Hertie School Data Science Society

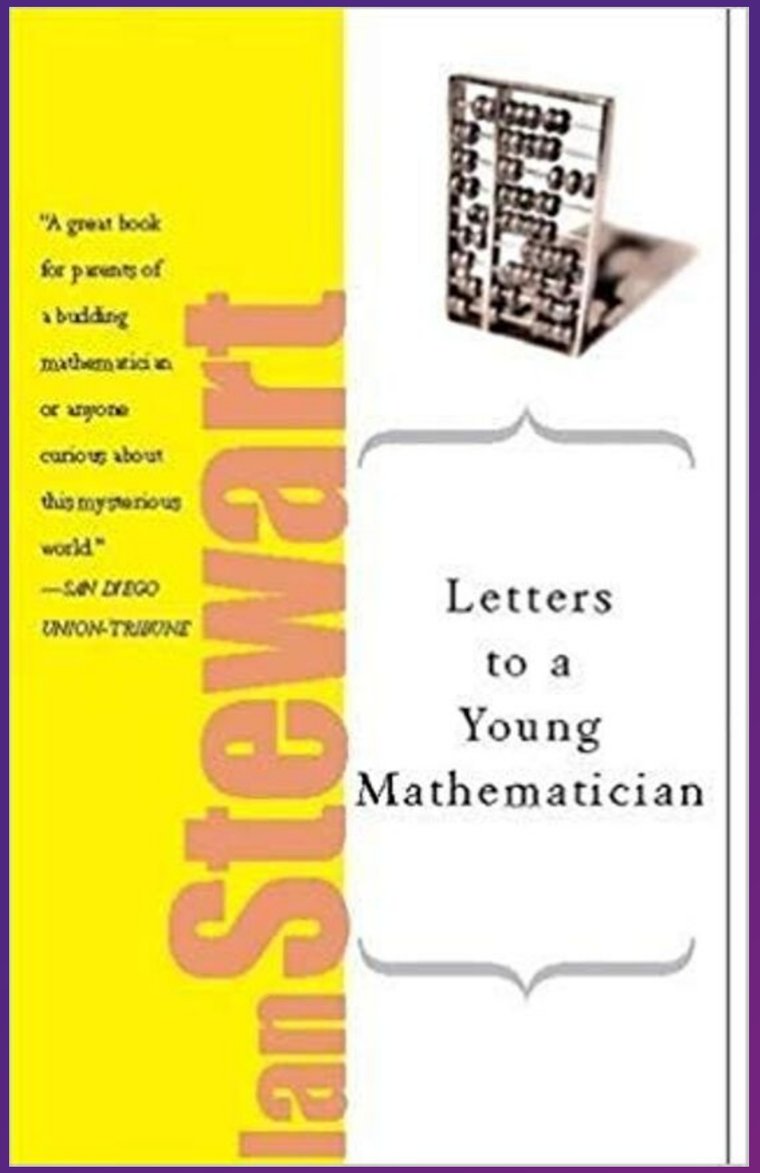
The Mathematics of Data Science

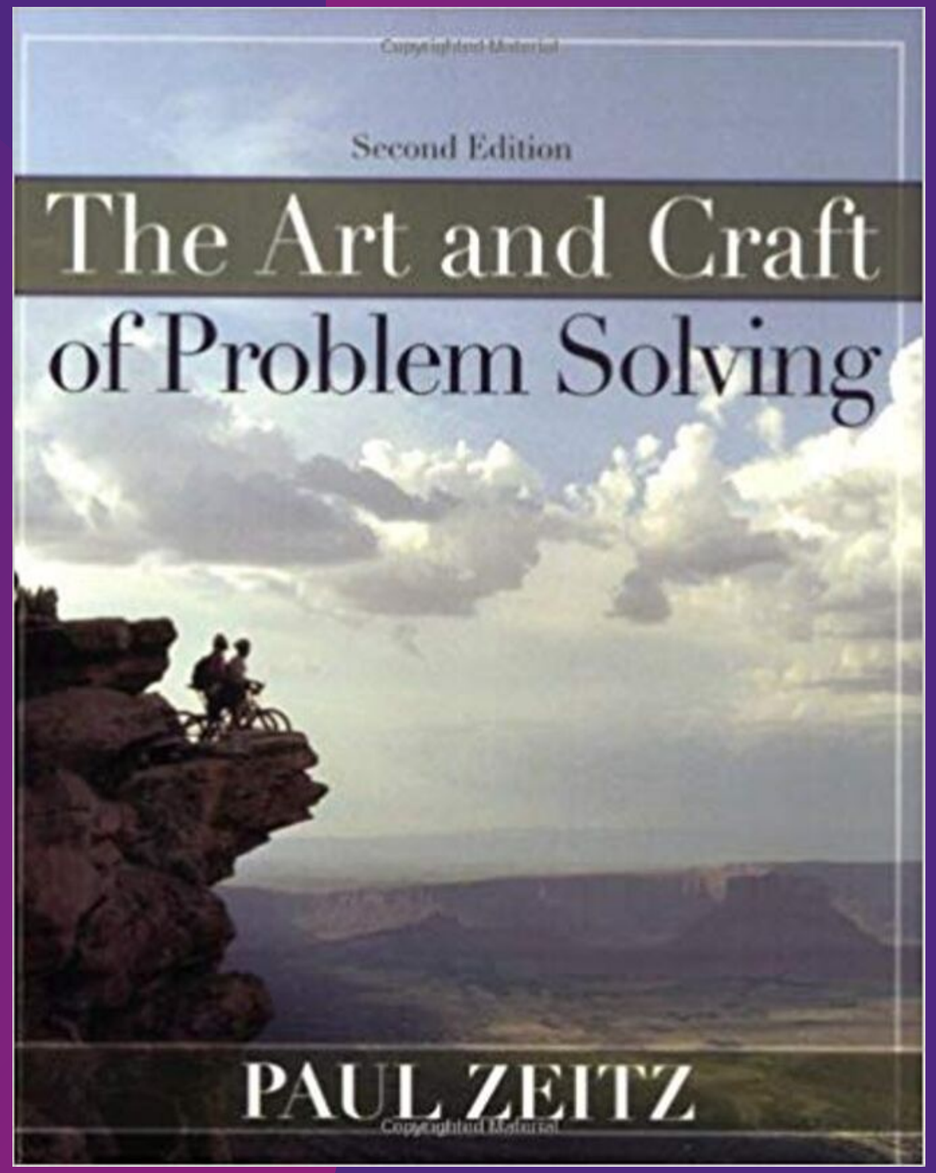
SESSION 3 - SEPTEMBER 30, 2019

TODAY'S WORKSHOP

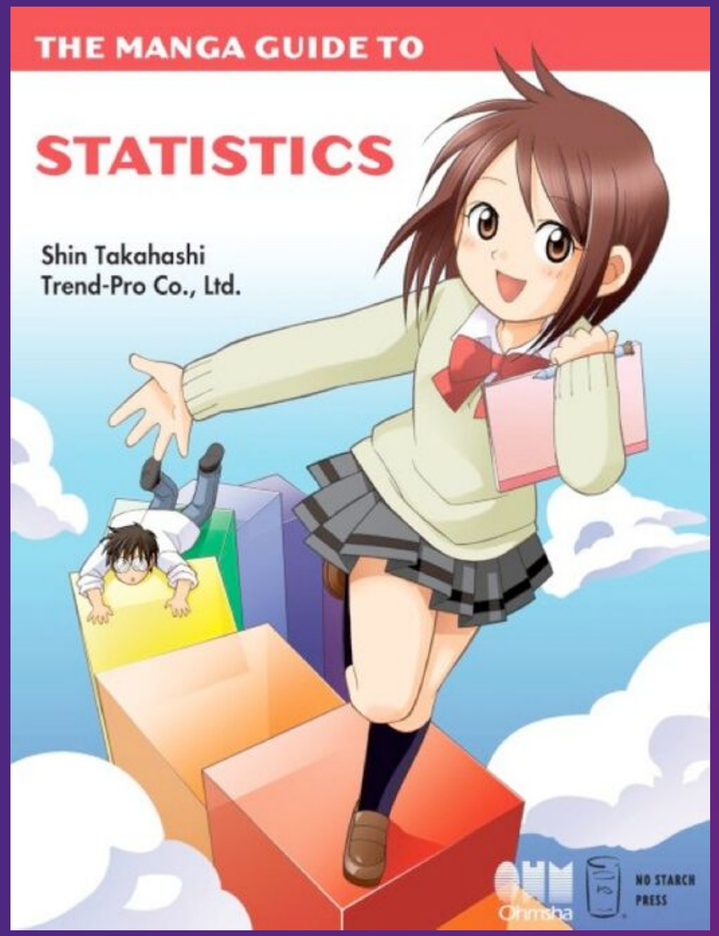
- The Maths: Statistics, Probability, Calculus & Linear Algebra
- A Quick Challenge
- Roadmap for Data Science

