Modelling & Big Data

Insight, not Data

Gerhard R. Joubert

a. Trigger

- Alex Woodie**: Big Data** (numerical, textual, multimedia) collected and stored on a world-wide scale are of strategic importance [1]
- Chris, Anderson: The **Data Deluge** makes the Scientific Method Obsolete [2]
- Mark, Graham: Big Data and the **end of Theory**? [3]
- Bill Blake (HPC2014): Data scientists claim: The larger the data the **simpler the algorithm**.
- Bill Gropp, 29.04.2014: Who needs Big Data? I'd be happy with **little data**.
- Horst Simon (ISC'14 panel summary): "Big Data" was named by some the most **annoying buzzword** of 2012.

Trigger (cont.)

- Big Data approaches often seem to focus on:

Data, not Insight

- Problem solutions need:

Insight, not Data.

b. Buzz Word: Big Data

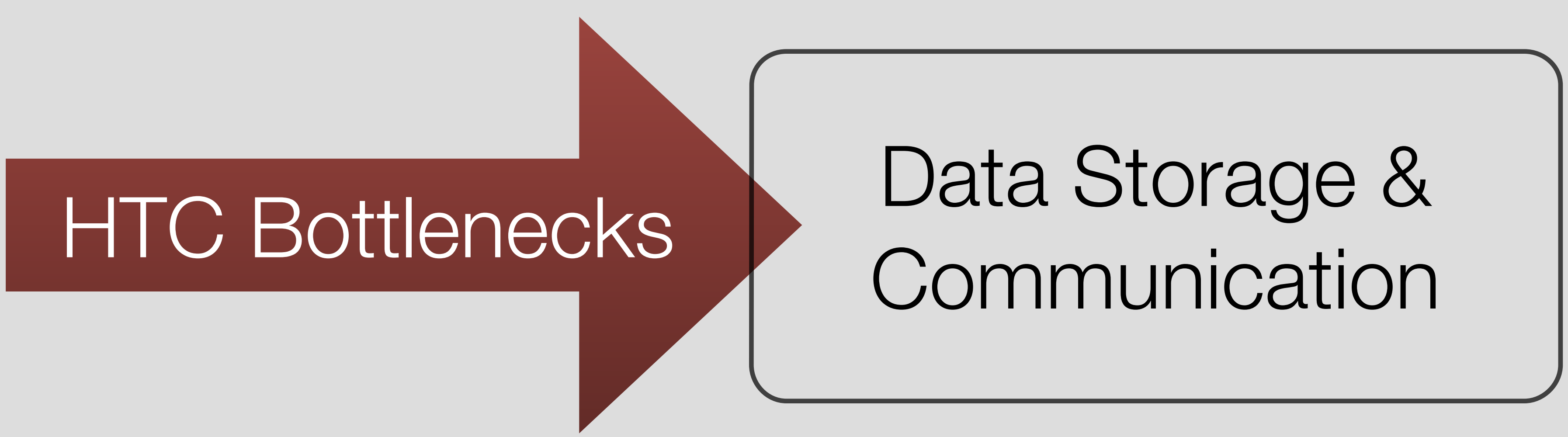
- **Positive:** Focus on the envisioned advantages offered by particular technologies
- **Negative**:
  - **Not or badly defined**: As buzz words mature and gain wide spread acceptance this leads to increased communication difficulties.
  - Unrealistic **overrated expectations** can lead to back lash. Example: Expert Systems in 1980's
- **Need:** <u>Clarification of the meaning of Big Data, and its interface with R&D processes</u>

c. Motivation

- Big Data is an **established** vague concept
- For clear communication between researchers and users It is essential that we clarify what we mean.
- What is the **difference** between **Data** and **Big Data**?
- Where are the **boundaries** and the **interface(s)**?
- What is the importance of **size**?
- etc.

d. Impact on Parallel Computing

- Solving most real world problems requires computationally complex and large scale problems to be solved.
- Compute platforms used: Parallel systems
  - Present focus: **HPC**
  - Future: Big Data needs **HTC, including Data Intensive Computing** (See comments by Jack Dongarra, Ian Foster, Geoffrey Fox, Bill Blake, Paul Coteus, Marcel Kunze, etc.)



