Managing Data Science | Lessons from the Field

Mac Steele

Director of Product | Domino Data Lab

mac@dominodatalab.com

@macsteele

DOMINO

aws

PARIVEDA

What You'll Learn Today

GOALS
What is the bar for data science teams

PITFALLS
What are common data science struggles

DIAGNOSES
Why so many of our efforts fail to deliver value

RECOMMENDATIONS
How to address these struggles with best practices

Lots of Legitimate Promises

Allstate

Saved $40M

In claims with predictive analytics

---

amazon.com

35% of Sales

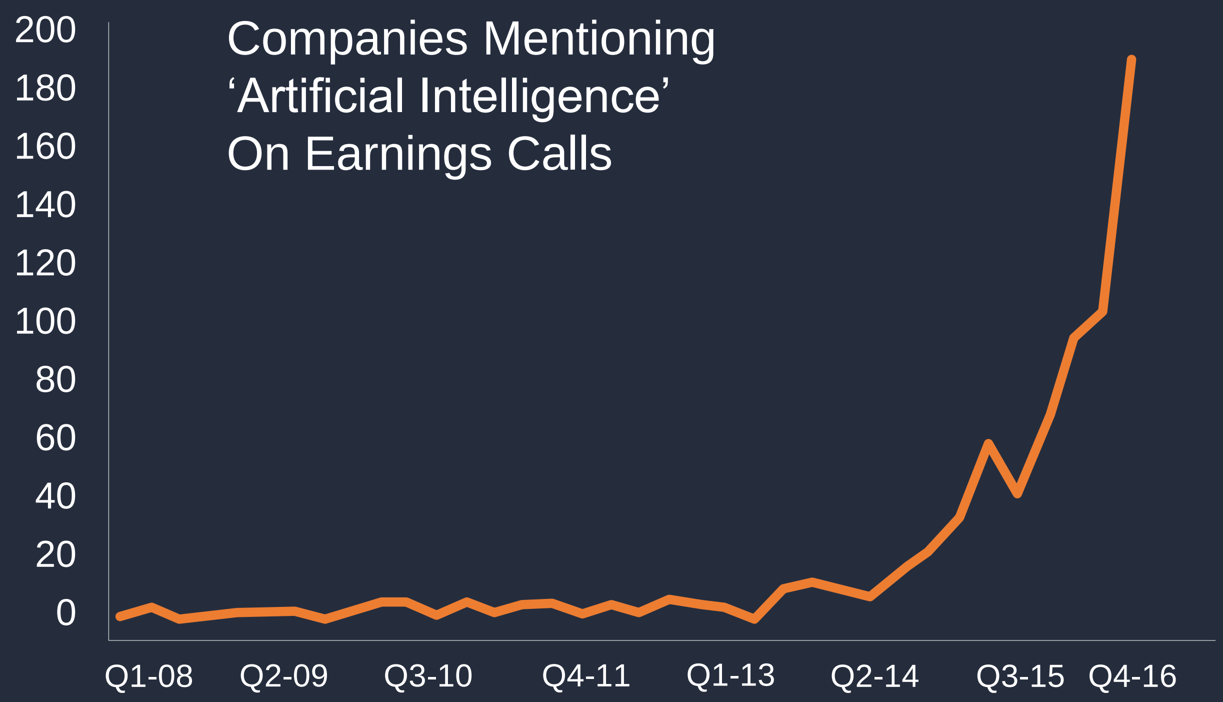
Come from product recommendations

NEW YORK STATE

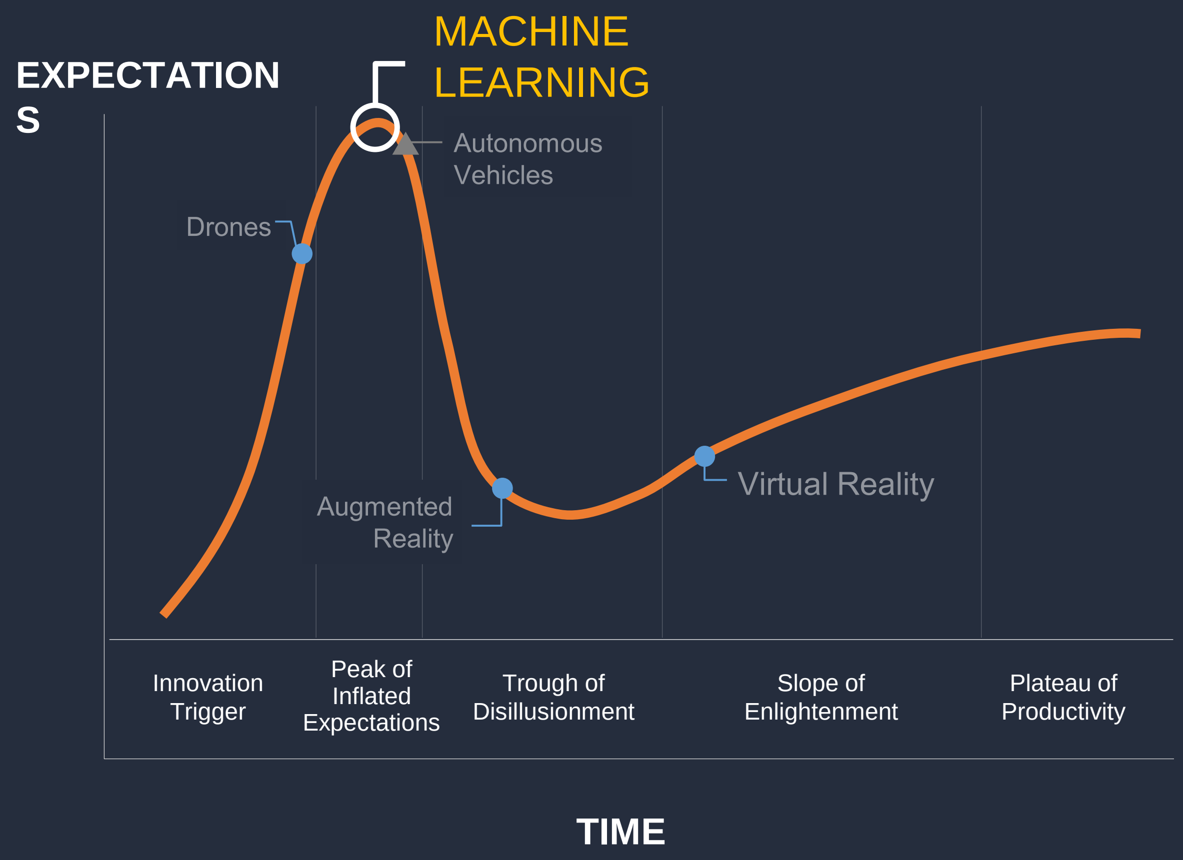
Saved $450M

By detecting fraudulent tax returns

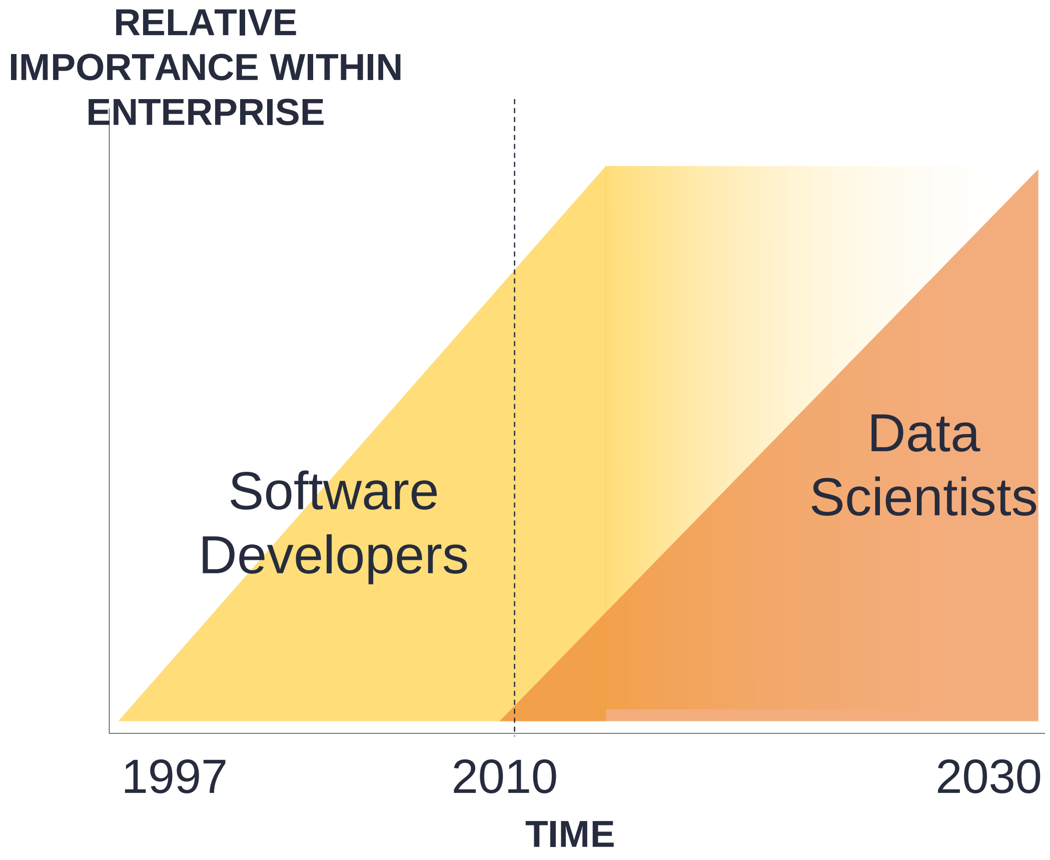
Lots of Hype



Lots of Risk of Disappointment



This Sounds Eerily Familiar



What is the Goal?

