# 1DrugAssist: Toward Intelligent Medicine Recommender System Framework Using Deep Machine Learning

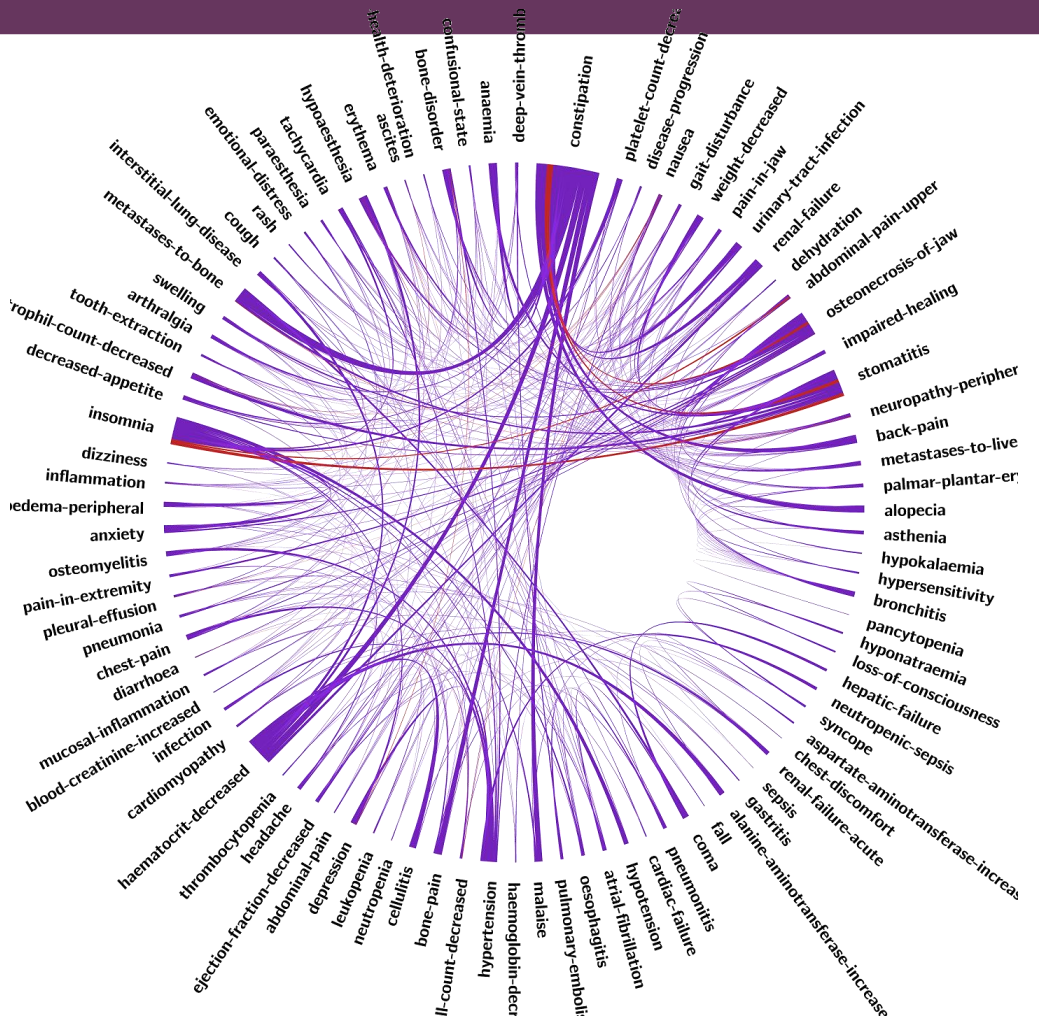
1DATA

**Majid Jaberi-Douraki, Jim Rivier, Gerald Wyckoff**
**Hossein Amini, Reza Mazloom, Heman Shakeri, Josh Staley**

K-State | 1Data, March 2019

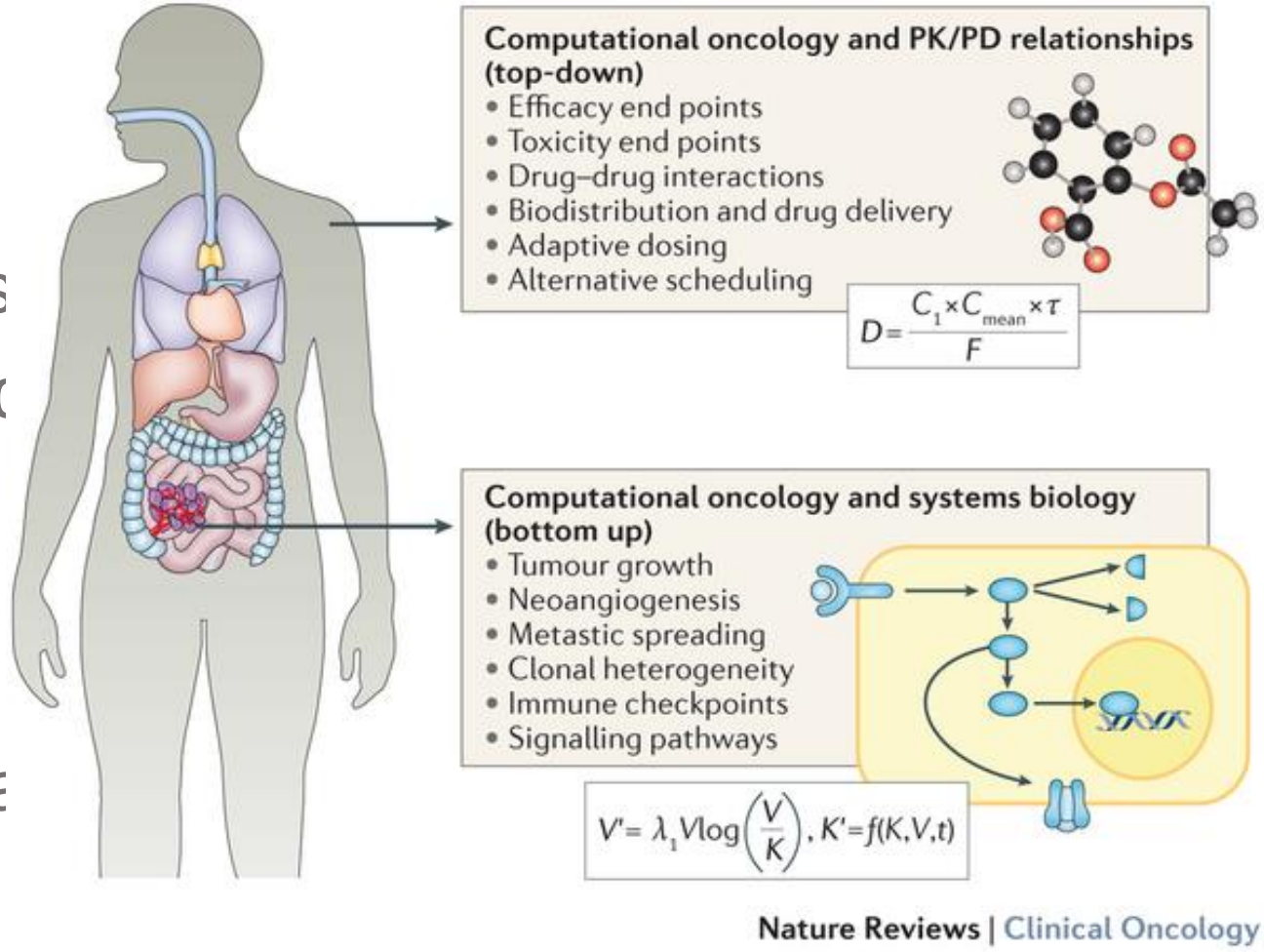
KANSAS STATE UNIVERSITY

Institute of Computational Comparative Medicine
ICCM
Kansas State University

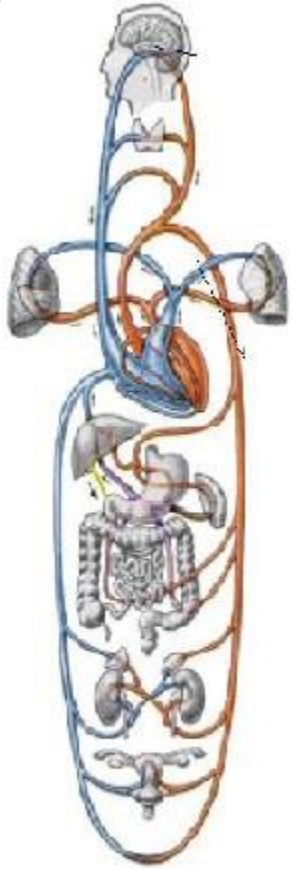


# Applications of Mathematical Modeling

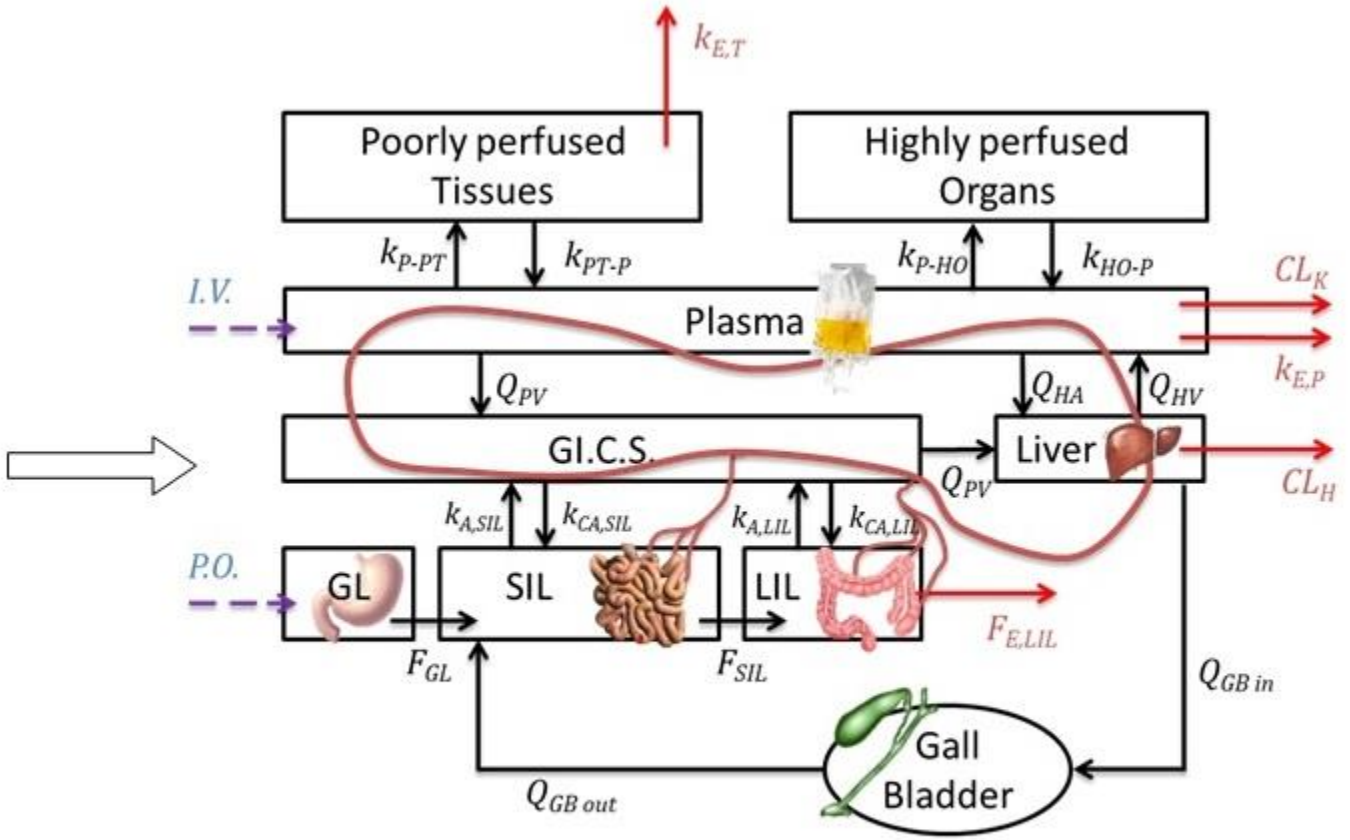
- Drug development
- Models for life sciences
- Aerospace, satellite, inc and gas, management,
- Prediction
- Parameter Estimation
- **Control** of diseases, pa



# Types of Mathematical Modeling



Nature

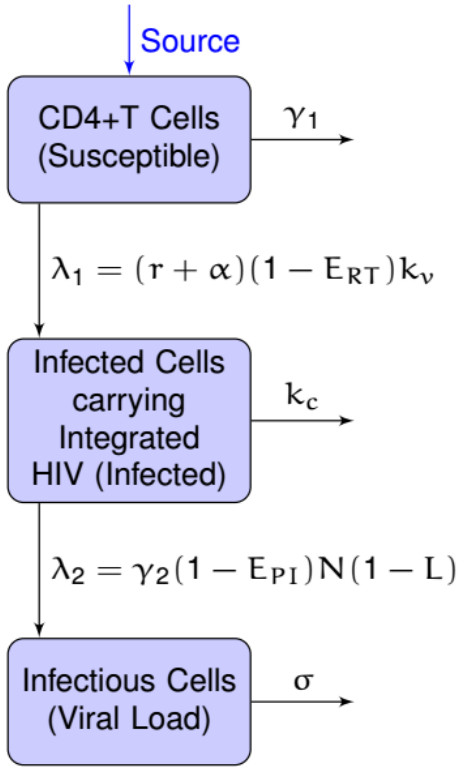



Structural representation

$$\begin{cases} \dfrac{d\boldsymbol{y}}{dt} = f(\boldsymbol{y}, t) \\ \boldsymbol{y}(0) = \boldsymbol{y}_0 \end{cases}$$

Model

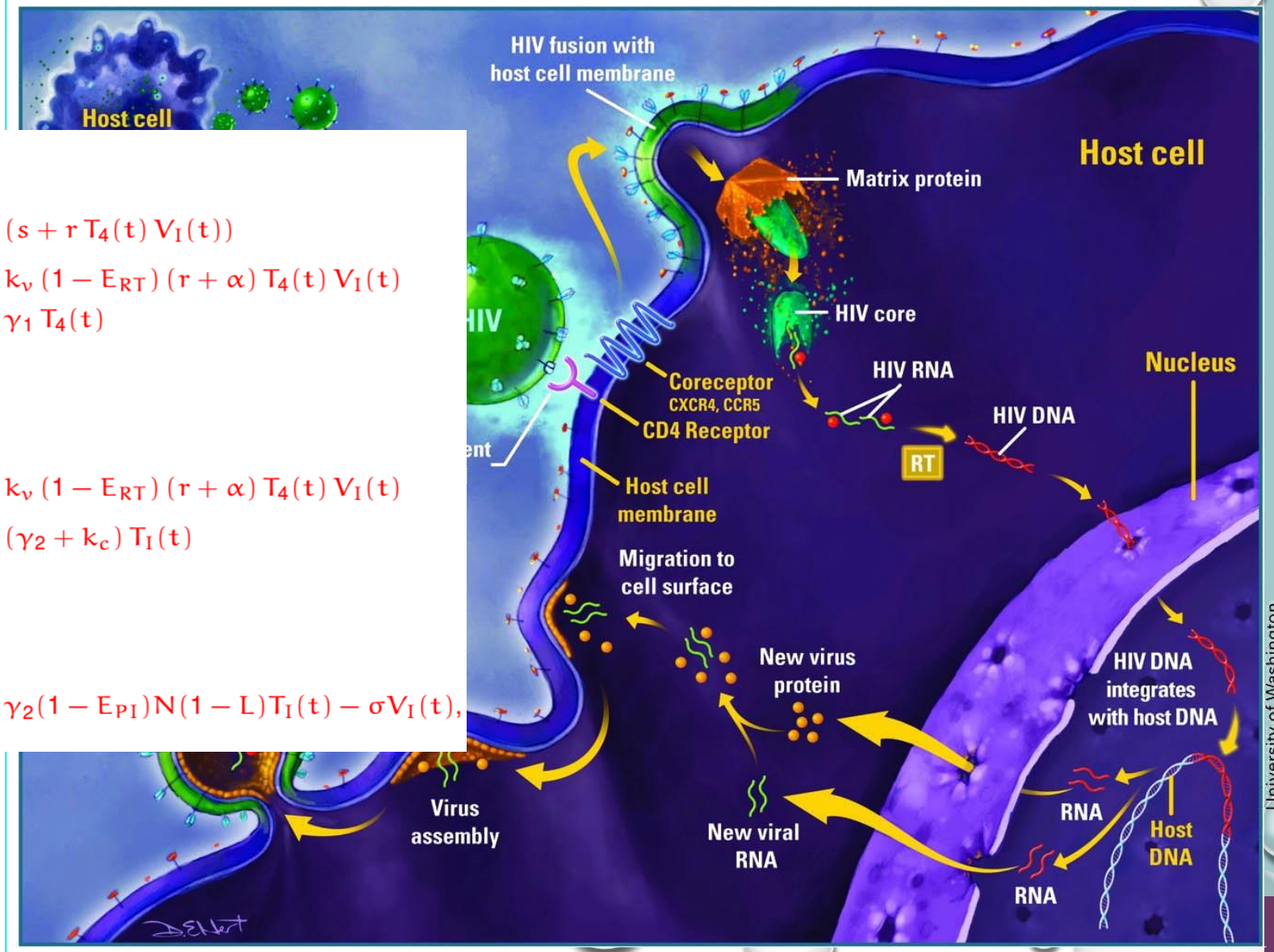
# Optimal Control (Example: HIV Dosing Regimens)




$$\frac{dT_4(t)}{dt} = (s + r\,T_4(t)\,V_I(t)) - k_v\,(1-E_{RT})\,(r+\alpha)\,T_4(t)\,V_I(t) - \gamma_1\,T_4(t)$$

