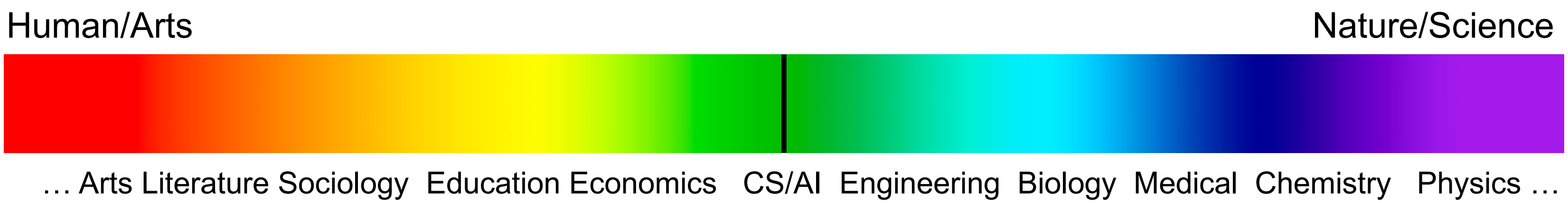
RUTGERS
UNIVERSITY | NEW BRUNSWICK

Explainable AI for Science

Yongfeng Zhang
Department of Computer Science, Rutgers University
yongfeng.zhang@rutgers.edu
http://yongfeng.me

AI helps in many Research Areas

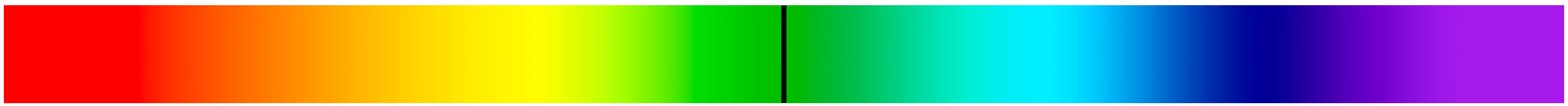
- A (very rough) spectrum of research discipline system



AI helps in many Research Areas

- A (very rough) spectrum of research discipline system

Human/Arts Nature/Science



… Arts Literature Sociology Education Economics CS/AI Engineering Biology Medical Chemistry Physics …

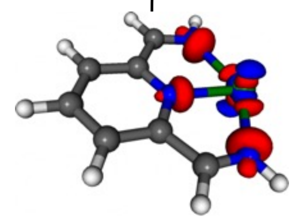
Computer Painting

WHAT STARTS OUT AS TRIUMPH SOON BECOMES MANIPULATED INTO A CACOPHONY OF FUTILITY, LEAVING ONLY A SENSE OF DREAD AND THE UNLIKELIHOOD OF A NEW ORDER.

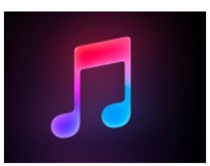
BY TEXT GENERATOR

Computer-generated poems/computer writing

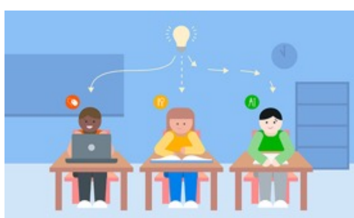
Computational Biology



Computational Chemistry



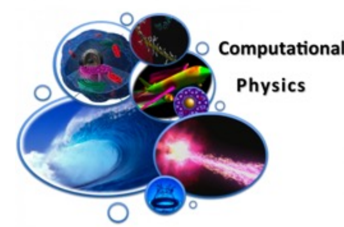
Computer Music Composing



Personalized education based on ML

Quantitative and Computational Econ

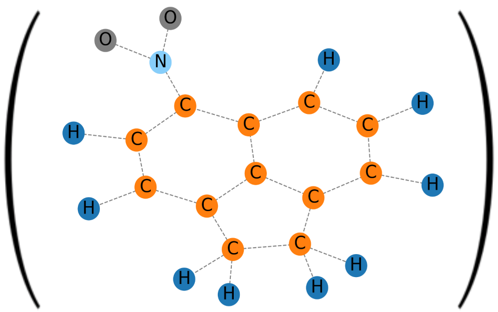
Computational drug discovery



Computational Physics

AI for Science: Some Examples

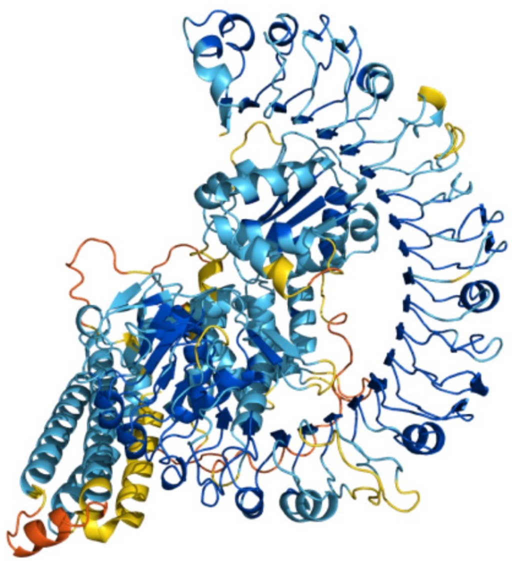
- AI for Drug Discovery
  - Molecule Generation and Property Prediction

NN  = Soluble? Toxic? Crosses the BBB?

- Protein Structure Prediction

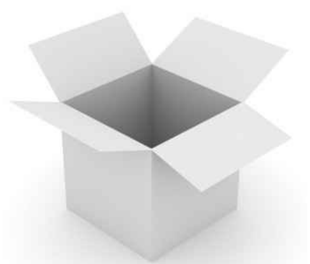
```
MAGELVSFAVNKLWDLLSHEYTLFQGVEDQVAELKSDLNL
LKSFLKDADAKKHTSALVRYCVEEIKDIVYDAEDVLETFV
QKEKLGTTSGIRKHIKRLTCIVPDRREIALYIGHVSKRIT
RVIRDMQSFGVQQMIVDDYMHPLRNREREIRRTFPKDNES
GFVALEENVKKLVGYFVEEDNYQVVSITGMGGLGKTTLAR
QVFNHDMVTKKFDKLAWVSVSQDFTLKNVWQNILGDLKPK
EEETKEEEKKILEMTEYTLQRELYQLLEMSKSLIVLDDIW
KKEDWEVIKPIFPPTKGWKLLLTSRNESIVAPTNTKYFNF
KPECLKTDDSWKLFQRIAFPINDASEFEIDEEMEKLGEKM
IEHCGGLPLAIKVLGGMLAEKYTSHDWRRLSENIGSHLVG
GRTNFNDDNNNSCNYVLSLSFEELPSYLKHCFLYLAHFPE
DYEIKVENLSYYWAAEEIFQPRHYDGEIIRDVGDVYIEEL
VRRNMVISERDVKTSRFETCHLHDMMREVCLLKAKEENFL
QITSNPPSTANFQSTVTSRRLVYQYPTTLHVEKDINNPKL
…
```

AlphaFold



The Explainability Crisis

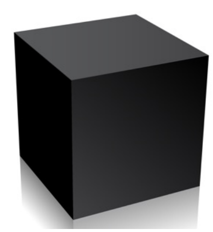
- A Key Problem with current AI models
  - Most AI prediction methods are not explainable
    - They can make good predictions based on massive data and complicated models, but are less capable of explaining the prediction results and reveal the insights to human scientists
    - They can produce prediction results, but hardly explains why the results are predicted they way they are
- Origin of the Problem
  - Difference from traditional methods: Whitebox vs. Blackbox models



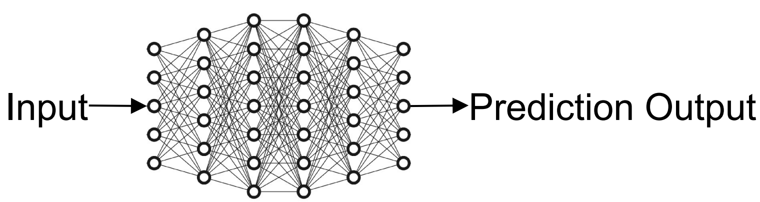
e.g., (partial) differentiable equation

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{\nabla P}{\rho} + \nu \nabla^2 \mathbf{u},$$

Navier-Stokes equation



e.g., deep neural networks



Multiple Layer Perceptron (MLP)

Why **Explainable AI** for Science?

- The essence of scientific research is to understand the "why"
  - Not only know **how** but also know **why**
    - Know how: Blackbox AI for Prediction; Know why: Explainable AI for Explanation
  - In many cases, understanding the "why" behind the result is even more important than just knowing the result itself, because knowing the why implies real growth of knowledge and helps in making critical decisions
  - Furthermore, if AI accumulates more and more dark knowledge that are not understandable to humans (which is already happening), it may eventually lead to a singularity where humans are lagged behind on the conquest of knowledge than machines

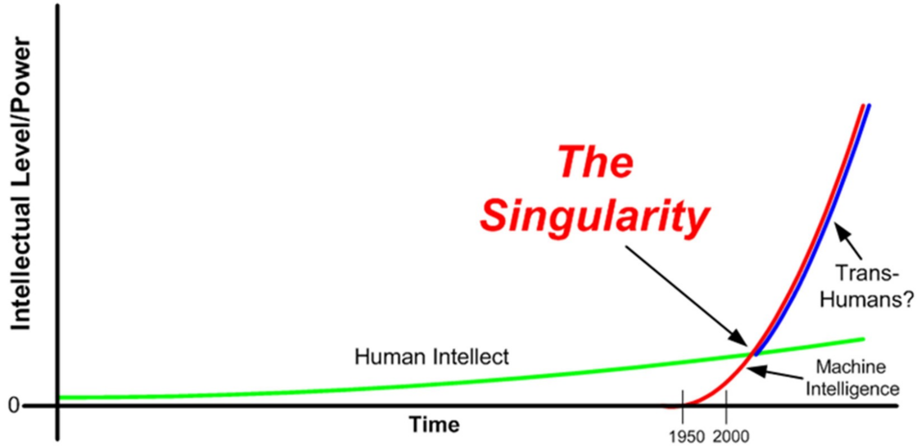


Image credit: https://www.webpages.uidaho.edu/vakanski/Technological%20Singularity.html

The Conquest of "Why" in Science

- The conquest of why has always been the key theme of science in human history
- A Legend Example
  - The Kepler's Laws of Planetary Motion
  - The Newton's Law of Universal Gravitation

Tycho Brahe (1546-1610) Johannes Kepler (1571-1630) Isaac Newton (1643-1727)

RUTGERS

Kepler’s Laws of Planetary Motion

We can **Obverse** it!

We can **Predict** it!

Tycho Brahe (1546-1610)
Demark astronomer

Good at astro-observation

Observed and recorded a lot of data about Mars movement.

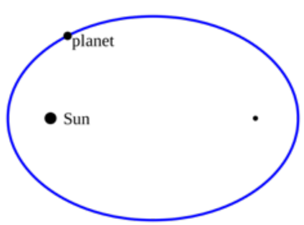
Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

Analyzed Tycho’s data, and discovered the rules hidden in the data.
The “Kepler’s laws of planetary motion”:

$$\frac{r^3}{T^2} = K$$

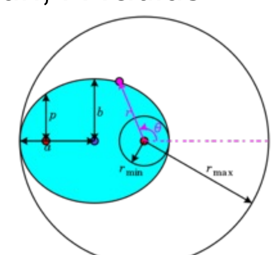
T: period of circling around the sun, r: radius

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)



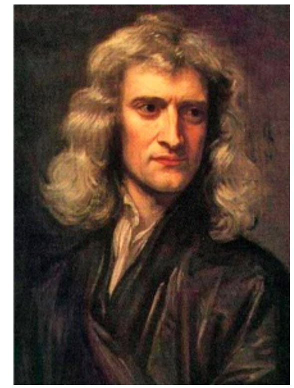
Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

$$\frac{r^3}{T^2} = K$$



Is the Story Over? No!

We can
Predict it!



We **Understand** it!
We know **Why**!

Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

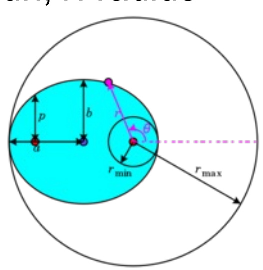
Analyzed Tycho's data, and discovered the rules hidden in the data.
The "Kepler's laws of planetary motion":

$$\frac{r^3}{T^2} = K$$

τ: period of circling around the sun, r: radius

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

$$\frac{r^3}{T^2} = K$$



Isaac Newton (1643-1727)
English mathematician, physicist, astronomer, theologian, and author.

Proposed the Newton's law of universal gravitation + differential calculus:

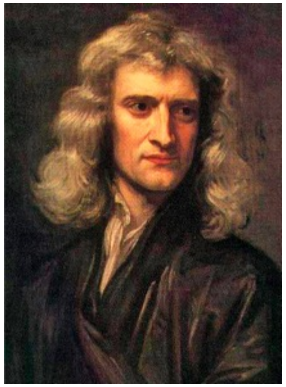
Naturally derives out the Kepler's laws of planetary motion!

$\frac{r^3}{T^2} = K$ is because $F = G\frac{m_1 m_2}{r^2}$

Three Key Roles in the Scientific Discovery Process

Tycho Brahe (1546-1610)

Johannes Kepler (1571-1630)



Isaac Newton (1643-1727)

Observation

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

Analyzation

$$\frac{r^3}{T^2} = K$$

Explanation

$$F = G\frac{m_1 m_2}{r^2}$$

What if Kepler had DL in the 16-17th Century?

We can
Obverse it!

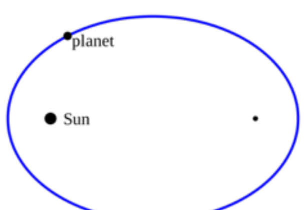
We can
Predict it!

Tycho Brahe (1546-1610)
Demark astronomer

Good at astro-observation

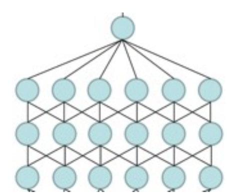
Observed and recorded a lot of data about Mars movement.

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)



Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

There could be some rules underlying the data.
I don't know what it is, but NN can fit any function.
So I'm going to train a NN to fit the data!



It fits the data pretty well!
I can make predictions!

$$r = some\ NN(T)$$

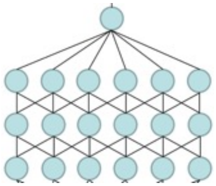
But wait: can this be called scientific discovery?
Science is not only about know HOW, but also know WHY!

Challenges in Modern Scientific Research

We can
Predict it!

Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

There could be some rules underlying the data.
I don't know what it is, but NN can fit any function.
So I'm going to train a NN to fit the data!



It fits the data pretty well!
I can make predictions!

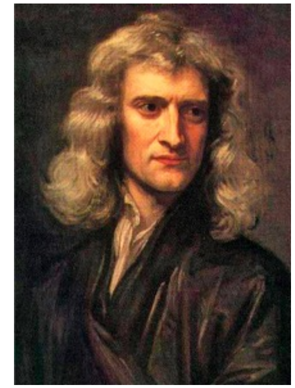
$$r = some\ NN(T)$$

But wait: can this be called scientific discovery?
Science is not only about know HOW, but also know WHY!

- However, manually analyzing data as Kepler did is very challenging in modern scientific research
  - Since the amount of data is huge
  - e.g., produced by astronomical telescope and particle colliders
- We indeed need AI for data analyses and model learning

Challenges in Modern Scientific Research

We can
Predict it!



We **Understand** it!
We know **Why**!

Johannes Kepler (1571-1630)
German astronomer, student of Tycho Brahe.

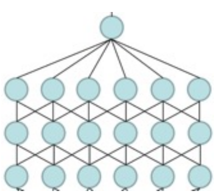
Isaac Newton (1643-1727)

Explainable AI (XAI) plays the role of Newton

There could be some rules underlying the data.
I don't know what it is, but NN can fit any function.
So I'm going to train a NN to fit the data!

Interpret and **explain** the learned (black-box) model, reveal its insights to human scientists

Help us better understand the nature.



It fits the data pretty well!
I can make predictions!

$$r = some\ NN(T)$$

But wait: can this be called scientific discovery?
Science is not only about know HOW, but also know WHY!

Three Key Roles in the Scientific Discovery Process

Using more CS/AI language

Tycho Brahe (1546-1610)

Observation

Data Collection

Time	Position
1	(a,b)
2	(c,d)
3	(e,f)

Almost automated

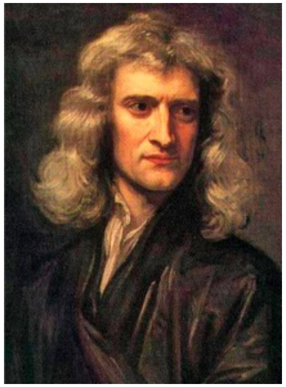
Johannes Kepler (1571-1630)

Analyzation

Model Learning

$$\frac{\tau^2}{r^3} = K$$

Many available methods



Isaac Newton (1643-1727)

Explanation

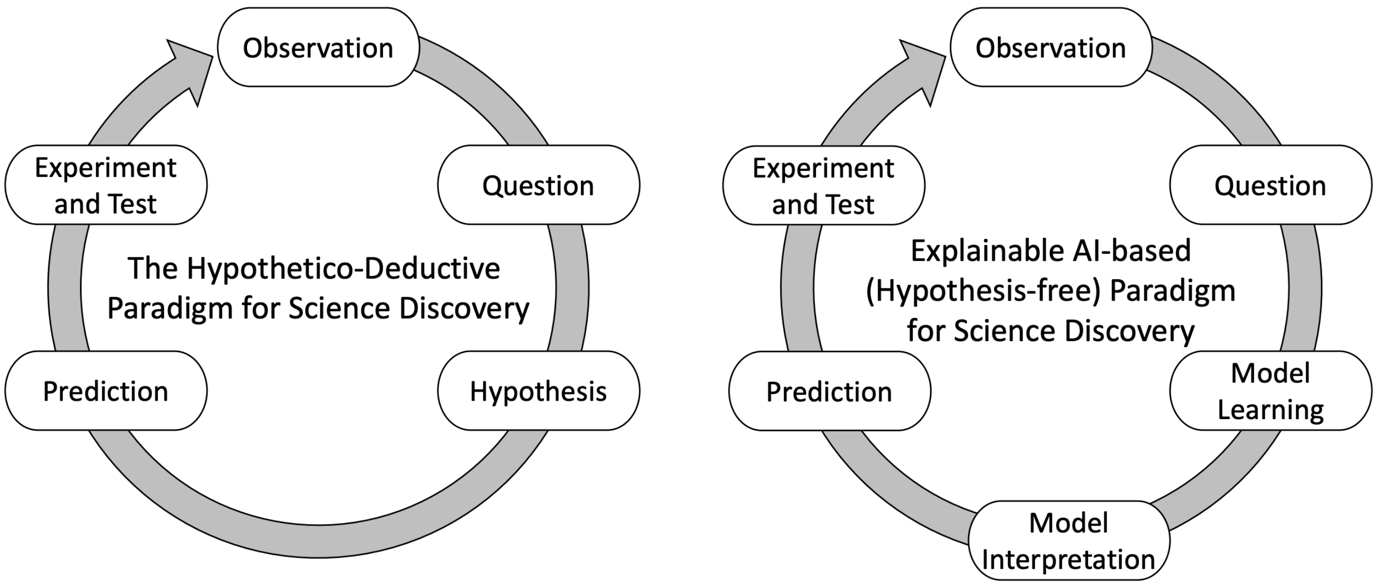
Model Interpretation (XAI)

$$F = G\frac{m_1 m_2}{r^2}$$

Still needs much exploration

A Paradigm Shift (again) for Scientific Research

- From Theory-driven to Data-driven (back to Kepler), but with Explainable AI (plus Newton)
  - Blackbox AI for Prediction (the Kepler model)
  - Explainable AI for Explanation (the Newton model)
- A Paradigm Shift in Scientific Discovery
  - Explainable AI replaces manual hypothesis generation



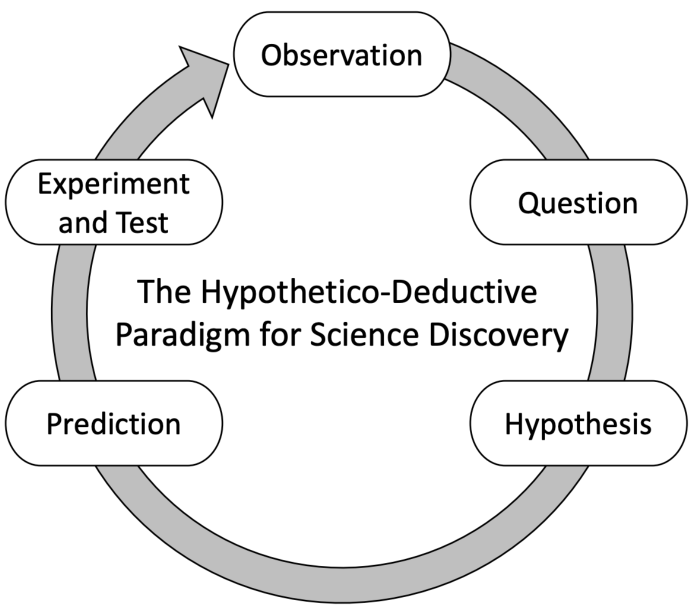
Three Examples on Explainable AI for Science

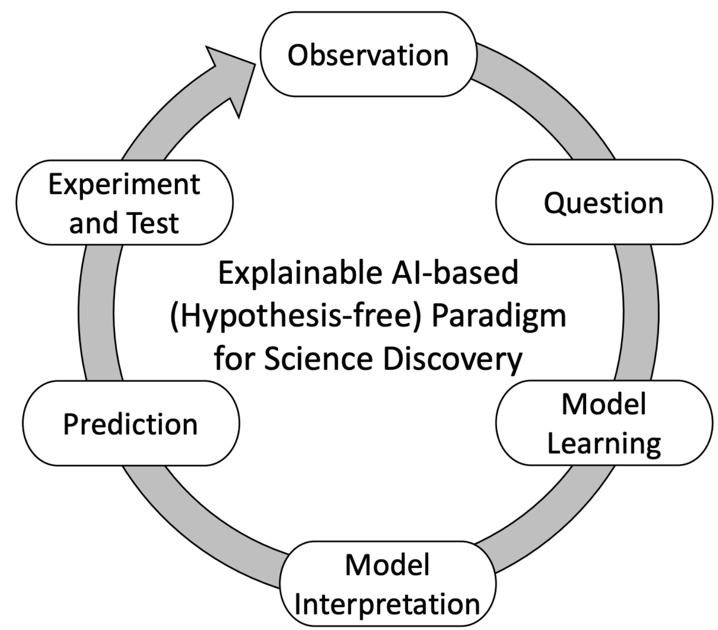
- Rediscover Kepler's laws and Newton's laws from Tycho's ancient data [1]
  - A good example to demonstrate the idea of XAI-driven scientific research
  - Pay our respect to some of the greatest minds in human history
- More "practical" Examples
  - Explainable AI for Molecular Property Prediction [2]
  - Explainable AI for Biodiversity Conservation [3]

- [1] Zelong Li, Jianchao Ji, and Yongfeng Zhang. "From Kepler to Newton: Explainable AI for Science Discovery." In ICML AI for Science 2022.
- [2] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In Proceedings of the ACM Web Conference 2022.
- [3] Meet Mukadam, Mandhara Jayaram, and Yongfeng Zhang. "A Representation Learning Approach to Animal Biodiversity Conservation." In Proceedings of the 28th International Conference on Computational Linguistics. 2020.

