WELCOME!

Summer School will begin at 8:30 a.m. PDT

8:30-8:35 a.m.	Introduction	Luke Durell
8:35-9:20	Data Science for ‘Omics Data	Kelly Stratton & Marija Velickovic
9:20-10:05	Data Science Workflows	Eva Brayfindley
10:05-10:15	Networking Break	
10:15-11:00	Data Science Communication	Moses Obiri
11:00-11:45	EMSL Data Science Tools	David Degnan

EMSL

Summer School Day 1: Data Science Tools, Workflows, and Best Practices

Natalie Winans & Luke Durell
Biostatistics & Data Science; Spatial Statistics

07.24.2023

8:30-8:35 a.m.	Introduction	Luke Durell
8:35-9:20	**Data Science for ‘Omics Data**	**Kelly Stratton & Marija Velickovic**
9:20-10:05	Data Science Workflows	Eva Brayfindley
10:05-10:15	Networking Break	
10:15-11:00	Data Science Communication	Moses Obiri
11:00-11:45	EMSL Data Science Tools	David Degnan

EMSL

Data Science for ‘Omics Data

Day 1: Tools of the Trade

Kelly Stratton & Marija Velickovic

Instructor Intro

Kelly Stratton

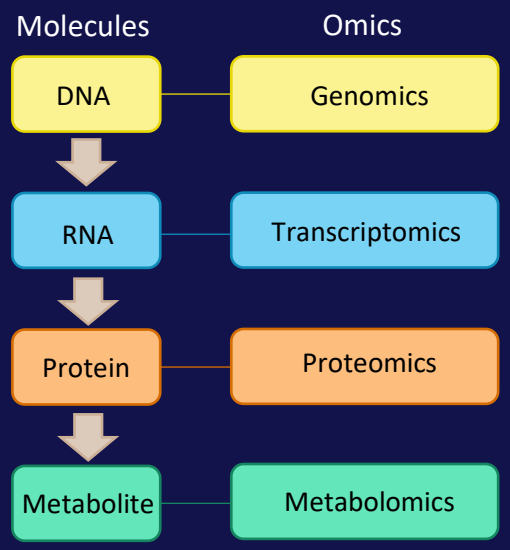
- Data Scientist, Data Transformations IRP Lead
- Statistics, R, visualization, analysis of ‘omics data
- Day 1: Data Science for ‘Omics Data
- kelly.stratton@pnnl.gov
- (509) 372-4349

Instructor Intro

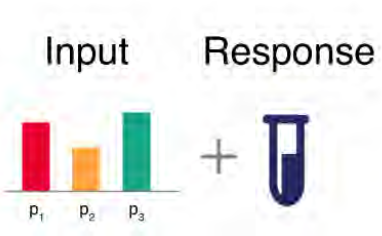
Marija Velickovic

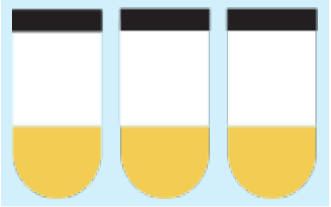
- Chemist
- Spatial proteomic and metabolomics sample prep
- Day 1: Data Science for ‘Omics Data
- marija.velickovic@pnnl.gov
- (509) 371-8867

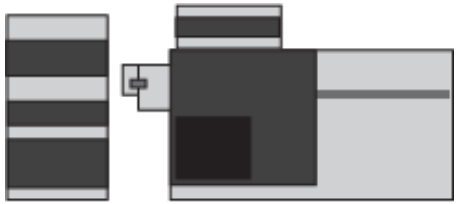
Many Opportunities for Data Science of ‘Omics Data

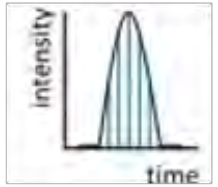


Experimental Design → Sample Collection → Instrument Runs → Molecule Identification

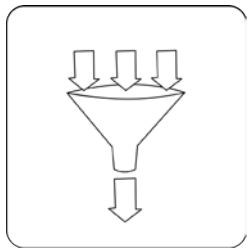


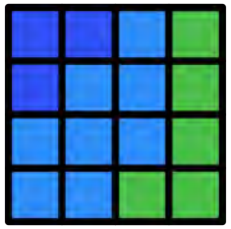


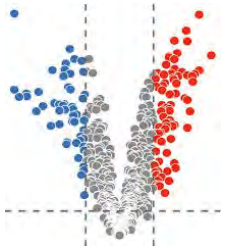




Annotation/ Assembly → Quality Control → Visualization/ Summarization → Statistical Analysis

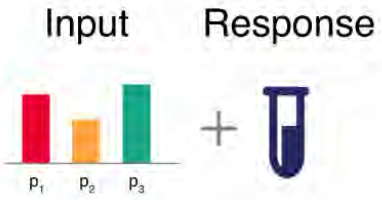


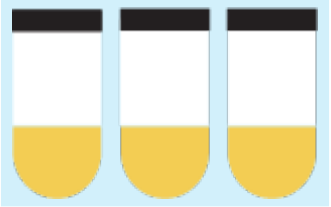


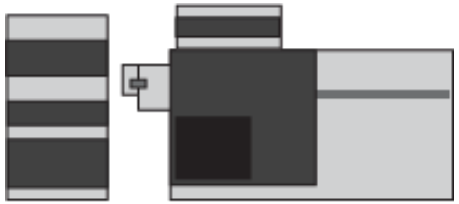


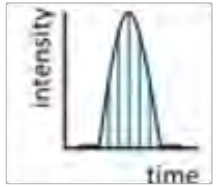
Many Opportunities for Data Science

Experimental Design → Sample Collection → Instrument Runs → Molecule Identification

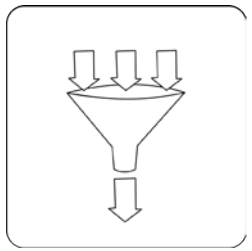


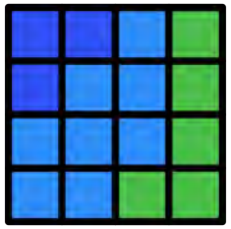


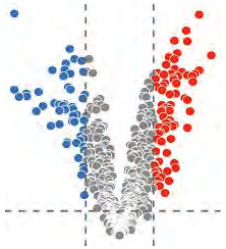




Annotation/ Assembly → Quality Control → Visualization/ Summarization → Statistical Analysis







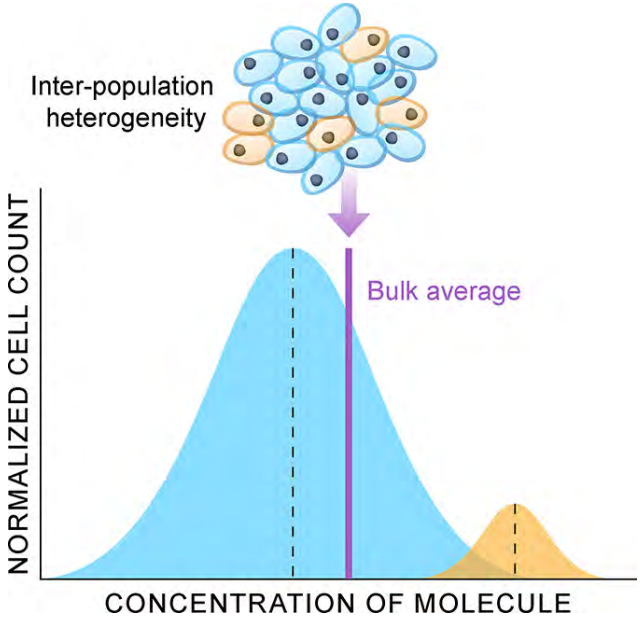
MS-based omics analyses

Single-cell vs bulk

Single-cell approaches have the advantage of dissecting cellular dynamics and heterogeneity, whereas traditional bulk technologies are limited to individual/population-level investigation.

Tissue sample

Bulk analysis

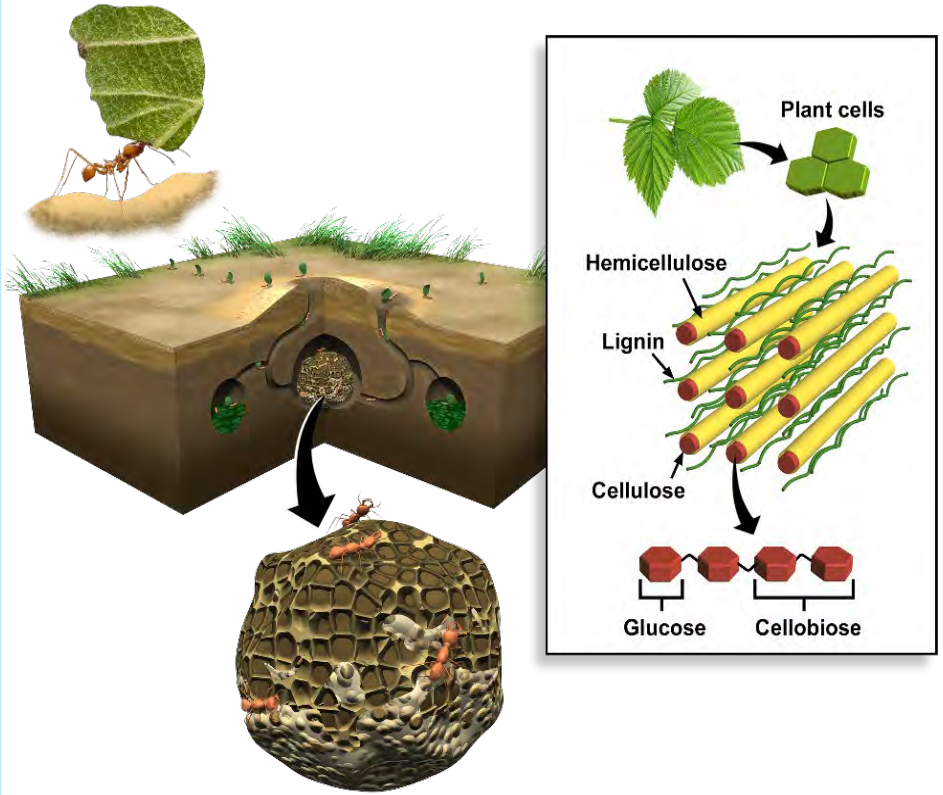


M. J. Taylor, J. K. Lukowski, and C. R. Anderton; J. Am. Soc. Mass Spectrom. 2021, 32, 4

Single cell analysis

Leaf-cutter ant fungal garden ecosystem spatial organization

Using fungal garden as an example for multi-omics analyses

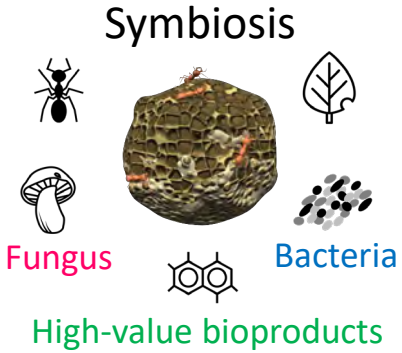


L. Khadempour et al., From Plants to Ants: Fungal Modification of Leaf Lipids for Nutrition and Communication in the Leaf-Cutter Ant Fungal Garden Ecosystem. mSystems 6, e01307-01320 (2021)


E. O. Wilson and Bert Holldobler's book The Leafcutter Ants

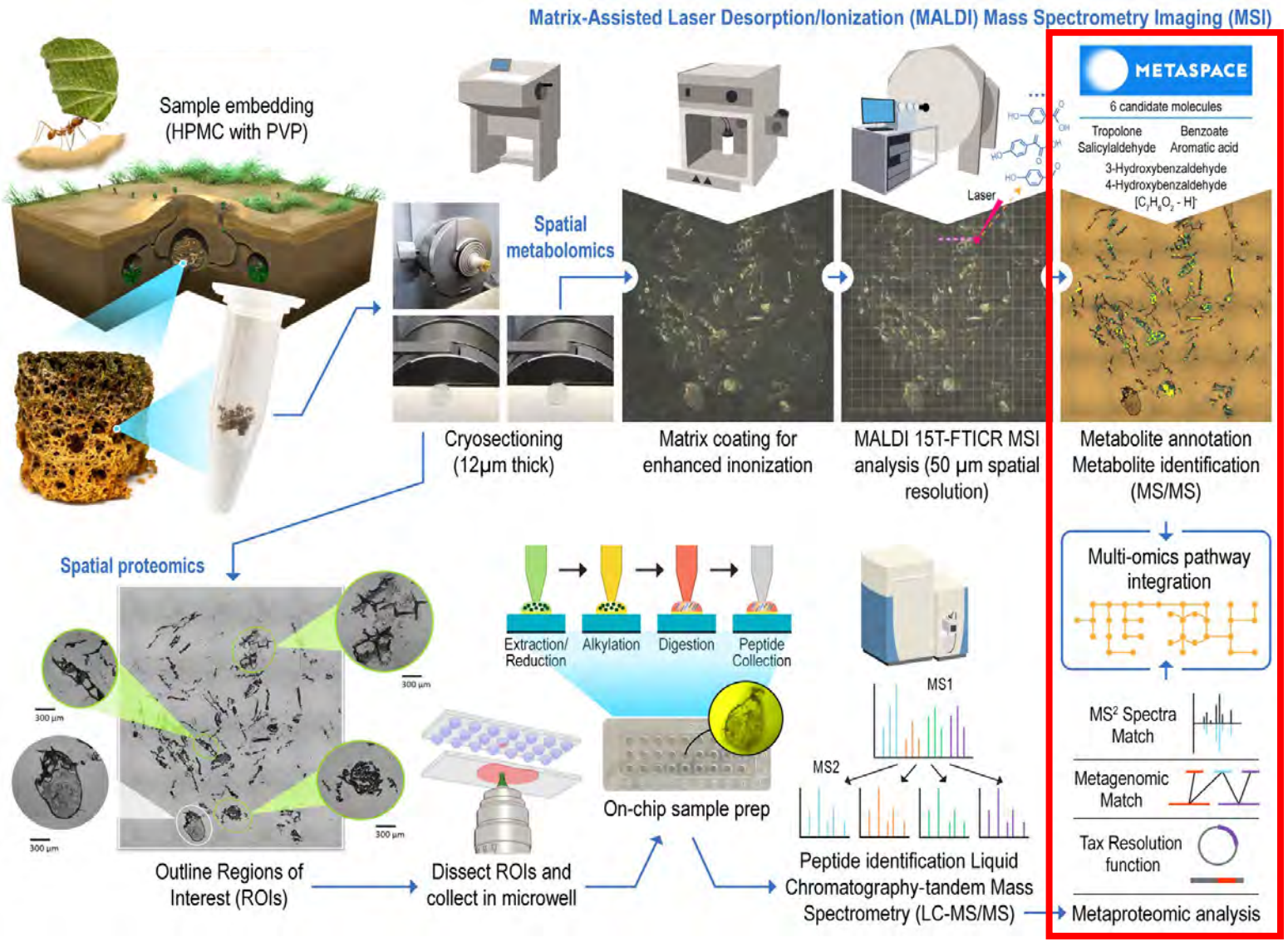

Cameron Currie Laboratory
University of Wisconsin–Madison


Cameron Currie Laboratory
University of Wisconsin–Madison



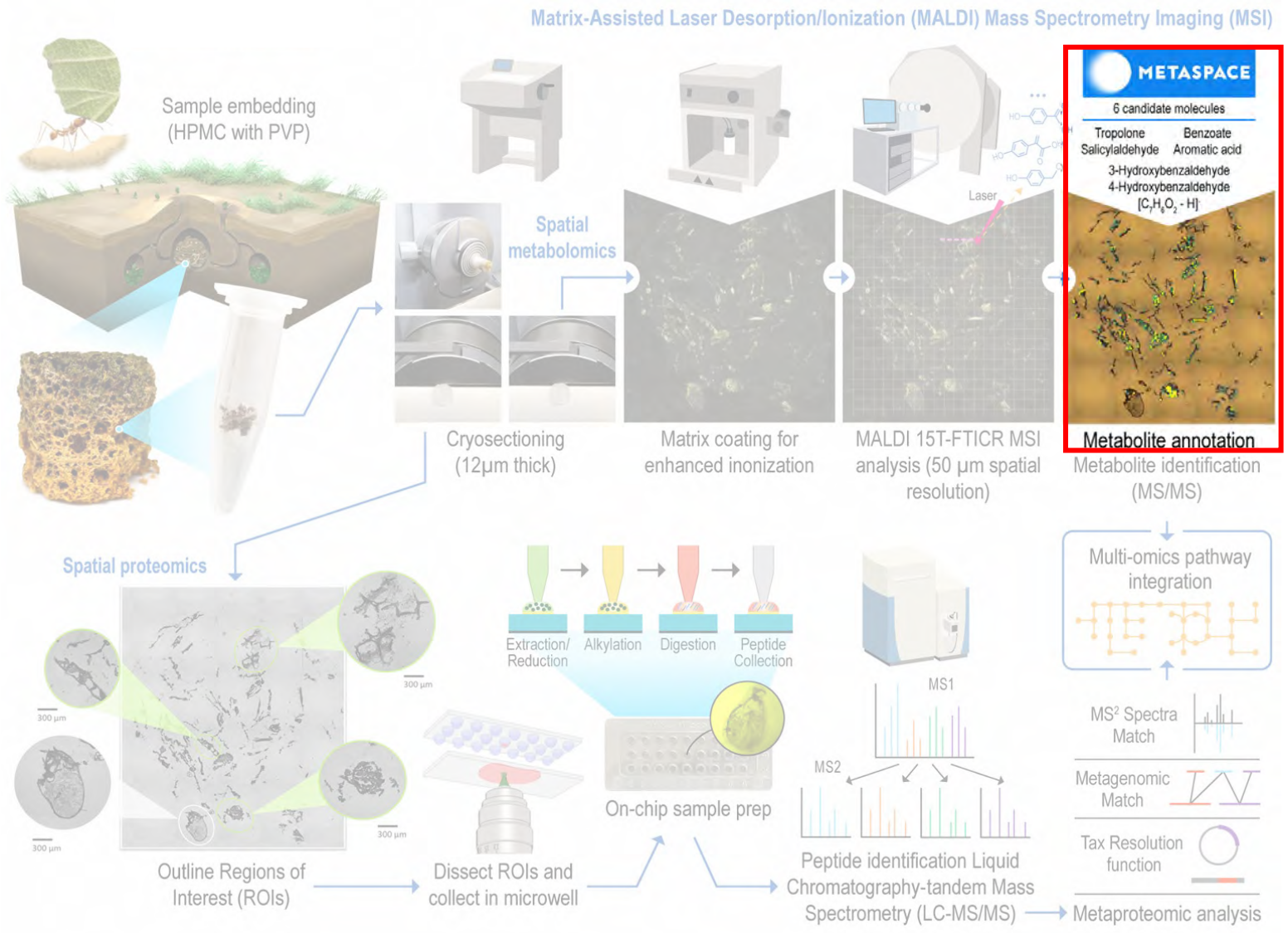
Source: Liquid Biofuels Market 2018-2025

Metabolome Informed Proteome Imaging (MIPI) Workflow



HPMC with PVP: hydroxypropyl methylcellulose with polyvinylpyrrolidone

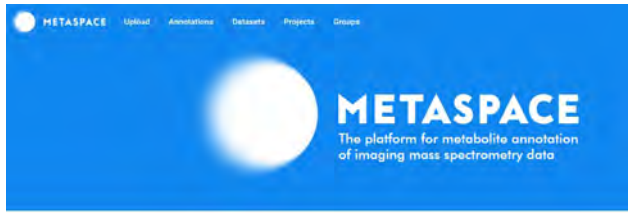
Molecular annotation of MS imaging datasets



Metabolite annotation using METASPACE

How METASPACE works?

https://metaspace2020.eu/



Metabolites, lipids

Molecular database → Theoretical MS properties

MSI dataset → Detected MS signals (images, spectra)

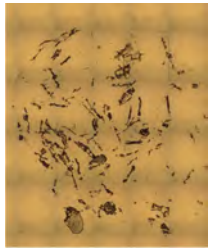
High mass resolving power (>50 k)

Quantifying:
(1) how close theoretical properties match detected ones;
(2) the quality of images

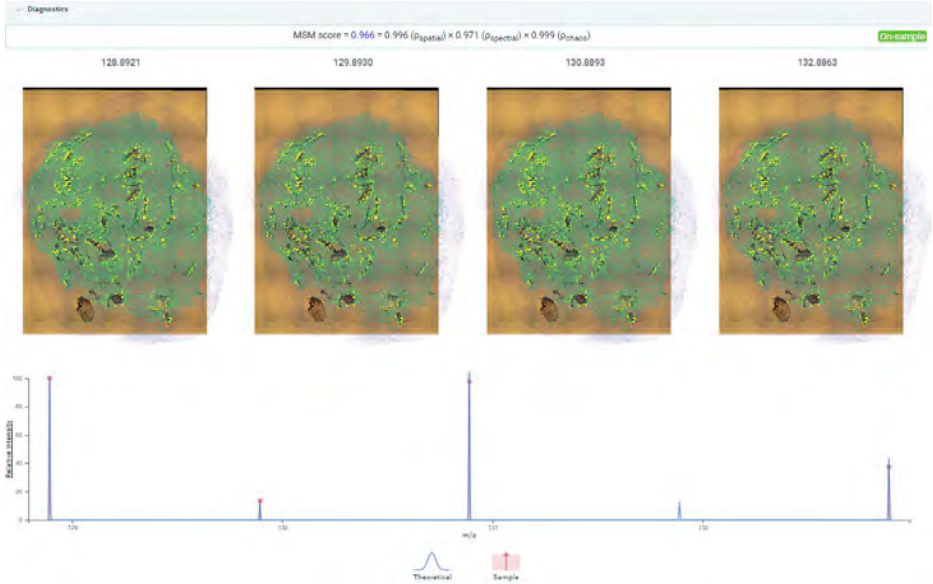
Molecule-Signal Match score for each molecule → FDR calculation (target-decoy) → Molecules with required FDR (e.g. < 10%)

