R-BASED TOOLS IN MASS SPECTROMETRY-BASED RESEARCH

Olga Vitek

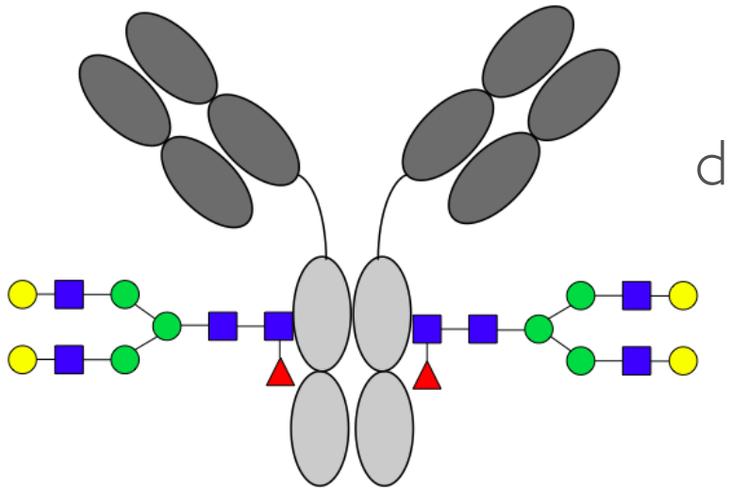
College of Computer and Information Science

