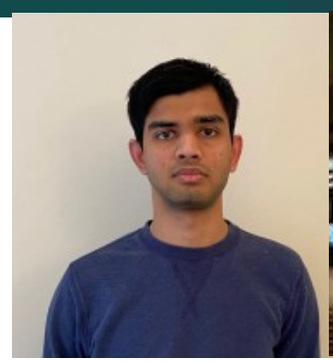
EE 274: Data Compression, Theory and Applications

(Aut 22/23)

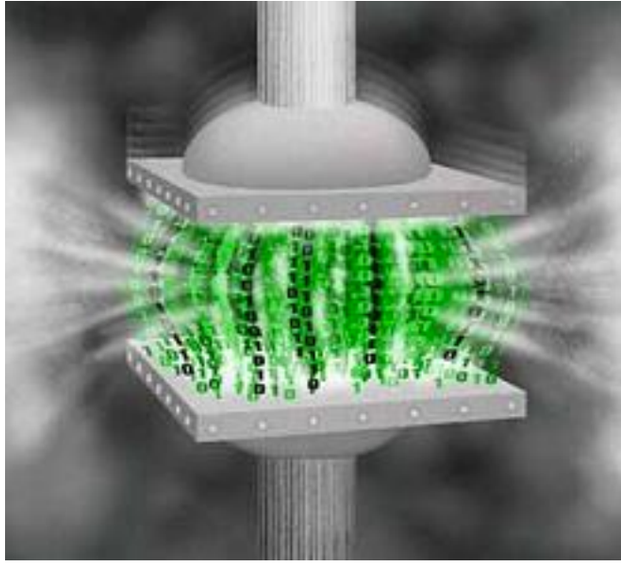
KEEP
CALM
AND
COMPRESS
DATA

Lecturers: Shubham Chandak, Kedar Tatwawadi and Tsachy Weissman



TA: Pulkit Tandon

what is data compression?



- succinct representation of information
- in bits
    - because we’re in the digital information age
    - not coincidentally or arbitrarily (more in 276)

why compression?

- storage is costly
- purest form of information processing:
  - distilling your information to its essence
  - communication without the noise
- compressed representations
  - ease search through the data
  - can simplify implementations
- compression leads to insights, modeling, prediction, etc.

this course

- theory
- general techniques and approaches
- concrete algorithms
- practical implementations

lossless and lossy compression

- lossless: text, code, genomic data, files…
- lossy: video, audio, EEG, sensor data, VR…

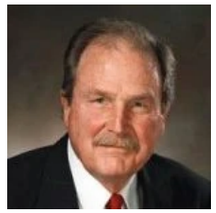
course outline

- lossless
- lossy
- beyond the vanilla settings

lossless part

- entropy and its role

- lossless compressors: Huffman, Arithmetic, Asymmetric Numeral Systems



- universal lossless compression

- Lempel-Ziv 77/78 schemes, GZIP, bzip2, zstd, etc.


