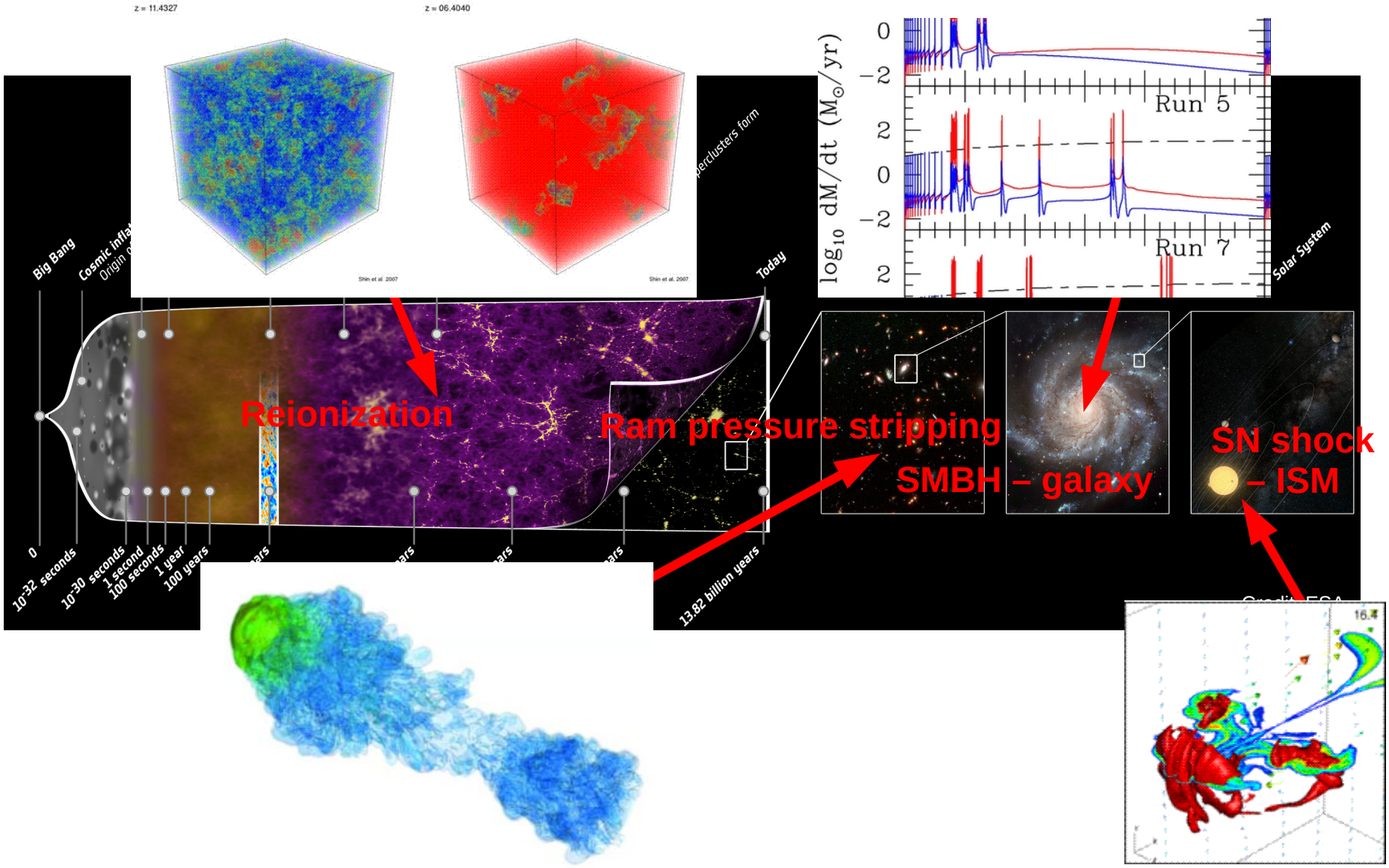
천문 우주 연구 분야에서 클라우드 컴퓨팅 활용 가능성

신민수
한국천문연구원
msshin@kasi.re.kr
http://astromsshin.github.io

HPC: 수치 모델 중심의 연구

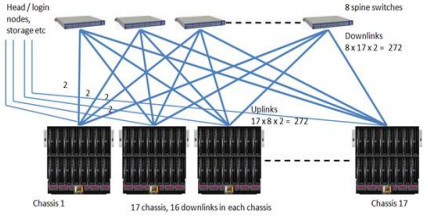


- Defining
  - *momentum density*: $\boldsymbol{\pi} \equiv \rho \mathbf{v}$,
  - *stress tensor*: $\mathbf{T} \equiv \rho \mathbf{v}\mathbf{v} + (p + \frac{1}{2}B^2)\,\mathbf{I} - \mathbf{B}\mathbf{B}$,
  - *total energy density*: $\mathcal{H} \equiv \frac{1}{2}\rho v^2 + \frac{1}{\gamma - 1}\,p + \frac{1}{2}B^2$,
  - *energy flow*: $\mathbf{U} \equiv (\frac{1}{2}\rho v^2 + \frac{\gamma}{\gamma - 1}\,p)\mathbf{v} + B^2\mathbf{v} - \mathbf{v} \cdot \mathbf{B}\,\mathbf{B}$,
  - *(no name)*: $\mathbf{Y} \equiv \mathbf{v}\mathbf{B} - \mathbf{B}\mathbf{v}$,

. . .

