Tony DiLoreto
SIM NYC - May 2019

The Pitfalls of the Cloud

...and how to prepare for them

Google Cloud

SIM
WHERE IT LEADERS CONNECT
NEW YORK METRO


About the
Speaker

1. **University at Buffalo**
Degrees in Mathematics
& Information Systems

2. **PWC**
Data Engineer; focus on
ETL & MDM

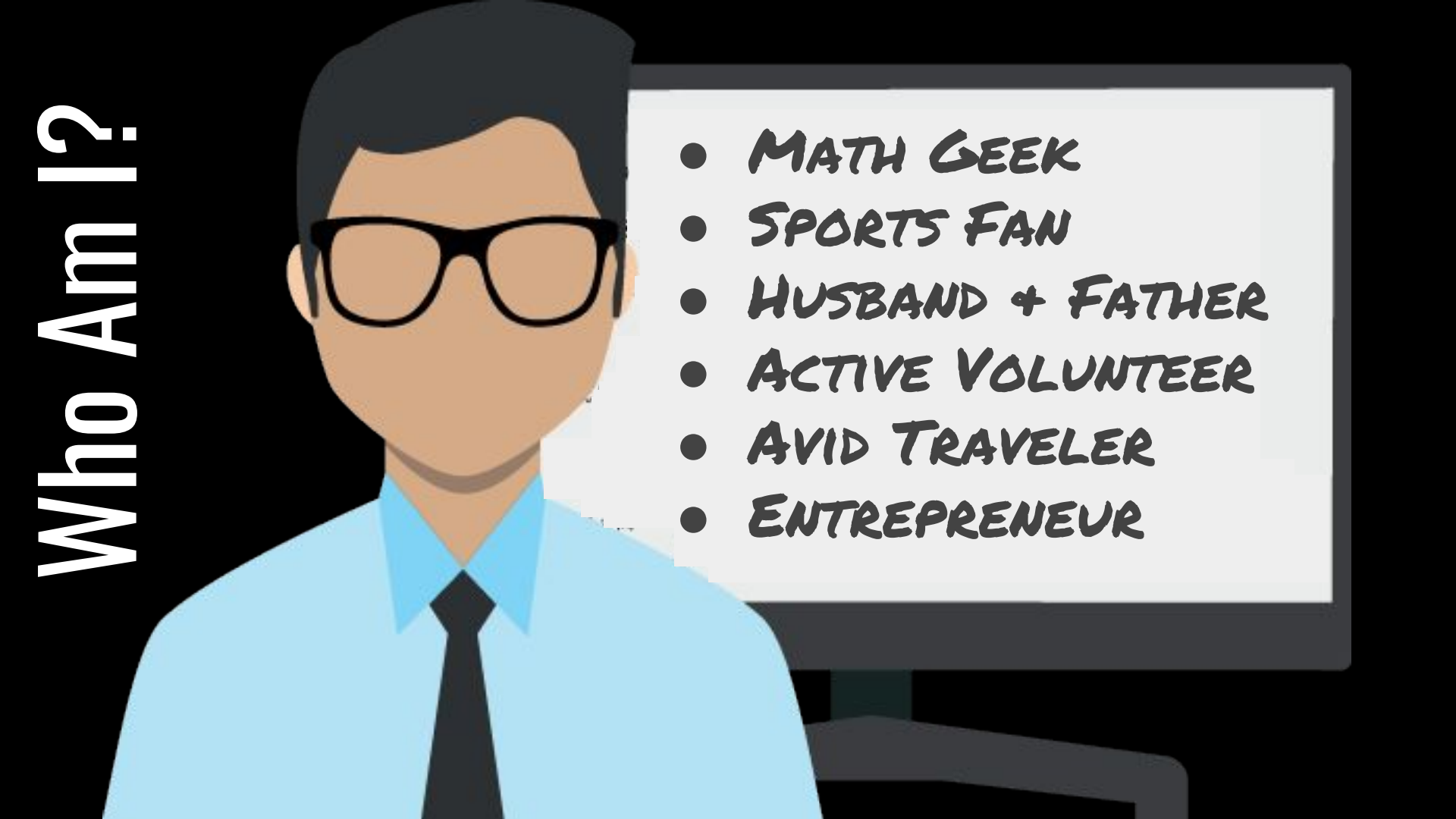
3. **Adobe**
Digital Marketing Solution
Architect / Data Scientist

4. **Northwestern U**
Master in Predictive Analytics /
Data Science

5. **Annalect (Omnicom)**
Director of Global Data
Platforms & Integrations

6. **Google**
Senior Cloud Consultant /
Architect

Who Am I?


