CS 295B/CS 395B Systems for Knowledge Discovery

Toward Sound Data Collection

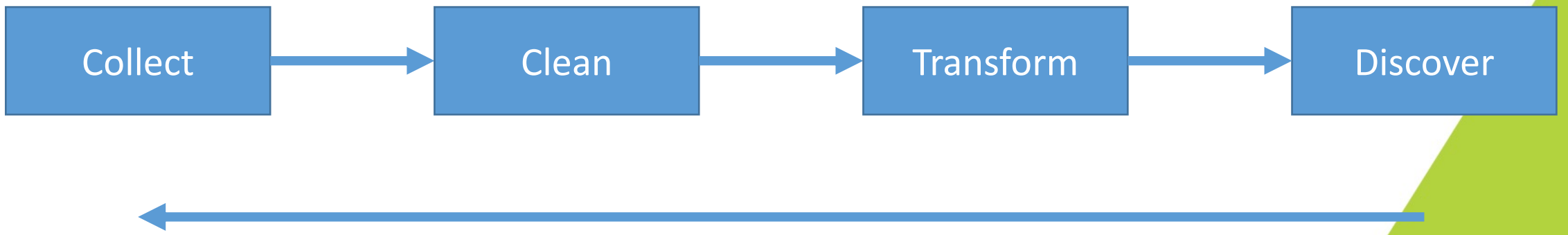
The University of Vermont

Topics for Today

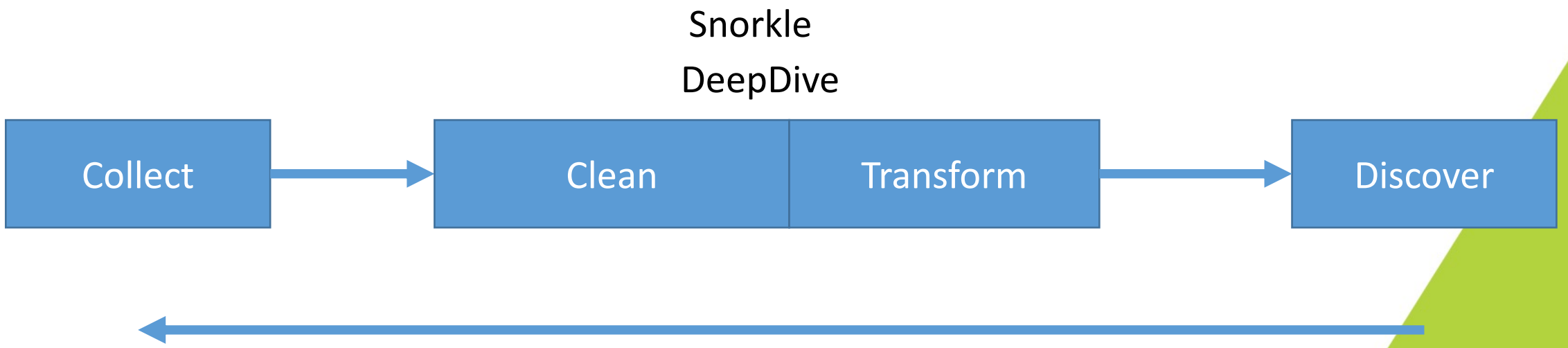
- ~~Parisa's presentation~~
- Contextualizing Friday's papers
- Background for Friday's papers
- New method: ablation studies
- Things to reflect on for this week's reviews and presentations

Context

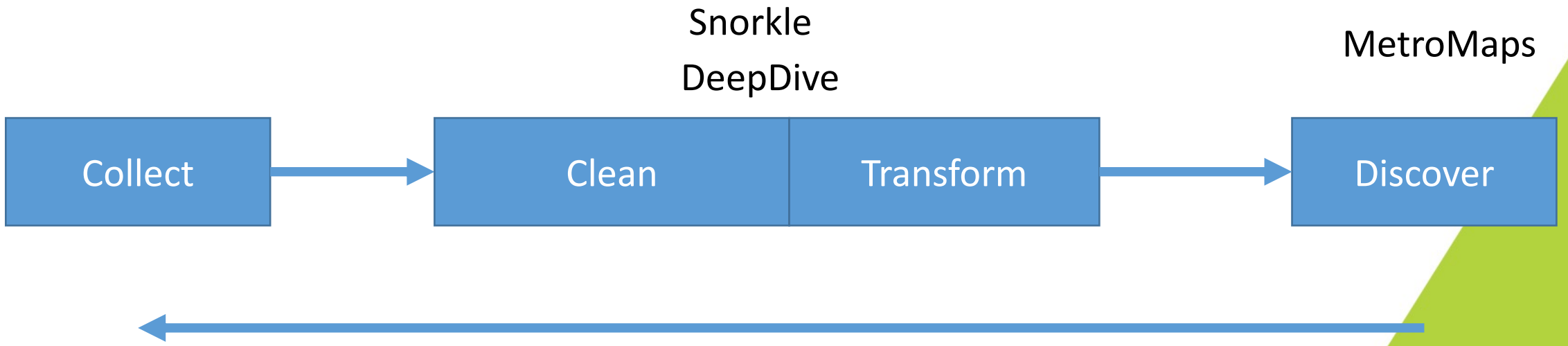
Remember this diagram…



Remember this diagram…



Remember this diagram…



Home Vision Slides People Members Projects Papers Seminar Blog

DAWN is a five-year research project to democratize AI by making it dramatically easier to build AI-powered applications.

Our past research–from Apache Spark to Mesos, DeepDive, and HogWild!–already powers major functionality all over Silicon Valley and the world. Between fighting against human trafficking, assisting in cancer diagnosis and performing high-throughput genome sequencing, we've invested heavily in tools for AI and data product development.

The next step is to make these tools more efficient and more accessible, from training set creation and model design to monitoring, efficient execution, and hardware-efficient implementation. This technology holds the power to change science and society–and we're creating this change with partners throughout campus and beyond.

We're proud to be supported by the following founding members:

