Improving Healthcare with Data Science

Janu Verma

IBM T.J. Watson Research Center, New York

http://jverma.github.io/

🐦 @januverma

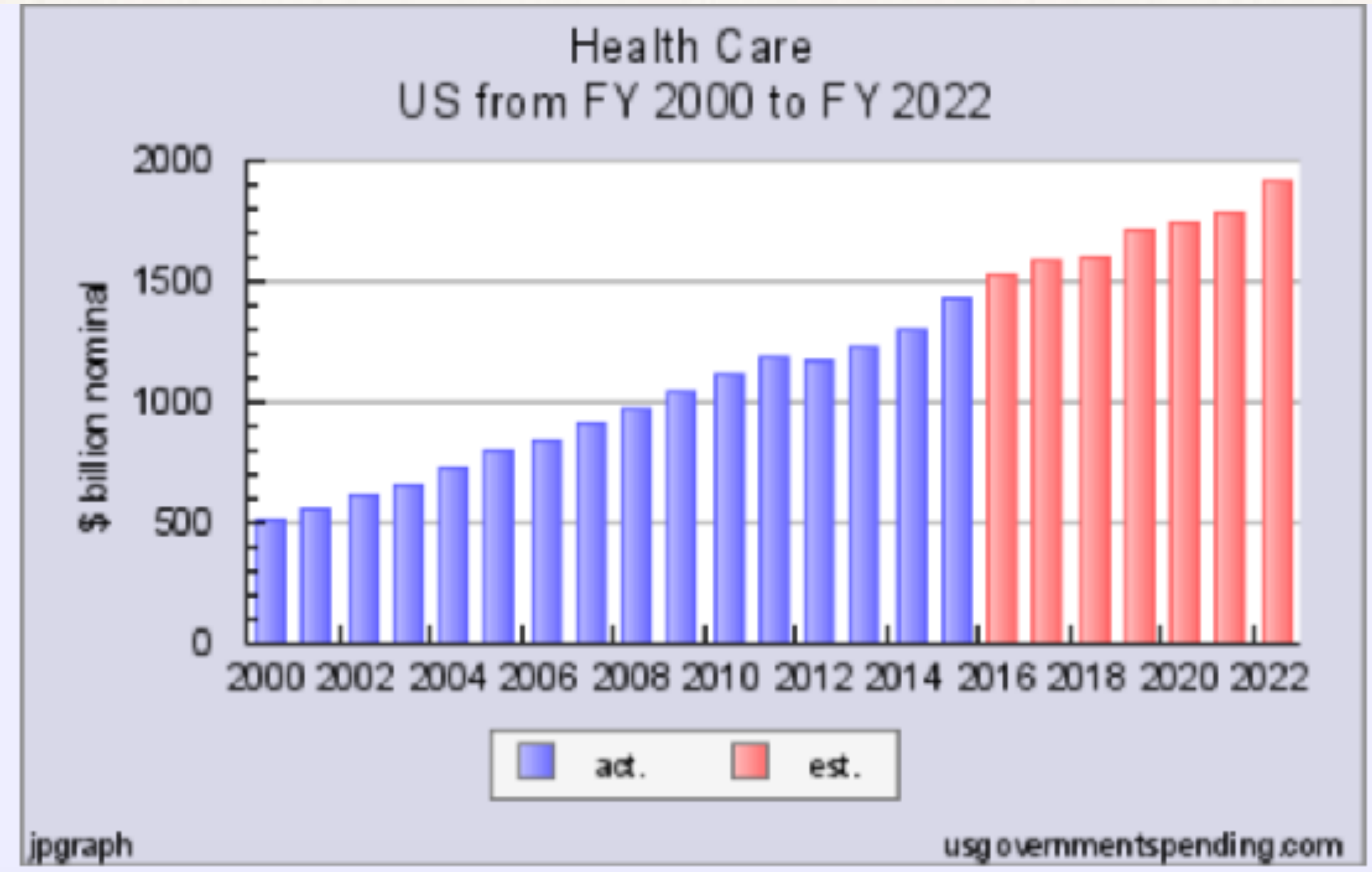
About me

- Researcher at IBM T.J. Watson Research Center in New York. *HealthCare Analytics Group*.
- Currently, on an assignment at IBM India Research Lab, New Delhi.
- Research: Machine leaning, Visual-analytics, data science, computational healthcare and genomics, mathematics.
- Advising and consulting: Data science machine learning, hiring and team building, teaching and speaking.
- Previous: Computational Genomics at Cornell University, pure mathematics at Kansas State University, theoretical physics at Cambridge University, TIFR, and JNCASR.
- Reach me: @januverma

Need for data driven healthcare

- Annual Healthcare spending in the US > $3 trillion and is increasing sharply at an unsustainable rate. 5% of GDP in 1960 to 17% in 2015.

US budget on healthcare



Need for data driven healthcare

- Healthcare spending in the US > $3 trillion and is increasing sharply at an unsustainable rate.

- A big part of spending (> $600M) is the treatments that are are totally ineffective or cause adverse reactions.

Need for data driven healthcare

- Healthcare spending in the US > $3 trillion and is increasing sharply at an unsustainable rate.

- A big part of spending ($600M) is the treatments that are are totally ineffective or cause adverse reactions.

- Healthcare providers rely on *standard-of-care-average-patient* treatment, something that works for one patient may not work for others, especially patients with other comorbidities. **Each patient is different!**

Need for data driven healthcare

- Healthcare spending in the US > $3 trillion and is increasing sharply at an unsustainable rate.

- A big part of spending ($600M) is the treatments that are are totally ineffective or cause adverse reactions.

- Healthcare providers rely on *standard-of-care-average-patient* treatment, something that works for one patient may not work for others, especially patients with other comorbidities. **Each patient is different!**

- Healthcare systems are still not very efficient and slow. **Early detection can save lives!!**

Need for data driven healthcare

- Healthcare spending in the US > $3 trillion and is increasing sharply at an unsustainable rate.

- A big part of spending ($600M) is the treatments that are are totally ineffective or cause adverse reactions.

- Healthcare providers rely on *standard-of-care-average-patient* treatment, something that works for one patient may not work for others, especially patients with other comorbidities. **Each patient is different!**

- Healthcare systems are still not very efficient and slow. **Early detection can save lives!!**

- Lot of admissions, readmissions, ER visits are preventable.

Need for data driven healthcare

- Healthcare spending in the US > $3 trillion and is increasing sharply at an unsustainable rate.

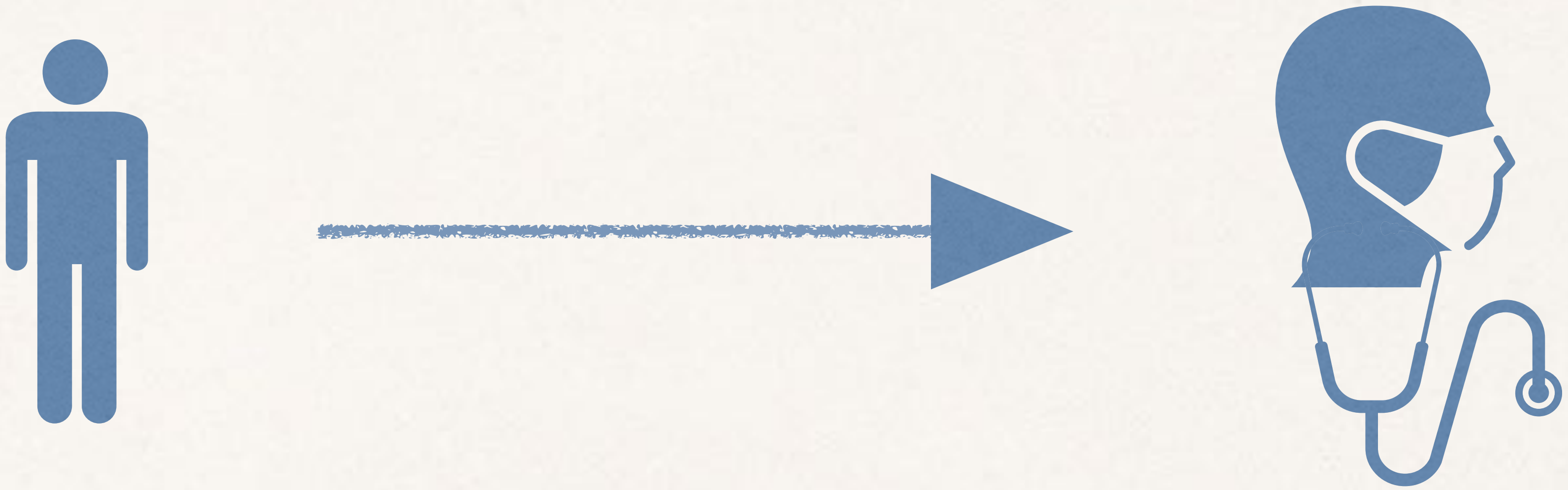
- A big part of spending ($600M) is the treatments that are are totally ineffective or cause adverse reactions.