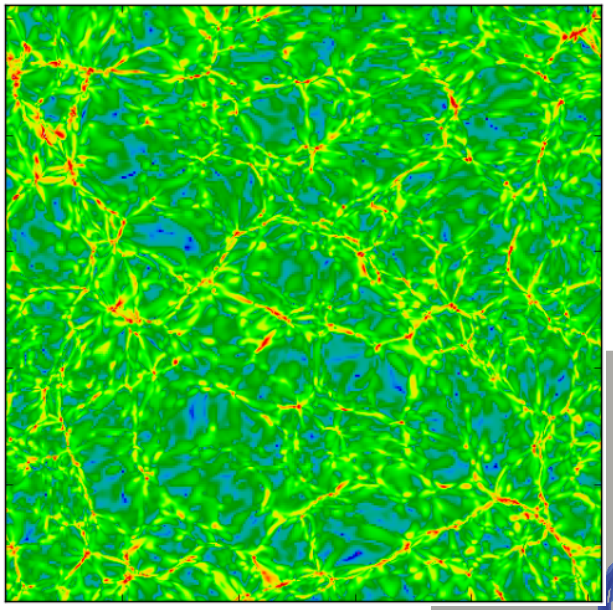
Cambridge COSMOS SHARED MEMORY Service
: the largest single system image in the UK
at 14.8TB of RAM. SGI cluster with Intel MIC.

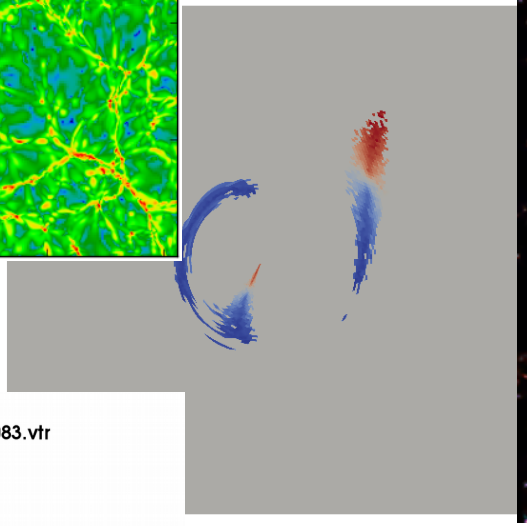
Edinburg 6144 node blue Gene/Q
: co-developed by Peter Boyle in Edinburg Physics and IBM.
100K cores.

