twiml

AI Enterprise Workflow Study Group

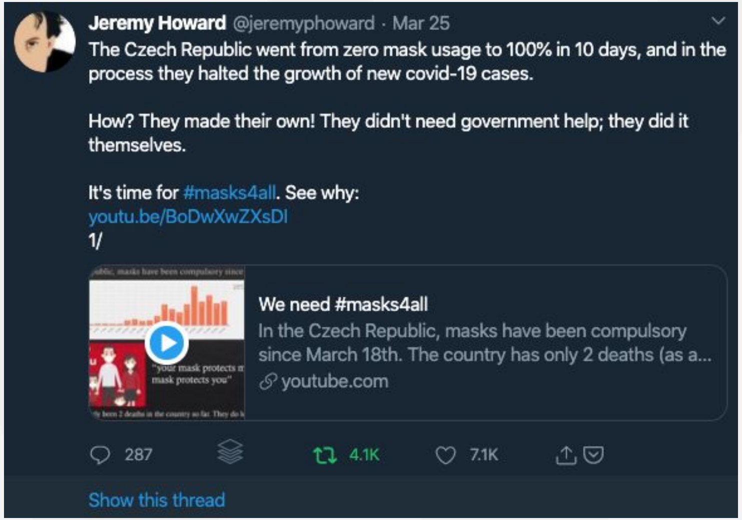
Course 3, Week 2

3/28/2020

Agenda

- Check in
- Discussion
- Next steps

#Masks4All



Jeremy Howard @jeremyphoward · Mar 25

The Czech Republic went from zero mask usage to 100% in 10 days, and in the process they halted the growth of new covid-19 cases.

How? They made their own! They didn't need government help; they did it themselves.

It's time for #masks4all. See why: youtu.be/BoDwXwZXsDI

1/

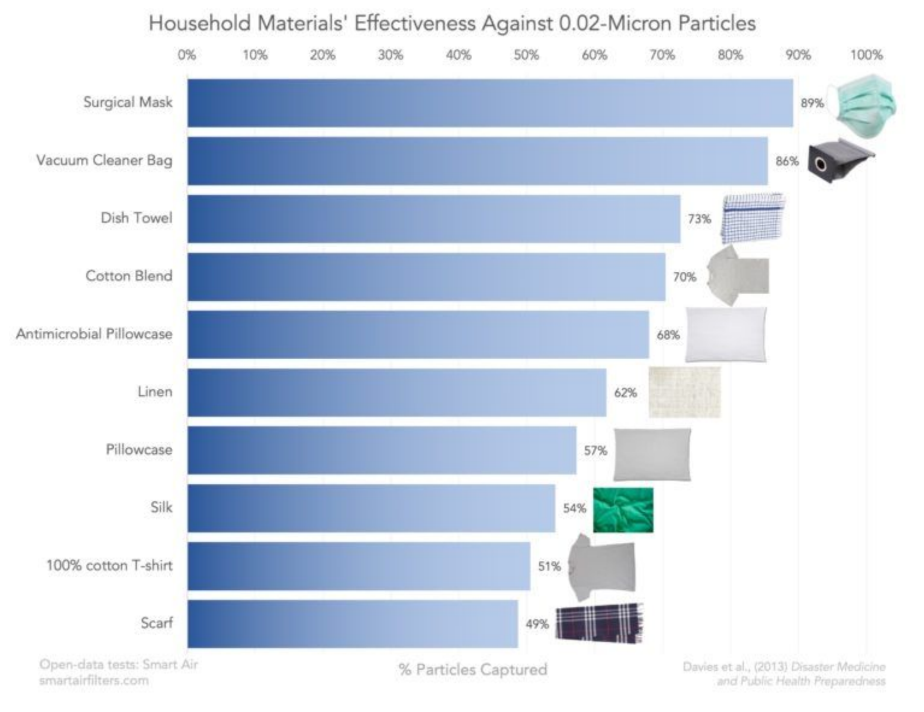
We need #masks4all

In the Czech Republic, masks have been compulsory since March 18th. The country has only 2 deaths (as a...

