CS 181AI
Lecture 25

Assorted Requested Topics!

Arthi Padmanabhan

Apr 24, 2023

Logistics

- Wednesday (4/26): Project presentations
  - 15 min each group

Today

- Assortment of requested topics + topics I think are good for you to know before leaving this class ☺
  - Themes
  - Life of an ML engineer
    - Interview
    - MapReduce & Hadoop
    - Scale of ML in industry, dev environments, etc.
  - PageRank algorithm

Themes from this course

- When evaluating a system, keep in mind:
  - Performance
  - Fault tolerance
  - Consistency
  - Availability
- Sometimes, these are at odds with each other, e.g., we might make a system more fault tolerant by stopping periodically to save a checkpoint…at the expense of performance

Themes from this course

- When evaluating a system, keep in mind:
  - Performance
  - Fault tolerance
  - Consistency
  - Availability
- **It's all about tradeoffs!**

Roles

Data Engineering
Modeling
Deployment
Business Analysis
Infrastructure
Data Scientist
Machine Learning Engineer
Data Analyst
Software Engineer-ML
Machine Learning Researcher
Software Engineer

Data Scientist: Job Description

- Identify valuable data sources and automate collection processes
- Undertake preprocessing of structured and unstructured data
- Analyze large amounts of information to discover trends and patterns
- Build predictive models and machine-learning algorithms
- Combine models through ensemble modeling
- Present information using data visualization techniques
- Propose solutions and strategies to business challenges
- Collaborate with engineering and product development teams

Data Scientist: Example

- Leverage data and business principles to create and drive large scale FB Data Center programs.
- Define and develop the program for metrics creation, data collection, modeling, and reporting the operational performance of Facebook's data centers.
- Collaborate with cross-functional data and product teams across business applications to define problem statements, access and manipulate data, build analytical models, explain data-gathering requirements, deliver analytics insights, and make recommendations.
- Define, compute, track, and continuously validate business metrics with descriptive and predictive analytics.
- Identify and implement streamlined processes for data reporting and communication, and use analytical models to identify insights to drive key decisions across leadership and the organization.
- Provide mentorship to other members of the team on best practices for design and implementation of cutting-edge analytics insights.
- Lead and support various ad hoc projects, as needed, in support of Facebook's Data Center strategy.
- Leverage tools like R, Tableau, Python, and SQL to drive efficient analytics.
- Degree in an analytical field (e.g. Computer Science, Engineering, Mathematics, Statistics, Operations Research, Management Science)
- 3+ years of experience in a role with data analysis and metrics development
- 3+ years of hands-on experience analyzing and interpreting data, drawing conclusions, defining recommended actions, and reporting results across stakeholders
- 3+ years of SQL development experience writing queries
- 3+ years of hands-on project management experience
- 3+ years of experience with data visualization tools
- 3+ years of experience with packages such as R, Tableau, SPSS, SAS, STATA, etc.
- 2+ years of experience with scripting in Python or PHP
- Experience leveraging data driven models to drive business decisions
- Experience using data access tools and building visualizations using large datasets and multiple data sources
- Experience thinking analytically
- Experience communicating data to all organizational levels
- Experienced with packages such as NumPy, SciPy, pandas, scikit-learn, dplyr, ggplot2
- Knowledge of statistics and optimization techniques
- Hands-on experience with medium to large datasets (i.e. data extraction, cleaning, analysis and presentation)
- Technical knowledge of data center operations

ML Engineer: Job Description

- Designing machine learning systems and self-running AI software.
- Transforming data science prototypes.
- Using data modeling and evaluation strategy to find patterns and predict unseen instances.
- Managing the infrastructure and data pipelines necessary for productionizing code.
- Finding available datasets online for training purposes.
- Optimizing existing ML libraries and frameworks.
- Running machine learning tests and interpreting the results.
- Implementing best practices to improve the existing machine learning infrastructure.
- Documenting machine learning processes.

