

# **Concepts of metabolomics**

- Dr. Cheng Zhang
- Senior Researcher
- Royal Institute of Technology, KTH

# Learning Objectives

***You will learn about***

**Metabolomics (targeted and untargeted)**

**Analytical technologies**

**Metabolomics applications & limits**

**Main processing pipelines**

**Tips, pitfalls and traps**

# Outline

**What is metabolomics?**

**Experimental Design**

**Data Generation / Analytical technologies**

**Applications and Limits**

**Preprocessing**

**Data cleaning / analysis**

# Outline

**What is metabolomics?**

Experimental Design

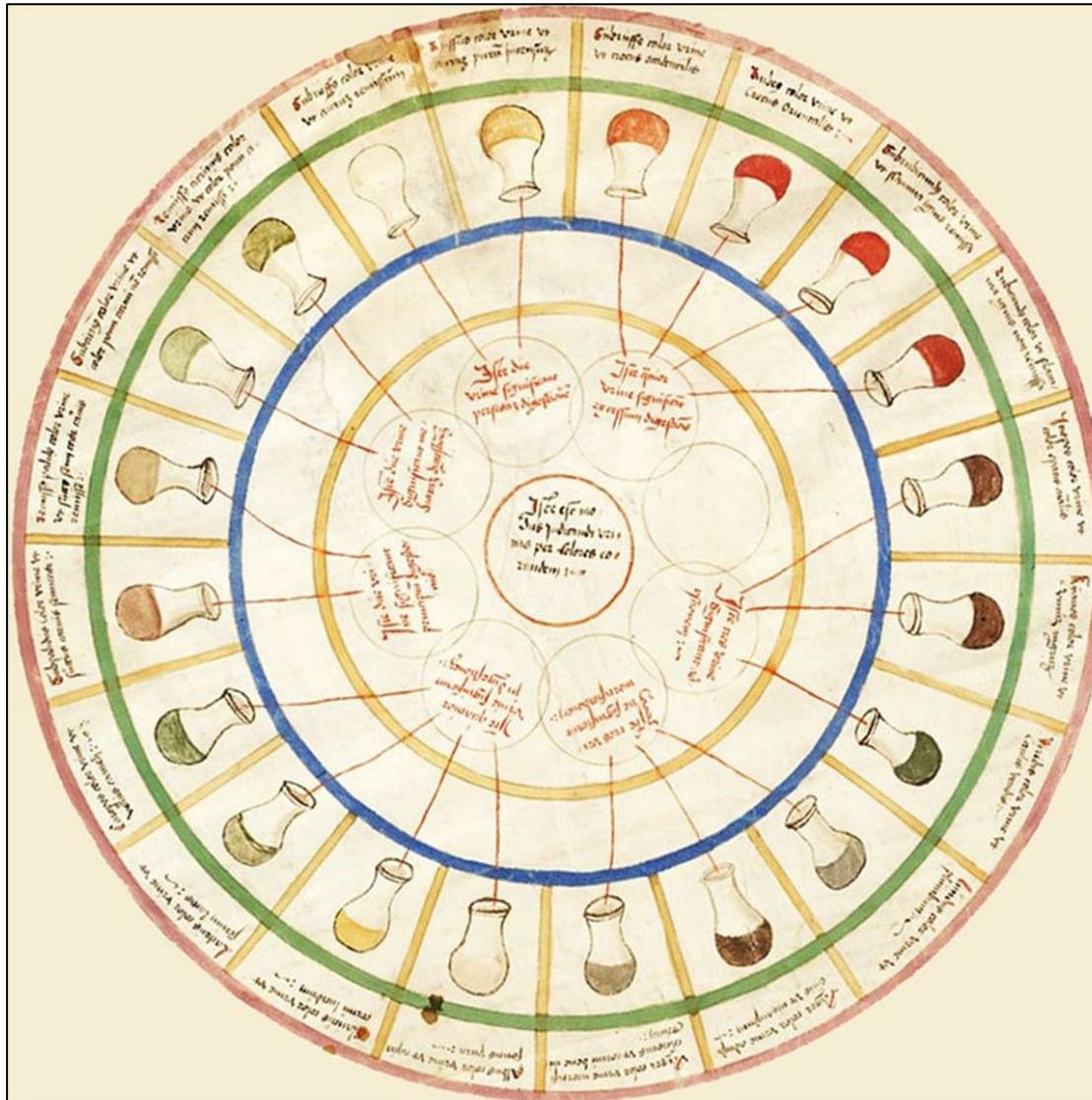
Data Generation / Analytical technologies

Applications and Limits

Preprocessing

Data cleaning / analysis

# The urine wheel



# Metabolomics

*What can happen*

Genomics



DNA

*What appears to be happening*

Transcriptomics



RNA

*What makes it happen*

Proteomics



Proteins

*What has happened and is happening*

Metabolomics



Metabolites

Amino acids, sugars, nucleotides  
lipids (Lipidome)

# What is a metabolite?

## Metabolites:

Small molecules (<1500 Da)

Ultimate support of the biological information

Includes human & microbial products

Endogenous metabolites: produced by the host organism

Exogenous metabolites: not produced by the host organism

## Metabolome:

refers to the complete set of metabolites in a biological sample

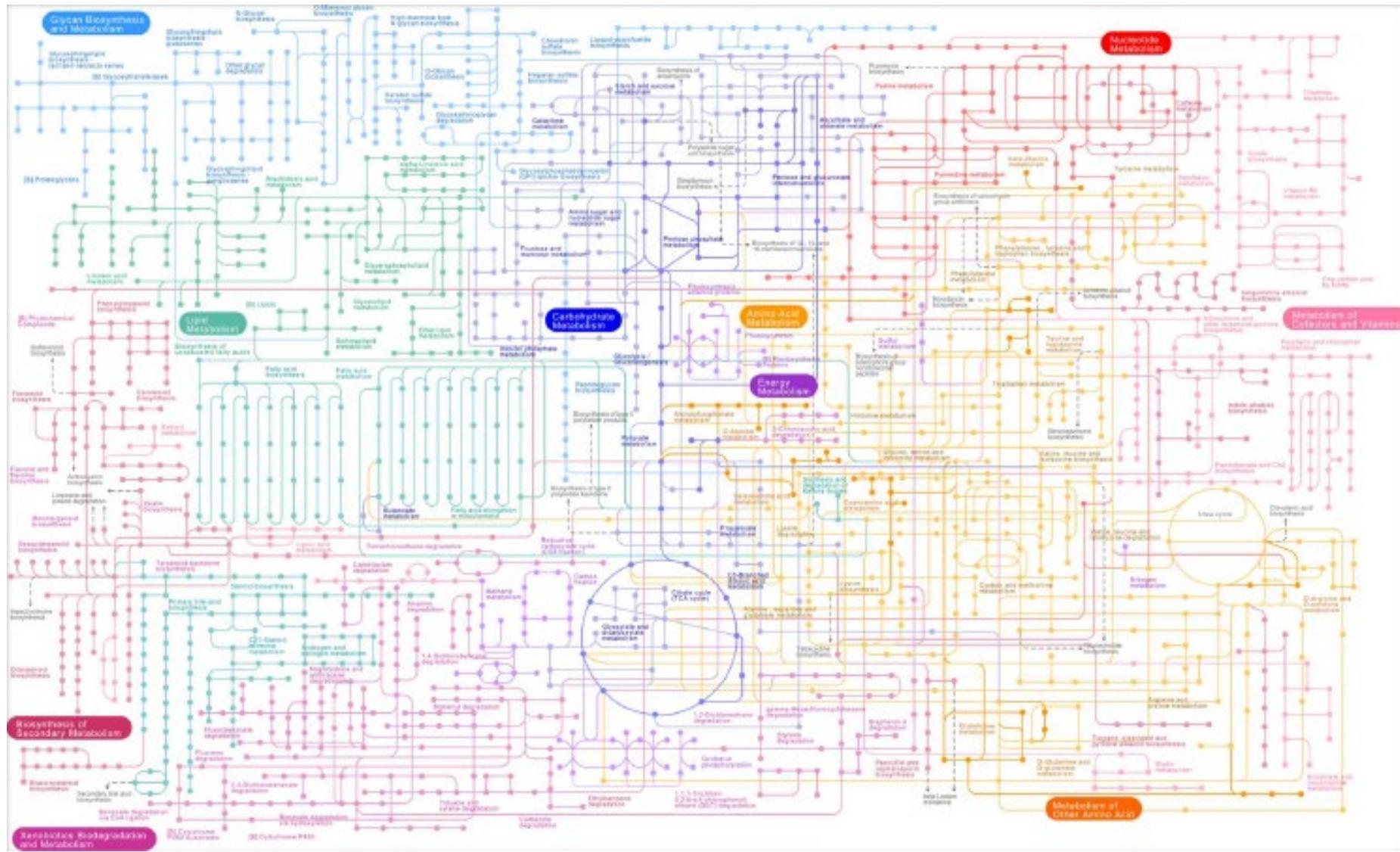
## MetaboLomics / MetaboNomics:

→ Metabolic Profiling / Metabotyping

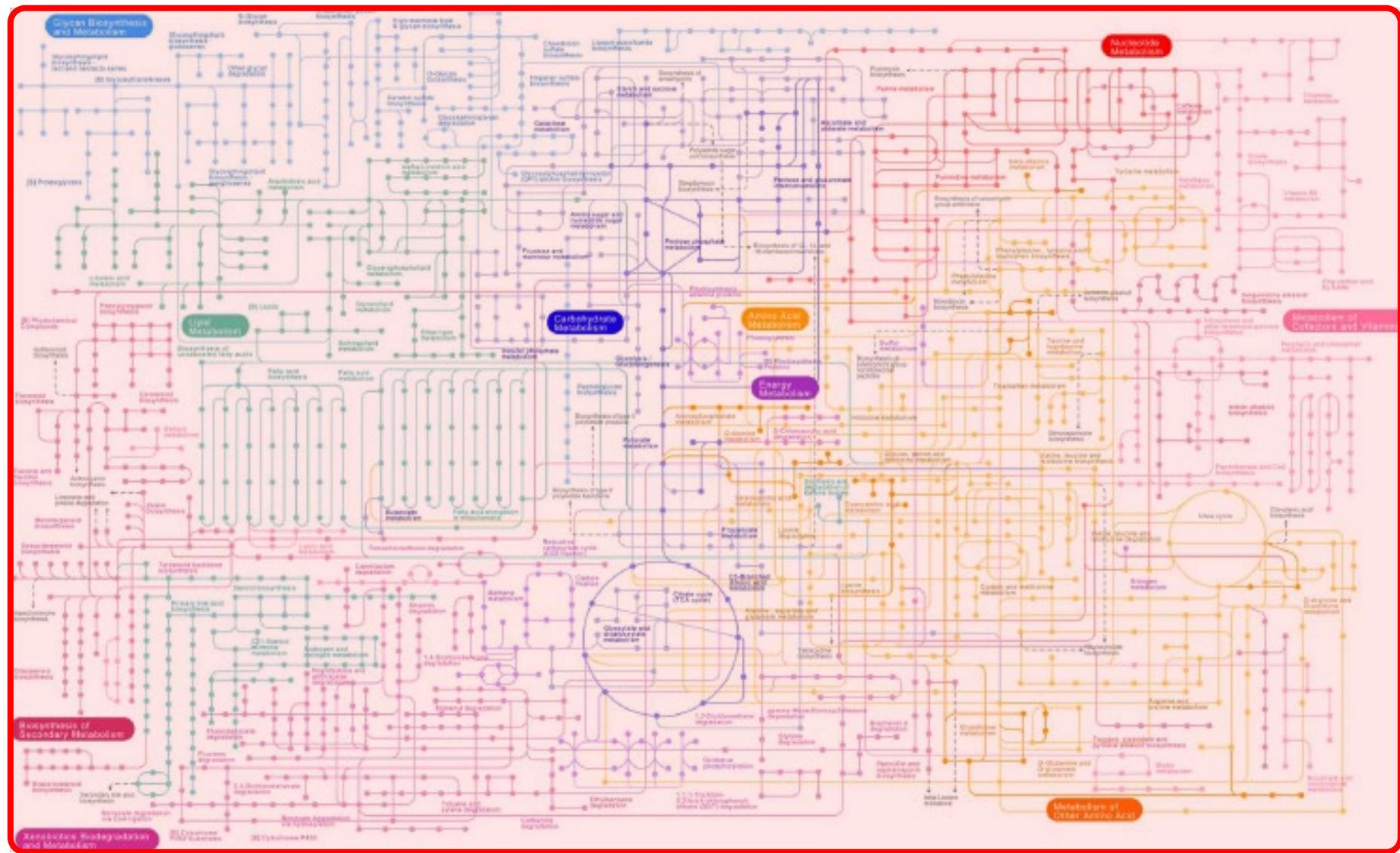
“the quantitative measurement of the metabolic **responses** of **complex systems** to a pathophysiological **stimulus** or genetic modification”.

(Nicholson, J. K., et al 1999, Xenobiotica, 29, 1181-89.)

