

Note to other teachers and users of these slides: We would be delighted if you found our material useful for giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. If you make use of a significant portion of these slides in your own lecture, please include this message, or a link to our web site: <http://cs224w.Stanford.edu>

Stanford CS224W: Relational Deep Learning

CS224W: Machine Learning with Graphs

Charilaos Kanatsoulis

Jure Leskovec, Stanford University

<http://cs224w.stanford.edu>



Announcements

- **Homework 2** is due today
- **Homework 3** will be released today
 - Recitation session details TBA on Ed
- **Project proposal grades are out now!**
 - Regrade requests are open until Thursday 11/6
- **No class next Tuesday 11/4 (Democracy Day)**
- **Colab 3** is due next Thursday 11/6
- **Project Milestone** is due next Thursday 11/6
- **Practice exam will be released next week**

Stanford CS224W: Relational Databases

CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



Deep Learning Revolution

Sequences

Statue of Liberty

Article Talk

From Wikipedia, the free encyclopedia

For other uses, see *Statue of Liberty* (disamb).

The Statue of Liberty (*Liberty Enlightening the World*; French: *La Liberté éclairant le monde*) is a colossal neoclassical sculpture on Liberty Island in New York Harbor in New York City, in the United States. The copper statue, a gift from the people of France, was designed by French sculptor *Frédéric Auguste Bartholdi* and its metal framework was built by *Gustave Eiffel*. The statue was dedicated on October



A spectrum of developmental disorders that includes autism, and Asperger syndrome. Signs and symptoms include poor communication skills, defective social interactions, and repetitive behaviors. Each child with autism spectrum disorder is likely to have a unique pattern of behavior. [...] Autism spectrum disorder has no single known cause. [...] Autism spectrum disorder affects children of all races and nationalities, but certain factors increase a child's risk. [...] There's no way to prevent autism spectrum disorder, but there are treatment options.

Risperidone is a second-generation antipsychotic (SGA) medication used in the treatment of a number of mood and mental health conditions including schizophrenia and bipolar disorder. The half-life is 3 hours in extensive metabolizers. Though its precise mechanism of action is not fully understood, current focus is on the ability of risperidone to inhibit the D2 dopaminergic receptors and 5-HT2A serotonergic receptors in the brain. [...] Risperidone and its active metabolite, 9-hydroxyrisperidone, are ~60% and ~77% protein-bound in human plasma, respectively. [...] The primary action of risperidone is to decrease dopaminergic and serotonergic pathway activity in the brain, therefore decreasing symptoms of schizophrenia and mood disorders.

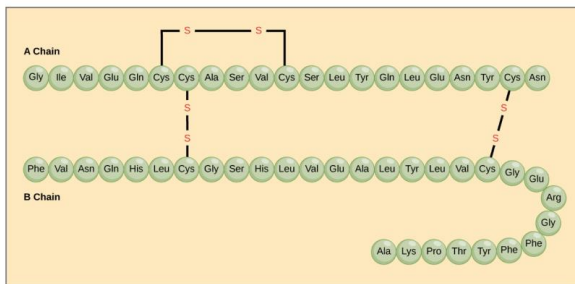


Image credit: OpenStax Biology.

Images

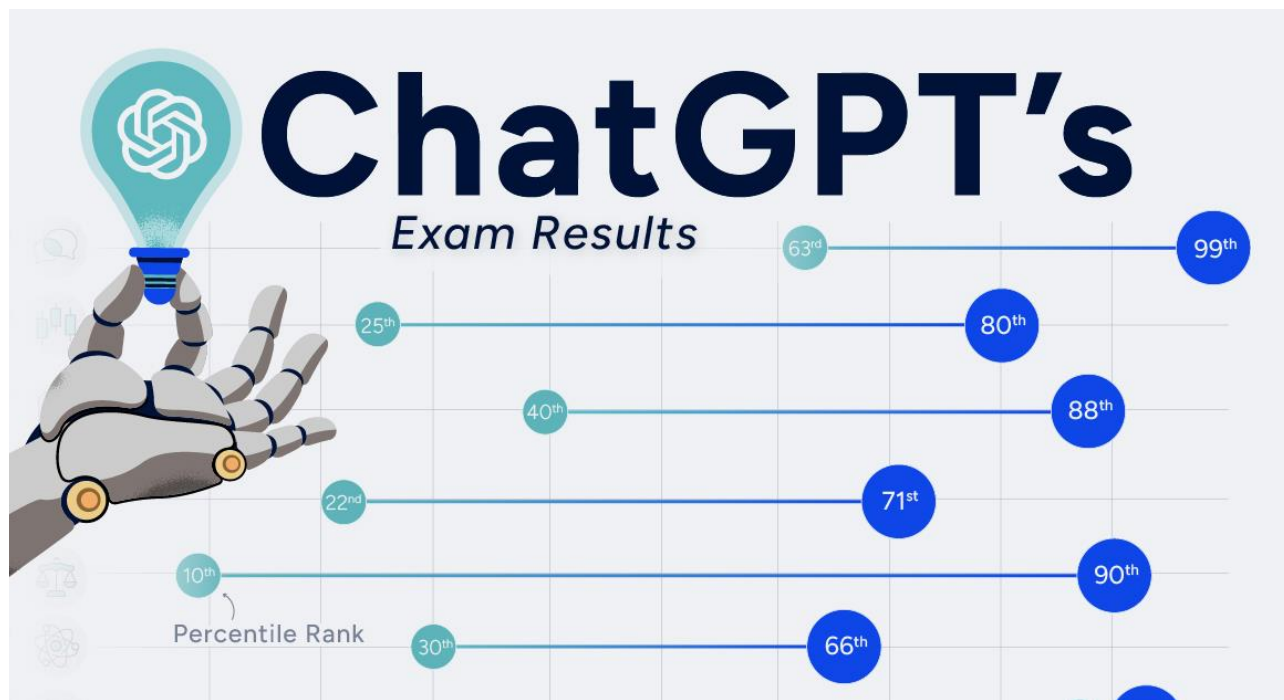
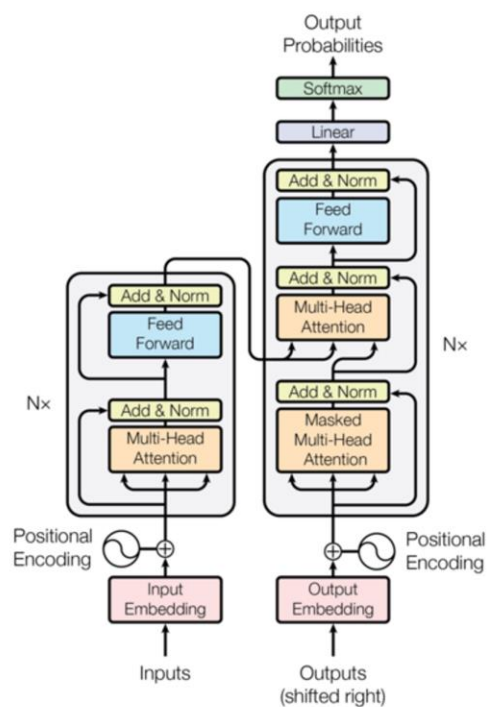


DirtyTesla Starlink Plz @Dirt... · 8h
If you experience any kind of traffic like this, you need Autopilot. It makes the experience relaxing instead of stressful.



Deep Learning Revolution

These breakthroughs are fueled by **Transformers**

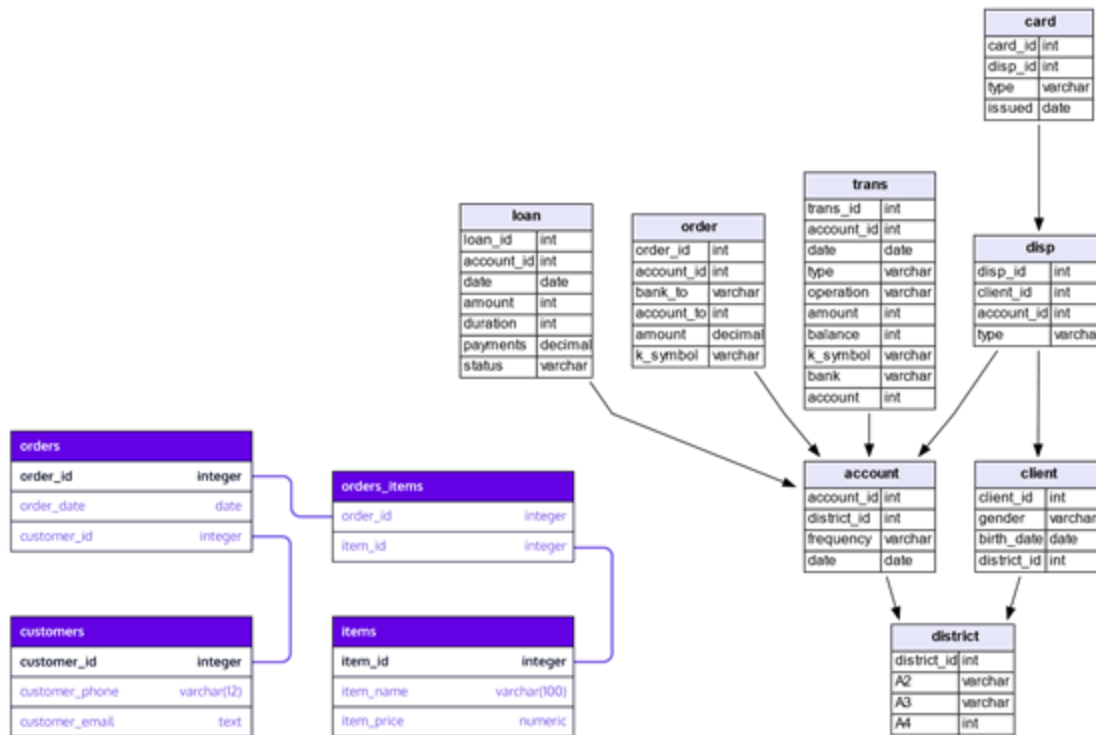


Deep Learning Revolution

These breakthroughs are fueled by **Data**

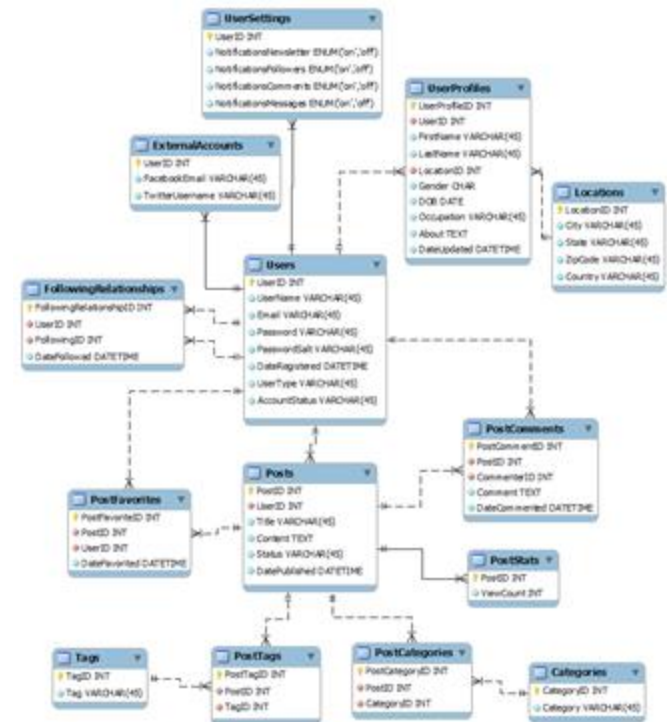


Data stored in Relational Databases



Commerce

Finance



Social Media

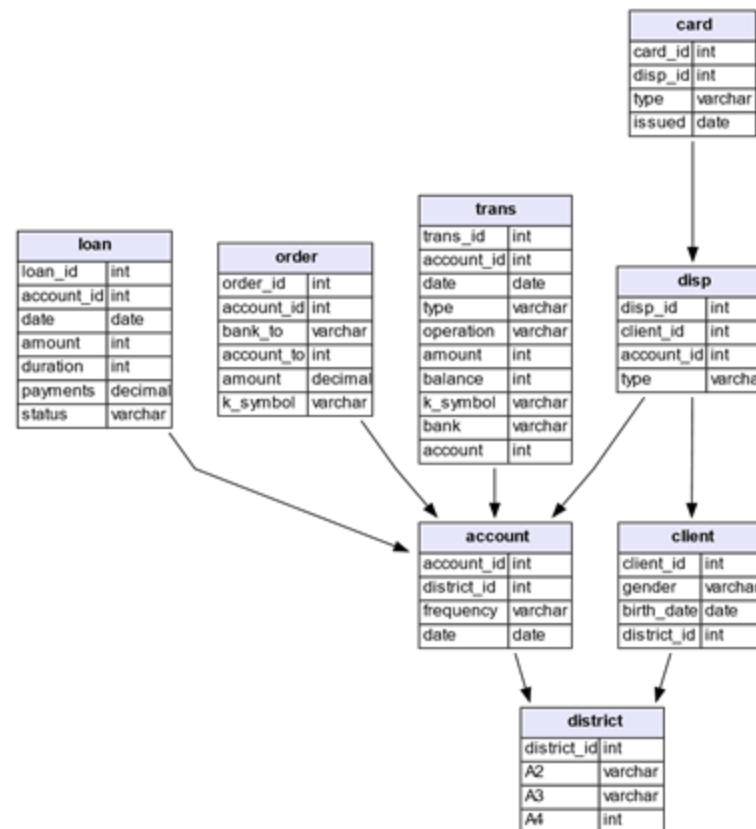
Predictions on Relational Data

- + Which **products** will a user **purchase** in the next 7 days?
- + Will an **active user** churn in the next 90 days?
- + What will be the **total sales** for each **product** in the next 30 days?



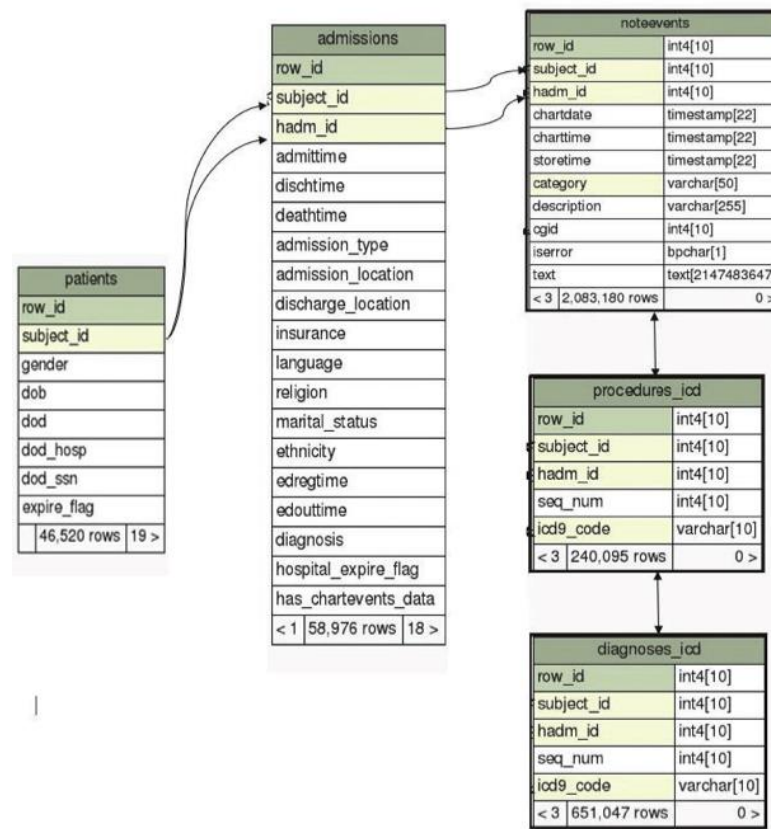
Predictions on Relational Data

- + What will be the **loan outcome** (good vs bad)?
- + What will be the **total number of failed loans** in the next year?

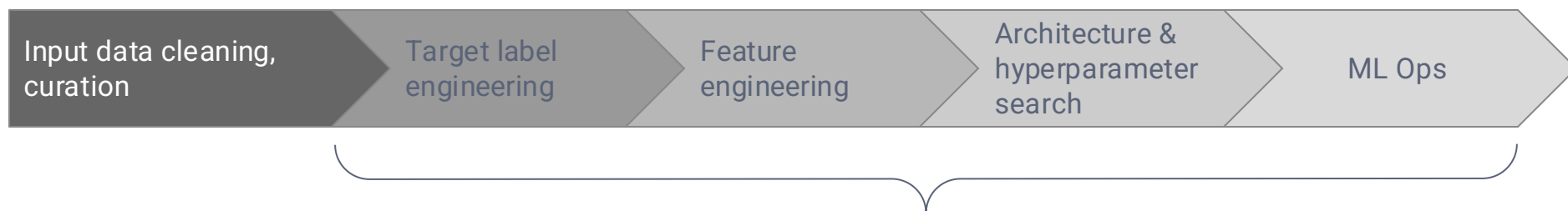


Predictions on Relational Data

- + Will a **patient return** if discharged from the hospital?
- + Which hospital admissions have **greatest risk to life** in the next 24 hours



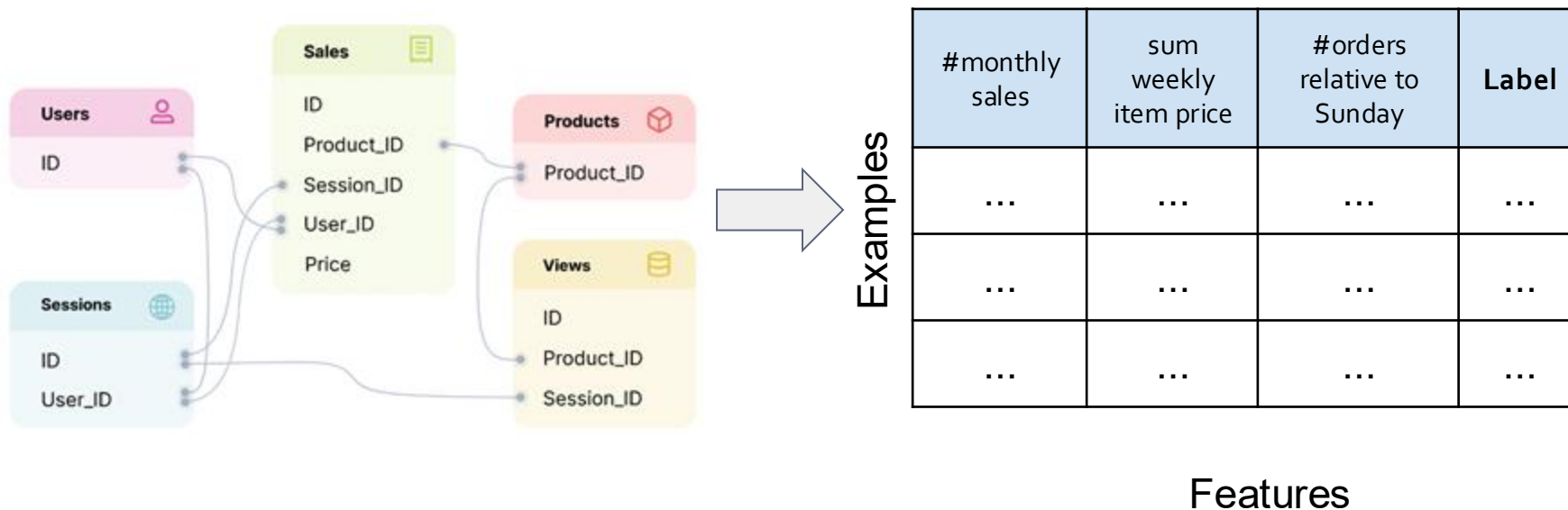
Doing AI is slow & complex



6-12 months of time for a team of data scientists,
data engineers and product engineers

Impedance Mismatch

Goal: Learn a **user churn** model based on their sales, purchased products and browsing behavior



- Features are chosen **arbitrarily** (e.g., aggregations, time windows)
- Only a **limited set of data** is used
- Issues with **point-in-time correctness/information leakage**

Key Question

How to leverage the data without going through the longest duration tasks (extraction, transformation, loading)?