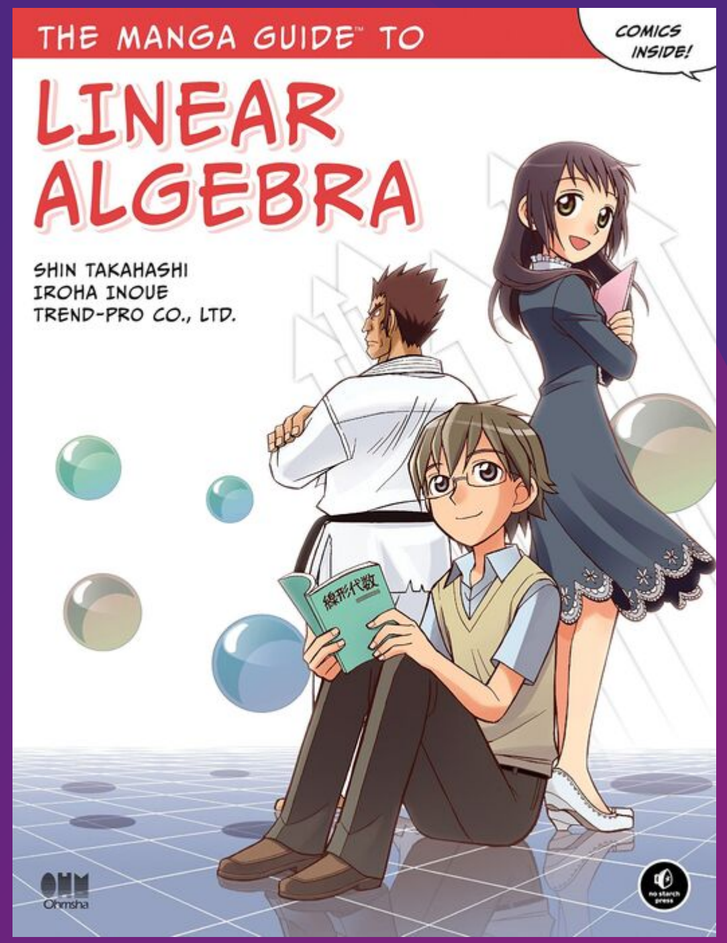
A GENTLE INTRODUCTION

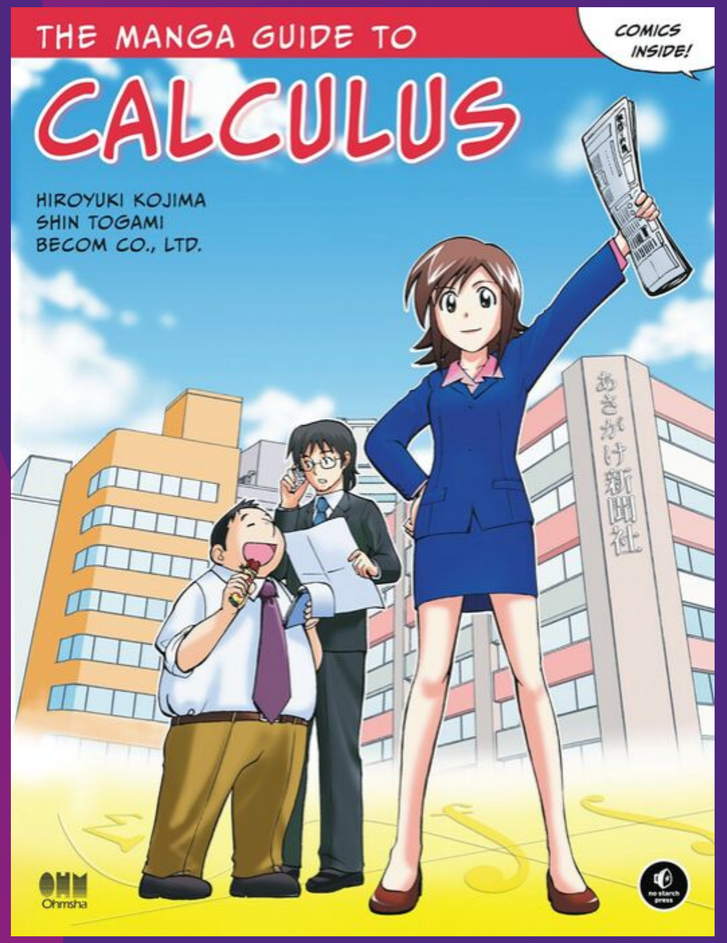




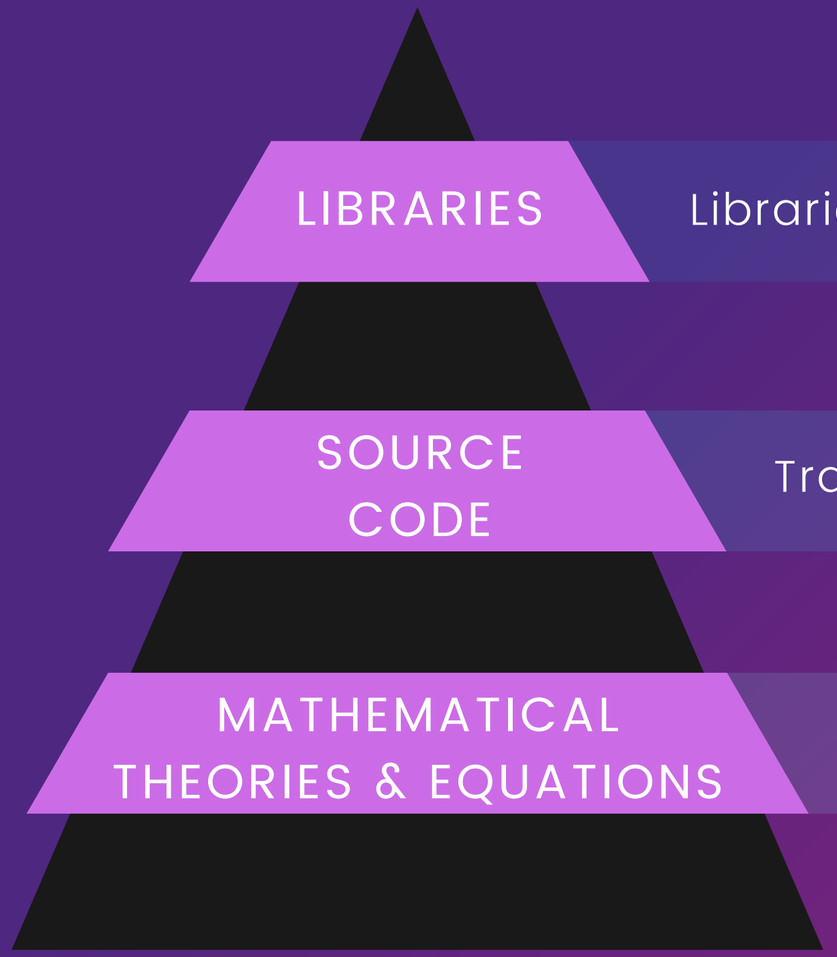
AN EVEN GENTLER INTRODUCTION







NECESSARY TO LEARN MATHS FOR DATA SCIENCE?

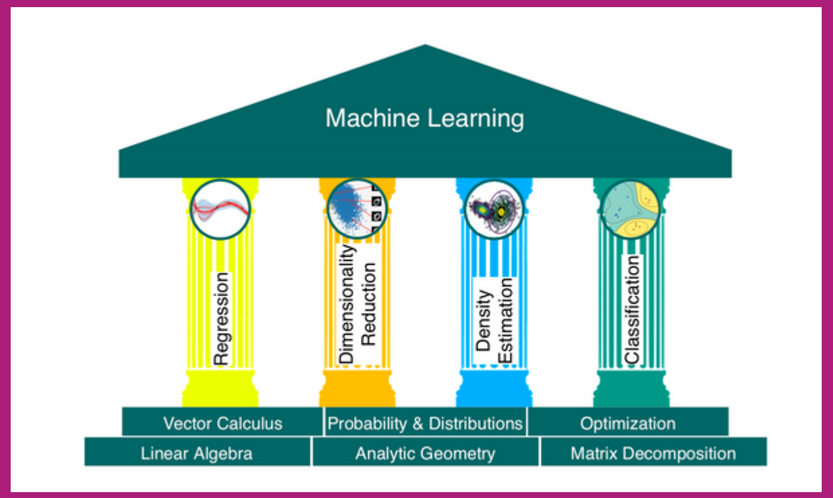


Libraries, packages and frameworks that simplify the coding process