ML Engineer: Example

- Develop highly scalable classifiers and tools leveraging machine learning, data regression, and rules based models
- Suggest, collect and synthesize requirements and create effective feature roadmap
- Code deliverables in tandem with the engineering team
- Adapt standard machine learning methods to best exploit modern parallel environments (e.g. distributed clusters, multicore SMP, and GPU)
- 2+ years of experience in one or more of the following areas: machine learning, recommendation systems, pattern recognition, data mining or artificial intelligence
- Proven experience to translate insights into business recommendations
- Experience with Hadoop/HBase/Pig or MapReduce/Sawzall/Bigtable
- Knowledge developing and debugging in C/C++ and Java
- Experience with scripting languages such as Perl, Python, PHP, and shell scripts
- Bachelor's in Computer Science or related quantitative field
- Experience with filesystems, server architectures, and distributed systems

Interview: Data Scientist

Technical:

- What is the difference between machine learning and deep learning (neural networks)?
- Define precision and recall
- What are some things you can do to deal with an imbalanced dataset?

Behavioral:

- Give me an example of how you’ve used your data analysis to change behavior. What was the impact, and what would you do differently in retrospect?
- Give an example of a problem you solved (or tried to solve) with machine learning.

Curiosity:

- Tell me about a recent paper you’ve read related to machine learning

Interview: Machine Learning Engineer

Technical (could include those from DS as well):

- Outline the design for a scheduler for ML jobs with this set of characteristics
- You're using an ML model and you find that the model uses too much memory for the device you want to run it on. What steps could you take?
- You deploy a model with 98% and come back the next day to find that it's at 80%. How would you start investigating?

Behavioral:

- Tell me about the latest dataset you've worked with. If (when) any problem came up in the data, how did you solve it?

Curiosity:

- Tell me about a recent paper you've read related to systems or big data

Question

- Regarding the final project, which has been the biggest challenge?
  - Data isn’t as clean as we had hoped
  - Models we were using didn’t produce expected result
  - Running everything is taking too long
  - It’s hard to figure out what metric in the tradeoff space to prioritize
  - Other