We want ML algorithms that can process data in its natural form!

Why not use an LLM?

Some have tried

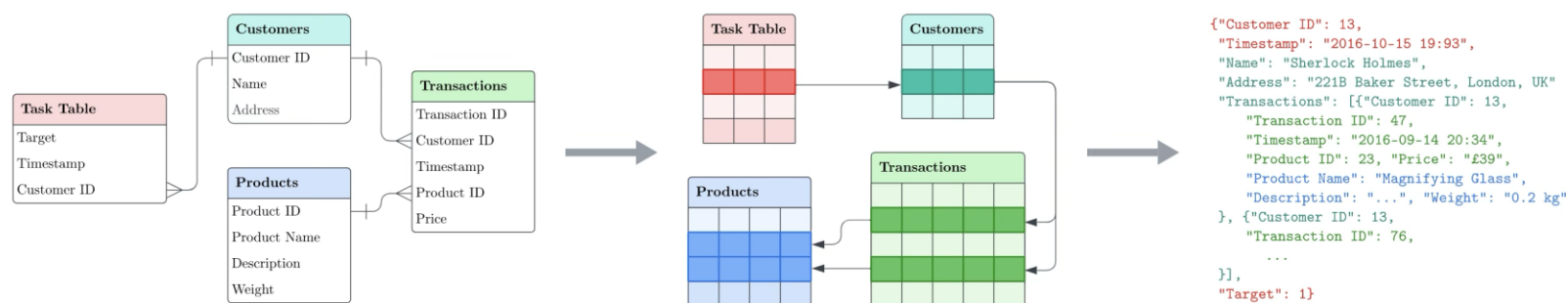


Figure 1: Process of constructing a single example for LLM-based inference.

Tackling prediction tasks in relational databases with LLMs

Marek Wydmuch^{*,†}

Łukasz Borchmann^{*}

Filip Graliński^{*,‡}

^{*} Snowflake AI Research

{first-name}.{last-name}@snowflake.com

[†] Poznan University of Technology / Poznań, Poland

[‡] Adam Mickiewicz University / Poznań, Poland

Prompt

[{"Necklace": "product_group_name": "Accessories", "graphical_appearance_no": 1010009, "graphical_appearance_name": "Glittering/Metallic", "colour_group_code": "5", "colour_group_name": "Gold", department_no": 7020, "department_name": "Conscious Exclusive", "index_code": "A", "index_name": "Ladieswear", "index_group_no": 1, "index_group_name": "Ladieswear", "section_no": "62", made from recycled metal. Chain cut with a sparkle stone at the bottom and a tear-shaped pendant covered in spars"]], [{"t_datetime": "2019-11-25T00:00:00", "customer_id": 198761, "article_id": "6", "age": "32.0", "postal_code": "7B4c17acc9a3033b5ef8e1a9eb76471d829851dbf92d43b0e95cb051", "add_transactions": [{"t_datetime": "2020-05-04T00:00:00", "customer_id": 198761, "article_id": "6", "product_type_name": "Trousers", "product_group_name": "Garment Lower body", graphical_appearance_no": 1010023, "graphical_appearance_name": "Denim", "colour_group_code": 71, "colour_perceived_colour_master_name": "Blue", "department_no": 1722, "department_name": "Trouser", "index_code": "A", "index_name": "Ladieswear", "index_group_no": 1, "index_group_name": "Ladieswear", "5-pocket jeans in washed, superstretch denim with regular waist, zip fly and belt, and skinny legs."}], [{"t_datetime": "2019-11-29T00:00:00", "customer_id": 1348061, "article_id": 68191, "price": 0, "postal_code": "be5edf195ddca193409dc9cc33242c1c86a039819bd20b82bf6f384ab60f817", "add_transactions": [{"t_datetime": "2020-04-11T00:00:00", "customer_id": 1348061, "article_id": 68191, "price": 0, "product_type_name": "Shirt", "product_group_name": "Garment Upper body", graphical_appearance_no": 1010101, "graphical_appearance_name": "Short-sleeved shirt with a collar, buttons down the front, yoke at the back and a gently rounded hem.", "colour_group_code": 9, "colour_group_name": "Blouse", "department_no": 1522, "department_name": "Blouse", "index_code": "A", "index_name": "Ladieswear", "index_group_no": 1, "index_group_name": "Ladieswear", "section_no": 15, "section_name": "Opening at the back with ties at the crotch of the neck and short, elasticated puff sleeves."}], [{"t_datetime": "2019-12-11T00:00:00", "customer_id": 556304, "article_id": 68191, "price": 0.033881355932033, "0.6439f990d5fe6831fdb32c4d82c5d6bc3426c42d7e0850e109398fd3c9", "add_transactions": [{"t_datetime": "2020-03-09T00:00:00", "customer_id": 556304, "article_id": 76457, "price": 0.016932032033, "product_group_name": "Garment Upper body", graphical_appearance_no": 1010016, "graphical_appearance_name": "Solid", "colour_group_code": 9, "colour_group_name": "Black", "perceived_676, "department_name": "Jersey Basic", "index_code": "A", "index_name": "Ladieswear", "index_group_no": 1, "index_group_name": "Ladieswear", "section_no": 16, "section_name": "Women's polo neck and long sleeves."}], [{"t_datetime": "2020-03-09T00:00:00", "customer_id": 556304, "article_id": 76456, "price": 0.016932033898305, "sales_channel_id": 2, "product_code": 785018, "1010016, "graphical_appearance_name": "Solid", "colour_group_code": 33, "colour_group_name": "Dark Orange", "perceived_colour_value_id": 2, "perceived_colour_value_name": "Medium Dusty", "index_code": "A", "index_name": "Ladieswear", "index_group_no": 1, "index_group_name": "Ladieswear", "section_no": 16, "section_name": "Women's Everyday Basics", "garment_group_no": 1002, "garment_2020-03-09T00:00:00", "customer_id": 556304, "article_id": 92998, "price": 0.0254067796610169, "sales_channel_id": 2, "product_code": 856109, "prod_name": "Sweaty", "product_type_n27, "graphical_appearance_name": "Check", "colour_group_code": 12, "colour_group_name": "Light Beige", "perceived_colour_value_id": 1, "perceived_colour_value_name": "Dusty Light", "index_code": "A", "index_name": "Ladieswear", "index_group_no": 1, "index_group_name": "Ladieswear", "section_no": 16, "section_name": "Women's Everyday Basics", "garment_group_no": 1002, "garment_group_name": "Jersey Basic", "detail_desc": "Detail desc: Trousers in 2019-09-25T00:00:00", "customer_id": 179592, "article_id": 68191, "price": 0.033881355932033, "sales_channel_id": 2, "FN": 1.0, "Active": 1.0, "club_member_status": "ACTIVE", "fashion_news2019-12-06T00:00:00", "customer_id": 1067378, "article_id": 68191, "price": 0.033881355932033, "sales_channel_id": 2, "FN": 1.0, "Active": 1.0, "club_member_status": "ACTIVE", "fashion_news2020-01-06T00:00:00", "customer_id": 50910, "article_id": 68191, "price": 0.033881355932033, "sales_channel_id": 2, "FN": null, "Active": null, "club_member_status": "ACTIVE", "fashion_news2020-07-09T00:00:00", "customer_id": 1063111, "article_id": 68191, "price": 0.0237118644067796, "sales_channel_id": 2, "FN": 1.0, "Active": 1.0, "club_member_status": "ACTIVE", "fashion_news["t_datetime": "2019-09-30T00:00:00", "customer_id": 1063111, "article_id": 76458, "price": 0.016932033898305, "sales_channel_id": 2, "product_code": 785018, "prod_name": "Shenzi LP", "product_1063111, "graphical_appearance_name": "Solid", "colour_group_code": 93, "colour_group_name": "Dark Green", "perceived_colour_value_id": 2, "perceived_colour_value_name": "Medium Dusty", "perceiveindex_name": "Ladieswear", "index_group_no": 1, "index_group_name": "Ladieswear", "section_no": 16, "section_name": "Women's Everyday Basics", "garment_group_no": 1002, "garment_group2019-11-29T00:00:00", "customer_id": 418521, "article_id": 68191, "price": 0.0271016949125242, "sales_channel_id": 2, "FN": null, "Active": null, "club_member_status": "ACTIVE", "fashion_news2019-10-06T00:00:00", "customer_id": 935373, "article_id": 68191, "price": 0.033881355932033, "sales_channel_id": 2, "FN": 1.0, "Active": 1.0, "club_member_status": "ACTIVE", "fashion_news["t_datetime": "2020-07-14T00:00:00", "customer_id": 935373, "article_id": 95999, "price": 0.0249999999999999, "sales_channel_id": 2, "product_code": 85624, "prod_name": "salt ss", "product_type_1063111, "graphical_appearance_name": "Solid", "colour_group_code": 10, "colour_group_name": "White", "perceived_colour_value_id": 3, "perceived_colour_value_name": "Light", "perceived_colour_masindex_name": "Ladieswear", "index_group_no": 2, "index_group_name": "Ladieswear", "section_no": 16, "section_name": "Women's Everyday Basics", "garment_group_no": 1002, "garment_group2020-02-08T00:00:00", "customer_id": 935373, "article_id": 88302, "price": 0.0146779661016949, "sales_channel_id": 1, "product_code": 832361, "prod_name": "Mona", "product_tyindex_name": "Solid", "colour_group_code": 12, "colour_group_name": "Light Beige", "perceived_colour_value_id": 1, "perceived_colour_value_name": "Dusty Light", "perceived_colour_master_id": 11, "perceiveindex_group_no": 1, "index_group_name": "Ladieswear", "section_no": 15, "section_name": "Women's Everyday Collection", "garment_group_no": 1005, "garment_group_name": "Jersey Fancy", "1510, "price": 0.008457627118644, "sales_channel_id": 2, "product_code": 849747, "prod_name": "Fashion offer Raymond 2PCS set", "product_type_no": 270, "product_type_name": "Garment Sindex_name": "Light Yellow", "perceived_colour_value_id": 7, "perceived_colour_value_name": "Medium", "perceived_colour_master_id": 8, "perceived_colour_master_name": "Yellow", "diindex_group_name": "Baby/Children", "section_no": 41, "section_name": "Baby Boy", "garment_group_no": 1005, "garment_group_name": "Jersey Fancy", "detail_desc": "Set with a vest top andindex_name": 935373, "article_id": 40589, "price": 0.033881355932033, "sales_channel_id": 2, "product_code": 665089, "prod_name": "Carla culotte", "product_type_no": 272, "product_type_nindex_group_code": 9, "colour_group_name": "Black", "perceived_colour_value_id": 4, "perceived_colour_value_name": "Dark", "perceived_colour_master_id": 5, "perceived_colour_master_nindex_group_name": "Ladieswear", "section_no": 15, "section_name": "Women's Everyday Collection", "garment_group_no": 1005, "garment_group_name": "Jersey Fancy", "detail_desc": "Wide, hindex_name": 935373, "article_id": 85831, "price": 0.0254067796610169, "sales_channel_id": 2, "product_code": 821148, "prod_name": "Corinne", "product_type_no": 258, "product_type_name": "Blouse", "prindex_group_name": "Light Orange", "perceived_colour_value_id": 3, "perceived_colour_value_name": "Light", "perceived_colour_master_id": 3, "perceived_colour_master_name": "Orange", "deindex_name": 935373, "article_id": 82777, "price": 0.016932033898305, "sales_channel_id": 2, "product_code": 80887, "prod_name": "Simona", "product_type_no": 254, "product_type_name": "Top", "product_tyindex_group_name": "White", "perceived_colour_value_id": 3, "perceived_colour_value_name": "Light", "perceived_colour_master_id": 9, "perceived_colour_master_name": "White", "departmenindex_group_no": 15, "section_name": "Women's Everyday Collection", "garment_group_no": 1005, "garment_group_name": "Jersey Fancy", "detail_desc": "Shorter, fitted top in ribbed cotton jersey witindex_name": 935373, "article_id": 662948, "prod_name": "Osman (II)", "product_type_no": 252, "product_type_name": "Sweater", "product_group_name": "Garment Upper body", "graphicaperceived_colour_value_id": 2, "perceived_colour_value_name": "Medium Dusty", "perceived_colour_master_id": 13, "perceived_colour_master_name": "Brown", "department_no": 1626, "deparindex_group_name": "Women's Everyday Collection", "garment_group_no": 1003, "garment_group_name": "Knitwear", "detail_desc": "Shorter polo-neck jumper in a rib knit containing some wool with index_name": 0.0677796610169491, "sales_channel_id": 2, "product_code": 763660, "prod_name": "Premium Pools down jkt (PQ)", "product_type_no": 262, "product_type_name": "Jacket", "product_gindex_group_name": "Greenish Khaki", "perceived_colour_value_id": 4, "perceived_colour_value_name": "Dark", "perceived_colour_master_id": 20, "perceived_colour_master_name": "Khaki greeindex_group_name": "Baby/Children", "section_no": 41, "section_name": "Baby Boy", "garment_group_no": 1007, "garment_group_name": "Outdoor", "detail_desc": "Jacket in woven fabric with ribindex_name": 935373, "article_id": 83620, "price": 0.0406610169491525, "sales_channel_id": 2, "product_code": 812683, "prod_name": "Notting Hill", "product_type_no": 265, "product_type_nindex_group_code": 12, "colour_group_name": "Light Beige", "perceived_colour_value_id": 1, "perceived_colour_value_name": "Dusty Light", "perceived_colour_master_id": 11, "perceived_colouindex_group_name": "Ladieswear", "section_no": 15, "section_name": "Women's Everyday Collection", "garment_group

Why not LLMs?

- LLMs are not trained to understand and effectively learn from the **relational nature** of Databases
- Database prediction is **NOT** sequence modeling
- LLMs are **NOT** trained to predict the future
- **Context size** is **limited**

Key Question

How to leverage the data without going through the longest duration tasks (extraction, transformation, loading)?

We want ML algorithms that can process data in its natural form!

Modern ML

- Today we want to design deep learning models that operate on relational tables
- But modern deep learning toolbox is designed for **different type of inputs**

Doubt thou the stars are fire,
Doubt that the sun doth move,
Doubt truth to be a liar,
But never doubt I love...

Text



Images

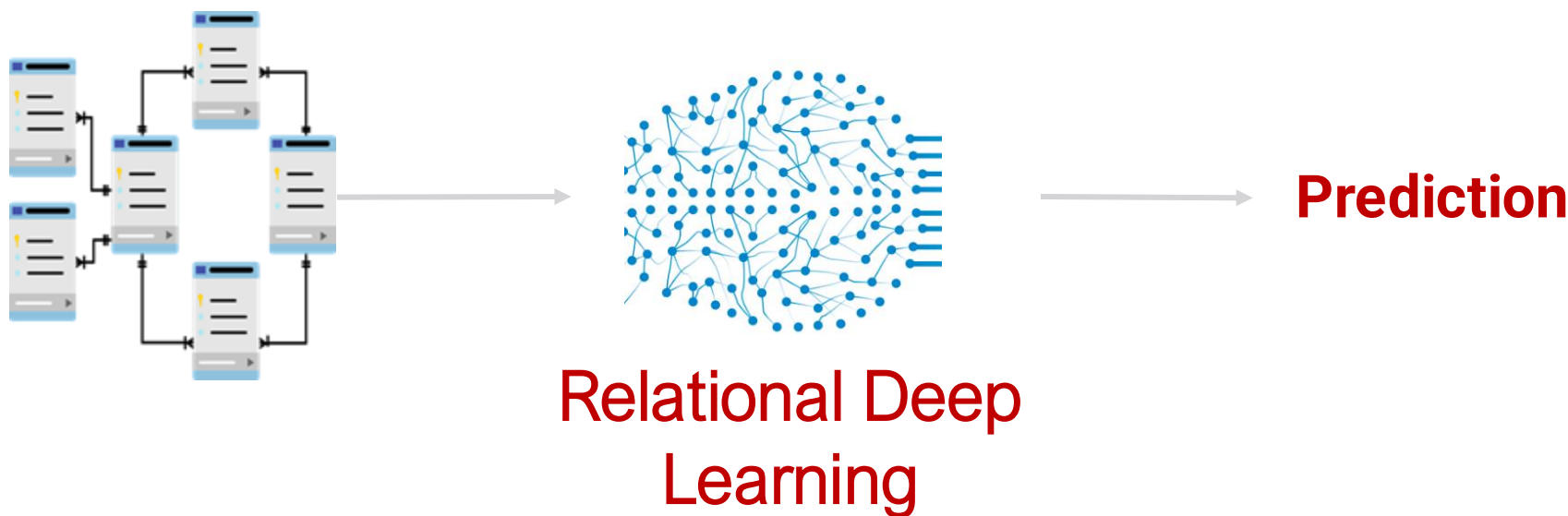


Stanford CS224W: Generalizing Deep Learning to Databases

CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



Deep Learning on Relational Tables



What is RDL?