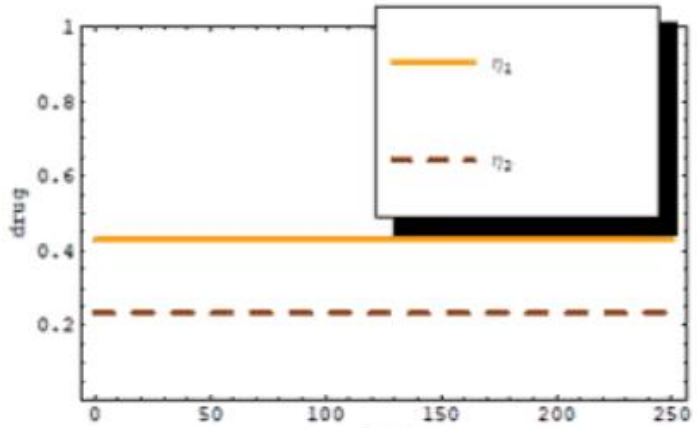
$$\frac{dT_I(t)}{dt} = k_v\,(1-E_{RT})\,(r+\alpha)\,T_4(t)\,V_I(t) - (\gamma_2 + k_c)\,T_I(t)$$

$$\frac{dV_I(t)}{dt} = \gamma_2(1-E_{PI})N(1-L)T_I(t) - \sigma V_I(t),$$

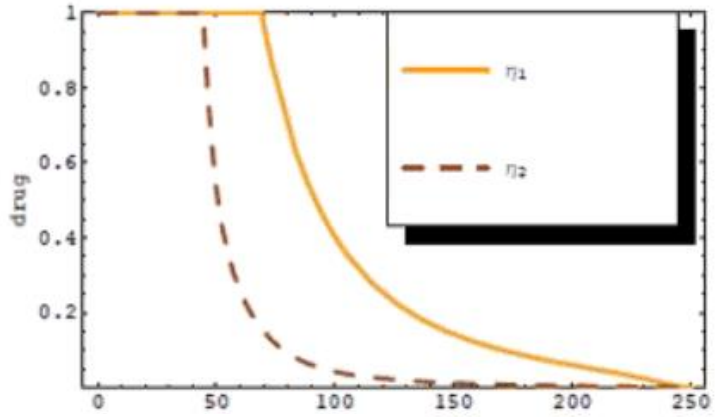


# Optimal Control (Example: HIV Dosing Regimens)

- Standard Therapy

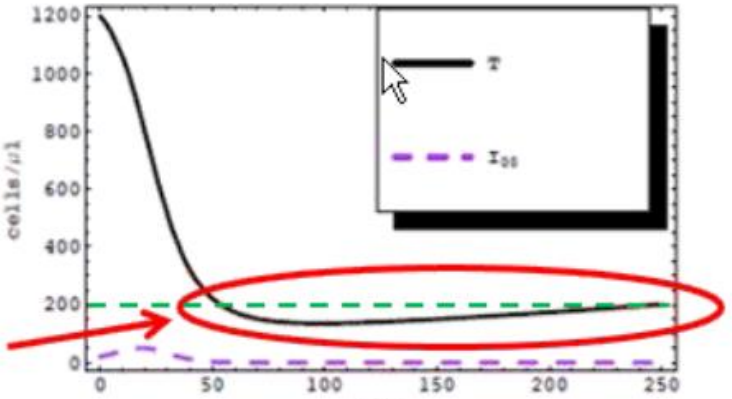


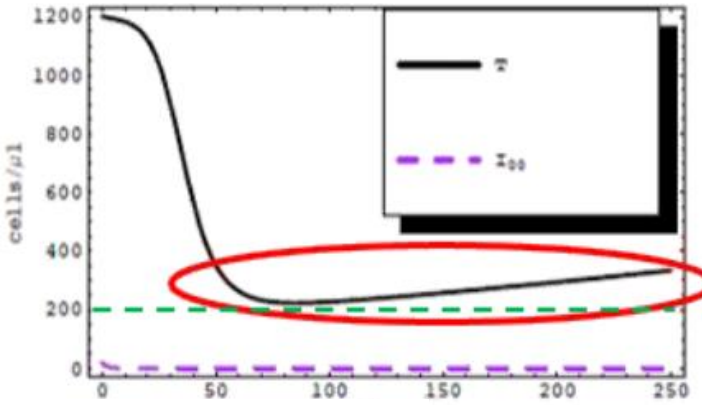

Optimal Therapy




- Total exposure (area under the curve) same for both regimens

Worse outcome

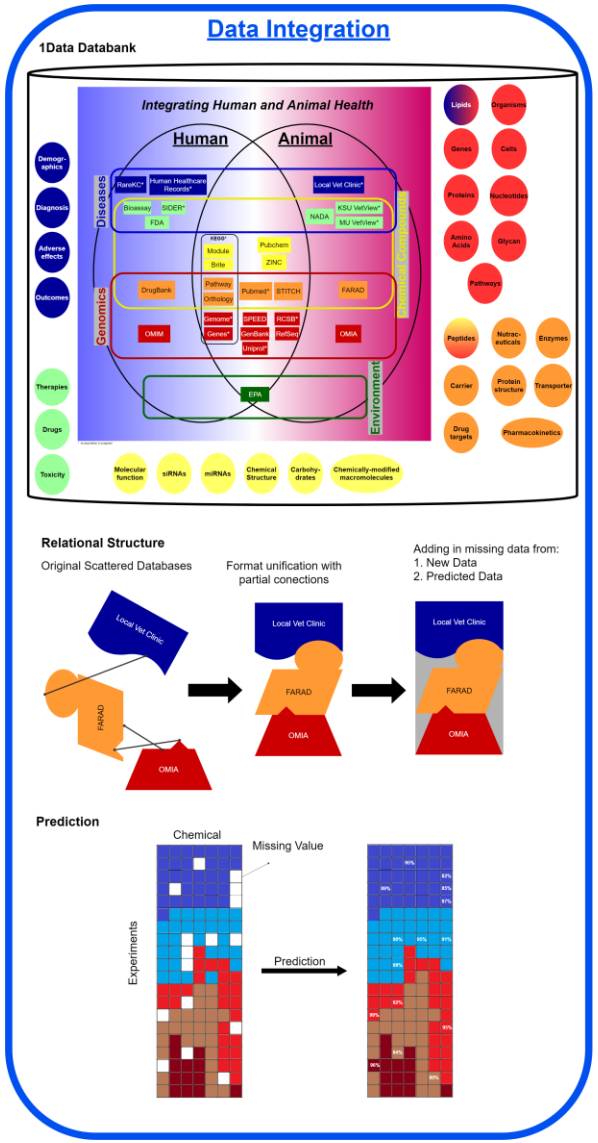


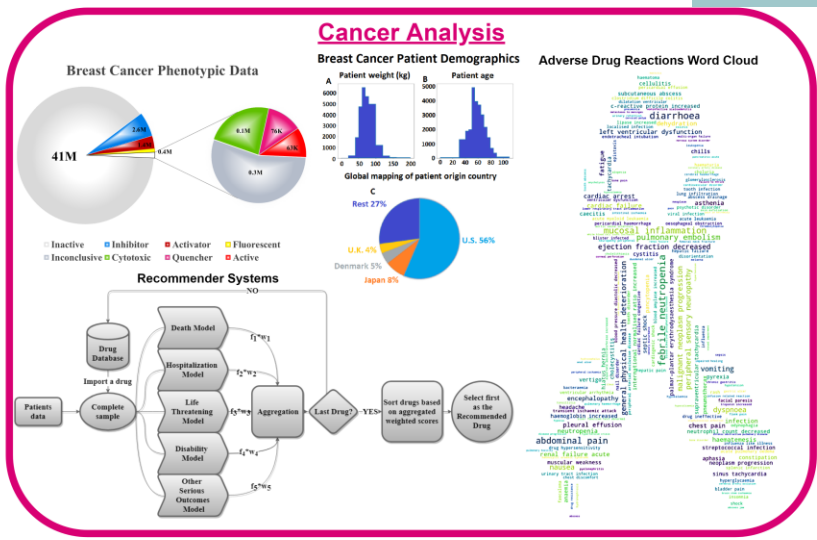

Better outcome

# 1Data

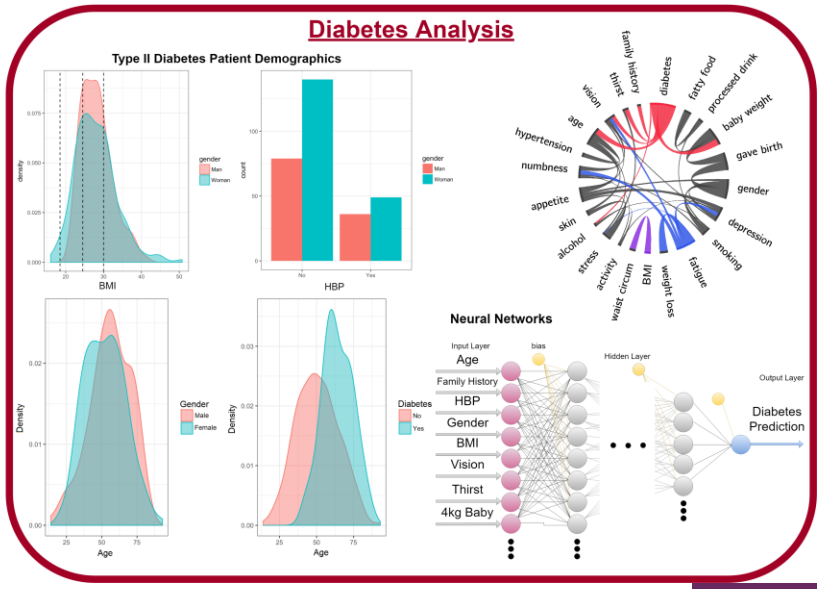
- **1Data** creates a *Structured Environment for Animal Data and Simulation* (SEADS), a key enabling technology for regional translational medicine efforts.
- **SEADS** brings together pre-clinical human and animal health data to develop, collect, and disseminate information to improve the quality of human and animal health.
- 1Data is the result of a unique partnership formed between K-State and UMKC. Other organizations also are part of this effort, including Children's Mercy Hospital, KUMED, Cerner, Aratana, MRIGlobal, KCALSI, St. Luke's Health System and many others.




Analysis



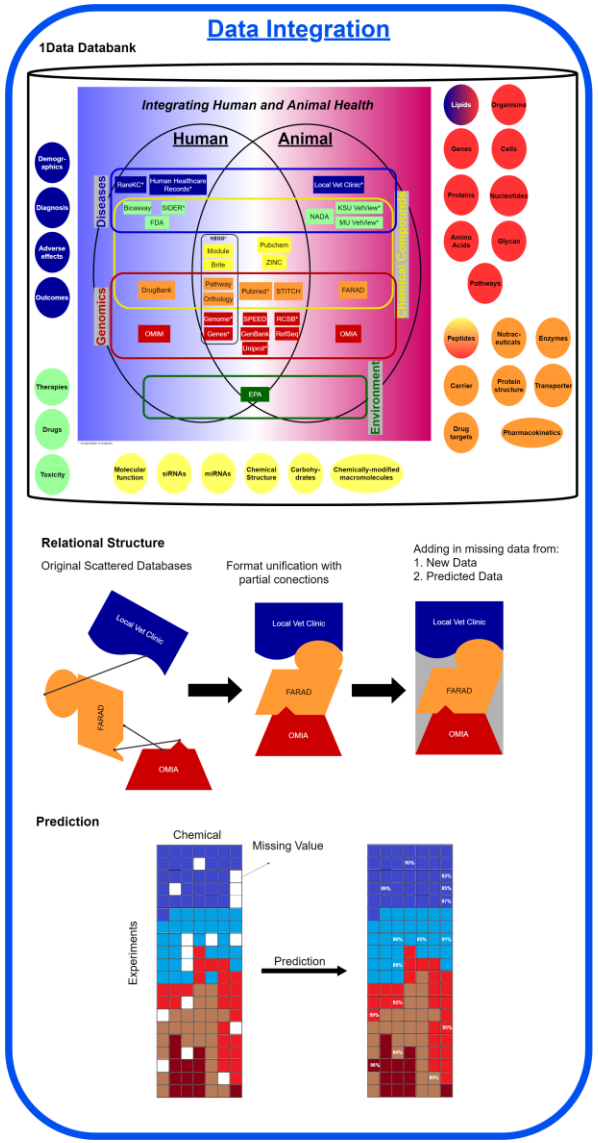

Analysis



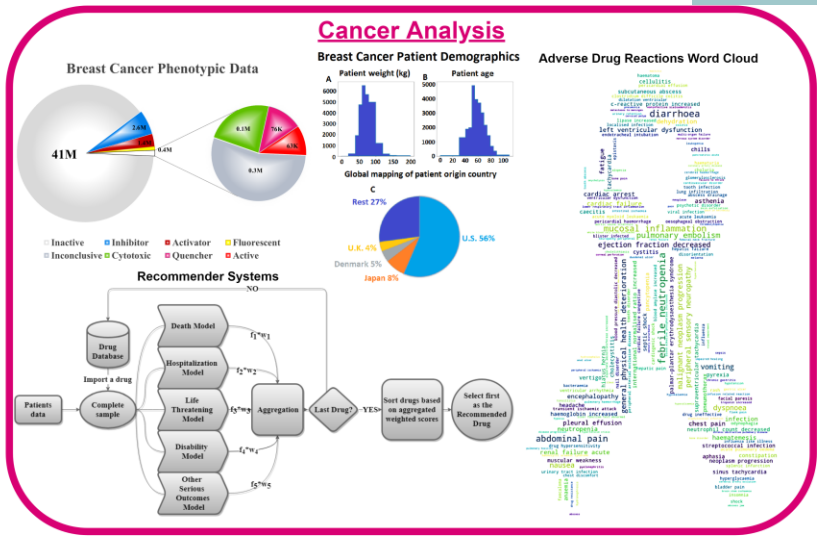
# Vision

The platform can be used by researchers, industry, health providers and community organizations to impact the drugs and technology available to help save lives and improve quality of life. 1Data allows for mining of shared data that can:

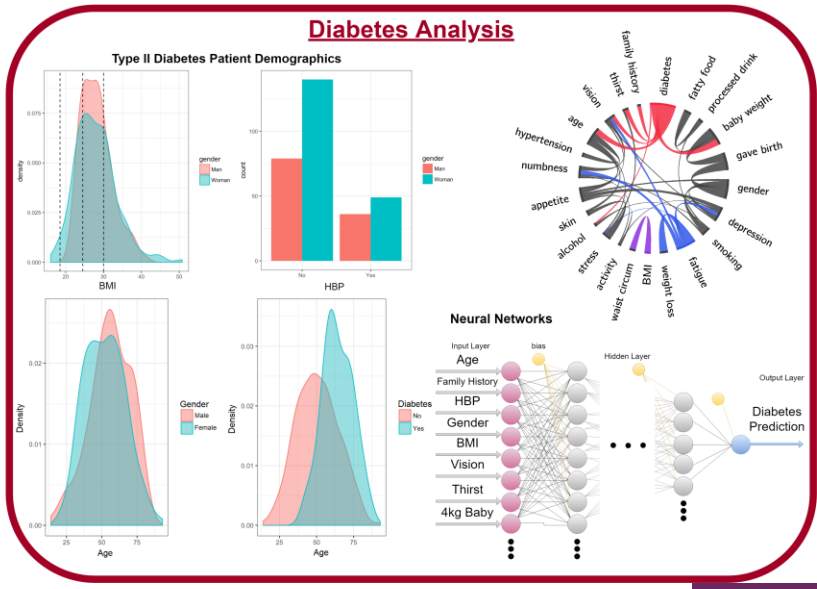
- Accelerate the development of human and animal drugs
- Enhance the regulatory approval process
- Decrease the use of animal testing




