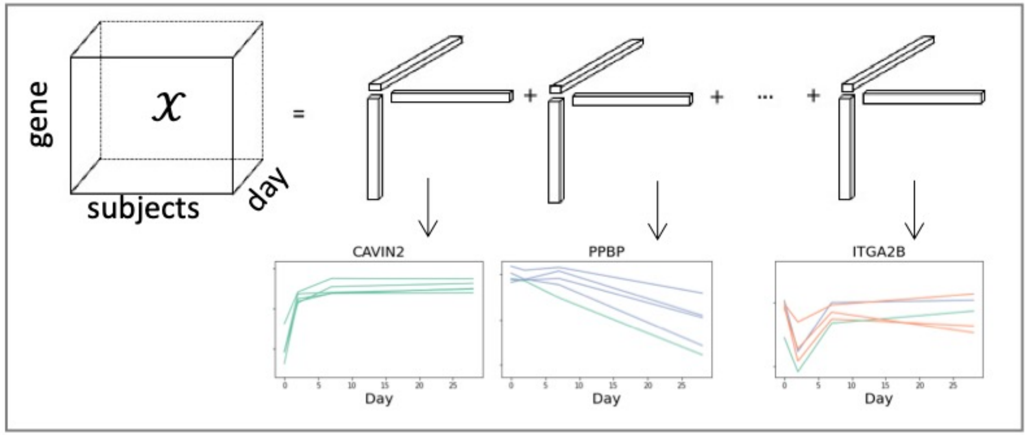
Decomposing a career

Identifying the latent variables in data and in a career trajectory



Anna Konstorum
Research Staff Member
Center for Computing Sciences, Institute for Defense Analyses
UCI Math Dept. (PhD 2015)

March 15, 2023

Overview

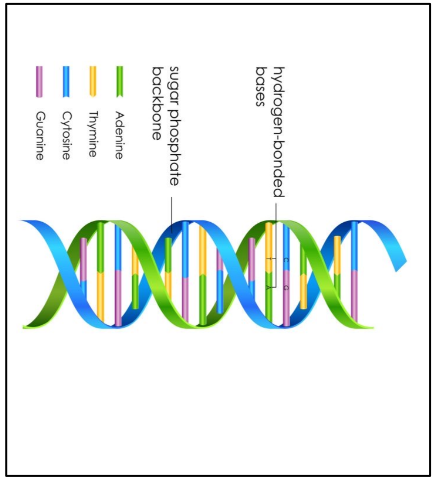
- Career journey
- The math along the way
    - Tensor decomposition of high-dimensional time-course data
    - Multi-view decomposition methods and software
- Background on the Center for Computing Sciences, Institute for Defense analyses
    - And similar opportunities for government-funded research
- Tips for grad school/career success (not just survival!)

Career journey

Education

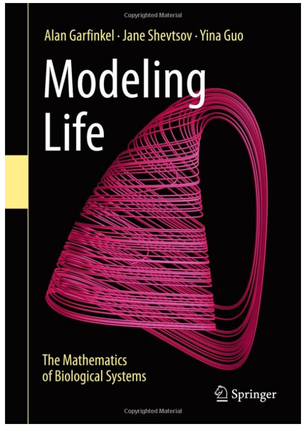
B.Sc. Biology (Hon)

- Took math classes 'for fun'



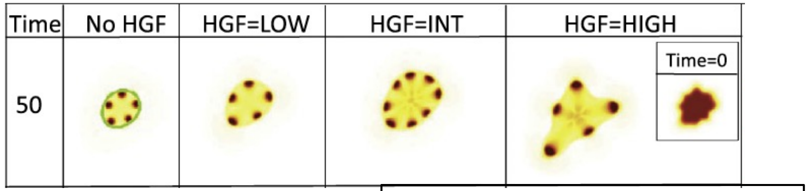
M.S. Physiological sciences

- Bioinformatics focus
- Took a course in *dynamical systems for the life sciences*


Copyrighted Material
Alan Garfinkel · Jane Shevtsov · Yina Guo
Modeling Life
The Mathematics of Biological Systems
Springer
Copyrighted Material

M.S. and Ph.D. in Mathematics

- Affiliated with Center for Complex Biological Systems (CCBS)
- Focus on PDE models in tumor growth
- Took a wide range of math classes (not just what I 'needed' for my research)



Time	No HGF	HGF=LOW	HGF=INT	HGF=HIGH
50				Time=0

(Konstorum and Lowengrub, 2018)

Career journey

PostDoctoral Fellow
(UConn Health)

- Worked on logical models of intracellular processes
- Started to 'return' to data science
  - Multi-omics data integraion
  - Comparison of dimension reduciton tehcniques for immune datasets

Career journey

IDA | CENTER FOR COMPUTING SCIENCES

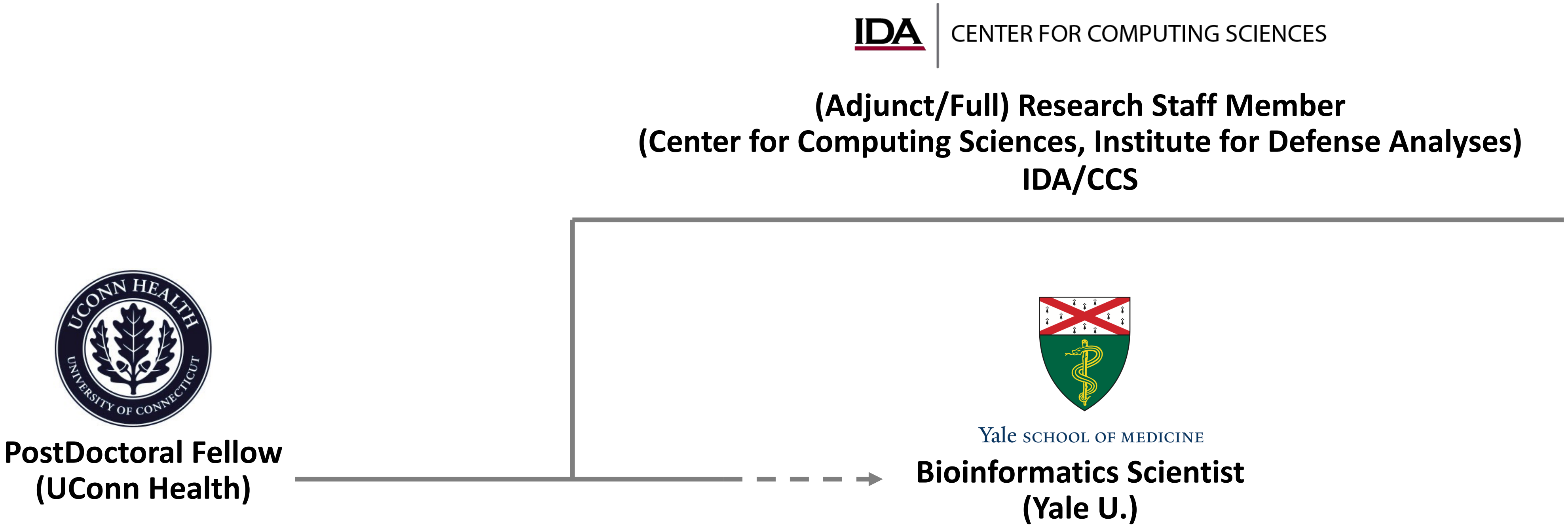
(Adjunct/Full) Research Staff Member
(Center for Computing Sciences, Institute for Defense Analyses)
IDA/CCS

- Deepened and broadened by research data science skills to include analysis of intelligence data using techniques including:
  - Natural Language Processing (NLP)
  - Network and graph analytics
  - Matrix & tensor decompositions

UCONN HEALTH
UNIVERSITY OF CONNECTICUT

PostDoctoral Fellow
(UConn Health)

Career journey



IDA | CENTER FOR COMPUTING SCIENCES

(Adjunct/Full) Research Staff Member
(Center for Computing Sciences, Institute for Defense Analyses)
IDA/CCS