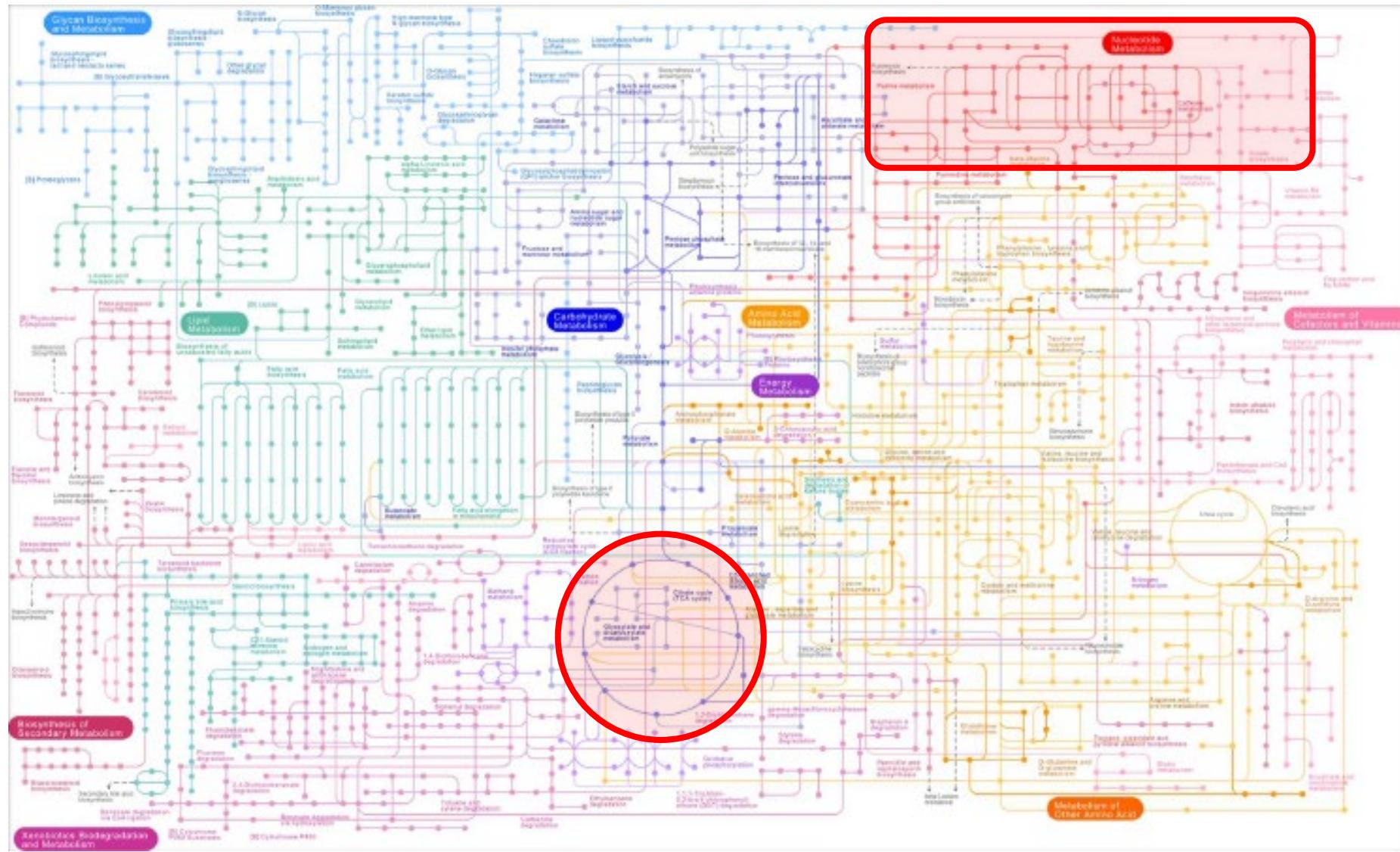
# Metabolomics and metabolism



# Untargeted metabolomics



# Targeted metabolomics





# Outline

**What is metabolomics?**

**Experimental Design**

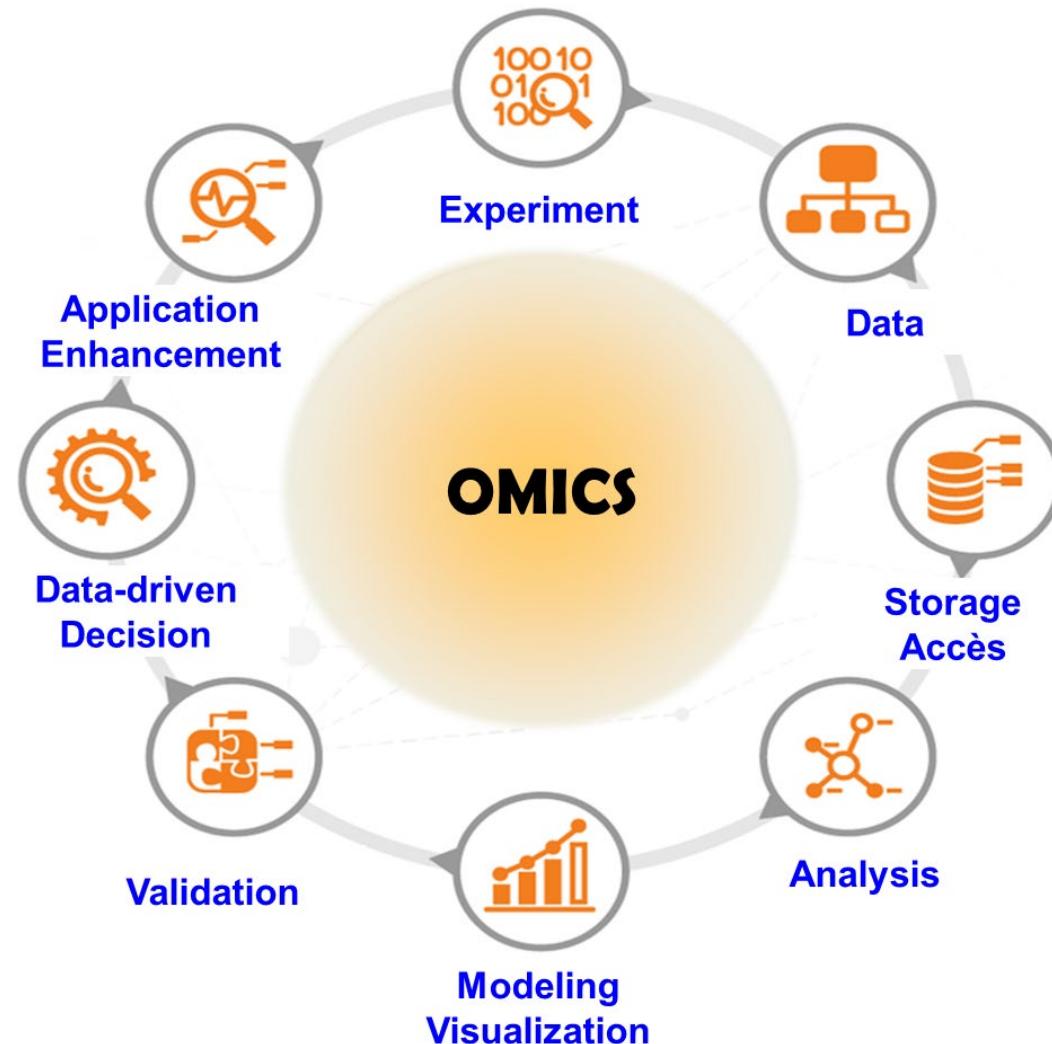
Data Generation / Analytical technologies

Applications and Limits

Preprocessing

Data cleaning / analysis

# Biological Information Generation



# Experimental Design

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.”

*Sir Ronald Fisher (1938)*





# Experimental Design **may help**

1. Clear and precise study objective
2. Sample type and size
3. Sampling and sample preparation strategy
4. Number of samples / Biological - Analytical replicates
5. Analytical technology(ies)
6. Collection of meta-data (Categorical, continuous, ordinal...)
7. Confounding factors
8. Randomization
9. Data analysis strategies (univariate vs. multivariate)
10. Biological Interpretation and insights
11. Validation (Biological/Analytical)
12. ...**Name it**

# Outline

**What is metabolomics?**

**Experimental Design**

**Data Generation / Analytical technologies**

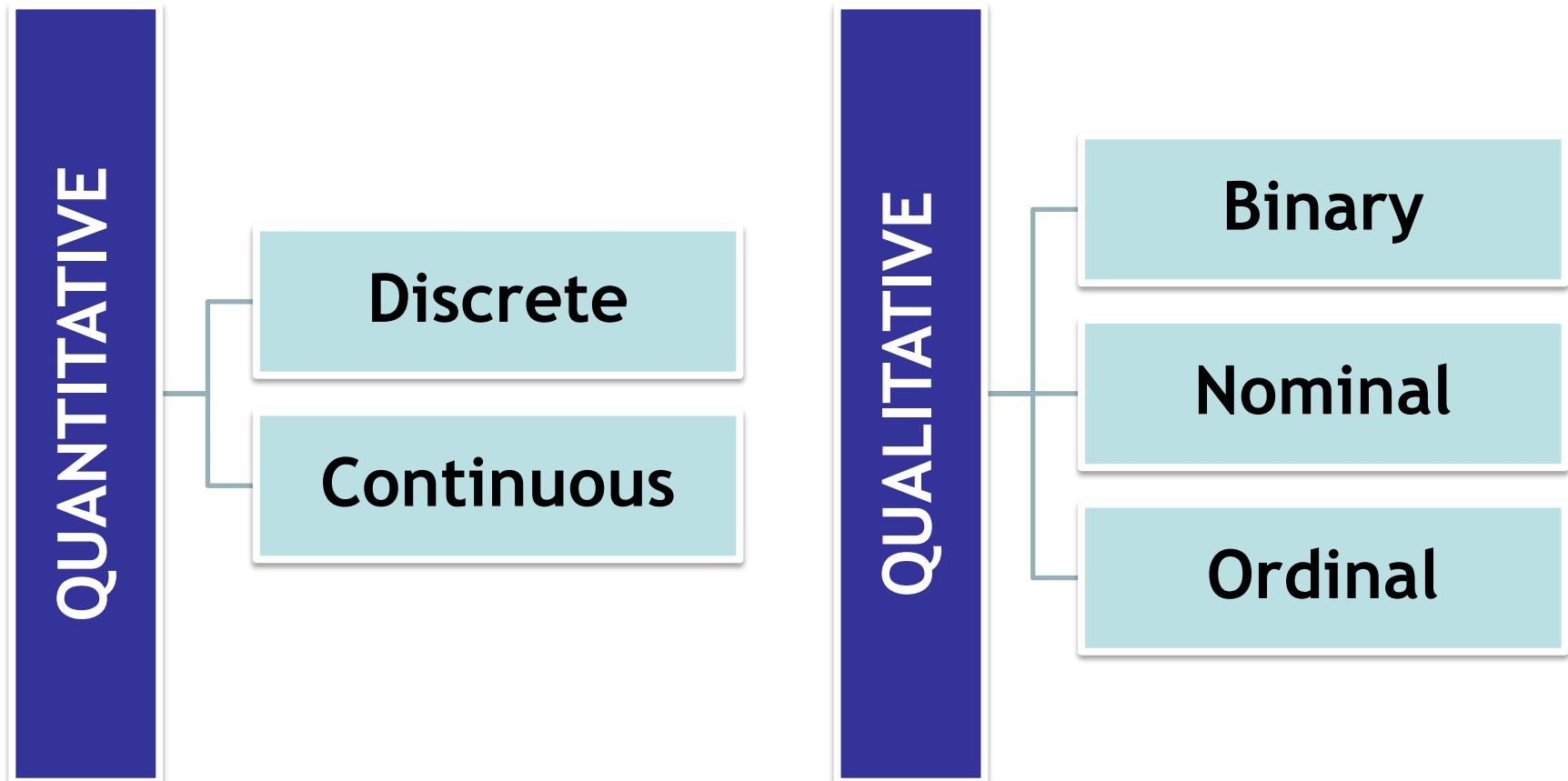
Applications and Limits

Preprocessing

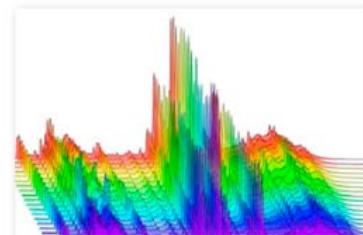
Data cleaning / analysis



# Know your data



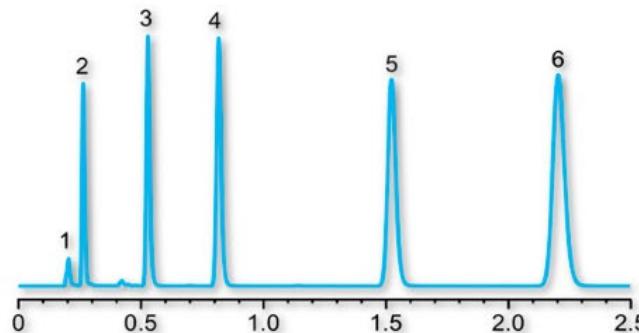
# Metabolomics



**Quantitative approach**

**Targeted approach**

(Set of Biomarkers Quantitation)

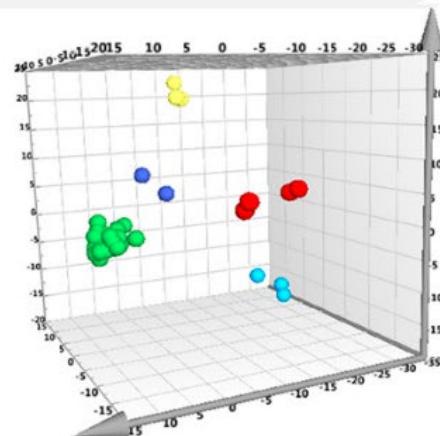


**Chemometric approach**

**Untargeted approach**

(Global Metabolic Fingerprints)

- Control
- Creatine deficiency
- Cystinuria
- Propionic aciduria
- Tyrosinemia



# Experimental Design

## BIOLOGICAL QUESTION (Experimental Design)

### Untargeted

### Targeted

#### SAMPLE PREPARATION

Global sample preparation  
Protein precipitation - Quenching  
Extraction

Metabolite specific preparation  
Enrichment - Extraction - Derivatization

#### DATA ACQUISITION

NMR  
HRMS (QToF, LTQ, FT-ICR)  
MS - MS/MS (DDA - DIA) - MS<sup>n</sup>

Targeted MS/MS  
(Triple Quadrupole - QToF)

#### DATA PREPROCESSING

Global  
Feature detection - Alignment  
Noise filtering - Deconvolution

Targeted  
Peak detection & integration  
(List of metabolites « m/z, tR, CCS »)

#### SCALING - TRANSFORMATION - NORMALIZATION

Scaling (e.g. Pareto)  
Transformation (e.g. Log)  
Normalization:  
Chemical: Internal standards - QC's  
Mathematical (e.g. PQN, LOESS, SVR)

Scaling (e.g. Pareto)  
Transformation (e.g. Log)  
Normalization: Internal standards

## BIOLOGICAL QUESTION (Experimental Design)

### Untargeted

### Targeted

#### QUANTIFICATION

Semi-Quantification

Absolute Quantification

#### DATA ANALYSIS

##### Univariate analysis

Parametric tests (e.g. t-tests, Pearson correlation)  
Nonparametric tests (e.g. Mann-Whitney U test, Spearman correlation)

##### Multivariate analysis

Unsupervised techniques: Clustering - Pattern recognition (e.g. Heat Maps, PCA, ICA)  
Supervised techniques: Predictive modeling - Regression - Classification (e.g. PLSDA, OPLSDA, IC-DA, SVM, RF)

#### METABOLITE IDENTIFICATION

Databases - Comparison with standards

No need - Already identified

#### FUNCTIONAL ANALYSIS

Over-Representation Analysis (ORA)  
Metabolite Set Enrichment Analysis (MSEA)  
Pathway & Network Analysis

# Metabolomics workflow

1

Question « Biology »

Experimental design

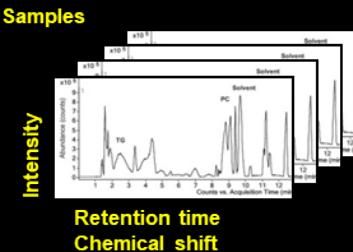
Sampling

Sample preparation

2

Instrumentation

MS / NMR

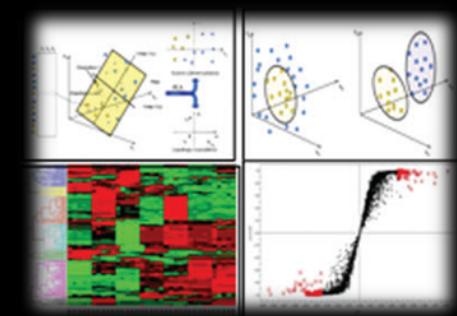
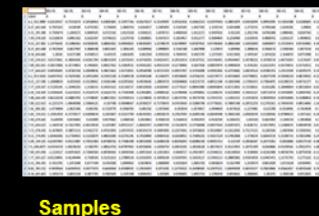


Data Acquisition

Data preprocessing

Data Analysis (Modeling)

(Filtration, detection, alignment, normalization)



3

Discriminant variables annotation

Metabolite Identification

Biological Insights

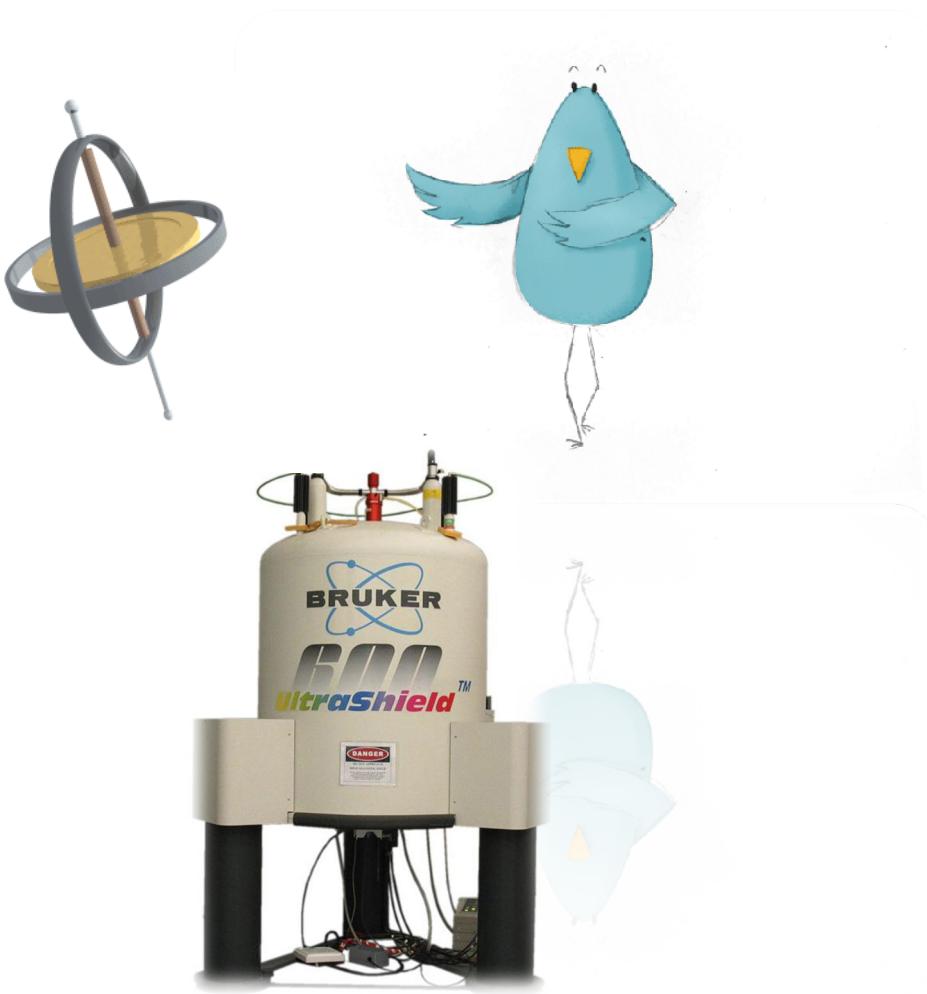
Putative identification

Comparison with standards

Pathway analysis  
Network analysis

Databases

# Biological Information Extraction



NMR spectroscopy



Mass spectrometry

# Biological Information Extraction

## NMR spectroscopy

Tissues, biofluids and extracts

Interaction of spin active nuclei ( $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{31}\text{P}$ ) with electromagnetic fields gives molecular information

