Information Processing. Real and Perceived Quality of Big Data. Challenges in Identifying the Useful Data. Panel. ComputationWorld. Namics.

A Merkle Company

VENICE, ITALY, MAY 6TH 2019

Hans-Werner Sehring. Panel Moderator.

Ole Kristian Ekseth. Panelist.

Jan Fesl. Panelist.

Birgitta Dresp-Langley. Panelist.

Marc Kurz. Panelist.

Seiichi Gohshi. Panelist.


Information Processing.

Theme: Real and Perceived Quality of Big Data


Topic: Challenges in Identifying the Useful Data
Best Intranet Site
Zuger Kantonsspital
Site built by Namics AG
SITE OF THE YEAR

Just some questions and quick thoughts.

Not starting the discussion…

Data Quality. Data Provenance.

Data quality might be judged by examining those who work on/with the data. Quite common in science (e.g., author of a paper, reviewers, …).

- **Who created the data? When? In which context?**
- **How often was the data used it until now? By whom? What for?**
- **Who modified the data? When? What was changed?**

What are we missing in CS?

Classical computation does not consider certain ways of reasoning.

- **Pragmatics.**
  Syntax and semantics (more or less) understood. How about this one?
- **Epistemology.**
  Claim: We only perceive, what we have a concept for.
  We only add concepts to our model if we find evidence.
- **Semiotics.**
  Is it the real thing, or just a representation of it?
- **Emotion.**
  Does certain data lead to a change of behavior?

N.b.: Database are a form of communication over time!

… and is there a difference?

Why?

Discussed on a previous panel: Most important question to ask about a result is “Why?”

- **Most important command of expert systems.**
- **Major shortcoming of, e.g., neuronal networks.**
- **Idea:**
  - **There are expectations towards the result.**
  - **From the explanation for a result, detect…**
    - the influence of parameters
    - significant system boundaries
    - etc.

hans-werner.sehring@namics.com. Panel Moderator.

Thank you. Namics.

A Merkle Company

© NAMICS AG 2019

UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Real and Perceived Quality of Big Data

Challenges in Identifying the Useful Data

Marc Kurz

Panel on Information Processing, Adaptive 2019, May 6th, 2019

Contact

Name

Dipl.-Ing. Dr. Marc Kurz
Professor for Mobile Software Systems

Contact

University of Applied Sciences Upper Austria
Department of Mobility & Energy

Softwarepark 11
4232 Hagenberg/Austria
Tel.: +43 (0)50804-22827

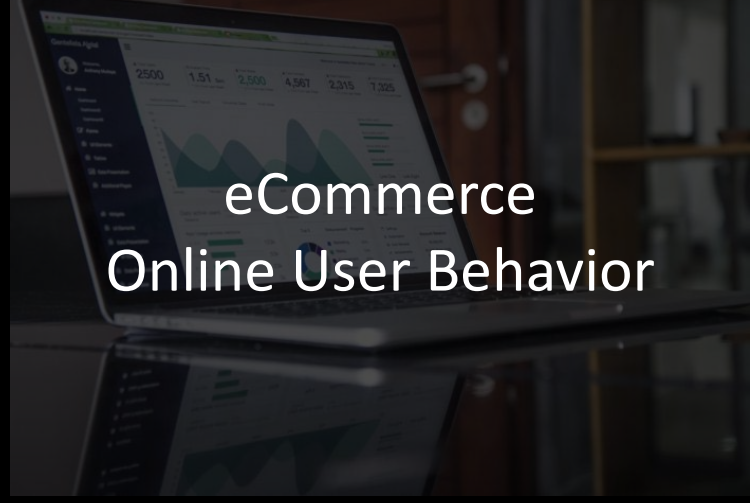
Mail: marc.kurz@fh-hagenberg.at
Web: https://www.fh-ooe.at/mc
FB.: https://www.facebook.com/MC.AC.ENI.fhooe/

UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA


SOS

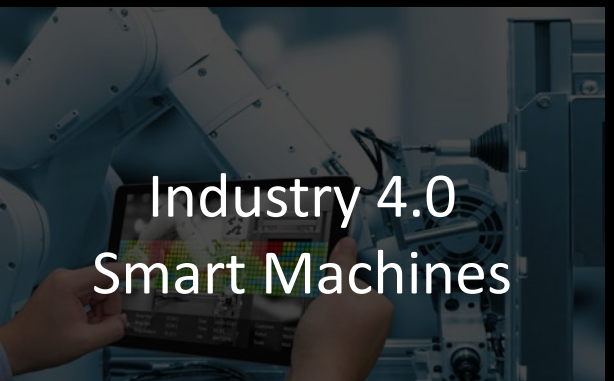

Smart Cities


SmartPhones
Mobile Systems


eCommerce
Online User Behavior

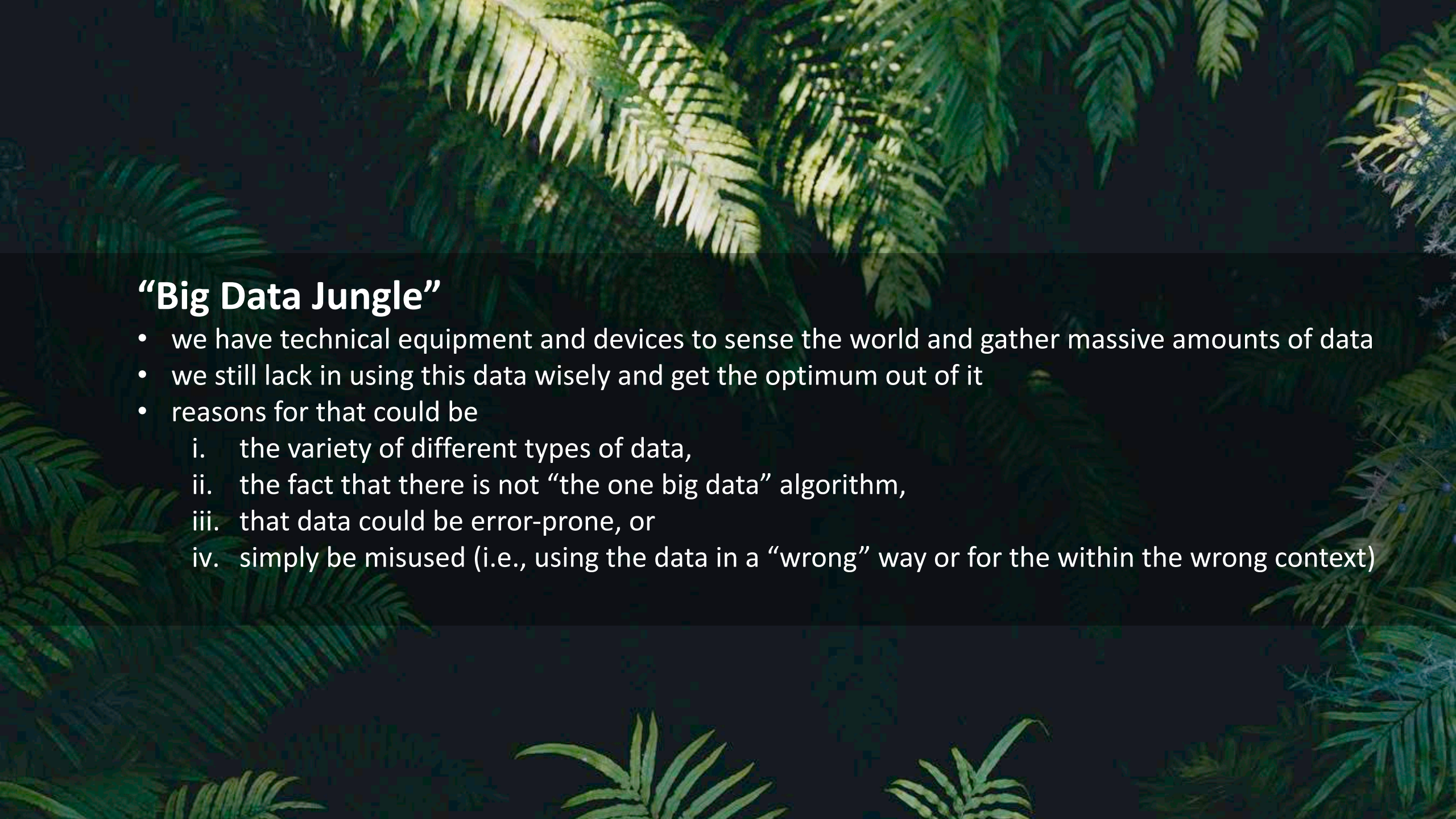

Smart Cars
Driving Behavior


Bio-Medicine


Industry 4.0
Smart Machines


Smart Homes


Connected Vehicles



“Big Data Jungle”

- we have technical equipment and devices to sense the world and gather massive amounts of data
- we still lack in using this data wisely and get the optimum out of it
- reasons for that could be
  i. the variety of different types of data,
  ii. the fact that there is not “the one big data” algorithm,
  iii. that data could be error-prone, or
  iv. simply be misused (i.e., using the data in a “wrong” way or for the within the wrong context)

Thank you! Any Questions?

Image Sources:

- http://www.nfp75.ch/de/News/Seiten/150916_news_nfp75_ausschreibung.aspx
- https://www.intelligenthq.com/industry-4-0-regulation-artificial-intelligence/
- https://www.csoonline.com/article/3228132/user-behavior-analytics-separating-hype-from-reality.html
- https://www.smartcitiesworld.net/news/news/council-launches-annual-readiness-challenge-3474
- https://www.analyticsinsight.net/big-data-3-3-billion-opportunity-automotive-industry-says-sns-telecom/
- https://chenjiazizhong.com/2014/03/30/biomedicine-healthcare-venture-capitals-investment-of-2014/
- https://towardsdatascience.com/smart-homes-safety-stability-and-trust-4ff1e270b2ee
- https://autorevue.at/ratgeber/connected-car-automobil-grenzenlos

https://www.facebook.com/MC.AC.ENI.fhooe/

What is going on with big data?

- Science and engineering changed in the past 10 years
- Big data, deep learning, IoT, and smart grid changed our world
- We have been chasing causality since Galileo Galilei
- Cause=> Analysis => Result
- Analyses is based on theories
- New technologies do not need the theories
- Big data changes data analysis technology from unintentional samples to 100% surveys
- It is the change from causality to correlation
- Generally after these discussions, next word is “How to do it?”
  - ✓ The appropriate computer languages are R, Python etc
  - ✓ Tools are SPSS, SAS etc

What is going on with big data? (2)

- Recently there are many young researches who take active part in conference
- More than ten years ago researchers had to learn many things
- It took time and young researchers were not able to submit papers
- Young researchers can stay in labs and cutting the sleeping time, as long as the physical strength continues, keep entering data into the computer
- Some of them can produce good results
- Behind their success there are many young researchers that were not able to reach the good results
- Generally we tend to only focus on good achievements

Pros and Cons

- In the future when you feel sick, the doctor might say “Let’s ask the black-box about your symptoms.” What do you think about it?
- Although it works well, there are no theoretical or reason based explanations
- Media always report cases that worked
- Deep learning is a good example
- Google Flu Trends did not work well
  - ✓ quality of the data is more important than quantity of data
- Privacy issue
- “And Yet It Moves” (Galileo Galilei)
- Our situation is different from that of Galilei
- Nobody is able to provide a logical explanation why deep learning works well and the big data indicate this result

Big data and deep learning

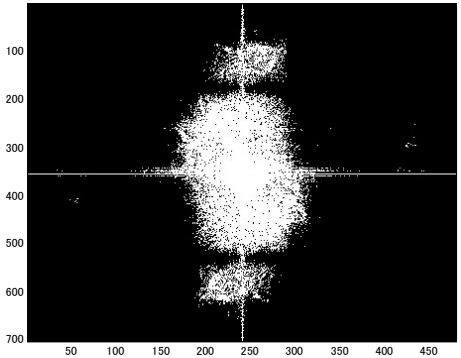
- There is a strong relationship between the big data and deep learning
- If we discuss this, we cannot avoid the important problem: “technical singularity.”


