Data science in troubled times

How the crisis affects data science activities and how Dataiku DSS can help

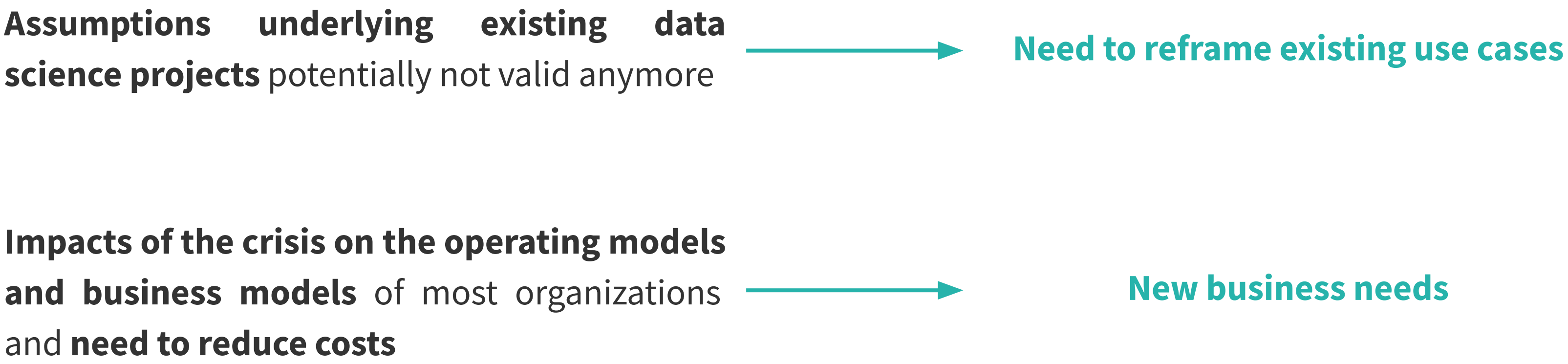
Five topics to discuss today

- Reframing data science projects
- Detecting broken models
- Retraining models
- Doing data science from afar
- Developing new skills

Reframing data science projects

Reframing data science projects

What are the impacts of the current crisis?



Assumptions underlying existing data science projects potentially not valid anymore → **Need to reframe existing use cases**

Impacts of the crisis on the operating models and business models of most organizations and **need to reduce costs** → **New business needs**

Reframing data science projects

Reviewing existing use cases (1/2)

Tip **Organize working sessions with business stakeholders and domain experts** to check the validity of the assumptions for the most impactful projects

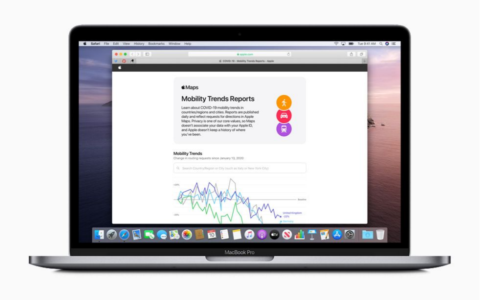
Examples of **questions to address**:

- How has the crisis changed the **business needs**?
- Has the crisis affected the **availability, reliability, and relevance of input data**?
- Should the **evaluation metric** be adjusted?
- Should the **predictions** be **consumed in a different manner**?
- What is the **plan B** for the business if the model is not valid anymore?

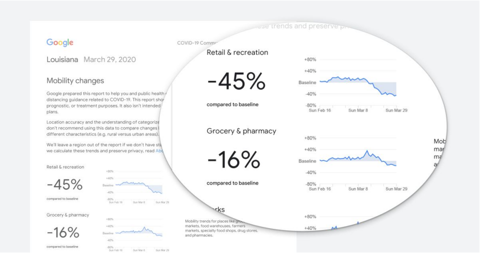
Reframing data science projects

Reviewing existing use cases (2/2)

Tip **Consider alternative data sources if existing data sources are insufficient or have become irrelevant**



Anonymized mobility data from [Google](#) and [Apple](#)


Google

Louisiana March 29, 2020

Mobility changes

Google prepared this report to help you and public health
distancing guidance related to COVID-19. This report sho
prognostic, or treatment purposes. It also isn't intended
plans.

Location accuracy and the understanding of categorize
don't recommend using this data to compare changes
different characteristics (e.g. rural versus urban areas).

