به نام آنکه به شماره موجودات آگاه است

آمار و رشته‌ی علوم داده‌ها

عادل محمدپور
دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)
AdelM.ir
adel@aut.ac.ir
Learning Data Science

- toolbox
- data munging
- data ingestion
- big data
- visualisation
- text mining/NLP
- machine learning
- programming
- fundamentals
- statistics
- skills
Skills

Curious, Creative, Tenacious, Resourceful, Inventive, Innovative, Deep Technical Skills, Communication Skills, Presentation Skills, Sees beyond the obvious
Fundamentals

- Hash Functions
- Binary Trees
- Matrices & Linear Algebra
- Relational Algebra
- SQL
- Sun Modelling
- Star Schemas
- Database Basics
- Joins (Inner, Outer, Cross)
- CAP Theorem
- Tabular Data
- Data Frames & Series
- Sharding
- OLAP (Online Analytical Processing)
- Multidimensional Modelling
- ETL (Extract-Transform-Load)
- Reporting vs BI vs Analytics
- JASON & XML
- NoSQL Fundamentals
- Regex (Regular Expressions)
- Env. Setup
- Vendor Ecosystem
Statistics

- Descriptive Statistics (Mean, Median, Mode, Standard Deviation)
- Exploratory Data Analysis
- Normal Distribution
- Histograms
- Percentiles & Quartiles
- Probability Theory
- Random Variables
- Cumulative Distribution Functions (CDF, Normal)
- Continuous Distributions
- Skewness
- ANOVA (Analysis of Variance)
- Probability Density Function (PDF)
- Central Limit Theorem
- Monte Carlo Simulation
- p-Value
- Chi-Square Test
- Estimation
- Confidence Interval
- Maximum Likelihood Estimation (MLE)
- Kernel Density Estimation
- Regression
Text Mining / Natural Language Processing

- Vocabulary Man.
- Classify Text
- Using NLTK
- Using Weka
- Feature Extraction
- Market Based Analysis
- Association Rules
- Support Vector Machines
- Term Frequency & Weight
- UIMA (Unstructured Information Management Architecture)
- Text Analysis
- Named Entity Recognition
- Term Document Analysis
- Corpus
- Social Networks
- Sentiment Analysis
- Clustering
Data Ingestion

- Using ETL
- How much Data?
- Google Open Refine
- Data Survey
- Transformation & Enrichment
- Data Fusion
- Data Integration
- Data Sources & Acquisition
- Summary of Data Formats
Data Munging

- Normalisation
- Data Scrubbing
- Handling Missing Values
- Unbiased Estimators
- Binning Sparse Values
- Feature Extraction
- De-noising
- Sampling
- Stratified Sampling
- Principal Component Analysis

Analysis Toolpak
The final step on this journey, is also your first...

Data Science never finishes, there are always new developments and new tools coming.

Keep Looking and Keep Learning!

And if you do find something interesting, let us know here so we can share it 😊

www.exploringdatascience.com