Osamu Tezuka's Original
ASTRO BOY

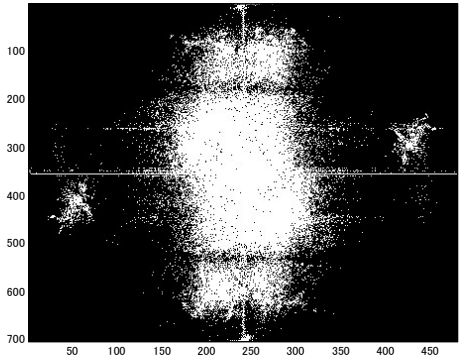

Google
facebook
amazon.com

https://www.techsparks.co.in/cloud-computing-fundamentals-its-basics-and-terminology/

Super resolution



Original image



SR processed image

On the perception of meaning in large data sets or « how much big data do we actually need »?

Birgitta Dresp-Langley

ICube UMR 7357 CNRS-Université de Strasbourg, France

❑ ***Storage of big data in the cloud - less is more - not all data are relevant !***

❑ *Detecting systematicities in large data sets is a challenge for Artificial Intelligence*

❑ *Deep learning in large-scale neural network architectures is the current trend, but not always the best solution in terms of costs vs benefits*

❑ *Interpretation of systematicity detected in big data: human expertise is needed more than ever to decide whether 1) the detected systematicity is meaningful 2) and, ultimately, relevant enough to keep the dataset*

Study example to illustrate this point:

Ultrafast automatic classification of large image sets showing CD4+ cells with varying extent of HIV virion infection

John M. Wandeto*+ and Birgitta Dresp-Langley+

**Department of Information Technology, Dedan Kimathi University of Technology, Nyeri, Kenya*
+ICube UMR 7357 CNRS-Université de Strasbourg, France

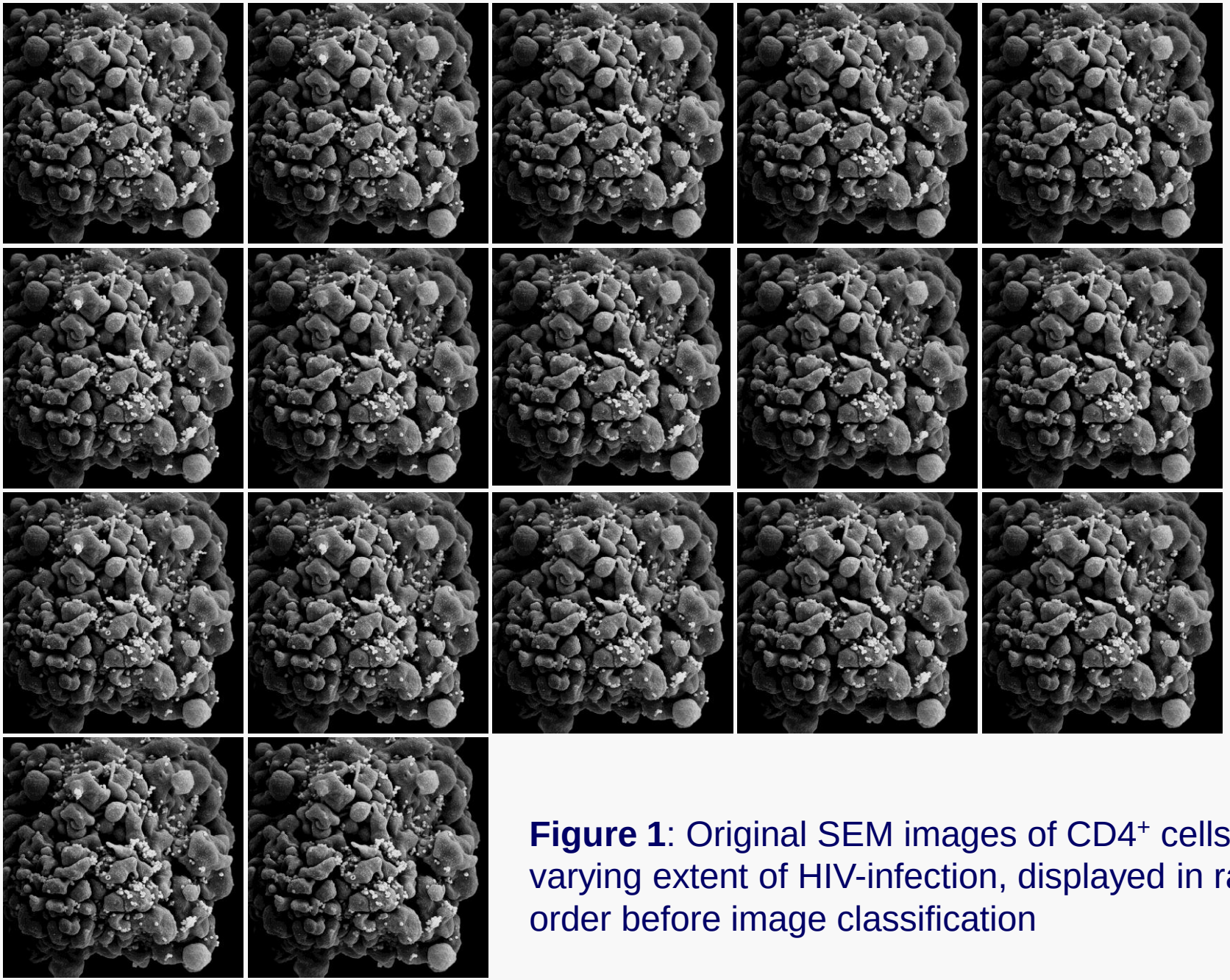


Figure 1: Original SEM images of $CD4^+$ cells with varying extent of HIV-infection, displayed in random order before image classification

