Political Science 15
Introduction to Research in Political Science
Lecture 10a: What is Data Science?

Alice Lépissier
University of California Santa Barbara

Agenda

- Broaden your horizons and explore your opportunities
- What is data science? What does a data scientist do?
- The data science workflow
  1. Data wrangling
  2. Data analysis
  3. Data visualization

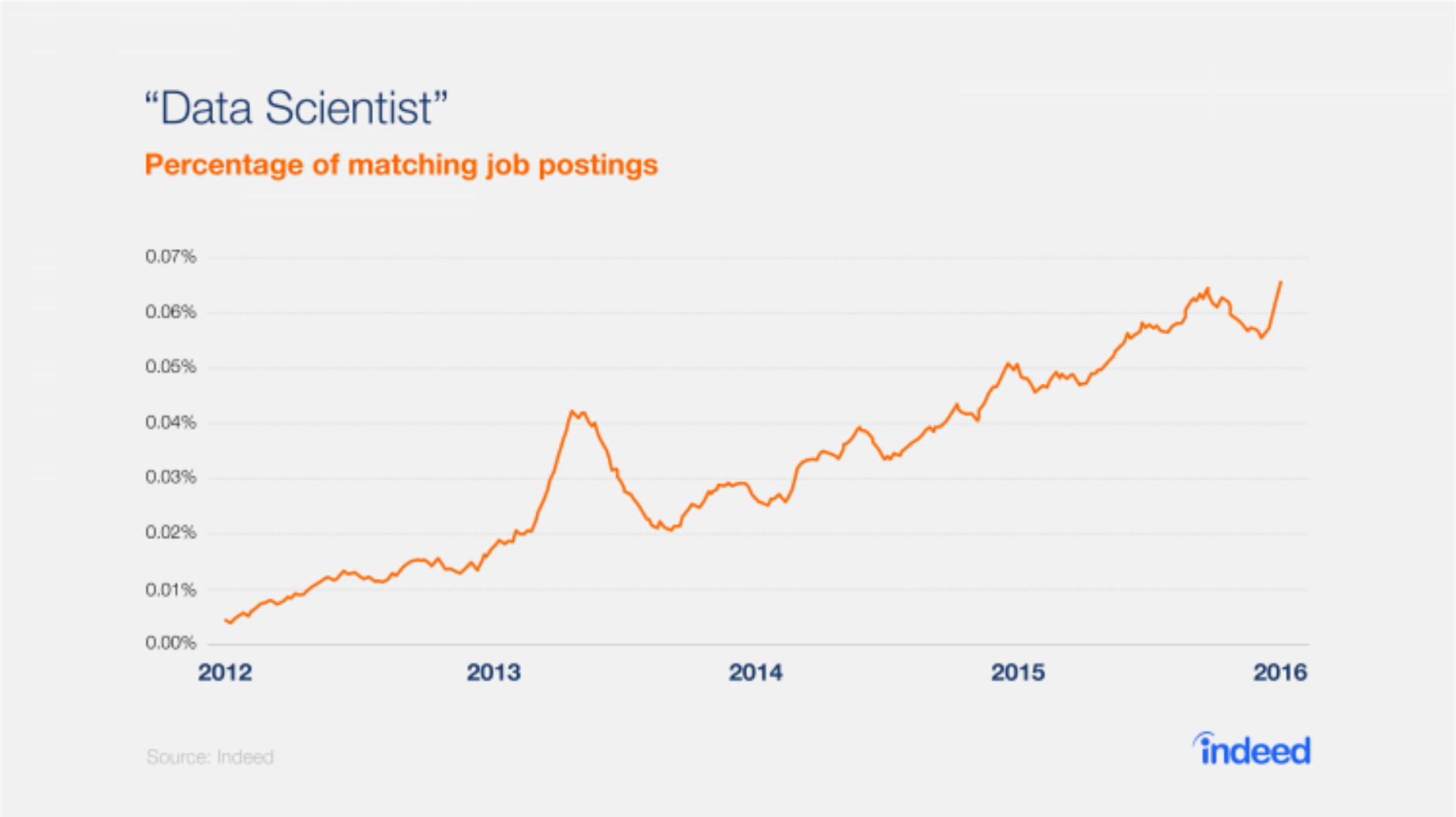
Data scientist: the sexiest job of the 21st century

In 2012, the Harvard Business Review called data scientist "the sexiest job of the 21st century" (https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century).

A data scientist is someone who can **extract insights from large amounts of unstructured data**.

Businesses, governments, and civil society are generating tons of data each day (e.g. ad clicks, satellite observations, tweets, etc.). The constraint to understanding how our world works is no longer having access to information, but having the analytical horsepower to process and generate knowledge from the data.

The demand for data scientists has exploded



Data scientist is listed as most popular job in America

50 Best Jobs in America for 2019

Best Jobs | 2019 | United States

	Job Title	Median Base Salary	Job Satisfaction	Job Openings
#1	Data Scientist	$108,000	4.3/5	6,510

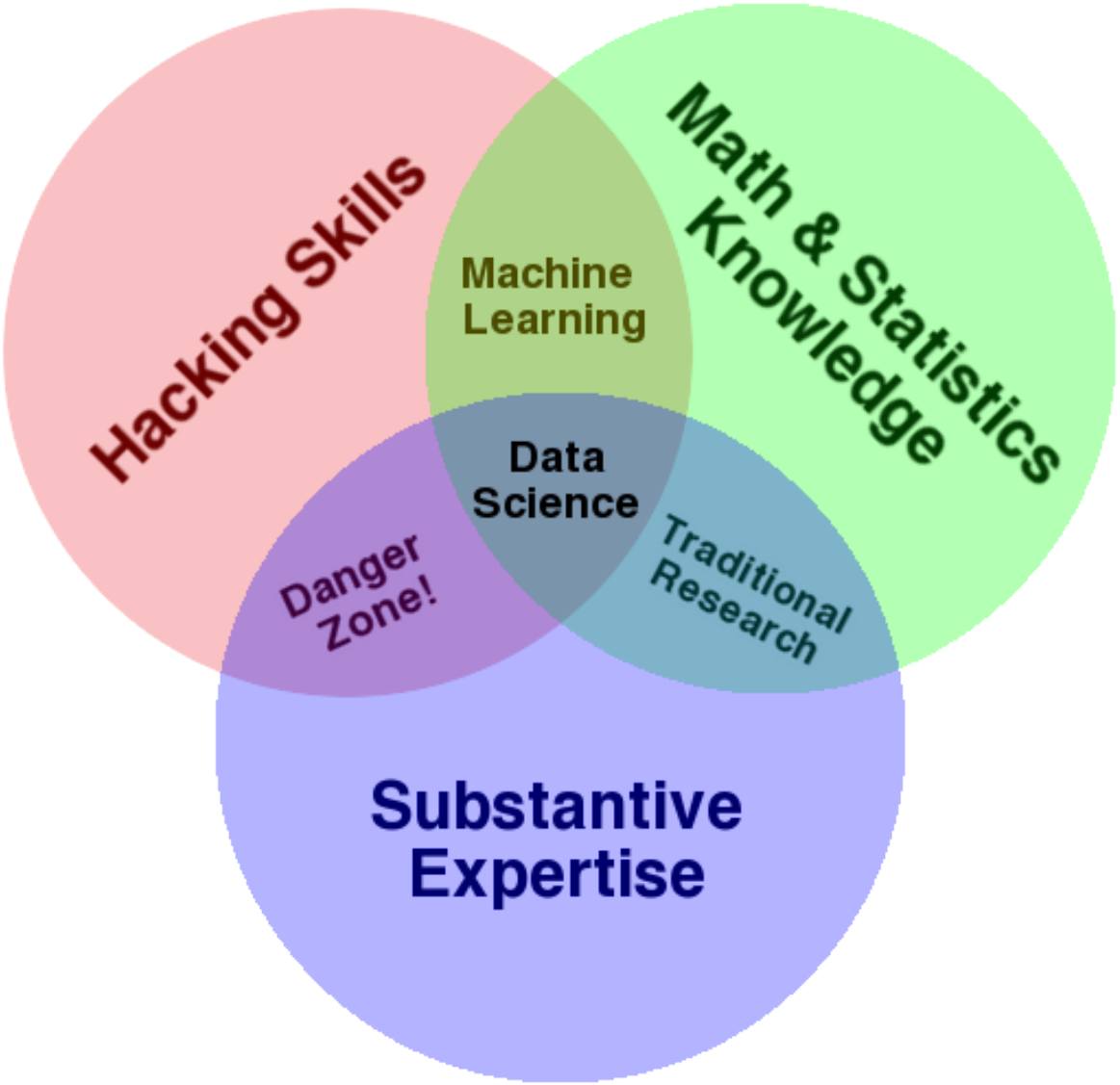
Source:
https://www.glassdoor.com/List/Best-Jobs-in-America-2019-LST_KQ0,25.htm.