Measurable
Your “quality” indicator.

Reliable
Your “hit rate.”

Scalable
Your “throughput.”

DATA SCIENCE PITFALLS

I SOLVED THE PROBLEM BUT…

It was the
wrong problem

Oops, already
solved by
someone else

Have the wrong
tools for this problem

Too slow for it
to matter

Solved the
wrong way

Results used
Wrong way

World changes
while solving
problem

Problems mulitply,
can’t tackle all
at once

DIAGNOSES


CYTOSINE
GUANINE
ADENINE
THYMINE
Cardiovascular diseases
Pulmonary disease
Diseases of the digestive system
Liver disease
Diseases of the musculoskeletal system
Neurological diseases

Data Science is Different from Software Development

- Research versus development focus
- No answer is a valid answer
- Traditional testing is insufficient given non-deterministic nature
- No generally accepted process metrics (e.g. story points)
- Data must be tracked

Forget About Other Stakeholders in the Process

For Data Science Managers

- Accelerate project delivery through reuse, knowledge management
- Mitigate key-man risk / accelerate onboarding
- Hire & retain top talent

For Business Leaders

- Understand real-world impact
- Reliable, predictable insights
- Minimize change to existing workflows

For Data Scientists

Access powerful infrastructure & preferred tools

For IT Leaders

- Ensure stability & security
- Leverage existing infrastructure
- Minimize operational burden

Fixation on Tools at the Expense of People and Process

Moonshot vs. Laps Around the Track

- Perfection as enemy of shipped
- Muddle “pure research” and “applied templates”

Disconnected from the Business

- Little familiarity with practical business constraints
- Limited ability to drive *adoption*