UCONN HEALTH · UNIVERSITY OF CONNECTICUT

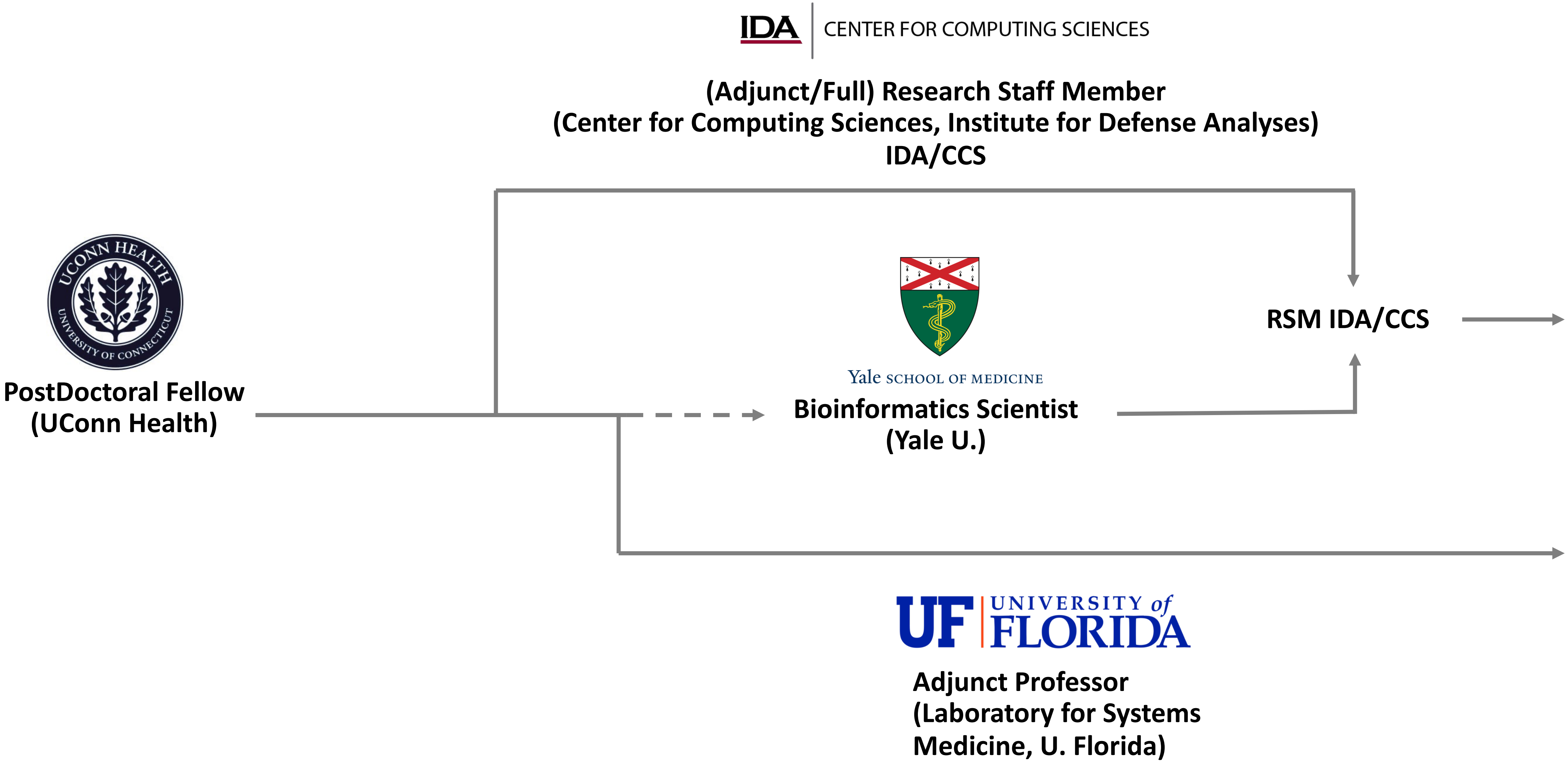
PostDoctoral Fellow
(UConn Health)

Yale SCHOOL OF MEDICINE

Bioinformatics Scientist
(Yale U.)

- Continued work on developing tensor decomposition methods for time-course, immune-profiling datasets
- Began working on multi-view decomposition algorithm and software development
- Involved in large collaborative projects for immune response to COVID19 and vaccination

Career journey



IDA | CENTER FOR COMPUTING SCIENCES
UCONN HEALTH UNIVERSITY OF CONNECTICUT
Yale SCHOOL OF MEDICINE
UF UNIVERSITY of FLORIDA

(Adjunct/Full) Research Staff Member
(Center for Computing Sciences, Institute for Defense Analyses)
IDA/CCS

PostDoctoral Fellow
(UConn Health)

Bioinformatics Scientist
(Yale U.)

RSM IDA/CCS

Adjunct Professor
(Laboratory for Systems
Medicine, U. Florida)

Career journey

IDA | CENTER FOR COMPUTING SCIENCES

RSM IDA/CCS

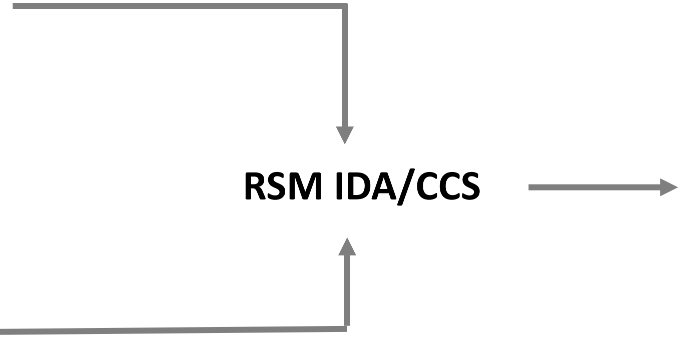
- Continue to work on algorithms for multi-view and tensor decomposition of complex data in defense and the natural sciences, as well as have expanded my research into other fields of interest to the CCS community.
- Continue to interact and collaborate with the academic community on methods development and application to biological questions.

UF | UNIVERSITY *of* FLORIDA

Adjunct Professor (Laboratory for Systems Medicine, U. Florida)

Career journey

IDA | CENTER FOR COMPUTING SCIENCES



RSM IDA/CCS

- Continue to work on algorithms for multi-view and tensor decomposition of complex data in defense and the natural sciences, as well as have expanded my research into other fields of interest to the CCS community.
- Continue to interact and collaborate with the academic community on methods development and application to biological questions.

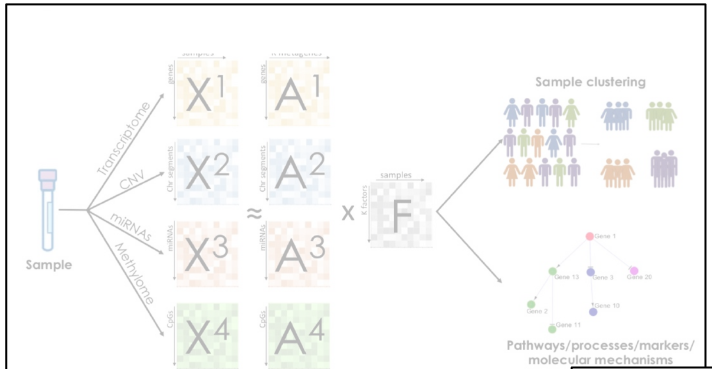
There exists a strong synergy by performing algorithm development in different types of research environments (academia, government) and working on different mission problems; there also exists a synergy by moving between the application communities (which have different standards and 'requests'), and the mathematics community.

We need to hear from and integrate both sides of the aisle!

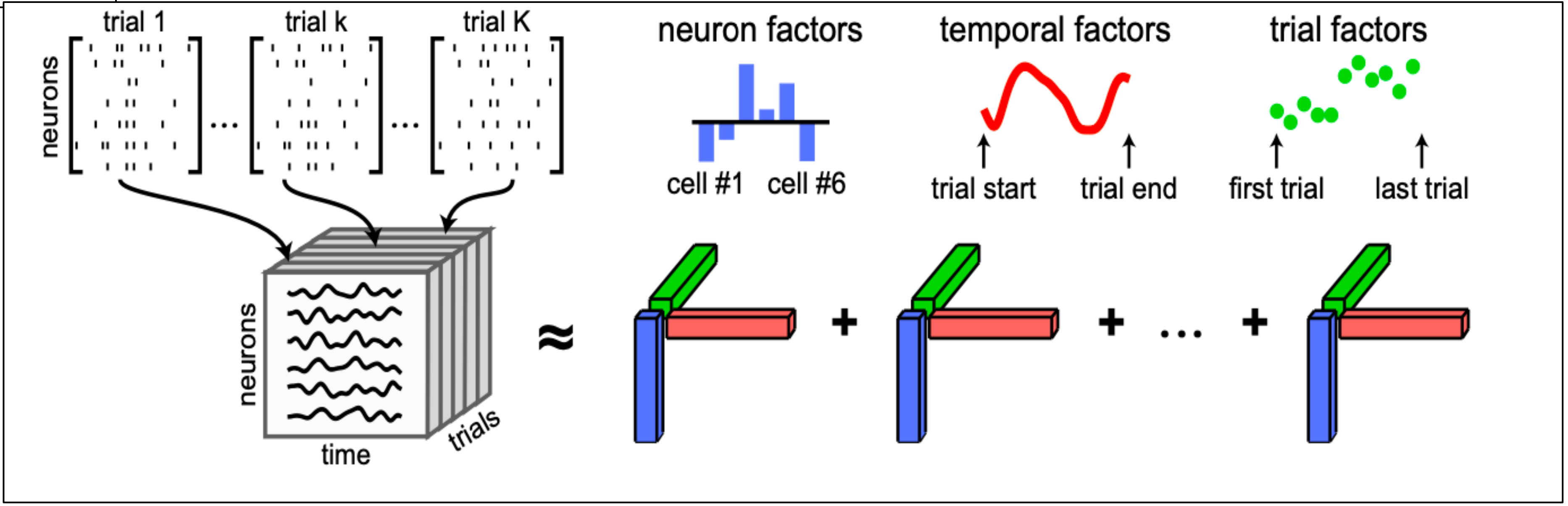
UF | UNIVERSITY of FLORIDA

Adjunct Professor (Laboratory for Systems Medicine, U. Florida)