Non-destructive  
Cross-instrument robustness

## Mass spectrometry

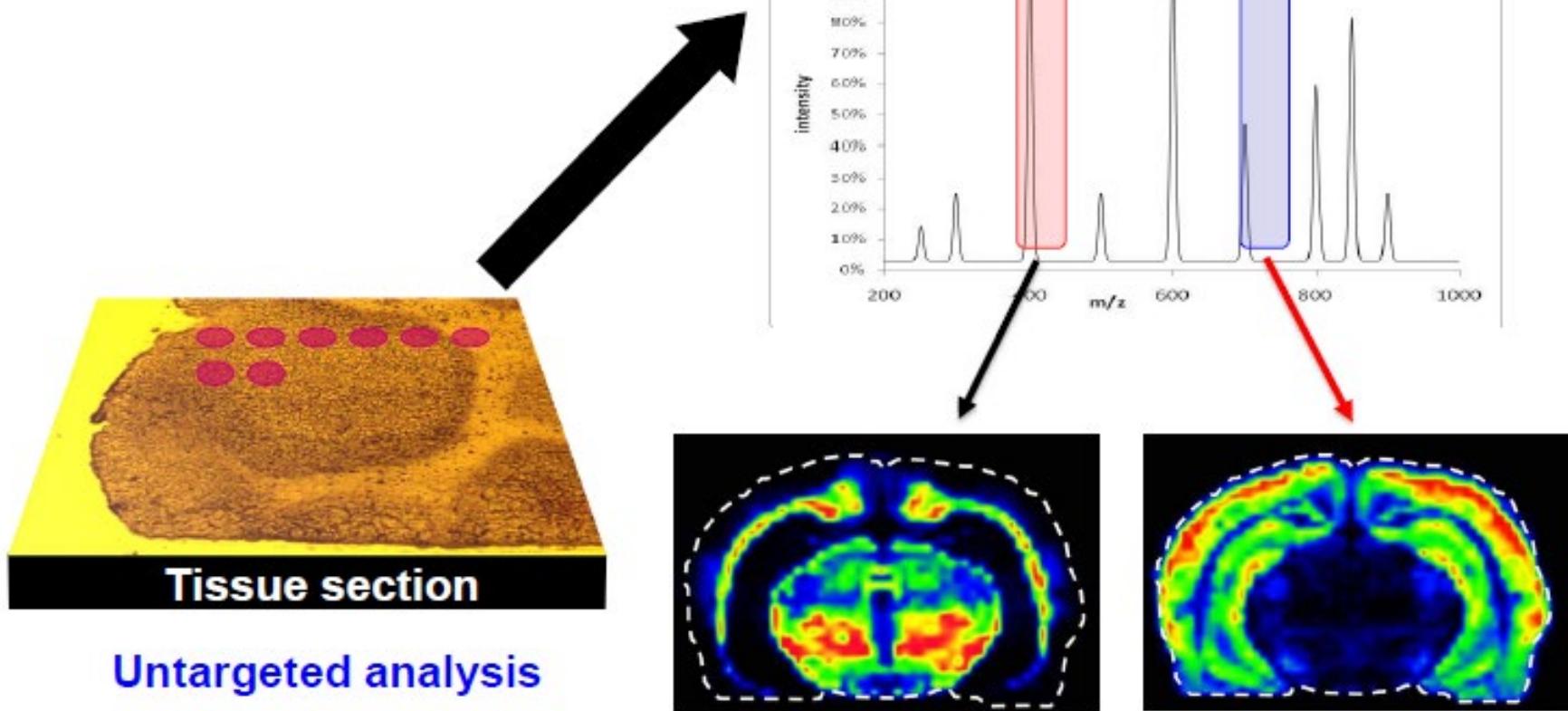
Tissues, biofluids and extracts

Mass to charge ratio ( $m/z$ )

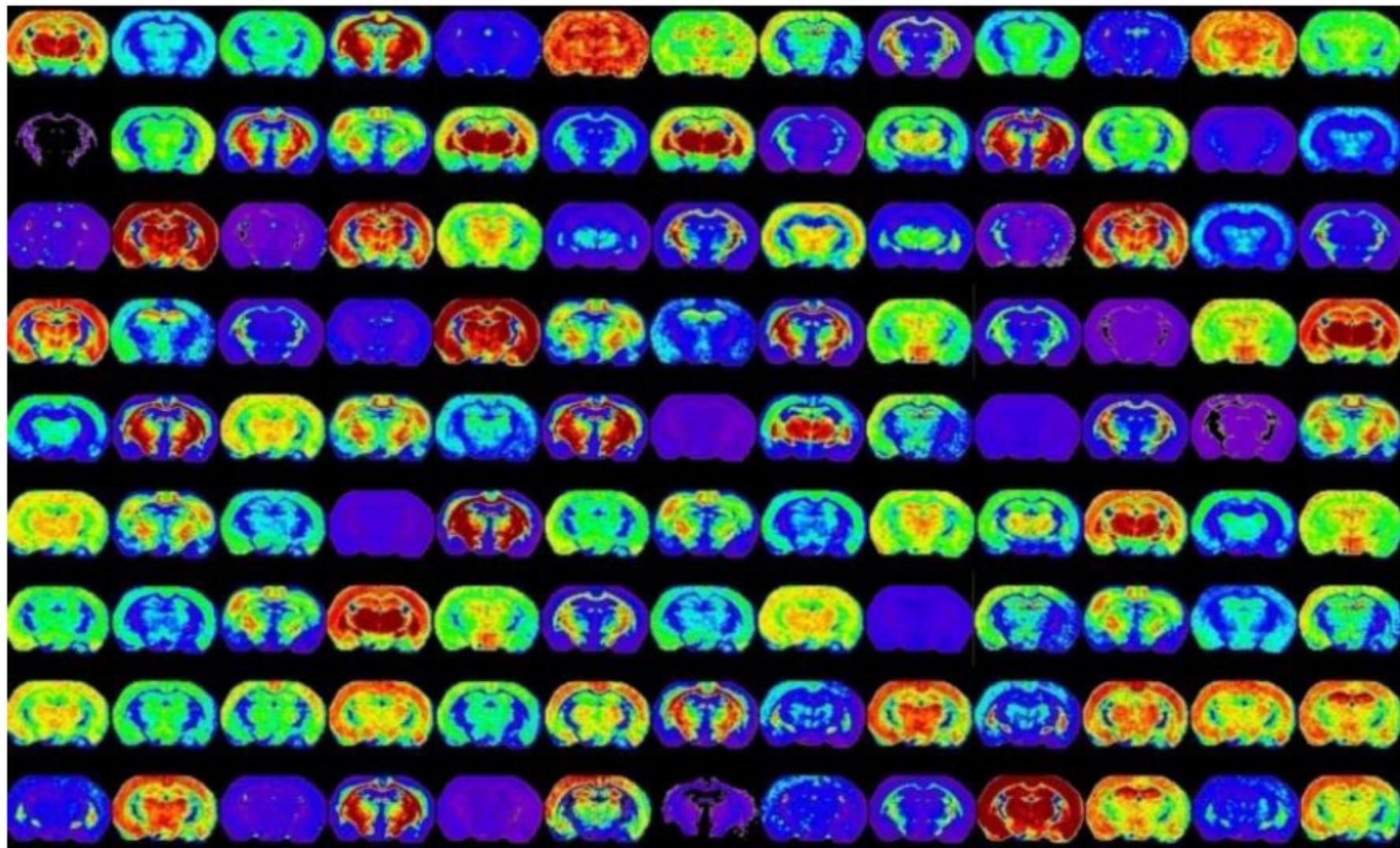
Sensitivity  
Higher metabolome coverage

# Metabolomics-based imaging

## Metabolomic Imaging



# Metabolomics-based imaging



Journal of Mass Spectrometry

Volume 46, Issue 2, pages 209-222, 24 JAN 2011 DOI: 10.1002/jms.1876  
<http://onlinelibrary.wiley.com/doi/10.1002/jms.1876/full#fig1>

# Outline

**What is metabolomics?**

**Experimental Design**

**Data Generation / Analytical technologies**

**Applications and Limits**

Preprocessing

Data cleaning / analysis

# Applications

## Metabolomics

### ENABLING TECHNOLOGIES

#### Bioinformatics / Machine Learning

Predictive Analytics  
Actionable visualization technologies  
Data integration - Network Analysis

#### Advanced Analytical Strategies

Nuclear Magnetic Resonance  
Mass spectrometry



### Next-Generation Diagnostics

Clinical Chemistry - Pathology - Precision Surgery - Microbiology

# Applications

## Metabolomics paths towards Precision Medicine



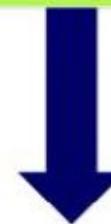
### Populational Profiling

Epidemiological stratification  
Disease-risk biomarker discovery  
Large-scale association studies  
Public health prevention



### Individual Profiling

Patient stratification  
Personalized therapies  
Pharmacometabonomics  
Nutritional assessment



### Drug Discovery

Proof of mechanism  
Proof of action  
Pharmacokinetics  
Pharmacodynamics

# Applications

Article



molecular  
systems  
biology

The gut microbiota modulates host amino acid and glutathione metabolism in mice

Adil Mardinoglu<sup>1,2,†,\*</sup>, Saeed Shoaei<sup>1,†</sup>, Mattias Bergentz<sup>1,4</sup>, Pouyan Ghaffari<sup>1</sup>, Cheng Zhang<sup>2</sup>, Erik Larsson<sup>3,4</sup>, Fredrik Bäckhed<sup>3,4</sup> & Jens Nielsen<sup>1,2</sup>

## Host-Microbiota interactions

*Int. J. Mol. Sci.* 2016, 17(7), 1167; doi:10.3390/ijms17071167

Review

**Clinical Metabolomics: The New Metabolic Window for Inborn Errors of Metabolism Investigations in the Post-Genomic Era**

Abdellah Tebani<sup>1,2,3</sup> □, Lenaig Abily-Donval<sup>2,4</sup> □, Carlos Alonso<sup>3</sup> □, Stéphane Marret<sup>2,4</sup> □ and Soumeya Bekri<sup>1,2,\*</sup> □

Open Access

## Inherited Metabolic Diseases

RESEARCH ARTICLE

CANCER DIAGNOSTICS

**Intraoperative Tissue Identification Using Rapid Evaporative Ionization Mass Spectrometry**

Júlia Balog,<sup>1\*</sup> László Sasi-Szabó,<sup>2\*</sup> James Kinross,<sup>3,4</sup> Matthew R. Lewis,<sup>3</sup> Laura J. Muirhead,<sup>3,4</sup> Kirill Veselkov,<sup>3</sup> Reza Mirnezami,<sup>4</sup> Balázs Dezső,<sup>5</sup> László Damjanovich,<sup>2</sup> Ara Darzi,<sup>4</sup> Jeremy K. Nicholson,<sup>3,†</sup> Zoltán Takáts<sup>3†</sup>

## Pathology and Cancer

**Pharmacometabolic Investigation of Dynamic Metabolic Phenotypes Associated with Variability in Response to Galactosamine Hepatotoxicity**

Muireann Coen,<sup>\*†</sup> Françoise Goldfain-Blanc,<sup>‡</sup> Gaëlle Rolland-Valognes,<sup>§</sup> Bernard Walther,<sup>||</sup> Donald G. Robertson,<sup>‡</sup> Elaine Holmes,<sup>†</sup> John C. Lindon,<sup>†</sup> and Jeremy K. Nicholson<sup>\*†</sup>

## Responder Non-responder Prediction

**An Integrative Approach for Identifying a Metabolic Phenotype Predictive of Individualized Pharmacokinetics of Tacrolimus**

PB Phapale<sup>1,2</sup>, S-D Kim<sup>2</sup>, HW Lee<sup>1,2</sup>, M Lim<sup>1,2</sup>, DD Kale<sup>1,2</sup>, Y-L Kim<sup>2,3</sup>, J-H Cho<sup>4</sup>, D Hwang<sup>4</sup> and Y-R Yoon<sup>1,2</sup>

## Human Drug Pharmacokinetics



ARTICLE

<https://doi.org/10.1038/s41467-019-10994-1> OPEN

**Assessing the causal association of glycine with risk of cardio-metabolic diseases**

Laura B.L. Wittemars<sup>1</sup>, Luca A. Lotta<sup>1</sup>, Clare Oliver-Williams<sup>2,3</sup>, Isobel D. Stewart<sup>1</sup>, Praveen Surendran<sup>2</sup>, Savita Karthikeyan<sup>2</sup>, Felix R. Day<sup>2</sup>, Albert Kouzman<sup>2,4</sup>, Fumiaki Imamura<sup>2</sup>, Lingyao Zeng<sup>5,6</sup>, Jeanette Erdmann<sup>7,8,9</sup>, Heribert Schunkert<sup>5,6</sup>, Kay-Tee Khaw<sup>10</sup>, Julian L. Griffin<sup>11</sup>, Nita G. Forouhi<sup>12</sup>, Robert A. Scott<sup>12</sup>, Angela M. Wood<sup>2</sup>, Stephen Burgess<sup>2,12</sup>, Joanna M.M. Howson<sup>2</sup>, John Danesh<sup>2,13</sup>, Nicholas J. Wareham<sup>1</sup>, Adam S. Butterworth<sup>2</sup> & Claudia Langenberg<sup>1</sup>

## Large-scale epidemiological profiling

# Metabolomics-based imaging

**ScienceDaily®**  
Your source for the latest research news

**News      Articles      Videos      Images      Books**

[Health & Medicine](#)   [Mind & Brain](#)   [Plants & Animals](#)   [Earth & Climate](#)   [Space & Tech](#)

**Science News**   ... from universities, journals, and other resea

**'Intelligent Knife' Tells Surgeon Which Tissue Is Cancerous**

July 17, 2013 — Scientists have developed an "intelligent knife" that can tell surgeons immediately whether the tissue they are cutting is cancerous or

**RESEARCH ARTICLE**

**CANCER DIAGNOSTICS**

**Intraoperative Tissue Identification Using Rapid Evaporative Ionization Mass Spectrometry**

Júlia Balog,<sup>1\*</sup> László Sasi-Szabó,<sup>2\*</sup> James Kinross,<sup>3,4</sup> Matthew R. Lewis,<sup>3</sup> Laura J. Muirhead,<sup>3,4</sup> Kirill Veselkov,<sup>3</sup> Reza Mirnezami,<sup>4</sup> Balázs Dezső,<sup>5</sup> László Damjanovich,<sup>2</sup> Ara Darzi,<sup>4</sup> Jeremy K. Nicholson,<sup>3†</sup> Zoltán Takáts<sup>3†</sup>

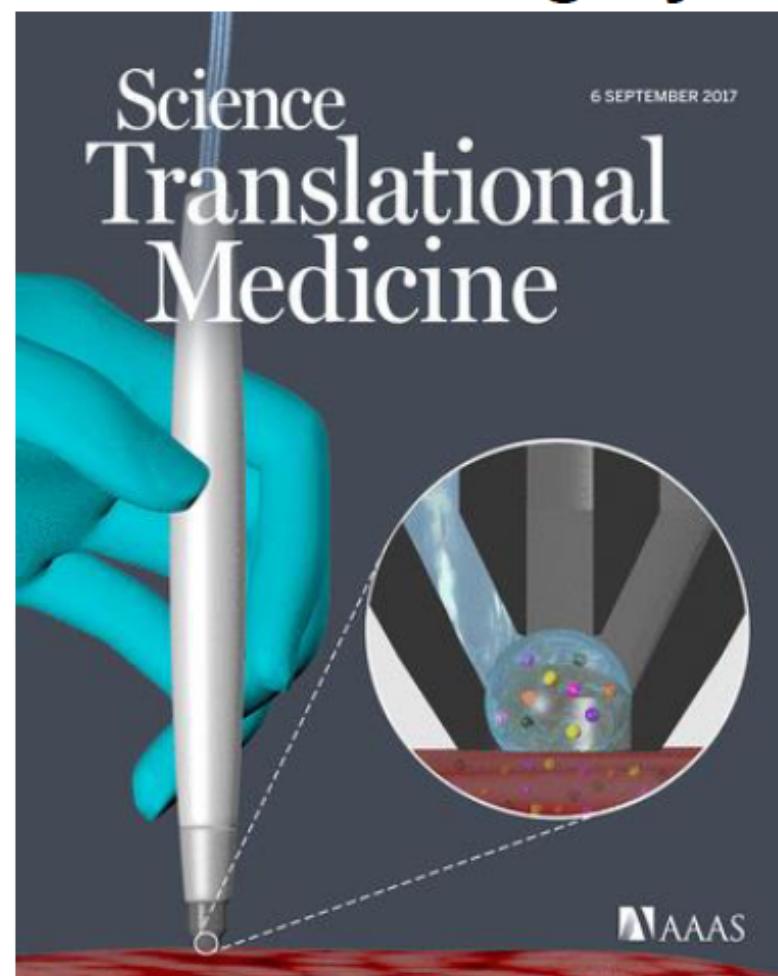
**SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE**