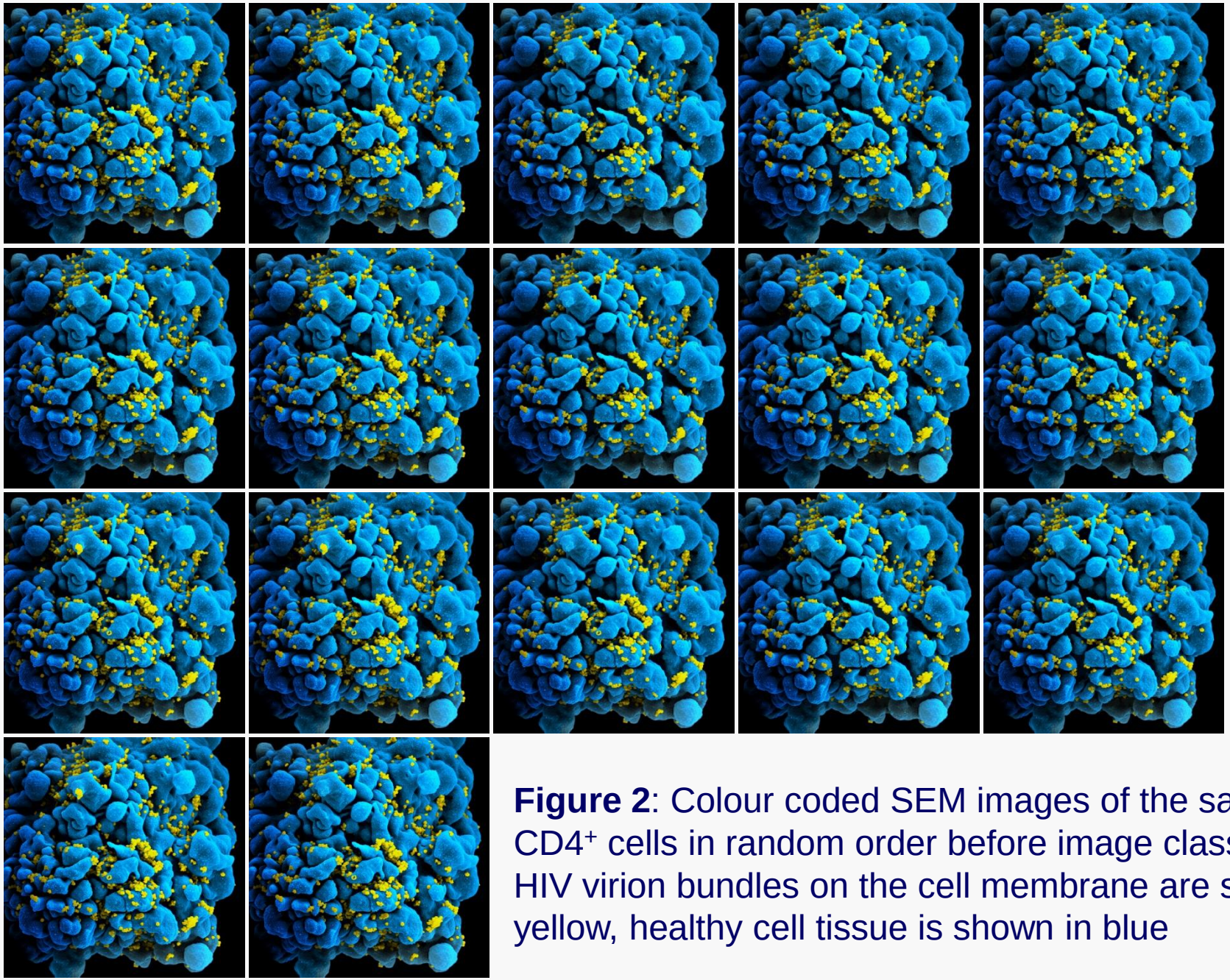


Figure 2: Colour coded SEM images of the same $CD4^+$ cells in random order before image classification. HIV virion bundles on the cell membrane are shown in yellow, healthy cell tissue is shown in blue

Digital image archiving *for public libraries in economically disadvantaged parts of the world will benefit from openly accessible and affordable computer algorithms that allow* ***sorting large sets of image data automatically*** *without human intervention in the process.*

In the case of ***cell images****, it may be required to order large sets of image data according to the spatial extent of specific local colour/contrast contents,* ***revealing different states of a cell in a certain order****, indicating progression or recession of a pathology, or the progressive response of the cell structure to a treatment.*

This can be achieved by exploiting the ***Quantization Error (QE) in the output of a Self-Organized neural network Map (SOM)*** *with minimal functional architecture.*

Easily implemented, the ***SOM learns the pixel structure of any target image in about two seconds (unsupervised "winner-take-all" feature learning)*** *and* ***detects local changes in contrast or colour in subsequent images*** *with a to-the-single-pixel precision in less than three seconds for a set of 20 images.*

We applied this method to the ***automatic classification of Scanning Electron Microscopy (SEM) images of CD4$^+$ T-lymphocytes*** *(so-called helper cells) with* ***varying extent of HIV virion infection.***

17 SEM images from two series were fed into SOM.

Unsupervised feature learning of the target image, and subsequent QE ***classification of the 16 other images in the correct order of magnitude took less than five seconds*** *on each of the two series, one unprocessed (original grayscale images), the other colour-enhanced.*