Matrix and tensor decomposition for complex data



- Tensors (multi-index arrays) can hold more complex data than matrices.
- Tensor decomposition can yield low-dimensional representations of the data.
- Very active research field, many challenges including:
  - Model selection, data imputation, computation, existence/uniqueness....



[Williams et al., 2018]

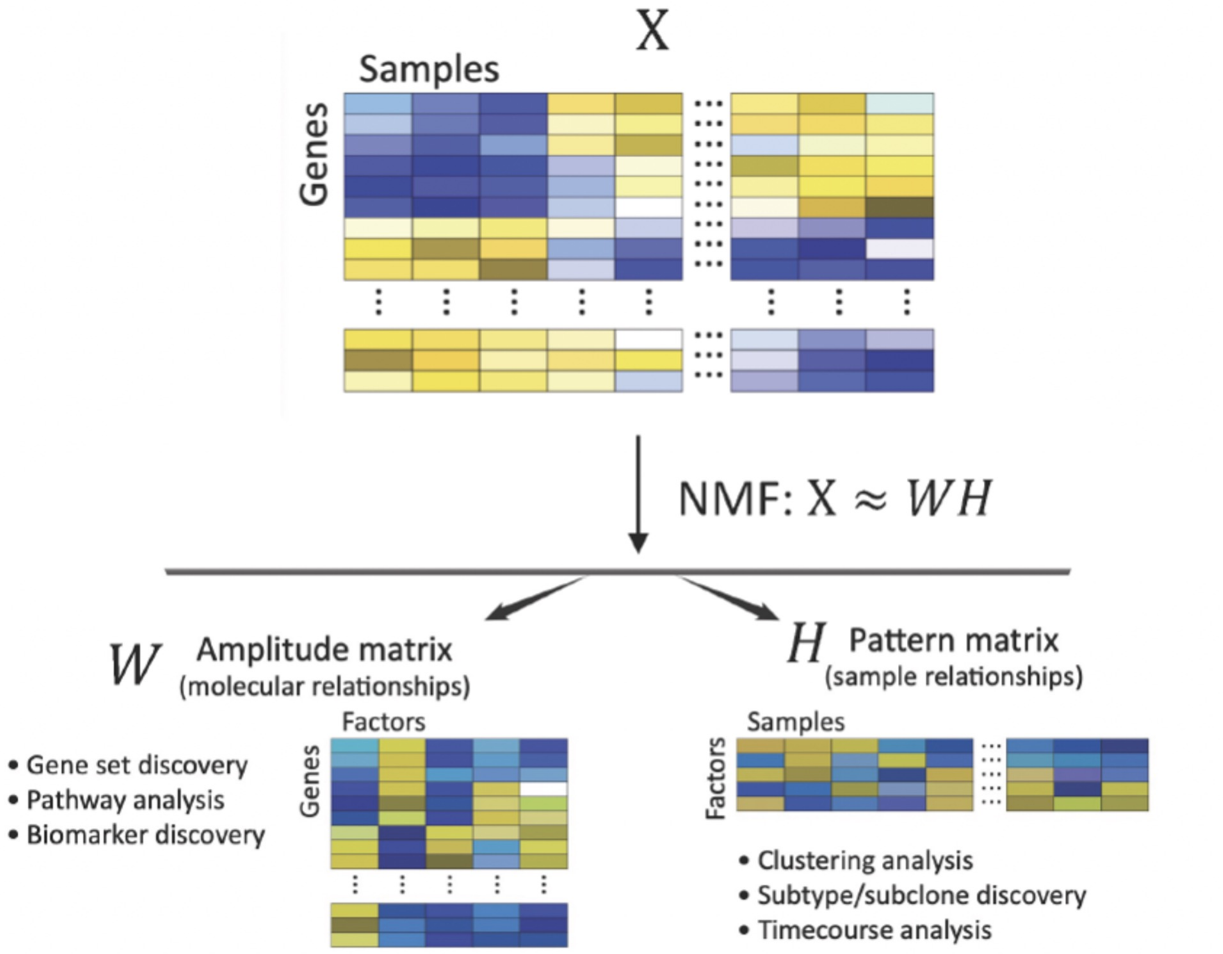
Non-negative matrix factorization (NMF)

Recall that for a non-negative matrix $\boldsymbol{X}_{nxm}$, non-negative matrix factorization (NMF) finds a rank *R* decomposition,

$$\mathbf{X}_{n\times m} \approx \mathbf{W}_{n\times R}\mathbf{H}_{R\times m} = \sum_{i=1}^{R} \mathbf{w}^i \circ \mathbf{h}_i$$

If $\boldsymbol{X}_{nxm}$ is a gene by sample matrix for an experiment, then for factor *i*,

- The column $\boldsymbol{w}^i$ represents a weighted set of co-expressed genes.
- The row $\boldsymbol{h}_i$ represents the strength of presence of that set in a given sample



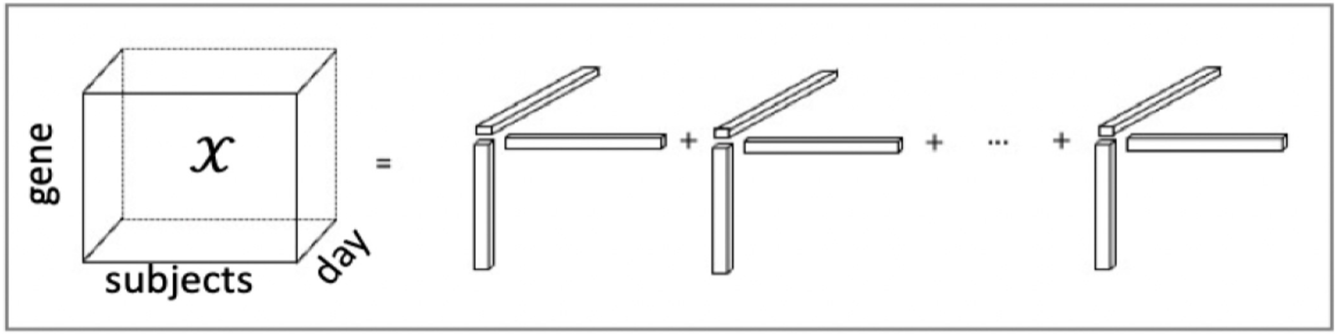
- Gene set discovery
- Pathway analysis
- Biomarker discovery

- Clustering analysis
- Subtype/subclone discovery
- Timecourse analysis

(Stein-O'Brien et al., 2018)

Non-negative CP tensor decomposition (NCPD)

- Non-negative CP decomposition (NCPD) extends the concept of representing a dataset as the sum of rank-one components to ≥ 2 dimensions.