Missing Some Key Personnel Muscles

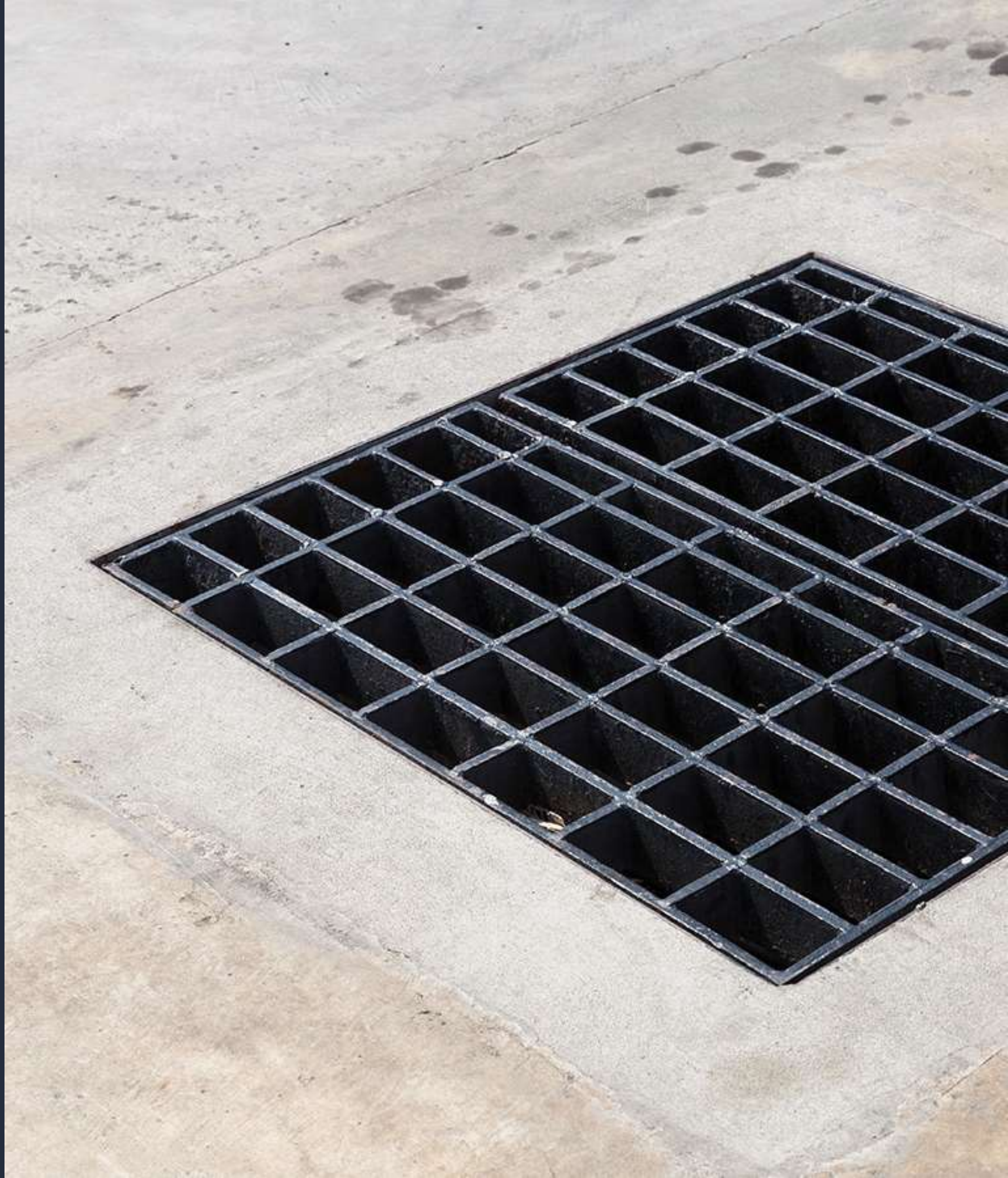
- The full stack data scientist is a myth
- Gap in ”soft” skills training

