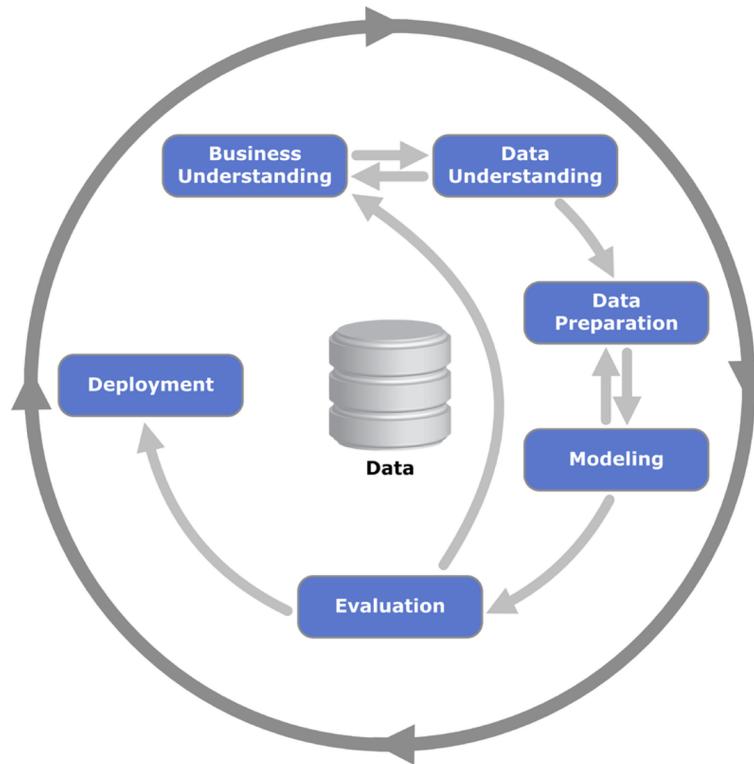


# DATA 601

Modeling in Data Science

# Data Science Life Cycle



# What is Modeling? Depends on Context...

- A [mathematical model](#) is a description of a system using mathematical concepts and language.
- A [statistical model](#) is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population). A statistical model represents, often in considerably idealized form, the data-generating process.
- A [machine learning model](#) is a file that has been trained to recognize certain types of patterns.

# Machine Learning Modeling

## The Traditional Programming Paradigm:



## Machine Learning



**Figure 1:** Machine learning vs. “classic” programming.

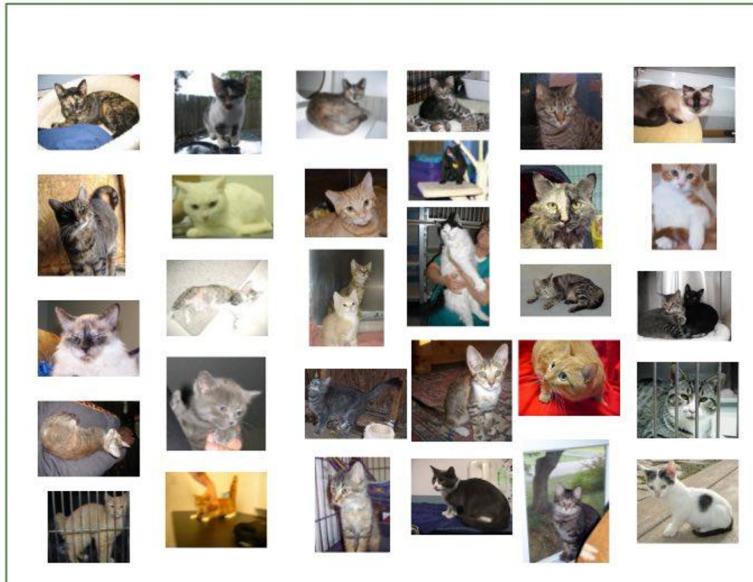
# Formal Definition of learning in ML:

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. (Tom Mitchell)

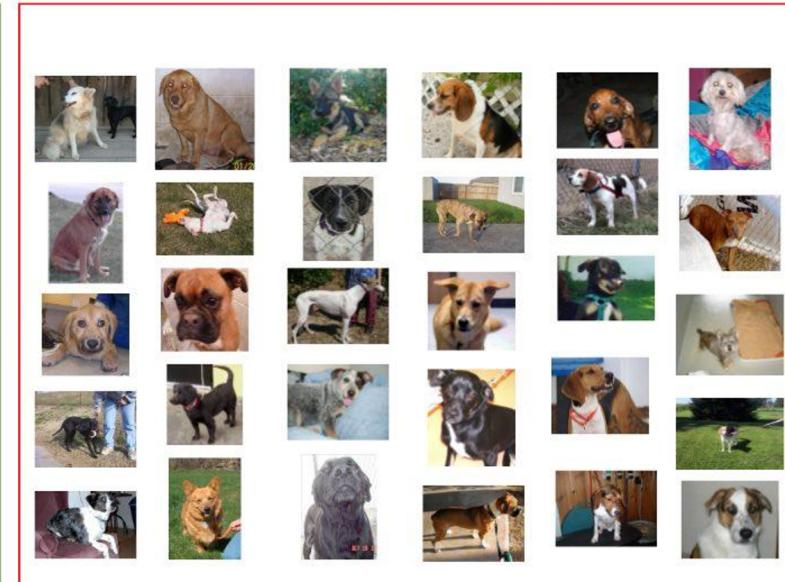


# Cats vs Dogs?

Cats



Dogs



Sample of cats & dogs images from Kaggle Dataset

# Types of Problems in ML

## Supervised Learning

- Data is labeled
- Goal: Predict (Values, classes, labels, etc.) for unseen cases.

