Introduction to the course

DS 503: Advanced Data Analytics

Instructor: Dr. Gagan Raj Gupta

What is data analytics?

- Deriving meaningful insights from data, solving problems
- Types of Data
  - Structured: Tabular data (Excel sheets), Time series,
  - Unstructured: Images, Videos, Audio, Transaction Logs, Text documents
  - Semi-structured: Graphs with edge and node attributes
- Data Analysis Examples
  - E-commerce: Product Display (best match), Pricing, Related Product, Popular Items, Distinct
  - Weather forecasting: Forecasting Rainfall, Temperature, Crop yields, Anomaly detection
  - Video analysis: Activity recognition, Crop disease, Rescue Operations
  - Natural Language (Thousands of words): Document classification, Sentiment analysis, Summarization, Fact Extraction

Data as a Matrix:

- **Rows:** Entities, instances, examples, records, transactions, objects, points, feature-vectors, tuples
  - Number of instances, n, is the size of the data
- **Columns:** Attributes, properties, features, dimensions, variables, fields
  - Number of attributes, d, is the dimensionality

Table 1.1. Extract from the Iris dataset

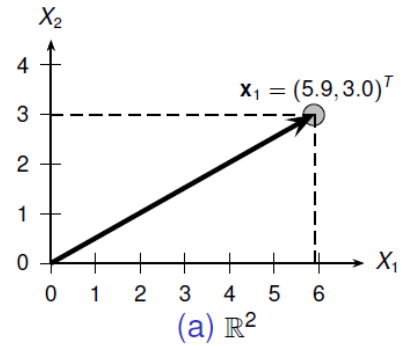
	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
⋮	⋮	⋮	⋮	⋮	⋮
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Numeric Attributes	Categorical Attributes
Interval-scaled: only differences are meaningful e.g., temperature	**Nominal:** only equality is meaningful e.g., domain(Gender) = { M, F}
Ratio-scaled: differences and ratios are meaningful e.g., Age	**Ordinal:** both equality (are two values the same?) and inequality (is one value less than another?) are meaningful e.g., domain(Education) = { High School, BS, MS, PhD}

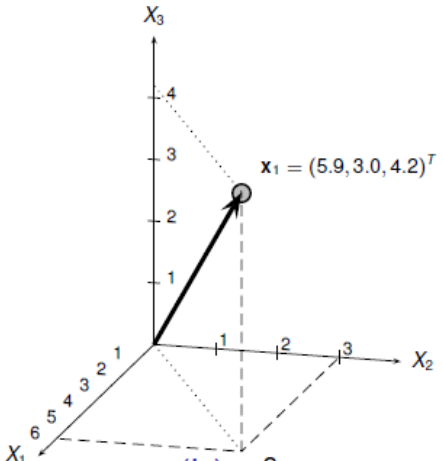
Algebraic and Geometric View of Data

- Geometric/Algebraic View
  - Numeric attributes can be represented as vectors/points
    - Each row is d-dimensional vector
    - Each column is a n-dimensional vector
  - Perform vector operations: Distance, Angle
  - Compute Mean, Variance
  - Compute the Rank, basis vectors, subspaces
  - Perform operations such as projections
  - Centering data (subtract the mean from each row)

$$\mathbf{x}_1 = (5.9, 3.0, 4.2, 1.5)^T$$



(a) $\mathbb{R}^2$



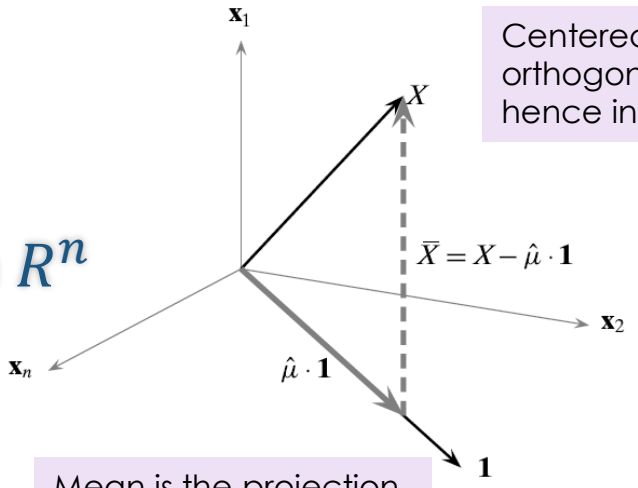
(b) $\mathbb{R}^3$

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \mu^T = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} - \begin{pmatrix} \mu^T \\ \mu^T \\ \vdots \\ \mu^T \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T - \mu^T \\ \mathbf{x}_2^T - \mu^T \\ \vdots \\ \mathbf{x}_n^T - \mu^T \end{pmatrix} = \begin{pmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{pmatrix}$$

Geometric Interpretation of Mean, Correlation and Covariance

- In general, sample mean = $\hat{\boldsymbol{\mu}} = \frac{1}{n} D^T \mathbf{1}$
- Let Z_1 and Z_2 denote centered attribute vectors in R^n

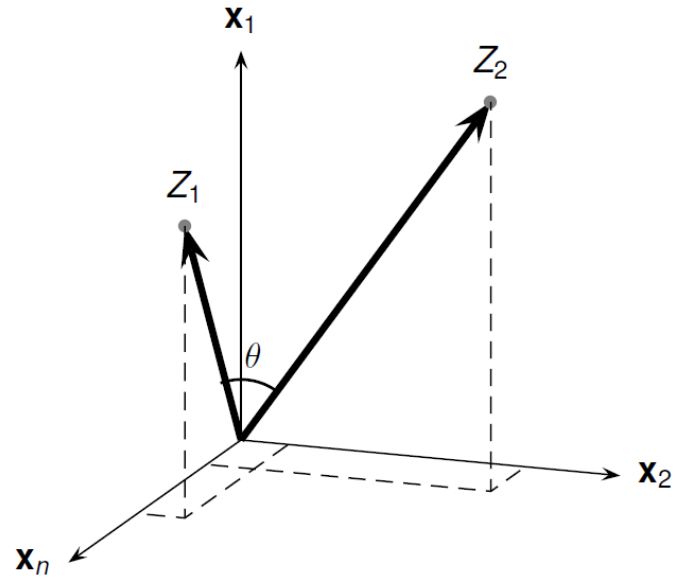
$$Z_1 = X_1 - \mathbf{1} \cdot \hat{\mu}_1 = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix} \qquad Z_2 = X_2 - \mathbf{1} \cdot \hat{\mu}_2 = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}$$



Centered attribute vector is orthogonal to ones vector and hence in n-1 dim subspace

Mean is the projection of X onto ones vector

- The sample covariance is given as: $\hat{\sigma}_{12} = \frac{Z_1^T Z_2}{n}$
- Sample Correlation is given as: $\hat{\rho}_{12} = \frac{Z_1^T Z_2}{||Z_1|| ||Z_2||} = \cos(\theta)$
- The correlation coefficient is simply the cosine of the angle between the two centered attribute vectors



Covariance matrix

- The variance-covariance matrix for 2 attributes X_1 and X_2 can be summarized in the 2x2 symmetric covariance matrix Σ
- $\Sigma = E[(X-\mu)(X-\mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$
- This can be generalized to d dimensions, and we obtain a symmetric positive defini $\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$
  - Its eigenvalues are real and non-negative

General form of covariance matrix

- The total variance is given as, $var(D) = tr(\Sigma) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2$
- The generalized variance is given as $|\Sigma| = \det(\Sigma) = \prod_{i=1}^{d} \lambda_i \; (= \sigma_1^2\sigma_2^2 - \sigma_{12}^2 \geq 0 \; in \; 2 \; dim)$
- Sample covariance matrix can be computed in terms of $\overline{D} = D - \mathbf{1}.\hat{\boldsymbol{\mu}}^T$, the centered data matrix
  - $\hat{\Sigma} = \frac{1}{n}(\overline{D}^T\overline{D})$ {pairwisedot products of centered attribute vectors}
  - $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \overline{x}_i \overline{x}_i^T$ {Sum of rank-1 matrices}

$$\frac{1}{n}\begin{pmatrix} \overline{X}_1^T\overline{X}_1 & \overline{X}_1^T\overline{X}_2 & \cdots & \overline{X}_1^T\overline{X}_d \\ \overline{X}_2^T\overline{X}_1 & \overline{X}_2^T\overline{X}_2 & \cdots & \overline{X}_2^T\overline{X}_d \\ \vdots & \vdots & \ddots & \vdots \\ \overline{X}_d^T\overline{X}_1 & \overline{X}_d^T\overline{X}_2 & \cdots & \overline{X}_d^T\overline{X}_d \end{pmatrix}$$

Data Normalization

If the attribute values are in vastly different scales, then it is necessary to normalize them.

