

Topic in Part III?  
(Research data infrastructure of Japan )

# Inductively Think about Impacts of Open Platforms on Research

オープンなプラットフォームが  
研究に与える影響を帰納的に考える

Department of Informatics

Kyushu University, Japan

Daisuke Ikeda

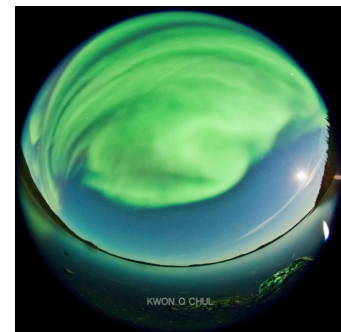
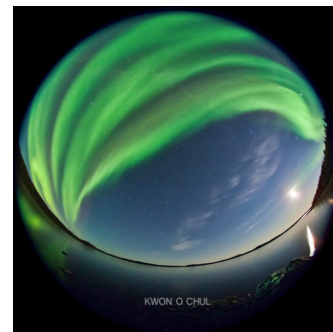
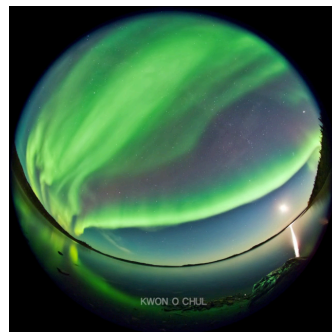
[daisuke@inf.kyushu-u.ac.jp](mailto:daisuke@inf.kyushu-u.ac.jp)



# My talk is about “What is e-Science?”

## Tentative title of my talk in the proposal:

- “E-Science and research data 2  
- in case of the physics of the upper atmosphere –”
- I’m expected to talk about the following research:
  - “Forecasting Aurora Substorms from Observed Data with a Supervised Learning Algorithm”, Tanaka et al., 11th IEEE Intl. Conf. on eScience, Munich, Germany Aug 31- Sept 04, 2015



<http://www.kwonochul.com/>

*Is this a Science?*



## Approach: Two Questions about e-Science

**What is a difference between impacts of the 3rd and 4th pillars of Science?**

**Is it an e-Science if we use data and analysis tools, such as machine learning?**

# Educational Background

**Bachelor:** *Dept. of Physics*, Faculty of Sciences, Kyushu Univ.

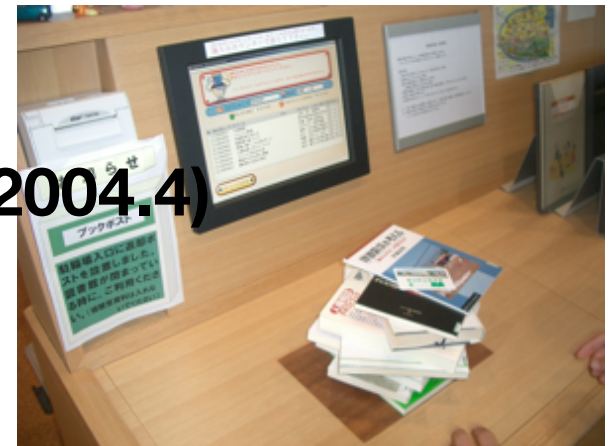
**Master:** Dept. of Information System, Interdisciplinary Grad. School of Engineering Sciences, Kyushu Univ.

**Doctor:** Dept. of Informatics, Grad. School of Information Science and Electrical Engineering, Kyushu Univ.

# Job & Research Career

## Kyushu Univ. Computer Center (1997.4~2004.4)

- HPC, Campus network and security
- Research: Web mining, text mining



## Kyushu Univ. Library (2004.5~2006.7)

- Infrastructure of scholarly communication, such as institutional repository, and Automatic identification (e.g., bar codes, RFID, IC cards)
- Research: Text mining (information extraction from the Web, spam detection)



## Dept. of Informatics, Kyushu Univ. (2006.8~)

- Research: Bioinformatics, e-Science, information retrieval, data infrastructures

