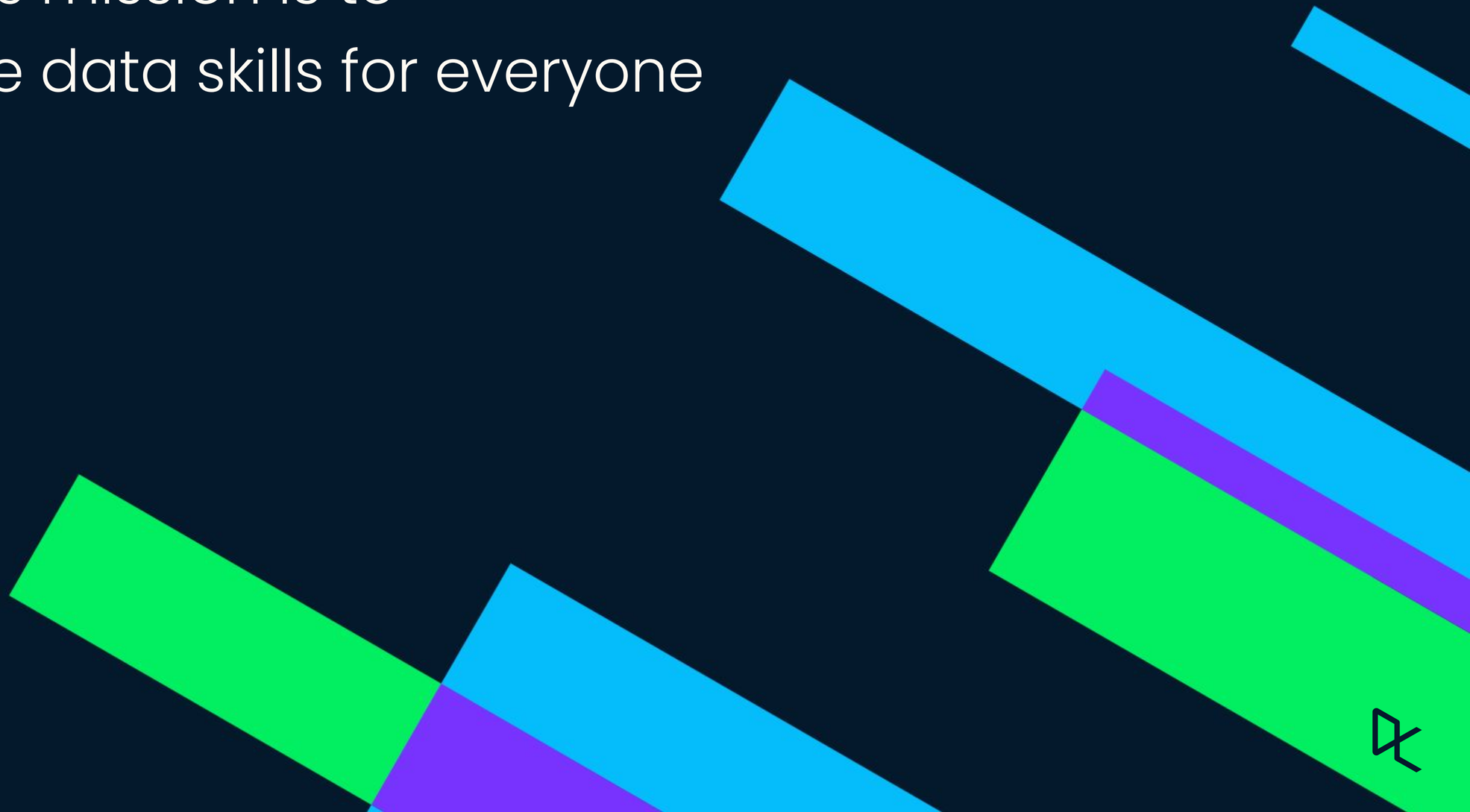
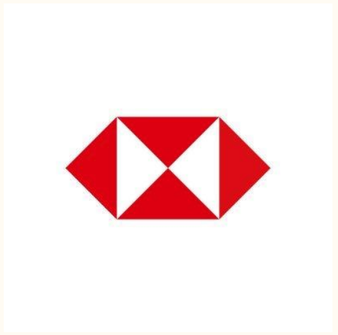
Manage Data Science Projects Effectively