Transform math equations into code and abstract the theories

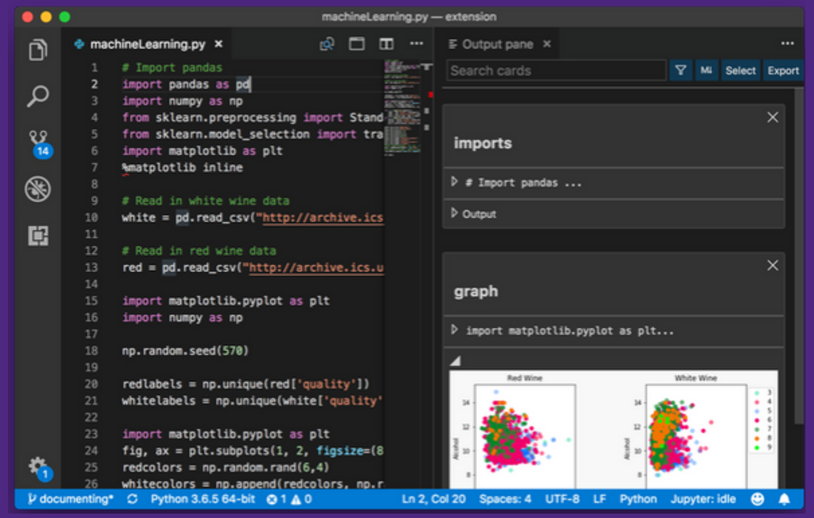
Equations that model and approximate the problems of interest

TWO APPROACHES



THEORIES THEN PRACTICE

Bottom-up: build a strong mathematical foundation and intuition before attempting complex concepts and projects