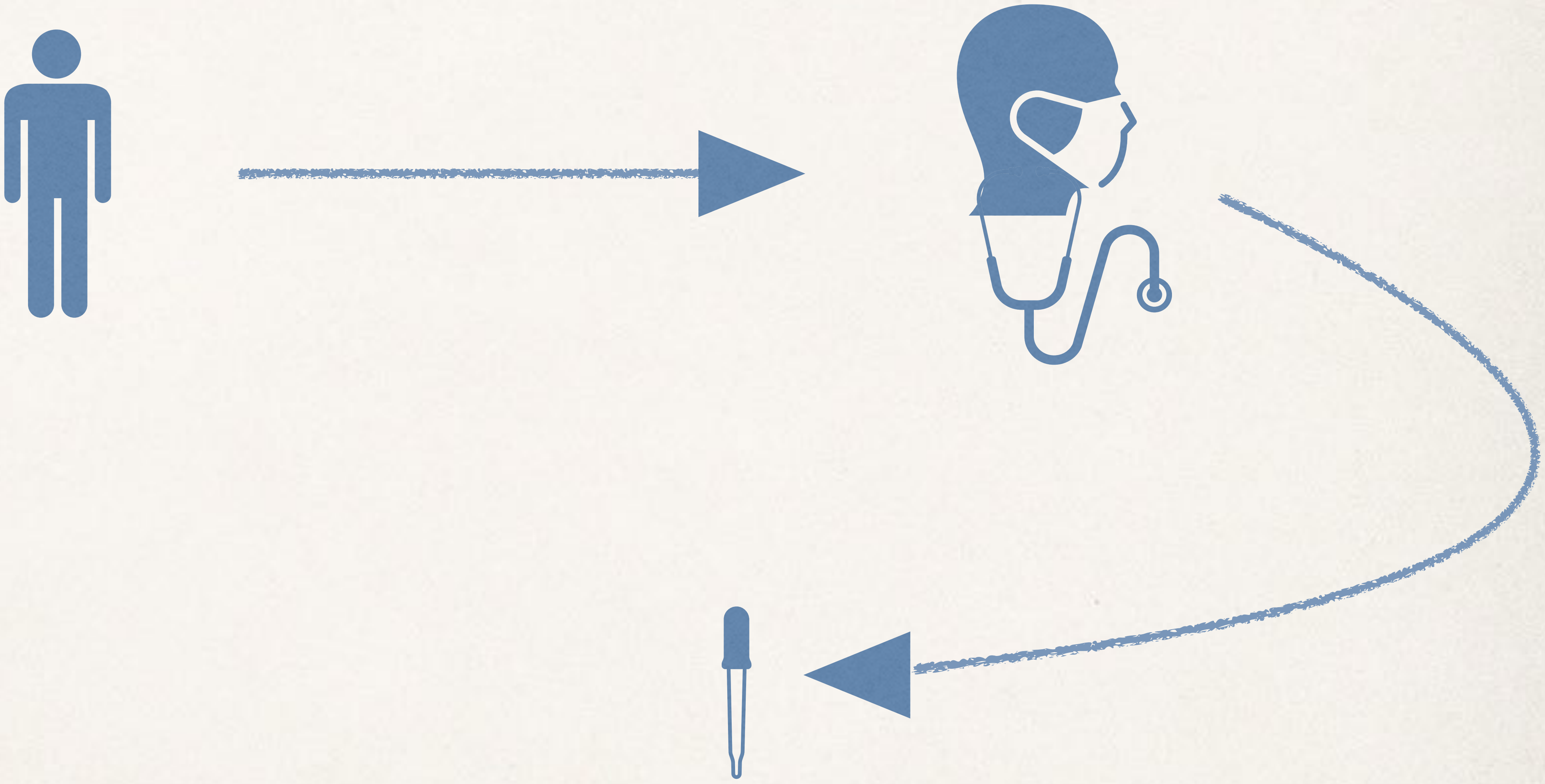
- Healthcare providers rely on *standard-of-care-average-patient* treatment, something that works for one patient may not work for others, especially patients with other comorbidities. **Each patient is different!**

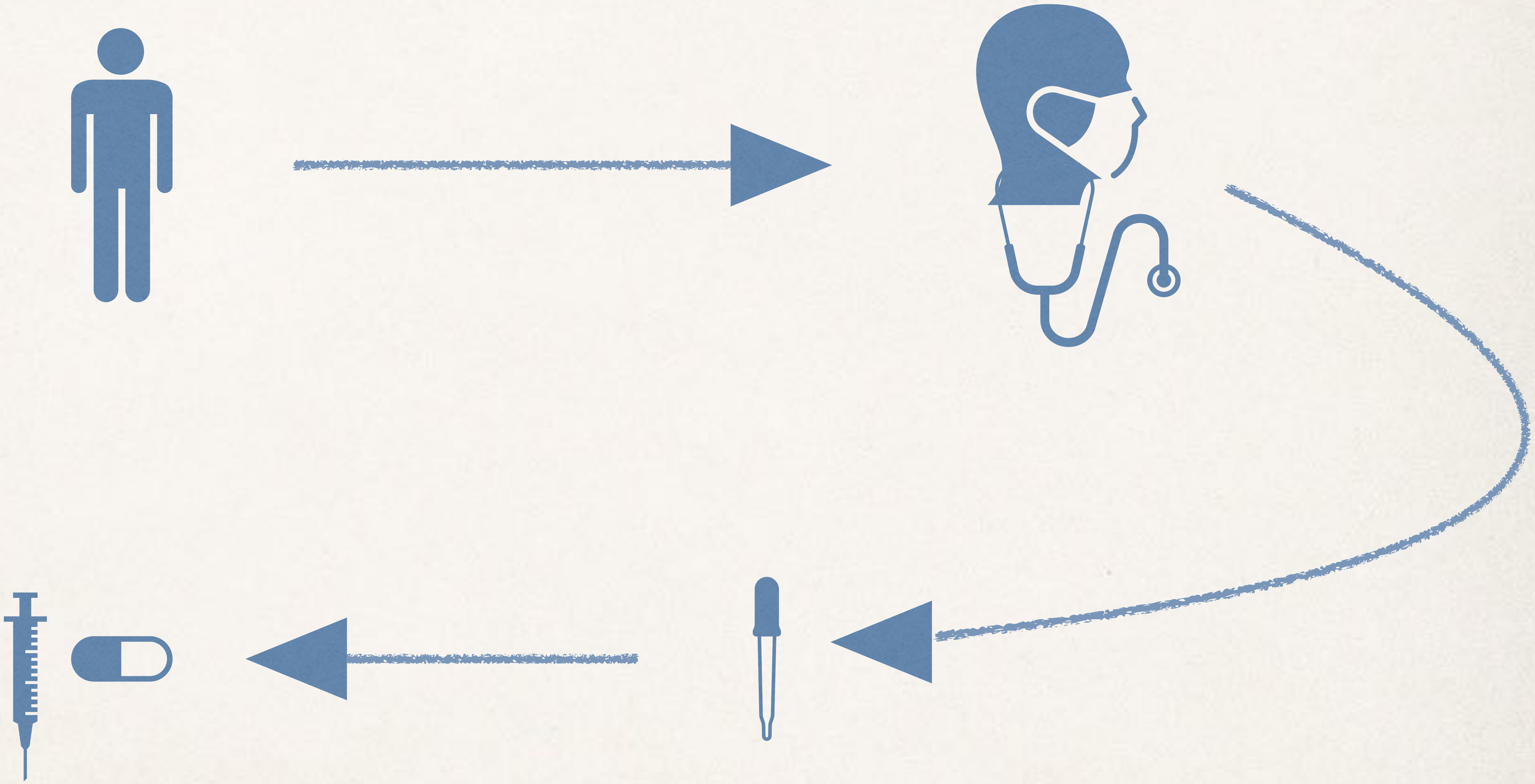
- Healthcare systems are still not very efficient and slow. **Early detection can save lives!!**