**CANCER DIAGNOSTICS**

**Nondestructive tissue analysis for ex vivo and in vivo cancer diagnosis using a handheld mass spectrometry system**

Jialing Zhang,<sup>1</sup> John Rector,<sup>1,2</sup> John Q. Lin,<sup>1</sup> Jonathan H. Young,<sup>1</sup> Marta Sans,<sup>1</sup> Nitesh Katta,<sup>2</sup> Noah Giese,<sup>1</sup> Wendong Yu,<sup>3</sup> Chandandeep Nagi,<sup>3</sup> James Suliburk,<sup>4</sup> Jinsong Liu,<sup>5</sup> Alena Bensussan,<sup>1</sup> Rachel J. DeHoog,<sup>1</sup> Kyana Y. Garza,<sup>1</sup> Benjamin Ludolph,<sup>1</sup> Anna G. Sorace,<sup>6</sup> Anum Syed,<sup>2</sup> Aydin Zahedivash,<sup>2</sup> Thomas E. Milner,<sup>2</sup> Livia S. Eberlin<sup>1\*</sup>

## Real time Metabolomics: Precision Surgery

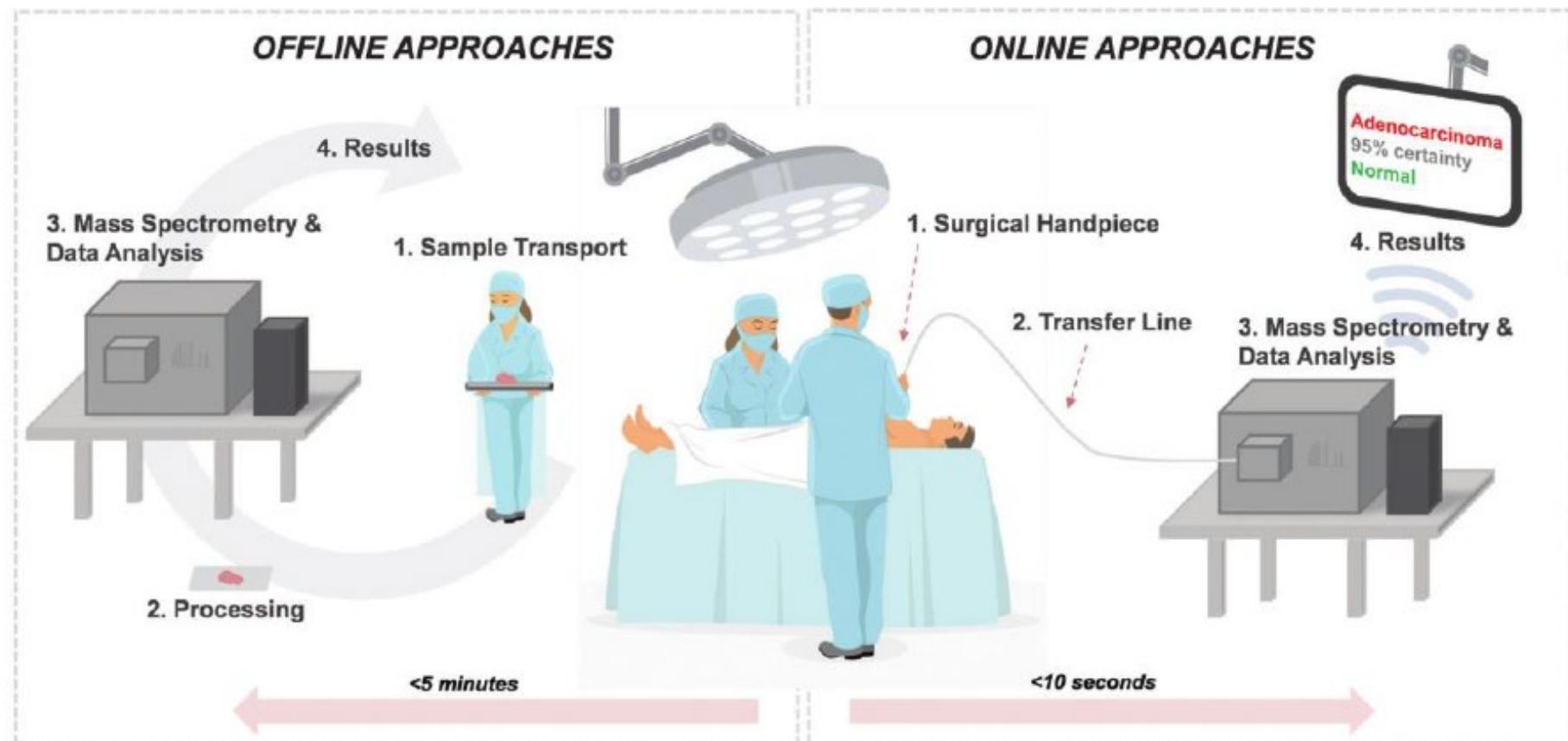


# Metabolomics-based imaging



# Metabolomics-based imaging

## Real time Metabolomics: Precision Surgery



Demian R. Ifa et al. Clin Chem 62:1 (2016)

Sci Transl Med 17 July 2013: Vol. 5, Issue 194, p. 194ra93

# Limits

**Metabolite Identification** are the main bottlenecks of metabolomics for large adoption in both translational and clinical context.

Lack of standardized **annotation** of the metabolome is important for functional analysis and integration with other omics through GEMs

More **absolute quantification** of metabolites is needed (targeted and untargeted) to achieve reliability and robustness

**Standardization and Harmonization** is a prerequisite for large adoption

**Miniaturization** will enhance high-throughput

**Automation, Data Visualization and Clinical Actionability** at different stages, instrument-, pre- and post-analytic levels including data processing, integration and interpretation are very important issues for large clinical adoption of any diagnostic innovation

# Outline

**What is metabolomics?**

**Experimental Design**

**Data Generation / Analytical technologies**

**Applications and Limits**

**Preprocessing**

**Data cleaning / analysis**

# Data preprocessing

Signal



Data

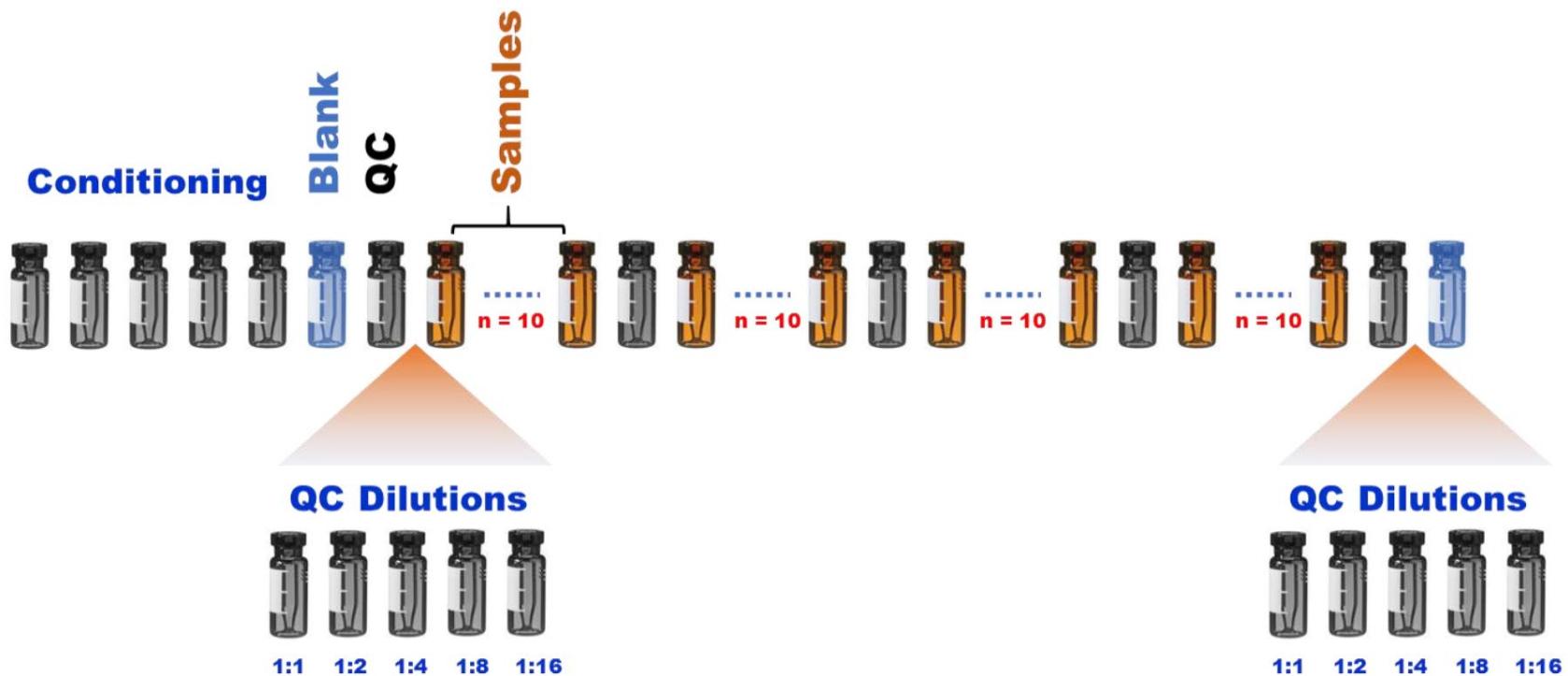


Samples

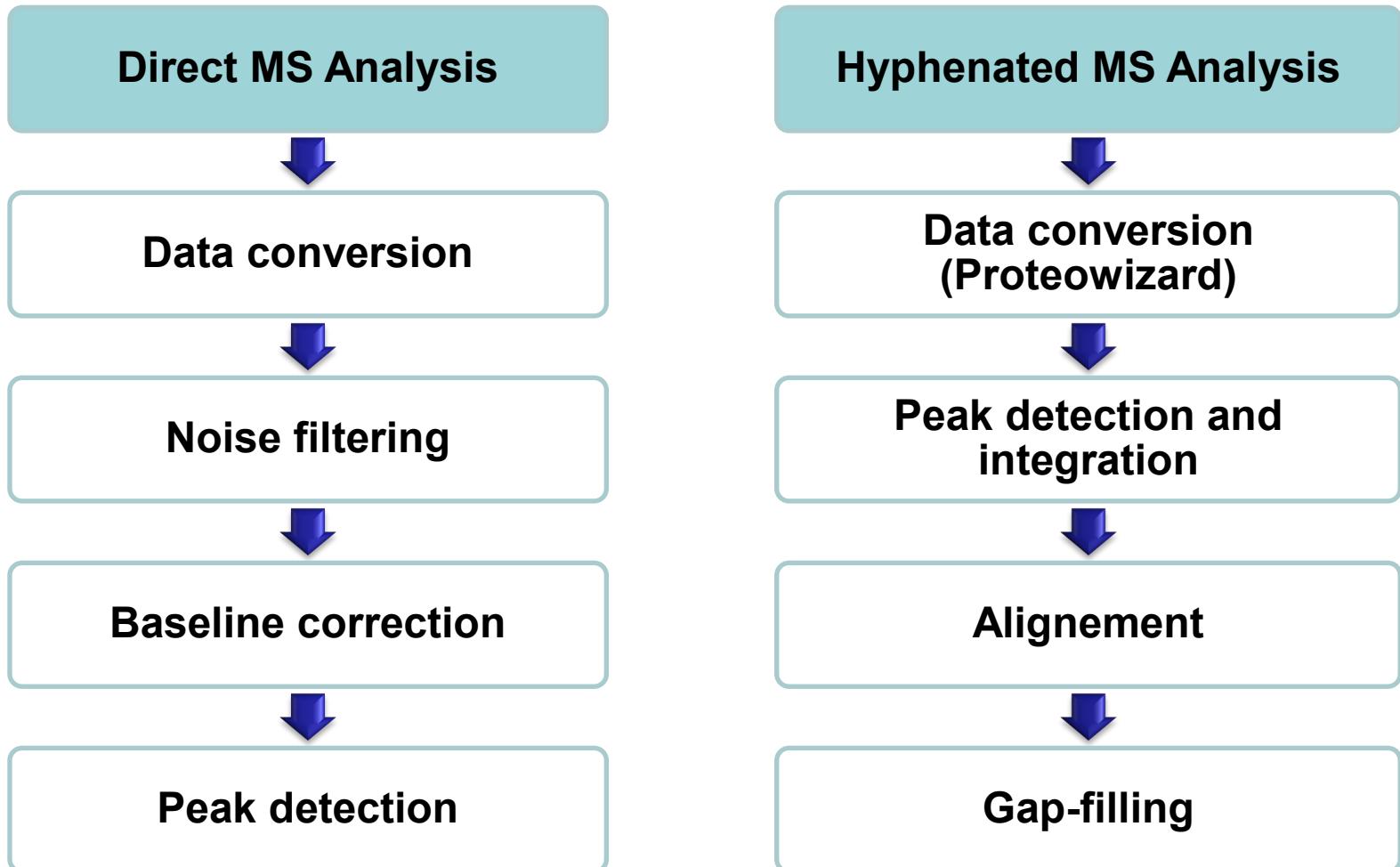
Variables

6	0	4	8	0	2	2	7	2	2
0	1	8	0	8	0	0	3	4	9
4	4	0	4	0	1	4	0	1	8
8	0	2	8	2	4	8	4	0	0
0	4	6	0	5	2	0	8	0	2
2	8	9	5	5	5	2	0	4	6
4	9	5	6	5	4	5	2	8	9
1	8	0	8	5	0	1	8	0	8
4	0	4	0	0	4	2	0	4	0
0	2	8	2	5	8	0	2	8	2

# Data preprocessing



# Data processing



# Data processing softwares

**Table 1** Software tools commonly used for the preprocessing of metabolomics data

Tool	Instrument data type	Software type	Website	References
XCMS	LC–MS, GC–MS	R Package	<a href="http://bioconductor.org/packages/release/bioc/html/xcms.html">http://bioconductor.org/packages/release/bioc/html/xcms.html</a>	Smith et al. (2006)
OpenMS—FeatureFinderMetabo	LC–MS	GUI	<a href="http://ftp.mi.fu-berlin.de/pub/OpenMS/release-documentation/html/TOPP_FeatureFinderMetabo.html">http://ftp.mi.fu-berlin.de/pub/OpenMS/release-documentation/html/TOPP_FeatureFinderMetabo.html</a>	Bertsch et al. (2010)
MetAlign	LC–MS	Windows GUI	<a href="http://www.wageningenur.nl/en/show/MetAlign-1.htm">http://www.wageningenur.nl/en/show/MetAlign-1.htm</a>	Lommen & Kools (2012)
MS-DIAL	LC–MS	Windows GUI	<a href="http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/index.html">http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/index.html</a>	Tsugawa et al. (2015)
mzMatch	LC–MS	R Package	<a href="http://mzmatch.sourceforge.net/index.php">http://mzmatch.sourceforge.net/index.php</a>	Scheltema et al. (2011)
IDEOM	LC–MS	Excel Template	<a href="http://mzmatch.sourceforge.net/ideom.php">http://mzmatch.sourceforge.net/ideom.php</a>	Creek et al. (2012)
AMDIS	GC–MS	Windows GUI	<a href="http://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:amdis">http://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:amdis</a>	Meyer et al. (2010)
MetaboliteDetector	GC–MS	CLI, GUI	<a href="http://md.tu-bs.de">http://md.tu-bs.de</a>	Hiller et al. (2009)
MET-IDEA	GC–MS	Windows CLI	<a href="http://bioinfo.noble.org/download">http://bioinfo.noble.org/download</a>	Broeckling et al. (2006)
MeltDB	LC–MS, GC–MS	Web App	<a href="https://meltdb.cebitc.uni-bielefeld.de/cgi-bin/login.cgi">https://meltdb.cebitc.uni-bielefeld.de/cgi-bin/login.cgi</a>	Kessler et al. (2013)
metaMS	GC–MS	R Package	<a href="http://bioconductor.org/packages/release/bioc/html/metaMS.html">http://bioconductor.org/packages/release/bioc/html/metaMS.html</a>	Wehrens et al. (2014)
MSeasy	GC–MS	R Package	<a href="https://cran.r-project.org/web/packages/MSeasy/index.html">https://cran.r-project.org/web/packages/MSeasy/index.html</a>	Nicolè et al. (2012)
SpectConnect	GC–MS	Web App	<a href="http://spectconnect.mit.edu">http://spectconnect.mit.edu</a>	Styczynski et al. (2007)
rNMR	NMR	R Package	<a href="http://rnmr.nmrfa.mw.vanderbilt.edu">http://rnmr.nmrfa.mw.vanderbilt.edu</a>	Lewis et al. (2009)