We'll leave a region out of the report if we don't have st
we calculate these trends and preserve privacy, read Abo

Retail & recreation

-45%

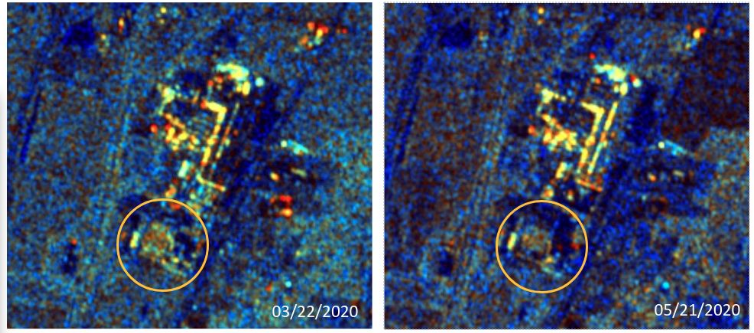
compared to baseline

Grocery & pharmacy

-16%

compared to baseline

Mobility trends for places like gro
markets, food warehouses, farmers
markets, specialty food shops, drug stores,
and pharmacies.



Satellite images analyzed by [Kayrros](#)
([partnership with Dataiku](#), webinar on 16 July)

Reframing data science projects

Detecting new business needs

Tip Identify the main challenges your organization currently faces and help business stakeholders identify data-driven ways to address them

Cf. our white paper or watch our webinar (in English or French) on **defining a successful AI project**.

DEFINING A SUCCESSFUL AI PROJECT

A Framework for Choosing the Right Use Cases

A WHITE PAPER BY DATAIKU

www.dataiku.com

Examples of current challenges and related data science use cases

Function	Need	Examples of use cases
Operations	Ensure that safety rules (e.g. social distancing) are complied with	Safety rules monitoring through computer vision
Operations	Automate manual tasks	Automated content moderation
Supply chain	Adapt to supply chain disruptions	Demand forecasting Inventory planning
Marketing	Detect sudden changes in consumer behavior	Social media analysis Consumer sentiment analysis
HR	Anticipate future workload and schedule human resources	Data-driven workforce planning

Detecting broken models

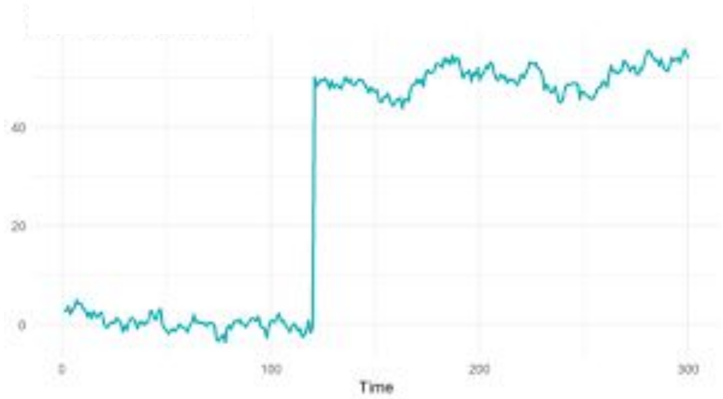
Detecting Broken Models

What are the impacts of the current crisis on ML pipelines?

An abrupt change at the onset of the crisis

By design, machine learning models learn patterns present in the training data.
Models trained on data prior to the crisis may become **irrelevant** if the underlying phenomena have significantly changed.

Such **drift** issues should be quickly detected, investigated and corrected for the models in production.



A rather abrupt change in data

ML Models at Risk

What does that mean for ML models ?

Forecasting Consumer Goods whether directly impacted (medicine, food,...) or impacted by lockdowns (fashion, cosmetics, books,...). ← Target Y changes: prior shift

Recommender model building upon buying patterns (fashion, cosmetics, books, movies,...). ← Features X changes: covariate shift

Churn Detection both B2C and B2B as many companies were temporally shutdown. ← Change of relationship X-Y: concept drift

Other examples: **Fraud Detection** for health insurance, ...

Not All Broken Models Are The Same

How fast is the feedback collected ?

Quick Feedback

The ground truth target is quickly collected so model performance can be measured and any degradation can be flagged.

Examples. Recommender systems

Methods. Thresholding, Statistical tests, Hoeffding drift,...

Existing Solutions. For advanced methods, streaming-oriented scikit-multiflow.

Delayed Feedback

The ground truth target cannot be quickly collected and model performance may only be measured weeks, months from scoring.

