



DEEP  
LEARNING  
INSTITUTE



DLI Accelerated Data Science Teaching Kit

# Lecture 5.2 - Data De-duplication

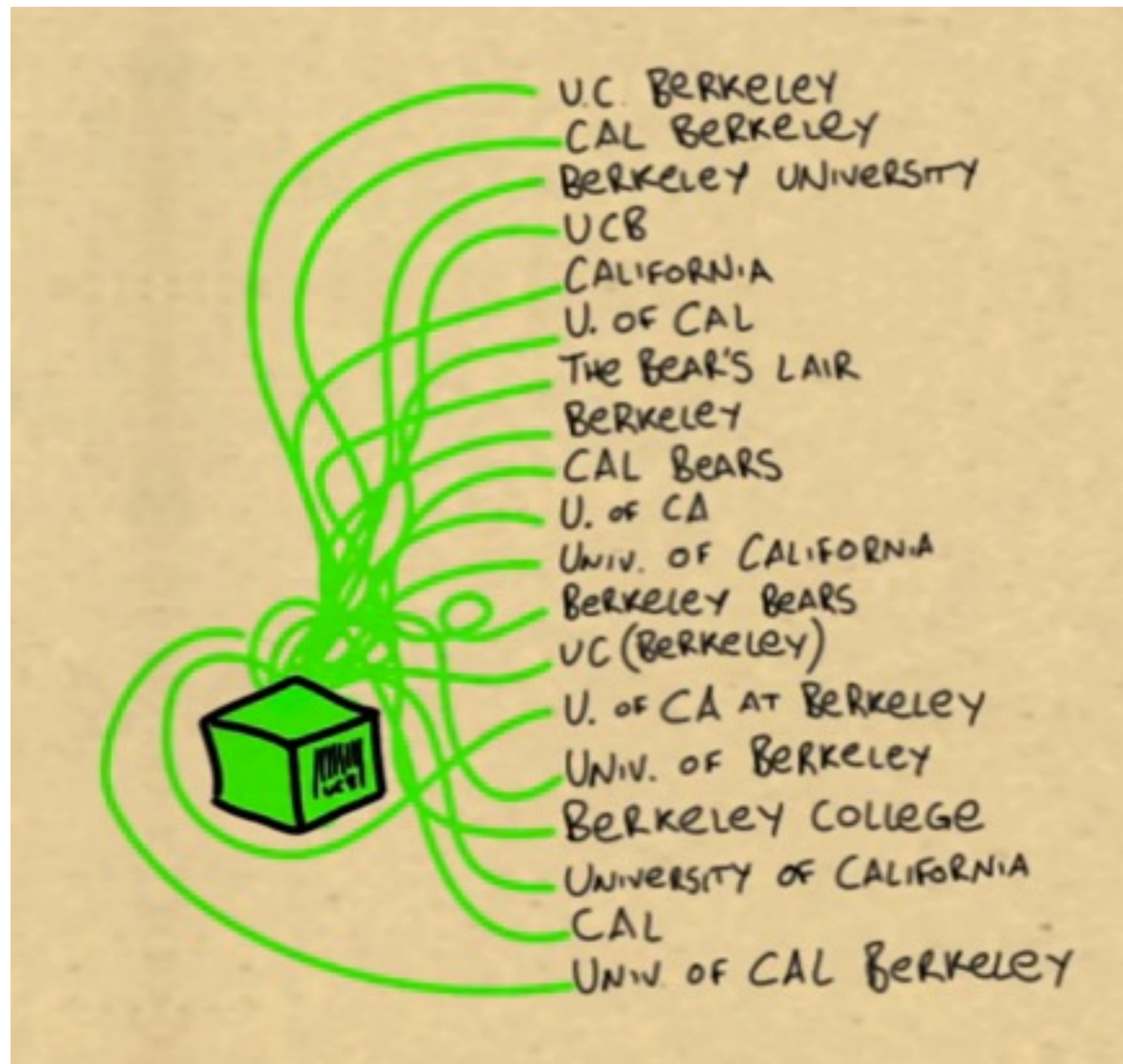


The Accelerated Data Science Teaching Kit is licensed by NVIDIA, Georgia Institute of Technology, and Prairie View A&M University under the [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

# What if we don't have the luxury of having IDs ?

A common problem in academia:

Polo Chau  
Duen Horng Chau  
D.H. Chau  
D. Chau



Then you need to do...