Northeastern University

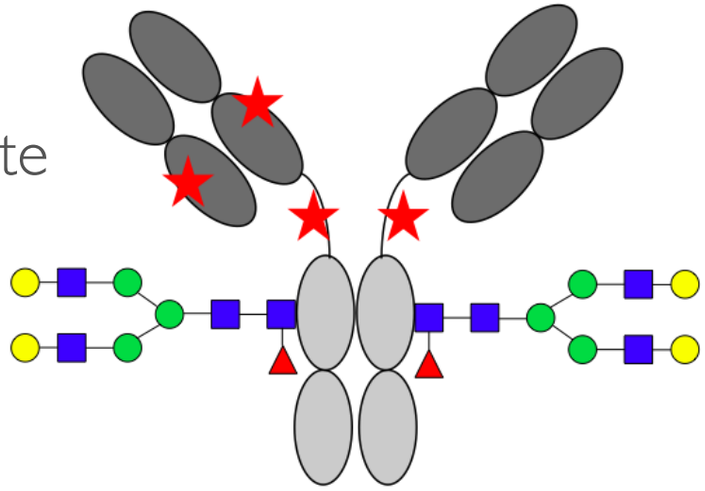
EXAMPLE PROBLEM

Characterization of therapeutic protein modifications

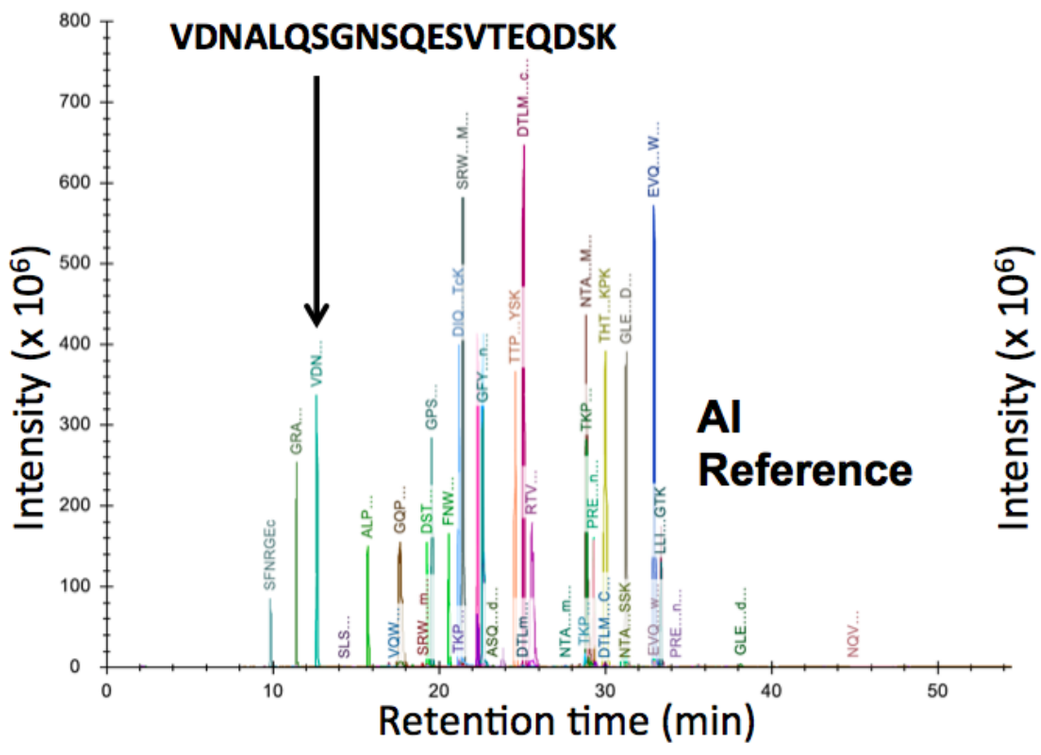
Antibody intermediate

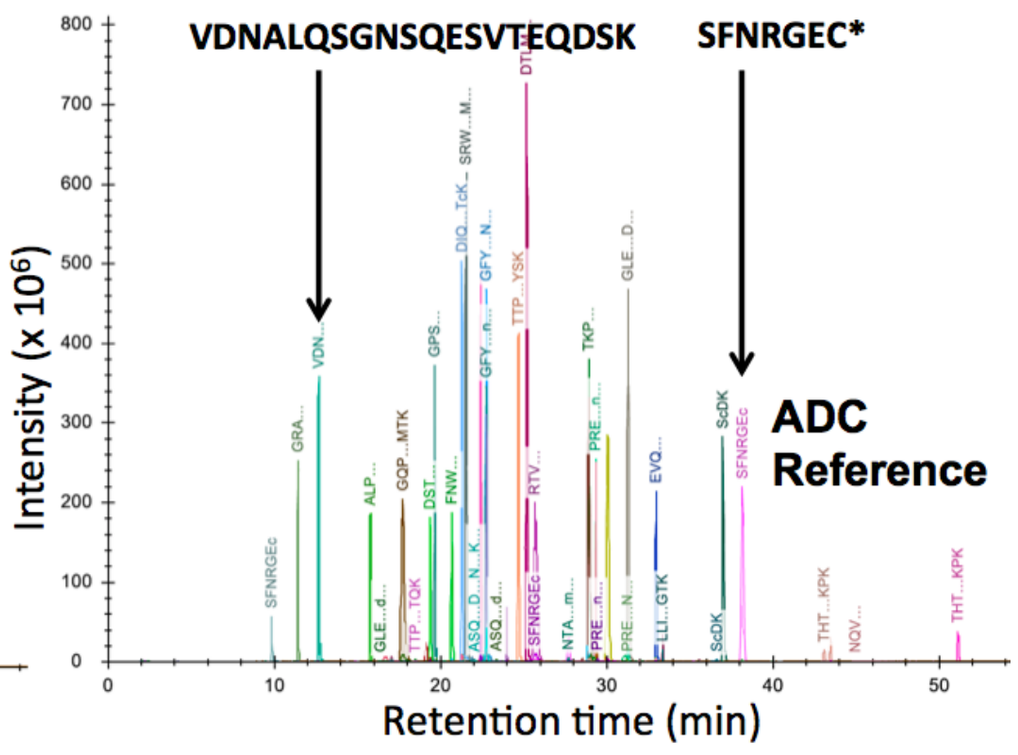


Antibody drug conjugate



Peptide mapping





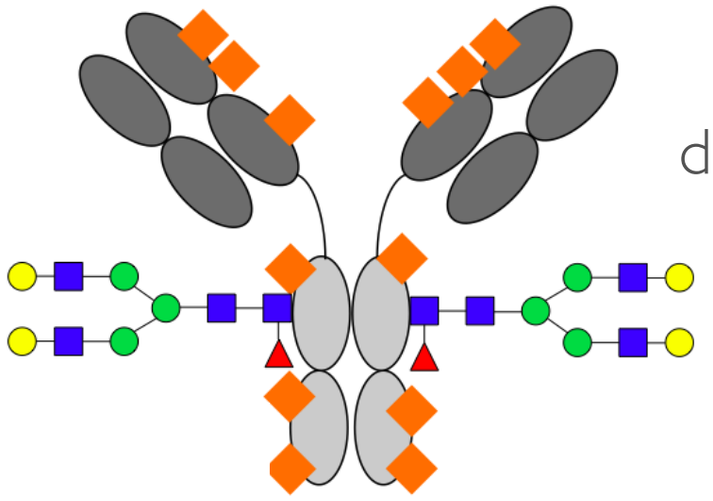
T.-H. Tsai *et al*. “Statistical characterization of therapeutic protein modifications”. *Scientific Reports*, 2017.

Collaboration with Genentech

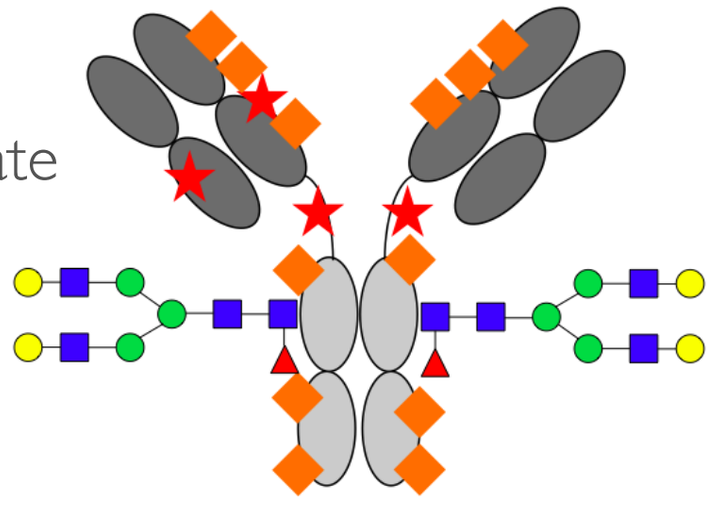
EXAMPLE PROBLEM

Characterization of therapeutic protein modifications

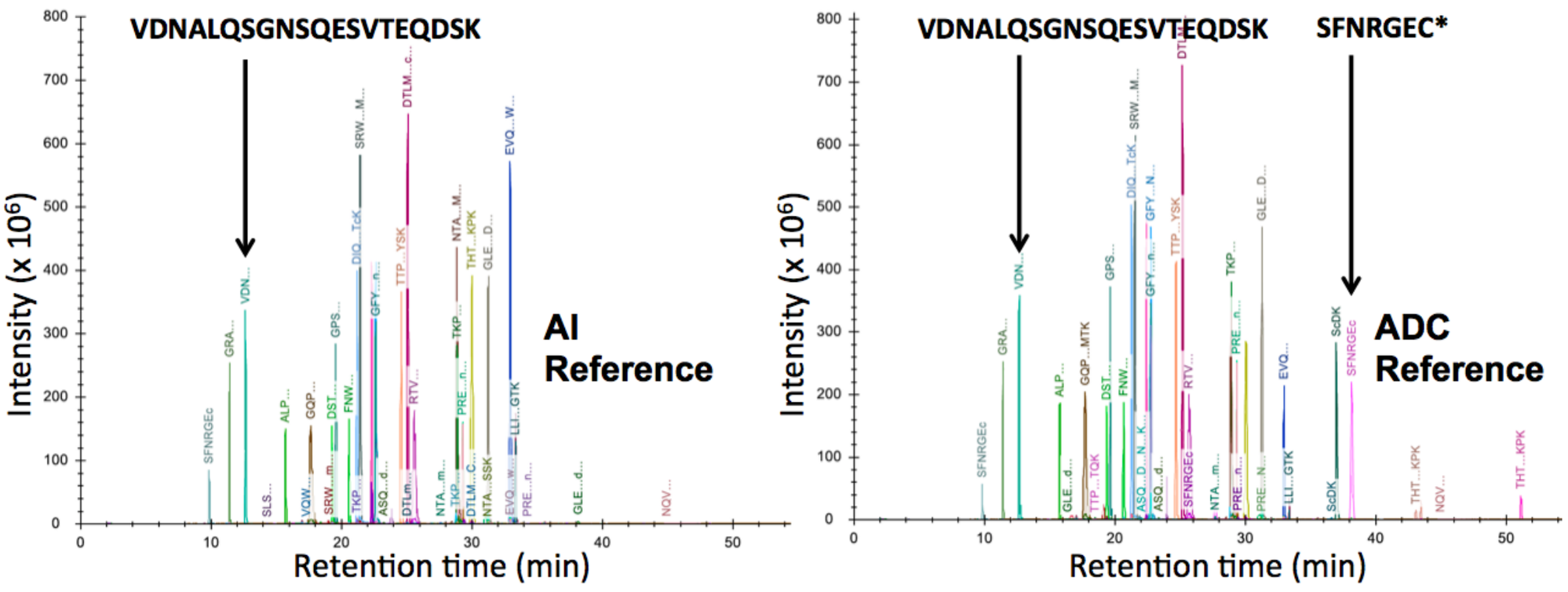
Antibody intermediate



Antibody drug conjugate



Peptide mapping



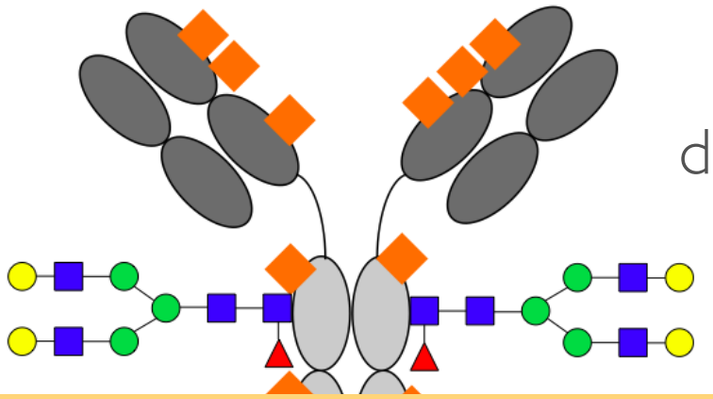
T.-H. Tsai *et al*. “Statistical characterization of therapeutic protein modifications”. *Scientific Reports*, 2017.