youtube.com

287 4.1K 7.1K

Show this thread



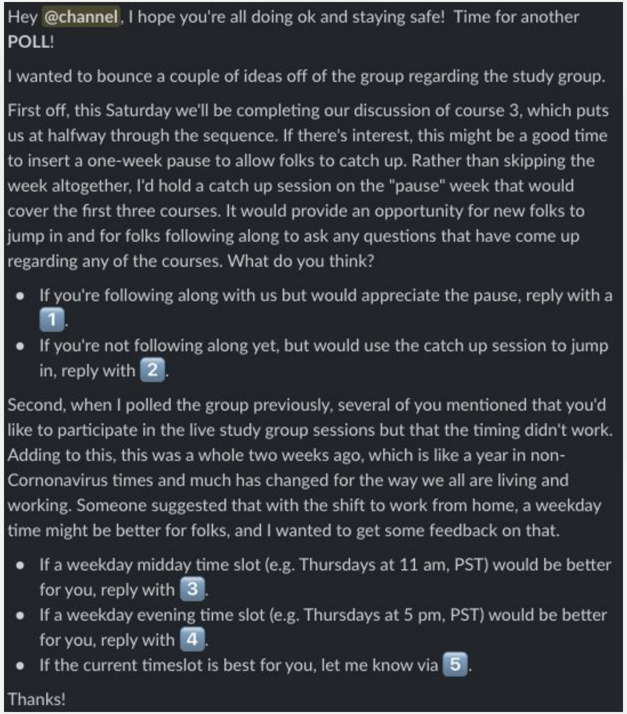
Household Materials' Effectiveness Against 0.02-Micron Particles

Open-data tests: Smart Air smartairfilters.com

Davies et al., (2013) *Disaster Medicine and Public Health Preparedness*

Check out: tiny.cc/maskswork

Recent Poll



Hey @channel, I hope you're all doing ok and staying safe! Time for another POLL!

I wanted to bounce a couple of ideas off of the group regarding the study group.

First off, this Saturday we'll be completing our discussion of course 3, which puts us at halfway through the sequence. If there's interest, this might be a good time to insert a one-week pause to allow folks to catch up. Rather than skipping the week altogether, I'd hold a catch up session on the "pause" week that would cover the first three courses. It would provide an opportunity for new folks to jump in and for folks following along to ask any questions that have come up regarding any of the courses. What do you think?

- If you're following along with us but would appreciate the pause, reply with a 1.
- If you're not following along yet, but would use the catch up session to jump in, reply with 2.

Second, when I polled the group previously, several of you mentioned that you'd like to participate in the live study group sessions but that the timing didn't work. Adding to this, this was a whole two weeks ago, which is like a year in non-Cornonavirus times and much has changed for the way we all are living and working. Someone suggested that with the shift to work from home, a weekday time might be better for folks, and I wanted to get some feedback on that.

- If a weekday midday time slot (e.g. Thursdays at 11 am, PST) would be better for you, reply with 3.
- If a weekday evening time slot (e.g. Thursdays at 5 pm, PST) would be better for you, reply with 4.
- If the current timeslot is best for you, let me know via 5.

Thanks!