What is data science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data (Dhar, 2013).

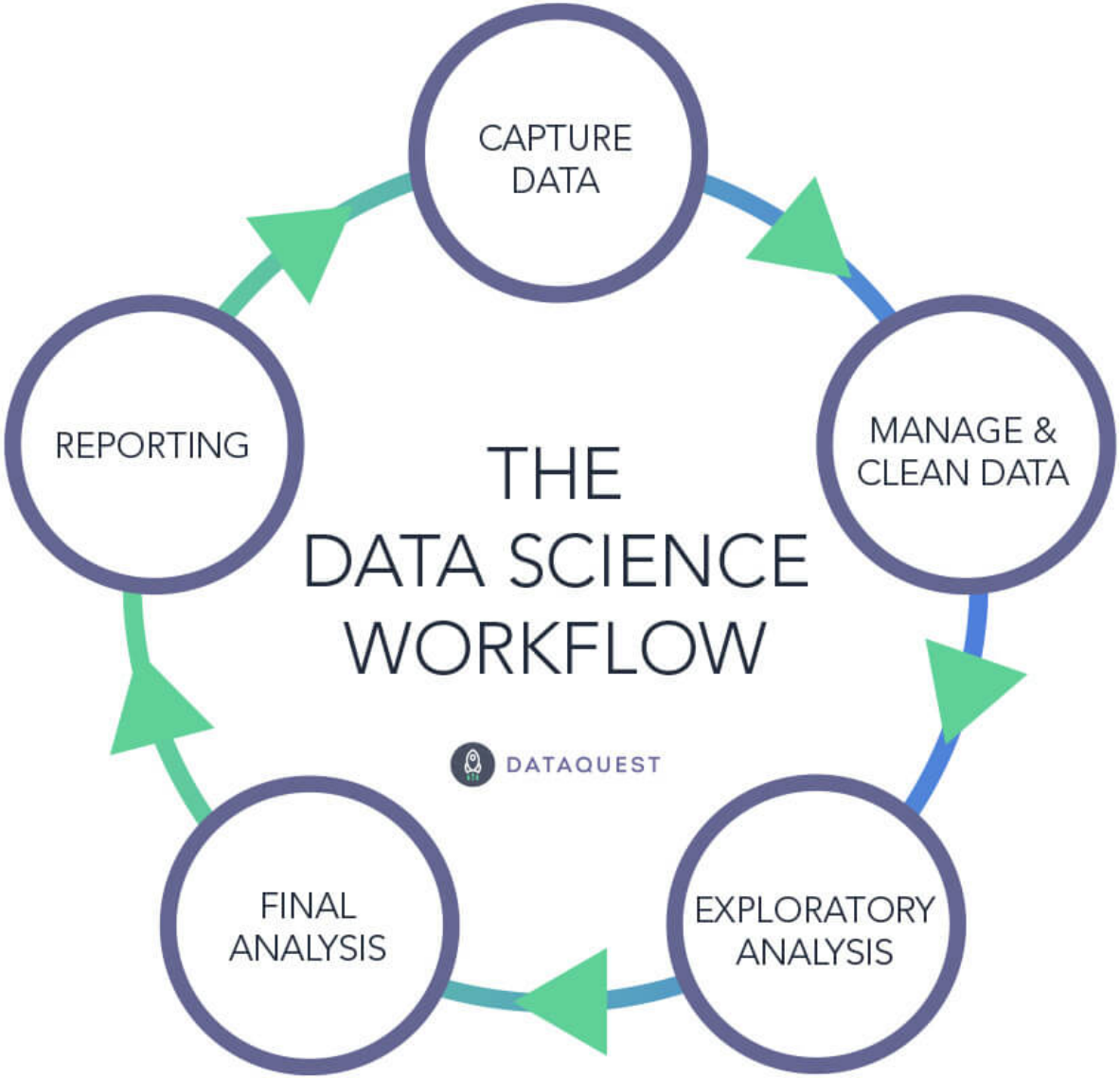
Data science has introduced **a new paradigm in science**. The deluge of data has created the Fourth Paradigm of science based on data-intensive discovery (Hey et al. 2009).

What skills does a data scientist have?



Source: Drew Conway.

What does a data scientist do?



Source: https://www.dataquest.io/blog/what-is-data-science/.

Political Science 15
Introduction to Research in Political Science
Lecture 10b: The Data Science Workflow