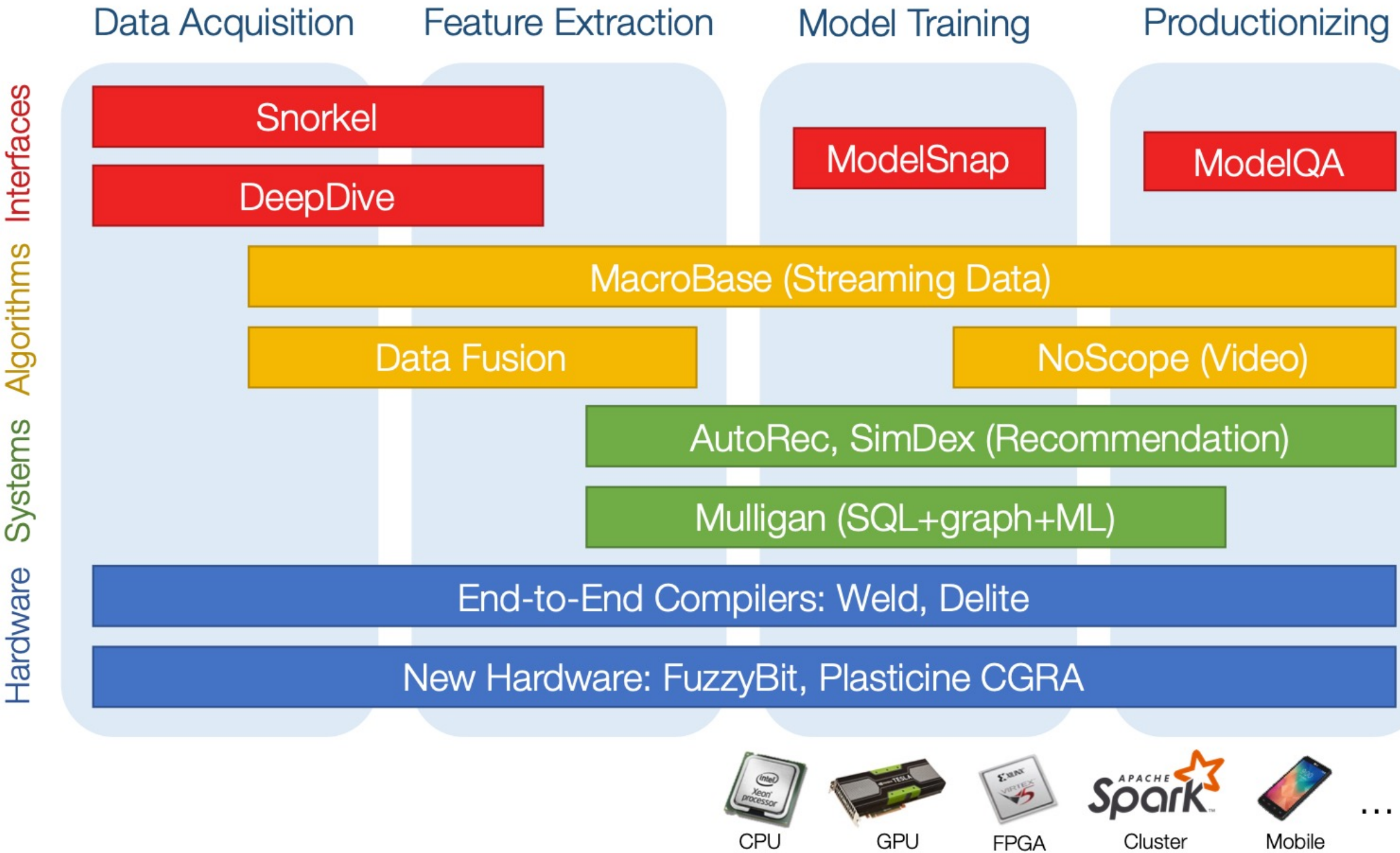
Background: DAWN project

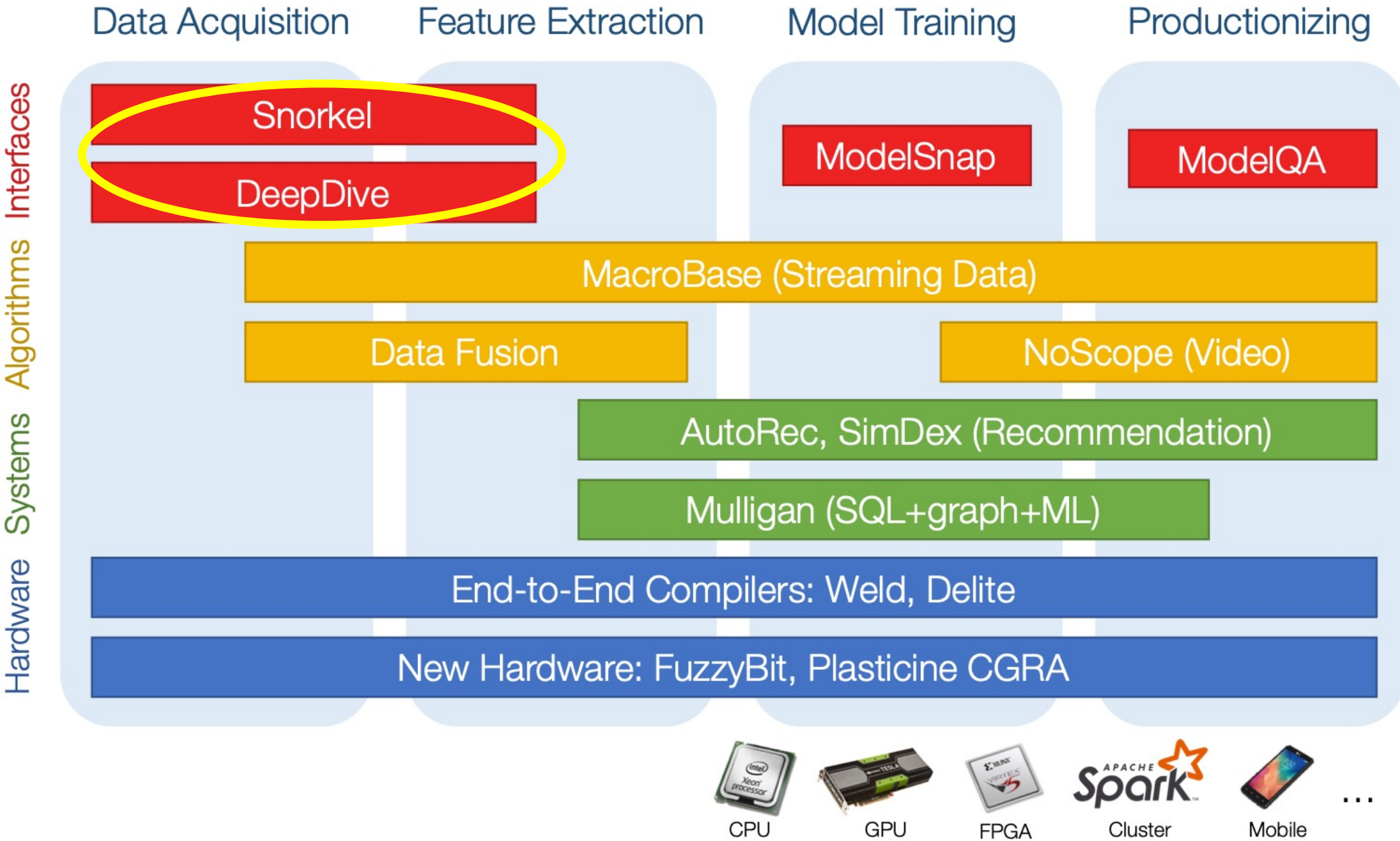
Goal: "To empower domain experts who are not ML experts"

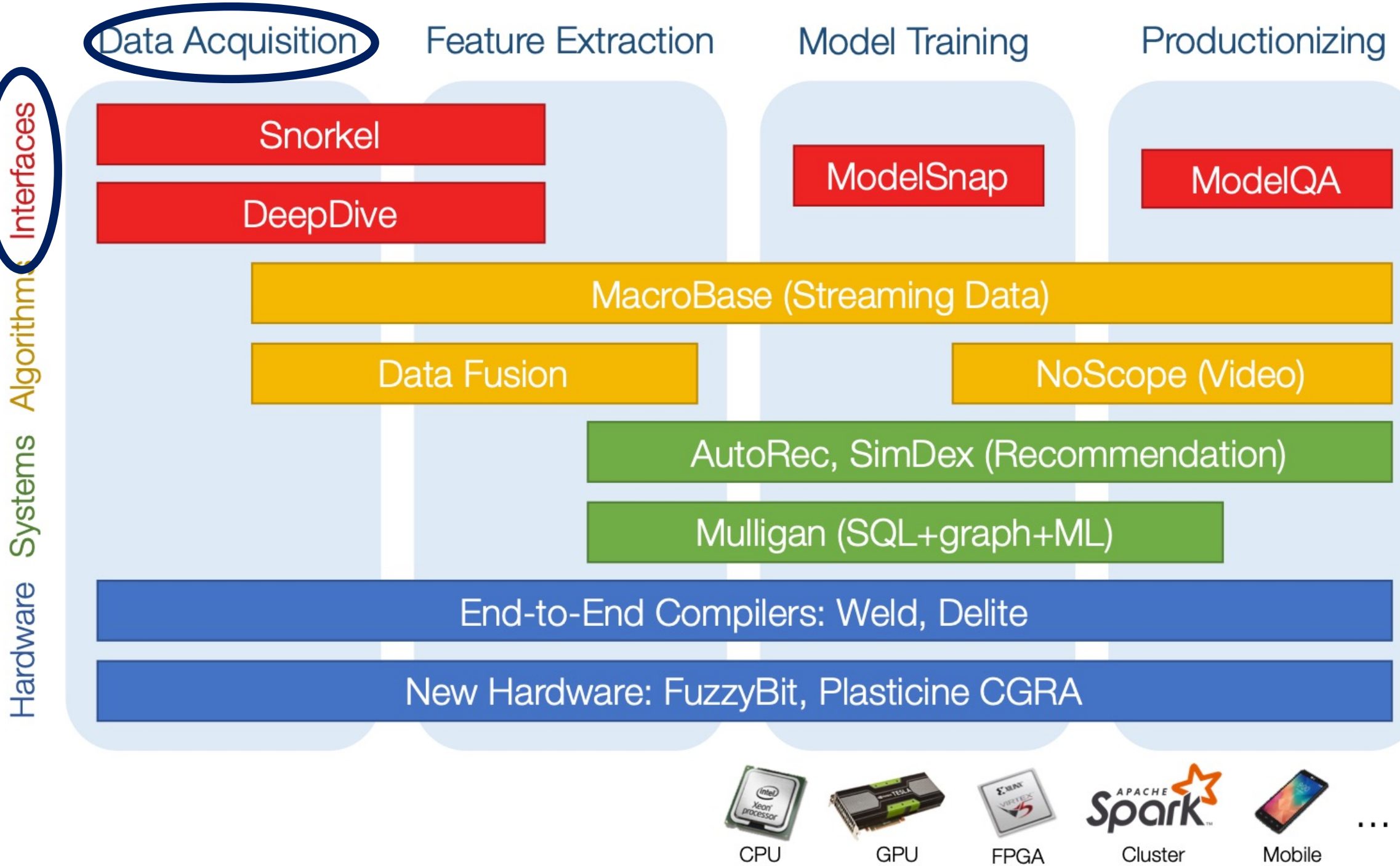
Achieved through many initiatives

Our focus: **Developing new interfaces that**

a. Make model specification easier
b. Explain results to humans
c. Make debugging easier
d. Make improving data quality easier