# Research Topics in My Lab

## Data Analysis (Inductive approach):

- Mining from search histories and access logs
- Graph mining
- Finding a pure pattern in genomic sequences
- Mining from geotagged text (Tweets)
- Forecasting latent popular hashtags
- Authorship estimation from a Japanese historical texts
- Simulation for Emerging New Words (deductive)

## Data Infrastructure:

- Context-aware information retrieval
  - create context vector from corpus
- Institutional Repository
- Infrastructure for Data Repository
- Database for Science

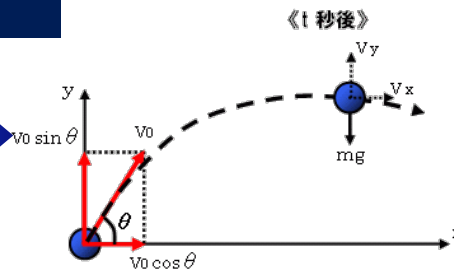
# Four Pillars for Science



# Traditional Pillars of Science

## 1. Theory

$$M \frac{d^2 x}{dt^2} = F$$



Deductive

General models  
→ Individual phenomena

- 速度
- x方向:  $v_x = v_0 \cos \theta$
  - y方向:  $v_y = v_0 \sin \theta - gt$
- 距離
- x方向:  $v_x = t \cdot v_0 \cos \theta$
  - y方向:  $v_y = t \cdot v_0 \sin \theta - \frac{1}{2}gt^2$

Inductive

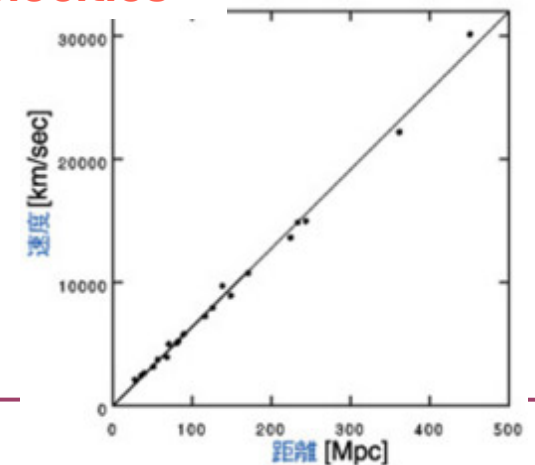
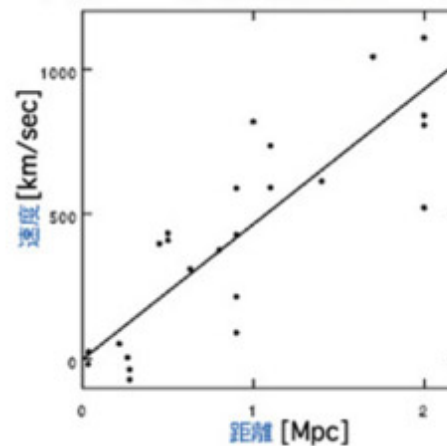
Individual instances  
→ General models