September 23, 2021

Our Mission

DataCamp's mission is to
democratize data skills for everyone





HSBC

Mercedes-Benz

PayPal

Colgate-Palmolive

Google

Deloitte

Uber

eBay

T-Mobile

Trusted by over 2,000 data-driven companies

DataCamp is transforming the way businesses prepare their employees for the future of work

Brian Campbell

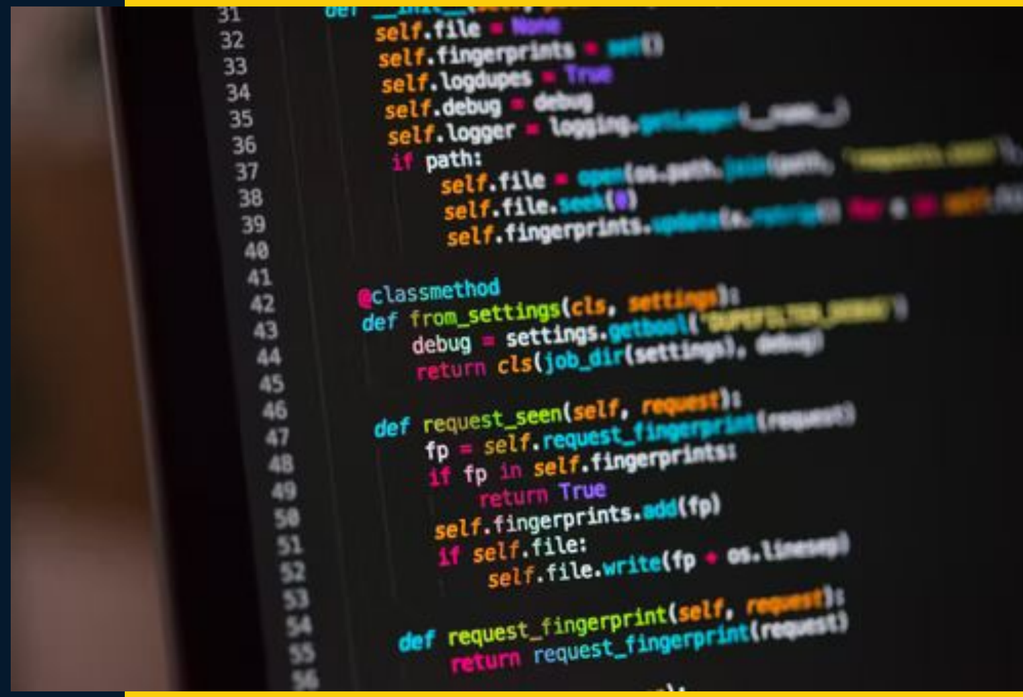
Brian Campbell

Engineering Manager
Lucid Software

- Former tech lead for “Data Infrastructure” team at Lucid
- Led data ingestion and deployment for 10 data science projects
- Currently manage Data Infrastructure and Data Science

What data scientists think they'll do

- ✓ Clean interesting data
- ✓ Build cool models
- ✓ Help get the model in customers hands



What data scientists think they'll do

- ✓ Find useful datasets
- ✓ Clean that data
- ✓ Make and share reports based on that data
- ✓ Build models
- ✓ Build deployment pipelines
- ✓ Integrate the model with a product
- ✓ Keep everyone up-to-date on all of the above

It’s too much for one person or even one team to do well.
A successful project requires collaboration

Agenda

1. Finding the right collaborators
2. Working with collaborators
3. Helping your team collaborate
4. Real life example
5. Closing notes and Q&A

1

Finding the right collaborators

Collaborate with a problem expert

Source problems from domain experts

Find people that know the business model and domain well and learn about their challenges.

Identify projects fit for data scientists

You and your team know your strengths best. Find how you can contribute to important problems

Befriend the problem experts

For each project, identify who understand and cares about the problem most. They will be your most valuable collaborator

Your data is the wrong place to look for projects

Collaborate with a problem expert

Source problems from domain experts

Find people that know the business model and domain well and learn about their challenges.

Identify projects fit for data scientists

You and your team know your strengths best. Find how you can contribute to important problems

Befriend the problem experts

For each project, identify who understand and cares about the problem most. They will be your most valuable collaborator

From problem to requirements

1 Understand your metrics

No model will operate with perfect accuracy. Work with your Problem Expert to understand how good a solution needs to be to be useful

2 Understand the domain's metrics

Throughput and latency can be just as important as accuracy. A good result an hour late isn't useful. Make sure you understand the all the requirements.