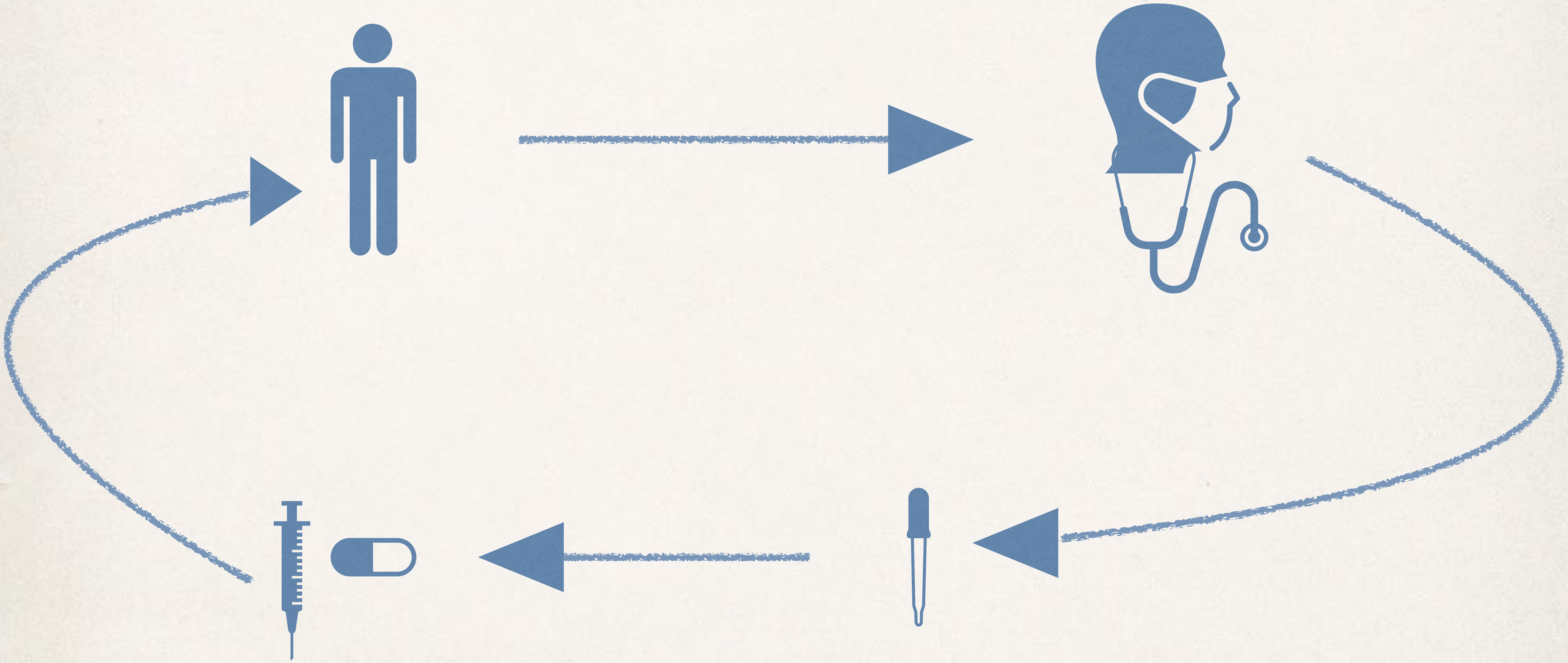
- Lot of admissions, readmissions, ER visits are preventable.

- Enormous cost for developing and testing new drugs.









"The best minds of my generation are thinking about how to make people click ads. That sucks."

–Jeff Hammerbacher,

Lead first ever data team at Facebook,

Co-coined the term 'data scientist',

Cofounder Cloudera,

Genomics Scientist at Mount Sinai Hospital, NY

Data Driven Healthcare: Opportunities

- Predict more accurately which treatments will be more effective for which patient, and which treatments won’t.
- Understanding relation between treatments, outcomes, and patients would lead us to personalised healthcare.

Data Driven Healthcare: Opportunities

- Predict more accurately which treatments will be more effective for which patient, and which treatments won't.

- Understanding relation between treatments, outcomes, and patients would lead us to personalised healthcare.

Instead of finding a treatment for all patients with a given problem, we need to find pairs (treatment, set of similar patients with the problem)

Data Driven Healthcare: Opportunities

- Predict more accurately which treatments will be more effective for which patient, and which treatments won't.

- Relation between treatments, outcomes, and patients would lead us to personalised healthcare.

- Predictive models that can identify those individuals who would benefit from early, targeted interventions have the potential to improve outcomes, reduce unnecessary utilization, and drive down spending.

  e.g. acute adverse events such as sepsis and cardiac arrest in hospitalized patients; these conditions significantly worsen outcomes, increasing length of stay and unnecessary spending.

Data Driven Healthcare: Opportunities

- Predict more accurately which treatments will be more effective for which patient, and which treatments won’t.

- Relation between treatments, outcomes, and patients would lead us to personalised healthcare.

- Predictive models that can identify those individuals who would benefit from early, targeted interventions have the potential to improve outcomes, reduce unnecessary utilization, and drive down spending.

- Models to predict readmissions. *Heritage Health Prize.*

Data Driven Healthcare: Opportunities

- Predict more accurately which treatments will be more effective for which patient, and which treatments won’t.

- Relation between treatments, outcomes, and patients would lead us to personalised healthcare.

- Predictive models that can identify those individuals who would benefit from early, targeted interventions have the potential to improve outcomes, reduce unnecessary utilization, and drive down spending.

- Models to predict readmissions. *Heritage Health Prize.*

- Better models for drug discovery, drug reposition, drug safety.

Data Driven Healthcare: Opportunities

- Predict more accurately which treatments will be more effective for which patient, and which treatments won't.

- Relation between treatments, outcomes, and patients would lead us to personalised healthcare.

- Predictive models that can identify those individuals who would benefit from early, targeted interventions have the potential to improve outcomes, reduce unnecessary utilization, and drive down spending.

- Models to predict readmissions. *Heritage Health Prize.*

- Better models for drug discovery, drug reposition, drug safety.

- Understand how disease progress over time.

Data

- Vast amount of data of various forms has been collected over the years.
  - Adoption of Electronic medical records (EMRs) instead of paper records by healthcare providers in 2009. This includes demographic, personal and family history, current and past treatments, history of allergic reactions, vaccination records, laboratory test results, imaging results, doctors notes etc.
  - Claims data by insurance providers.
  - Drugs: chemical data, clinical trails, biological pathways, adverse reactions etc.
  - Gene expression data, DNA sequence data, proteomics, and metabolomics.
  - Vast literature available.
  - Online communities, social media, discussion forums.
  - Wearable devices - activities, sleep, heart rate etc.
  - Human behavioural data.

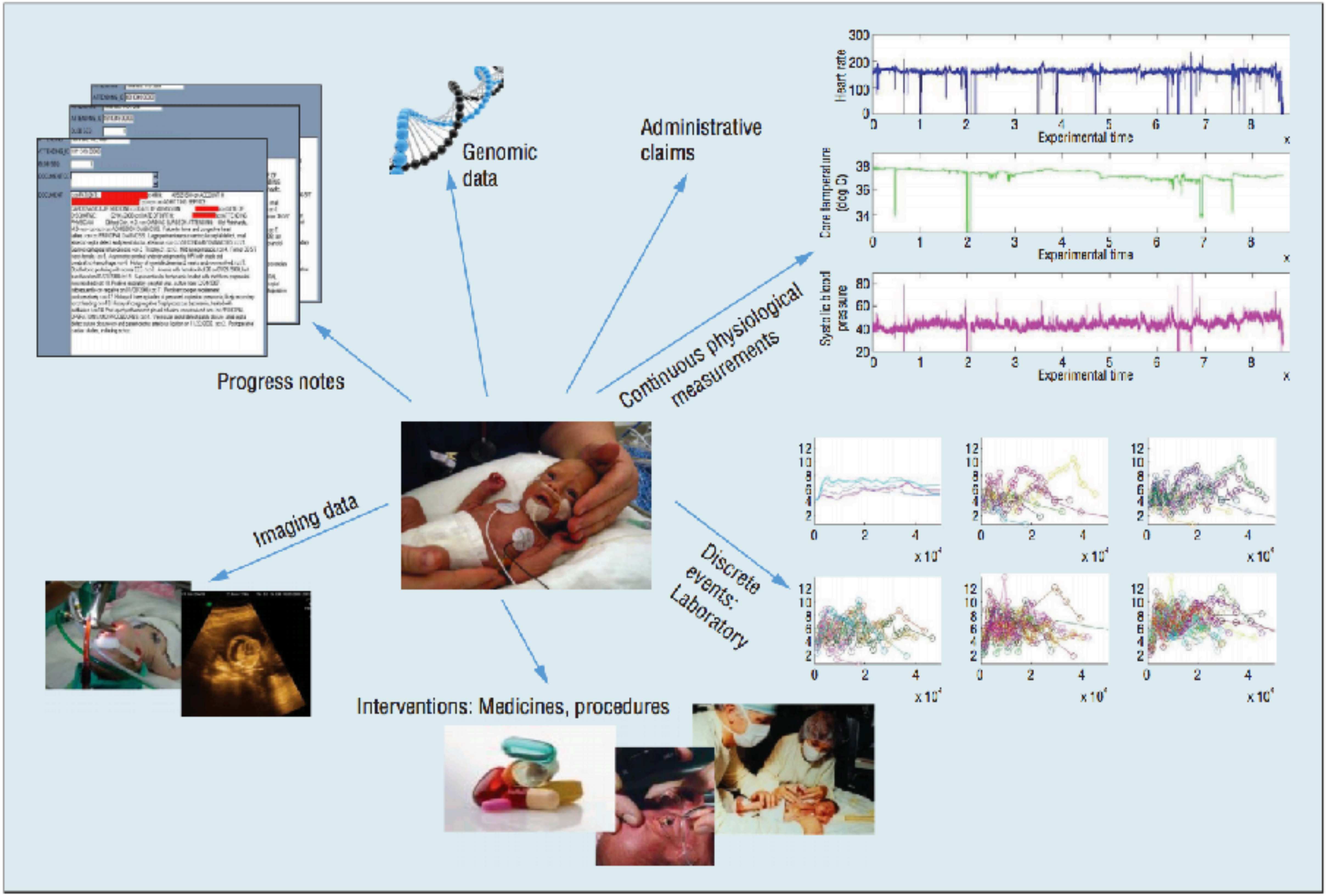


Figure 1. An illustration of the diverse electronic health data (EHD) that are routinely collected, including physiological measurements, laboratory test results, medications administered, imaging test results, progress and discharge reports, genomic profiles, and administrative claims.