Robert M. Gray

lossy part

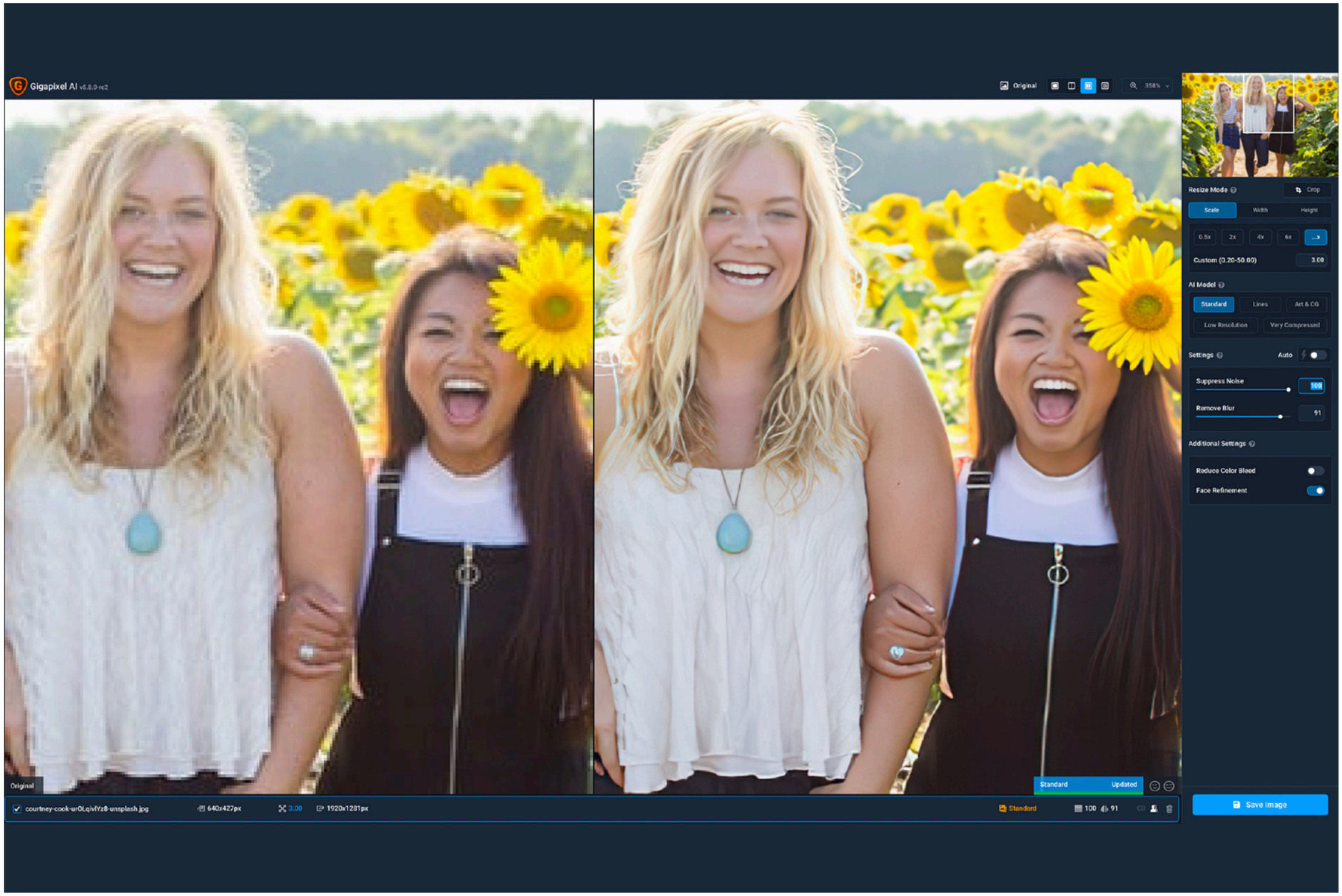

Toby Berger

- rate distortion theory, mutual information and the r-d function
- quantization
- Gaussian r-d function, water pouring, justification for compression in the transform domain
- transform coding: FFT, wavelets, orthonormal transformations, predictive coding, non-linear transforms
- JPEG, MP3, MPEG
- neural network based

beyond vanilla

- successive refinement
- distributed compression
- compression and NNs
- random access
- lossy compression and denoising
- interplay between compression and inference
- compression for perceptual distortion metrics
- genomic data compression
- physical data compression: storage in DNA