Range Normalization: Let X be an attribute and let $x_1, x_2, \ldots, x_n$ be a random sample drawn from X. In *range normalization* each value is scaled by the sample range $\hat{r}$ of X:

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\hat{r}} = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

After transformation the new attribute takes on values in the range $[0, 1]$.

Standard Score Normalization: Also called z-normalization; each value is replaced by its z-score:

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

where $\hat{\mu}$ is the sample mean and $\hat{\sigma}^2$ is the sample variance of X. After transformation, the new attribute has mean $\hat{\mu}' = 0$, and standard deviation $\hat{\sigma}' = 1$.

Probabilistic View of Data

- Probabilistic View
  - Each numeric attribute is a random variable
  - Compute joint distributions, estimate parameters etc.
- Given dataset **D**, the n data points x_i (with $1 \leq i \leq n$) constitute a d-dimensional multivariate random sample drawn from the vector random variable

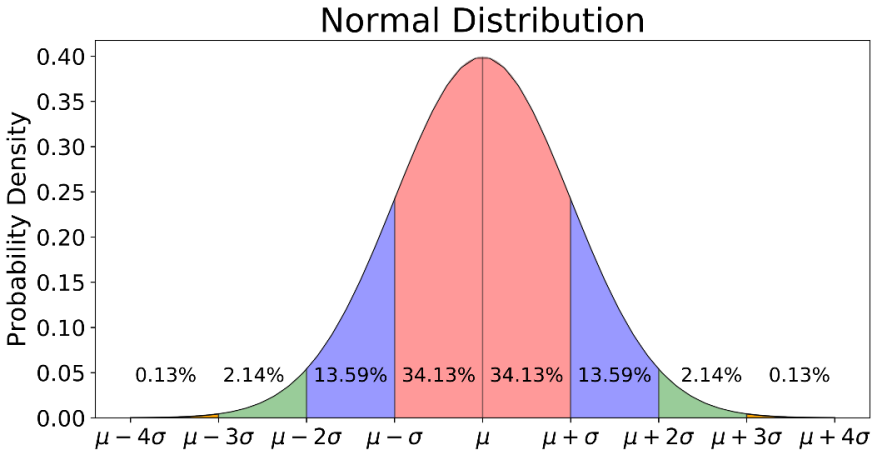
$$X = (X_1, X_2, \ldots, X_d)$$

- If we assume x_i to be independent and identically distributed, their joint distribution is given as $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$ where f_X is the probability mass or density function for X.
- Assuming that the d attributes $X_1, X_2, \ldots, X_d$ are statistically independent, the joint distribution for the entire dataset is given as:

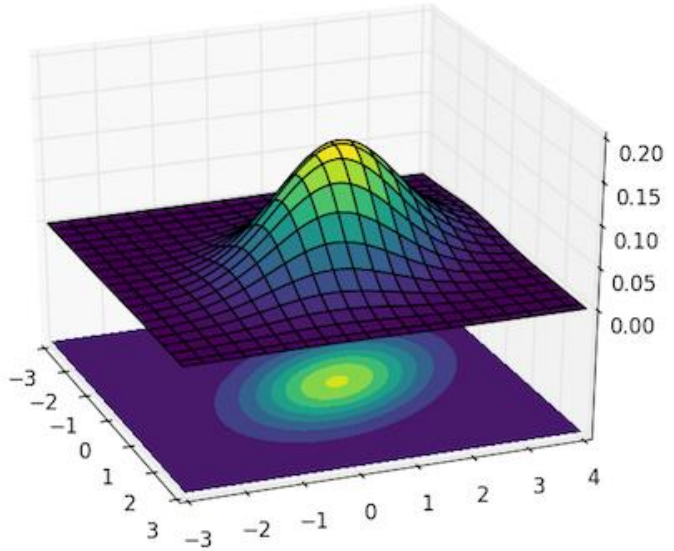
$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \prod_{j=1}^{d} f_{X_j}(x_{ij})$$

Gaussian Distribution

- Gaussian or Normal density function: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\,\sigma^2}}$
  - μ is the mean of the random variable X, and σ is its std. deviation
  - A one-dimensional Gaussian has its mass close to its mean
  - Probability density decreases exponentially with squared distance



- Bivariate Normal: modeling joint distribution for X1 and X2
  - $f(x) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{\frac{-(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}}$
  - $\mu = (\mu_1, \mu_2)$ is the mean
  - Σ is the covariance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$
    - $\{\sigma_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \mu_i)(x_{kj} - \mu_j)\};\ \sigma_i^2 = \sigma_{ii}$
  - $(x-\mu)^T\Sigma^{-1}(x-\mu)$ is the Mahalanobis distance of x from mean



$$\Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1.5 \end{pmatrix}$$

Geometry of the Multi-variate Normal

- Compared to the standard multivariate normal, the mean $\boldsymbol{\mu}$ translates the center of the distribution, whereas the covariance matrix $\boldsymbol{\Sigma}$ scales and rotates the distribution.
- The eigen-decomposition of $\boldsymbol{\Sigma}$ is given as $\boldsymbol{\Sigma}\mathbf{u_i} = \lambda_i \boldsymbol{u_i}$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ are the eigenvalues and $\boldsymbol{u_i}$ the corresponding eigenvectors.
- This can be compactly written as $\boldsymbol{\Sigma} = \mathbf{U}\,\boldsymbol{\Lambda}\mathbf{U}^{\mathbf{T}}$ where

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} \qquad \mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix}$$

- The eigenvectors represent the new basis vectors, with the covariance matrix given by $\boldsymbol{\Lambda}$ (all covariances become zero).
- Since the trace of a square matrix is invariant to similarity transformation, such as a change of basis, we have

$$\text{var}(\mathbf{D}) = tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{d} \sigma_i^2 = \sum_{i=1}^{d} \lambda_i = tr(\boldsymbol{\Lambda})$$

What are some common problems in data analysis?

- Getting data is becoming easier day by day, but we have too much to analyze
- Data has errors of various types (missing, incorrect etc.) and is hard to clean
- Data is usually high-dimensional (involving lot of columns or features)
- Data is incomplete (matrix completion, compressed sensing, signal re-construction)
- Data may have complex correlations (e.g. graph data, time-series data)
- Data is being generated at a great speed and it is too expensive to store all of it
- Data on the network in encrypted

