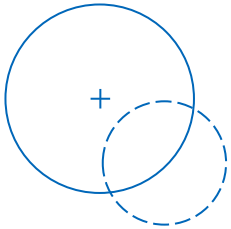
data iku | ACADEMICS

FALL 2021 HACKATHON

data iku | ACADEMICS

Agenda



Welcome & introduction

Who is Dataiku?

Hackathon overview

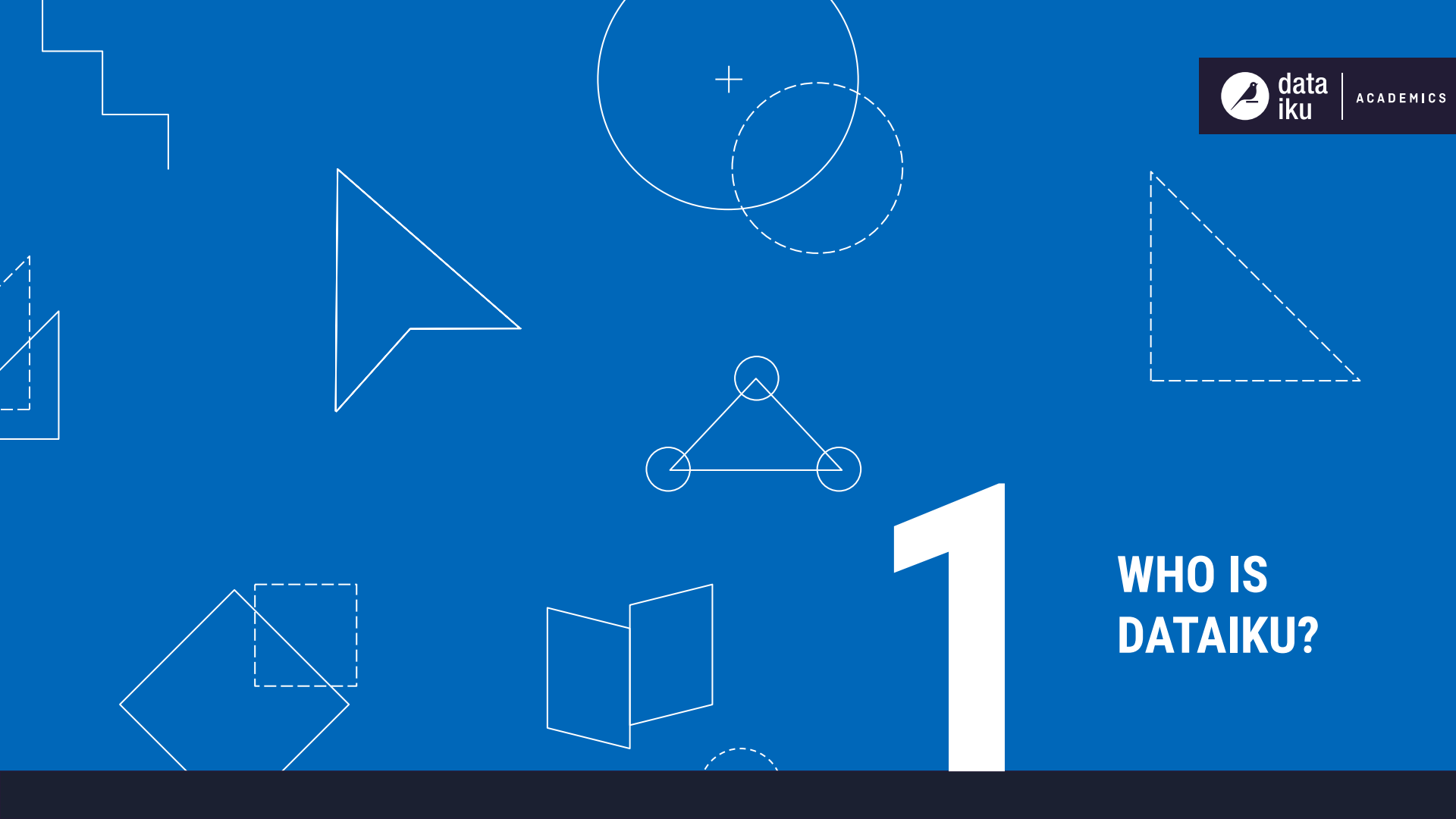
Hackathon logistics

Q&A

Wrap-up

dataiku | ACADEMICS

WELCOME!


data iku
ACADEMICS
1
WHO IS DATAIKU?