Contents

1. Advent of Big Data
2. Scientific Method: Hypotheses & Models
3. Purposed Data and Data Pools
4. Summary

1. Advent of Big Data

Large Data Collections

- Today very large data sets (Petabytes+) are collected in many areas.
- Examples
  - Internet (Google, Amazon, Social Networks, etc.)
  - Medical data
  - Geodata, incl. satellite images
  - Physics (LHC)
  - Astronomy (SETI, SKA)

When is Data Big?

No acceptable definition of Big Data:

1. Volume, Velocity & Variety (3V's) [5,6] (See also Sudip Dosanjh, Marcel Kunze)

2. All data available in an organisation [5]

3. Large data sets compared to memory size of topical computer systems:

   - Yesterday: Terabytes (TB) < 2010
   - Today: Petabytes (PB) 2010 - 2019
   - By ca. 2020: Exabytes 2020 - 2029
   - Then: Zetabytes, etc. 2030 +

Big Data Analytics

- **Analytics** = Process to detect patterns (relationships) in data sets (See Geoffrey Fox)
- Patterns can give insight, e.g.
  - Searching/buying behaviour (Google, Amazon, ...)
  - Medical data (causes of ailments, treatments, ...)
- Discovered patterns: Interpreted as (possible) solutions to real-world problems
- Requirements: HTC / Data Intensive systems

Patterns = Problem Solutions?

- **Notion:** If enough data is available patterns can be detected to solve (all) problems **=> End of Theory**.
- Even if this is true in some cases, this **ignores fundamental aspects** of a sound scientific approach:
  - How were data sources chosen and data selected?
  - Are these representative?
  - Can the results be reproduced/Instantiated with renewed/additional data?
  - What insight is gained into the true nature of the problem considered?

Patterns = Support Solution of Problems

- **Fact:** Detected patterns can greatly support the solution of real-world problems
- For this sufficiently large data sets must be available
- Great care must be taken that erroneous solutions are not considered as correct.

Example: Data Centric Weather Prediction

- **Step 1:** Collect data on today's weather
- **Step 2:** Detect pattern: Next day's weather often = today's
- **Step 3:** Prediction => Tomorrow's weather the same as today
- **Result:** On average ca, 65% correct
- **Problem**: How to obtain longer term forecasts?

Example: Lucio's Improved Weather Prediction

- **Step 1:** Can one see Stromboli from the hotel terrace?
- **Step 2:** If yes: The weather will be worse tomorrow
- **Step 3:** If no: Tomorrow's weather will be the same as today

Conclusion

Solutions based on data/observations without insight into the real nature of the problem can lead to erroneous or less than optimal solutions.

ON THE OTHER HAND

Such data centric solutions may (collectively) contribute to a better understanding of the nature of the problem considered.

2. Scientific Method: Hypotheses & Models

Scientific Method

The scientific method or procedure consists in:

- **Data** collection through systematic observation, experiment and measurement, and the

- Formulation, testing and modification of **hypotheses**.

This method applies in practice to all real world problems.

The order in which the two components - hypothesis formulation and data collection - are applied can be interchanged.