# Entity Resolution

(A hard problem in data integration)

# Why is **entity resolution** so difficult?

Let's understand it through  
**shopping for an iPhone on**  
**Apple, Amazon and eBay**

[Mac](#)[iPad](#)[iPhone](#)[Watch](#)[TV](#)[Music](#)[Support](#)

## iPhone 12

From \$33.29/mo. or \$799 before trade-in\*

### Special Trade-in Offers at Apple

[See all offers](#)



Pay from \$0/mo. after  
AT&T bill credits.<sup>Δ</sup>

T-Mobile



Get an additional \$100 trade-in  
credit from T-Mobile/Sprint.<sup>°</sup>

verizon

Pay from \$11.95/mo. after  
Verizon bill credits.<sup>ΔΔ</sup>



Fast, free, no-contact  
delivery



Free Personal Session



Free and easy returns



Have questions about buying an iPhone?

[Chat with an iPhone Specialist](#)

New

## Buy iPhone 12

Get up to \$220 off with Apple Trade In\*

[See how trade-in works](#)

### Choose your model.

[Which model is right for you?](#)

#### iPhone 12 mini

5.4-inch display<sup>1</sup>

From \$29.12/mo.  
or \$699 before trade-in\*

#### iPhone 12

6.1-inch display<sup>1</sup>

From \$33.29/mo.  
or \$799 before trade-in\*

### Choose your finish.



White



Black



Blue



Green



- Eligible for Free Shipping**  
☐ Free Shipping by Amazon  
All customers get FREE Shipping on orders over \$25 shipped by Amazon
- Delivery Day**  
☐ Get It Today  
☐ Get It by Tomorrow
- Department**  
Cell Phones  
    Unlocked Cell Phones  
    Carrier Cell Phones  
Cell Phone Cases & Covers  
    Cell Phone Basic Cases  
Electronics  
Cell Phone Accessories  
[See All 5 Departments](#)
- Avg. Customer Review**  

★★★★★

& Up

★★★★☆


& Up

★★★☆☆


& Up

★★☆☆☆


& Up
- Brand**  
☐ Apple  
☐ Samsung Electronics  
☐ OnePlus  
☐ SAMSUNG  
☐ GooPhone
- Price**  
Up to \$50  
\$50 to \$100  
\$100 to \$150




iPhone 12. Blast past fast.  
[Shop Apple >](#)




New Apple iPhone 12 (256GB, Green) [Locked] + Carrier Subscription  
★★★★★ 53  
prime



New Apple iPhone 12 (128GB, Blue) [Locked] + Carrier Subscription  
★★★★★ 53  
prime




New Apple iPhone 12 (128GB, Black) [Locked] + Carrier Subscription  
★★★★★ 53  
prime



Sponsored  
New Apple iPhone 12 (128GB, Blue) [Locked] + Carrier Subscription  
★★★★★ 53  
\$879<sup>00</sup>  
prime Get it as soon as **Tomorrow, Feb 25**  
FREE Shipping by Amazon

Display Size	Memory	Color	Brand
6.1 inches	128 GB	Blue	Apple



Sponsored  
New Apple iPhone 12 Pro (256GB, Pacific Blue) [Locked] + Carrier Subscription  
★★★★★ 77  
\$1,099<sup>00</sup>  
Display Size  
6.1 inches  
Memory  
256 GB  
Color  
Pacific Blue  
Brand  
Apple





Shop by category

iphone 12

Cell Phones & Sm...