Alice Lépissier
University of California Santa Barbara

The data science workflow

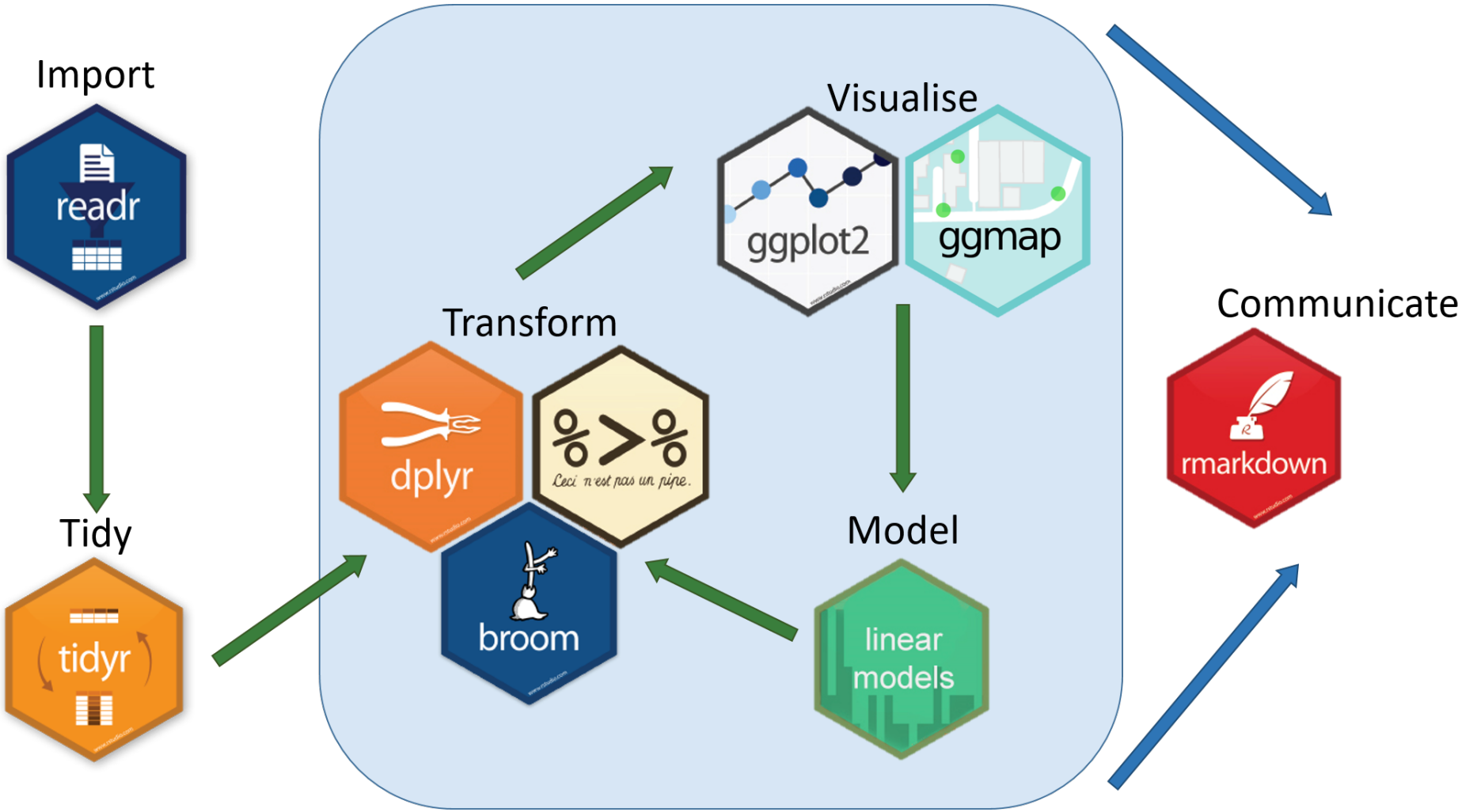
1. **Data wrangling** ($\approx$ 70% of your time)
   - Cleaning, merging, processing, validating data, ..., and much more
2. **Data analysis**
   - Statistical modeling, machine learning, text mining, regression analysis, ..., and much more
3. **Data visualization**
   - Reporting results, interactive graphs, interactive dashboards, ..., and much more

Why data wrangling is so important

- Real-world data is messy, unstructured, and cannot usually be analyzed directly.
- We need a systematic approach to tidying the data and getting it ready for analysis.
- Common tasks include:
  - **Discovering**: summarizing and understanding the data.
  - **Structuring**: organizing raw data into a usable structure.
  - **Cleaning**: fixing coding mistakes, spelling consistencies, null values, formatting.
  - **Enriching**: merging datasets to add other variables that come from outside sources.
  - **Validating**: summarizing and visualizing the data to make sure that the values make sense and that there are no errors from inputting, merging, etc.
- This "janitorial work" is incredibly important. You won't extract useful insights if you haven't wrangled your data effectively.

Welcome to the tidyverse

The tidyverse is an opinionated collection of R packages designed for data science. Install with `install.packages("tidyverse")`.



Let's skip to data visualization

Communicating your results in an effective and informative manner is both an art and a science.

It depends who your audience is: other scientists, C-level executives, policy-makers, the general public, etc.

A useful package to do beautiful data visualizations in R is `ggplot2`, which comes loaded in the `tidyverse`.

Principles of effective visualization

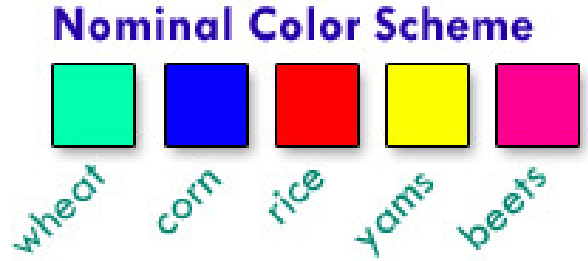
Data visualization involves **encoding data using visual cues**, that is, "mapping" data onto variations in shapes, size, color, etc.

Use the appropriate visualization for what you are trying to show.

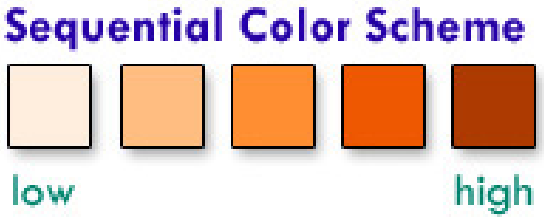
- **Distribution**: violin plot, density plot, histogram, boxplot, ridgeline
- **Correlation**: scatter plot, heatmap, correlogram, bubble chart
- **Ranking**: barplot, spider/radar chart, wordcloud, parallel chart, lollipop, circular barplot
- **Part of a whole**: grouped and stacked barplots, treemap, doughnut/sunburst chart, pie chart, dendrogram
- **Evolution**: line chart, area and stacked area charts, time series
- **Geography**: map, choropleth, hexbin map, cartogram, flow map, bubble map
- **Flow**: chord diagram, network graph, Sankey

Principles of effective visualization

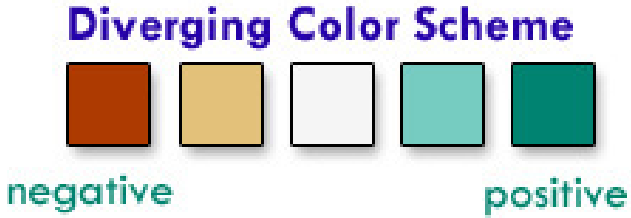
Use the appropriate color scheme for your data.



different hues that keep lightness and saturation constant should be used for **nominal data** (i.e., un-orderable categories, not numerical data).