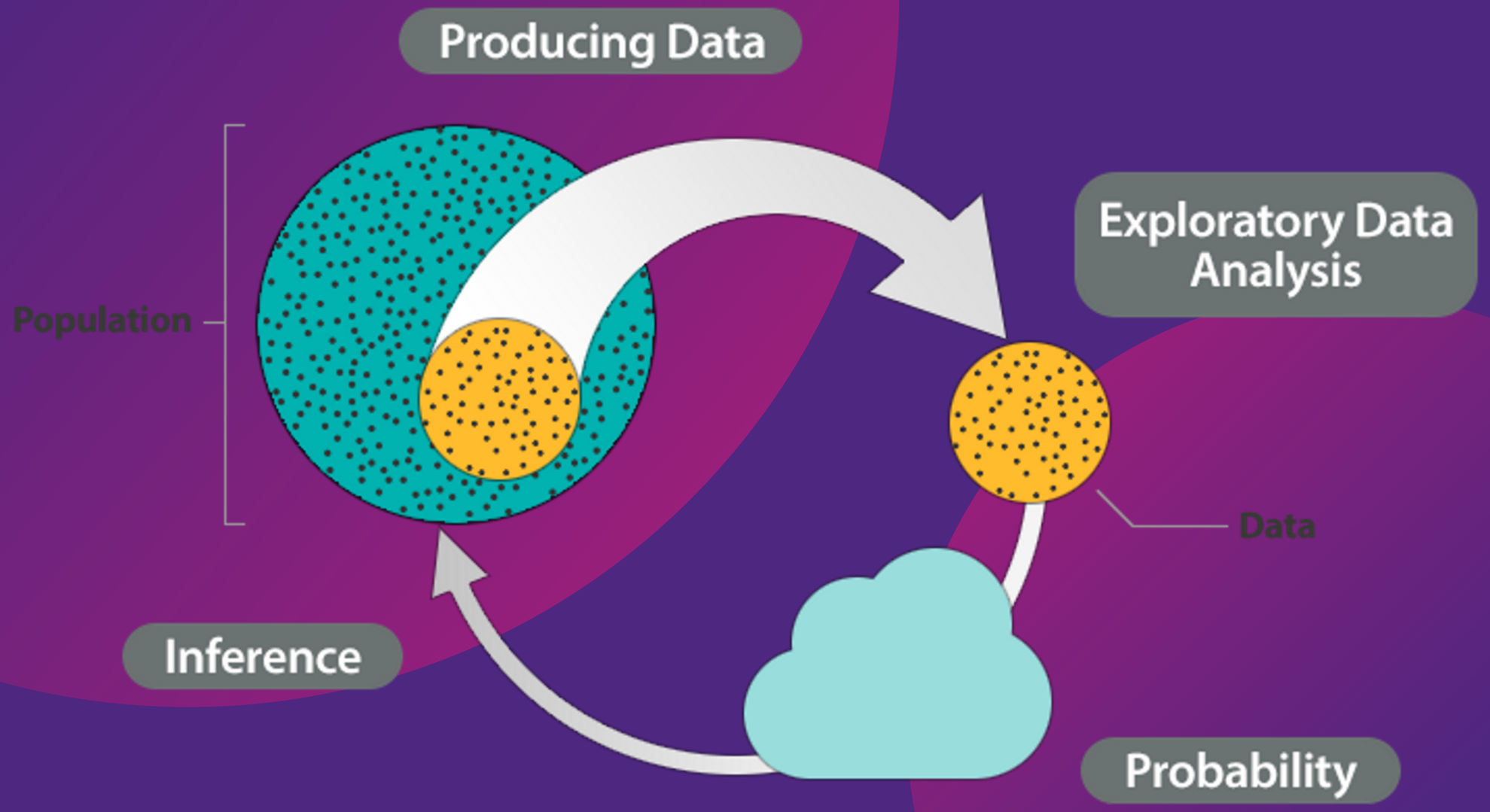


PRACTICE AND THEORIES

Top-down: tackle a project/concept head-on, reverse-engineer concepts and codes and gradually build your mathematical intuition as you go along