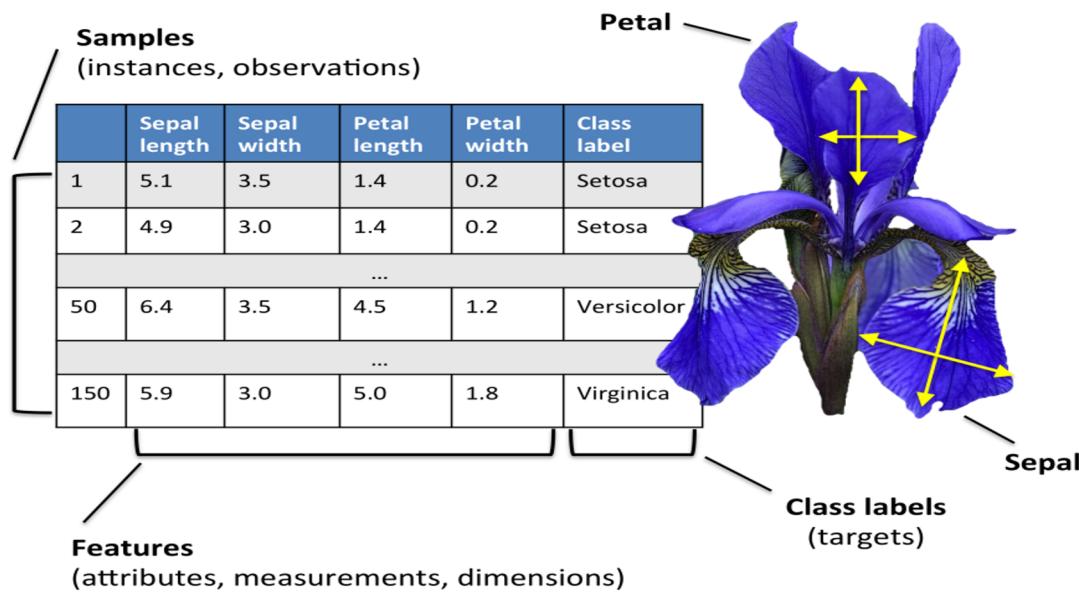
## Unsupervised Learning

- Data is unlabeled.
- There is no ground truth.
- Goal: Detect hidden/Previously unknown structure in data.

## Reinforcement Learning

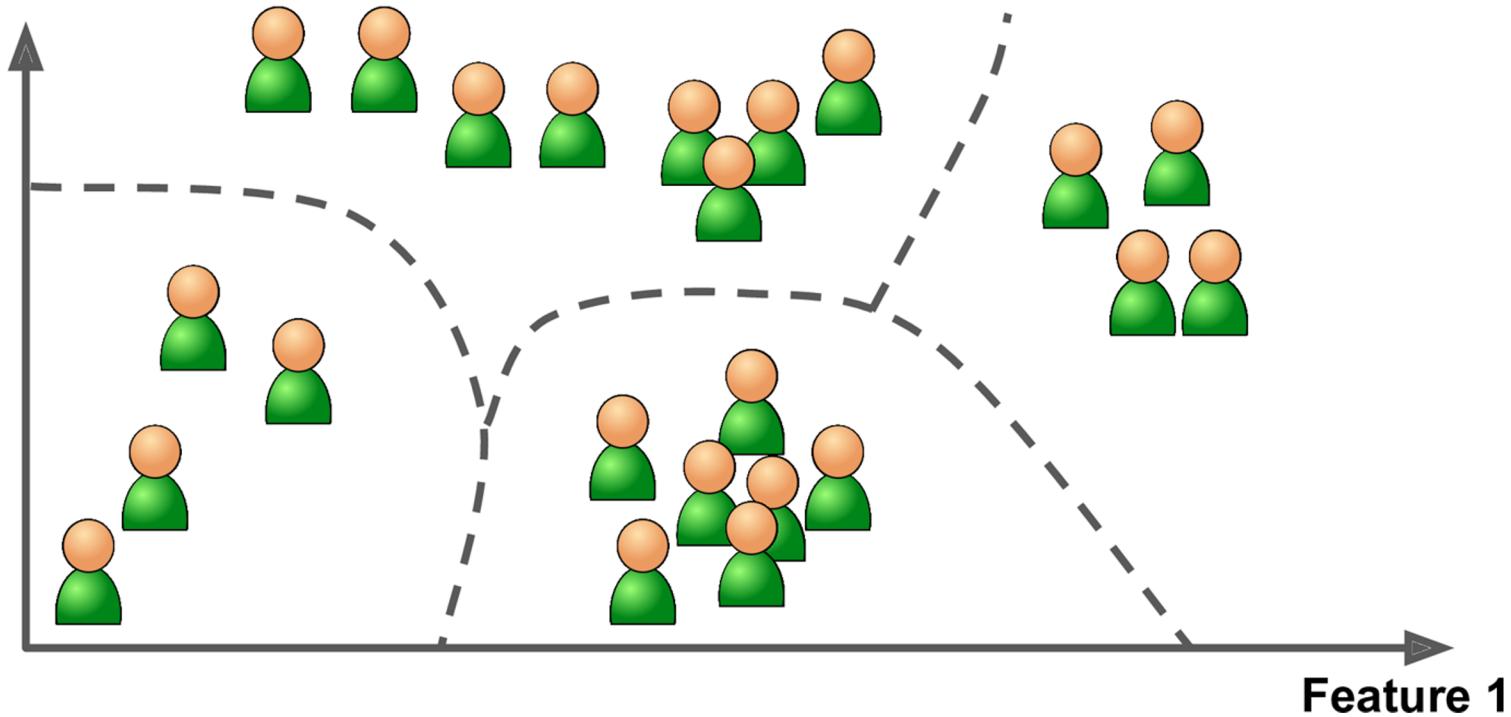
- Data collected by an agent during the learning procedure.
- Learning happens with reward/punishments.
- Goal: Find actions that maximizes total reward.