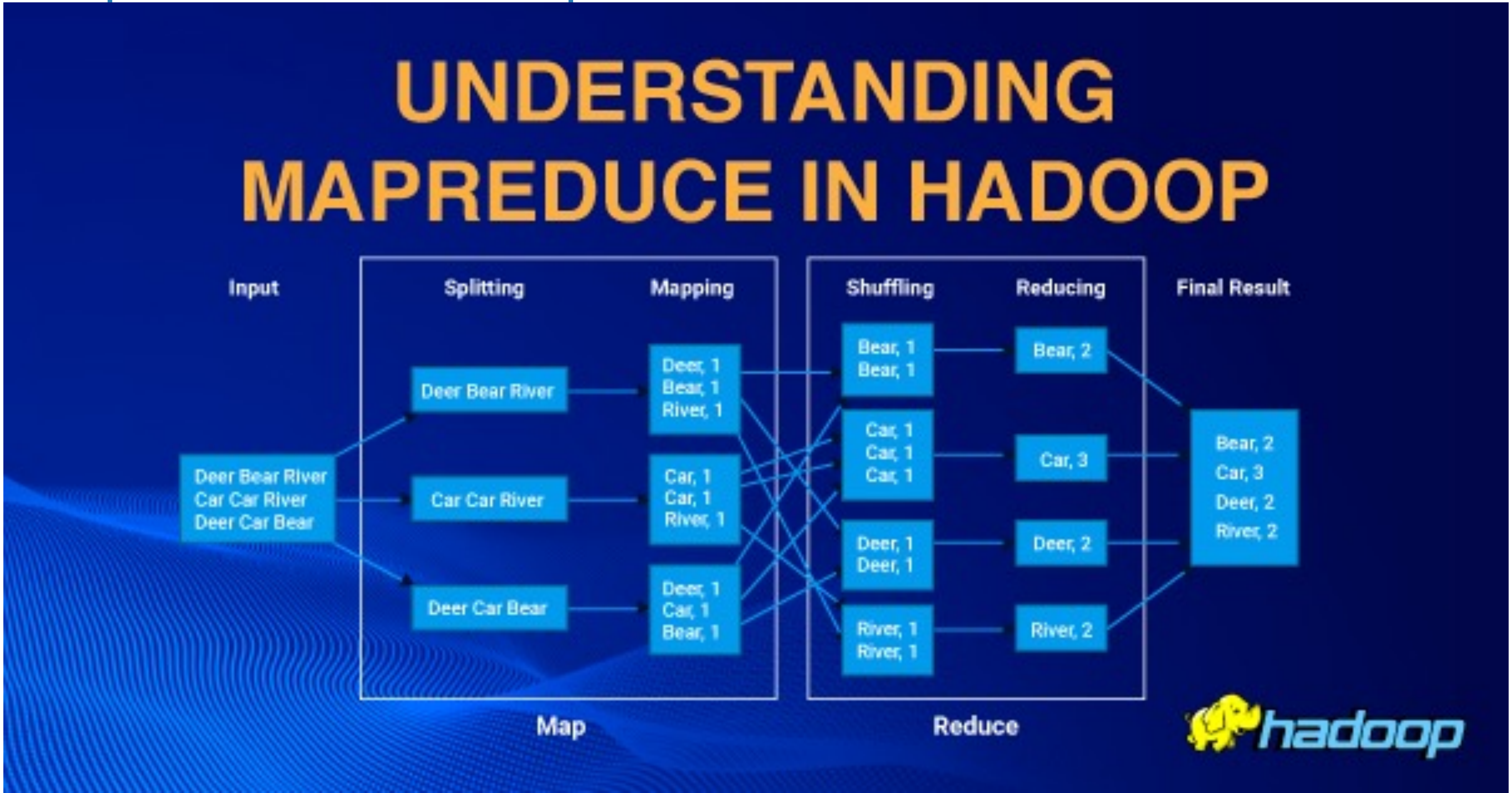
MapReduce and Hadoop

- You should be able to identify when a system uses MapReduce
- There are several iterations on MapReduce, but it was the first such approach to processing large volumes of data
- MapReduce and its variations are still widely used, e.g., Netflix recommendation system