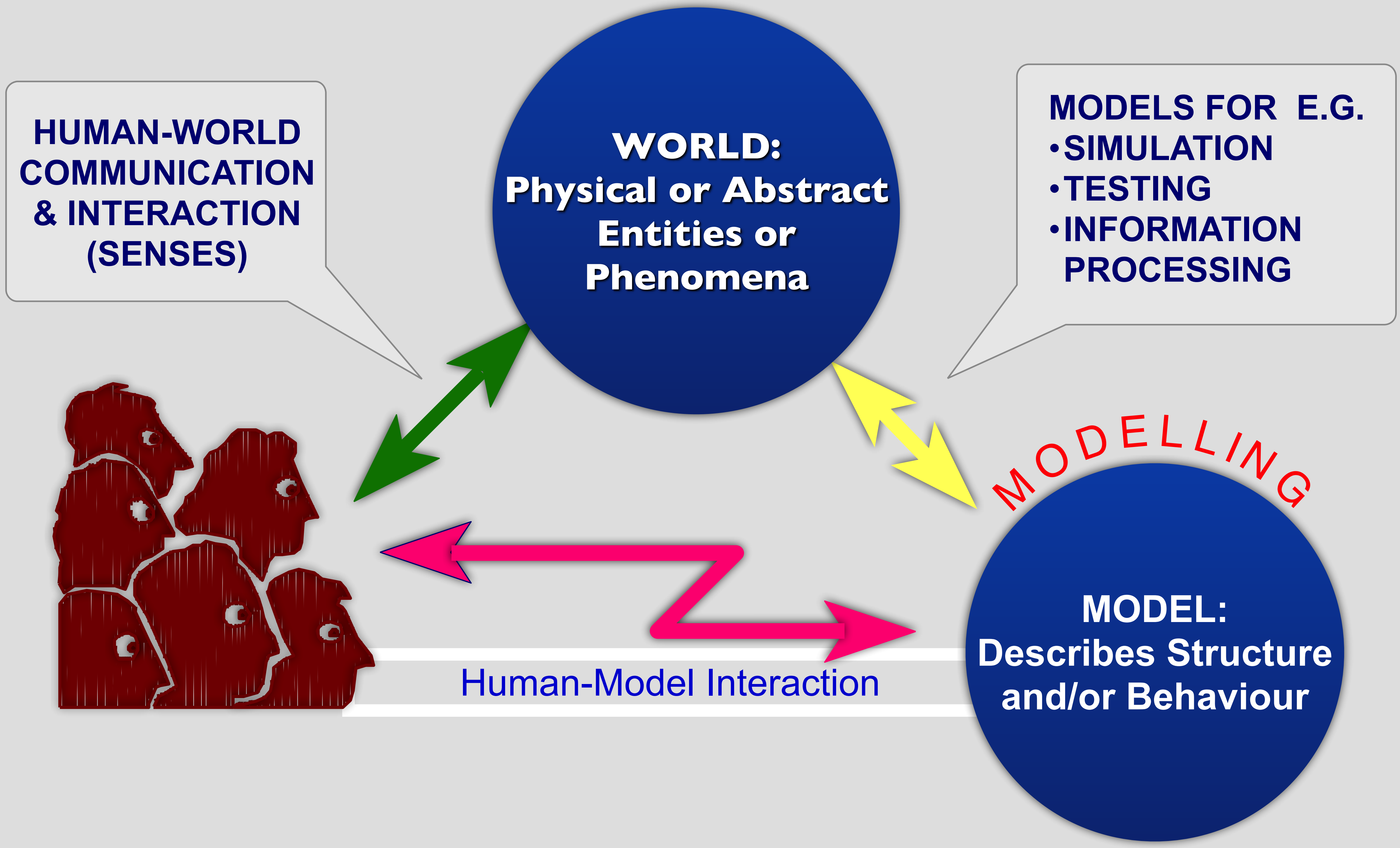
Hypotheses & Models

- Models describe hypotheses about phenomena (physics, chemistry, engineering, economics, ..) in mathematical terms
  - Static phenomena (Babylonian Algorithms ca. 6000 years ago [4], Pythagoras, etc.)
  - Dynamic phenomena (Newton, Leibniz, etc.)

Model Construction

- To construct a model the problem to be solved must be understood - at least in part.
- Analysis of collected data (observations, experiments) regarding a phenomenon can determine structures that enable or enhance insight to describe this phenomenon.
- If a problem is ill-defined, not understood or ill-posed: no suitable model can be defined.