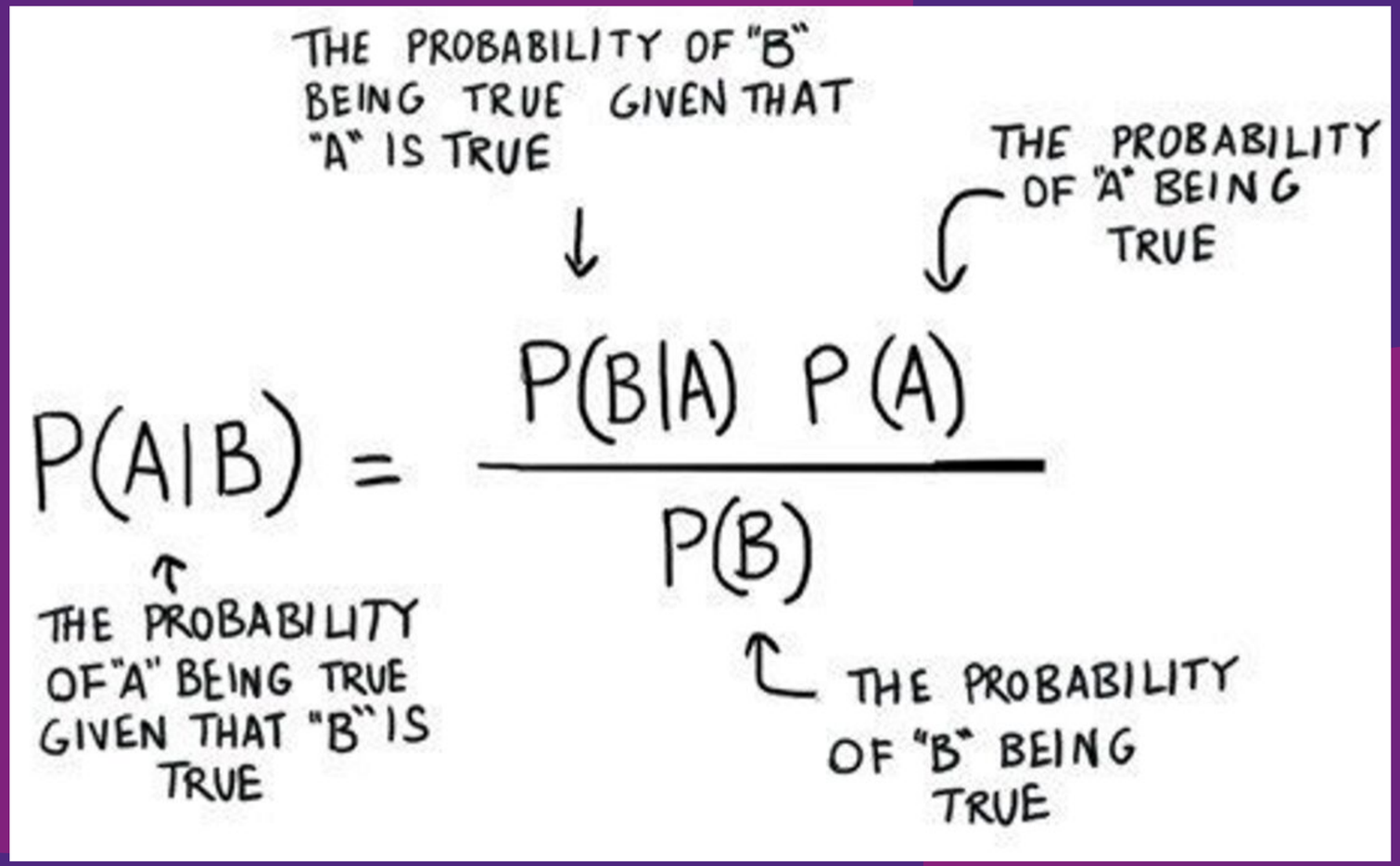
PROBABILITY & STATISTICS

"STATISTICS IS THE GRAMMAR OF SCIENCE" - KARL PEARSON



PROBABILITY & STATISTICS

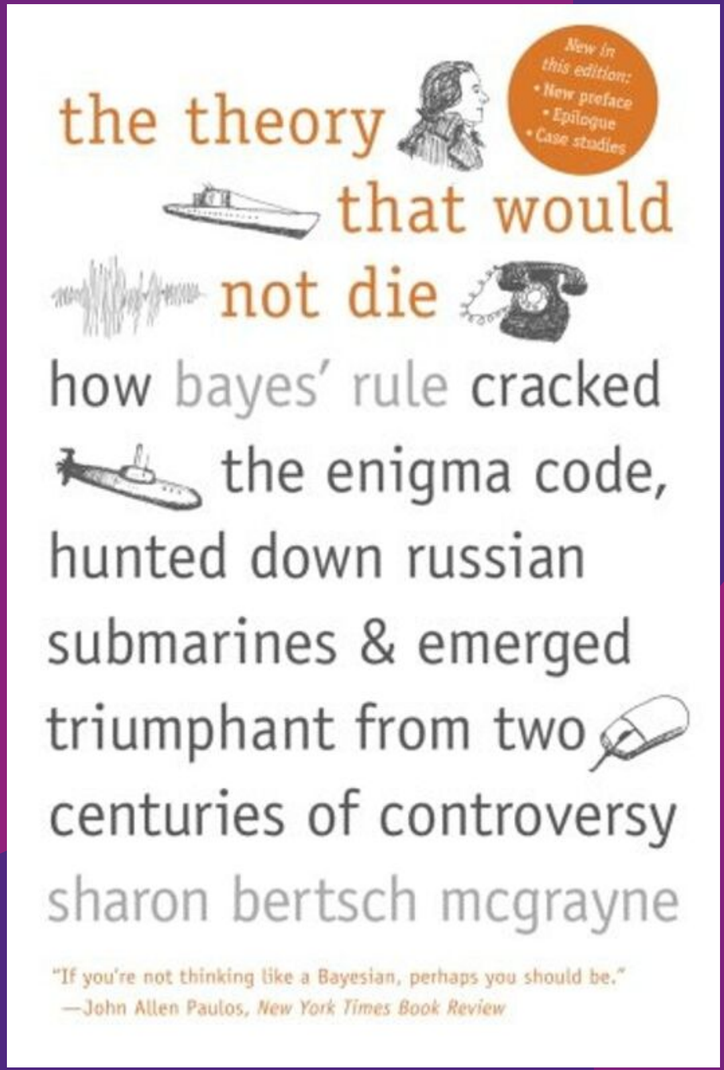
BAYES' THEOREM



PREDICT THE LIKELIHOOD OF AN EVENT OCCURING

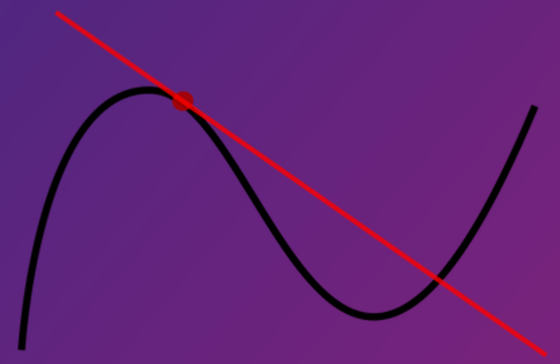
PROBABILITY & STATISTICS

BAYES' THEOREM

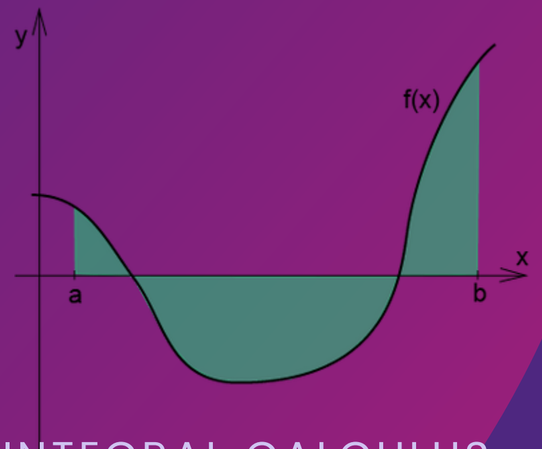


CALCULUS