From Kepler to Newton: A Case Study

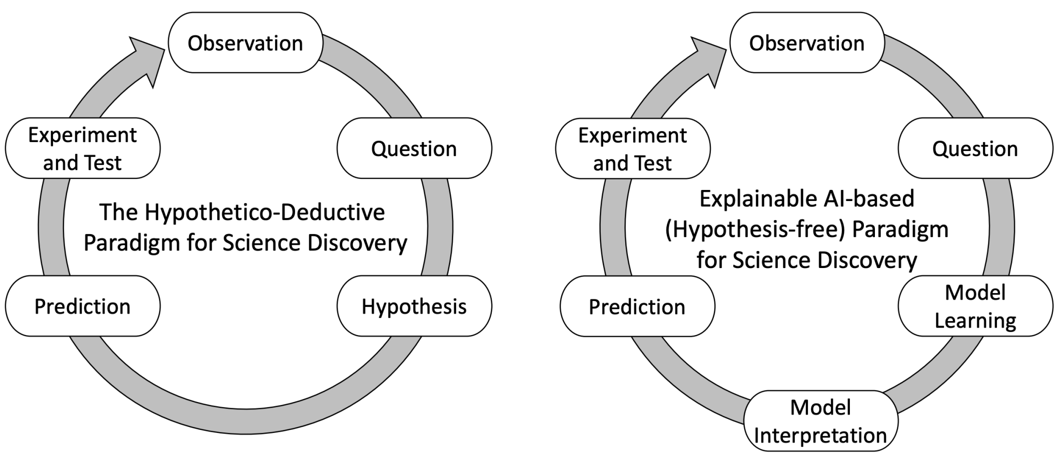
- A New Paradigm for Scientific Discovery
  - Model Learning and Interpretation automatically generates hypothesis
- Use the paradigm to rediscover:
  - Kepler's Laws of Planetary Motion
  - Newton's Law of Universal Gravitation





Kepler's Reasoning Process

- At Kepler's time, there were three models of planetary motion
  - The Ptolemaic, Copernican and Tychonic systems
  - Kepler mentioned that these three systems all had high prediction accuracy in the near term, but diverged and failed to fit historical and future observations in the long term
  - Propose a new hypothesis: the orbit of a planet is an ellipse with the Sun at one of the two foci (Kepler's first law of planetary motion)
  - Then he used the observation data to test his hypothesis
- We show the hypothesis-free scientific discovery process based on Explainable AI
  - We directly start from data to rediscover the Kepler's laws.



RUTGERS

Dataset: Ancient Mars Data from Tycho Brahe

Time YYYY/MM/DD	Mars' Position in Ecliptic	Sun-Mars Distance	Difference
1582/11/23 16:00	90.70306°	1.58852	+1′30″
1582/12/26 08:30	106.12167°	1.62104	+3′49″
1582/12/30 08:10	107.94222°	1.62443	+5′50″
1583/01/26 06:15	120.10667°	1.64421	−2′33″
1584/12/21 14:00	123.86250°	1.64907	+1′04″
1585/01/24 09:00	138.78556°	1.66210	−3′32″
1585/02/04 06:40	143.56139°	1.66400	−3′08″
1585/03/12 10:30	159.38722°	1.66170	−2′29″
1587/01/25 17:00	158.22778°	1.66232	−0′10″
1587/03/04 13:24	174.94722°	1.64737	−0′59″
1587/03/10 11:30	177.59833°	1.64382	0′0″
1587/04/21 09:30	196.74750°	1.61027	+1′30″
1589/05/08 16:24	196.92056°	1.61000	−2′43″
1589/04/13 11:15	214.03056°	1.57141	+1′40″
1589/04/15 12:05	215.02806°	1.56900	+0′37″
1589/05/06 11:20	225.51000°	1.54326	+0′57″
1591/05/13 14:00	252.12722°	1.47891	−4′24″
1591/06/06 12:20	265.64667°	1.44981	−3′15″
1591/06/10 11:50	267.94694°	1.44526	−4′39″
1591/06/28 10:24	278.49222°	1.42608	−5′39″
1593/07/21 14:00	320.02722°	1.38376	−2′31″
1593/08/22 12:20	340.25694°	1.38463	−0′36″
1593/08/29 10:20	344.62083°	1.38682	−2′19″
1593/10/03 08:00	6.32750°	1.40697	−0′16″
1595/09/17 16:45	22.82194°	1.43222	−1′27″
1595/10/27 12:20	45.59389°	1.47890	−0′29″
1595/11/03 12:00	49.44250°	1.48773	+0′03″
1595/12/18 08:00	73.04139°	1.54539	−0′59″

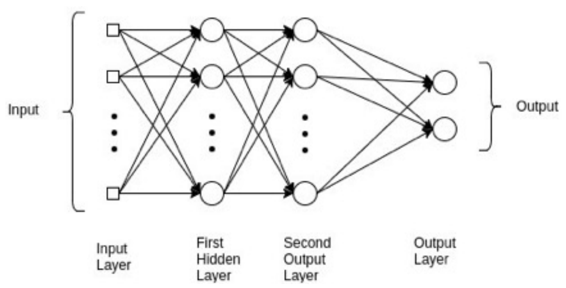
Table 1: Position of Mars when orbiting the Sun

- Data copied from Kepler's book **Astronomia Nova** (1609)
- Three main variables
  - Time: t
  - Mars angular position: θ
  - Sun-Mars distance: r

Blackbox and Whitebox Models

- The black-box model for Prediction and Data Augmentation
  - Simple Multiple Layer Perceptron (MLP) neural network

$$y = \sigma(\underbrace{\boldsymbol{w}_3^T \sigma(\underbrace{\boldsymbol{W}_2^T \underbrace{\sigma(\boldsymbol{W}_1^T \boldsymbol{x} + \boldsymbol{b}_1)}_{1^{st}\text{ Layer}} + \boldsymbol{b}_2)}_{2^{nd}\text{ Layer}} + b_3)}_{3^{rd}\text{ Layer}}$$



- The white-box model for Explanation
  - Symbolic Regression: Transform the MLP neural network into a symbolic equation

RUTGERS

Rediscover Kepler's Laws based on Explainable AI

- Black-box Model (DNN) for Prediction and Data Augmentation

$$r = NN(\theta)$$

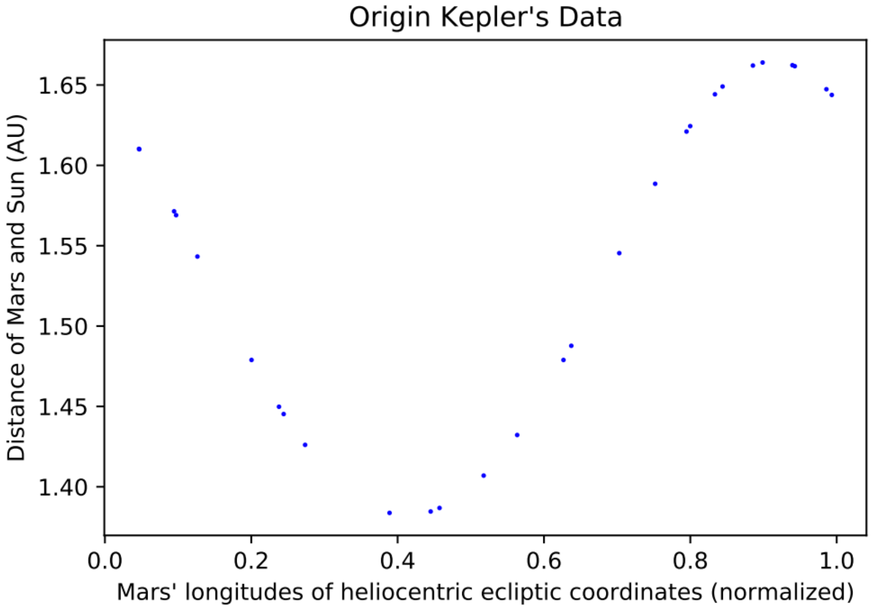


Figure 3: Data Visualization before Training

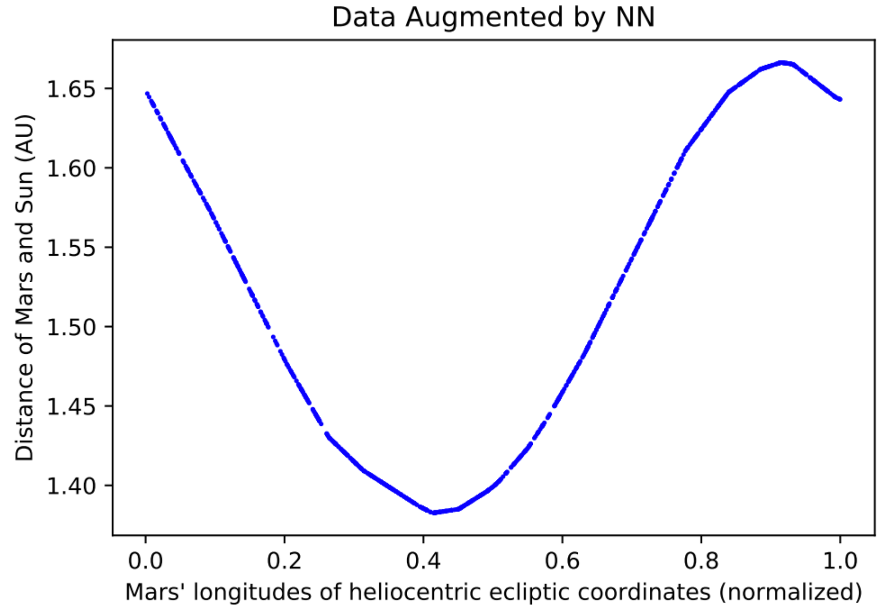


Figure 4: Data Visualization after Training

Use 90% data points for training and 10% for validation.
MSE on training data: 4×10^{-11}; MSE on validation data: 7×10^{-8}
Blackbox neural networks can already make accurate predictions, though we don't understand the insight

Rediscover Kepler’s Laws based on Explainable AI

- White-box Model for Explanation
  - Symbolic regression based on the augmented data

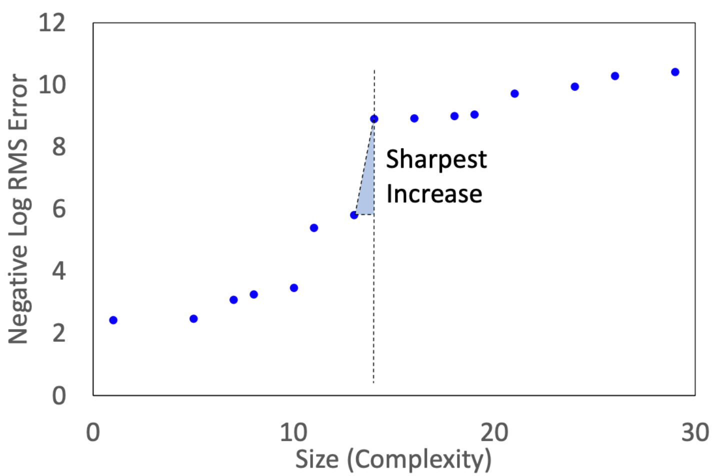


Figure 5: Size and Negative log Error

Size	Error	Function
1	0.088419	1.54806
5	0.084370	$1.54329 + 0.0130577 \cdot \theta$
7	0.045791	$1.45537 + 0.021878 \cdot \theta \cdot \theta$
8	0.038594	$1.53256 - 0.101048 \cdot \cos\theta$
10	0.031201	$1.65411 - \frac{0.321963}{1.21921+\theta\cdot\theta}$
11	0.004519	$1.51578 - 0.142019 \cdot \cos(\theta + 0.542453)$
13	0.003003	$1.51836 - 0.141285 \cdot \cos(0.979081 \cdot (-0.544189 - \theta))$
14	0.000136	$\frac{1.51977}{1.00625+0.0932972\cdot\cos(\theta+0.544536)}$
16	0.000133	$\frac{1.51975}{1.00625+0.0933058\cdot\cos(1.00017\cdot\theta+0.544619)}$
18	0.000124	$\frac{1.5221}{1.0078+\sin(0.0935495\cdot\cos(\theta+0.544689))}$
19	0.000118	$1.51016 - \frac{0.0794197}{0.0536393+\frac{0.567314}{\cos(1.00052\cdot(0.544488+\theta))}}$
21	0.000060	$\frac{1.51978}{1.00625+0.0932649\cdot\cos(\theta+0.544414+\frac{0.000322752}{\theta-1.48167})}$
24	0.000048	$1.51031 - \frac{0.0793261}{0.052716+\frac{0.56737}{\cos(0.543701+\theta)}-\frac{0.000771507}{1.48757-\theta}}$
26	0.000034	$1.51023 - \frac{0.0793521}{0.0531939+\frac{0.56753}{\cos(0.543898+1.00028\cdot\theta)}-\frac{0.00067777}{1.49235-\theta}}$
29	0.000030	$1.51032 - \frac{\cos(-0.543588-\theta)}{7.14743+0.668919\cdot\cos(0.55992-\frac{0.00769976}{1.58368-\theta}+\theta)}$

