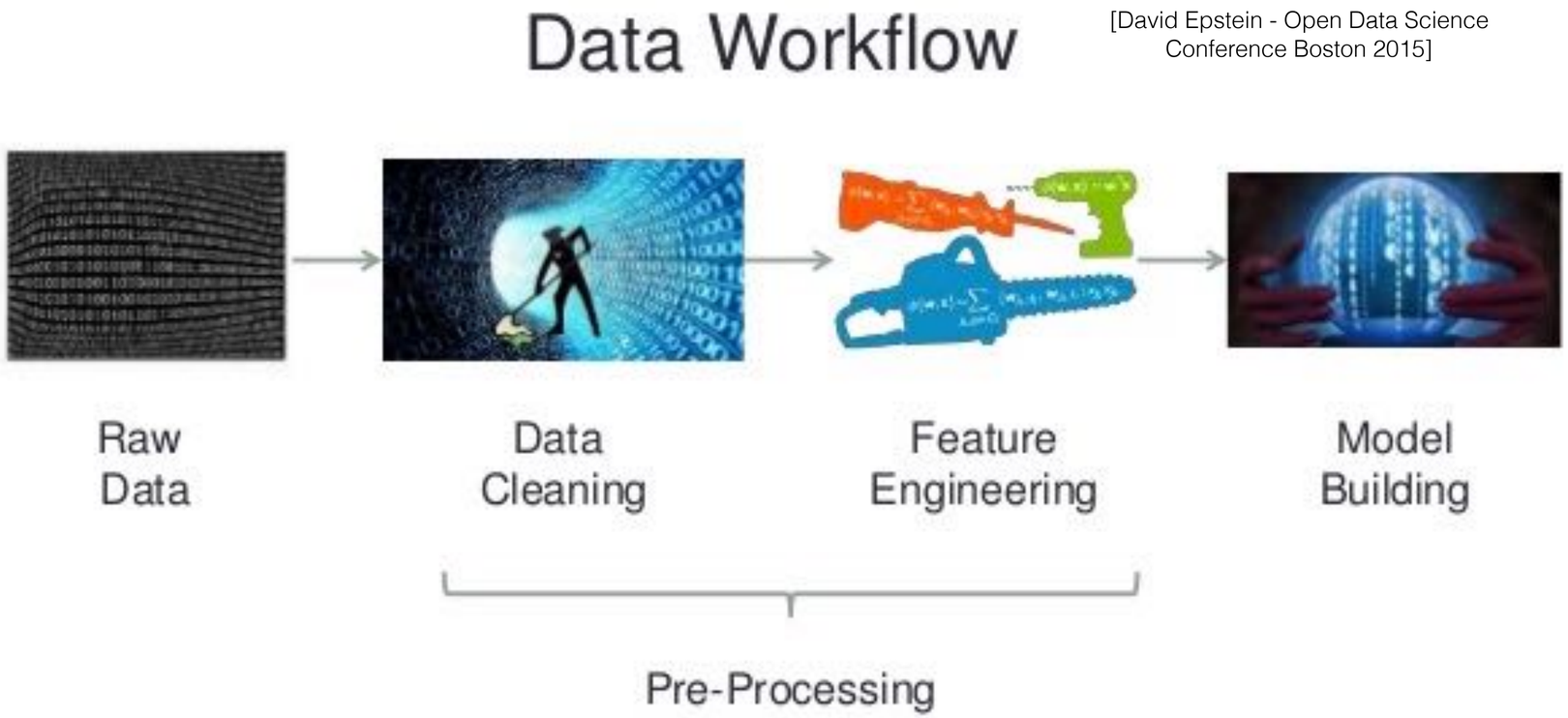
FIRST STEPS IN FEATURE ENGINEERING



[David Epstein - Open Data Science Conference Boston 2015]

Karlsruhe Machine Learning University Group
February 23 2017

CHRISTIAN STAUDT - Researcher | Developer | Consultant - Independent Data Scientist

DEFINITION

"Feature engineering is the process of **using domain knowledge of the data to create features that make machine learning algorithms work**. Feature engineering is fundamental to the application of machine learning, and is both **difficult and expensive.** "
— Wikipedia

"You have to turn your inputs into things the algorithm can understand"
— answer to "What is the intuitive explanation of feature engineering in machine learning?" on quora.com

IMPORTANCE

"Feature engineering is another topic which doesn't seem to merit any review papers or books, or even chapters in books, but it is absolutely vital to ML success. […] **Much of the success of machine learning is actually success in engineering features** that a learner can understand.