Analysis

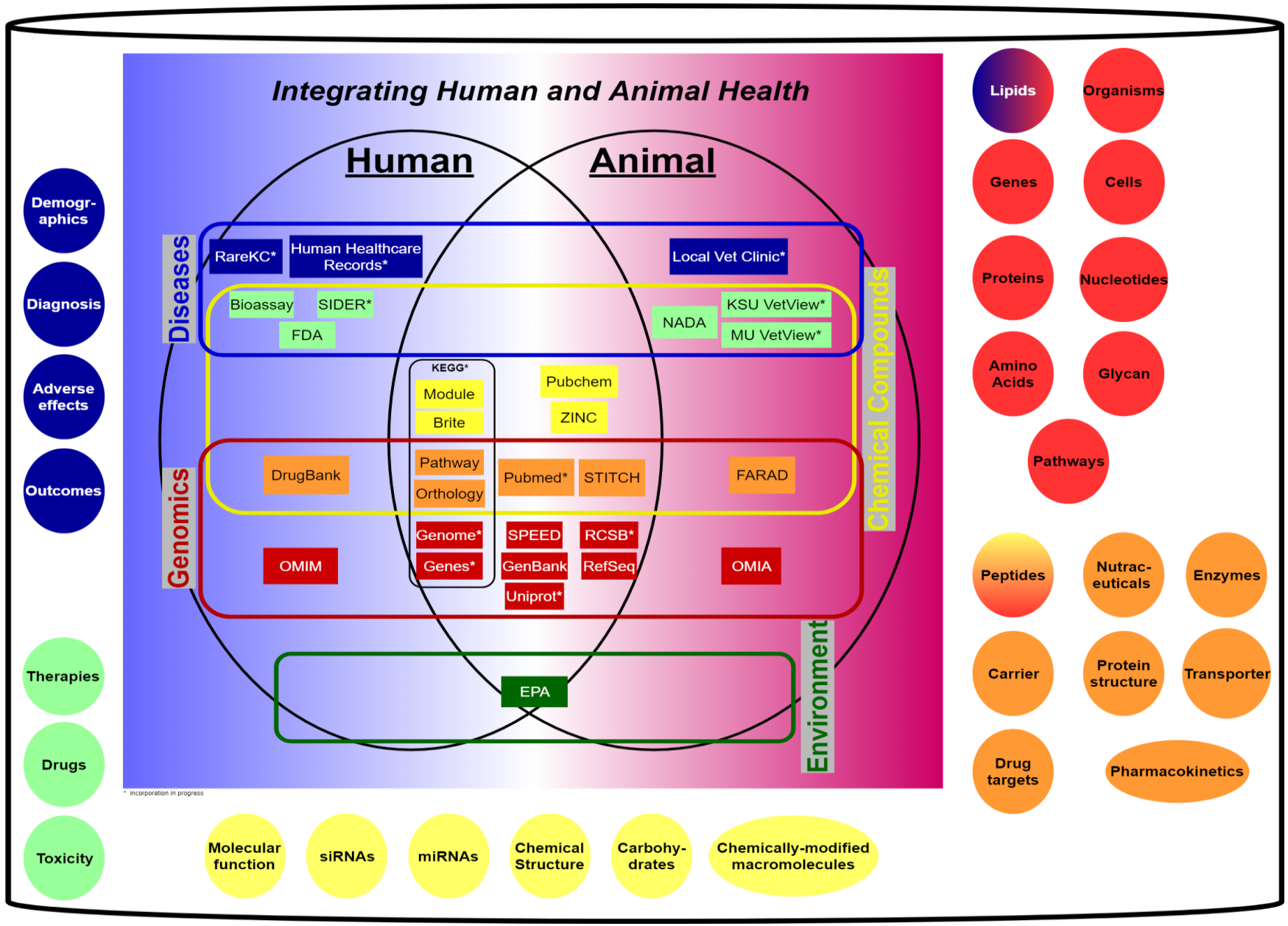



Analysis



# Database

## 1Data Databank



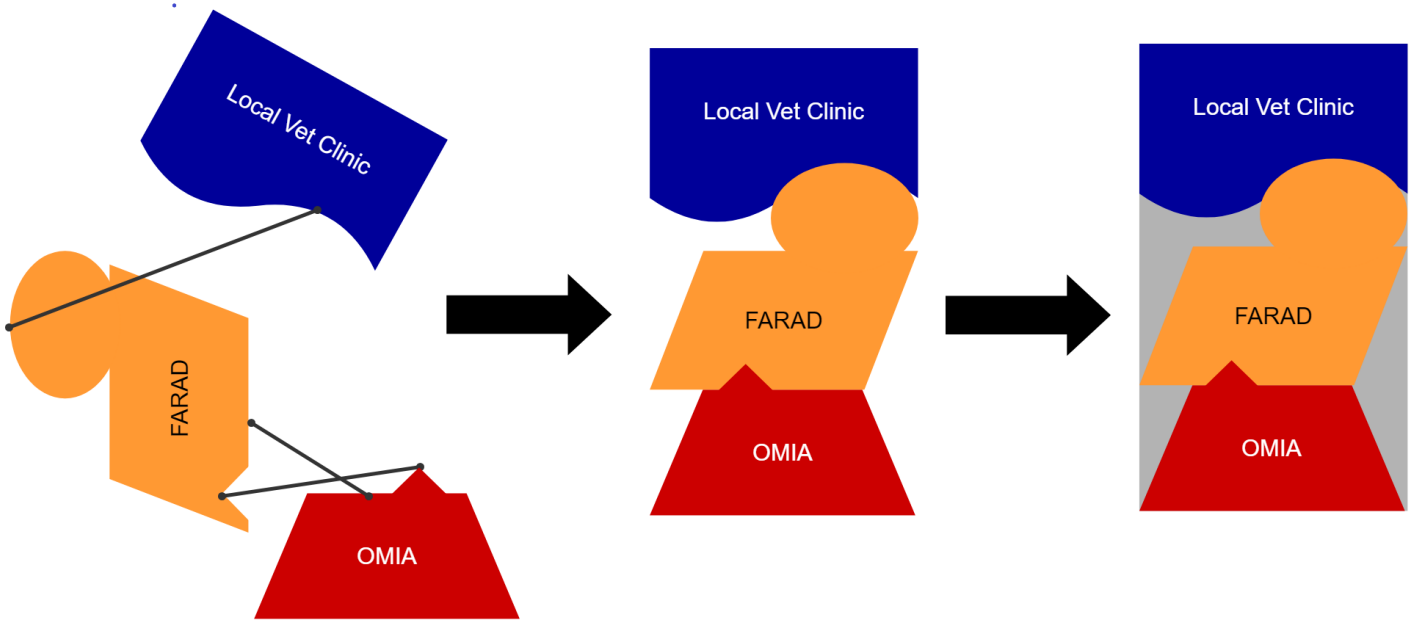
# Data Curation

**Relational Structure**

Original Scattered Databases

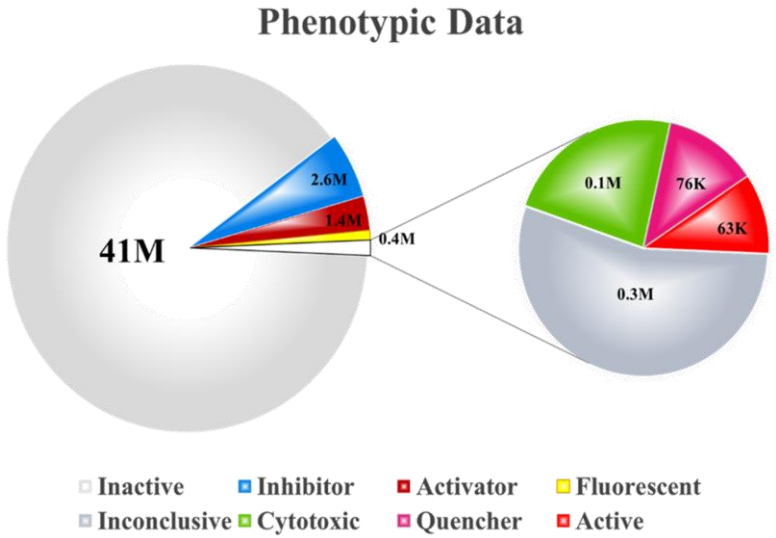
Format unification with partial conections

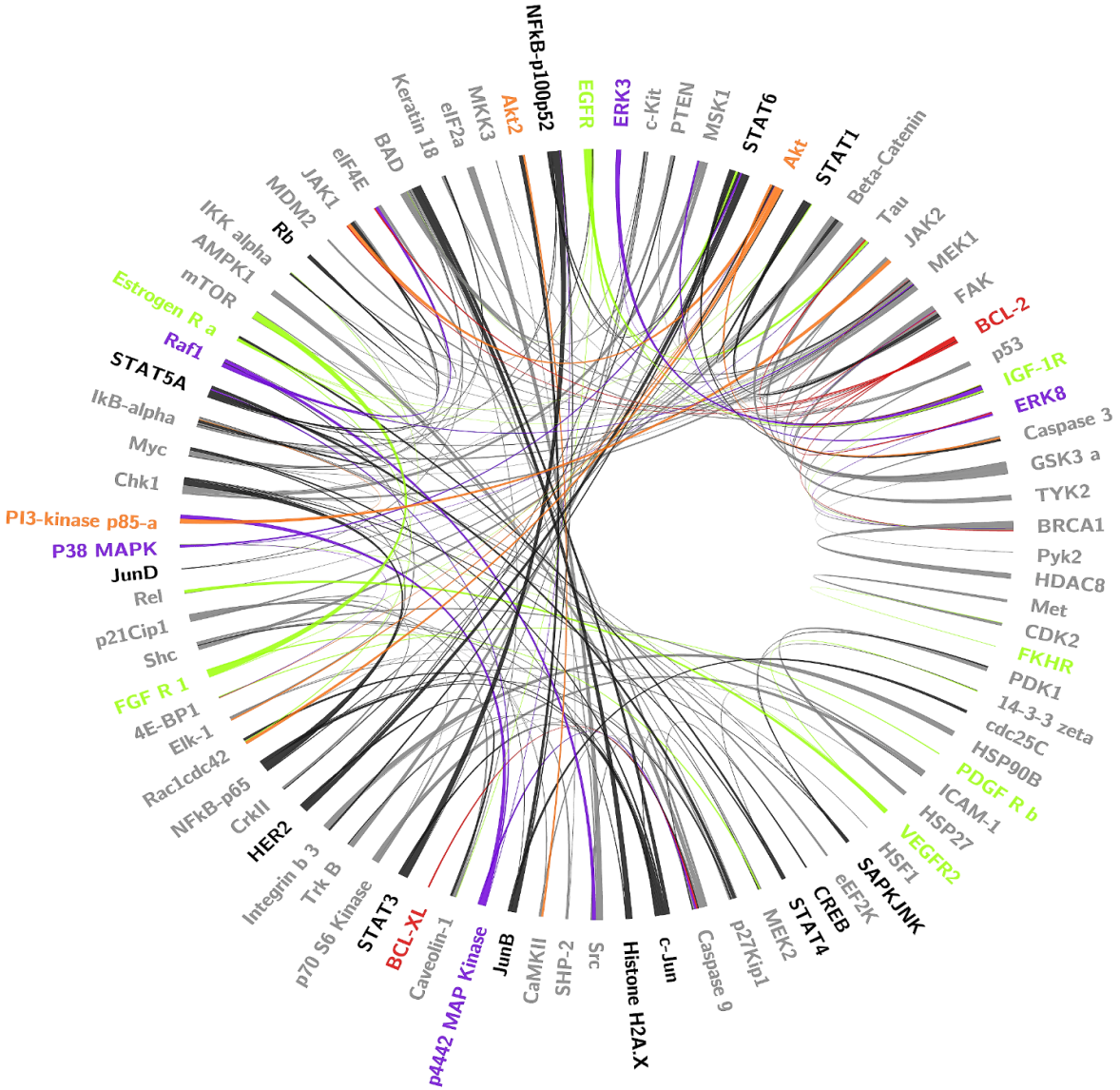
Adding in missing data from:
1. New Data
2. Predicted Data



# Predictive Modeling of the PubChem Bioassay Database

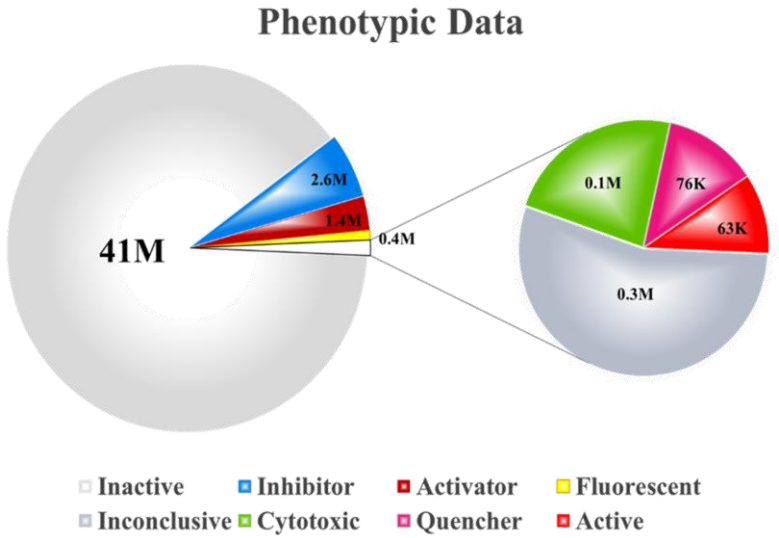
- 28 chemicals common in the inhibition of the ADAM family of proteases, an enzyme responsible for amplification of HER2 signal, present in 20-30% of breast cancer patients.

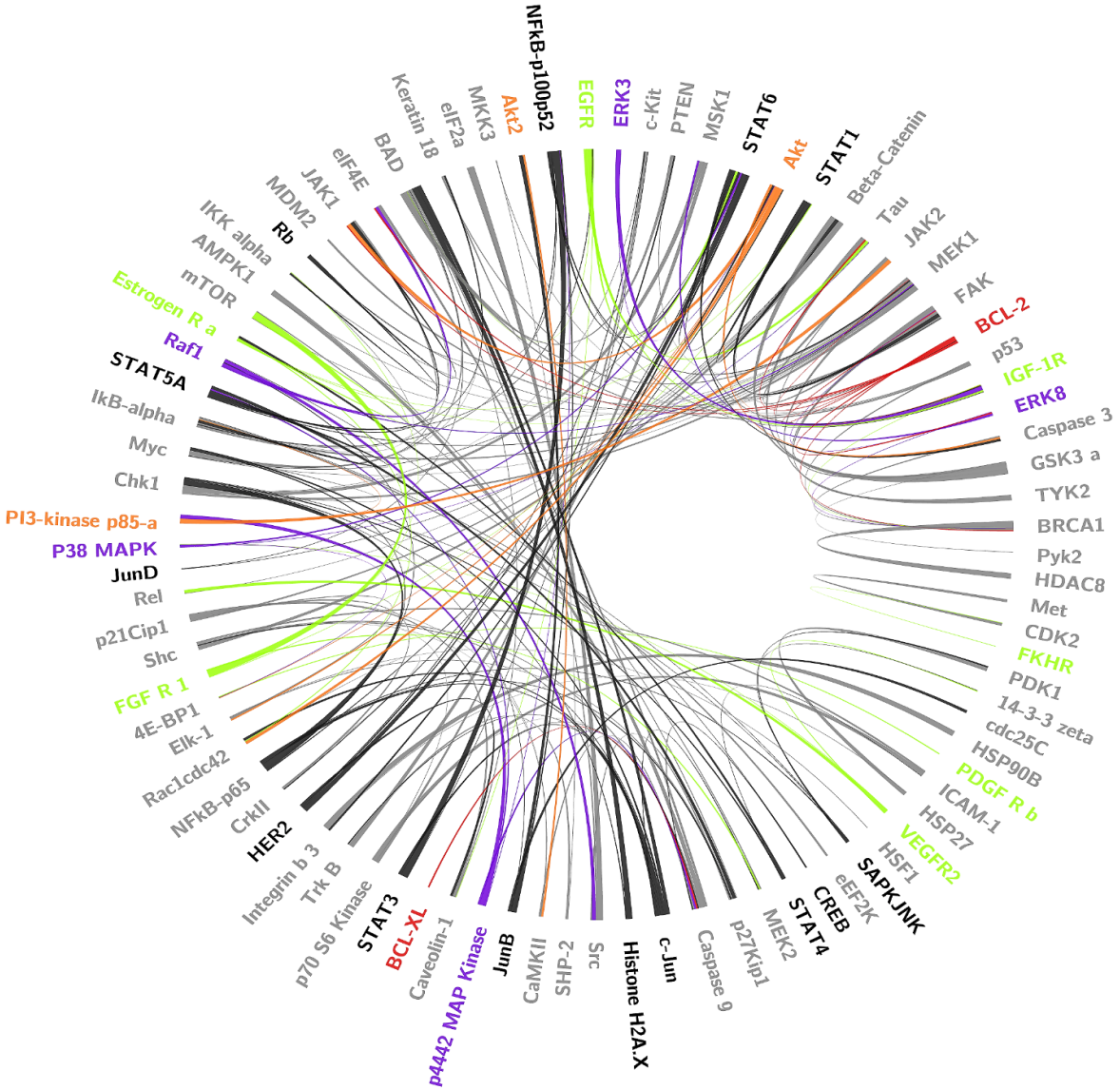


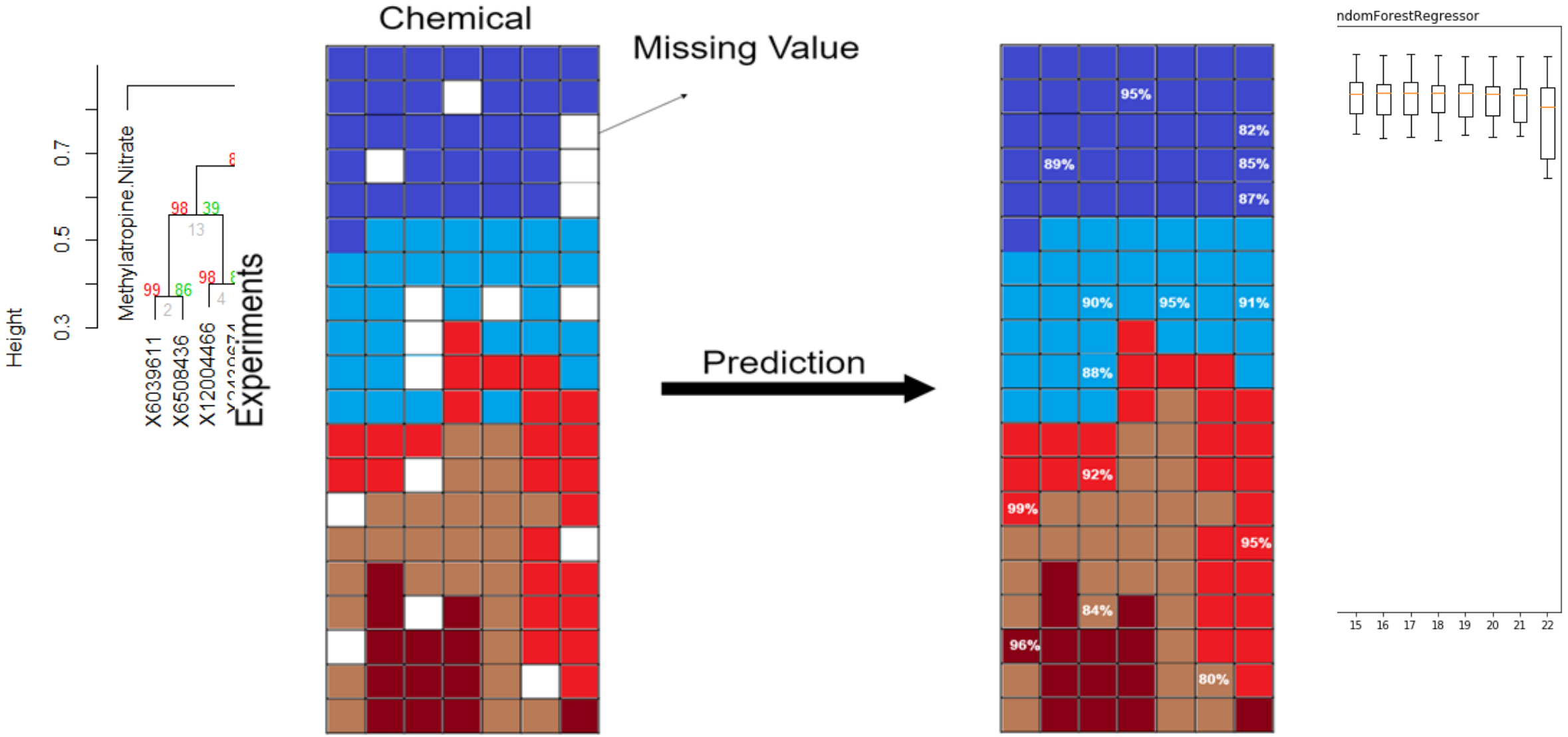
# Predictive Modeling of the PubChem Bioassay Database

