UML – Wrap Up:

Brock Tibert
btibert@bu.edu

BU
Boston University Questrom School of Business

I'm a #datascientist who used to believe that machine learning was everything.

Now I realize that ML takes its rightful place alongside other equally important pieces of the puzzle including:

• Working with stakeholders

• Formulating business problems as #data problems

• Discovering relevant data sets

• Connecting to data

• Cleaning and preparing data

• Exploring and understanding data

• Defining key metrics

• Analyzing & experimenting with data

• Visualizing data

• Incorporating the ML model into a broader AI workflow

• Converting the predictions of the model into prescriptive recommendations

• Incorporating business logic to handle the recommendations of the system in a realistic way

• Accounting for exceptions & edge cases

• Automating as much of the AI workflow as possible & making the rest human-friendly

• Deploying the AI workflow

• Driving adoption of the AI workflow by decision makers

• Monitoring and repairing the AI workflow

• Iteratively improving upon the AI workflow

• Measuring the value of the AI workflow throughout its life cycle

Most relevant

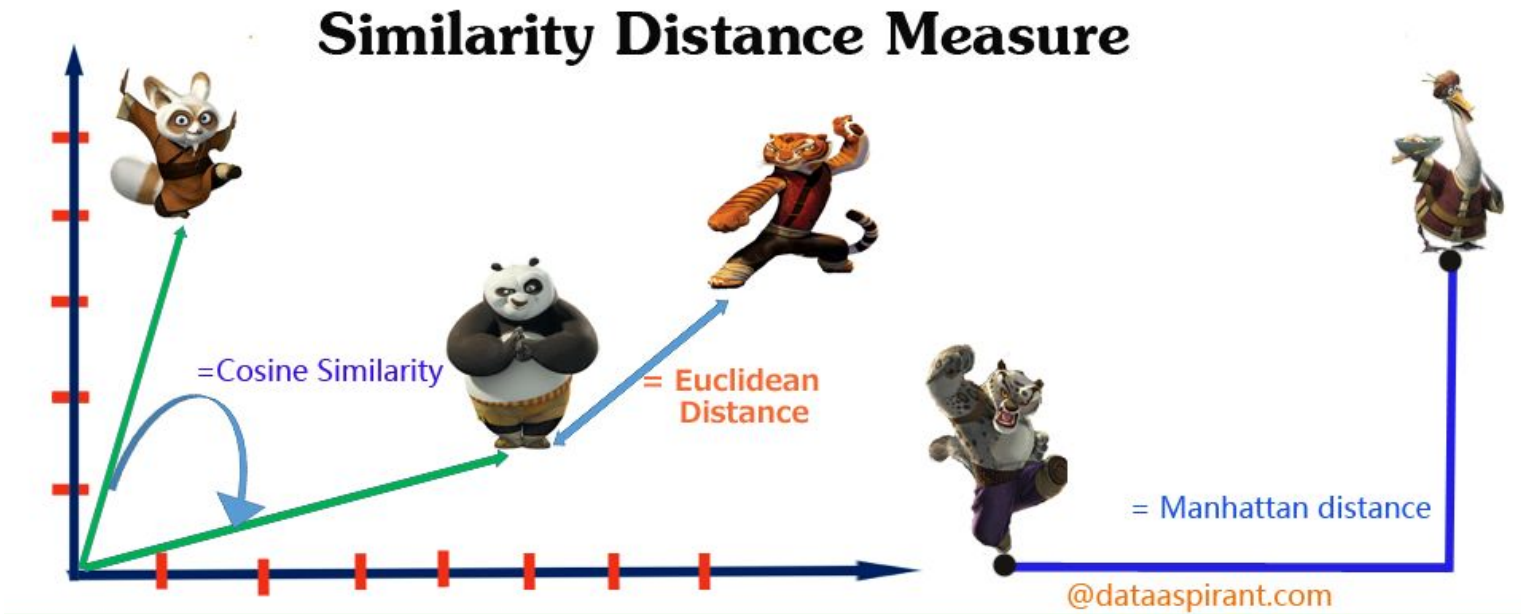
Kirk Mettler • 1st
Chief Data Scientist and R guy at IBM
2d •••

When I started it is all about the models. Slowly I learned that was probably the smallest and easiest part of the gig.