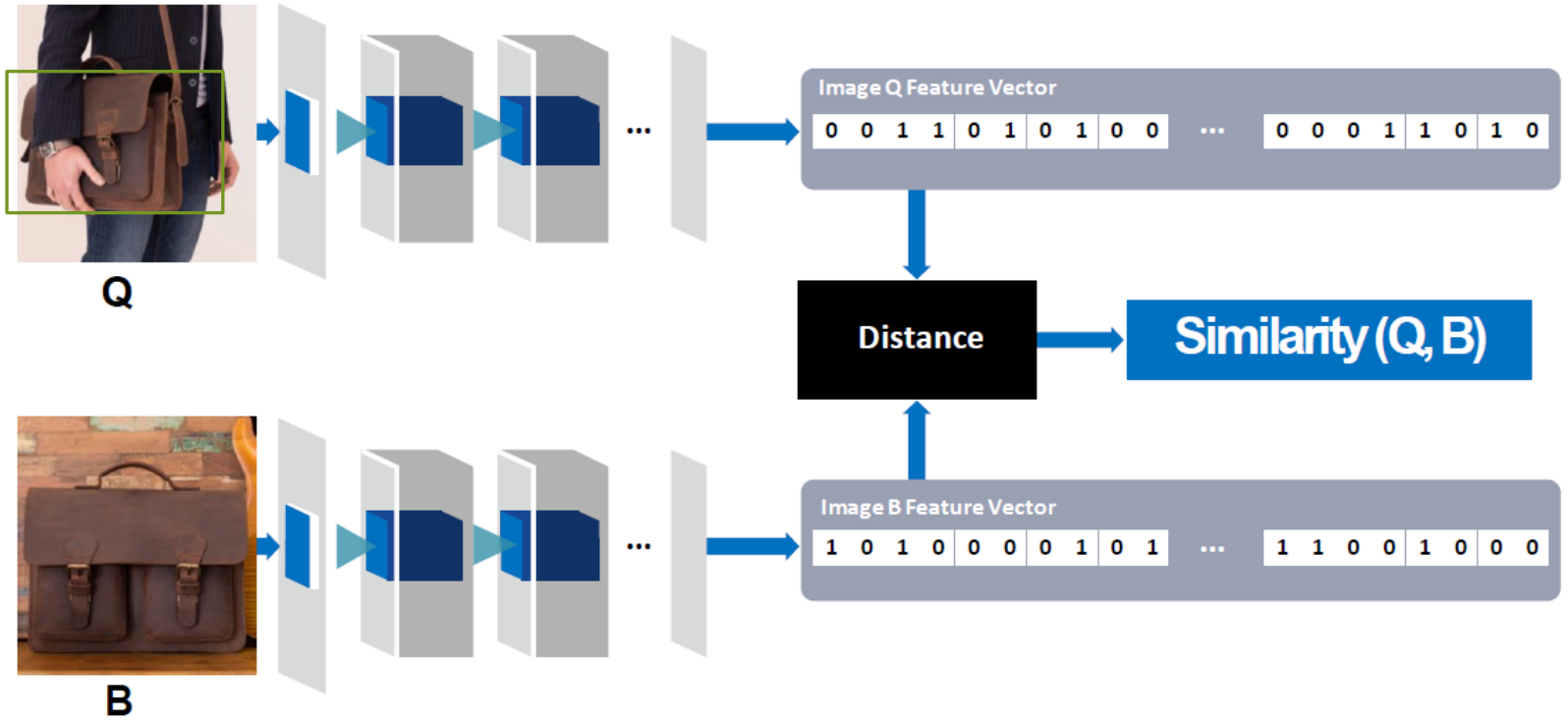
We are often asked to answer difficult questions from this **messy data and make decisions (often in real-time)**. This course teaches some **techniques** to handle these challenges. In fact, these can be used to our **advantage**!!

Various themes in the course

Theme	Techniques	Lectures
High-dimensional Data Analysis	Projections to low dimensions, Random Projections, LSH, Matrix Completion, Compressed Sensing	1-6 (F1)
Streaming Data Analysis	Random Samples, Frequent Items, Bloom Filters, Count Unique, Moments Estimation	7-9 (F2)
Large Scale ML	Online Learning, Decision Making, Decision Trees, Neural Networks, SVMs, Online Clustering, Model and Data Parallel	10-15 (F3)
Graph Data Analysis	Graph Kernels, Embeddings, Neural Nets, Attention, Knowledge Graphs, Graph Sampling, Completion	15-18 (F4)
Time Series Analysis	Matrix Profile and Applications, LSTM	19-22 (F5)

- Course assignments will aim to reinforce the techniques taught in the class
  - Mix of theory and programming problems in each assignment
- Major project should try to combine multiple techniques or study one technique in depth
  - Read research papers, implement and improve/ Solve new problem (present in F6)
  - Analyze a new data-set of your interest

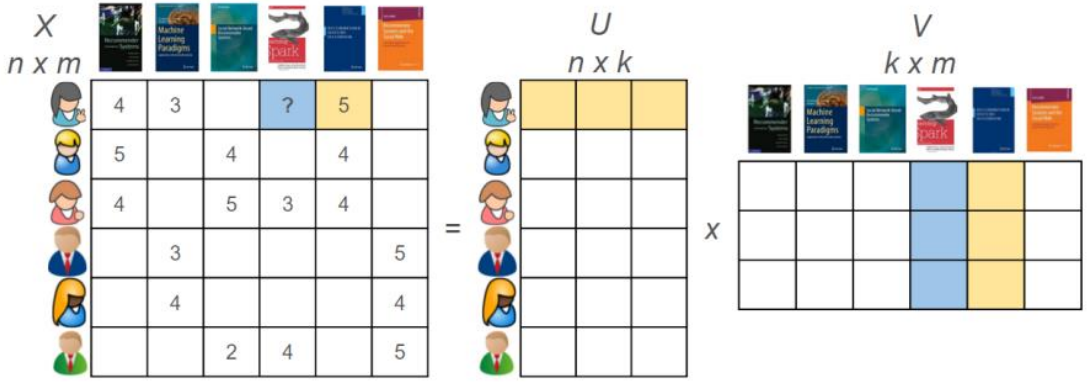
Application: Finding similar objects/items



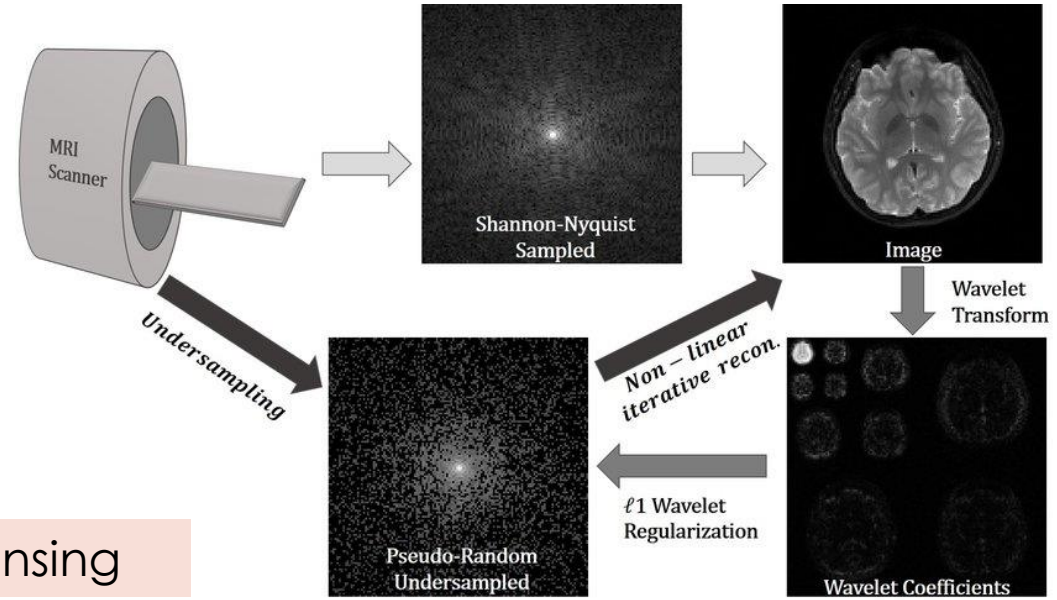
- Pinterest Visual Search
  - Given a query image patch (Q) , find similar images among the collection
  - Determine a feature vector and find nearest neighbors

Applications: Data Completion

Image Completion Problem



Collaborative Filtering

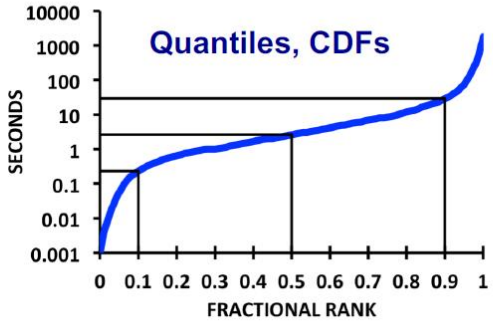


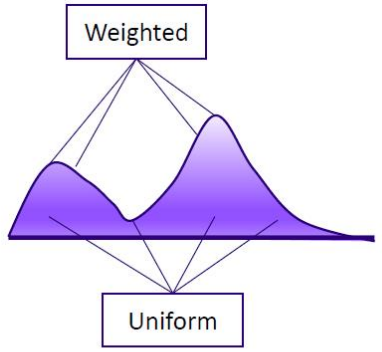
MRI: Compressed Sensing

Applications: Streaming Data Analysis

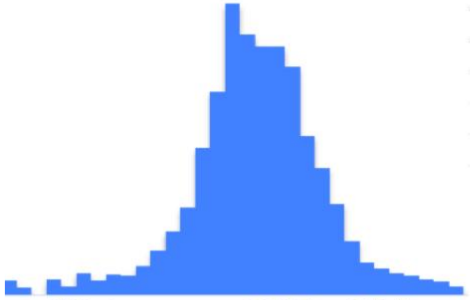
Some Very Common Queries …

Counting Unique Identifiers

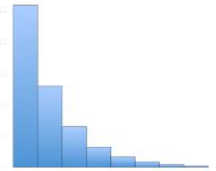




Reservoir Sampling



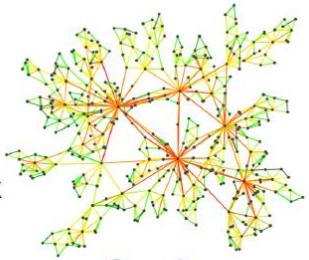
Histograms, PMFs, CDFs



Frequent Items

$$\begin{Bmatrix} 5 & \cdots & 2 \\ \vdots & \ddots & \vdots \\ 4 & \cdots & 3 \end{Bmatrix}$$

Vector & Matrix Operations: SVD, etc.