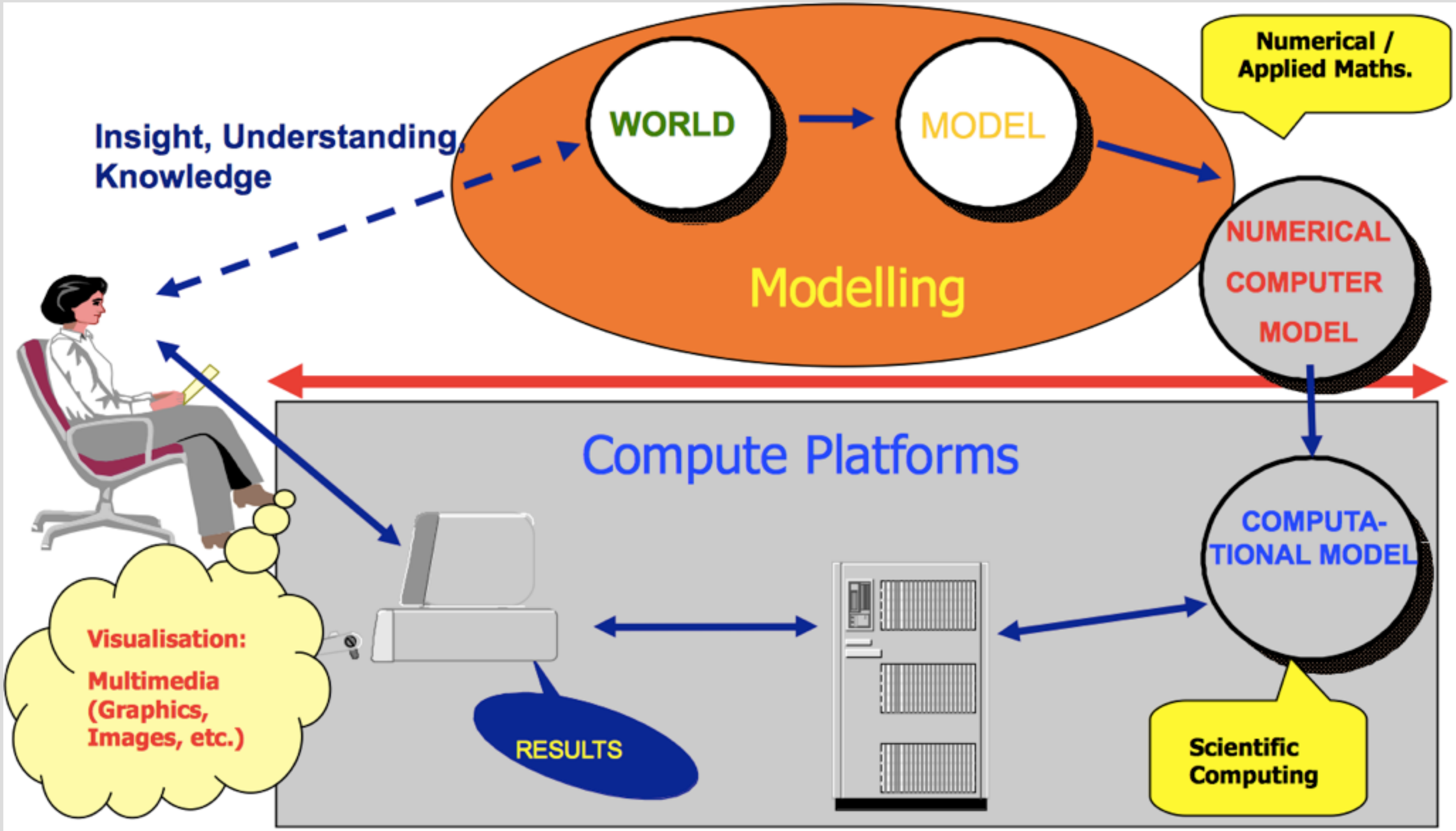
Human-World Interaction



Computational Models

- Models of complex real-world problems often too complex to solve analytically
- Such models can be approximated by numerical (computer) models

Computational Models



Models: Disadvantages

- Tedious and complex to construct
- Numerical approximations can be difficult to define: accuracy, stability, scalability
- Software implementations: often complex and time consuming