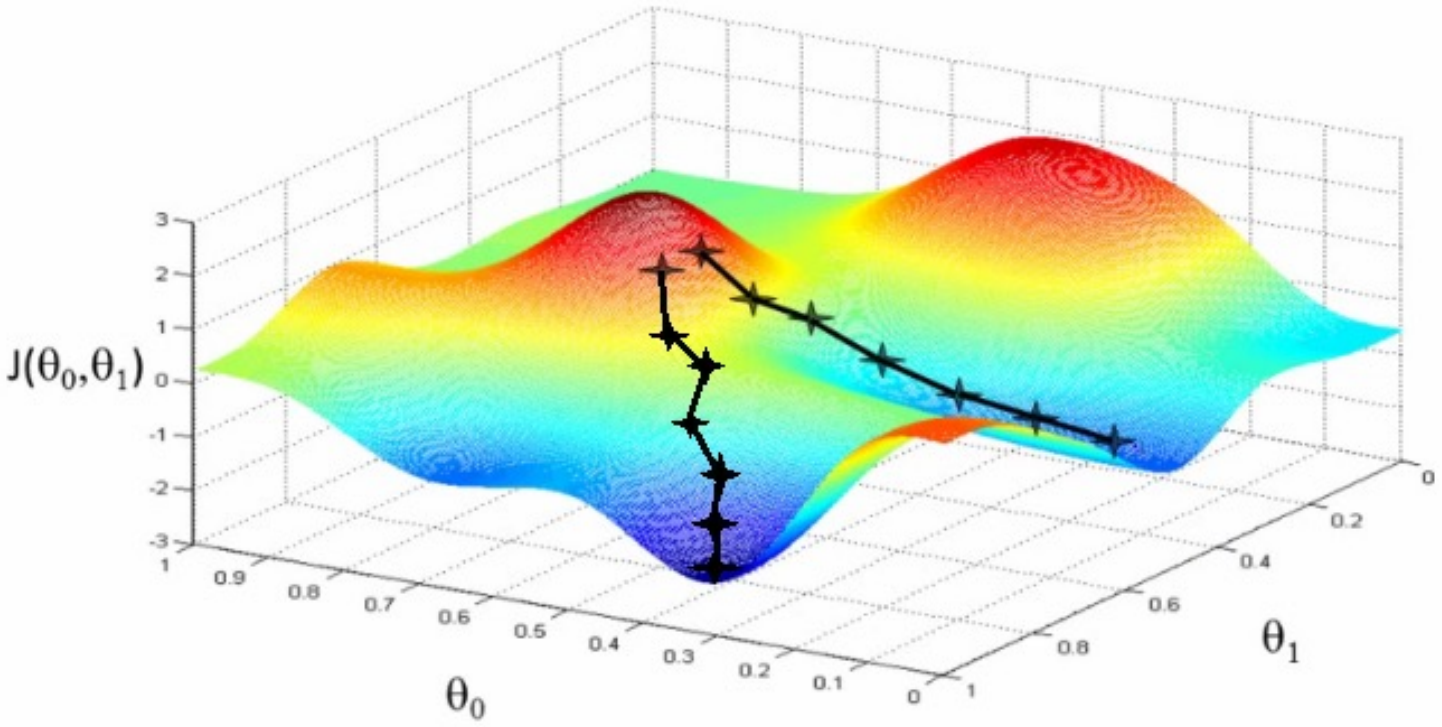

Data Acquisition
Feature Extraction
Model Training
Productionizing
Interfaces
Algorithms
Systems
Hardware
Snorkel
DeepDive
ModelSnap
ModelQA
MacroBase (Streaming Data)
Data Fusion
NoScope (Video)
AutoRec, SimDex (Recommendation)
Mulligan (SQL+graph+ML)
End-to-End Compilers: Weld, Delite
New Hardware: FuzzyBit, Plasticine CGRA
CPU
GPU
FPGA
Cluster
Mobile
...


Data Acquisition
Feature Extraction
Model Training
Productionizing
Interfaces
Algorithms
Systems
Hardware
Snorkel
DeepDive
ModelSnap
ModelQA
MacroBase (Streaming Data)
Data Fusion
NoScope (Video)
AutoRec, SimDex (Recommendation)
Mulligan (SQL+graph+ML)
End-to-End Compilers: Weld, Delite
New Hardware: FuzzyBit, Plasticine CGRA
CPU
GPU
FPGA
Cluster
Mobile
...


Data Acquisition
Feature Extraction
Model Training
Productionizing
Interfaces
Algorithms
Systems
Hardware
Snorkel
DeepDive
ModelSnap
ModelQA
MacroBase (Streaming Data)
Data Fusion
NoScope (Video)
AutoRec, SimDex (Recommendation)
Mulligan (SQL+graph+ML)
End-to-End Compilers: Weld, Delite
New Hardware: FuzzyBit, Plasticine CGRA
CPU
GPU
FPGA
Cluster
Mobile
...

Questions?