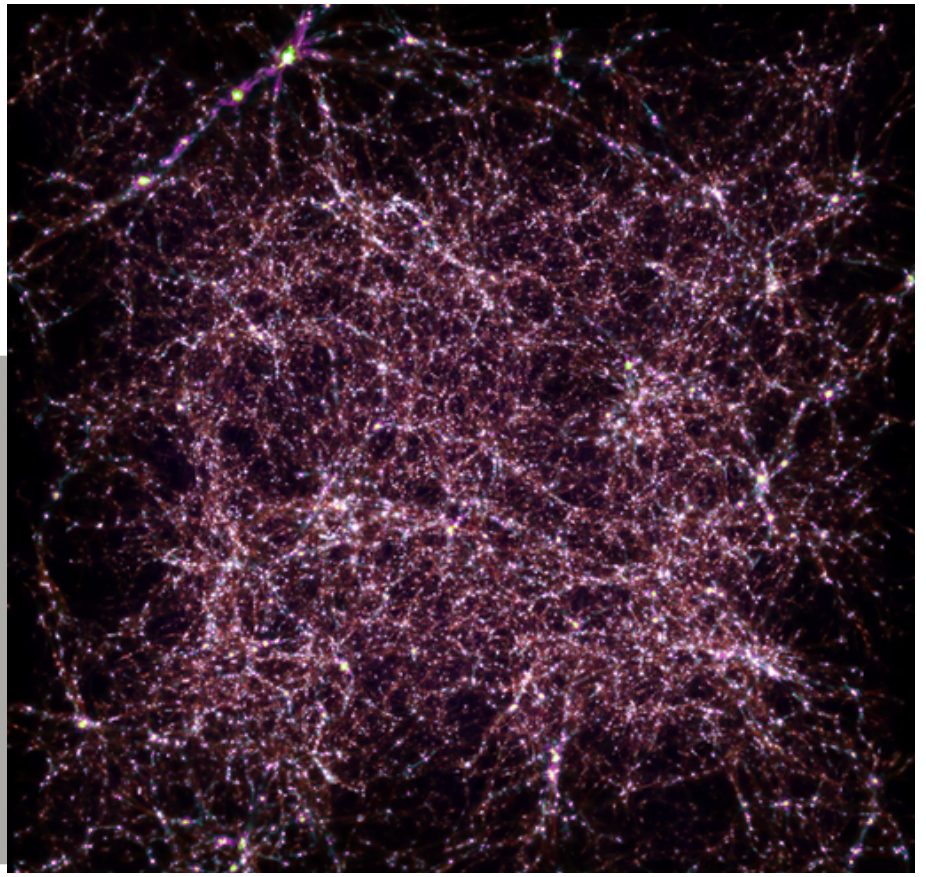


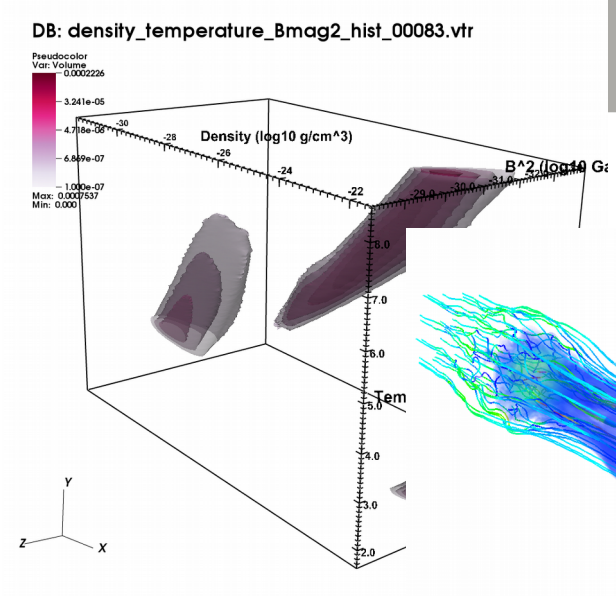
Leicester IT Services: Complexity Cluster
: targetted towards simulations with a large memory Footprint but difficult load balancing by using a new switching arthitecture designed by Leicester and HP.