Collaboration with Genentech

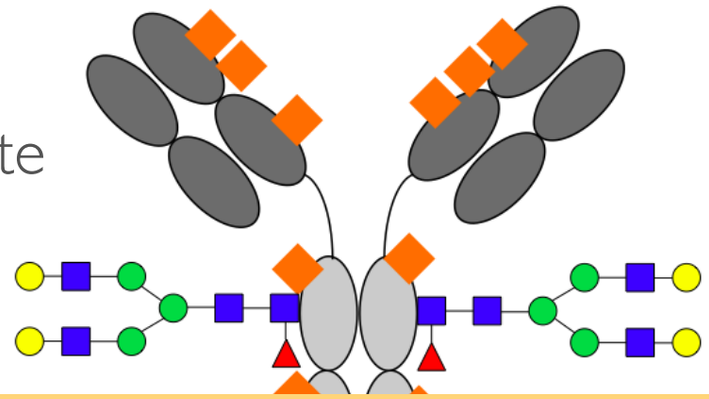
EXAMPLE PROBLEM

Characterization of therapeutic protein modifications

Antibody intermediate

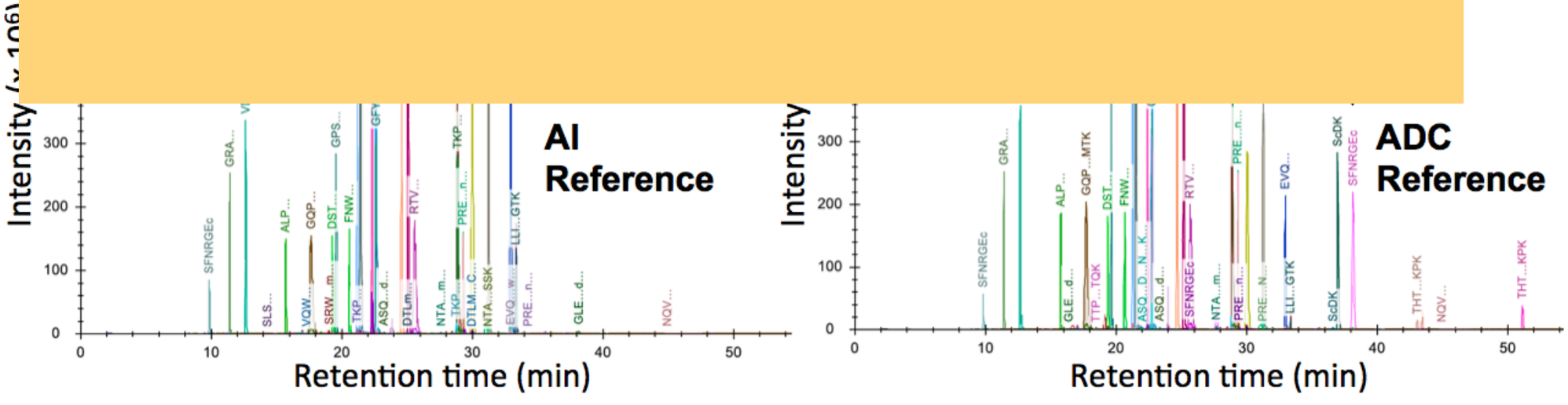


Antibody drug conjugate



- Characterize site occupancy of modifications
- Does drug conjugate change functional form?
- What is the min number of replicates?

Peptide mapping

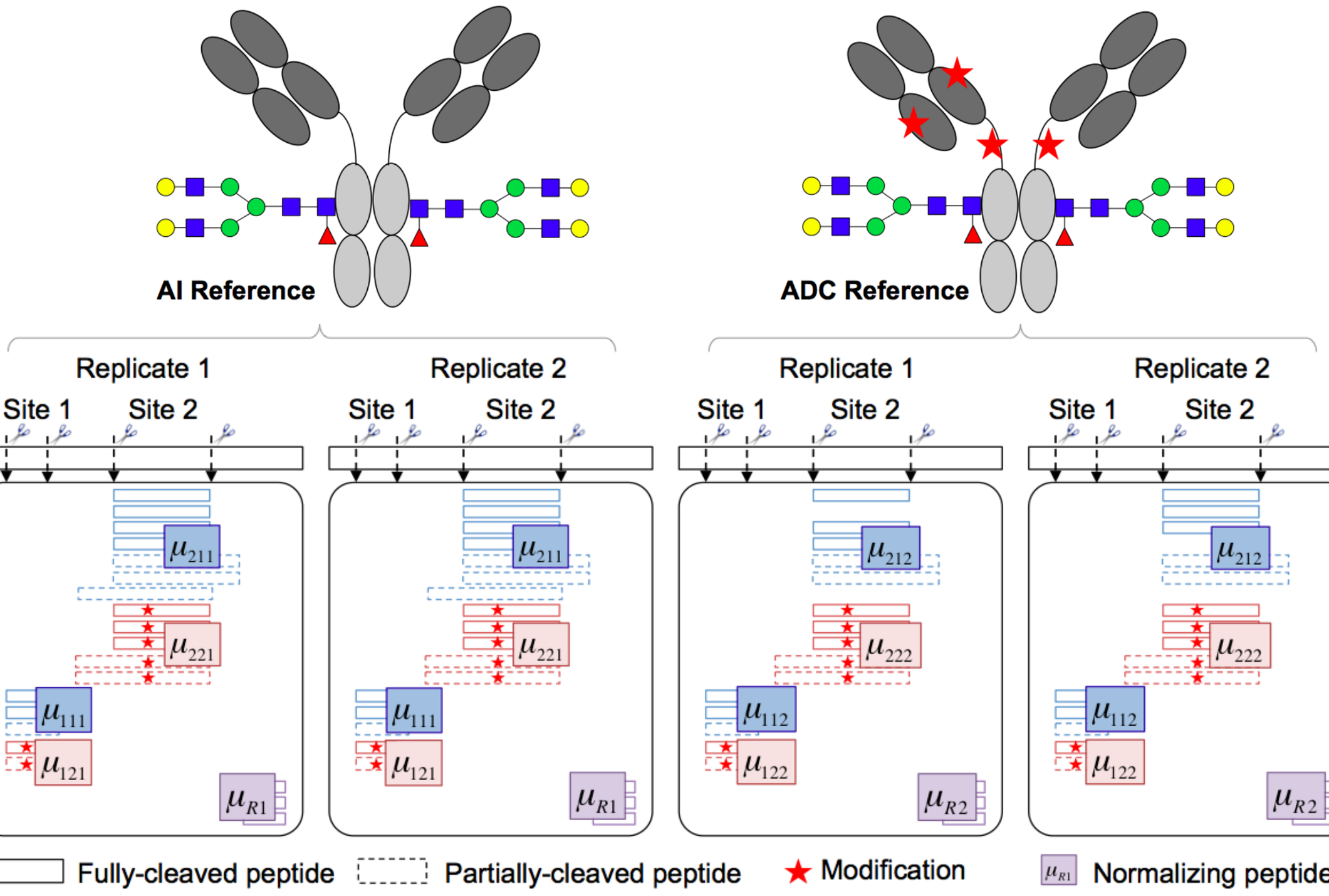


T.-H. Tsai *et al*. “Statistical characterization of therapeutic protein modifications”. *Scientific Reports*, 2017.