SGD and infinite data



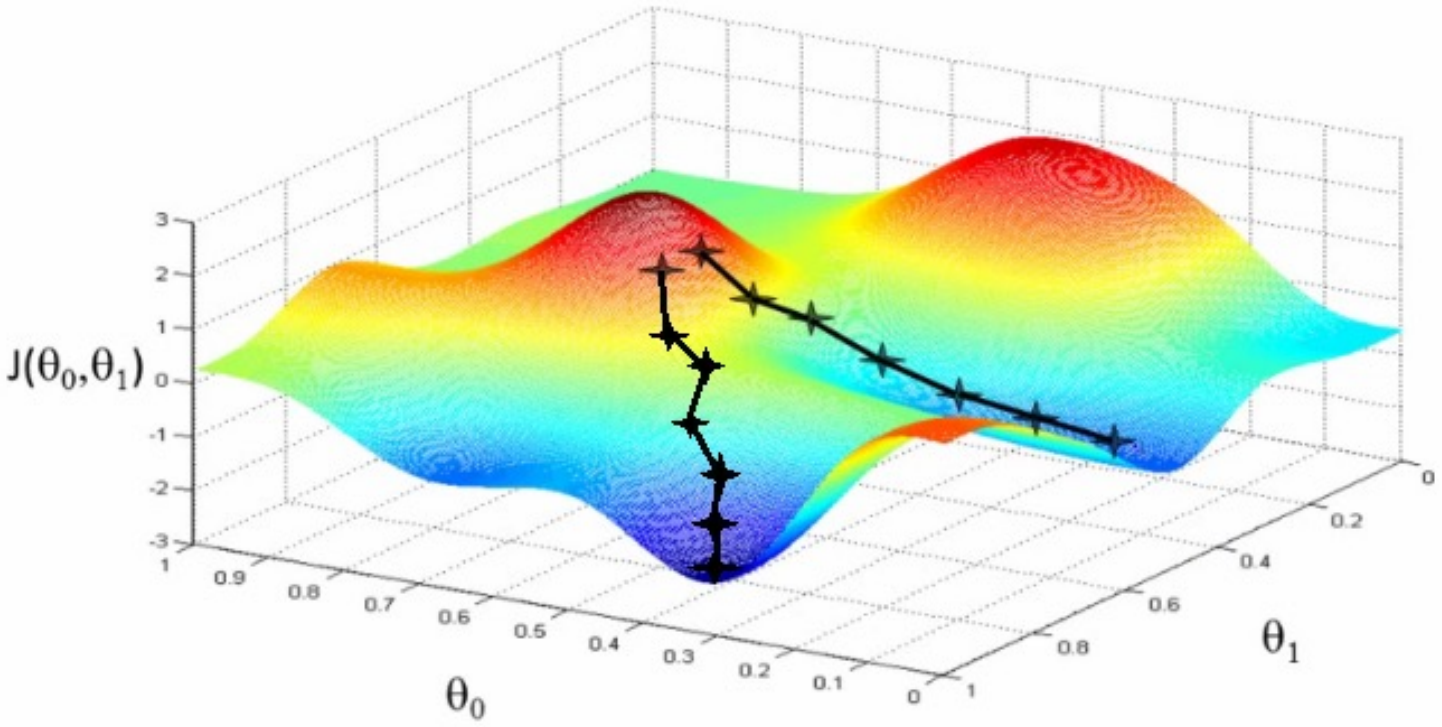
Stochastic randomly

Gradient follow the derivative

Descent toward the minimum

See Andrew Ng's videos for details

SGD and infinite data

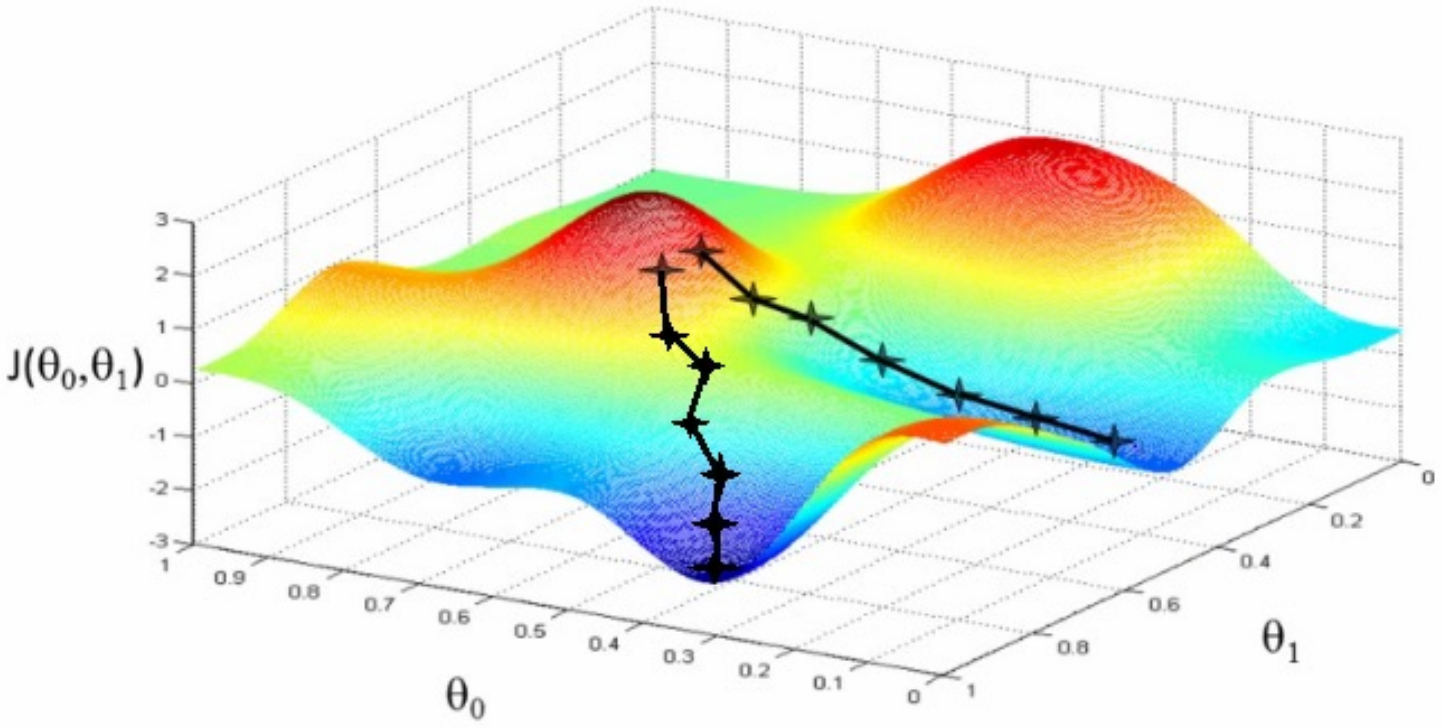


SGD powers all modern machine learning…

… especially deep learning …

Idea: more data is better

SGD and infinite data



But how to get more data?

Better question:

How to get more quality data?

Gold standard:

Quality *labelled* data

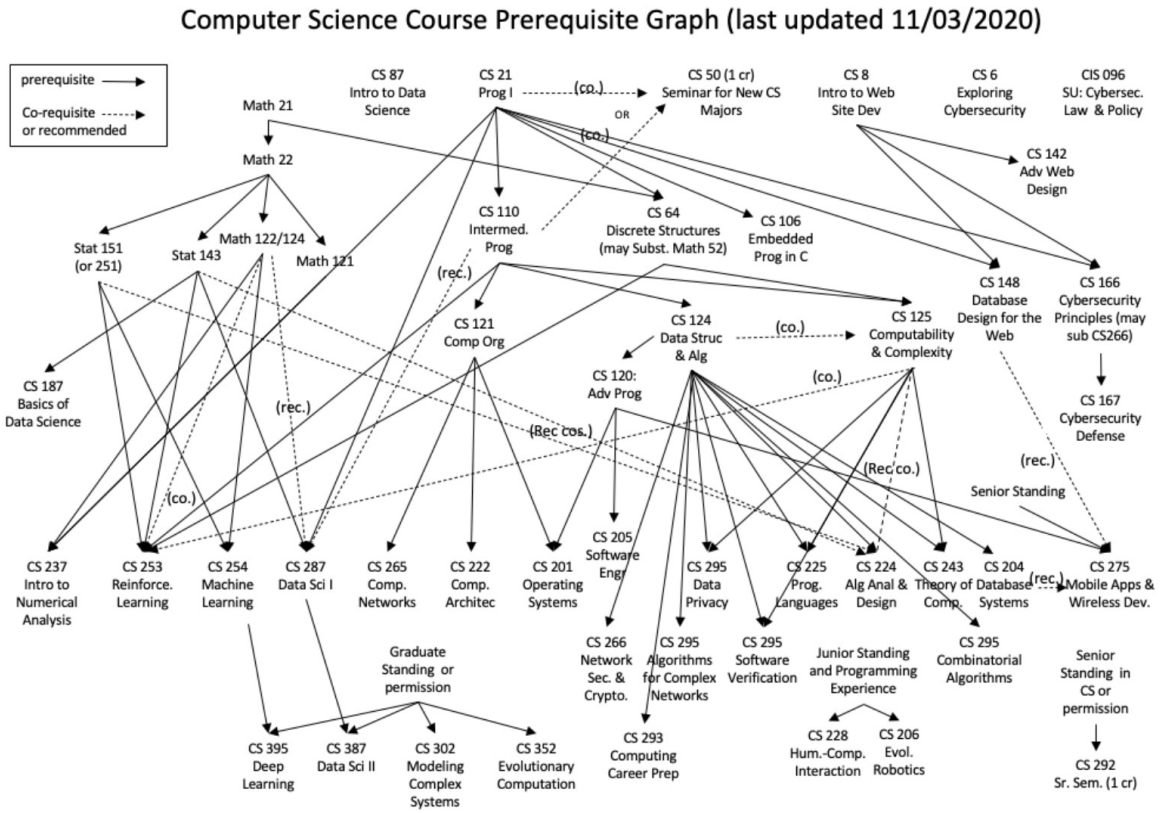
- Expensive
- DeepDive + Snorkle

Questions?

Datalog

Snorkle/DeepDive interfaces include Datalog

- A domain-specific language (DSL) for *logic programming*
- Declaratively write out facts and relations
- Can then query this set of rules



Datalog Example

This is the CEMS CS major pre-req chart

(It is out of date)

We can express with datalog:

```
Required("CS50")

Prereq("CS121", "CS222")

IsNot(Standing(student), "graduate") :-
HasPermission(student, "CS352")

?- CanTake(student, "CS204")

?- CanTake(student)
```

Questions?

Some prior work

Connecting the Dots Between News Articles

Dafna Shahaf
Carnegie Mellon University
dshahaf@cs.cmu.edu

Carlos Guestrin
Carnegie Mellon University
guestrin@cs.cmu.edu

ABSTRACT

The process of extracting useful knowledge from large datasets has become one of the most pressing problems in today's society. The problem spans entire sectors, from scientists to intelligence analysts and web users, all of whom are constantly struggling to keep up with the larger and larger amounts of content published every day. With this much data, it is often easy to miss the big picture.

