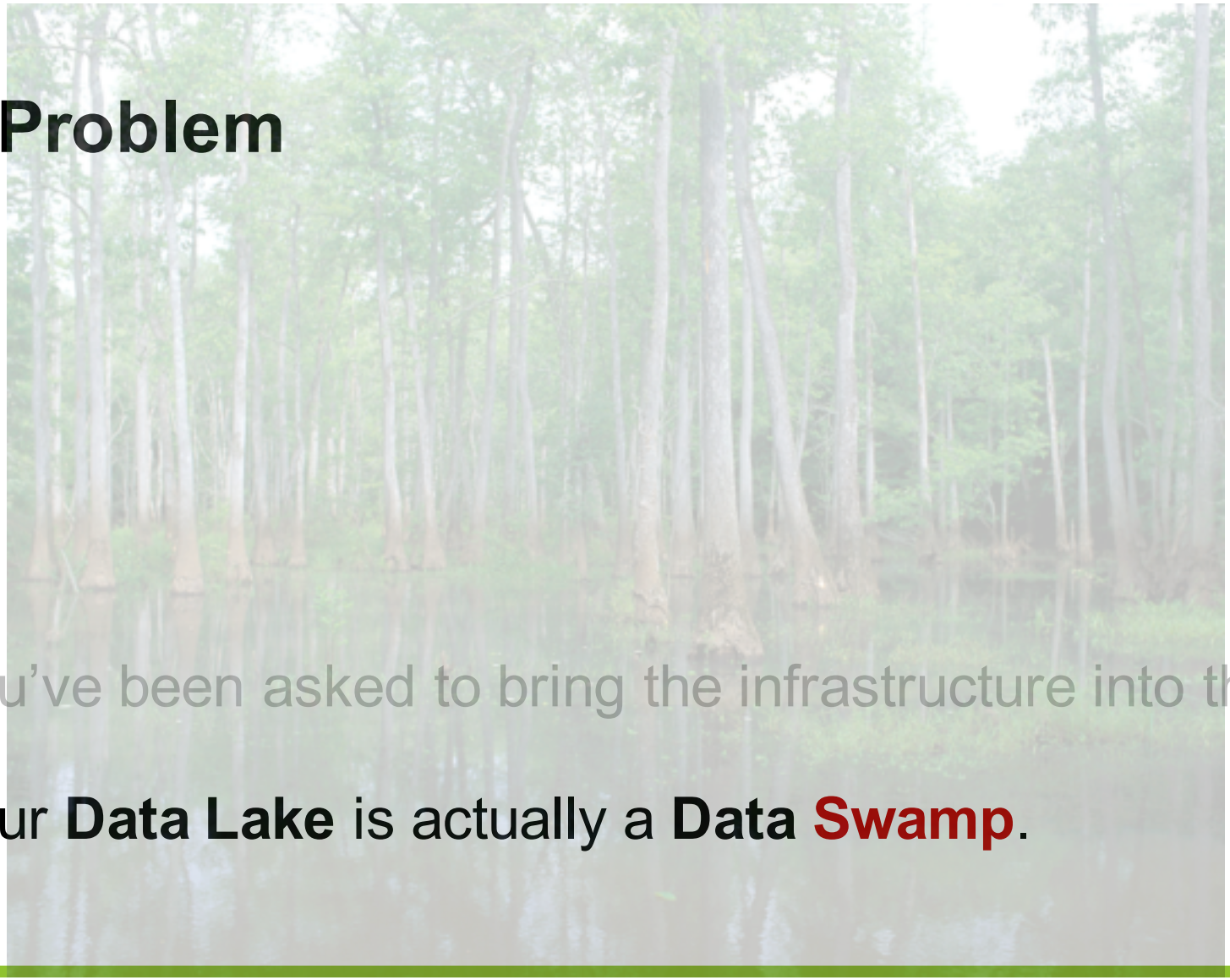
Industrializing DataScience Workflows

… and associated Idiosyncratic Operating Principles.

Sean Downes
Sr DataScientist @ Expedia, Inc.

The Problem



So you've been asked to bring the infrastructure into the cloud.

So your **Data Lake** is actually a **Data Swamp**.

Expedia

SPARK SUMMIT 2017

The Problem

Every **MicroService** has a **log**

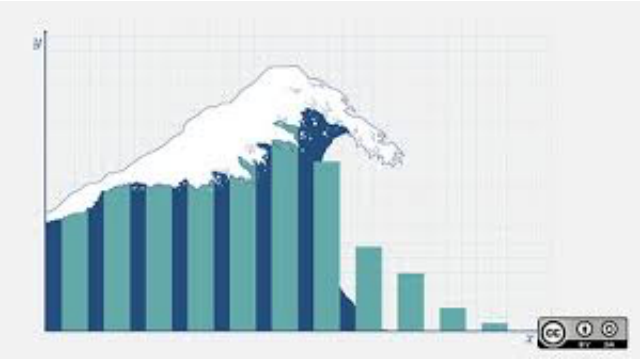
Login
Impressions
Clicks
Purchases

And you want to **A|B Test**

Every **Line of Business** has its own **Structure**

The Problem

Can you *please* turn..