- Find all experiments that contained activity score data for the 28 chemicals.
- Data from different species including Human, Cattle, Dog, Horse, Swine, …



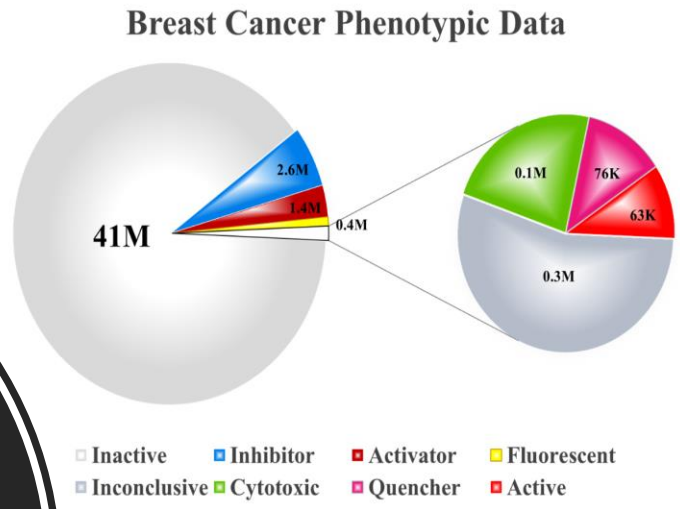



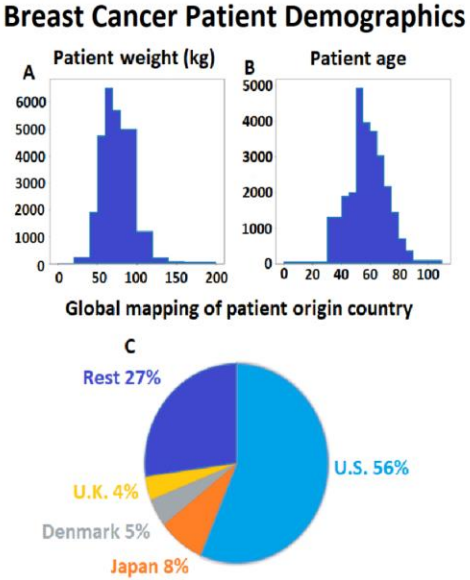
# Predictive Modeling of the PubChem Bioassay Database

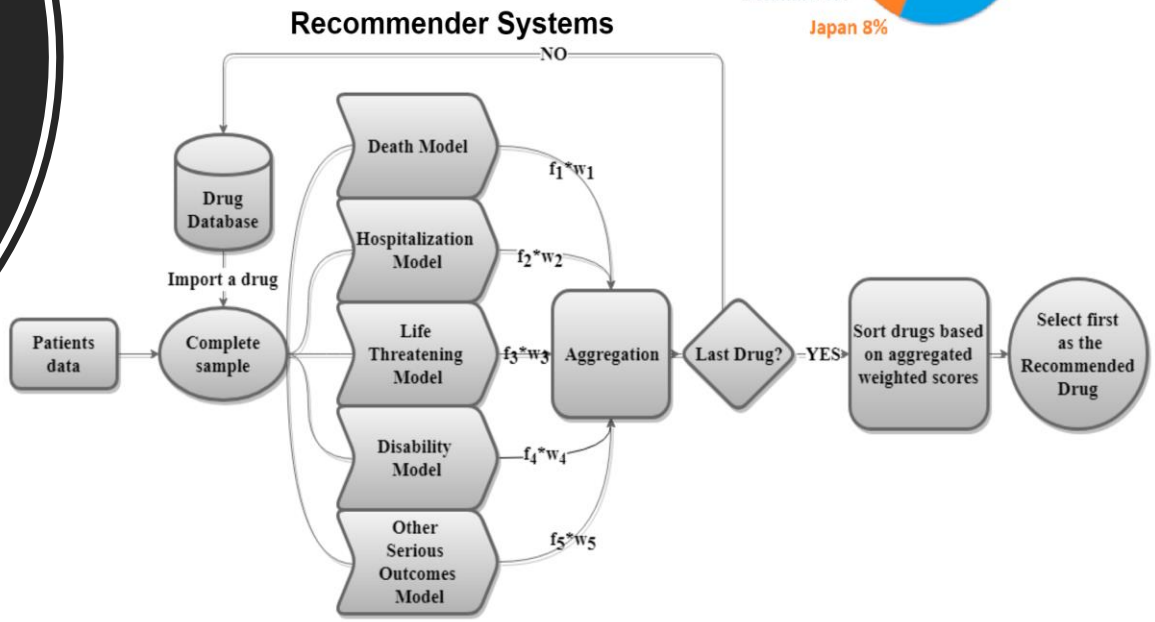


# Predictive models for survivability and adverse drug reactions

# Predictive models for survivability and adverse drug reactions



# Background

According to global cancer statistics for 2018 (GLOBOCAN) there is an estimation of 9.6 million death records due to cancer in 2018.

Lung cancer has been seen as the leading cancer among both men and women. In the second place, breast cancer is the major cancer among all type of cancers and is the first deadliest cancer between women in US.

Breast cancer does not target only females as nearly 2000 male cases were diagnosed between 2003 and 2008.

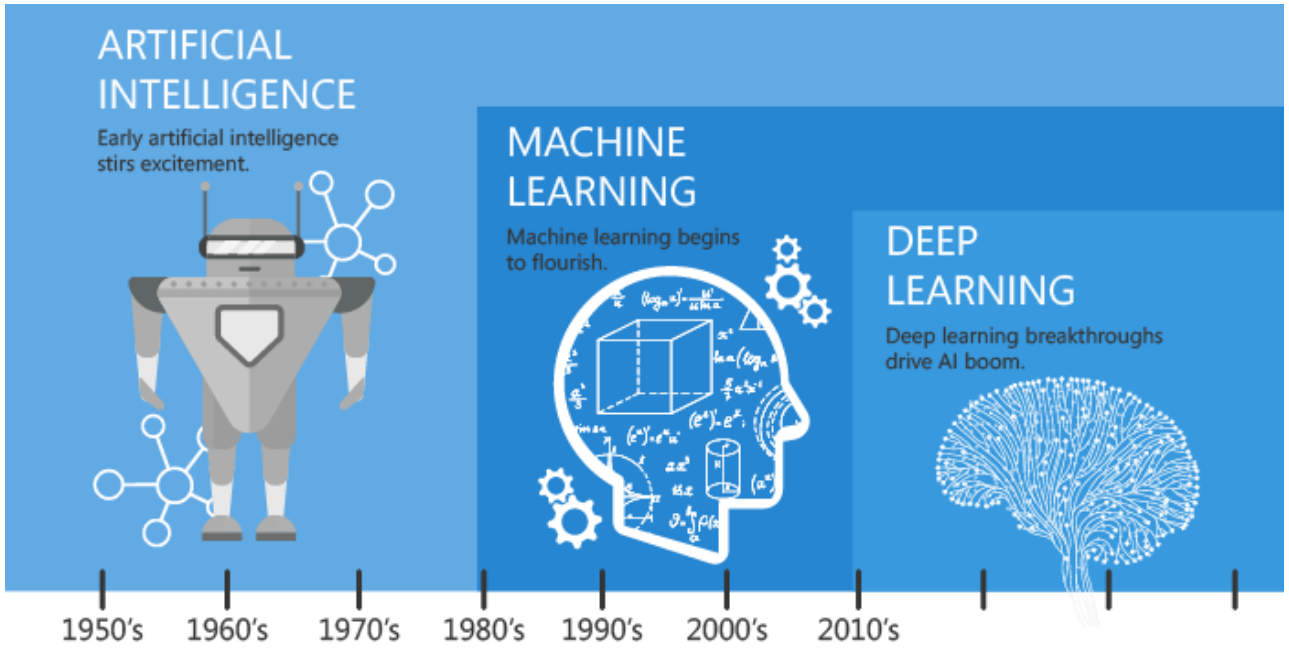
Detecting the risk level associated with the patient can help the treatment procedure as risky patients could be monitored more.

# Goal

- To increase the accuracy of expert's opinion to reduce wrong decisions which in cases could be fatal.

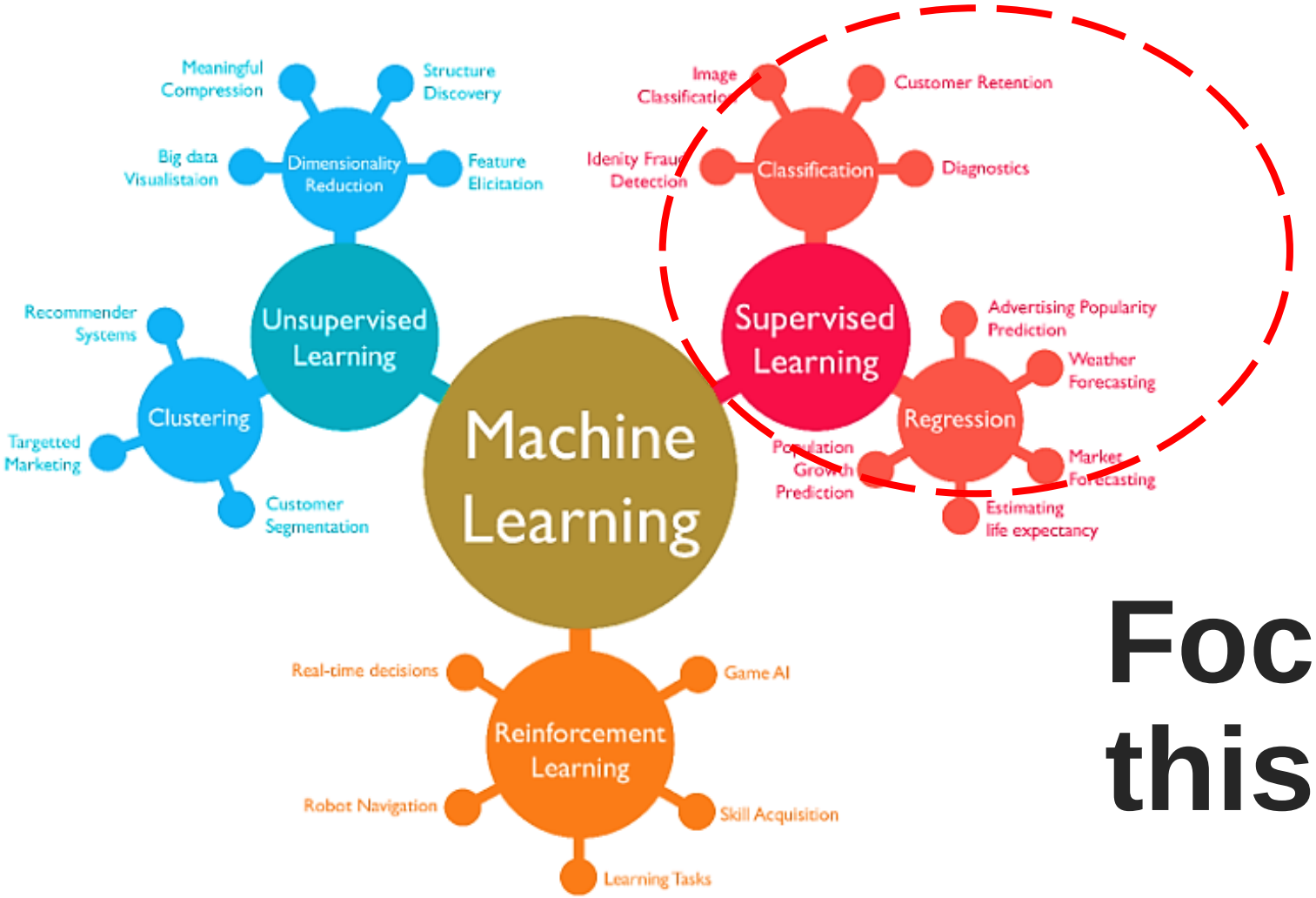
"I knew it was a rash decision, but I went with it anyway."

# Background




- Machine learning is a subset of artificial intelligence (AI) where data is used to make decisions.

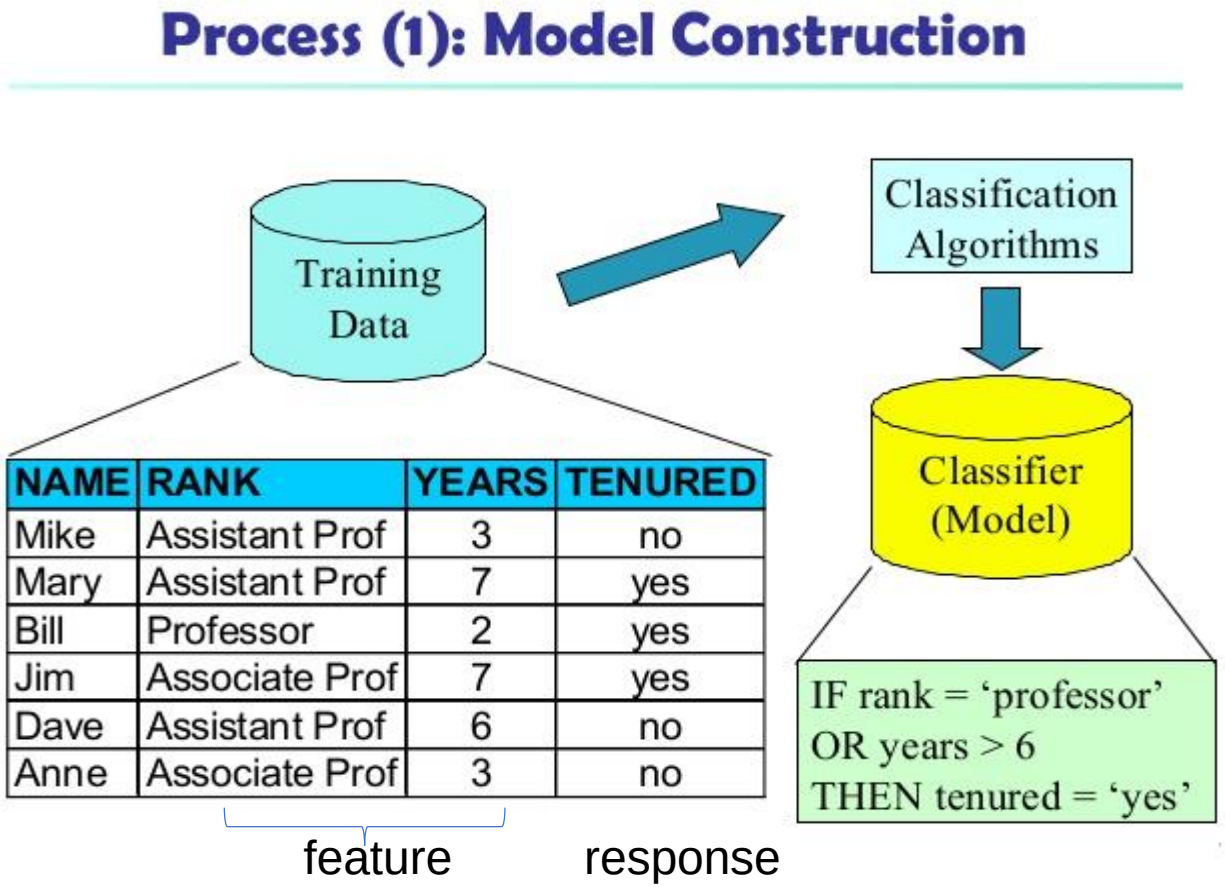
Pic Credit: Linked IN | Machine Learning vs Deep Learning