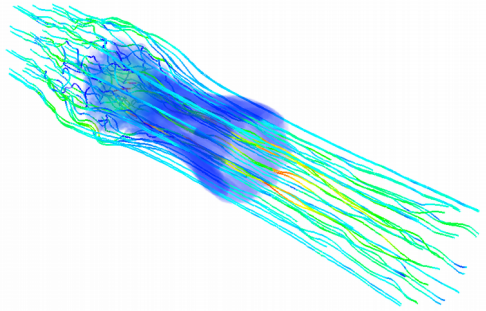
가시화 역시 중요한 역할을 수행



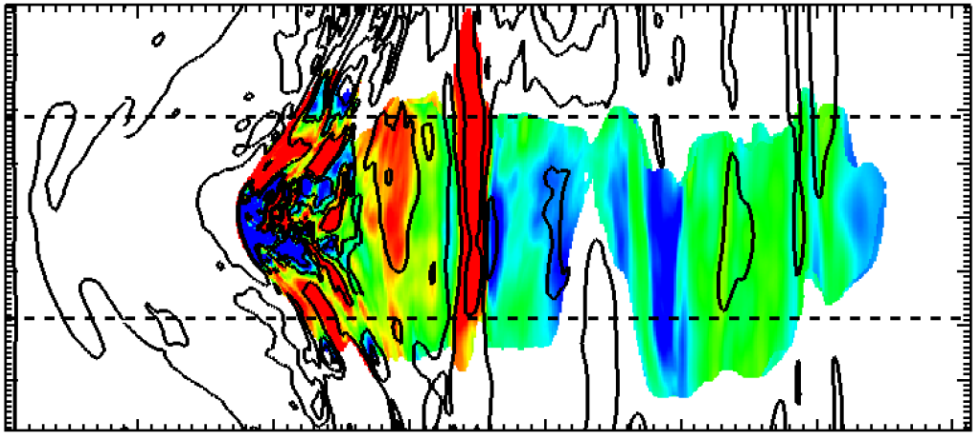




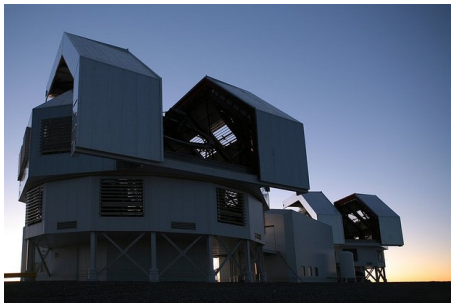




1.7e-06 1.9e-06 2.1e-06 2.3e-06 2.5e-06



HTC: 천문학 자료 분석 중심의 연구



관측 기기

원천 자료 저장 + 기본 자료 처리를 통한 분석 자료 생산

공개용 자료 저장 + 자료 분석을 위한 추가 자료 처리

지역의 계산 자원을 이용한 대용량 자료 분석 중심의 연구

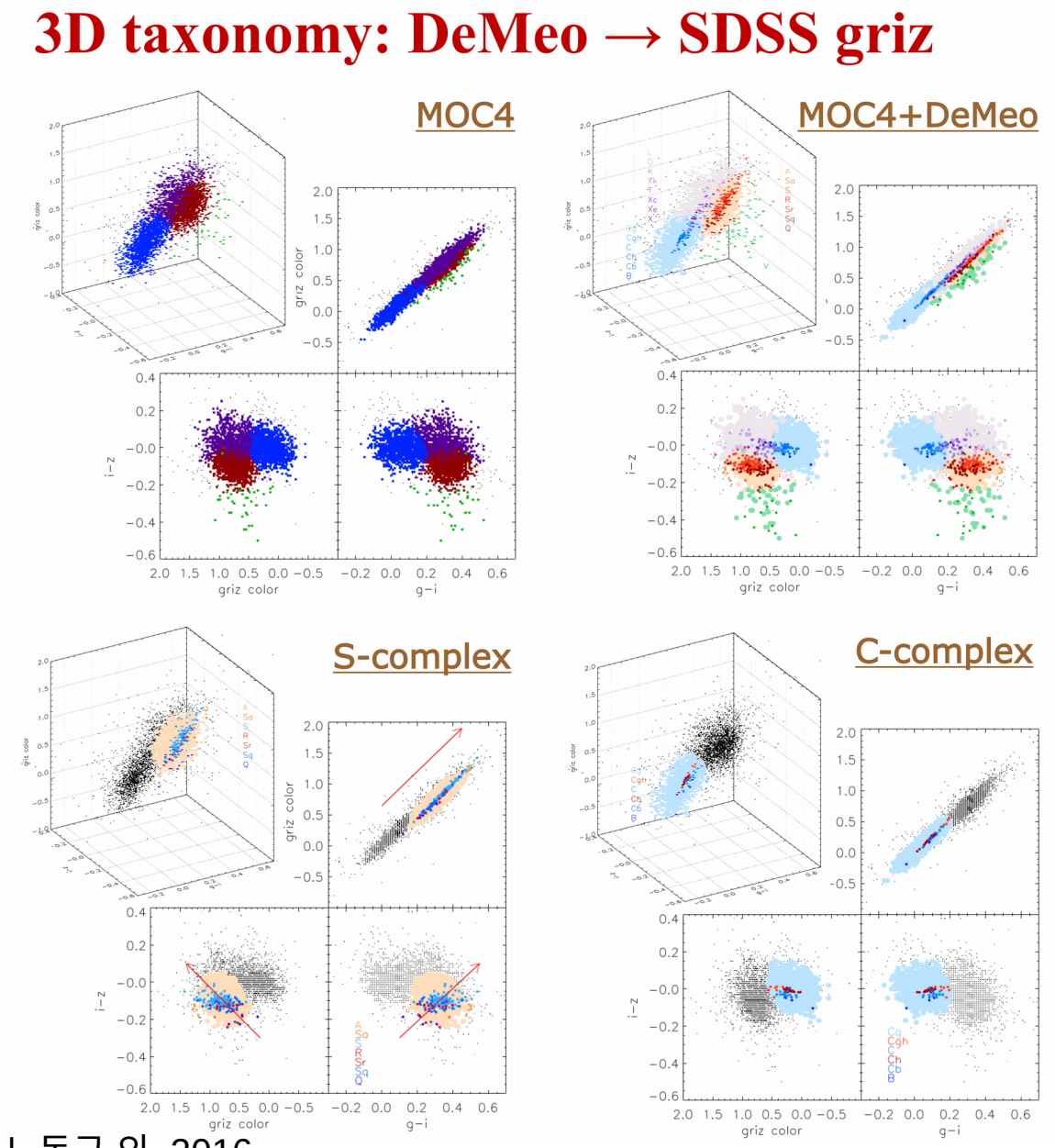
데이터 베이스나 파일 등을 통한 대용량 관측 자료 및 관련 부속 자료 저장

자료 처리 및 분석의 방법들은 자료의 크기가 커짐에 따라, 그 적용이 어려워지며 비용 또한 증가.

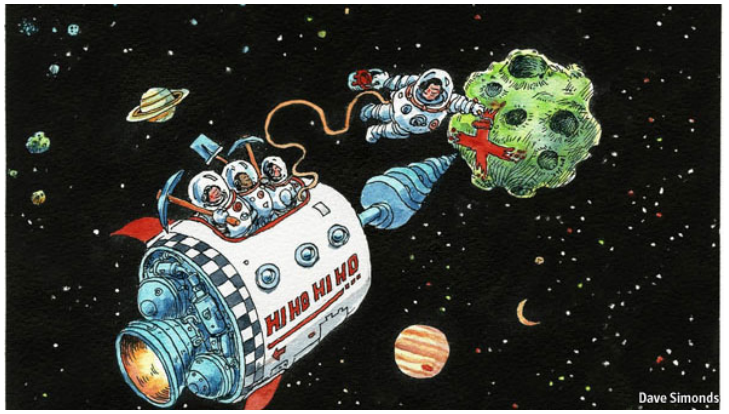
천문학 자료 분석의 몇 가지 유형들

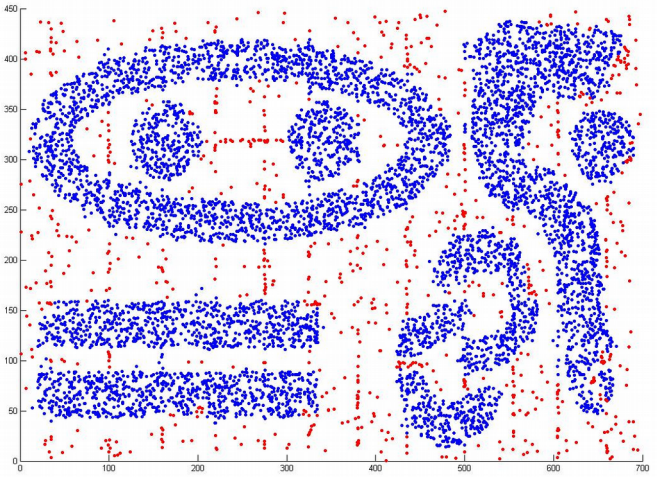
- **이미지 자료 분석.**
  - N^2 혹은 NlogN 연산이 많이 요구됨.
  - 타 분야의 이미지 분석의 경우와 유사한 방법들이 활용됨.
  - 대부분 file system을 통해서 자료 I/O이 발생.
- **시계열 및 목록 자료 분석.**
  - 역시 N^2 혹은 NlogN 연산이 많음.
  - 대부분 file system이나 database system을 통해서 자료 I/O이 발생.
- **모델과 관측 자료의 비교가 가장 흔하며, 일반적인 통계적 분석(statistical inference)이 이에 해당.**
  - Frequentist or Bayesian analysis.
  - Markov Chain Monte Carlo method는 활발히 이용됨.
- **기계학습법(machine learning)을 활용한 분석으로, classification, clustering, regression이 가장 활발히 이용되고 있으나, 아직 statistical inference가 주도적으로 활용됨.**
  - Optimization이나 MCMC에 많은 연산이 요구됨.