# Supervised Learning Problem

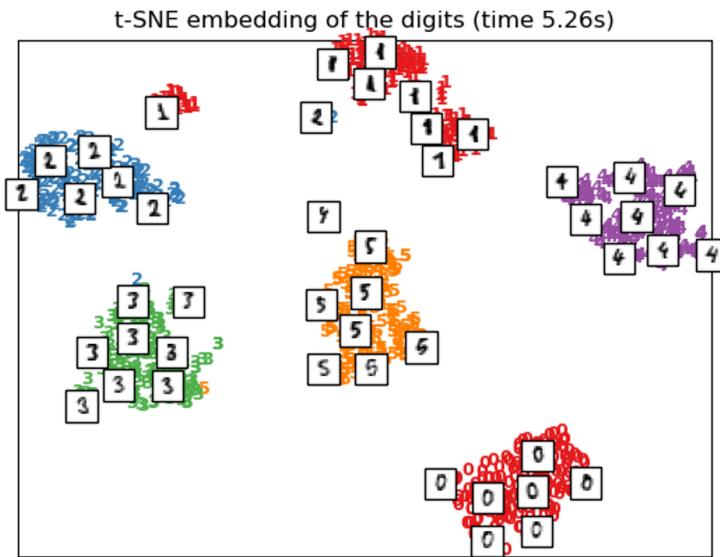


# Unsupervised Learning Problem: Clustering

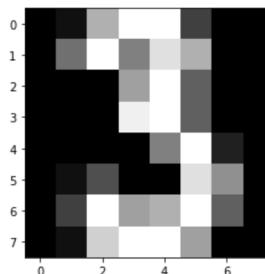
Feature 2



# Unsupervised Learning: Dimension Reduction

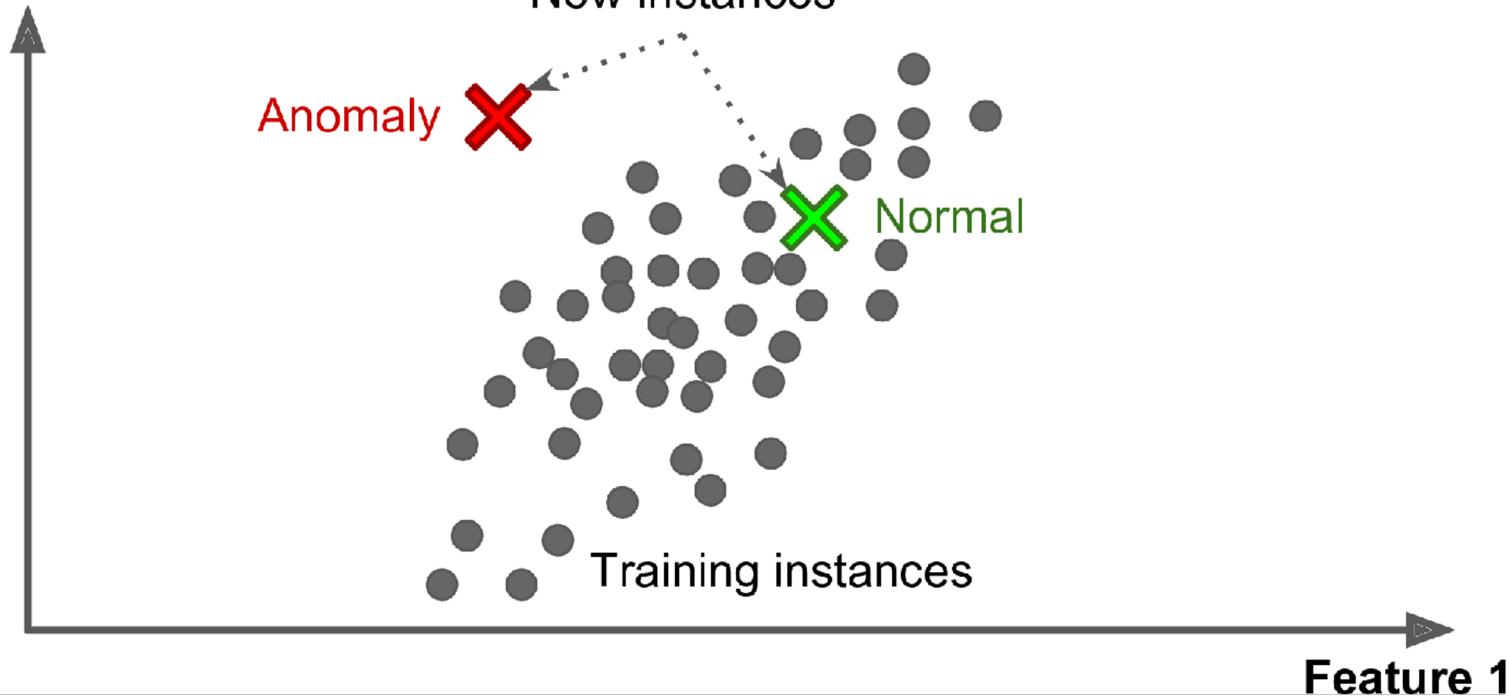


```
[ 0.  1. 11. 16. 16. 4. 0. 0. 0. 0. 7. 16. 8. 14. 11. 0. 0. 0. 0.  
 0. 10. 16. 6. 0. 0. 0. 0. 15. 16. 6. 0. 0. 0. 0. 0. 0.  
 8. 16. 2. 0. 0. 1. 5. 0. 0. 14. 9. 0. 0. 4. 16. 10. 11. 16.  
 6. 0. 0. 1. 13. 16. 16. 10. 0. 0.] 3
```