ACADEMIC PROGRAM

Our **completely free** program supports the activities of educators, students, researchers and higher education leaders, and we connect student talent to our 300+ customers and partners.

A Brief History of Dataiku

2013 2014 2015 2016 2017 2018 2019 2020 2021

Company

Dataiku Created

20th employee

100th employee

400th employee

750th employee

Office in NY

Office in London

Offices in Singapore & Australia

$4M raised

$14M Series A

$28M Series B

$101M Series C

$100M Series D

$400M Series E

$4.6B Valuation

Customers

First customer

First 10+ users customers

50th customer

First 100+ users customers

EGG CONFERENCE

100th customer

First 500+ users customers

1st CAB Customer Advisory Board

First 1000+ users customers

300th customer

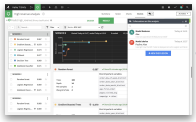
1000+ EGG Visitors

Software

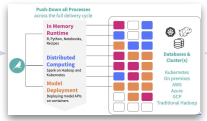
Version 2
Real-time collaboration
Spark integration

Version 3
Scenarios, automation and version control

Version 4
Advanced scalability, cloud support

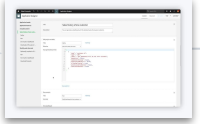


Version 5
Deep learning, Containerization, Project wikis

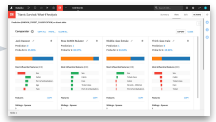


Version 6
Enterprise Scale Elastic AI with Kubernetes

Version 7
Explainable AI and Advanced Visual Statistics



Version 8
Applications-as-Recipes, Flow Zones, Model Doc Gen



Version 9
ML Diagnostics, Model Assertions, What-If Analysis, Fuzzy Join

Dataiku: The Unified Platform to Deliver Impactful AI at Enterprise Scale

Connect and use existing data and compute

Google Cloud Platform, Hadoop, MySQL, PostgreSQL, Teradata, Vertica, Amazon Redshift, Snowflake, IBM DB2, memSQL, MongoDB, Amazon S3, Cassandra, SQL Server, Avro, Netezza, Salesforce, Pivotal Greenplum, SAP HANA, Oracle, Exasol, JSON, XML, CSV, Excel, esri

Empower and upskill all user populations




No code | Low code | Full code | Programmatic control

- Cataloging and connectivity
- Data exploration and preparation
- (Auto) machine learning
- Insight sharing, visualization applications
- Scheduling Orchestration Deployment Monitoring

Systemic reuse and collaboration

Integration and staging

Centralized governance and audit

Horizontal and vertical scalability

Leverage all best available technologies

Alation, tableau, esri, Qlik, Power BI, CARTODB

Make the most of available cloud stack and open source technologies, transparently

Google Cloud Platform, aws, Microsoft Azure, On Premise

Empower Everyone From Low- to Full-Code

Python expert

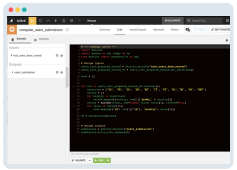
Data Scientist

Statistician

Analyst

Spreadsheet wizards

High code



Code your way

- Code freely with Jupyter, R, Python IDEs, any package, isolated environments, and full Git integration
- Have full programmatic control with full-fledged API for models, pipelines, automation

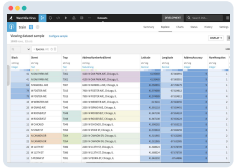
Don't get distracted

- Self-provision of compute resources including cloud-based elastic processing for large volumes of data, users or services
- Expedite wrangling with facilitated connections to SQL, HDFS, cloud storage, NoSQL, HDFS, APIs,...