Artisan Thinking vs. Modular System Thinking

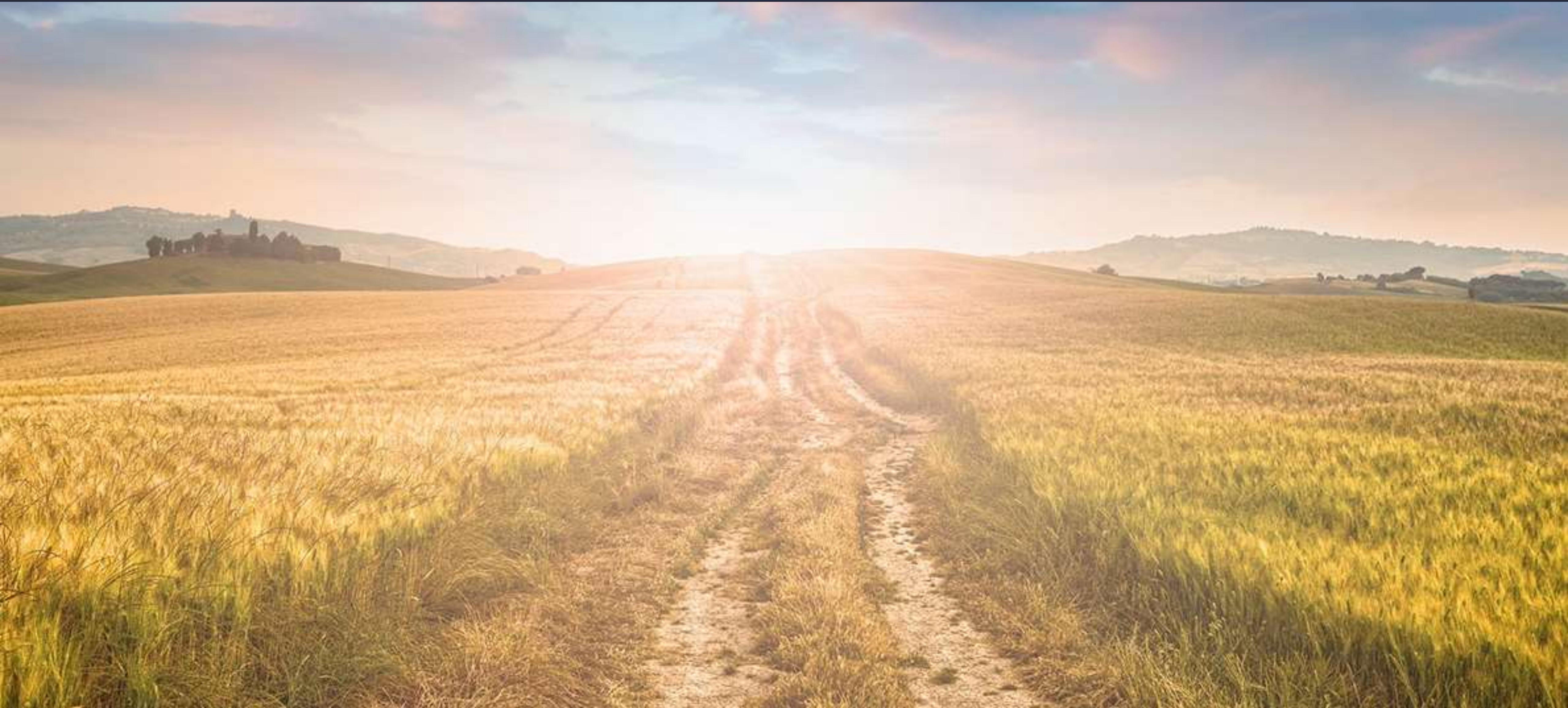
- Limited culture of re-use and compounding
- Not planning for future iterations (e.g., no reproducibility / documentation)

Bad Incentive Structures

- Key responsibilities fall between gaps
- Significant information loss in project transitions