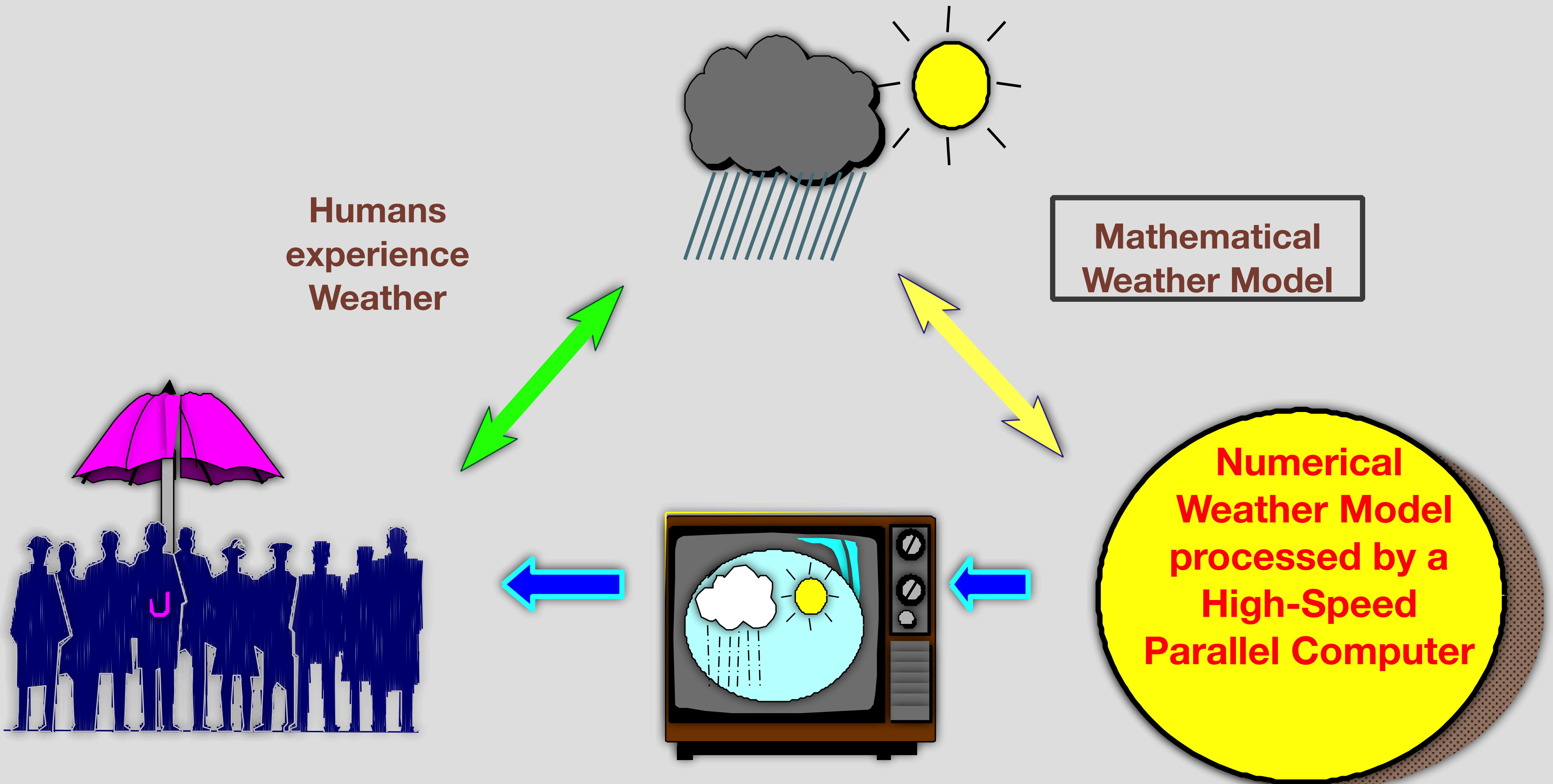
Models: Advantages

- Construction of models give insight into problems considered
- Models allow for example to:
  - Do sensitivity analyses
  - Prioritize salient factors
  - Define data needed and/or additional measurements required

Model Based Weather Forecast

- To compute a multi-day weather prediction appropriate weather data must be collected (atmospheric pressure, temperature, satellite images, ...)(See Jean Luis Vazquez-Poletti)
- An appropriate model describing the dynamic nature of weather progression must be developed (Complex, compute intensive -> parallel platforms.)
- Model: Improves insight into weather behaviour
- Advantage: Models can be improved.