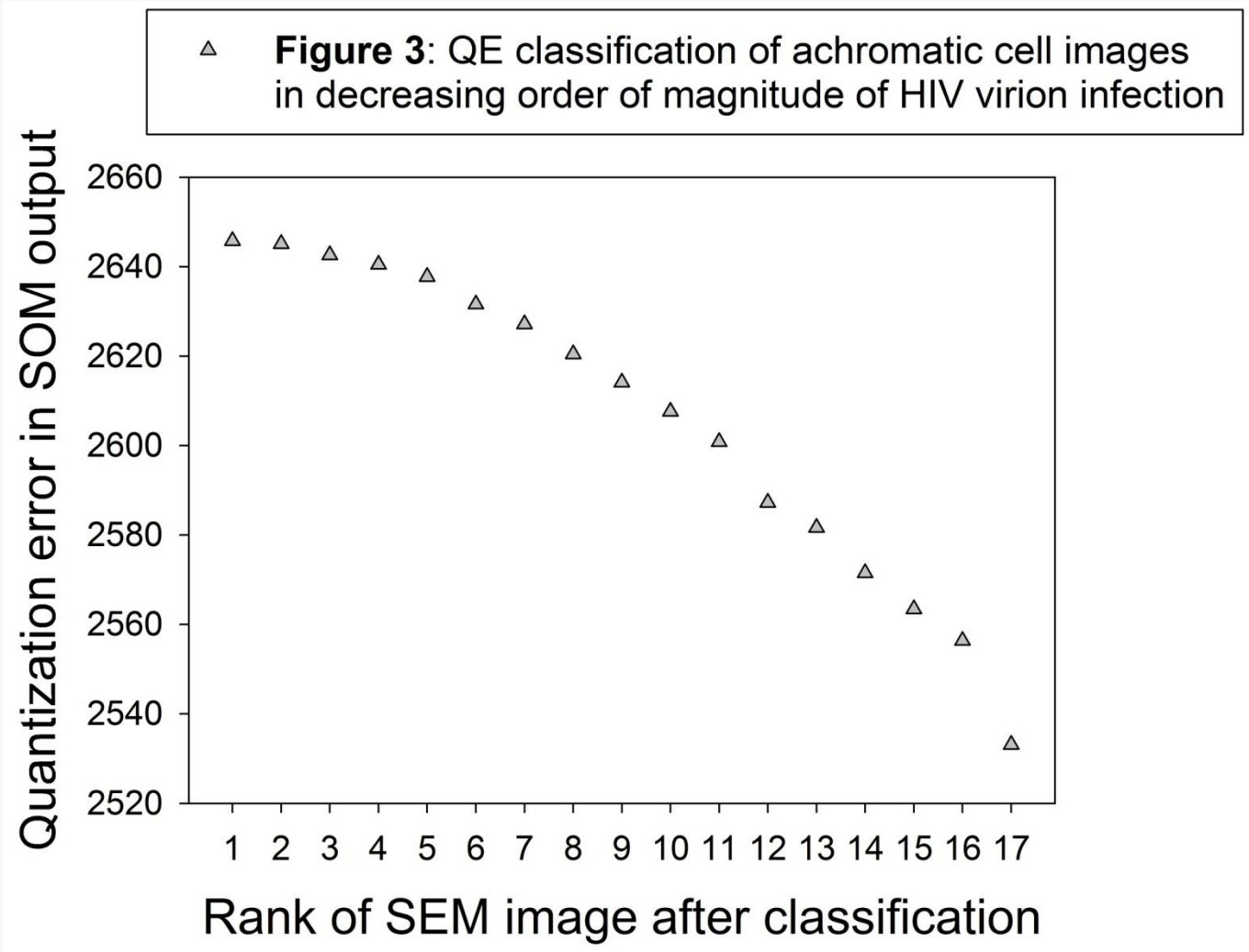


Figure 3: QE classification of achromatic cell images in decreasing order of magnitude of HIV virion infection
Quantization error in SOM output
2660
2640
2620
2600
2580
2560
2540
2520
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Rank of SEM image after classification