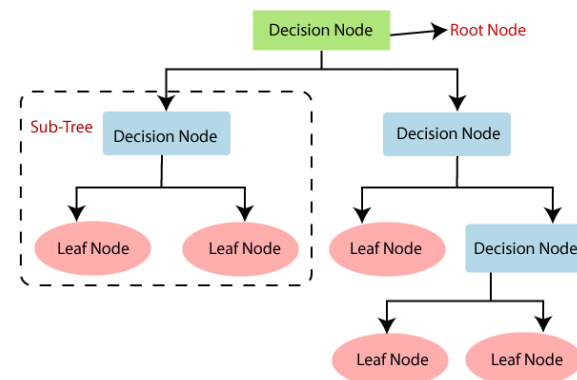
- End-to-end **deep learning on relational tables** (*i.e.*, databases)
- Works directly on relational tables, no transformations, no feature engineering
- Casts predictive tasks as **graph representation learning** problems

Impact and Consequences

- **More accurate models**
(no feature engineering)
- **More robust models**
(model-learned features automatically “update” over time)
- **Shorter time to models**
(no mundane ETL work)
- **Simpler infrastructure**
(no pipelines, no feature stores, etc.)

Related Work: Tabular ML

- Great for building models on **one table**
 - But most data is not a single table!
- Deep learning is not dominant:
 - **Hypothesis:** Because single table already contains features engineered from other tables (**loss of information**)
- **RDL accounts for relations between multiple tables, unlike tabular ML!**

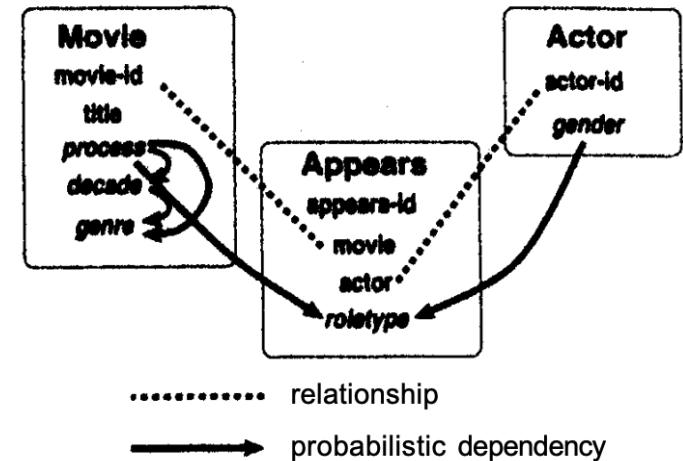


Related Work: Stat. Rel. Learning

[Friedman et al., 1999]

Learns a **distribution**
(graphical model) **over**
the relational structure

- Predictions via world model + inductive logic programming
- Only works with categorical variables
- Does not scale to large data
- RDL a **scalable, expressive inheritor** of Statistical Relational Learning



Stanford CS224W: Relational Deep Learning

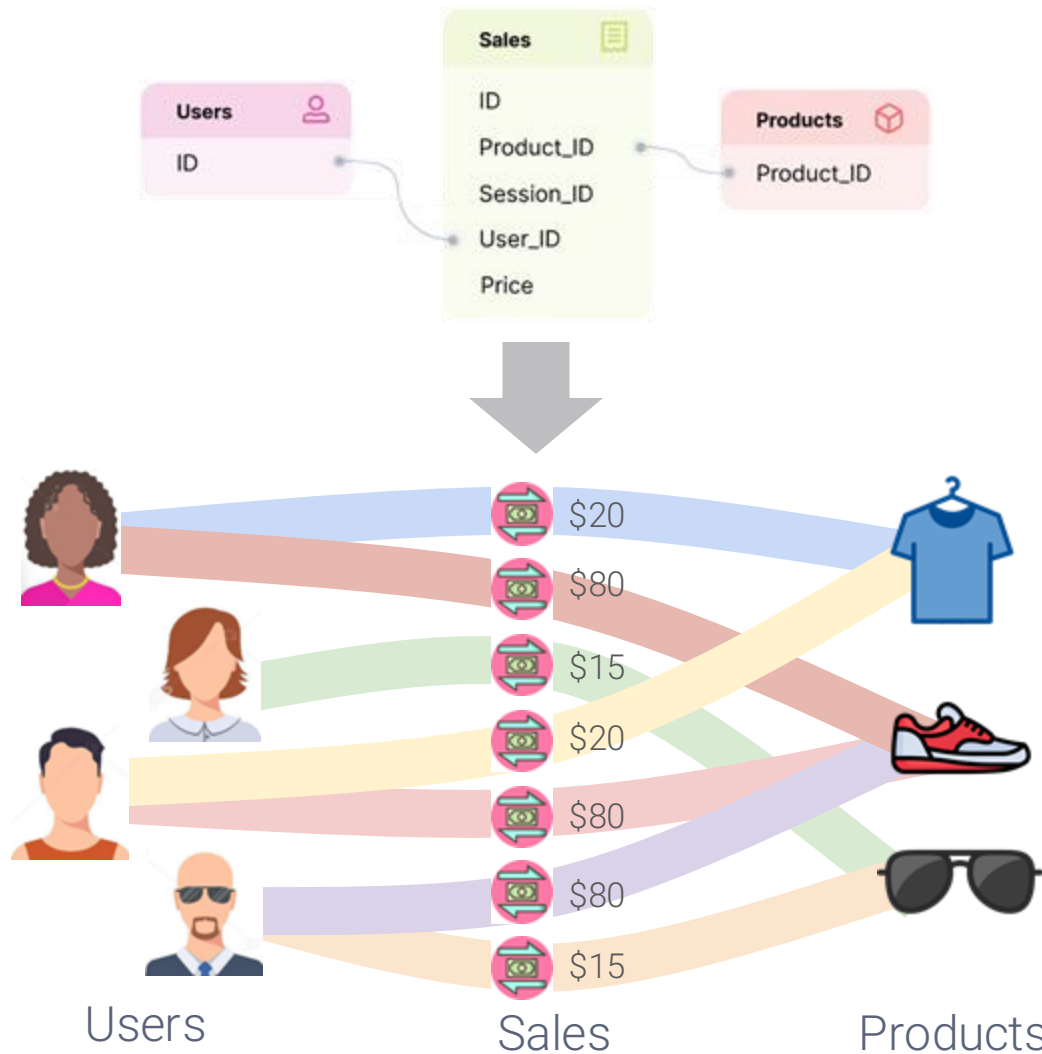
CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



Insight: A Data is a graph!



A Database is a graph!



Just do ML on a Graph!

ML in the language of graphs:

- **Node-level:**

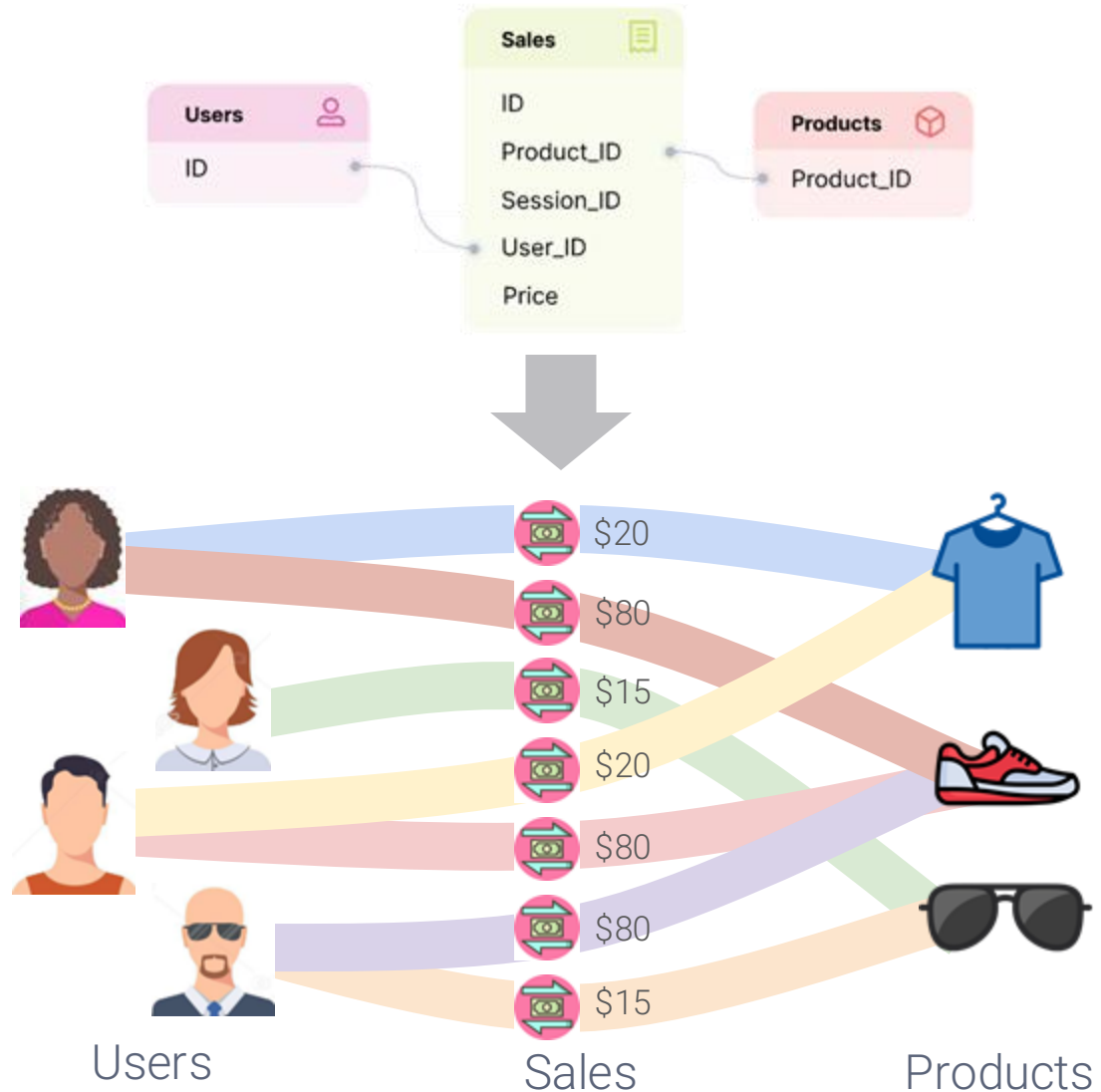
- Churn
- Life-time value
- Next best action

- **Link-level:**

- Product affinity
- Recommendations

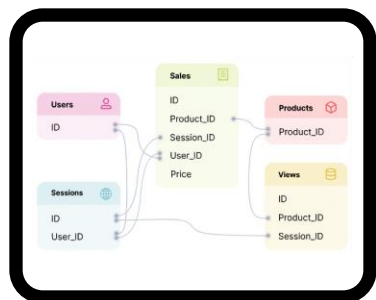
- **Graph-level:**

- Fraud, money laundering

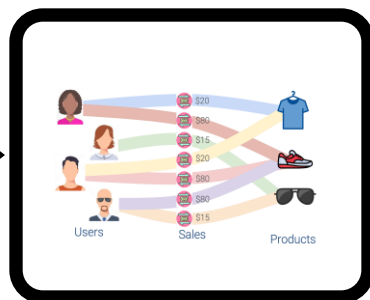


Graph ML Problem Solving Pipeline

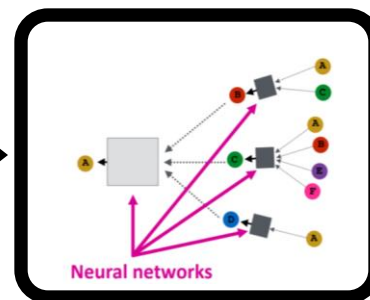
Relational DB



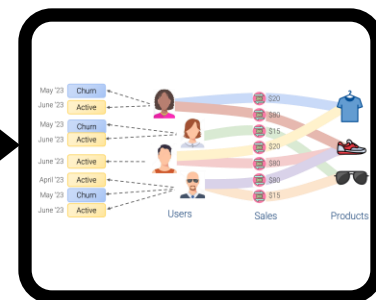
Graph Problem



Graph ML



Solutions



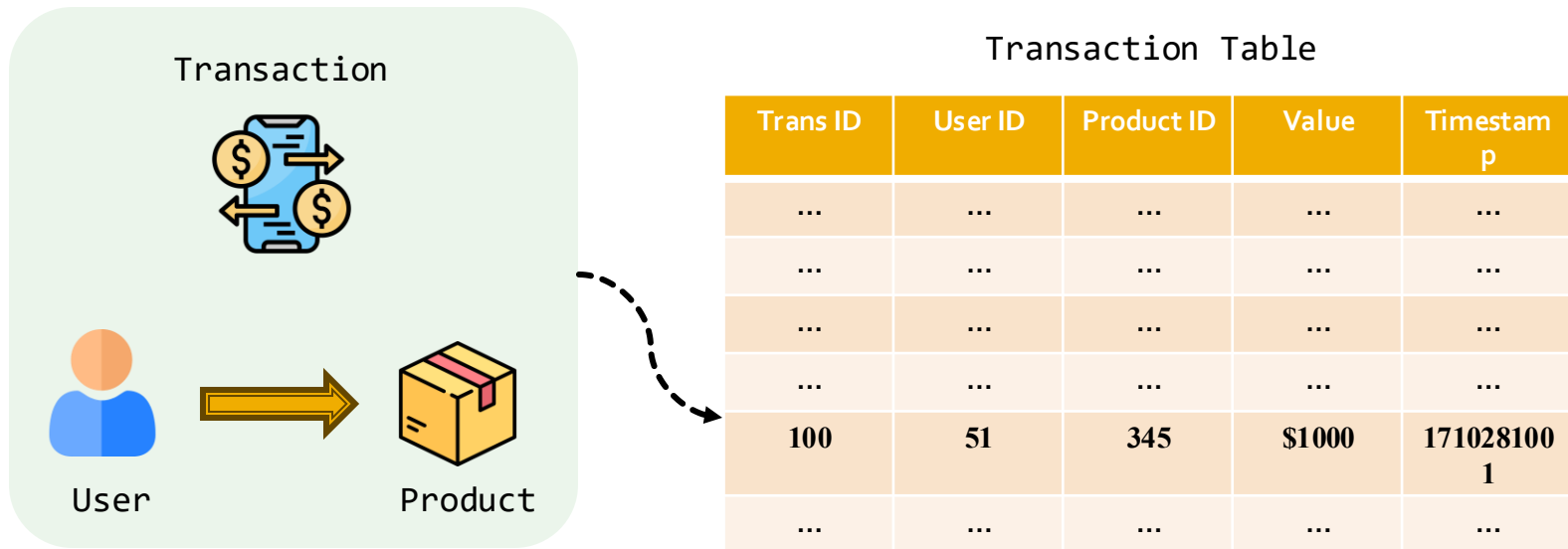
Stanford CS224W: Relational Database Graph

CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



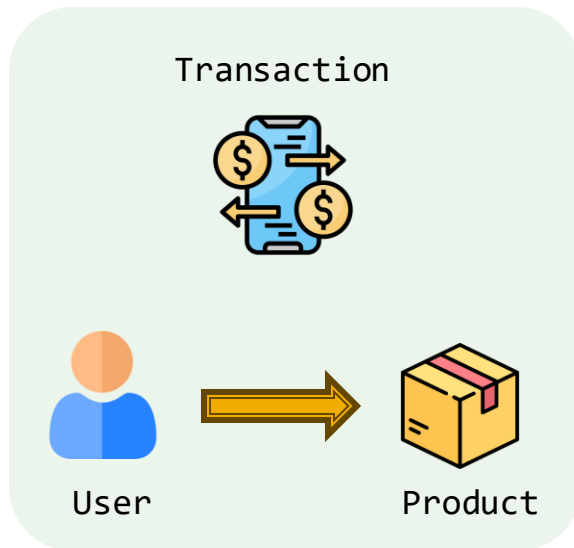
Databases

Real-world data are stored in databases



Relational Databases

Databases are often relational



Transaction Table

Trans ID	User ID	Product ID	Value	Timestamp
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
100	51	345	\$1000	1710281001
...	...	...	...	...