Examples. Fraud, Churn, Forecasting,...

Methods. Monitor changes of distributions of features X as well as distribution of predictions.

Existing Solutions. Data Validation with TensorFlow Extended (tfx).

data
iku

Detecting broken models in DSS

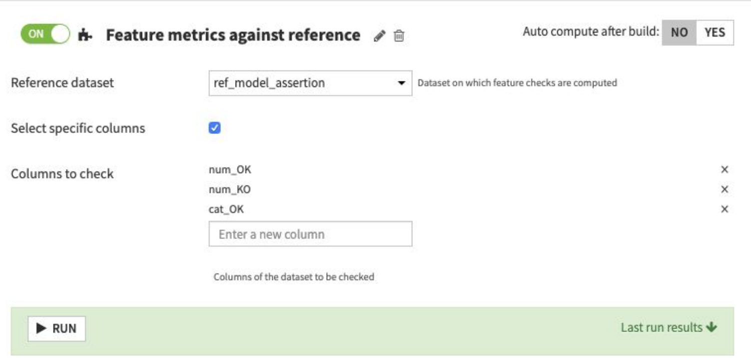
Anticipating potential drift

Tip **Monitor the performance of models in production** (this should already be the case!).

It can be done by leveraging metrics, checks and scenarios in Dataiku DSS. New Interactive Statistic features can help set up rigorous statistical test.

Tip **Check data compliance with past data.**

Based on the training dataset of ML model, automatically check new incoming data by putting bounds on values (min/max /frequency).



ON **Feature metrics against reference**

Auto compute after build: NO YES

Reference dataset: ref_model_assertion — Dataset on which feature checks are computed

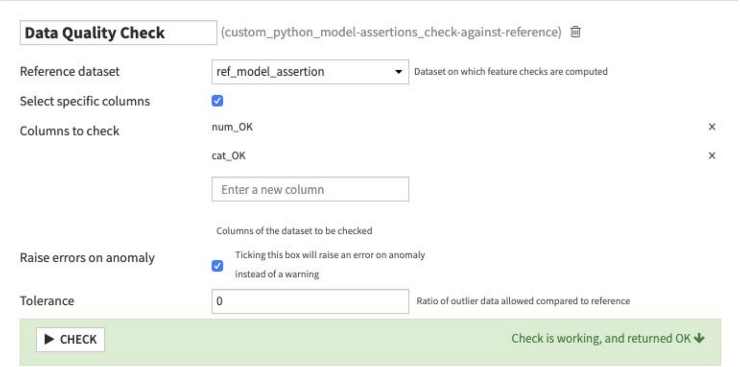
Select specific columns: ☑

Columns to check: num_OK ×, num_KO ×, cat_OK ×

Enter a new column

Columns of the dataset to be checked

▶ RUN — Last run results



Data Quality Check (custom_python_model-assertions_check-against-reference)

Reference dataset: ref_model_assertion — Dataset on which feature checks are computed

Select specific columns: ☑

Columns to check: num_OK ×, cat_OK ×

Enter a new column

Columns of the dataset to be checked

Raise errors on anomaly: ☑ Ticking this box will raise an error on anomaly instead of a warning

Tolerance: 0 — Ratio of outlier data allowed compared to reference

▶ CHECK — Check is working, and returned OK

Detecting broken models in DSS

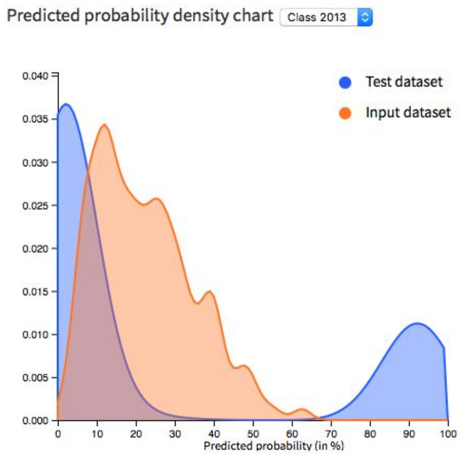
Measuring drift in the case of delayed feedback