CLI command line interface, GUI graphical user interface

# Data processing softwares

Electrophoresis 2019, 40, 227–246

227

Biswapiya B. Misra<sup>1</sup>   
Subhashree Mohapatra<sup>2</sup>

<sup>1</sup>Department of Internal Medicine, Section of Molecular Medicine, Medical Center Boulevard, Winston-Salem, NC, USA

<sup>2</sup>Independent Researcher, 151 Edgeway Drive, Winston-Salem, NC, USA

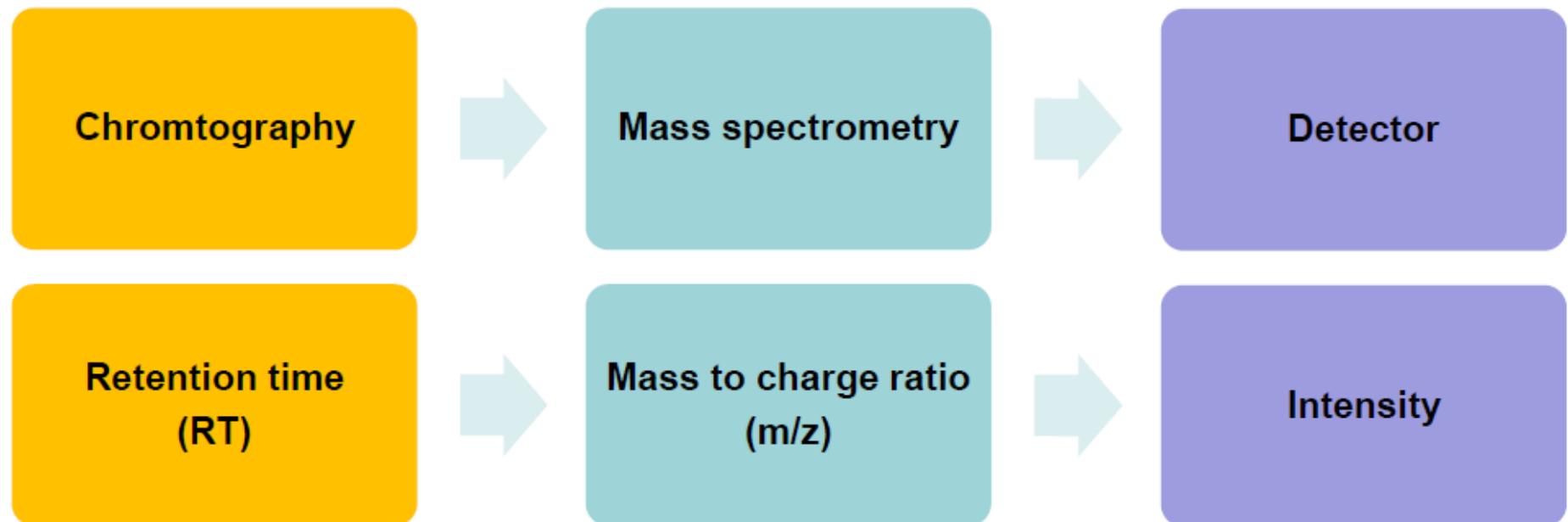
Received October 11, 2018  
Revised November 9, 2018  
Accepted November 9, 2018

## Review

### Tools and resources for metabolomics research community: A 2017–2018 update

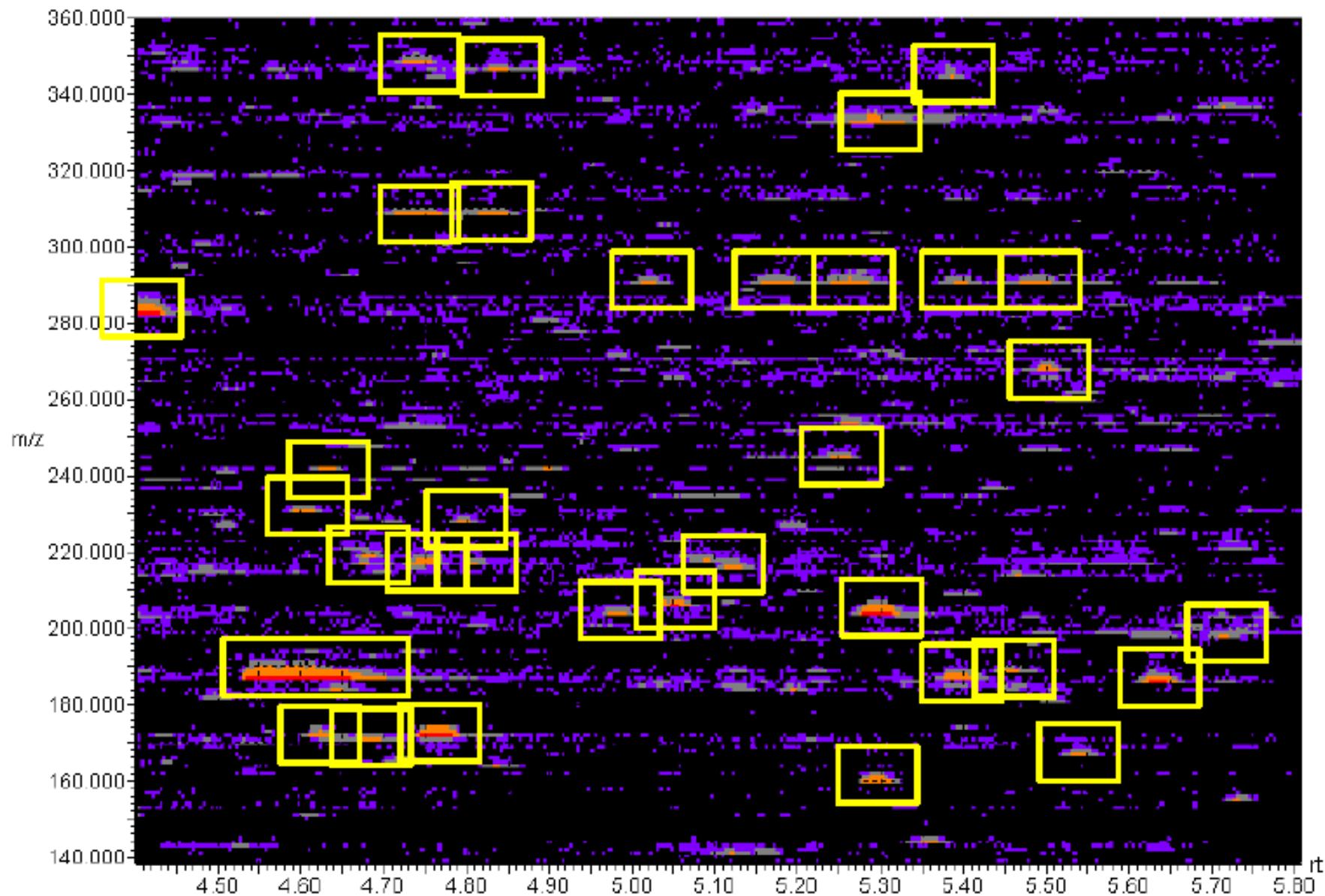
The scale at which MS- and NMR-based platforms generate metabolomics datasets for both research, core, and clinical facilities to address challenges in the various sciences—ranging from biomedical to agricultural—is underappreciated. Thus, metabolomics efforts spanning microbe, environment, plant, animal, and human systems have led to continual and concomitant growth of *in silico* resources for analysis and interpretation of these datasets. These software tools, resources, and databases drive the field forward to help keep pace with the amount of data being generated and the sophisticated and diverse analytical platforms that are being used to generate these metabolomics datasets. To address challenges in data preprocessing, metabolite annotation, statistical interrogation, visualization, interpretation, and integration, the metabolomics and informatics research community comes up with hundreds of tools every year. The purpose of the present review is to provide a brief and useful summary of more than 95 metabolomics tools, software, and databases that were either developed or significantly improved during 2017–2018. We hope to see this review help readers, developers, and researchers to obtain informed access to these thorough lists of resources for further improvisation, implementation, and application in due course of time.

# Output data structure

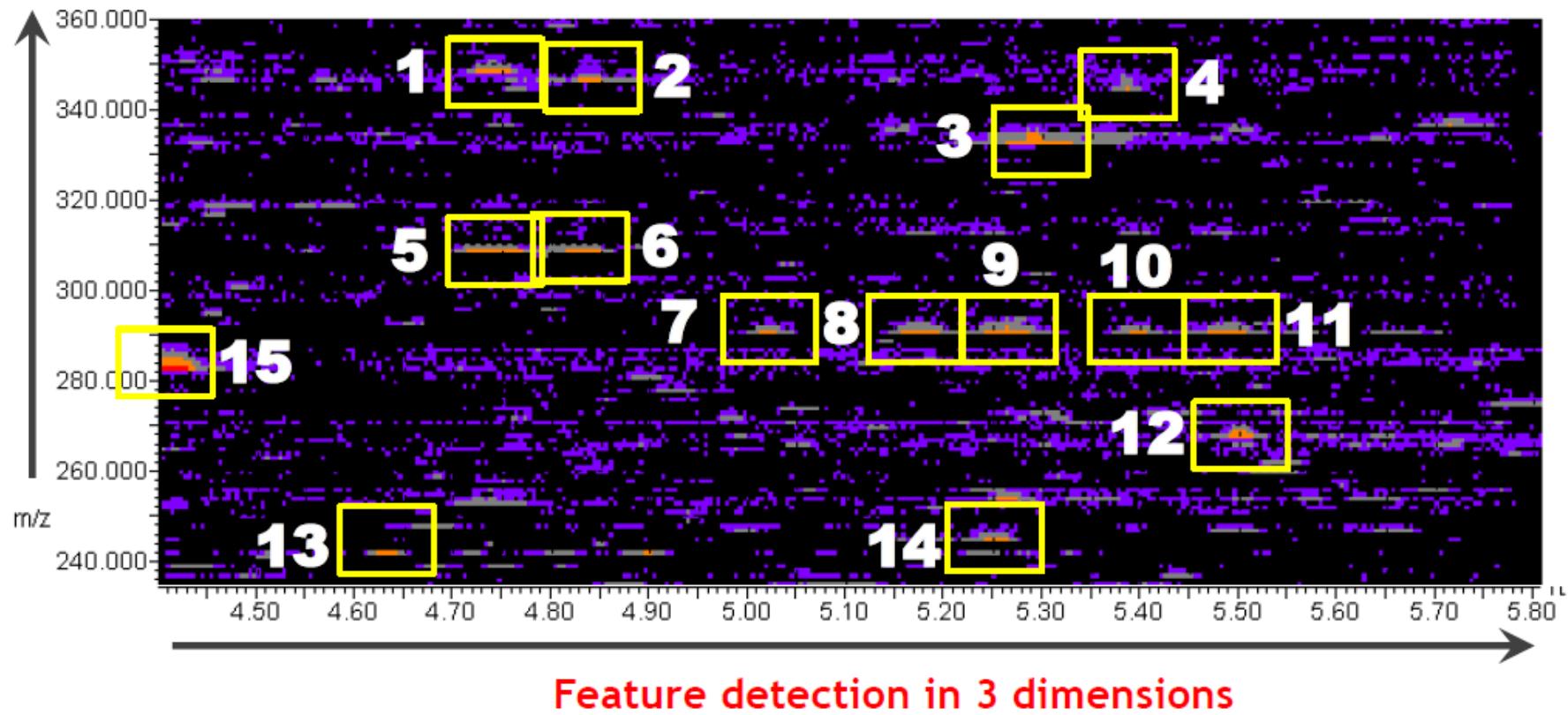




# Feature detection in 3-dimensional data



# Feature detection in 3-dimensional data



Concatenated to  
single term  
representing each  
feature

Feature = RT\_mz

1. Mass ( $m/z$ )
2. Chromatographic retention time (RT)
3. Intensity (“counts”)

# Output data structure

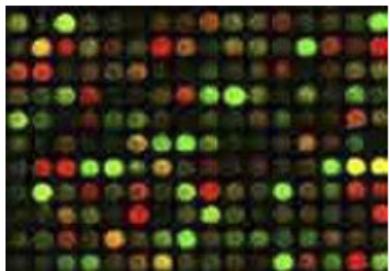
Feature Identifier

Sample identifier

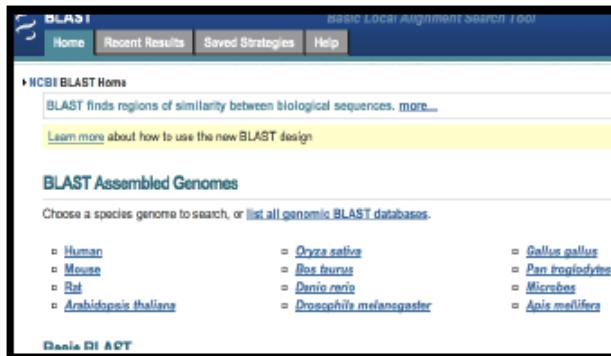
Feature.name	mz	rt	QC.nor.rsd	R2	X180501_1805235_T_NEG	X180501_1805236_F_NEG	X180501_1805237_T_NEG	X180501_1805238_F_NEG	X180501_1805239_QC_NEG
Label	mz	rt	QC.nor.rsd	R3	T	F	T	F	QC
13.22_855.6905m/z	855,6905	13,22	2,78	0,90	1,595756608	1,127168446	0,90987049	1,502157626	1,5242749
12.78_1175.8173m/z	1175,8173	12,78	3,75	0,89	1,012068895	1,924264871	0,941052344	1,24659619	0,700269515
13.11_832.6373m/z	832,6373	13,11	2,58	0,90	1,895778753	0,878647757	0,668815029	3,531989245	1,739529775
13.84_857.7065m/z	857,7065	13,84	2,50	0,91	1,446563144	1,459917821	0,95356938	1,72439594	1,702321205
5.97_339.2102m/z	339,2102	5,97	2,61	0,96	0,462711866	0,319314738	0,528005243	0,457088002	0,340008931
13.20_829.6733m/z	829,6733	13,20	3,62	0,86	1,694567338	1,470567656	0,934753971	1,746620672	1,462169722
12.41_899.6548n	898,6100	12,41	5,13	0,93	1,651790124	0,927857874	0,937857223	1,247611732	1,462013507
9.87_745.5764m/z	745,5764	9,87	4,30	0,93	1,218828015	1,276430502	1,070145633	1,561077878	1,428686119
11.62_828.6054m/z	828,6054	11,62	2,83	0,94	1,503987724	0,941322885	0,941501372	1,483197405	1,458004305
11.49_807.5658n	852,6062	11,49	5,46	0,91	1,031303648	0,798662614	0,683650659	2,074864895	1,016034483
10.74_747.5927m/z	747,5927	10,74	1,41	0,91	1,426032378	0,958717603	1,165625673	0,822211366	1,333198095
10.54_824.5737m/z	824,5737	10,54	2,97	0,92	1,02955304	0,68418959	0,920950584	1,168736087	2,084530777
12.31_854.6227m/z	854,6227	12,31	1,90	0,91	0,922779377	0,893270646	0,801607214	2,58524926	0,979421296
13.84_925.6952m/z	925,6952	13,84	2,97	0,91	1,541002608	1,549550309	0,964287935	1,703201186	1,722283028
11.21_826.5906m/z	826,5906	11,21	2,25	0,91	0,883019997	1,045390523	0,81326039	1,939078305	1,314143816
11.30_802.5890m/z	802,5890	11,30	1,88	0,93	1,3456936	1,026355676	1,006204223	0,940871989	1,513826994
11.30_870.5778m/z	870,5778	11,30	3,85	0,91	1,311794457	0,989558753	1,212741308	0,797155748	1,214987377
13.02_803.6564m/z	803,6564	13,02	5,33	0,91	1,430486792	1,911681286	1,338784094	1,130707907	1,540004947
13.12_900.6262m/z	900,6262	13,12	5,84	0,86	2,015477313	0,820250179	0,604150649	3,738009603	1,8536893
7.16_303.2418m/z	303,2418	7,16	4,66	0,91	1,379045696	0,893287758	1,514480954	1,358690786	2,126265644
11.64_934.6489n	915,6310	11,64	5,34	0,75	2,603452294	1,936584452	2,406736471	1,606282757	1,385685385
9.47_915.6317m/z	915,6317	9,47	7,79	0,88	2,289866624	1,380027238	1,588154586	0,952194138	1,786313204
13.86_993.6850m/z	993,6850	13,86	5,55	0,77	1,390865724	1,187803446	0,929504794	1,249722929	1,466552135
12.06_804.6054m/z	804,6055	12,06	2,78	0,92	1,709644369	0,933713733	0,696277652	1,561432257	1,782679331
6.03_566.3656m/z	566,3656	6,03	2,90	0,86	1,165464794	0,949195488	0,790945442	1,089173428	1,673271311
12.72_856.6375m/z	856,6375	12,72	3,32	0,92	1,420020977	0,930475089	0,884645361	2,959682093	0,998135122

Feature intensity

# The biggest challenge is Annotation



DNA/RNA



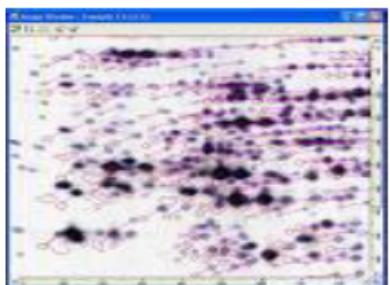
NCBI BLAST Home  
BLAST finds regions of similarity between biological sequences. [more...](#)  
Learn more about how to use the new BLAST design.