RECOMMENDATIONS



Best Practices Take Many Forms

Process

Both a *single* project and *portfolio* of projects

People

Types of capabilities and org design

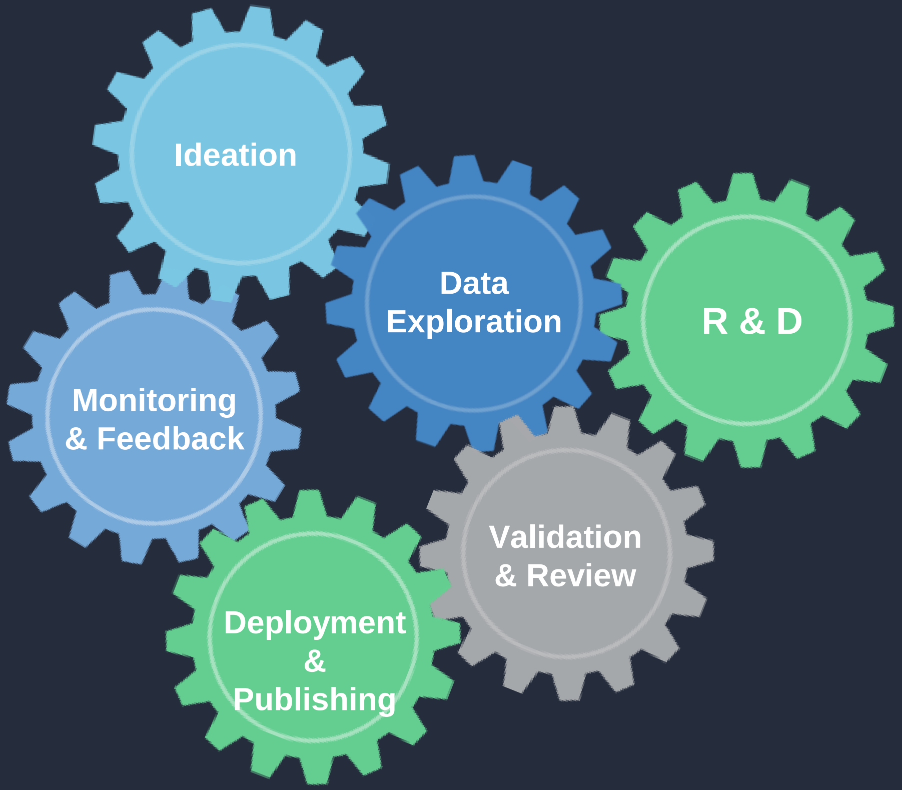
Technology

Flexible infrastructure and tooling without the wild west

Data science system at many levels

Single Project

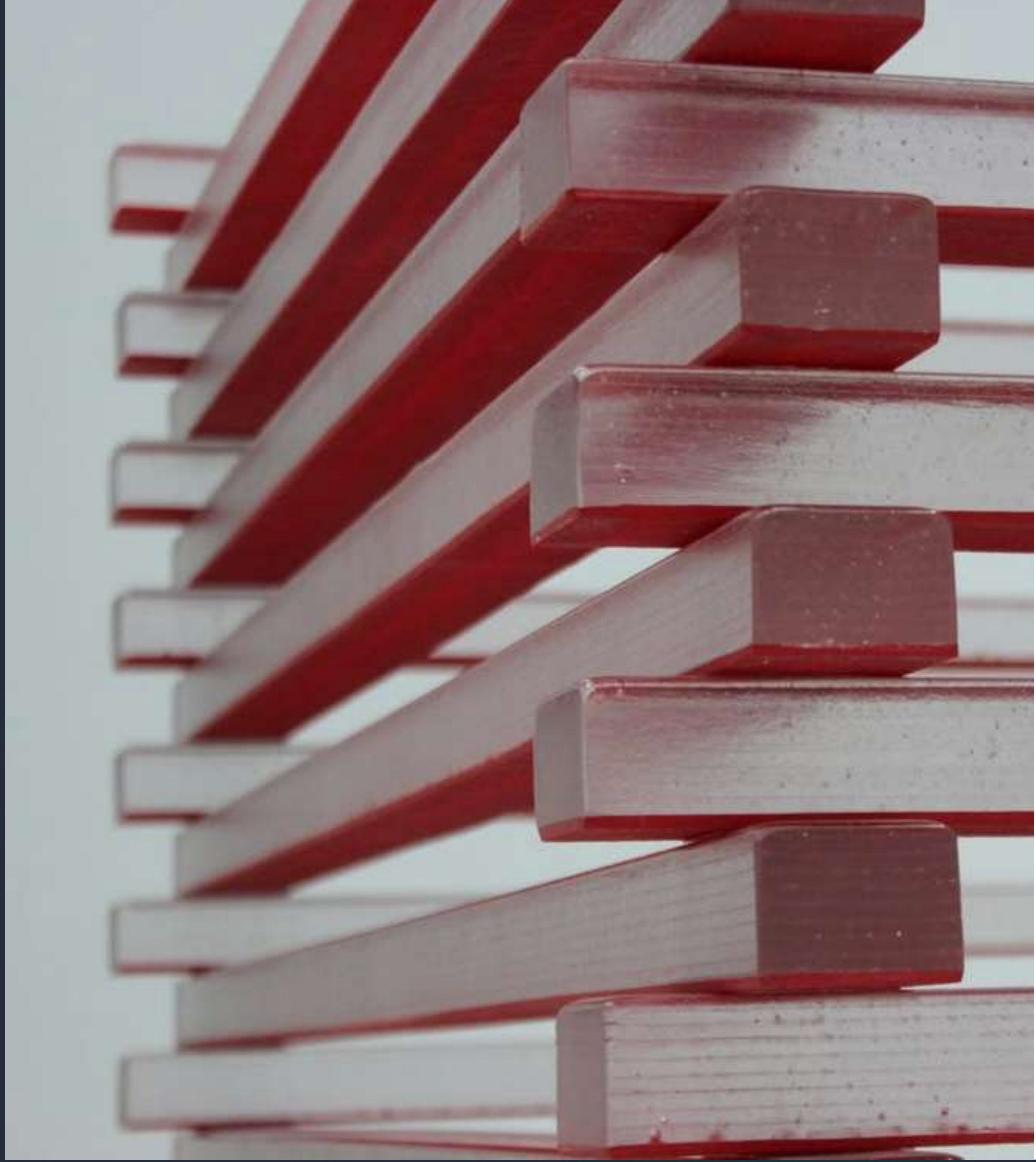
Single Step



Portfolio of Projects