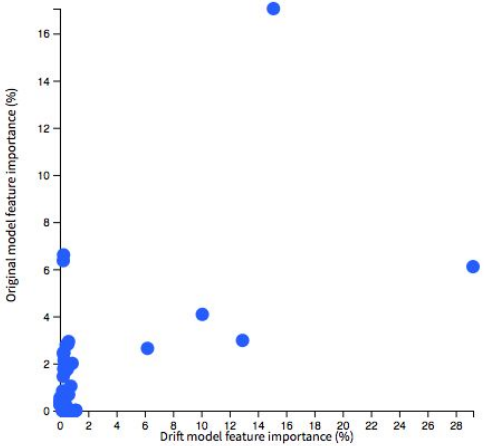
Tip **Use Dataiku DSS plugin for model drift monitoring**, especially when the ground truth labels are not quickly available

The model drift monitoring plugin allows to compare recent data with the data on which the model was evaluated. If these datasets are too different, the model may need to be retrained.

The plugin takes as inputs a deployed model that we want to monitor and a dataset containing new data the model is exposed to. It provides:

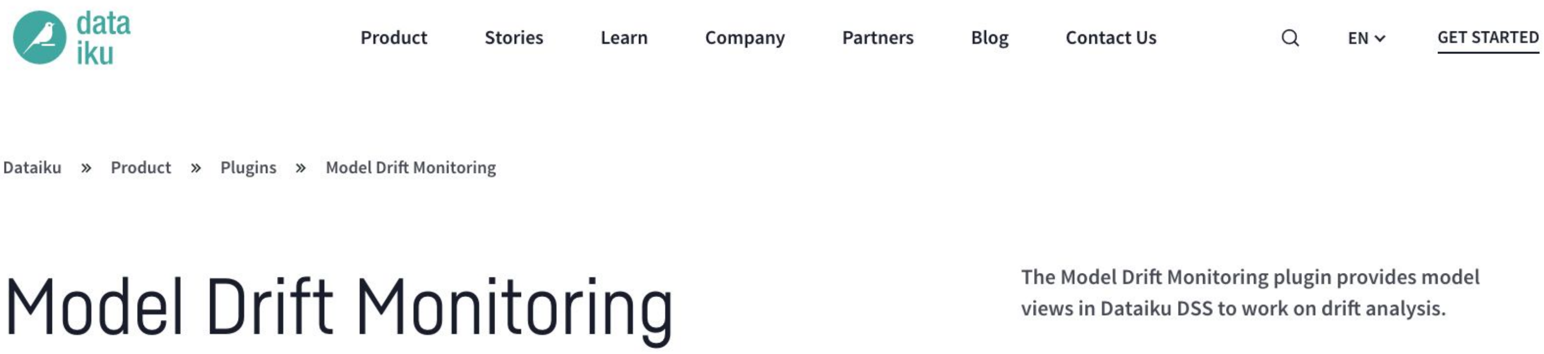
- A **drift score**
- A table and a chart **comparing predictions for each class** when scoring **with both the test and input datasets**
- A chart showing the **importance of individual features both for the original model and the data drift**





Data Drift Plugin in DSS

Demo if time allows...



Model Drift Monitoring

The Model Drift Monitoring plugin provides model views in Dataiku DSS to work on drift analysis.

dataiku

Retraining models

Rescuing Drifted Models

Can the previous model be saved ?

Past data and models are still relevant

The concept hasn't changed and previously labeled data is still relevant. New data can be incorporated to learn a new model.

If the new data is labeled, deep learning model can be recycled with **transfer learning** and **fine-tuning**.

Otherwise, if the pool of new unlabeled data is large enough, **semi-supervised learning** offers an interesting alternative. More sophisticated **domain adaptation** techniques can also be used.

Few directly relevant data

The past data labels are irrelevant, there is a concept shift. The old model is of no use.

If the new data is labeled, in sufficient quantity, an option is to discard all data and learn a new model from scratch.

Otherwise, it is important to first label the new data (and optionally the old data). This is where **Active Learning** techniques can be leveraged.