## 2. Experimentation



a falling body and its mass

distance to astral bodies vs.  
their recessional velocities





# New Pillars of Science

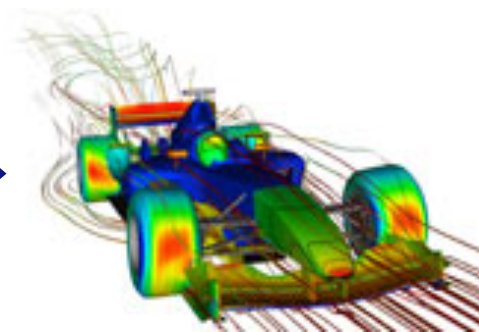
Deductive

**1. Theory**

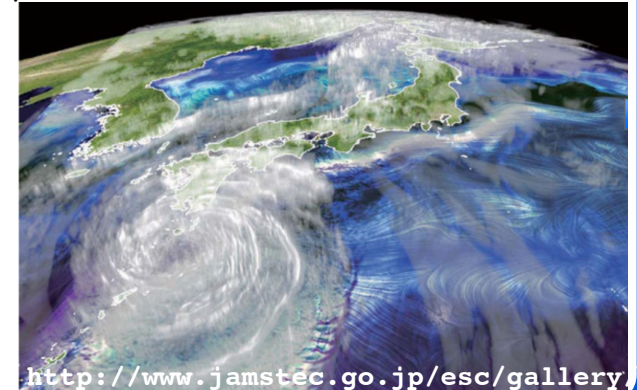
$$M \frac{d^2 x}{dt^2} = F$$



<http://www.fltechnical.net/news/7720>



**3. Simulation**



<http://www.jamstec.go.jp/esc/gallery>

**2. Experimentation**



**4. Data Science**

High-speed network  
sensing technologies  
data analysis tools

Inductive

# Examples in 3rd and 4th Pillars

## 3rd Pillar (simulation):

- Computer Simulation of Typhoons
- Computer Simulation of Atomic Bombs
- Computer Simulation of Forest Fires
- Computer Simulation of Stock Markets

→ **Same Research Style (closed-style) with HPC**

## 4th Pillar (data science):

- Computer Shougi (Japanese chess) Programs
  - Machine learning with records of professional games
- Automatic Driving
  - pattern recognition, such as image recognition
- Forecasting Influenza Epidemics with Search Records

→ **High accurate recognition and prediction, but without (scientific) mechanism**

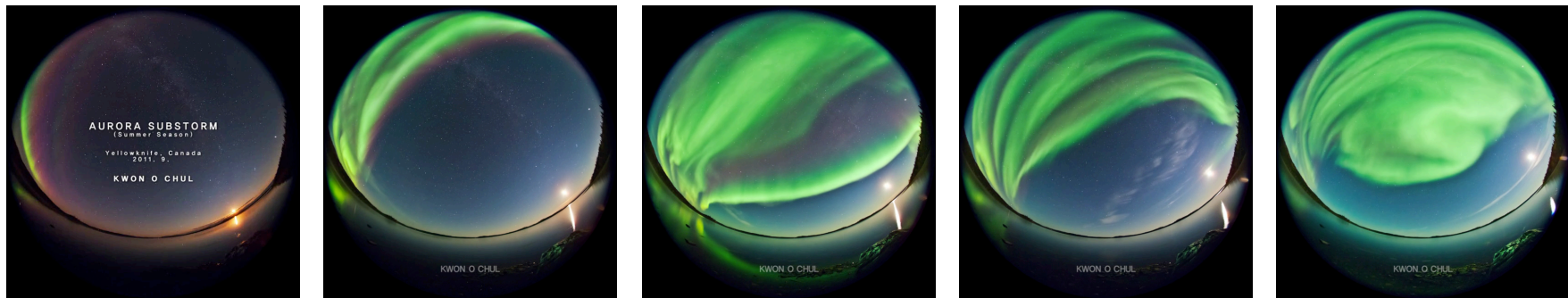
# Impact of Open Platforms to Research



# Example of e-Science: Auroral Substorms

## Auroral Substorm:

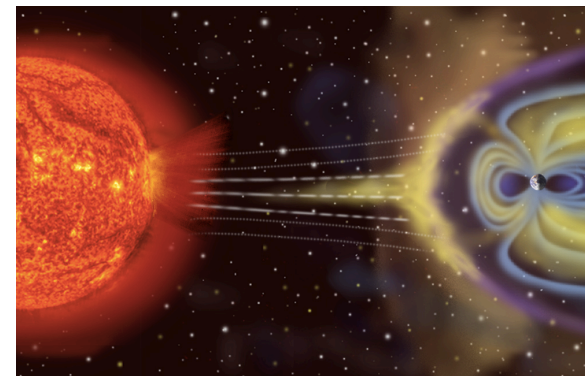
- sudden brightening and increased movement of auroral arcs
- detailed physical models still remains a subject of dispute.



<http://www.kwonochul.com/>

## (Supervised) Machine learning from solar wind and geomagnetic field data

- training data: data with labels, that is, data collected when the phenomena happen or not.
- parameter tuning
- algorithm creates a *model* to classify unlabeled data.