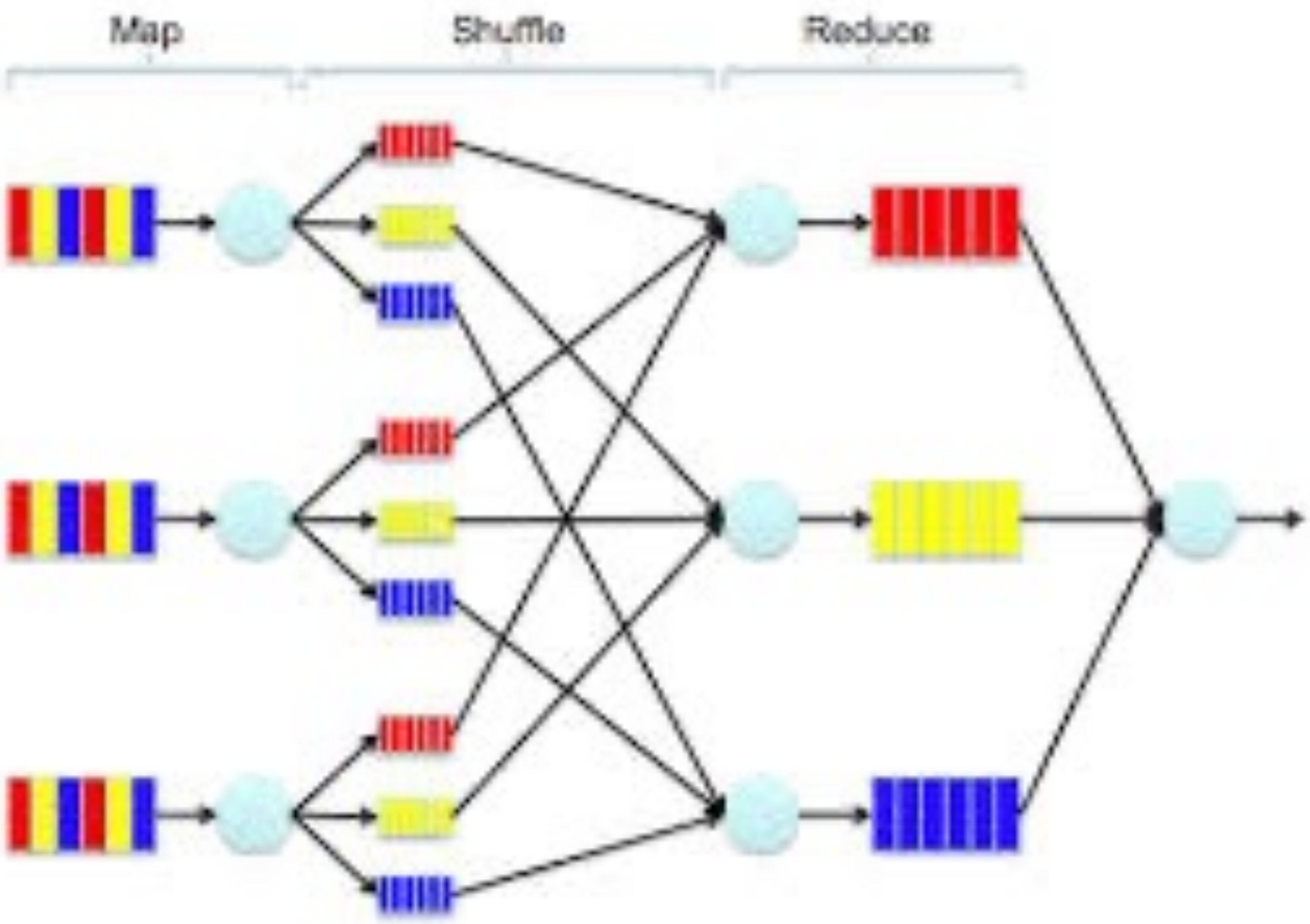
MapReduce

- 1990’s: lots of data being collected, data centers were created with lots of machines, but coordinating how the machines process data had high overhead
- Insight: most tasks could be split into three steps: map, shuffle, and reduce
  - Map: Each worker (machine) applies the map function to its portion of the data to generate a set of <key, value> pairs and writes these to temporary storage
  - Shuffle: Workers move data based on keys so that all key-value pairs with the same key are at the same worker
  - Reduce: Workers process each key in parallel

MapReduce Example



MapReduce Visualization



Hadoop

- Hadoop is a framework to process large datasets. It has three parts:

1. Data storage: It uses Hadoop Data File System (HDFS)
    - Separates data into blocks and stores them on different workers. To ensure fault-tolerance, it stores each block at three different locations
2. Map-Reduce
3. YARN: yet another resource negotiator
    - Processes job requests and manages cluster resources (CPU, bandwidth, RAM, etc)

MapReduce and Hadoop

- You should be able to identify when a system uses MapReduce
- There are several iterations on MapReduce, but it was the first such approach to processing large volumes of data
- MapReduce and its variations are still widely used, e.g., Netflix recommendation system

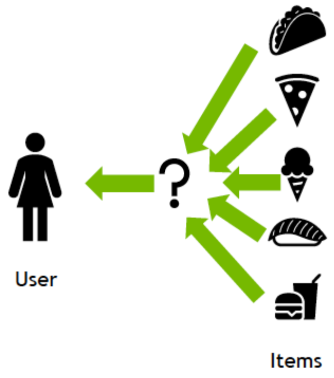
Scale of ML in Industry

- Most common is in the middle

One model; low severity of wrong predictions, not necessary to run extremely fast