Collaborate to get data

1 Understand the data you need and have

Decide what data would be useful for solving this problem. Then review what your team has access to and find the gaps

2 Find if and where the data you need exists

Some of what you need may live in a 3rd-party or silo. Some of it may not be collected yet. Find out who knows

3 Find if and where the data you need exists

Enlist the help of whoever can help get access to the data or can start collecting the data you'll need

Collaborate to implement

1 Let people know about your project early

Once you have a rough idea of a solution, or have narrowed down to a few options, start involving people that will help implement. Don't forget, they have their own roadmaps and priorities

2 Think about how people will use your work

Is it going to be a feature in a product? Is it going to be part of an API? If you need to integrate with something that already exists, you'll want to work with the people responsible for it

3 Don't forget about infrastructure

Data science projects often require unique infrastructure to run. Make sure whoever owns infrastructure at your organization is aware well ahead of time.

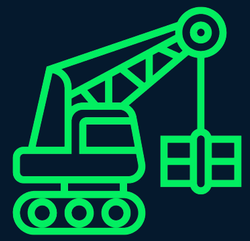
Who do we collaborate with?

Problem Experts

The person most affected by the problem. Key stakeholder through the whole process

Data Experts

Know how to get the data you need and can help you understand it



Implementation Experts

Have the skills to successfully get your solution in to the world

2

Working with collaborators

Organizations with highly-effective communication see 80% of their projects reach their goals, compared to 52% for organizations with minimally effective communication

The High Cost of Low Performance: The Essential Role of Communication by PMI

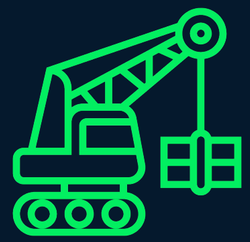
Setting clear expectations

Problem Experts

Frequent check-ins throughout your project

Data Experts

Frequent check-ins at the beginning then occasional updates



Implementation Experts

Occasional updates then frequent check-ins near deployment

These expectations apply to you as well as your partners

Timelines are hard

1 Timelines are difficult but important

Your partners need to know when to act. The timelines will be rough, but should give people the right idea

2 Use milestones to communicate needs

Be clear about what phases in the project you'll need help and what phase of the project you're currently in so collaborators will know they'll be involved soon even if they don't know exactly when

3 Use milestones to update timelines

As the project progresses, you should have a clearer picture of how long the different phases will take. Keep updating the timeline so your partners have the the clearest view possible

An example data science timeline

Problem Expert

Gather requirements for project

2W

Data Expert

Collect data for project

2W

Clean and analyze data

2W

Build model

2W

Implementation Expert

Integrate model with product

2W

Implementation Expert

Go live

You tell your implementation partners that you'll need their help in 2 months

Problem Expert

Data Expert

Implementation Expert

Implementation Expert

Gather requirements for project

2W

Collect data for project

2W

Clean and analyze data

2W

Build model

2W

Integrate model with product

2W

Go live

Then things start to slip..