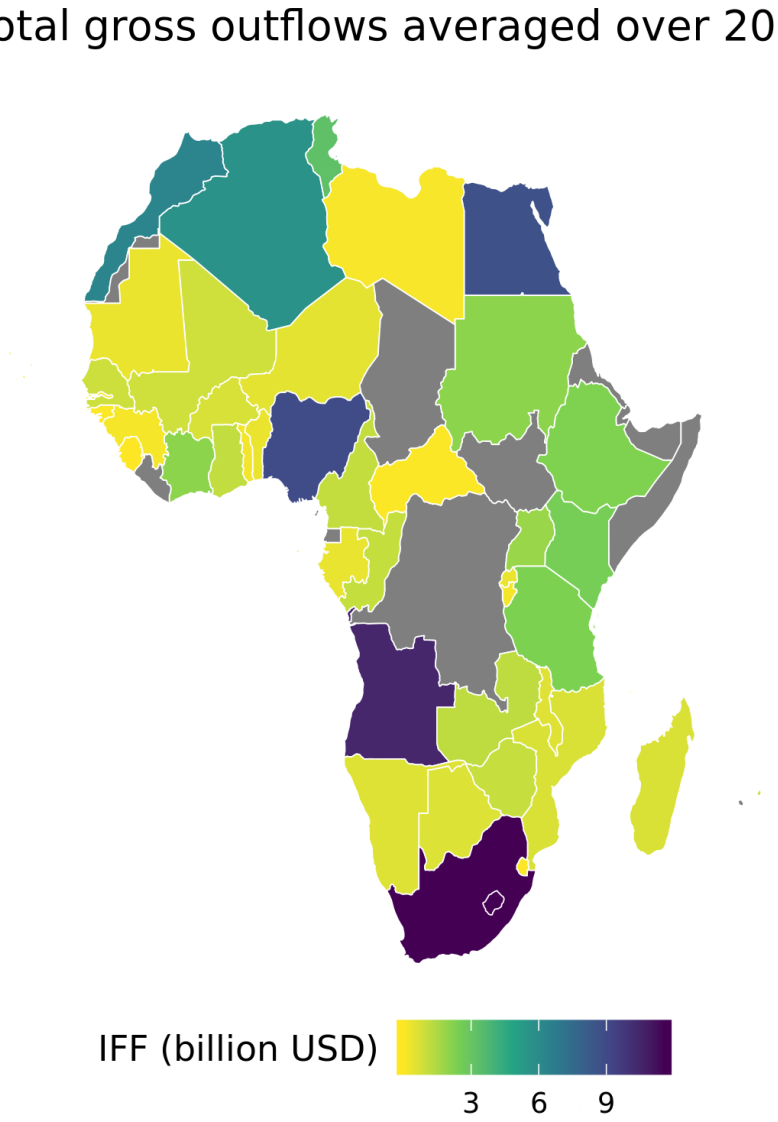
any sequence that is **dominated by changes in lightness** can be used with orderable (rankable) categories (low/med/high) or with numerical data.



any numerical data **that can be divided meaningful at a mid-point** (e.g., national average, zero) can use a diverging scheme: the data are split in two around the lightest, middle color/class.

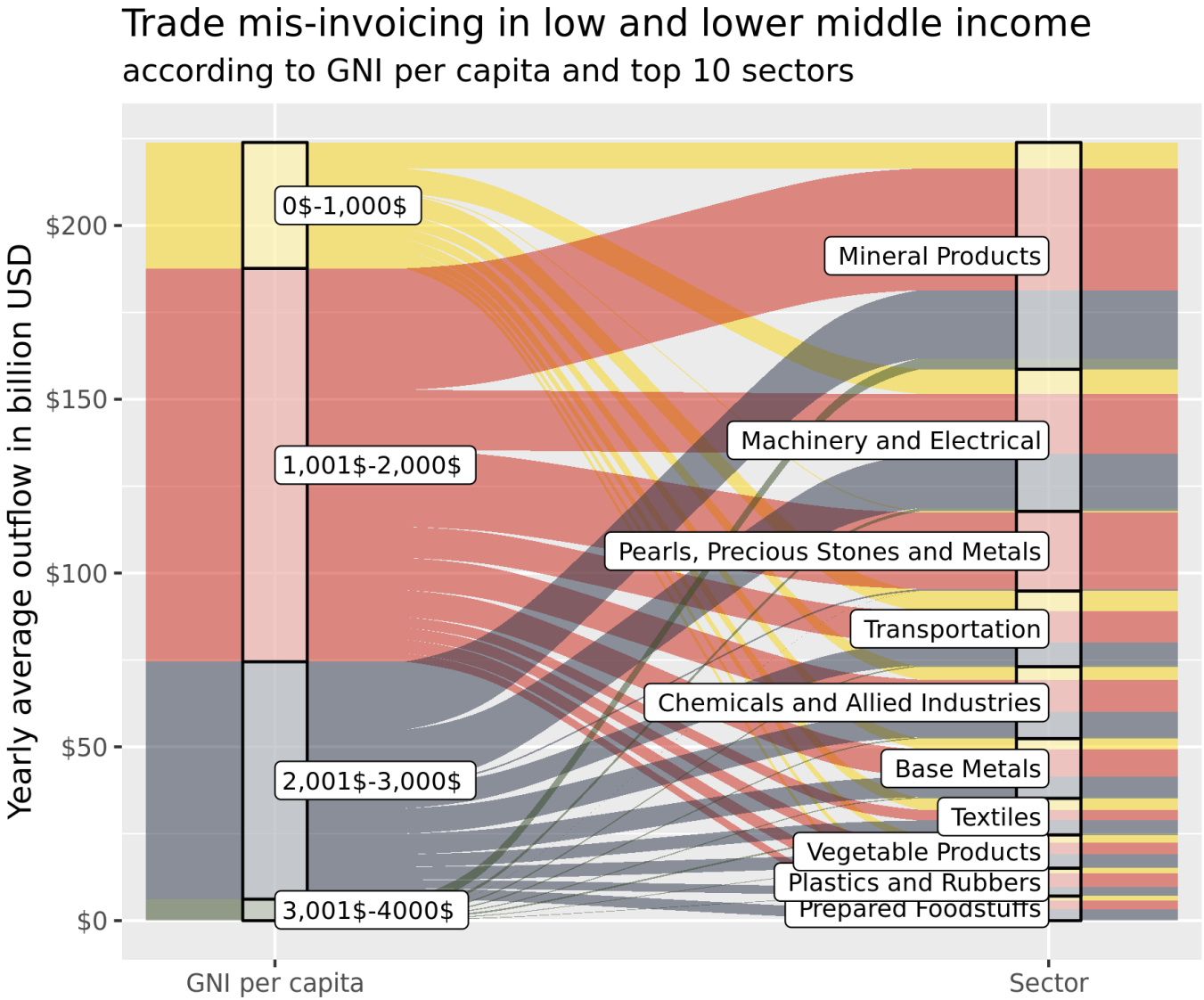
Source: https://www.axismaps.com/guide/general/using-colors-on-maps/.

Choropleth example: to see proportions on a map



Source: Lépissier (2019)