Table 2: Symbolic Regression Results

$$r = f(\theta) = \frac{1.51977}{1.00625 + 0.0932972 \cdot \cos(\theta + 0.544536)} = \frac{1.51033}{1 + 0.0927177 \cdot \cos(\theta + 0.544536)}$$

Rediscover Kepler's Laws based on Explainable AI

- Physical Interpretation of the Results
  - Mars orbit is an ellipse, and AI-derived eccentricity is 0.0927177
  - Very close to Kepler's result 0.09264 (relative error < 0.1%) and modern result 0.09341233 (relative error < 0.7%)

$$r = f(\theta) = \frac{1.51977}{1.00625 + 0.0932972 \cdot \cos(\theta + 0.544536)} = \frac{\boxed{1.51033}}{1 + 0.0927177 \cdot \cos(\theta + 0.544536)}$$

  - $r_{min} = f[\theta = -0.544536\ (-31.2°)]$, indicating closest Mars Opposition in August, which is consistent with historical observations
    - 31.2/360 × 365 ≈ 32 days ahead of the fall equinox, thus in August

Year (AD)	Date	Earth-Mars Distance in AU
1561	Aug. 07	0.37325
1640	Aug. 20	0.37347
1687	Aug. 09	0.37434
1719	Aug. 25	0.37401
1766	Aug. 13	0.37326
1845	Aug. 18	0.37302
1924	Aug. 22	0.37285
2003	Aug. 27	0.37272
2050	Aug. 15	0.37405

Table 5: Closest Approaches of Mars Oppositions in History

Rediscover Newton's Laws based on Explainable AI

- Black-box Model for Prediction and Data Augmentation
- We already have $r = f(\theta)$, we want $\theta = g(t)$
  - So we can predict the position of Mars (θ, r) for any given time t

$$\theta = NN(t)$$

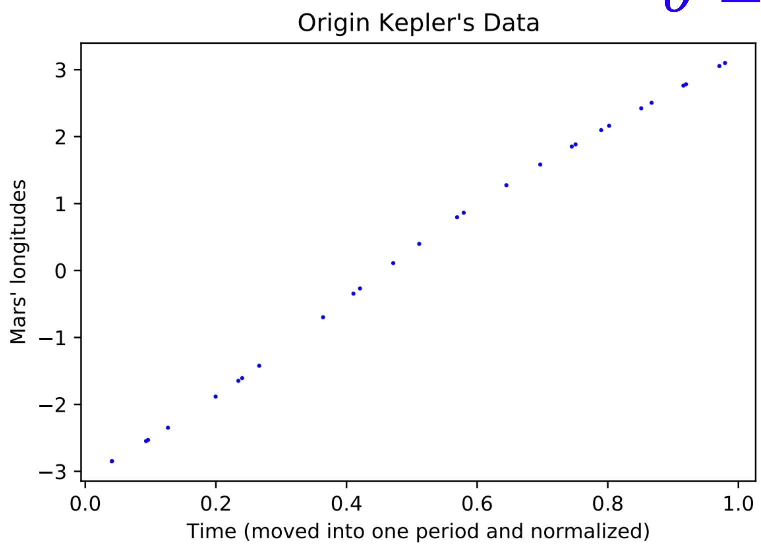


Figure 6: Data Visualization before Training

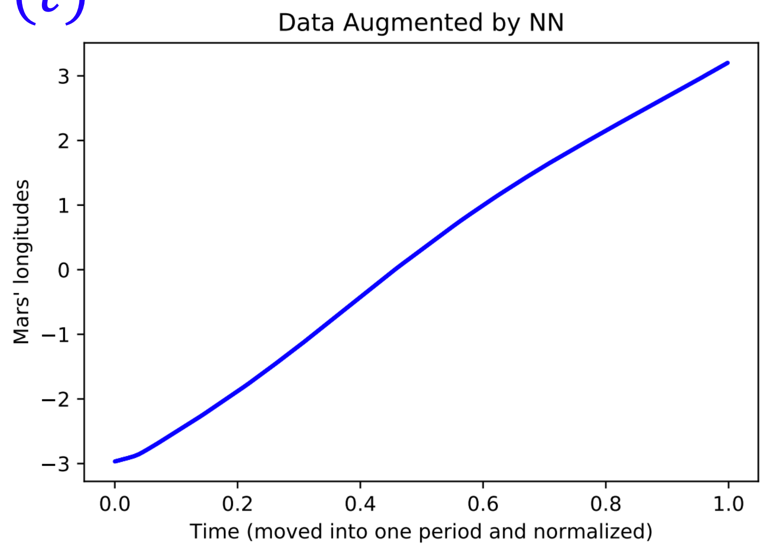


Figure 7: Data Visualization after Training

Use 90% data points for training and 10% for validation.
MSE on training data: 7×10^{-8}; MSE on validation data: 1.5×10^{-5}
Blackbox neural networks can already make accurate predictions, though we don't understand the insight

Rediscover Newton's Laws based on Explainable AI

- Deep Learning for Prediction and Data Augmentation
  - The simple experiment implies a significant role of machine learning (especially deep learning) in scientific discovery

- The real $t - \theta$ relation based on advanced math tools and deeper understandings of planetary motion:

$$\frac{2\pi}{T}t = 2\tan^{-1}\left(\sqrt{\frac{1-\epsilon}{1+\epsilon}}\tan\left(\frac{\theta}{2}\right)\right) - \frac{\epsilon\sqrt{1-\epsilon^2}\sin(\theta)}{1+\epsilon\cos(\theta)}$$

  - i.e., we can express t as a function of θ, i.e., $t = h(\theta)$, however, we can hardly find a function to express θ as t, i.e., $\theta = g(t)$, since it is a transcendental equation

RUTGERS

Rediscover Newton's Laws based on Explainable AI

- However, we still want some θ-as-t relationship
  - We already have $r = f(\theta)$, if we have $\theta = g(t)$, then we can predict the position of Mars (r, θ) for any time t
- We can adopt deep neural networks to learn a black-box predictor $\theta = NN(t)$
  - Universal Approximation Theorem (UAT) [4,5,6]
    - A network containing a finite number of neurons can approximate arbitrarily well any real-valued continuous functions on compact subsets of $\mathbf{R}^n$.
  - $\theta = NN(t)$ is differentiable!
    - We can conduct mathematical analysis on the θ-as-t relationship
    - $\omega = \frac{\mathrm{dNN}(t)}{\mathrm{d}t}$, $a = \frac{\mathrm{d}^2\mathrm{NN}(t)}{\mathrm{d}t^2}$
    - Makes it possible to analyze the relationship between many variables that are otherwise difficult to calculate

[4] Balazs Csanad Csaji (2001). Approximation with Artificial Neural Networks. Faculty of Sciences, Etvs Lornd University, Hungary 24(48:7).
[5] Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. Mathematics of Control, Signals, and Systems, 2(4):303–314.
[6] Hornik, Kurt (1991). Approximation capabilities of multilayer feedforward networks. Neural networks, 4(2): 251-257.

Rediscover Newton's Laws based on Explainable AI

- White-box Model for Explanation
  - Variable augmentation $(t_i, \theta_i, r_i, \omega_i)$
  - $\theta_i = NN(t_i), r_i = f(\theta_i) = f\big(NN(t_i)\big), \omega_i = \frac{NN(t_i+\delta)-NN(t_i-\delta)}{2\delta}$
  - Augment variable without prior assumption: $(t_i, \theta_i, r_i, r_i^2, r_i^3, \omega_i, \omega_i^2, \omega_i^3)$

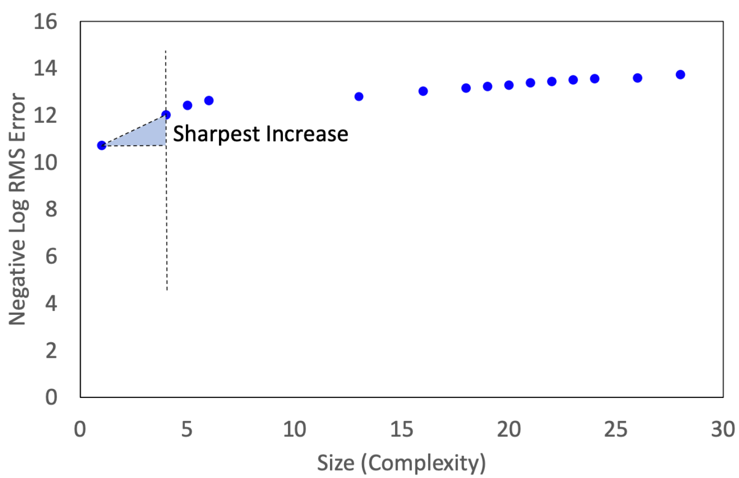


Figure 8: Size and Negative log Error for r and ω

Size	Error	Function
1	0.000022	8.18954×10^{-5}
4	0.000006	$\boxed{\frac{0.000298491}{r_3}}$
5	0.000004	$0.000218591 \cdot (1.92033 - r)$
6	0.000003	$-2.65592 \times 10^{-5} + \frac{0.000390417}{r_3}$
13	0.000003	$-8.50685 \times 10^{-5} + \frac{0.000395123}{r_2 - \frac{0.000290053}{r-1.48997}}$
16	0.000002	$\frac{0.000100316}{-1.08788-0.0590273\cdot\cos(-2147483648\cdot r_3)+r_2}$
22	0.000001	$\frac{0.000448514}{\left(\frac{0.0460772}{r_3-3.36628}\cdot\cos\left(\frac{r_3}{-3.79879\times10^{-5}}\right)+r_3\right)\cdot r}$

Table 6: Symbolic Regression Result for r and ω

$$\omega^2 = \frac{0.000298491}{r^3}, \quad \text{or} \quad r^3\omega^2 = c = 0.000298491\,AU^3 day^{-2}$$

RUTGERS

Rediscover Newton's Laws based on Explainable AI

- Physical Interpretation of the Results

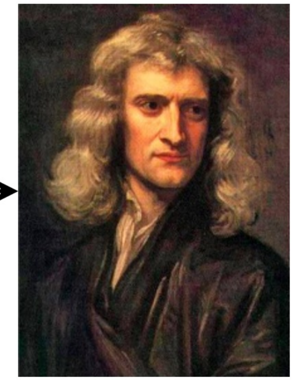
$$\omega^2 = \frac{0.000298491}{r^3}, \quad \text{or} \quad r^3\omega^2 = c = 0.000298491 AU^3 day^{-2}$$

  - $r^3\omega^2$ is close to modern result: $r^3\omega^2 = GM = 2.96\times10^{-4}\ AU^3 day^{-2}$
    - Relative error < 0.8%
  - Acceleration $a = r\omega^2 = \frac{0.000298491}{r^2} \propto \frac{1}{r^2}$
  - Leading to the inverse-square law of acceleration and gravitation
  - Also Kepler's third law
    - $\frac{r^3}{T^2} = \frac{c}{4\pi^2} = 7.56086\times10^{-6} AU^3 day^{-2}$
    - Close to Kepler's result $7.5\times10^{-6} AU^3 day^{-2}$
    - Relative error < 0.82%

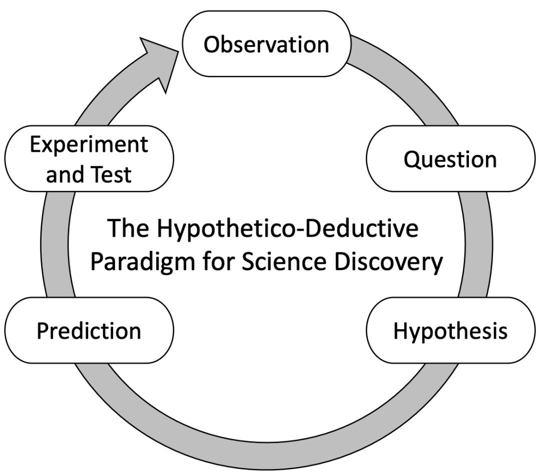
Recap

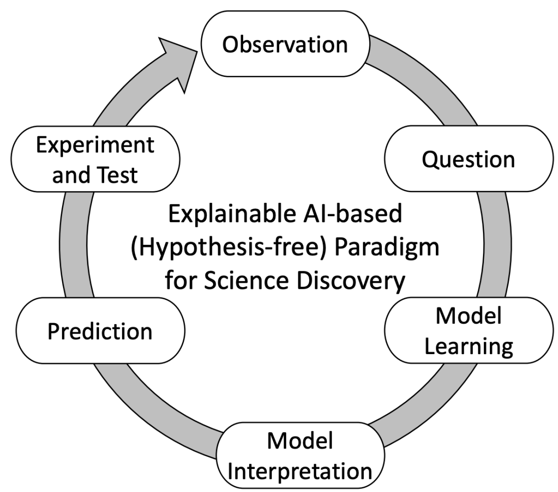
Tycho Brahe (1546-1610)
Observation
Data Collection