PCA and beyond

- PCA wraup
- Tsne
- UMAP
- UML and SML together in a data challenge!

Exam 1 next week - All Together - Different Building on syllabus



First-half Review

What have we covered …. And why

- Association Rules
  - Intuitive pattern recognition in form of "if **THIS** then **THAT**"
  - Offline method: we collect the data, build/filter rules, and then use that work to drive insights and actionable recommendations
- Cluster Analysis/Data Segmentation (Operate on "Rows")
  - Find insights and patterns within our data when we do not have/use a target variable
  - The clusters themselves are strategic outcomes (marketing personas) or can be used to create a new column (cluster assignment) for use of profiling or other analysis tasks
- Dimensionality Reduction (Operate on "columns")
  - When we have lots of columns, or highly correlated features, we can shrink down our search space
  - PCA looks to minimize the number of features, maximize variance retained, and recreate our numeric inputs into ***better*** columns that are uncorrelated, but projected onto a new coordinate system
  - Supports compression and data viz

Alternatives to PCA for Dimensionality Reduction

t-SNE (2008)

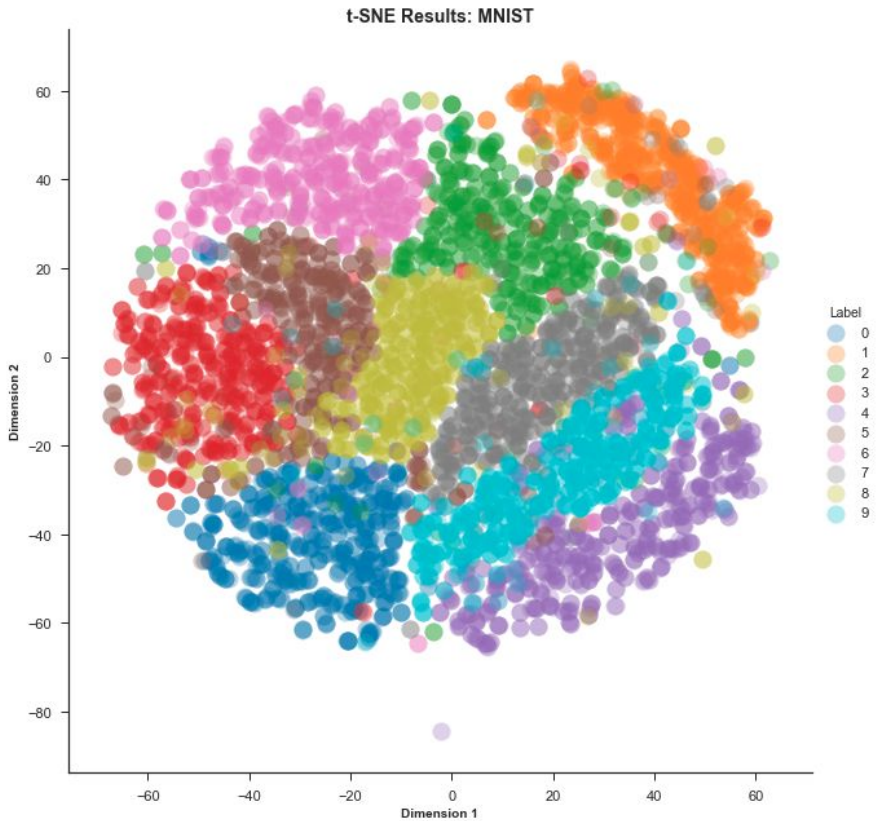
Remember that PCA is a linear combination of all of the features we use?

t-SNE uses a non-linear approach, and just like MDS, we can use the technique to project our information onto 2 dimensions.

This is a data reduction/viz technique, but the output of the process can be used downstream in other analytical tasks like clustering and classification models.

The approach can be **computationally intensive** as it attempts to minimize the fit within the data (keep similar points together) as it reduces the dimensionality. It can be slow (sklearn version)!

Popular with extremely high dimensional data like images, genomics, etc.



UMAP (2018)

Directionally similar to t-SNE, but arguments can be made that UMAP is able to better preserve the global structure of the original data. Moreover, UMAP tends to better separate clusters (categories) within the data.