Sankey example: to visualize flows

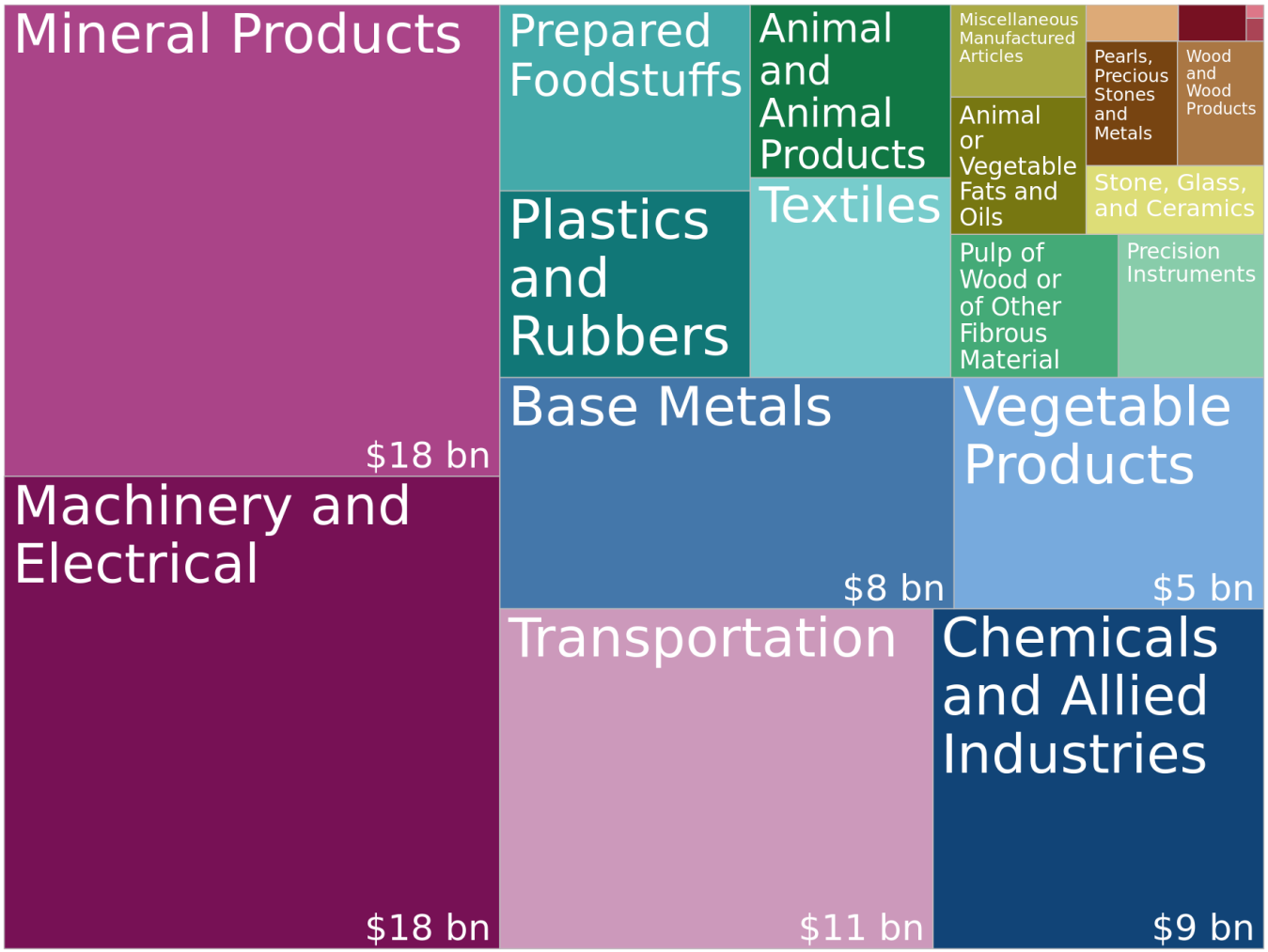


Source: Lépissier (2019)

Treemap example: to show parts of a whole

Top sectors in Africa

Average gross yearly outflow during 2000-2016

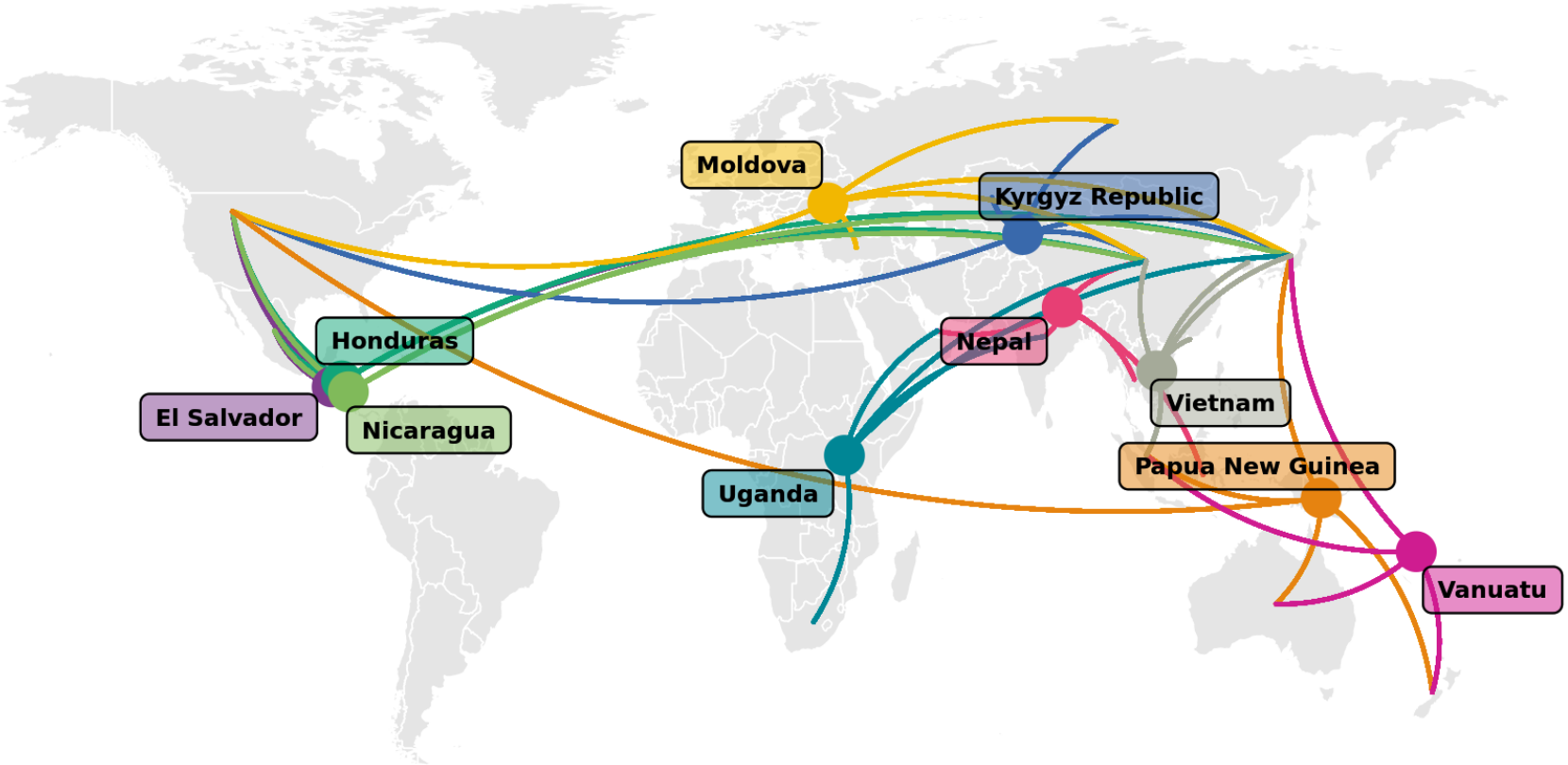


Source: Lépissier (2019)

Flow map example: to show geographical connections

Destinations of top origins in low and lower middle income

Top 10 origin countries by % of GDP



Source: Lépissier (2019)

Static versus interactive visualizations

See more examples of types of plots on https://www.r-graph-gallery.com/.

Interactive visualizations allow you to create dashboards, widgets, and other apps where **your audience interacts with your data**.

You can embed interactive HTML graphs using RMarkdown!