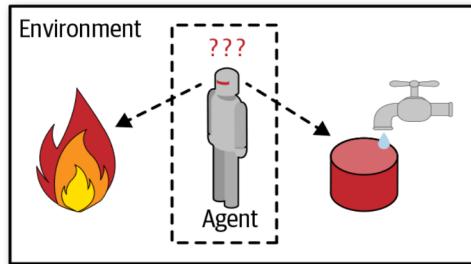


# Unsupervised Learning: Anomaly Detection

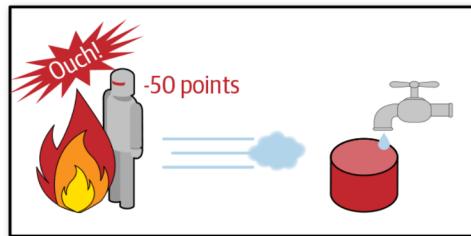
Feature 2



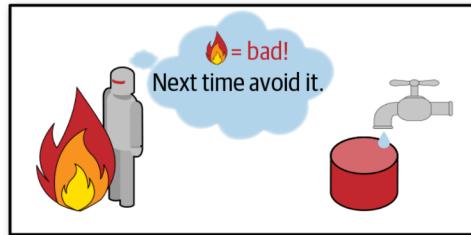
# Reinforcement Learning



- 1 Observe
- 2 Select action using policy



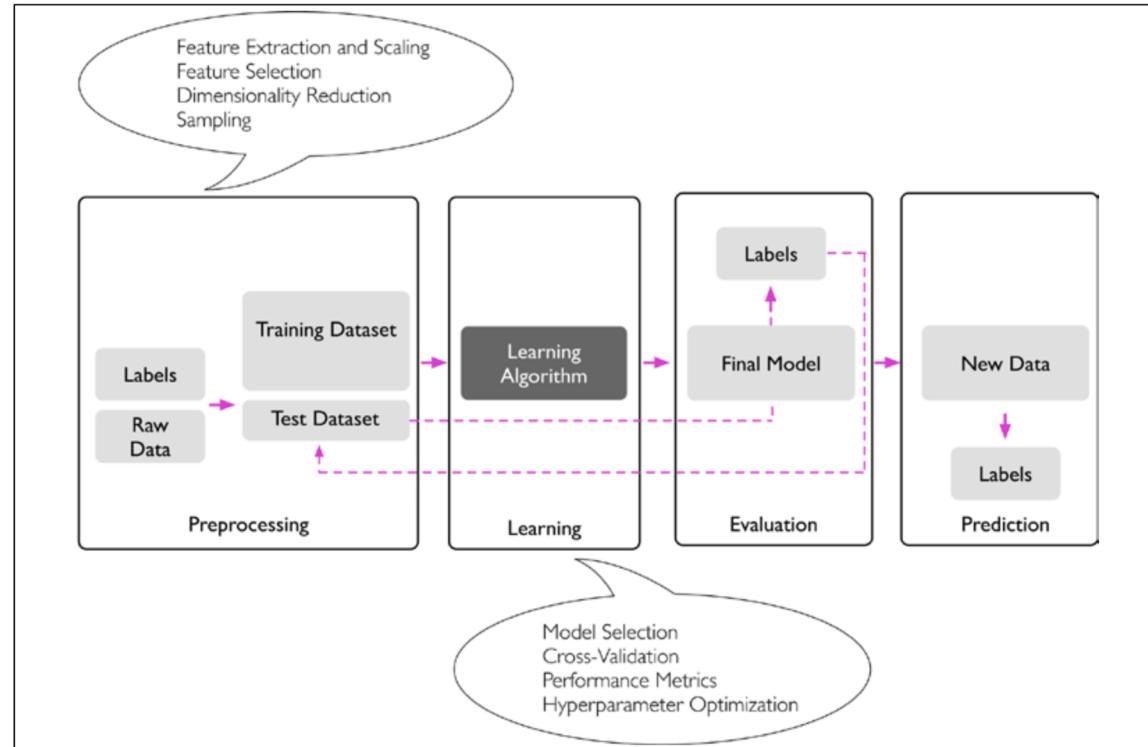
- 3 Action!
- 4 Get reward or penalty



- 5 Update policy (learning step)
- 6 Iterate until an optimal policy is found



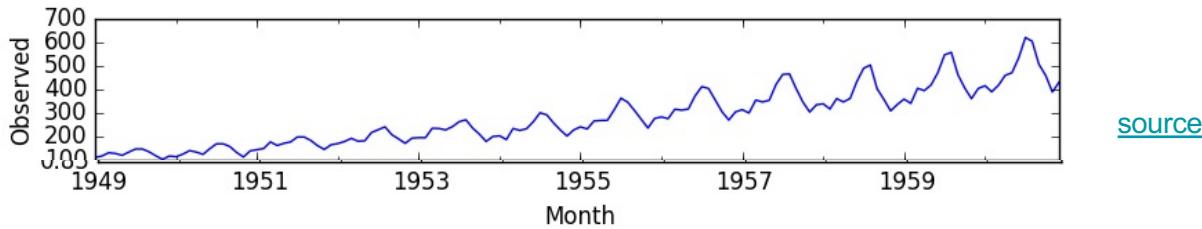
# Procedures to Create ML Models





# Time Series Continued

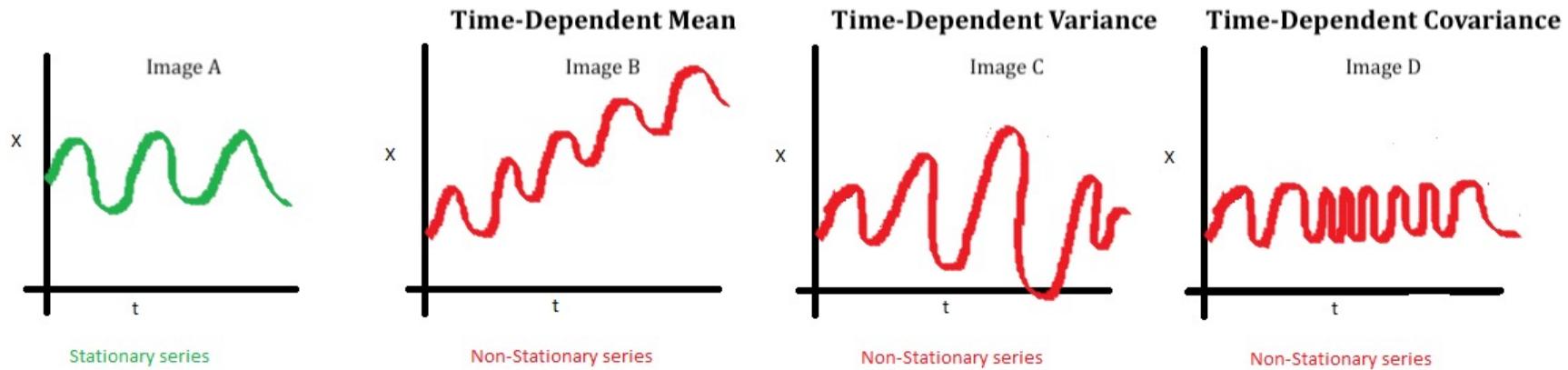
# Temporal data has multiple factors present



"The Airline Passengers dataset describes the total number of airline passengers over a period of time.

The units are a count of the number of airline passengers in thousands. There are 144 monthly observations from 1949 to 1960."

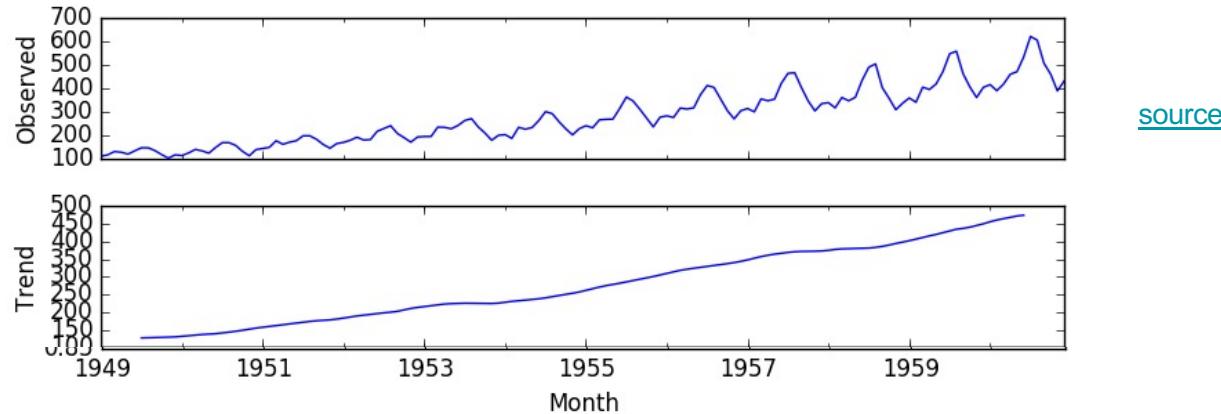
# Stationary = mean, variance, covariance, do not vary with time