THE STUDY OF CHANGE

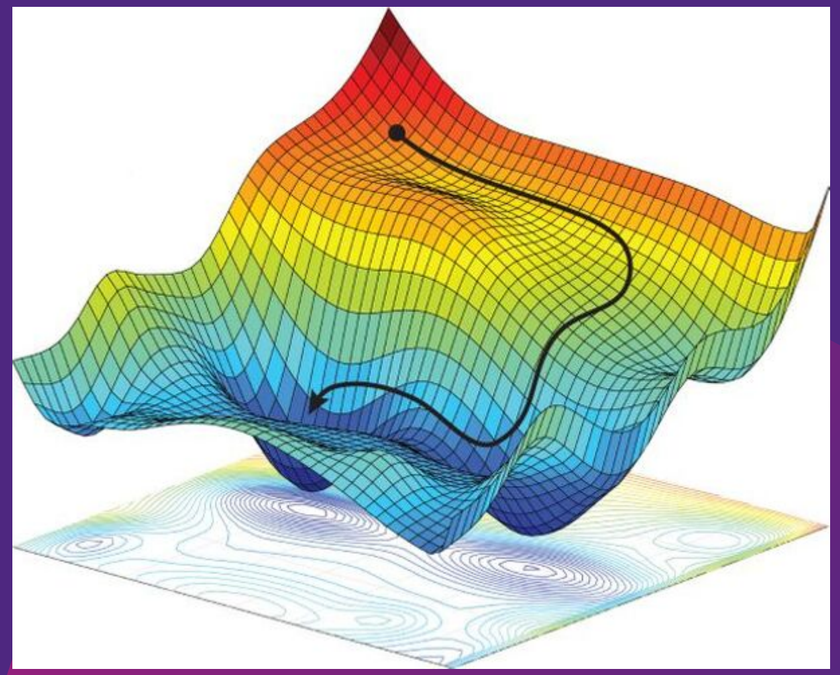


DIFFERENTIAL CALCULUS



INTEGRAL CALCULUS

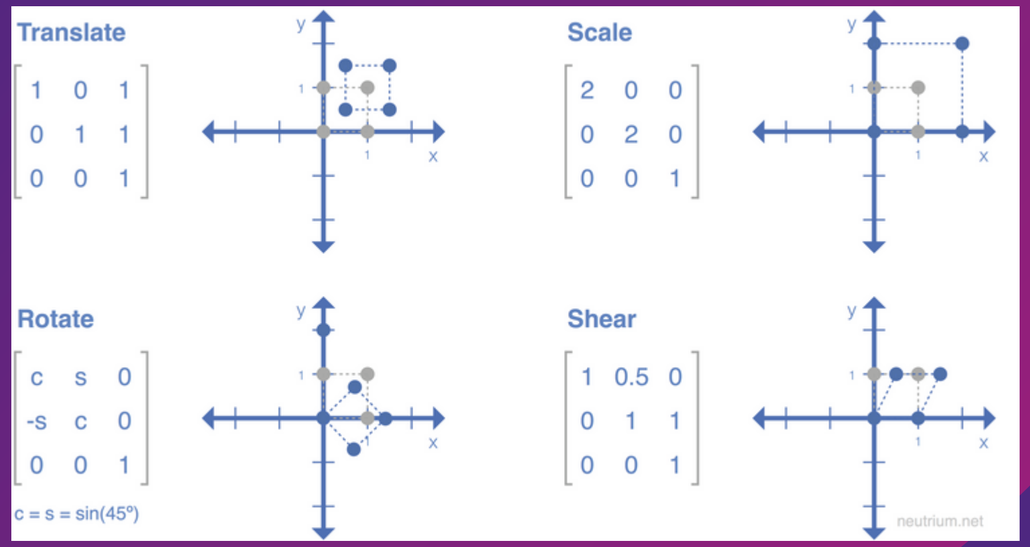
GRADIENT DESCENT



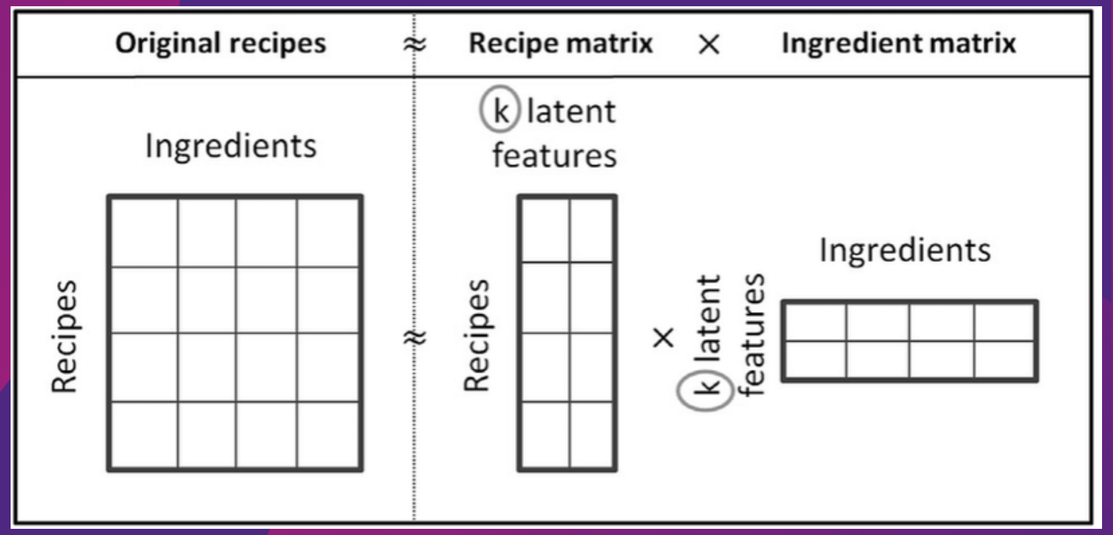
TELLS US HOW TO OPTIMIZE FOR OUR MODELS;
HELPS US FIND THE DIRECTION OF CHANGE: IN WHAT DIRECTION SHOULD WE CHANGE OUR VARIABLES SO THAT OUR PREDICTION IS MORE OPTIMAL AND CLOSER TO THE TRUTH?