GB – TB of data daily; 10s – 100s of data scientists/ML engineers

Several models - all must run accurately and quickly

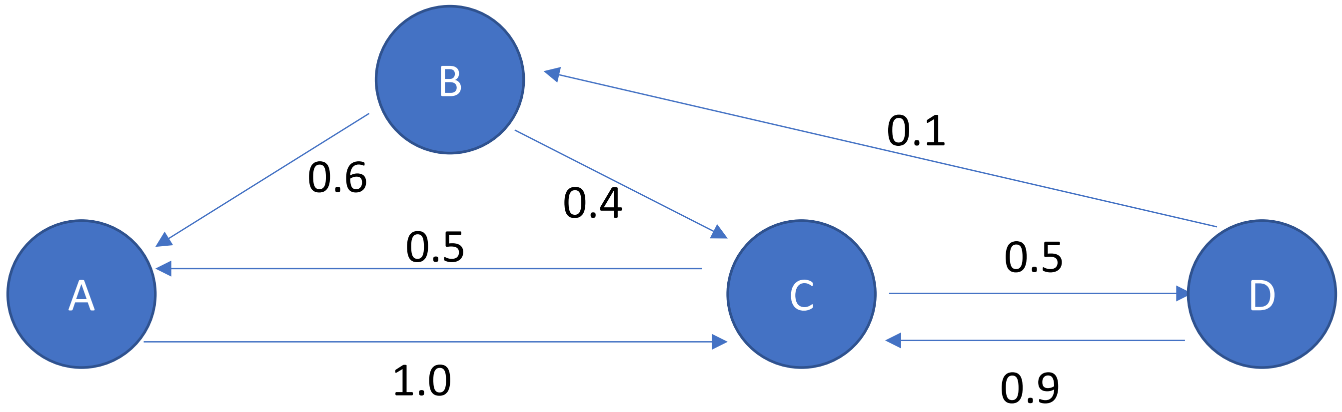


Largest Model Used in Companies

- 60%: 1 GB – 500 GB
- 14%: 500 GB – 10 TB
- 5%: >10 TB
- 21%: <1 GB

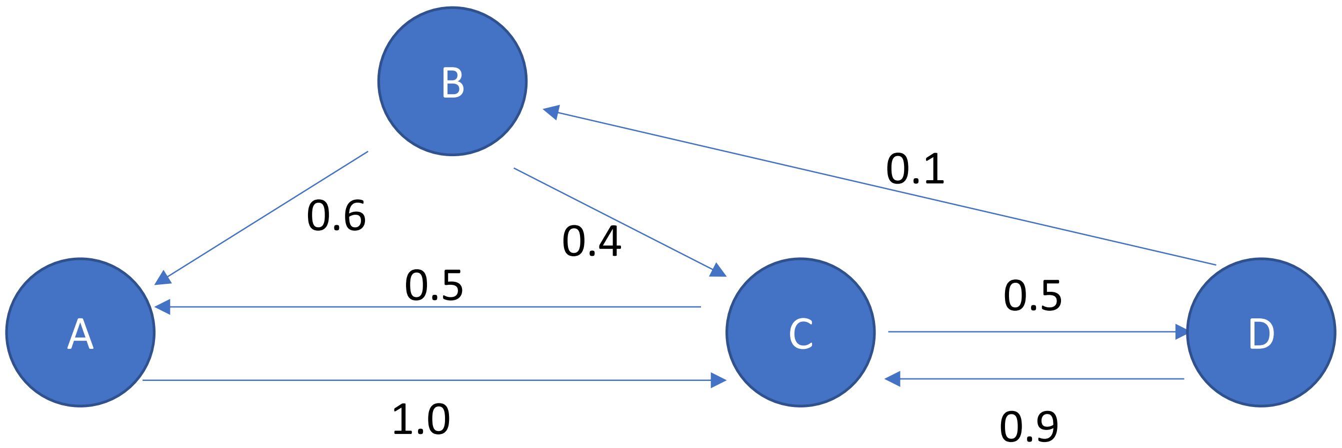
PageRank

- Good common algorithm to know for big data interviews – Google uses this (now with additions) to rank web pages in search
- Suppose we have three web pages, some with links other. The probabilities that people follow each link are marked