Math Geek
Sports Fan
Husband & Father
Active Volunteer
Avid Traveler
Entrepreneur

Street Cred

400+

Petabytes of Data Migrated

35

Fortune 500 Companies Consulted

12

Years of Cloud & Digital Marketing Experience

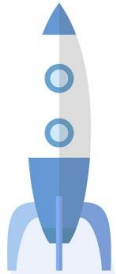
Street Cred

2

Year old baby girl


Why is Cloud
Different?

Cloud vs On-Prem Differences



When migrating to the cloud, there are a couple core differences IT administers must keep in mind:

- Virtualization of services
- Software-defined networks
- Elastic (ephemeral) workloads
- Containerization *everywhere*

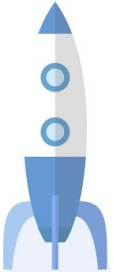
Google Cloud

Moving from Hardware to Software Defined

- As companies move away from managing their own infrastructure, pitfalls will shift from hardware-based to software-based.
- Platform vendors are now creating “managed” versions of their applications.
- This will change skill sets required to implement and debug, as well as hiring policies.

Google Cloud

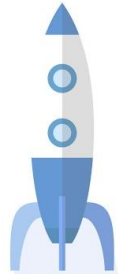
Shared Responsibility Model



- Both cloud providers **AND** companies using cloud technologies have a responsibility with respect to security.
- Understand your cloud provider's **SLA/SLOs** and if they align to your internal requirements.
- Follow the *principle of least privilege (PoLP)*

Google Cloud

Core Cloud Categories



Big Data & Analytics

Data storage, transformation and visualization

Application Development

Applications developed for and in the cloud (containers)