User Table

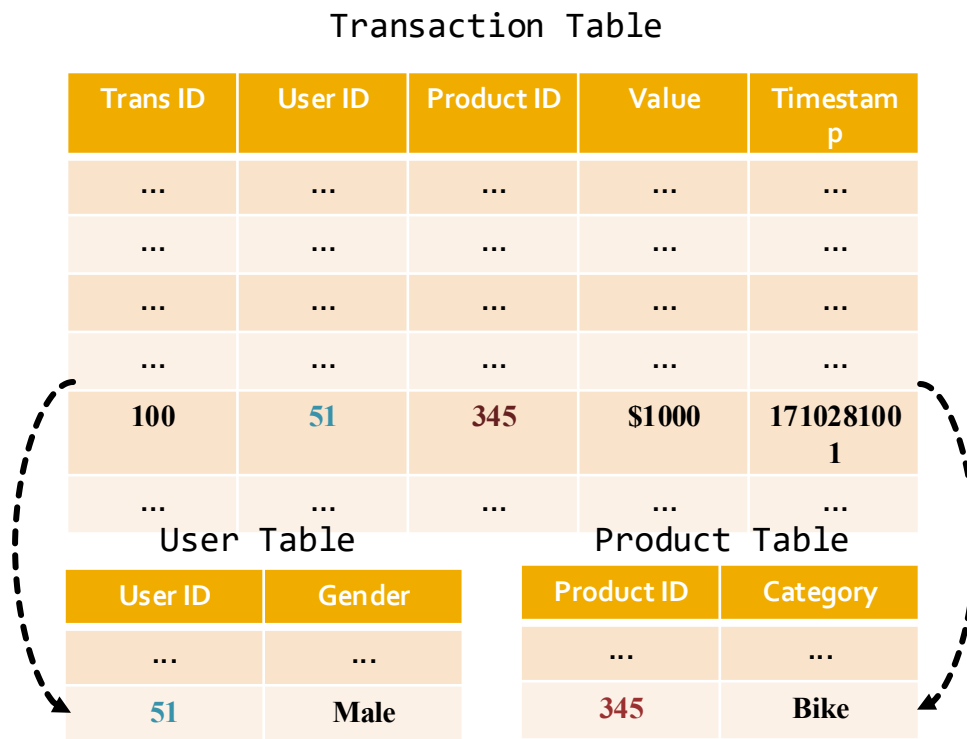
User ID	Gender
...	...
51	Male

Product Table

Product ID	Category
...	...
345	Bike

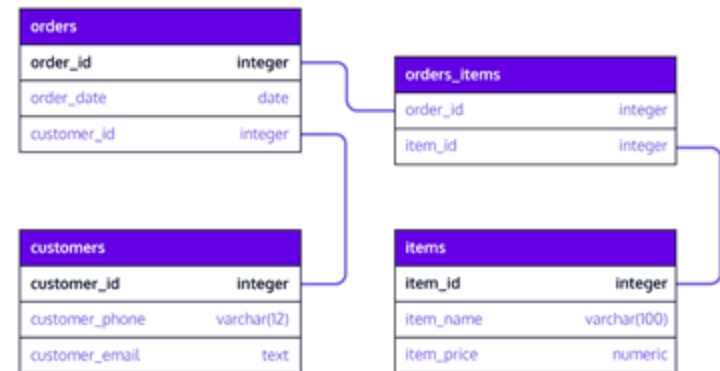
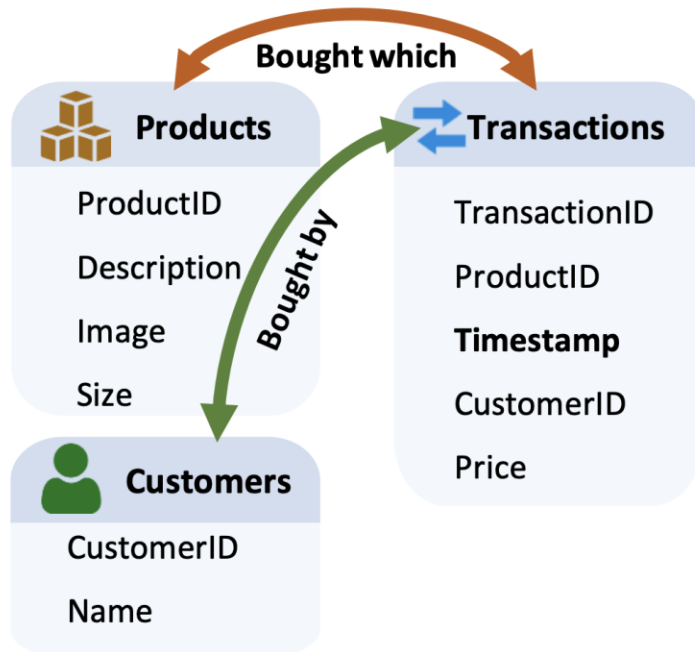
Relational databases as heterogeneous graphs

Mathematically...



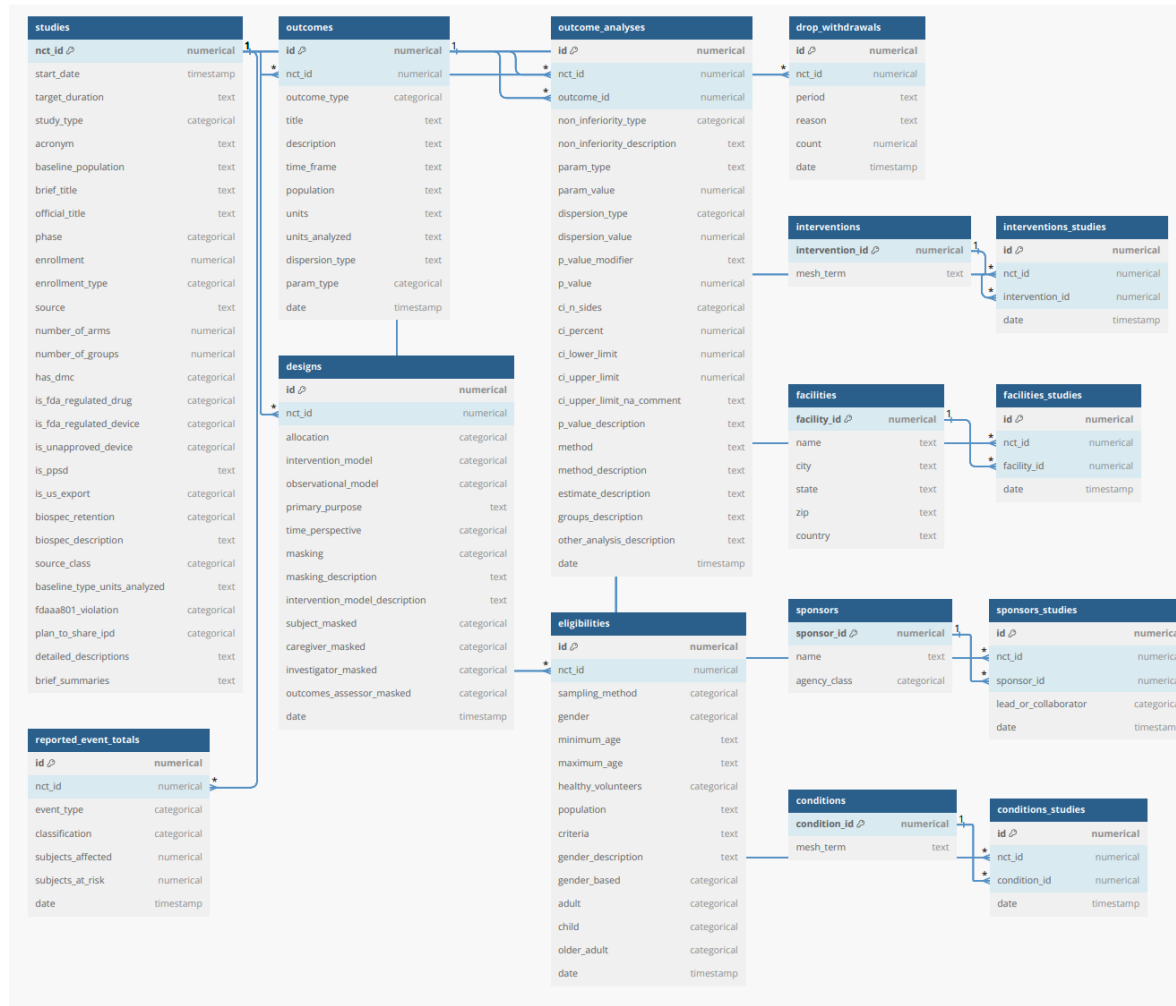
- A database is a set of tables $\mathcal{T} = \{T_1, \dots, T_n\}$ and
- Links between tables $\mathcal{L} \subseteq \mathcal{T} \times \mathcal{T}$

Schema Graph



- The schema graph represents the high-level structure of the heterogeneous graph

Rel-trial Schema Graph

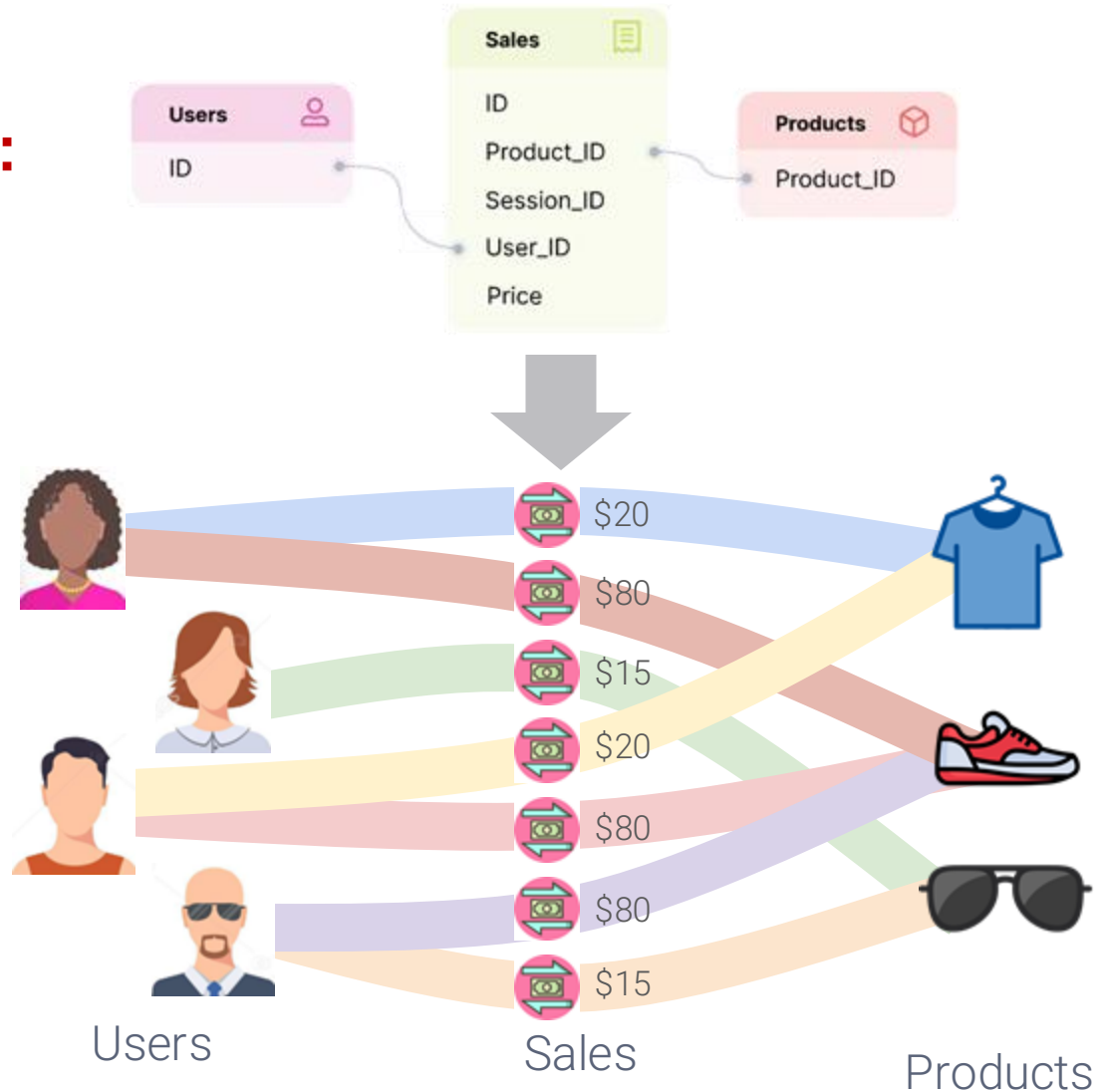


- The schema graphs can be more complex

Relational Entity Graph

Relational Entity Graph:

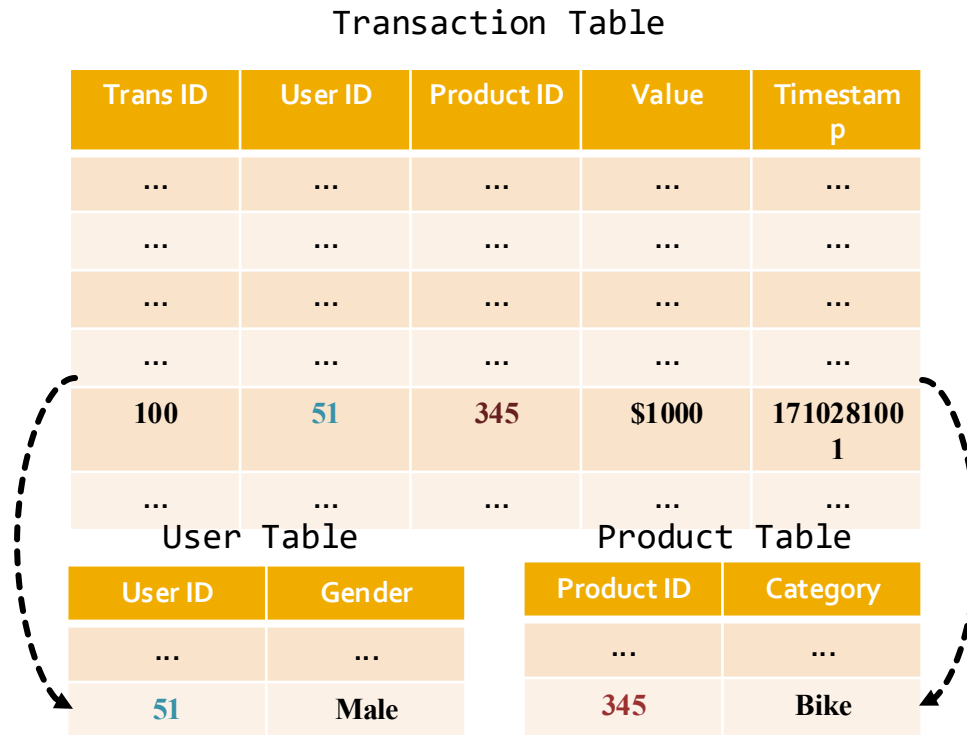
Create connections via primary-foreign keys



Mathematically...

- Given database of tables $\mathcal{T} = \{T_1, \dots, T_n\}$ and relations $\mathcal{L} \subseteq \mathcal{T} \times \mathcal{T}$
- Each table is a set of entities $v \in T_i$ possessing a primary key and optional foreign keys
- The **relational entity graph** is such that
 - the **set of nodes** is defined by all rows in all tables
$$\mathcal{V} = \bigcup_{T \in \mathcal{T}} T$$
 - The **set of edges** is defined by connecting two entities v_1, v_2 whose primary and foreign keys match

Entities in Relational Entity Graph



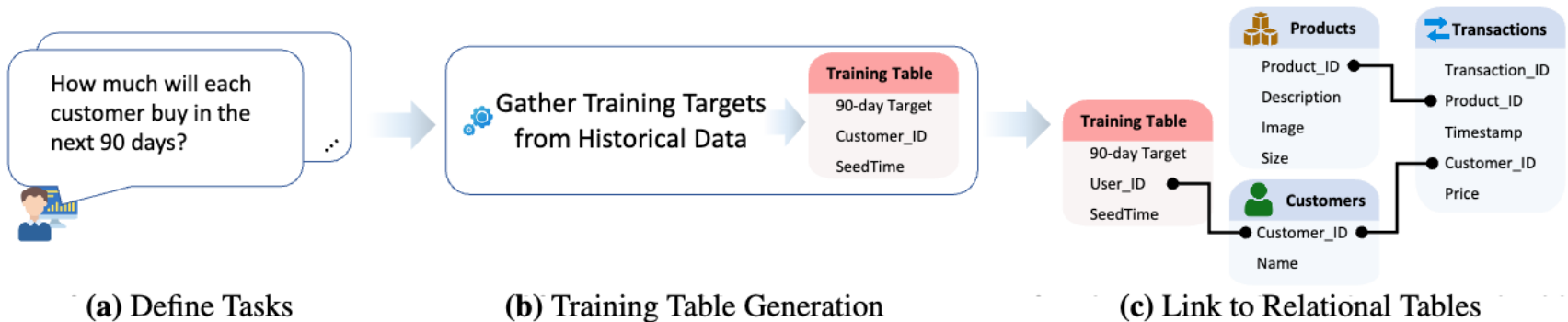
- Entities also have features (different than KGs)

Stanford CS224W: Predictive Tasks in Relational Databases

CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



Overview of the Approach



- Next:
 - Define task(s)
 - Relational Deep Learning

Defining a Task

+ **Example:** Predict whether a user is going to churn in the next 30 days?

+ Most tasks are **temporal**:

✧ Customer's **label changes over time**

✧ Database changes over time

✧ For every time, a feature needs to be recomputed

+ **To define a task, we need:**

✧ Entity

✧ Label

✧ Time



Temporal tasks are especially challenging because features are time dependent:

- For every time, a feature needs to be recomputed
- Entity's label can change between time steps

Defining a Task: Training Table

- **Training Table:** A special table containing training labels
 - (entity ID, time, labels)
 - Classification, Regression, Multi-class
 - time column is essential for **temporal prediction tasks**
 - An entity may have different labels at different times
 - Only use information up to the time of label

Defining a Task: Training Table

- **Training Table:** A special table containing training labels
 - (entity ID, time, labels)
 - Classification, Regression, Multi-class

Training Table

Entity ID	Timestamp	Label
99	10172024	1
99	10182024	1
...	...	...
100	10172024	1
100	10182024	0
...	...	...

Example: Churn

- Schema:



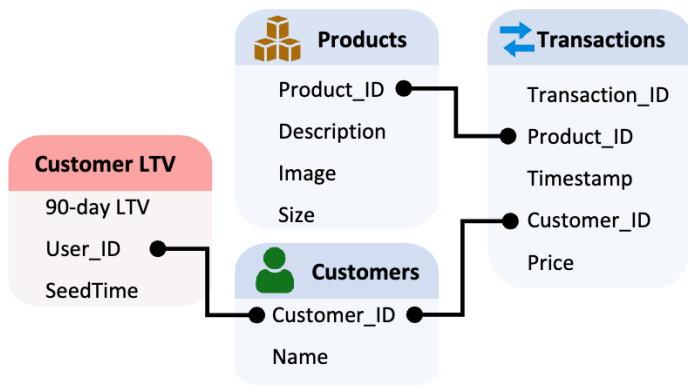
- Example prediction task:

- Predict whether a user is going to churn
 - Zero sales in the next 30 days.