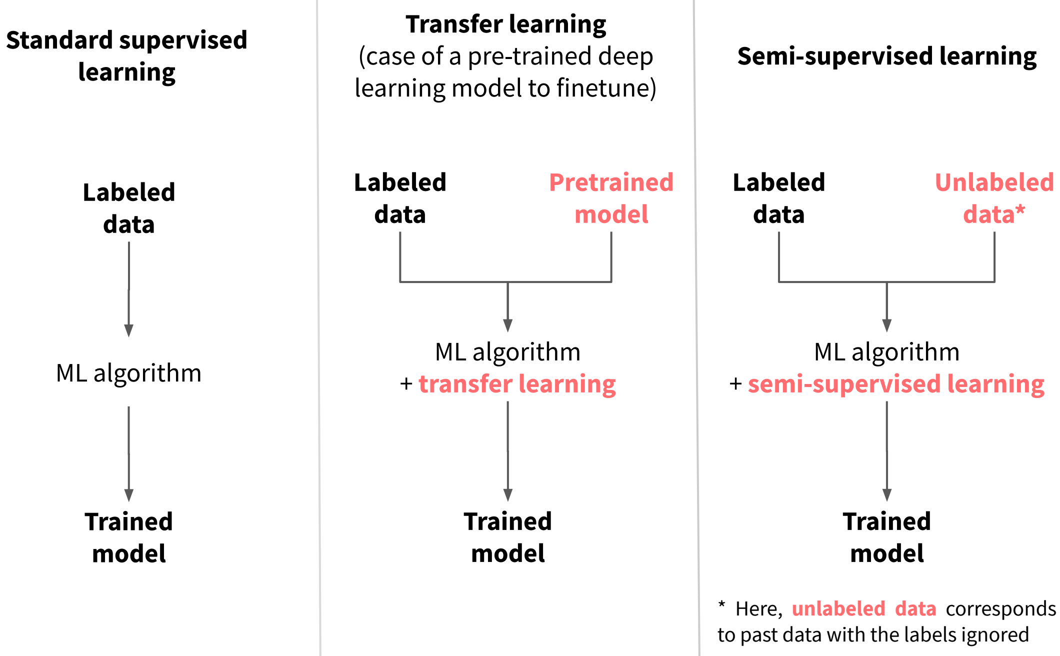
Retraining Models with New Data

Reusing past data or past models

Tip Consider using transfer learning or semi-supervised learning, when past data or past models exist

Even if the context has significantly changed, past data or past models may still be useful **if the input data distribution has not been strongly impacted**.

Beyond transfer or semi-supervised learning, other techniques include **importance reweighting** (to give more to new data).



* Here, **unlabeled data** corresponds to past data with the labels ignored

Training Models with Small Data

ML-assisted Labeling

Tip When only recent data can be used, apply the usual good practices for training models with few data

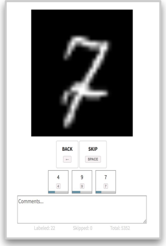
This includes:

- Favoring **less expressive models (i.e. regularization)**
- Using **data augmentation** for deep learning models
- Being especially cautious with outliers, data imbalance, or when **evaluating performance** through cross-validation.

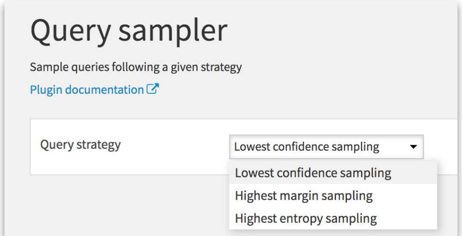
Focus on Active Learning in Dataiku DSS

ML algorithms require high quality labeled data but **labeling can be tedious, time consuming, and expensive**. **Dataiku DSS** reduces time and efforts to create training datasets by:

- Making **human-in-the-loop data labeling** easy (whether your data is tabular, images, or sounds)
- Using active learning to smartly select the **best samples for annotators to label next** (instead of randomly selecting them)



