



Data Science for Researchers and Scholars

Vasant G. Honavar

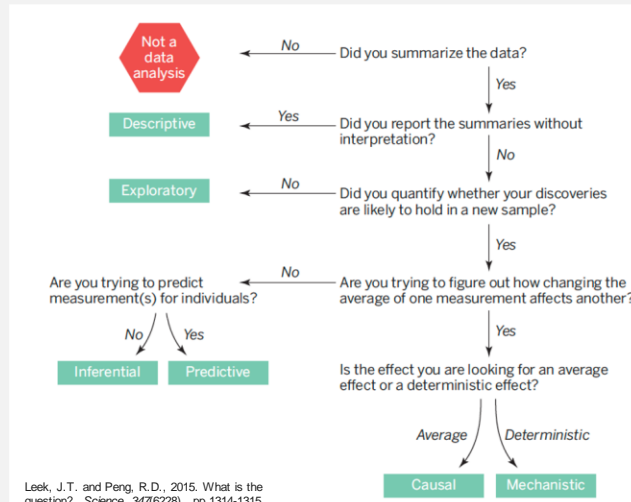
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics,
Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
<http://faculty.ist.psu.edu/vhonavar>
<http://ailab.ist.psu.edu>

Course Project

- Phase 1
 - Define research questions
 - You may draw on your own research or
 - Attempt to fill a gap in published studies
 - Identify suitable data set(s) that will help you answer your research questions
 - Prepare and submit project abstract – Due October 5, 2023
- Phase 2
 - Design your study and analysis
 - Analyses can be descriptive, predictive and/or causal
 - Identify suitable analysis methods
 - Prepare and submit your project proposal – Due October 26, 2023
- Phase 3
 - Conduct your study
 - Analyze and interpret your results
 - Prepare and submit your project report – Due November 30, 2023

Data Science Starts with a Question



Leek, J.T. and Peng, R.D., 2015. What is the question?. *Science*, 347(6228), pp.1314-1315.

Data science begins with a question

- Questions come in many forms

Question type	Description	Example
Descriptive	A question about summary characteristics of a data set without interpretation (i.e., report a fact).	How many students are enrolled at Penn State in Fall 2023?
Exploratory	A question about patterns, trends, or relationships within a single data set. Often used to propose hypotheses for future study.	Do political party preferences change with indicators of wealth in a collected sample of 2000 individuals US?

Data science begins with a question

- Questions come in many forms

Question type	Description	Example
Predictive	A question about prediction of an outcome of interest, but not what causes the outcome.	What political party will Joe Sixpack vote for in the next US Presidential election?
Inferential	A question about patterns, trends, or relationships in a single data set and quantification of how applicable these findings are to the wider population.	Do political party preferences change with indicators of wealth for all people living in the US?

Data science begins with a question

- Questions come in many forms

Question type	Description	Example
Causal	A question about whether changing one factor will lead to a change in another factor, on average, in the wider population.	Does college education causally impact voting for a certain political party in the US elections?
Mechanistic	A question about the underlying mechanism of the observed patterns, trends, or relationships (i.e., how does it happen?)	How do wealth lead to voting for a certain political party in the US elections?

- Mechanistic questions are beyond the scope of this course

Assembling data

- Good data scientists spend significant effort
 - Assembling the data needed to answer their research questions
 - Preparing data for analyses
- The rest spend whine about lack of data

Notebook Environments

- **Expect** to have to redo your analysis from scratch, so build your data analysis workflows to make it possible.
- Notebook environments
 - ❖ Mix code, data, computational results, and documentation into easy-to-maintain data analysis workflows
 - ❖ Make projects
 - Self-contained
 - Reproducible
 - Modifiable
 - Extensible
 - Shareable
 - Documented

Standard Data Formats

Common domain-agnostic data formats

- **CSV files:** for tables like spreadsheets
- **XML:** for structured but non-tabular data
- **JSON:** Javascript Object Notation for APIs
- **SQL databases:** for multiple related tables
- **RDF triple stores:** for scalable analysis
- **Pickle:** serialize and de-serialize complex objects

Domain-specific data standards

- **Omics:** FASTA, MIAME, MIAPE, MIAMET, PDB ..
- **EHR:** OMOP
- **Chemistry:** SMILES, PubChem, XYZ ..