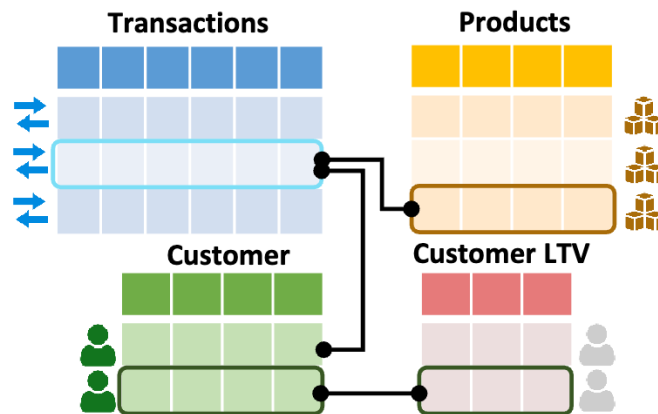
- Training table:

- (user, time, churn label)

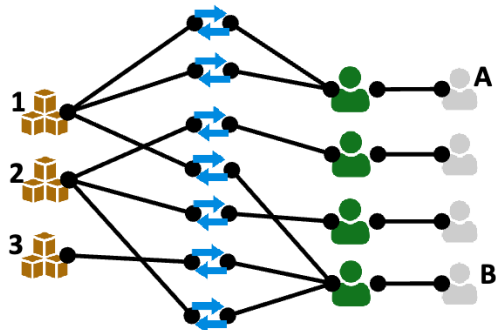
Relational Deep Learning



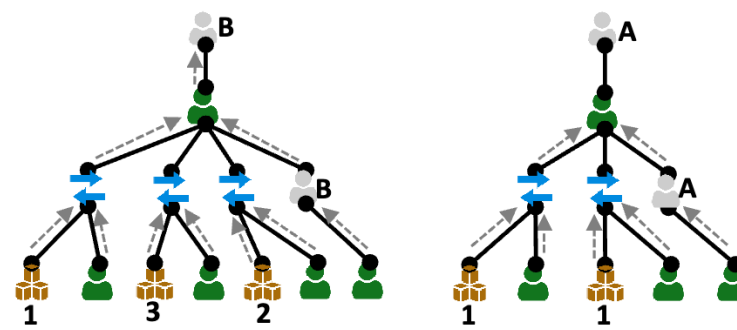
(a) Rel. Tables with Training Table



(b) Entities Linked by Foreign Keys



(c) Relational Entity Graph

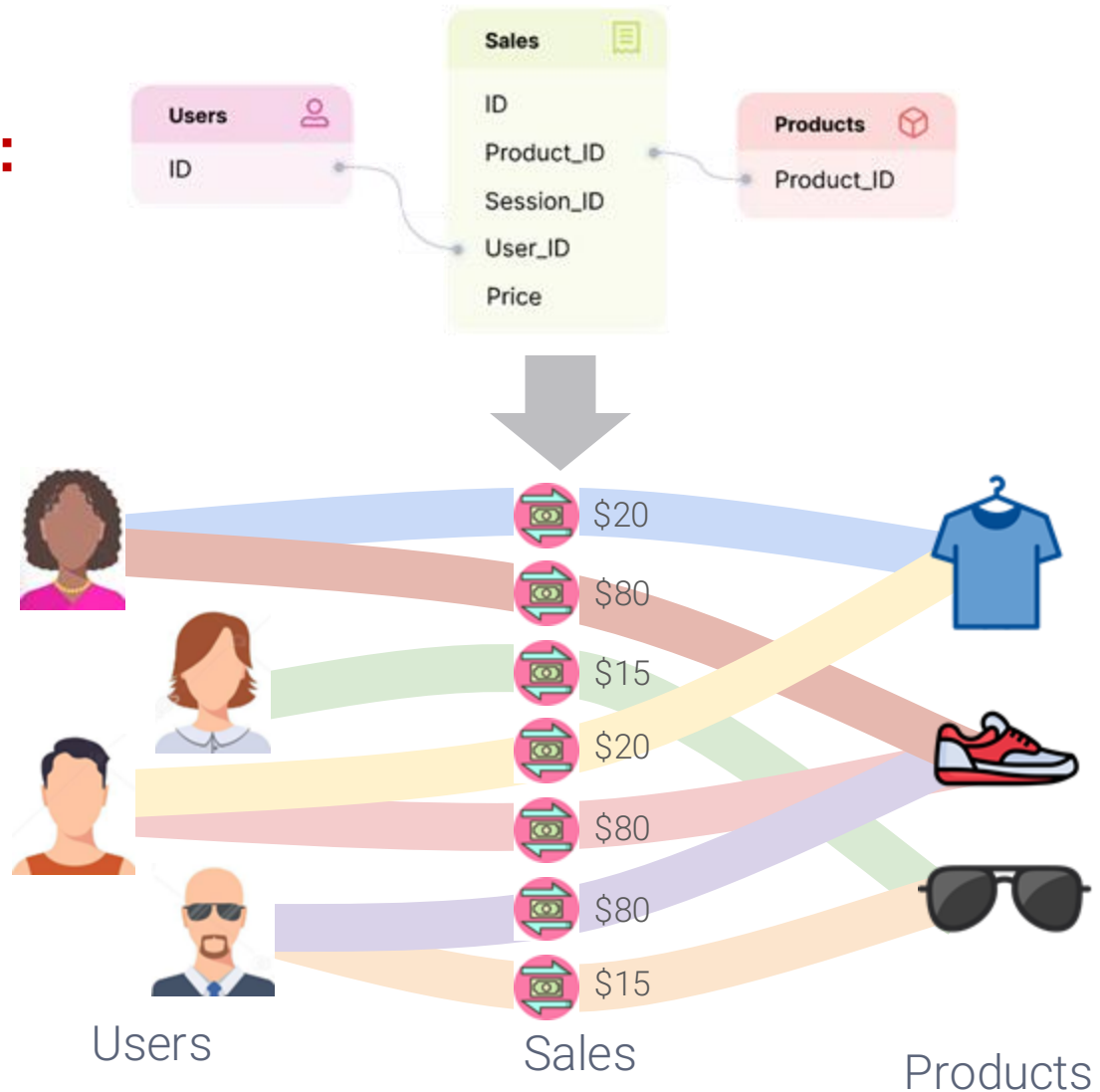


(d) Graph Neural Network

Relational Entity Graph

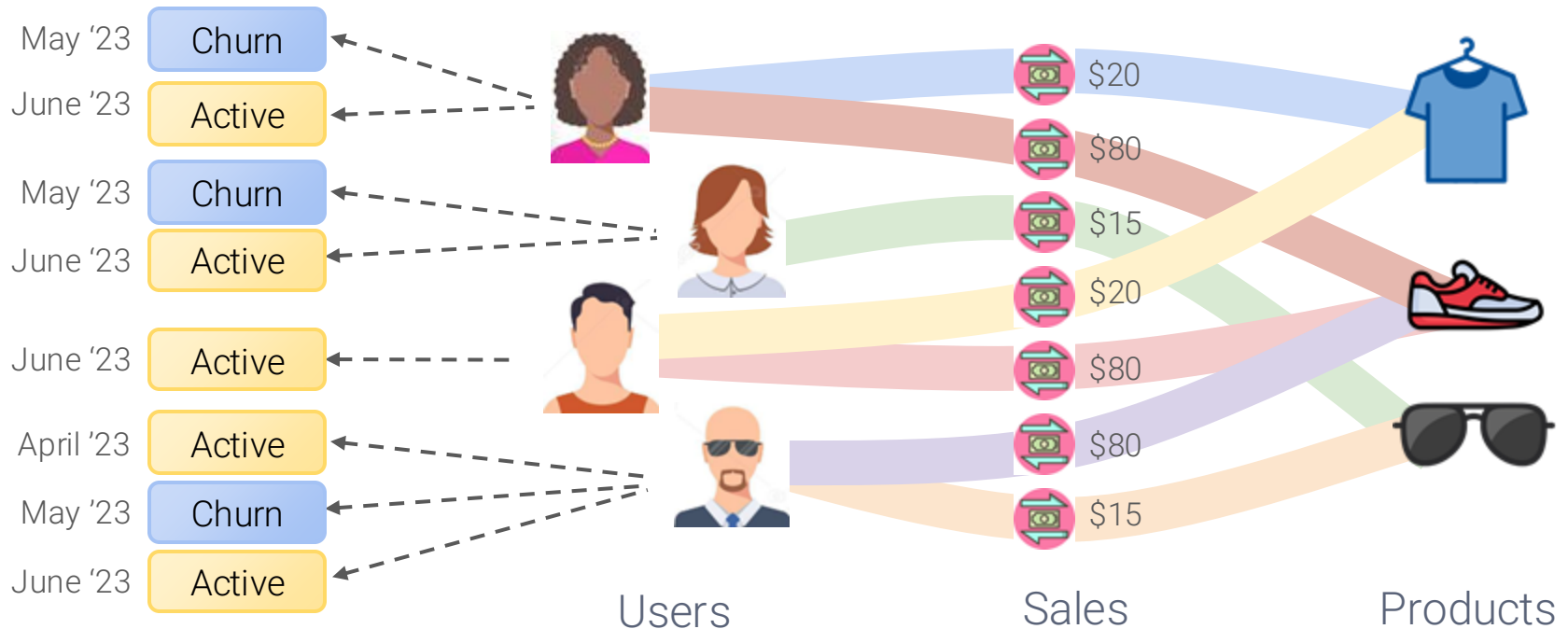
Relational Entity Graph:

Create connections via primary-foreign keys



Connect the Training Table

Training labels together with timestamps are attached to the graph

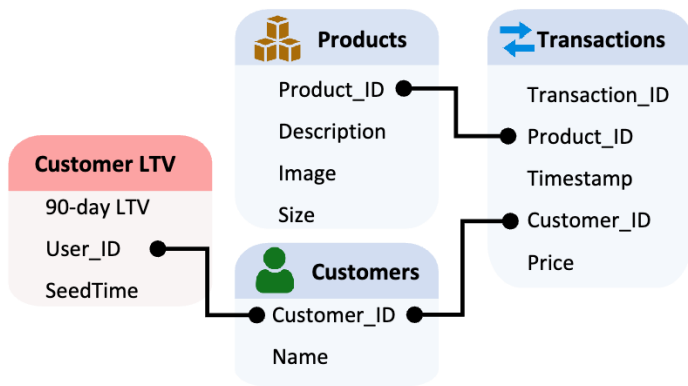


Stanford CS224W: Relational GNN

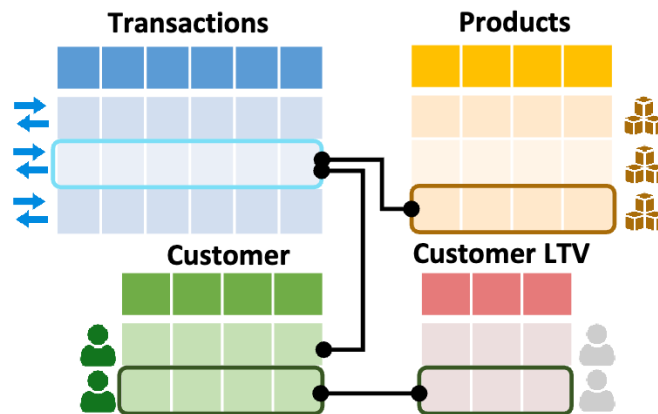
CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



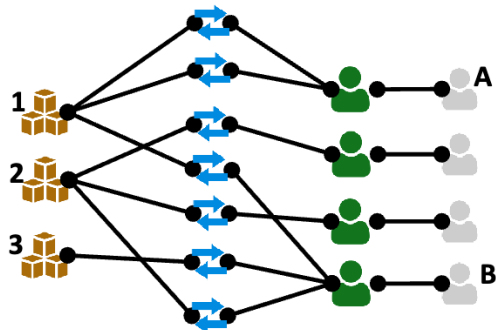
Relational Deep Learning



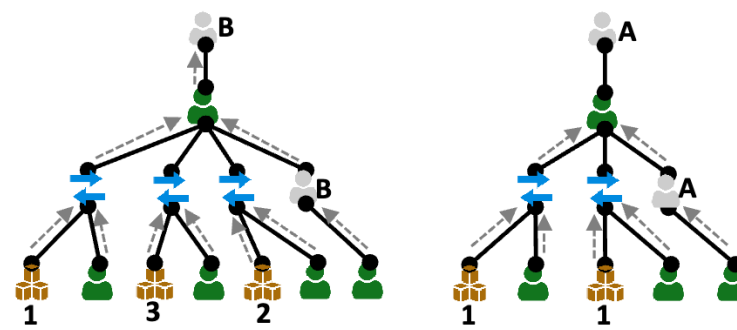
(a) Rel. Tables with Training Table



(b) Entities Linked by Foreign Keys



(c) Relational Entity Graph

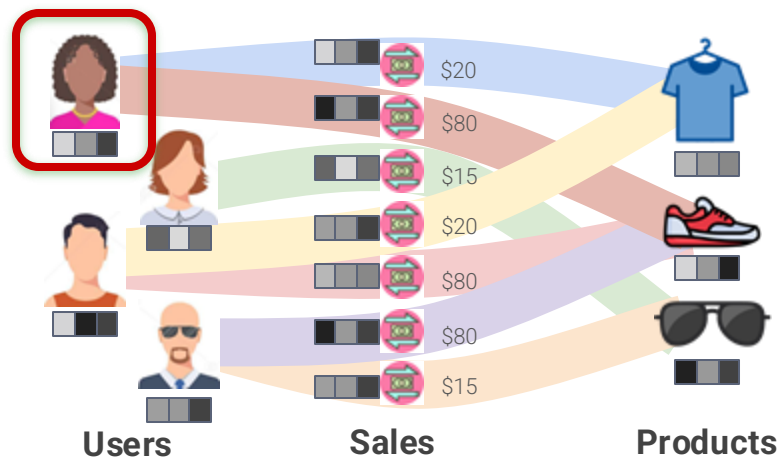


(d) Graph Neural Network

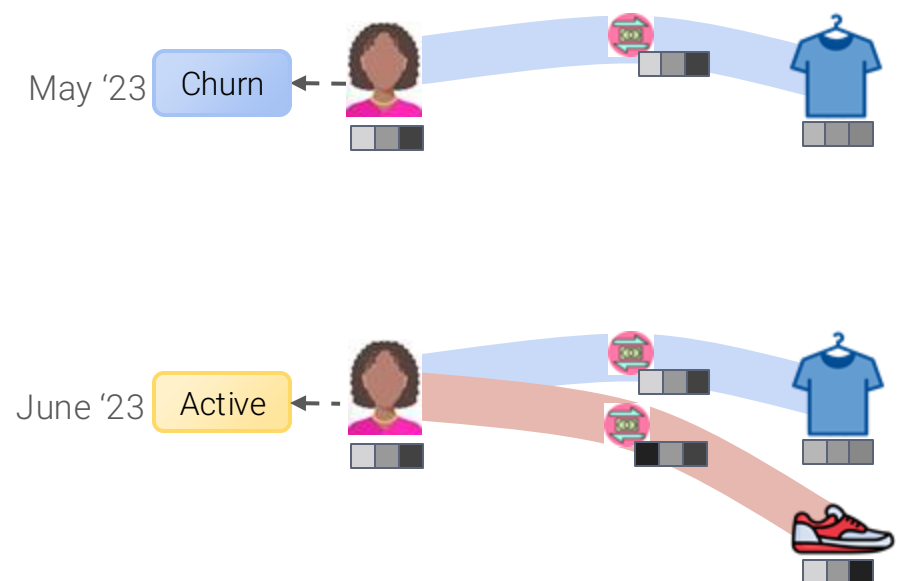
GNN on the Entity Graph

Node's neighborhood defines a computation graph

Nodes learn how to *optimally* use information from neighbors to obtain enhanced node representations

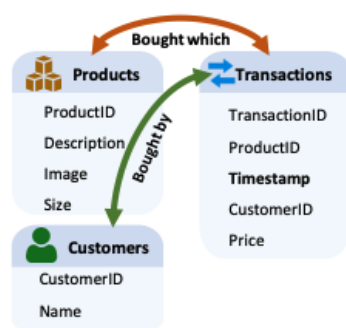


Entity Graph

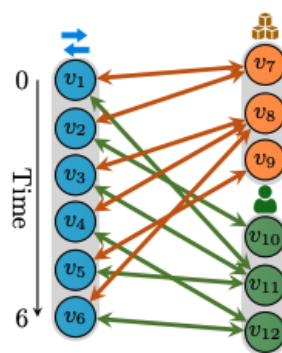


GNN computation graphs

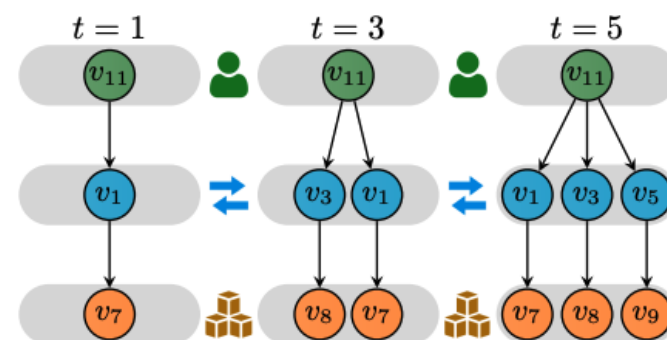
GNNs on Temporal Graphs



(a) Schema Graph



(b) Relational Entity Graph

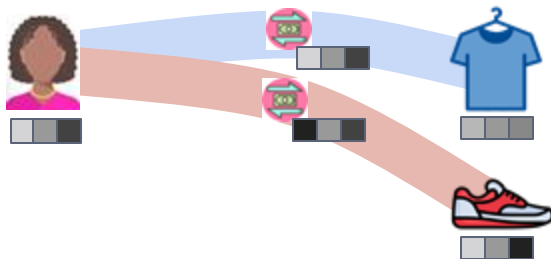


(c) Computation Graphs for different time t

- The computation graph for each node is **time-dependent**
- Message+Aggregation becomes **time-dependent**
- Sampling over neighbors is **time-dependent**

GNN vs Feature Engineering

GNN-based features:



vs.

Hand-engineered features:

<code>SUM(TRANSACTIONS.Price)</code> over <code>(-30, 0)</code> days	<code>AVG(TRANSACTIONS.Price)</code> over <code>(-30, 0)</code> days	...
---	---	-----

GNN aggregation is **learnable version of hand-crafted features!**

GNNs give better performance by learning optimal features.