LINEAR ALGEBRA

MATRIX TRANSFORMATION



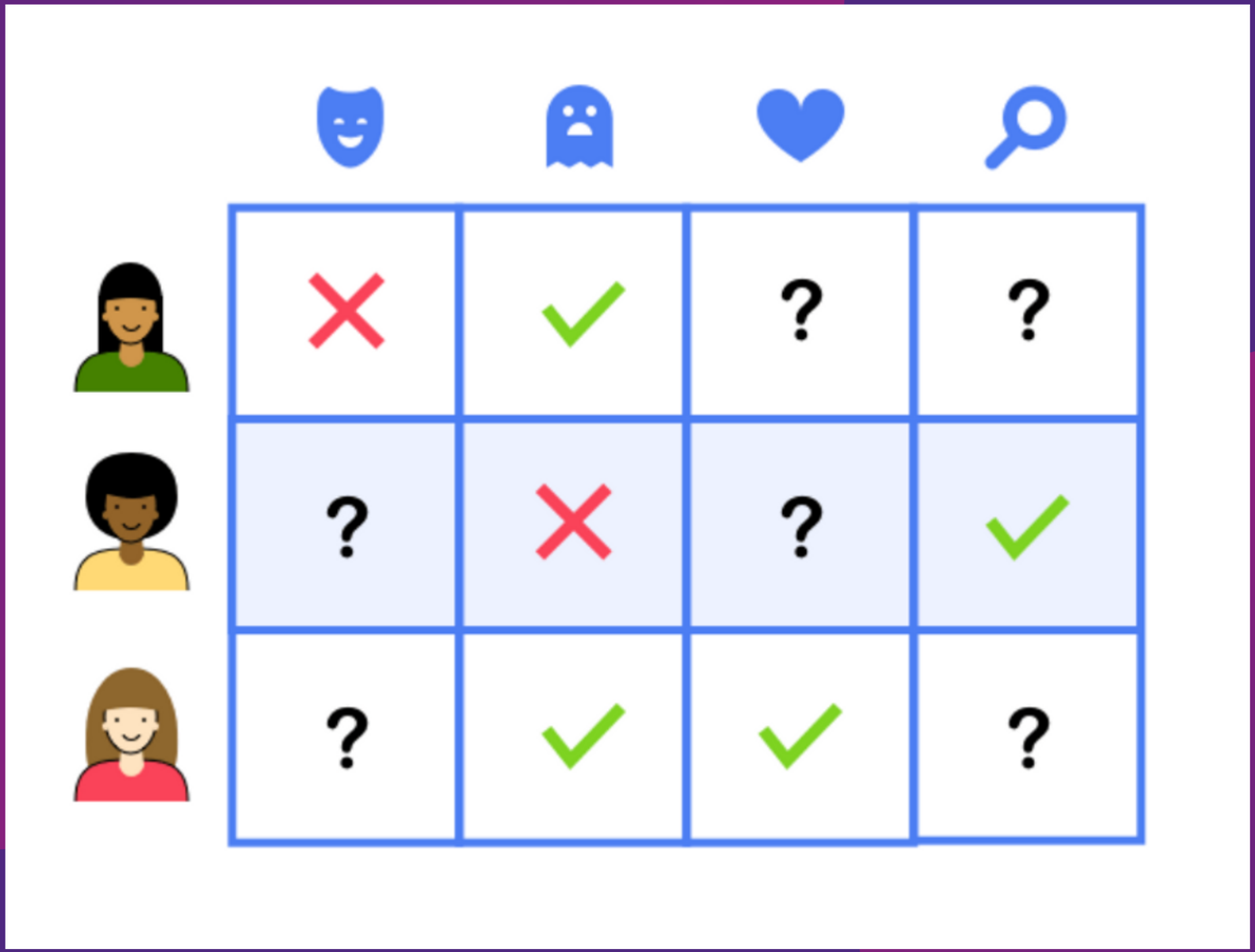
MATRIX DECOMPOSITION



MAKES RUNNING ALGORITHM FEASIBLE ON MASSIVE DATASET; GIVES US OPERATIONS THAT WE CAN PERFORM ON MATRICES TO UNDERSTAND PATTERNS AND TRENDS IN DATA

4 DISCIPLINES TOGETHER

MOVIE RECOMMENDATION SYSTEM FOR NETFLIX

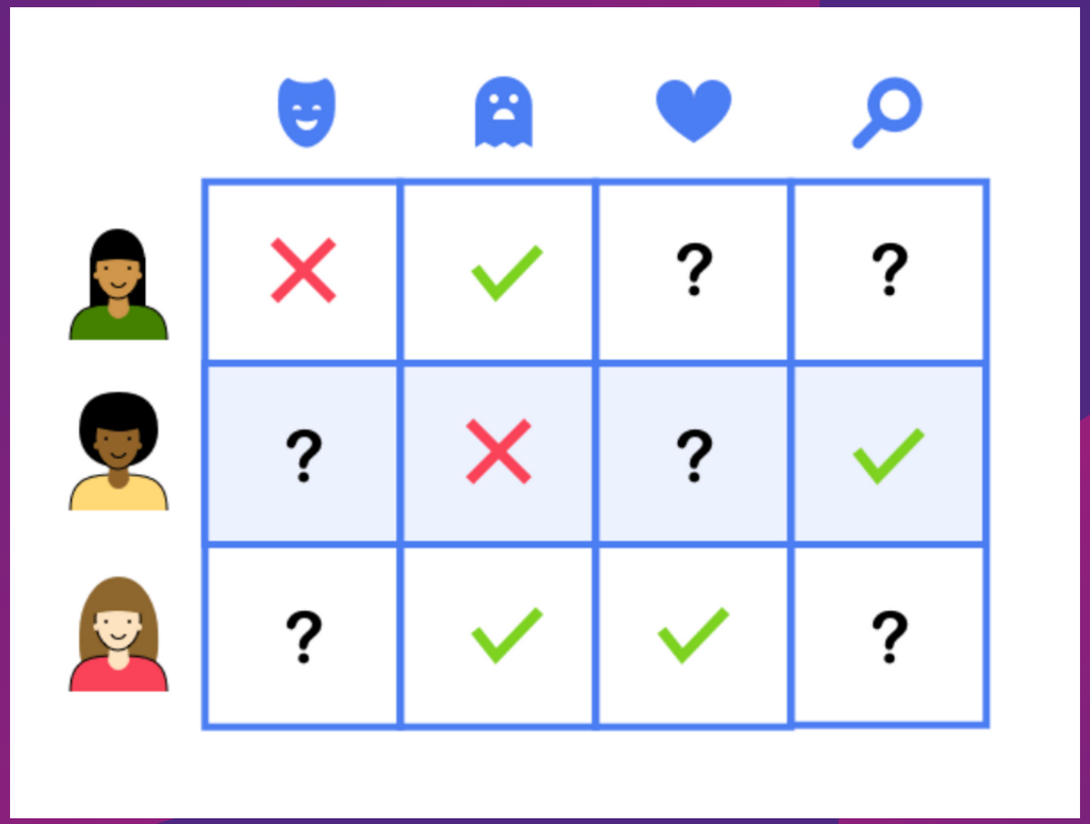


4 DISCIPLINES TOGETHER

MOVIE RECOMMENDATION SYSTEM FOR NETFLIX

STATISTICS: discover what factors influence the sentiment of users towards a movie

CALCULUS: optimize the search for the right movies/shows



PROBABILITY: discover the likelihood of a user liking or disliking a movie

LINEAR ALGEBRA: help to run this personalized recommendation for millions of people/data points

INDEPENDENT STUDY RESOURCES

DATA SCIENCE THEORIES

- **Probability**: Khan's Academy | MIT Opencoursewave
- **Calculus**: TrevTutor Calculus 1 | TrevTutor Calculus 2 | MathTutor | Professor Leonard's Calculus 1 | Professor Leonard's Calculus 2 |
- **Multivariable Calculus**: Khan's Academy | MIT Opencoursewave | 3Blue1Brown | TheTrevTutor |
- **Linear Algebra**: MathTutor Vol 1 | MathTutor Vol 2 | MIT Opencoursewave | TrevTutor | 3Blue1Brown |
- **The Mathematics for Machine Learning**: Course Specialization |
- **Essential Maths for Machine Learning**: Microsoft's EDX Course

A LITTLE CHALLENGE

RULES:

- 10-MINUTE DEADLINE;
- GOOGLE IS ALLOWED (BUT NOT ENCOURAGED);
- RANDOM GUESS IS FINE (BUT NOT ENCOURAGED);
- 2 BEST GUESSES GET THE PRIZES (THE CLOSEST GUESS AND THE BEST NON-RANDOM GUESS).