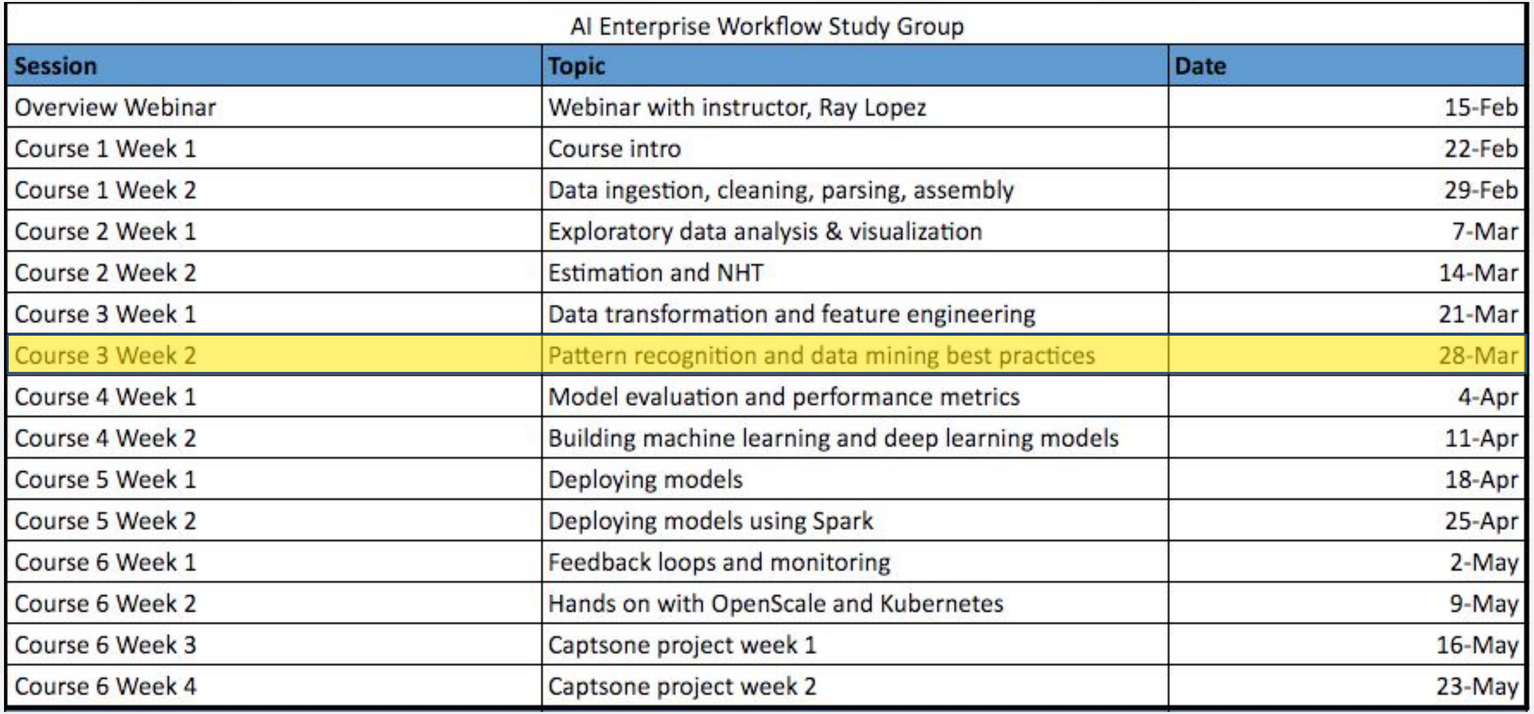
Pause:

- Seven votes pro-catch up
- Next two weeks catch up sessions 4/4 & 4/11

Time Slot:

- Move to midday: 2
- Move to weekday: 2

Course & Study Group Schedule



AI Enterprise Workflow Study Group

Session	Topic	Date
Overview Webinar	Webinar with instructor, Ray Lopez	15-Feb
Course 1 Week 1	Course intro	22-Feb
Course 1 Week 2	Data ingestion, cleaning, parsing, assembly	29-Feb
Course 2 Week 1	Exploratory data analysis & visualization	7-Mar
Course 2 Week 2	Estimation and NHT	14-Mar
Course 3 Week 1	Data transformation and feature engineering	21-Mar
Course 3 Week 2	Pattern recognition and data mining best practices	28-Mar
Course 4 Week 1	Model evaluation and performance metrics	4-Apr
Course 4 Week 2	Building machine learning and deep learning models	11-Apr
Course 5 Week 1	Deploying models	18-Apr
Course 5 Week 2	Deploying models using Spark	25-Apr
Course 6 Week 1	Feedback loops and monitoring	2-May
Course 6 Week 2	Hands on with OpenScale and Kubernetes	9-May
Course 6 Week 3	Captsone project week 1	16-May
Course 6 Week 4	Captsone project week 2	23-May

Course 3 Week 2 learning objectives

1. Employ IBM AI Fairness 360 libraries to detect bias in models
2. Employ outlier handling best practices in high dimension data
3. Employ outlier detection algorithms as a quality assurance tool and a modeling tool
4. Employ unsupervised learning techniques using pipelines as part of the AI workflow
5. Employ basic clustering algorithms

AI Fairness 360

Sources of bias:

- Bias already in original training dataset
- Algorithm that creates model may be biased (towards particular features)
- Test set may be biased