Finding data

- Finding the right data set is critical for answering your research questions
 - Collect the data that you need for your study
 - Repurpose the existing data
 - The latter is increasingly common with the advent of big data
 - **Example:** Electronic Health Records data collected during the course of healthcare delivery used for understanding health disparities, predicting health risks, etc.
- Repurposing data requires
 - Digging into the “story” behind the data
 - Imagination as well as care

Sources of Bias in Data Sets

- **Selection Bias:** when the sample fails to reflect the entire target population
 - Surveys done on Internet exclude people without internet access
 - Analyses of current employees of a corporation won't tell us whether there is discrimination in hiring
- **Historical Bias:** when the world changes during data collection (e.g. Twitter data)
- **Survivorship Bias:** when only successes are recorded in the data
 - Data about COVID survivors alone can't tell us factors that contribute to survival
- **Dunning Kruger effect:** Those least competent in a subject area overestimate their competence the most and vice versa

...

Finding data

- Proprietary data
- Government data
- Academic data
- Web search
- Sensor data
- Crowd-sourced data
- Self-curated data

<https://datasetsearch.research.google.com>

Proprietary Data

- Facebook, Google, Amazon, Blue Cross, Highmark, etc. have exciting user/transaction/log data sets
- Most organizations have internal data sets of interest to their business
- Getting access as an outsider is usually impossible
- Getting access is sometimes possible
 - Collaboration
 - Internships
- Companies, e.g., Google sometimes offer rate-limited APIs
- The commercial promise of generative AI (e.g., large language models) has prompted once open data providers like X (formerly known as Twitter) and Reddit to tighten up
- Data may be contaminated with AI-generated data

Government Data

- City, State, and Federal governments are increasingly committed to open data.
 - Data.gov (233,957 open data sets)
 - Data.Europa.eu
 - Data.gov.in
 - Data.gov.au
 - dataportal.opendataforafrica.org
- The Freedom of Information Act (FOIA) enables you to request data if it is not open
- Privacy protection and national security considerations impact data availability

Academic Data

- Data sharing is now a requirement for federally funded projects and publication in many fields.
- You can often find such data if you look hard enough
- Track down from relevant papers, and ask.
- Google topic and “Open Science” or “data”
- Major data repositories are available in some fields
 - PDB for structural biology
 - Genbank for genomics
 - Materials cloud for material science
 - NACC for Alzheimer’s
 - Neurodata for brain connectomics
 - Ourworldindata for assorted data sets for addressing global problems
 - GDELT contains geo-indexed data about human society

Web Scraping

- Web scraping refers to extracting text/data from webpages
- Python Libraries (BeautifulSoup, Selenium), help parse/scrape the web
 - Are APIs available from the source?
 - Did someone previously write a scraper?
- Terms of service limit what you can legally do.

Sensor Data

The “Internet of Things”

- Image/video data
 - Crowd dynamics at major events
 - Crop health
 - Flooding
 - Vegetation cover
- Sensor data
 - Wearable sensors (smart watches) for health monitoring
 - Aircraft meteorological reports
 - Video surveillance data

Build logging systems: storage is cheap!

Crowdsourcing

Many open data resources have been built by crowdsourcing

- Wikipedia/WikiData
- IMDB
- E-bird
- Zooniverse
- ...

Crowdsourcing platforms like Amazon Turk enable you to pay for armies of people to help you gather data, like human annotation.

Self-curated data

- Sometimes you may have to assemble data on your own
- Much historical data still exists only on paper or PDF, requiring manual/semi-automated data entry/curation
- At one record per minute, you can enter 1,000 records in only two work days
- Crowdsourcing can amplify your effort
- Unique data sets that help answer interesting questions are important scientific contributions
- Example:

Cleaning Data: Garbage In, Garbage Out

Many issues arise in ensuring the sensible analysis of data from the field, including:

- Distinguishing errors from artifacts
- Data compatibility / unification
- Imputation of missing data
- Identifying outliers
- ...