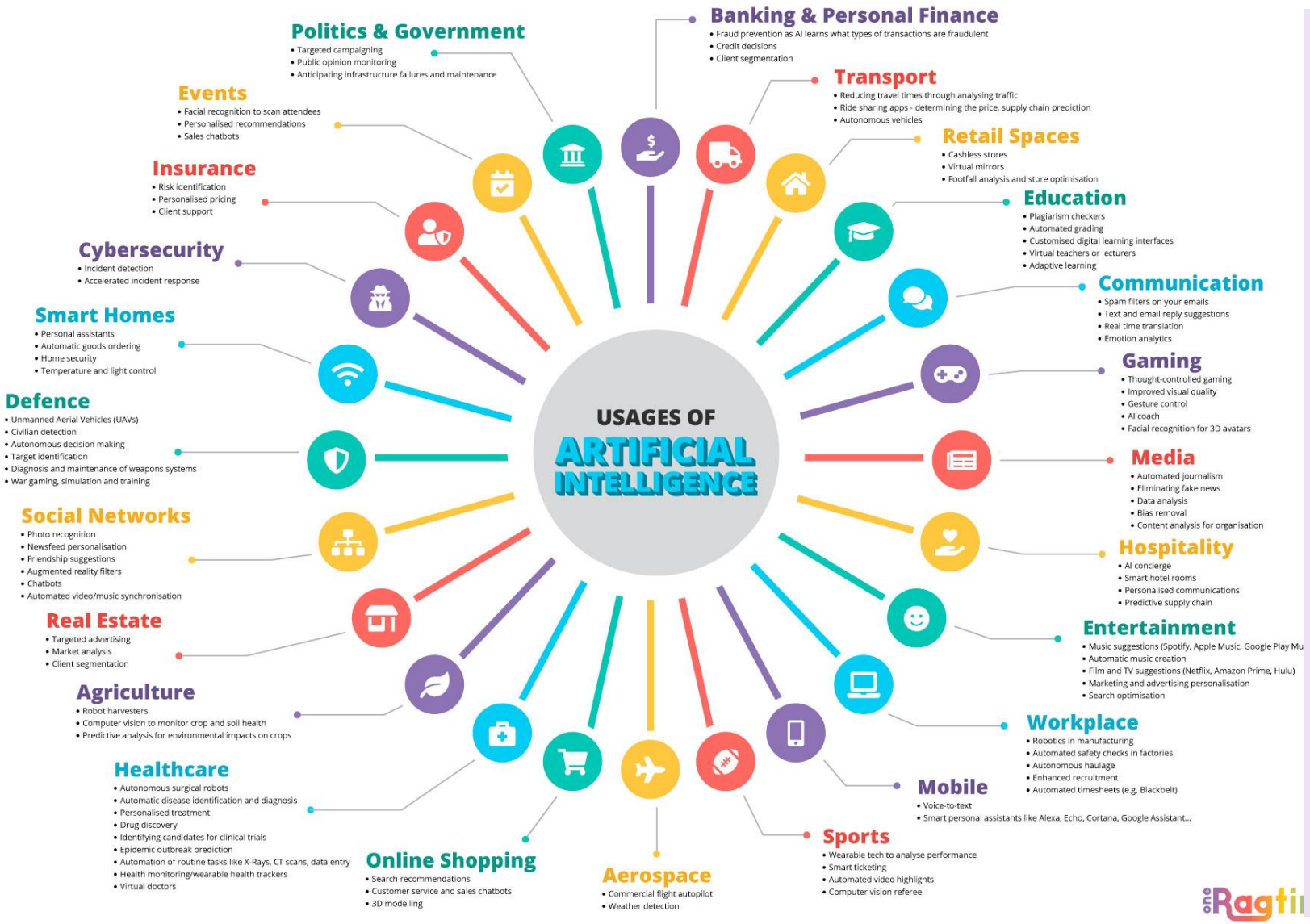


Graphs

… All Are Computationally Difficult

Computation tasks on data streams (Google Queries, Tweets, Ad click streams, Image data, Telephone calls, Network packets, Industry Automation, Cloud Infra Monitoring)

Applications: Large Scale ML

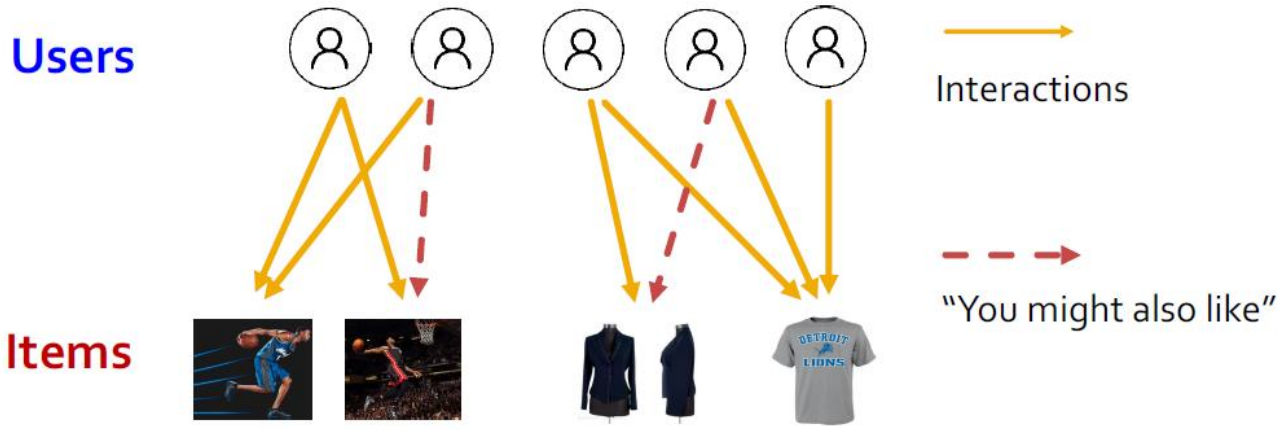


Facial Recognition
Speech, handwriting recognition
Chatbots
Incident detection
Home security
Target identification
Friendship suggestions
Advertising
Market Analysis
Crop monitoring
Drug Discovery
Weather prediction
Music suggestions
Bias removal in media
Spam filters, Sentiment analysis
Fraud prevention
Credit decisions
Ride sharing apps

Source

GNN Application: Recommendation Systems

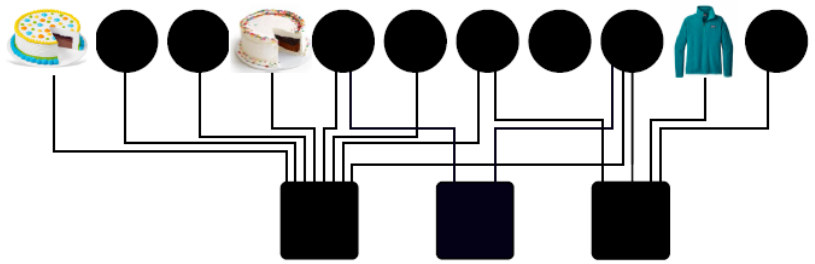
- **Users interacts with items**
- Watch movies, buy merchandise, listen to music
- **Nodes:** Users and items
- **Edges:** User-item interactions
- **Goal: Recommend items users might like**