# Focus of this work

Pic Credit: chatbotsmagazine

# Background

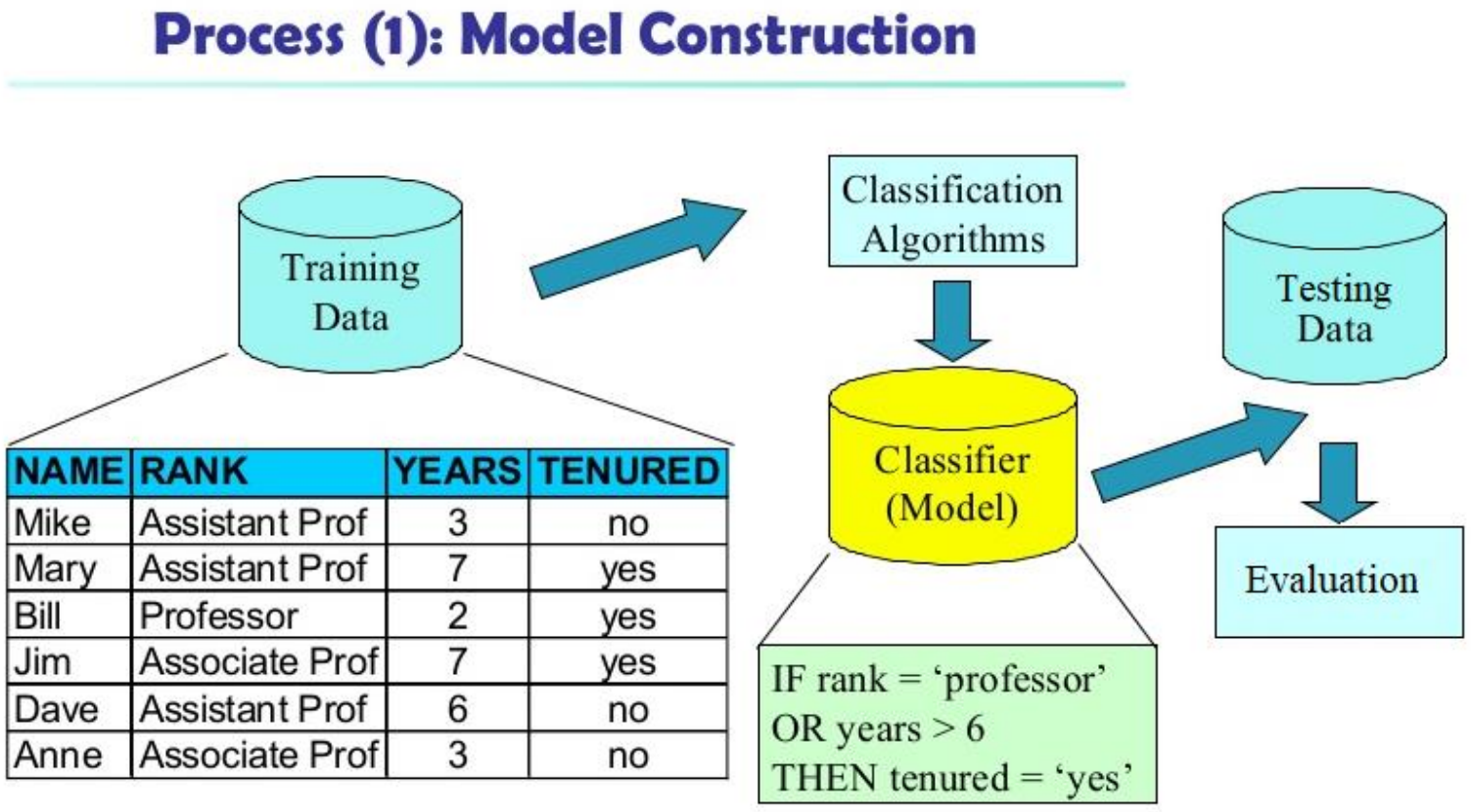



| NAME | RANK | YEARS | TENURED |
|---|---|---|---|
| Mike | Assistant Prof | 3 | no |
| Mary | Assistant Prof | 7 | yes |
| Bill | Professor | 2 | yes |
| Jim | Associate Prof | 7 | yes |
| Dave | Assistant Prof | 6 | no |
| Anne | Associate Prof | 3 | no |

$$\widehat{Y}_i = f(X_{i1}, X_{i2}, \ldots, X_{im})$$

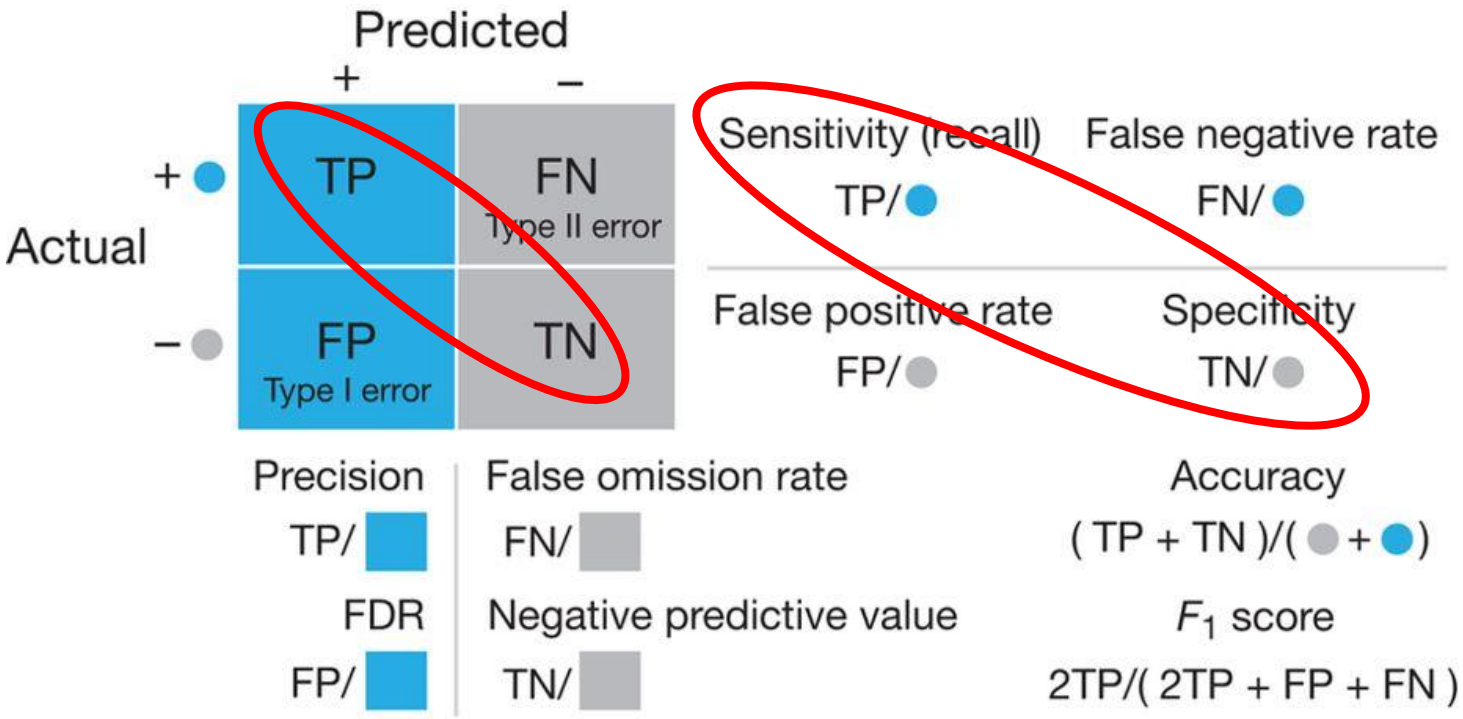
- Decision Tree
- SVM
- Bayes Models
- Neural Networks
- Logistic Regression
- ...

# Background




| NAME | RANK | YEARS | TENURED |
|---|---|---|---|
| Mike | Assistant Prof | 3 | no |
| Mary | Assistant Prof | 7 | yes |
| Bill | Professor | 2 | yes |
| Jim | Associate Prof | 7 | yes |
| Dave | Assistant Prof | 6 | no |
| Anne | Associate Prof | 3 | no |

# Confusion Matrix



openFDA

receivedate from 1989-12-07 to 2018-06-30

Dashboard

Product

Drug Variable

patient.drug.openfda.generic_name

Select Drug...

Drug Name: *ASPIRIN*

Match drug name:

Exactly  Any Term

Product Summary

Table | Dotchart | Piechart

| | Serious | Case Counts | Code | % |
|---|---|---|---|---|
| 5 | Lawyer | 3,725 | 4 | 1.4% |
| 4 | Pharmacist | 21,303 | 2 | 7.7% |
| 3 | Other Health Professional | 55,090 | 3 | 20.0% |
| 2 | Physician | 78,296 | 1 | 28.4% |
| 1 | Consumer or non-health... | 116,818 | 5 | 42.4% |

Table | Dotchart | Piechart

| | Serious | Case Counts | % |
|---|---|---|---|
| 1 | Congenital Anomaly | 430 | 0.15% |
| 3 | Disability | 11,001 | 3.76% |
| 5 | Life Threatening | 15,011 | 5.13% |
| 2 | Death | 27,298 | 9.33% |
| 6 | Other | 113,833 | 38.91% |
| 4 | Hospitalization | 124,945 | 42.71% |

Adverse Events and Concomitant Medications

Events | Concomitant Medications | Indications | Other Apps | Data Reference | About

Tables | Word Cloud

# Dataset

- FAERS (FDA Adverse Event Report System, 2004-2018)
- FDA CVM Adverse Drug Experience (ADE) Reports (Between 1987 And April 30, 2013)

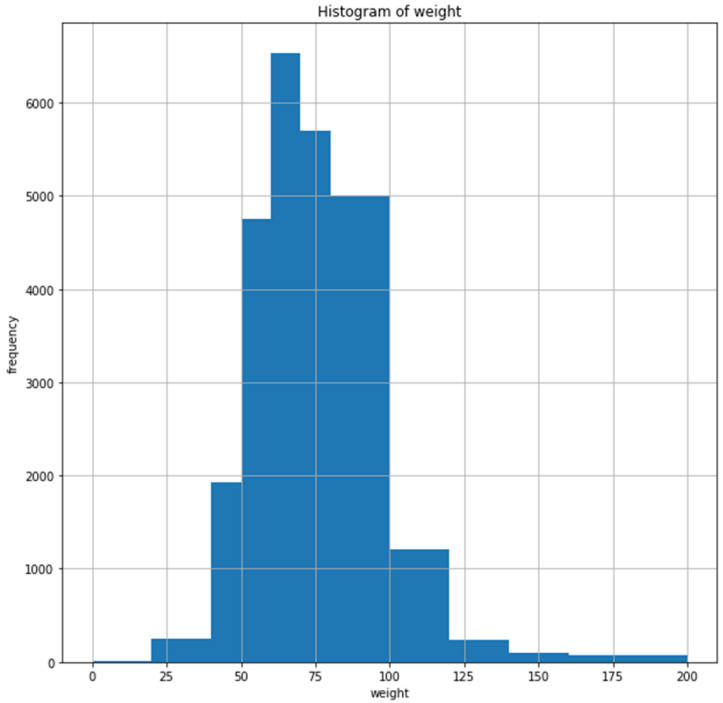
# Dataset

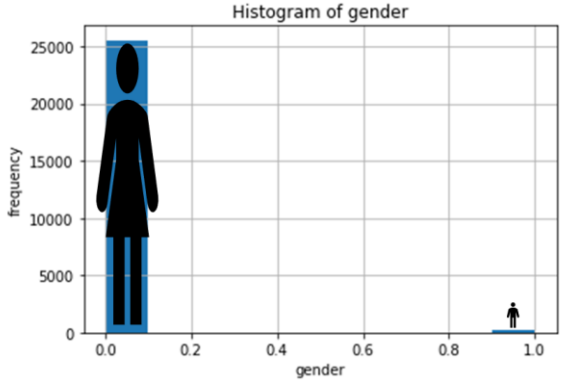
Data cleaning is >80% of the job

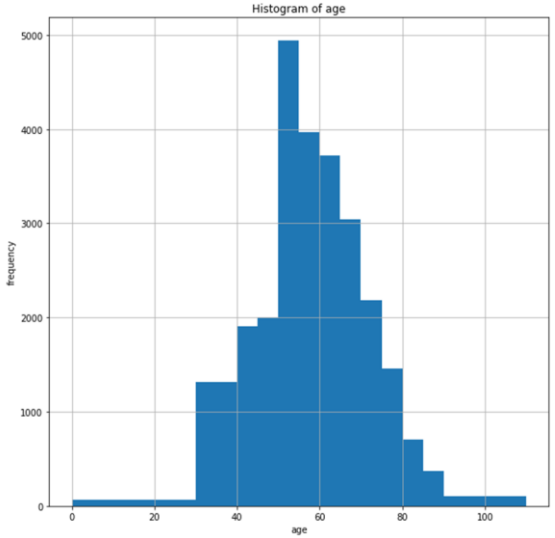
| Patient | Drug | ation | … |
|---|---|---|---|
| X1 | doce | | … |
| X2 | faslo | | … |
| X3 | Arim | | … |

FAERS (FDA Adverse Event Report System) dataset contains:

- Patient info: patient id, age, gender, weight, country,…
- Drug info: name, id, sequence, dosage, company,…
- Visit info: visit id
- Outcomes: death, hospitalization, life-threating, disability, other serious issues,..

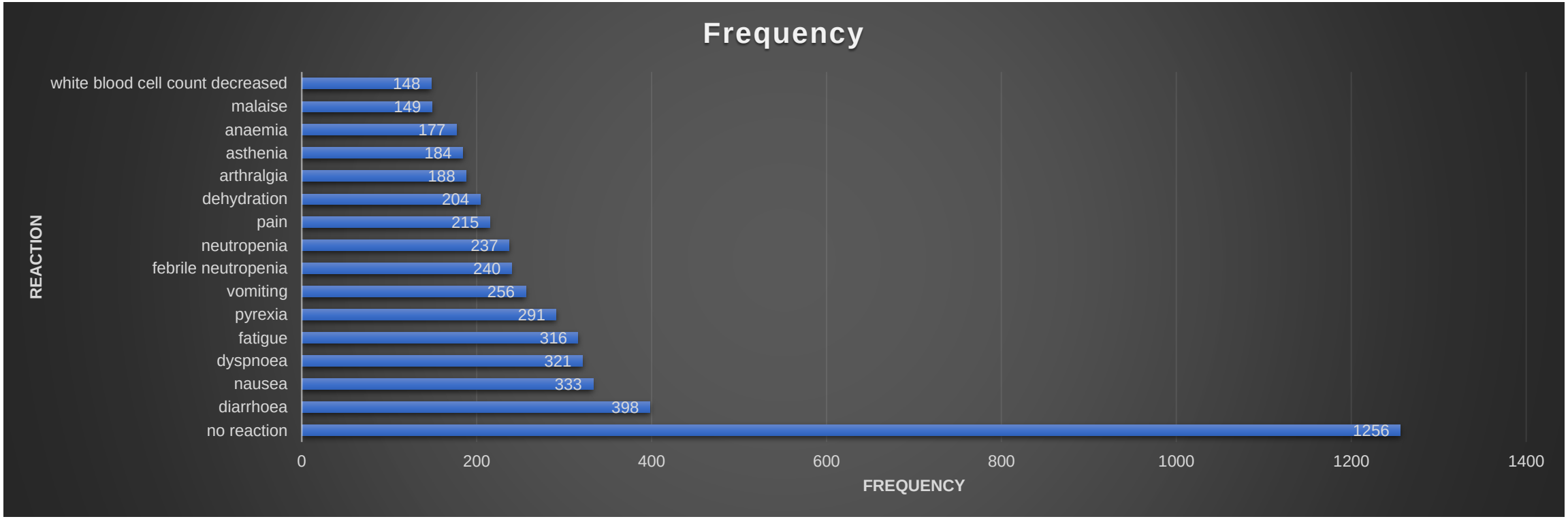


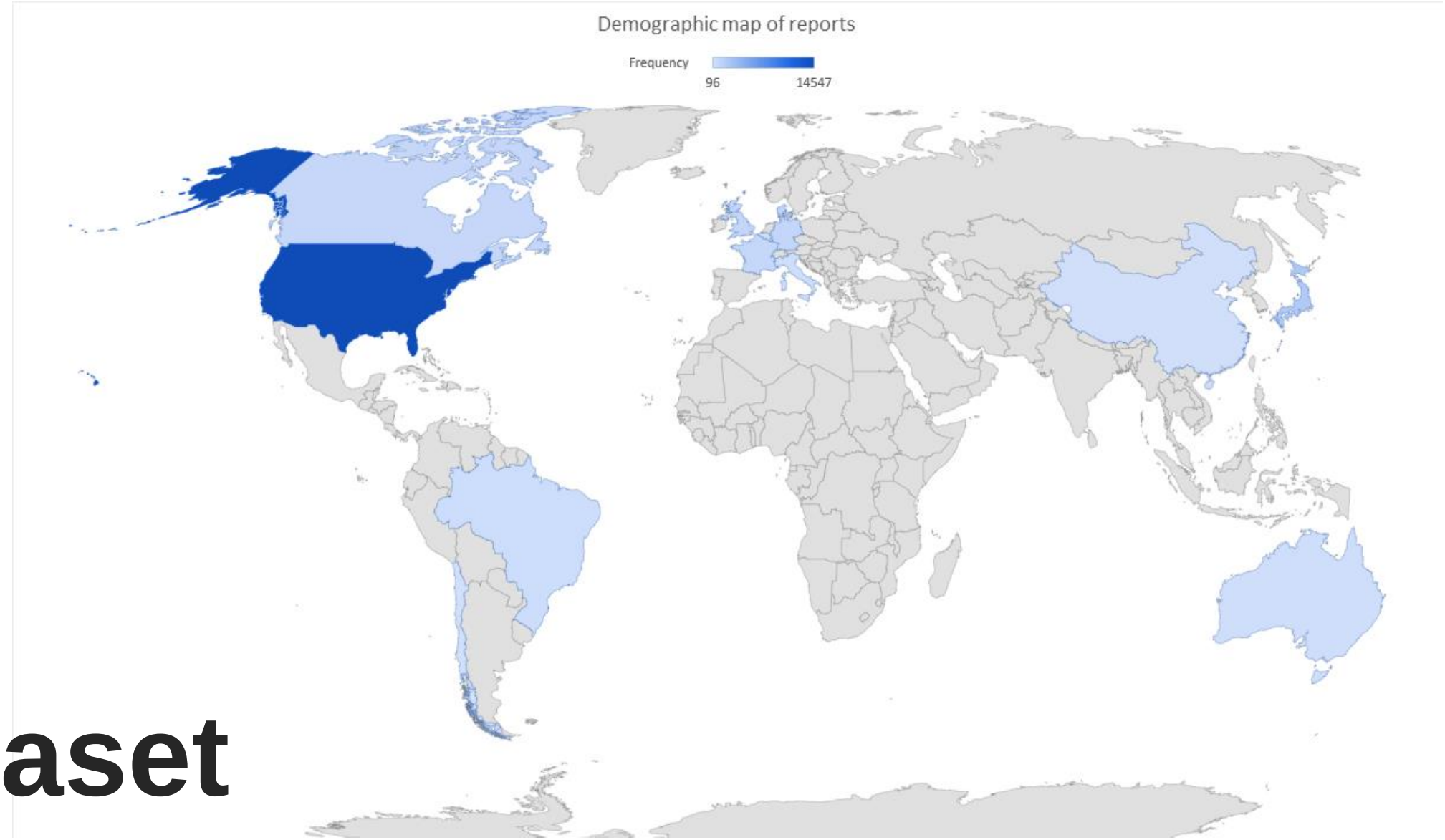

# Dataset

Histograms of ages, weights, and genders of patients




# Dataset

## PLOT OF REACTIONS




# Dataset

DEMOGRAPHIC MAP

# Predictive models – survivability

Predictive models have used features as:

- Age (numeric)
- Gender (categorical)
- Weight (numeric)
- Prescribed drugs (categorical)
- Outcomes (categorical) (In survivability model is fixed)
- Death record is the response :0 or 1

# Predictive models - survivability

Predictive models do not work with strings (like reaction “nausea”).