Johannes Kepler (1571-1630)
Analyzation
Model Learning



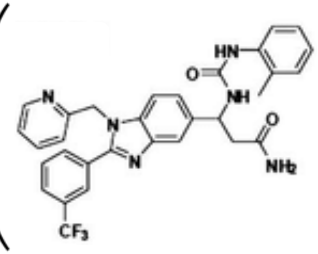
Isaac Newton (1643-1727)
Explanation
Model Interpretation



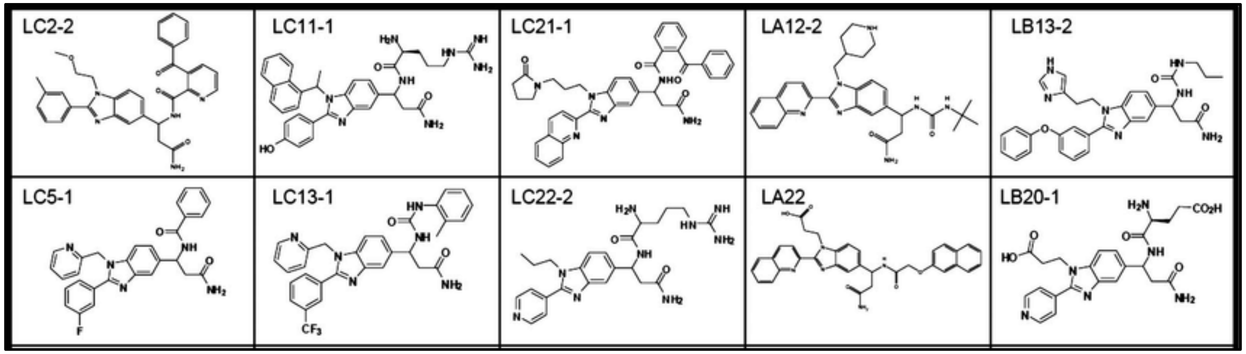


The Molecule Classification Problem

- Predicting the function of molecules
  - A fundamental problem in many chemistry/biological/medical research tasks, e.g., drug discovery
- Mathematically, molecule is a graph
  - Current approaches use Graph Neural Networks (GNN) for prediction
  - E.g., Predict if a molecule is soluble, toxic, or can pass the Blood-Brain Barrier (BBB)
  - A binary classification problem

GNN () = Yes / No

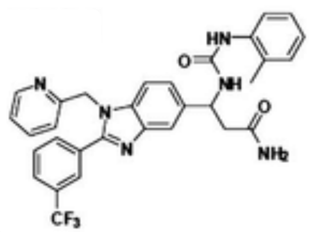
- However, we want to know **why** the model believes in the classification result

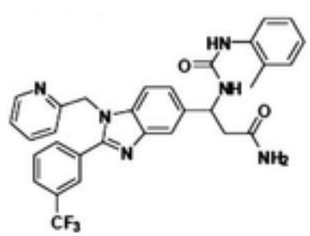


[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y, Li, Y. Zhang. " Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning." WWW 2022.

Explainable Graph Neural Networks

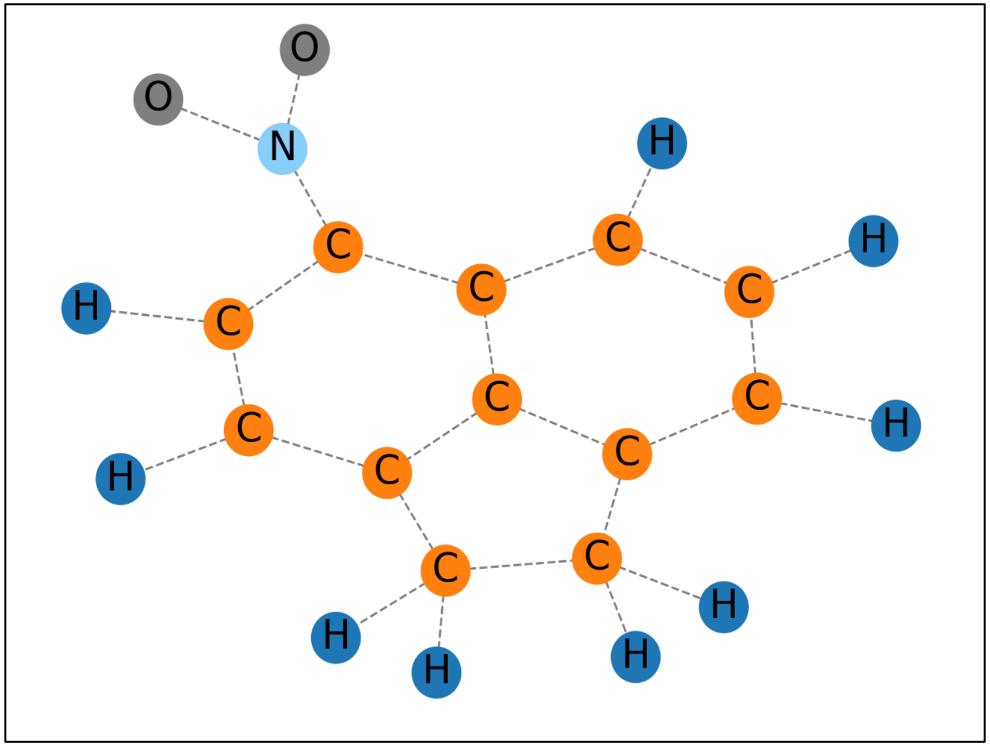
- Our goal is to develop Explainable Graph Neural Networks (XGNN)

GNN () = Yes / No

XGNN () = Yes / No + Explanation

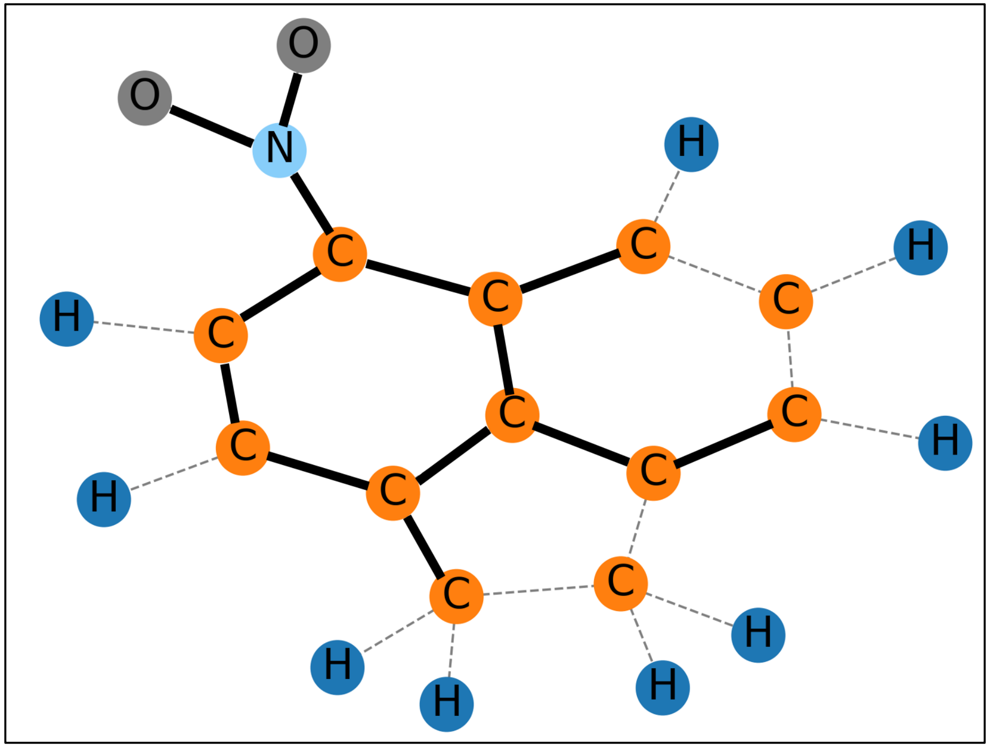
Factual and Counterfactual Explanations

- Example: Molecule mutagenetic prediction
  - If the GNN model predicts the molecule as mutagenetic, why?



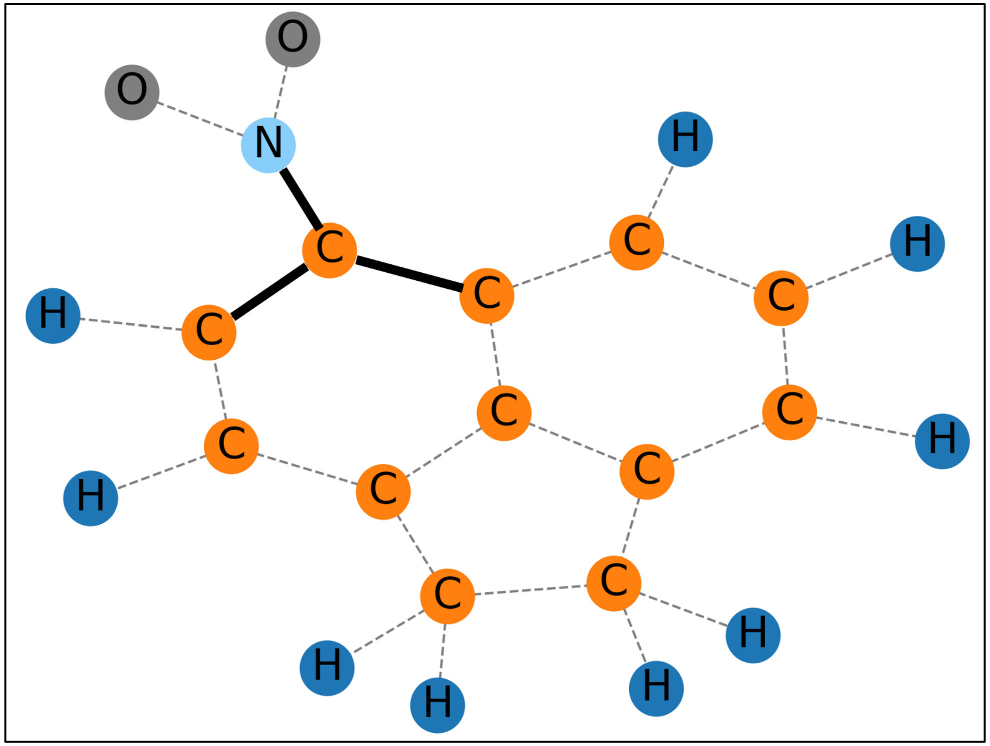
Factual and Counterfactual Explanations

- Factual explanation seeks a sufficient condition
  - The molecule will be mutagenetic with the highlighted bonds



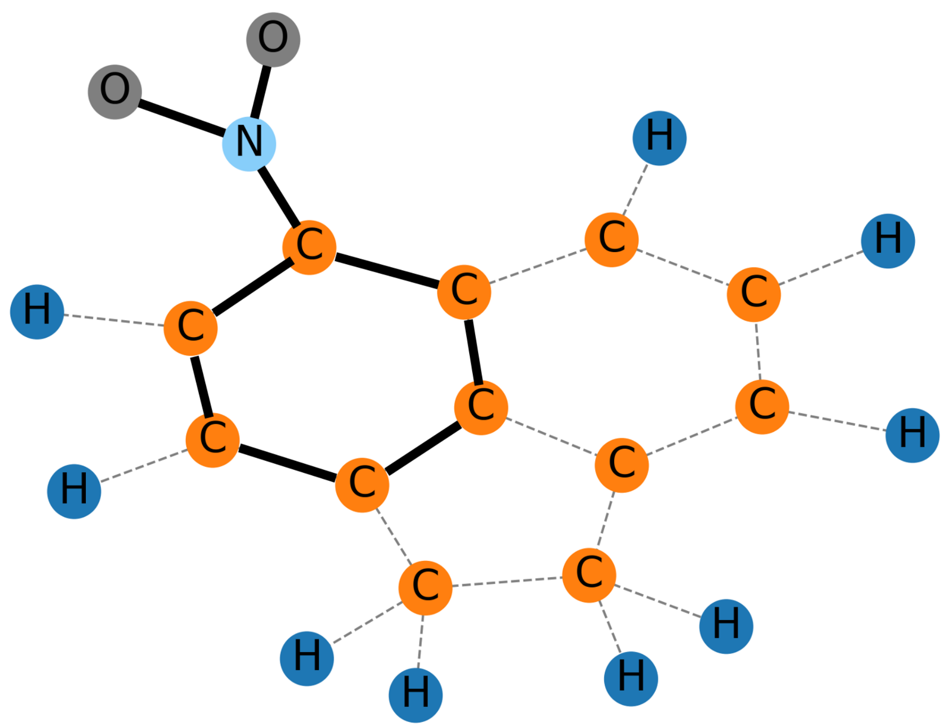
Factual and Counterfactual Explanations

- Counterfactual explanation seeks a necessary condition
  - The molecule will not be mutagenetic without the highlighted bonds



Factual and Counterfactual Explanations

- Factual and Counterfactual explanation seeks a compact (both sufficient and necessary) condition
  - The molecule will be mutagenetic with the highlighted bonds
  - The molecule will not be mutagenetic without the highlighted bonds
  - No more, no less, just OK



How to Find the Explanations?