successive refinement

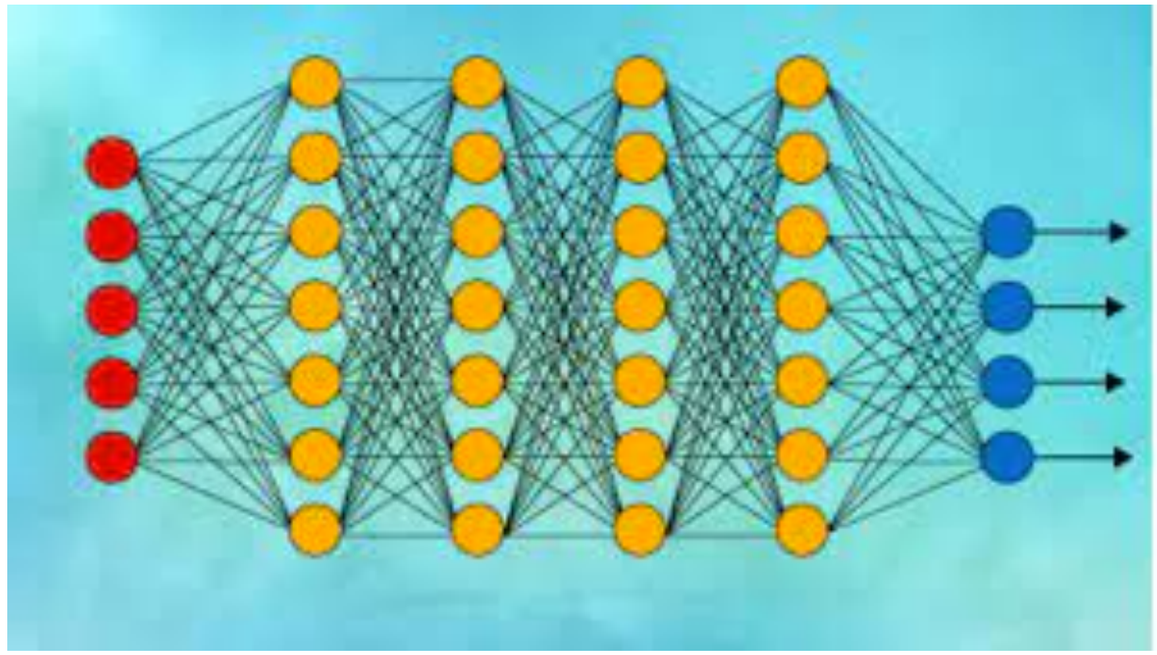


distributed compression

- lossless (Slepian-Wolf)
- lossy
- Gaussian
- (full of fun surprises)

compression and neural nets

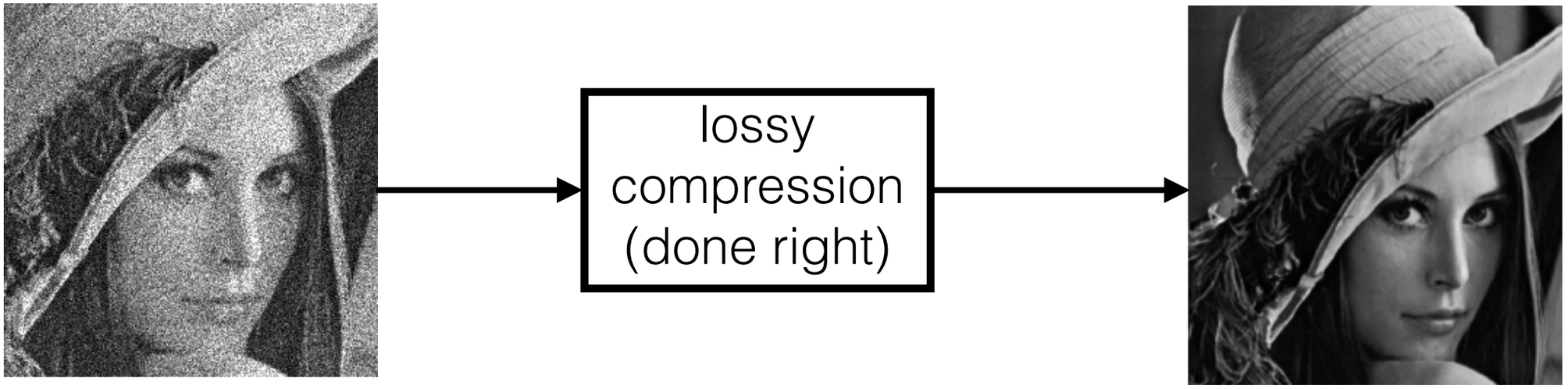
- compression of NNs
- compression for NNs
- compression with NNs
- and combinations thereof…



compression with random access

- retrieving parts of the data without decompressing it all
- queries and computations in the compressed domain
- tabular compression with fast retrieval of rows/columns