Expedia

into

using

Qubole

Spark

amazon web services

hadoop

Microsoft Azure

Preview

Context / Disclaimer

Lightning Review of Data Platforms

(**idiosyncratic**) Organizing Principles

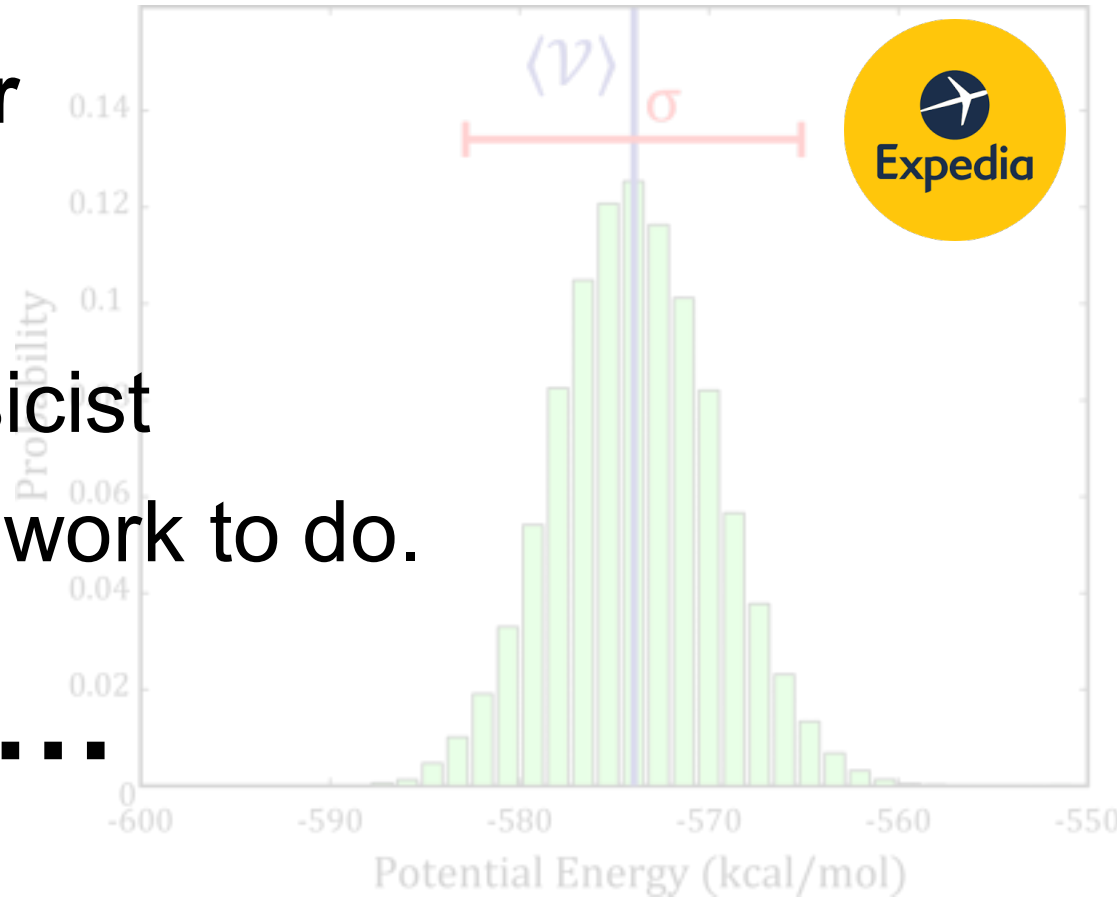
Context / Disclaimer

Academic

Theoretical Physicist

We've got some work to do.