Estimating the ratio of false positives in the results

Palmer et al, *Nature Methods* 2017

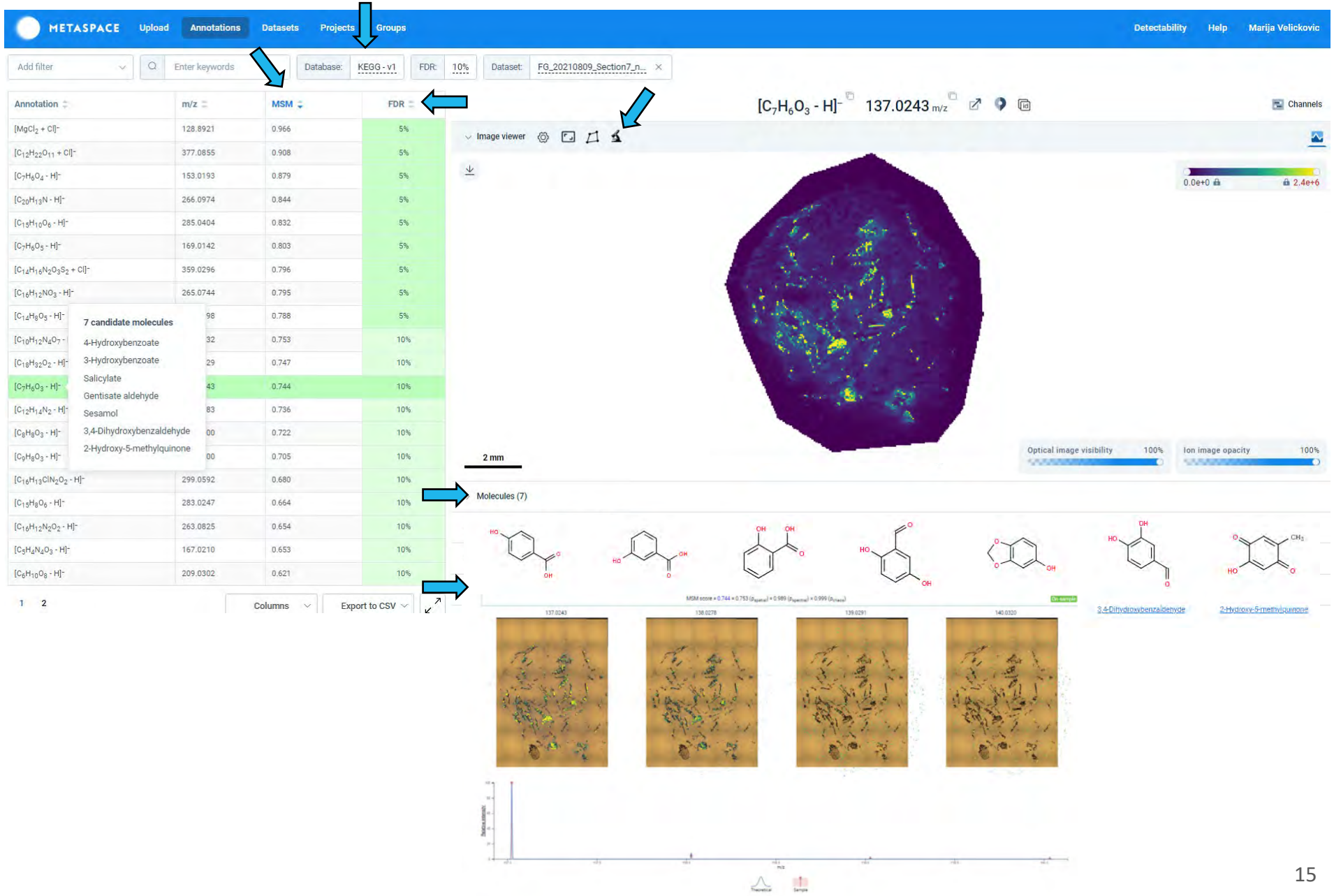


Optical image

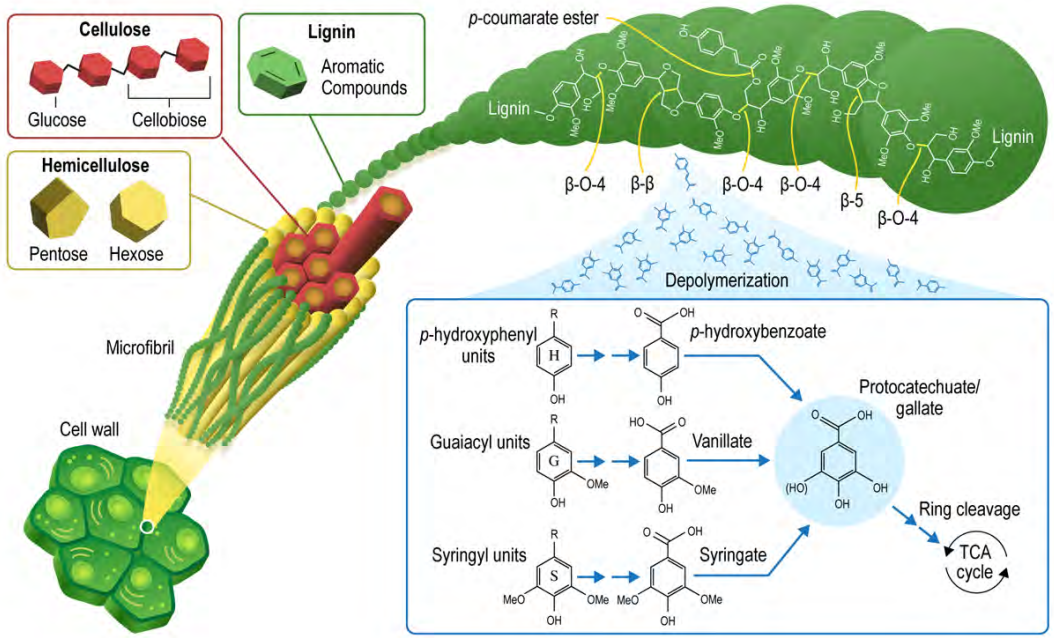


METASPACE annotation example on fungal garden sample

KEGG: Kyoto Encyclopedia of Genes and Genomes

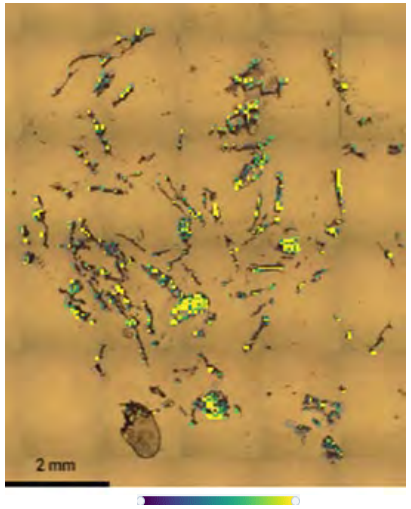


MALDI-MSI mapped the presence of low molecular weight lignin products



Hydroxybenzoate

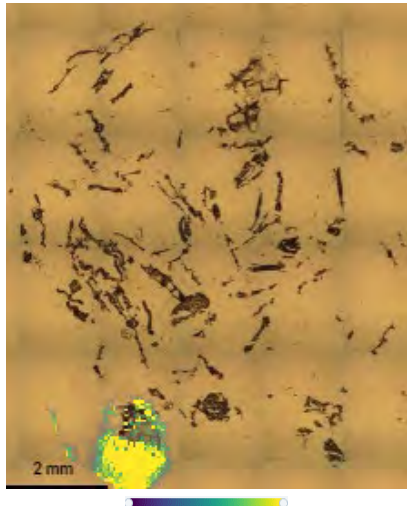
$[C7H6O3 - H]^- = 137.0243 m/z$



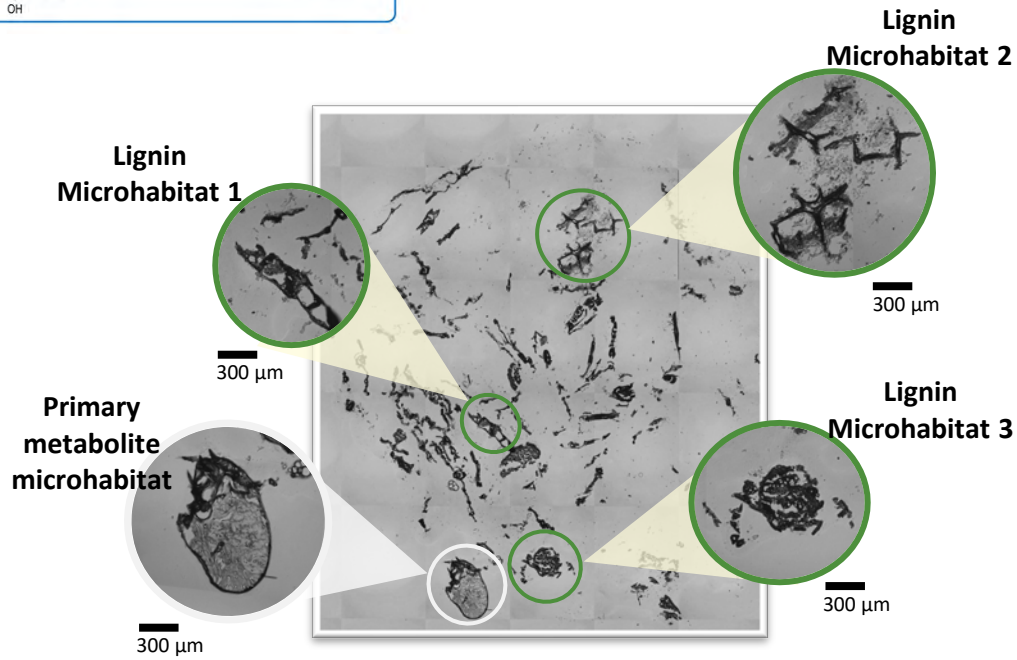
rel. signal intensity

Glutamine

$145.0618\ m/z = [C_5H_{10}N_2O_3 - H]^-$

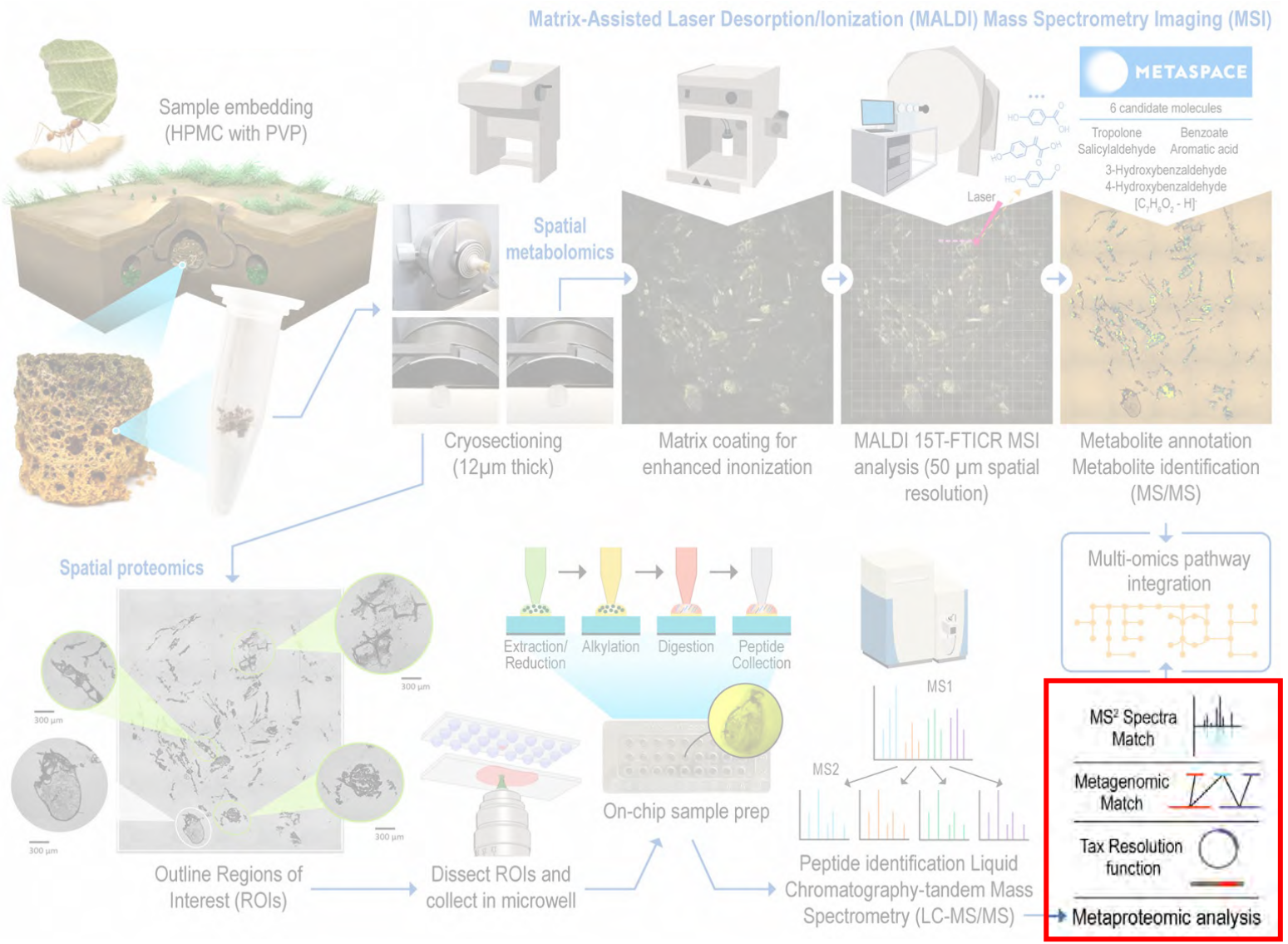


rel. signal intensity



Spatial proteomics

- Metaproteomics data analyses



Metaproteomic data analysis

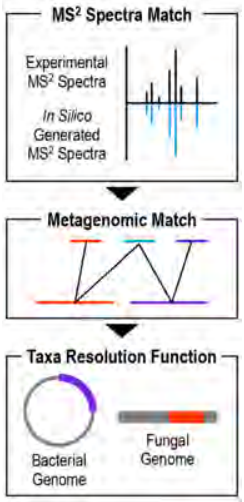
- Metaproteomics - functional annotation and taxonomic assignments

1. Reference database was curated from 50 million proteins of known members in the consortium.

2. Proteins were grouped into >24 million clusters based on sequence similarity.

3. The cluster proteins were annotated using the KEGG database via the 'Functional Annotation' and 'Taxonomic annotation' modules of the JGI metagenome workflow.
Clum, A., et al. DOE JGI Metagenome Workflow. Msystems 6(2021).

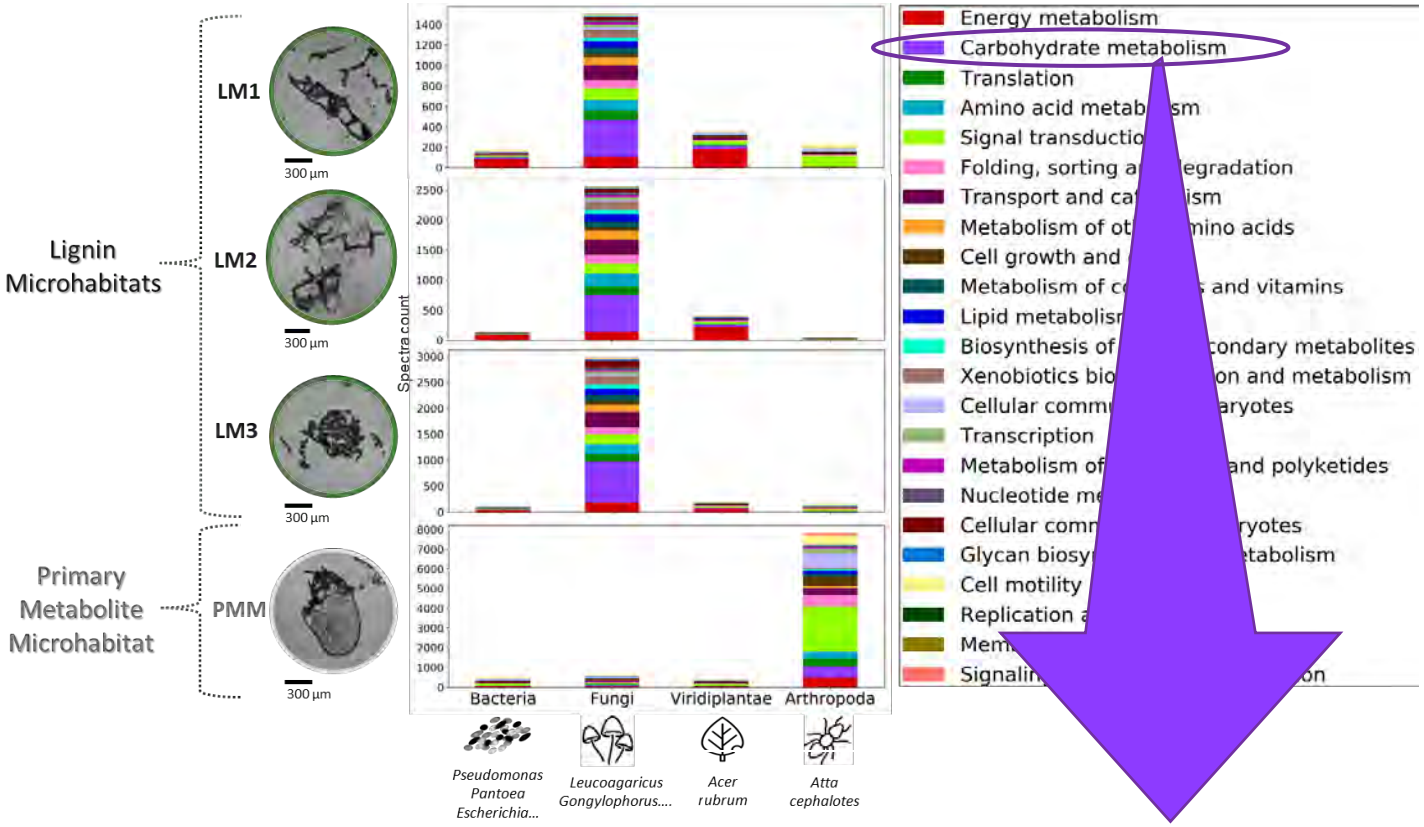
4. Peptide-spectrum matches (PSMs) using MS-GF+.

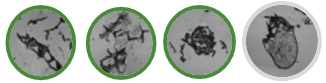


				Lignin microhabitat 1 (LM1)			Lignin microhabitat 2 (LM2)			Lignin microhabitat 3 (LM3)			Primary metabolite microhabitat (PMM)		
Protein Cluster	Peptide	Taxonomic assignment	KO number	Intensity LM1_R1	Intensity LM1_R2	Intensity LM1_R3	Intensity LM2_R1	Intensity LM2_R2	Intensity LM2_R3	Intensity LM3_R1	Intensity LM3_R2	Intensity LM3_R5	Intensity PMM_R1	Intensity PMM_R2	Intensity PMM_R3
Cluster483495.1	APSIEPGALSPDVTR	Fungi	K01576	3.24E+07		2.29E+07	3.46E+07		2.94E+07	6.51E+07	1.06E+07	1.02E+07			
	EVM*EEELNDETFR			1.09E+07			2.38E+07	4.26E+07		3.58E+07		2.25E+06			
	GLQFATSEPK			2.10E+07			2.52E+07	3.78E+07	2.84E+07						
	IATALLTAQYPLIITSR			1.20E+07		1.25E+07	2.18E+07	4.52E+07	2.85E+07	3.40E+07	6.60E+06	1.24E+07			
	REVM*EEELNDETFR			1.04E+07				2.76E+07	2.04E+07						

Contact:
Yuqian Gao, *Yuqian.Gao@pnnl.gov*
Ruonan Wu, *ruonan.wu@pnnl.gov*

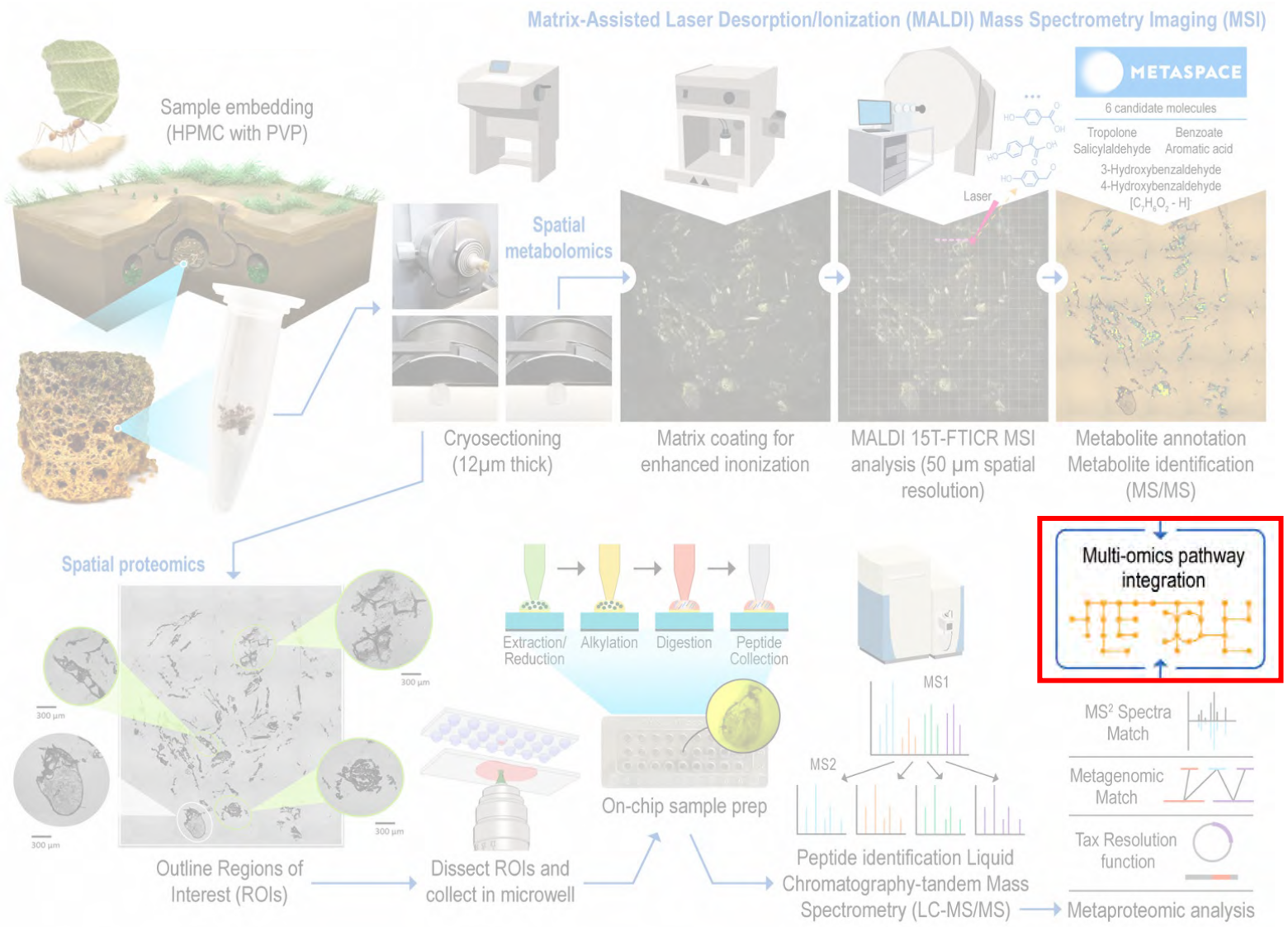
Metaproteomics data unveiled a complex community





Substrate specificity	KO number	EC number	Enzyme	Taxonomic annotation(s)	LM1	LM2	LM3	PMM
Lignin	K20929	1.2.3.15	Glyoxal/methylglyoxal oxidase	Fungi	O	O	O	X
Hemicellulose	K01209	3.2.1.55	α-L-arabinofuranosidase	Fungi/Leucoagaricus	O	O	O	O
Hemicellulose	K15920	3.2.1.37	Exo-1,4-β-xylosidase	Fungi/Leucoagaricus	O	O	O	X
Hemicellulose/ Cellulose	K05349	3.2.1.21	β-glucosidase	Fungi/Leucoagaricus	O	O	O	X
Starch	K01178	3.2.1.3	Glucoamylase	Fungi/Leucoagaricus	O	O	O	X
Pectin	K01051	3.1.1.11	Pectinesterase	Fungi/Leucoagaricus	O	O	O	X
Pectin	K15530	3.1.1.86	Rhamnogalacturonan acetylesterase	Fungi/Leucoagaricus	O	O	O	X
Pectin	K18106	1.1.1.-	D-galacturonate reductase	Fungi/Leucoagaricus	O	O	O	X
Cellulose	K01179	3.2.1.4	Endoglucanase	Fungi/Leucoagaricus	O	O	O	X
Cellulose	K01225	3.2.1.91	Cellulose 1,4-β-cellobiosidase	Fungi/Leucoagaricus	O	O	O	X

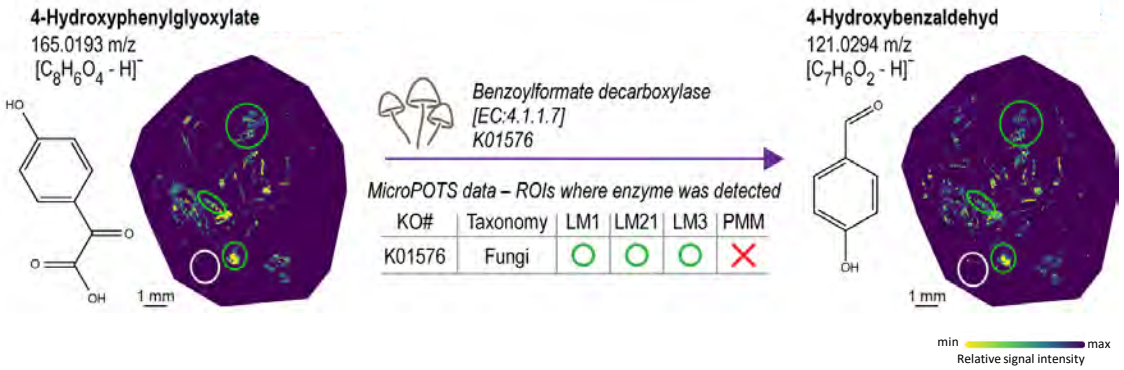
Multi-omics data integration



Multi-omics integration results

- Integration of our multimodal MS approaches provided crucial information on underlining molecular mechanisms in distinct microhabitats, providing an integrated pathway-level view of lignin degradation.
- The reconstruction of spatial microbial activities was achieved by mapping the detected metabolites and the paired enzymes to the respective KEGG metabolic pathways.

Aromatic compound degradation pathway



MicroPOTS data – ROIs where enzyme was detected

KO#	Taxonomy	LM1	LM21	LM3	PMM
K01576	Fungi	○	○	○	✕

KEGG

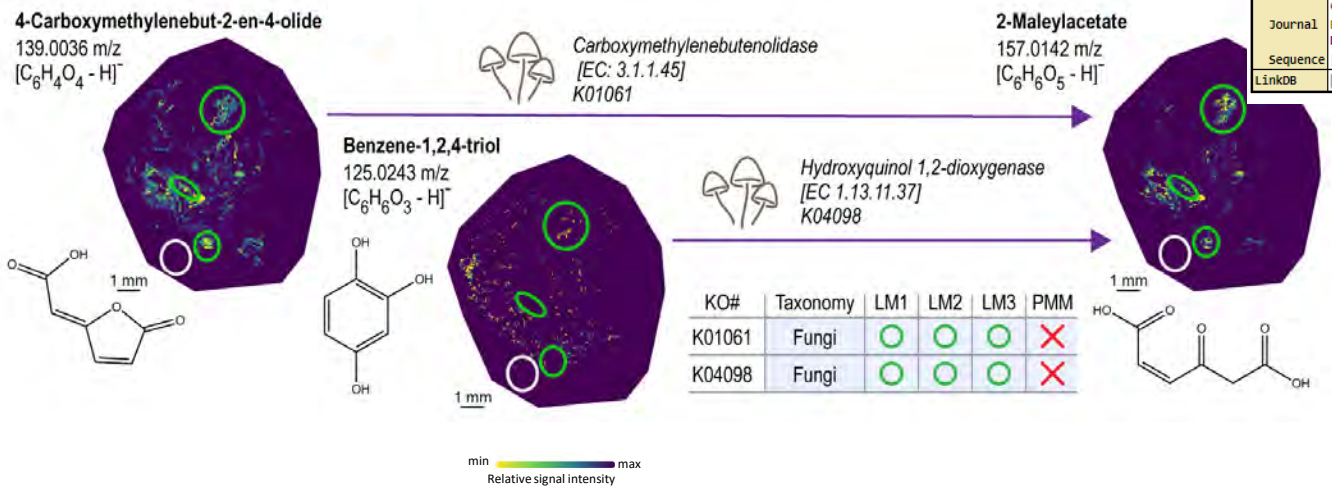
Entry
Name
Definition
Equation
Reaction class
Enzyme
Pathway
Orthology
LinkDB

KEGG ORTHOLOGY: K01576

Help

Entry	K01576 KO
Symbol	mdlC
Name	benzoylformate decarboxylase [EC:4.1.1.7]
Pathway	map00627 Aminobenzoate degradation map01120 Microbial metabolism in diverse environments
Brite	KEGG Orthology (KO) [BR:ko00001] 09100 Metabolism 09111 Xenobiotics biodegradation and metabolism 00627 Aminobenzoate degradation K01576 mdlC; benzoylformate decarboxylase Enzymes [BR:ko01000] 4. Lyases 4.1 Carbon-carbon lyases 4.1.1 Carboxy-lyases 4.1.1.7 benzoylformate decarboxylase K01576 mdlC; benzoylformate decarboxylase BRITE hierarchy
Other DBs	RN: R01764 R02672 COG: COG0028 GO: 0050695
Genes	ENL: A3UG_05905 ECLN: ECNIH4_16870 ECLI: ECNIH5_05690 ECLA: ECNIH3_05700 ECLC: ECR091_05680 EAU: DI57_12945 EKB: BFV64_05600 ENO: ECENHK_05830 ERN: BFV67_05375 ECLS: LI67_006645 » show all Taxonomy UniProt
Reference	PMID:2271624
Authors	Tsou AY, Ransom SC, Gerlt JA, Buechter DD, Babbitt PC, Kenyon GL
Title	Mandelate pathway of Pseudomonas putida: sequence relationships involving mandelate racemase, (S)-mandelate dehydrogenase, and benzoylformate decarboxylase and expression of benzoylformate decarboxylase in Escherichia coli.
Journal	Biochemistry 29:9856-62 (1990) DOI:10.1021/bi00494a015
Sequence	[ppun:PP4_48080]
LinkDB	All DBs

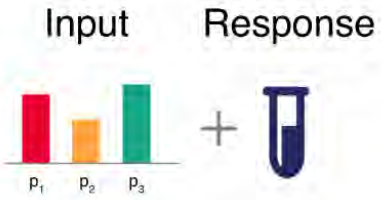
Ring cleavage pathways

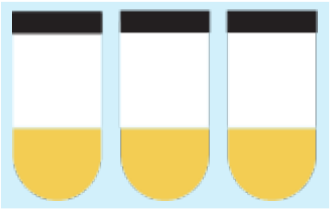


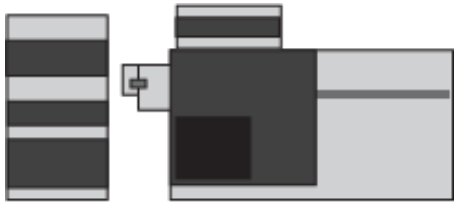
KO#	Taxonomy	LM1	LM2	LM3	PMM
K01061	Fungi	○	○	○	✕
K04098	Fungi	○	○	○	✕

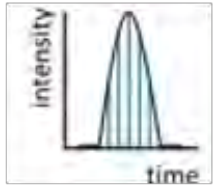
Many Opportunities for Data Science

Experimental Design → Sample Collection → Instrument Runs → Molecule Identification

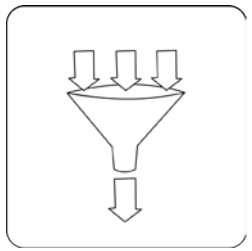


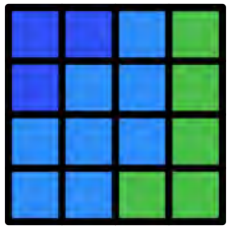


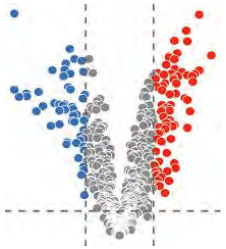




Annotation/ Assembly → Quality Control → **Visualization/ Summarization** → **Statistical Analysis**







Fungal Garden Tubes

How to model infection gradient?

Native fungal garden
(mainly *Leucoagaricus*)

Escovopsis
(Pathogenic fungus)

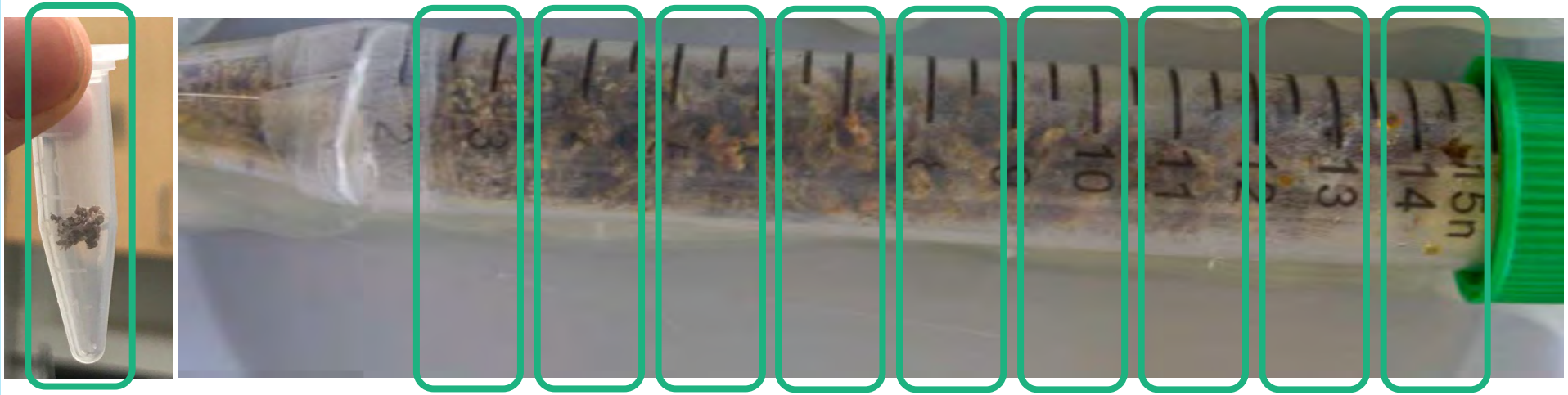
Infection spreads

Less *Escovopsis*

More *Escovopsis*

Fungal Garden Tubes

How to model infection gradient?



Native fungal garden
(mainly *Leucoagaricus*)

Escovopsis
(Pathogenic fungus)

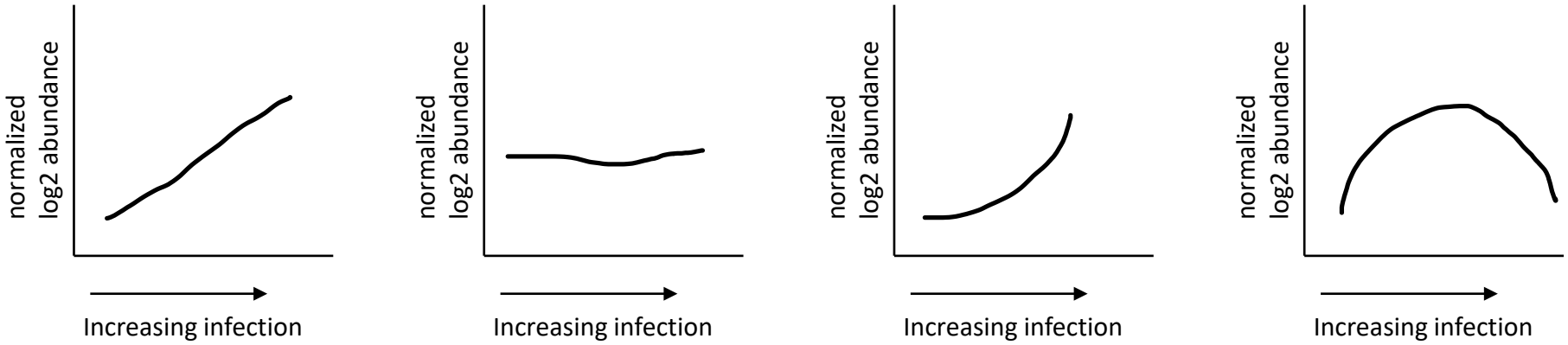
Control sample

Samples collected across the fungal garden tube

Fungal Garden Tubes

How to model infection gradient?