— Scott Locklin, in "Neglected machine learning ideas"

"The results you achieve are a factor of the model you choose, the data you have available and the features you prepared. "

— Jason Brownlee, 2014 on http://machinelearningmastery.com

IMPORTANCE

"[With good feature engineering] you can choose 'the wrong models' (less than optimal) and 'the wrong parameters' (less than optimal) and still get good results. Most models can pick up on good structure in data.

— Jason Brownlee, 2014 on http://machinelearningmastery.com

EXAMPLE: TITANIC SURVIVORS

Classification Task: Predict who survived the sinking of the Titanic

Hypothesis: Could it be that social class/status had something to do with it?

Hypothesis: Maybe social class/status is reflected in the passengers names?

Feature Engineering: Extract titles from names to create new feature

EXAMPLE: TITANIC SURVIVORS

```
> table(combi$Title)
        Capt          Col          Don         Dona           Dr     Jonkheer         Lady
           1            4            1            1            8            1            1
       Major       Master         Miss         Mlle          Mme           Mr          Mrs
           2           61          260            2            1          757          197
          Ms          Rev          Sir  the Countess
           2            8            1            1
```

http://trevorstephens.com

EXAMPLE: ENCODING TIME OF DAY

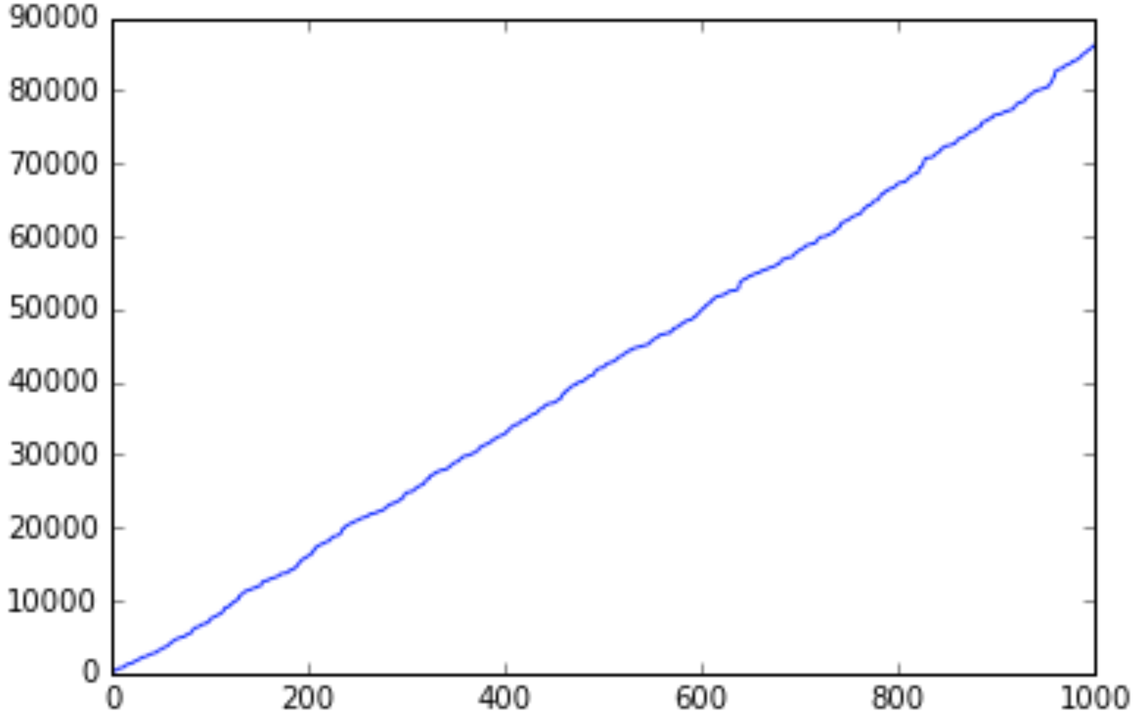
Raw attribute: time of day as string, e.g.