Problem Expert

Gather requirements for project

2W

Data Expert

Collect data for project

4W

Clean and analyze data

2W

Build model

4W

Implementation Expert

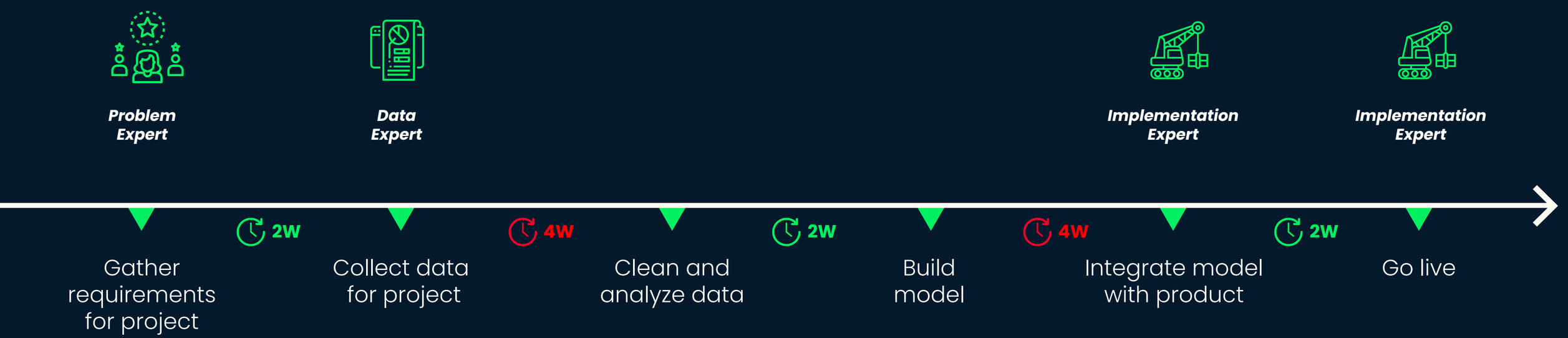
Integrate model with product

2W

Implementation Expert

Go live

Implementation shows up a month before you're ready and now their roadmap is ruined



Keep everyone up to date

Keep track of your timelines

Tell all partners the revised timeline at each milestone

Keep your closest partners updated

When a partner is expected to help with the next milestone, update them on timeline changes frequently

3

Helping your team collaborate

Find potential problems fast by learning fast

Work with baseline models

1 Build a baseline model