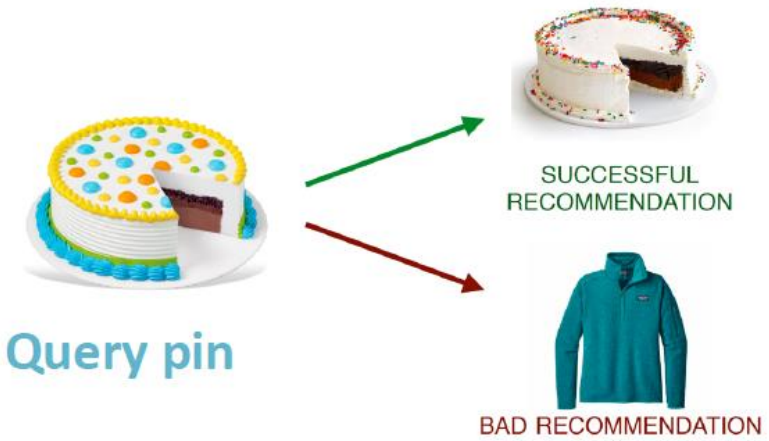


Task: Recommend related pins to users

Task: Learn node embeddings z_i such that

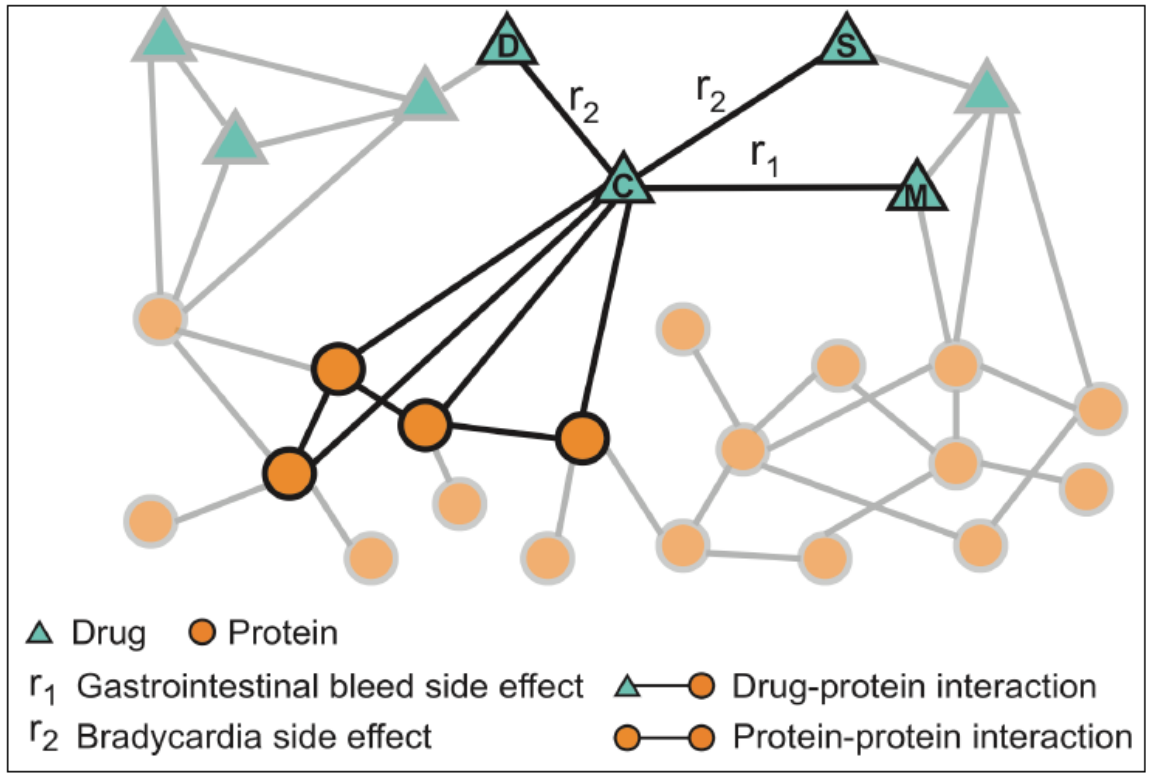
$$d(z_{cake1}, z_{cake2}) < d(z_{cake1}, z_{sweater})$$



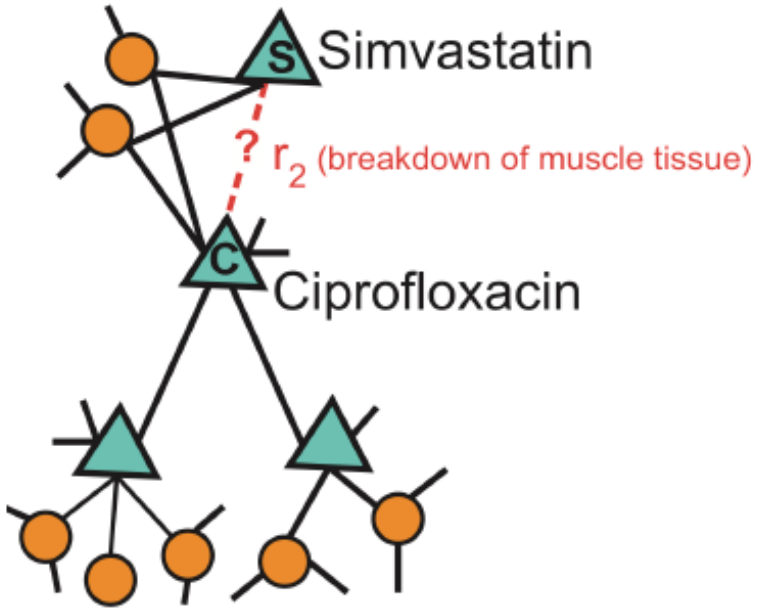


GNN Application: Biomedical Graph Link Prediction

- **Nodes**: Drugs & Proteins
- **Edges**: Interactions

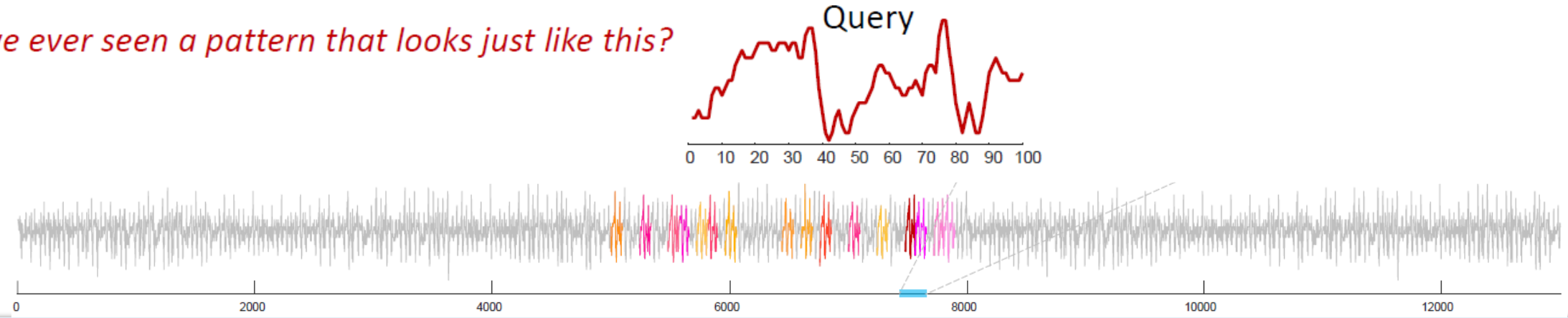


Query: How likely will Simvastatin and Ciprofloxacin, when taken together, break down muscle tissue?



Time Series Analysis Applications

Have we ever seen a pattern that looks just like this?