In this paper, we investigate methods for automatically connecting the dots – providing a structured, easy way to navigate within a new topic and discover hidden connections. We focus on the news domain: given two news articles, our system automatically finds a coherent chain linking them together. For example, it can recover the chain of events starting with the decline of home prices (January 2007), and ending with the ongoing health-care debate.

We formalize the characteristics of a good chain and provide an efficient algorithm (with theoretical guarantees) to connect two fixed endpoints. We incorporate user feedback into our framework, allowing the stories to be refined and personalized. Finally, we evaluate our algorithm over real news data. Our user studies demonstrate the algorithm's effectiveness in helping users understanding the news.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning; G.3 [**Probability and Statistics**]

General Terms: Algorithms, Experimentation

1. INTRODUCTION

"Can't Grasp Credit Crisis? Join the Club", stated David Leonhardt's article in the New York Times. Credit crisis had been going on for seven months by that time, and had been extensively covered by every major media outlet throughout the world. Yet many people felt as if they did not understand what it was about.

Paradoxically, the extensive media coverage might have been a part of the problem. This is another instance of the *information overload* problem, long recognized in the computing industry. Users are constantly struggling to keep up with the larger and larger amounts of content that is being published every day; with this much data, it is often easy to miss the big picture.

For this reason, there is an increasing need for techniques to present data in a meaningful and effective manner. In this paper, we investigate methods for automatically *connecting the dots* – providing a structured, easy way to uncover hidden connections between two pieces of information. We believe that the ability to connect dots and form a logical, coherent story lies at the basis of understanding a topic.

We focus on the news domain: given two news articles, our system automatically finds a coherent *story* (chain of articles) linking them together. For example, imagine a user who is interested in the financial crisis and its effect on the health-care reform. The user vaguely recalls that the financial crisis is related to the decline of home prices in 2007. The user would then choose representative articles for those two topics and feed them to our system. An output chain may look like this (parenthesized text not part of output):

1.3.07 **Home Prices Fall** Just a Bit
3.4.07 **Keeping Borrowers Afloat**
(Increasing delinquent mortgages)
3.5.07 **A Mortgage Crisis** Begins to Spiral, ...
8.10.07 **... Investors Grow Wary** of Bank's Reliance on Debt.
(Banks' equity diminishes)
9.26.08 **Markets Can't Wait for Congress** to Act
10.4.08 **Bailout Plan** Wins Approval
1.20.09 **Obama's Bailout Plan Moving Forward**
(... and its effect on health benefits)
9.1.09 Do Bank Bailouts **Hurt Obama on Health?**
(Bailout handling can undermine health-care reform)
9.22.09 Yes to Health-Care Reform, but Is This the Right Plan?

The chain mentions some of the key events connecting the mortgage crisis to healthcare, including the bailout plan. Most importantly, the chain should be *coherent*: after reading it, the user should gain a better understanding of the progression of the story.

To the best of our knowledge, the problem of connecting the dots is novel. Previous research (e.g., [19, 13, 18, 17, 4, 6]) focused on organizing news articles into hierarchies or graphs, but did not address the notion of output coherence.

Our main contributions are

- Formalizing characteristics of a good story and the notion of *coherence*.
- Formalizing *influence* with no link structure.
- Providing an efficient algorithm for connecting two fixed endpoints while maximizing chain coherence (with theoretical guarantees).

KDD'10, July 25–28, 2010, Washington, DC, USA.

Metro Maps of Science

Dafna Shahaf
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
dshahaf@cs.cmu.edu

Carlos Guestrin
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
guestrin@cs.cmu.edu

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA
horvitz@microsoft.com

ABSTRACT

As the number of scientific publications soars, even the most enthusiastic reader can have trouble staying on top of the evolving literature. It is easy to focus on a narrow aspect of one's field and lose track of the big picture. Information overload is indeed a major challenge for scientists today, and is especially daunting for new investigators attempting to master a discipline and scientists who seek to cross disciplinary borders. In this paper, we propose metrics of influence, coverage, and connectivity for scientific literature. We use these metrics to create structured summaries of information, which we call *metro maps*. Most importantly, metro maps explicitly show the relations between papers in a way which captures developments in the field. Pilot user studies demonstrate that our method can help researchers acquire new knowledge efficiently: map users achieved better precision and recall scores and found more seminal papers while performing fewer searches.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5 [**Information Interfaces and Presentation**]

Keywords

Metro maps, Information, Summarization

1. INTRODUCTION

"Distringit librorum multitudo" (the abundance of books is a distraction), said Lucius Annaeus Seneca; he lived in the first century.

A lot has changed since the first century, but Luciusʼ problem has only become worse. The surge of the Web brought down the barriers of distribution, and the scientific community finds itself overwhelmed by the increasing numbers of publications; relevant data is often buried in an avalanche of publications, and locating it is difficult.

Search engines have been relied upon in recent years accessing the scientific literature, and investments have been made to create special academic search and retri tools. However, the search and browsing experience m be best characterized as providing keyhole views onto literature: while search engines are highly effective i trieving scientific publications, the task of fitting those lications into a coherent picture remains difficult.

In contrast, we are interested in methods that *exp* show the relationships among publications in a wa captures the main developments in the discipline. V lieve that such methods can allow a user to explore complex topic and discover hidden connections effe We consider as a sample motivation the creation of v literature exploration tools that could help people en new field, such as new graduate students or experts beyond their traditional disciplinary borders.

Several tools already exist for summarizing and vi scientific literature (see [Borner, 2010] for a comp However, the output of these systems is often not su a starting researcher. Some systemsʼ level of grar too coarse: Boyack et al. [2009] provide a graph-su chemistry research, where each node corresponds ter of disciplines ('Biology-Zoology-Ecology'). Ba and Zitt [1999] produce a hierarchical graph, w correspond to clusters of journals.

We believe that in order to allow researcher stand how a field is organized, a finer level of is needed. For this reason, we chose *papers* as analysis. Most current tools that work at this le larity provide visualizations of citation (or co-c works, where papers are nodes [Chen, 2004; I 2010]. Importantly, edges between papers are *cal* computation: the edges are selected beca some threshold, or belong to a spanning tree. ods, there is no notion of coherent *lines of* believe that the notion of story lines is essent tates users' knowledge acquisition and compr frontier and evolutionary history of ideas in

Several systems have attempted to create pecially in the news domain [Swan and Jer et al., 2011; Allan et al., 2001]. However, t marization only works for simple stories, w nature. In contrast, research fields display behaviour: lines of research branch like a ta with side stories, dead ends, and intertwini order to explore these stories, one needs through unfamiliar territory.

The metro map metaphor has been used play abstract knowledge. For example, Nesbitt's map shows

KDD'12, August 12–16, 2012, Beijing, China.

Trains of Thought: Generating Information Maps

Dafna Shahaf
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
dshahaf@cs.cmu.edu

Carlos Guestrin
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
guestrin@cs.cmu.edu

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA
horvitz@microsoft.com

ABSTRACT

When information is abundant, it becomes increasingly difficult to fit nuggets of knowledge into a sigle coherent picture. Complex stories spaghetti into branches, side stories, and intertwining narratives. In order to explore these stories, one needs a map to navigate unfamiliar territory. We propose a methodology for creating structured summaries of information, which we call *metro maps*. Our proposed algorithm generates a concise structured set of documents which maximizes coverage of salient pieces of information. Most importantly, metro maps explicitly show the relations among retrieved pieces in a way that captures story development. We first formalize characteristics of good maps and formulate their construction as an optimization problem. Then we provide efficient methods with theoretical guarantees for generating maps. Finally, we integrate user interaction into our framework, allowing users to alter the maps to better reflect their interests. Pilot user studies with a real-world dataset demonstrate that the method is able to produce maps which help users acquire knowledge efficiently.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5 [**Information Interfaces and Presentation**]

Keywords

Metro maps, Information, Summarization

1. INTRODUCTION

As data becomes increasingly ubiquitous, users are often overwhelmed by the flood of information available to them. Although search engines are effective in retrieving nuggets of knowledge, the task of fitting those nuggets into a single coherent picture remains difficult.

We are interested in methods for building more comprehensive views that *explicitly* show the relations among retrieved nuggets. We believe that such methods can enable people to navigate new, complex topics and discover previously unknown links. We shall focus on the news domain; for example, the system described in this paper can be used by a person who wishes to understand the debt crisis in Europe and its implications.