Machine learning 의 예제들



노동구 외. 2016





소행성들의 색상 정보들을 이용해서 유사한 것들의 군집을 확인.

Star–galaxy classification using deep convolutional neural networks

Edward J. Kim[1]★ and Robert J. Brunner[1,2,3,4]
[1]Department of Physics, University of Illinois, Urbana, IL 61801, USA
[2]Department of Astronomy, University of Illinois, Urbana, IL 61801, USA
[3]Department of Statistics, University of Illinois, Champaign, IL 61820, USA
[4]National Center for Supercomputing Applications, Urbana, IL 61801, USA

딥 러닝의 활용 사례: 관측된 이미지에서 별과 별이 아닌 천체의 모습을 구별해서 분류.

Accepted 2016 October 12. Received 2016 October 4; in original form 2016 June

ABSTRACT

Most existing star–galaxy classifiers use the reduced summary information from catalogues, requiring careful feature extraction and selection. The latest advances in machine learning that use deep convolutional neural networks (ConvNets) allow a machine to automatically learn the features directly from the data, minimizing the need for input from human experts. We present a star–galaxy classification framework that uses deep ConvNets directly on the reduced, calibrated pixel values. U Canada–France–Hawaii Telescope to produce accurate and well-calibr conventional machine learning tech success with current and forthcomi and the Large Synoptic Survey Tel manual feature engineering.

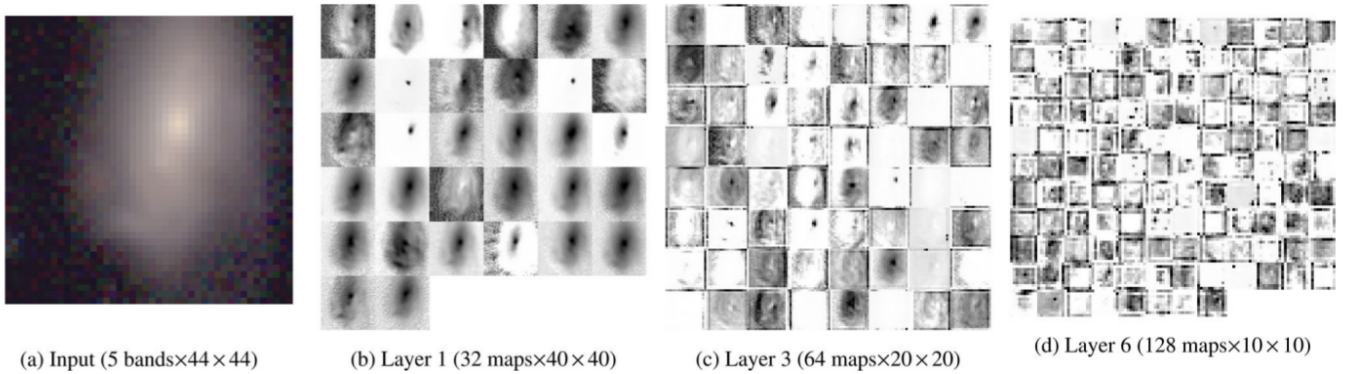


Figure 6. (a) A sample 44 × 44 RGB image of a galaxy in the CFHTLenS data set. The RGB image is created by mapping R → *i*-band magnitude, G → *r*-band magnitude, and B → *g*-band magnitude. (b) Activations on the first convolutional layer when a 5 × 44 × 44 image is fed into the network. (c) Activations on the third convolutional layer. (d) Activations on the sixth convolutional layer. Each image in (b), (c), and (d) is a feature map corresponding to the output for one of the learned features.

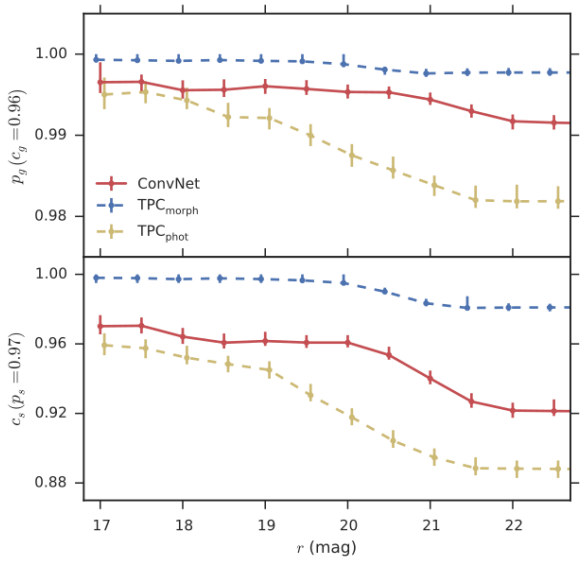


Figure 9. Galaxy purity and star completeness as functions of the *r*-band magnitude for the integrated counts in the SDSS data set.