- Goal: Determine how individual biomolecules change across the infection gradient
- For a given biomolecule, we can imagine various patterns occurring across the infection gradient



Fungal Garden Tubes

How to model infection gradient?

- Goal: Determine how individual biomolecules change across the infection gradient

- Different patterns require different statistical models
- Could look at the 100s of biomolecules in our dataset, but difficult to parse through just by visual inspection of so many plots

→ Trelliscope

Trellis Displays

Why Trellis is Effective

» Edward Tufte's term for panels in Trellis Display is *small multiples*:

– "The same graphical design structure is repeated for each slice of a data set"

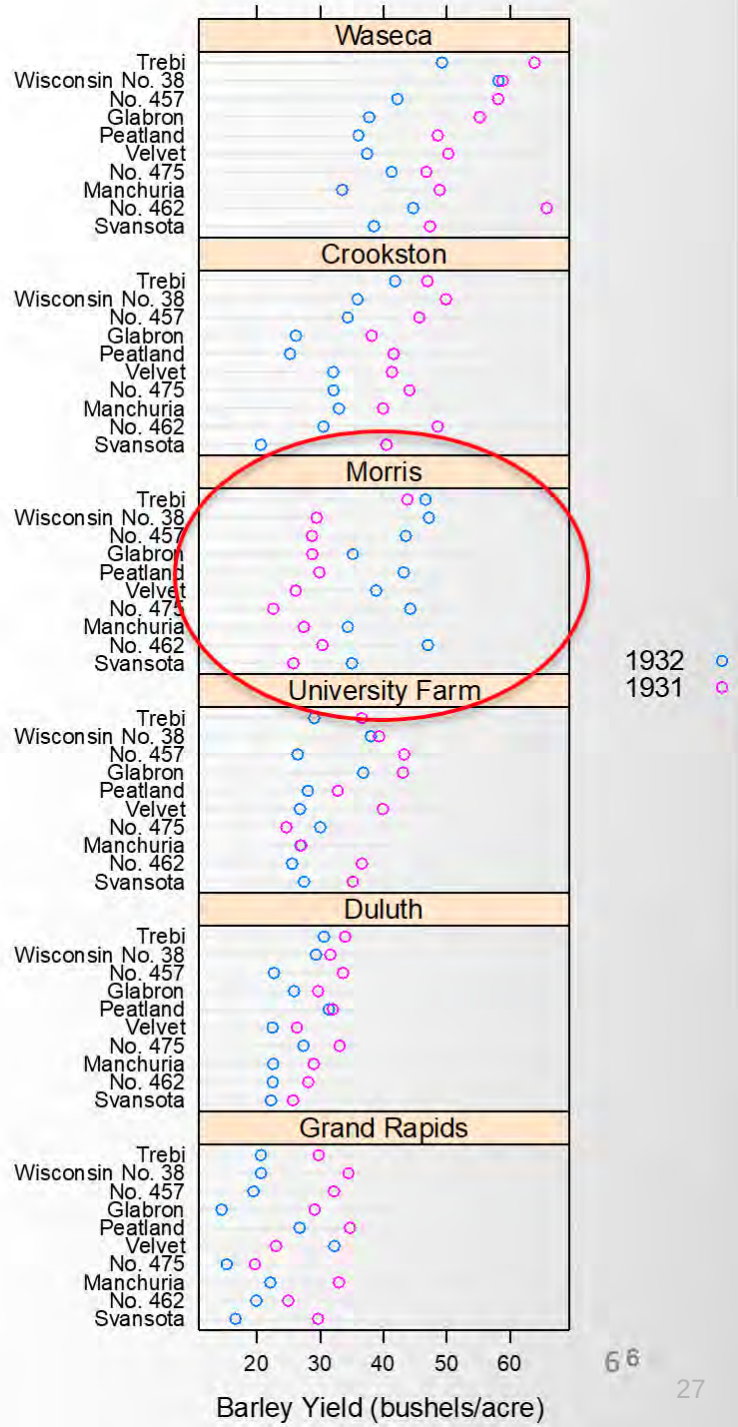
– Once a viewer understands one panel, they have immediate access to the data in all other panels

– Small multiples directly depict comparisons to reveal repetition and change, pattern and surprise

» Fisher barley data example

– Average barley yields for 10 varieties at 6 sites across 2 years

– A glaring error in the data went unnoticed for nearly 60 years



The Visual Display of Quantitative Information, Tufte
Visualizing Data, Cleveland

Trellis Displays

Trelliscope

- Look at all the things you want to see in a manageable way, real-time sorting and filtering
- Visualize and discover trends in metadata that are not immediately obvious
- Link to relevant data sources or other displays for further analysis
- Share content easily with other collaborators

Trellis Displays

Example Trelliscope Display

Trelliscope

Making a Display

- R programming
  - *trelliscopejs* package for the most flexibility (also most effort required to build a display)
    - https://cran.r-project.org/web/packages/trelliscopejs/index.html
  - *pmartR* package for simple generation of pre-defined displays
    - https://github.com/pmartR/pmartR
- User interface
  - Multiomics Data Exploration (MODE)
    - https://map.emsl.pnnl.gov/app/mode-classic
  - Multiomics Analysis Portal (MAP) version
    - https://map.emsl.pnnl.gov/app/map

Acknowledgments

Kristin Burnum-Johnson
Kristin.Burnum-Johnson@pnnl.gov

Funded by:
U.S. DEPARTMENT OF ENERGY | Office of Science
Biological and Environmental Research
Early Career Research Program

PNNL:
Ruonan Wu, Yuqian Gao, Lisa M Bramer, Dusan Velickovic, Nathalie Munoz Munoz, Rui Zhao, Carrie D Nicora, Jennifer E Kyle, Sarai Williams, Matthew E Monroe, Ronald J Moore, Bobbie-Jo M Webb-Robertson, Daniel Orton, Aivett Bilbao Pena, Priscila M Lalli, Kevin Zemaitis, Rosalie K Chu, Chaevien S Clendinen, Ying Zhu, and Paul D Piehowski.

University of Wisconsin-Madison:
Margaret Thairu, Cameron Currie

Questions?

8:30-8:35 a.m.	Introduction	Luke Durell
8:35-9:20	Data Science for ‘Omics Data	Kelly Stratton & Marija Velickovic
9:20-10:05	**Data Science Workflows**	**Eva Brayfindley**
10:05-10:15	Networking Break	
10:15-11:00	Data Science Communication	Moses Obiri
11:00-11:45	EMSL Data Science Tools	David Degnan

EMSL

Data Science Workflow

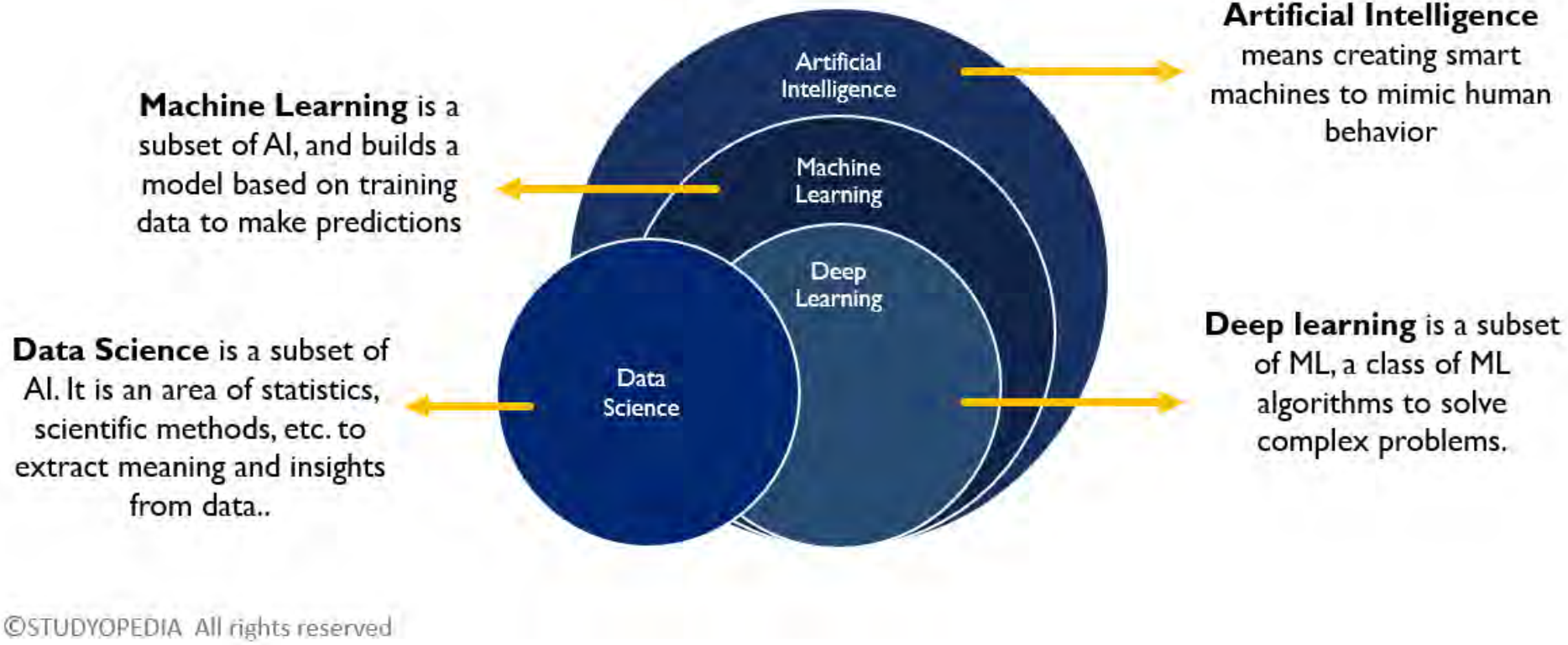
Eva Brayfindley
Senior Data Scientist
Chemical and Nuclear Defense

Presentation developed from slides created by Sam Dixon, Sam Erwin, and Alex Hagen

Overview

- Data Science, AI and ML
  - What's the difference?
  - What sorts of things can data science or ML do?
- Data science project workflow
  - Data cleaning
  - Model building
  - Model evaluation
- Important principles
  - Reproducibility
  - Ethics
- Useful tools

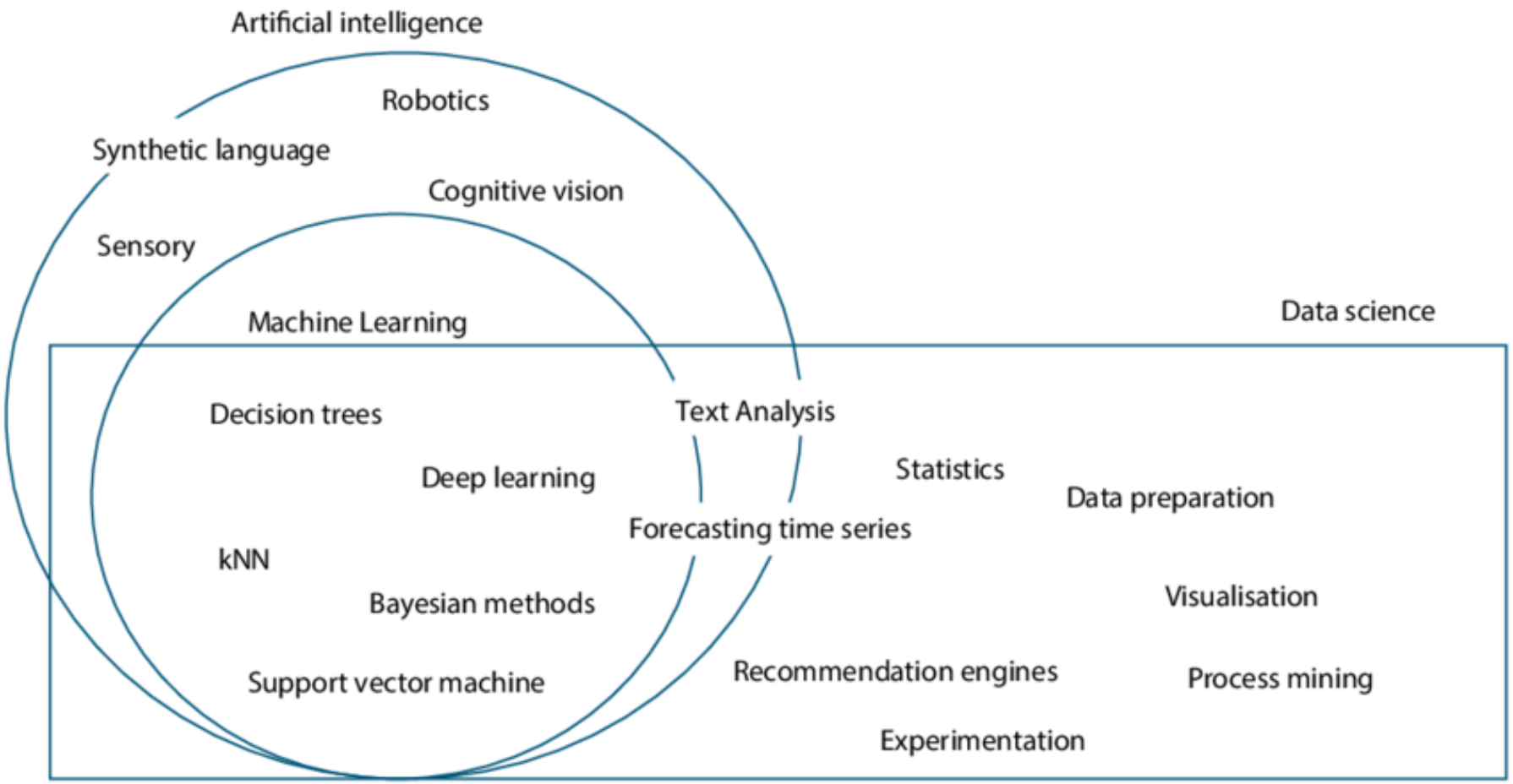
Data Science, AI and ML: What's the difference?



https://studyopedia.com/data-science/difference-datascience-machinelearing-ai-dl/

Artificial intelligence, machine learning and data science. Retrieved from "Data Science: Concepts and Practice" by Kotu, V., Deshpande, B. (2019). (2 ed.): Morgan Kaufmann. p. 3.

Data Science, AI and ML: What's the difference?

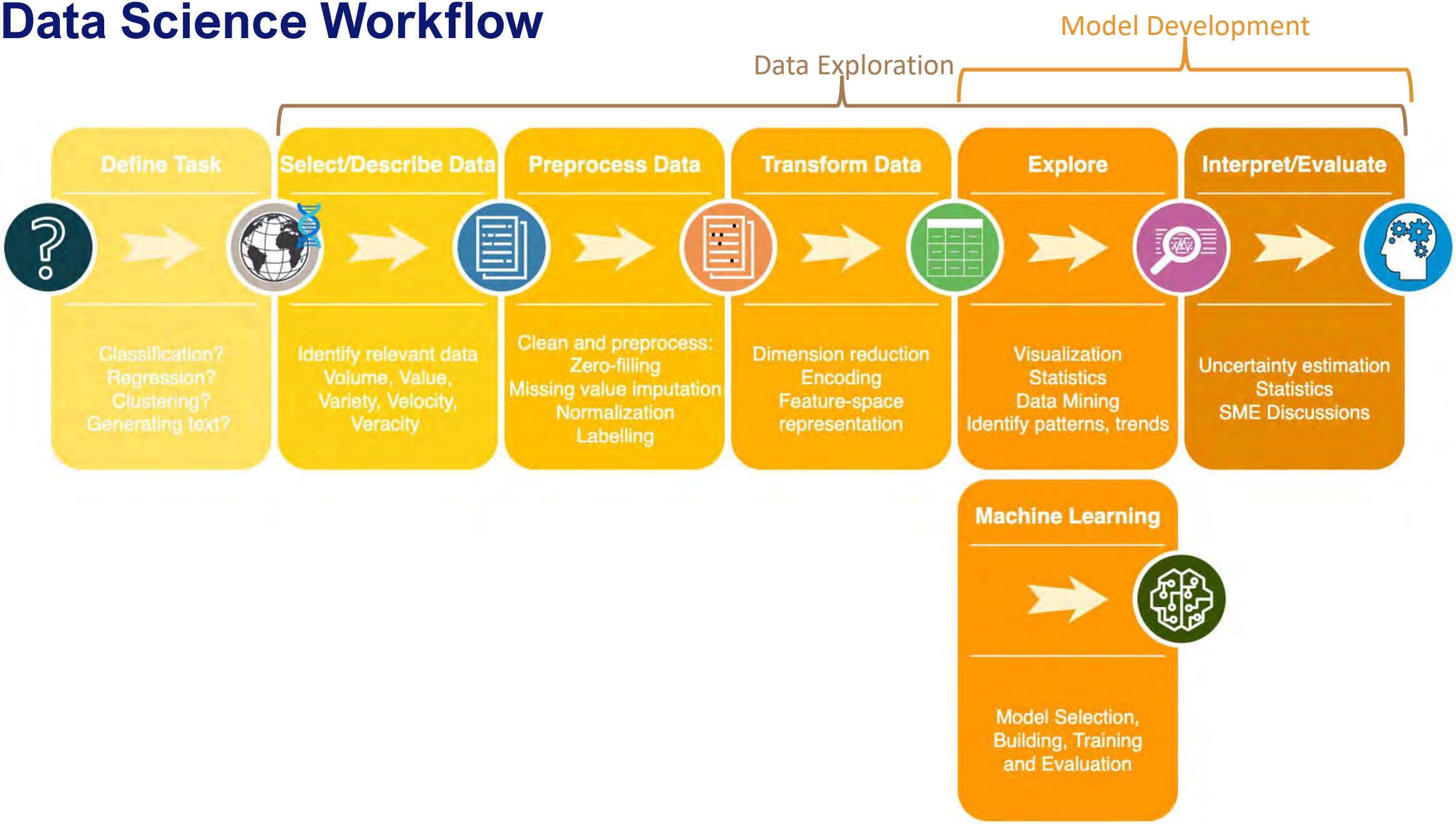


https://studyopedia.com/data-science/difference-datascience-machinelearing-ai-dl/

Artificial intelligence, machine learning and data science. Retrieved from "Data Science: Concepts and Practice" by Kotu, V., Deshpande, B. (2019). (2 ed.): Morgan Kaufmann. p. 3.

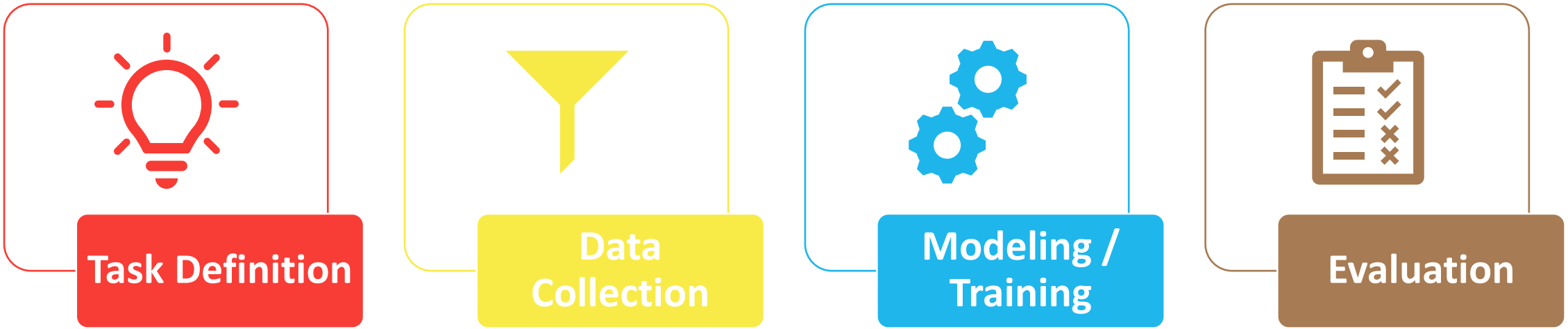
Project Workflow

Data Science Workflow

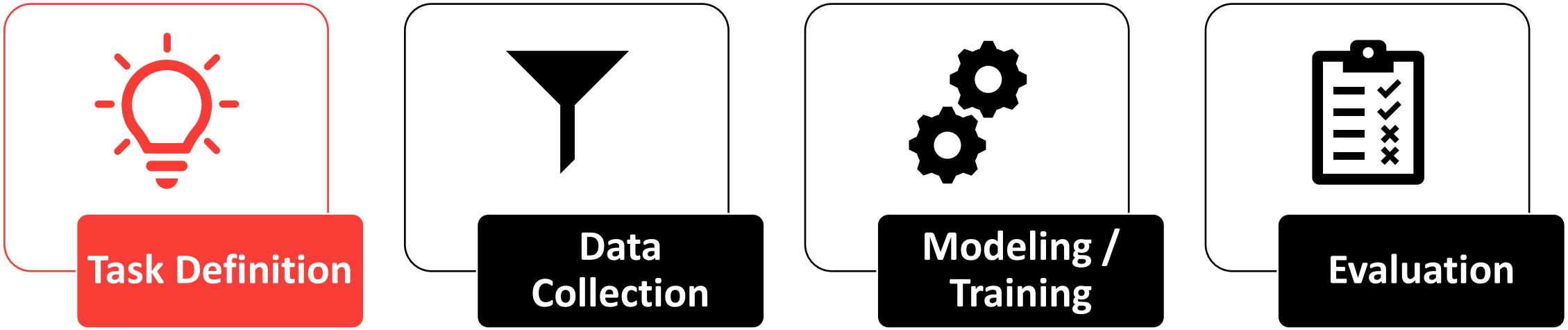


Machine Learning Model Workflow

- Each task is highly related and can often organically flow from one to another **and back again** (e.g., after you start training you may realize you lack sufficient data or need to explore other model architectures)



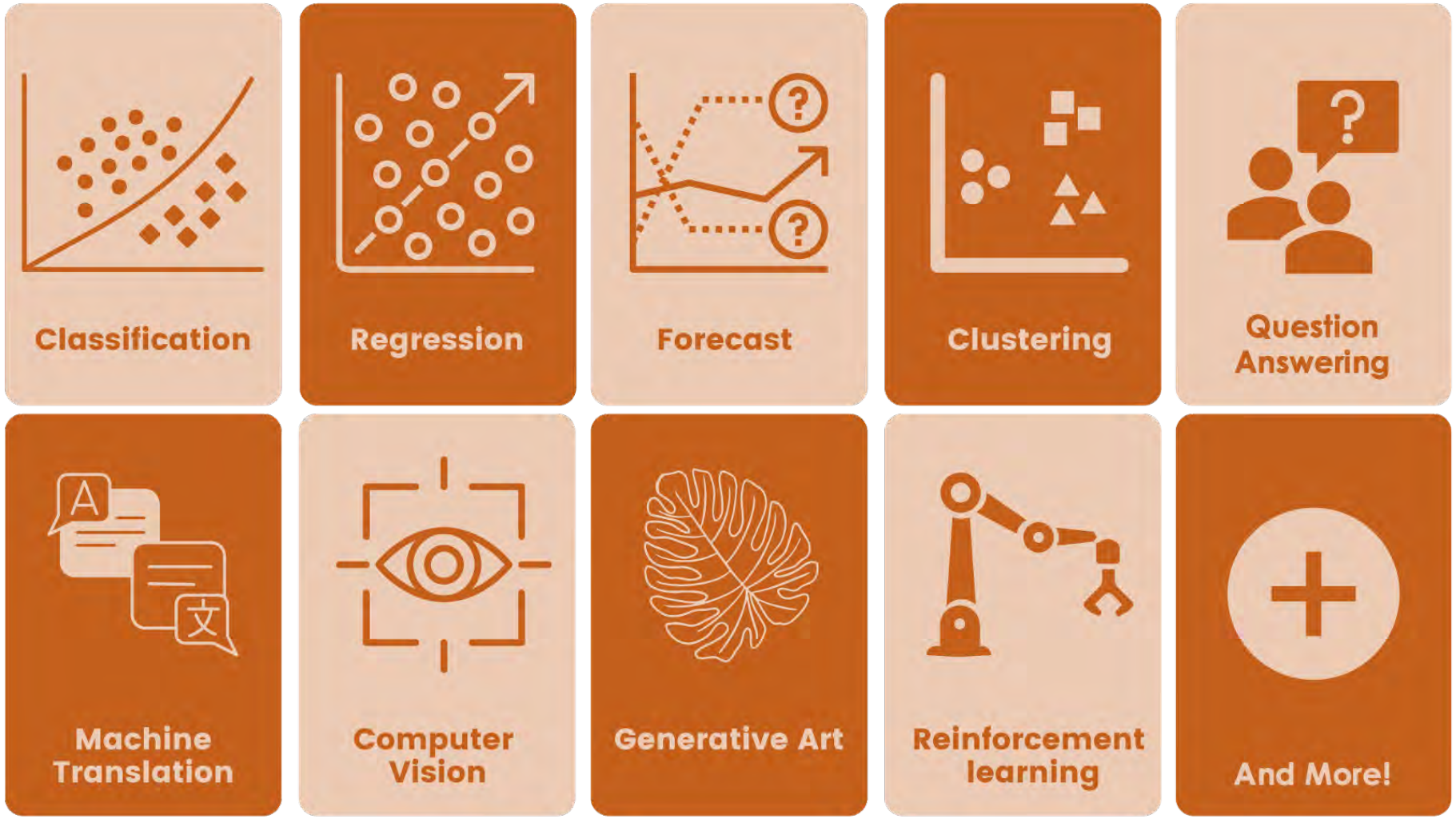
Defining your task



- It’s not the best idea to throw data at a model without an understanding of what you are trying to accomplish
- Identifying the task and starting to think about what data or modeling may be necessary is essential to a successful workflow

What can we do with data science?

What can we do with ML?