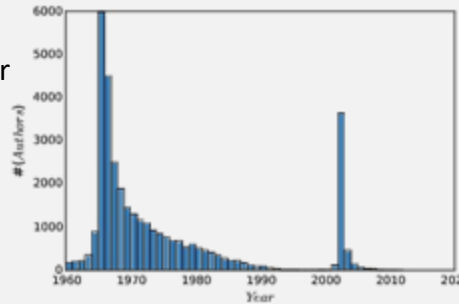
Errors vs. Artifacts

- **Data errors** arise due to uncorrectable problems arising from data acquisition
 - Example: human data entry error in EHR
 - Hard to fix after the fact
- **Artifacts** are systematic problems arising during data acquisition or processing
 - Example: Systematic measurement error
 - Example: Background noise in a video recording
 - The key to detecting artifacts is examining the data closely enough to detect potential artifacts
 - Possible to fix if we know enough about the source of the artifacts

Perils of fishing for wisdom in oceans of data

Cautionary Tale 1: Scientific Authors by Year*

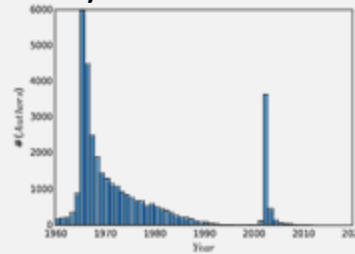
- A bibliographic study analyzed PubMed data to identify the year of first publication for the 100,000 most frequently cited authors yielding the results shown
- What *should* the distribution of the year of first publication of top authors by year look like?
 - Relatively flat
 - Perhaps gradually increasing as population of researchers expands



*Source: Skiena, S. Data Science Design Manual

Cautionary Tale 1: Scientific Authors by Year*

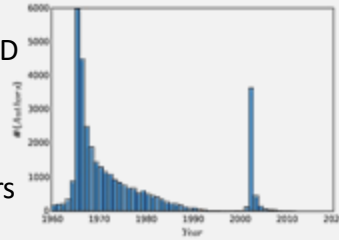
- **Do you see anything fishy?**
 - There is a huge bump around 1965
 - And another significant bump around 2002
 - Almost no stars emerging between 1990 and 2000!
- **What might explain the 1965 bump?**
 - PUBMED started systematically indexing publications starting in 1965
 - Top cited authors who had been publishing before 1965 will appear in the database only around 1965



*Source: Skiena, S. Data Science Design Manual

Cautionary Tale 1: Scientific Authors by Year*

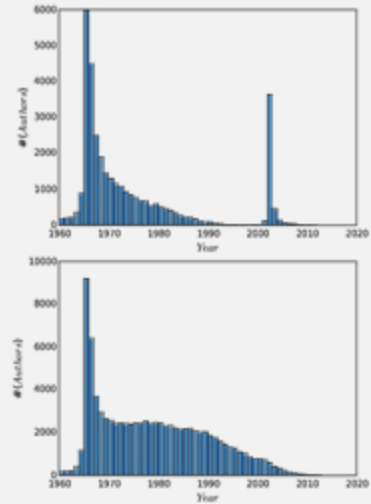
- **What might explain the 2002 bump?**
 - Sudden rise in scientific output? PUBMED started using first names starting in late 2001
 - V. Honavar became Vasant Honavar
 - Unusually large number of “new” authors suddenly appeared in the database, and some eventually became top cited, leading to a bump around 2002
- **What explains the paucity of stars between 1990 and 2000?**
 - A star author entering in 1990 is unlikely to show up on our plot because by 2001, their name would change, reducing their citation count



*Source: Skiena, S. Data Science Design Manual

Cautionary Tale 1: Scientific Authors by Year*

- What if we merge the records of authors by examining last names, first initials, first names, and other relevant information?
- The 1968 bump does not disappear (but can be explained)
- The 2001/2002 bump disappears



*Source: Skiena, S. Data Science Design Manual

Perils of fishing for wisdom in oceans of data

- It makes no sense to compare weights of 123.5 against 78.9, when one is in pounds and the other is in kilograms.
- It makes no sense to directly compare the movie gross of Gone with the Wind against that of Avatar, because 1939 dollars are 15.43 times more valuable than 2009 dollars.
- It makes no sense to compare the price of gold at noon today in New York and London, because the time zones are five hours off, and the prices affected by intervening events.
- It makes no sense to compare the stock price of Microsoft on February 17, 2003 to that of February 18, 2003, because the intervening 2-for-1 stock split cut the price in half, but reflects no change in real value.
- NASA lost the \$125 million MARS climate orbiter on September 23, 1999 due to a unit conversion error.