- Are there any repeated patterns in my data? (Medical applications/monitoring patients)
- What are three most unusual days in the last 3 months? (Demands/Network/Capacity Planning)
- Is there any pattern that is common to these two time series? (Music plagiarism)
- How do these two time series differ in terms of alignment? (Songs/Music)
- If you had to summarize this long time series with just two shorter examples, what would they be?
- How do I quickly search this long dataset for this pattern, if an approximate search is acceptable?
- What is most likely to happen next (Shape based prediction)? (Power demand/Industrial applications)

Weightage details

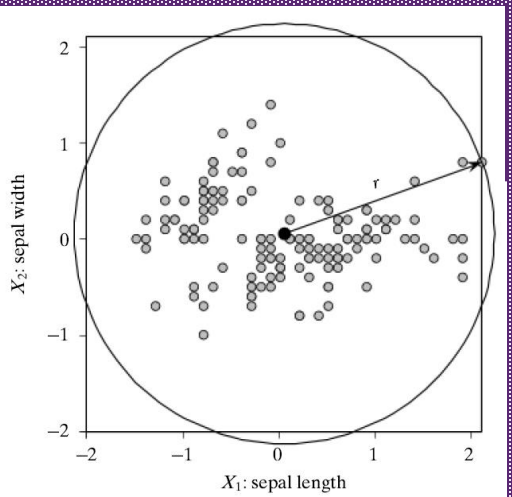
Assignments	Quizzes	Exams	Project Proposals	Project Presentation	Project Report	Class participation
5x5=25	5	2x15=30	5	5	20	10

- **Assignments need to be individual.**
  - 6 late days allowed in the semester. Use them judiciously. No credit after that.
  - **0 for any copying detected for all parties involved**
  - **Copying from internet is also copying**
  - **2 instances will lead to F or D in the course**
- Best 5 out of x quizzes will be used to grade (no pre-announcement)
- **Exams** will likely have a component of **programming**
- **Projects in groups of 2 students**
- Class Participation includes
  - Marks for **scribe/notes and detecting typos in slides/course materials**
    - Contact instructor/TAs if you are interested in scribing/lecture notes
  - First correct solution to **bonus problems** in lectures + first correct quiz solutions

Project Planning

- **Projects in groups of 2 students**
  - Keep a record of participation of your partner
  - Equal participation/contribution required
- Choose a project topic early. **Final proposal deadline (Sep 20, 2021)**
  - List of interesting projects will be posted/updated
- Project proposal needs to explain the problem and provide references (template provided)
- In project presentation, explain the problem and related work done on it in the past
- All project presentations in Nov15-Nov30

High dimensional geometry



- Let D be a n x d data matrix.
  - High-dimensional data is very different from the 2-3 dimensional spaces
- **Hyper-rectangle:** The data space is a d-dimensional hyper-rectangle
  - $R_d = \prod_{j=1}^{d} [min(X_j), \max(X_j)]$
- **Hypercube:** Assume data is centered, and the value of the maximum attribute be m
  - The data hyperspace can be represented by: $\boldsymbol{H_d(2m)} = \{\boldsymbol{x} = (x_1, x_2, \dots, x_d)^T | \forall i, x_i \in [-m, m]\}$
  - The unit hypercube has all sides of length 1 and is denoted as $H_d(1)$
- **Hypersphere:** Assume data is centered, and the max. magnitude among all points be r
  - **Hyperball:** $\boldsymbol{B_d(r)} = \{\boldsymbol{x}| \ \ ||\boldsymbol{x}|| < \boldsymbol{r}\}$
  - **Surface of Hyperball is called Hypersphere:** $\boldsymbol{H_d(r)} = \{\boldsymbol{x}| \ \ ||\boldsymbol{x}|| = \boldsymbol{r}\}$

Hyperplanes

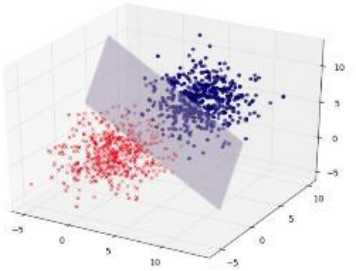
- A **hyperplane** in d dimensions is given as the set of all points $x \in R^d$ that satisfy the equation h(x)= 0, where h(x) is the hyperplane function, defined as follows:

$$h(x) = w^T x + b = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$$

  - Here w is the d dimensional weight vector, that is normal (orthogonal) to the hyperplane
  - b is the bias, which controls the intersection points of the hyperplane with the axes
- A hyperplane in d-dimensions has dimension d −1
- A hyperplane splits the original d-dimensional space into two half-spaces.
  - Points on one side satisfy h(x) >0
  - Points on the other side satisfy h(x) <0
  - Points on the hyperplane satisfy the condition h(x) = 0.

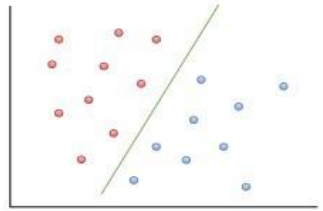
$\mathbf{w}^T\mathbf{x} = 0$

Hyperplane



$y = ax + b$

Line



High dimensional geometry (Volume)

- Volume of hypercube in with side length l is $\boxed{vol(H_d(l)) = l^d}$
- **FoDS Lemma 2.6:** Volume of hyperball and its corresponding hypersphere is

$$\boxed{vol(S_d(r)) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r^d}$$

- **Bonus Q:** Surface area of hypersphere can be obtained by differentiating volume w.r.t. r

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$$

$$\Gamma\left(\frac{d}{2}+1\right) = \begin{cases} (\frac{d}{2})! & \text{if } d \text{ is even} \\ \sqrt{\pi}\,\frac{d!!}{2^{(d+1)/2}} & \text{if } d \text{ is odd} \end{cases}$$

$$d!! = \begin{cases} 1 & \text{if } d = 0 \text{ or } d = 1 \\ d \cdot (d-2)!! & \text{if } d \geq 2 \end{cases}$$

- Consider a unit ball in d dimensions $B_d(1)$: the set of all points x such that |x|<= 1
  - Volume increases up to a point, then decreases and finally vanishes as d goes to infinity.
  - Volume is concentrated near its surface and is also concentrated at its equator
- These properties have important consequences, for problems such as nearest neighbor search

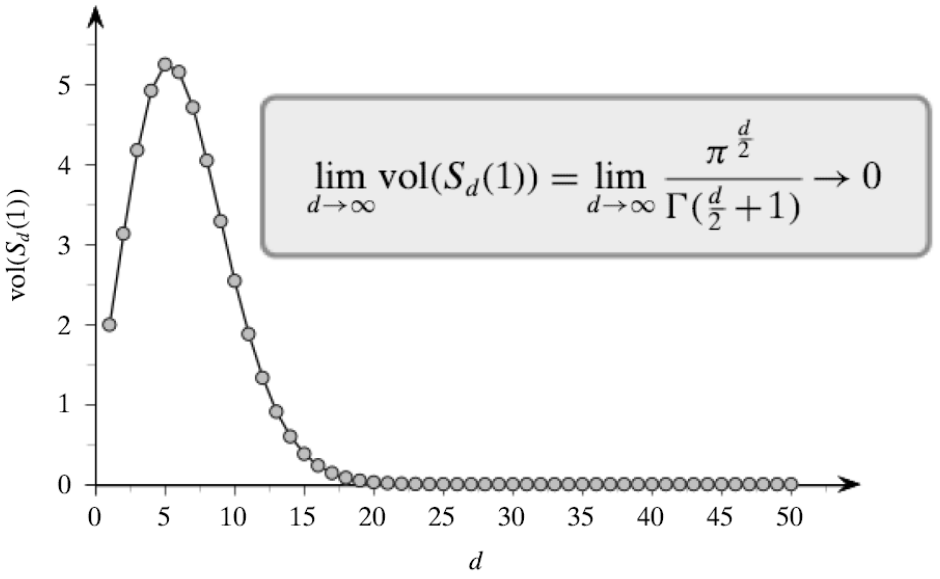


Figure 6.3. Volume of a unit hypersphere.

https://en.wikipedia.org/wiki/Gamma_function

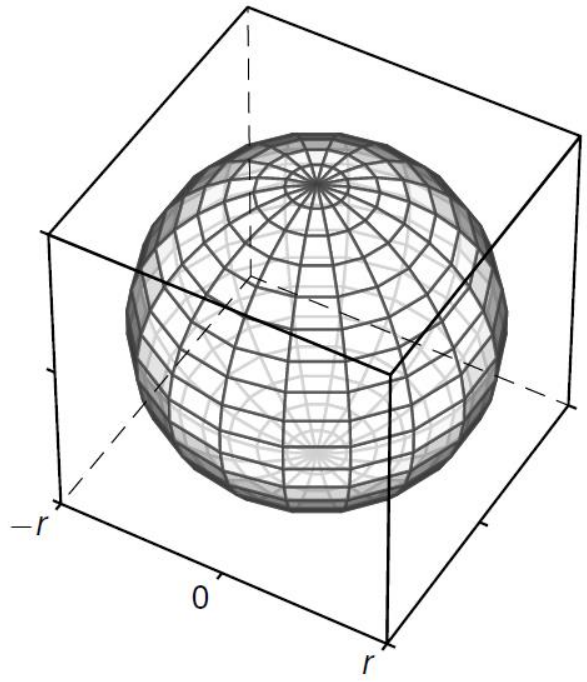
Hypersphere within Hypercube

- Consider the space enclosed within the largest hypersphere that can be accommodated within a hypercube (which represents the dataspace).
- Ratio of the volume of the hypersphere of radius r to the hypercube with side length l=2r is given as