Types of Machine Learning

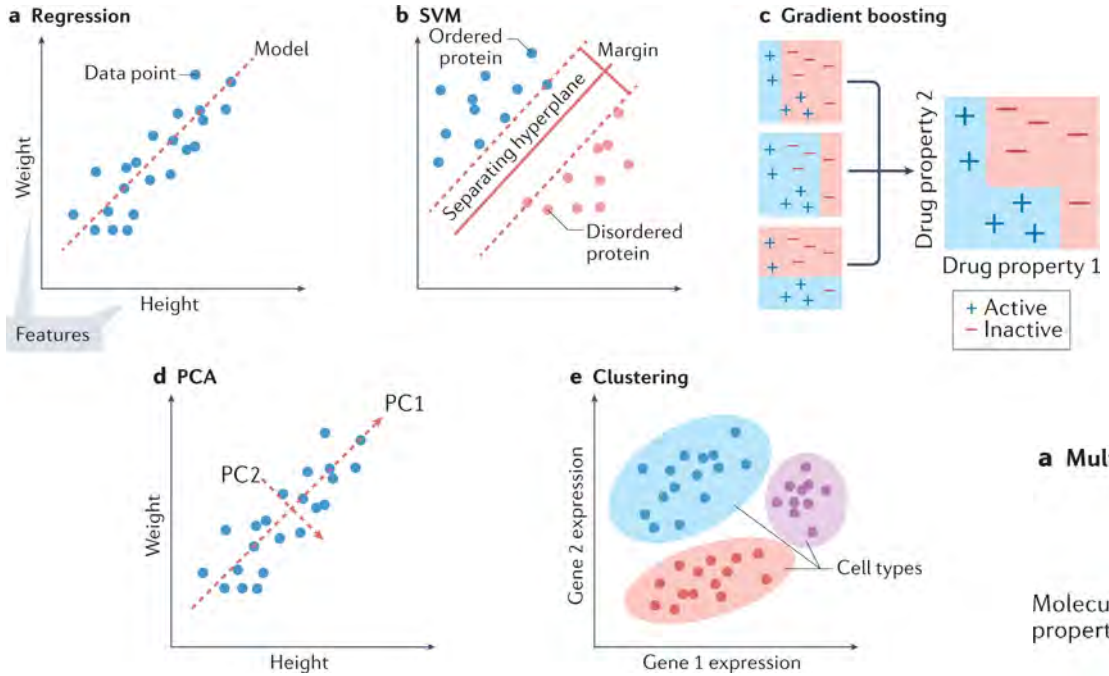
Supervised

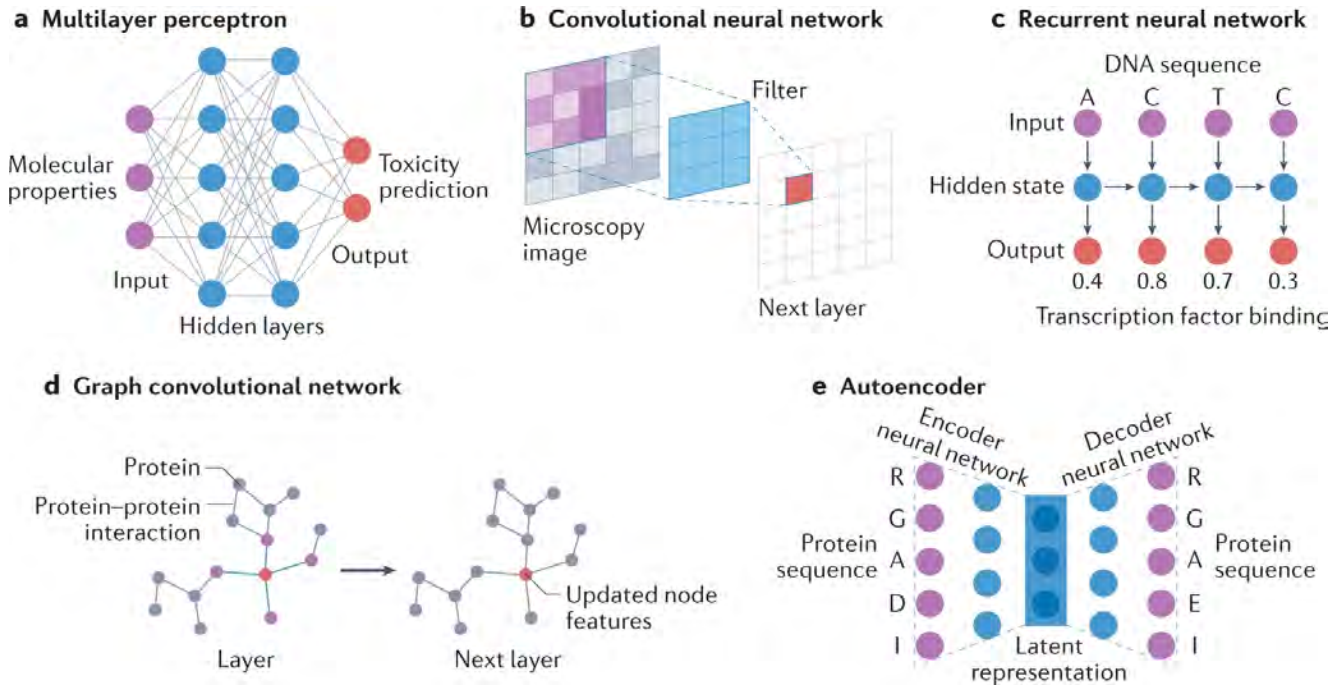
- Goal: learn to predict an outcome
- Requires labeled data
- Requires representative data in training phase
- Common tasks:
  - Classification
  - Regression

Unsupervised

- Goal: understand the data
- No labels
- Example from biology: PCA!
- Common tasks:
  - Clustering
  - Anomaly detection
  - Autoencoders (capturing regularities through data compression)

What do you want to do?

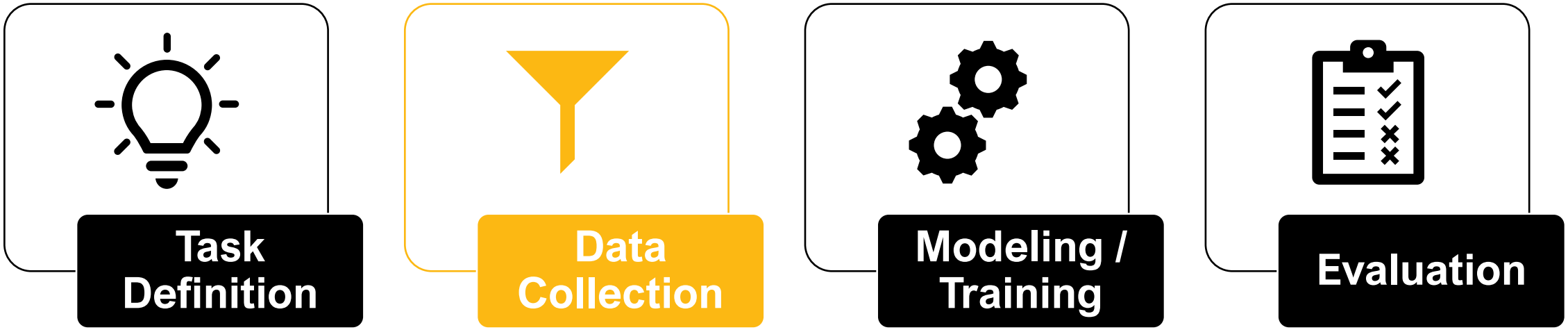




Task definition: Key Takeaways

- *Don't skip this step!!!* Having a problem to solve is **not** the same as *framing that problem for data science*
- Always talk to subject matter experts as you frame your problem
- Data may impact your task definition

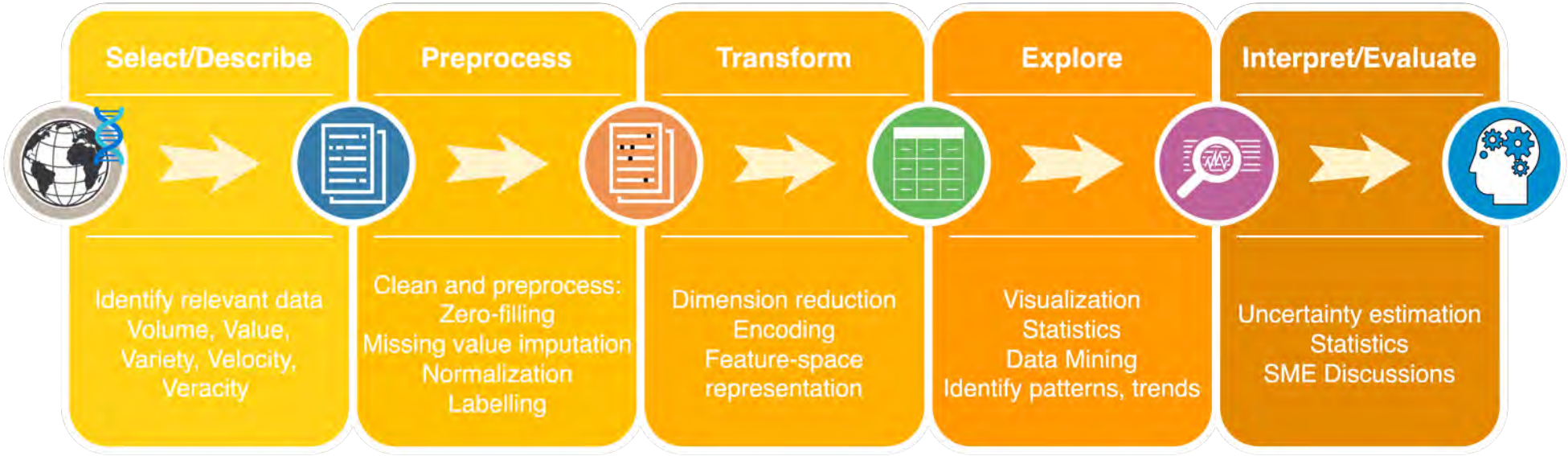
Machine Learning Model Workflow



- We are humans using human-designed tools to collect a *sample* of data
  - Like we touched on last week, this will produce an approximation of the data space
- Exploring and analyzing the data is important to determine if the data is relevant to the task and to identify/mitigate any present biases

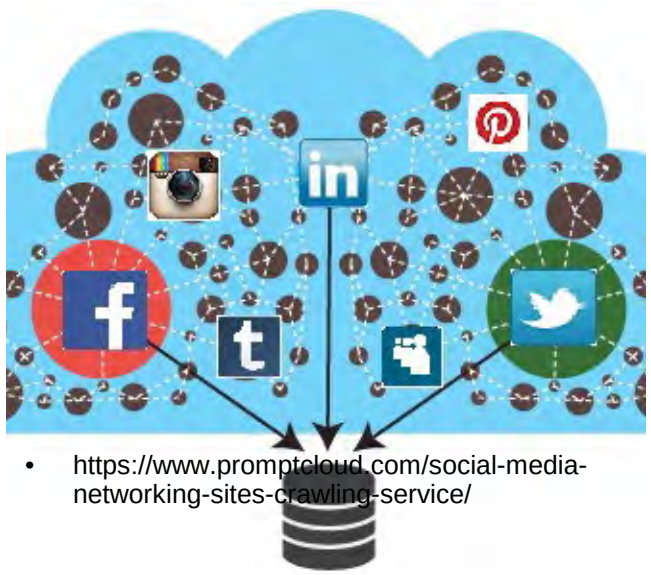
Key steps in Data Collection stage

- Identify relevant data
- Describe data
- Identify data formats and representations
- Clean and preprocess data as needed
- Explore your data to identify any initial trends that may impact model selection



What Is Data?

- Traditional data came from experimentation
- Today, everything is data

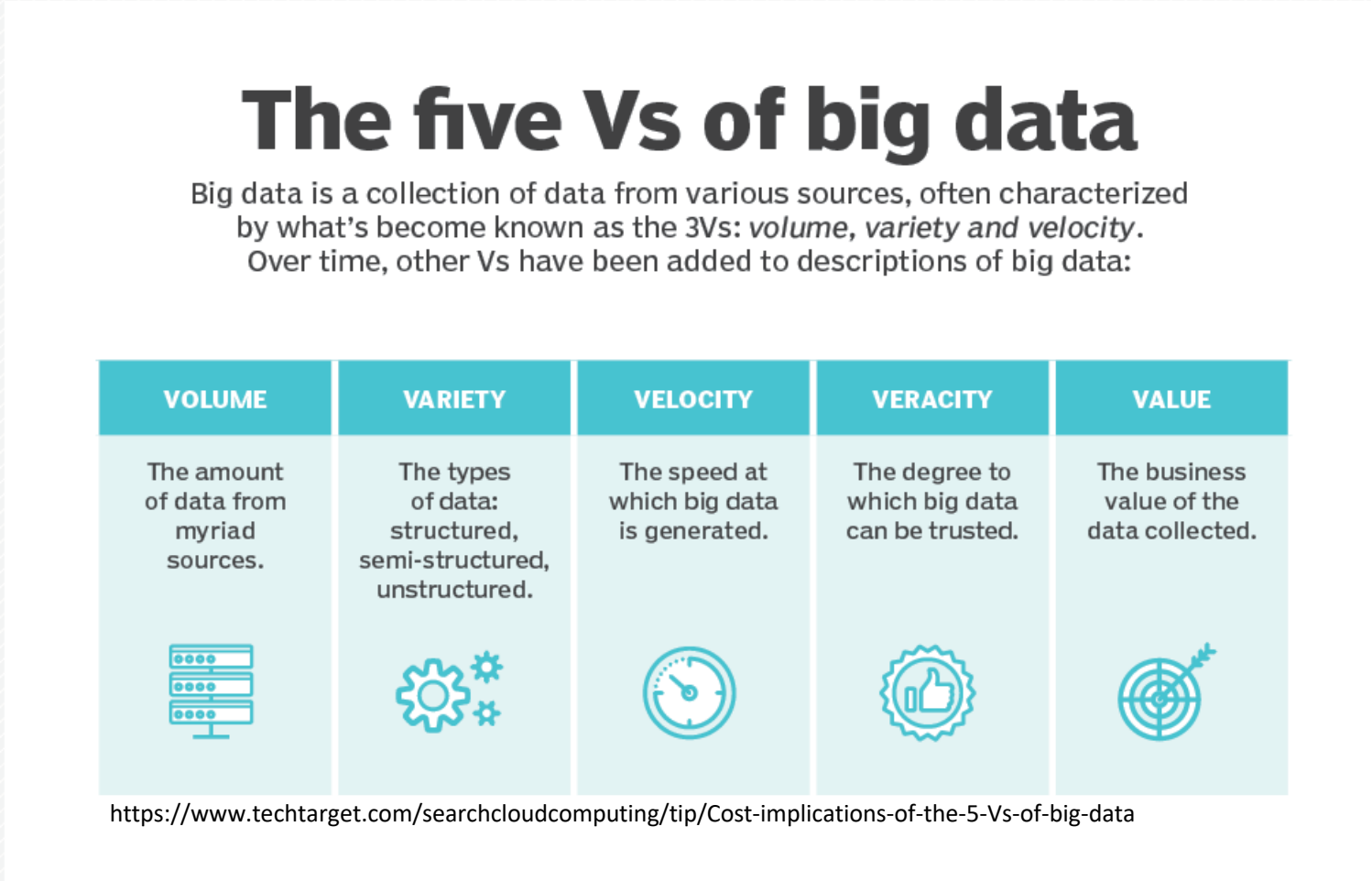


- https://www.promptcloud.com/social-media-networking-sites-crawling-service/

- https://www.robosoftin.com/blog/wearable-technology-in-fintech

Describing Data

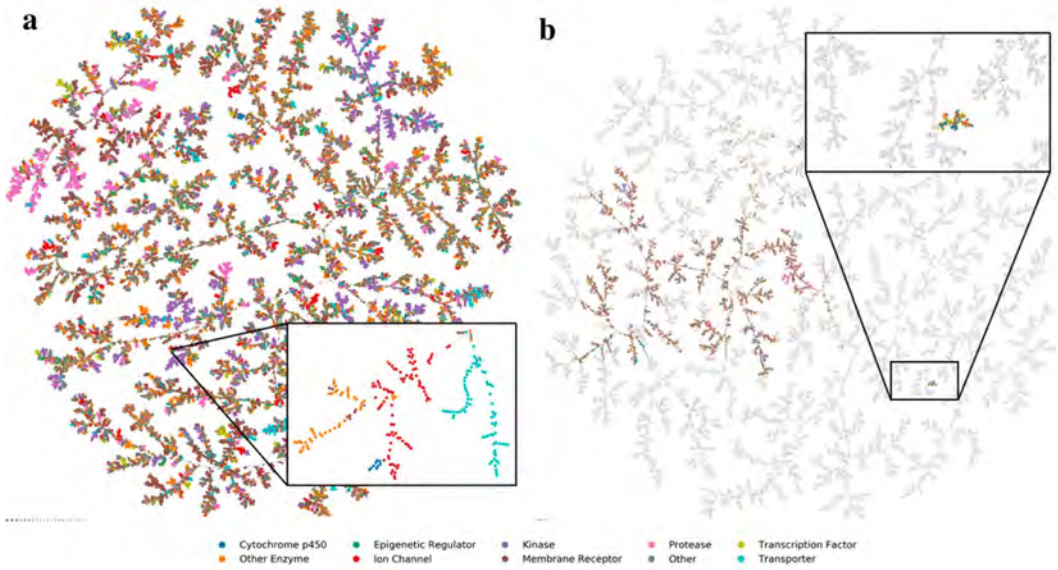
- Data needs differ by use case
- Use case can be driven by data
- How can we describe the qualities of the data we have or need?



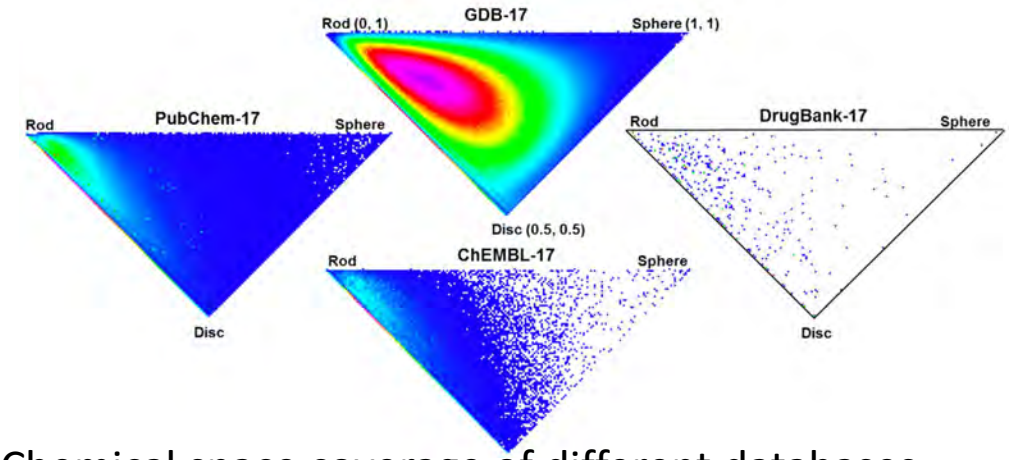
https://www.techtarget.com/searchcloudcomputing/tip/Cost-implications-of-the-5-Vs-of-big-data

Does the Data Answer My Question?

- Why was the data collected?
  - Data collected for one purpose may be incomplete for another
  - Example: data that verifies engineering specifications may not predict performance
- Does data sample the entire space of interest?
  - Sparse or uneven sampling is common
  - Example: different databases have different emphases—DrugBank focuses on drug-like compounds. Is that the space of interest?
- Does the training environment look like the deployed environment?



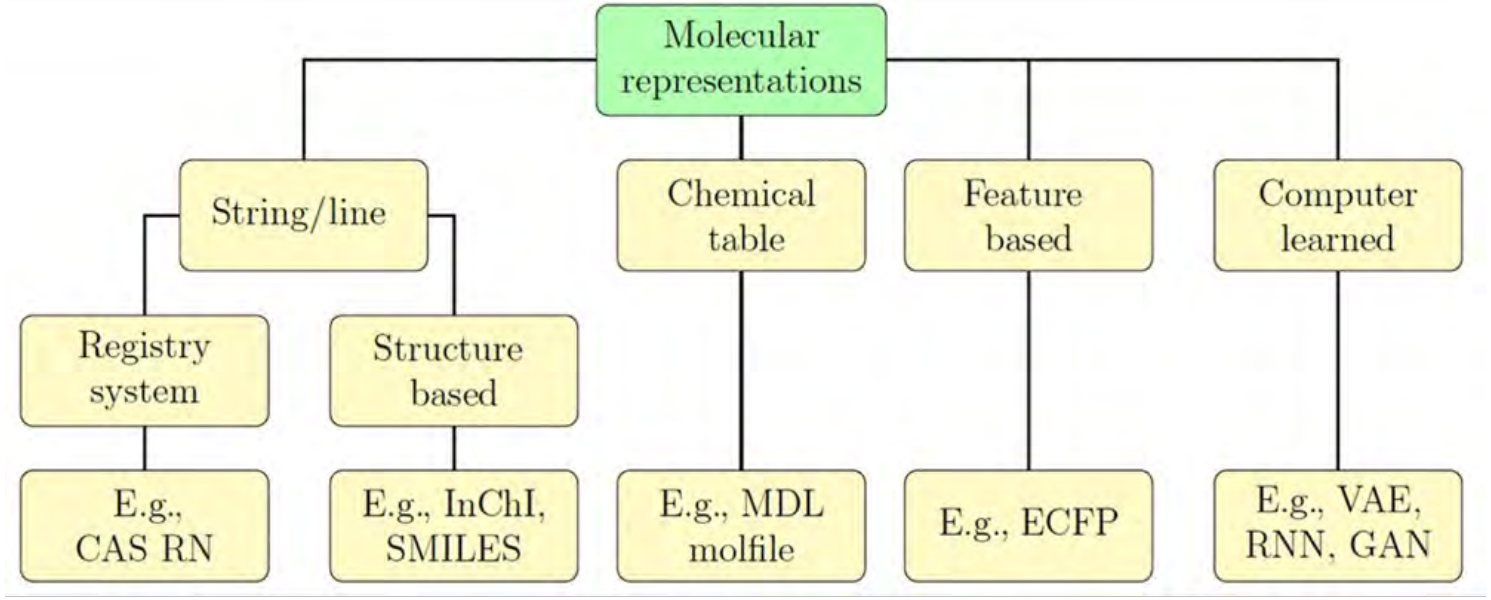
(a) ChEMBL TMAP (b) ChEMBL (color) within FDB17 (gray)



Chemical space coverage of different databases

Data Formats and Representations

- There are many possible data formats
  - Choose the best format you can
  - Each representation has different strengths
  - "Ideal" representation may not be obvious
- **Example:** Molecular structure representations
  - Each representation is a different way to represent structural information
  - The different representations may require different model types (e.g., string or image)



WIREs Comput Mol Sci, Volume: 12, Issue: 5, First published: 18 February 2022, DOI: (10.1002/wcms.1603)

Data Prep

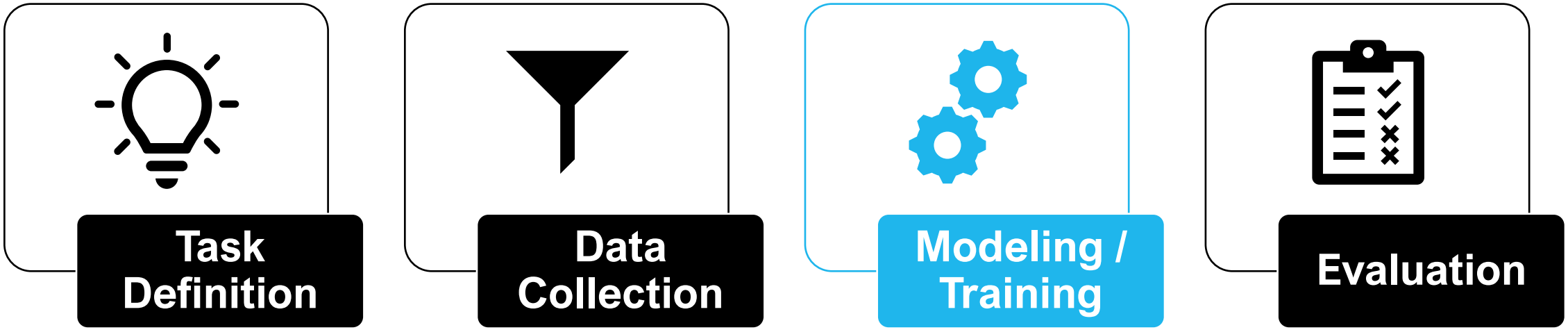
- Common standards for common data formats
- Chemical structure representations
    - Strings, graphs, feature sets as tables, etc.
- Instrument-specific data
    - Mass spectra, UV-Vis, IR, etc.
    - May need (or already include) calibration

- ***Preprocessing done by physical scientists may be different that what is needed for a data scientist or may have actually destroyed features***
    - Example: Thresholds for noise removal may be useful for an experimentalist to make conclusions, but may make ML brittle or unable to generalize
    - Example: normalization. Mathematically, scaling to mean=0 and standard deviation=1 can improve ML training… but may make no sense scientifically

Data:
Key Takeaways

- ***A model is only as good as your data!***
- Select a data format appropriate for your task
- Subject Matter Experts can provide insight into key features in data
- Knowing and addressing data properties is crucial to success

Machine Learning Model Workflow



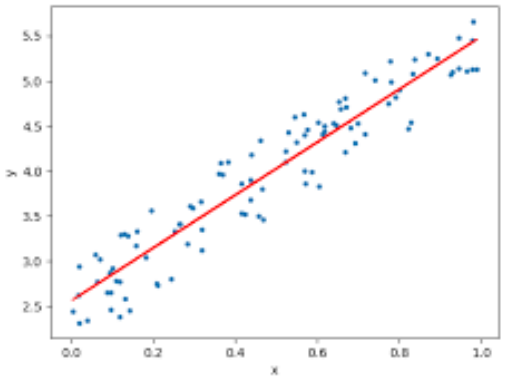
- Start simple, move to more complex architectures as needed
- Baseline models/methods exist for most problem formulations start with them before developing more complex models

Key steps in Modeling/Training stage

- Model Selection
  - Consider:
    - Input types
    - Output goals
    - Data features
    - Data quality
    - Literature norms
- Model training parameters
  - Vary with models/architectures
  - Can be fine-tuned forever—start with defaults until you have a better idea of what you need to change

How well ML can do is based on:

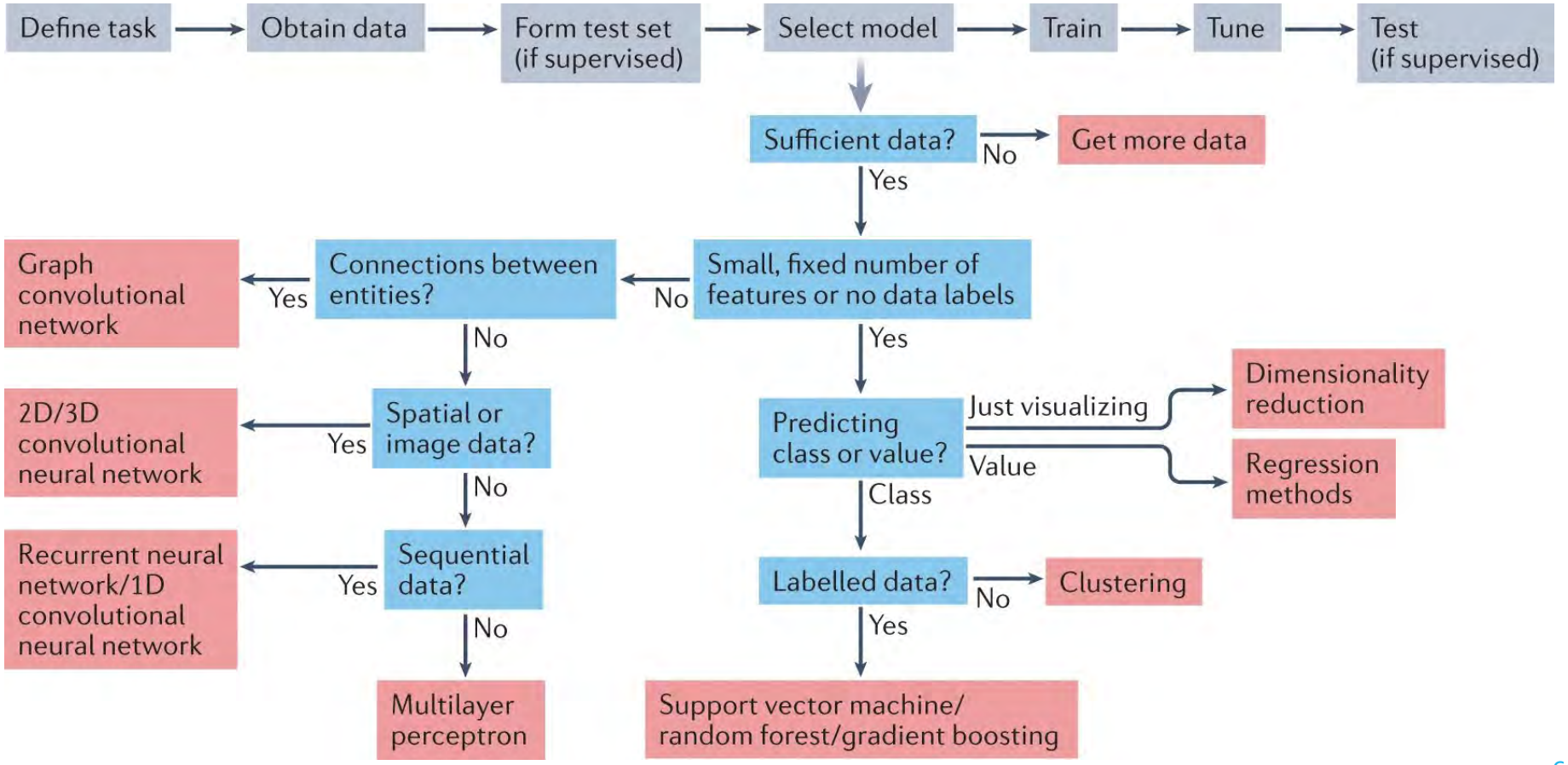
- Available data (quantity and quality)
  - Do we have enough to describe the behavior of interest?
- “Capacity” of model
  - How complex of a function can it represent?
  - How many learnable parameters it has
    - Linear Regression 2 parameters
    - GPT-4 – billions of parameters



Linear Regression

Model Selection

- Machine learning modeling is about modeling observed behavior given your data
- The model you select depends on your data type, quality, and quantity