- A Given graph $G_k = \{\mathcal{V}_k, \mathcal{E}_k\}$. Adjacency matrix $A_k \in \{0,1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$. Node feature matrix $X_k \in \mathbb{R}^{|\mathcal{V}_k| \times d}$.
- The ground-truth class label is $y_k \in C$ (mutagenetic, non-mutagenetic).
- The GNN will predict the estimated label $\hat{y}_k$ for G_k by:

$$\hat{y}_k = \arg\max_{c \in C} P_\Phi(c \mid A_k, X_k)$$

- Generate edge mask $M_k \in \{0, 1\}^{|\mathcal{V}_k| \times |\mathcal{V}_k|}$, feature mask $F_k \in \{0, 1\}^{|\mathcal{V}_k| \times d}$.
- Explanation: Sub-graph $A_k \odot M_k$, sub-features $X_k \odot F_k$.

Explanation Sub-Graph

[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li and Y. Zhang. “Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning”, WWW2022.

How to Find the Explanations?

- Factual Reasoning: "Given A already happened, will B happen?".
- Factual Condition:

$$\underset{c \in C}{\arg\max}\, P_\Phi(c \mid \underbrace{A_k \odot M_k}_{\text{The remaining edges}}, X_k \odot F_k) = \hat{y}_k$$

- Counterfactual Reasoning: "If A did not happen, will B still happen?"
- Counterfactual Condition:

$$\underset{c \in C}{\arg\max}\, P_\Phi(c \mid \underbrace{A_k - A_k \odot M_k}_{\text{The removed edges}}, X_k - X_k \odot F_k) \neq \hat{y}_k$$

[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li and Y. Zhang. "Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning", WWW2022.

RUTGERS

What are good Explanations? Simple and Effective

- Occam's Razor Principle
  - If two explanations are equally effective in explaining the results, we prefer the simpler explanation than the complex one.

- To character Simpleness
  - Explanation Complexity

$$C(M,F) = \|M\|_0 + \|F\|_0$$

How many edges are included in the explanation

How many features are included in the explanation

- To character Effectiveness
  - Factual Explanation Strength

$$S_f(M,F) = P_\Phi(\hat{y}_k \mid A_k \odot M_k, X_k \odot F_k)$$

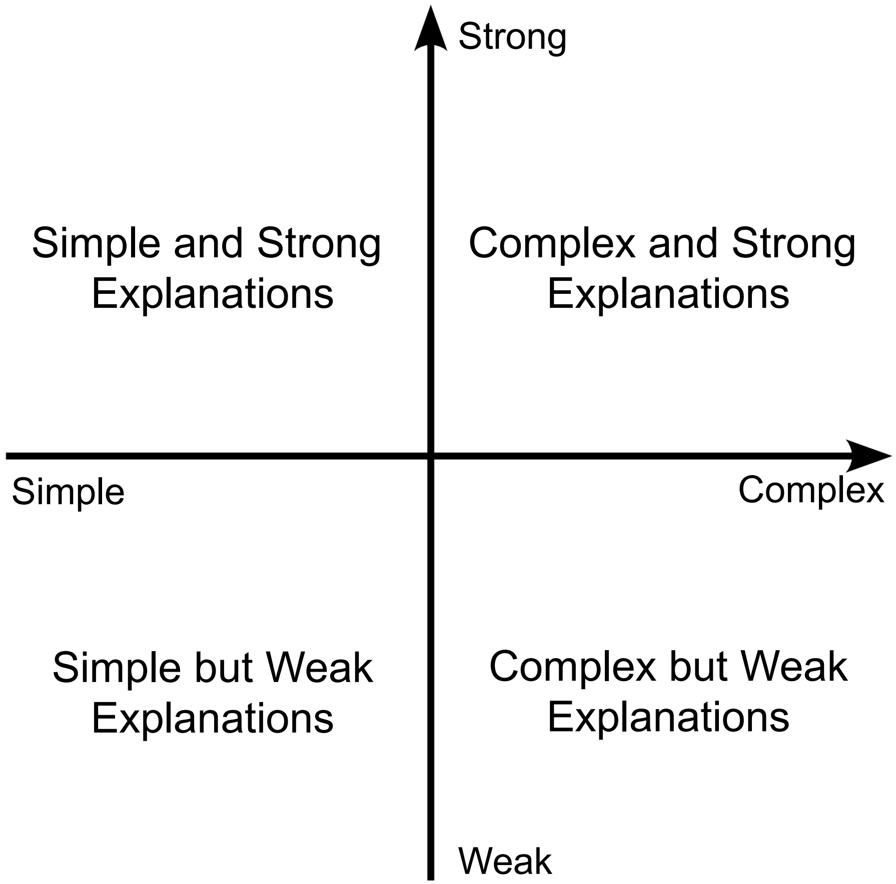
  - Counterfactual Explanation Strength

$$S_c(M,F) = -P_\Phi(\hat{y}_k \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k)$$

Both should be large enough to satisfy the conditions

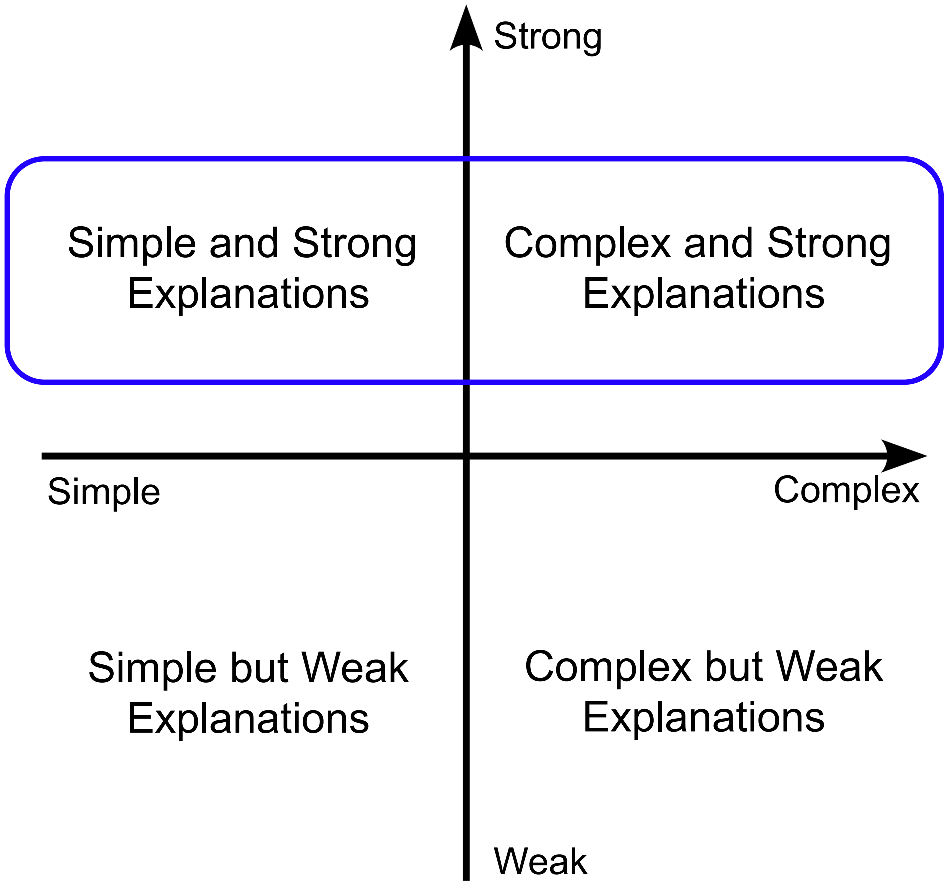
Complexity vs. Strength

- Two orthogonal concepts



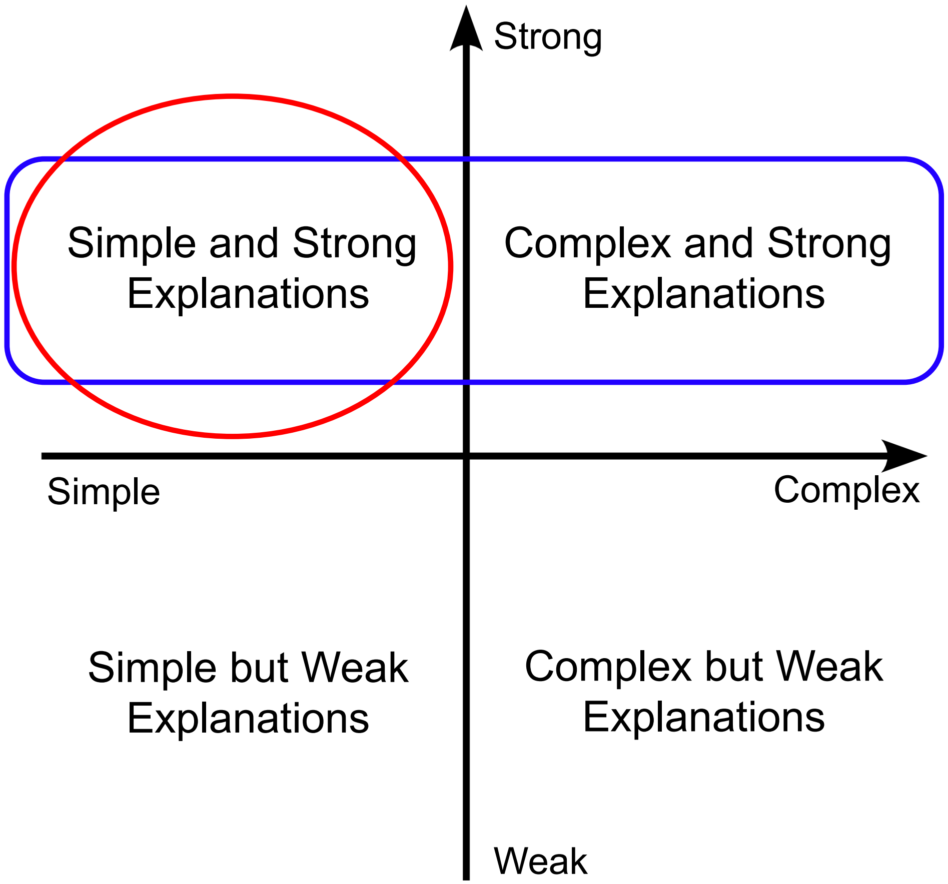
Complexity vs. Strength

- Two orthogonal concepts



Complexity vs. Strength

- Two orthogonal concepts



Counterfactual Learning and Reasoning

- Seek simple and effective explanations

$$\text{minimize} \;\; \text{Explanation Complexity}$$
$$\text{s.t.,} \;\; \text{Explanation is Strong Enough}$$

$$\begin{aligned} &\text{minimize} \;\; C(M_k, F_k) \\ &\text{s.t.,} \; S_f(M_k, F_k) > P_\Phi(\hat{y}_{k,s} \mid A_k \odot M_k, X_k \odot F_k), \\ &S_c(M_k, F_k) > -P_\Phi(\hat{y}_{k,s} \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k) \end{aligned}$$