Search

Advanced

Related: [iphone 11](#) [iphone 11 pro max](#) [iphone 12 pro](#) [iphone 11 pro](#) [iphone 12 case](#) [iphone 12 pro max](#) [iphone x](#) [iphone xr](#) [iphone 12 mini](#) [iphone 12 unlocked](#) ... ☐ Include description

## Category

All

< Cell Phones & Accessories

Cell Phone Accessories

Cell Phones & Smartphones

Phone Cards & SIM Cards

Cell Phone & Smartphone Parts

More

Computers/Tablets & Networking

Consumer Electronics

Clothing, Shoes & Accessories

eBay Motors

Cameras & Photo

Show More

## Network

- ☐ Unlocked (2,563)
- ☐ AT&T (1,596)
- ☐ T-Mobile (1,528)
- ☐ Verizon (1,437)
- ☐ Sprint (1,043)
- ☐ TracFone (1,027)
- ☐ SIMPLE Mobile (1,024)
- ☐ Google Fi (976)

See all

## Storage Capacity

- ☐ 128 GB (1,095)
- ☐ 64 GB (1,078)
- ☐ 256 GB (566)
- ☐ 32 GB (809)
- ☐ 512 GB (193)
- ☐ 16 GB (431)
- ☐ 8 GB (20)
- ☐ 9 KB (20)

See all

All Listings

Accepts Offers

Auction

Buy It Now

Condition

Delivery Options

Sort

Best Match

2,371 results for **iphone 12**

Save this search

Shipping to: 30318

## Price

Under \$700.00

\$700.00 to \$900.00

Over \$900.00

## Storage Capacity

128 GB

64 GB

256 GB

512 GB



GREAT PRICE

Apple iPhone 12 Mini - 128gb - Unlocked - Factory Sealed -...

\$814.00

Trending at \$829.00

Buy It Now

Free 3 day shipping

Free returns

69+ sold



Sponsored

Apple iPhone 7 32GB, 128GB, 256GB CDMA/GSM Unlocked...

★★★★★ (95)

\$129.99 to \$194.99

Buy It Now

Free 4 day shipping

Free returns

Extra 10% off



Sponsored

Apple iPhone 6 Plus Smartphone GSM Unlocked 16GB 64GB 128G...

★★★★★ (59)

\$114.00 to \$144.00

Buy It Now

Free 4 day shipping

Free returns

567+ sold



Apple iPhone 12 mini - 64GB - Black (Verizon)

\$600.00

or Best Offer

Free 3 day shipping



Sponsored

TracFone

Keep in touch for \$60/year

No contract. Best networks. Dependable phone plan.