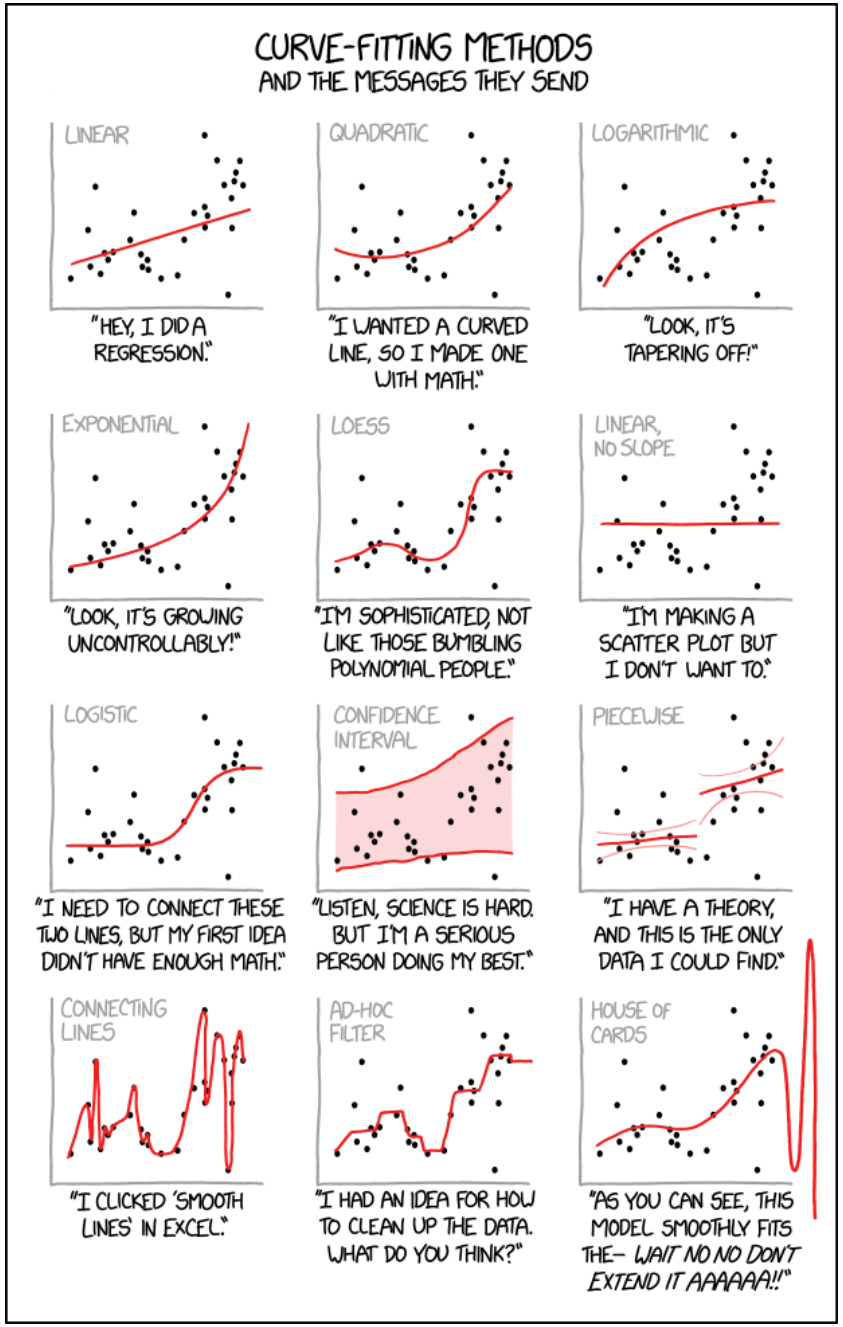


Common models for biological problems

Input data	Example prediction tasks	Recommended models	Challenges
Gene sequence	DNA accessibility[14] 3D genome organization[58] Enhancer–promoter interactions[40]	1D CNNs RNNs Transformers	Repetitive regions in genome Sparse regions of interest Very long sequences
Protein sequence	Protein structure[23,55] Protein function[132] Protein–protein interaction[133]	2D CNNs and residual networks using co-variation data Multilayer perceptrons with windowing Transformers	Metagenome data stored in many places and therefore hard to access Data leakage (from homology) can make validation difficult
Protein 3D structure	Protein model refinement[134] Protein model quality assessment[135] Change in stability upon mutation[136]	GCNs using molecular graph 3D CNNs using coordinates Traditional methods using structural features Clustering	Lack of data, particularly on protein complexes Lack of data on disordered proteins
Gene expression	Intergenic interactions or co-expression[137] Organization of transcription machinery[138]	Clustering CNNs Autoencoders	Unclear link between co-expression and function High dimensionality High noise
Mass spectrometry	Detecting peaks in spectra[139] Metabolite annotation[140]	CNNs using spectral data Traditional methods using derived features	Lack of standardized benchmarks[141] Normalization[a] required between different datasets
Images	Medical image recognition[24,62] Cryo-EM image reconstruction[60,142] RNA-sequencing profiles[143]	2D CNNs and residual networks Autoencoders Traditional methods using image features	Systematic differences in data collection affect prediction Hard to obtain large datasets of consistent data
Molecular structure	Antibiotic activity[73] Drug toxicity[54] Protein–ligand docking[39] Novel drug generation[144]	GCNs using molecular graph Traditional methods or multilayer perceptrons using molecular properties RNNs using text-based representations of molecular structure such as SMILES Autoencoders	Experimental data available for only a tiny fraction of possible small molecules
Protein–protein interaction network	Polypharmacology side effects[77] Protein function[145]	GCNs Graph embedding	Interaction networks can be incomplete Cellular location affects whether proteins interact High number of possible combinations

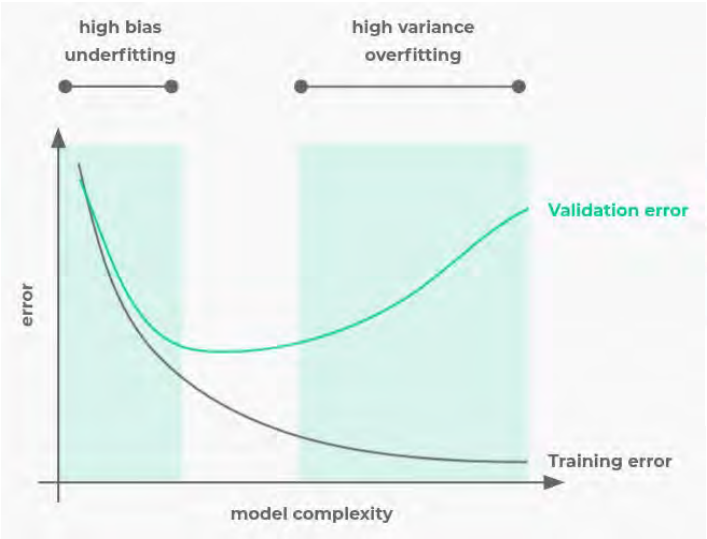
Common Pitfalls

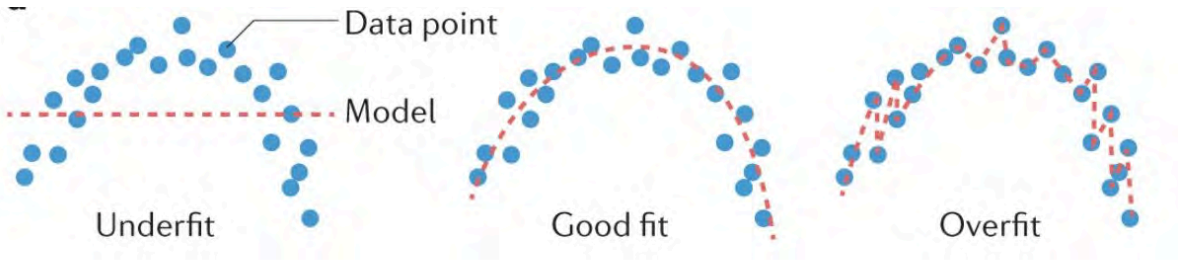
- Measuring the wrong thing
- Confusing correlation and causation
- Failing to generalize (overfitting)
- Underestimating data
- Losing sight of the objective



Underfitting and Overfitting

- Bias is the amount that a model's prediction differs from the target value, compared to the training data
- Variance indicates how much a random variable differs from its expected value
- Fitting your training data perfectly
  - Getting a perfect score is scary—your model may have memorized the training data and won't be able to generalize!





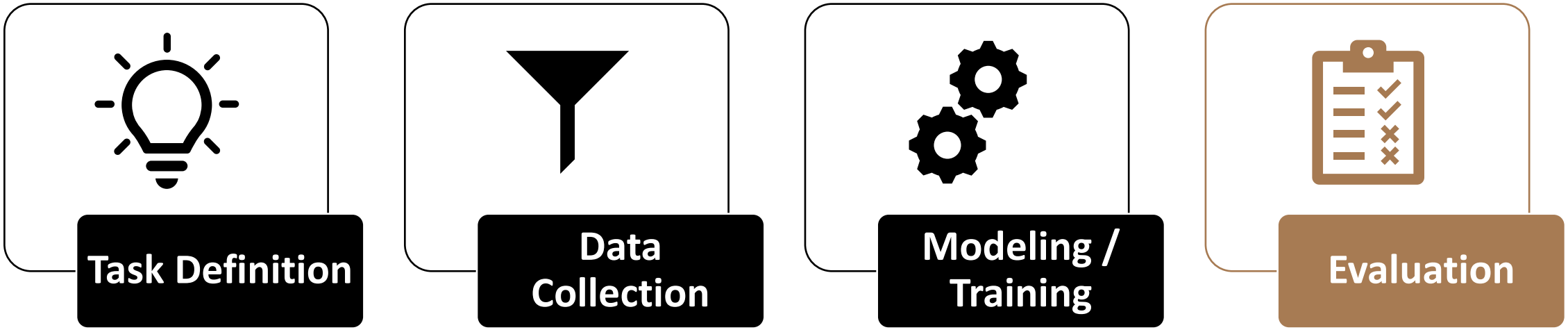
https://www.mastersindatascience.org/learning/difference-between-bias-and-variance/

Training: Key Takeaways

- Be careful when selecting your model
- Make sure you have both enough and the right type of data for the model you choose
  - Check the literature! See what other people have done
  - Talk to SMEs, other data scientists if you're not sure

- Beware underfitting and overfitting!

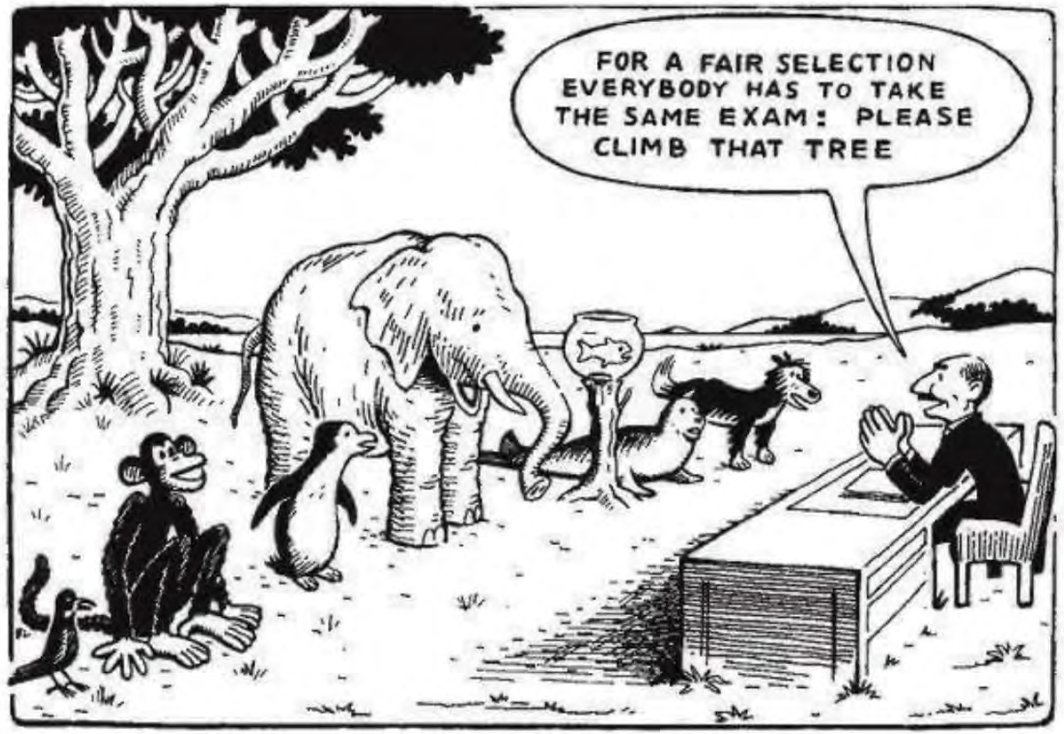
Machine Learning Model Workflow



- Identify both quantitative and qualitative results
- Ensure model output is unbiased / fair
- Reminder: may need to go back a step or two or three

Metrics

- Depend on the problem formulation
- Can be useful to the model, people, or both
- Can also be very misleading based on your data or problem formulation

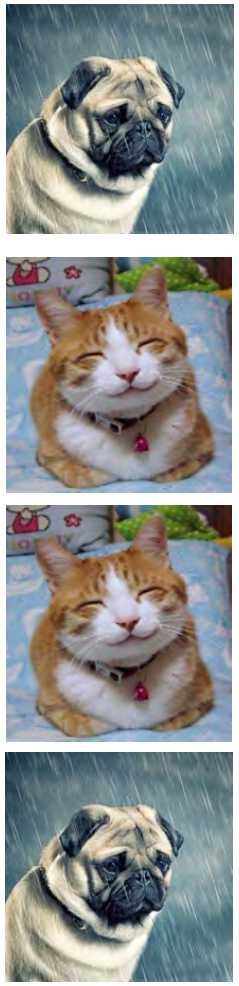


Evaluating Performance on Classification Problems

- Relevant performance metrics are dependent on the task
- Classification relies on discrete labels so we can categorize performance with:
  - True Positives, False Positives, True Negatives, and False Negatives
- A confusion matrix visualizes and summarizes how well a model classifies data.
- Goal: high True Positives and True Negatives, and low False Positives and False Negatives

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Calculating Counts



Model

Dog TP: Actual dog, Predicted dog

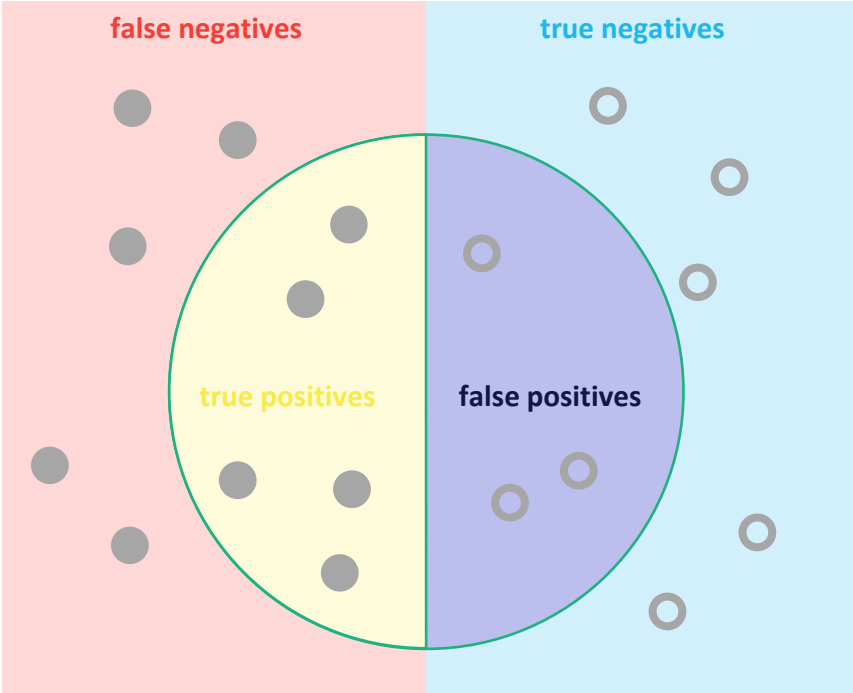
Dog FP: Not a dog, Predicted dog

Cat TN: Not a dog, Not predicted as dog

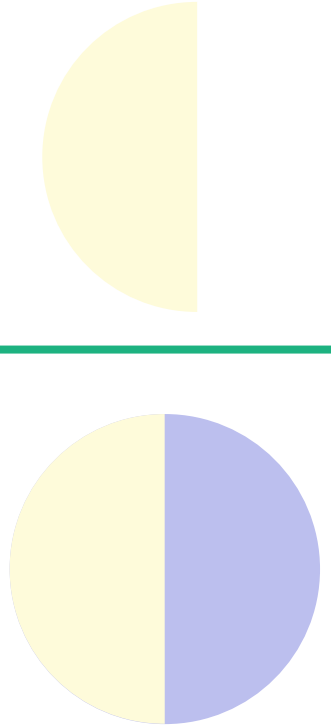
Cat FN: Actual dog, Not predicted as dog

Precision: How many selected items are relevant?

- Of all positive predictions, precision counts the percentage that is correct
  - E.g., Number of images accurately labeled as ‘dog’ relative to total number of images labeled as dog



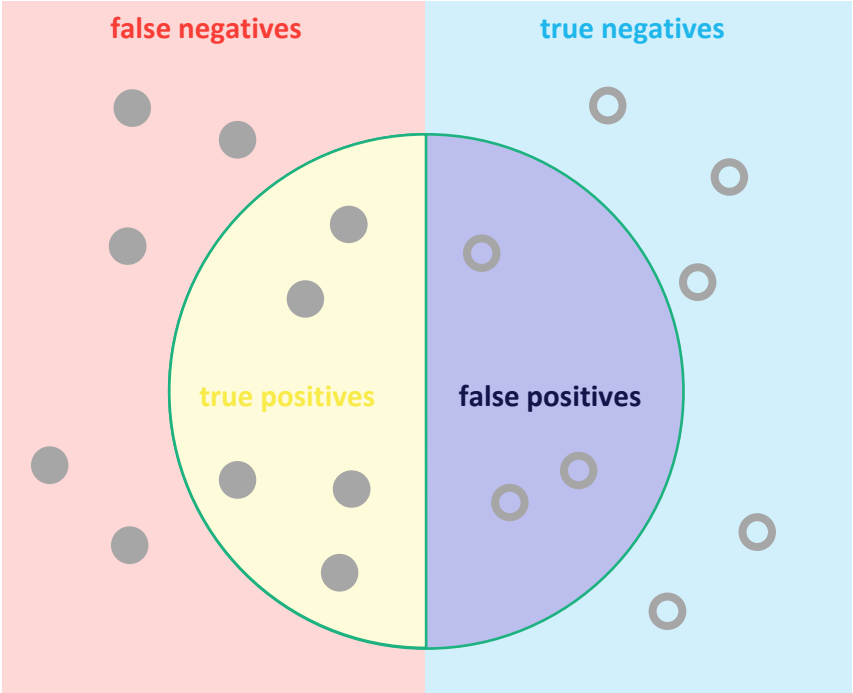
$$\text{Precision} = \frac{\quad}{\quad}$$



https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021

Recall: How many relevant items are selected?

- Referred to as “Sensitivity” or True Positive rate
  - E.g., Number of images accurately labeled as ‘dog’ relative to total number of dog images



$$\text{Recall} = \frac{\quad}{\quad}$$

https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021

Evaluation: Key takeaways

- Performance can be calculated with a variety of metrics
  - Select your metric carefully! Accuracy is not always the best choice
- Beware misleading metrics!
  - If you have a database of Gatorade drink colors, and a sample of Windex, it'll be identified confidently as blue Gatorade. This is a big problem if you plan to drink it.

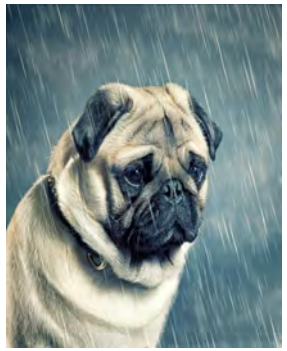
Metric	**Calculation**
Recall/ Sensitivity/ TP Rate	$\frac{TP}{TP+FN}$
Precision/ Positive Predictive Value	$\frac{TP}{TP+FP}$
Specificity/ TN Rate	$\frac{TN}{TN+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F1 score	$\frac{2TP}{2TP+FP+FN}$

https://stackoverflow.com/questions/50666091/true-positive-rate-and-false-positive-rate-tpr-fpr-for-multi-class-data-in-py

Important Principles

Misleading Evaluations

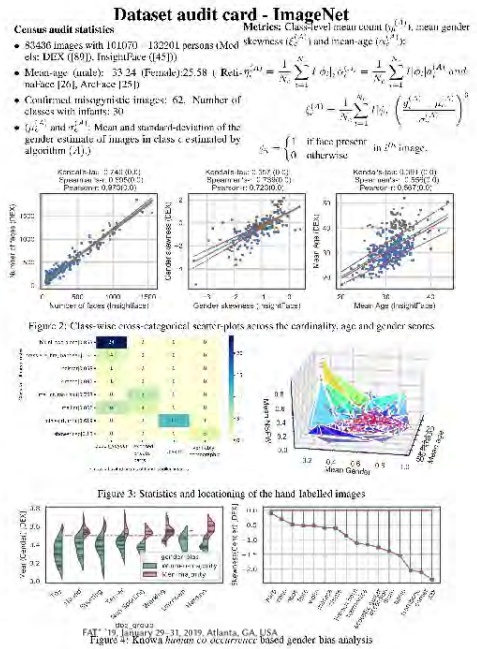
- Incomplete evaluation
  - A model that chooses true 100% of the time has 100% Recall
  - A model that gets 1 right out of 1000 can have 100% Precision
  - You may need more than one metric!
- Misleading data
  - Data collected for evaluation is not realistic or representative
  - If you have 99 cat images and one dog image,
    even a poor classifier will achieve 99% accuracy



vs

Important Topics to Consider

- Reproducibility! Just like with a lab notebook
- Data reproducibility
  - Where does the data come from?
  - How did you process the data?
- Model reproducibility
  - Code, frameworks, packages
  - Training environment, hyperparameters
  - Trained model
- Explainability/Interpretability
  - How interpretable are the model's decisions to a human?
  - How interpretable do they need to be?



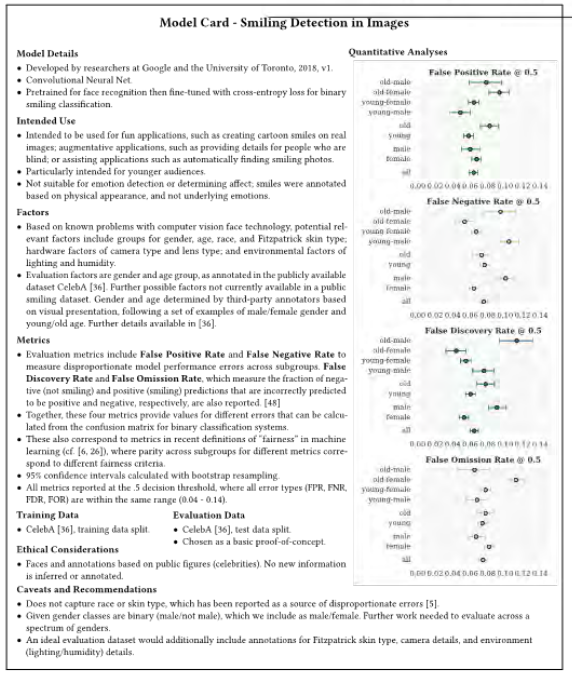


Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

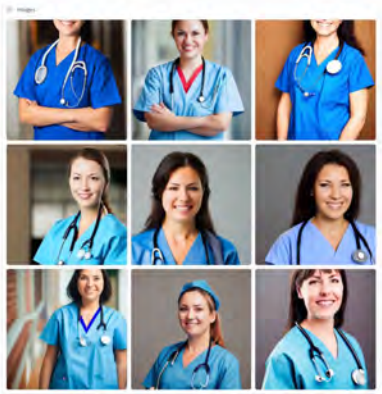
Important Topics to Consider

- Ethics
  - Responsible: ill-posed questions (or inappropriate-for-ML questions) can lead to large consequences
  - Equitable: minimize unintended bias
  - Traceable: document, document, document. Your model, your data sources, your designs, your methods. Same as you would a lab notebook—so that someone can reproduce your results
  - Reliable: explicit, well-defined uses and testing across the ML lifecycle
  - Governable: don't release your chatbot on Twitter

"The Kalashnikov 'combat module' will include 7.62-millimeter machine gun coupled with a camera attached to a computer system. According to TASS, the module uses 'neural network technologies that enable it to identify targets and make decisions'."

Doctor

Nurse



https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer

Useful Tools

Useful Tools

- Python
  - Data packages: Pandas, Numpy,
  - Modeling packages: SciKitLearn, PyTorch, Keras, HuggingFace
  - Visualization packages: matplotlib, plotly
- R
  - Data packages: Tidyverse, data.table
  - Modeling packages: tidymodels, keras, caret
  - Visualization packages: ggplot, plotly, trelliscopejs
- GitHub
  - Version control!
- Are you stuck?
  - StackOverflow, Reddit

https://github.com/

The AI community building the future.

Build, train and deploy state of the art models powered by the reference open source in machine learning.

https://huggingface.co/

https://www.tidyverse.org/

Questions?

Networking Break

10:05 – 10:15 a.m.

8:30-8:35 a.m.	Introduction	Luke Durell
8:35-9:20	Data Science for ‘Omics Data	Kelly Stratton & Marija Velickovic
9:20-10:05	Data Science Workflows	Eva Brayfindley
10:05-10:15	Networking Break	
10:15-11:00	**Data Science Communication**	**Moses Obiri**
11:00-11:45	EMSL Data Science Tools	David Degnan

EMSL

Data Science Communication: Visualizing and Interpreting Data Ethically

Moses Obiri, PhD
Data Scientist/Statistician

Outline

- Introduction to data science communication
  - Data visualization
  - Types of plots and their uses
  - Misinterpretation of plots and misleading information
- Principles of effective data science communication
- Ethics in data science

Data Science Communication

Data Science Communication

Introduction to data science communication

- Data science communication is the process of turning complicated, often large amounts of data into information that is clear and easy to understand. It includes graphics, storytelling, and the use of data interpretation techniques

Importance

- Clarity/decision making
  - Assists in transforming complex data into useful insights, improving the usability of information for decision-making
- Accessibility
  - Makes data or results from complex models more understandable and interactive for non-experts
- Modeling/Future predictions
  - Illustrates trends that inform modeling, forecasts and future planning

Data Visualization

What is data visualization?

- Data visualization is the graphical representation of data and information.