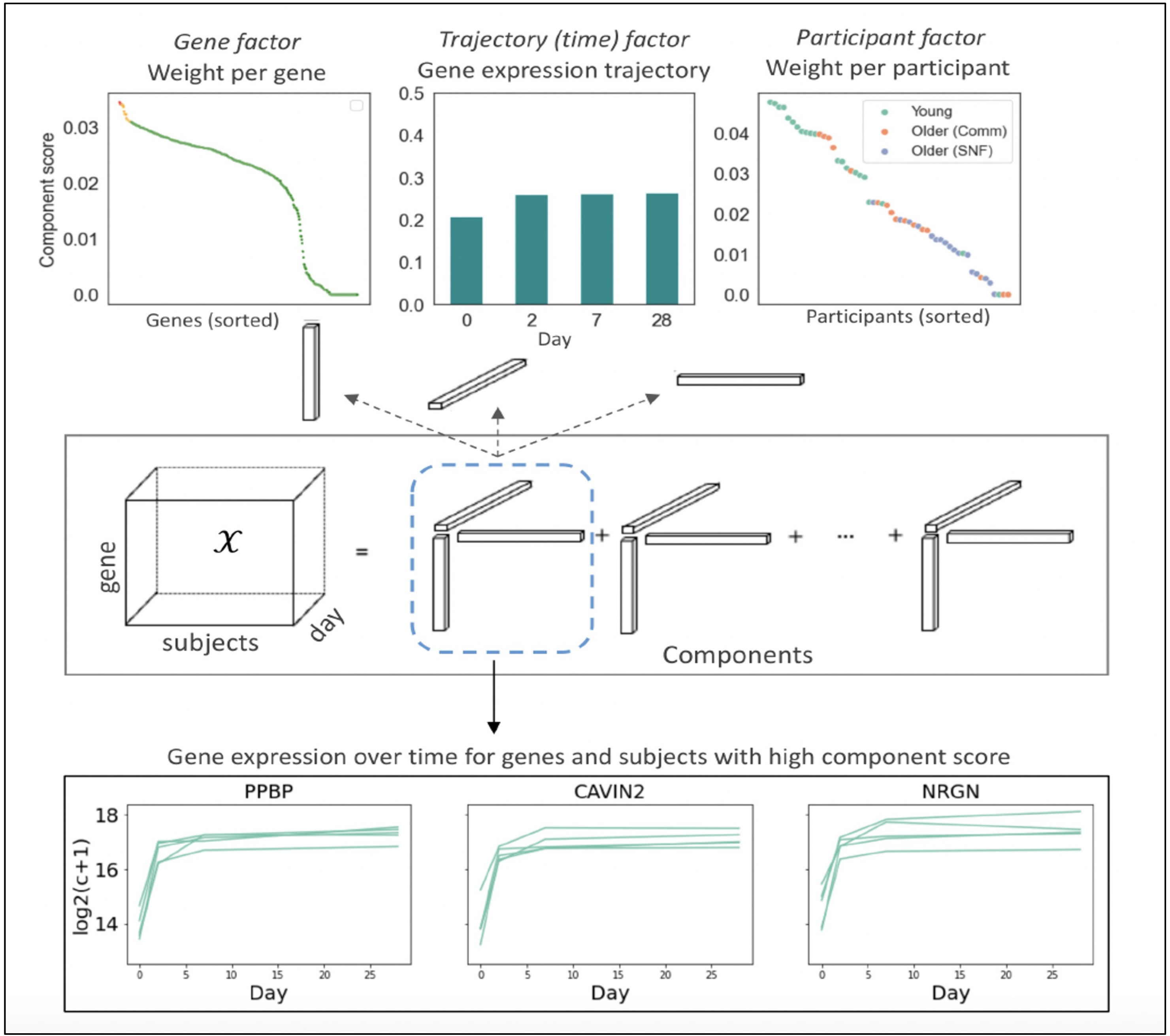
- For a tensor $\mathcal{X}$ of feature x subject x time data, we can represent $\mathcal{X}$ as

$$\mathcal{X} \approx [[\lambda; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}]] \equiv \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \mathbf{a}_r^{(3)},$$

where $\mathbf{a}_r^{(i)} \geq 0$ and $||\mathbf{a}_r^{(i)}||_2 = 1$ for $i = \{1, 2, 3\}$, and λ is the normalization constant.

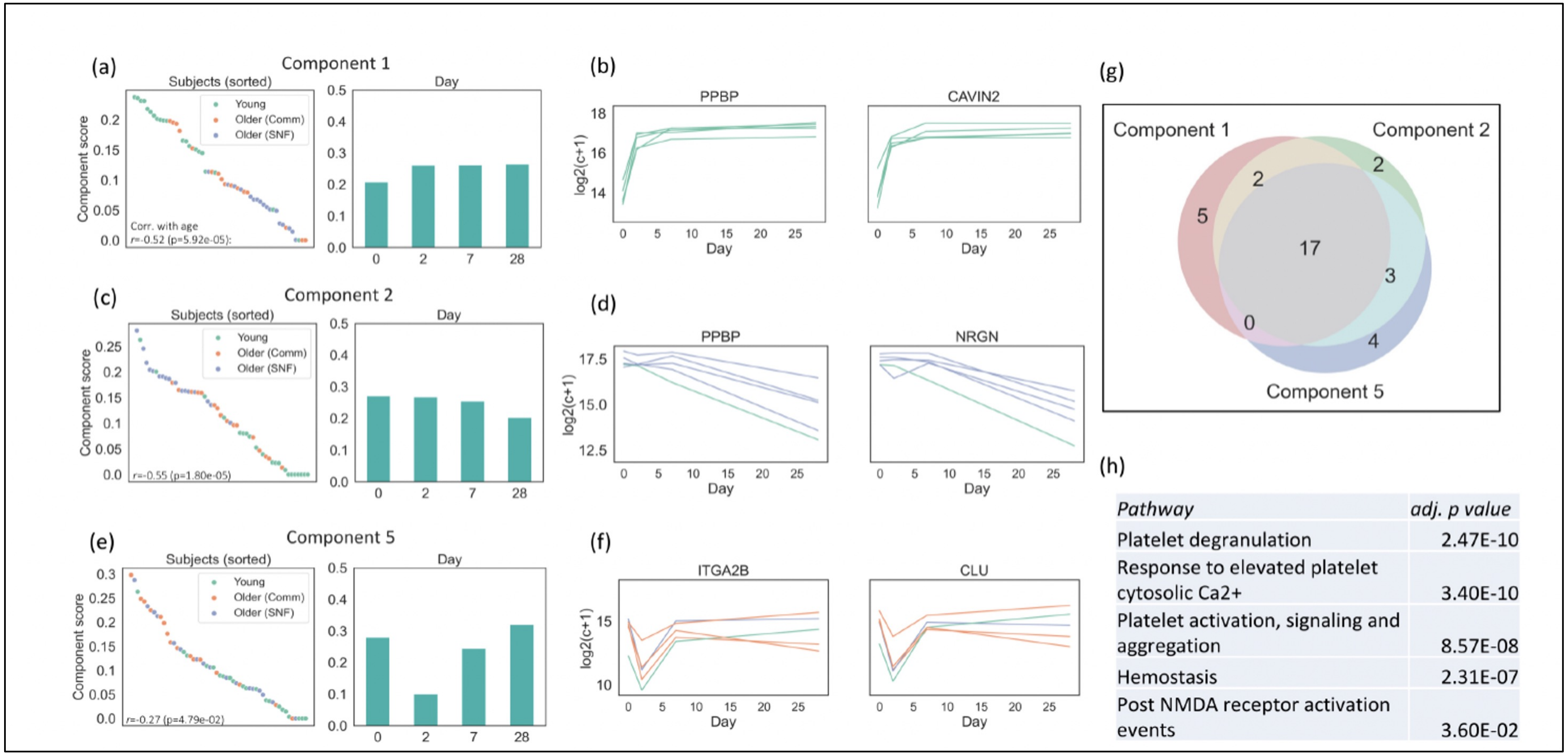
Non-negative CP tensor decomposition (NCPD)

- One can interpret each component from an associated NCPD as follows:
    - The time factor vector gives a *canonical trajectory*.
    - The gene factor vector scores the set of genes (*canonical features*) that follow the canonical trajectory.
    - The participant factor vector scores the participants whose *canonical features* follow the *canonical trajectory*.
- Powerful way to identify similar feature trajectories that may be associated with specific groups of individuals.

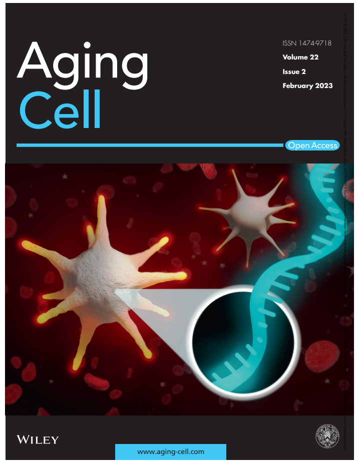


NCPD in the wild

- An NCPD of platelet transcription data from vaccinated individuals from different age groups showed that a canonical feature-set of platelet-activation genes (g,h) responded with different canonical trajectories in different age groups (a-f).
- These differences may contribute to the observed increase in pro-inflammatory diseases in older adults.