Ensure impact

- Low effort CI / CD through orchestrated pipelines with optional automatic checks
- Create deployment artifacts
- Deploy your models as containerized APIs
- Showcase insights with webapps (Shiny, Flask, Bokeh) and deploy in Kubernetes package for reuse by the target population

Low code



Leverage advanced tech

- No need to master the underlying infrastructure
- Connect, explore and navigate
- Combine point and click with R, SQL, Python and leverage code snippets
- Packaging of cloud services

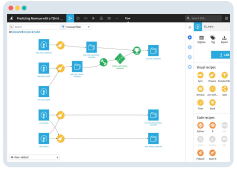
Orchestrate advanced flows

- Build check, metrics, and warnings on top of your data flows
- Orchestrate simple and complex data flows in minutes for continuous value generation

Upskill

- Gain exposure to more advanced usage of data within a controlled environment

No code



Point & click

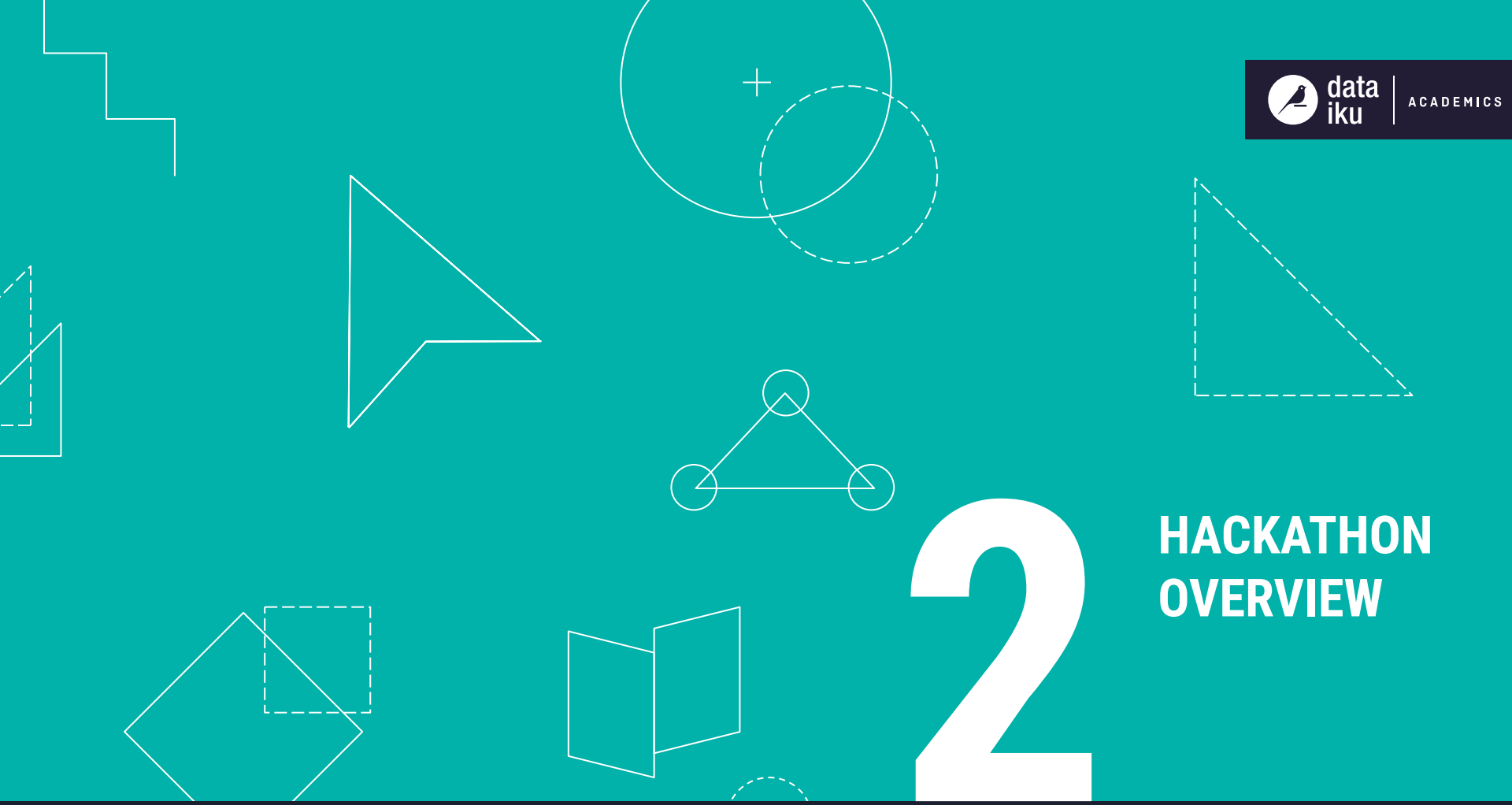
- Search, connect to, and explore data from preconnected or local systems
- Data wrangling, (auto) machine learning directly through visual and click interfaces
- Dashboarding, advanced statistics and data visualization