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

Previous news summarization systems with structured output [17, 18, 2] have focused mostly on *timeline* generation. However, this style of summarization only works for simple stories, which are linear in nature. In contrast, complex stories display a very non-linear structure: stories spaghetti into branches, side stories, dead ends, and intertwining narratives. To explore these stories, one needs a *map* to guide them through unfamiliar territory.

In this paper, we investigate methods for automatically creating *metro maps* of information. Metro maps are concise structured sets of documents maximizing coverage of salient pieces of information; in addition, the maps make explicit the various ways each piece relates to the others. Due to the sparsity of the output, it naturally lends itself to many visualization techniques. We chose to follow the metro-map metaphor: a metro map consists of a set of lines which have intersections or overlaps. Each line follows a coherent narrative thread; different lines focus on different aspects of the story. This visualization allows users to easily digest information at a holistic level, and also to interact with the model and make modifications.

Figure 1 shows a simplified metro map representing the debt crisis in Greece. The middle (blue) line details the chain of events leading from Greece's debt 'junk' status to the Greek bailout. The L-shaped (red) line is about strikes and riots in Greece. Both lines intersect at an article about the austerity plan, since it plays an important role in both storylines: it was a key precondition for Greece to get bailout money, but it also triggered many of the strikes.

To the best of our knowledge, the problem of constructing metro maps is novel. We believe that metro maps can serve as effective tools to help users cope with information overload in many fields. For example, maps can be a great vehicle for scientists exploring the research landscape. Our main contributions are as follows:

Figure 1: Greek debt crisis: a simplified metro map

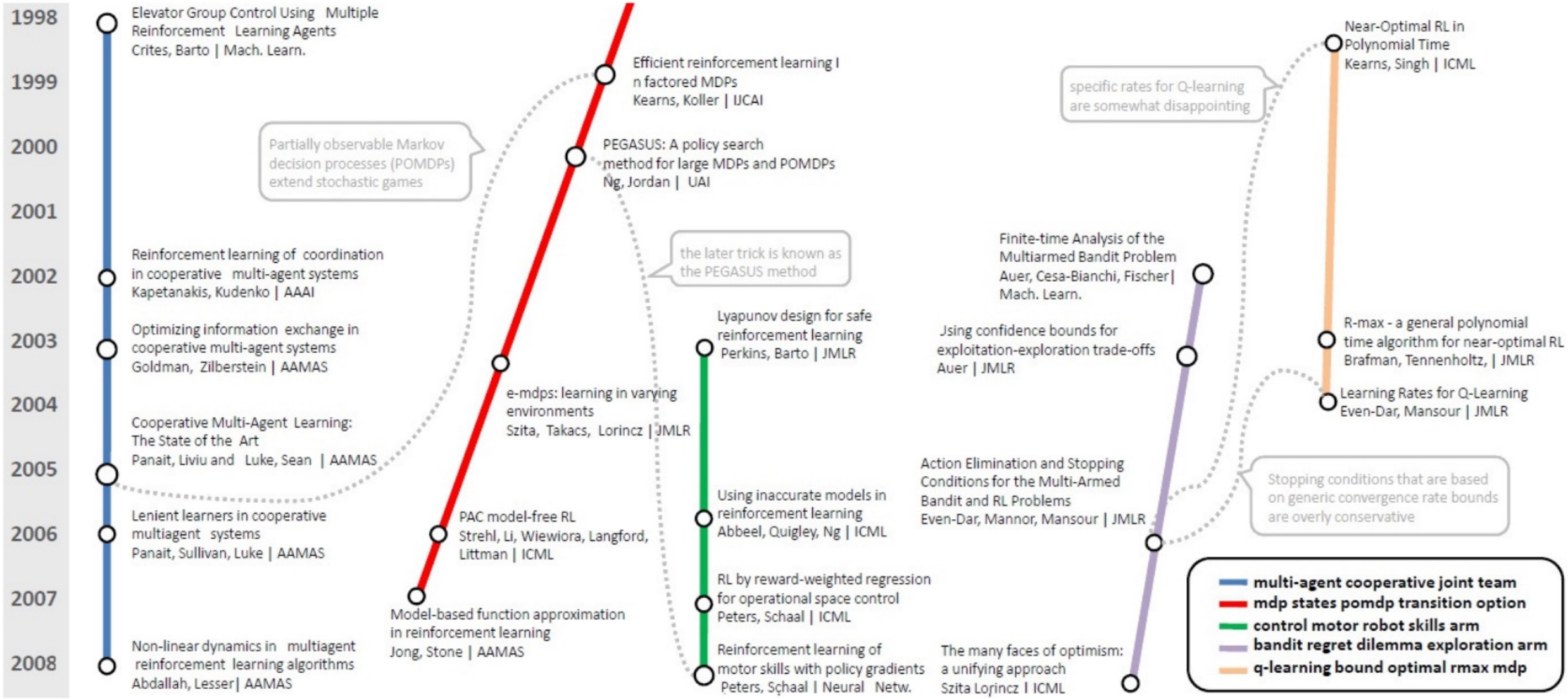


Figure 8: Part of the map computed for the query 'Reinforcement Learning'. The map depicts multiple lines of research (see legend at the bottom). Interactions between the lines are depicted as dashed gray lines, and relevant citation text appears near them.

Why bring this up?

- DAWN project's version of this is for business
    - For them, explanation/discovery == business analytics
- Exercise: think about what you could do with these tools for general knowledge
- MetroMaps an example of general knowledge extraction in this context
- Similar in effort to other methods, e.g. topic models
    - Big DARPA project (FUSE)
    - DAWN also influenced by a big DARPA project
    - Discuss/reflect: funding <-> project choices

Questions?

Topics for Today

- ~~Parisa's presentation~~
- ~~Contextualizing Friday's papers~~
- Background for Friday's papers
- New method: ablation studies
- Things to reflect on for this week's reviews and presentations

Background

Reminder: ETL

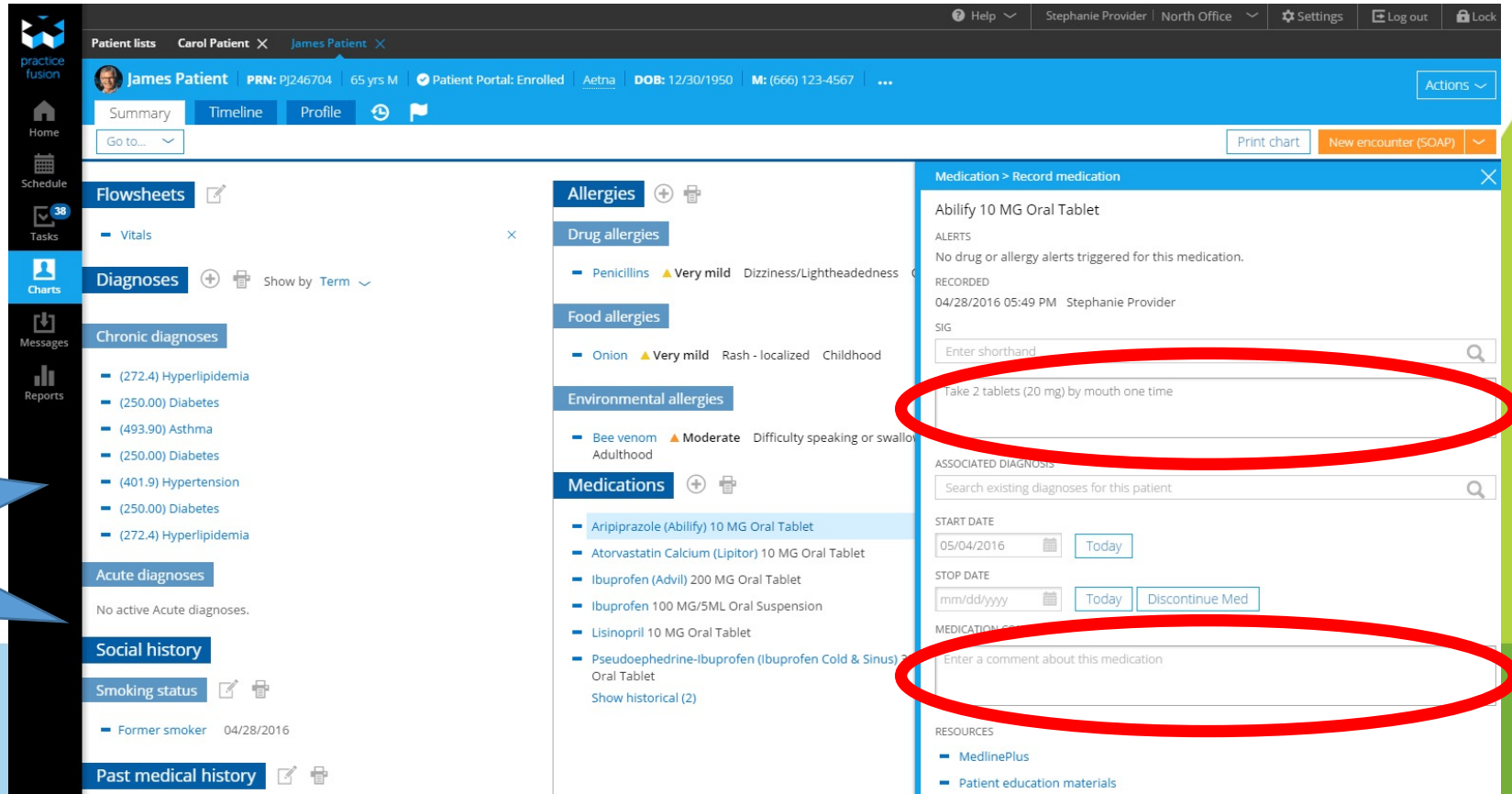
Snorkle/DeepDive as an alternative to traditional ETL

Extract

Transform

Load

Often partially loaded in DBs



Questions?