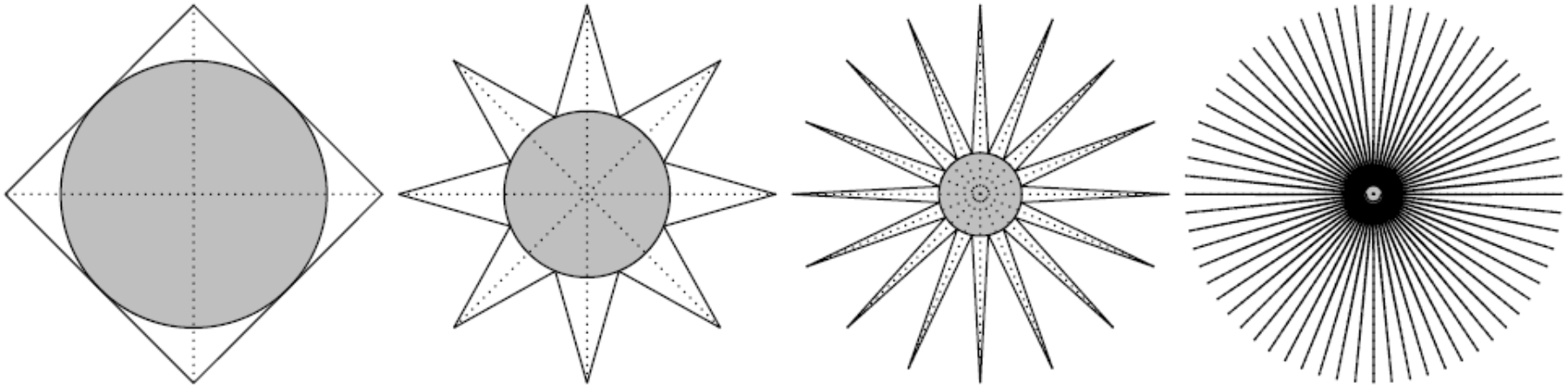
$$\text{In 2 dimensions: } \frac{\text{vol}(S_2(r))}{\text{vol}(H_2(2r))} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} = 78.5\%$$

$$\text{In 3 dimensions: } \frac{\text{vol}(S_3(r))}{\text{vol}(H_3(2r))} = \frac{\frac{4}{3}\pi r^3}{8r^3} = \frac{\pi}{6} = 52.4\%$$

$$\text{In } d \text{ dimensions: } \lim_{d\to\infty} \frac{\text{vol}(S_d(r))}{\text{vol}(H_d(2r))} = \lim_{d\to\infty} \frac{\pi^{d/2}}{2^d \Gamma(\frac{d}{2}+1)} \to 0$$

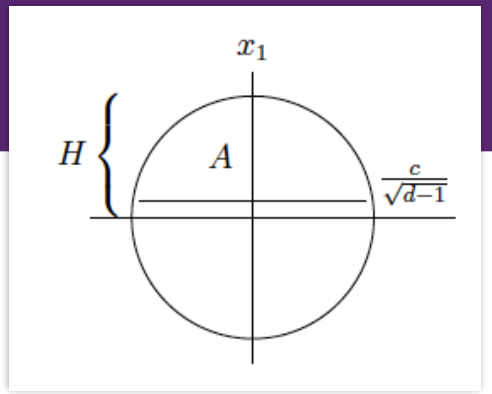


Geometry of the unit ball in high dimensions

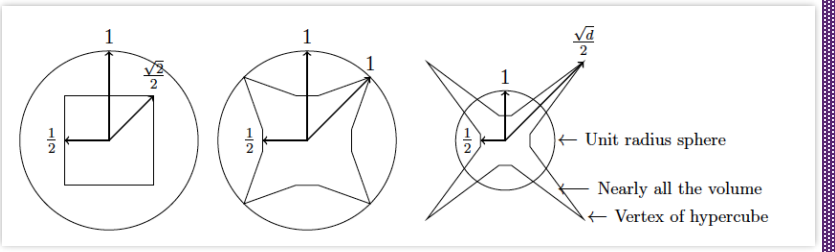


- Nearly all the points are near the surface and at the same time are within a small box
- This is because points on the surface of the ball satisfy $x_1^2 + x_2^2 + \cdots + x_d^2 = 1$, so for each coordinate the typical value will be $\frac{1}{\sqrt{d}}$
- If we draw n points at random from the unit ball, w.h.p. they will be close to unit length and will be pair-wise orthogonal

Volume near the equator

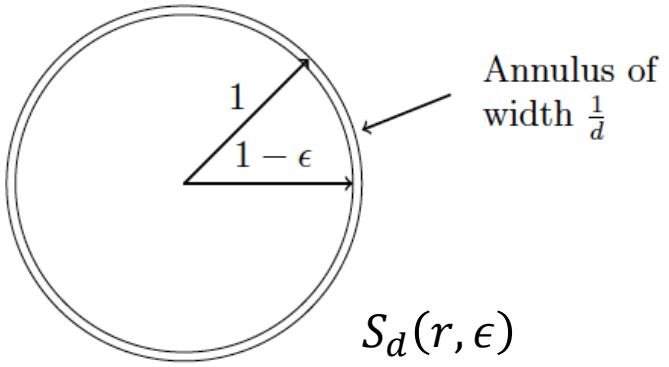


- Most of the volume of unit ball in high dimensions is concentrated near "equator"
- Take x_1 = arbitrary coordinate. Most of the volume has $|x_1| = O\left(\frac{1}{\sqrt{d}}\right)$
- **FoDS Thm 2.7:** For $c \geq 1$ and $d \geq 3$ at least $1 - \frac{2}{c}\, e^{-\frac{c^2}{2}}$ fraction of the volume of the d-dimensional unit ball has $|x_1| \leq \frac{c}{\sqrt{d-1}}$
- As $d \to \infty$, volume of unit ball goes to 0.
- Consider a small box centered at origin with side length $\frac{2c}{\sqrt{(d-1)}}$
  - For c = $2\sqrt{(\ln d)}$ this box contains over half of the volume of the ball.
  - Because, for $x_1 \geq \frac{c}{\sqrt{d-1}}$, less than $\frac{1}{d^2\sqrt{(\ln d)}}$ fraction of the volume is outside.
  - Considering this in each dimension, $\frac{1}{d}$ fraction of the volume is outside this box
  - As $d \to \infty$, the volume of this box = $O\left(\left(\frac{\ln d}{d-1}\right)^{\frac{d}{2}}\right)$ goes to 0



Studying volumes in high dimensions

- Almost all volume near the surface:
  - Take arbitrary body $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{d}}$
  - Shrink to $(1-\epsilon)\boldsymbol{A} = \{(1-\epsilon)x | x \in \boldsymbol{A}\}$
  - Volume change ratio: $\frac{volume((1-\epsilon)\boldsymbol{A})}{volume(\boldsymbol{A})} = (1-\epsilon)^{\boldsymbol{d}} \leq e^{-\epsilon \boldsymbol{d}}$



$S_d(r, \epsilon)$
Volume of hypershell

- Let $B_{\boldsymbol{d}}$ (1) = unit $\boldsymbol{d}$-dimensional ball
  - At least $1 - e^{-\epsilon \boldsymbol{d}}$ fraction of its volume is in the annulus of width $\boldsymbol{\epsilon}$
  - $\boldsymbol{\epsilon} = O\left(\frac{1}{\boldsymbol{d}}\right)$: most of the volume in the annulus

$$\lim_{d\to\infty} \frac{\mathrm{vol}(S_d(r,\epsilon))}{\mathrm{vol}(S_d(r))} = \lim_{d\to\infty} 1 - \left(1-\frac{\epsilon}{r}\right)^d \to 1$$

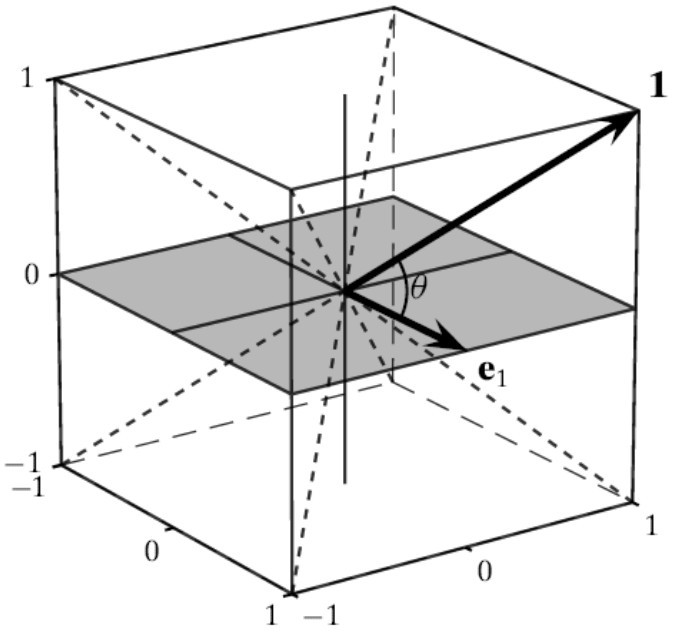
We used the fact that: $1 - x \leq e^{-x}$
If the ball is of radius r, then the annulus width is $O\left(\frac{r}{d}\right)$

It can also be shown that in the limit as d goes to infinity, the volume of the ball goes to 0.