BLAST Assembled Genomes  
Choose a species genome to search, or [list all genomic BLAST databases](#).

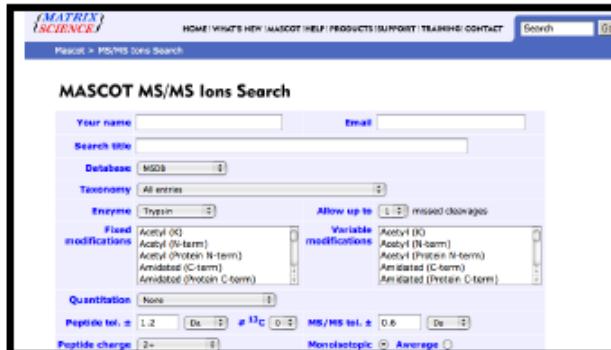
<input type="checkbox"/> Human	<input type="checkbox"/> <i>Oryza sativa</i>	<input type="checkbox"/> <i>Gallus gallus</i>
<input type="checkbox"/> Mouse	<input type="checkbox"/> <i>Bos taurus</i>	<input type="checkbox"/> <i>Pan troglodytes</i>
<input type="checkbox"/> Rat	<input type="checkbox"/> <i>Danio rerio</i>	<input type="checkbox"/> <i>Microbes</i>
<input type="checkbox"/> <i>Arabidopsis thaliana</i>	<input type="checkbox"/> <i>Drosophila melanogaster</i>	<input type="checkbox"/> <i>Anolis mitchellii</i>

Basic BLAST

Gene IDs +  
Transcript  
Abundance



Proteomics



MASCOT MS/MS Ions Search

Your name: [ ] Email: [ ]

Search title: [ ]

Database: MSDB [ ]

Taxonomy: All entries [ ]

Enzyme: Trypsin [ ] Allow up to [ ] missed cleavages

Fixed modifications: Acetyl (K) [ ] Acetyl (N-term) [ ] Amidated (C-term) [ ] Amidated (Protein C-term) [ ]

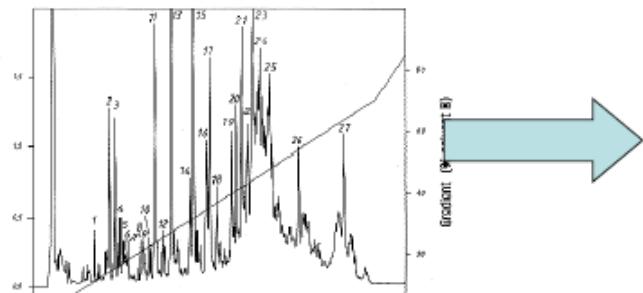
Variable modifications: Acetyl (K) [ ] Acetyl (N-term) [ ] Acetyl (Protein N-term) [ ] Amidated (C-term) [ ] Amidated (Protein C-term) [ ]

Quantitation: None [ ]

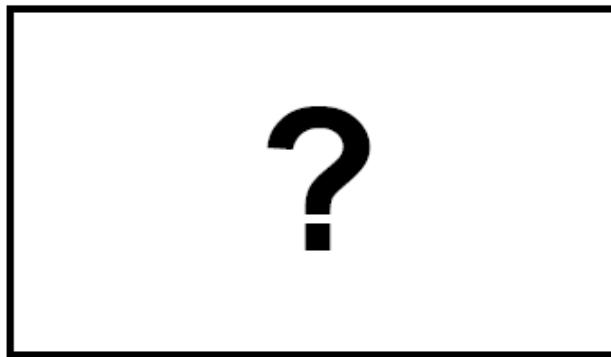
Peptide rel. #: 1.2 [ ] 10 [ ] 100 [ ] MS/MS rel. #: 0.6 [ ] 10 [ ]

Peptide charge: 2+ [ ] Monoisotopic: Average [ ]

Protein IDs +  
Concentrations



Metabolomics



Metabolite IDs +  
Concentrations

# Metabolome Databases

<http://metabolomicssociety.org/resources/metabolomics-databases>



[www.hmdb.ca](http://www.hmdb.ca)



[www.drugbank.ca](http://www.drugbank.ca)



[www.ymdb.ca](http://www.ymdb.ca)



[www.phenol-explorer.eu](http://www.phenol-explorer.eu)



[www.ecmdb.ca](http://www.ecmdb.ca)



[www.foodb.ca](http://www.foodb.ca)



[www.cowmetdb.ca](http://www.cowmetdb.ca)



[www.t3db.ca](http://www.t3db.ca)



[www.smpdb.ca](http://www.smpdb.ca)



[www.csfmetabolome.ca](http://www.csfmetabolome.ca)



[www.serummetabolome.ca](http://www.serummetabolome.ca)



[www.urinemetabolome.ca](http://www.urinemetabolome.ca)

# The biggest challenge is Annotation

## Levels of Metabolite Identification in MS

### 1. Positively identified compounds

Confirmed by match to known standard

### 2. Putatively identified compounds

Match to MS + RT or MS/MS + RT

### 3. Compounds putatively identified in a compound class

### 4. Unknown compounds

# Processing tools

## Commercial tools

**Agilent MassHunter Profinder**

**Bruker's ProfileAnalysis**

**Thermo SIEVE™**

**Waters' Progenesis QI**

**SkyLine**

## Free options

**XCMS Online**

**MZmine**

# Processing tools

## XCMS

The first open source tool for spectra processing

Does peak picking, peak matching and retention time alignment

Available as a program and a server

Accepts multiple formats: [mzXML](#), [mzData](#), [.cdf \(NetCDF\)](#), [.d folders \(Agilent; Bruker\)](#), [.wiff files \(AB SCIEX\)](#)

Metabolite identification is not the focus in XCMS (linked to Metlin)

# METLIN

Home\*  isoMETLIN Simple Search Advanced Search Batch Search Fragment Similarity Search Neutral Loss Search MS/MS Spectrum Match Search MRM\*  Logout [ tebanidz ]

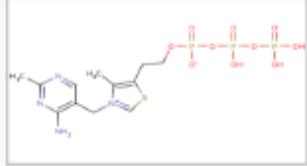
## Simple Search

Mass

Tolerance  PPM

Charge Neutral  Positive  Negative

Adducts M+H  M+NH4  M+Na  M+H-2H2O  M+H-H2O  M+K  M+ACN+H  M+ACN+Na  M+2Na-H  M+2H  M+3H  M+H+Na  M+2H+Na  M+2Na  M+2Na+H  M+Li  M+CH3OH+H

Show 10 entries							Search:
METLIN ID	Mass	ΔPPM	Name	KEGG	MS/MS	Structure	
3586	[M+H] <sup>+</sup> m/z 505.0107	21	Thiamin triphosphate <i>Formula:</i> C12H19N4O10P3S <i>CAS:</i> 3475-65-8	C03028	NO		

METLIN ID	Mass	ΔPPM	Name	KEGG	MS/MS	Structure
-----------	------	------	------	------	-------	-----------

Showing 1 to 1 of 1 entries

Previous 1 Next

Peptides

Remove Peptides from Search

Activate Windows

# Annotation Conversion

CTS - The Chemical Translation Service

 Simple Conversion

 Batch Conversion

 Services

## Simple Conversion

To convert a single identifier, enter it in the box below, select source and target types, and hit the Convert button.

Chemical Name

Convert ➔

InChIKey

Enter ID for conversion

Issues? Let us know on [BitBucket](#).

Finley and King Labs, Harvard Medical School

FLUKA

ForeChem

Fragments

Georganics

GlaxoSmithKline (GSK)

GLIDA, GPCR-Ligand Database

GNF / Scripps Winzeler lab

Golm Metabolome Database (GMD), Max Planck Institute of Molecular Plant Physiology

Hangzhou APIChem Technology

Hangzhou Trylead Chemical Technology

HDDH Pharma

Human Metabolome Database

HUMGENEX

IBCH RAS

IBM

ICCB-Longwood/NSRB Screening Facility, Harvard Medical School

Immunology Lab, Department of Biotechnology, Calicut University

InChI Code

# Outline

**What is metabolomics?**

**Experimental Design**

**Data Generation / Analytical technologies**

**Applications and Limits**

**Preprocessing**

**Data cleaning / analysis**

# Data cleaning and analysis

## Input

A matrix containing numerical values

- Concentrations (Targeted)

- Peak intensities (Untargeted)

Meta-data

- Class labels, experimental factors

## Output

Discriminant features

Clustering patterns

Biological Inference

Biomarkers

Predictive models

# Data cleaning and analysis

Samples

## Variables

Metabolites

6	0	4	8	0	2	2	7	2	2
0	1	8	0	8	0	0	3	4	9
4	4	0	4	0	1	4	0	1	8
8	0	2	8	2	4	8	4		0
0	4	6	0	5	2	0	8	0	2
2	8	9	5	5	5	2	0	4	6
4	9	5	6	5	4	5	2	8	9
1	8	0	8	5	0	1	8	0	8
4	0	4	0	0	4	2	0	4	0
0	2	8	2	5	8	0	2	8	2

# Data cleaning and analysis

**Variables**

Metabolites

**Samples**

**Information**  
+  
**Noise**

# Data Analysis / Data cleaning

**Remove as much as possible noise**

**Extract as much as possible information**



# Data Analysis / Data cleaning



## Data cleaning

- Missing values imputation
- Filtering (Min, IQR, RSD, CV, R2 ...)
- Normalization

# Data Analysis / Data cleaning

Cut-off



Feature.name Label	mz	rt	QC.nor.rsd	R2	X180501_1805235_T_NEG T	X180501_1805236_F_NEG F	X180501_1805237_T_NEG T	X180501_1805238_F_NEG F	X180501_1805239_T_NEG T
	mz	rt	QC.nor.rsd	R3					
13.22_855.6905m/z	855,6905	13,22	2,78	0,90	1,595756608	1,127168446	0,90987049	1,502157626	1,5242749
12.78_1175.8173m/z	1175,8173	12,78	3,75	0,89	1,012068895	1,924264871	0,941052344	1,24659619	0,700269515
13.11_832.6373m/z	832,6373	13,11	2,58	0,90	1,895778753	0,878647757	0,668815029	3,531989245	1,739529775
13.84_857.7065m/z	857,7065	13,84	2,50	0,91	1,446563144	1,459917821	0,95356938	1,72439594	1,702321205
5.97_339.2102m/z	339,2102	5,97	2,61	0,96	0,462711866	0,319314738	0,528005243	0,457088002	0,340008931
13.20_829.6733m/z	829,6733	13,20	3,62	0,86	1,694567338	1,470567656	0,934753971	1,746620672	1,462169722
12.41_899.6548n	898,6100	12,41	5,13	0,93	1,651790124	0,927875874	0,937857223	1,247611732	1,462013507
9.87_745.5764m/z	745,5764	9,87	4,30	0,93	1,218828015	1,276430502	1,070145633	1,561077878	1,428686119
11.62_828.6054m/z	828,6054	11,62	2,83	0,94	1,503987724	0,941322885	0,941501372	1,483197405	1,458004305
11.49_807.5658n	852,6062	11,49	5,46	0,91	1,031303648	0,798662614	0,683650659	2,074864895	1,016034483
10.74_747.5927m/z	747,5927	10,74	1,41	0,91	1,426032378	0,958717603	1,165625673	0,822211366	1,333198095
10.54_824.5737m/z	824,5737	10,54	2,97	0,92	1,02955304	0,68418959	0,920950584	1,168736087	2,084530777
12.31_854.6227m/z	854,6227	12,31	1,90	0,91	0,922779377	0,893270646	0,801607214	2,58524926	0,979421296
13.84_925.6952m/z	925,6952	13,84	2,97	0,91	1,541006208	1,549550309	0,964287935	1,703201186	1,722283028
11.21_826.5906m/z	826,5906	11,21	2,25	0,91	0,883019997	1,045390523	0,81326039	1,939078305	1,314143816
11.30_802.5890m/z	802,5890	11,30	1,88	0,93	1,3456936	1,026355676	1,006204223	0,940871989	1,513826994
11.30_870.5778m/z	870,5778	11,30	3,85	0,91	1,311794457	0,989558753	1,212741308	0,797155748	1,214987377
13.02_803.6564m/z	803,6564	13,02	5,33	0,91	1,430486792	1,911681286	1,338784094	1,130707907	1,540004947
13.12_900.6262m/z	900,6262	13,12	5,84	0,86	2,015477313	0,820250179	0,604150649	3,738009603	1,8536893
7.16_303.2418m/z	303,2418	7,16	4,66	0,91	1,379045696	0,893287758	1,514480954	1,358690786	2,126265644
11.64_934.6489n	915,6310	11,64	5,34	0,75	2,603452294	1,936584452	2,406736471	1,606282757	1,385685385
9.47_915.6317m/z	915,6317	9,47	7,79	0,88	2,289866624	1,380027238	1,588154586	0,952194138	1,786313204
13.86_993.6850m/z	993,6850	13,86	5,55	0,77	1,390865724	1,187803446	0,929504794	1,249722929	1,466552135
12.06_804.6054m/z	804,6055	12,06	2,78	0,92	1,709644369	0,933713733	0,696277652	1,561432257	1,782679331
6.03_566.3656m/z	566,3656	6,03	2,90	0,86	1,165464794	0,949195488	0,790945442	1,089173428	1,673271311
12.72_856.6375m/z	856,6375	12,72	3,32	0,92	1,420020977	0,930475089	0,884645361	2,959682093	0,998135122

# Data Analysis / Data cleaning

## Missing values imputation

Replace by

mean

median

min

...

k-nearest neighbour (KNN)

PCA based methods

...