Build on shoulders of giants

- Use packaged plug-ins and full apps to deliver data scientist-grade projects with business expert insights
- Copy and paste past projects, data and models for reuse

Human and readable

- Understand projects, from data connection to trained models



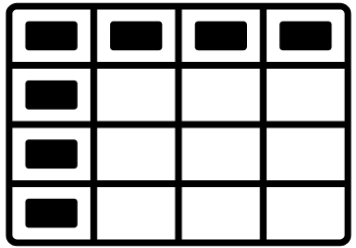
data iku | ACADEMICS

2

HACKATHON OVERVIEW

Build NYC Housing Price Prediction Model

Data obtained using the Zillow API



Data

Processing, EDA, Modeling and Prediction
in Dataiku

Insights

More details in the problem statement.

Sharing your results

Report and Presentation

data iku | ACADEMICS

Submit Predictions

- CSV format
- Export from Dataiku

ID bigint Integer	prediction double Decimal
2	2942737.9667038815
3	1241677.2784606314
10	637690.0875658464
14	938159.7923913429
17	1685577.1492876082
22	734224.0386626173

Written Report

Think of if as a blog!

- Description of the problem, analysis, models, results
- ~2000 words
- Include charts, images, code (if applicable)

Slide Presentation

- Concise summary of your report
- 8-10 minutes
- Technical+Non-technical audience

Judging Criteria

Predictions, Report and Presentation

Best RMSE $\neq$ **Best Project !**

What will the judges be looking for?

- Organization, quality of visuals, communication (verbal and written), clarity of insights, claims supported by evidence.
- Details on the rubric can be found in problem statement.