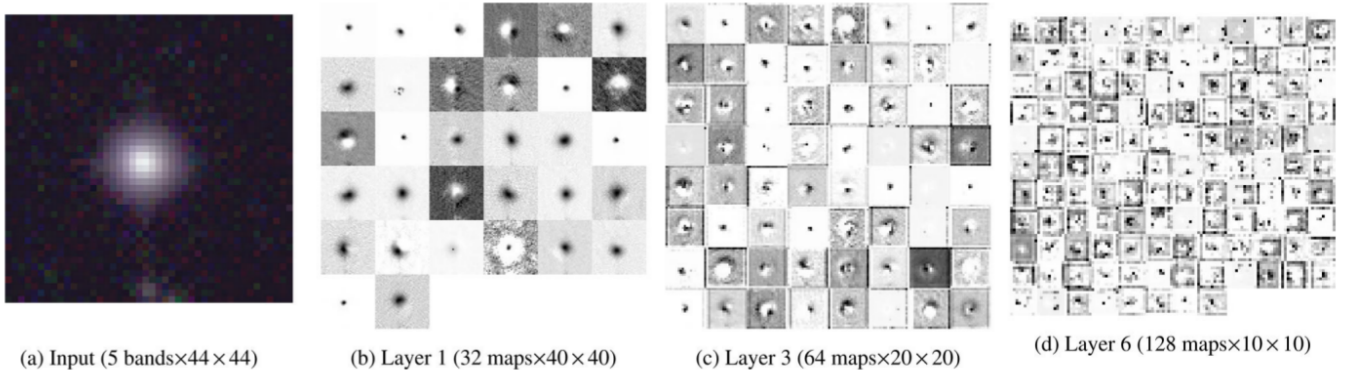


Figure 7. Similar to Fig. 6 but for a star in the CFHTLenS data set.

천문학에서의 클라우드 활용

- 기본적으로 저비용으로 유연하게 필요에 맞추어 전산 자원을 활용하고자 하는 목적.
  - 새로운 전산 자원 활용 실험의 경우도 포함 (e.g., Chameleon Cloud; https://www.chameleoncloud.org/).
  - Hadoop ecosystem (Hbase, GeoMesa, OpenTSDB).
- 천문학의 operational demand와 analysis demand로 구별할 때, analysis demand가 빈번하나 필요 자원의 규모로는 operational demand가 클 것으로 예상됨.
  - 항시성/시급성의 유무.
- 천문학의 연구 특성을 고려할 때 다양한 특수성이 반영된 software stack이 활용됨.
  - PaaS나 IaaS에 대한 수요가 주가 됨.

몇 가지 활용 사례

- Gemini Observatory Archive
  - 미국, 캐나다, 브라질, 아르헨티나 등의 국제 컨소시엄이 운영하는 천문대의 관측 자료 저장/배포 서비스는 AWS 활용.
  - 전산 자원의 구매/유지 및 관련 운영 요원 유지 등의 비용 절감의 목적.

8 August 2016

The new Gemini Observatory archive: a fast and low cost observatory data archive running in the cloud

Paul Hirst; Ricardo Cardenes;

Author Affiliations +

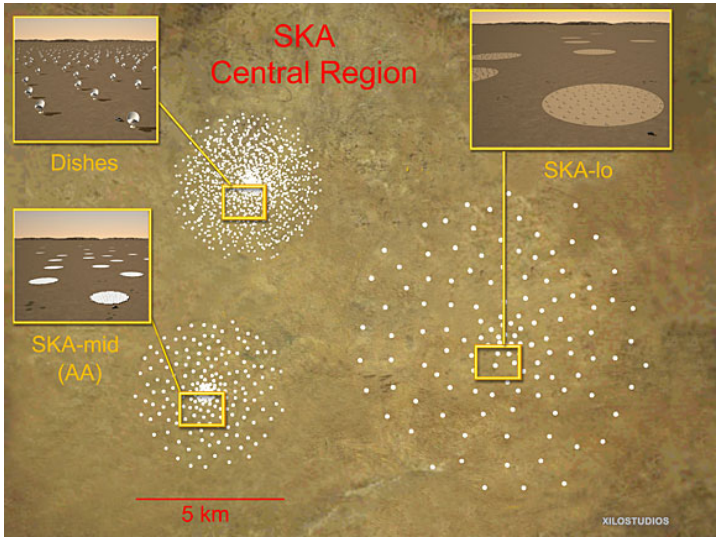
Proceedings Volume 9913, Software and Cyberinfrastructure for Astronomy IV; 99131E (2016); doi: 10.1117/12.2231833

몇 가지 활용 사례

- National Radio Astronomy Observatory의 CASA
  - 전파 천문 자료 분석을 위해 개발해서 배포하는 CASA 프로그램으로 일시적으로 대용량 관측 자료 분석을 수행하고자 하는 연구자들을 위한 AWS image 제공 등.
  - Parallel computing을 위한 AWS EC2 활용.
  - 연구자가 접근할 수 있는 전산 자원이 제한적일 때를 고려한 지원.
  - https://casa.nrao.edu/casa_aws_introduction.shtml

몇 가지 활용 사례

- AWS – SKA의 AstroCompute in the Cloud
  - SKA (Square Killometre Array) 프로젝트의 준비를 위하여 백만 달러에 해당하는 AWS 자원 기부.
  - 연구 제안서를 통한 AWS 자원 할당.
  - AWS S3에 최대 1PB까지 저장 지원 및 EC2를 같이 활용하는 조건.