The strings need to be represented by numbers. However, this simply makes confusion in predictive models.

Therefore, we have further treated the categorical features as binary features. For example, a feature such as disease with thousands of records, has been broken into pieces like:

- cancer-yes, cancer-no, osteopenia-yes, osteopenia-no ,…
  - Advantage: Easy to understand
  - Disadvantage: Higher memory usage, large matrix

Other approaches: Feature hashing, high/less importance categories, hybrid models…

- Advantage: Less memory usage
- Disadvantage: Hard to understand

This makes a huge matrix with many columns (computationally columns are more complex than rows!)

- It has increased the number of columns from 13 to **617**

# Predictive models - survivability

- Predictive model for **Breast cancer** (male & female):
  - A **Decision Tree** has been trained on 90% of all available data. 10% of the records have been reserved for evaluation of the model. (**for all stages**)
  - The predictive model acquires 95.6% accuracy with confusion matrix:

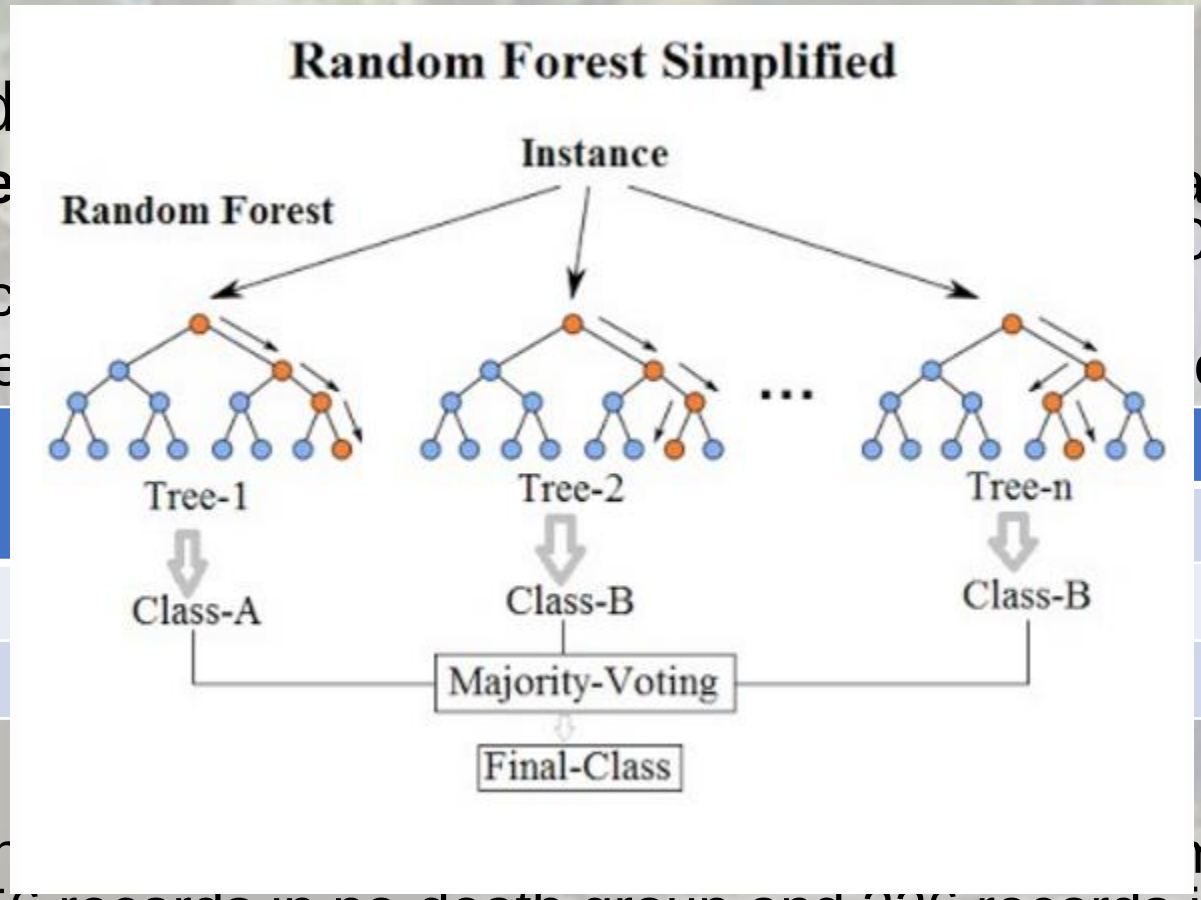
| **Actual** | **Prediction** | |
|---|---|---|
| | No death | Death |
| No death | 2240 | 86 |
| Death | **25** | 226 |

  - This means that, in the testing dataset (2577 records), the model correctly predicted 2240 records in no-death group and 226 records in death group. However, the model has misclassified 86 records in death group (while they are not actually in death group) and 25 records in no-death group (while they are actually in death group).

# Predictive models - survivability

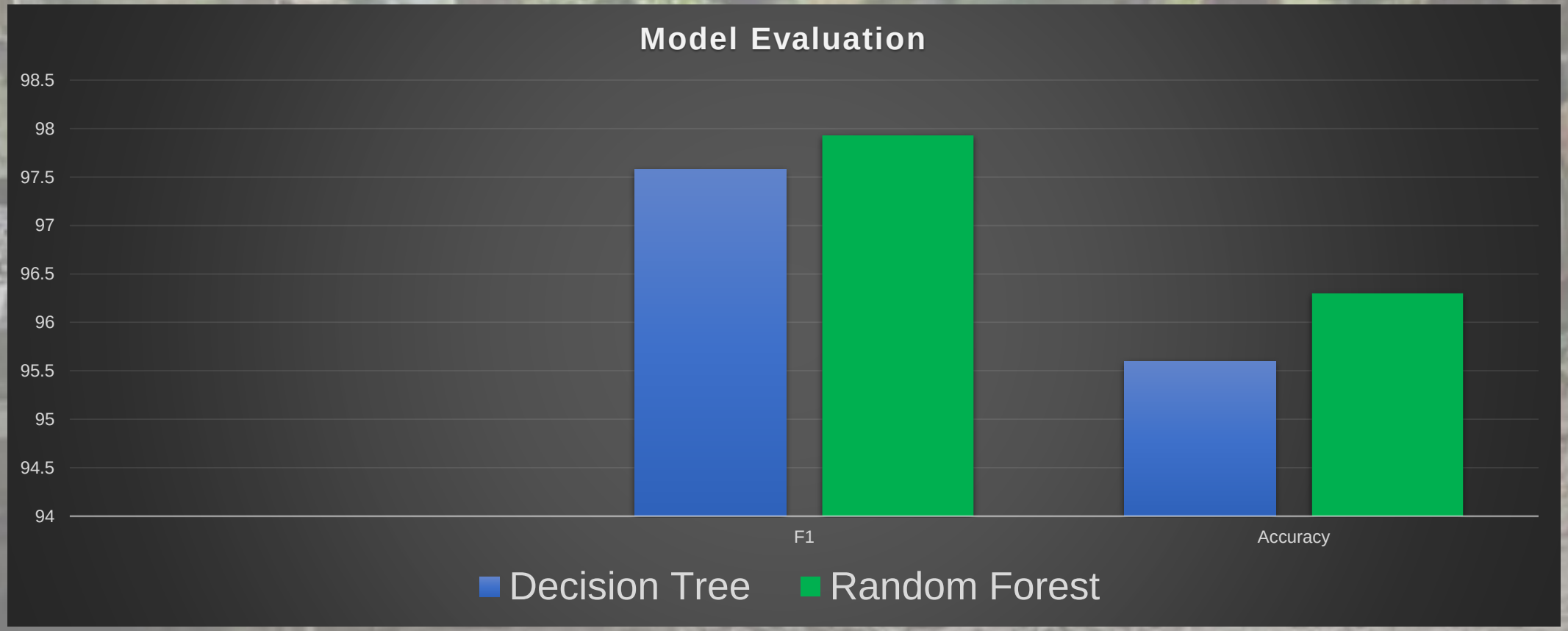
- Predictive mod
  - **Random fore** ata. 10% of the records have 00 trees, class-weight-balanc
  - The predictive on matrix:

| Actual | | |
|---|---|---|
| No death | | |
| Death | | |




  - This means th model correctly predicted 2256 records in no-death group and 226 records in death group. However, the model has misclassified 70 records in death group (while they are not actually in death group) and 25 records in no-death group (while they are actually in death group).

# Predictive models - survivability



# Correlation Analysis

- Correlation analysis on patients having **breast cancer** (**stage 1**):
  - Using data mining approaches, a correlation of records with **serious outcomes** (no death record) and drugs, weight, gender, age, and reactions has been obtained.
  - Drugs:
    - Drug Aromasin has the highest correlation (58%)
      - Aromasin has been repeated 2 times in total(20) and only to patients with serious issues.
        - Arimidex has been repeated 3 times in patients with serious issues.
    - Drug Herceptin has the lowest correlation (-21%).
      - Herceptin has been repeated 3 times only with patients with no serious issue.
  - Reactions:
    - “trigeminal neuralgia” has the highest correlation (40%)
    - “palpitations” has the least correlation (-17%).
  - Age and weight have -7% and -35% correlation and gender is NAN (no male).

# Correlation Analysis

- Correlation analysis on patients having **breast cancer** (**stage 2**):
  - Using data mining approaches, a correlation of records with **serious outcomes** (no death record) and drugs, weight, gender, age, and reactions has been obtained.
  - Drugs:
    - Drug trastuzumab emtansine has the highest correlation (44%)
      - Trastuzumab emtansine has been repeated 4 times only with patients with serious issues.
    - Arimidex has the lowest correlation (-24%).
      - Arimidex has been repeated 29 times in 74 reports and 27 of those times have been for patients with no serious issues.
  - Reactions:
    - “neutrophil count decreased” has the highest correlation (44%)
    - “arthralgia” has the least correlation (-11%).
  - Age and weight have -2% and 33% correlation and gender is NAN (no male).

# Correlation Analysis

- Correlation analysis on patients having **breast cancer (stage 3**):
  - Using data mining approaches, a correlation of records with **serious outcomes** (no death record) and drugs, weight, gender, age, and reactions has been obtained.
  - Drugs:
    - Drug faslodex has the highest correlation (48%)
      - Faslodex has been repeated 2 times only for patients with serious outcomes.
    - radiation therapy has the lowest correlation (-10%).
      - Radiation therapy has been repeated 7 times only to patients with no serious issues.
  - Reactions:
    - “blood oestrogen decreased” has the highest correlation (34%)
    - “arthralgia” has the least correlation (-7%).
  - Age and weight have 20% and -0.5% correlation and gender is NAN (no male).

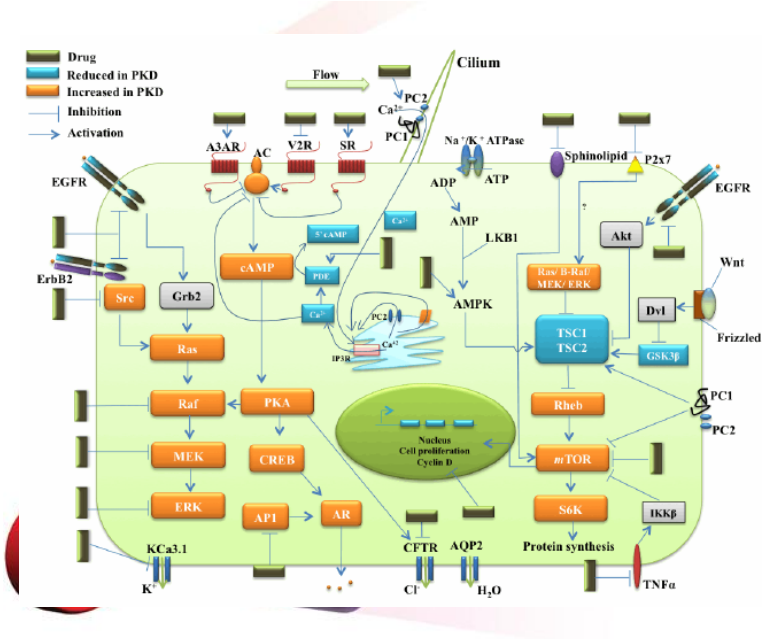
# Correlation Analysis

- Correlation analysis on patients having **breast cancer** (**stage 4**):
  - Using data mining approaches, a correlation of **death records** with drugs, weight, gender, age, and reactions has been obtained.
  - Drugs:
    - Drug docetaxel has the highest correlation (81%)
      - Docetaxel has been repeated 2 times only for patients with death record.
    - Aredia has the lowest correlation (-3%).
      - Aredia has been repeated 81 times only for patients with no death record.
  - Reactions:
    - “Procedural complication” has the highest correlation (57%)
    - “fatigue” has the least correlation (-1.4%).
  - Age and weight have 9% and 1% correlation and gender is close to 0.

# Solution

- Add ingredients… (Features)

Drug Target Pathway




➢A metabolic pathway is a linked series of chemical reactions occurring within a cell. The reactants, products, and intermediates of an enzymatic reaction are known as metabolites, which are modified by a sequence of chemical reactions catalyzed by enzymes.

➢A drug target is a molecular structure (chemically definable by at least a molecular mass) that will undergo a specific interaction with chemicals that we call drugs because they are administered to treat or diagnose a disease.

# Predictive models on reactions

These information have been extracted from DrugBank.ca

The steps to extract:

- Mapping the drugname in FDA database with DrugBank.
- Extracting the ID from DrugBank and PubCHem for drugname in FDA.
- Extracting target and pathway of drugs from DrugBank.ca

DRUGBANK

PubChem

NCBI National Center for Biotechnology Information

MedDRA Web-Based Browser

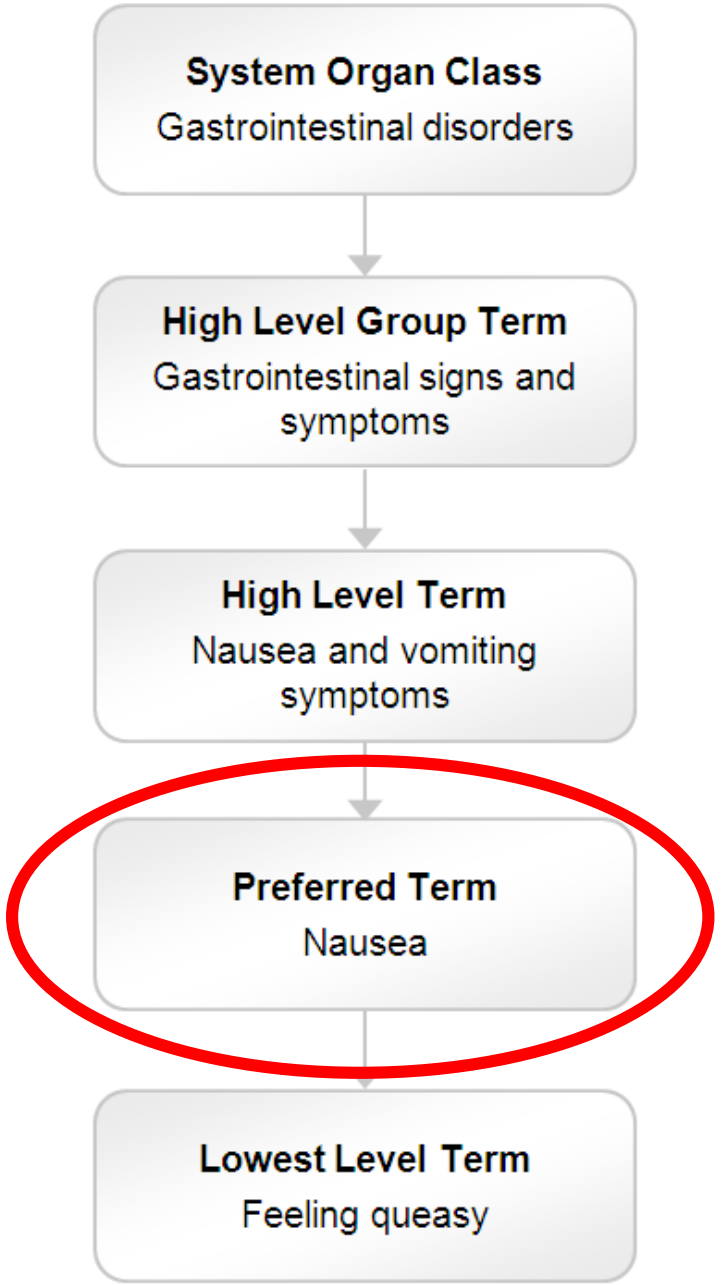
Medical Dictionary for Regulatory Activities