SQL joins vs Graph edges

Definitions:

- A table R is a set of entities $R = \{r_1, \dots, r_n\}$
- Each entity is a tuple $r_i = (r_{i1}, r_{i2}, \dots, r_{in})$
- Given two tables, R, S , a join operation is a **subset** of a Cartesian product:
- Aggregation will specify which rows are kept

Ex.1: Cross Join

Employee table		Department table	
LastName	DepartmentID	DepartmentID	DepartmentName
Rafferty	31	31	Sales
Jones	33	33	Engineering
Heisenberg	33	34	Clerical
Robinson	34	35	Marketing
Smith	34		
Williams	NULL		

```
SELECT *  
FROM employee INNER JOIN department ON 1=1;
```

Employee.LastName	Employee.DepartmentID	Department.DepartmentName	Department.DepartmentID
Rafferty	31	Sales	31
Jones	33	Sales	31
Heisenberg	33	Sales	31
Smith	34	Sales	31
Robinson	34	Sales	31
Williams	NULL	Sales	31
Rafferty	31	Engineering	33
Jones	33	Engineering	33
Heisenberg	33	Engineering	33
Smith	34	Engineering	33
Robinson	34	Engineering	33
Williams	NULL	Engineering	33
Rafferty	31	Clerical	34
Jones	33	Clerical	34
Heisenberg	33	Clerical	34
Smith	34	Clerical	34
Robinson	34	Clerical	34
Williams	NULL	Clerical	34
Rafferty	31	Marketing	35
Jones	33	Marketing	35
Heisenberg	33	Marketing	35
Smith	34	Marketing	35
Robinson	34	Marketing	35
Williams	NULL	Marketing	35

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

SQL joins vs Graph edges

Connection between SQL Joins and graph edges

Ex.2: Inner
Join

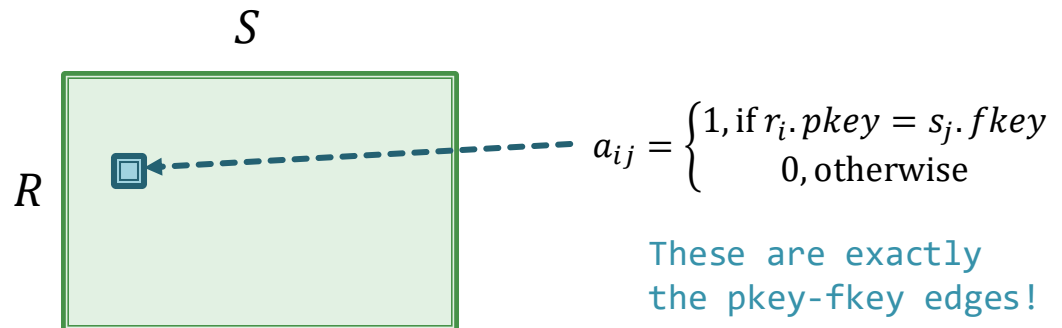
$$R \bowtie S := \{r \cup s \mid r \in R \wedge s \in S \wedge \text{Fun}(r, s)\}$$

$$\text{Fun}(r, s): r.pkey = s.fkey$$

Employee table		Department table	
LastName	DepartmentID	DepartmentID	DepartmentName
Rafferty	31	31	Sales
Jones	33	33	Engineering
Heisenberg	33	34	Clerical
Robinson	34	35	Marketing
Smith	34		
Williams	NULL		

```
SELECT employee.LastName, employee.DepartmentID, department.DepartmentName
FROM employee
INNER JOIN department ON
employee.DepartmentID = department.DepartmentID;
```

Employee.LastName	Employee.DepartmentID	Department.DepartmentName
Robinson	34	Clerical
Jones	33	Engineering
Smith	34	Clerical
Heisenberg	33	Engineering
Rafferty	31	Sales



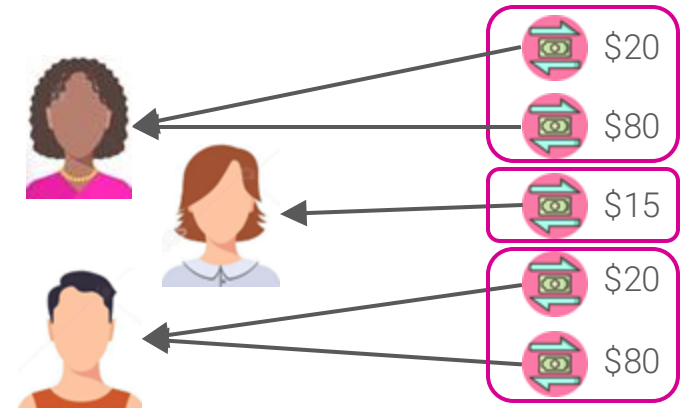
https://en.wikipedia.org/wiki/Relational_algebra

GNNs perform a JOIN+AGG

Input:

USER_ID
1
2
3

USER_ID	VALUE	DATE
1	20	01-01
1	80	01-02
2	15	01-01
3	20	01-02
3	80	01-03



$$\bigoplus_{w \in \mathcal{N}(v)} \mathbf{t}(t_w) \cdot \mathbf{x}_w$$

GNN can learn:

```
SELECT SUM(VALUE)
FROM SALES
WHERE DATE > 01-01
GROUP BY USER_ID
```

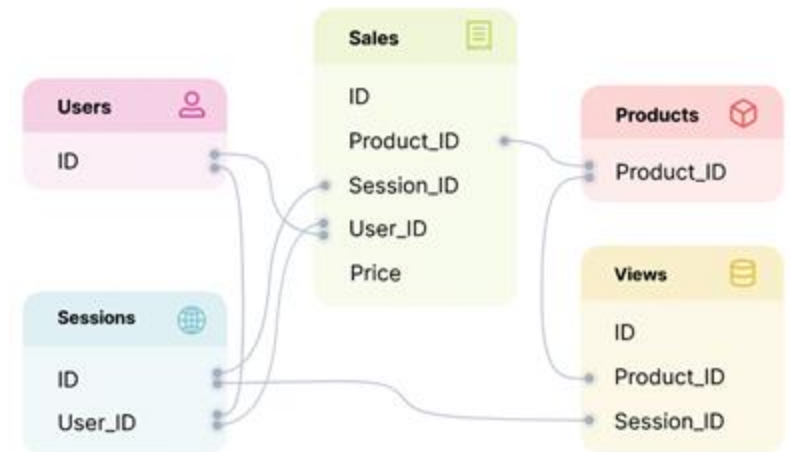
USER_ID	SUM(VALUE)
1	80
2	0
3	100

Learnable Aggregation $\bigoplus$
 Temporal embedding $\mathbf{t}(t_w)$
 Fact representation $\mathbf{x}_w$

Benefits of GNNs

GNNs learn how to aggregate information:

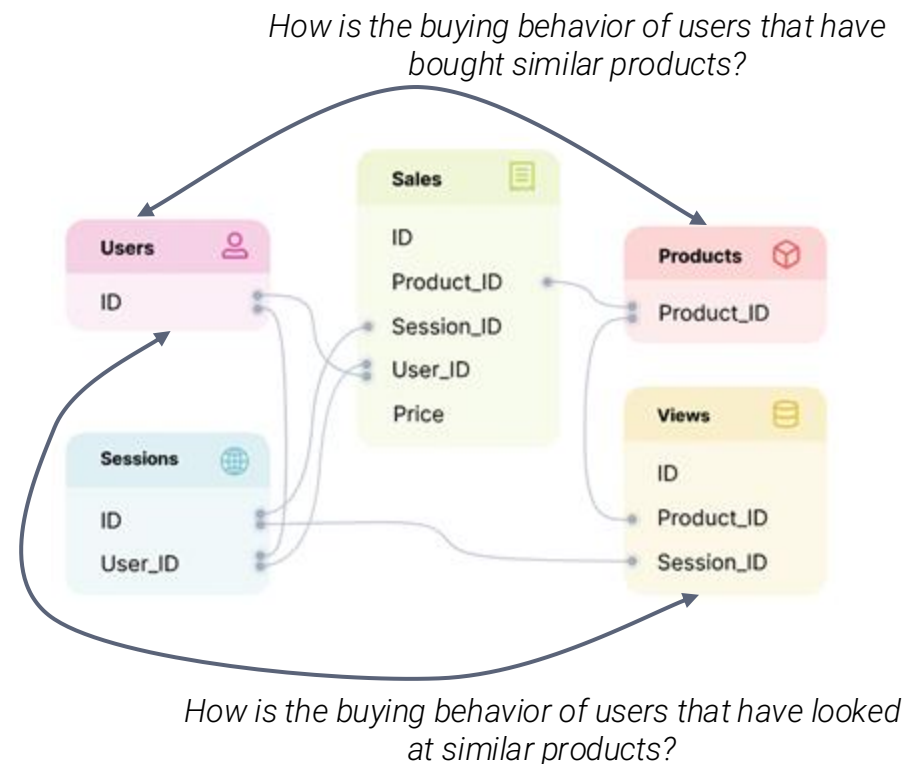
- They can discard neighboring node information that is irrelevant for the given downstream task
- They can detect fine-grained patterns within local neighborhoods (e.g., buying behavior over the last year)



Benefits of GNNs

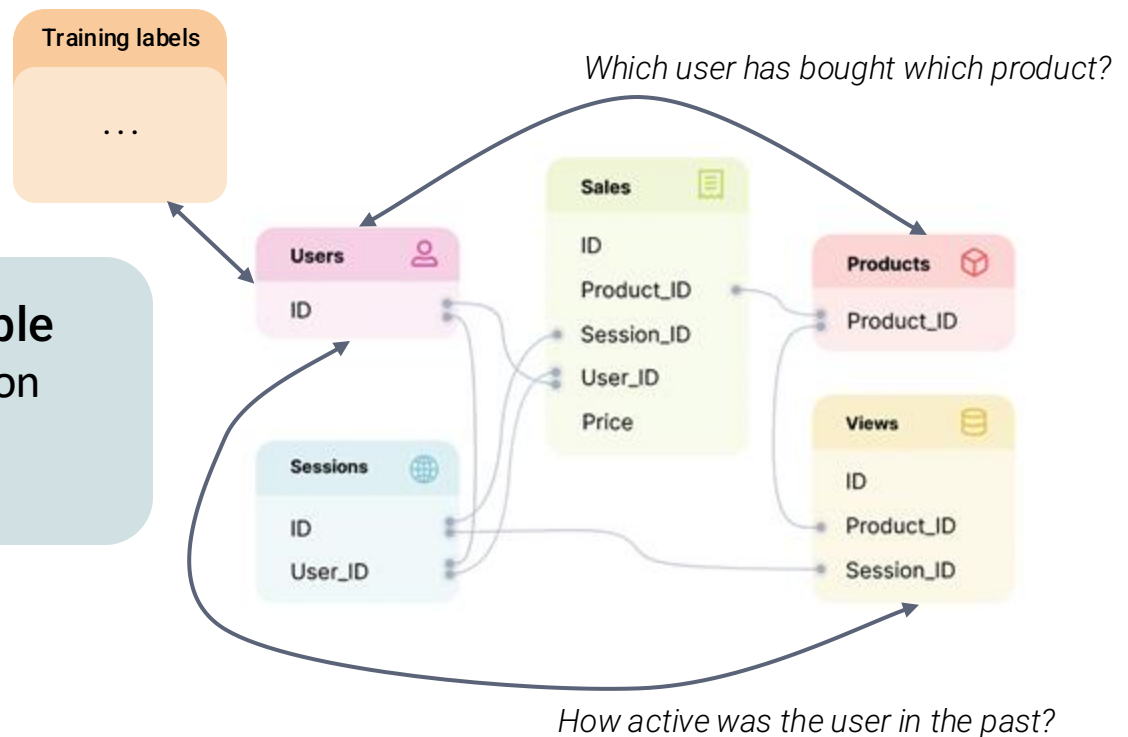
GNNs can exchange information *across* training examples:

- ✗ Instead of treating examples as isolated, there now exists an inter-dependency *between* entities (e.g., users with similar features, users with similar behavior)
- ✗ GNN can *use* these features to enrich an entity's representation



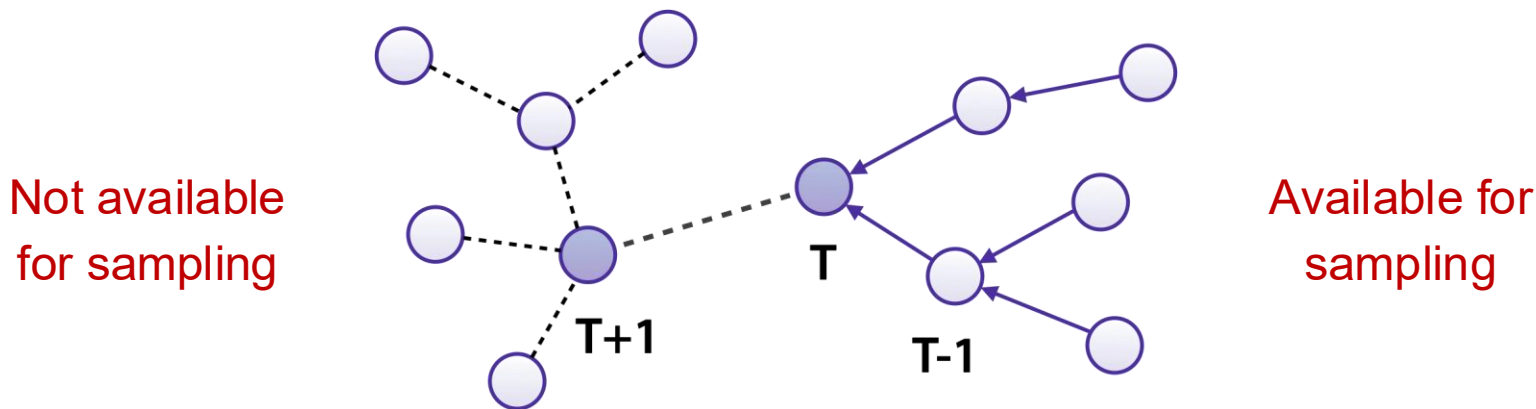
Benefits of GNNs

Multi-hop reasoning *across table boundaries* can catch information which is **hard** to pre-compute *beforehand*



Benefits of GNNs

- First-class **temporal support**
 - Capture fine-grained relative and seasonal features via temporal embeddings
 - Avoid data leakage via temporal sampling directly during data loading



Full Vision Described in Paper

Relational Deep Learning: Graph Representation Learning on Relational Tables

**Matthias Fey^{2,*}, Weihua Hu^{2,*}, Kexin Huang^{1,*}, Jan Eric Lenssen^{2,3,*}, Rishabh Ranjan^{1,*},
Joshua Robinson^{1,*}, Rex Ying⁴, Jiaxuan You⁵, Jure Leskovec^{1,2}**

*Equal contribution. Listed in alphabetic order.

¹Stanford University

²Kumo.AI

³Max Planck Institute for Informatics

⁴Yale University

⁵University of Illinois at Urbana-Champaign

Available at: <https://relbench.stanford.edu>



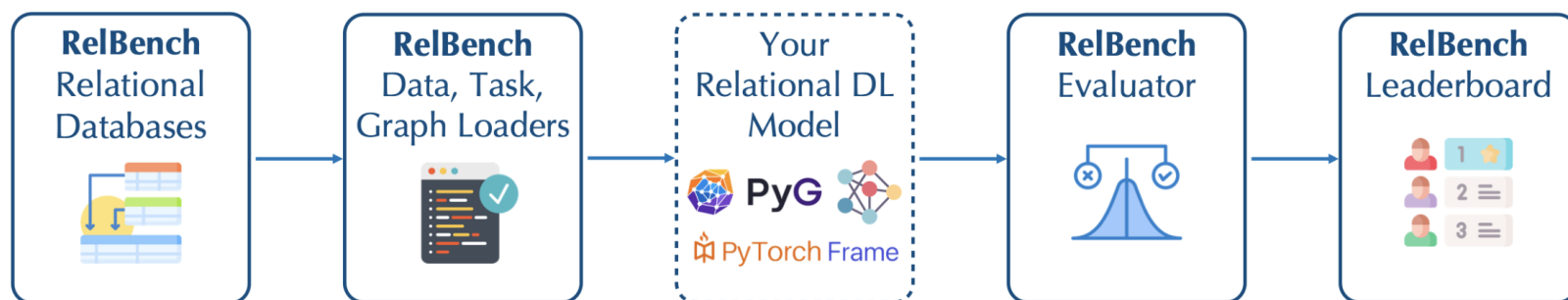
Stanford CS224W: RELBNCH

CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



Enabling Research on RDL

- Relbench is more than just a collection of Databases



- Automatically download datasets
- Load database and task tables
- Standardized evaluation protocol:
 - Prevents temporal leakage from test set
- Framework-agnostic data structures: use your favorite ML stack!