“23:55”
“00:05”
“17:30”

Representation 1: convert strings to number of seconds after 00:00

Now 23:55 and 00:05 look 23 hours and 50 minutes apart to the algorithm!

How to preserve the cyclical nature of these timestamps?



[Ian London’s Blog]

EXAMPLE: ENCODING TIME OF DAY

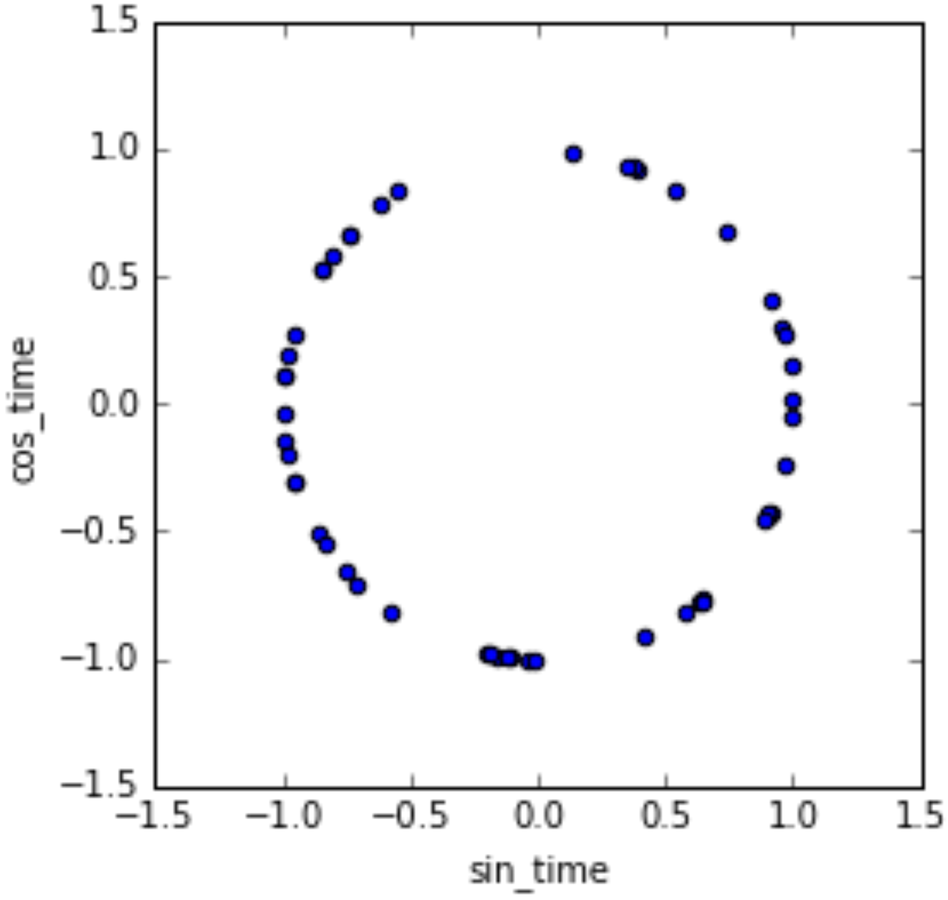
Representation 2: Apply sin/cos transformation to seconds to create 2 new features

```
seconds_in_day = 24*60*60

df['sin_time'] = np.sin(2*np.pi*df.seconds/seconds_in_day)
df['cos_time'] = np.cos(2*np.pi*df.seconds/seconds_in_day)
```