See more examples of interactive visualizations here:

- https://shiny.rstudio.com/gallery/movie-explorer.html
- https://www.gapminder.org/tools/

Political Science 15
Introduction to Research in Political Science

Lecture 10c: Machine Learning

Alice Lépissier
University of California Santa Barbara

What is machine learning?

- Machine Learning (ML) is a set of techniques and approaches where the statistical model *learns* from the data
- Often focused on **predictive** inference, rather than causal inference
  - Examples: what is the likelihood that a borrower will default on her mortgage?; is a financial transaction likely to be fraud?; predict diagnosis of person in ER on the basis of her symptoms
- Also used for **classification problems**
  - Examples: is this a picture of a cat or a raccoon?; is this email spam or not?
- Can be used to **cluster** observations

Applications of machine learning

- Netflix recommendations
- Computer vision, e.g. recognizing surface-to-air missiles from satellite images
- Anomaly detection, e.g. outliers, fraud
- Natural language processing to identify the authors of the Federalist Papers
- and so much more

Two types of machine learning

Supervised learning

- Has "training labels", i.e. information on the outcome
  - Whether the borrower defaulted, what the diagnosis of the patient was, whether this was a picture of a cat or a raccoon
- Goal: to predict as best as we can what the outcome would be for an "out-of-sample" observation

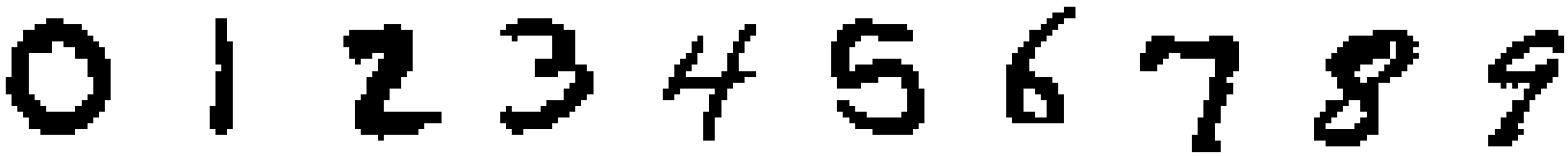
Unsupervised learning

- We don't have training labels: we only have the inputs (e.g. patient demographics), not the output (e.g. diagnosis)
- Goal: to find structure and patterns in the data, to cluster/group observations
- Also used for "dimensionality reduction", i.e. if we have many, many variables, what is the subset of variables that best explains the data?

Example: genetic algorithm

- A genetic algorithm (GA) is a type of **evolutionary algorithm** from artificial intelligence
- Inspired by the Darwinian process of evolution
- Mimics the process of evolution by which only the "fittest" individuals survive over many generations
- Starting from a randomly generated population, a GA applies a variety of "genetic operators" so that individuals in the population *reproduce*, *mutate*, and *clone* themselves in order to optimize an objective function called the "fitness function"

Genetic algorithm to recognize hand-written digits



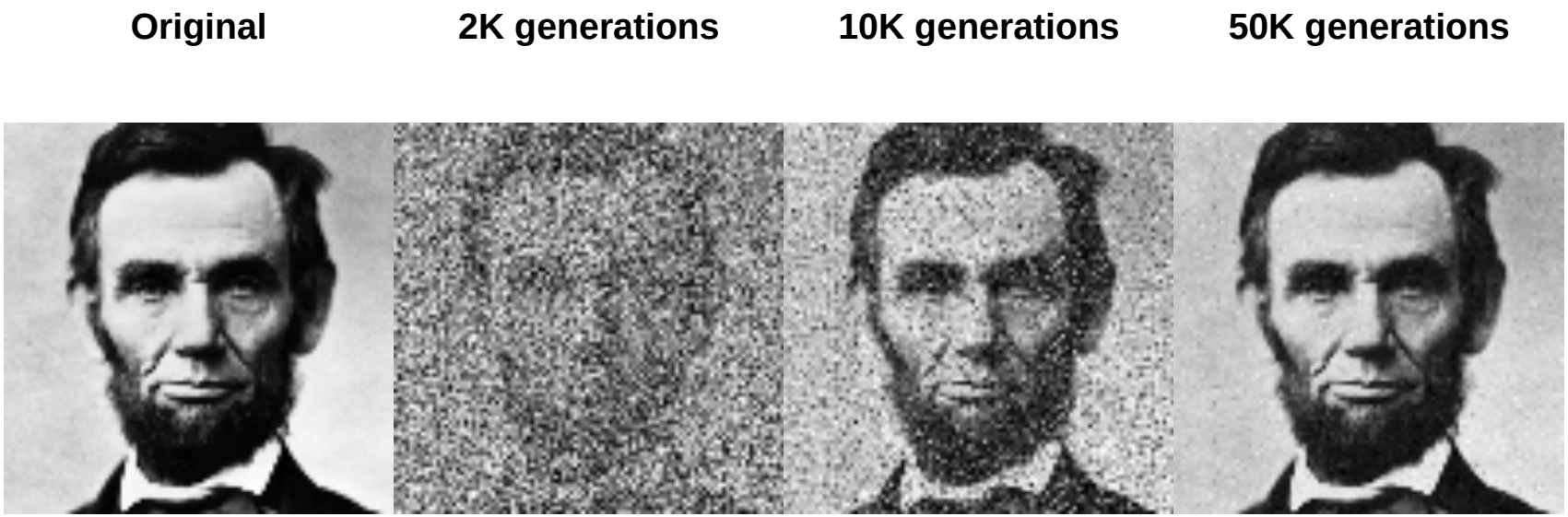
Real vs. reconstructed digits

Source: Lépissier. Code: https://github.com/walice/Genetic-Algorithm.

Genetic algorithm for image reconstruction

We have the original picture. Starting from a randomly-generated set of pixels (e.g. noise), we can reconstruct the original image over several "generations".

Abraham Lincoln



Source: Lépissier. Code: https://github.com/walice/Genetic-Algorithm.