Cautionary Tale 2: Protein Annotation Using Machine Learning*



- Exponential increase in protein sequences
- Experimental determination of structure and function lags behind
- Automated methods for protein function annotation
 - Allow high-throughput annotation of thousands of sequences
 - Increase the risk of error propagation

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

- **Goal: Automate the subclassification of protein kinases**
- **Protein Kinases are among**
 - **The most well-studied proteins**
 - **The most popular drug targets**
- **Two broad classes (some have dual specificity)**
 - **Serine/Threonine kinases**
 - **Tyrosine kinases**
- **Protein serine/threonine phosphorylation regulates virtually every signaling pathway in the eukaryotic cell**
- **Tyrosine phosphorylation modulates key biological events associated with cancer, diabetes, and inflammation**
- **Accurate annotation extremely important**

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

Data Set: Human and Mouse Protein Kinases with Gene ontology annotations (www.geneontology.org) via AmiGO

GO:0003674 : molecular_function (121801)

- GO:0003824 : catalytic activity (41632)

- GO:0016740 : transferase activity (13210)

- GO:0016301 : kinase activity (5613)

- GO:0004672 : protein kinase activity (3415)

- GO:0004674 : protein serine/threonine kinase activity (2077)

- GO:0004713 : protein-tyrosine kinase activity (771)

Data retrieved in 2007 [Andorf et al., BMC Bioinformatics, 2007]

Source: Andorf, Carson, Dena Dibbs, and Vasant Honavar. "Exploiting Inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

- Initial goal: **predicting protein kinase subclasses using machine learning**
- Machine learning algorithms
 - Naïve Bayes: Amino acid composition
 - NB(k): Extension of Naïve Bayes to k th order Markov model
 - SVMs using these data representations
 - A hybrid algorithm that combines the above with an annotation transfer based on sequence homology (BLAST)
- Initial Question: **how effective are these methods on classifying kinases?**

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

- **Data: 244 mouse and 330 human protein sequences**
 - GO families GO0004674 (Serine/Threonine Kinase)
 - GO0004713 (Tyrosine Kinase)
- **Reference class labels: annotations returned by AmiGO**
 - 71 mouse and 233 human proteins are labeled with GO0004674
 - 106 mouse and 90 human proteins are labeled with GO0004713
 - 67 mouse and 7 human proteins had both labels
- **Train classifier on human data and test on mouse data**

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

- Train classifier on human data and test on human data
- Classifier trained on human data and tested on human data (cross validation)
 - 89.1% accuracy with a 0.85 correlation coefficient
 - Good! 😊
- But really, what we want is a classifier that can be used on data from new species as they are sequenced
- How would the classifier perform on mouse kinases?

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

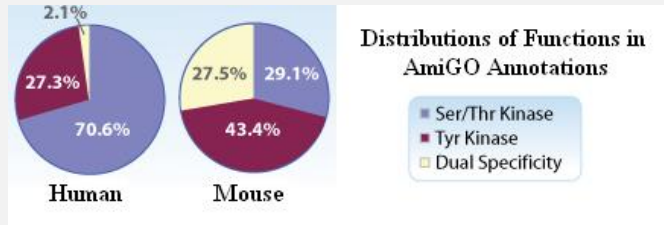
Cautionary Tale 2: Protein Annotation Using Machine Learning*

- Train classifier on human data and test on mouse data
- Classifier trained on human data and tested on mouse data
 - 15.1% accuracy and a -0.42 correlation coefficient
 - Really, really, Bad! ☹️
- Result surprising because
 - Human and mouse kinases share common origin (homologues)
- Question: **How could this be?**

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

Surprising discrepancy in relative distribution of kinase subtypes



Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

Reexamining the kinase subtype annotations in the data

Annotations came with different levels of confidence (evidence codes)

- 211 of the 244 mouse protein kinases had a RCA (inferred from reviewed computational analysis) evidence code
- Of the 33 mouse proteins that did not have a RCA evidence code, 28 were classified correctly by the classifier trained on human data
- Question: What is special about the 211 mouse proteins with GO function labels with RCA evidence code?

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

Reexamining the kinase subtype annotations in the data

Annotations came with different levels of confidence (evidence codes)

- 211 of the 244 mouse protein kinases had a RCA (inferred from reviewed computational analysis) evidence code
- Of the 33 mouse proteins that did not have a RCA evidence code, 28 were classified correctly by the classifier trained on human data
- Question: What is special about the 211 mouse proteins with GO function labels with RCA evidence code?