An illustration of the principles of stationarity, Source: [BeingDatum](#)

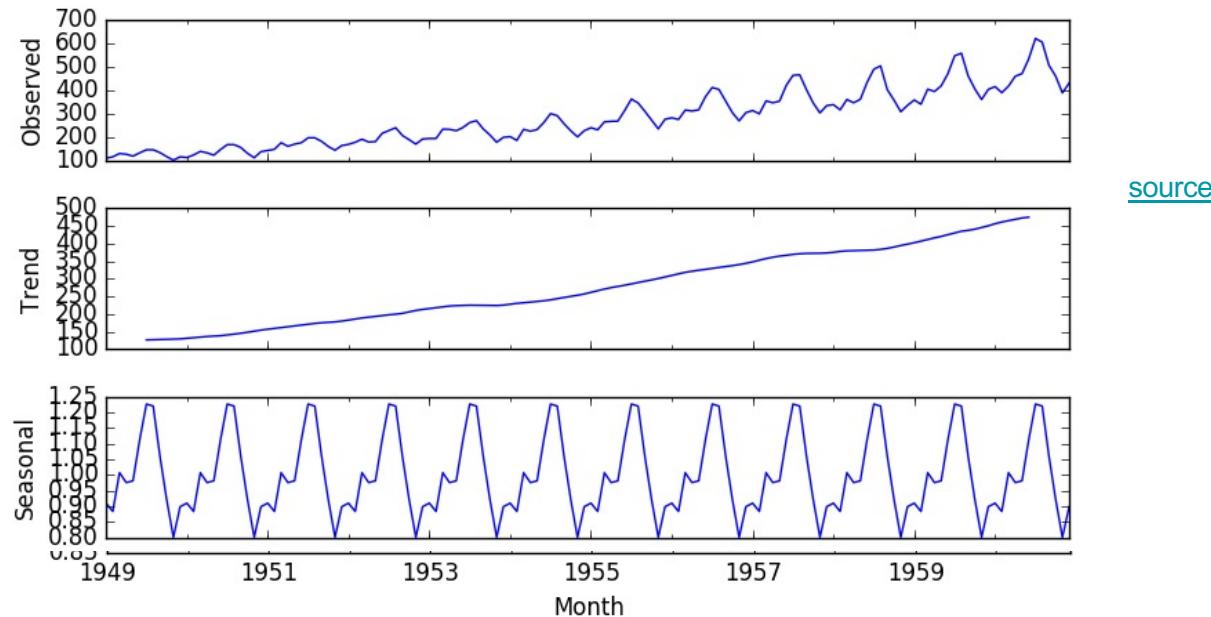
# Motivation: decomposing temporal data

Step 1: separate the trend

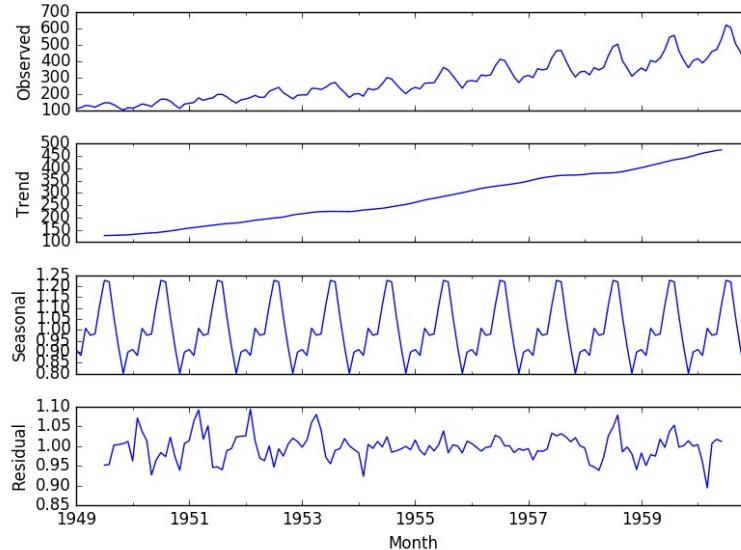


# Motivation: decomposing temporal data

Step 2: separate seasonality



# Motivation: decomposing temporal data



[source](#)