Data Driven Healthcare

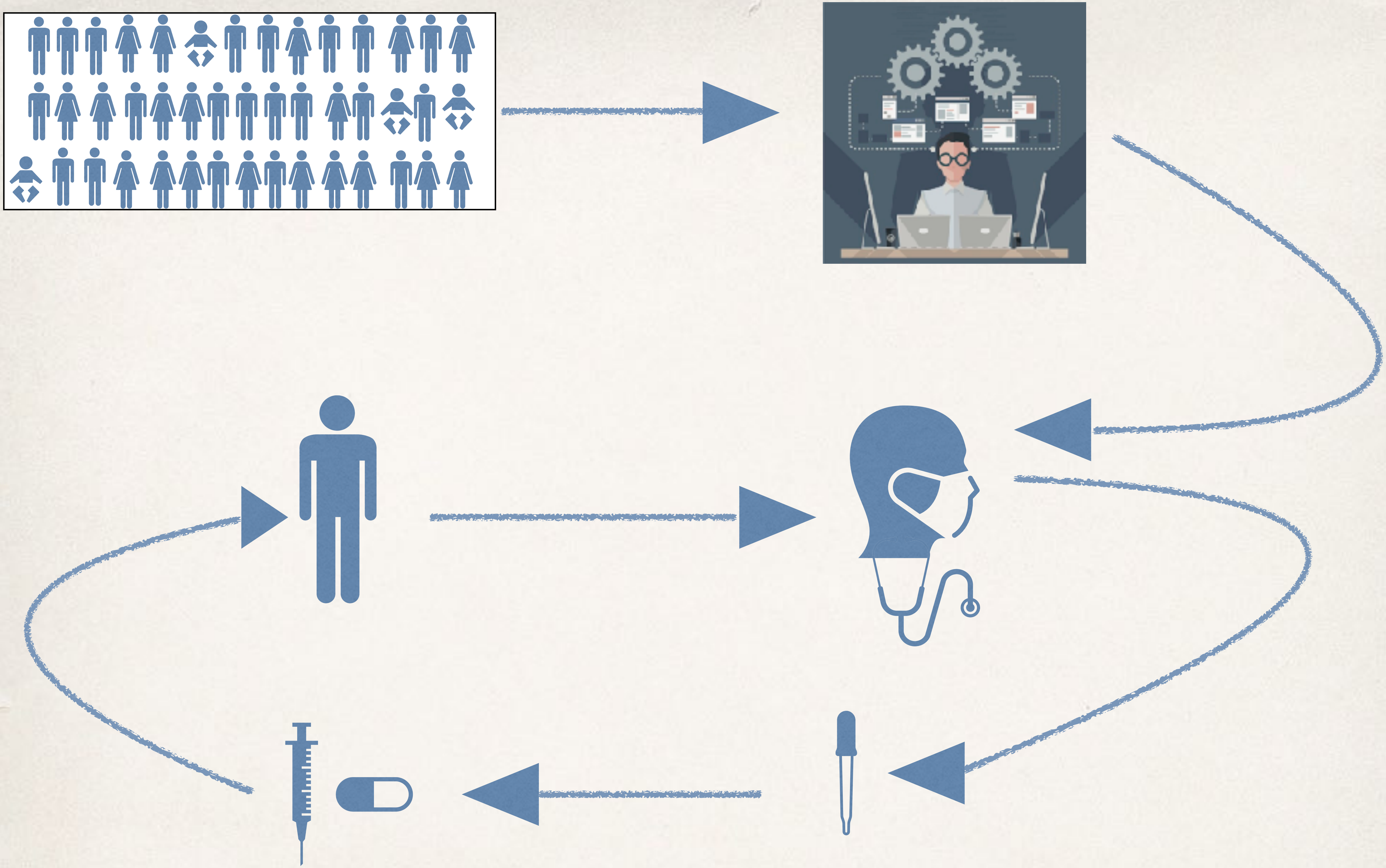
- Presents an unprecedented opportunity for data scientists, machine learning researchers, statisticians, software developers.

- At IBM, we do research at the interface of data science and healthcare incorporating machine learning, statistical analysis, visual-analytics, genomics,

- Data: Numerical, ordinal, boolean, temporal, sequential, textual, images, bio-chemical ….

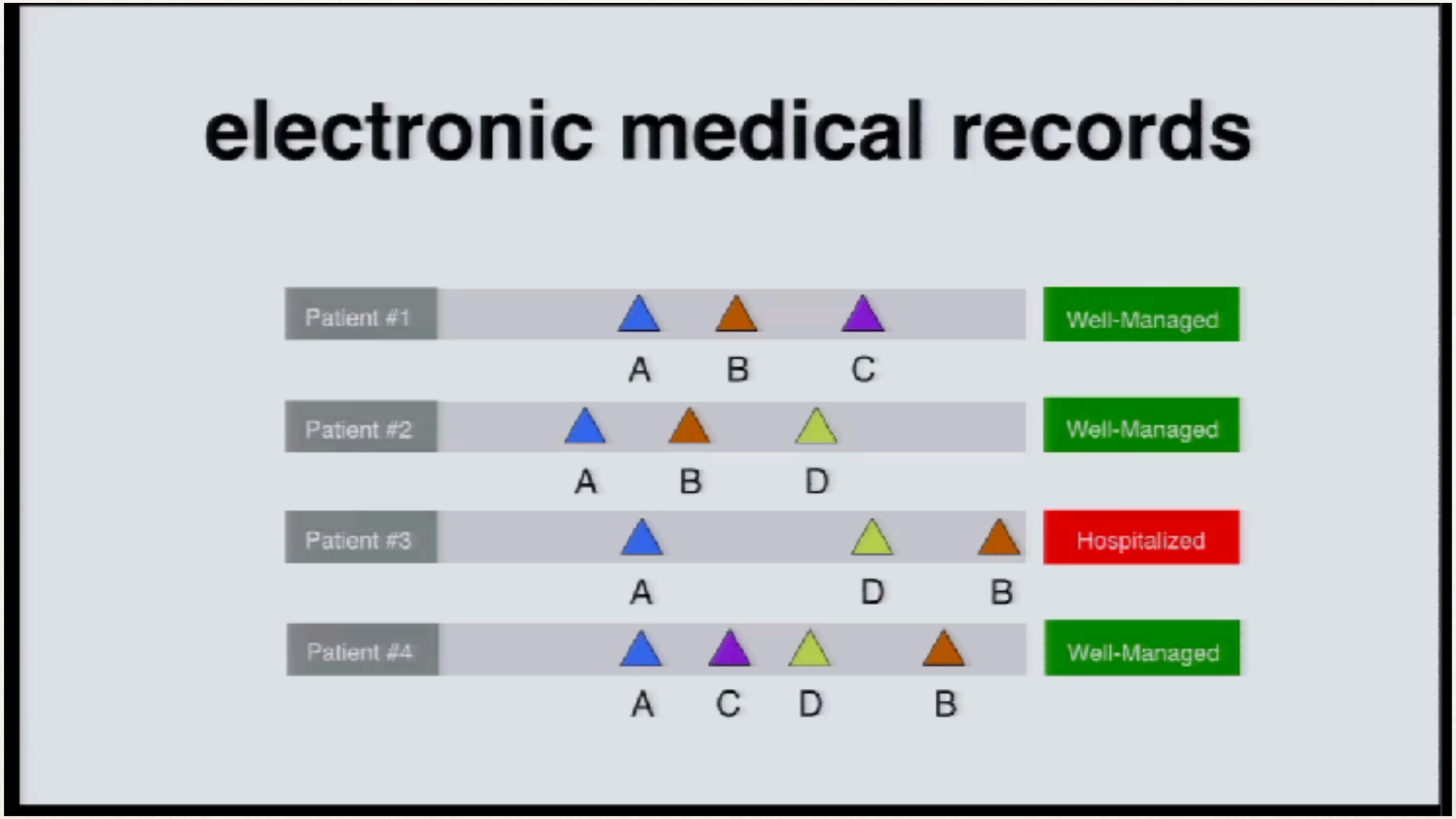
- My research touches on each of these avenues. Largely, machine learning applications and visual-analytics to understand the models.

“Goal is **not** to remove the clinical experts from the loop but work with them, help them understand the vast amount of data, build tools that they can use and make better decisions. Towards a human-in-the-loop analysis.”



Event Sequence Mining

- EMRs contain all the medical history of patients which can be seen as a *temporal event sequence*.