denoising and lossy compression



compression and inference

- compressor <-> probabilistic model
- compressibility as an inferential tool

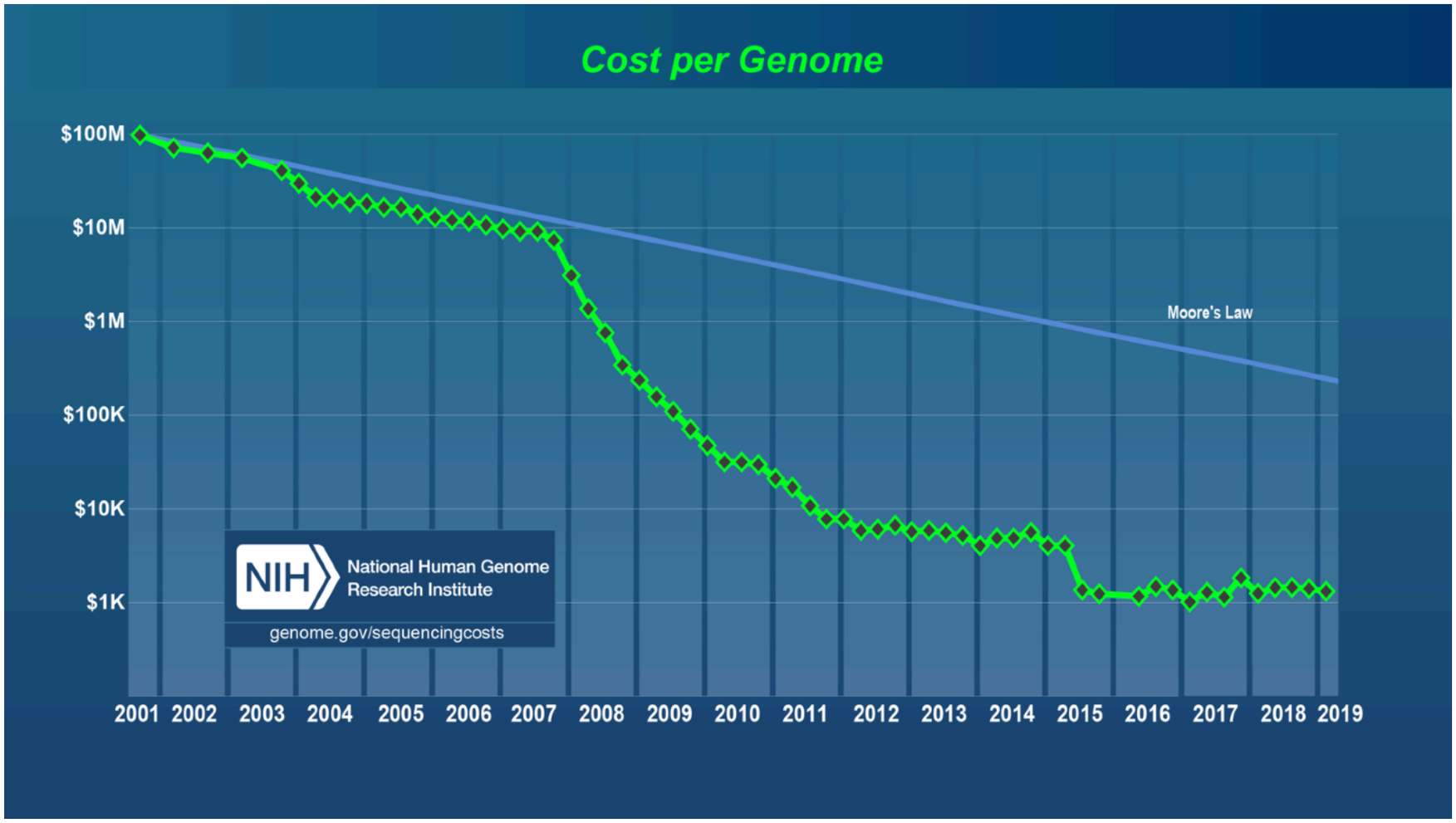
compression under perceptual distortion metrics

- Structural Similarity Index (SSIM)

- learned metrics

genomic data compression

- whole sequenced genomes
- raw reads



physical compression

- storage of data in DNA

we'll only scratch the surface

- but will give you tools and hope to whet your appetite to delve deeper
    - independently
    - in more advanced courses such as the 376 series
- and appreciate (retrospectively or concurrently or in future) courses such as 276, 278, 261

have fun!

(and let us know if you're not)