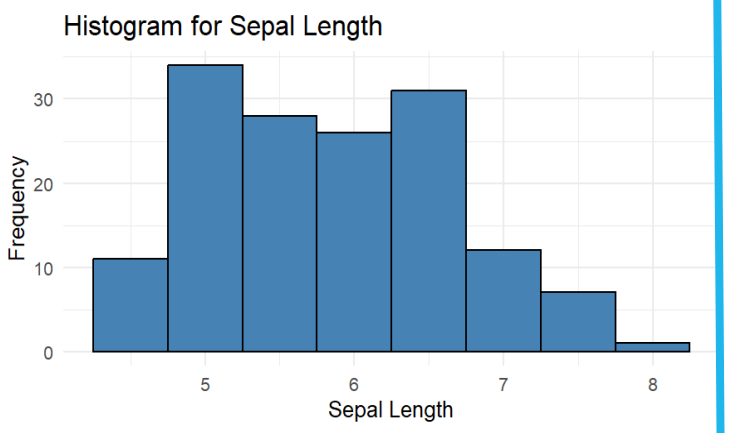
Purpose

- Simplify complex data
  - Breaks down complex datasets into simpler, visual formats that are easier to understand
- Highlight trends and patterns
  - Bring out hidden patterns, trends, and insights in the data
- Engage the audience
  - Well-crafted visualizations can grab the attention of the audience and make the communication more impactful
- Facilitate quick decision making
  - Presenting data in an understandable way can help decision-makers draw conclusions quickly and make informed decisions

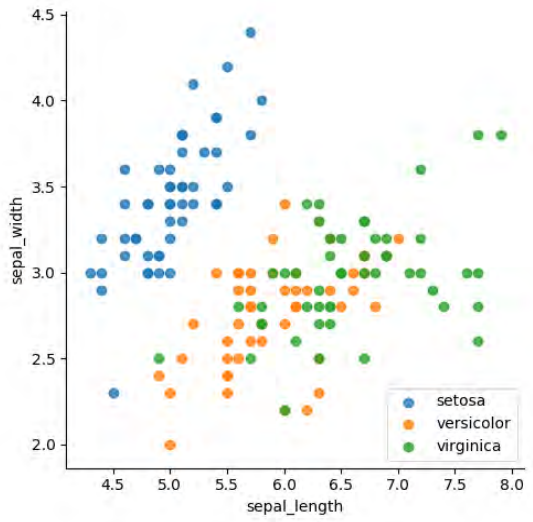
Types of plots

- Numerical data

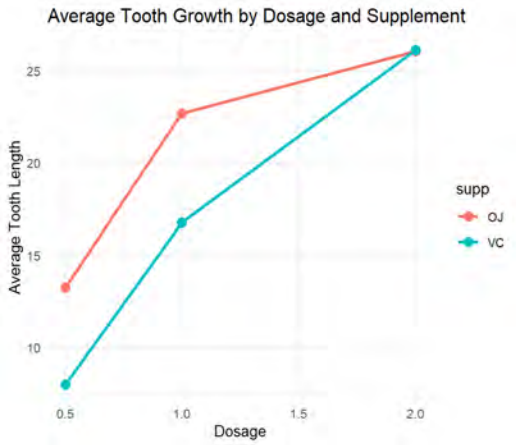
Histograms: To represent the distribution of a continuous numerical variable



Scatter plots: Visualize the relationships b/n 2 numerical variables



Line graphs: Show a trend b/n 2 numerical variables

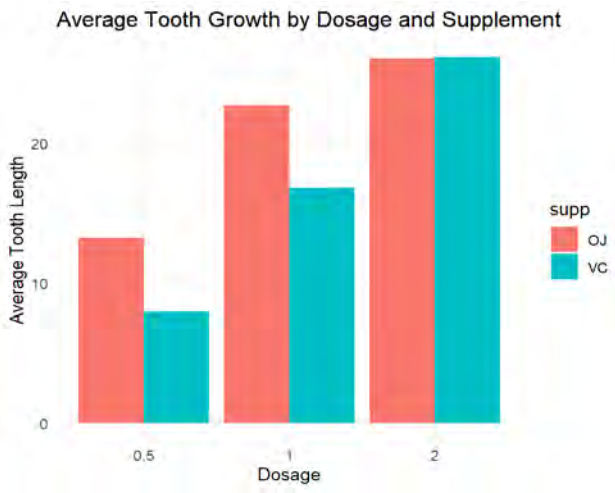


Types of plots

- Categorical data

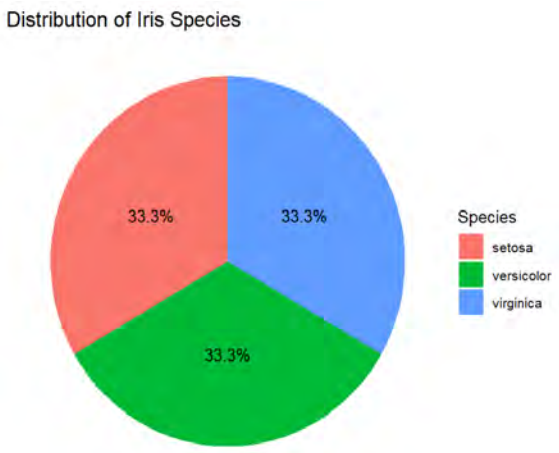
Bar graphs:

To compare the frequency, count, or other measures for different categories



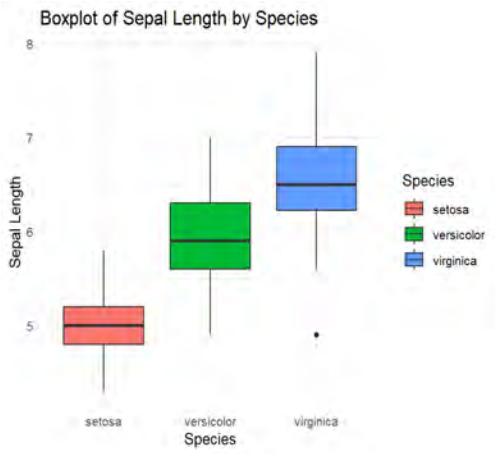
Pie charts:

To display proportion of each category as part of a whole



Boxplots:

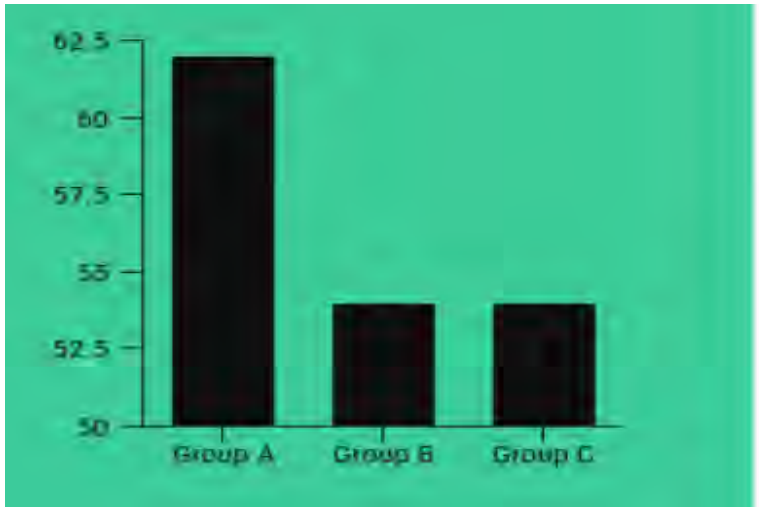
To show the distribution of a numerical variable for different categories

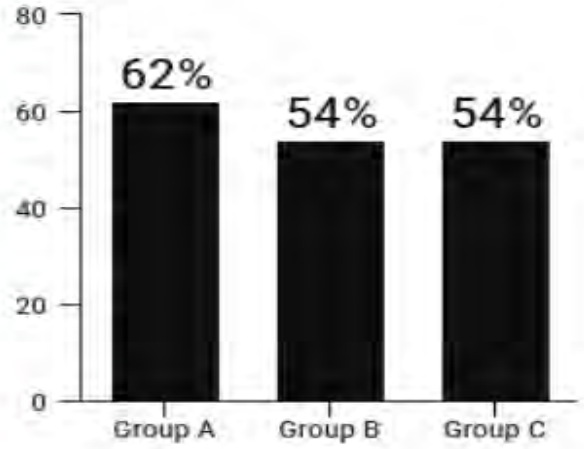


Misinterpretation of plots

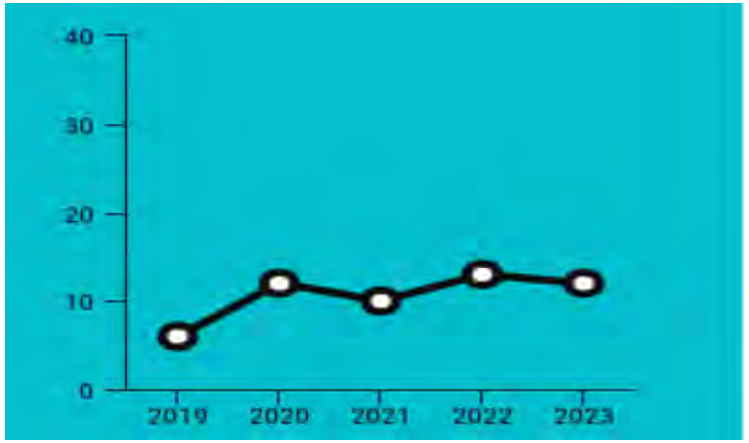
Scale and Axes Manipulation

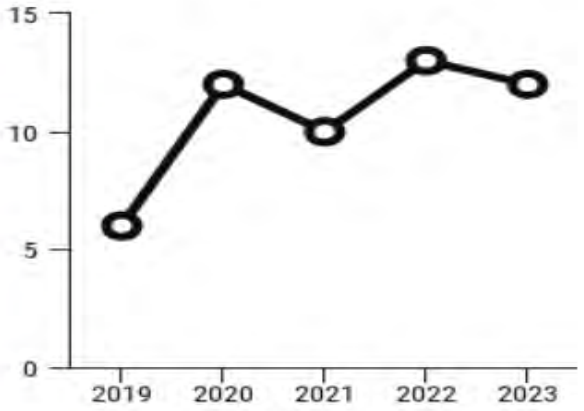
- Omitting baseline





- Manipulating Y-axis

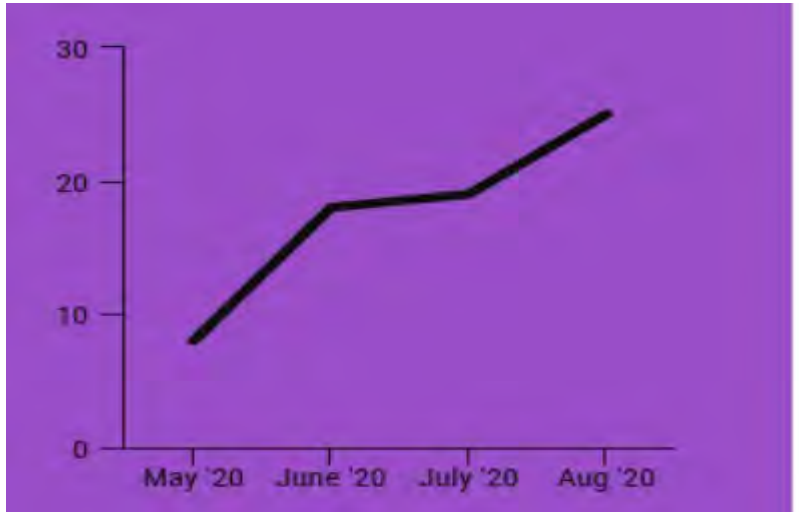


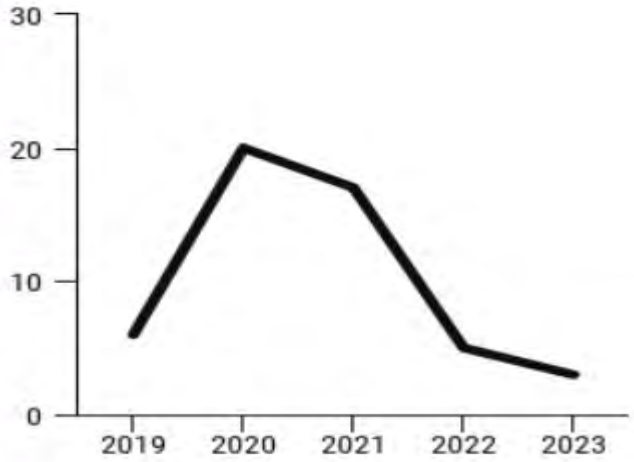


Misinterpretation of plots

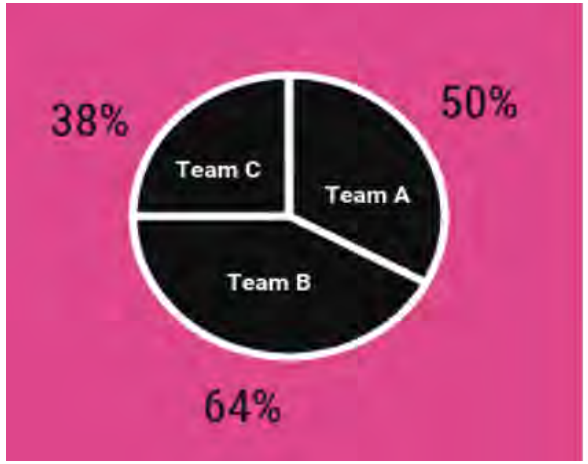
Data manipulation/wrong graph

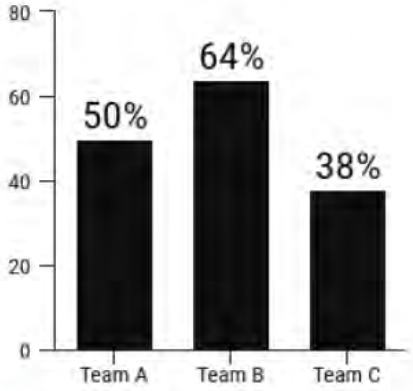
- Data omission/truncation





- Wrong graph



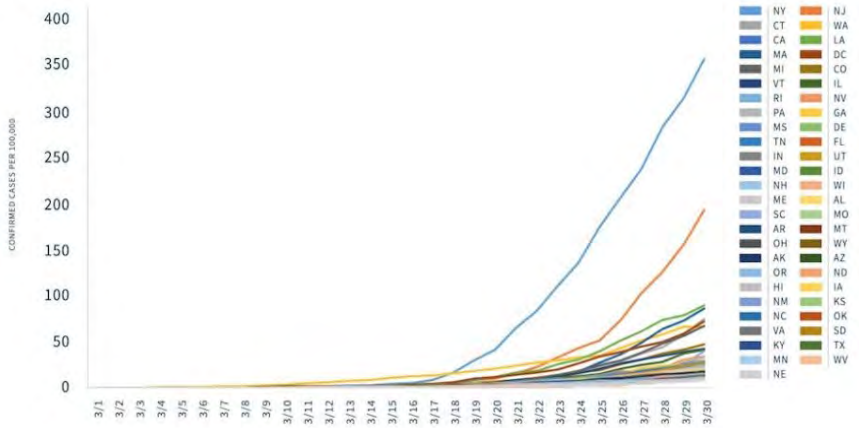


Communication Principles

Communication Principles

Good visualization

- Know your audience: Understanding who you are communicating with is fundamental. The level of technical detail, the type of plot and context required can vary depending on the audience
  - Communicating a technical audience (like data scientists, biologists) vs a non-technical audience (like executives or the general public, sponsors)
- Keep it simple: The best visualizations are simple and not cluttered. They communicate the main idea efficiently
- Hard to read the rise in cases across states
- Graph a subset of states at a time



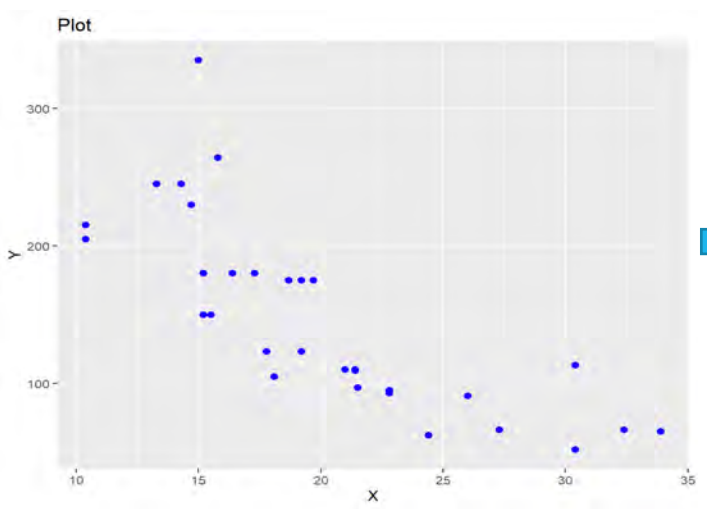
Communication Principles

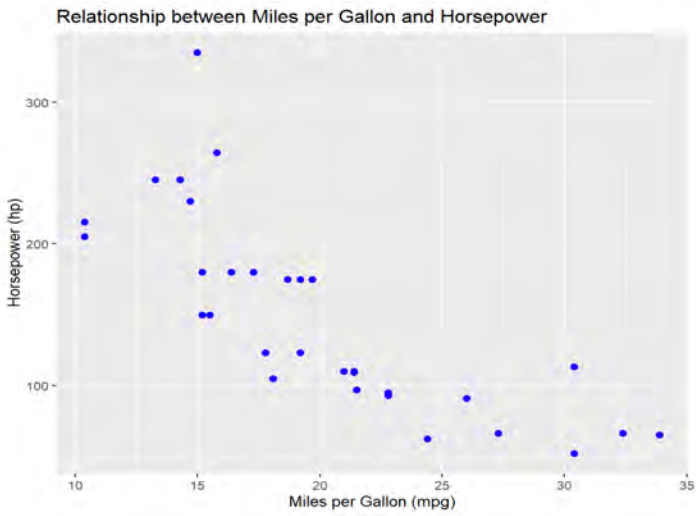
Accuracy

- Visualizations should accurately represent the data. Misleading scales, cherry-picked data, or other distortions should be avoided.
  - $Lie\ factor = \frac{effect\ shown\ in\ a\ graph}{effect\ seen\ in\ the\ data} \approx 1.$

Clarity

- Visualizations should be easily understandable. Use clear labels, legends, and annotations. Avoid confusing or overly complex visuals

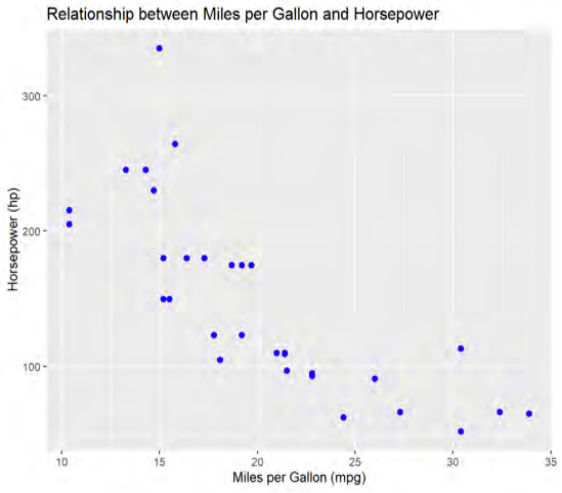


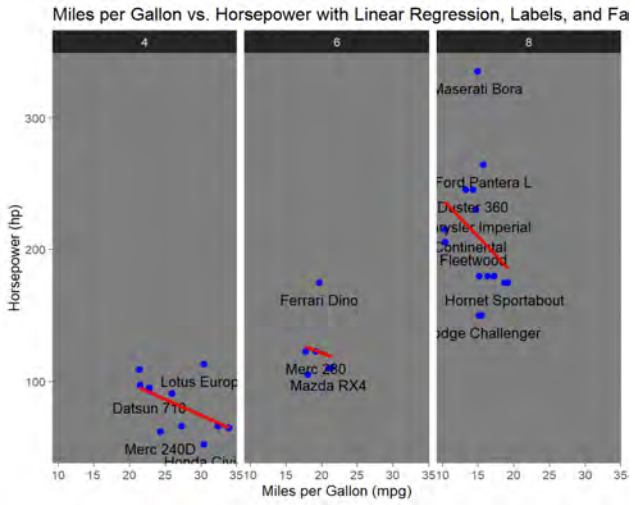


Communication Principles

Relevancy

- Plots should be directly relevant to the content being communicated. Avoid unnecessary decoration or unrelated elements





Added regression lines, changed the theme, facetted by the number of cylinders to make the plot too busy

Accessibility

- Ensure visualizations are accessible to all users. This includes considering color blindness and providing text alternatives

Storytelling

- A good visualization tells a story. It should guide the audience through the data and support the main narrative

Data Science Ethics

Data Science Ethics

Ethics

- Transparency
  - This involves clearly explaining the methodology and tools used to gather and analyze the data. Transparency also includes making the results and the processes leading to them accessible and understandable to relevant stakeholders
- Privacy and Confidentiality
  - Subjects' privacy must be protected. Data should be de-identified, securely stored, and collected with informed consent
- Accuracy
  - Data scientists must strive to provide accurate and reliable results. This includes rigorously validating models, acknowledging uncertainties or limitations, and correcting errors promptly when they are discovered

Data Science Ethics

Ethics

- Fairness and Bias
  - Scientists must be aware of and seek to reduce biases in their data and models, whether they are related to data collection, dataset composition, or modeling. They should strive for impartiality in their work
- Accountability
  - Scientists must own their work and its effects. They should accept criticism, fix errors, and examine the impact of their studies and models

Conclusion

Review of key points

- Understanding types of plots and when to use them effectively communicate the insights in our data
- Awareness of misinterpretations and misleading information in plots helps ensure our visualizations are accurate and ethical
- Following good principles of data science communication ensures our work is clear, engaging, and impactful
- Adhering to data science ethics helps ensure our work is transparent, fair, accurate, and respectful of privacy and confidentiality
- Effective data science communication and ethical practice are key to making data science truly beneficial for our communities, industries, and societies

References

- Driessen, J. E., Vos, D. A., Smeets, I., & Albers, C. J. (2022). Misleading graphs in context: Less misleading than expected. *Plos one*, *17*(6), e0265823.
- *Create Amazing Animated Graphs in Python with this 2-Liner Code in Python*. (n.d.). Retrieved June 11, 2023, from https://www.analyticsvidhya.com/blog/2021/04/animated-bar-graph-data-science-project/.
- *5 Ways Writers Use Misleading Graphs To Manipulate You [INFOGRAPHIC] - Venngage*. (n.d.). Retrieved June 11, 2023, from https://venngage.com/blog/misleading-graphs/#Misleading-Coronavirus-graphs.
- Wickham, Hadley, et al. "dplyr: A Grammar of Data Manipulation. R package version 0.7. 6." *Computer software]. https://CRAN. R-project. org/package= dplyr* (2018).

Questions?

8:30-8:35 a.m.	Introduction	Luke Durell
8:35-9:20	Data Science for ‘Omics Data	Kelly Stratton & Marija Velickovic
9:20-10:05	Data Science Workflows	Eva Brayfindley
10:05-10:15	Networking Break	
10:15-11:00	Data Science Communication	Moses Obiri
11:00-11:45	**EMSL Data Science Tools**	**David Degnan**

EMSL

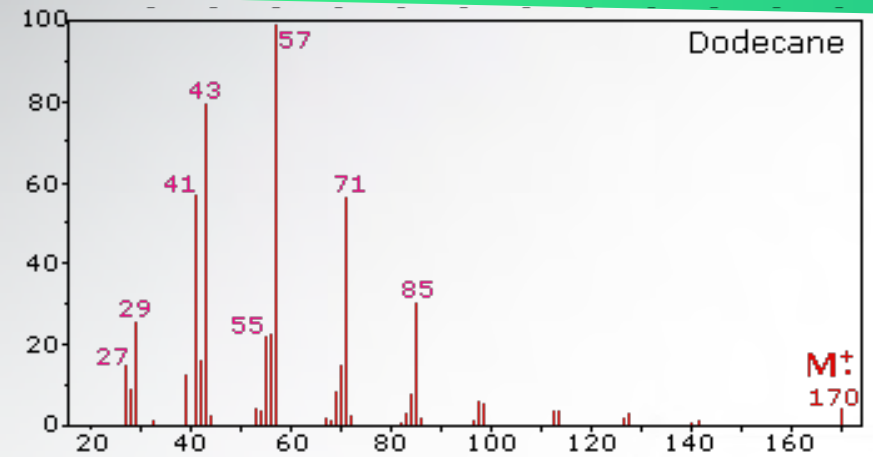
A Tour of EMSL Packages & Web Tools

David Degnan
Biological Data Scientist

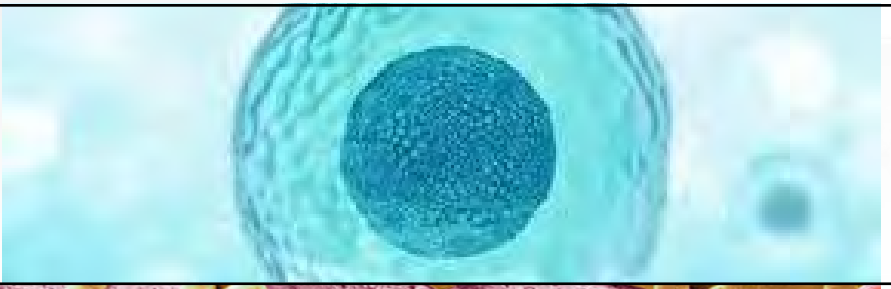
Rules of Today's Tour

- Welcome to your personalized tour of EMSL tools & applications!
- Like every good tour, we have a few rules to cover:
  1. In no way is this a comprehensive tour of every tool & application currently developed or supported at EMSL
  2. For clarity, we will be presenting tools within the "bulk omics" general workflow. There is always nuanced variability in how analyses are conducted.
  3. No tool is a one-size-fits-all, however, almost all tools presented here are under *active* and *continued* development, which may open the door for collaboration opportunities.
  4. I will do my best to credit staff involved with each of these tools, and I am sorry if anyone is missed.

Types of Omics Workflows Being Developed at PNNL

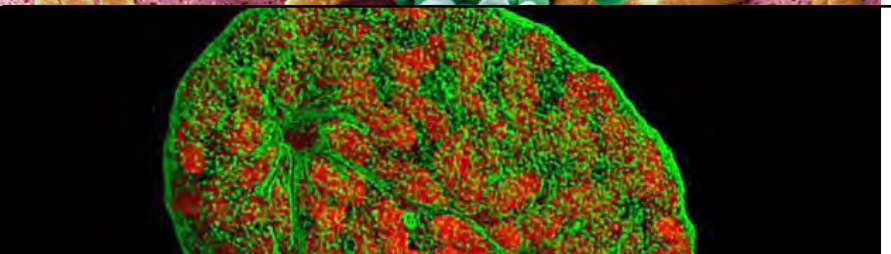


- Here, we will refer to tools in MS and NMR-based omics applications. There are a few tools that also include RNA-seq data (transcriptomics).
- **Bulk MS Omics:** The traditional omics pipeline where typical GC/LC-MS or NMR is used. ← Focus of this talk



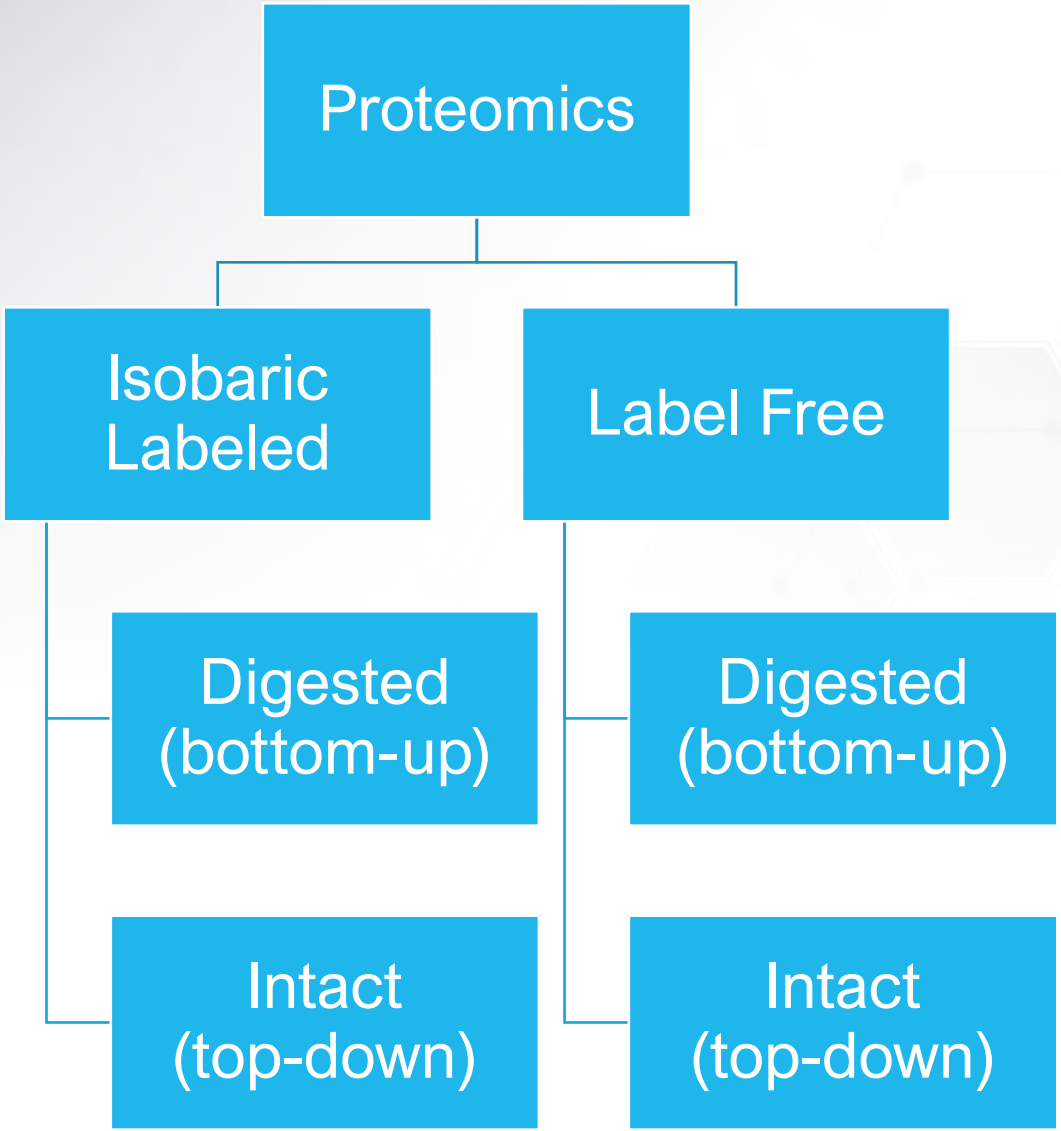
- **Single Cell MS Omics:** A burgeoning technology that uses microPOTs and nanoPOTs to detect biomolecules at the single cell level

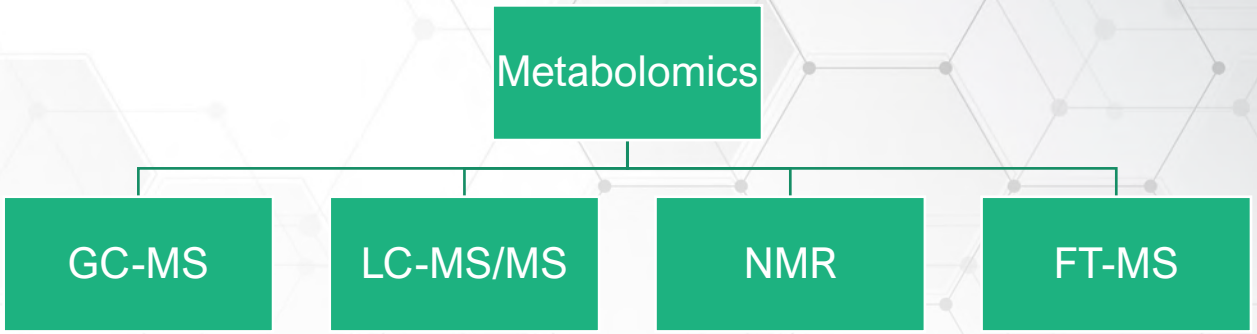
- **Meta MS Omics:** Adapting traditional bulk omics to multi-species microbial communities. Under development.



- **Spatial MS Omics:** Using new 3D mass spectrometry to understand biomolecule dispersement across tissues and sediments.

Which “omics” are included in bulk omics?