Pathway	*adj. p value*
Platelet degranulation	2.47E-10
Response to elevated platelet cytosolic Ca2+	3.40E-10
Platelet activation, signaling and aggregation	8.57E-08
Hemostasis	2.31E-07
Post NMDA receptor activation events	3.60E-02



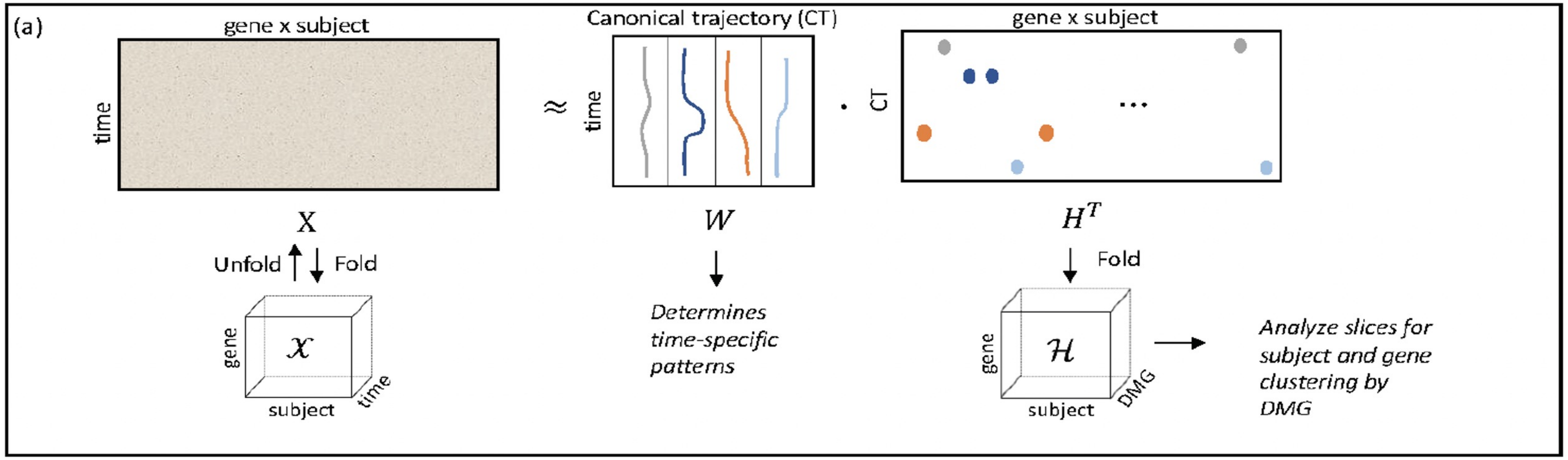
Featured on the Feb., 2023 cover of *Aging Cell*

Konstorum, A. et al. (2023). Platelet response to influenza vaccination reflects effects of aging. *Aging Cell*, 22, e13749. https://doi.org/10.1111/acel.13749

NCPD: current challenges and future directions

- Finding an *optimal** NCPD model for a dataset if very difficult. The model needs to succeed with the standard optimization criteria, but also be interpretable.
  - Working on improved criteria for model evaluation that emphasizes interpretability.
  - Building a generative data model in order to mathematically define concepts such as *patterns* and *trajectories*, and then using this new language to understand what is the *optimal/best* decomposition approach.
  - Creating novel decompositions of data to maximize interpretability inspired by the generative model structure.
  - Asking deeper questions:
    - There exists a globally optimal** NCPD solution[1] , how do we incorporate this into our more holistic consideration of optimality?

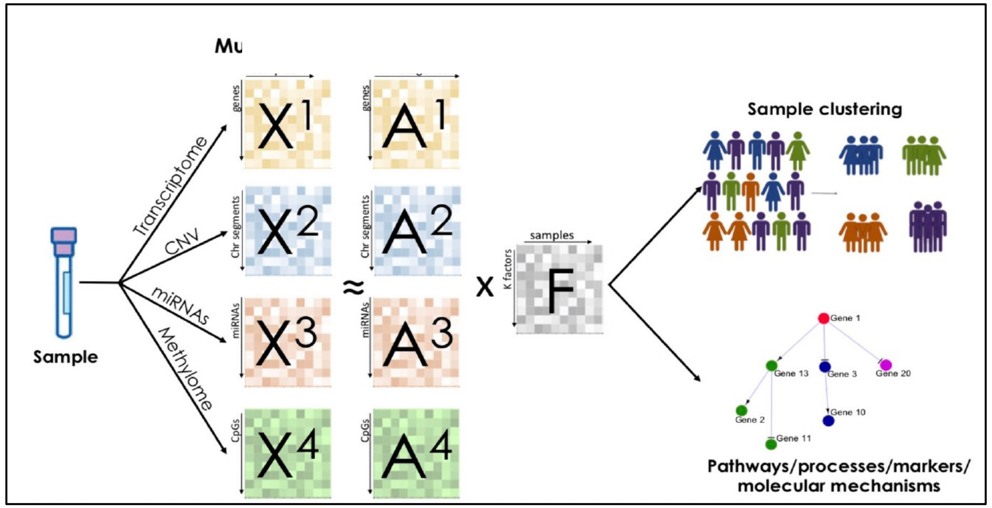
My favorite kind of math: math that is still being developed hand-in-hand with application needs....



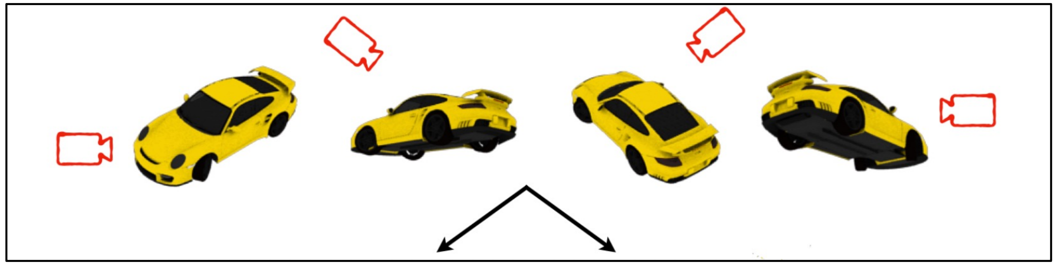
[1](Lim and Comon, 2009)

Matrix and tensor decomposition for complex data

What is multi-view embedding?

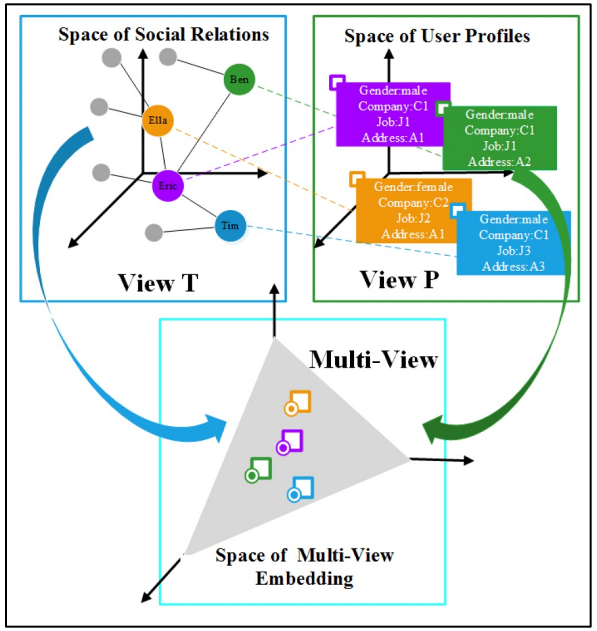


[Cantini et al., 2021]



[https://bair.berkeley.edu/blog/2017/09/05/unified-3d/]

- Multi-view embedding can be thought as an extension to Principal Components Analysis (PCA) for situations where one has multiple datasets (*views)* of a sample.
- As in PCA, the goal of multi-view embedding is to embed the samples and features into a common space for follow-on analysis (clustering, important feature extraction...)