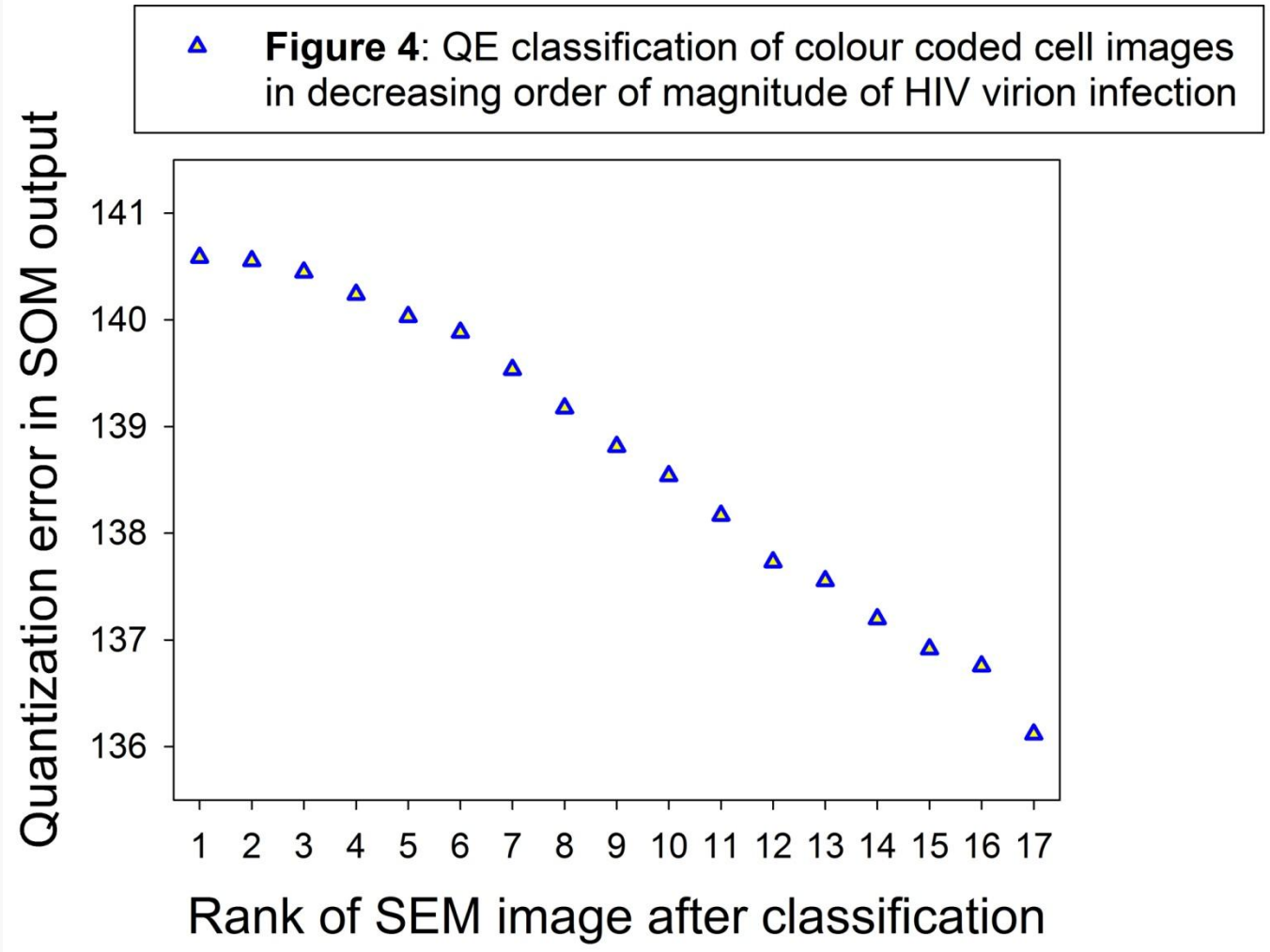


Figure 4: QE classification of colour coded cell images in decreasing order of magnitude of HIV virion infection
Quantization error in SOM output
141
140
139
138
137
136
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Rank of SEM image after classification

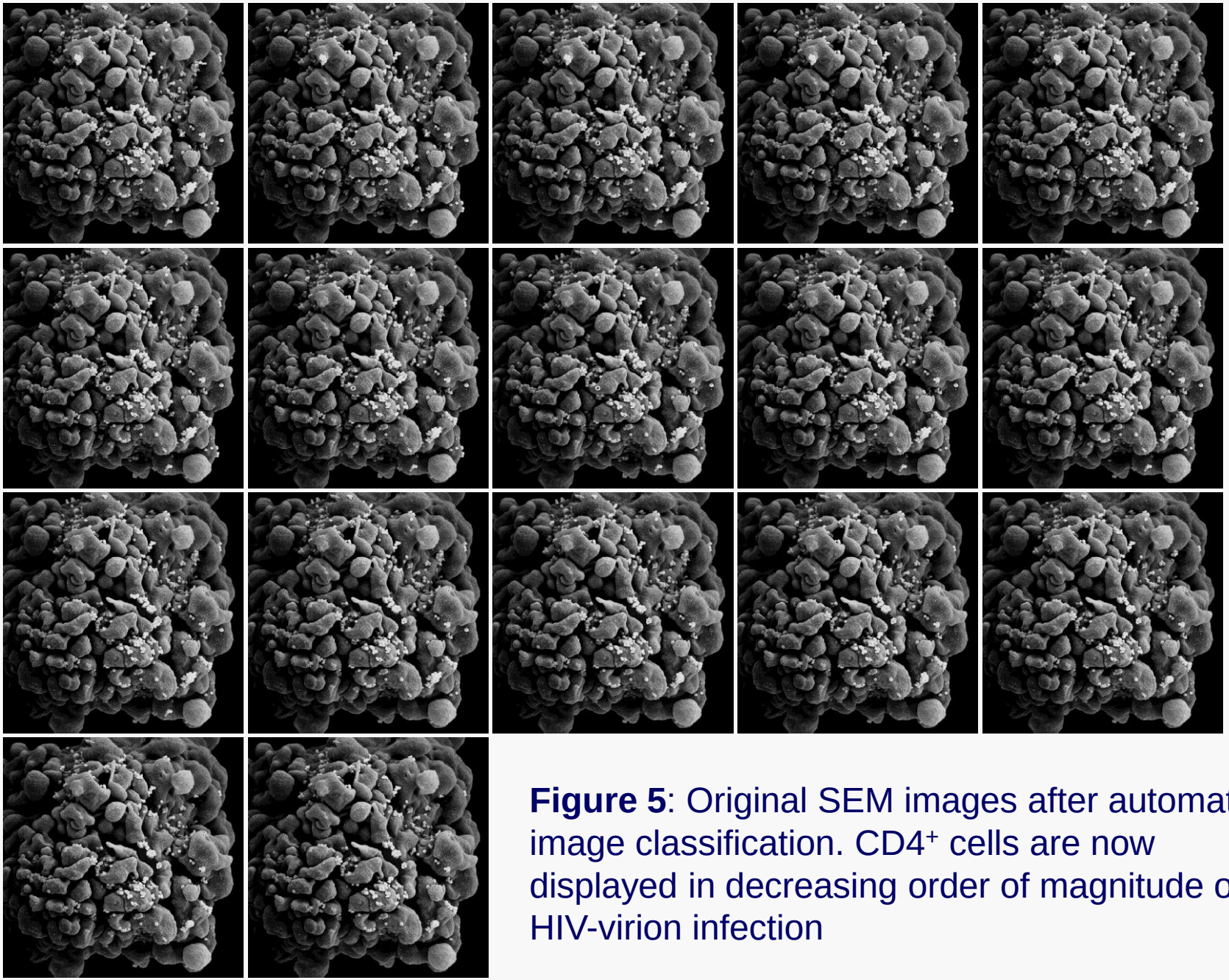


Figure 5: Original SEM images after automatic image classification. $CD4^+$ cells are now displayed in decreasing order of magnitude of HIV-virion infection