Shop now →





# D-Dupe

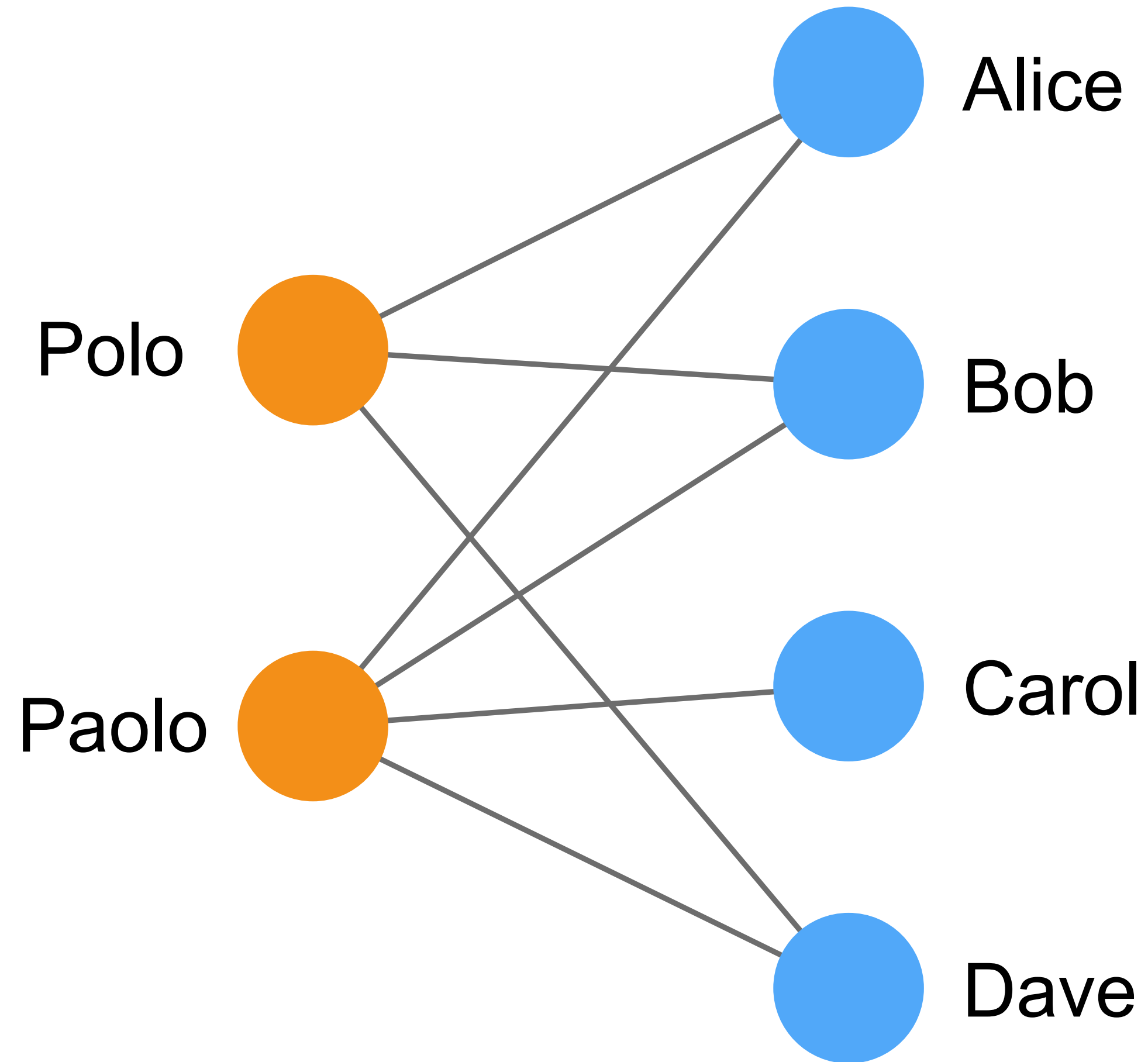
Interactive Data Deduplication and Integration  
TVCG 2008

University of Maryland

Bilgic, Licamele, Getoor, Kang, Shneiderman

Retrieved: <https://linqpub.soe.ucsc.edu/basilic/web/Publications/2006/bilgic:vast06/>







# Core components: **Similarity functions**

Determine how two entities are similar.

D-Dupe's approach:

**Attribute similarity** + **relational similarity**



$$sim(e_i, e_j) = (1 - \alpha) \times sim_A(e_i, e_j) + \alpha \times sim_R(e_i, e_j),$$

$$0 \leq \alpha \leq 1,$$

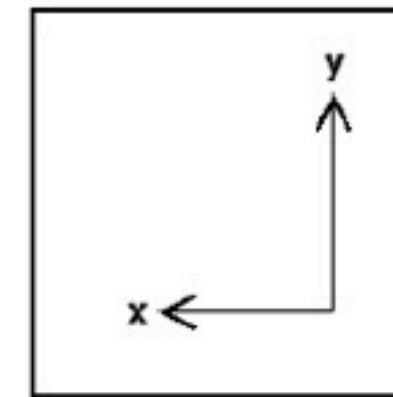


**Similarity score** for a pair of entities

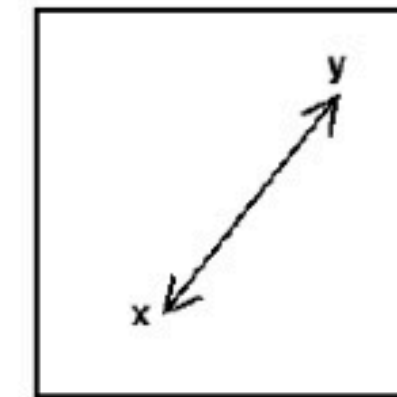
# Numerous **similarity** functions

Excellent read: <http://infolab.stanford.edu/~ullman/mmds/ch3a.pdf>

- Euclidean distance  
Euclidean norm / L2 norm
- TaxiCab/Manhattan distance



Manhattan



Euclidean

- Jaccard Similarity (e.g., used with w-shingles)  
e.g., overlap of nodes' #neighbors