<http://sdo.gsfc.nasa.gov/mission/spaceweather.php>

# Process of “Auroral Substorms”

## 0. Hearing from domain experts

- available data and its basic characteristics

## 1. Find data related to auroral activities

- all-sky images from NIPR (National Inst. of Polar Research)  
(<http://polaris.nipr.ac.jp/~acaaurora/aurora/Tromso/>)
- solar wind from CDAWeb ([http://cdaweb.gsfc.nasa.gov/istp\\_public/](http://cdaweb.gsfc.nasa.gov/istp_public/))
- geomagnetic field data from World Data Center for Geomagnetism, Kyoto  
(<http://wdc.kugi.kyoto-u.ac.jp/>)

## 2. Create Training Data

- manually identify when the phenomena happened

## 3. Apply some machine learning algorithms

- LibSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>)

## 4. Evaluate the results with domain experts

Open Platforms



*Collaboration with unknown people*



# Open Software for Science

## Impact for business: LAMP

- Linux (OS), Apache (Web server), MySQL (DB), PHP/Perl/Python (Programming language)
- Programming skills are necessary to use LAMP.

## Impact for Science

- Programming Language (including R)
- Scientific Libraries
  - Linear Algebra
  - Fourier Transform
  - Plotter (Graphics)
  - Machine Learning
  - Data Mining
  - Statistics
  - Natural Language Processing

***Easy to use state-of-the-art data analysis tool***

# Fine-grained Unit of Scholarly Communication

## Data Journals

Ecological Researchは日本生態学会が刊行する英文誌で、水域・陸域を問わず、あらゆる生態学の領域において、生態学についての理解を本質的に発展・変化させる論文を掲載します。単純な記載論文や、既存の研究論文の単純な繰り返しや改変は査読対象となりません。出版される論文のタイプは、Original articles, Current topics in ecology (総説), Special features, Technical reports, Notes and comments, **Data papers**, Forumです。また、「Biodiversity in Asia—アジアにおける生物多様性」についての論文を掲載したいと計画しています。

## Database/Software Papers

## Ecological Research (Ecological Soc. of Japan)

3. 論文の種類 : 本論文誌は次の3種類の論文を受け付けます。なお、いずれの論文に対しても査読が行われます。

【オリジナル論文(original paper)】

バイオ情報学に関するオリジナルな研究成果について述べた論文

【サーベイ論文 (survey paper)】

バイオ情報学に関する既存研究のサーベイを行った論文。ただし、読者にとって有用性のあるサーベイである必要があります。

【データベース・ソフトウェア論文(database/software paper)】

開発し、かつ、公開したデータベースもしくはソフトウェア (Webサーバーを含む) を紹介する論文です。方式自体は既存のものを用いていても構いませんが、他のデータベースやソフトウェアと比較して有用性があることが必要です。最終フォーマットに換算して3ページ以上の長さで

**Inter- & Trans-disciplinary Researches**

IT CC Trans. on Bioinformatics

# The 4th Pillar for Complex Phenomena

**Complex phenomena are left to be solved, such as environmental issues and geoscience.**

**In such a field, the approach of the 4th pillar would be useful.**

- even though it does not directly yield a mechanism of phenomena.

**Habu, Yoshiharu (top professional of Shougi):**

**from now, human beings are tested if we can understand a move derived by a computer algorithm and we can derive the same move.**

- He said after a retired top professional took a beating in a match with a computer algorithm.



# Conclusion

**e-Science is important for understanding complex phenomena.**

**To do that, the followings are important:**

- collaboration with domain experts
  - find data, creating labels, and evaluating results
- collaboration with data scientists
- open data and open software

## One more thing ...

**Creating a common mental model of “Data” is important.**

- mental model of “Paper” seems to be clear.

**For that, keep in mind there exists two types of approach**

### **1. DB like approach:**

- This approach requires that DB is abstracted by metadata.
  - But, “Data” is difficult to be abstracted in general.

### **2. IR (Information Retrieval) like approach:**

- This approach requires text index.
- PR: KAKENHI (Grants-in-Aid for Scientific Research) project  
Project Number: 15H02787  
Basic idea: data = list of words  
data1 = [“aurora”, “substorm”, “geomagnetic”, ...]