# Data Analysis / Normalization

## Sample normalization (row-wise)

To remove systematic variation between experimental conditions unrelated to the biological differences (i.e. dilutions, mass)

Total signal, sum of signals

Reference compound: Internal standards, endogenous metabolites

Reference sample: QC's, Controls

## Feature normalization (column-wise)

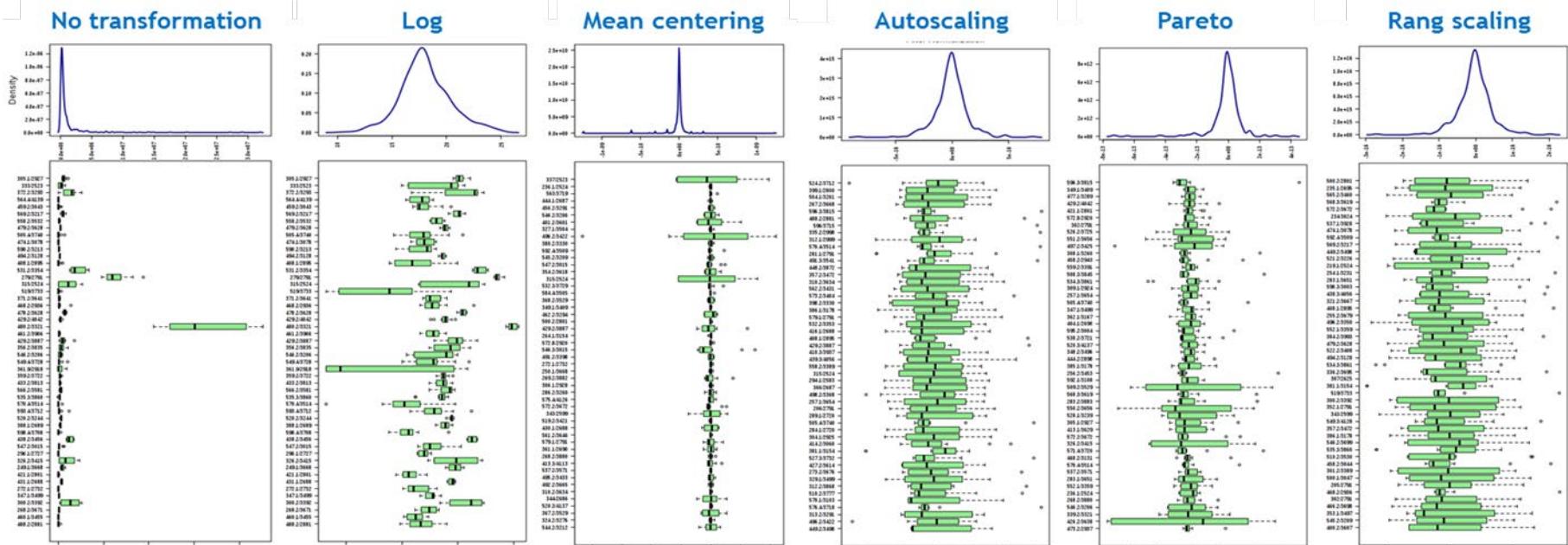
To bring variances of all features close to equal

Log transformation

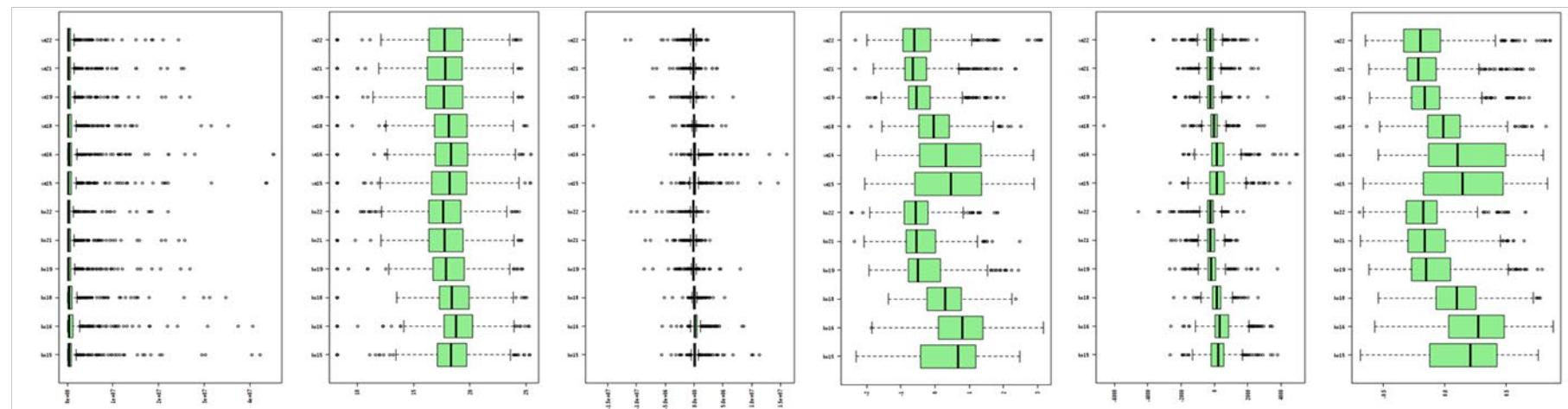
Scaling

# Data Analysis / Normalization

Variables



Samples



# Data Analysis / Normalization

W162–W170 *Nucleic Acids Research*, 2017, Vol. 45, Web Server issue  
doi: 10.1093/nar/gkx449

Published online 19 May 2017

## NOREVA: normalization and evaluation of MS-based metabolomics data

Bo Li<sup>1,†</sup>, Jing Tang<sup>1,†</sup>, Qingxia Yang<sup>1,2,†</sup>, Shuang Li<sup>1</sup>, Xuejiao Cui<sup>1</sup>, Yinghong Li<sup>1</sup>, Yuzong Chen<sup>3</sup>, Weiwei Xue<sup>1</sup>, Xiaofeng Li<sup>1</sup> and Feng Zhu<sup>1,2,\*</sup>

Preprints ([www.preprints.org](http://www.preprints.org)) | NOT PEER-REVIEWED | Posted: 3 July 2018

[doi:10.20944/preprints201807.0059.v1](https://doi.org/10.20944/preprints201807.0059.v1)

Peer-reviewed version available at *Metabolites* 2018, 8, 47; [doi:10.3390/metabo8030047](https://doi.org/10.3390/metabo8030047)

*Review*

## Data Normalization in NMR-based Metabolomics

Helena U. Zacharias<sup>1</sup>, Michael Altenbuchinger<sup>2</sup> and Wolfram Gronwald<sup>3\*</sup>

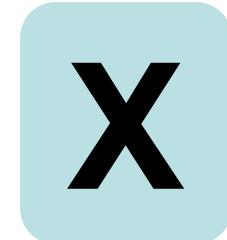
# Data Analysis

Tow main objectives

## **Descriptive data analysis (Unsupervised learning)**

Mining massive datasets to discover hidden  
*data structures*  
*hidden relationships*  
*patterns, trends and clusters*  
*outliers*

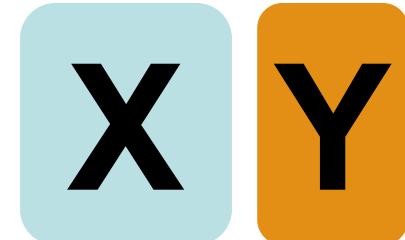
Dimension reduction



## **Predictive data analysis (Supervised learning)**

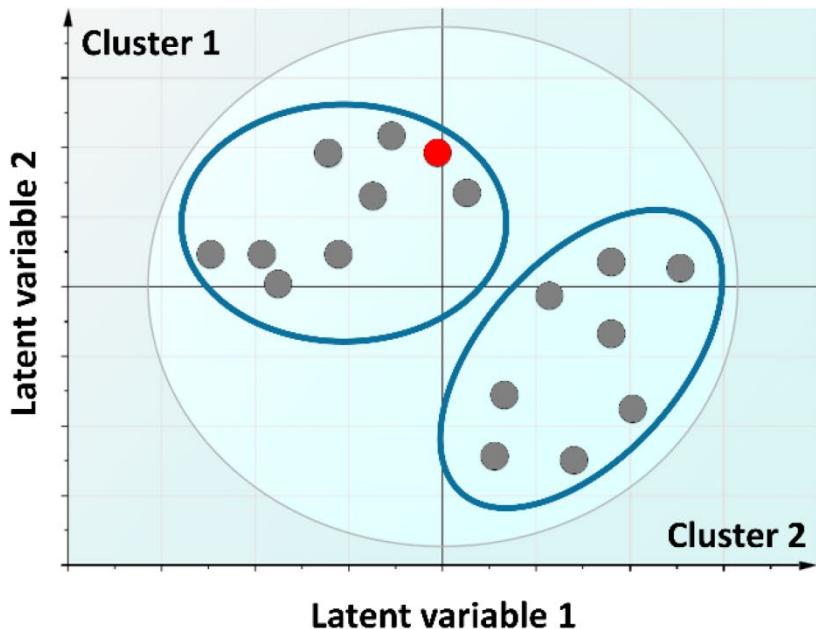
Building models for specific tasks using training datasets  
*regression*  
*classification,*  
*pattern recognition*  
*machine learning tasks*

Assessing the predictive accuracy of the models using new datasets

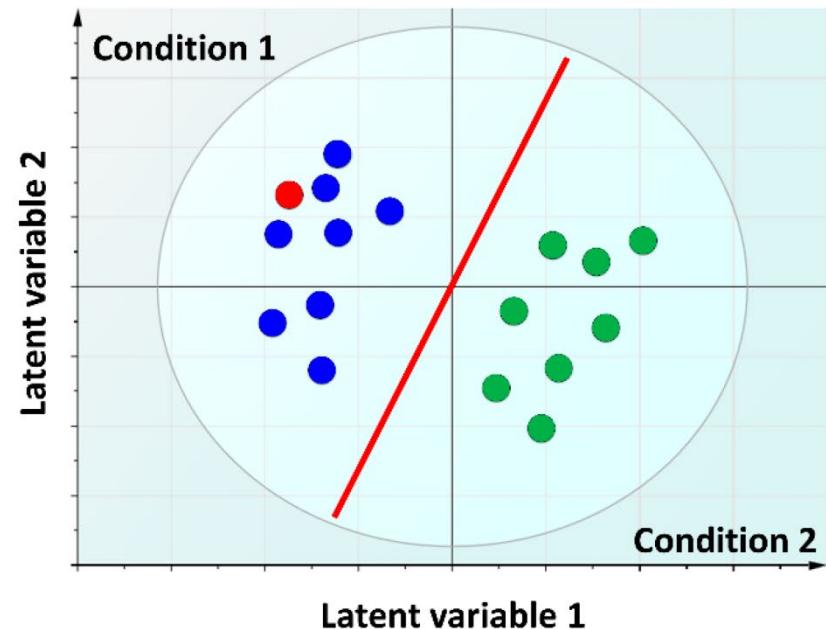


# Data Analysis

## Unsupervised learning



## Supervised learning



# Data Analysis / **Unsupervised learning**

## **Clustering**

Organize the 1000s of variables into blocks

Variables in each block are more homogenous

Key parameter: similarities (Distance, Spearman, Pearson ...)

Similarity between samples - Similarity between clusters

Visualization using Heatmaps

- *K-means*
- *Hierarchical Methods*

## **Dimension reduction**

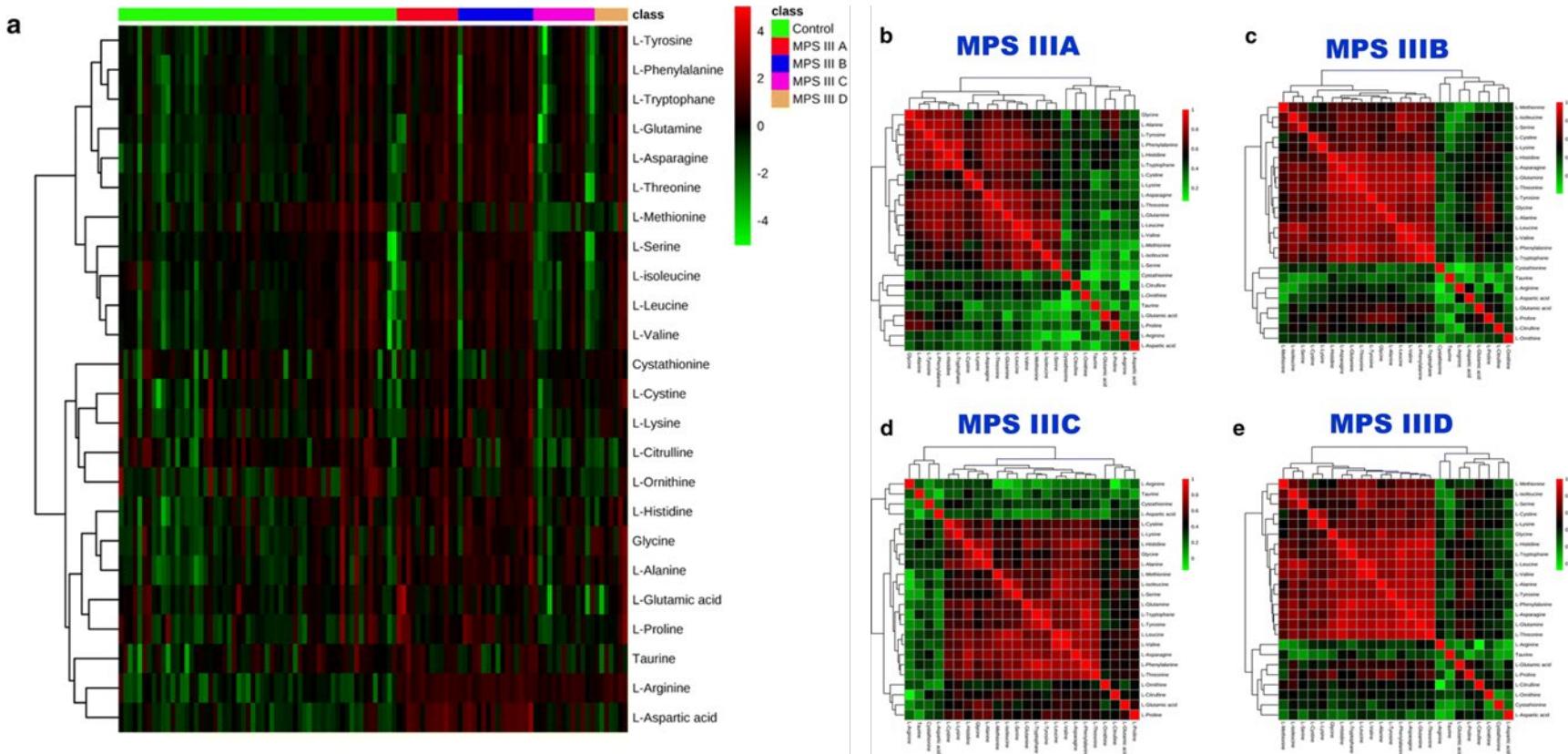
Reduce the high-dimensional data

1000s into low-dimensions (Latent variables)

- *Principal component analysis (PCA)*

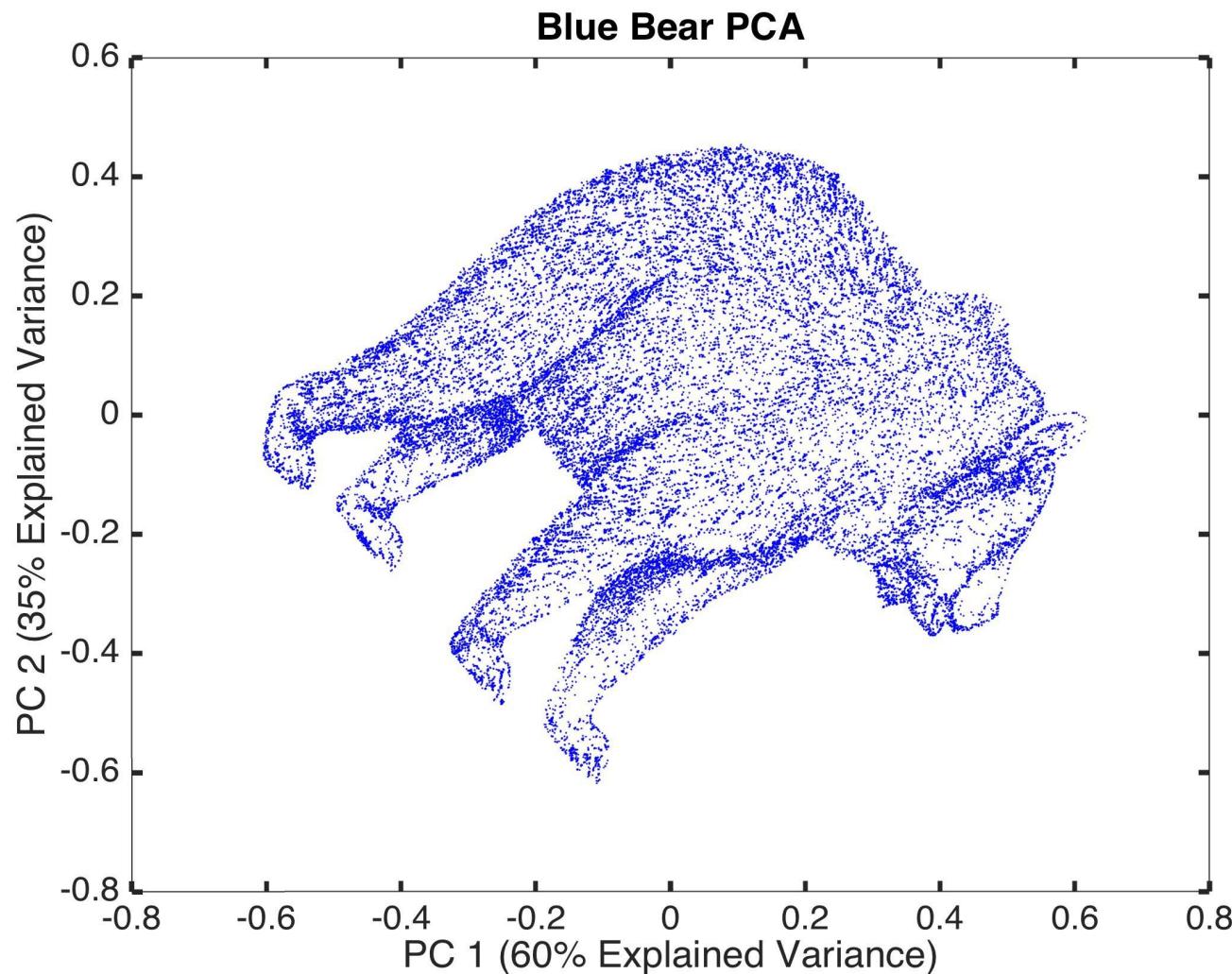
# Data Analysis / Unsupervised learning

## Correlation Heatmap



Tebani A et.al. JTRM 2018

# Data Analysis / Unsupervised learning (PCA)



# Data Analysis / Supervised learning

**Linear Discriminant Analysis**

**Partial Least Squares**

**k-Nearest Neighbors**

**Random Forest**

**Support Vector Machines**

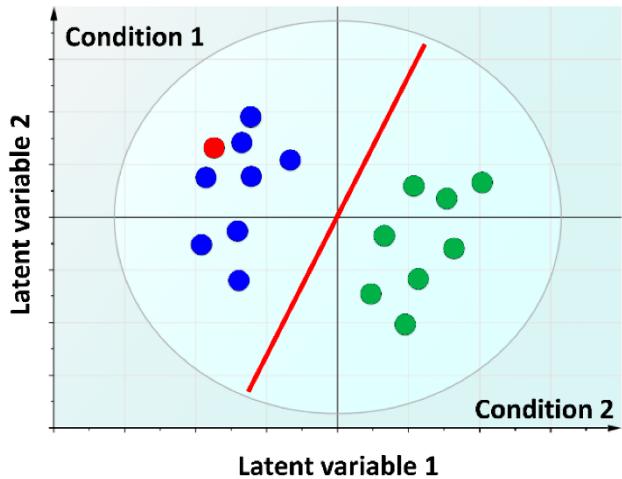
**Bayesian networks**

**Neural Networks**

**... name it**

# Data Analysis / Supervised learning

## Supervised learning



$$f(\mathbf{x}) = \mathbf{y}$$

*Supervised = algorithm does know true answer*

Attempt to find rule which predicts output for given input

*Two cases*

Classification - output is discrete (Class label)