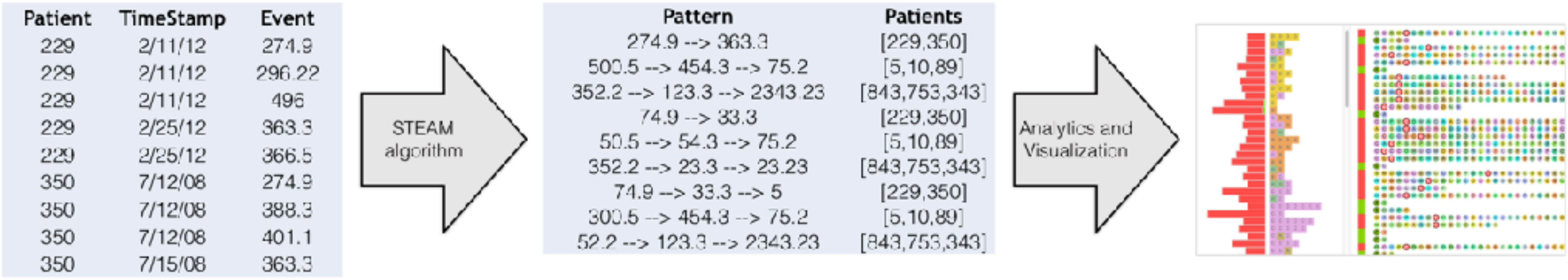


- Question: Can we apply machine learning and data mining to patient record sequences to automatically discover care pathways?
- This helps to -
    - Understand how diseases progress over time.
    - Understand the effects of interventions.
    - Identify subgroups of patients based on temporal patterns.
    - Feature extraction for machine learning.
- We want to be able to extract most informative subsequences in cohort with respect to the outcomes.

Temporal sequence mining

- The standard sequence mining algorithms are not very effective in this case due to:
    - The large scale of the data - too many patients and sequences.
    - Huge number of events.
    - Temporality in data.
    - Inability to associate the sequences with the outcomes.
    - Non-uniformity in the gap between two events.
    - Concurrency of events.
    - Hierarchy of events.