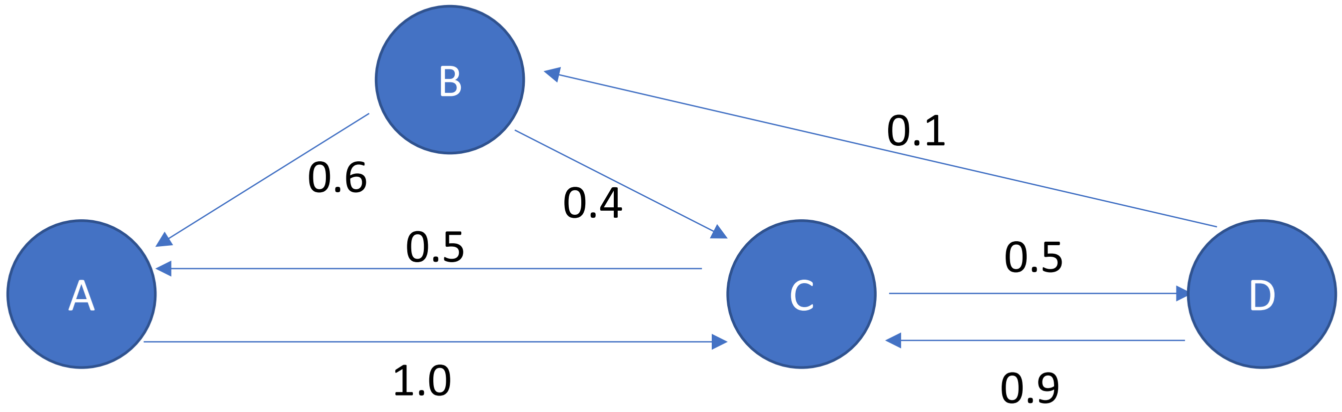
PageRank Algorithm

- Aims to find the relative importance of websites
- Stationary distribution: proportion of people at each site stops changing (though the people themselves might)
- Our goal is to find the stationary distribution
- Intuition: if a site is pointed to by many other sites, in particular other important sites, it is important