PyF and PyG Integration

- Load as a PyG graph
- Train GNN end-to-end
- Temporal neighbor sampling
- Use PyTorch Frame to encode tables



RelBench Datasets

7 Diverse Datasets



E-Commerce

- rel-amazon
- rel-avito
- rel-hm



Social

- rel-event
- rel-stack



Sports

- rel-f1

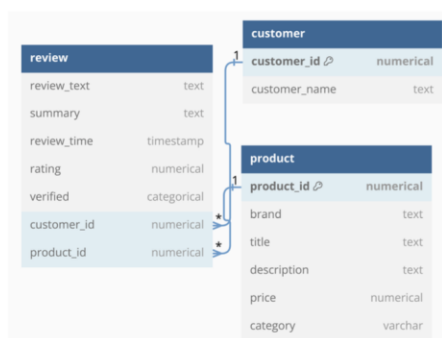


Medical

- rel-trial

RelBench Datasets

Rich Schemas



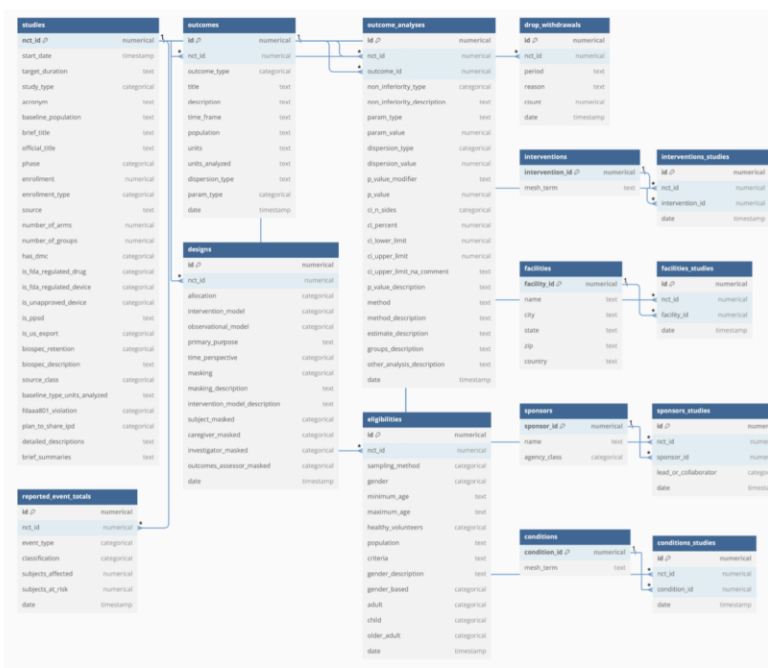
rel-amazon

3 to 15 tables

74k to 41M rows in a DB

15 to 140 columns in a DB

Time span from **2 weeks** to **55 years**



rel-trial

RelBench Tasks

30 Real-World Predictive Tasks

Entity Classification

- rel-amazon
 - user-churn
 - item-churn
- rel-stack
 - user-badge
- rel-trial
 - study-outcome

...

Entity Regression

- rel-amazon
 - user-ltv
 - item-ltv
- rel-avito
 - ad-ctr
- rel-f1
 - driver-position

...

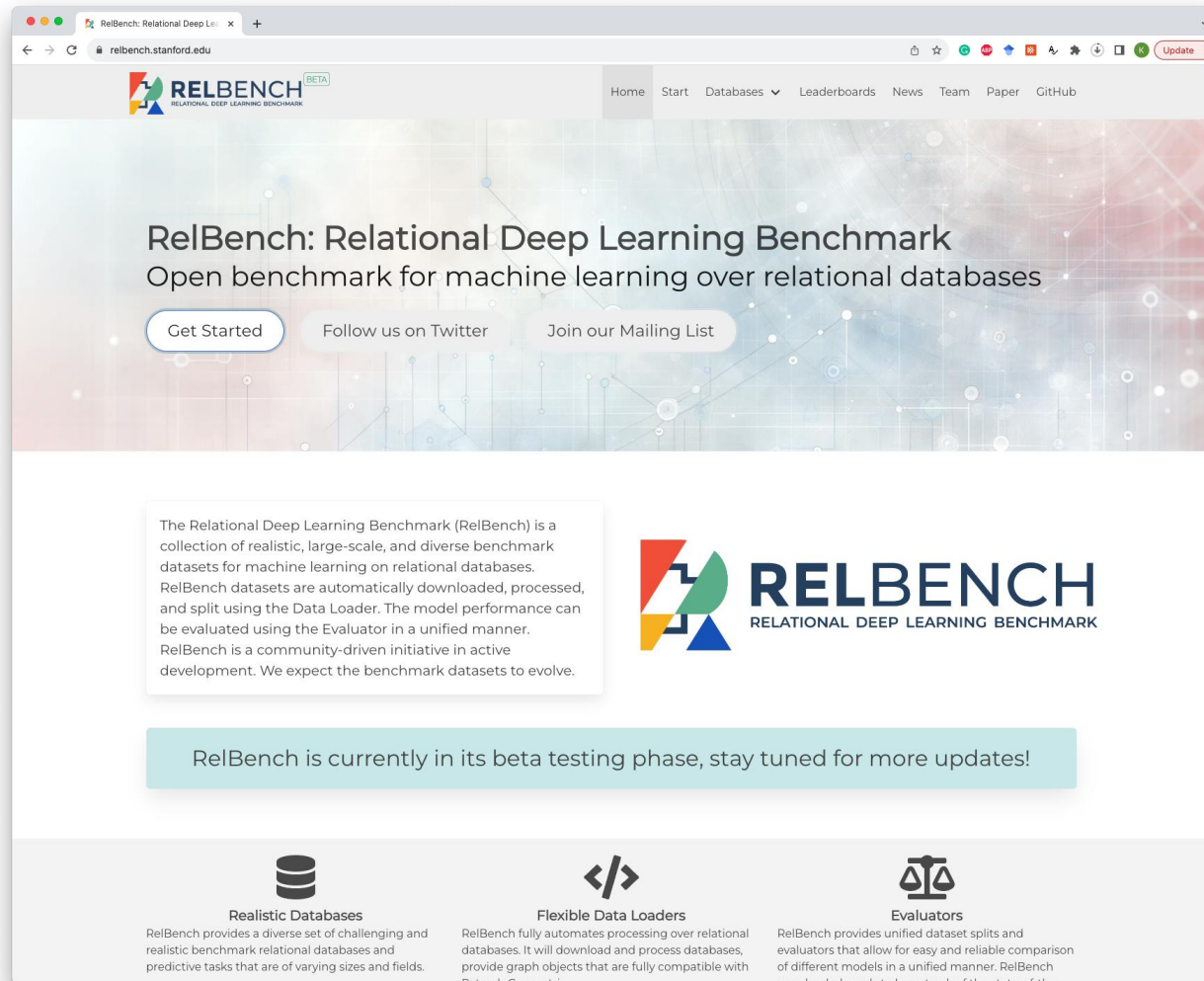
Recommendation

- rel-amazon
 - user-item-purchase
- rel-avito
 - user-ad-visit
- rel-stack
 - user-post-comment
 - post-post-related

...

See website to get started

<https://relbench.stanford.edu>



Stanford CS224W: GNN vs expert Data Scientist

CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



Expert Data Scientist



(Alejandro)

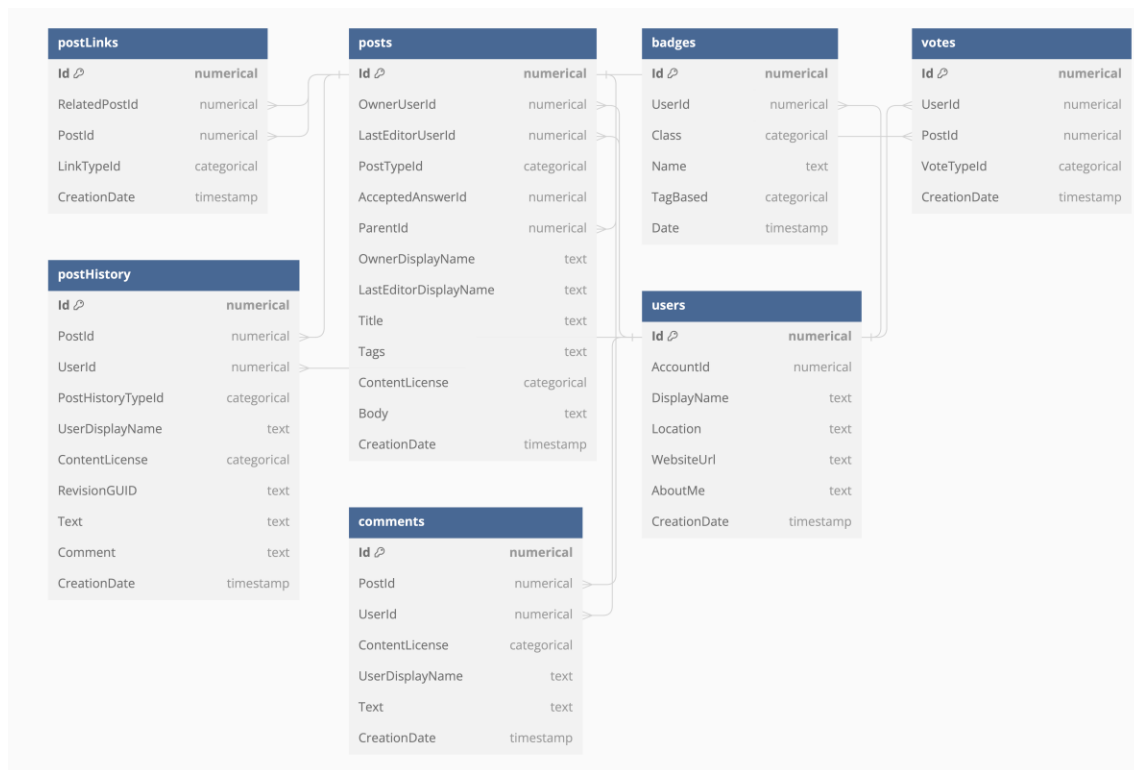
- Recruited Experienced Data Scientist
- 5 years in industry (*specializing in financial databases*)
- Responsible for full model building lifecycle (*more detail next*)



Representative Relbench task

Q: Will a user be active in the next 6 months?

Stack Exchange Database



Expert Data Scientist Workflow

Task: Will a user be active in the next 6 months?

Expert Data Scientist Workflow



Exploratory Data Analysis
(EDA)

Manual work

4hrs

Task: Will a user be active in the next 6 months?

Expert Data Scientist Workflow



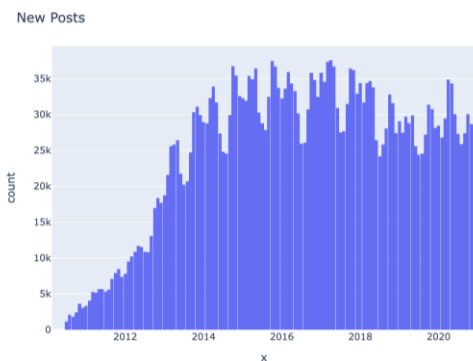
Exploratory Data Analysis
(EDA)

Manual work

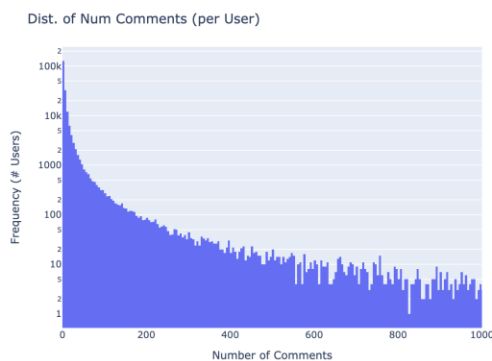
4hrs

Task: Will a user be active in the next 6 months?

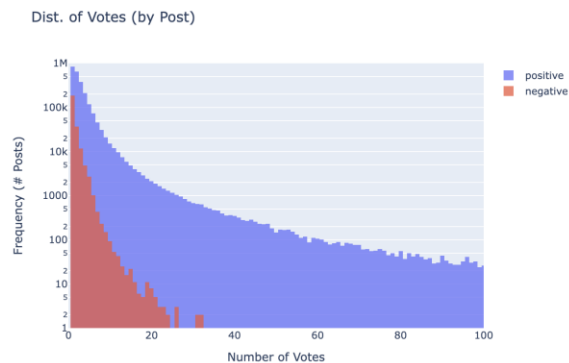
Example observations



Activity is seasonal

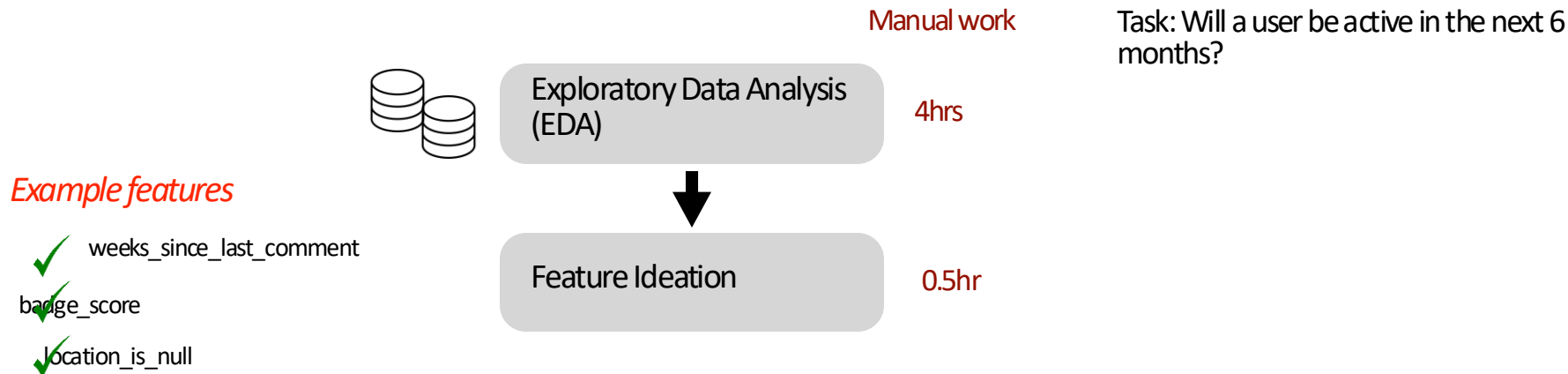


Comments follow power law

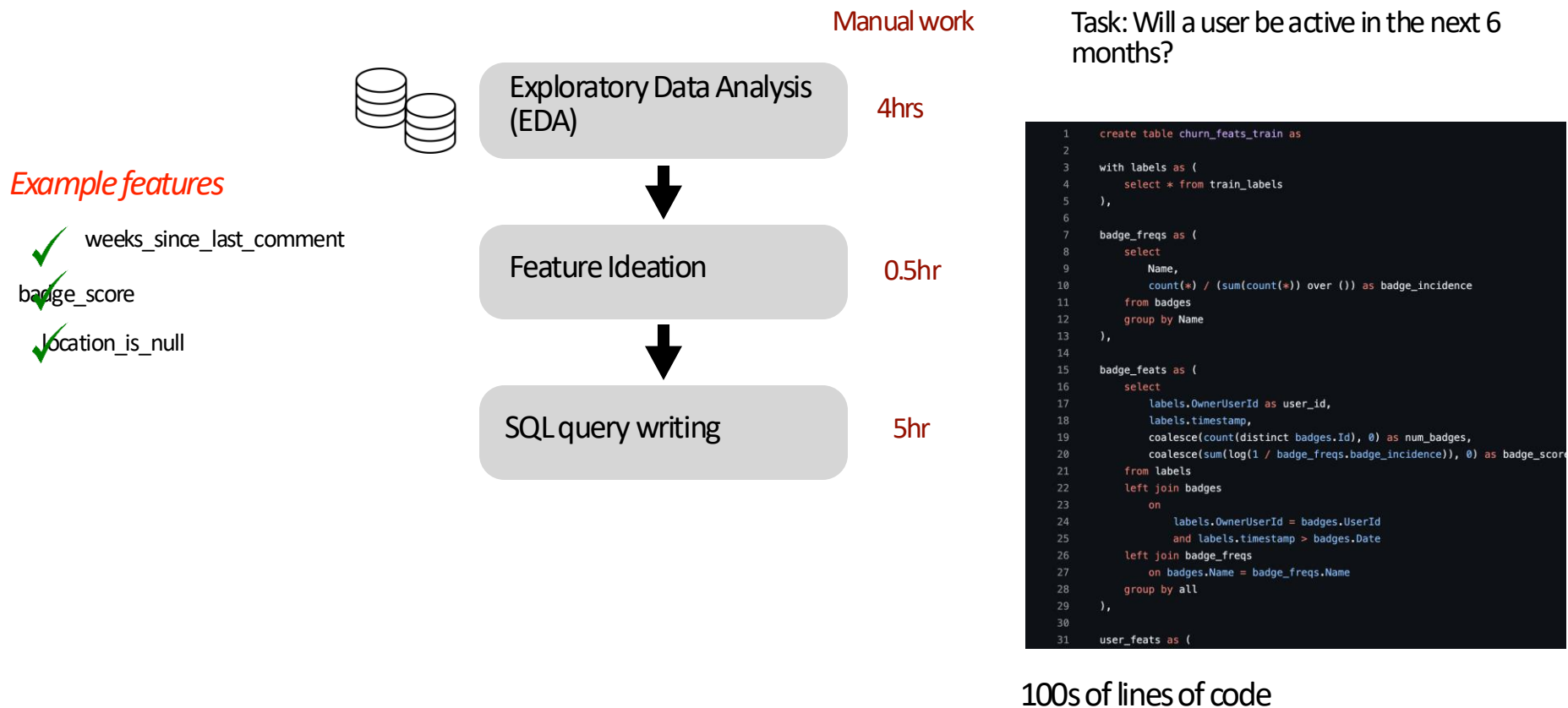


Negative votes are infrequent

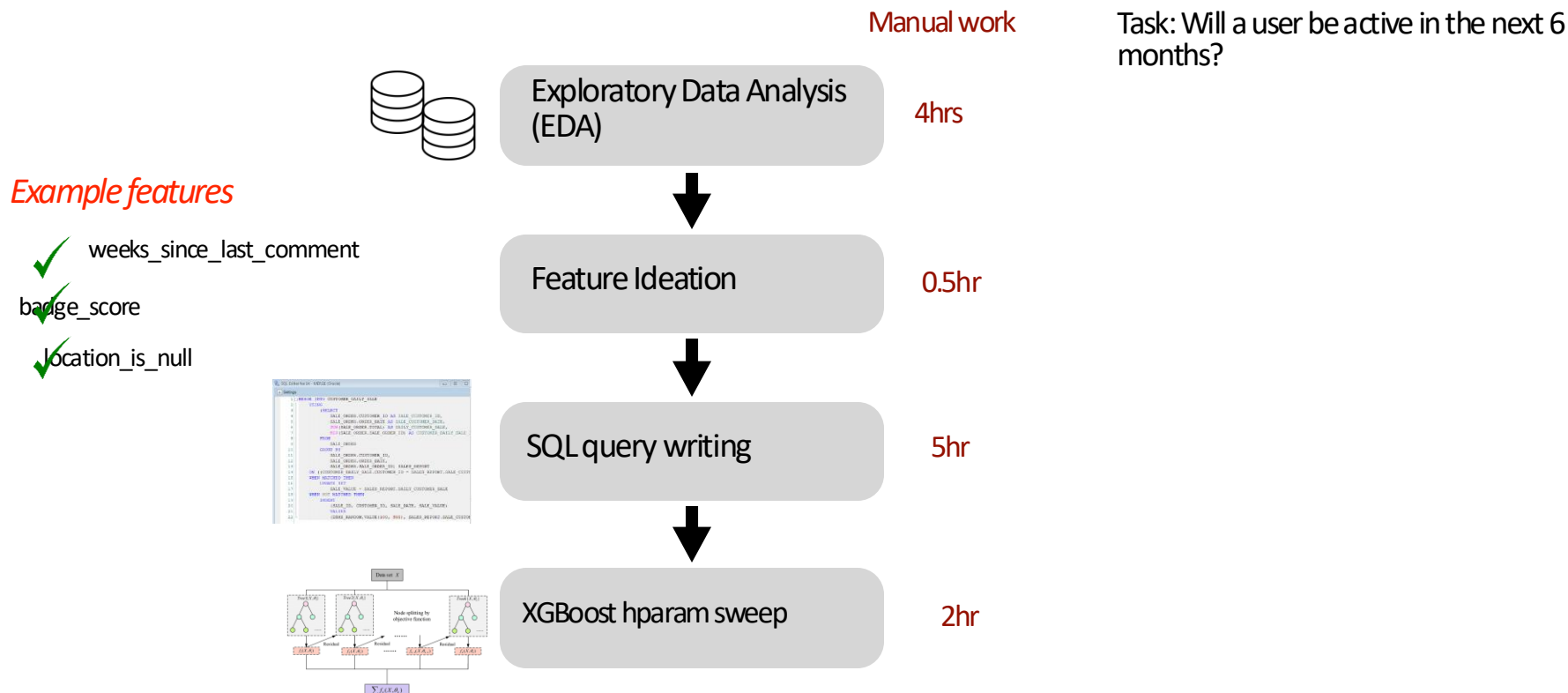
Expert Data Scientist Workflow



Expert Data Scientist Workflow



Expert Data Scientist Workflow



Expert Data Scientist Workflow

Manual work

Task: Will a user be active in the next 6 months?