  - $\hat{y}_{k,s}$ is the label of the second largest prediction probability
  - Idea: Find minimal components of a molecule which is both sufficient and necessary

- Relaxed Optimization based on Lagrange Multiplier for model learning

$$\text{minimize} \;\; \|M_k^*\|_1 + \|F_k^*\|_1 + \lambda(\alpha \underline{L_f} + (1-\alpha)\underline{L_c})$$

$$\begin{aligned} L_f = &\text{ReLU}(\gamma + P_\Phi(\hat{y}_{k,s} \mid A_k \odot M_k^*, \mathcal{X}_k \odot F_k^*) \\ &- S_f(M_k^*, F_k^*)) \end{aligned}$$

$$\begin{aligned} L_c = &\text{ReLU}(\gamma - S_c(M_k^*, F_k^*) \\ &- P_\Phi(\hat{y}_{k,s} \mid A_k - A_k \odot M_k^*, \mathcal{X}_k - \mathcal{X}_k \odot F_k^*)) \end{aligned}$$

Counterfactual Learning and Reasoning

- Seek simple and effective explanations

$$\begin{aligned}&\text{minimize} \;\; \text{Explanation Complexity}\\ &\quad \text{s.t.,} \;\; \text{Explanation is Strong Enough}\end{aligned} \Rightarrow \begin{aligned}&\text{minimize} \;\; C(M_k, F_k)\\ &\text{s.t.,} \; S_f(M_k, F_k) > P_\Phi(\hat{y}_{k,s} \mid A_k \odot M_k, X_k \odot F_k),\\ &S_c(M_k, F_k) > -P_\Phi(\hat{y}_{k,s} \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k)\end{aligned}$$

  - $\hat{y}_{k,s}$ is the label of the second largest prediction probability
  - Idea: Find minimal components of a molecule which is both sufficient and necessary

Objectives	Simple (Complexity)	Effective (Strength)	
Measure	#edges, #features	Sufficiency	Necessity
Method	Minimization	Factual	Counterfactual

Sufficiency and Necessity of Explanations

- $S \Rightarrow N$: S is a sufficient condition for N
- $\neg N \Rightarrow \neg S$: N is a necessary condition for S

[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li and Y. Zhang. “Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning”, WWW2022.

Sufficiency and Necessity of Explanations

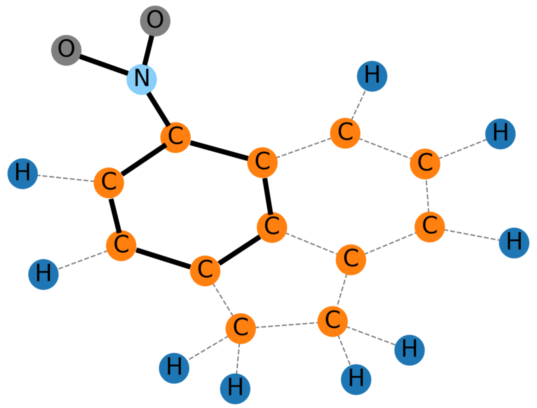
- S ⇒ N: S is a sufficient condition for N
- ¬N ⇒ ¬S: N is a necessary condition for S

- Probability of Sufficient (PS): If we only keep the nodes/edges in the explanation, the prediction result will be the same, then we say the explanation is sufficient

- PS: percentage of molecules whose explanation sub-graph is sufficient

$$\mathrm{PS} = \frac{\sum_{G_k \in \mathcal{G}} \mathrm{ps}_k}{|\mathcal{G}|}, \text{ where } \mathrm{ps}_k = \begin{cases} 1, & \text{if } \hat{y}'_k = \hat{y}_k \\ 0, & \text{else} \end{cases}$$

$$\text{where } \hat{y}'_k = \arg\max_{c \in C} P_\Phi(c \mid A_k \odot M_k, X_k \odot F_k)$$



[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li and Y. Zhang. "Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning", WWW2022.

RUTGERS

Sufficiency and Necessity of Explanations

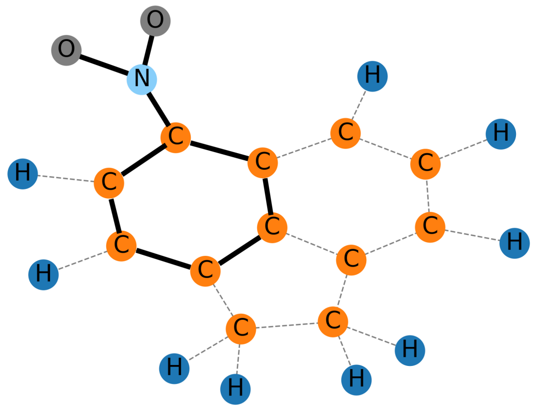
- S ⇒ N: S is a sufficient condition for N
- ¬N ⇒ ¬S: N is a necessary condition for S

- Probability of Necessity (PN): If we remove the nodes/edges in the explanation, the prediction result will change, then we say the explanation is necessary

- PN: percentage of molecules whose explanation sub-graph is necessary

$$\mathrm{PN} = \frac{\sum_{G_k \in \mathcal{G}} \mathrm{pn}_k}{|\mathcal{G}|}, \text{ where } \mathrm{pn}_k = \begin{cases} 1, & \text{if } \hat{y}'_k \neq \hat{y}_k \\ 0, & \text{else} \end{cases}$$

$$\text{where } \hat{y}'_k = \arg\max_{c \in C} P_\Phi(c \mid A_k - A_k \odot M_k, X_k - X_k \odot F_k)$$



[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li and Y. Zhang. "Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning", WWW2022.

Datasets for Evaluation

Dataset	#graph	#ave n	#ave e	#class	#feat	task	gt
BA-Shapes	1	700	4100	4	-	node	✓
Tree-Cycles	1	871	1950	2	-	node	✓
Mutag	4337	30.32	30.77	2	14	graph	
$Mutag_0$	2301	31.74	32.54	2	14	graph	✓
NCI1	4110	29.87	32.30	2	37	graph	
CiteSeer	1	3312	4732	6	3703	node	

[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li and Y. Zhang. “Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning”, WWW2022.

Evaluate Explanation Quality with PN, PS

(without ground-truth explanation)

Models	BA-Shapes				Tree-Cycles				Mutag_0			
	PN%	PS%	F_{NS}%	#exp	PN%	PS%	F_{NS}%	#exp	PN%	PS%	F_{NS}%	#exp
GNNExplainer†	72.19	45.62	55.91	6.00	100.00	59.72	74.78	6.00	71.79	97.44	82.67	15.00
CF-GNNExplainer	75.34	41.10	53.18	5.79	100.00	31.94	48.42	3.44	96.26	7.48	13.88	7.72
Gem†	61.36	52.27	56.45	6.00	100.00	29.89	46.02	6.00	83.01	76.42	79.58	15.00
CF^2	76.73	68.22	**72.07**	6.21	100.00	81.94	**90.08**	5.81	97.44	100.00	**98.70**	14.95

Models	NCI1				CiteSeer (edge)				CiteSeer (feature)			
	PN%	PS%	F_{NS}%	#exp	PN%	PS%	F_{NS}%	#exp	PN%	PS%	F_{NS}%	#exp
GNNExplainer†	92.13	62.16	74.24	15.00	66.67	90.05	76.61	5.00	71.64	99.50	72.79	60.00
CF-GNNExplainer	97.14	31.43	47.49	7.75	69.50	82.00	75.23	2.58	72.14	92.54	81.07	72.91
Gem†	99.03	52.15	68.32	15.00	61.05	72.67	66.36	5.00	-	-	-	-
CF^2	100.00	63.81	**77.91**	17.70	71.00	94.50	**81.08**	3.18	74.63	95.02	**83.60**	62.73

[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li and Y. Zhang. “Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning”, WWW2022.

Evaluate Explanation Quality with Accuracy

(with ground-truth explanation)

Models	BA-Shapes				Tree-Cycles				Mutag_0			
	Acc%	Pr%	Re%	F_1%	Acc%	Pr%	Re%	F_1%	Acc%	Pr%	Re%	F_1%
GNNExplainer†	95.25	60.08	60.08	60.08	92.78	68.06	68.06	68.06	96.96	59.71	85.17	68.85
CF-GNNExplainer	94.39	67.19	54.11	56.79	90.27	<u>87.40</u>	47.45	59.10	96.91	<u>66.09</u>	39.46	47.39
Gem†	**96.97**	64.16	64.16	64.16	89.88	57.23	57.23	57.23	96.43	63.12	47.11	54.68
CF^2	96.37	<u>73.15</u>	<u>68.18</u>	**66.61**	**93.26**	84.92	<u>73.84</u>	**75.69**	**97.34**	65.28	<u>88.59</u>	**72.56**

Kendall's τ and Spearman's ρ correlation scores

Models	BA-Shapes		Tree-Cycles		Mutag_0	
	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$
F_{NS} & F_1	1.00	1.00	1.00	1.00	1.00	1.00
F_{NS} & Acc	0.66	0.79	1.00	1.00	0.66	0.79

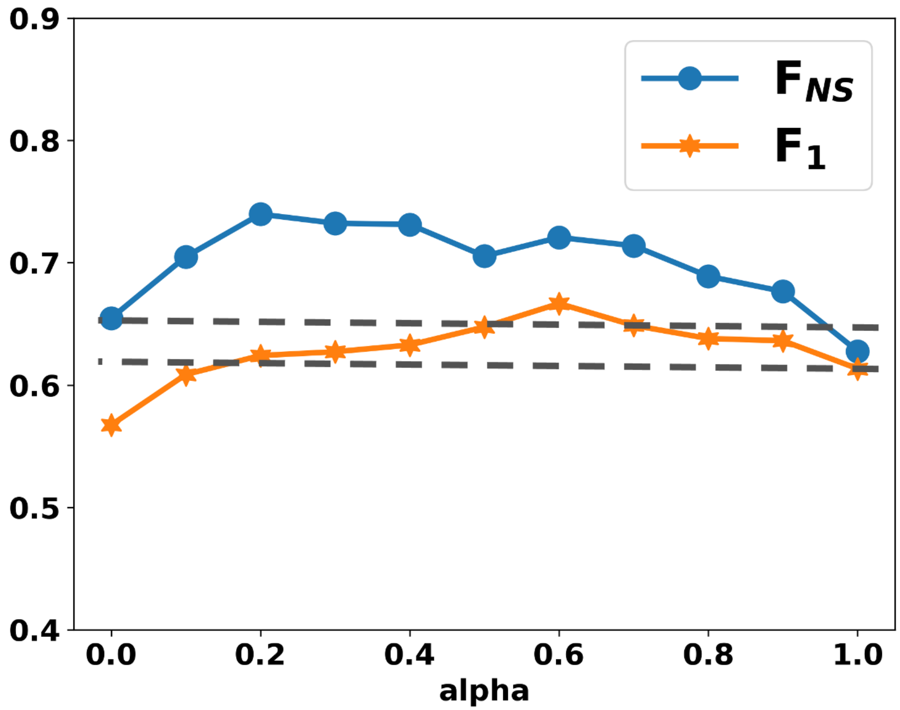
$$F_{NS} = \frac{2PN \cdot PS}{PN + PS}$$

PN/PS-based evaluation is highly consistent with ground-truth-based evaluation.
We can use PN/PS to evaluate explanations when ground-truth is not available

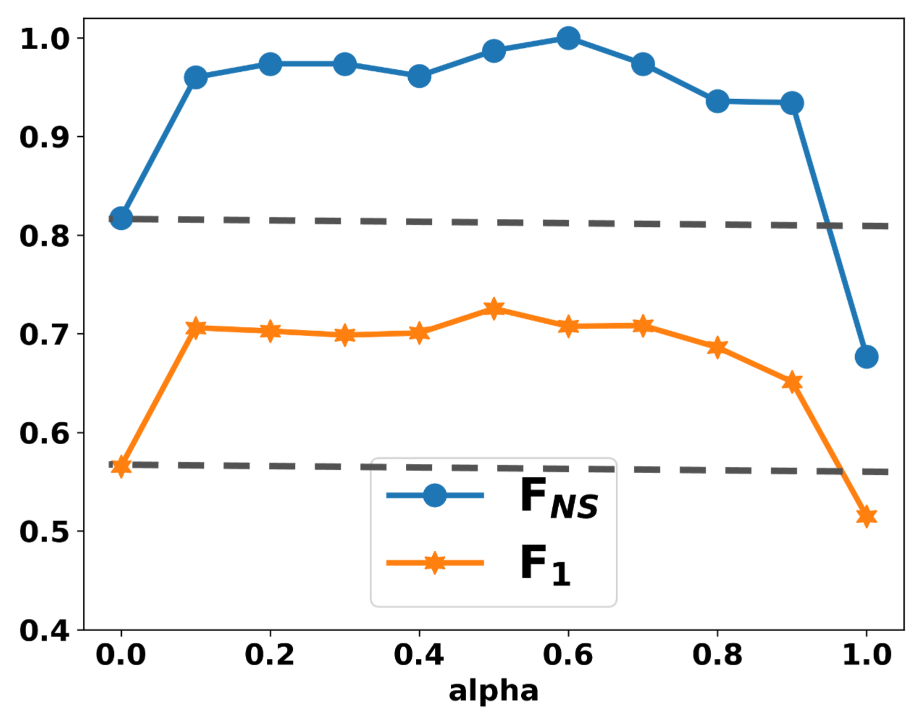
[2] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li and Y. Zhang. "Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning", WWW2022.