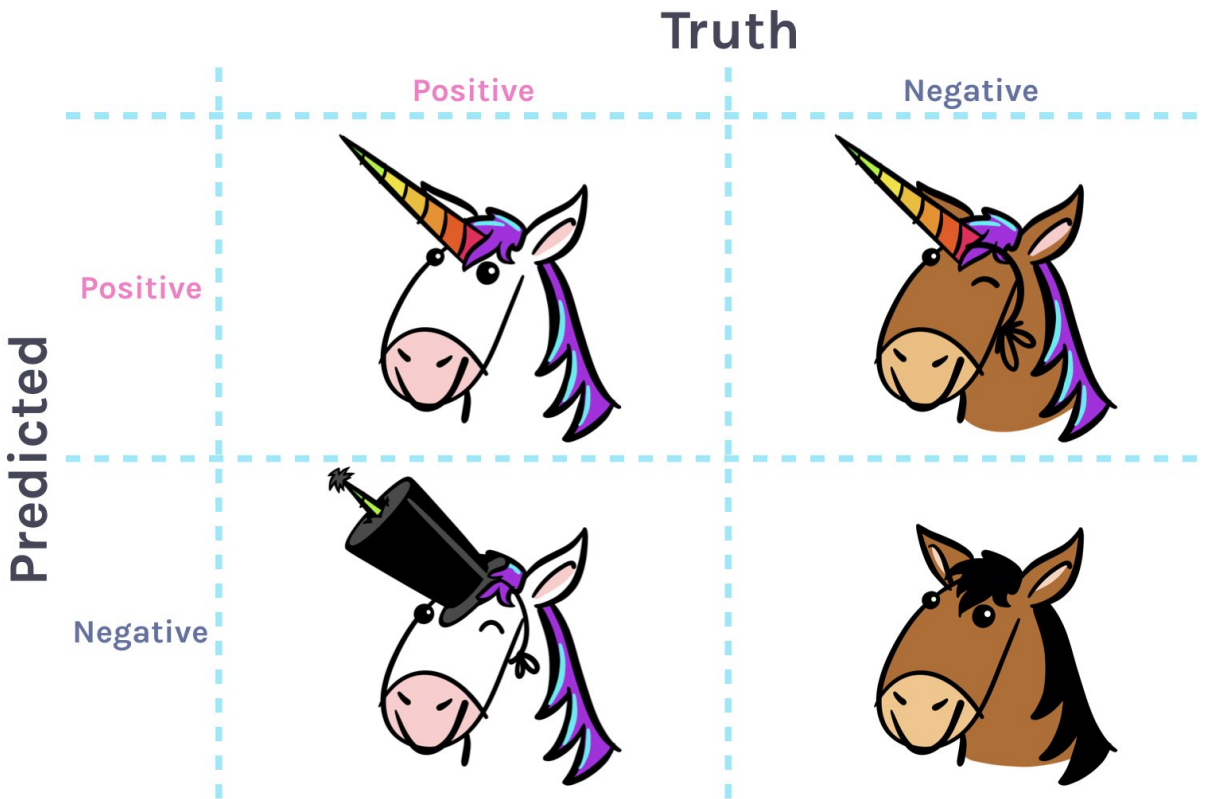
Classification problems

1. Start with **training data**: a sample with the outcome Y, and several features $X_1, X_2, \ldots, X_p$.
2. **Train** a model on this sample. Usually, do this so as to maximize "goodness-of-fit". The model will *learn* the parameters.
3. Use this model to predict the outcome for an observation from the **test data**, i.e. a sample of observations on which the model was not trained.

Classification problems: confusion matrix

How can we assess the model's performance on the training data?

We draw a **confusion matrix** which allows us to calculate the rate of misclassifications.



Credit: allison_horst

Neural networks

- Neural networks are inspired by biological brains
- Allows computers to learn from the data
- Applications: facial recognition, Google translate, advanced robotics
- A type of deep learning: will solve highly abstract and complex problems