Collaboration with Genentech

EXAMPLE PROBLEM

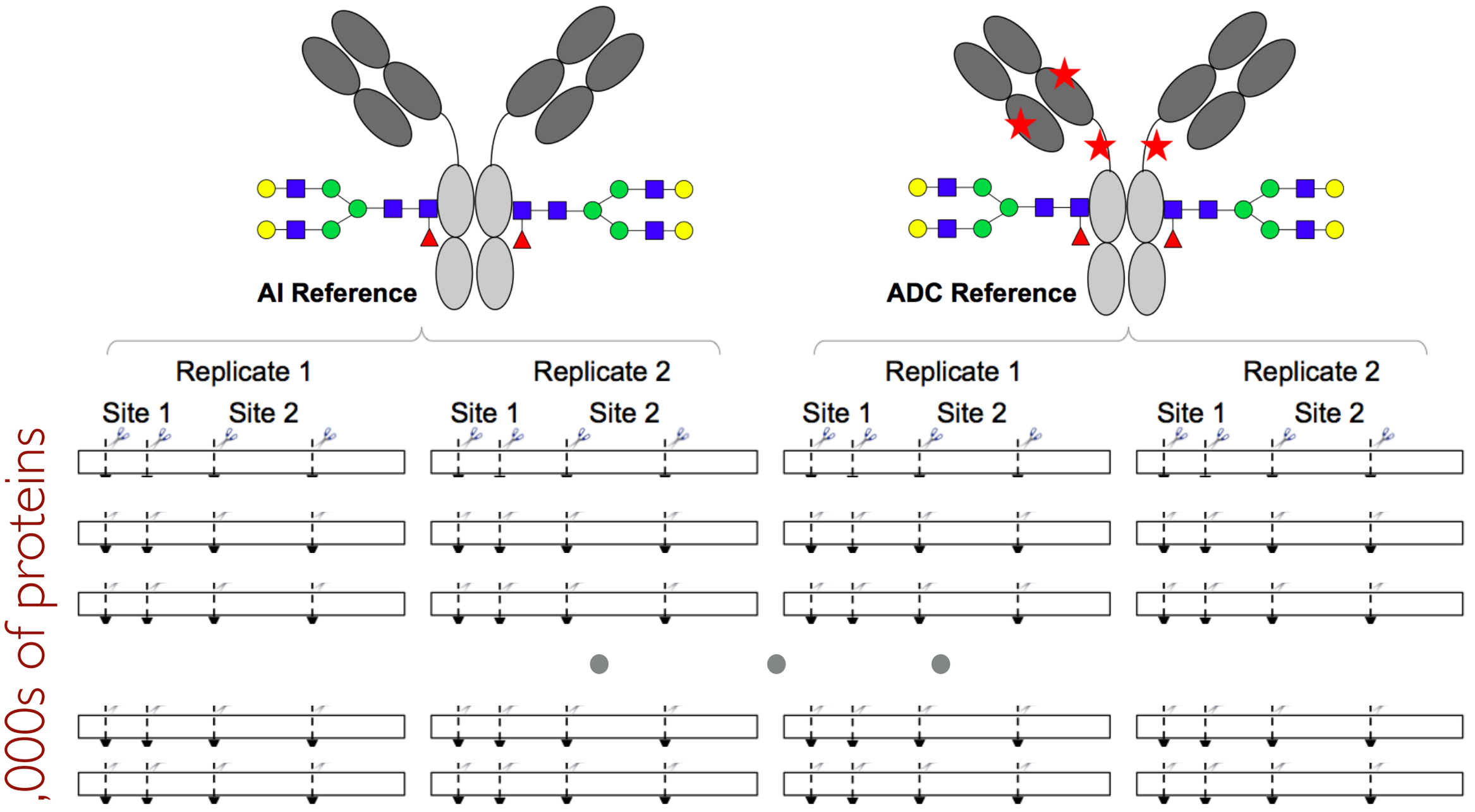
Statistical model formulation



Linear mixed effects models

EXTENSION TO SIMILAR PROBLEMS

Higher throughput: mass spectrometry-based proteomics



Diverse biological and technological aspects of experimental designs

Collaboration with Hoffman-La Roche

MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

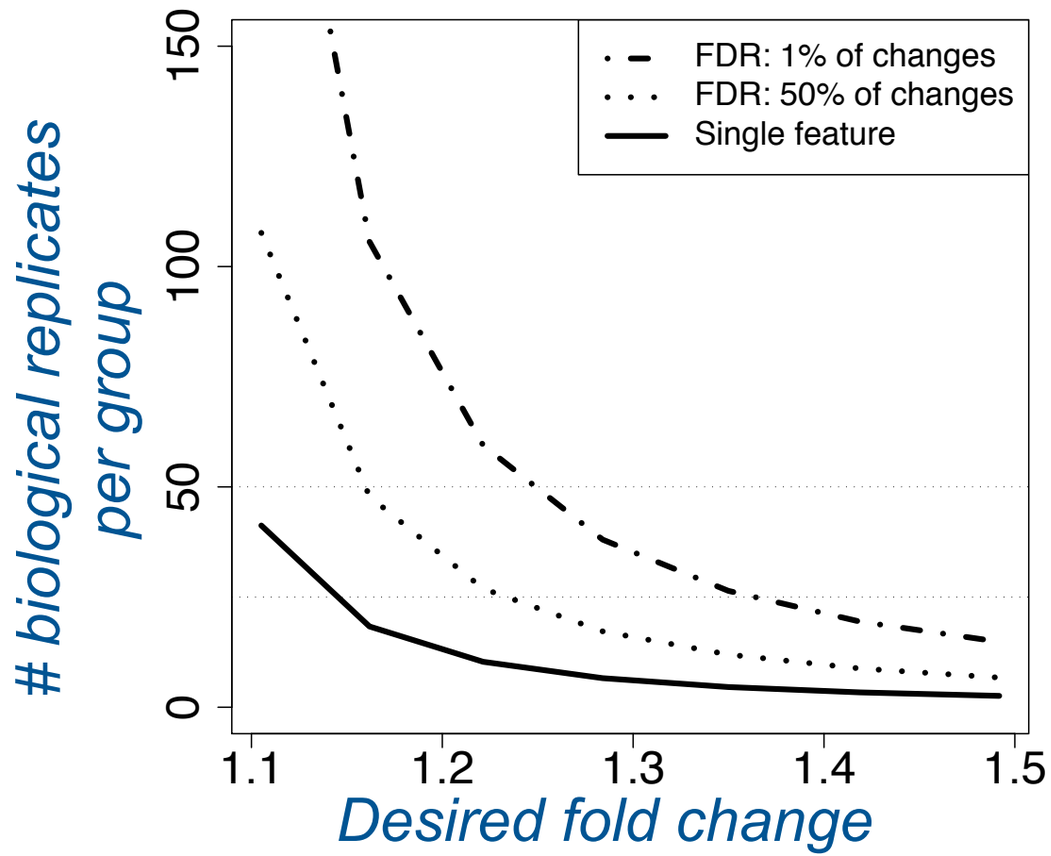
Data processing

↓

Statistical analysis

↓

Experimental design



MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition

Quality control

↓

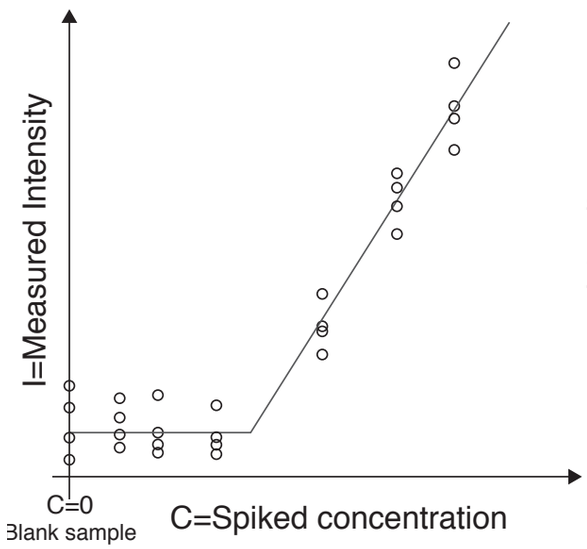
Data processing

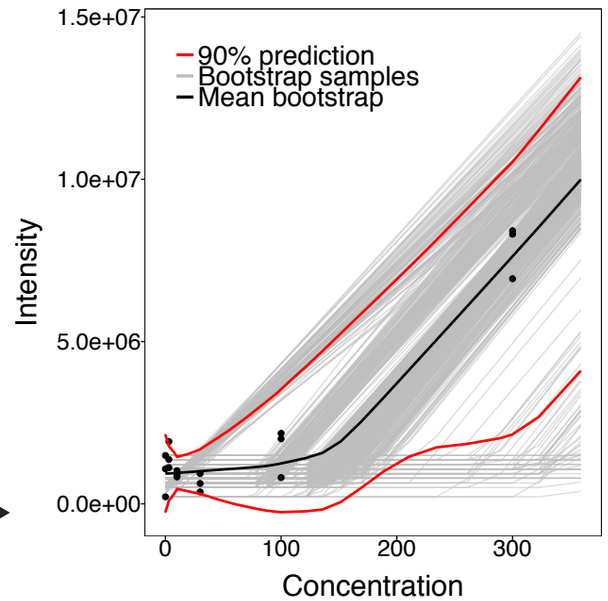
↓

Statistical analysis

↓

Experimental design





MS EXPERIMENT: STATISTICIAN'S VIEW

Analytical method validation

↓

System suitability testing