Labeled: 22 Skipped: 0 Total: 5352


Query sampler

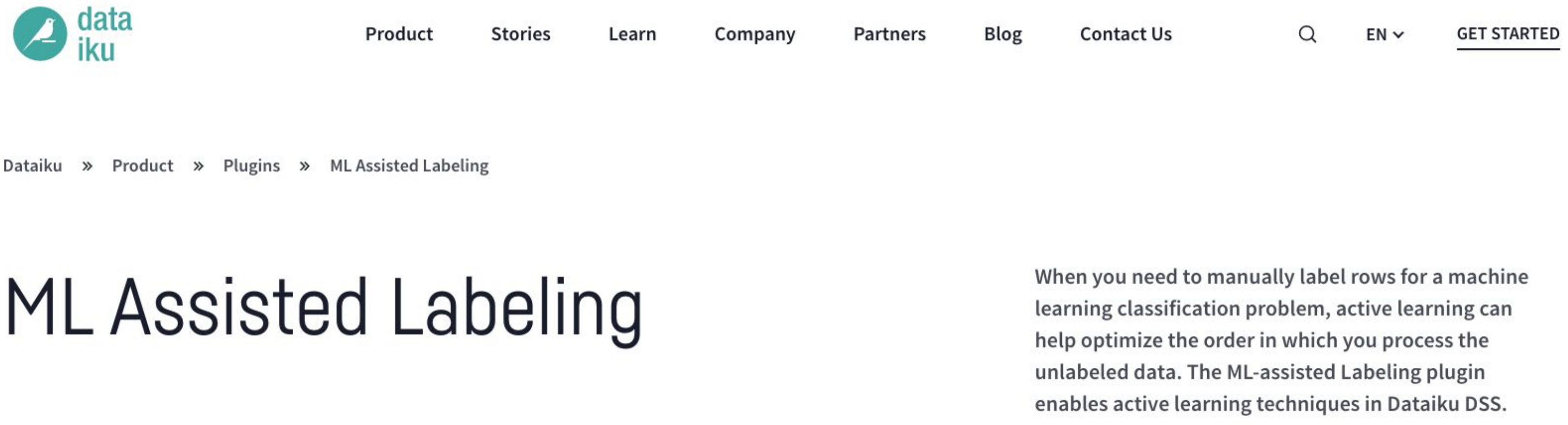
Sample queries following a given strategy

Plugin documentation

Query strategy: Lowest confidence sampling

ML-assisted Labeling in DSS

Demo if time allows...



ML Assisted Labeling

When you need to manually label rows for a machine learning classification problem, active learning can help optimize the order in which you process the unlabeled data. The ML-assisted Labeling plugin enables active learning techniques in Dataiku DSS.

dataiku

Doing data science from afar

Doing data science from afar

What are the impacts of the current crisis?

More written communication → **Risks of misunderstanding or information overload**

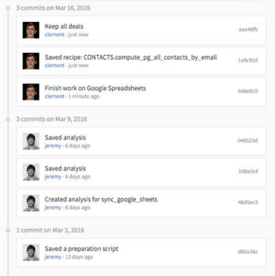
Less opportunities for informal discussions → **Harder to stay-up-to-date or get help**

Doing data science from afar

Using Dataiku DSS collaborative features more extensively

Team activity

- **Every action is versioned** through an integrated Git repository
- **Follow each action in the timeline**
- Active and inactive projects
- **Notification of changes** to team members



3 commits on Mar 16, 2016

Keep all deals
clement - just now
aae48fb

Saved recipe: CONTACTS.compute_pg_all_contacts_by_email
clement - just now
1a9c92d

Finish work on Google Spreadsheets
clement - 1 minute ago
648e810

3 commits on Mar 9, 2016

Saved analysis
jeremy - 6 days ago
046523d

Saved analysis
jeremy - 6 days ago
b9be5cf

Created analysis for sync_google_sheets
jeremy - 6 days ago
48d5ec3

1 commit on Mar 3, 2016

Saved a preparation script
jeremy - 13 days ago
d85a36c

Wiki for centralized documentation

- Create **team documentation**
- Centralize and organize the shared resource in a hierarchical manner
- Create a project new entry point with structured documentation