Factual vs. Counterfactual Explanations

$$\text{minimize } \|M_k^*\|_1 + \|F_k^*\|_1 + \lambda(\alpha L_f + (1-\alpha)L_c)$$



(a) Influence of α on BA-Shapes

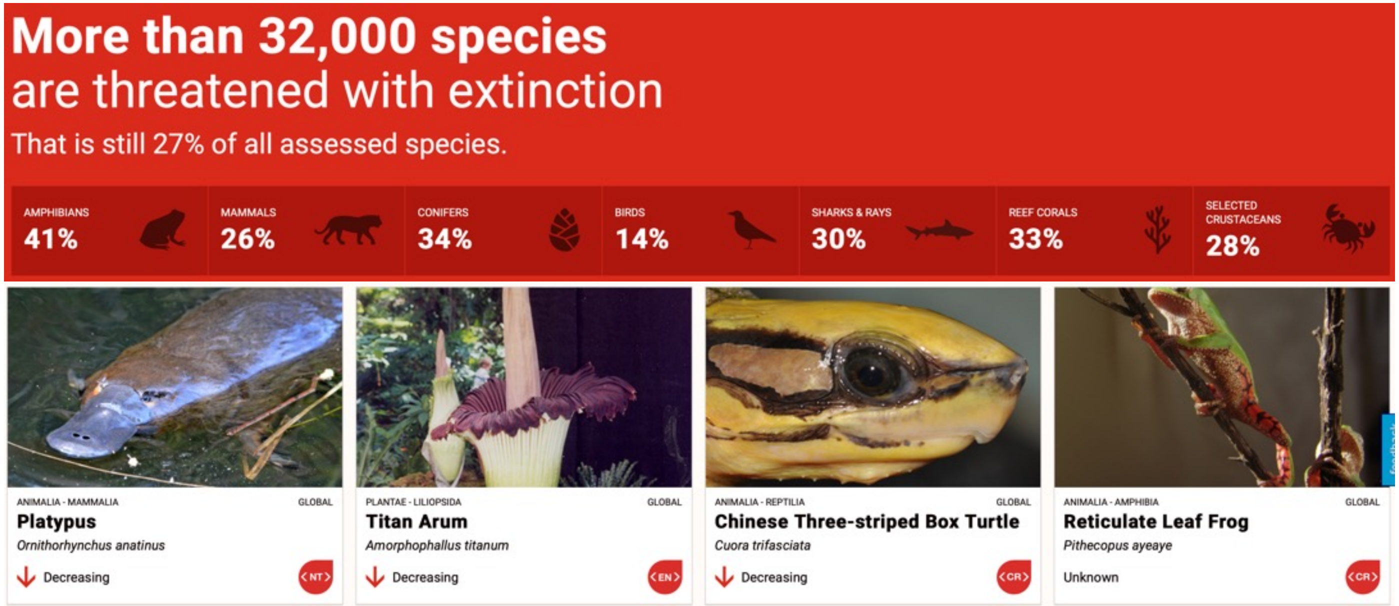


(b) Influence of α on Mutag_0

Both factual and counterfactual reasoning are important

Explainable AI for Biodiversity Conservation

- Task: Predict if a species is endangered or not [15]
  - An important nature-oriented task
  - A dynamic task: species that were not endangered may become endangered now, and vice versa
  - Needs dynamic monitoring and fast reaction
    - E.g., IUCN Red List maintains the status for animal species
    - Critically Endangered, Endangered, Extinct, Extinct in the Wild, Least Concern, Low Risk, Threatened, Vulnerable

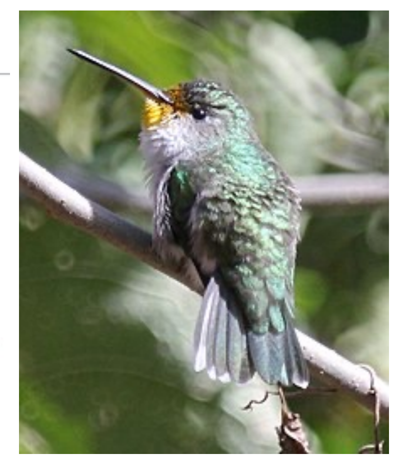


From the IUCN (International Union for Conservation of Nature) Red List of Threatened Species https://www.iucnredlist.org

[19] M. Mukadam, M. Jayaram, and Y. Zhang. “A Representation Learning Approach to Animal Biodiversity Conservation”. COLING 2020.

Explainable AI for Biodiversity Conservation

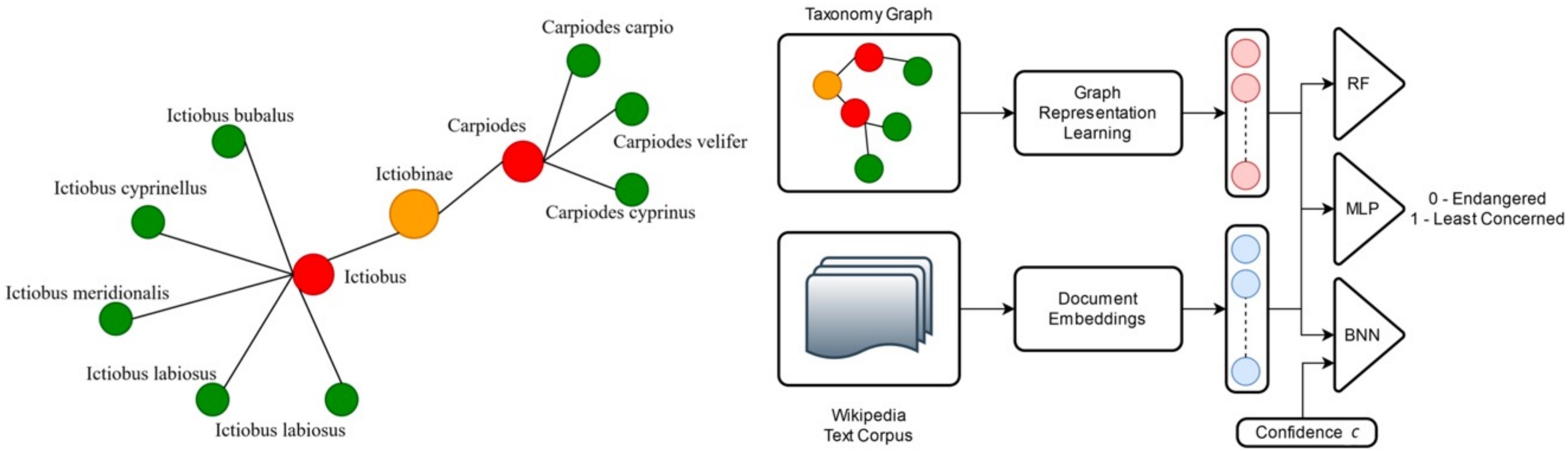
- Machine learning may help as an assistive tool
  - Why Machine Learning may work?
  - Intuition: Species become endangered mostly because habitat destruction due to human activities
  - If we know one species in a habitat is endangered, other species in the same habitat may too
- Habitat (and other useful information) can be found in Wikipedia
  - Information is dynamic/up-to-date due to real-time edits



Explainable AI for Biodiversity Conservation

- Wikipedia text is not enough
  - Due to Interspecific Competition, one species get endangered may imply another competitor species get more populated

- Solution: Graph-Text Co-Learning for Animal Biodiversity Conservation [15]
  - Animal taxonomy graph shows the relationship between species



[19] M. Mukadam, M. Jayaram, and Y. Zhang. “A Representation Learning Approach to Animal Biodiversity Conservation”. COLING 2020.

RUTGERS

Explainable AI for Biodiversity Conservation

Metric	Value
Total animal species	45,170
Endangered animal species	10,947
Least-concern animal species	27,053
Data-Deficient species	7,170
Average length of Wikipedia documents in training corpus (number of words)	146
Documents with length more than average length	13,970
Documents with length less than average length	31,200
Documents that explicitly contain Red List status information	14,253
BNN training data points	14,083
BNN test data points ($c \geq 0.75$)	3,521

Dataset statistics (data collected from Wikipedia, IUCN, and ITIS)

Model	F1 Score	-AUC
RF w/ node2vec	0.862	0.784
RF w/ doc2vec	**0.869**	0.820
RF w/ node2vec + doc2vec	0.860	**0.827**
MLP w/ node2vec	0.843	0.729
MLP w/ doc2vec	**0.886**	0.864
MLP w/ node2vec + doc2vec	0.885	**0.873**
BNN w/ node2vec + doc2vec + $c \geq 0.75$	0.856	0.868
BNN w/ node2vec + doc2vec + $c \geq 0.9$	**0.889**	**0.911**

Prediction accuracy

Nycticryphes semicollaris - The South American painted-snipe (Nycticryphes semicollaris), or lesser painted-snipe, is a shorebird in the family Rostratulidae. There are two other species in its family, the Australian painted-snipe and the greater painted-snipe. Measurements: 19–23 cm in length; 65–86 g in weight. Vocalizations: A hoarse, hissing “wee-oo” has been recorded from birds in captivity. Distribution and habitat: The species is found in the southern third of South America, from southern Brazil, Paraguay, and Uruguay to Chile and Argentina. It inhabits lowland freshwater wetlands, including wet grasslands. Breeding: South American painted-snipes are monogamous and breed semi-colonially. The nest is a shallow cup on the ground in a wetland, with a clutch of 2-3 eggs. Breeding has been recorded mainly from July to February. Feeding: The South American painted-snipe is omnivorous, feeding by probing in mud and shallow water for small animals and seeds, often at dusk.

Attention-based Explanation

IUCN: International Union for Conservation of Nature (for ground-truth label)
ITIS: Integrated Taxonomic Information System (for Animal taxonomy graph) https://www.itis.gov

Summary

- Rediscover Kepler's laws and Newton's laws from Tycho's ancient data [1]
  - A good example to demonstrate the idea of XAI-driven scientific research
  - Pay our respect to some of the greatest minds in human history
- More "practical" Examples
  - Explainable AI for Molecular Property Prediction [2]
  - Explainable AI for Biodiversity Conservation [3]

- [1] Zelong Li, Jianchao Ji, and Yongfeng Zhang. "From Kepler to Newton: Explainable AI for Science Discovery." In ICML AI for Science 2022.
- [2] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. "Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning." In Proceedings of the ACM Web Conference 2022.
- [3] Meet Mukadam, Mandhara Jayaram, and Yongfeng Zhang. "A Representation Learning Approach to Animal Biodiversity Conservation." In Proceedings of the 28th International Conference on Computational Linguistics. 2020.

Yongfeng Zhang
Department of Computer Science, Rutgers University
yongfeng.zhang@rutgers.edu