↓

Experimental design

↓

Data acquisition
Quality control

↓

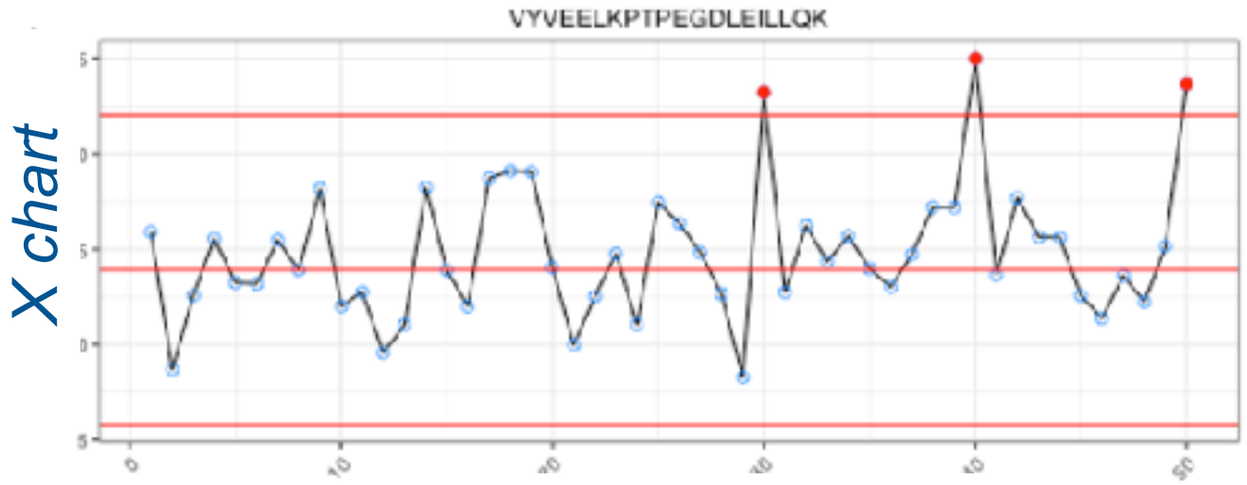
Data processing

↓

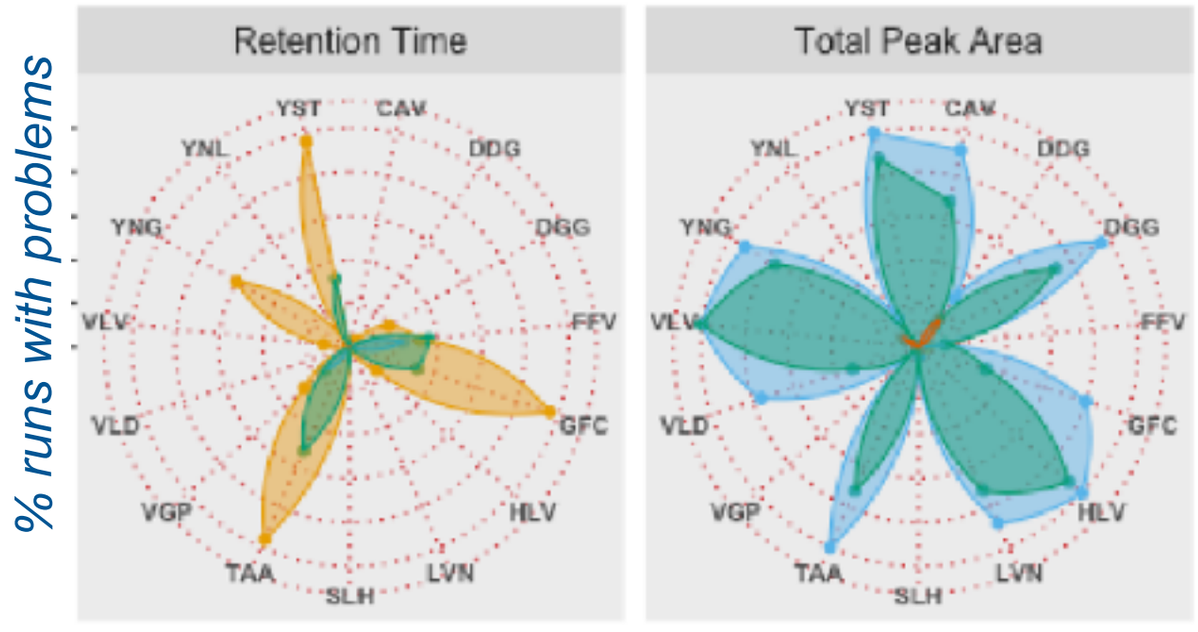
Statistical analysis

↓

Experimental design



Radar plots: sustained drift in mean & variance



MSSTATS

Open-source statistical software for MS-based proteomics

www.msstats.org

MSstats

Statistical Tool For Quantitative Mass Spectrometry-Based Proteomics

HOME MSSTATS LOB/LOD MSSTATSQC DATASETS TRAINING CONTACT

Overview

MSstats is an open-source R package for statistical relative quantification of proteins and peptides in global, targeted and data-independent proteomics. It provides workflows for

News

- The upcoming NEU course (Computation and Statistics for Mass Spectrometry) on May 1-12, 2017. The detailed information will follow shortly.