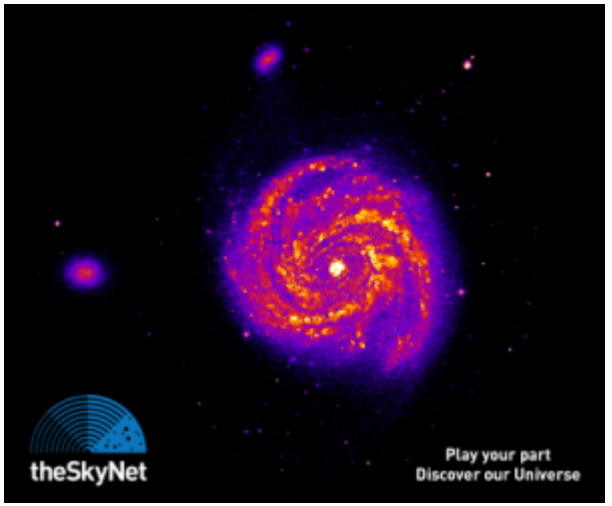


몇 가지 활용 사례

- 호주의 ICRAR SkyNet
  - 은하들의 이미지 자료로부터 물리량을 추론하는 crowd-sourcing 프로젝트 수행.
  - AWS Grant in Education으로 시작하여 진행되고 있으며, crowd-sourcing 참여자 규모에 맞추어서 on-demand로 Route 53, EC2, EBS, Glacier, S3 등의 자원 활용.

theSkyNet

theSkyNet.org



몇 가지 활용 사례

- STScI의 AWS를 통한 자료 제공/분석
  - 공개된 우주 망원경 자료를 AWS에 미리 준비해 둔 상태에서 연구자들의 분석에 AWS 자원 활용 예상.
  - 실제 활용 방법 등의 사례는 https://mast-labs.stsci.io/ 참고.
  - Space Telescope Science Institute Uses AWS to Increase Accessibility of Astronomical Data https://youtu.be/vK1dWqwyI9M

MAST Labs

Home

Experiments with software & computing, astronomical archives, and data science. Brought to you by the team @ MAST.

6 DEC, 2018

TESS data available on AWS

> *tl;dr - Sectors 1 & 2 from TESS are available on Amazon Web Services (AWS). In this first post, we'll introduce a basic method for accessing the data programmatically through the `astroquery.mast` client library.*

With the release of TESS sectors 1 & 2, we're making calibrated and uncalibrated full frame images, two-minute cadence target pixel and light curve files, and co-trending basis vectors, and FFI cubes (for the Astrocut tool) available in the `s3://stpubdata/tess` S3 bucket on AWS.

These data are available under the same terms as the public dataset for Hubble, that is, if you compute against the data from the AWS US-East region, then data access is free.

Accessing the data

In what follows, we are going to assume you already have an AWS account, have created AWS secret access keys and are able to create an authenticated session using the `boto3` Python package with these keys.

Astroquery & Boto3

AWS Government, Education, & Nonprofits Blog

Hubble Space Imagery on AWS: 28 Years of Data Now Available in the Cloud

on 06 SEP 2018 | in Education, Government, Public Sector | Permalink | Share

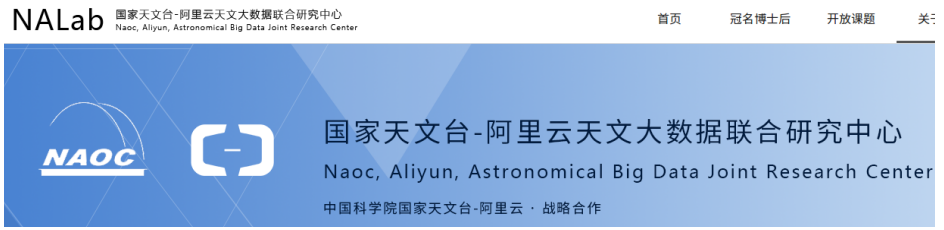
Since going live in 1990, the Hubble Space Telescope has delivered groundbreaking images to broaden our understanding of the universe. Each image captured by the telescope is archived and made publicly available, free of cost, by NASA through the Space Telescope Science Institute (STScI).

The Hubble images archive is used by a global community of astronomers, researchers, and engineers and has led to the discovery of distant galaxies and nebulae. "The legacy is a treasure trove of data that can be mined in the future," Arfon Smith, head of data science at STScI, said.

몇 가지 활용 사례

- 중국 National Astronomical Observatories, Chinese Academy of Sciences (NAOC)와 Alibaba Cloud의 협력
  - LAMOST (Large Sky Area Multi-Object Fiber Spectroscopic Telescope)에서 얻어지는 수천만개 천체들의 분광 스펙트럼 저장/공유/분석에 클라우드 환경 활용.
  - 매일 약 50TB 규모의 자료가 획득 되는 FAST 전파 망원경에 대해서도 클라우드 환경 사용.
  - 2017년에 NAOC-Alibaba Cloud Astronomical Big Data Joint Research Center 활동 시작.

