Epilogue

Wrap-up

Goals of the course

- To know the most common/important matrix factorisation methods
  - their advantages and disadvantages
  - their use in data mining
- To understand the theoretical foundation behind the techniques
- To be able to use the techniques to solve real-world data analysis problems

What have we learned

- Factorization methods
  - SVD
  - NMF
    - CX, CUR, and NNCX
    - ICA
- Others
  - Optimization
  - Spectral methods

On factorizations: SVD

- Can be computed in polynomial time
- Minimal reconstruction error
- Explains the direction of the variance
- Otherwise often hard to interpret
- Does not preserve sparsity or nonnegativity

On factorizations: NMF

- Preserves nonnegativity
- Parts-of-whole interpretation
- Can be regularized to be sparse
- NP-hard to compute
- The factors are not orthogonal
  - Though variants exist for that

On factorizations: CX et al.

- ***C*** preserves the original columns
  - In NNCX, also ***X*** is nonnegative
  - Some quality guarantees
  - Projection is easy, selecting the columns not
  - To be limited to the columns is often very restricting

On factorizations: ICA

- Finds the original signals
  - Provided they’re not Gaussian...
- In many ways orthogonal to SVD
- No dimensionality reduction
  - Though can be done in whitening
- No exact algorithms
  - Though FastICA works well in practice

Which method to use? (1)

- Horses for courses!
- Start with SVD/PCA (normalization?!)
  - Gives the most powerful first glance
- If data is mixed-sign, PCA can be followed with ICA for a different take
  - CX or CUR only if preserving the original columns/rows is important

Which method to use? (2)

- If nonnegativity is important, NMF is the way to go
  - Many re-starts, different algorithms
  - NNCX only if selecting some columns is important
- ICA can also be used, but often looses the nonnegativity

Which method to use? (3)

- Different methods find different structure
  - No correct answer, just different answers
- Embrace diversity!
  - And be afraid of over-fitting

Matrix factorizations that don’t factorize

- Matrix factorizations can be used to solve many problems in data mining/analysis, e.g.
  - Clustering (spectral or *k*-means)
  - Frequent itemset mining (not in this course)
  - Clique detection/social network analysis
  - Topic models
- Linear algebra is a powerful language to present your problems

On exam

Format & basic info

- Written exam
- 24 July 2017 from 14:00–16:00
  - Times are sharp!
- Lecture hall 001, building E1.3
- Remember: you must be registered to HISPOS

What you can and cannot bring

- You can (must) bring
  - writing equipments & student ID
  - one (1) A4-sized “cheat sheet” paper
- You cannot bring (use)
  - electronic devices (incl. phones and pocket calculators and electric pencil sharpeners)
  - any other notes than the cheat sheet (incl. lecture slides, assignments, etc.)

Cheat sheet

- **Must contain your name!**
- A4-sized paper, text can be on both sides
- Any content is OK (as long as it is legal)
  - Use your discretion what you think is important or hard for you
- Can be made with computer or be hand-written (or with typewriter)

What is covered in the exam?

- All lectures
  - Lecture on 17 July is also included
- All problem sheets and analysis assignments
- The chapters of books and articles cited in the lecture slides

What kind of questions are there in the exam?

- Simple mathematical proofs
  - Similar to those in problem sheets
- Developing variations of presented algorithms
  - “Explain how would you compute ABC decomposition with the following constraints”
- Short texts or longer essays comparing different decomposition methods and/or explaining their use cases and interpretations
  - “What are the main differences between ABC and XYZ?” “Given this-and-that kind of data, how would you interpret its ABC decomposition?”
- Short questions about features and properties of decompositions and methods
  - “Explain briefly the main idea behind algorithms computing ABC.” “True or false: computing the optimal XYZ decomposition (w.r.t. the Frobenius norm) is NP-hard.”

Exam checking day

- 26 July from 10:15 to 11:45
  - room 024, building E1.4
  - your only chance!

Re-exam

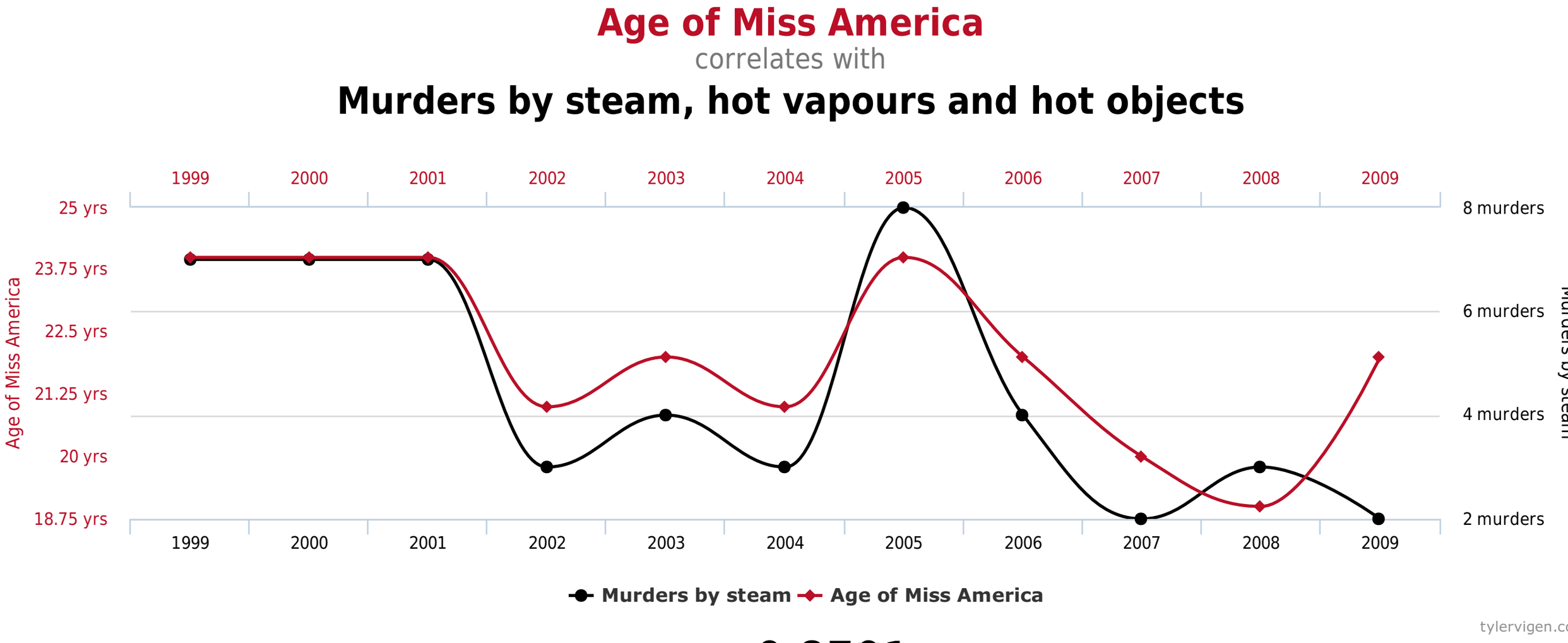
- Hopefully not needed
  - If needed, will happen towards the begin of the winter semester 2017–2018 (i.e. early October)
- You must let us know if you want to attend
- **Bonus points do not apply for re-exam**

Follow-up course on tensors?

- Block course on tensors
  - First weeks of October
    - Tensors extend matrices, lots of nice math!
    - Prob’ly no programming (too short time)
- Alternatively a block seminar
  - Prob’ly two days in January

Ask Me Anything

Spurious correlations



http://www.tylervigen.com/spurious-correlations

Thank You!