Integration with collaboration tools

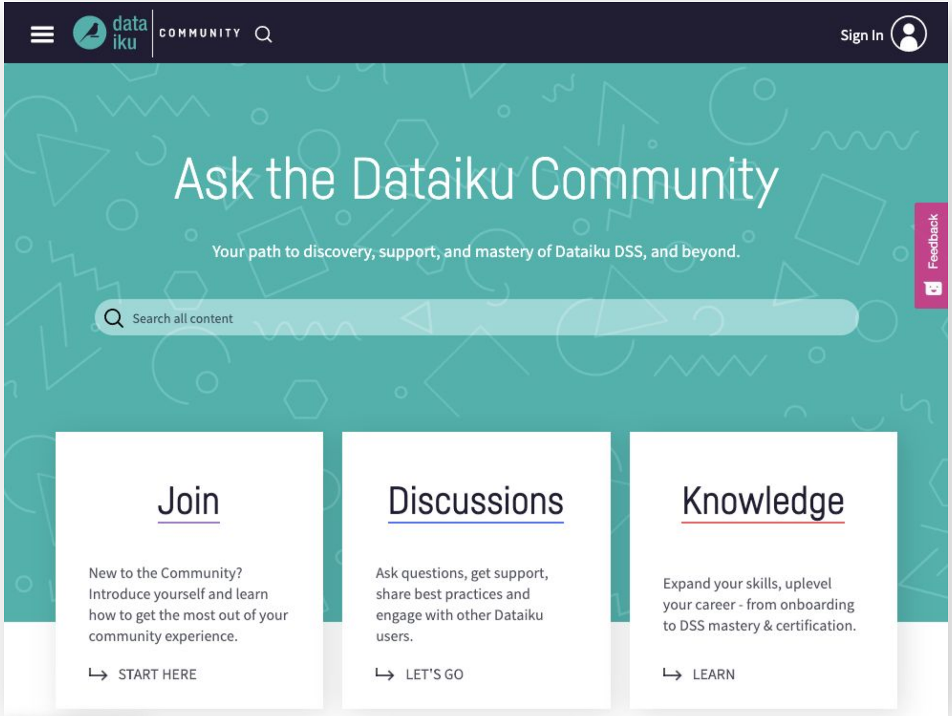
- Send **scenario updates** to **Microsoft Teams**, **Slack**, **Twilio** or **emails**
- Use **remote repositories** (e.g. GitHub for projects and/or plugins)

Tip **Take advantage of DSS collaborative features**

Doing data science from afar

Leveraging the Dataiku Community

Tip **Join the Dataiku Community for peer-to-peer support**



Ask the Dataiku Community

Your path to discovery, support, and mastery of Dataiku DSS, and beyond.

Search all content

Join

New to the Community? Introduce yourself and learn how to get the most out of your community experience.

↳ START HERE

Discussions

Ask questions, get support, share best practices and engage with other Dataiku users.

↳ LET'S GO

Knowledge

Expand your skills, uplevel your career - from onboarding to DSS mastery & certification.

↳ LEARN

Developing new skills

Developing new skills

What are the impacts of the current crisis?

New needs for new projects?
More time to learn? → **Needs or additional time to develop new skills**

Traditional training sessions in a class setting not possible anymore → **Switch to remote learning**

Developing new skills

Using free online resources

Tip Take advantage of the training resources made available during the crisis

Examples:

- MOOC platforms such as Coursera and Udacity
- Publishers such as Springer Nature and Cambridge University Press
- Dataiku’s “Data Science from Home Calendar”

DATA SCIENCE FROM home

DATA SCIENTIST CALENDAR

Tip Visit Dataiku Academy, the new online and self-paced Dataiku training and certification platform

Welcome to

dataiku ACADEMY

Developing new skills

Organizing remote training sessions

Tip Look for online resources on organizing remote training sessions

Many resources have recently been made available to help educators transition to remote teaching. For example:

- Many **universities**, such as Stanford, UCLA or Penn, have published guidelines for their instructors
- **Tech companies**, such as Zoom or Google, offer guidance on how best to leverage their tools
- **Online learning platforms**, such as Coursera or Khan Academy, also provide tips for remote teaching

A Quick-Start Guide to Teaching from Home by Andrew Ng

May 15, 2020 ♡9

Tips & Tricks:
Teachers Educating on Zoom

Starting a Classs

- If it's your first time using Zoom, visit the Educating Guide: Getting Started on Zoom and complete the steps in advance of your class to get you set up.
- Schedule your class in the Zoom application for your desired date/time and copy the invitation details to send to your students.
*Please note that students will not need to register for an account to join.
- Join your class a couple minutes early to ensure a proper connection then follow the below tips for a quality online learning experience.

Tips and Tricks for Virtual Lessons

- For your first class, set aside some time to introduce your students to Zoom and ensure that they're able to connect their audio and video.

Lessons from our remote training sessions

In March 2020, we converted our training sessions for customers and partners to a remote format. Here are the lessons we drew from this:

- Split training sessions in **shorter segments** (≤ 4 hours)
- Reduce the **number of trainees** (≤ 10 persons)
- Make sure the attendees have a **similar background** (to the extent possible)
- Be extra careful about the **logistics** (video conferencing software, proper equipment - microphone, webcam… - especially for trainers, time zones...)
- Make the **training sessions as interactive as possible**, in particular by using the features of your video conferencing software (e.g. "raise hands", "break-out rooms", "polls")
- Share the **slides** at the beginning of the training session
- Take **breaks** (~15 minutes every hour)

dataiku

Thanks for your attention