Regression - output is continuous

## Samples

Variables	6	0	4	8	0	2	2	7	2	2
Variables	0	1	2	0	8	0	0	3	4	9
Variables	4	4	0	0	0	1	0	1	8	
Variables	8	0	2	0	0	0	0	4	0	0
Variables	0	4	6	0	0	0	8	0	2	
Variables	2	8	9	5	0	0	2	0	4	6
Variables	4	9	5	0	0	0	5	2	8	9
Variables	1	8	0	5	0	0	8	0	0	8
Variables	4	0	0	0	4	0	0	4	0	
Variables	0	2	8	2	5	8	0	2	8	2

**Y**

## Model validation

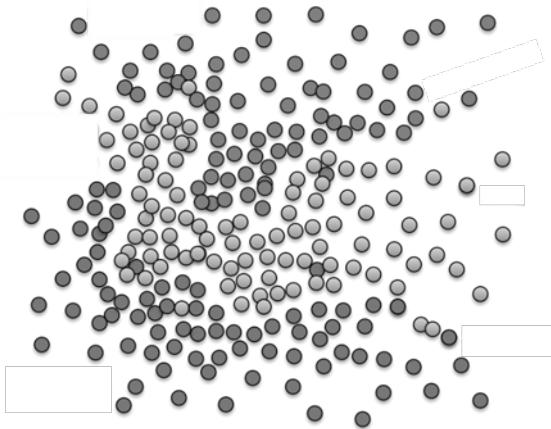
A *statistically valid* model

- Has **good fit** to the data
- Is ***predictive*** of new data

Validation methods

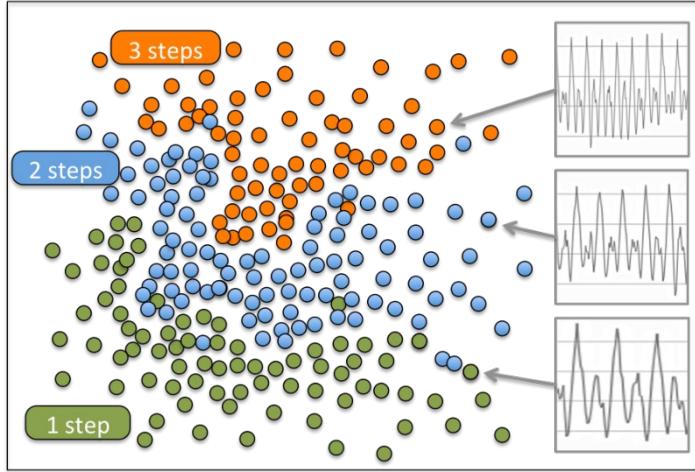
- Training set / Testing set
- Cross-validation
- Permutation test

# Data Analysis / Supervised learning

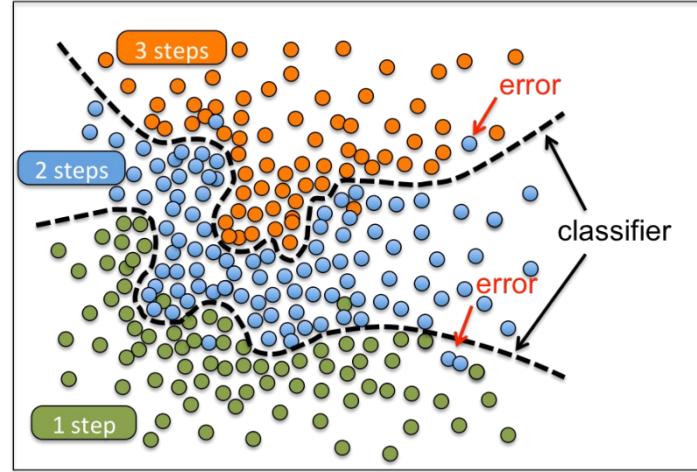


# Data Analysis / Supervised learning

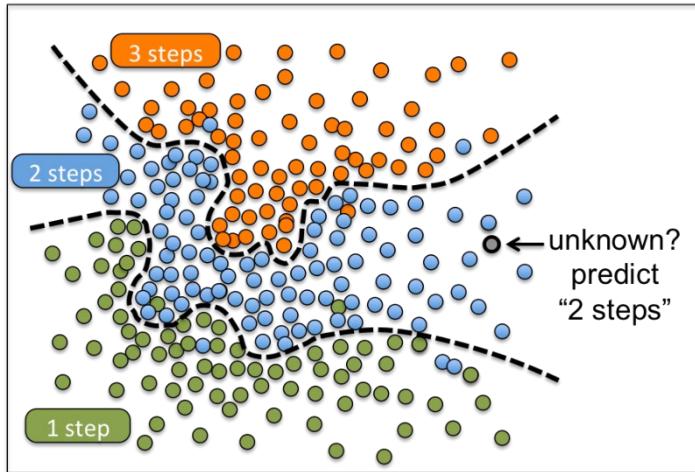
1



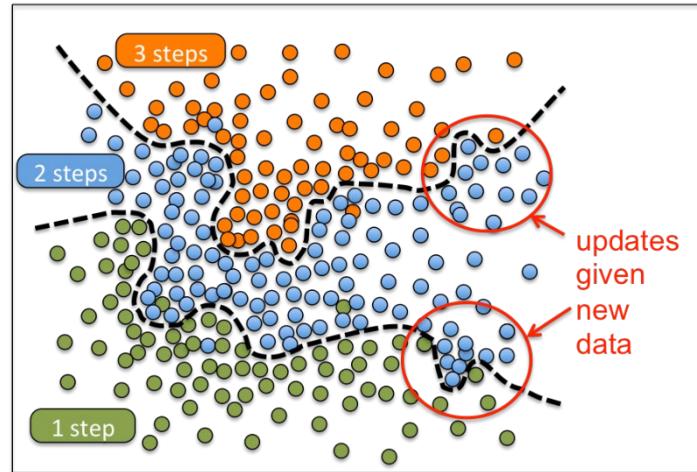
2



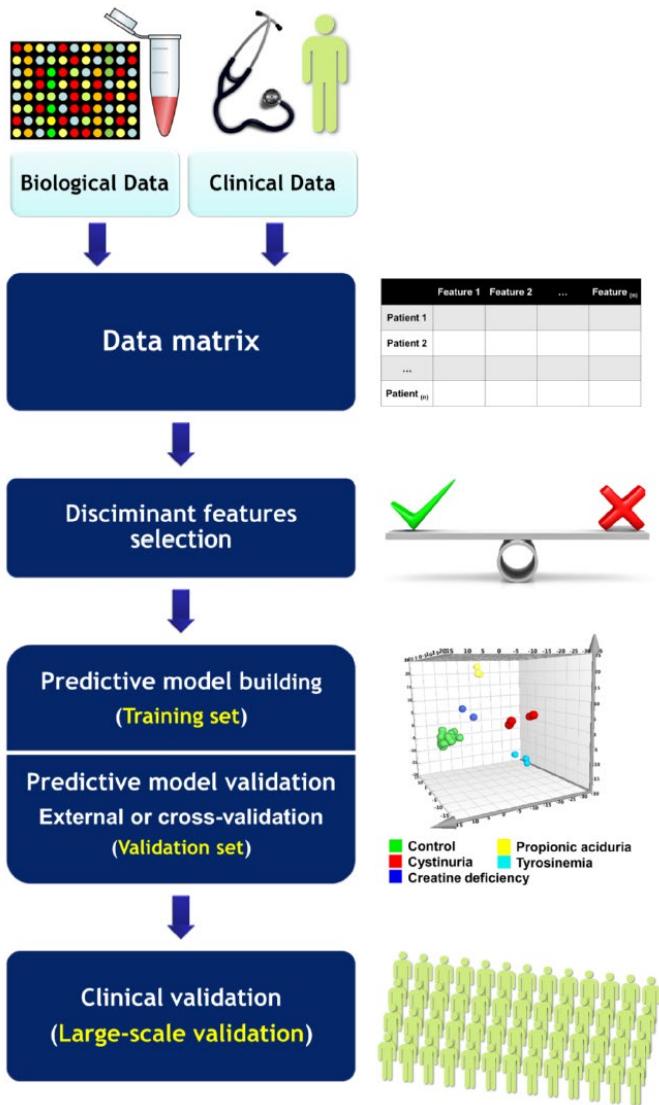
3



4



# Data Analysis / Supervised learning

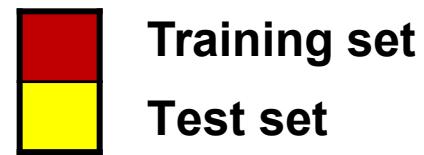


**Training dataset:** a set of examples used to build the predictive model

**Validation dataset:** a set of examples used to refine the model parameters and estimate the error.

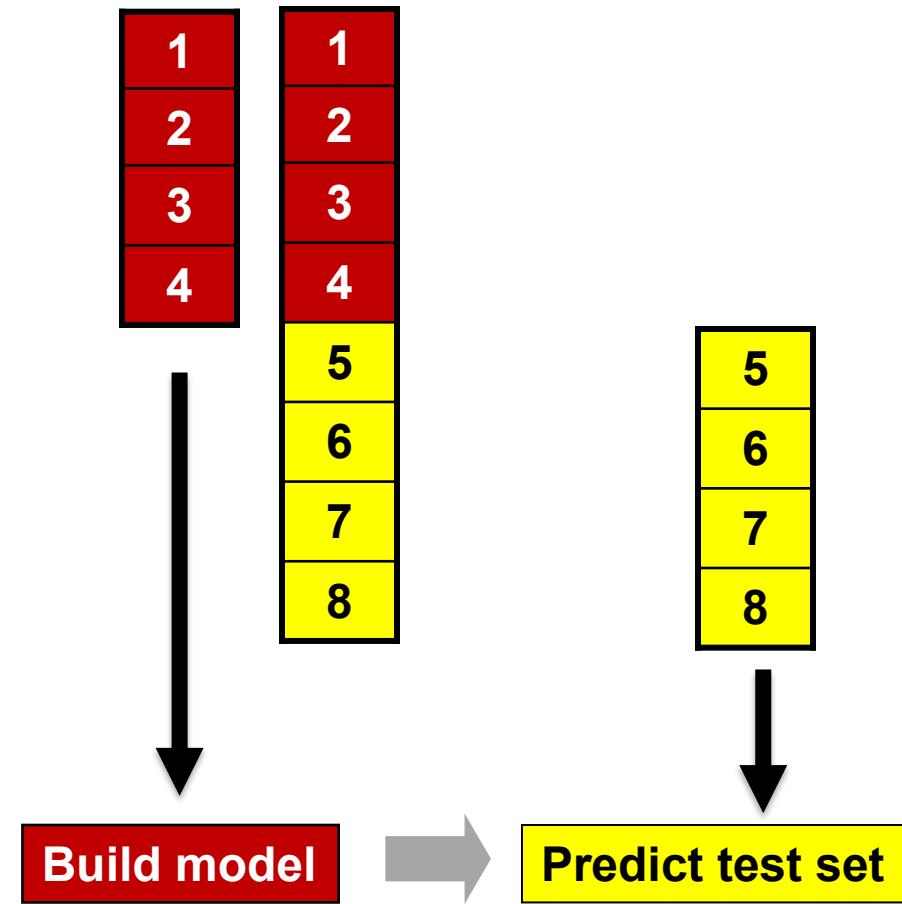
**Test dataset:** used only to evaluate the predictive performance of the model. They are never used during the learning or testing process.

# Data Analysis / Supervised learning



## *Validation: split the data*

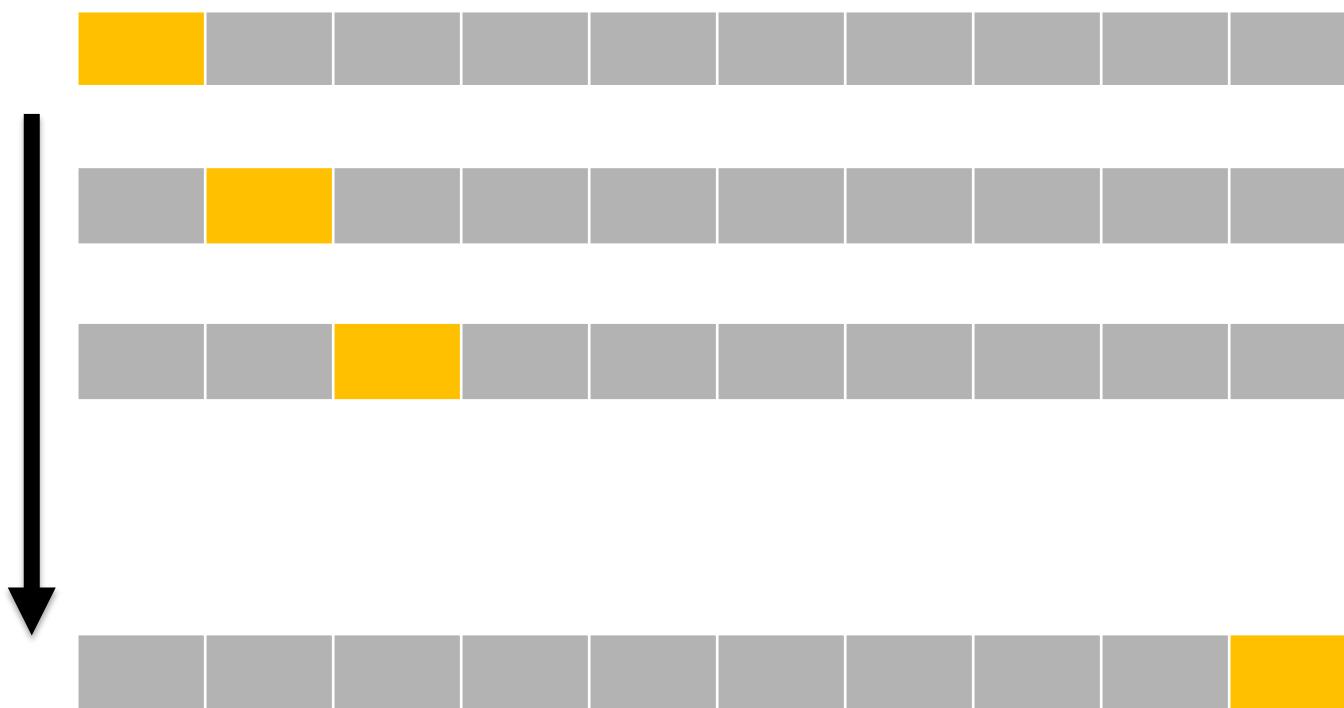
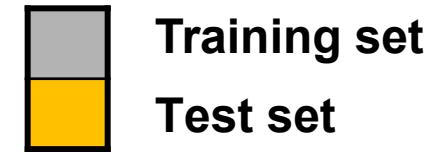
- Training set - to build the model
- Test set - to validate the model



# Data Analysis / Supervised learning

## Cross-Validation

- Divide the dataset into  $k$  parts
- Use  $k-1$  for training and 1 for testing
- Determine an optimization metric (RMSE, Accuracy, ROC, ...)
- Iterate



# Take home message

Experimental design ++++++

Know your data

Main technologies are MS and NMR

Annotation is challenging in untargeted metabolomics

# Acknowledge & Questions



Dr. Abdellah Tebani