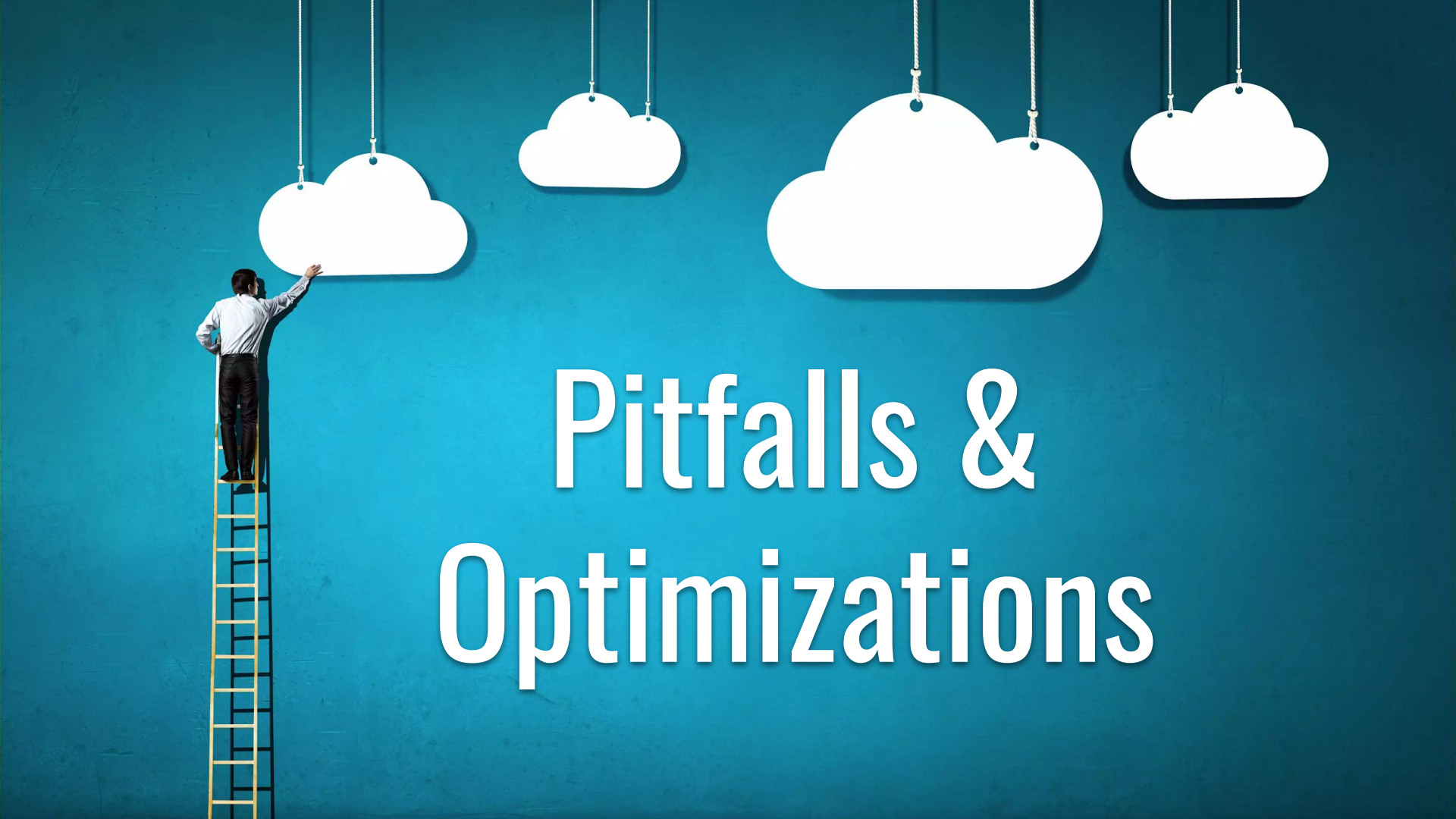
Infrastructure

Corporate connectivity and security management

Machine Learning

Advanced data science & statistical applications run on cloud infrastructure

Google Cloud


Pitfalls &
Optimizations

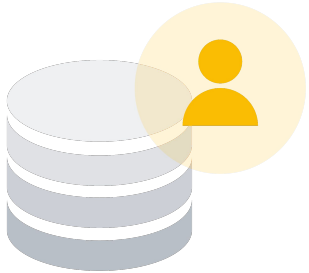
Data & Analytics

Google Cloud

Common Pitfalls

1. **Understand the “heat” of your data**
   Not all data needs to be copied to all regions in an on-demand state.

2. **Query the smallest dataset possible**
   Analysts want all data at all times, adding latency to queries and cost. Minimize to what is really needed.

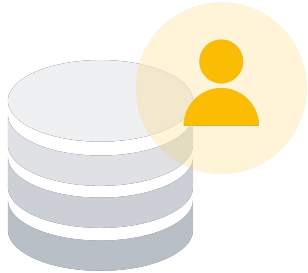


3. **Analysis Paralysis**
   Too much data = too many options = too much time = not enough value. Focus on 80/20.

Cost Optimizations

1. **Partition Your Data**
   Regardless of the storage engine, partitioning normally allows for faster, less expensive queries, especially when time-partitioning.

2. **Automate Object Storage Classification**
   All cloud providers offer ways to automatically reclassify the storage "class" of file stores, allowing you to save money on less-frequently accessed data.



3. **Focus on Business Needs, Not DB Capabilities**
   There's normally more than one way to skin a cat (store data), which offer trade-offs between latency, durability, availability, and cost.

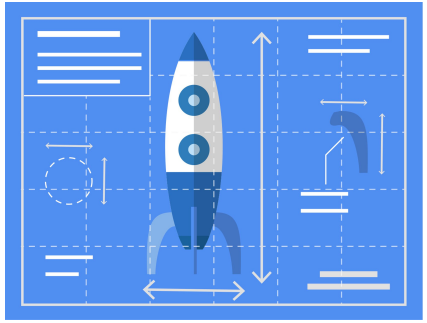
Infrastructure

Google Cloud

Common Pitfalls

1. **Under or Over-provisioning**
   Many cloud users like to either take an extremely restrictive or liberal approach to network provisioning. Start with what you need.

2. **Improperly Designed Networks**
   Network administrators sometimes fail to take advantage of or understand how cloud networking differs from physical on prem.



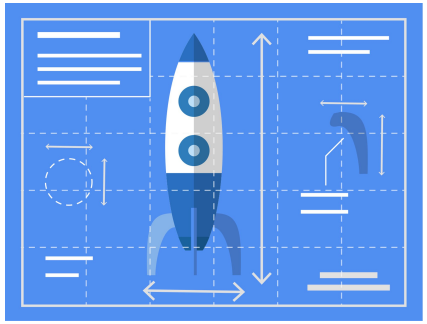
3. **Separation of Duties between Firewalls and Identity & Access Management (IAM)**
   Many network designers do not understand or take advantage of the differences between packet protection (Firewall) and user protection (IAM) rules.

Cost Optimizations

1. **Plan Plan Plan**
   Networking costs are normally focused on the labor, so a proper network plan can save $$$ in management costs.

2. **Sharing is not always Caring**
   Newer features like Shared VPCs (shared deployable network) allow for greater communication between devices, but could come at a cost.

3. **Keep in mind the Cost of Poor Security**
   Security is one of the largest potential costs to a business, both designing for or paying for it after a breach.



