Experimental Projects on Web Algorithms

Yury Lifshits
http://yury.name

CalTech, Fall'07
Invited lecture at CS141a

Invitation to CS101.2

New Caltech course
Algorithmic Problems Around the Web:

- http://yury.name/algoweb.html
- MW 11:00-11:55, Jorgensen 287
- Lectures: algorithms for nearest neighbor search
- Projects: adjusting above algorithms to web technologies
- Datasets: friendship graph, users-ads graph

Course Philosophy

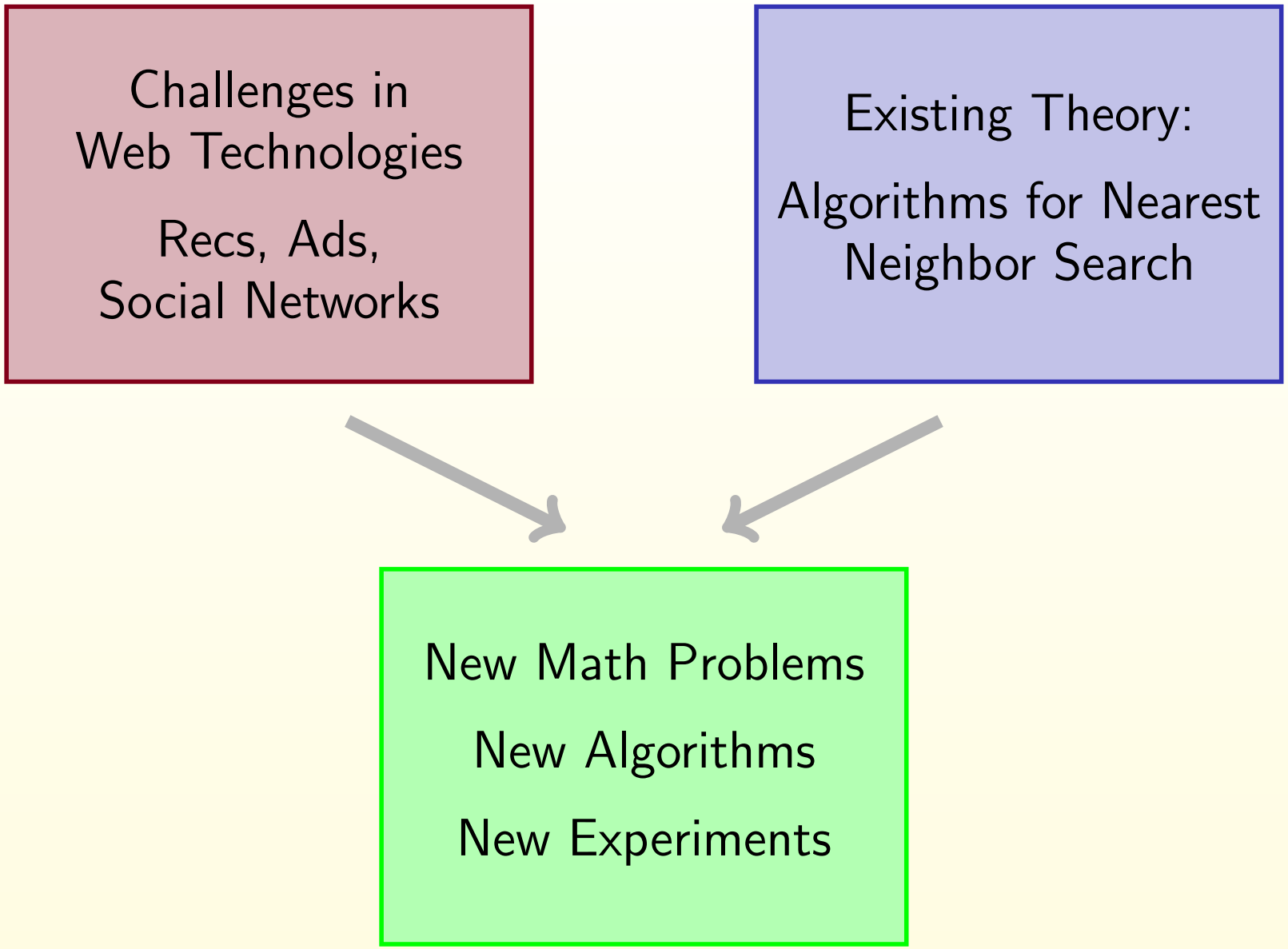
Challenges in Web Technologies

Recs, Ads, Social Networks

Existing Theory:

Algorithms for Nearest Neighbor Search

Course Philosophy



Outline

1. Challenges in Web Technologies

Outline

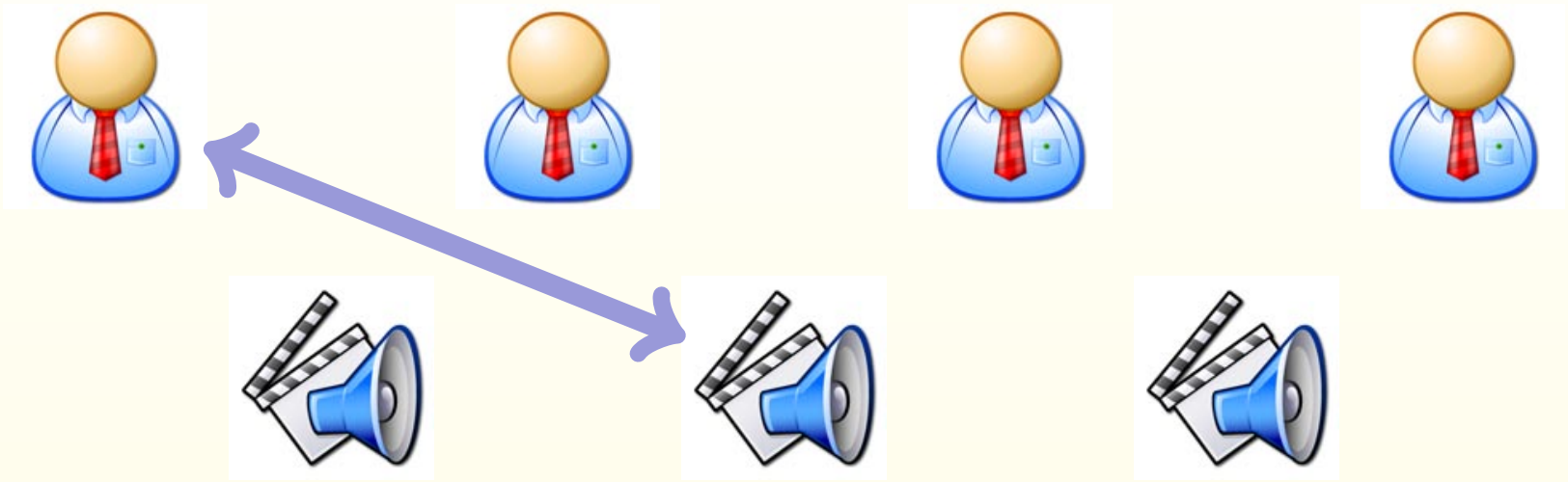
1. Challenges in Web Technologies
2. Existing Theory: Nearest Neighbors

Outline

1. Challenges in Web Technologies
2. Existing Theory: Nearest Neighbors
3. Topics for Experimental Projects

Part I
Challenges in Web Technologies

Recommendation Systems



Approaches:

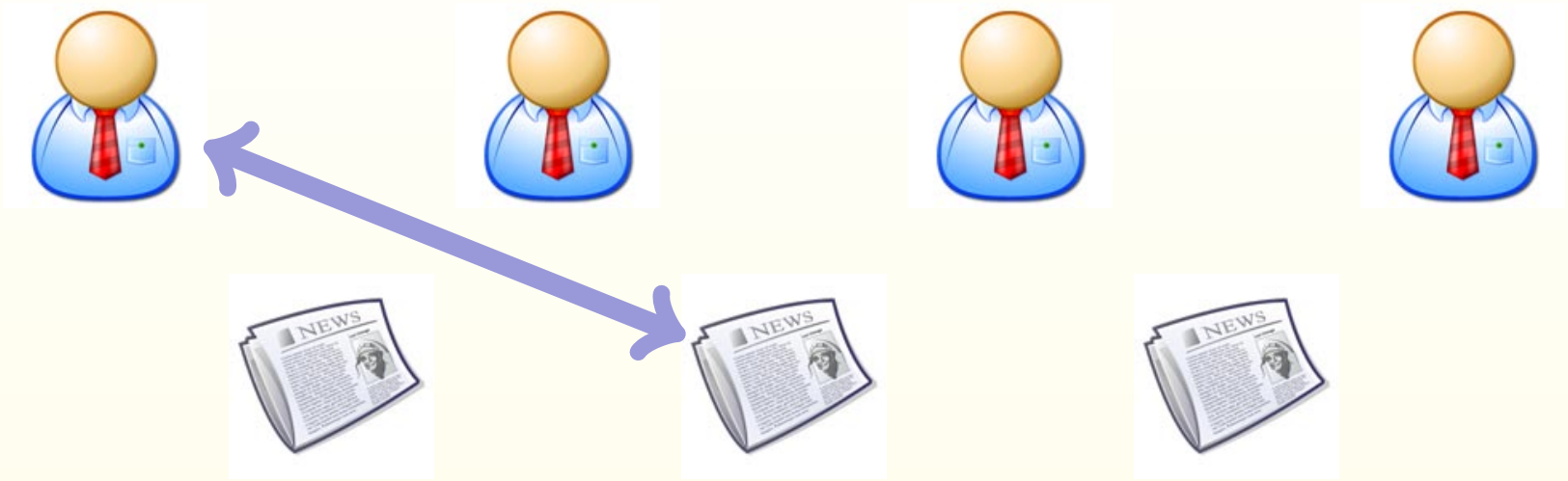
- Content-based
- Collaborative filtering

Behavioral Targeting

Ad targeting:

- Ancient: broadcasting
- Current: contextual
- Future: behavioral

Personalized News Aggregation



Factors to take into account:

- Friendship
- Content
- Feedback (previous ratings)
- Popularity (votes, comments, hyperlinks)

Social Networks Analysis

Social network:

- Nodes
- Edges

Examples of relations: financial exchange, friends, dislike, conflict, trade, web links, sexual relations, disease transmission, airline routes, etc.

Social Networks Analysis

Social network:
- Nodes
- Edges

Examples of relations: financial exchange, friends, dislike, conflict, trade, web links, sexual relations, disease transmission, airline routes, etc.

Our focus
- Community discovery
- Burst detection

Part II Theory of Nearest Neighbors

Nearest Neighbors Informally

To preprocess a database of n objects
so that given a query object,
one can effectively determine
its nearest neighbors in database