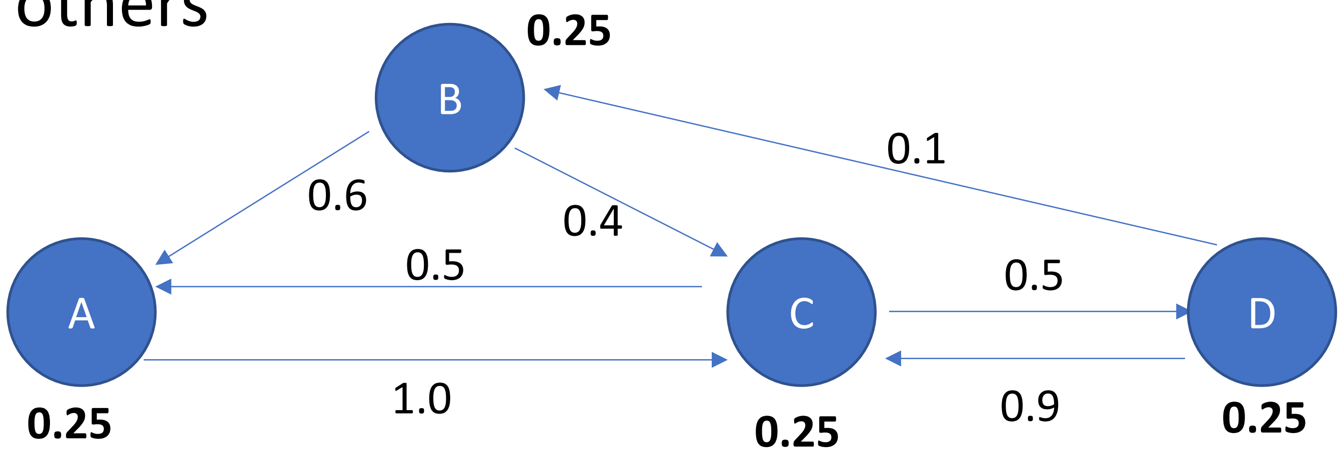
PageRank Algorithm

- We'll start by initializing the distributions uniformly
- P(A) = P(B) = P(C) = P(D) = 0.25



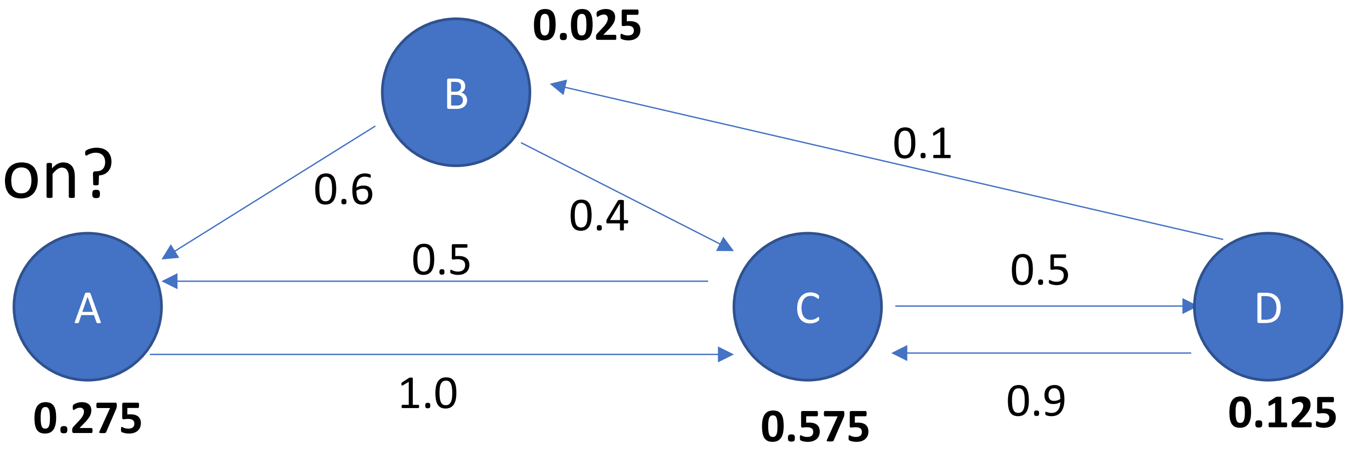
PageRank Algorithm

- Then we start taking steps. For a node *s*, the probability that a user is at *s* in the next step is the probability that the user was at another node *n* and then stepped from *n* to *s*, for every other node *n*
- $P(A)_{t+1} = P(B)_t * P(B \rightarrow A) + P(C)_t * P(C \rightarrow A) + P(D)_t * P(D \rightarrow A)$
- $P(A)_{t+1} = 0.25*0.6 + 0.25*0.5 + 0.25*0 = 0.275$
- Calculate the others



PageRank Algorithm

- $P(A)_{t+1}$ = 0.25*0.6 + 0.25*0.5 + 0.25*0 = 0.275
- $P(B)_{t+1}$ = 0.25*0 + 0.25*0 + 0.25*0.1 = 0.025
- $P(C)_{t+1}$ = 0.25*1 + 0.25*0.4 + 0.25*0.9 = 0.575
- $P(D)_{t+1}$ = 0.25*0 + 0.25*0 + 0.25*0.5 = 0.125
- Notice that the sum is still 1.0
- New distributions ->
- Is this the stable distribution?



Acknowledgments

- Chip Huyen, Stanford Machine Learning System Design