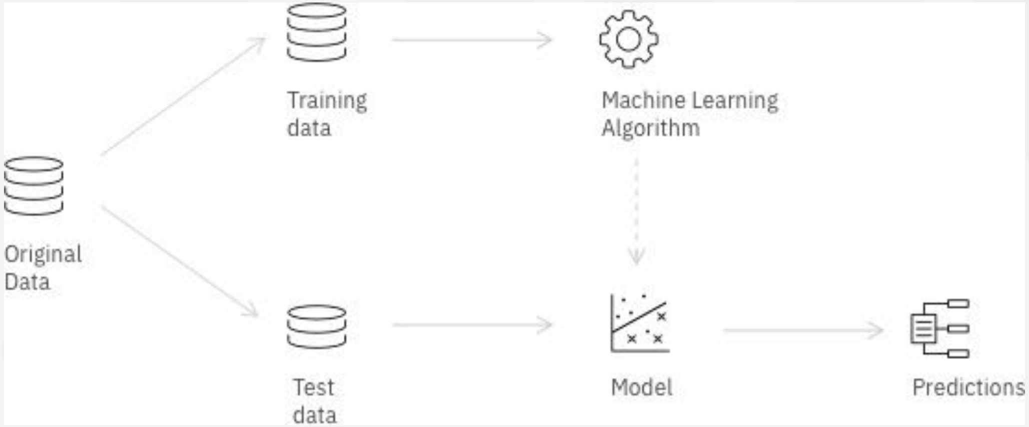
⇒ Pre-Processing

⇒ Processing

⇒ Post-Processing

Tool:

- Fairness metrics
- Bias mitigators



https://aif360.mybluemix.net/

Outlier Handling

Outliers: Data points or observations that fall outside of expected distribution or pattern.

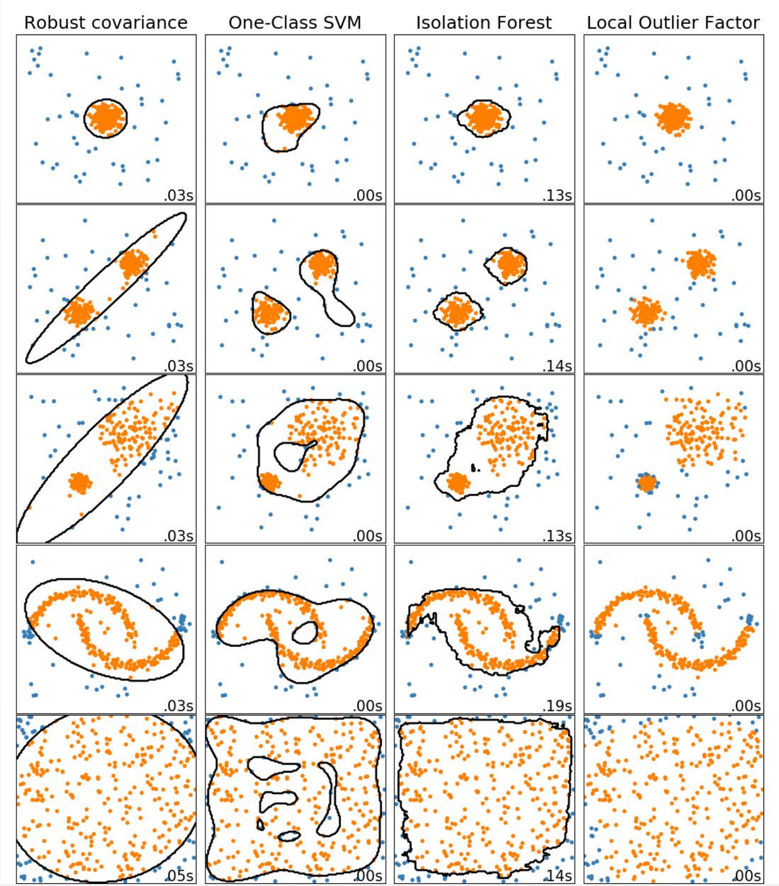
Outlier discovery generally occurs during EDA.