Generative vs. Discriminative Models

Suppose we want to predict class C from features X1, X2, and X3.

All classifiers learn P(C | X1, X2, X3)

Generative:

Learn the "whole story"

A generative model learns the probability distribution P(C, X1, X

P(C | X1, X2, X3) can be derived

Requires labelled data to get P(X1, X2, X3 | C) → Use Bayes' rule

Pros: few parameters, fast, robust if correct
Cons: lots of assumptions

Generative vs. Discriminative Models

Suppose we want to predict class C from features X1, X2, and X3.

All classifiers learn P(C | X1, X2, X3)

Discriminative:

Only learns what's necessary: P(C | X1, X2, X3)

Requires labelled data to minimize error

Pros: few assumptions, lower barrier-to-entry
Cons: may not generalize, data-hungry

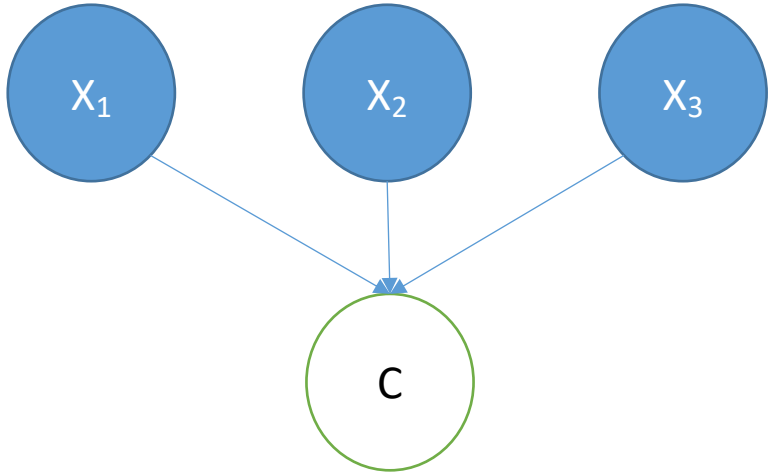
Examples of Generative vs. Discriminative Models

Generative	Discriminitive
Naïve Bayes Classifier	Classic Deep Neural Networks
Hidden Markov Models	Regression
Latent Dirichlet Allocation (a kind of topic model)	Conditional Random Fields
Gaussian Mixture Models	Random Forests

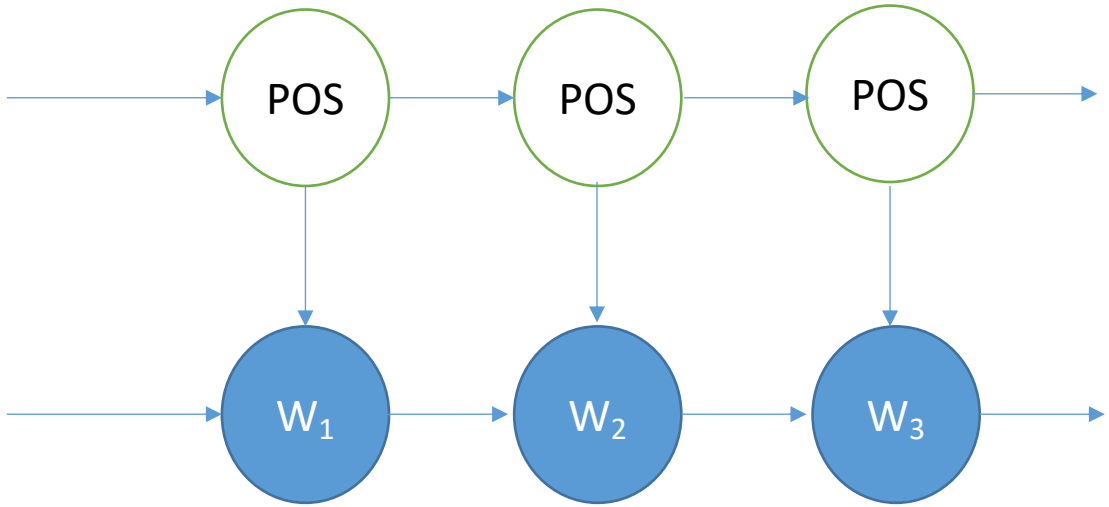
What is a latent variable?

Simple answer:

a variable whose values are not/cannot be observed



More interesting example



Questions?

Topics for Today

- ~~Parisa's presentation~~
- ~~Contextualizing Friday's papers~~
- ~~Background for Friday's papers~~
- New method: ablation studies
- Things to reflect on for this week's reviews and presentations

Ablation Studies

Ablation study: a method we haven't seen before!

We have not previously talked about ablation studies

Idea:

You've designed a solution that performs better. Yay!

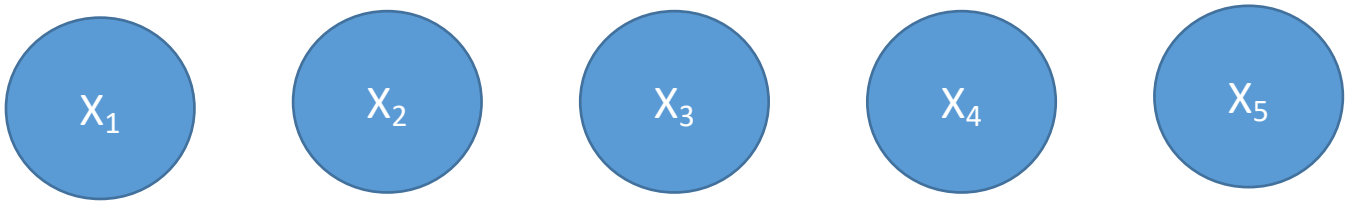
But that solution is *complex*. Which parts really contributed to the improved performance?

Run a kind of experiment (**control** is the complete solution; **treatment** is removing one part)

Very common in complex machine learning, especially deep learning.

Ablation study: what it can tell us

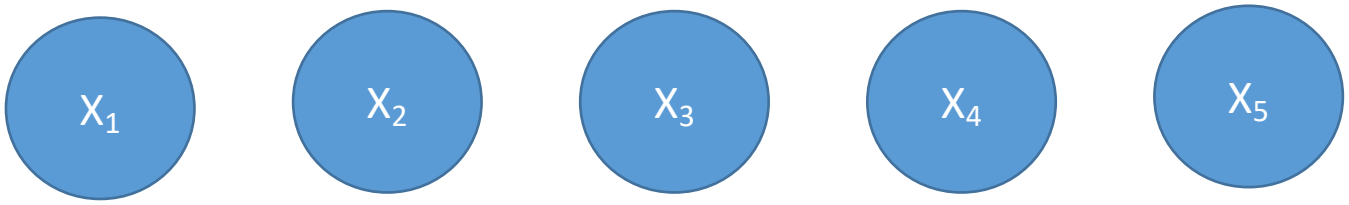
Only tests whether a feature **independently** does not contribute to the final performance



$$P_{solution} > P_{SOTA}$$

Ablation study: what it can tell us

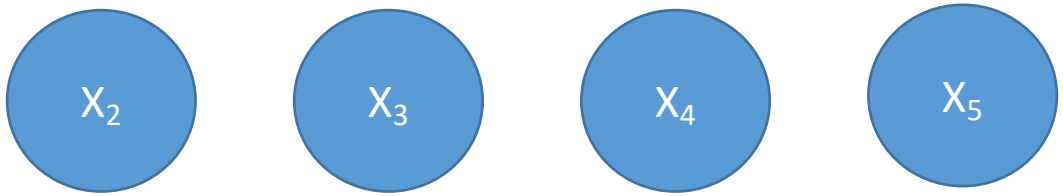
Only tests whether a feature **independently** does not contribute to the final performance