Image classification with a neural network

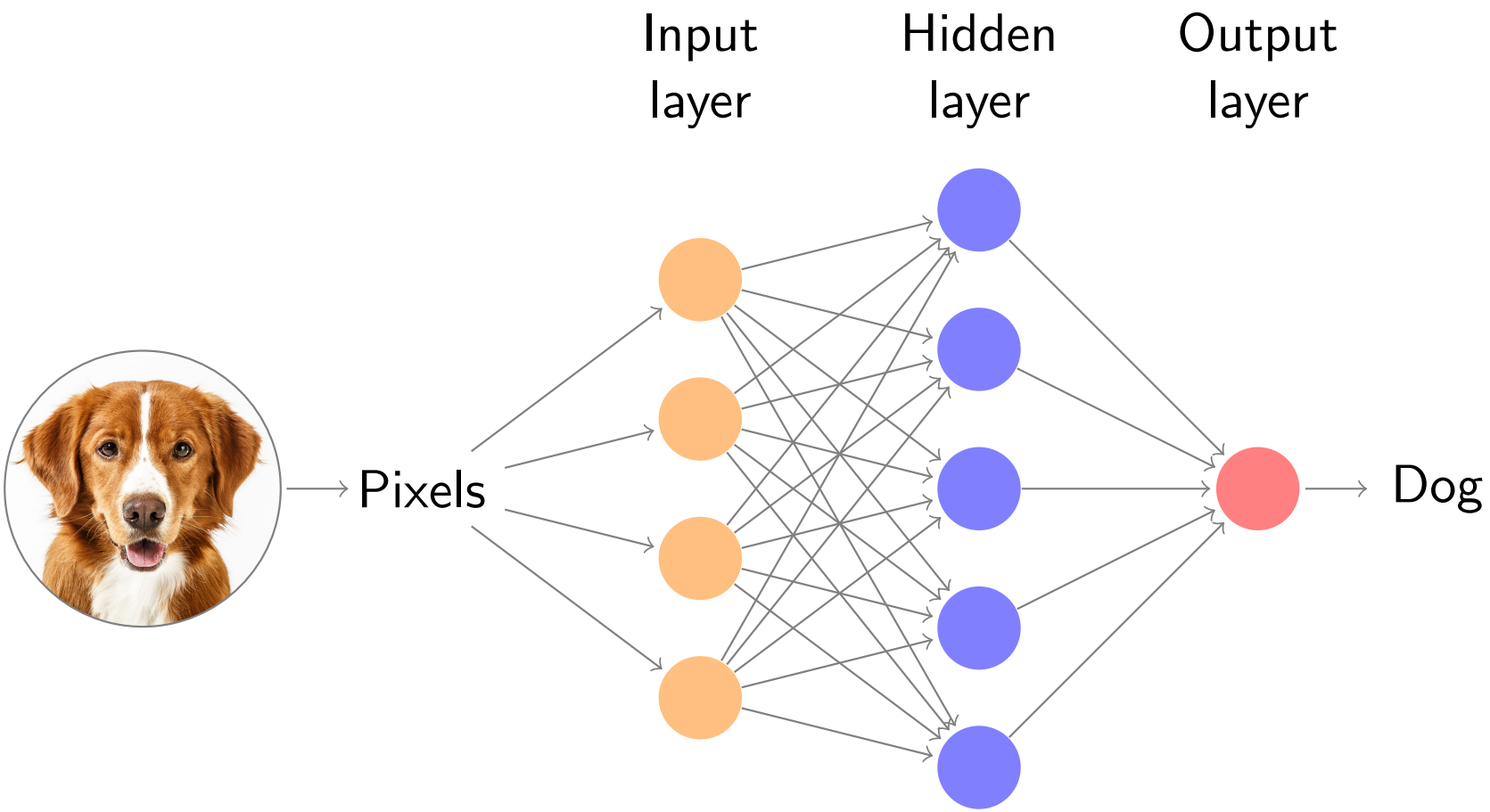


Image classification with a neural network

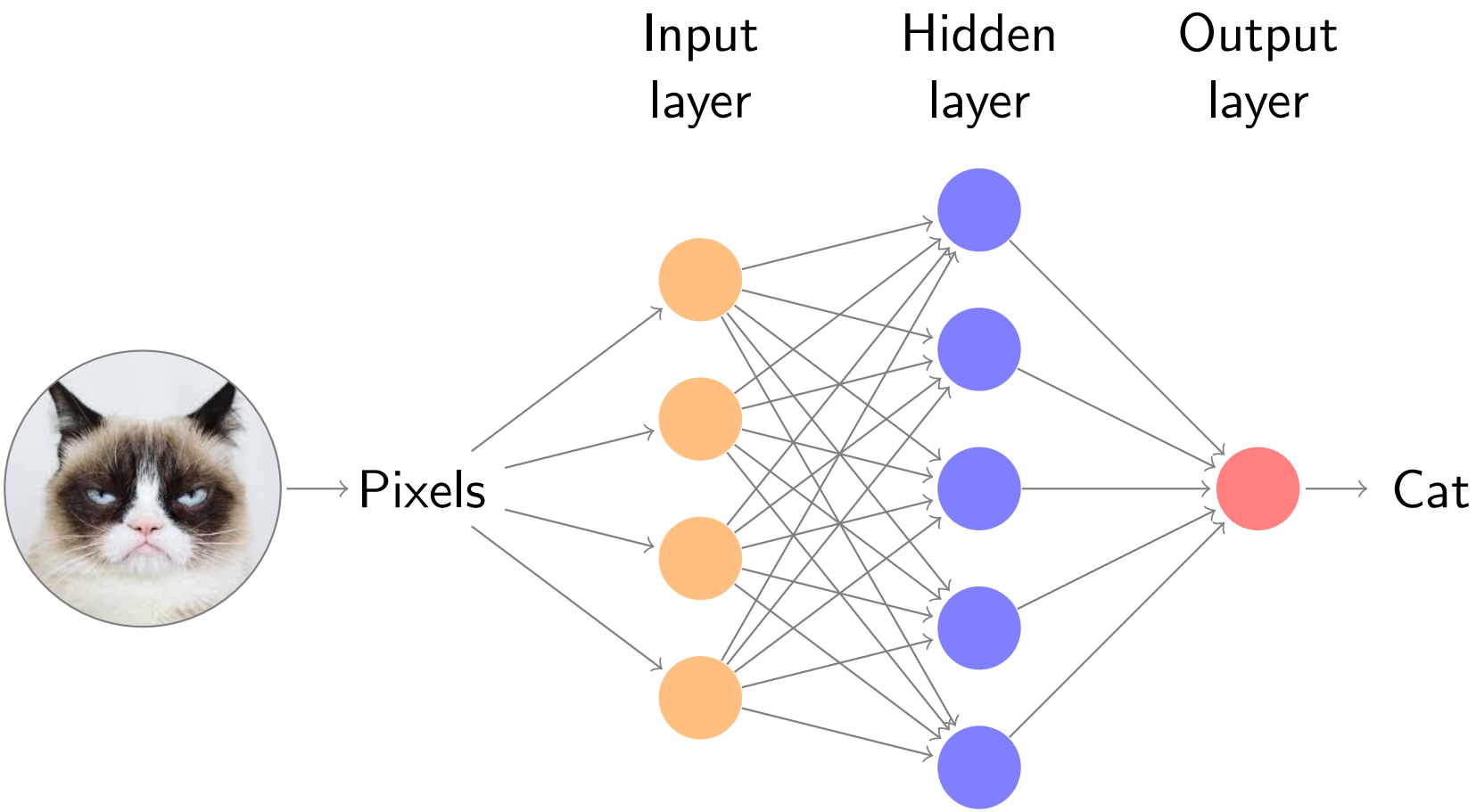
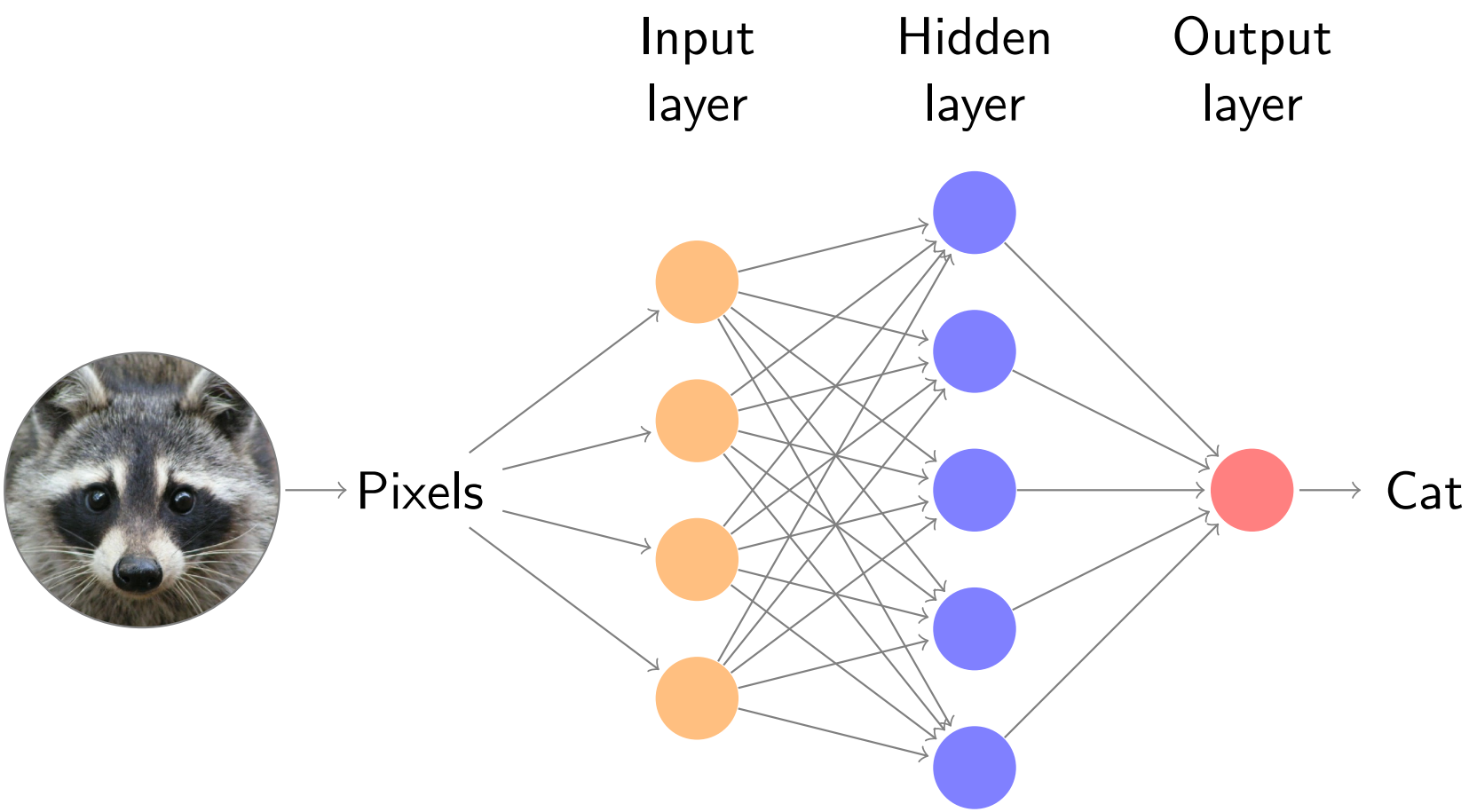


Image classification with a neural network - OOPS



Causal inference versus predictive inference

- In political science, we often ask **causal questions**. We care about estimating causal effects in an unbiased and consistent manner.
- But **prediction questions** are also important. In that case, we care about the accuracy of our predictions.
- Real-life example: algorithms used by cities to allocate health inspectors. We want our model to accurately predict whether an establishment will violate health code based on risk factors (predictive inference), but we also want to understand the causal effect of an establishment receiving a health inspection (causal inference).
- Prediction problems are more frequent in social sciences than is commonly thought.

Final thoughts

- Machine learning solves different types of problems
- Increasingly used in cutting-edge political science research
- Most common software languages: R or Python