data iku
ACADEMICS
3
HACKATHON
LOGISTICS

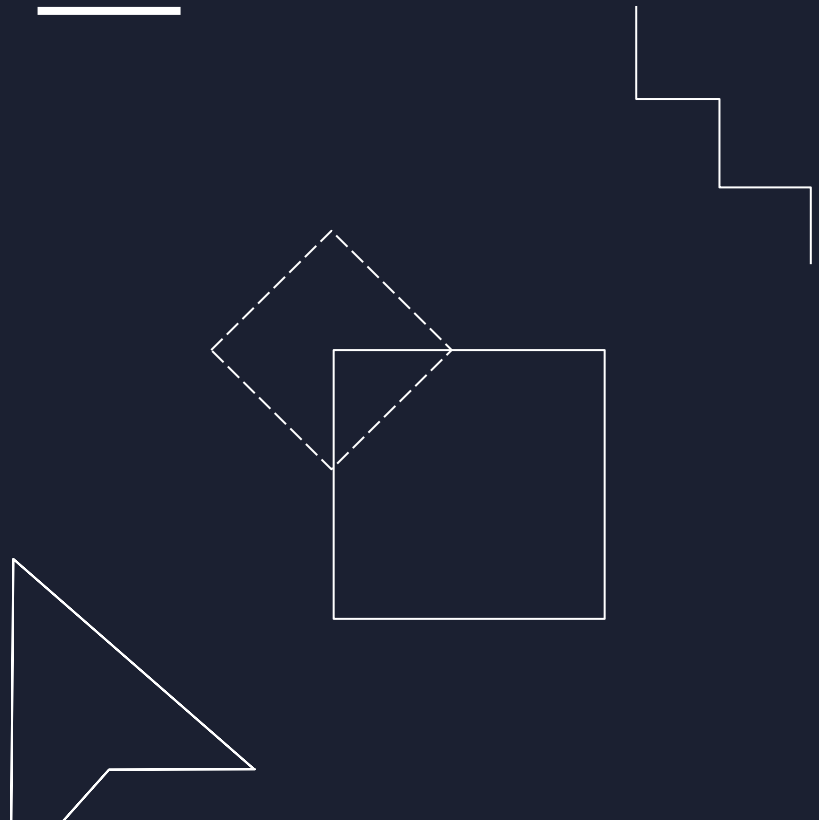
IMPORTANT DATES & DEADLINES

- October 12th: Virtual kick-off
- October 18th, 4pm Eastern: Storytelling with Data webinar
- October 22nd: Registration cut-off
- November 19th: Hackathon closes and submissions due

Office Hours

- October 19th, 3pm Eastern
- November 4th, 11am Eastern

OFFICE HOURS

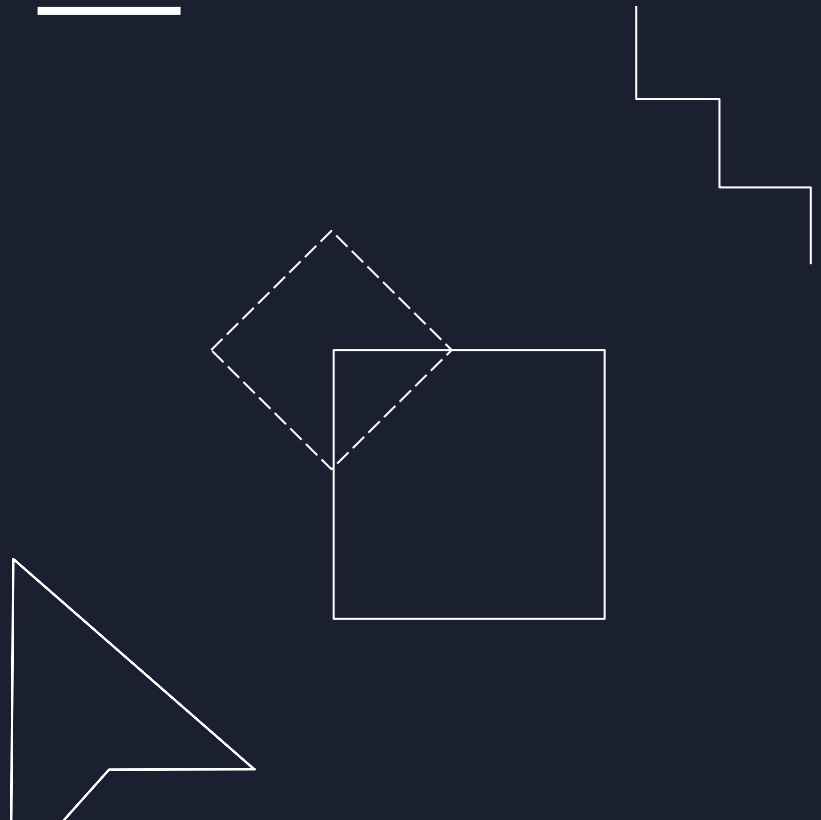


Drop in and ask questions!