App Dev

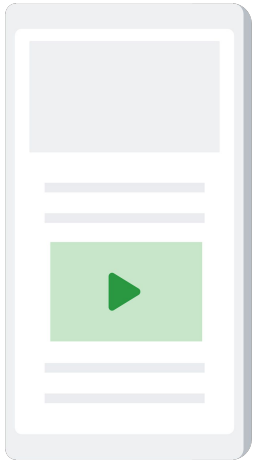
Google Cloud

Common Pitfalls

1. **No (or improperly planned) CI/CD pipeline**
   CI/CD pipelines do not need to be “logo soup”, but should ensure all major tasks (build, test, deploy) are setup.

2. **IaC FTW**
   Use the flexibility, elasticity, and ephemerality of the cloud to your advantage. Spin up/down when needed with minimal work.

3. **DevOps = Developers Best Friends**
   Failure to align the goals of the Development and DevOps teams slows down deployment cycles and costs real money.

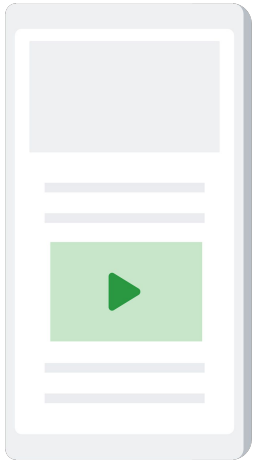


Cost Optimizations

1. **Monolith → (Micro) Services**
   Management costs alone for monolith applications adds technical debt, multiple points of failure and inflexibility. Use microservices where possible.

2. **Managed Services :-)**
   All cloud providers produce managed services of most every major platform type. Live to deploy, not manage.

3. **Containerization**
   Investing time in learning container management/orchestration software (i.e. Kubernetes) will save your company $$ in the long run.



Machine Learning

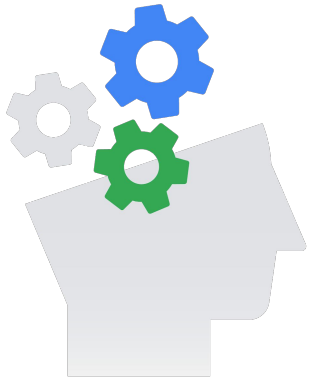
Google Cloud

Common Pitfalls

1. **Jumping to ML**
   Every company wants to go straight to ML without laying the data foundation (cleansing, storage, preparation).

2. **Trying to Overcomplicate Solutions**
   90% of Machine Learning problems can (initially) be solved by basic regression/classification models. Pick the right problem & start simple.

3. **Reinventing the Wheel**
   Most cloud providers (especially Google) create ready to use models for common AI tasks. If image recognition is not core to the application, start with pre-trained models instead of building from scratch.

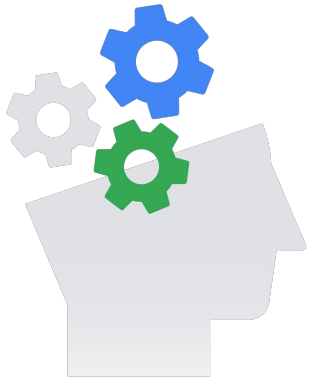


Cost Optimizations

1. **Sample! Sample! Sample!**
   Data acquisition costs are said to be the largest chunk of Machine Learning. Sample first, and cultivate when needed.

2. **80/20 Rule is King**
   Many Data Scientists will spend untold hours trying to improve a model's performance a fraction of a %. Is the equity worth the sweat?

3. **Corporate Model Sharing**
   Because so many problems can be solved with similar approaches, companies are now creating "model sharing" platforms to allow for quicker iteration across the org.



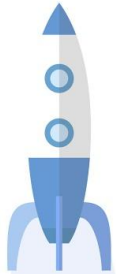

Final Thoughts

Prepare for the Cloud

1. **Do Your Homework**
   Not all data needs to be copied to all regions in an on-demand state.

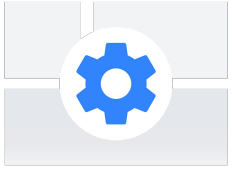
2. **Understand Differences**
   Analysts want all data at all times, adding latency to queries and cost. Minimize to what is really needed.

3. **Test it Out!**
   Review insights from the workshop and next steps for proceeding with a cloud solution.



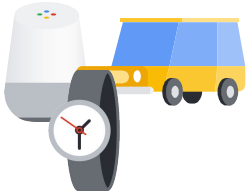
Google Cloud

What's Next?



Hardware Commoditization and one-click Infrastructure Deployments

AI as a Service (AIaaS)



Total Lifestyle Connectivity (Beyond IoT)