Exploratory Data Analysis (EDA)

4hrs



Feature Ideation

0.5hr



SQL query writing

5hr



XGBoost hparam sweep

2hr



SHAP (feature importance analysis)

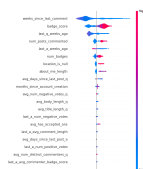
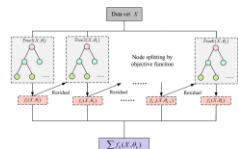
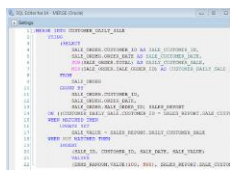
1hr

Example features

weeks since last comment

badge score

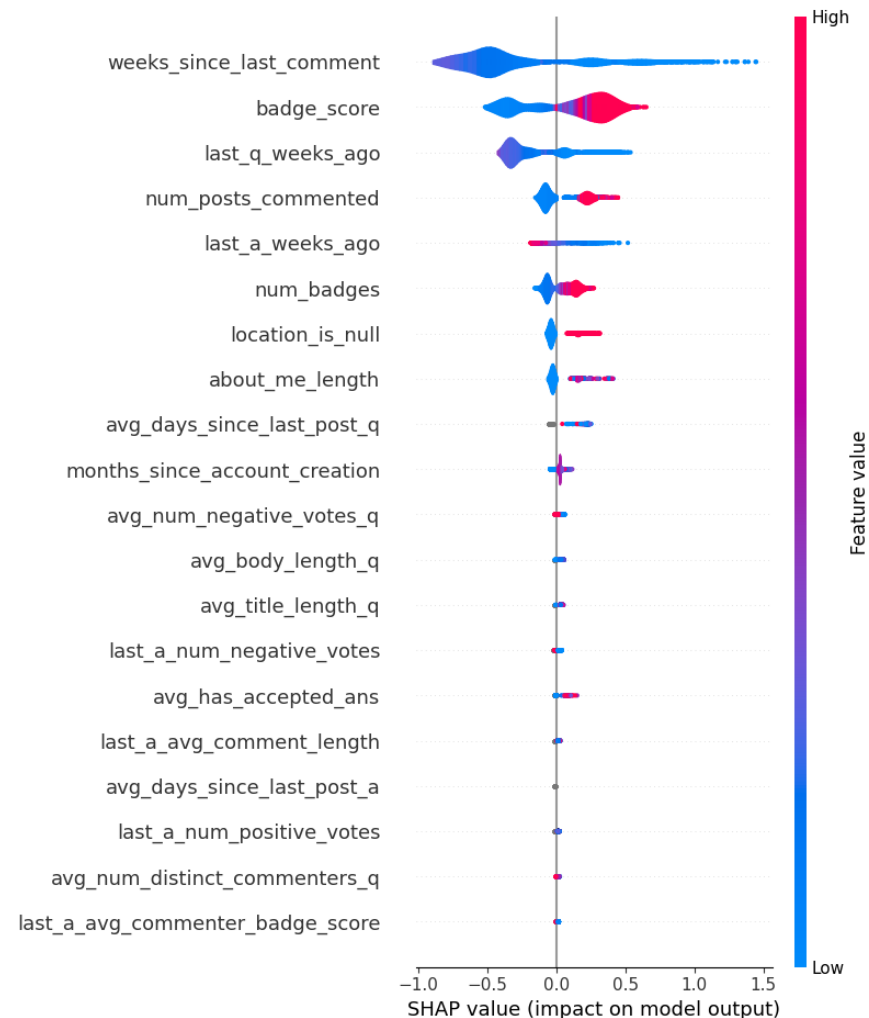
location_is_null



SHAP feature importance analysis

Selected Observations

- Website “seniority” **predictive** (total number of comments, badge score etc.)
- Time since last active / commented **is predictive**
- Completed bio (about me, location etc.) **is predictive**
- Number of positive/negative votes **not predictive**
- Interacting with “senior” community members **not predictive**



Final list of features

~12 hours of high-quality expert work

221:

	Label Corr.	Label MI	NaN %
num_badges	0.229	0.057	0.0%
num_questions_last_6mo	0.209	0.067	89.7%
badge_score	0.206	0.069	0.0%
ans_acceptance_rate	0.195	0.094	90.4%
avg_comment_length	0.135	0.045	0.0%
num_posts_commented	0.131	0.061	0.0%
avg_num_tags	0.131	0.077	89.7%
avg_has_accepted_ans	0.130	0.106	89.7%
num_comments	0.130	0.057	0.0%
num_answers_last_6mo	0.125	0.037	90.4%
last_q_num_tags	0.111	0.014	13.6%
avg_num_positive_votes_a	0.108	0.070	90.4%
last_a_is_accepted_ans	0.106	0.049	73.3%
last_q_body_length	0.096	0.007	13.6%
about_me_length	0.096	0.010	0.0%
avg_num_comments_a	0.084	0.088	90.4%
last_q_has_accepted_ans	0.076	0.009	13.6%
avg_num_distinct_commenters_a	0.073	0.073	90.4%
avg_body_length_q	0.067	0.051	89.7%
avg_num_positive_votes_q	0.060	0.078	89.7%
last_a_body_length	0.056	0.023	73.3%
avg_body_length_a	0.034	0.033	90.4%
last_q_title_length	0.023	0.004	13.6%
last_a_num_comments	0.023	0.040	73.3%
last_q_num_comments	0.015	0.003	13.6%
avg_avg_comment_length_a	0.012	0.047	90.4%
last_q_avg_comment_length	0.009	0.001	13.6%
display_name_is_null	0.004	0.001	0.0%
last_a_num_distinct_commenters	0.002	0.057	73.3%
last_q_num_distinct_commenters	-0.002	0.006	13.6%
last_q_num_positive_votes	-0.006	0.002	13.6%
last_a_num_positive_votes	-0.009	0.030	73.3%
last_a_avg_comment_length	-0.010	0.022	73.3%
avg_num_comments_q	-0.011	0.078	89.7%
avg_commenter_badge_score_q	-0.015	0.048	89.7%
avg_title_length_q	-0.022	0.064	89.7%
last_q_num_negative_votes	-0.026	0.002	13.6%
avg_num_distinct_commenters_q	-0.027	0.096	89.7%
avg_avg_comment_length_q	-0.043	0.057	89.7%
last_q_avg_commenter_badge_score	-0.046	0.006	13.6%
last_a_avg_commenter_badge_score	-0.048	0.023	73.3%
avg_commenter_badge_score_a	-0.060	0.032	90.4%
website_url_is_null	-0.061	0.032	0.0%
last_a_num_negative_votes	-0.065	0.051	73.3%
months_since_account_creation	-0.113	0.011	0.0%
avg_num_negative_votes_q	-0.114	0.100	89.7%
location_is_null	-0.136	0.051	0.0%
avg_days_since_last_post_a	-0.168	0.035	97.1%
avg_days_since_last_post_q	-0.169	0.055	93.5%
avg_num_negative_votes_a	-0.170	0.088	90.4%
last_q_weeks_ago	-0.346	0.073	13.6%
weeks_since_last_comment	-0.392	0.091	34.3%
last_a_weeks_ago	-0.394	0.049	73.3%
last_a_num_tags	nan	0.065	73.3%

Stanford CS224W: Results

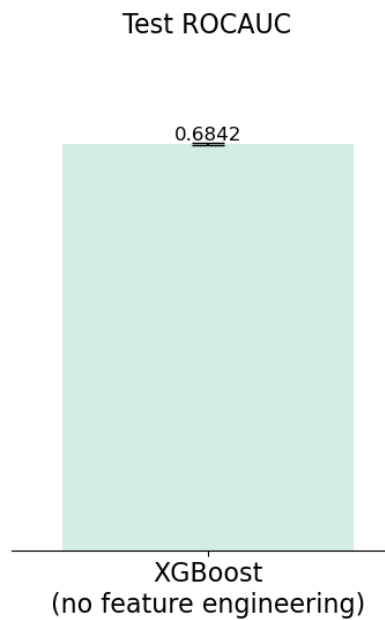
CS224W: Machine Learning with Graphs
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



Results

Task: Will a user be active in the next 6 months?

Naive baseline



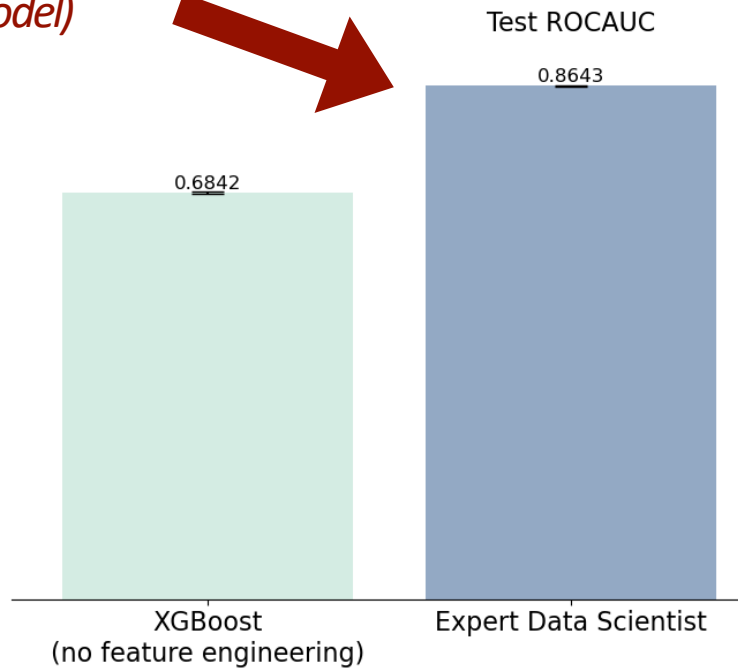
users	
Id 	numerical
AccountId	numerical
DisplayName	text
Location	text
WebsiteUrl	text
AboutMe	text
CreationDate	timestamp

Results

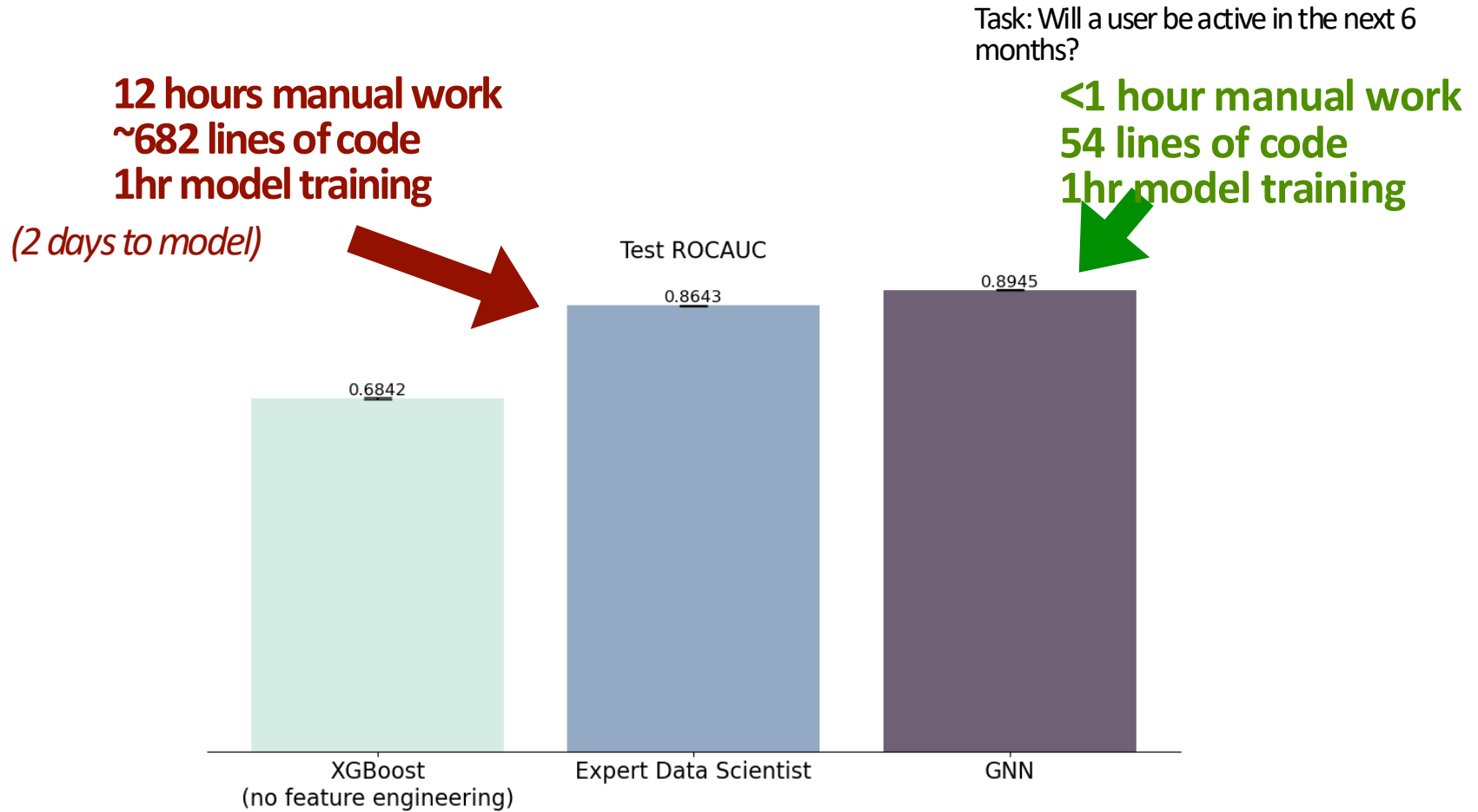
12 hours manual work
~682 lines of code
1hr model training
(2 days to model)



**Work measured as the marginal effort to solve a new task*



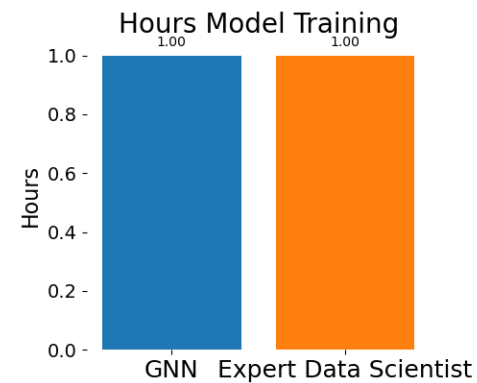
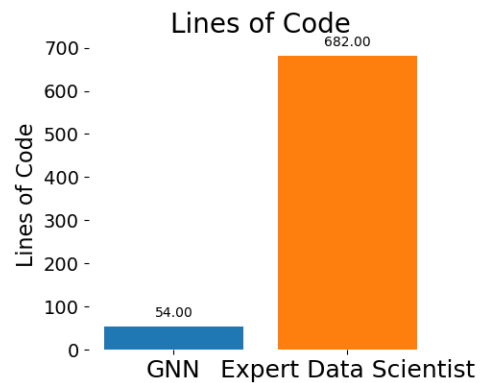
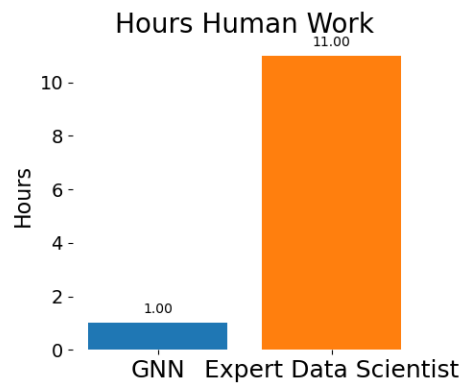
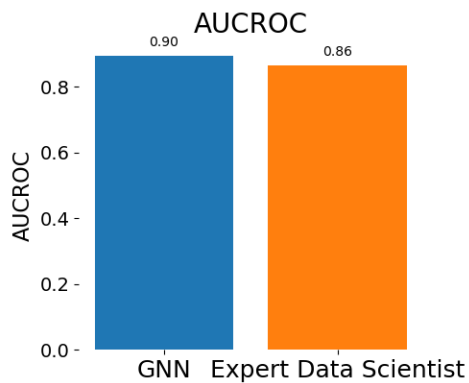
Results



Results

Task: Will a user be active in the next 6 months?

Performance Comparison



More Relbench tasks

Performance Comparison