Managing the lifecycle

- Expect and embrace iteration
- Enable compounding collaboration
- Ensure auditability and reproducibility, even if you’re not regulated (yet)



Ideation

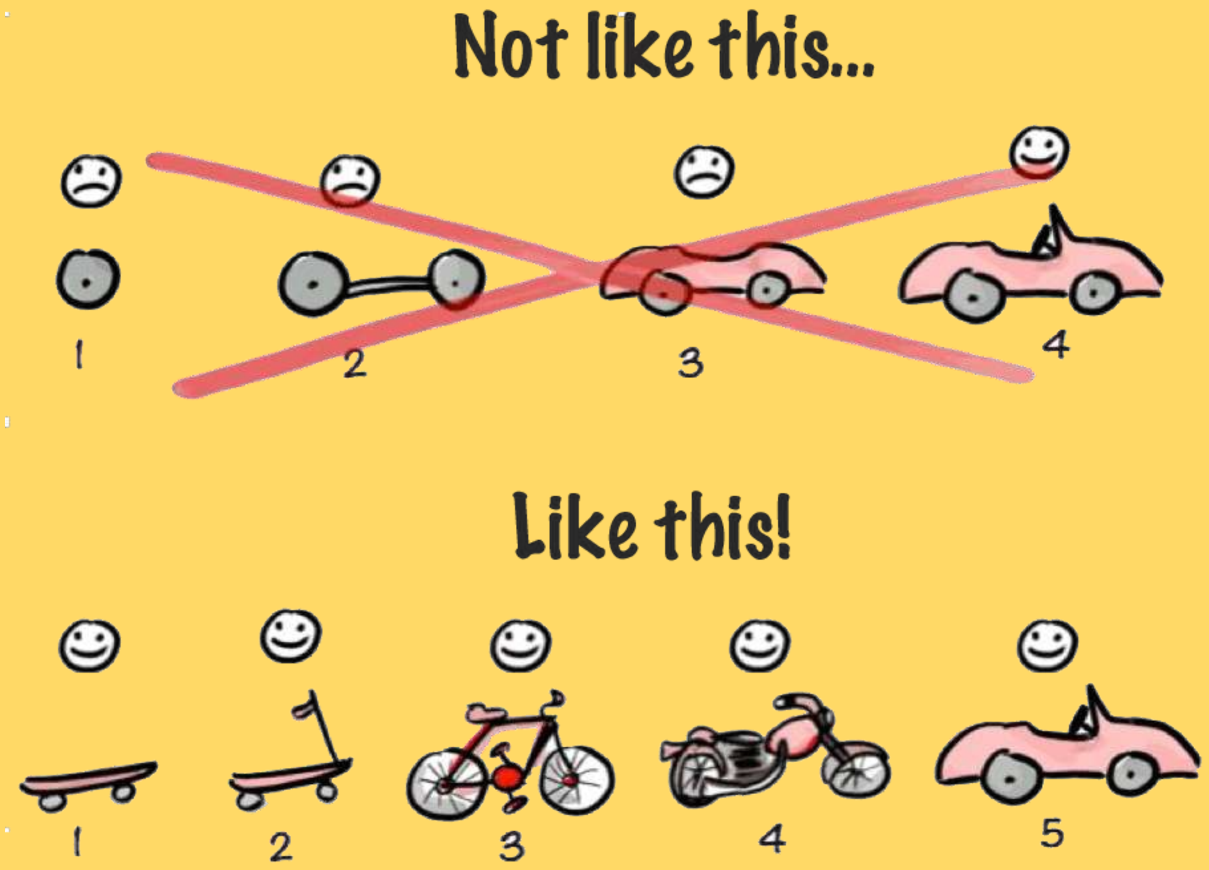
- Problem first, not data first
- Practice and master order of magnitude ROI math
- Maintain repo of past work
- Create and enforce templates for MRDs
- Maintain a stakeholder-driven backlog