[Yang et al., 2017]

Multi-view embedding as an optimization problem

X_k (*n x l_k*) correspond to a sample - by - feature matrix for view *i. (n*:=number of samples*, l_i*:=number of features for kth block)
F_k (*n x r*) correspond to 'local' (or, block) embeddings of samples, akin to a principal components (PCs) for each block (r*:=number of components*)
Q_k (*l x r*) are the loadings on the Xi: indicate the weight ('importance') of the contribution of each feature to the embedding.

$$\begin{tikzcd}
X_1 \approx F_1 Q_1^T \arrow[dr] & X_2 \approx F_2 Q_2^T \arrow[d] & \cdots & X_K \approx F_K Q_K^T \arrow[dll] \\
 & X \approx F Q^T & &
\end{tikzcd}$$

Want to find a global loadings matrix *Q*, such that the local embeddings F*$_i$* contribute *optimally* to the global embedding *F*

X (*n x (l_1+l_2+...l_K)*) is the concatenated data matrix ***X* = [*X_1* | *X_2* | ... | *X_K***]
F *(n x r)* is the global scores embeddings matrix
Q *((l_1+l_2+...l_K) x r) is the global loadings matrix*

Multi-view embedding as an optimization problem

X_k (*n x l_k*) correspond to a sample - by - feature matrix for view *i. (n*:=number of samples*, l_i*:=number of features for kth block)
F_k (*n x r*) correspond to 'local' (or, block) embeddings of samples, akin to a principal components (PCs) for each block (r*:=number of components*)
Q_k (*l x r*) are the loadings on the Xi: indicate the weight ('importance') of the contribution of each feature to the embedding.

$$\begin{tikzcd}
F_1 \approx X_1 Q_1 \arrow[dr] & F_2 \approx X_2 Q_2 \arrow[d] & \cdots & F_K \approx X_K Q_K \arrow[dll] \\
 & F \approx XQ & &
\end{tikzcd}$$

Want to find a global loadings matrix *Q*, such that the local embeddings F*$_i$* contribute *optimally* to the global embedding *F*

X (*n x (l_1+l_2+...l_K)*) is the concatenated data matrix ***X* = [*X_1* | *X_2* | ... | *X_K***]
F *(n x r)* is the global scores embeddings matrix
Q *((l_1+l_2+...l_K) x r) is the global loadings matrix*

Multiple co-Inertia Analysis (MCIA)

X_k (*n x l_k*) correspond to a sample - by - feature matrix for view *i. (n*:=number of samples*, l_i*:=number of features for k^th block)
F_k (*n x r*) correspond to 'local' (or, block) embeddings of samples, akin to a principal components (PCs) for each block (r*:=number of components*)
Q_k (*l x r*) are the loadings on the Xi: indicate the weight ('importance') of the contribution of each feature to the embedding.

$$\begin{tikzcd} F_1 \approx X_1 Q_1 \arrow[dr] & F_2 \approx X_2 Q_2 \arrow[d] & \cdots & F_K \approx X_K Q_K \arrow[dll] \\ & F \approx XQ & & \end{tikzcd}$$

Want to find a global loadings matrix *Q*, such that the local embeddings F*$_i$* contribute *optimally* to the global embedding *F*

The MCIA algorithm interprets the problem as the following optimization function:

$$argmax_{\mathbf{q}_1^i \ldots \mathbf{q}_k^i \ldots \mathbf{q}_K^i} \sum_{k=1}^{K} cov^2\left(\overset{\boldsymbol{F_k^i}}{\overset{\downarrow}{\mathbf{X}_k^i \mathbf{q}_k^i}}, \overset{\boldsymbol{F^i}}{\overset{\downarrow}{\mathbf{X}^i \mathbf{q}^i}}\right)$$

i corresponds to the ith component

Multiple co-Inertia Analysis (MCIA)

Current implementations of MCIA have at least some of the following *features*:

- Are not optimized for scale
- Poorly documented
- Are incorrect in some preprocessing steps
    - Have few choices for pre-processing
- Do not have modular code
    - Do not allow for algorithmic extension if novel theoretical ideas are proposed for MCIA
- Are not maintained

$$argmax_{\mathbf{q}_1^{\mathrm{i}} \ldots \mathbf{q}_k^{\mathrm{i}} \ldots \mathbf{q}_K^{\mathrm{i}}} \sum_{k=1}^{K} cov^2(\mathbf{X}_k^i \mathbf{q}_k^i, \mathbf{X}^i \mathbf{q}^i)$$

Multiple co-Inertia Analysis (MCIA)

Current implementations of MCIA have at least some of the following *features*:

- Are not optimized for scale
- Poorly documented
- Are incorrect in some preprocessing steps
  - Have few choices for pre-processing
- Do not have modular code
  - Do not allow for algorithmic extension if novel theoretical ideas are proposed for MCIA
- Are not maintained

$$\cancel{argmax_{\mathbf{q}^i_1..\mathbf{q}^i_k..\mathbf{q}^i_K} \sum_{k=1}^{K} cov^2(\mathbf{X}^i_k\mathbf{q}^i_k, \mathbf{X}^i\mathbf{q}^i)}$$

nipalsMCIA: an efficient implementation of MCIA

In *nipalsMCIA*[1], we are developing a software package that is:

- Closely aligned with the theory behind MCIA
- Efficient
    - Can handle large volume multi-view data, such as single cell data
- Has multiple options for pre-processing and deflation parameters that are well-explained.
- Is designed for both *use* and *extension*
    - We hope that its development can help with extension of the theory to handle application topics such as missing data or higher-complexity data sets
- Well-supported

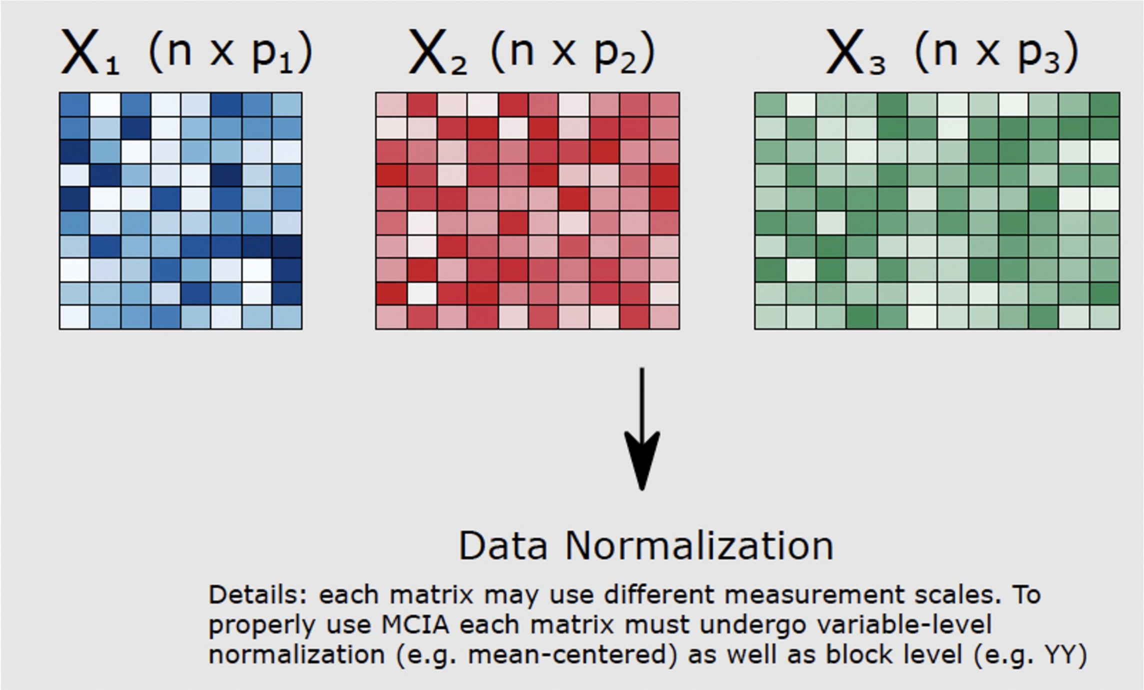
Highly-collaborative effort between labs at Yale U., UCSD, Tufts, Northwestern, CCS (mostly student-driven!)

[1] https://github.com/Muunraker/nipalsMCIA

nipalsMCIA: an efficient implementation of MCIA

nipalsMCIA uses an ad-hoc extension to the Non-linear Iterative Partial Least Squares (NIPALS) algorithm, which can find the first *n* principal components without performing an SVD and has been proven to converge when used for MCIA.