A LITTLE CHALLENGE

HOW MANY TRAINS ARE RUNNING IN BERLIN AT THIS VERY MOMENT?

SOLUTION

BREAK THE ORIGINAL PROBLEM DOWN INTO SMALLER PROBLEMS

How many train lines (both S-bahn and U-bahn) are there in Berlin?

How long are the average track for each of these lines?

On average, how many stations are there for each line?

Is it rush-hour now or not?

What is the average time between each train?

What is the average speed of the train (S-bahn and U-bahn individually)?

On average, how long does it take one train to finish one average line?

How many trains are running on one line right now?

SOLUTION

BREAK THE ORIGINAL PROBLEM DOWN INTO SMALLER PROBLEMS

S-BAHN

One average S-bahn line:

- 22km
- 11 stations
- 5 minutes interval
- Average speed of train: 40km/h
- Time for one train to finish one track: 33 minutes + 1 minute wait time at each station = 44 minutes

Conclusion: 8.8 trains/line => 132 S-bahn trains

U-BAHN

One average U-bahn line:

- 14.6km
- 17.3 stations
- 5 minutes interval
- Average speed of train: 30.7 km/h
- Time for one train to finish one track: 28.5 minutes + 1 minute wait time at each station = 45.8 minutes

Conclusion: 9.2 trains/line => 92 U-bahn trains

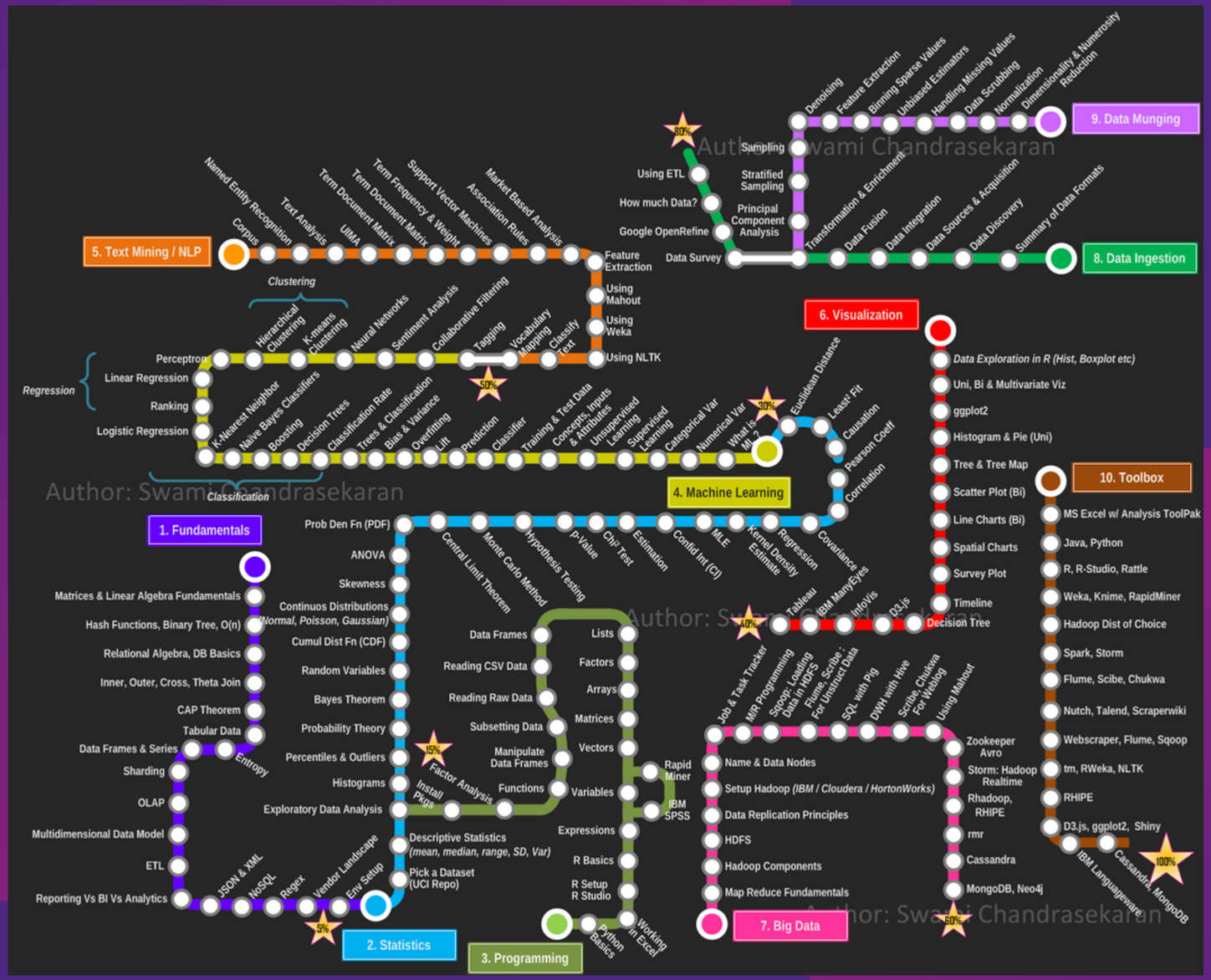
REGIONAL & INTERNATIONAL TRAINS

Let's assume there's 10 regional and international trains running concurrently on the tracks of Berlin

TOTAL

234 trains currently running in Berlin at this moment

DATA SCIENCE ROADMAP - HARDCORE



DATA SCIENCE ROADMAP - HERTIE EDITION

1

Statistics I
+ Programming Skills
+ Data Science Society

2

Winter Break:
Linear Algebra +
Calculus + Review
Statistics & Probability

3

Statistics II: Causal
Inference
+ Data Science Practice
Challenges & Hackathon

4

Summer Break:
Learn SQL & work with raw data
Machine Learning Course by
Andrew NG

5

Machine Learning &
Natural Language Processing
Projects
+ Lead Data Science Society

6

Winter Break:
Polish up your projects,
knowledge and skillsets in
data science and analytics

7

Apply for jobs & internship
Start-up & make cool shit together