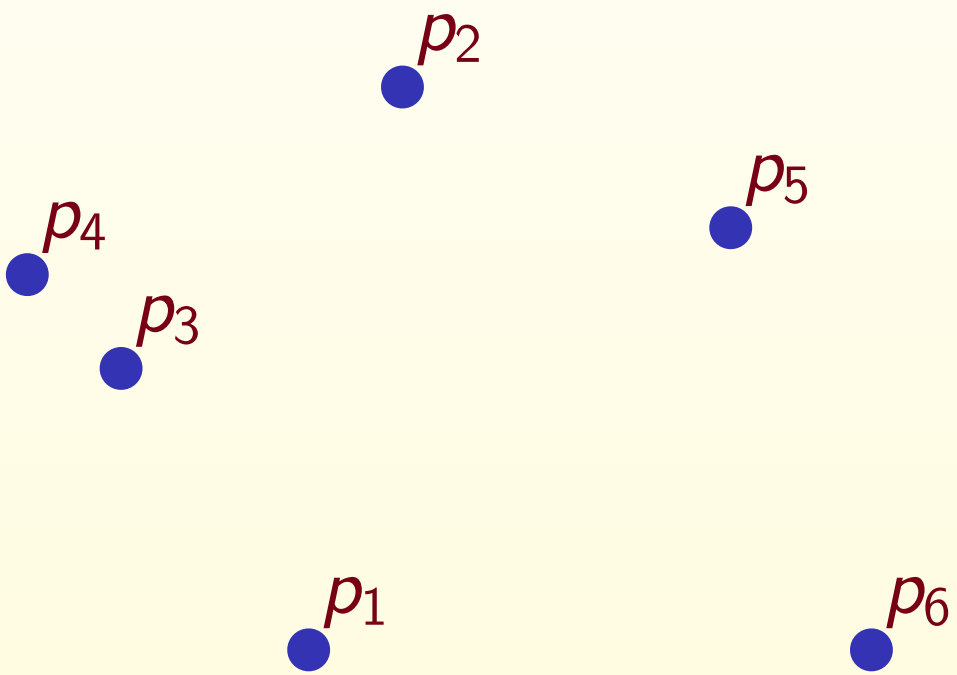
More Formally

Search space: object domain $\mathbb{U}$, similarity function σ

Input: database $S = \{p_1, \ldots, p_n\} \subseteq \mathbb{U}$

Query: $q \in \mathbb{U}$

Task: find $\text{argmax}_{p_i}\ \sigma(p_i, q)$



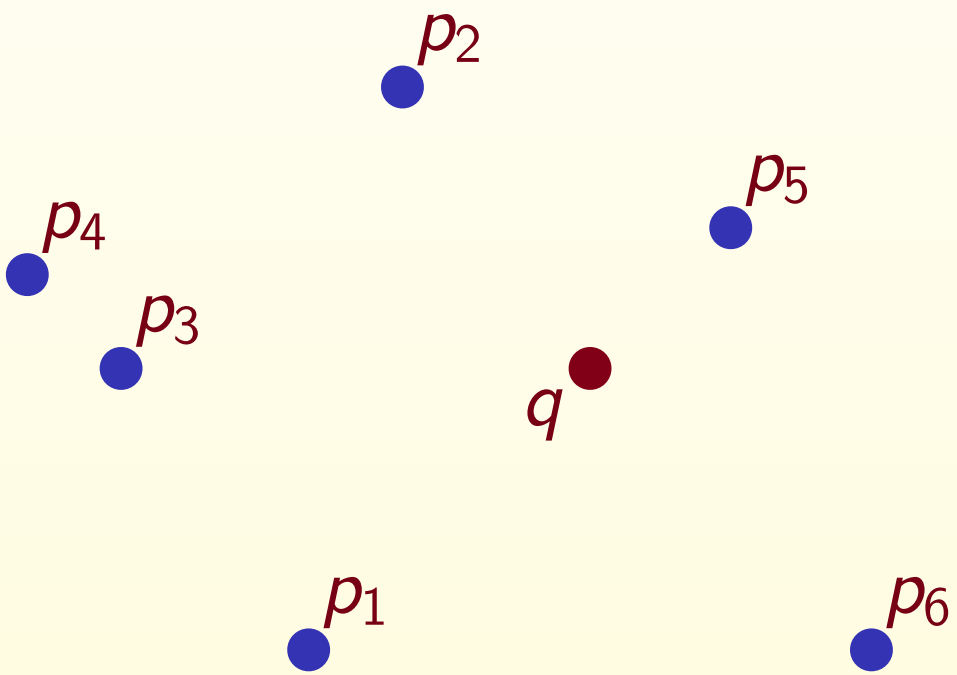
More Formally

Search space: object domain $\mathbb{U}$, similarity function σ

Input: database $S = \{p_1, \ldots, p_n\} \subseteq \mathbb{U}$

Query: $q \in \mathbb{U}$

Task: find $\text{argmax}_{p_i}\ \sigma(p_i, q)$



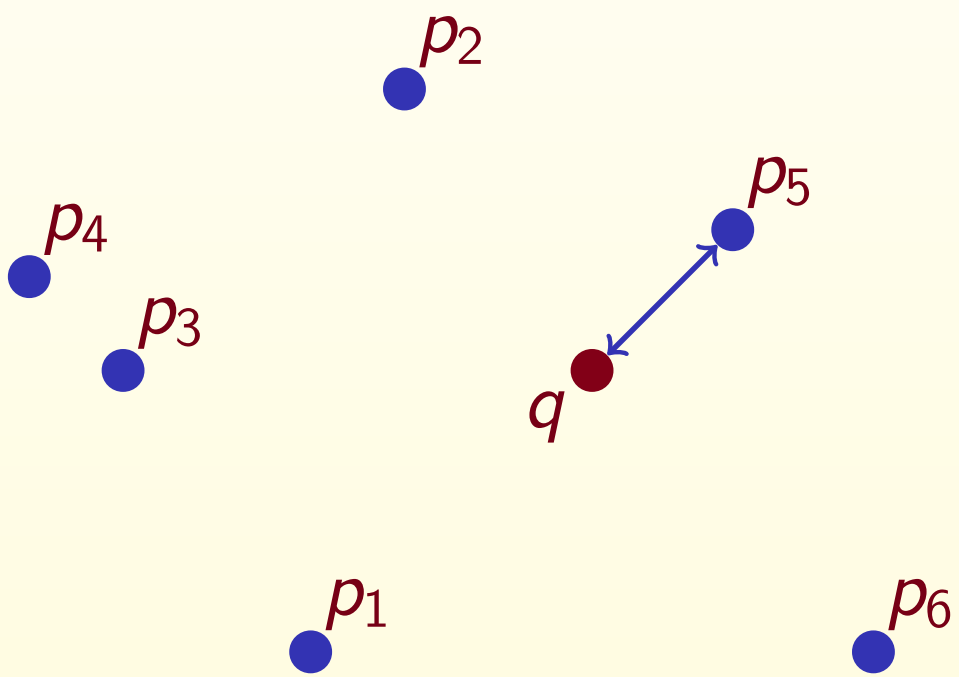
More Formally

Search space: object domain $\mathbb{U}$, similarity function σ

Input: database $S = \{p_1, \ldots, p_n\} \subseteq \mathbb{U}$

Query: $q \in \mathbb{U}$

Task: find $\text{argmax}_{p_i}\ \sigma(p_i, q)$



Some Solutions for NN Problem

Sphere Rectangle Tree
Orchard's Algorithm
LAESA
k-d-B tree
Geometric near-neighbor access tree
Excluded middle vantage point forest
mvp-tree
Fixed-height
fixed-queries tree
AESA
Vantage-point
tree
R*-tree
Burkhard-Keller tree
BBD tree
Navigating Nets
Voronoi tree
Balanced aspect ratio tree
Metric tree
vp^s-tree
M-tree
Locality-Sensitive Hashing
SS-tree
R-tree
Spatial approximation tree
Multi-vantage
point tree
Bisector tree
mb-tree
Generalized hyperplane tree
Hybrid tree
Slim tree
Spill Tree
Fixed queries tree
X-tree
k-d
tree
Balltree
Quadtree
Octree
Post-office tree

Part III
Topics for Experimental Projects

E1 Recommendations for Blog Posts

Available information:

- Friendship graph
- Comments, hyperlinks
- Keywords of interests, post content

Task: For every user recommend 10 posts from last day that seems to be the most interesting for him/her

E2 CTR Prediction

Available information:

Click-or-not bipartite graph

Task: Predict click-through rate for given pair “user-ad”

E3 Social Networks Visualization

Input:

Friendship graph

Similarity:

Number of joint friends

Length of shortest path

E3 Social Networks Visualization

Input:

Friendship graph

Similarity:

Number of joint friends

Length of shortest path

Task:

Construct embedding into 2D
that put similar people close to each other

E4 Disorder Analysis

Disorder inequality for some constant D:

$$\forall p, r, s \in \{q\} \cup S : \quad \text{rank}_r(s) \leq D \cdot (\text{rank}_p(r) + \text{rank}_p(s))$$