- Faster than t-SNE
- Can generate a larger embedding space, whereas t-SNE is (nearly always) two dimensional
- Results heavily influenced by the parameters we choose
- We can configure the distance metric used

For both t-SNE and UMAP, some approaches will apply PCA beforehand, but UMAP tends to scale much better and may not need it.

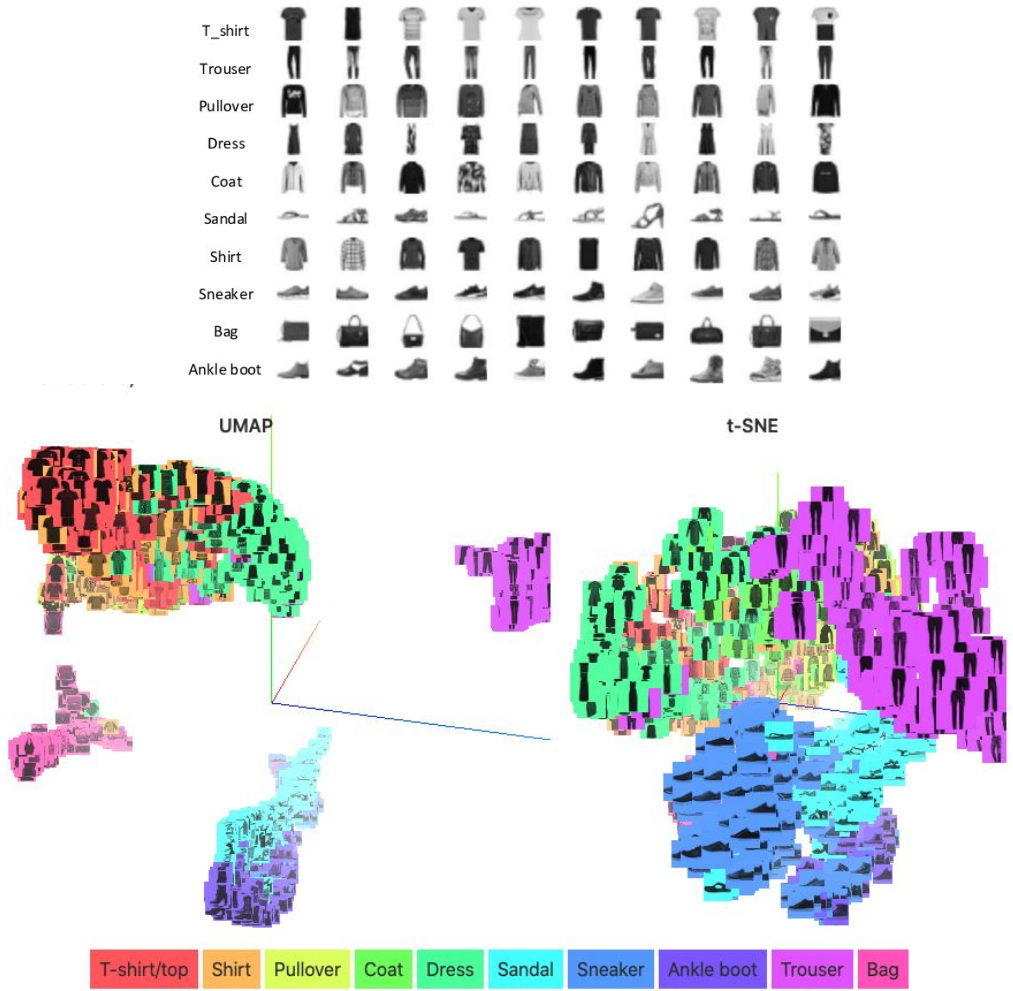


Figure 2: Dimensionality reduction applied to the Fashion MNIST dataset. 28x28 images of clothing items in 10 categories are encoded as 784-dimensional vectors and then projected to 3 using UMAP and t-SNE.

Reco Engines

Recommendation Engines

- turicreate (formerly graphlab) is a great toolkit for all sorts of ML tasks
  - Graphlab was purchased by apple to embed ML ops
- Going back to distance
  - scipy **cdist** to calculate distances between two "databases"
  - Intuition: A reference database and an "item" of interest
    - Calculate the distance between item and the database
    - Get the N closest items from the database
    - Those N items are what you would recommend
    - Can think of this as information retrieval
  - The feature space can be represented by embeddings!

Let’s write some code!