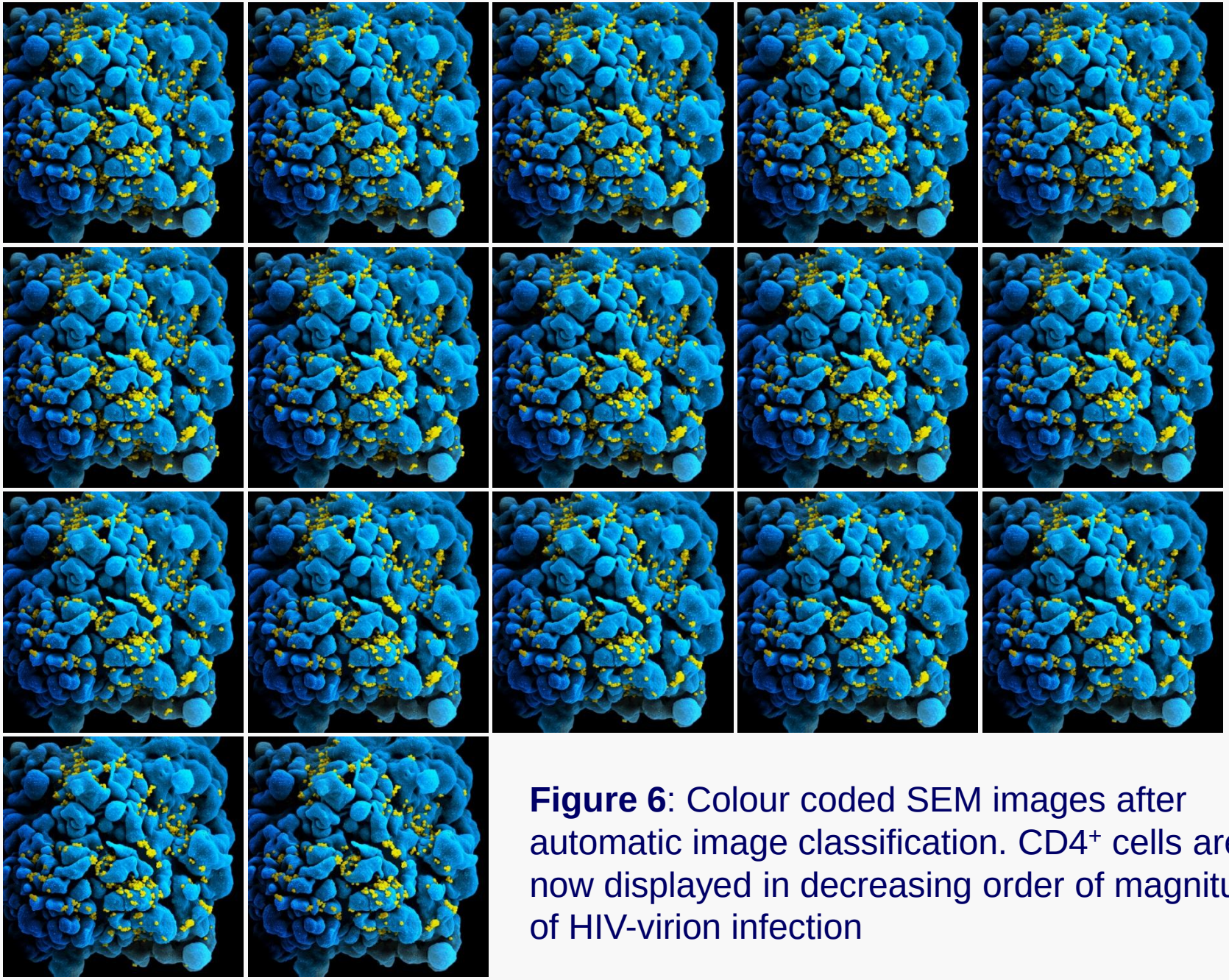


Figure 6: Colour coded SEM images after automatic image classification. $CD4^+$ cells are now displayed in decreasing order of magnitude of HIV-virion infection

The QE in the SOM output permits to scale the spatial magnitude and the direction of change (+ or -) in localized pixel contrast/colour with a reliability that exceeds that of any human.

*However, only the **human expert** is able to decide whether the detected systematicity has clinical **relevance** (meaning), and whether the dataset is useful and should be kept.*

In Big Data Analytics:

No algorithms fits the task: why?

Ole Kristian Ekseth (NTNU and Eltorque)

Norwegian University of Science and Technology

--- ***AI*** algorithms (eg, NN) does **Not work** on Big Data

What *is* Big Data Analytics?

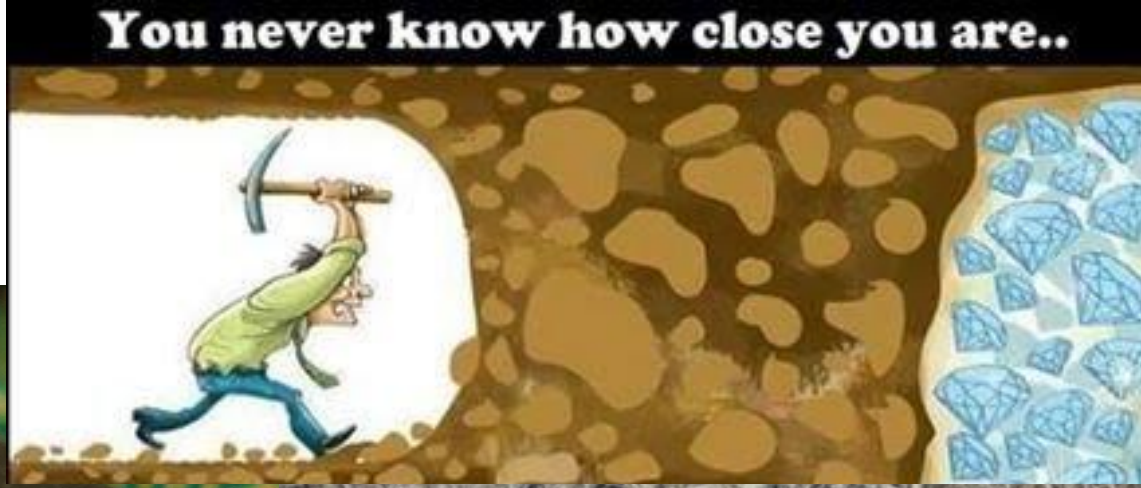
Examples of Big Data:

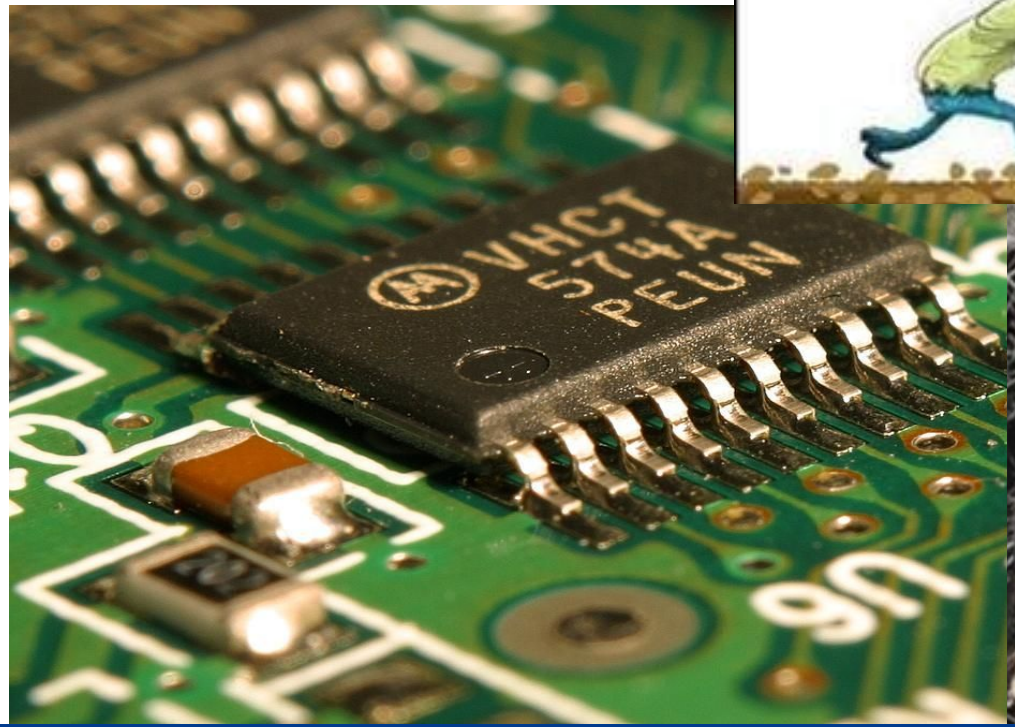
1. Biomedicine:
   a. data: manual, auto-collected;
   b. data quality: error-prone;
   c. usage: new drugs;
2. IOT sensors:
   a. data-volumes requires filtering;
3. Power Grids:
   a. data: auto-collected;
   b. adjusted at random points;
   c. usage: need large time-series;

Software Issues:

1. Execution Time?
   a. poor-written software;
2. Why slow performance?
   a. generic databases slow;
   b. heuristics rarely hold;
3. Why quality issues?
   a. impossible to visually validate;
   b. heuristics needs always to hold;

Fast and Accurate Big Data Mining





NTNU

Do as everyone --- follow the Sheep Pack

1. Algorithms: follow the buz-words:
   a. Neural Networks, Deep Learning, PCA, Semantics, Minimum Spanning Trees, etc.;
2. Machine learning: use a software by Google, Apple, … ← 100x accuracy loss
   a. mistake: algorithm is wrong for your problem
   b. example: Norway's largest energy producer (Statkraft) has tried this; fails utterly;
   c. fix: understand the problem …
3. Software: ← 100x cost increase
   a. always use external libraries;
   b. write performance critical parts in Python, Java, poor C/C++, ...

Big-Data: Common Misunderstandings

1. More data increases quality:
   a. → decreased trustworthiness, increased misunderstandings;
2. store all data in same format:
   a. → developers are in a hurry → semantics becomes broken;
3. reduce the data-size:
   a. → results becomes inaccurate;
4. when an algorithm is 2x better than all others, then always use it:
   a. → ‘2x’ is meaningless (as difference depends on input data);
5. free lunch when established algorithms are used:
   a. → assumptions invalid for most use-cases;

How addressing the problems? --- rock-climbing

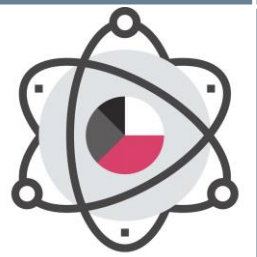
1. joyfulness:
    a. socialize (with experts); ← identify valid shortcuts;
    b. understand + grasp ← avoid rediscovering wheels
2. courage: ask questions:
        i. what are the users interested in?
        ii. possible to use an old-fashioned algorithm (eg, Newton's method)?
        iii. when is the result good enough --- convergence
3. Laziness and craftsmanship:
    a. Occam's razor;
    b. find + fix bottlenecks

Future: Where are we heading?

1. issues may be fixed by experts;
2. researchers lacks resources alleviating issues for particular data collections;
3. → need for:
    a. new methodology, and software, which may automatically analyze data in:
    b. feasible time, and with
    c. accurate predictions;

How to evaluate the Big Data Quality?

Jan Fesl



Is it worth to solve such question?

- **Motivation:** To process interesting data only, but…
- Big Data is really **big and unique** (petabytes – exabytes)

  ↓

  Do we have multiple different data sources ?

- The data analysis consume time and computing resources are expensive

  ↓

  To inspect the data quality can correlate with its complete processing…

  (the quality can be seen after but not before)

How to read the data?

- **Do we know something about the data ?**

  Data Storage Formats and Structures (textual vs binary)

  Data Coding (encrypted ?)

  Data Values (ordering, sorting → distributions)

- **Where is the data stored ?**

  NOSQL, Graph-Oriented type databases

  Distributed File Systems

  Clouds → Content Delivery Networks

  The way of the data storing implicates the possibilities of its processing or evaluation

How to preprocess the data ?

- Is it possible to reduce the data volume ?

  e.g. to select the representative smaller sub-sample set?

  The data is mostly unstructured → sorting (expensive)

- Is it possible to identify outliers → misinterpretations ?

  common known values (before processing) or sorting (expensive)

- Data normalization (is it possible for a data type?)

How to evaluate the data quality ?

- For some types of the data, it is probably impossible (speech samples, pictures)
- Are the data values according to some distribution?

  If not, does it mean the bad quality ?
- For known data types with predicatable values, it is easy to use the traditional statistical or data mining methods
- Which part of the low quality data in the set is still acceptable ?