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

- Annotations came from the Mouse Genome Informatics Database (MGI)
- The MGI annotations came from the Fantom2 (Functional Annotation of Mouse) Database
- Each of the 211 mouse proteins had at least one RCA from FANTOM Consortium and the RIKEN Genome Exploration Research Group (Okazaki et al, Nature, 420, 563-573, 2002)
- Are there other independent annotations for these proteins?
 - Fortunately Yes - UniProt

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

- AmiGO RCA annotations for 201 of the 211 mouse proteins were inconsistent with UniProt annotations

KINASE FAMILY	AmiGO Ser/Thr	AmiGO Tyr	AmiGO dual specificity
UniProt Ser/Thr	10	105	35
UniProt Tyr	54	0	3
UniProt dual specificity	0	4	0

- A search of the Mouse Kinome Database shows that 154 of the 244 mouse kinases have a human ortholog with sequence similarity greater than 90%!
- So this must be an easy problem for machine learning
- Why does machine learning fail on this problem?

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8, 1 (2007): 284.

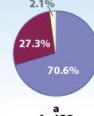
PennState
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

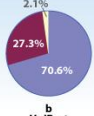
PennState
Clinical and Translational
Science Institute

Cautionary Tale 2: Protein Annotation Using Machine Learning*

HUMAN

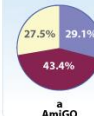


a
AmiGO

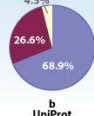


b
UniProt

MOUSE



a
AmiGO



b
UniProt

■ Ser/Thr Kinase
■ Ty Kinase
■ Dual Specificity

**Comparison of the Distributions of Functions
in AmiGO and UniProt Annotations**

Distribution of kinase subfamilies look similar a cross human and mouse based on the UniProt annotations!

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

PennState
Institute for Computational
and Data Sciences

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023

Cautionary Tale 2: Protein Annotation Using Machine Learning*

Story So far

- When reference annotations are from AmiGO:
 - Classifier trained on human kinases and tested on human kinases – good
 - Classifier trained on human kinases and tested on mouse kinases – bad
- AmiGO RCA and UniProt annotations inconsistent
- Questions:
 - Could the AmiGO RCA annotations be incorrect?
 - How does the classifier trained on human and tested on mouse perform when the reference annotations are from UniProt?

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar: "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

- Use UniProt labels instead of original (AmiGO) labels as reference
- Train classifier using Human proteins and test on mouse proteins
- Test accuracy on mouse proteins: 97%!
- 205 of the 211 proteins that were mislabeled with respect to AmiGO reference labels were correctly labeled with respect to UniProt reference labels

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar: "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

- There is no reason to expect that the relative distribution of the Ser/Thr kinases and Tyr kinases in human and mouse would be dissimilar
- The machine learning approach used is sound, and found effective in other macromolecular sequence classification tasks
- **Could it be the case that the annotations returned by AmiGO for the 211 mouse protein kinases (nearly 95% of the 244 mouse protein kinases) are incorrect?**

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar: "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

Implications

- To the best of our knowledge, the problematic mouse kinase annotations with RCA evidence code
 - Came from Okazaki et al, Nature, 420, 563-573, 2002
 - Were propagated to MGI through the Fantom2 (Functional Annotation of Mouse) Database
 - And from MGI to AmiGO
- Examination of GO annotation is often the first step in many high throughput studies e.g., gene expression analysis
- Question: **How far did these annotations propagate?**

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar: "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

Implications

- 136 rat protein kinase annotations from AmiGO had:
 - ISS - **inferred based on sequence or structural similarity-evidence code**
 - **Functions assigned based on some of the 201 potentially incorrectly annotated mouse proteins**
 - 94 Ser/Thr kinase proteins mislabeled as either a Tyr kinase or dual specific
 - 42 Tyr kinase proteins mislabeled as a Ser/Thr kinase or a dual specific
- **201 mouse and 136 rat protein kinase annotations are probably incorrect!**
- **Not to mention annotations of kinases in other species and analyses that relied on these erroneous annotations!**

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar: "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Cautionary Tale 2: Protein Annotation Using Machine Learning*

Implications

- The apparent *failure* of a machine learning approach helped us discover potential errors in annotations
- This story underscores the need for better protocols and tools for
 - Algorithm testing and validation, bug tracking, etc.
 - Multiple checks for consistency of annotations – especially in the case of annotations with RCA and ISS evidence codes
 - Better methods for tracking propagation of annotations across databases
 - Reproducible (and correctible) computational workflows
- The erroneous mouse kinase annotations were traced to errors in annotation scripts used and have since been fixed by the MGI 😊

Source: Andorf, Carson, Drena Dobbs, and Vasant Honavar: "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.