- Earlier methods relied on a full eigendecomposition for each component.
- Can theoretically handle missing data
- Is fast



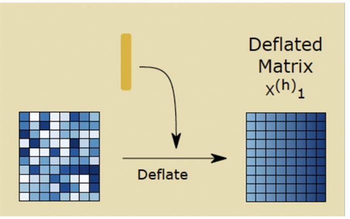
Details: each matrix may use different measurement scales. To properly use MCIA each matrix must undergo variable-level normalization (e.g. mean-centered) as well as block level (e.g. YY)

Image credit: Joaquin Reyna (UCSD)

nipalsMCIA: an efficient implementation of MCIA

$$argmax_{\mathbf{q}_1^i..\mathbf{q}_k^i..\mathbf{q}_K^i} \sum_{k=1}^{K} cov^2(\mathbf{X}_k^i \mathbf{q}_k^i, \mathbf{X}^i \mathbf{q}^i)$$

After convergence, deflate data matrices, and restart process



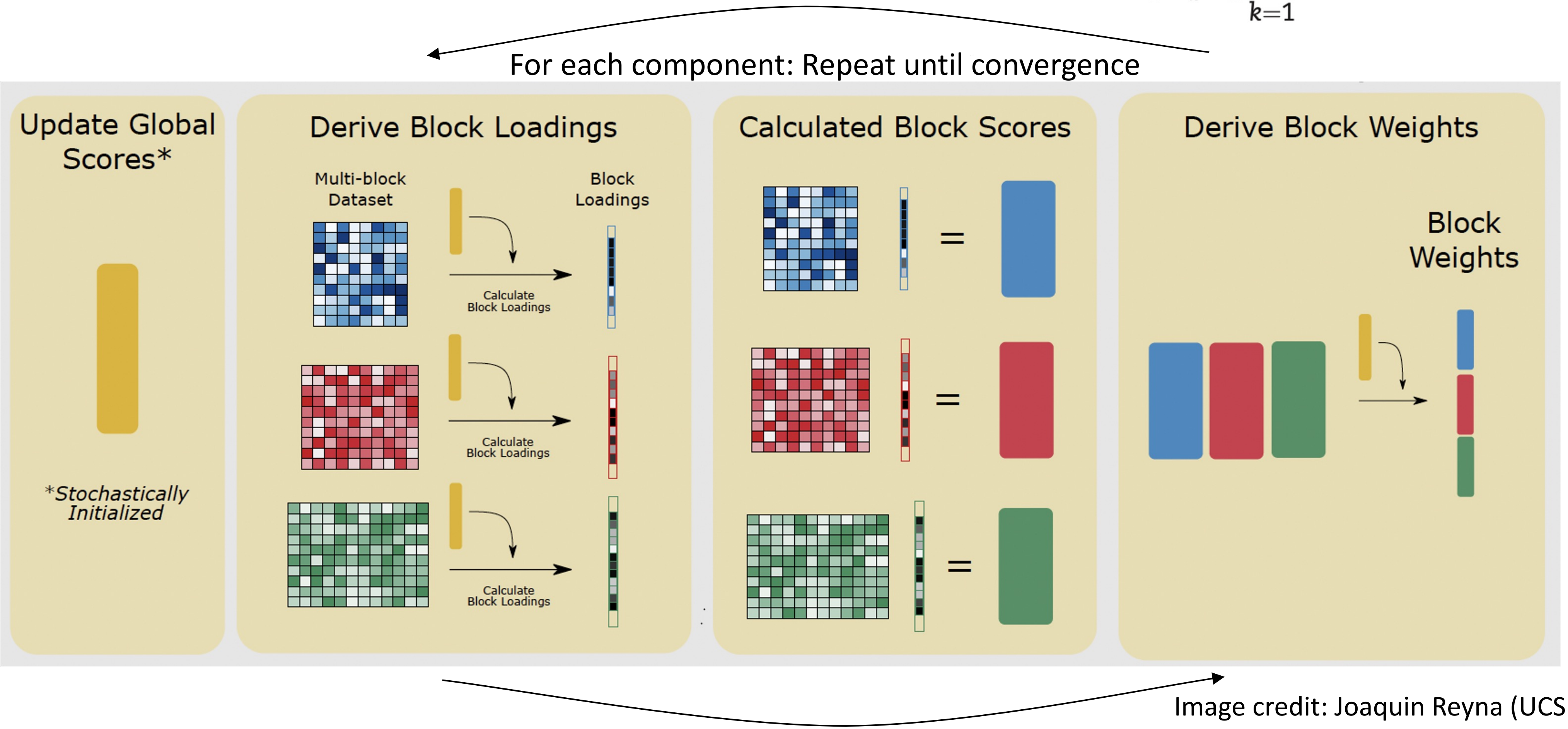
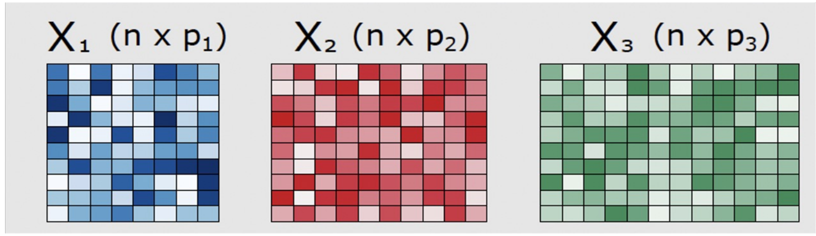


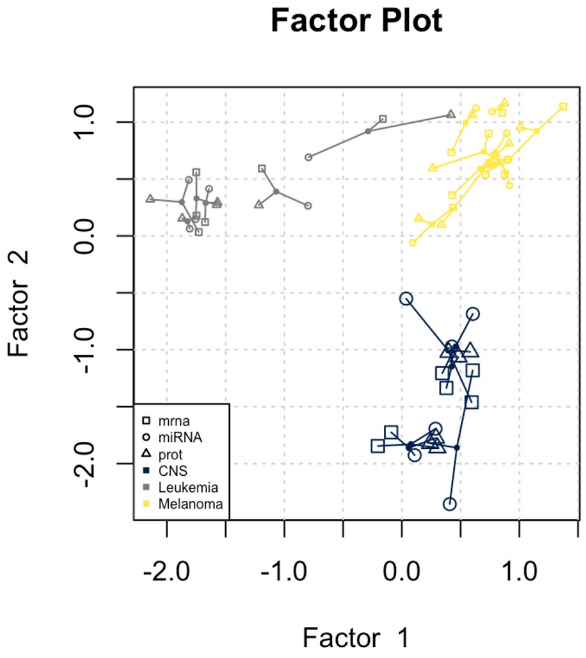
Image credit: Joaquin Reyna (UCSD)

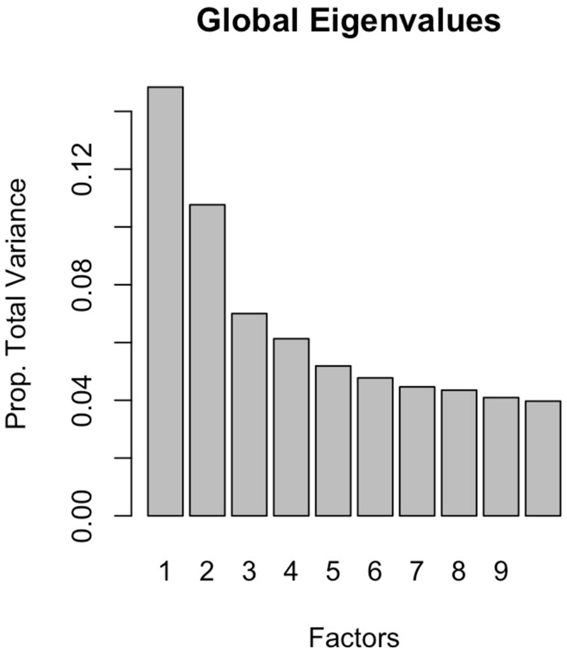
nipalsMCIA: an efficient implementation of MCIA

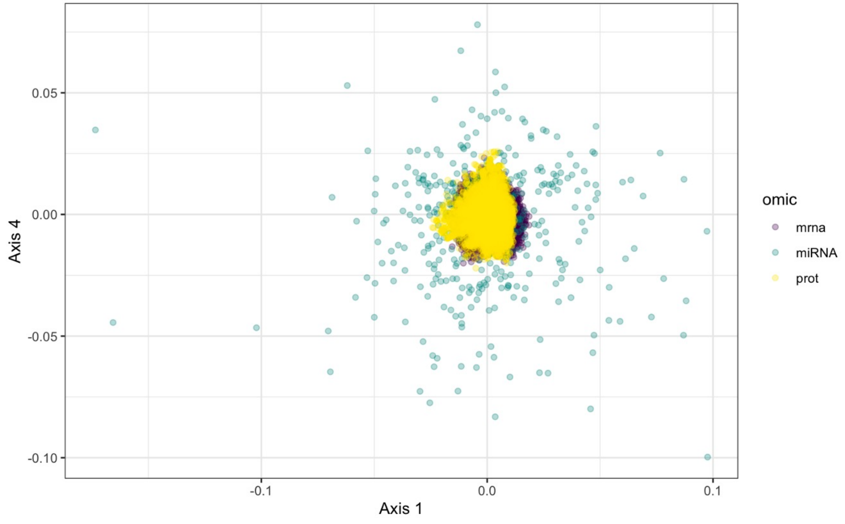


Global and block factor scores correspond to output $\boldsymbol{F}^i$ and $\boldsymbol{F}^i_k$ for the *i*th comp

Plotting the global loadings $\boldsymbol{q}_i$ can show which features/view have the strongest contributions to the associated factors.