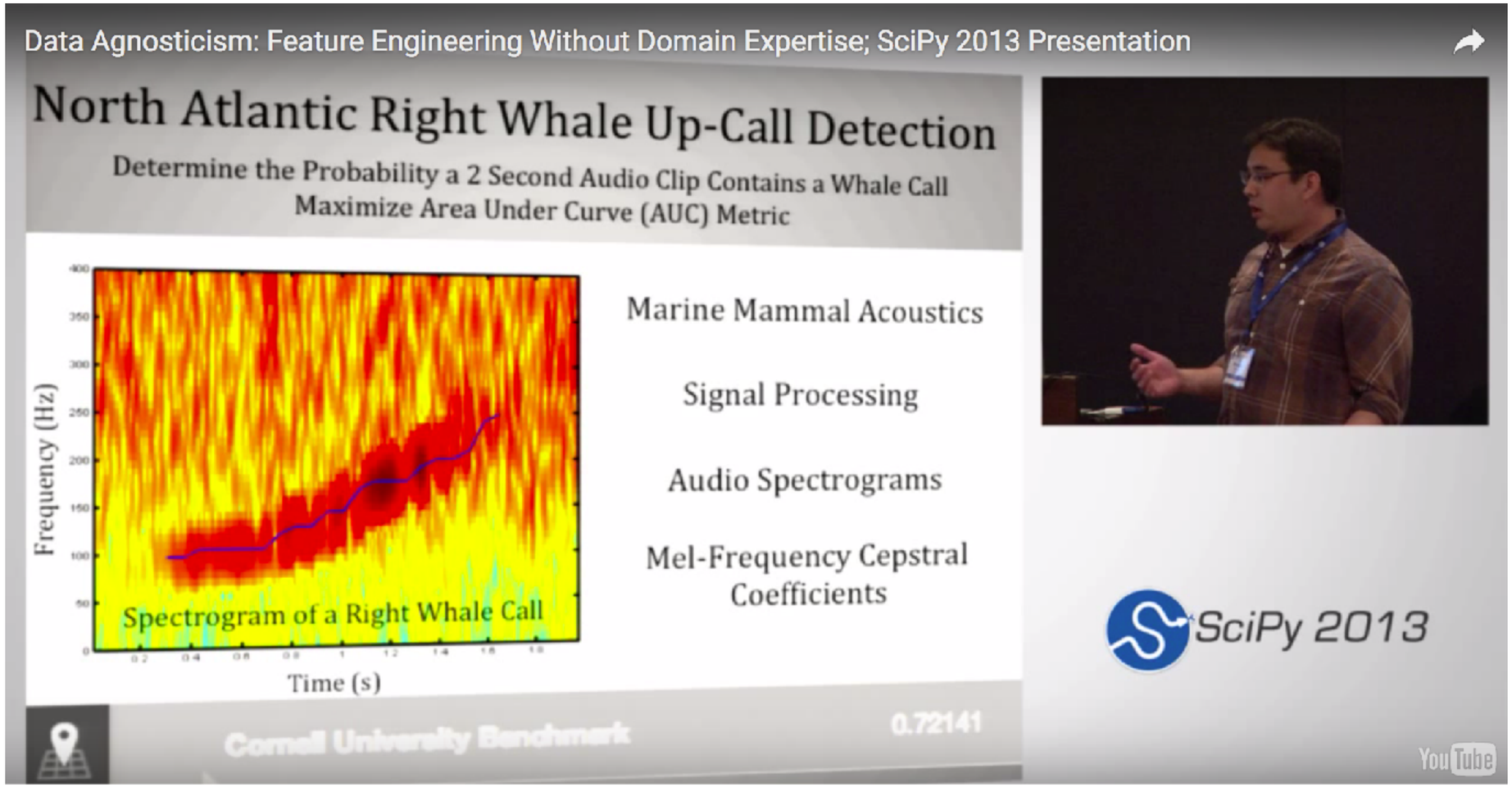
Intuitively, arrange time of day timestamps on a circle

Now, 23:55 and 00:05 are close together.