"A novel sequence mining algorithm for this case, and a visual-analytical system that a physician can use to navigate through the large number of pathways to reach at a meaningful hypothesis."



Patient	TimeStamp	Event
229	2/11/12	274.9
229	2/11/12	296.22
229	2/11/12	496
229	2/25/12	363.3
229	2/25/12	366.5
350	7/12/08	274.9
350	7/12/08	388.3
350	7/12/08	401.1
350	7/15/08	363.3

Pattern	Patients
274.9 --> 363.3	[229,350]
500.5 --> 454.3 --> 75.2	[5,10,89]
352.2 --> 123.3 --> 2343.23	[843,753,343]
74.9 --> 33.3	[229,350]
50.5 --> 54.3 --> 75.2	[5,10,89]
352.2 --> 23.3 --> 23.23	[843,753,343]
74.9 --> 33.3 --> 5	[229,350]
300.5 --> 454.3 --> 75.2	[5,10,89]
52.2 --> 123.3 --> 2343.23	[843,753,343]

Input:
Dataset of Patient
Time-stamped events

Output:
List of frequent event sequences and
the patients that have them

UI:
Visual Analytics to explore and
interpret pathways

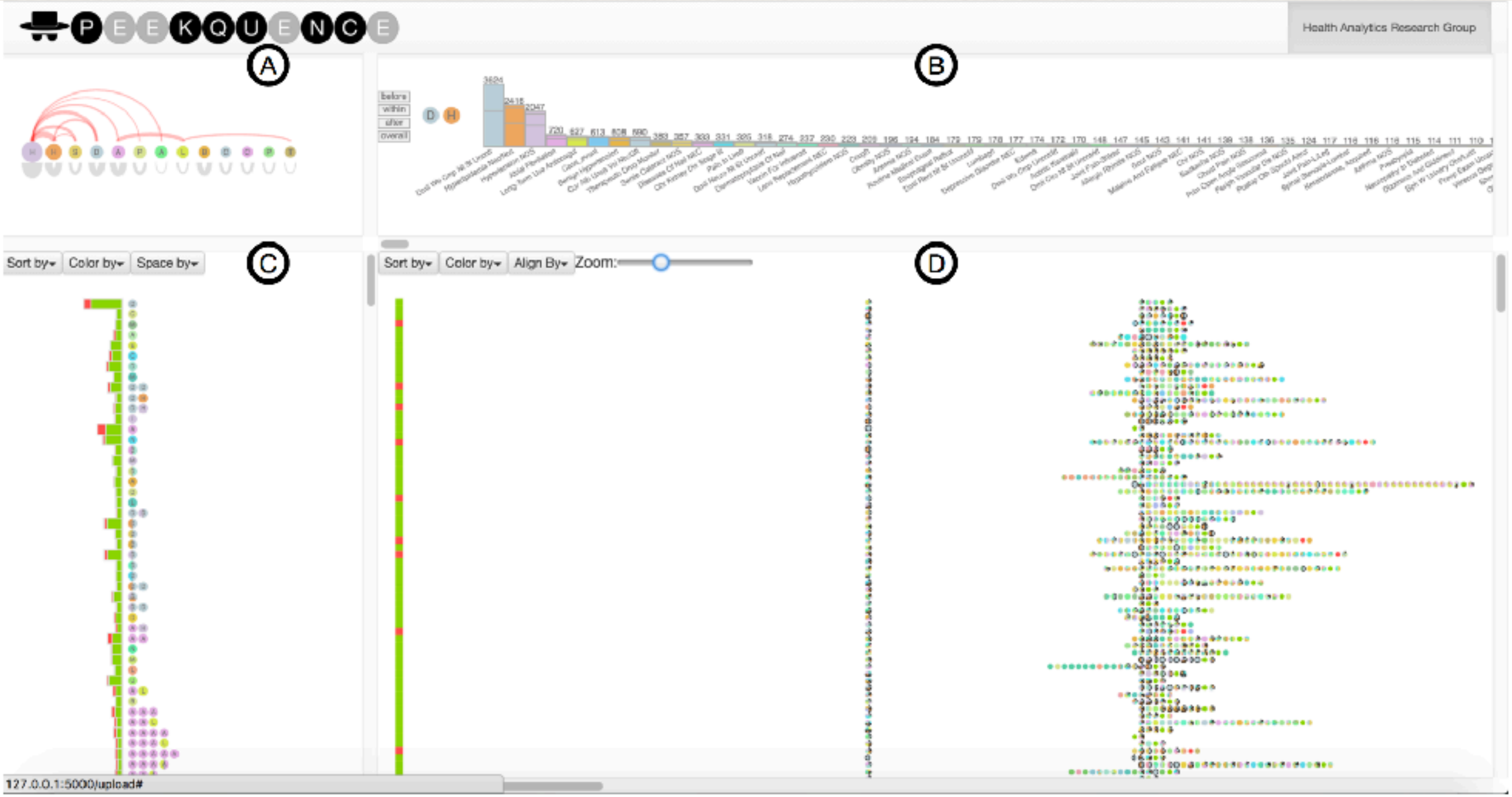
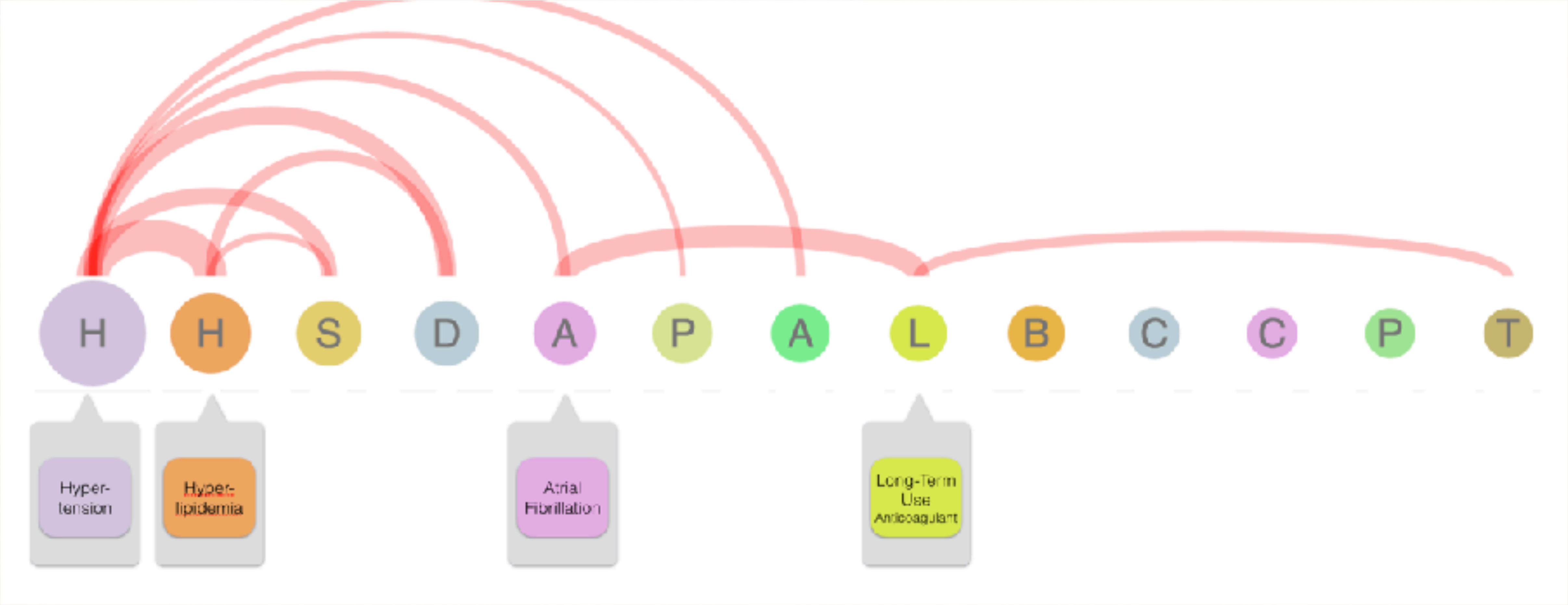
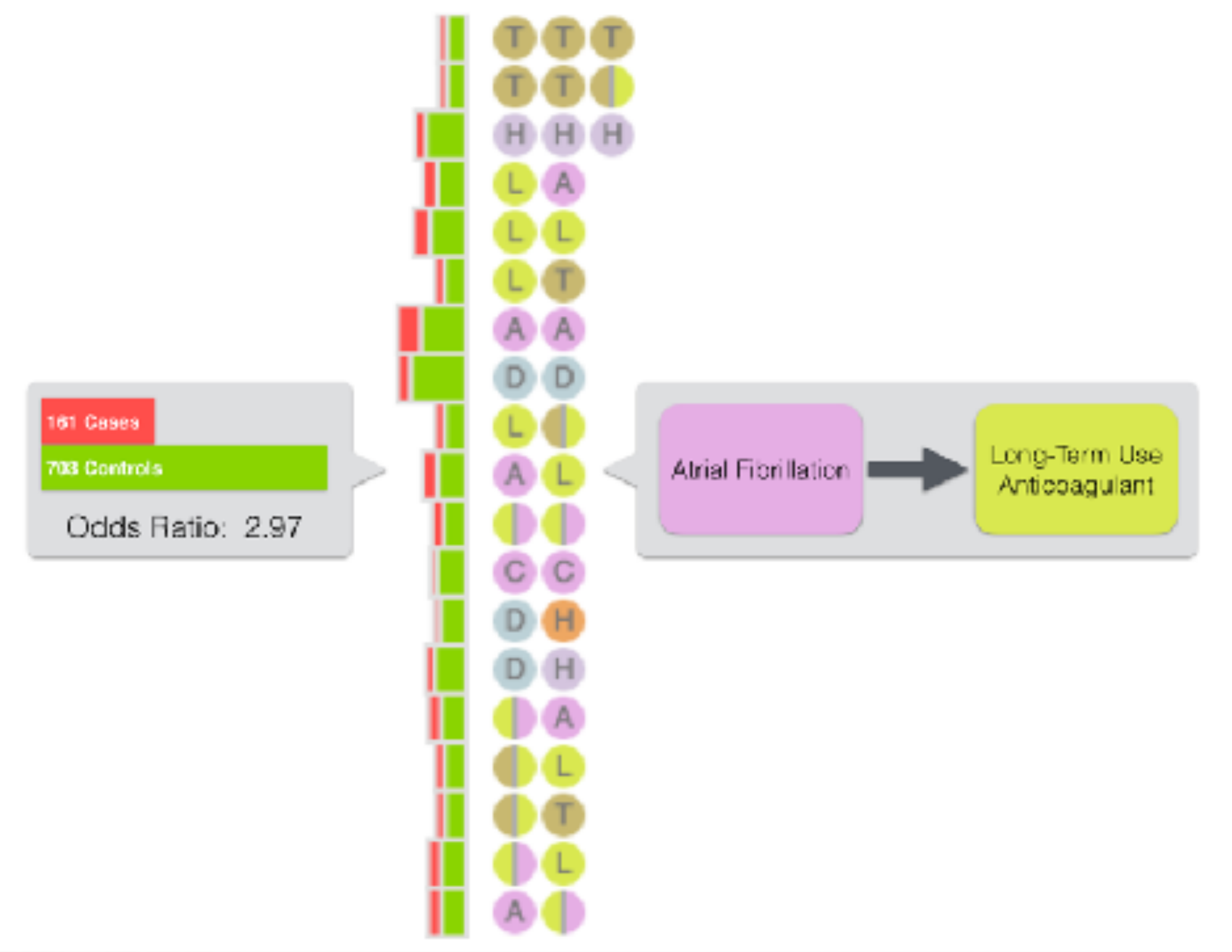
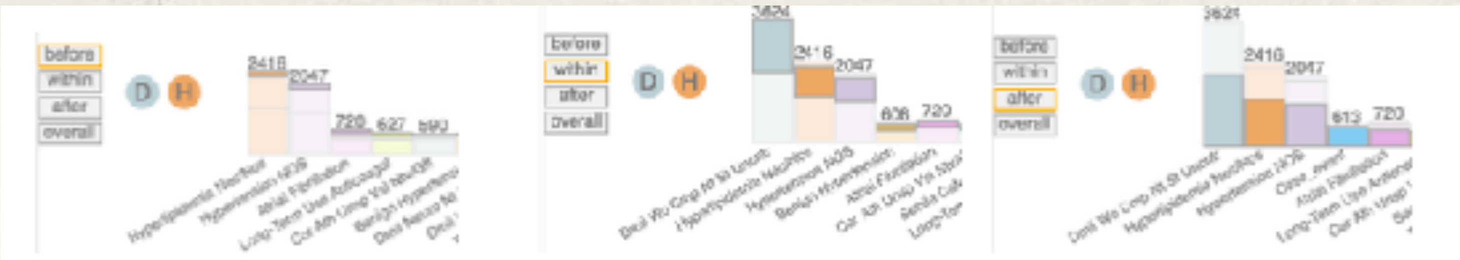