Skyline external tool

Download MSstats
Downloaded: 13557

Tool Information

Organization: Vitek Lab, Purdue University

Authors: Meena Choi, Cyril Galitzine, Tsung-Heng Tsai, Olga Vitek

Languages: R(3.3.1), C#

More Information: http://www.msstats.org/

14,000 lines of code in base MSstats

Bioconductor

7,775 unique IP downloads since 2013

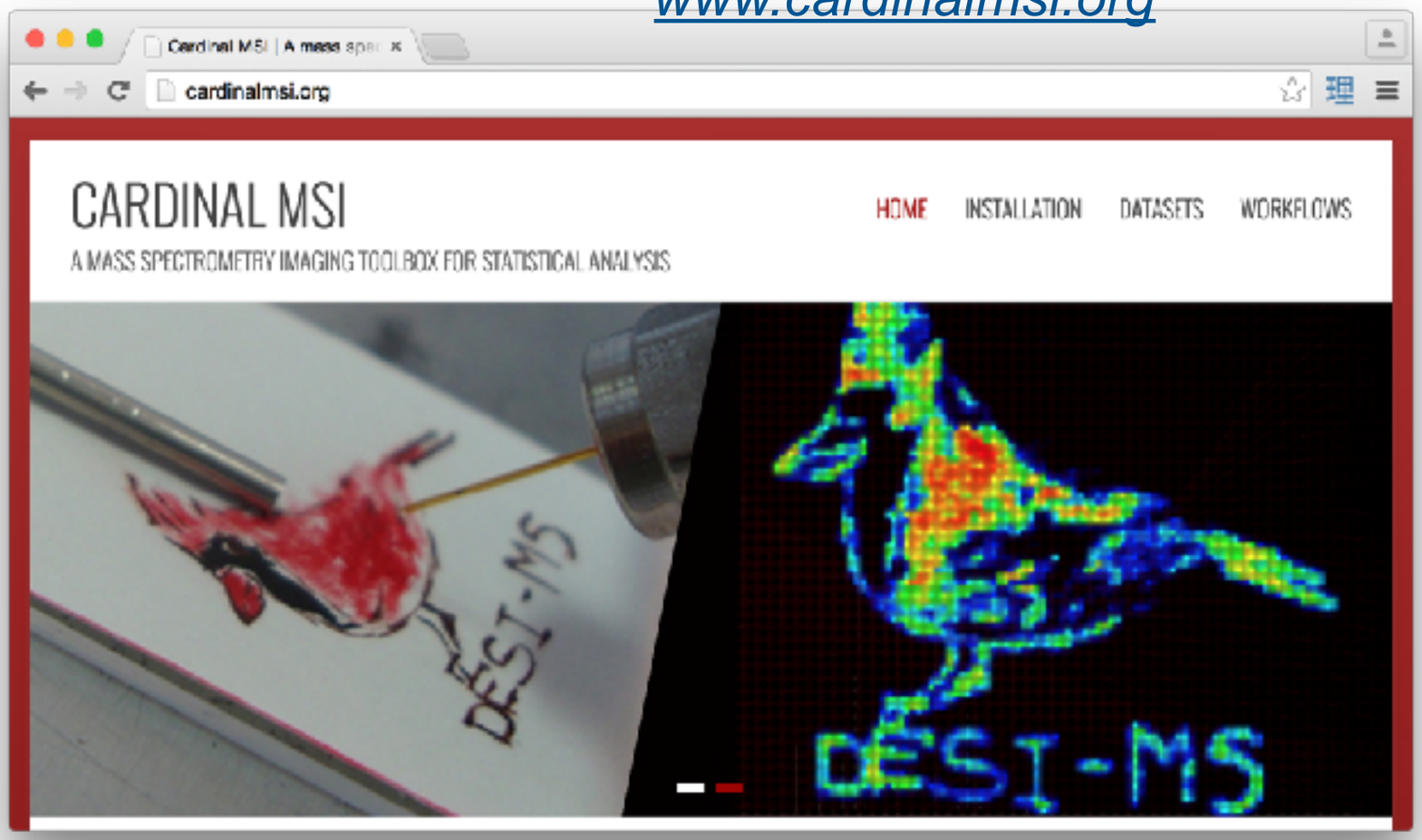
platforms all | downloads top 20% | build ok | commits 0.33

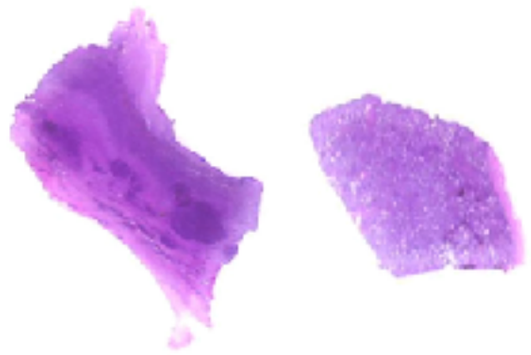
M. Choi *et all Bioinformatics*, 2014.

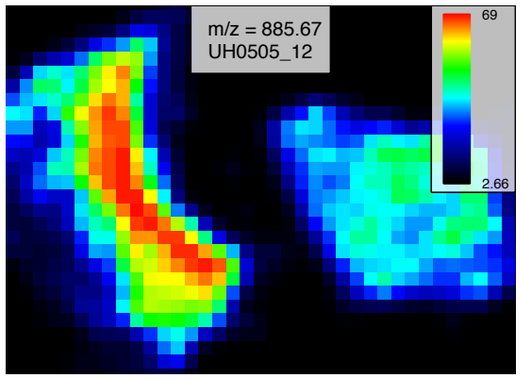
CARDINAL

Open-source statistical software for MSI

www.cardinalmsi.org

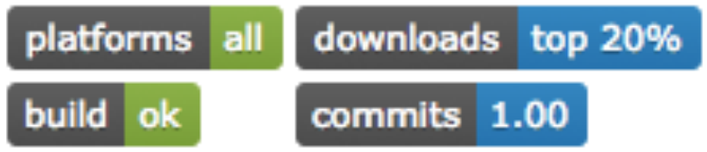






- Over 4,000 unique downloads since public release in 2015
- 2015 John M. Chambers Statistical Software Award

Bioconductor

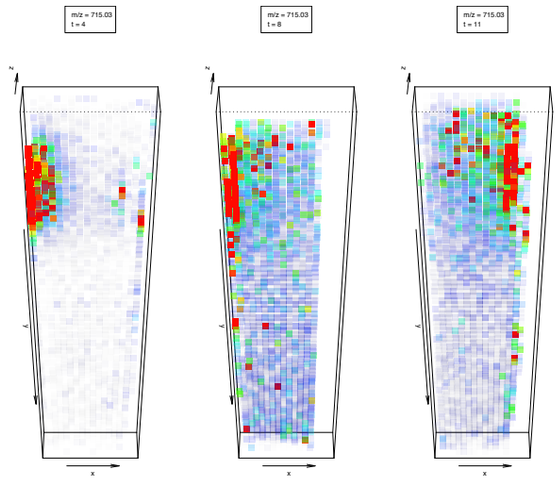


Software	K. Bemis *et all Bioinformatics*, 2015.
Statistical methods	K. Bemis *et all Molecular and Cellular Proteomics*, 2016.
Larger-than memory data	K. Bemis *et all Bioinformatics*, 2017.

PROBLEM: LARGER-THAN-MEMORY DATA

challenges statistical method development

- MS imaging experiments rapidly advancing
  - Increasing mass and spatial resolutions
  - Larger sample sizes, multiple files
- Growing data size poses difficulty for statistics
  - Need to test methods on larger-than-memory data
  - Need to work with domain-specific formats
  - Current R solutions are inflexible



120GB on disk

130GB RAM does not make it

I have a RAM with 32 GB but this will be insufficient.