E4 Disorder Analysis

Disorder inequality for some constant D:

$$\forall p, r, s \in \{q\} \cup S : \quad \text{rank}_r(s) \leq D \cdot (\text{rank}_p(r) + \text{rank}_p(s))$$

Tasks:

- Compute disorder values for various datasets
- Implement disorder-based algorithms for NNS
- Study their performance

Last Slide

Challenges in Web Technologies

Recs, Ads, Social Networks

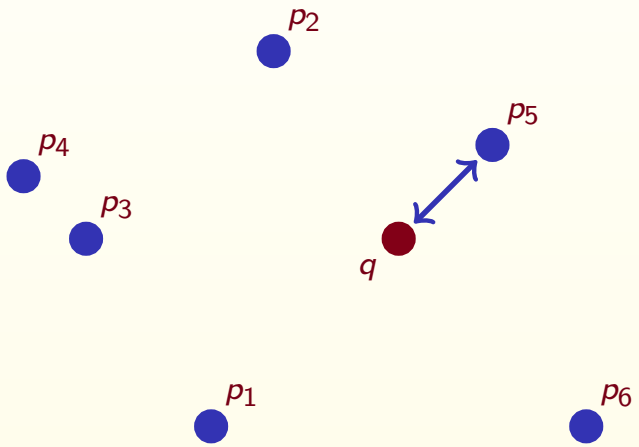
Existing Theory:

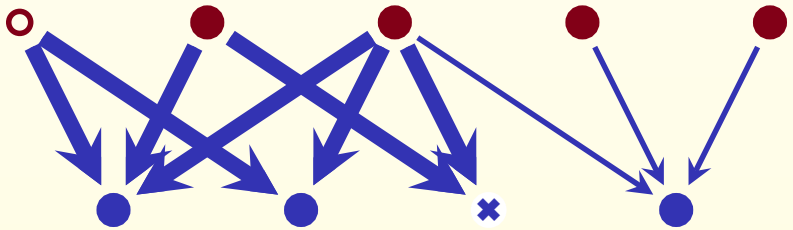
Algorithms for Nearest Neighbor Search

New Math Problems

New Algorithms

New Experiments





Last Slide

Challenges in Web Technologies

Recs, Ads, Social Networks

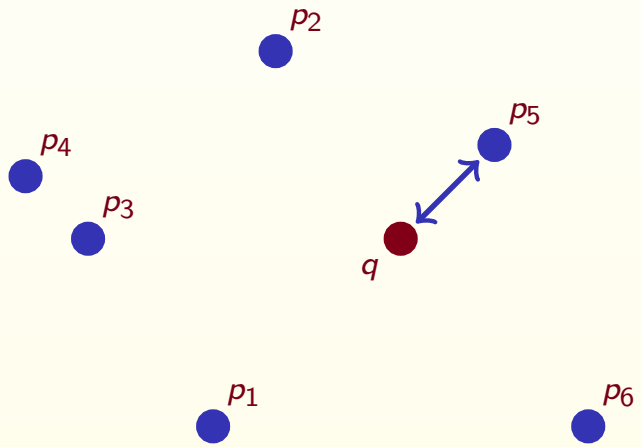
Existing Theory:

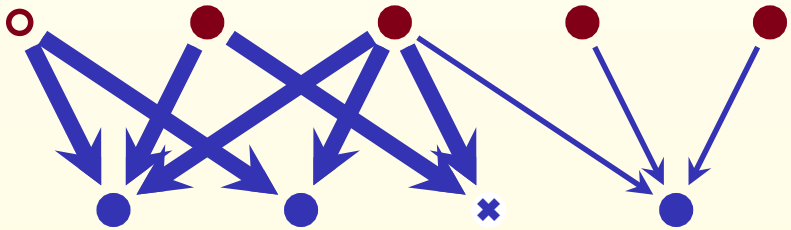
Algorithms for Nearest Neighbor Search

New Math Problems

New Algorithms

New Experiments





Thanks for your attention! Questions?