Use what data you have at the start, heuristics, or even just random data to act as a baseline

2 Use that baseline

Iterate on potential models quickly by having a baseline to compare against

3 Learn from the baseline

By understanding the gap between your baseline and your requirement, you can prioritize the most valuable data and solutions

Prototypes

1 Build a prototype

Use the baseline model to prototype your solution

2 Use that prototype

Problem experts can see how well the solution matches the problem and refine requirements

3 Learn from the prototype

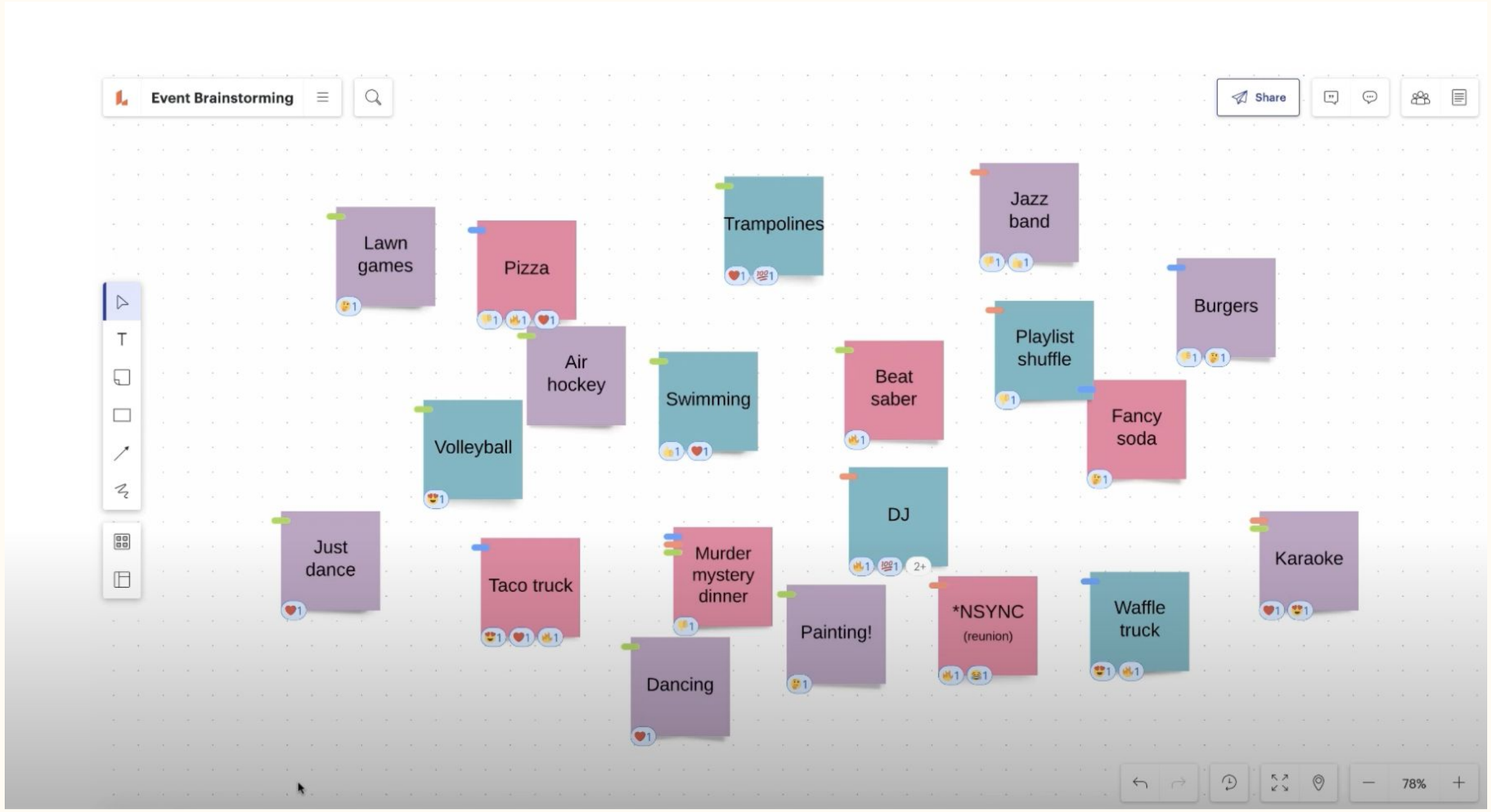
Implementation partners can see how a particular solution will perform and reality and iterate or recommend changes

