Welcome to DSCI 101

Introduction to Data Science

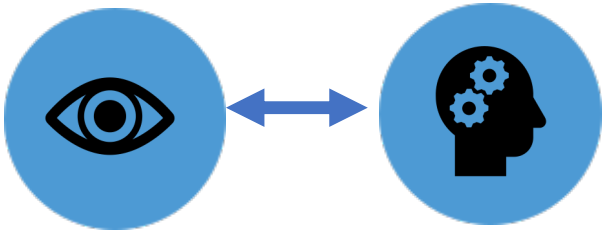
Week 6 Recap

- Data visualization and basic plots:
  - Bar plots: one categorical variable
  - Histogram: one numerical variable
  - Box plots: one categorical and one numerical variable
  - line and scatter plots: two numerical variable
- More advanced plots:
  - Overlaid plots, contour plot, 3-d plot, heatmap, maps…
- Data visualization examples
  - Df.plot

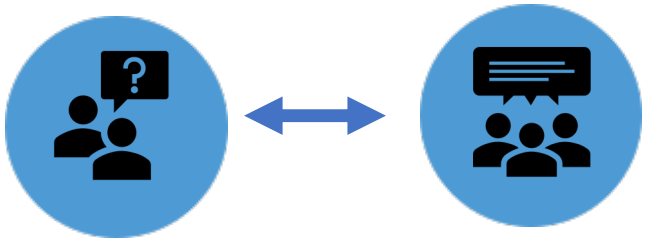
Week 7 Preview

- Exploratory Data Analysis (EDA)
  - half-way of the data science pipeline
- EDA goals and techniques
  - key properties to explore
  - more Pandas for EDA
- Real data example:
  - police traffic stop data for Houston

What is EDA?



- **A first look at the data:**
  - Not a formal process
  - No strict set of rules
  - Gain understanding
  - Maximize insights

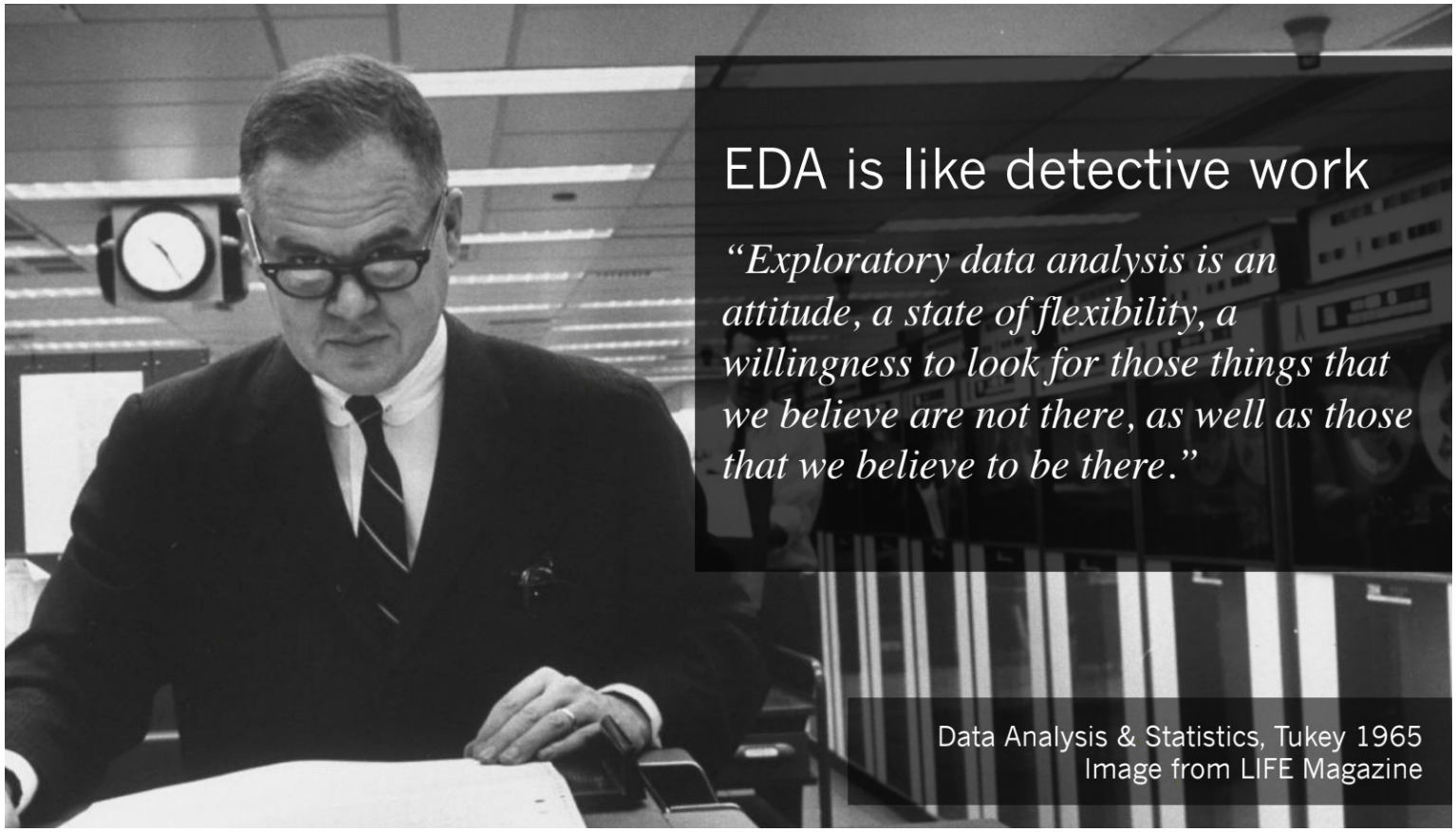


- **An iterative process:**
  - Generate questions
  - Search for answers
  - Refine questions
  - Generate new questions