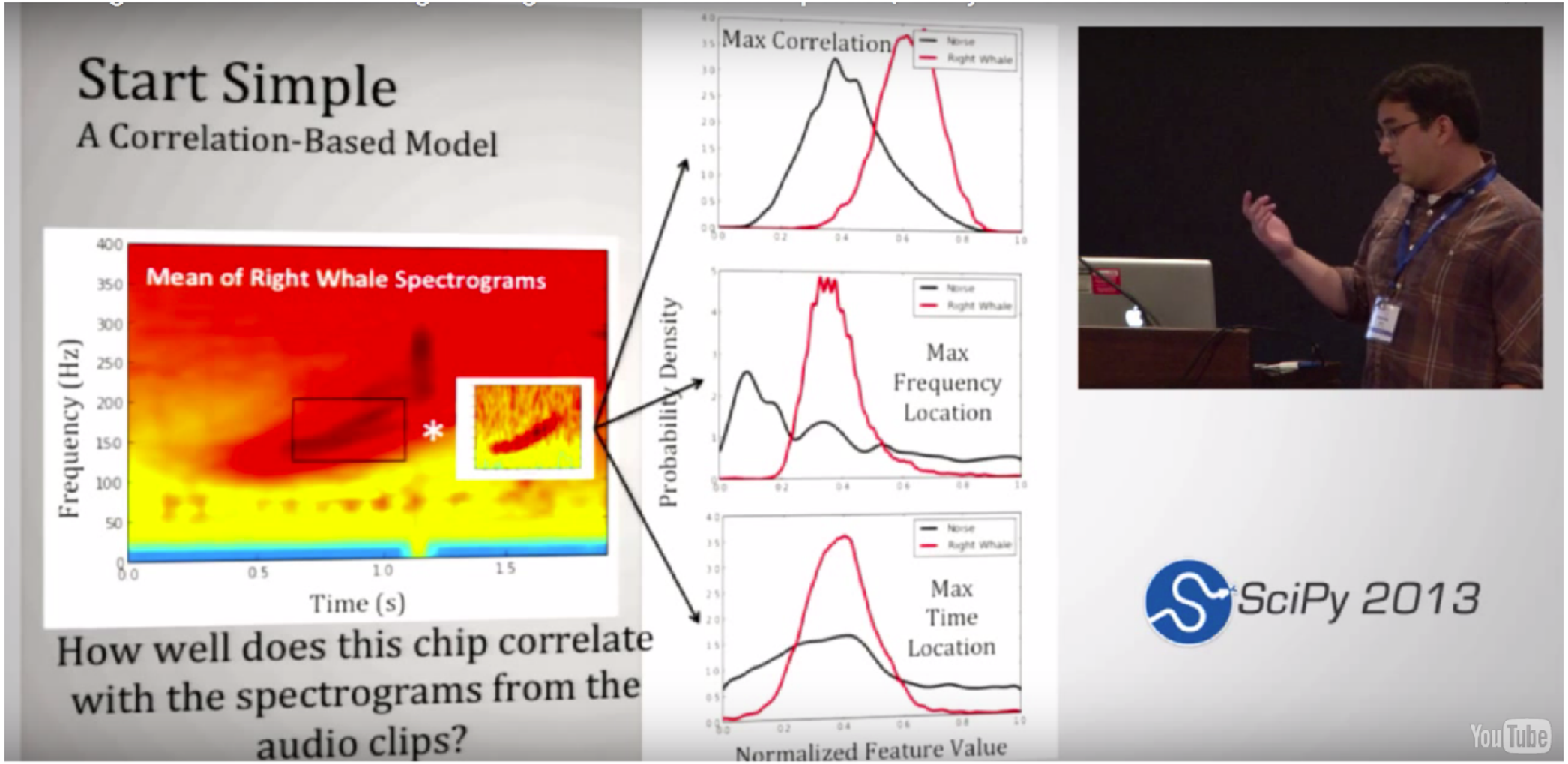
[Ian London's Blog]

EXAMPLE: DETECTING WHALE CALLS



Nicholas Kridler @ SciPy 2013 - team won Marinexplore Right Whale Detection Kaggle challenge

EXAMPLE: DETECTING WHALE CALLS



EXAMPLE: DETECTING WHALE CALLS

- audio clips (waveforms)
- create spectrograms
- average spectrograms
- compare spectrogram with average
- generate features like
  - max correlation
  - max frequency location
  - max time location
- feed to ML algorithm (random forest)

CREDITS

SOURCES

- **http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/**
- **https://www.youtube.com/watch?v=bL4b1sGnILU**
- **http://trevorstephens.com/kaggle-titanic-tutorial/r-part-4-feature-engineering/**
- **https://ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/**

CONTACT

+4917623351772
mail@clstaudt.me
http://clstaudt.me

CONNECT

linkedin.com/in/christian-staudt/
researchgate.net/profile/Christian_Staudt
stackoverflow.com/users/626537/
github.com/clstaudt