# Medical Dictionary for Regulatory Activities

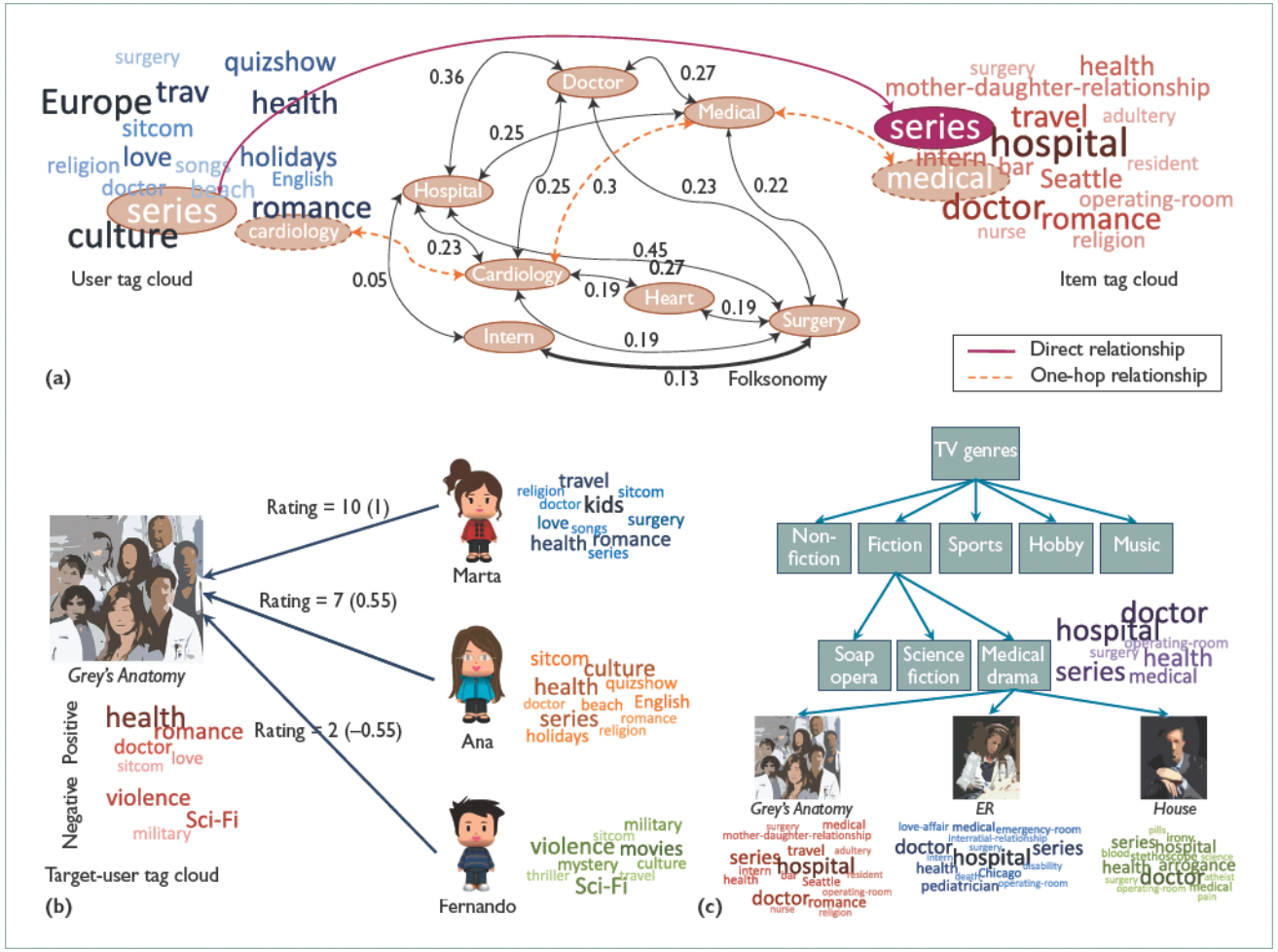
- In the late 1990s, the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) developed MedDRA, a rich and highly specific standardised medical terminology to facilitate sharing of regulatory information internationally for medical products used by humans.

# Medical Dictionary for Regulatory Activities

- In developing and continuously maintaining MedDRA, ICH endeavours to provide a single standardised international medical terminology which can be used for regulatory communication and evaluation of data pertaining to medicinal products for human use. As a result, MedDRA is designed for use in the registration, documentation and safety monitoring of medicinal products through all phases of the development cycle (i.e., from clinical trials to post-marketing surveillance).




MedDRA Web-Based Browser
Medical Dictionary for Regulatory Activities



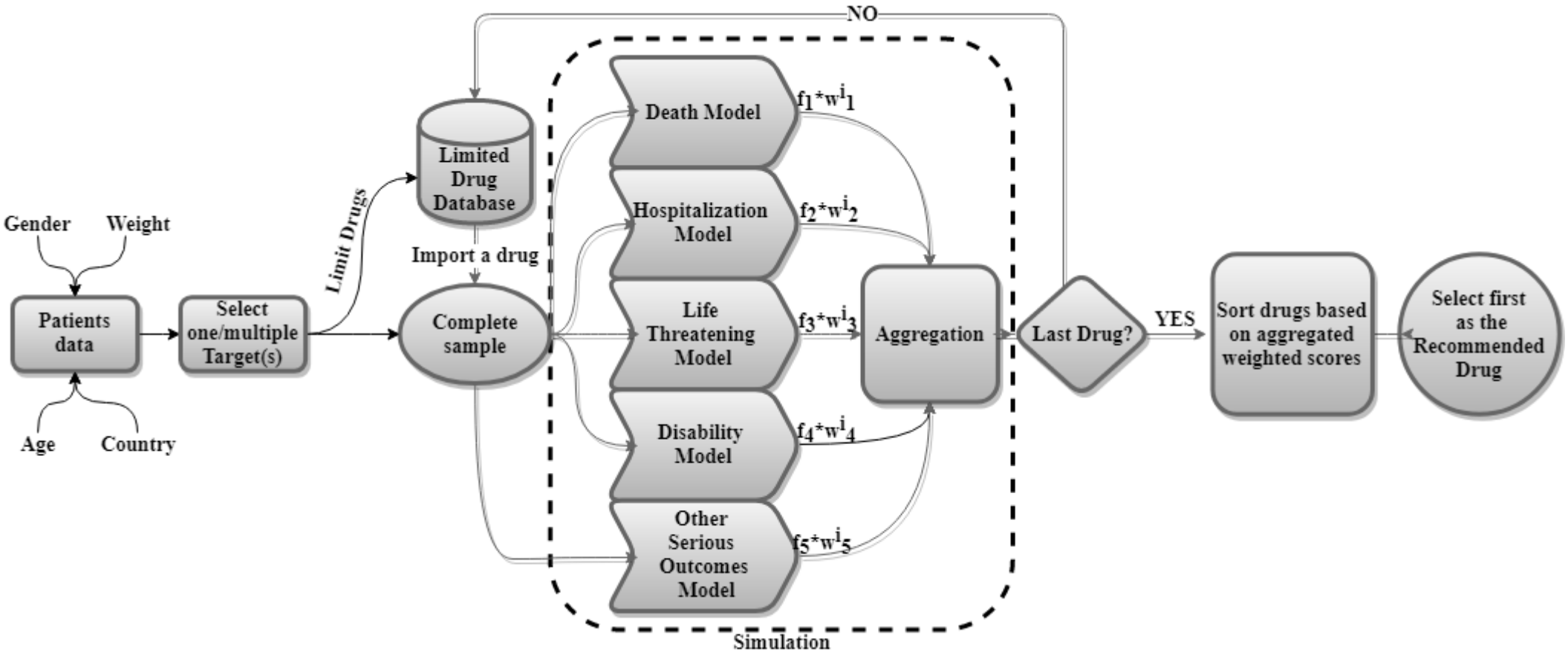

```
Death model
oob score: 0.9677474991376337
0.9631354287931704
[[2256   70]
 [  25  226]]
================================
Life threating model
oob score: 0.9885736460848569
0.9879705083430346
[[2468   13]
 [  18   78]]
================================
Hospitalization model
oob score: 0.8979389444636081
0.9057043073341094
[[1625  140]
 [ 103  709]]
================================
Disability model
oob score: 0.9917212832011039
0.9934031819945673
[[2505    8]
 [   9   55]]
================================
Other Serious issues model
oob score: 0.919454984477406
0.9181218471090415
[[1632  134]
 [  77  734]]
```

# Recommender System

# Recommender System

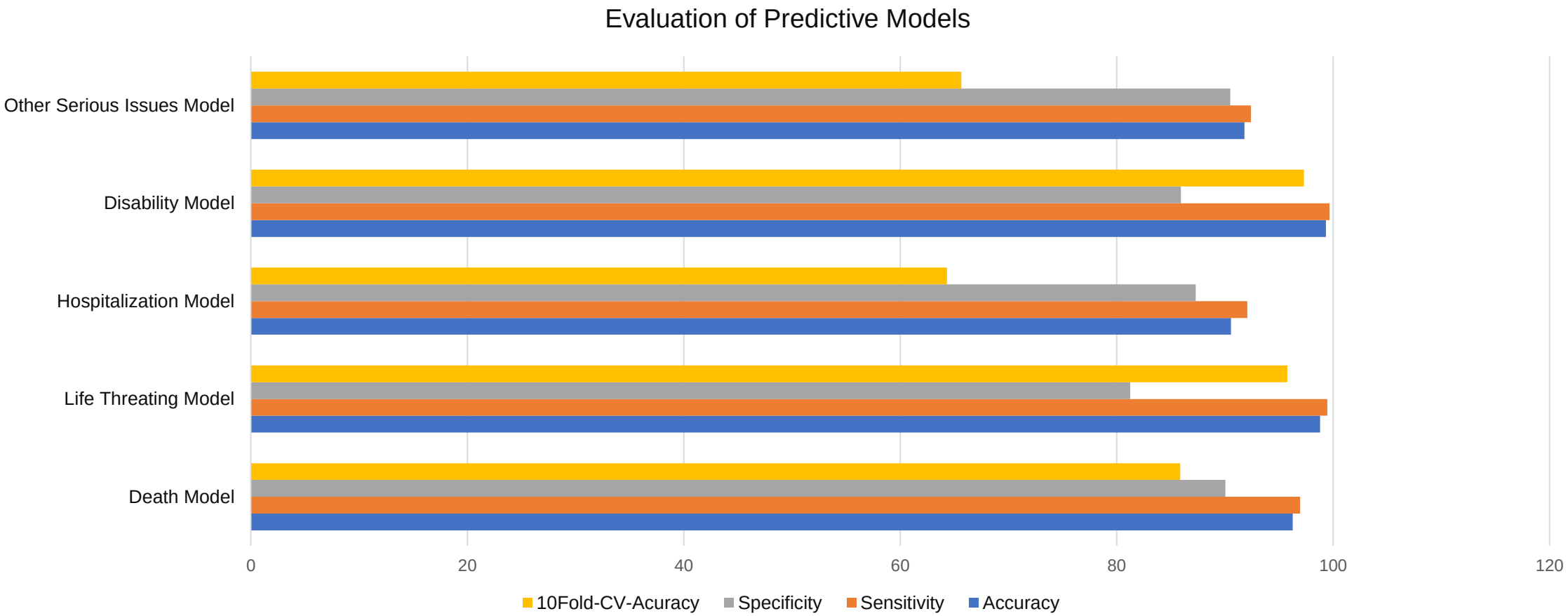
- A drug recommender system on patients with “**Breast Cancer** “:
  - Using predictive models, a recommender system on patients suffering breast cancer is made based on outcome of drugs administrated (outcomes are records such as death, serious issue, hospitalization, …).
  - The model is built using age, gender, weight, and drugs. Based on these features, a score (weighted summation of death probability, hospitalization probability , …) is generated.
  - Weights are as: (subjective)
    - w1 (=1) for death
    - w2 (=0.8) for life threating
    - w3 (= 0.5) for hospitalization
    - w4 (= 0.7) for disability
    - w5 (= 0.5) for other serious issues
  - Then, based on given age, weight, and gender and having a knowledge of disease (**breast cancer**), a drug is recommended to a patient.
  - In addition, a ranked top 5 listed of drugs is shown as well.

# 1DrugAssist



DrugAssist

# Evaluation of classifiers on all outcomes

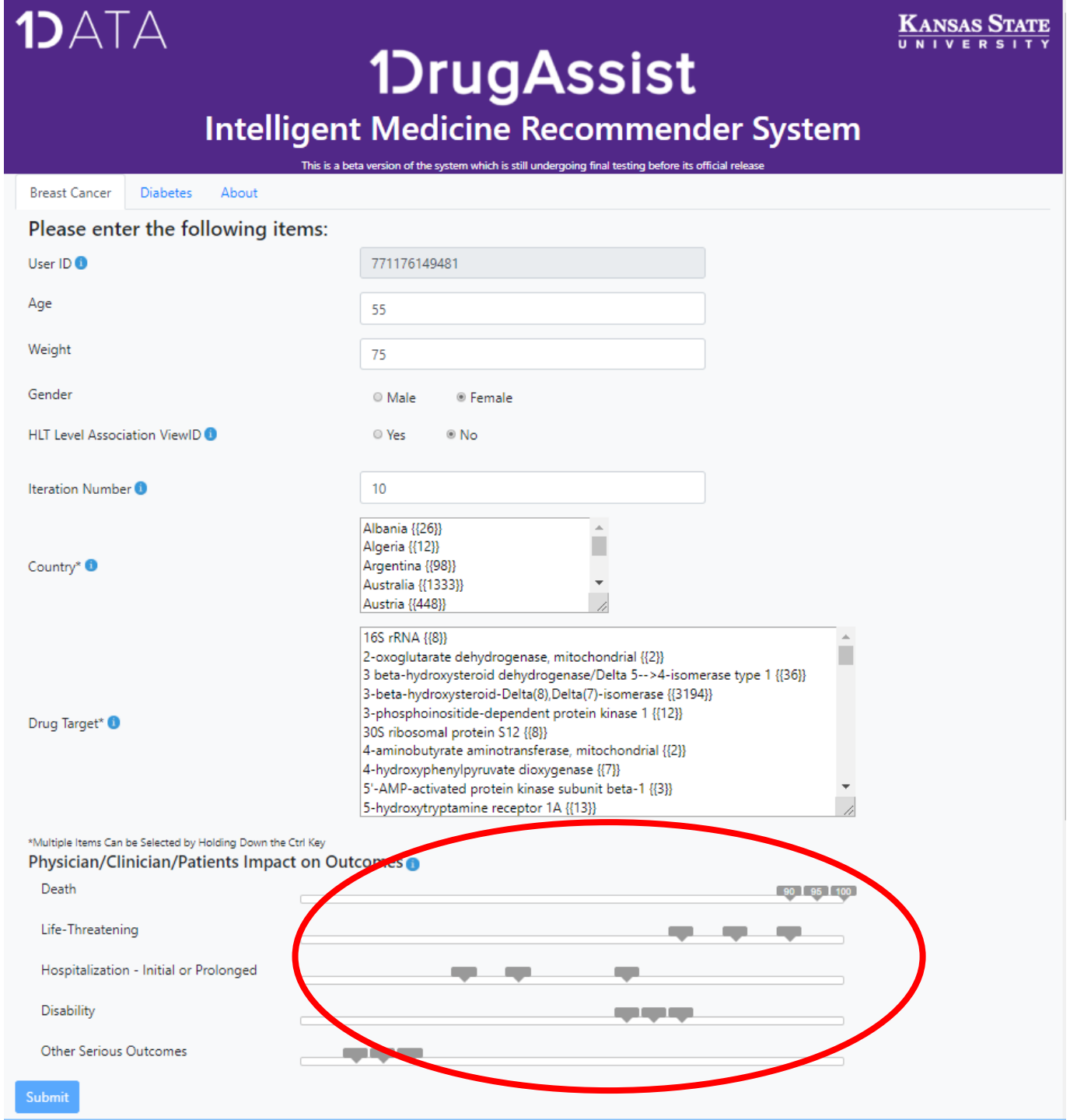


# 1DrugAssist

[1data.olathe.ksu.edu/drugassist/](1data.olathe.ksu.edu/drugassist/)

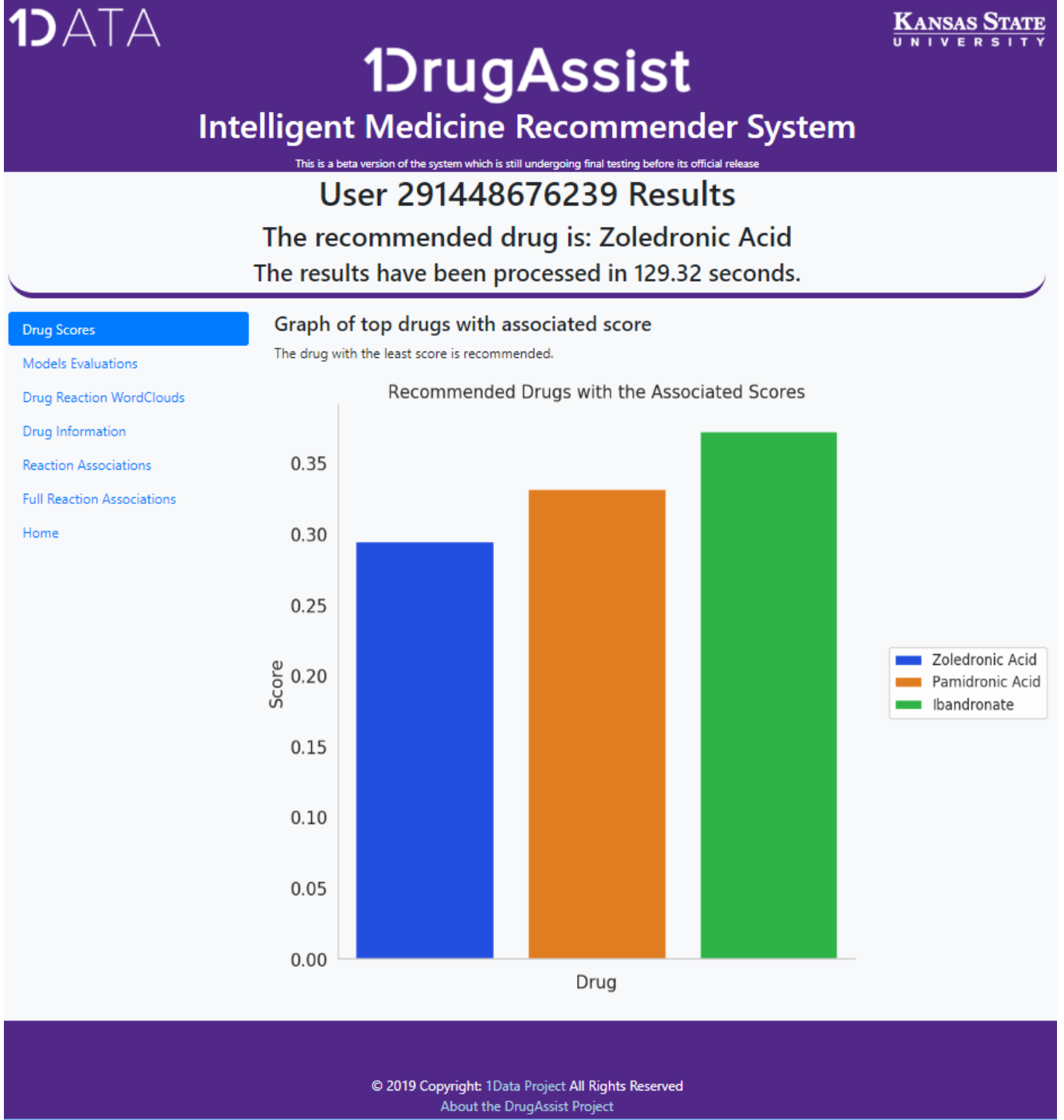
# 1DrugAssist

[1data.olathe.ksu.edu/drugassist/](1data.olathe.ksu.edu/drugassist/)