20 GB, 2 slices

Greg Drazek

Hi Kyle,

I have data from flexImaging 4.1 (20 GB, 2 slices)

fmcp1979@gmail.com

Hi Kyle,
I would like to make a supervised analysis of a "big" set of samples. I have a RAM with 32 GB but this will be insufficient.

Uli Wellner

Dear all,

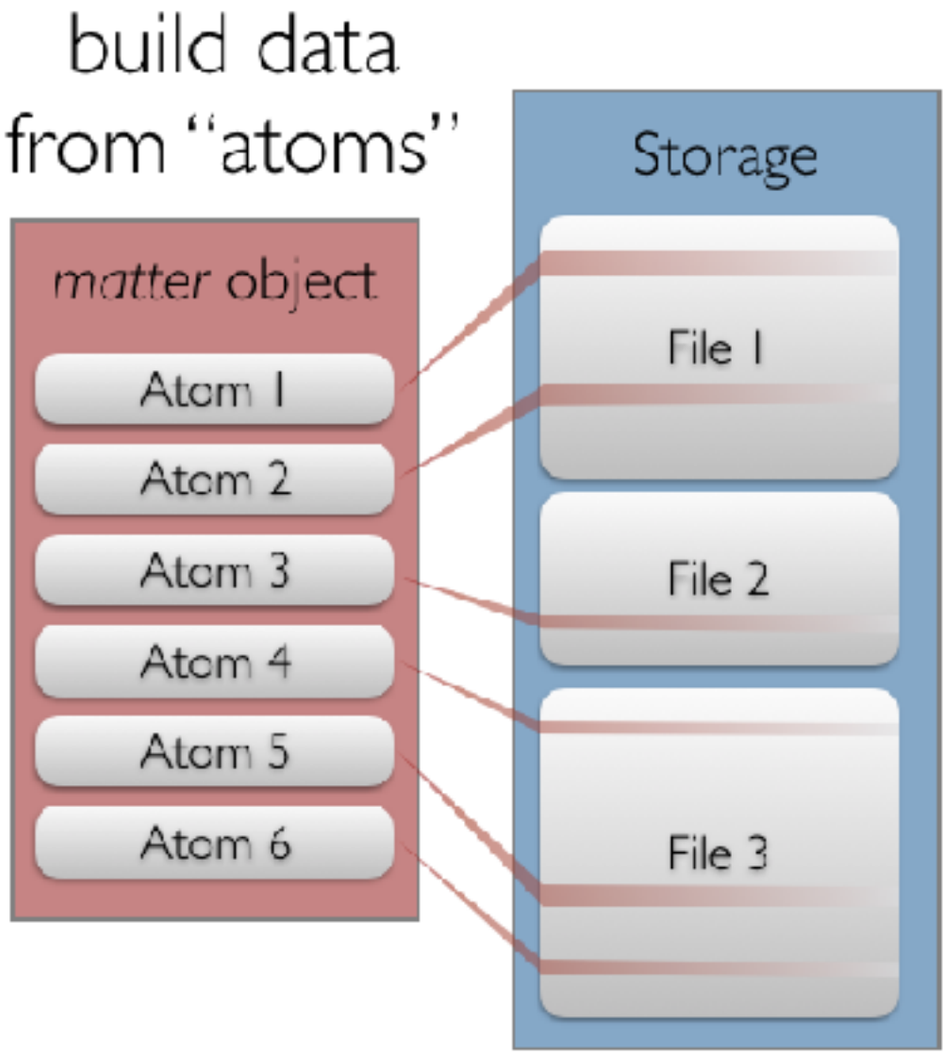
we are trying to ~~load large processed imzml files~~ (120GB on disk) our server with 130GB RAM does not make it

Cardinal help Google group

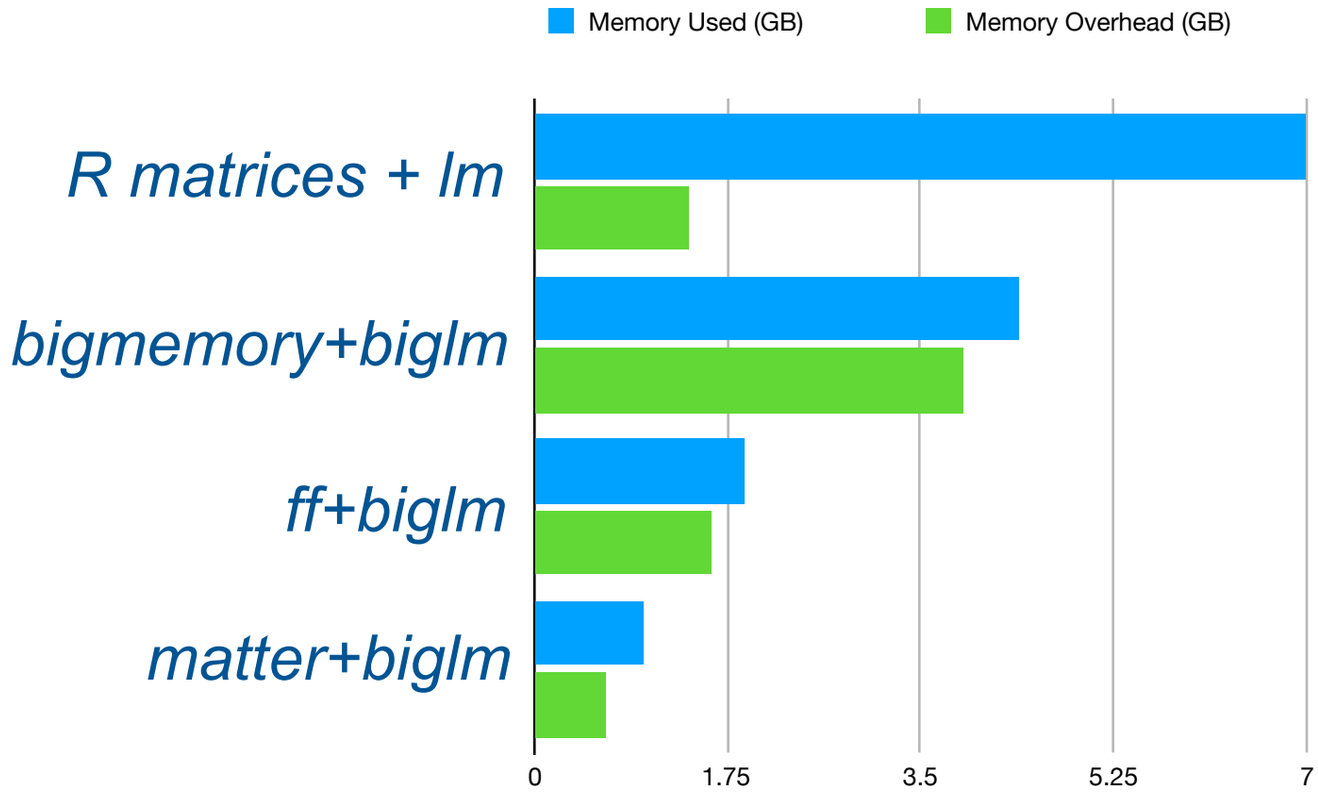
MATTER

Open-source statistical computing with data on disk

- Emphasizes flexibility with a minimal memory footprint



- 1.2 GB dataset
  - N = 15,000,000 observations
  - P = 9 variables
- Linear regression with biglm