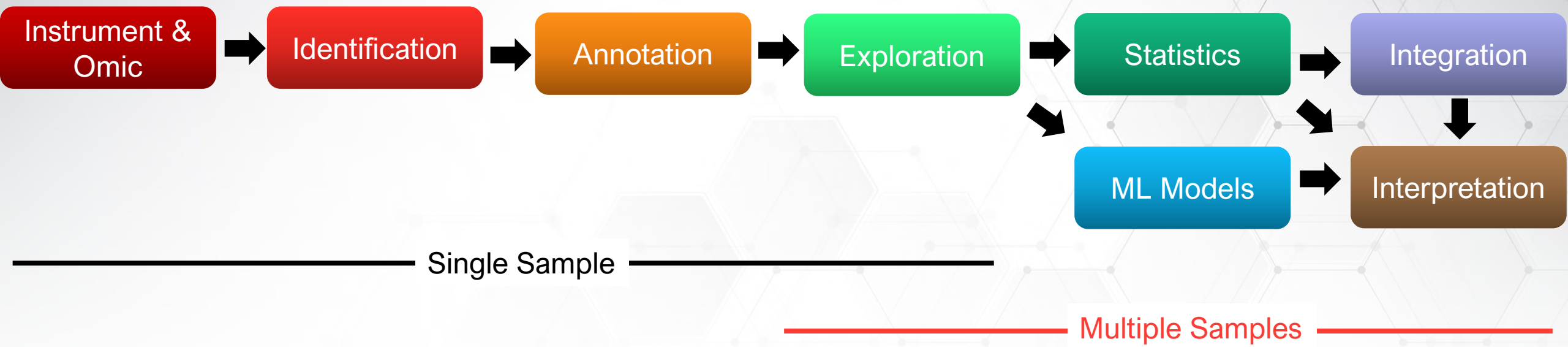




Others include:

- Lipidomics (typically MS)
- Transcriptomics (Illumina or PacBio)
- MALDI-MS (any MS omic)

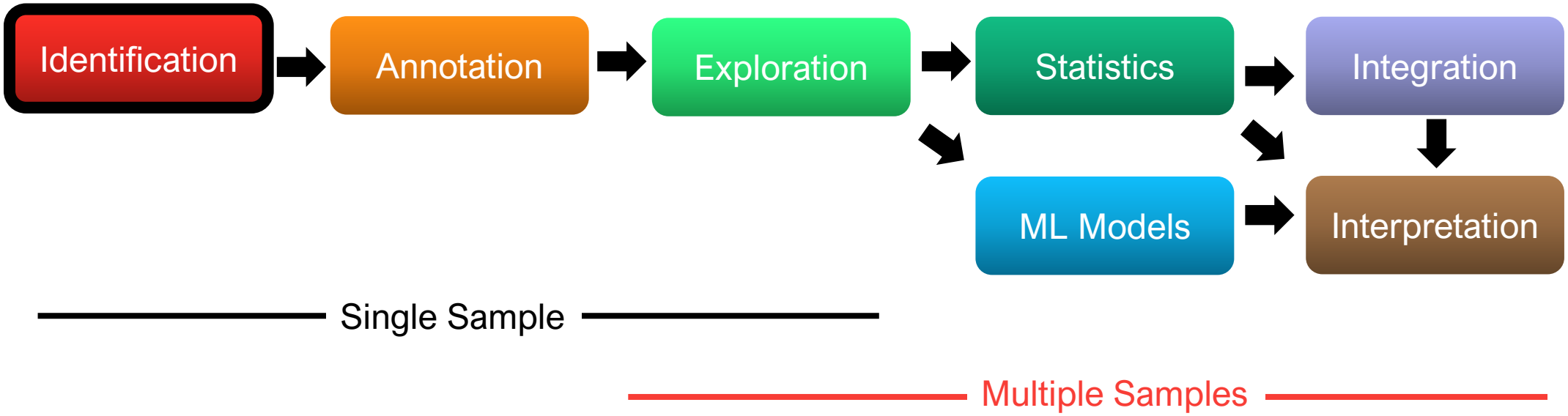
Here are the stops on our tour of bulk omics tools!



Here are some examples:

Omic Type	Identification	Annotation	Exploration	Statistics	Integration	ML Models	Interpretation
Label-Free Top-Down Proteomics	MSPathFinder TopPIC	PSpecteR	MODE	PMart	iPMart	SLOPE	IsoForma
NMR Metabolomics	NMRAnalysis	N/A	MODE	PMart	iPMart	SLOPE	N/A
FT-MS Metabolomics	CoreMS	CoreMS	FREDA	N/A	N/A	SLOPE	N/A

Identification Tools



Most are command line interfaces (CLIs) that require users to have some understanding of the instruments and the studied biological system.

Proteomics Identification: MS-GF+

Tool Type: Command Line Interface (CLI)

MSGFPlus/
msgfplus

MS-GF+ (aka MSGF+ or MSGFPlus) performs peptide identification by scoring MS/MS spectra against peptides derived from a protein sequence database.

10 Contributors | 60 Issues | 59 Stars | 36 Forks

- **Omics:** Bottom-up (digested) peptides for both labeled and label-free datasets
- **Description:** A fully automated CLI tool written in Java for identifying peptides. Works with several ionization modes and is widely-used. Limited to input database.
- **Status:** Maintenance. https://github.com/MSGFPlus/msgfplus
- **Development team:** Though initially developed by Kim & Pevzner 2014, it is currently maintained at PNNL by Bryson Gibbons – Bryson.Gibbons@pnnl.gov

Proteomics Identification: InformedProteomics / MSPathFinder

Tool Type: Command Line Interface (CLI)

PbfGen
Condense MS data

↓

ProMex
Feature Identification

↓

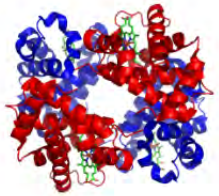
MSPathfinderT
Map long peptide sequences and proteins to features

- **Omics:** Label-free top-down (intact) proteomics
- **Description:** A fully automated C# CLI tool for top-down mass spectrometry.
- **Status:** Maintenance. https://github.com/PNNL-Comp-Mass-Spec/Informed-Proteomics
- **Development team:** Many PNNL scientists, see https://www.nature.com/articles/nmeth.4388. Maintained by Matthew Monroe – Matthew.Monroe@pnnl.gov

Proteomics Identification: MASIC

Tool Type: Command Line Interface (CLI)

PNNL-Comp-Mass-Spec/**MASIC**



MASIC generates selected ion chromatograms (SICs) for all of the parent ions chosen for fragmentation in an LC-MS/MS analysis, characterizing...

2 Contributors | 3 Issues | 10 Stars | 4 Forks

- **Omics:** Labeled proteomics, either top-down (intact) or bottom-up (digested).
- **Description:** Fully automated C# CLI tool to identify features across isobaric labeled proteomics data. Usually combined with identification data from another tool, like MS-GF+.
- **Status:** Maintenance. https://github.com/PNNL-Comp-Mass-Spec/MASIC
- **Development team:** Many PNNL scientists. Currently maintained by Matthew Monroe & Bryson Gibbons

Proteomics Identification: MSnID, PlexedPiper

Tool Type: R package

vladpetyuk/
PlexedPiper

Isobaric Tag Processign Pipeline

7 Contributors | 5 Issues | 2 Stars | 5 Forks

- **Omics:** Labeled proteomics
- **Description:** A series of R packages for analyzing and combining labeled proteomics data (i.e. MASIC and MS-GF+)
- **Status:** Maintenance. https://github.com/PNNL-Comp-Mass-Spec/MSnID
- **Development team:** Vlad Petyuk – Vladislav.Petyuk@pnnl.gov, Tyler Sagendorf – Tyler.Sagendorf@pnnl.gov, and others

Proteomics Identification: TopPICR

Tool Type: R package

- **Omics:** Label free top-down mass spectrometry.
- **Description:** R package to extend TopPIC capability to label free top-down mass spectrometry.
- **Status:** Completed and published. https://github.com/PNNL-Comp-Mass-Spec/TopPICR
- **Development team:** Vlad Petyuk, James Fulcher – James.Fulcher@pnnl.gov, Mowei Zhou, Matt Monroe, and Evan Martin

Metabolomics Identification: CoreMS

Tool Type: Python Library & Web Application

CoreMS

- **Omics:** GC-MS, FT-MS, and LC-MS/MS (in development)
- **Description:** Python library & GUI. A platform approach to metabolomics.
- **Status:** Initial version released. Continued development. https://github.com/EMSL-Computing/CoreMS
- **Development team:** Yuri Corilo – Corilo@pnnl.gov, Will Kew – William.Kew@pnnl.gov, Lee Ann McCue, Anastasiya Prymolenna, and others



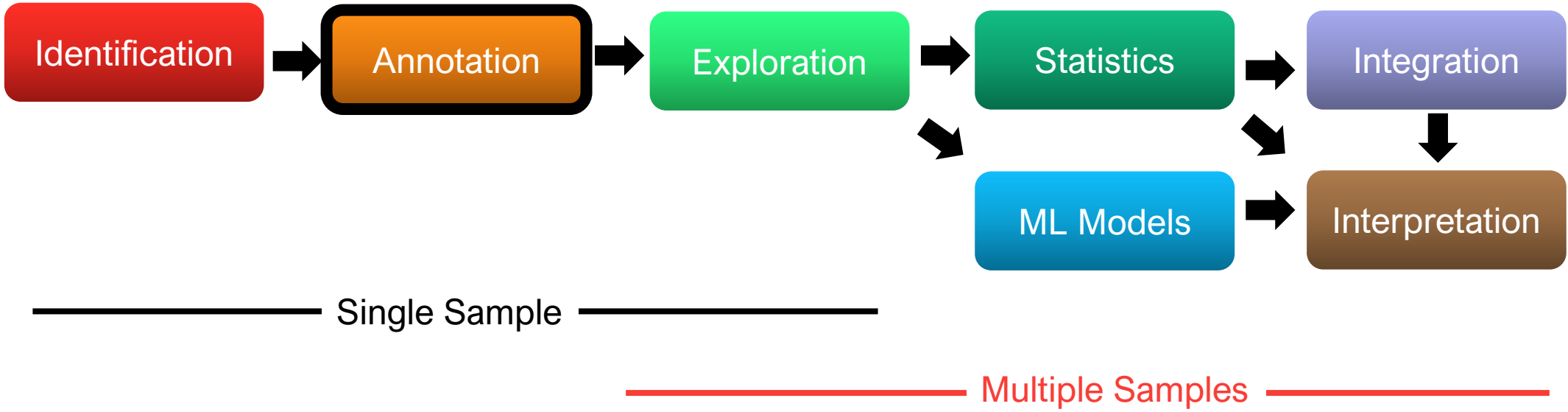

Other Identification Tools

Metabolomics: MetaboliteDetector, MS-Dial

Lipidomics: Liquid (EMSL), MS-Dial, MZMine

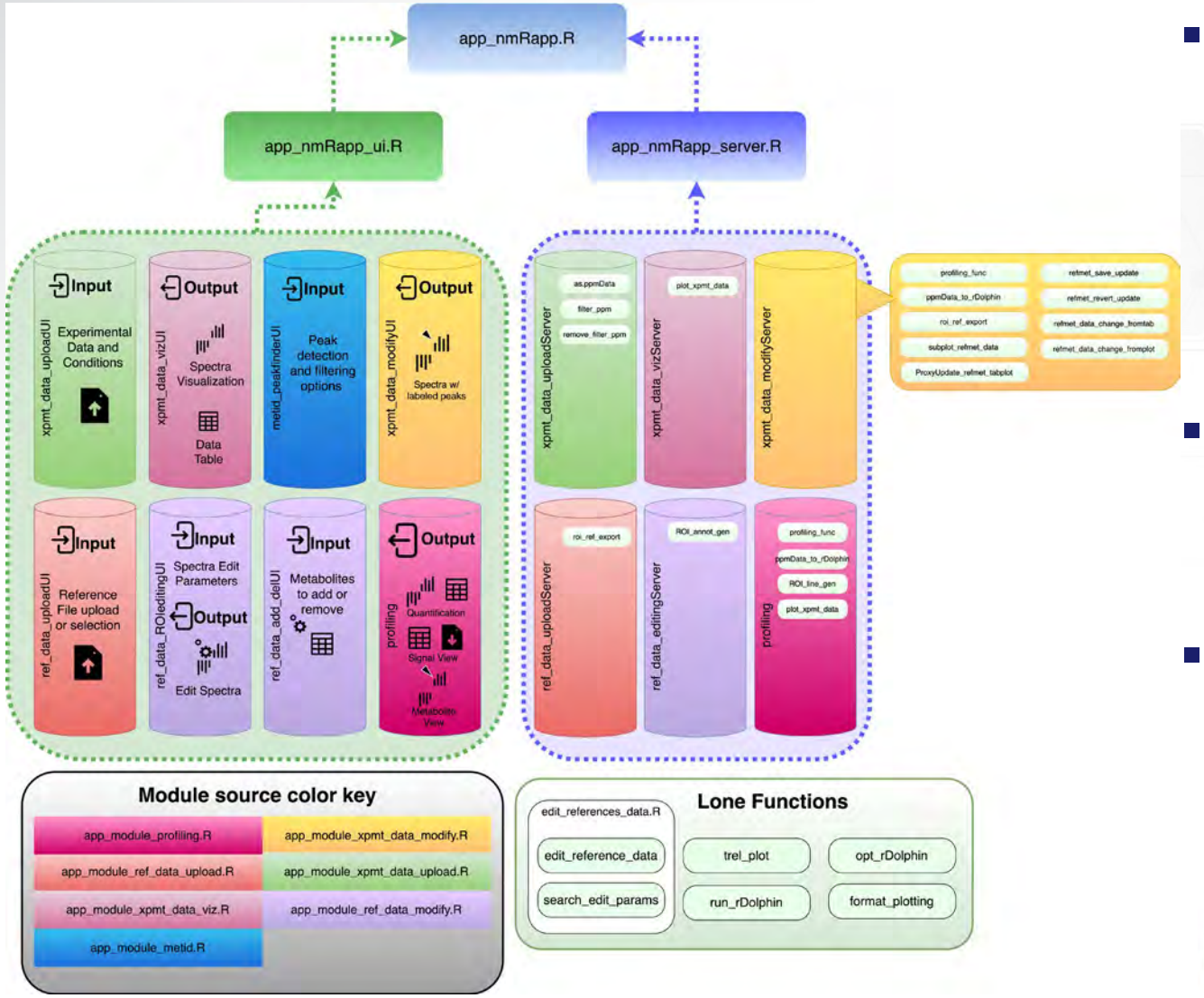
Proteomics: TopPIC, ProSight

Annotation Tools



Metabolomics Identification & Annotation: NMRanalysis

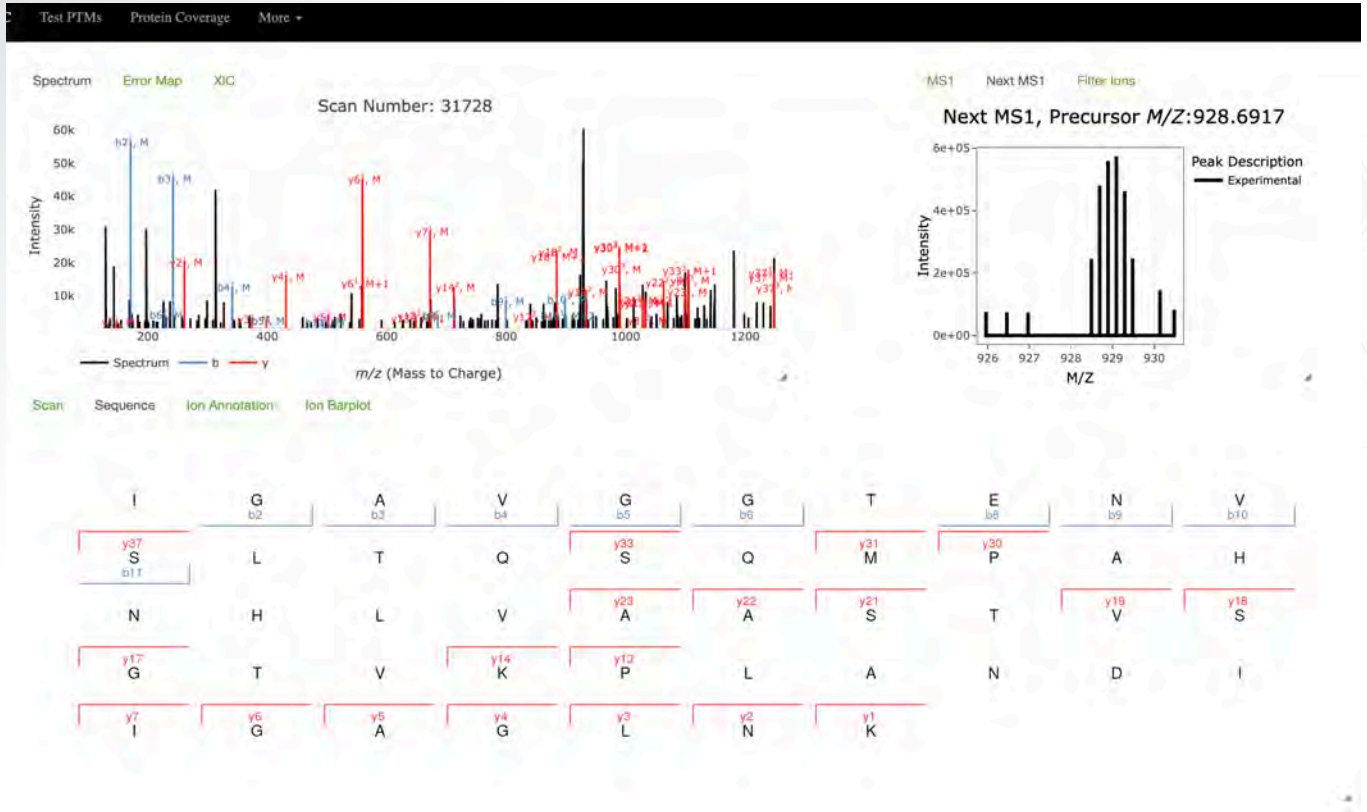
Tool Type: R Package & Shiny Web Application



- **Omics:** NMR metabolomics.
- **Description:** GUI & R package. Semi-automated metabolite annotation tool for NMR metabolomics. Stores all annotations for auto-filling as time goes on. More use = faster annotation. Reminder that not all tools perfectly fit!
- **Status:** Initial version released. Continued development. https://github.com/EMSL-Computing/nmRanalysis
- **Development team:** Javier Flores – Javier.Flores@pnnl.gov, Will Kew, Anastasiya Prymolenna, Natalie Winans, Logan Lewis, and others

Proteomics Annotation: PSpecteR

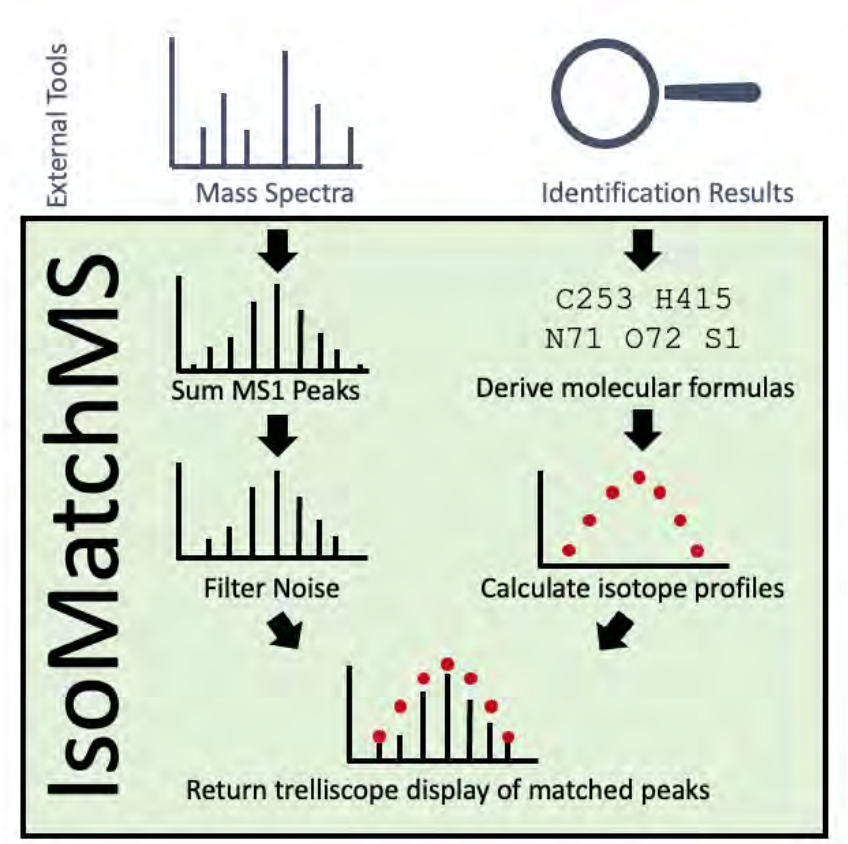
Tool Type: R Package & Shiny Web Application



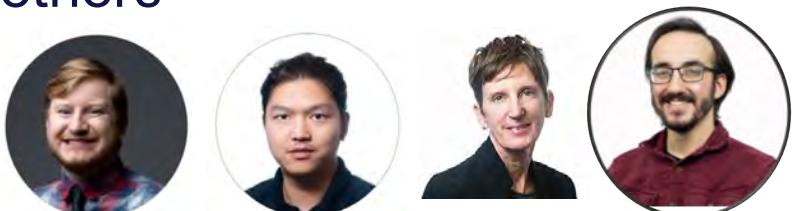
- **Omics:** Label-free intact and digested proteomics.
- **Description:** An R package & GUI for visualizing and testing peptides, modified peptides, and proteins.
- **Status:** Maintainence. https://github.com/EMSL-Computing/pspecterlib
- **Development team:** David Degnan – David.Degnan@pnnl.gov, Lee Ann McCue, Aivett Bilbao, Mowei Zhou, Lisa Bramer

MALDI-MS Annotation: IsoMatchMS

Tool Type: R Package



- **Omics:** Most MS-based omics, including label free proteomics,
- **Description:** An R package for matching identify high quality annotations from MALDI-MS data for downstream technology (like mass spec imaging). Results are visualized in trelliscope displays.
- **Status:** Recently released. https://github.com/PNNL-HubMAP-Proteoform-Suite/IsoMatchMS
- **Development team:** David Degnan, Mowei Zhou, Ljiljana Pasa-Tolic - Ljiljana.PasaTolic@pnnl.gov, Logan Lewis, and others

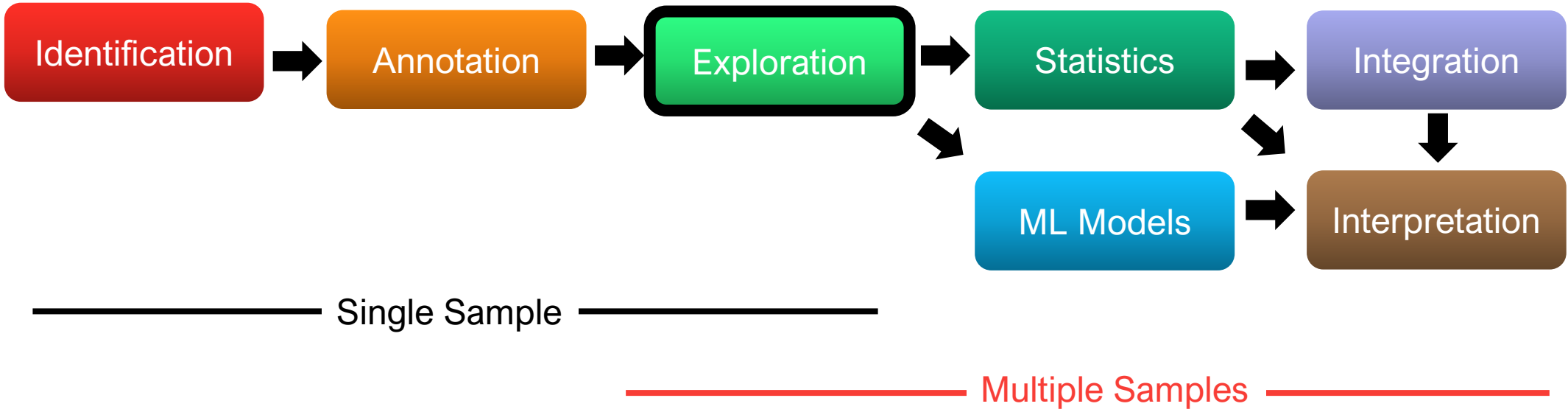


Other Annotation Tools

Metabolomics: AMDIS

Several MS Omics: Skyline

Exploration Tools



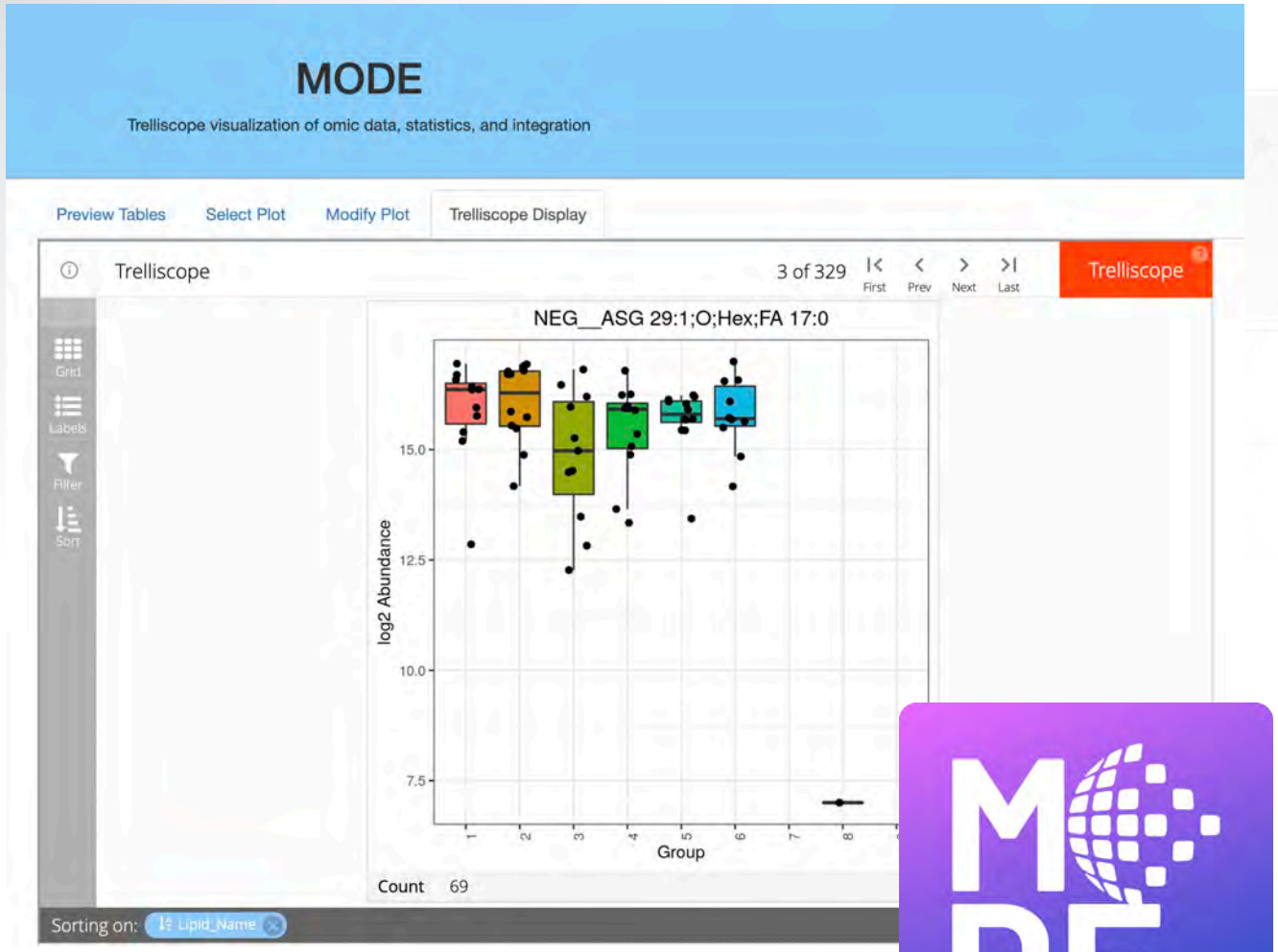
FT-MS Exploration: FREDA

Tool Type: R Package & Shiny Web Application

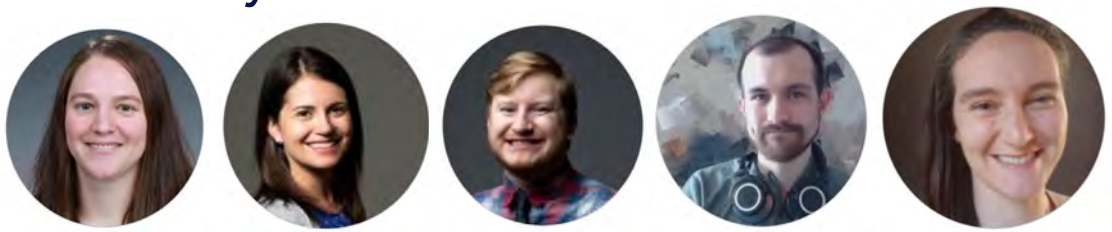
- **Omics:** FT-MS metabolomics
- **Description:** An R package and GUI for FT-MS data cleaning and exploration.
- **Status:** Maintenance. https://github.com/EMSL-Computing/ftmsRanalysis
- **Development team:** Daniel Claborne – Daniel.Claborne@pnnl.gov, Lee Ann McCue, Lisa Bramer, Kelly Stratton – Kelly.Stratton@pnnl.gov, and others