# 1DrugAssist

[1data.olathe.ksu.edu/drugassist/](1data.olathe.ksu.edu/drugassist/)



# 1DrugAssist

[1data.olathe.ksu.edu/drugassist/](1data.olathe.ksu.edu/drugassist/)

The recommended drug is: Zoledronic Acid

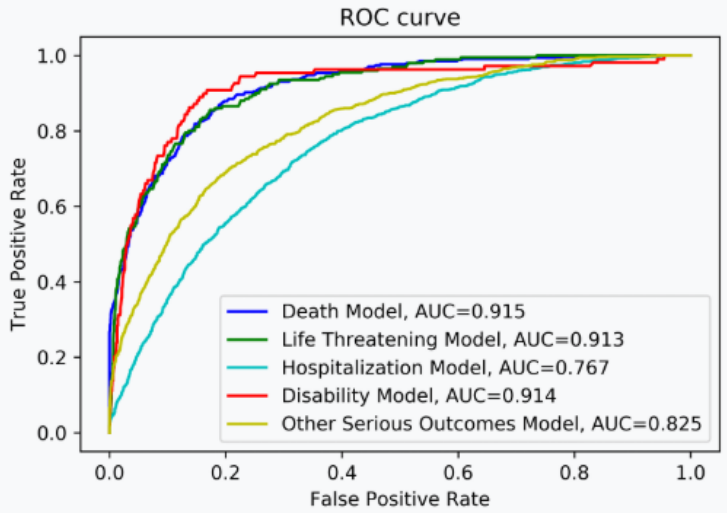
The results have been processed in 129.32 seconds.

- Drug Scores
- Models Evaluations
- Drug Reaction WordClouds
- Drug Information
- Reaction Associations
- Full Reaction Associations
- Home

## Graph of Models Evaluations

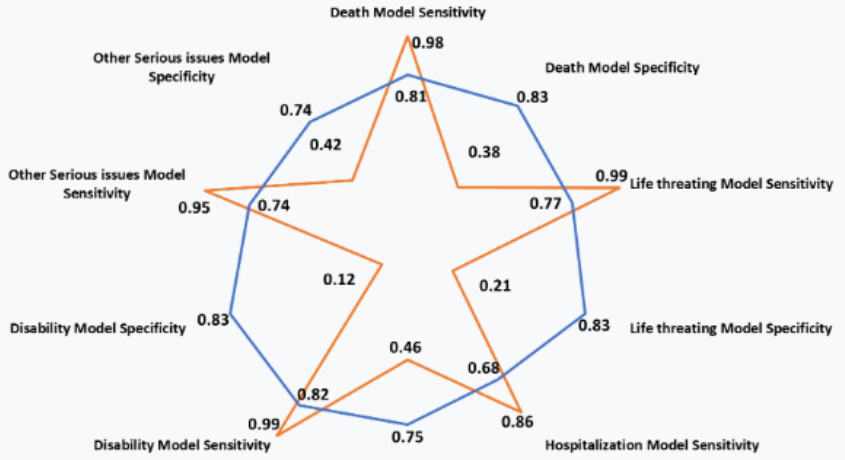
### ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve) score.

ROC and AUC are the best tools to show the evaluation of a model.



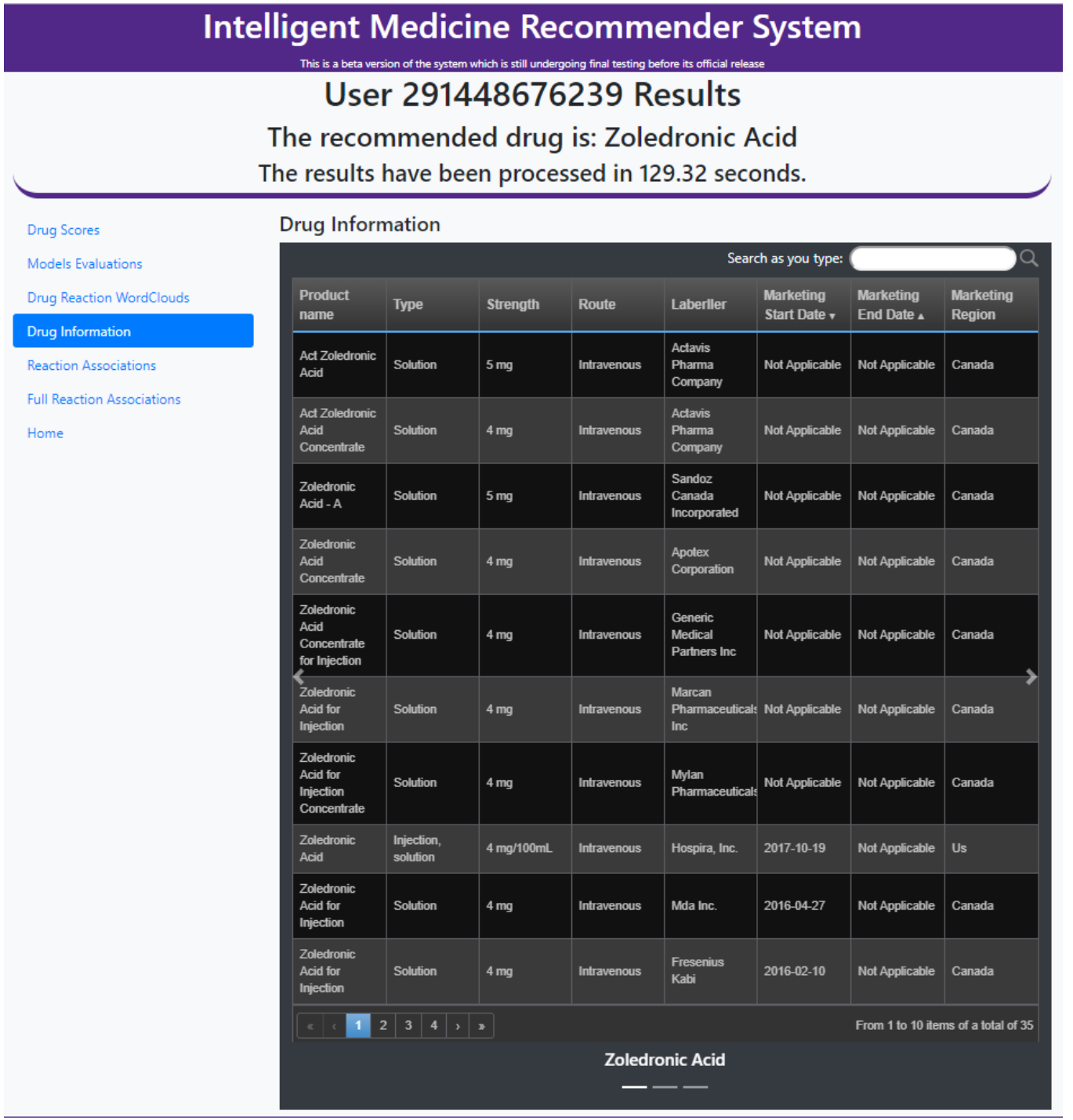

### Under Sampling versus Without Sampling in terms of Specificity and Sensitivity scores.

Specificity and Sensitivity show how accurate is the model in terms of detecting of both classes. Predicting using the whole dataset provides higher sensitivity score and very low specificity score, while undersampling makes a balanced score.



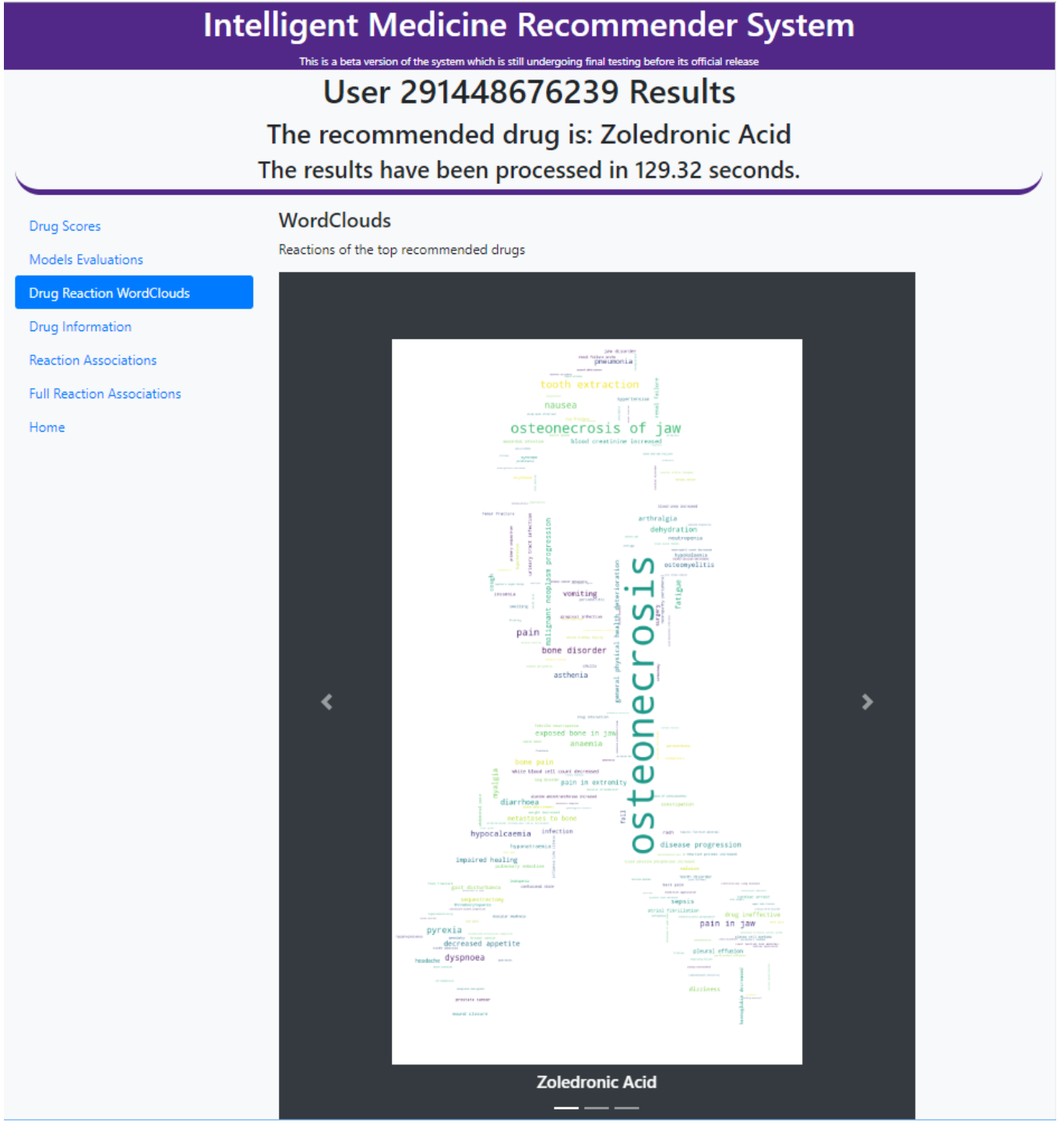
# 1DrugAssist

[1data.olathe.ksu.edu/drugassist/](1data.olathe.ksu.edu/drugassist/)



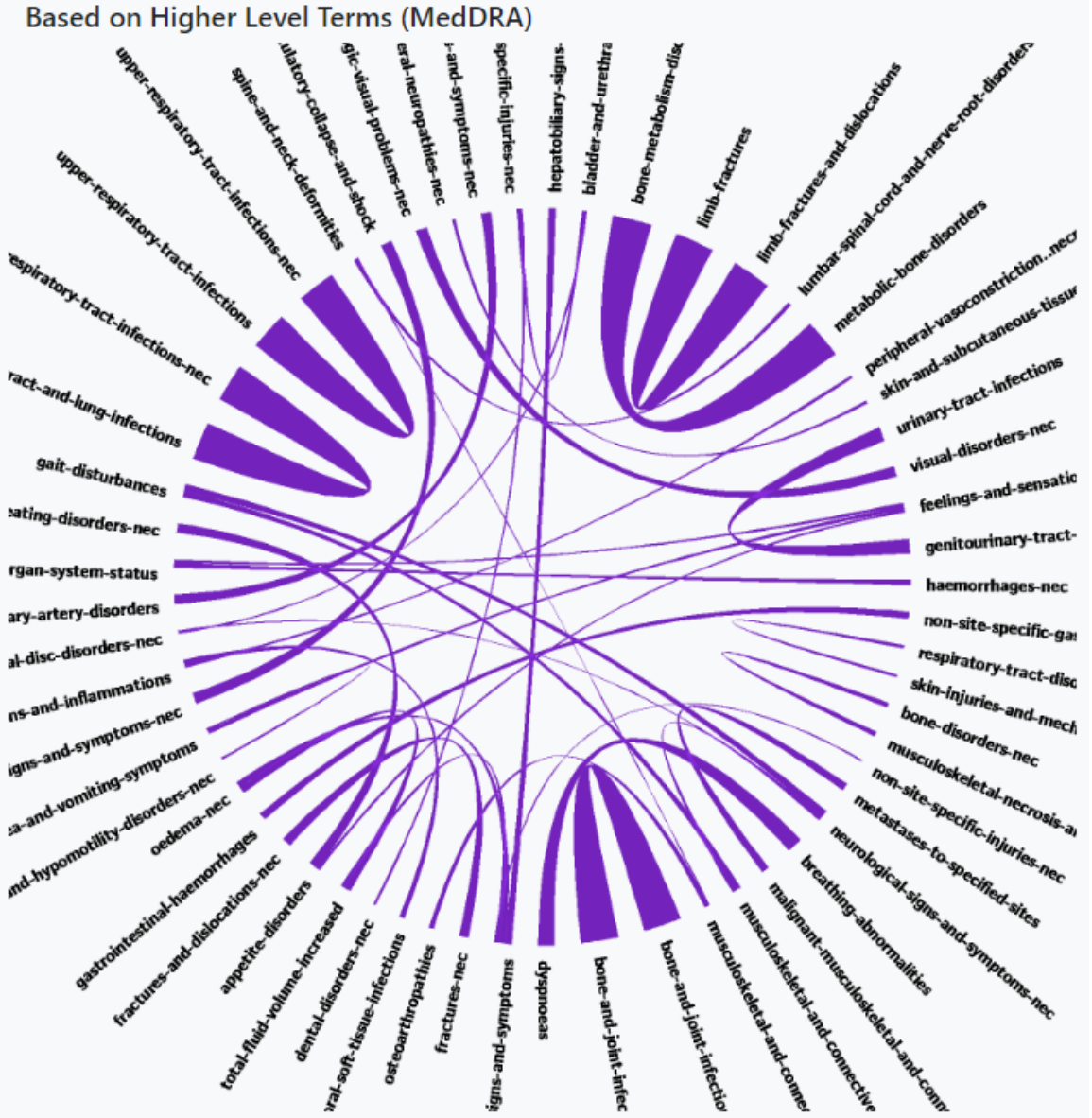
# 1DrugAssist

[1data.olathe.ksu.edu/drugassist/](1data.olathe.ksu.edu/drugassist/)



# 1DrugAssist

[1data.olathe.ksu.edu/drugassist/](1data.olathe.ksu.edu/drugassist/)




© 2019 Copyright: 1Data Project All Rights Reserved

About the DrugAssist Project

# Case study …

| Patient | Age | Weight | Gender | Recommended drug (First Physician/Clinician/... Impact on Outcomes) | Recommended drug (Second Physician/Clinician/... Impact on Outcomes) |
|---|---|---|---|---|---|
| 1 | 58 | 78 | Female | Endoxan | Endoxan |
| 2 | 75 | 80 | Female | Arimidex | Endoxan |
| 3 | 45 | 70 | Male | Lapatinib | Cytoxan |
| 4 | 80 | 60 | Female | Arimidex | Aclasta |
| 5 | 20 | 57 | Female | Pamidronate Disodium | Cytoxan |

**1Data Team:**

Jim Riviere

Jerry Wyckoff

Debbie Kirckoff

Ralph Richardson

Reza Mazloom

Josh Staley

Hossein Amini

Heman Shakeri

Jessica Kawakami

**Thanks to Funding Agencies:**

USDA

FARAD

KCALSI (Now BioNexus KC)

K-State Olathe

Elanco