Remove or replace with an appropriate substitute

Ablation study: what it can tell us

Only tests whether a feature **independently** does not contribute to the final performance



X_1 had no effect

$$P_{solution/x1} \approx P_{solution}$$

Ablation study: what it can tell us

Only tests whether a feature or **hidden unit** does not contribute to the final performance

X_1 X_2 X_3 X_4 X_5

Continue for X_3, X_4, X_5

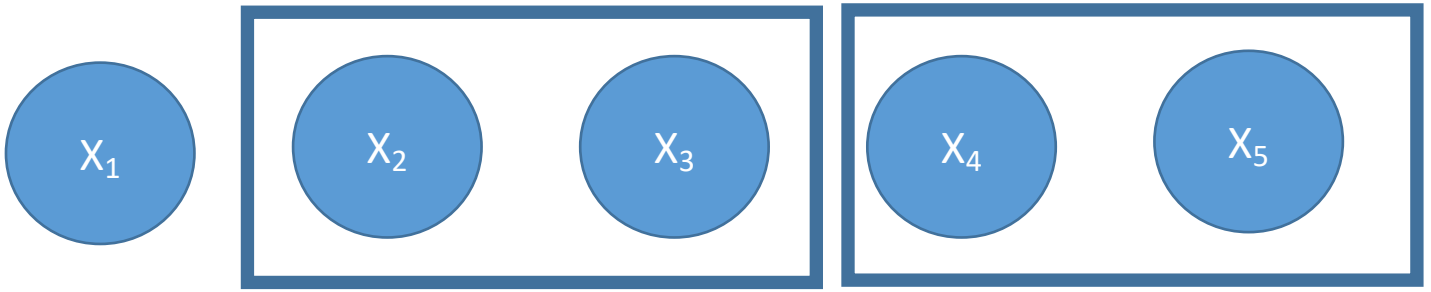
Assume you find all but X_1 had an effect.

$P_{solution/x2} < P_{solution}$

X_2 had an effect!

Ablation study: what it can tell us

Only tests whether a feature **independently** does not contribute to the final performance

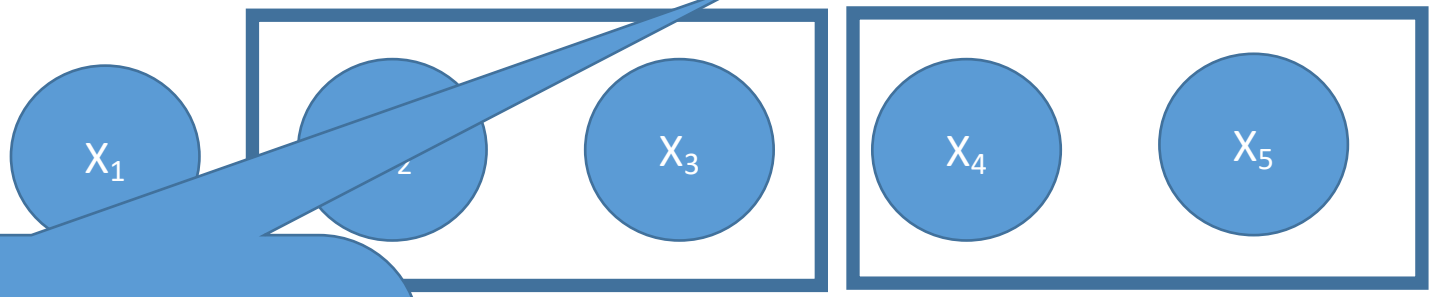


But what if certain parts/features only have an effect when they are together?

Can’t detect this!

Ablation study: what it can tell us

Only tests whether a feature **independently** does not contribute to the final performance



This is what we mean by "independently does not contribute"

in parts/features only have an effect when they are together?

Can't detect this!

Ablation study: example

Domain: classifying Wikipedia pages as literary or not

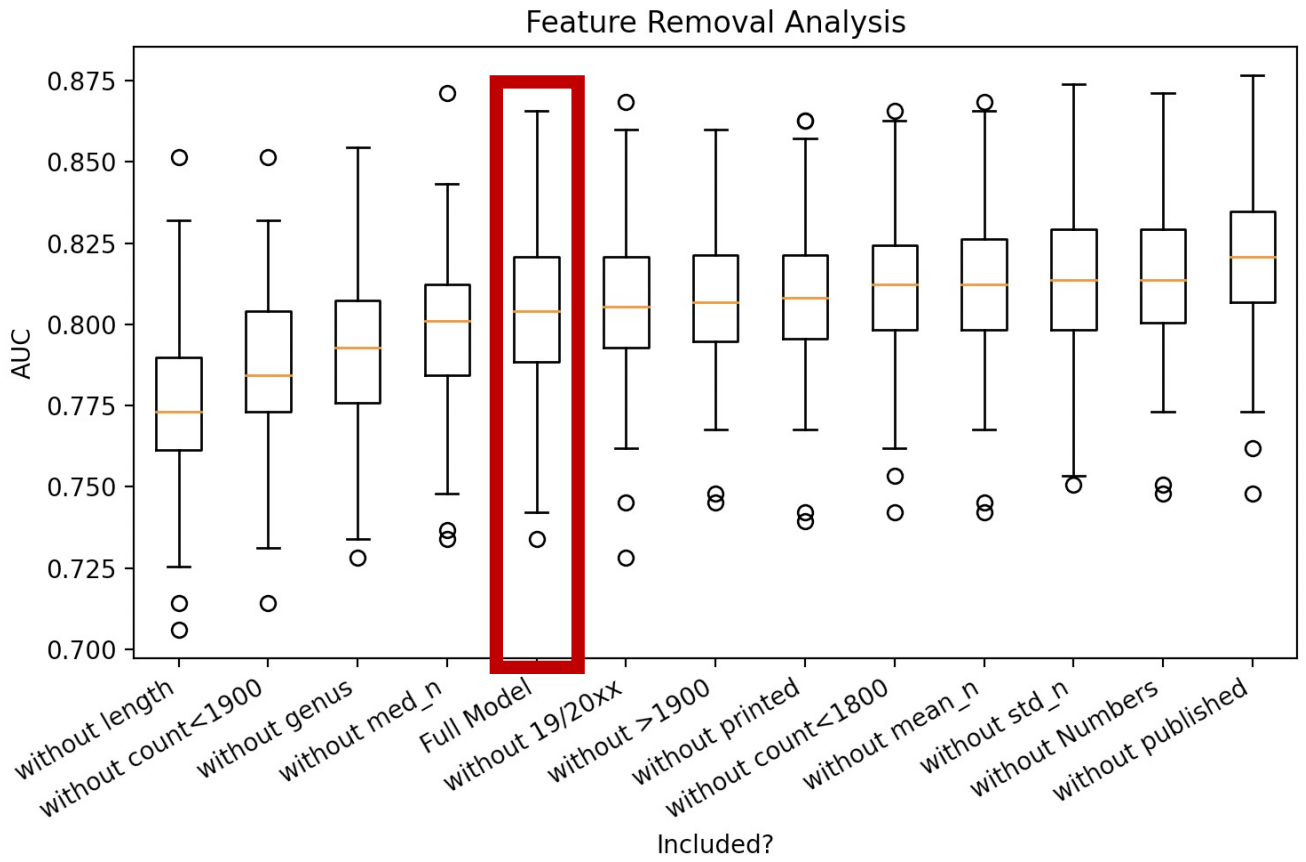
To left of Full Model:

important features that contribute to performance

To the right of Full Model:

Probably just randomness

Definitely possible to do better without!



Questions?

Topics for Today

- ~~Parisa's presentation~~
- ~~Contextualizing Friday's papers~~
- ~~Background for Friday's papers~~
- ~~New method: ablation studies~~
- Things to reflect on for this week's reviews and presentations

Things to think about

As you read…

DeepDive vs. Snorkle

- What's new?
- Which was published first?
- Following the citations: how many papers are by the same authors?

What is the ethical obligation of the annotator? Do you think human-in-the-loop solutions are a viable antidote to uncontrolled machine learning applications?