- Currently in review for inclusion in open-source bioinformatics library *Bioconductor* with associated article in *Bioinformatics*.
- You can try it here: https://github.com/Muunraker/nipalsMCIA

Center for Computing Sciences, Institute for Defense Analysis (CCS/IDA)

IDA | CENTER FOR COMPUTING SCIENCES

- The Institute for Defense Analyses (IDA) is a private, nonprofit corporation headquartered in Alexandria, VA.
    - IDA’s mission is to answer the most challenging U.S. security and science policy questions with objective analysis leveraging extraordinary scientific, technical, and analytic expertise.[1]
- IDA's Center for Communications and Computing performs fundamental research in support of the National Security Agency's mission in cryptology.
    - Comprises three centers: Center for Communications Research in Princeton, New Jersey (CCR-P) and La Jolla, California (CCR-LJ) and the Center for Computing Sciences in Bowie, MD (CCS). Each center has a unique focus.
- At CCS, research staff focus on intelligence-related problems of importance to national security and on tackling problem sets of interest to the entire computational science world[2]. Active areas of research include:
    - High-performance computing - components, systems, algorithms, languages, and applications
    - Network security and related cybersecurity issues, such as cryptography
    - Emerging algorithmic and mathematical techniques for analyzing extremely complex data sets

[1] https://www.ida.org/about-ida,

[2] https://www.ida.org/en/ida-ffrdcs/center-for-communications-and-computing/center-for-computing-sciences

Tips for grad school/career success (general)

- **Find your mission**
  - It can change, but it will help in the daily grind, and setting your compass.
  - Figure out *what you want to do*, and this will help you in finding what *you want to be* (professor, data scientist, etc.)
  - Insider tip on figuring out what you want to do: what do you read, study, think about after you've done all your work for the day? What do you wish you had more time for? This can help in understanding your own compass.
    - Unfortunately, people will be all too happy to tell you your mission, so better to figure it out first!

Tips for grad school/career success (general)

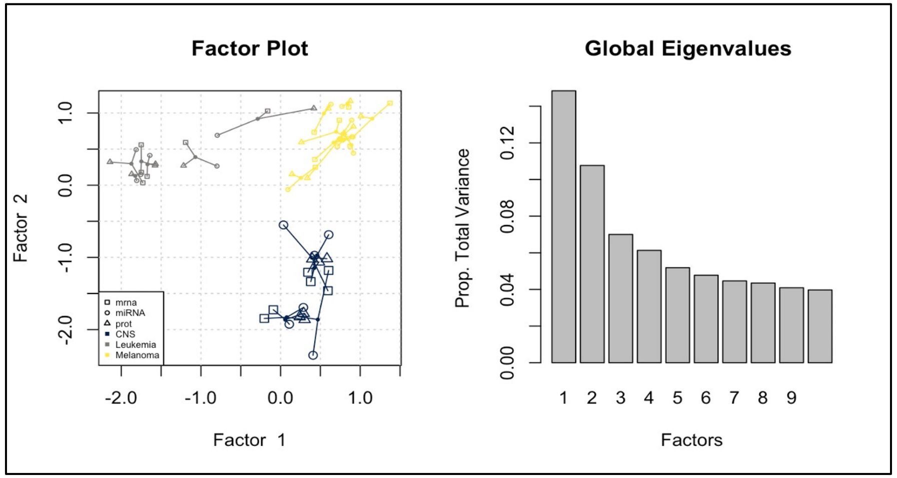
- **Find your mission**
  - It can change, but it will help in the daily grind, and setting your compass.
  - Figure out *what you want to do*, and this will help you in finding what *you want to be* (professor, data scientist, etc.)
  - Insider tip on figuring out what you want to do: what do you read, study, think about after you've done all your work for the day? What do you wish you had more time for? This can help in understanding your own compass.
    - Unfortunately, people will be all too happy to tell you your mission, so better to figure it out first!
- **It's *ok/normal/healthy* to think about finances, location, and work/life balance**
  - These are not signs of weakness or a lack of dedication.
  - We all come in to graduate school with different components 'in the bank', this can be actual material goods, but also reserves of emotional strength and network support.
  - In graduate school, it can be difficult to build up these banks – so it's legitimate to prioritize (re)building these banks after graduate school.

Tips for grad school/career success (resources)

- Scholarly societies!
  - Societies like the American Mathematical Society (AMS), Society for Industrial and Applied Mathematics (SIAM), American Association for the Advancement of Sciences (AAAS), Society for Mathematical Biology (SMB) host regular meetings (regional and annual), and provide career exploration and employment resources.
  - Membership and conference fees are relatively low for students
  - Don't even have to present a talk – just go, you *do* belong there!
- Non-academic government research centers often offer summer programs for grad students and/or post-doctoral programs
  - National Labs (17 National Labs in the country, including Sandia, Lawrence Livermore, Argonne...)
  - National Institute for Standards and Technology (NIST), National Security Agency (NSA), Applied Physical Laboratory (APL)
  - IDA and its centers (CCR-P, CCR-LJ, CCS)
- Government contractors offer additional employment opportunities
  - MITRE, METRON, Daniel J. Wagner, Booz Allen Hamilton
- usajobs.gov is a fantastic resource to search for positions funded by the federal government
- Resources devoted to communicating paths and challenges for mathematicians, including
  - MEET a Mathematician (), Her Math Story (), Mathematically Gifted and Black (https://mathematicallygiftedandblack.com/)

Thank you!

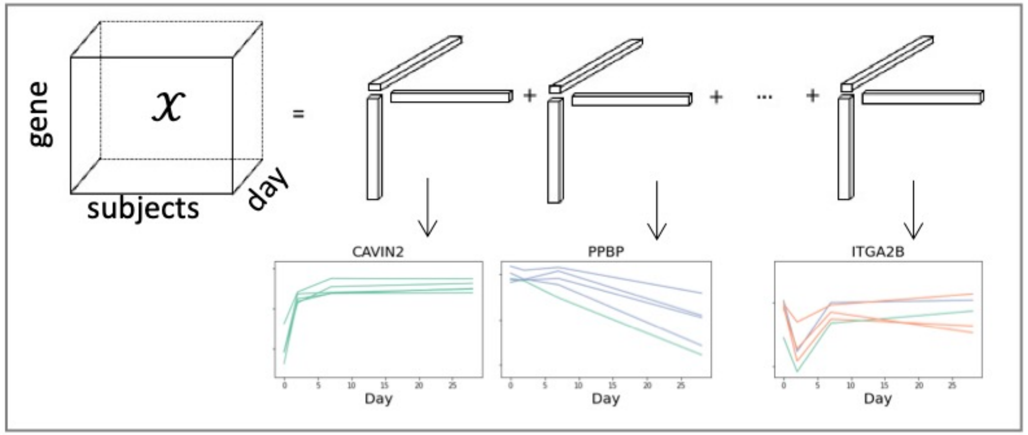
Questions?



Contact:

Email: akonsto@super.org
akonstor@uci.edu

Web: https://akonstorum.owlstown.net/
Linkedin: https://www.linkedin.com/in/akonst/



Non-Academic Career Opportunities (Government)

- Institute for Defense Analyses FFRDCs (Federally Funded Research and Development Centers) (ida.org; https://www.ida.org/ida-ffrdcs/center-for-communications-and-computing)
- National Labs (sponsored by the Department of Energy): https://energy.gov/about-national-labs
- NIST (National Institute for Standards and Technology)
  - Postdoctoral program: https://www.nist.gov/iaao/nist-postdoctoral-research-associateships-program
  - Undergraduate Research Program: https://www.nist.gov/surf
- National Security Agency (Fort Meade, Maryland)
  - They have internships for undergraduate and graduate students.: https://www.intelligencecareers.gov/icstudents.html?Agency=NSA
  - For more information: http://www.nsa.gov
- Applied Physics Laboratory (Laurel, Maryland)
  - For more information: http://www.jhuapl.edu/employment/
- Some Government Contractors
  - MITRE: https://www.mitre.org/
  - METRON: http://www.metsci.com/
  - Daniel J. Wagner Associates: http://www.wagner.com/
  - Booz Allen Hamilton: http://www.boozallen.com/

Acknowledgment: Kelly Yancey (CCS/IDA) for organizing this list