Figure 1. Peekquence consists of four views: (A) the sequence network view showing the frequency of event sequence occurrences within patterns mined from SPAM; (B) the event co-occurence histogram view showing the frequency of events co-occuring with a pattern selected ("S", "H" in this example); (C) the pattern list view showing patterns mined from SPAM with event sequences (colored circles with letters) as well as bars of patients with the ratio of case and control labels (diagnosis of a disease); (D) the patient timeline view showing patients' event sequences aligned with respect to the pattern selected ("S" and "H" events are vertically algined in this example).



- The radius of the circle corresponds to the number of patients that have the event in their sequence.
- The circles are coloured according to a category of the event according to the ICD-9 International Classification of Diseases for clinical events.
- The edges represent the co-occurrence of the the events in patients, the thickness of the edge corresponds to the frequency of co-occurrence.
- User can click on any of the nodes of the edges to filter the other panels accordingly.





(a) Before (b) Within (c) After

Figure 4. The histogram view sorted by before, within, and after the pattern.

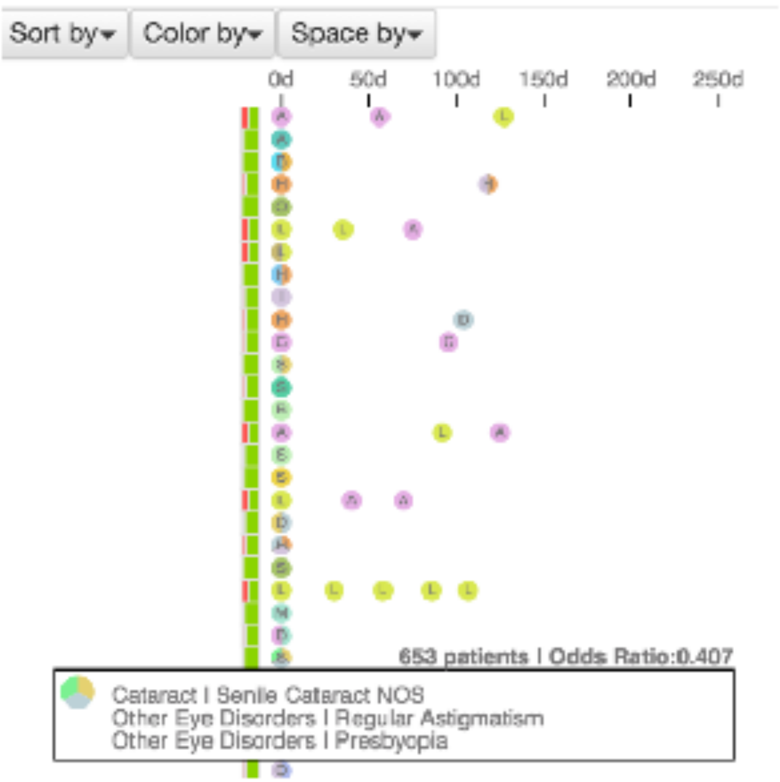
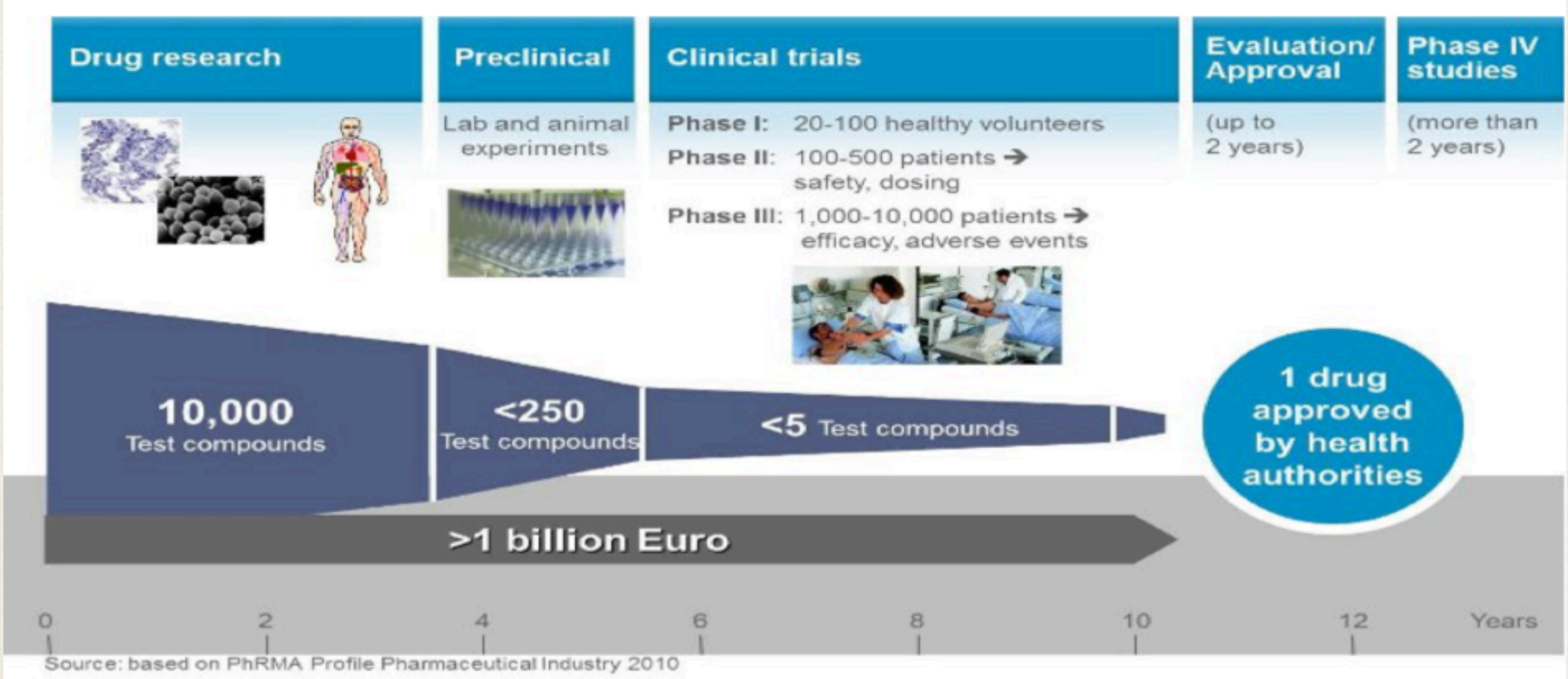

Sort by▾
Color by▾
Space by▾
0d
50d
100d
150d
200d
250d
653 patients | Odds Ratio:0.407
Cataract | Senile Cataract NOS
Other Eye Disorders | Regular Astigmatism
Other Eye Disorders | Presbyopia

Figure 3. The pattern list view showing patterns of events spread out based on average time duration between events.

Drug Discovery



Drug research

Preclinical

Lab and animal experiments

Clinical trials

Phase I: 20-100 healthy volunteers

Phase II: 100-500 patients → safety, dosing

Phase III: 1,000-10,000 patients → efficacy, adverse events

Evaluation/ Approval

(up to 2 years)

Phase IV studies

(more than 2 years)

10,000 Test compounds

<250 Test compounds

<5 Test compounds

1 drug approved by health authorities

>1 billion Euro

0 2 4 6 8 10 12 Years

Source: based on PhRMA Profile Pharmaceutical Industry 2010

Drug Adverse Reactions

- Adverse drug reactions (ADRs) are the clinical conditions resulted from taking medications at normal doses

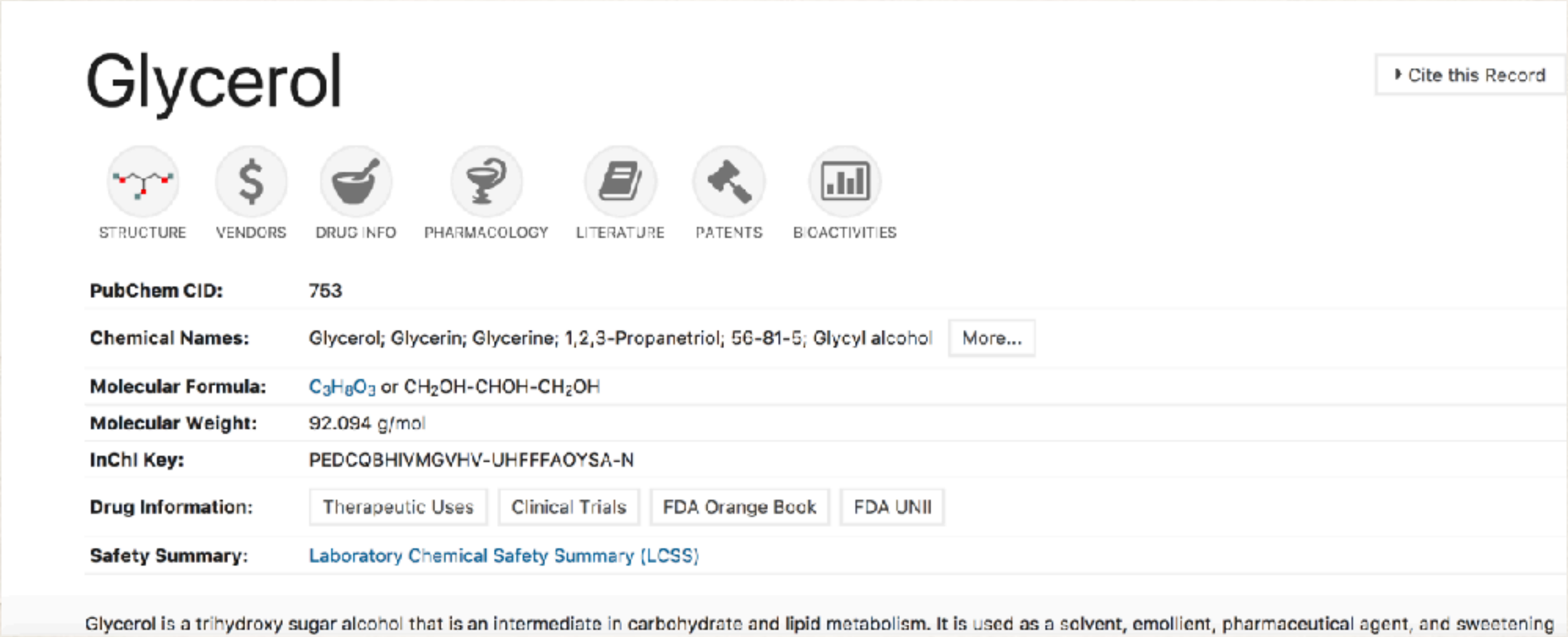
- They cause 700,000 emergency department visits and 120,000 hospitalizations per year and are one of the major causes of death among hospitalized patients

- Very important to be able to predict the ADRs for drugs even in the early development stage

- A challenging problem is to build machine learning model to predict the adverse reactions for the drug.

ML model for ADR prediction

- Use publicly available known ADR information for drugs as the training data.
- Data munging and expert advice to prepare training data with strong evidence of drug-ADR relation.
- Features: 881 structural descriptors.
- Train machine learning model and use it for prediction of new drug-ADR pairs.
- Also important is understanding of statistics to be able to access significance of the association. (STATS!!)