Agile methods make timelines more complicated, but produce better results faster

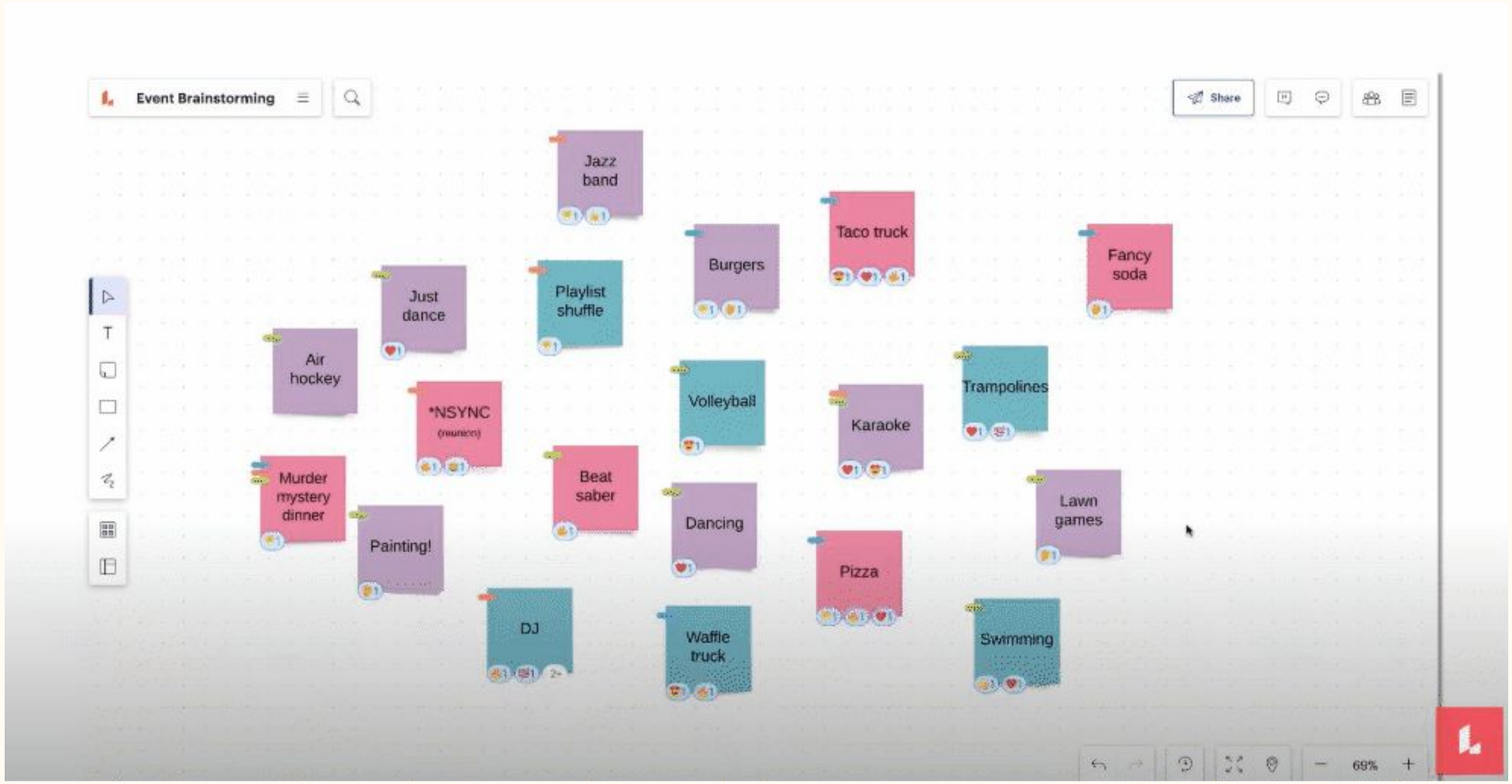
4

A real-life example

Cluster related ideas from a brainstorm session



Cluster related ideas from a brainstorm session

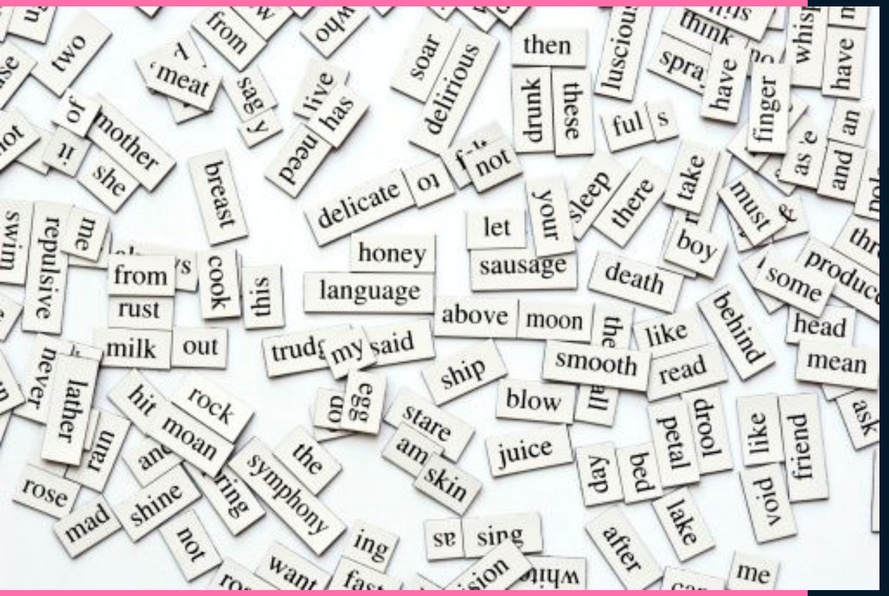


Brainstorm lots of ideas by throwing sticky notes on a Lucidspark board and get similar ideas gathered together in to groups

From problem to requirements

- ✓ Problem experts were product development team
- ✓ Roadmaps aligned
- ✓ Implementation partners spread between data science and product teams