A little bit of history

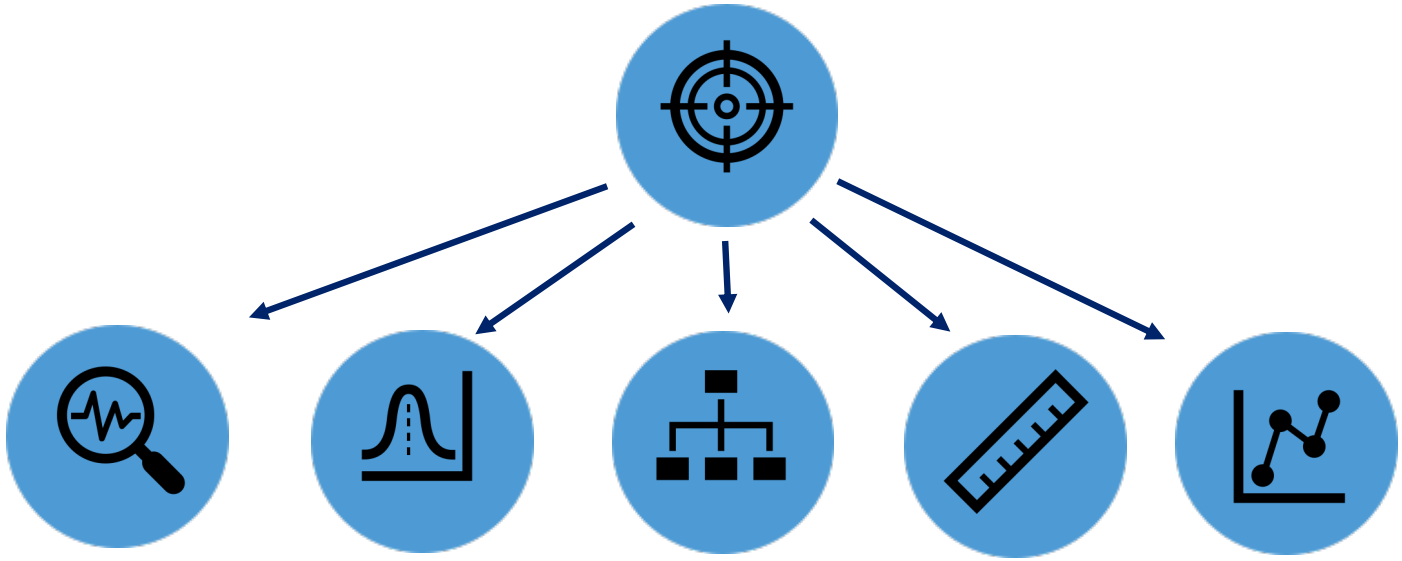
- Principally developed by John Tukey since 1970
- “to examine the data before applying a model”
- Princeton mathematician & statistician, also introduced
  - Fast Fourier transform
  - “bit”: binary digit
  - Box plot!

John Tukey


EDA is like detective work
"Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there."
Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine

EDA goals

- **General goals**
- **Specific goals**



detect outliers / anomalies/ missing

check distributions of single variable

assess relation between variables

engineer features

select appropriate models

Key properties to look at

- Structure
  - the shape of your data
- Granularity
  - the resolution of your data
- Scope
  - the features of your data
- Temporality
  - how does the data sit in time
- Origin
  - how was the data collected

Three Aspects to Explore

- **Variation within one variable (univariate)**
  - visualizing distributions
  - typical values vs. extreme values
  - missing values
- **Covariation between two variables (bivariate)**
  - association between the two
  - marginal vs. joint distribution
- **Data patterns (multivariate)**
  - signal or noise?
  - relation implied by the pattern?
  - pattern within subgroups of the data?

Some EDA Techniques

- **Summary, tables and graphics**
  - summary statistics, pivot table
  - bar plot, histogram, box plot
  - scatter plot, heat map…
- **Data pattern**
  - dimension reduction
  - clustering
- **Feature engineering**
  - manipulate the columns of your data

Feature engineering

- Filter / select features

- Create / transform new features based on existing ones
  - useful during the modelling phase, more on this later
  - “the process of transforming the representation of model inputs to enable better model approximation”
  - encode non-numeric features

Some examples of feature engineering

- Take average of closely related variables
- Split strings into multiple features
- Categorize numerical variables
- Create dummy variables from categorical variables
- Extract features from free text, image, signal…

Take away message

- **Think about the big picture**
  - what is the ultimate goal
  - how does EDA fit into the pipeline

- **Ask quality questions**
  - start with some questions
  - refine your questions along the way

- **Be creative**
  - think out of the box
  - art and science

Demo