Artifact Selection

- Leverage rapid prototyping and design sprint methodology
- Create multiple mock-ups of different deliverable types
- Consider creating synthetic data with baseline models

Research & Development

- Establish standard software configurations, but give flexibility to experiment
- Abstract away compute provisioning
- Build simple models first
- Set a cadence for delivering insights
- Ensure business KPI tracked consistently over time



Validation

- More than just code review, get stakeholder and IT sign-off
- Ensure reproducibility and clear lineage
- Use automated validation checks to support human inspection
- Preserve results (even nulls) to central repo

WHAT INFLUENCES A RESULT?

Results

Depend on

The statistical analyses selected

Depend on

The R scripts that implemented the analyses

Depend on

The R libraries that implement the statistical functions

Depend on

The C libraries that perform the mathematical computations

Depend on

The operating system running the computational framework

Depend on

Reduced data

Depend on

Scripts that reduce the data

Depend on

Raw data

Delivery

- Support for many deliverable artifacts (reports, dashboards, apps, batch APIs, real-time APIs)
- Define a promote-to-production workflow
- Flag upstream and downstream dependencies

Monitoring

- Build ROI testing into all major deliverables
- Require monitoring plans before considering “done”
- Integrate with tools where people spend most of their time (e.g., email / Slack)
- Anticipate risk and change management burdens

Keeping all the balls in the air

- Measure everything, including yourself
- Focus on reducing time to iterate
- Socialize aggregate portfolio impact

The many hats of data science

ROLE	PRIORITIES	PITTFALLS WITHOUT THEM
Data Scientist	Generating and communicating insights, understanding the strengths and risks	Naïve or low power insights
Data Infrastructure Engineer	Building scalable pipelines and infrastructure that make it possible to do the higher levels of needs.	Insight generation is slow, because DS is spending their time doing infrastructure work
Data Product Manager	Articulating the business problem, translating to day-to-day work, ensuring ongoing engagement.	Projects miss the mark, don't translate into tangible business value
Business Stakeholder	Vetting the priortization and ROI, providing ongoing feedback	ROI decisions aren't made sensibly, not knowing when to pull the plug
Data Storyteller	Creating engaging visual and narrative journeys for analytical solutions	Low engagement and adoption from end users

Organizational Design Dilemmas

- False centralization / decentralization dichotomy
- Most evolve as they scale and as business demands shift
- Technology can help bridge the gap

	CENTRALIZATION	DECENTRALIZATION
Pros	• Community and mentorship • easier transparency for managers and IT • More passive technical knowledge sharing	• Deeper understanding of business processes and priorities • Easier change management
Cons	• Isolation on data science island • Loss of credibility with business • Frustrated data scientists	• Less technical knowledge compounding • Harder to codify best practices • Risk of shadow IT

What We Covered Today

GOALS

What is the bar for data science teams

PITFALLS

What are common data science struggles

DIAGNOSES

Why so many of our efforts fail to deliver value

RECOMMENDATIONS

How to address these struggles with best practices

QUESTIONS?

Check out dominodatalab.com or find us in the AWS Marketplace

WELCOME TO THE

DATA SCIENCE INNOVATION SUMMIT

9:15	Coffee & Light Breakfast
9:30	Welcome, Carlos Escapa, AWS
9:40	Joe Spisak, AWS
10:15	Mac Steele, Domino Data Lab
11:15	Break
11:30	Sean Beard, Pariveda
12:30	Lunch
1:15	Panel Discussion/Q&A
2:00	Connect with others

DOMINO aws PARIVEDA