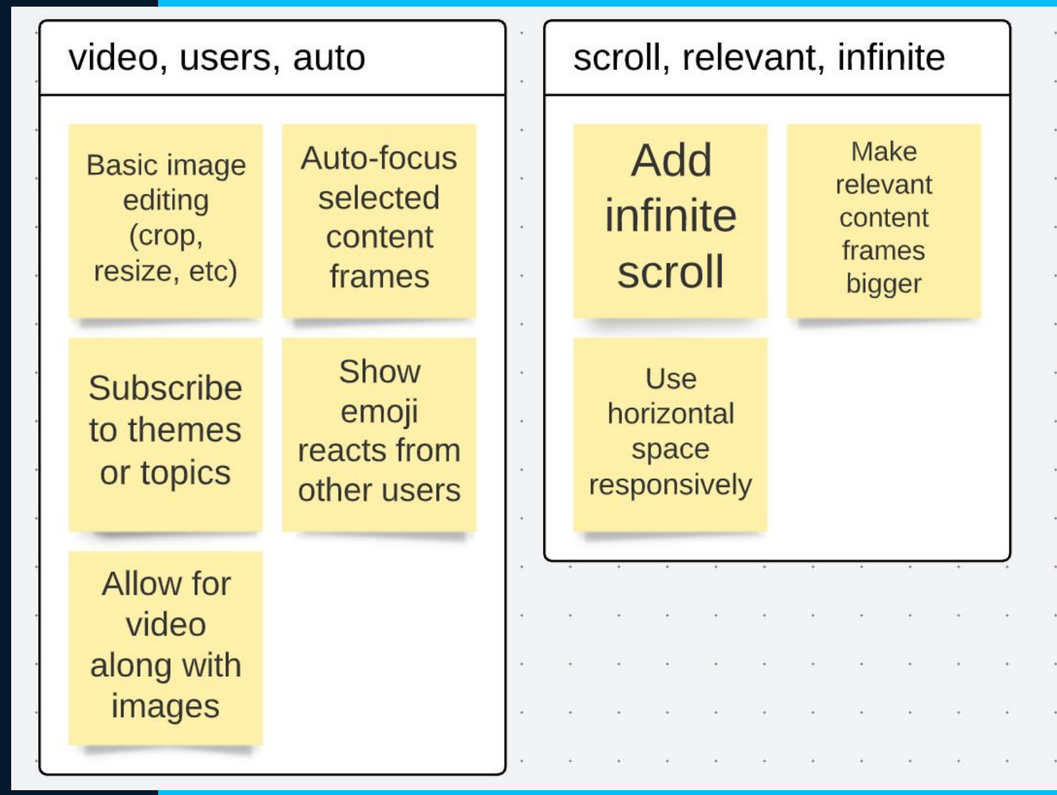
Getting data



- ✓ Data requirements were too complex
- ✓ ETL team became data partner
- ✓ Usage of standard corpora

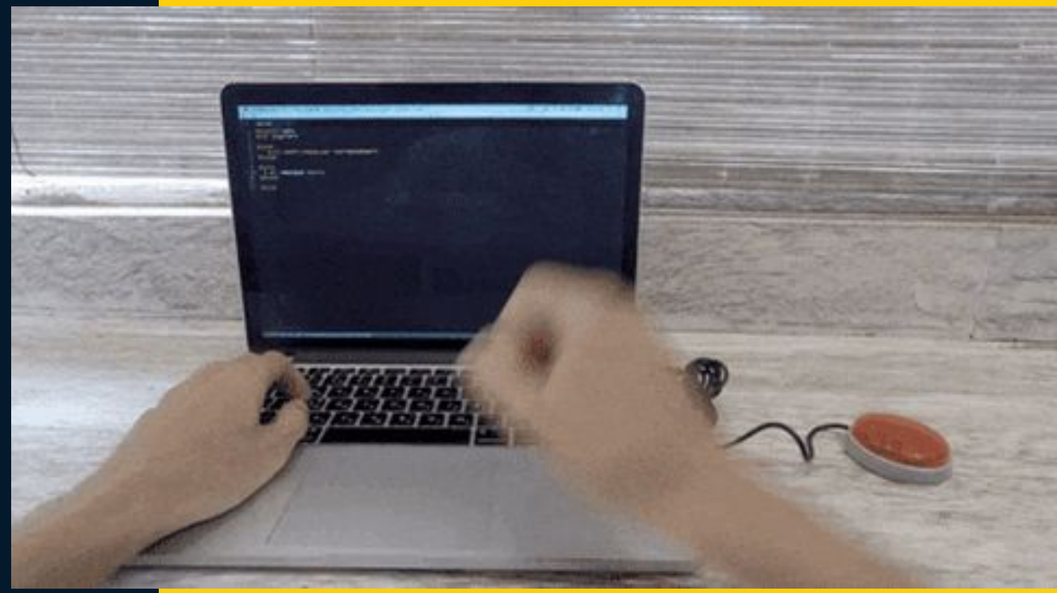
Building a prototype

- ✓ Built requirements for inputs and outputs
- ✓ Build a baseline based on randomness and iterated from there
- ✓ Plug-in model to product experience



Deployment time

- ✓ AWS Lambda to use for deployment
- ✓ Low maintenance and low budget option
- ✓ Infrastructure team set up instance, and stopped paying attention