Pair-wise distances in high-dimensions

- Generate n points at random in d-dimensions where each coordinate is a zero mean, unit variance Gaussian.
  - For sufficiently large d; with high probability the distances between all pairs of points will be essentially the same.
- Let x and y be two such random points with the square distance between them:
  - $|x-y|^2 = \sum_i (x_i - y_i)^2$: this can be viewed as sum of d independent samples of $z_i = (x_i - y_i)^2$
  - For large d, Law of Large Numbers says: w.h.p., the average value will be close to the expectation
  - In other words, the sum is close to expectation of the sum which is :
    - $\mathrm{d}(Var(x_i) + Var(y_i) - 2E[x_i y_i]) = \ 2d$
- Not only this, x and y are approximately orthogonal to each other
  - Apply Pythagoras theorem
  - So if, we generate a vector x and call this the north pole, the much of the surface area of the unit ball must lie near the equator.
- **Bonus Q : Does this property hold for other distributions?**

Diagonals in Hyperspace



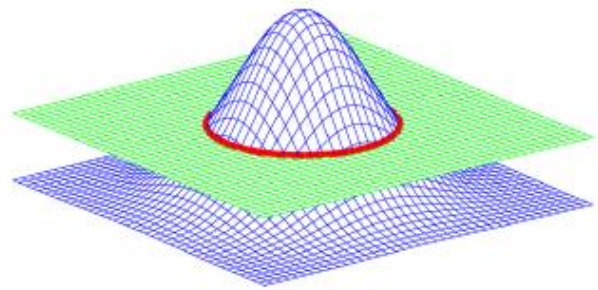
- Assume we have a d-dimensional hypercube $H_d(2)$
  - Range of each dimension in [-1,1]
  - Corners of the hyperspace are given by $(\pm 1_1, \pm 1_2, \ldots \pm 1_d)^T$
- Let $\mathbf{1} = (1_1, 1_2, \ldots 1_d)^T$ be the d-dimensional ones vector
- Let $e_i = (0_1, 0_2, \ldots, 1_i, 0_{i+1}, \ldots, 0_d)^T$ be the canonical unit vector in dimension i
  - Angle between ones vector and standard basis vector = $\cos^{-1}\left(\frac{1}{\sqrt{d}}\right)$
  - As $d \rightarrow \infty$, the angle goes to 90 degrees
- 2^d corners in a d-dimensional hyperspace and 2^d diagonal vectors from origin to each of the corners
  - Same result holds for any diagonal vector and the principal axis vectors e_i
  - In high dimensions all of the diagonal vectors are perpendicular (or orthogonal) to all the coordinates axes!
  - We have 2^{d-1} new axes orthogonal to the d principal coordinate axes
- **Consequence:** If there is a group of points, say a cluster of interest, near a diagonal, these points will get projected into the origin and will not be visible in lower dimensional projections.

Multi-variate Gaussians

- The *multivariate Gaussian distribution* is fully characterized by a mean vector μ and a covariance matrix Σ (DxD)

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right)$$

- The special case of zero mean and Identity Covariance is called Standard Normal.
- Product of Gaussian densities is a Gaussian scaled with a parameter
- If X and Y are two independent Gaussian variables, then X+Y is also Gaussian
- Marginals and Conditionals of Gaussians are Gaussians
- For N(**0**,**I**), the marginal probability density is $p(r) \propto r^{d-1} e^{-\frac{r^2}{2}}$



- To get the maximum mass, set derivative to zero.

Quiz: https://forms.gle/njB2NYCbhGmgNUtd7

Gaussian Annulus Theorem

- Gaussian in $\boldsymbol{d}$ dimensions $(N_{\boldsymbol{d}}(0^{\boldsymbol{d}}, 1))$:

$$\Pr[\boldsymbol{x} = (z_1, \ldots, z_{\boldsymbol{d}})] = (2\pi)^{-\frac{\boldsymbol{d}}{2}} e^{-\frac{z_1^2 + z_2^2 + \cdots + z_{\boldsymbol{d}}^2}{2}}$$

Nearly all mass is in annulus of radius $\sqrt{\boldsymbol{d}}$ and width $O(1)$:

- **Thm. (FoDS Ch2)** For any $\boldsymbol{\beta} \leq \sqrt{\boldsymbol{d}}$, $\sqrt{\boldsymbol{d}} - \boldsymbol{\beta} \leq \left|\left|\boldsymbol{x}\right|\right|_2 \leq \sqrt{\boldsymbol{d}} + \boldsymbol{\beta}$ for constant $\boldsymbol{c}$, **with probability** $1 - 3e^{-\boldsymbol{c}\boldsymbol{\beta}^2}$
  - In higher dimensions the probability density around the mean decreases very rapidly as one moves away from the mean. In essence the entire probability mass migrates to the tail regions.

Useful Results (Concentration Inequalities)

Markov's inequality

- If $\boldsymbol{X}$ is a non-negative r.v. then for every $c > 0$: $\Pr[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]] \leq \frac{1}{c}$
- **Corollary** $(c' = c\ \mathbb{E}[\boldsymbol{X}])$ **:** For every $c' > 0$: $\Pr[\boldsymbol{X} \geq c'\ ] \leq \frac{\mathbb{E}[\boldsymbol{X}]}{c'}$
- **Pro**: Always works! You just need to know the Expected value of the random variable.
- **Cons**:
  - Not very precise
  - Doesn't work for the lower tail: $\Pr[\boldsymbol{X} \leq c\ \mathbb{E}[\boldsymbol{X}]]$
- **Example:** Roll the dice 100 times. What is the probability that the sum is greater then 500?
  - $\Pr[Value \geq 500] \leq \frac{\mathbb{E}[Value]}{3} = \frac{350}{500} \approx 0.7$ {Actual probability is much smaller}

Chebyshev's Inequality

- For every $c > 0$: $\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$
- Proof: $\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] = \Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2 \geq c^2 Var[\boldsymbol{X}]\right]$ (by squaring)

  $= \Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2 \geq c^2 \mathbb{E}[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2]]$ (def. of Var)

  $\leq \frac{1}{c^2}$ (by Markov's inequality)
- **Corollary** ($c' = c\sqrt{Var[\boldsymbol{X}]}$): For every $c' > 0$: $\Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c'] \leq \frac{Var[\boldsymbol{X}]}{c'^2}$

- **Example:** Roll the dice 100 times. What is the probability that the sum is greater then 500?
  - *Step 1:* Calculate variance. Variance of one die roll $= 2.91$. For 100 rolls, it is $= 291$
  - $\Pr[Value_n \geq 500 \text{ or } Value_n \leq 200] \leq \frac{n2.91}{n^2 \cdot 1.5^2} \approx \frac{1.3}{n}$

Chernoff bound

- Let $X_1 \dots X_{\boldsymbol{t}}$ be independent and identically distributed r.vs with range [0,1] and expectation μ.
  - Then if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$, $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu \boldsymbol{t} \delta^2}{3}\right)$
- Let $X_1 \dots X_t$ be independent and identically distributed r.vs with range [0, **c**] and expectation μ.
  - Then if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$, $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3\boldsymbol{c}}\right)$
- Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq z] = O\left(\frac{1}{t}\right)$
- Chernoff: $\Pr[|\boldsymbol{X} - \mu| \geq z] = e^{-\Omega(t)}$

So is Chernoff always better for us?

- Yes, if we have i.i.d. variables.
- No, if we have dependent or only pairwise independent random variables.
- If the variables are not identical – Chernoff-type bounds exist.