Outlier vs Novelty:

- Outliers defined from perspective of training data
- Novelty detection assumes training data does not contain outliers

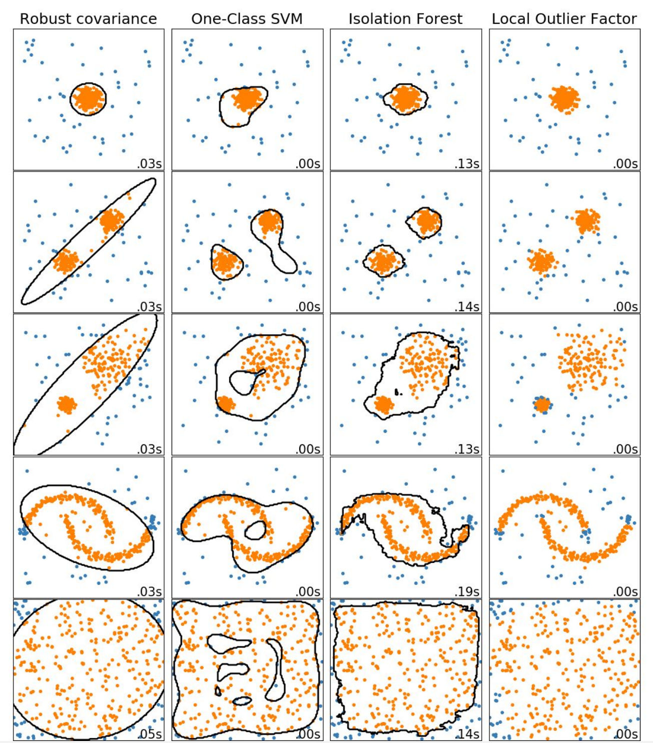
Outlier Detection Algorithms

- **Elliptic envelope (Outlier).** May break or not perform well in high-dimensional settings. (Use dim reduction first.)
- **One class SVM (Outlier, Novelty).** Requires choice of kernel and scalar parameter to define boundary.
- **Local Outlier Factor (Outlier, Novelty).** Works well on high dimensional datasets. Usually radial basis function kernel.
- **Isolation Forest (Outlier).** Works like Random forests w/ many single decision stumps. Several tunable params.



https://scikit-learn.org/stable/auto_examples/plot_anomaly_comparison.html

Comparing Outlier Detection Algorithms



https://scikit-learn.org/stable/auto_examples/plot_anomaly_comparison.html

Note: Robust covariance = EllipticEnvelope

Clustering

Unsupervised learning: Patterns in features themselves, vs in relation to labels

Clustering algorithms:

- Combinatorial algorithms
  - **k-Means** and Hierarchical clustering
  - Based on geometric perspective: treating observations as points in space
- Mixture modeling
  - Data grouped probabilistically, drawn from finite number of distributions
  - **Gaussian Mixture Model**, Dirichlet Process Mixture Model
- Mode seeking, e.g. Mean shift
- Ranking systems/transformations of data
  - Spectral clustering
  - Affinity propagation

Evaluating Clustering Performance

Number of clusters K is parameter that needs to be estimated/optimized. Some methods like Dirichlet Process Mixture Model don't need.

Can't use cross-validation, so choosing what's "good enough" will always be a judgment call.

Evaluating clustering performance:

- Adjusted Rand Index and Normalized mutual information score, if cluster labels are known
- **Silhouette score** and Davis-Bouldin, if labels unknown
- Elbow method mentioned, but not usually recommended

Note: When comparing across algorithms, it's important that comparisons be between results w/ roughly same number of clusters.

Silhouette Score

Method of interpretation and validation of consistency within clusters of data.

Silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation), ranging from -1 to 1.

- -1: incorrect clustering
- 0: highly overlapping clusters
- 1: dense well-separated clusters

Additional Discussion

What did you learn?

What stumbling blocks did you run into?

How do these lessons relate to your experience?

What did you learn/find interesting in this week's lesson?

What are you doing as homework?

What interesting resources have you found?

Other?

Reminder

Next two weeks 4/4 and 4/11 will be catch up sessions. Please join us even if you're caught up as the discussion will help us go deeper into the material.

Will start Course 4 on 4/18.

Chime in on Slack if you run into any issues or want to share any observations

Prepare your questions, discussion points, etc. for next week's meetup twiml