So…



Expedia

SPARK SUMMIT 2017

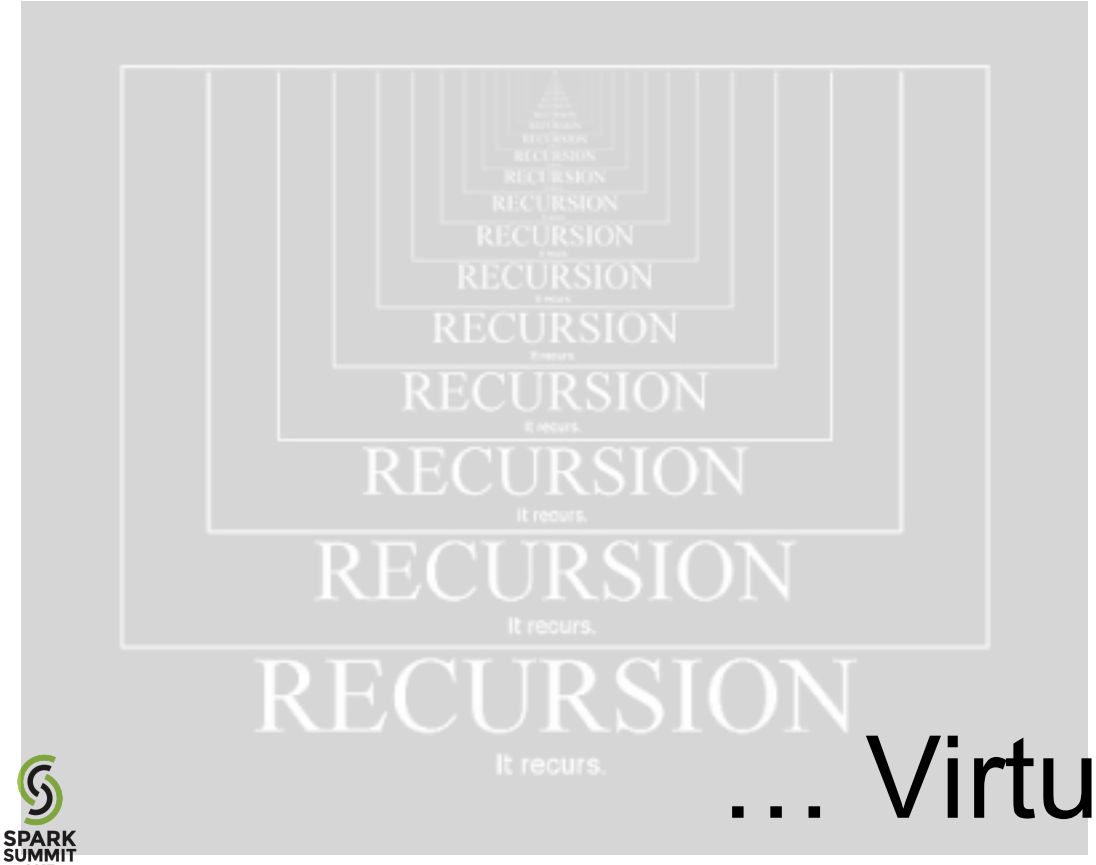
I'm ~~implicitly~~ assuming this talk will be **one** in an **ensemble of opinions**

Lightning Review of Data Platforms

Lightning Review of Data Platforms

… the Commerical Data Center…

Lightning Review of Data Platforms



1. Assign Tasks their own **virtual** hardware
2. Expend / Contract Resources by **demand**
3. Real-time HotSwapping
4. Software Updates Built In
5. Etc Etc Etc.

… Virtualized Everything

idiosyncratic Organizing Principles

iOP1) Clarity

iOP2) Engineers are not Data Scientists

iOP3) PMs are not Data Scientists

iOP4) Data Scientists are not Engineers

iOP5) Close the Data Loop

iOP1: Data Clarity

"big data, big noise"

Where is what data?

Who owns what field?

What is this this field?

Where did this field go?

Why is this field NULL?

PUBLISH THIS INTERNALLY!

iOP1: Data Clarity

"Expect Data Science"

Minibatch streaming into **nested** JSON?

O(10kB)?

GZip?

Spark

databricks™

O(50-500 MB)

Parquet.

Snappy.

Big Thanks to **Jason Pohl** @ DB!

And **Charles Pritchard**!

iOP2: Engineers are not Data Scientists

"why would you need to do that?"

Scratch Space

Cluster Bootstrap **Permissions**

Access S3 Buckets

Sandbox Clusters

Share Notebooks Across **Accounts**

We **DO NOT SPEAK IAM** Role/Anything

iOP2: Engineers are not Data Scientists

"why would you need to do that?"

if possible:

Write your **own** Pipelines.

else:

Explain Data Science.

iOP3: PMs are not Data Scientists

"you don't need that!"

Once upon a time in the **Flight DataLake**…

Only 10% of a Search Impression was recorded

Worse: It was only the **Cheapest** 10%

Many of the **bookings** where **not included** in this list!

iOP4: Data Scientists are Not Engineers

"we need to support models in @#&%? format"

Pick a **Robust Standard** and Stick to It. **jPMML**

If you're **big enough** to worry about this, you can **commit code**

Everybody Use **Git**. **Now**. Yes **You**.

Production Code Matters. **Format**. Document.

Pipelines Count as Production Code.

iOP5: Close that Data Loop

What is your data **doing**?

Big Data? Set up a **learning** problem.

New Data? Consider **Bandits**!

Empower by **Design**.

Empower by Design.

Thank You.

Contact information or call to action goes here.