# Example

- historical power df\_analysis.ipynb

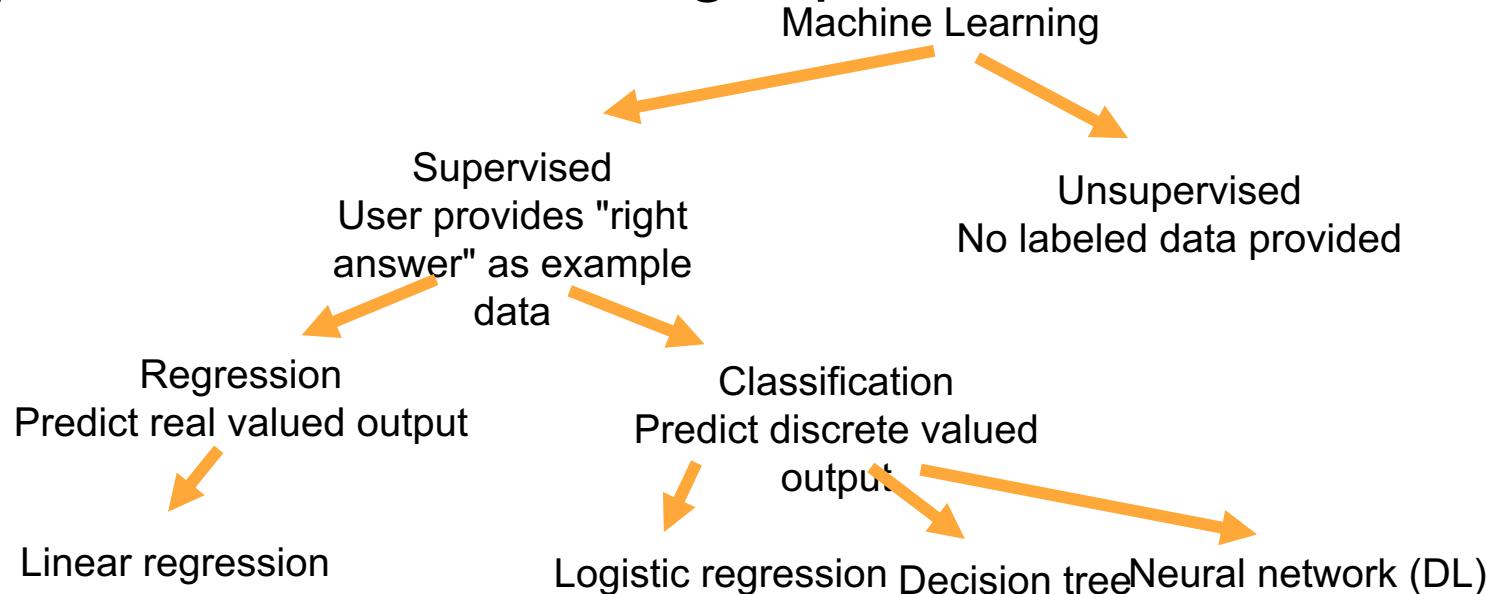
# Methods to decompose temporal data

Trend removal

- Differencing
- Rolling average
- Linear regression

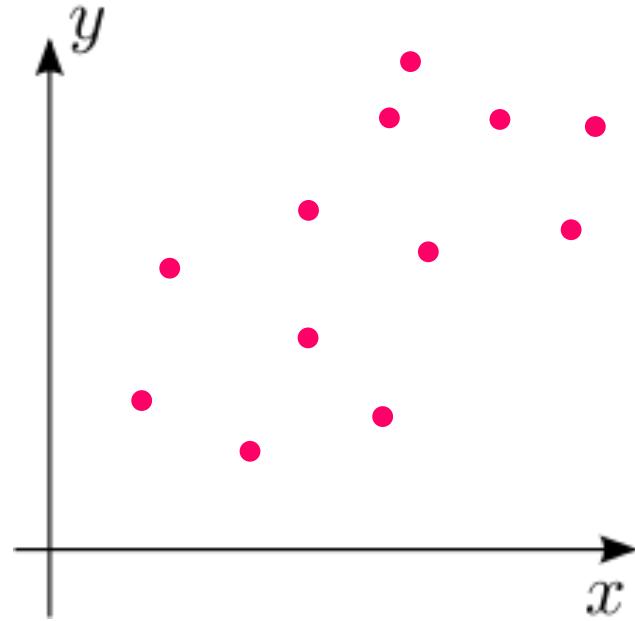
- Decomposing time series
- Linear regression for trend analysis
- Fourier transform
- Homework

# Map of Machine Learning topics



## Activity: Draw best fit straight line on plot

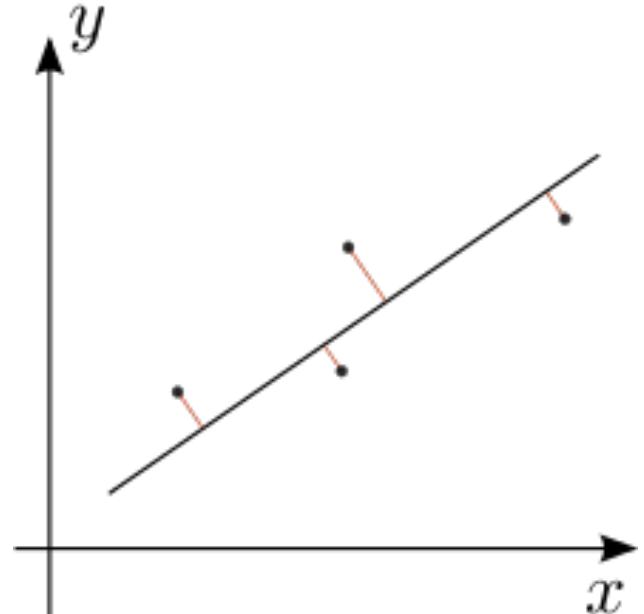
Need a volunteer who is not in Data 602



# What you (probably) did

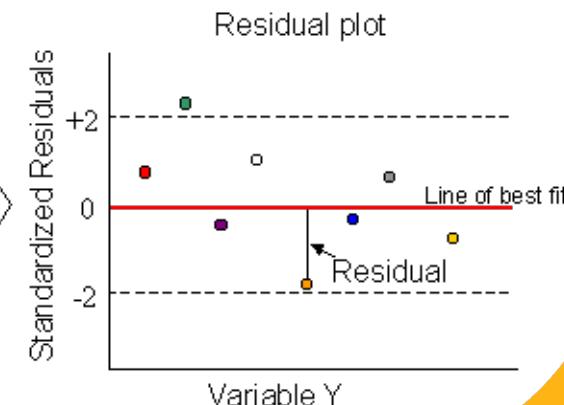
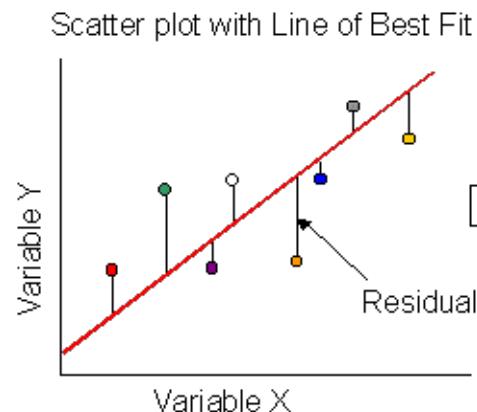
If there is uncertainty in both the  $x$  and  $y$  coordinates, then use an approach which admits variation in both

[https://en.wikipedia.org/wiki/Deming\\_regression](https://en.wikipedia.org/wiki/Deming_regression)



# Linear Regression concepts

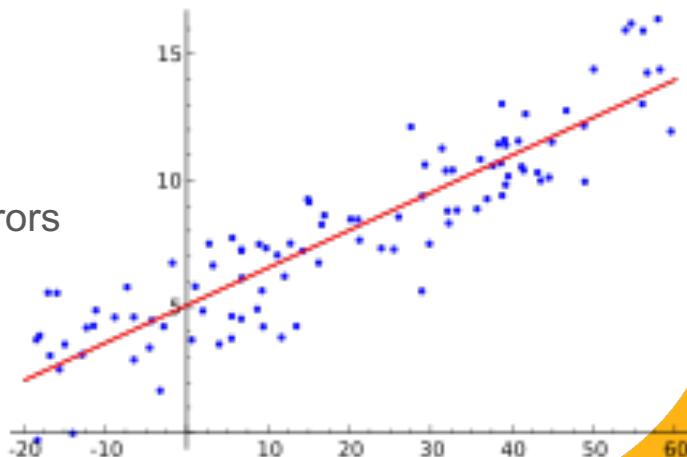
- Predict a target variable by fitting the *best linear relationship* between the dependent (Y) and independent (X) variable.
- The *best fit* is done by making sure that the sum of all the distances between the shape and the actual observations at each point is as small as possible.