*Jaccard similarity* of sets  $S$  and  $T$  is  $|S \cap T| / |S \cup T|$

- String edit distance  
e.g., “Polo Chau” vs “Polo Chan”

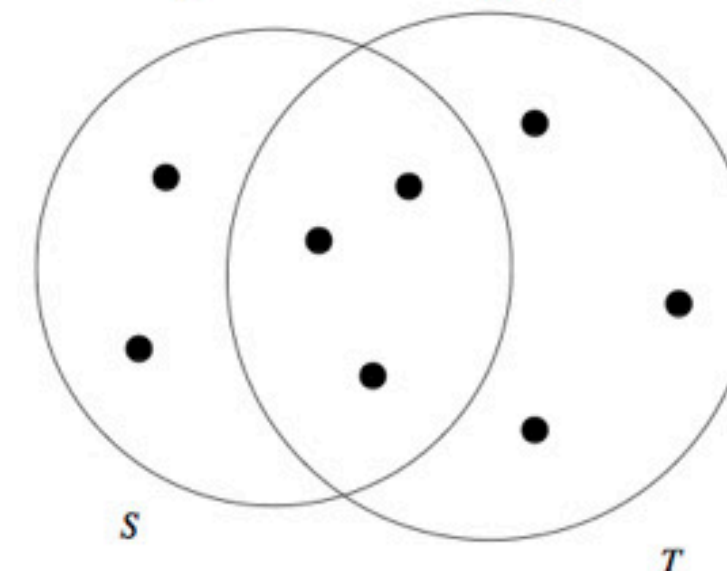


Figure 3.1: Two sets with Jaccard similarity 3/8

27



# Distance and Similarity Measures

Different measures of distance or similarity are convenient for different types of analysis. The Wolfram Language provides built-in functions for many standard distance measures, as well as the capability to give a symbolic definition for an arbitrary measure.

## ▼ Reference

### Numerical Data

**EuclideanDistance** ▪ **SquaredEuclideanDistance** ▪ **NormalizedSquaredEuclideanDistance** ▪  
**ManhattanDistance** ▪ **ChessboardDistance** ▪ **BrayCurtisDistance** ▪ **CanberraDistance** ▪  
**CosineDistance** ▪ **CorrelationDistance** ▪ **BinaryDistance** ▪ **TimeWarpingDistance**

### Boolean Data

**HammingDistance** ▪ **JaccardDissimilarity** ▪ **MatchingDissimilarity** ▪ **DiceDissimilarity** ▪  
**RogersTanimotoDissimilarity** ▪ **RussellRaoDissimilarity** ▪ **SokalSneathDissimilarity** ▪  
**YuleDissimilarity**

### String Data

**EditDistance** ▪ **DamerauLevenshteinDistance** ▪ **HammingDistance** ▪  
**SmithWatermanSimilarity** ▪ **NeedlemanWunschSimilarity**

### Images & Colors

**ImageDistance** ▪ **ColorDistance**

### Geospatial & Temporal Data

**GeoDistance** ▪ **DateDifference**



# Excellent Tutorial on Entity Resolution

[http://www.umiacs.umd.edu/~getoor/Tutorials/ER\\_KDD2013.pdf](http://www.umiacs.umd.edu/~getoor/Tutorials/ER_KDD2013.pdf)

by Lise Getoor and Ashwin Machanavajjhala



DEEP  
LEARNING  
INSTITUTE



DLI Accelerated Data Science Teaching Kit

# Thank You