Glycerol

PubChem CID:	753
Chemical Names:	Glycerol; Glycerin; Glycerine; 1,2,3-Propanetriol; 56-81-5; Glycyl alcohol More...
Molecular Formula:	$C_3H_8O_3$ or $CH_2OH-CHOH-CH_2OH$
Molecular Weight:	92.094 g/mol
InChI Key:	PEDCQBHIVMGVHV-UHFFFAOYSA-N
Drug Information:	Therapeutic Uses Clinical Trials FDA Orange Book FDA UNII
Safety Summary:	Laboratory Chemical Safety Summary (LCSS)

Glycerol is a trihydroxy sugar alcohol that is an intermediate in carbohydrate and lipid metabolism. It is used as a solvent, emollient, pharmaceutical agent, and sweetening

https://pubchem.ncbi.nlm.nih.gov/compound/glycerol#section=Top

Biological support for the ML model

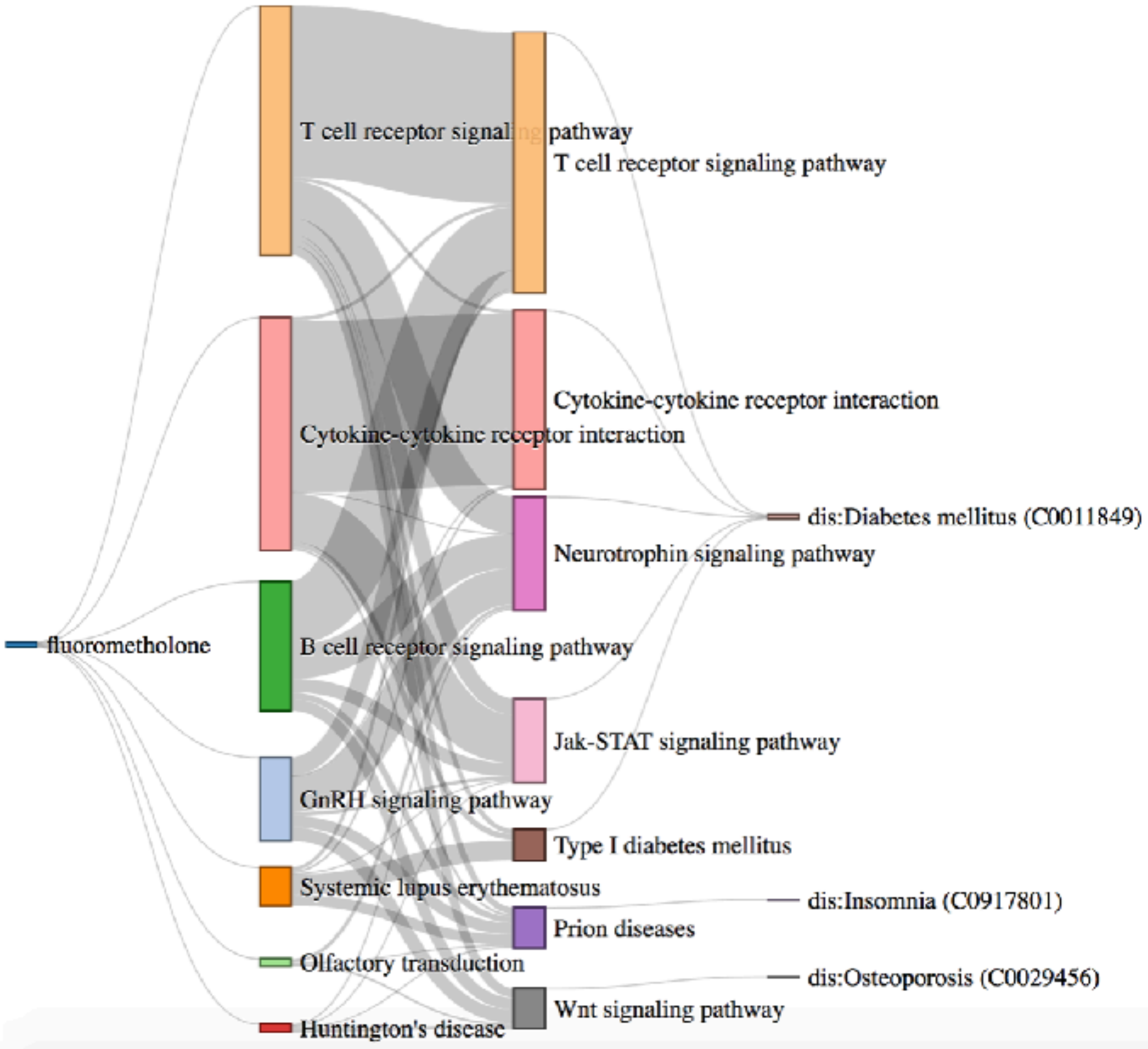
- Biological interpretation and understanding of machine learning based predictive models are highly desirable in healthcare analytics. Why should we trust a black-box model ???

- DrugPathSeeker, a novel interactive user interface that integrates the machine learning model, database query API, statistical analysis, and visualization for exploring and understanding of the association between drugs and ADRs.

- **Idea:** The gene pathways of action of the drugs and those of ADRs hold biological information about how drugs and ADRs interact with human body. They can be exploited for better understanding of the mechanism of action of the Drug-ADR association.

- This gives a method to generate a hypothesis of underlying mechanism of action between drugs and ADRs.

fluorometholone-induced diabetes


fluorometholone
T cell receptor signaling pathway
Cytokine-cytokine receptor interaction
B cell receptor signaling pathway
GnRH signaling pathway
Systemic lupus erythematosus
Olfactory transduction
Huntington's disease
T cell receptor signaling pathway
Cytokine-cytokine receptor interaction
Neurotrophin signaling pathway
Jak-STAT signaling pathway
Type I diabetes mellitus
Prion diseases
Wnt signaling pathway
dis:Diabetes mellitus (C0011849)
dis:Insomnia (C0917801)
dis:Osteoporosis (C0029456)

Text Classification

- International Conference on Healthcare Informatics (ICHI) 2016 data challenge.

- Data: Real messages on healthcare forums.

- Two different data files were provided in tab separated format for training and testing, respectively. The training data has 8000 messages each with the title text, contents text, and a category. The challenge provided seven different types of categories (tags) - Demographic, Disease, Social, Family, Treatment, Pregnancy, Goal-oriented.

- Goal: Classify the messages into appropriate categories.

- Our solution won the challenge with first prize.

Model

- Extract textual features from the titles and the contents of the messages i.e. tf-idf vectors for the n-grams and word2vec representation for the words in the text.
- We used pre-trained word2vec from Google.
- Model is an ensemble of the Support Vector Machine (SVM) on tf-idf vectors and a Convolution Neural Network (CNN) on the word2vecs.

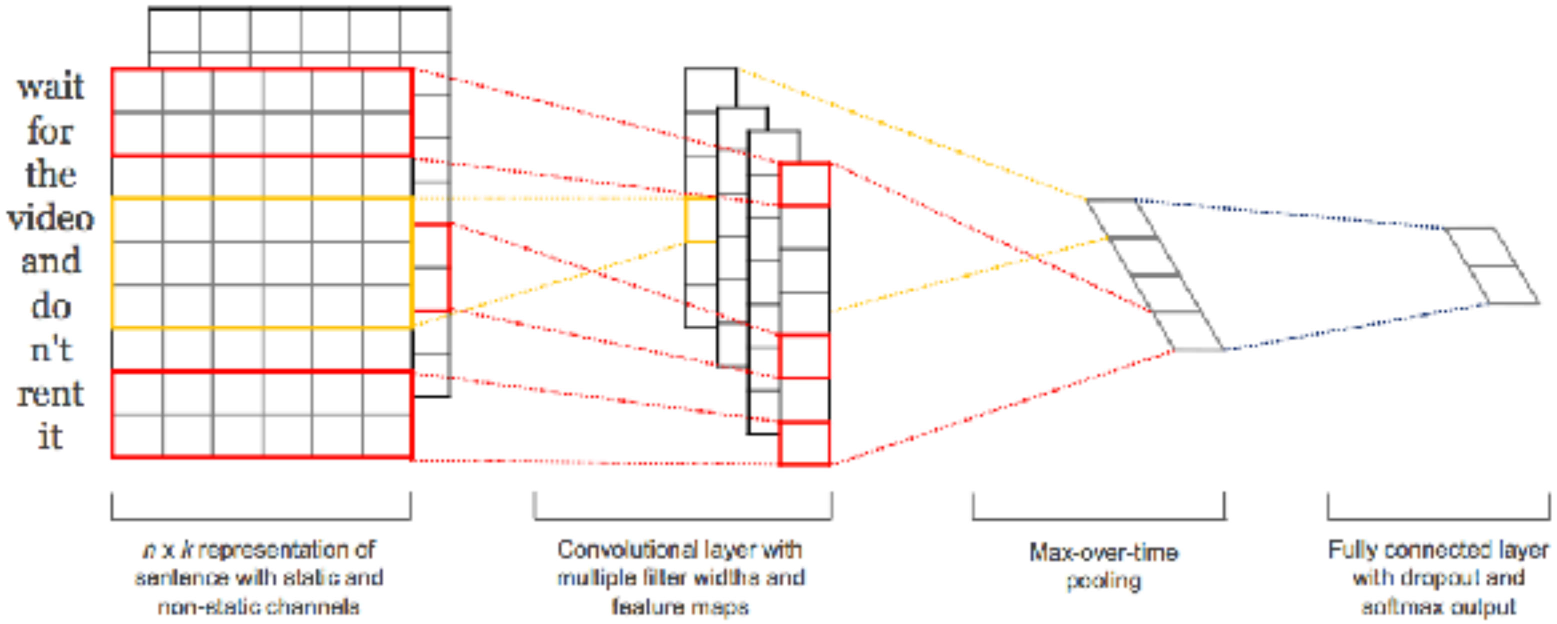

wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 1: Model architecture with two channels for an example sentence.

“Thank you!”