Residual = difference between what the current model gives us and the "right" output

Best approach to minimize the residual depends on your data

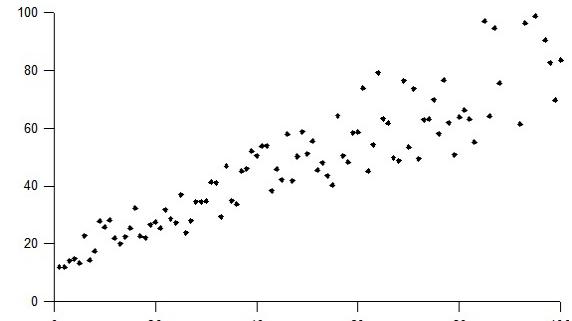
- ordinary least squares for independent and identically distributed errors
- generalized least squares for arbitrary covariance
- weighted least squares for heteroskedastic errors  
(error in y varies with x)
- feasible generalized least squares with autocorrelated errors  
(function repeats)



Residual = difference between what the current model gives us and the "right" output

Best approach to minimize the residual depends on your data

- ordinary least squares for independent and identically distributed errors
- weighted least squares for heteroskedastic errors  
(error in y varies with x)
- generalized least squares for arbitrary covariance
- feasible generalized least squares with autocorrelated errors  
(function repeats)



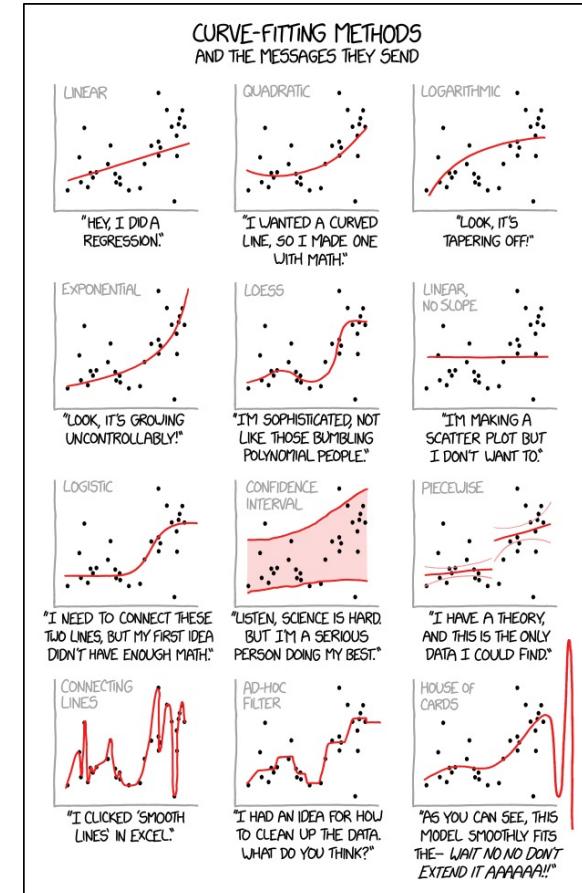
Residual = difference between what the current model gives us and the "right" output

Best approach to minimize the residual depends on your data

- ordinary least squares for independent and identically distributed errors
- weighted least squares for heteroskedastic errors  
(error in y varies with x)
- generalized least squares for correlated residuals
- feasible generalized least squares with autocorrelated errors  
(function repeats)



<https://xkcd.com/2048>



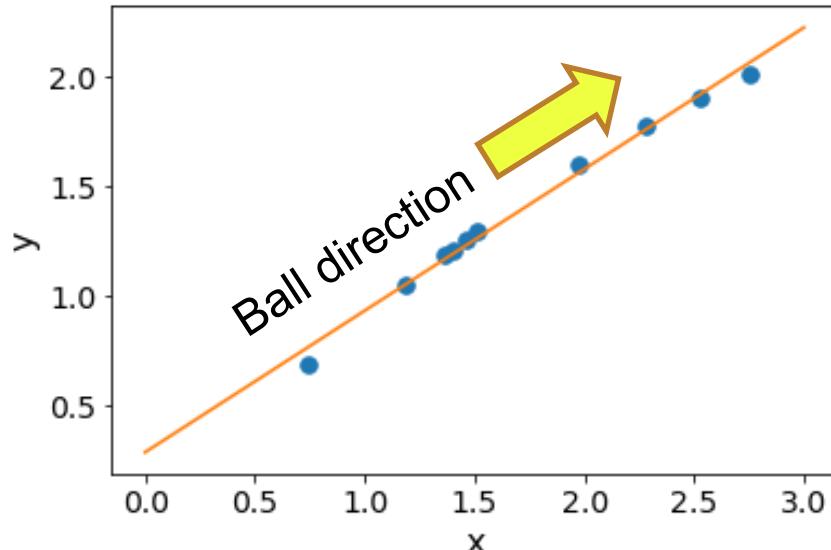
# Linear regression = fit data, but for what?

- Prediction
  - Extrapolation = independent variable (x) being evaluated is outside the range of values you already have information on
  - Interpolation = value you are evaluating is within the range of values you already know
- Fitting Trends
- Measure Correlation of two variables