Many Omics Exploration: MODE

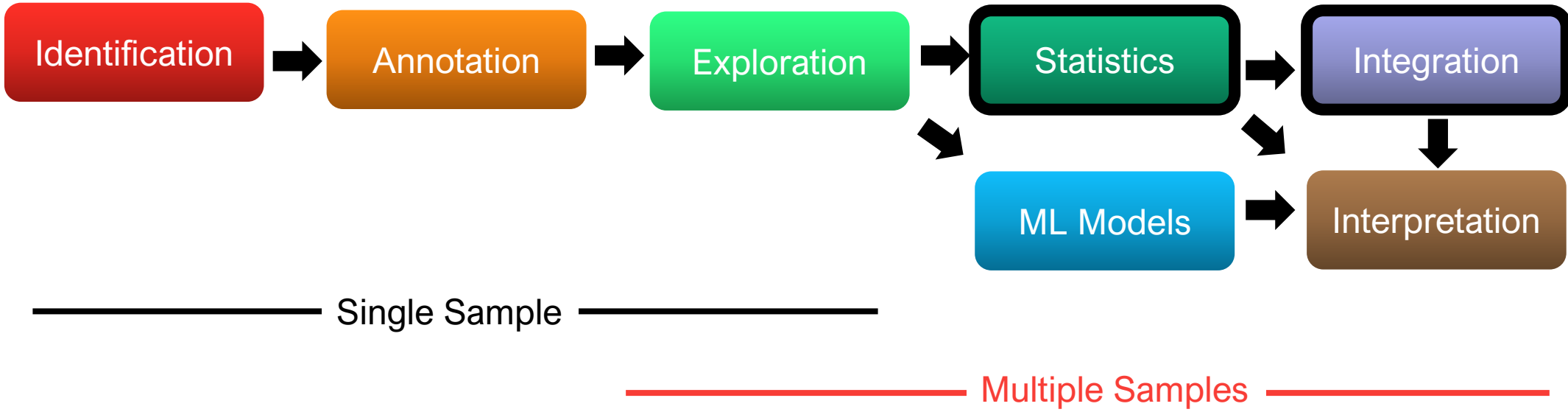
Tool Type: R Package & Shiny Web Application



- **Omics:** Labeled & label-free digested and intact proteomics, GC/LC-MS and NMR, metabolomics, lipidomics.
- **Description:** A GUI for visualizing trends in data the molecule, sample, and molecule class (i.e. protein) level.
- **Status:** Release & publication underway.
- **Development team:** MAP portal team, headed by Lisa Bramer



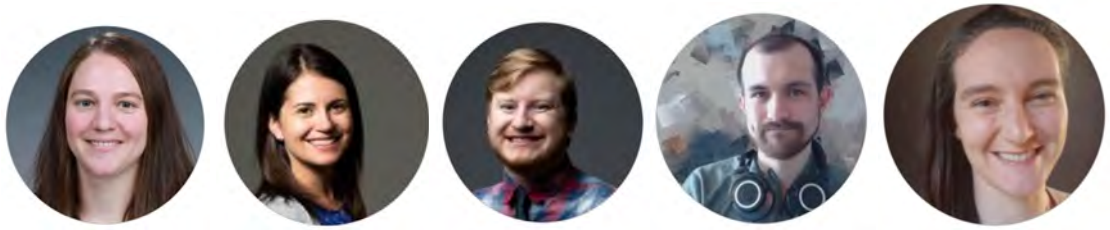
Statistics & Integration Tools



Many Omics Statistics: PMart

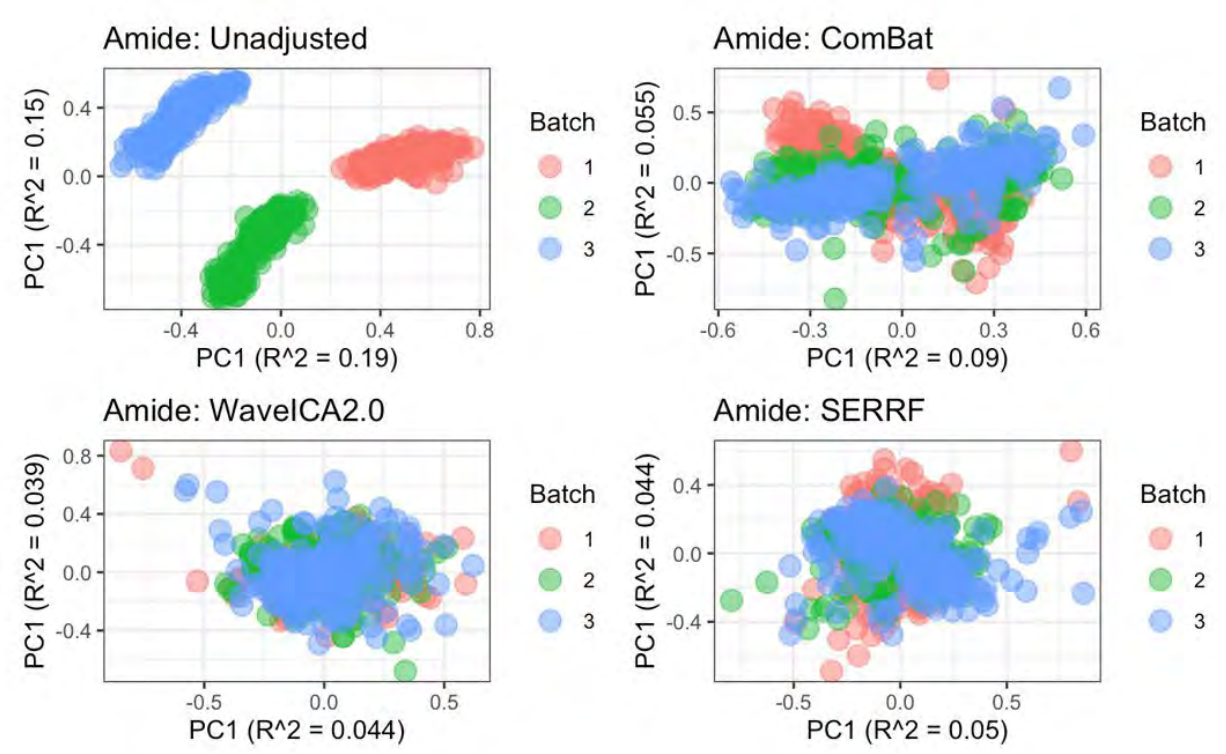
Tool Type: R Package & Shiny Web Application

- **Omics:** Labeled & label-free digested and intact proteomics, GC/LC-MS and NMR, metabolomics, lipidomics, transcriptomics.
- **Description:** An R package and GUI for differential expression and abundance filtering, normalization, and statistics.
- **Status:** Published and continually expanding.
- **Development team:** MAP portal team, headed by Lisa Bramer

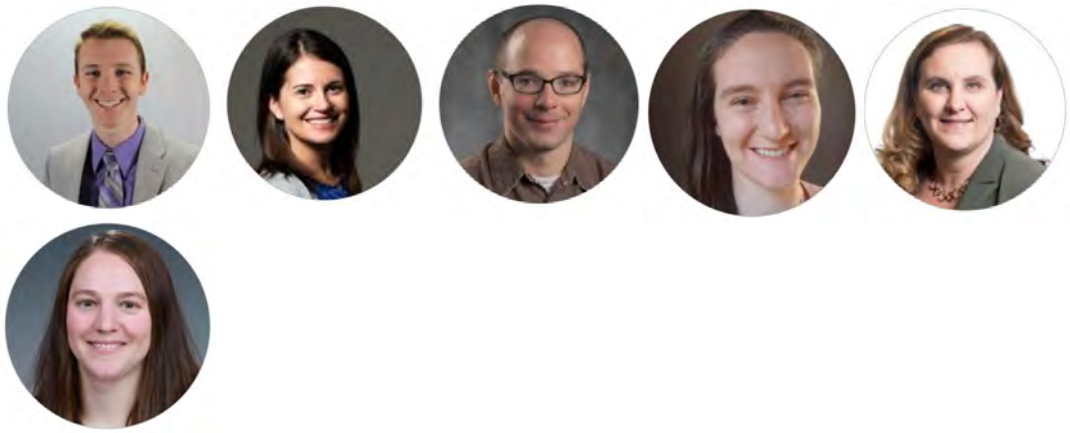


Many Omics Statistics: malbacR

Tool Type: R Package

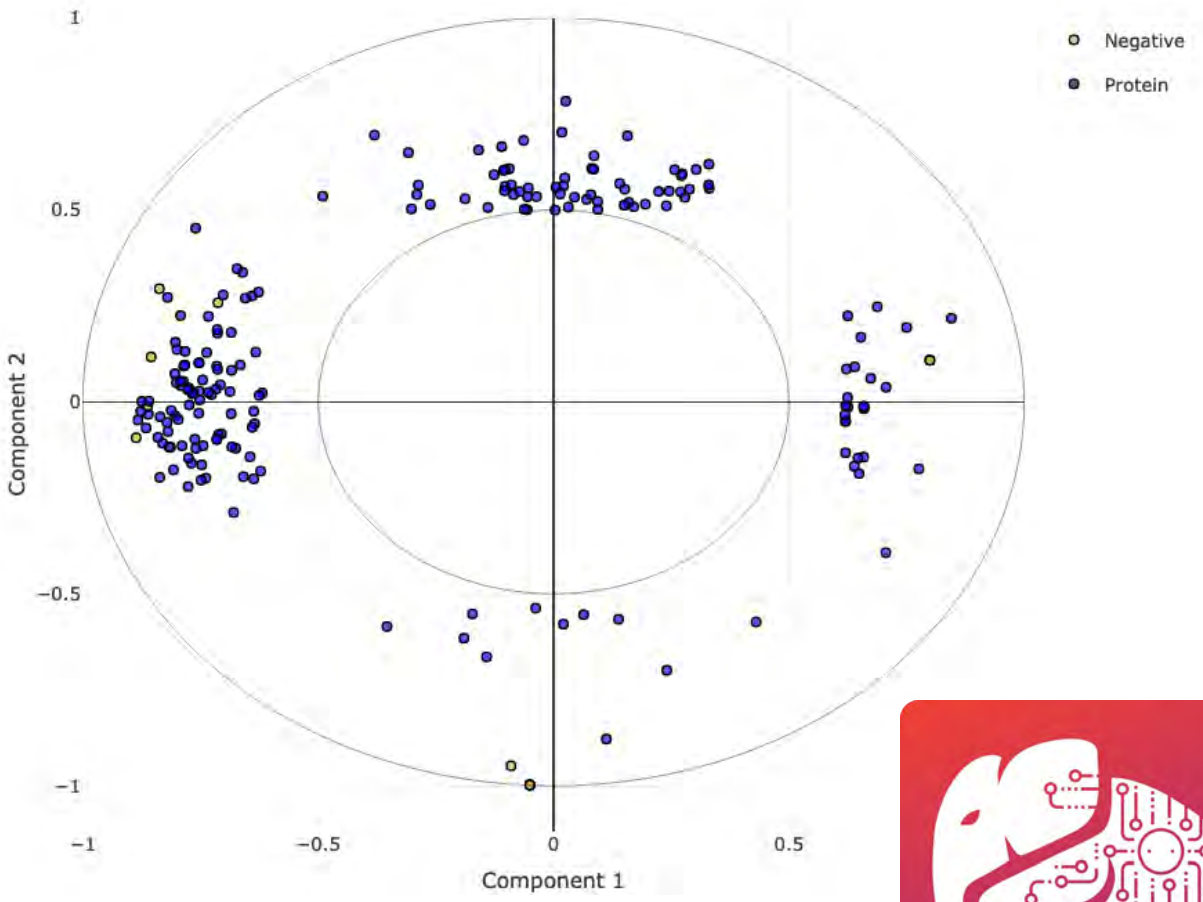


- **Omics:** Proteomics
- **Description:** Corrects batch effects in proteomics data.
- **Status:** Released.
- **Development team:** Damon Leach, Kelly Stratton, Jan Irvahn, Rachel Richardson, Bobbie-Jo Webb-Robertson, Lisa Bramer

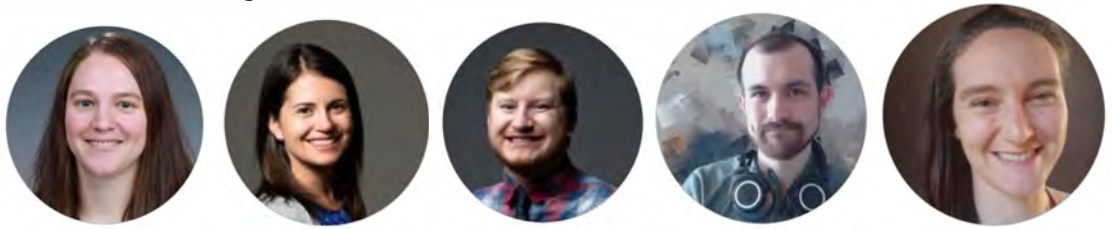


Many Omics Integration: iPMart

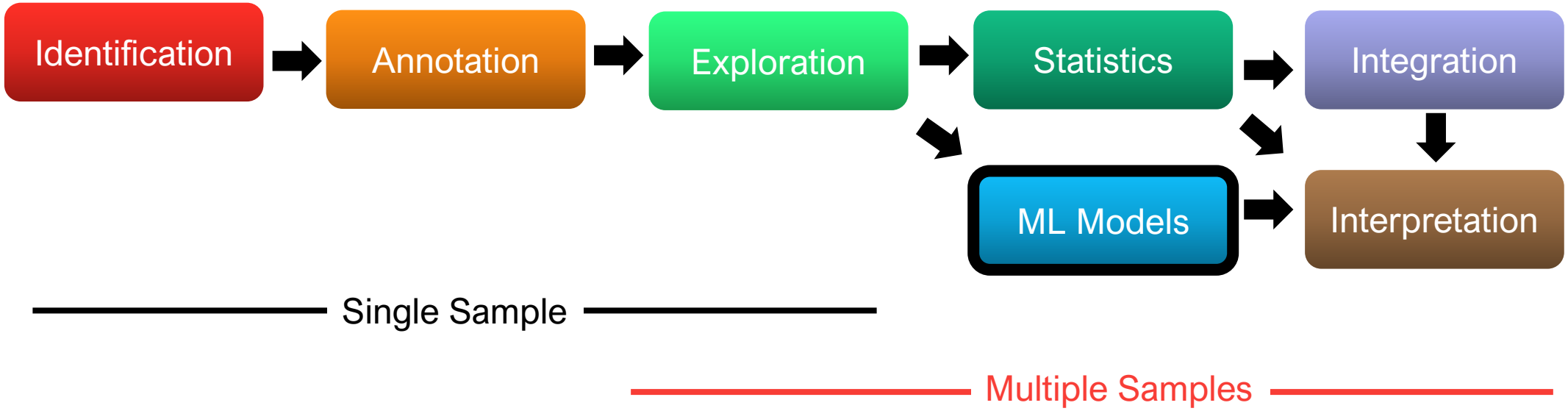
Tool Type: R Packages & Shiny Web Application



- **Omics:** Labeled & label-free digested and intact proteomics, GC/LC-MS and NMR, metabolomics, lipidomics, transcriptomics
- **Description:** A GUI for integrating several omics datatypes together.
- **Status:** Completed. Publication underway.
- **Development team:** MAP portal team, headed by Lisa Bramer

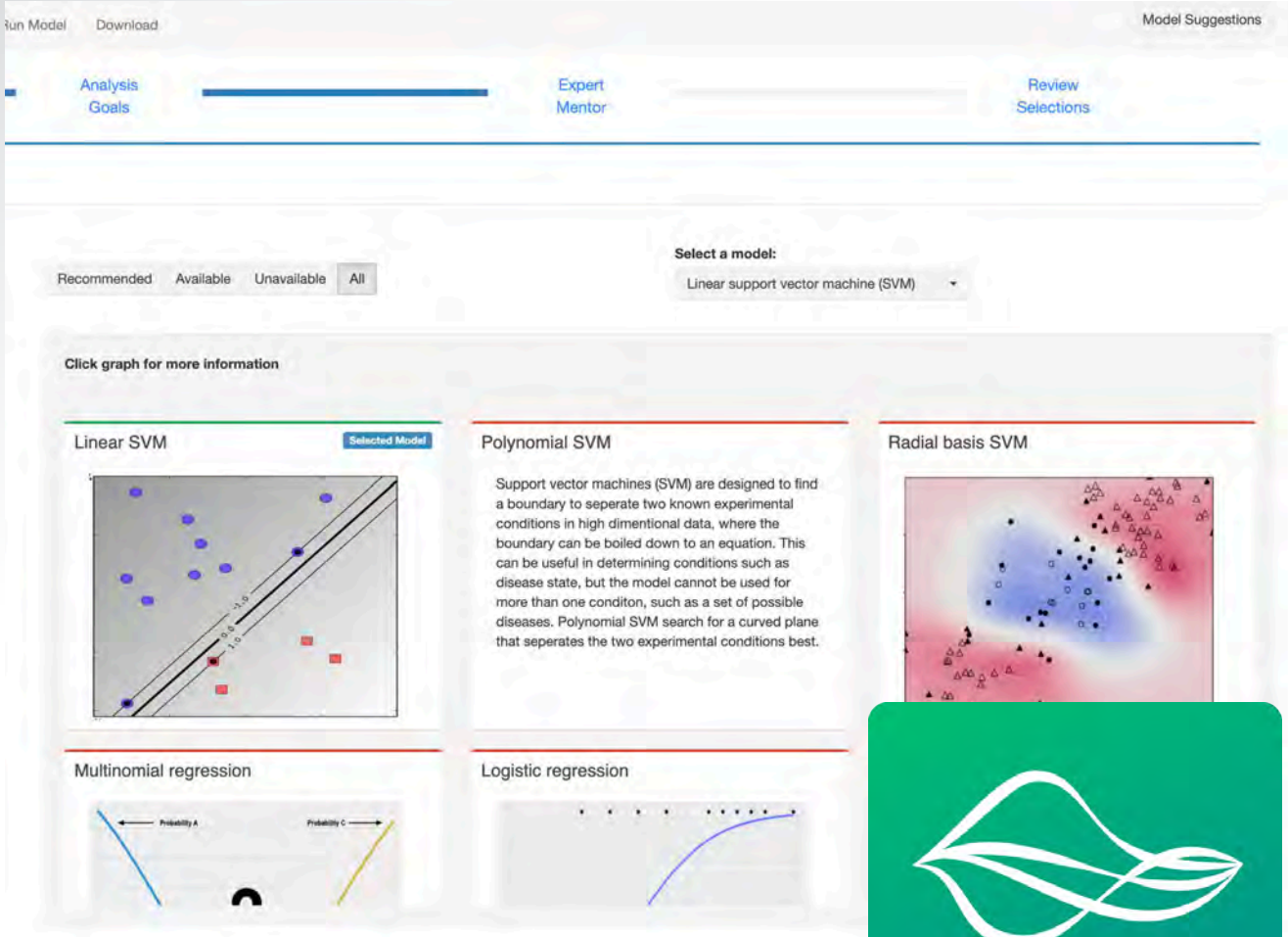


ML Tools

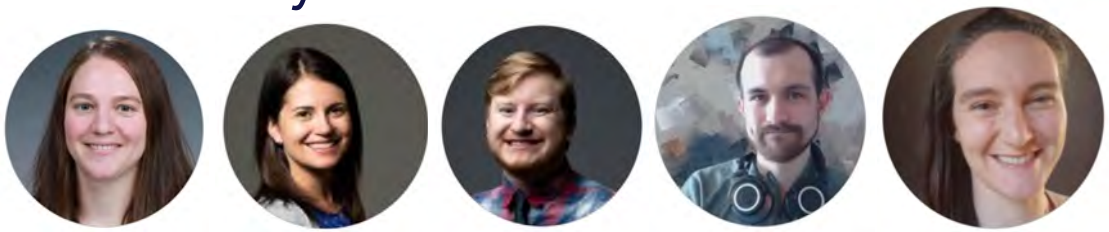

Identification
Annotation
Exploration
Statistics
Integration
ML Models
Interpretation
Single Sample
Multiple Samples

Many Omics ML: SLOPE

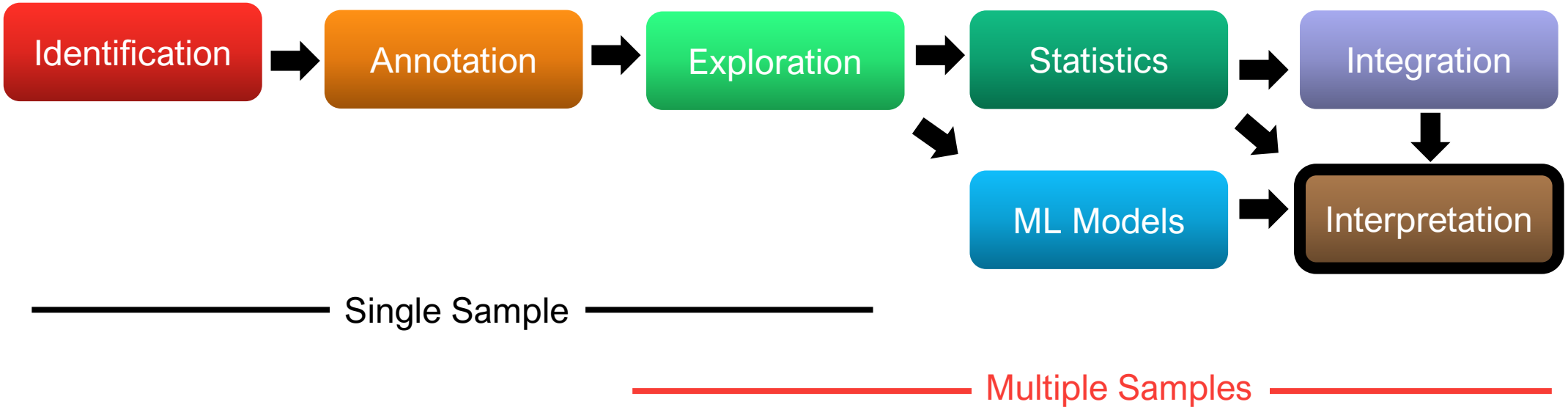
Tool Type: R Package & Shiny Web Application



- **Omics:** Any bulk omics!
- **Description:** An R package and GUI for running statistical machine learning methods on omics data, including response prediction, variable selection, and unsupervised clustering methods.
- **Status:** In development
- **Development team:** MAP portal team, headed by Lisa Bramer

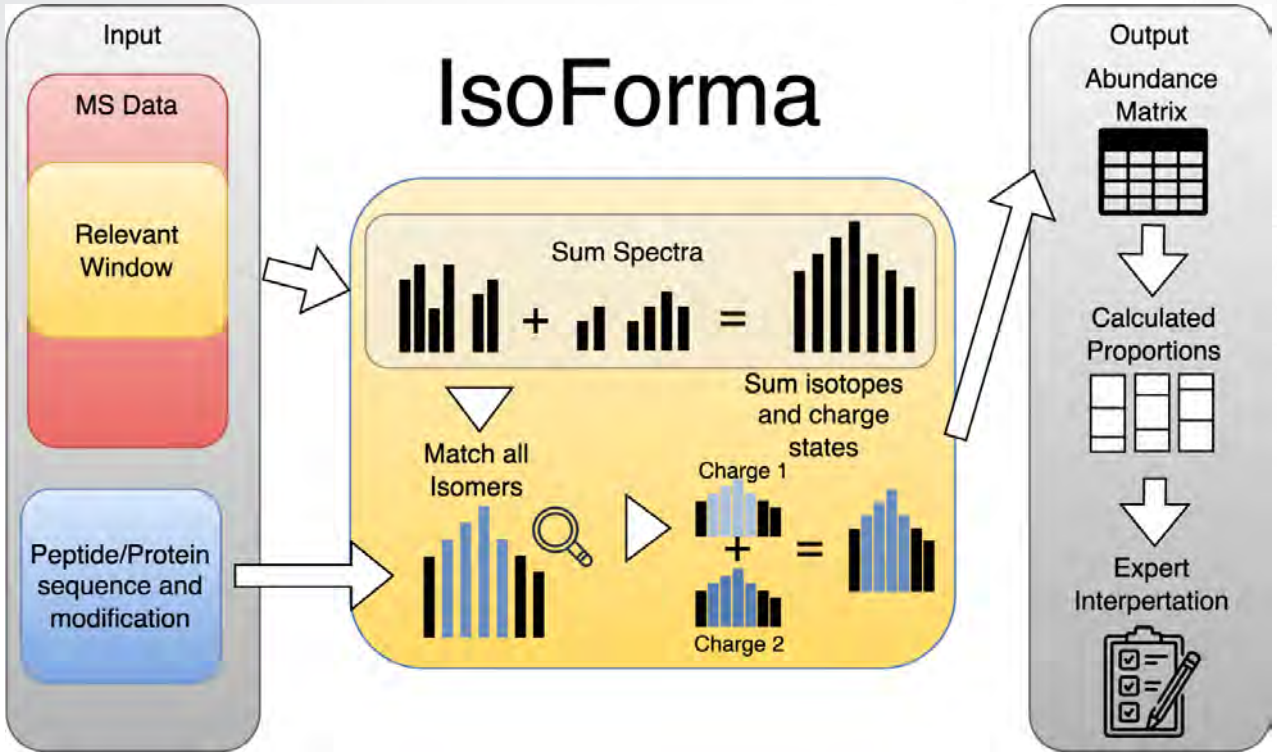


Interpretation Tools



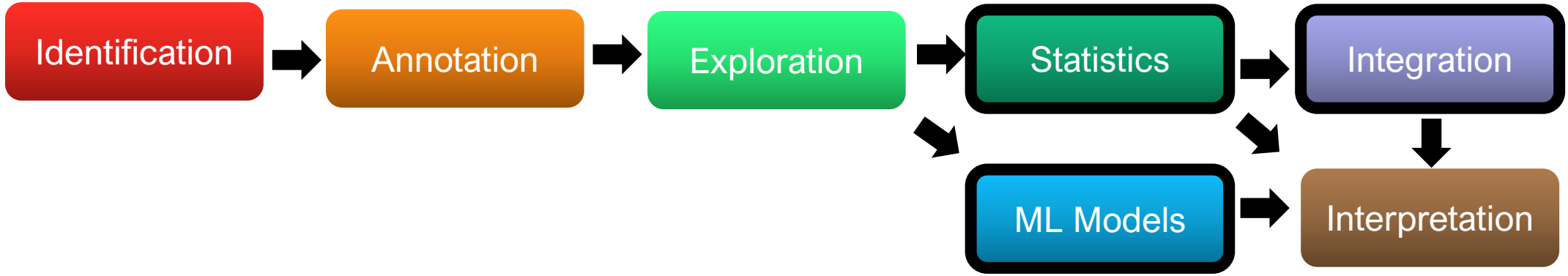
Proteomics Interpretation: IsoForma

Tool Type: R Package



- **Omics:** Labeled or label-free intact proteomics.
- **Description:** An R package for quantifying positional isomers in relationships to each other. Helpful for distinguishing epigenetic regulation, etc.
- **Status:** Completed. Publication underway.
- **Development team:** David Degnan, Mowei Zhou, Aivett Bilbao, Logan Lewis

A note on portals



So at this point, you may be thinking “that was a nice talk, but it seems like too much to track” which to that I’d respond:

1) Engage bioinformaticians, statisticians, and data scientists to be your tour guides.
2) We are working on portal-based applications, like the Multiomics Analysis Portal (MAP, relevant steps bolded above), to *guide* users through analyses, seamlessly port data between applications, and generate analytic reports.

Support your computational teams and projects!

This concludes our tour

- We covered one of the many developing omics pipelines – bulk omics. There are many tools and technologies being developed for new types of omics and MS technologies.
- For clarity, we presented these tools as pieces of a pipeline, but tool selection will vary depending on the analysis and goals.
- There are many more tools that could fit into this pipeline. The only way people know to use an analysis or tool is to talk about it! Take advantage of the contact information here and *always use GitHub issues pages.*

Special thanks to Kelly Stratton and Lisa Bramer for all their research in preparing this topic.

Thank you!

Summer School will resume tomorrow at 8:30 a.m. PDT

Afternoon Session

1:15-1:35 p.m.	Getting started with R	Natalie Winans
1:35-2:20	Data Wrangling	Luke Durell
2:20-2:30	Break	
2:30-3:00	Functions and Functionals	Luke Durell
3:00-3:45	Visualization	Natalie Winans
3:45-4:00	Statistical Modeling	Luke Durell