Tuesday, October 19th from 3PM - 3:50PM Eastern
https://dataiku.zoom.us/j/93075157386?pwd=dGFhKzBjWnhzQ3dZMjR6NStidXQwZz09

Thursday, November 4th from 11AM - 11:50AM Eastern
https://dataiku.zoom.us/j/94089457883?pwd=UjZuWkVQRXhiZTlxenkzdkN6YXdwUT09

STORYTELLING

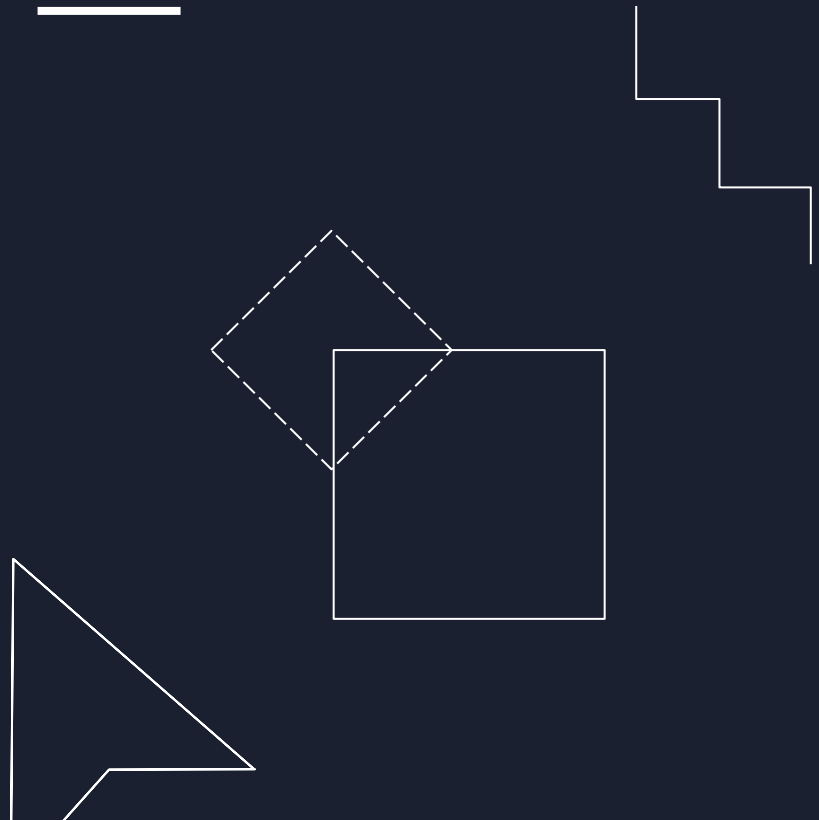


Paul Smith, best-selling author of the books Lead with a Story, Sell with a Story, and The 10 Stories Great Leaders Tell, will share what he's learned from his extensive research in storytelling.

You'll learn why you should be using more storytelling in your work, when you should be telling stories (and when not), as well as actionable tools for how to craft the most effective and engaging business stories. In the second half of the discussion, Paul will explain how to apply those same storytelling techniques to data.

Monday, October 18th at 4PM Eastern
Register here!

HOW DO I...



Form a team?
Coordinate internally! (Teams of 3-4)

Register?
Fill out the registration form (one per team)

Access my Dataiku instance?
After October 15th, Dataiku will provide access information to teams via email

Get help?
Join office hours, and/or post your questions to Slack:
#pace-hackathon-21

BE PREPARED

data iku | ACADEMICS

Technical

- Prepare: Complete Basics 101-103 and Machine Learning and Scoring Basics on the Dataiku Academy

Logistics

- An email will be sent out with instructions for connecting to Dataiku
- Attend office hours if you can
- Utilize the Slack channel: #pace-hackathon-21
- Event hashtag is the same: #pace-hackathon-21

Q&A

DATAIKU ACADEMICS

Interested in more events?

Visit our Academics Event Portal to stay up-to-date!

EVENTS PORTAL