# Physics experiment: ski jump



# Projectile motion after ball launches

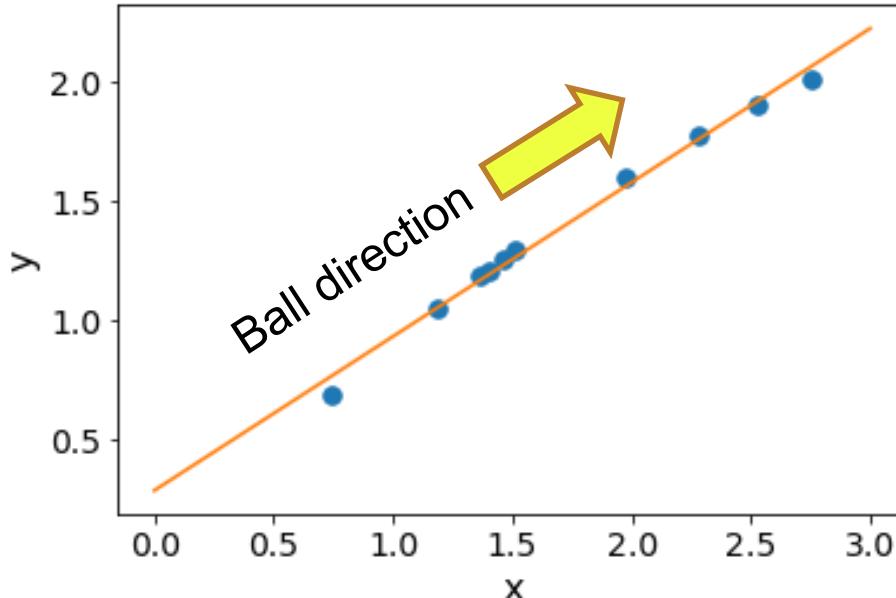


Ramp  
ends here

X and Y correlation = 0.996179

$Y = 0.64672026 \cdot X + 0.28288071$

## Projectile motion *activity*: identify 2 problems



Ramp  
ends here

X and Y correlation = 0.996179

$Y = 0.64672026 \cdot X + 0.28288071$

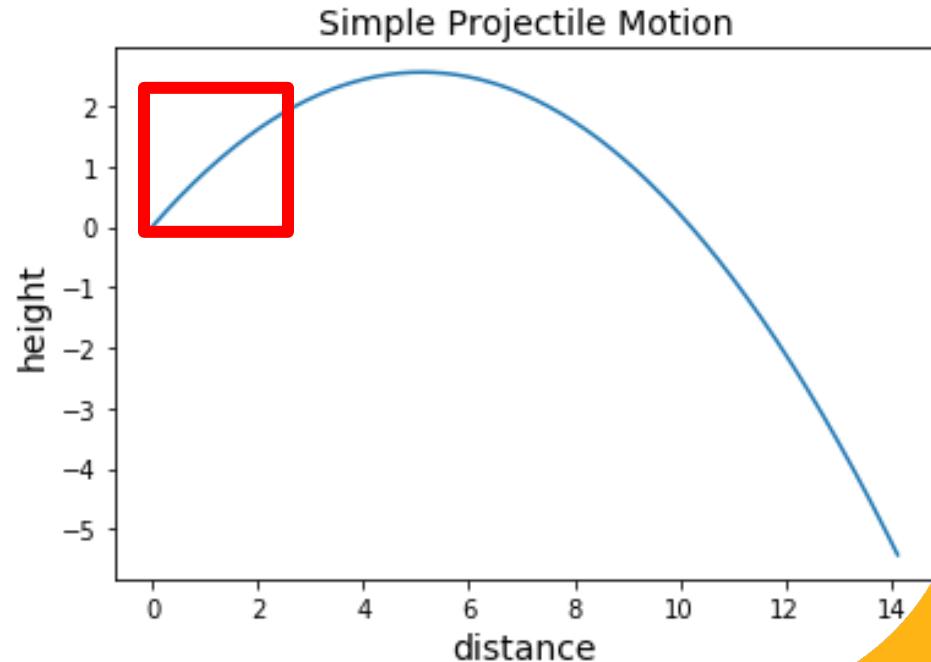
Raise your hand if  
you have an  
observation

# Issues

- Insufficient sampling in relevant range

While points within the sampling window are correct, the range of interest is the full arc

- Fitting a line to the data appears adequate, but misses the underlying physics of projectile motion: a parabolic path is a second order polynomial



## *Lesson: Simply trusting data leads to invalid conclusion*

- Consistently question your data and your own analysis
- Investigate how data was collected, by whom, for what purpose
- Additional data often helps. Can relevant data be gathered?

# statistics emphasizes inference machine learning emphasizes prediction

- Statistics: *infer the process by which data you have was generated.*
  - Machine learning: *predict what future data will look like with respect to some variable.*
- > Knowing how the data was generated matters to the story you tell

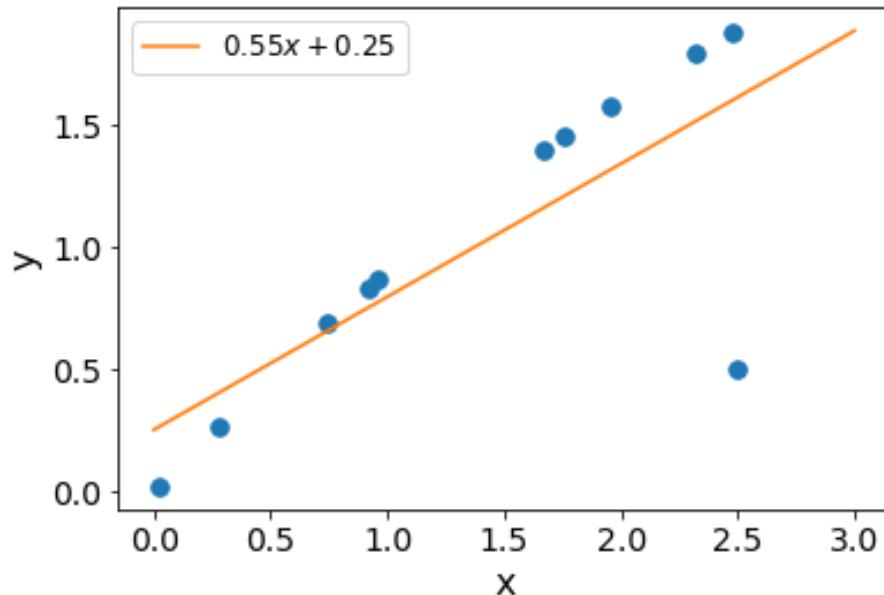
# When does linear regression apply?

- Is the data numerical? (Necessary but not sufficient.)
- The stronger the correlation coefficient is,  
the more of a linear relationship exists.

## Recall from previous lecture

- **Variance** measures width of a distribution
- **Covariance** is the measure of variance for two random variables (joint variability)
- **Correlation** is the normalized covariance, from  $-1$  to  $1$

# Danger to blindly applying linear regression: Outliers



Correlation without outlier: 0.996  
Correlation with outlier: 0.76

- Linear regression for trend analysis
- Fourier Transform
- Homework

## Fourier Transform:

- FFT only applies to data in which the timestamp is uniform

`fourier_transform.ipynb`