Weather Forecast



3. Purposed Data & Data Pools

Purposed Data

- Problems/phenomena often recognised, but not well understood => no hypotheses, no models possible
- Scientific approach:
    - Collect data (observe, measure, collect and use standardised metrics, formats)
    - Analyse: detect patterns (structures)
    - Insight -> Problem formulation
    - Hypothesis definition -> Model construction
- Data collected with a particular goal: **Purposed Data**.

Recorded Data

- Data often collected as **recorded data** that are not aimed at solving a particular problem
- Routinely collected data e.g.:
  - Human behaviour (Internet, supermarket)
  - Patient data (hospitals, etc.)
  - Traffic flow (cities, highways, air & sea)
  - Satellites (weather, crops, movements)
  - Insurance, financial data
- Collected data stored: **Data Pool** (Satoshi Matsuoka: **Data Silo**).

Data Pool

- Collected data stored in a **Data Pool** may be used for other purposes than those originally intended.
- Such data can be analysed from different points of view
- Examples:
  - Medical records: correlations between various symptoms (ailments) and treatment results
  - Geodata: Data on water, sewerage, communication networks -> city planning, transport systems, etc.

Purposed Data v. Data Pools

- A fundamental difference exists between
  - **Purposed Data** collected for solving a particular problem, with defined formats and data elements, meeting specific quality and accuracy requirements and
  - Data in **Data Pools** used for previously unintended and unplanned purposes, such as searching for previously unknown and unforeseen patterns or integration with other similar data sets.
- Data in a Data Pool may have uncertain quality, accuracy, etc. <u>with respect to new (alternate) pattern searches</u> (analytics).
- Note: All data considered here may be **structured** or **unstructured**,

Big Data = Data Pools?

- Consider the two views:
  - **Purposed Data** collected for solving a particular problem,
  - **Data Pools** used for other purposes than originally intended

From an application point of view

Purposed Data = Data

Data Pools = Data Silos = Big Data

The data set size must meet the requirements of the problem(s) to be solved, and may be relatively small.

Why then **BIG** Data? **Big in value, not in size?**

Big Data = Data Pools

- **Problems related to Data Pools, i.e. Big Data:**
  - **Integration/Combination** of different data sets
  - **Quality**
  - **Formatting**
  - **Metrics,** etc.
  - See comments by e.g. Ian Foster, Geoffrey Fox, Sudip Dosanjh, Marcel Kunze, etc.

4. Summary

Summary

- **Big Data** does not obviate formal analysis and modelling
- The **size** of a data set does not change the standard problem solving paradigm.
- **Data (Purposed Data)** is an inseparable component in modelling complex real-world problems
- **Data Pools (Data Silos, Big Data)** offer new insights through **Analytics**

Conclusions

- **Data** repositories will continue to increase dramatically
- What is big today will be small tomorrow
- To record, process and store these expansive data collections **HTC** parallel systems are needed.

References

- [1] Woodie, Alex: The Storytelling Mandate of Big Data (2014) http://www.datanami.com/2014/06/13/storytelling-mandate-big-data/
- [2] Anderson, Chris: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, Wired Magazine 16.07 (2008)
- [3] Graham, Mark: Big Data and the end of Theory? The Guardian (2012) (see also wikipedia.org) http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory/
- [4] Knuth, Donald E.: Ancient Babylonian Algorithms, Communications of the ACM, 15 (1972), 671-677
- [5] Lopez, I.: On Algorithm Wars and Predictive Apps (2013) http://www.datanami.com/2013/05/15/on_algorithm_wars_and_predictive_apps/
- [6] Big Data, http://www.wikipedia.org