Deployment time
Unknown unknowns

- ✓ NLP libraries used were too big for AWS Lambda
- ✓ Real-life usage different than testing use-cases
- ✓ Performance was misaligned with desired product experience

Iteration and re-launch

- ✓ Project was moved to internal alpha
- ✓ Firm but realistic expectations for speed and stability
- ✓ Collaboration with partners to ensure every goal was met
- ✓ Faster model adopted, and deployment method was changed

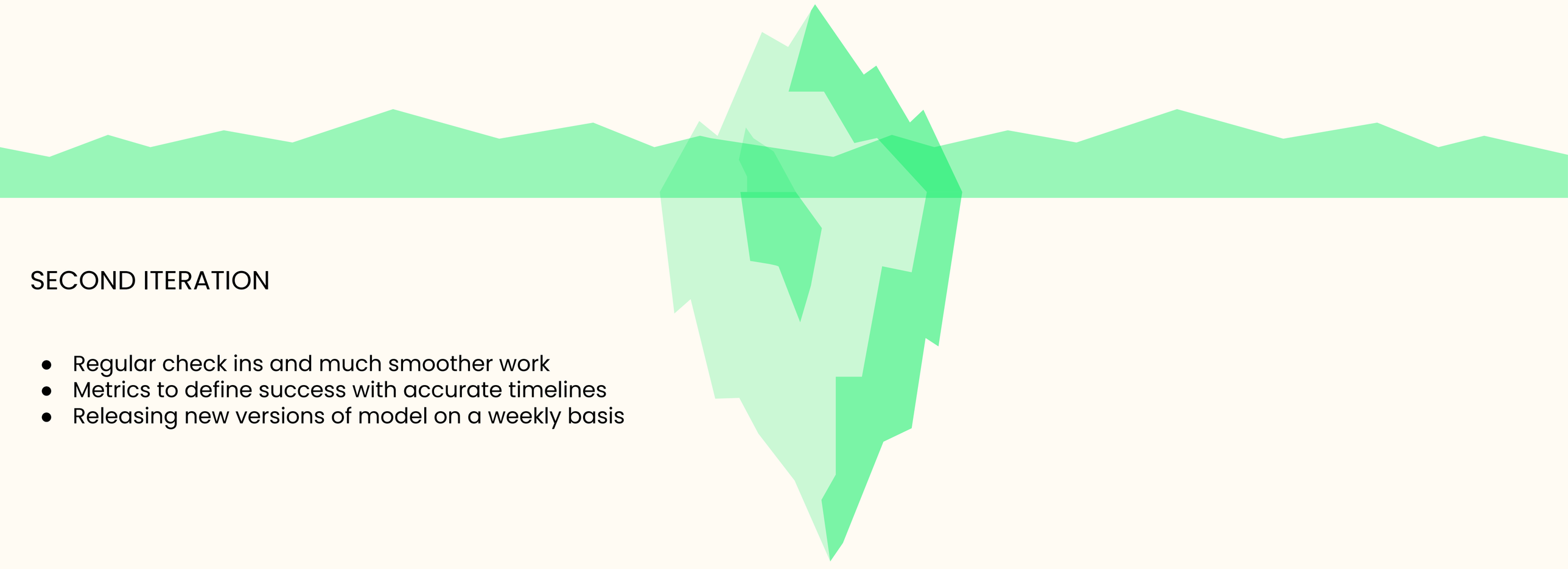
5

Closing notes and Q&A

First iteration vs second iteration

FIRST ITERATION

- Solve for everything from the upfront with no interaction during
- Optimizing for model, not for the solution



SECOND ITERATION

- Regular check ins and much smoother work
- Metrics to define success with accurate timelines
- Releasing new versions of model on a weekly basis

To succeed we:

1. Chose good collaborators
2. Communicated frequently and effectively
3. Used milestones to inform and revise
4. Put what we could in the hands of our partners and customers as quickly as we could
5. Iterated towards a goal

What questions can I answer for you?

Additional Resources


Connect with Brian on LinkedIn


Learn more about DataCamp for Business


WHITE PAPER: Your Organization's Guide to Data Maturity


WHITE PAPER: The L&D Guide to Data Fluency


Register for one of our upcoming webinars


ON-DEMAND: Train your Workforce to Thrive in a Data-Driven Age

Thank you

Adel Nehme
Data Science Evangelist
adel@datacamp.com