- K. A. Bemis, O. Vitek. "matter: an R package for rapid prototyping with larger-than-memory datasets on disk". Bioinformatics, 2017

LESSONS LEARNED

- Many interesting problems!
  - Basic and computational research; pharma
  - PhD students can take time to think
- R is a key benefit
  - Users
  - Developers
  - Open science
- R scripts over Shiny apps
  - Balance broad applicability and specificity
  - Best lab notebook
- Building a community of competent users
  - Programming skills
  - Machine learning skills
  - Understanding biotechnological problems

May Institute
Computation and statistics for mass spectrometry and proteomics

April 30–May 11, 2018, Northeastern University, Boston MA

Organizers : Meena Choi, Brendan MacLean and Olga Vitek

Program Overview

- April 30 – May 2, 2018 : Targeted proteomics with Skyline
- April 30 – May 2, 2018 : Proteomics and metabolomics with OpenMS
- May 2-4, 2018 : Beginner's statistics in R
- May 2-4, 2018 : Advanced R
- May 7-9, 2018 : Statistics for quantitative mass spectrometry
- May 7-9, 2018 : Visualization of biomolecular data
- May 9-11, 2018 : Capstone – case studies in data-independent acquisition (DIA)

NIH deNBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

http://computationalproteomics.ccis.northeastern.edu/

JJ Allaire

Ruedi Aebersold

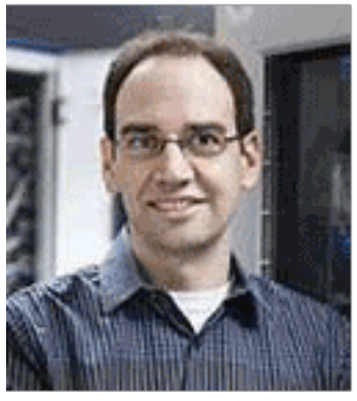
Kylie Bemis

Meena Choi

Ben Collins

Laurent Gatto

Nils Gehlenborg



Oliver Kohlbacher

Mike MacCoss

Brendan MacLean

Olga Vitek

MS IN DATA SCIENCE

Intro to linear algebra/probability; programming/data structures

CS 5800 - Algorithms >

DS 5110 - Introduction to Data Management and Processing >

DS 5220 - Supervised Machine Learning and Learning Theory >

DS 5230 - Unsupervised Machine Learning and Data Mining >

DS 5500 - Information Visualization: Applications in Data Science >

3 electives from area of interest

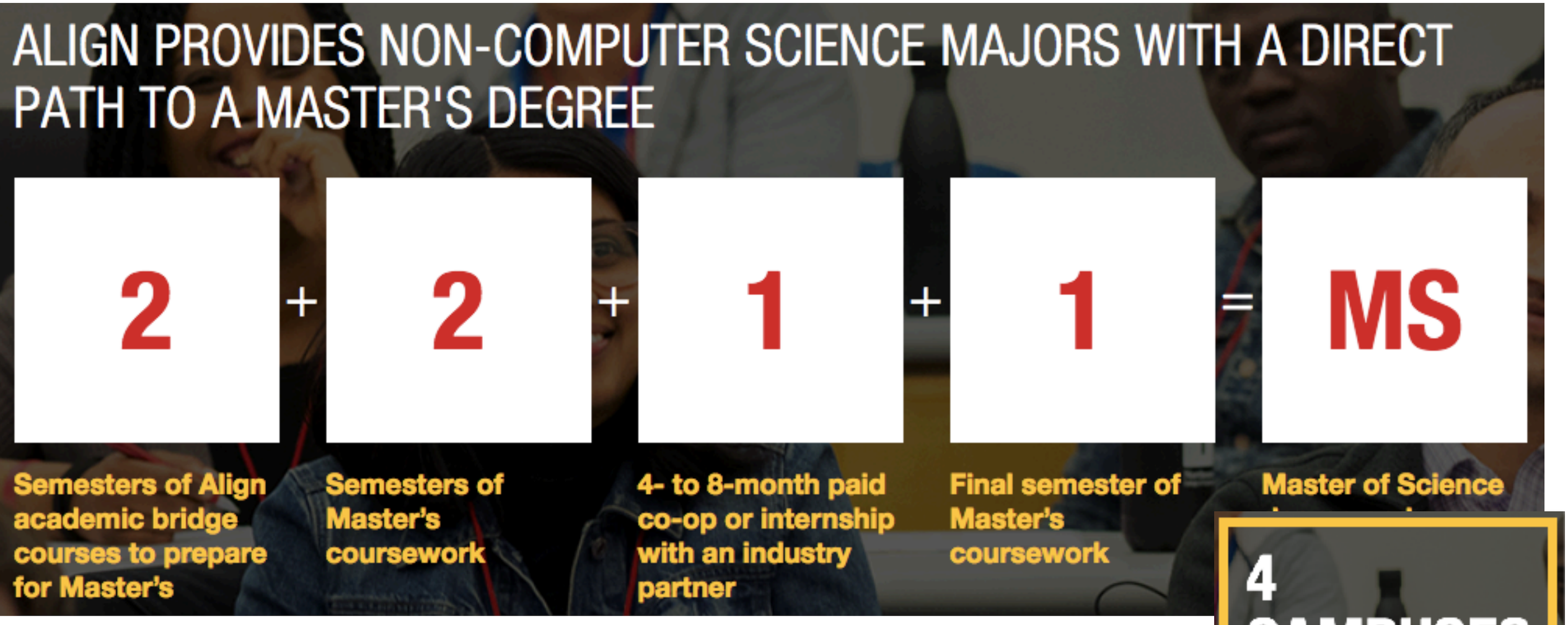
MS IN DATA SCIENCE

Data management and processing

Course expectations, introduction to data science, reproducible research, R Markdown.	Tidying text data and text mining in R
Introduction to ggplot2	Sentiment analysis and topic modeling in R
Grammar of graphics	Introduction to OOP in R, S3 versus S4
Introduction to dplyr	Building R packages, unit tests, reproducibility
Visualization + transformation	
Importing data, tidy data	Working with big data in R
	Parallelization and distributed computing in R
Web scraping, tidying HTML and XML data	Interactivity, advanced graphics, Shiny
Relational data in R, SQL	
Connecting to RDBMS's with dbplyr	Final Projects - presentations in class
Data wrangling	
Introduction to data modeling	
Evaluating models	
Advanced modeling, cross-validation	

ALIGN

MS in CS or DS for non-CS majors



Northeastern University

ACKNOWLEDGEMENTS

Northeastern University

Kylie Bemis
Meena Choi
Eralp Dogu
Dan Guo
April Harry
Ting Huang
Cyril Galitzine
Robert Ness
Sara Taheri
Tsung-Heng Tsai

Purdue University

Graham Cooks
Livia Eberlin
Julia Laskin

University of Washington

Michael MacCoss
Brendan MacLean
Jarrett Egertson

ETH Zurich

Ruedi Aebersold
Tiannan Guo
Ruth Huttenhain
Paola Picotti
Silvia Surinova
Bernd Wollscheid

Uppsala University

Ingela Lanekoff

Mugla University

Eralp Dogu

Support:

NSF
NIH
Sternberg Chair
Canary Center
Roche
Genentech
Eli Lilly