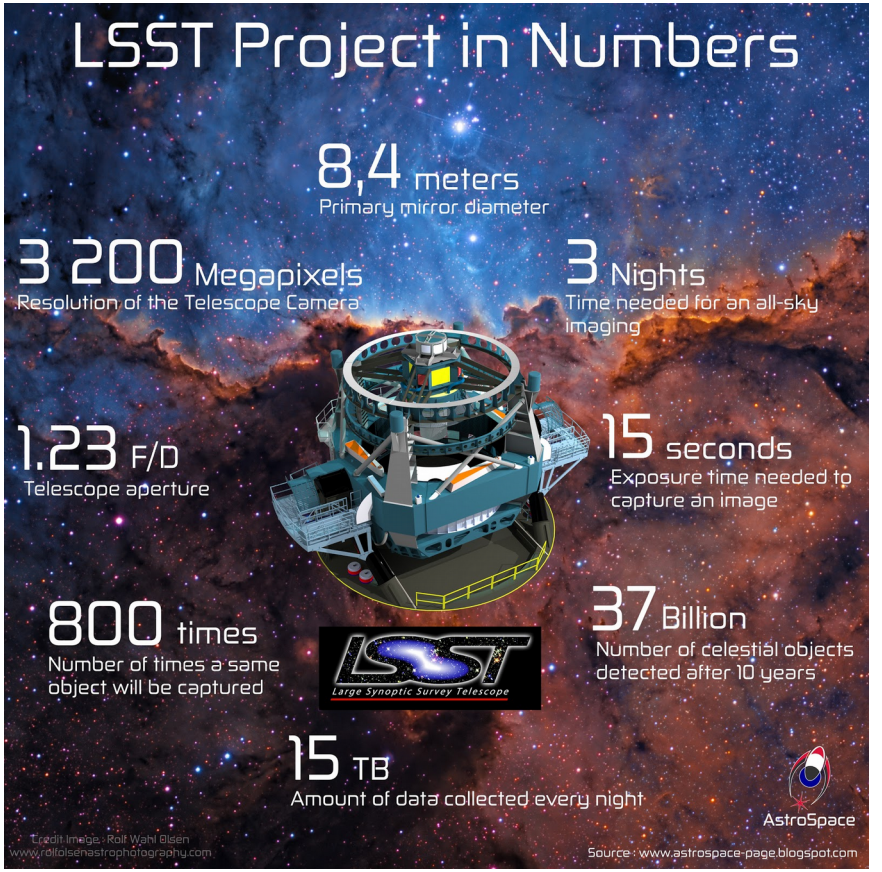
    http://nalab.china-vo.org/



몇 가지 활용 사례

- LSST의 자료 처리 등을 위한 클라우드 컴퓨팅 활용 가능성 탐색
  - LSST (https://www.lsst.org) 의 대용량 자료 처리 및 공유를 위한 전산 자원으로서의 클라우드 컴퓨팅 활용 가능성 검토 중.
  - AWS와 GCP를 이용한 proof-of-concept sutdies 수행 및 그 결과 검토 중.
    - Cloud technical assesment https://dmtn-072.lsst.io/
    - Potential proofs of concept for cloud deployment https://dmtn-078.lsst.io/
  - 참고 자료:

    https://agenda.hep.wisc.edu/event/1325/session/10/contribution/51/material/slides/0.pdf



결론: 천문학에서의 클라우드 활용 문제

- 천문학 연구 자금의 대부분이 공적 자금으로, 국가나 지역 슈퍼 컴퓨팅 센터와 같은 공공 전산 자원의 이용이 용이. 그런데 왜 cloud computing?
- 공공 전산 자원의 PaaS나 IaaS 형태로의 실시간 elastic usage이 보편적이지 않음. Cloud computing의 강점이 있는 부분이 존재.
- 일부 특정 경우에는 자료의 손실에 대한 피해가 크므로, 이에 대한 클라우드 활용에서의 기술적/정책적 구성 필요.
  - 인류 전체의 자료로서의 보존 가치 큼. 자료의 안전한 이송과 보관 및 법적 책임의 명확한 협의가 필요.
- 공공 연구의 특성을 고려할 때 연구비를 민간 클라우드 활용에 이용하는데 여러 현실적인 문제들 존재. Where is elasticity?
  - 연구비, 구매/계약 관리 및 집행 체계: 예, 단가 기준 계약 vs. 총 금액 기준 계약.
  - 정책/연구비 유연성 부재: 예, 실사용 요금 집행 vs. 고정된 계약 금액.
  - 다양한 경우(예, 보안성, 사용 목적 등)를 고려한 정책 설정 부재.
  - 민간 클라우드 컴퓨팅 활용 자체를 조건으로 하는 연구비 제공 프로그램 부재: 미국 NSF의 다양한 연구 프로그램 참고.

미국 NSF의 민간 클라우드 활용 프로그램 예들

Leading cloud providers join with NSF to support data science frontiers

Amazon Web Services, Google Cloud Platform, Microsoft Azure contribute initial tranche of up to $9 million in cloud credits through NSF's BIGDATA program

NSF's awards are paired with support from leading cloud computing providers.
Credit and Larger Version

February 7, 2018

The National Science Foundation (NSF) is providing nearly $30 million in new funding for research in data science and engineering through its Critical Techniques, Technologies and Methodologies for Advancing Foundations and Applications of Big Data Sciences and Engineering (BIGDATA) program.

NSF's awards are paired with support from Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure, which have each committed up to $3 million in cloud resources for relevant BIGDATA projects over a three-year period, beginning with this year's awards. A key goal of this collaboration is to encourage research projects to focus on large-scale experimentation and scalability studies.

"NSF's participation with major cloud providers is an innovative approach to combining resources to better support data science research," said Jim Kurose, assistant director of NSF for Computer and Information Science and Engineering (CISE). "This type of collaboration enables fundamental research and spurs technology development and economic growth in areas of mutual interest to the participants, driving innovation for the long-term benefit of our nation."

"... Beginning in 2017, NSF's BIGDATA program has included support in the form of access to cloud computing resources ("cloud credits") from **Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure; IBM** signed on in early 2018. **Each cloud computing provider has committed up to $3 million in cloud computing resources for BIGDATA projects over a three- or four-year period, beginning with awards in 2017**. Key goals of this public-private collaboration have been to encourage research projects to focus on large-scale experimentation and scalability studies, and to provide researchers access to cutting-edge, scalable platforms for developing new methods and solutions for today's most pressing data challenges. ..." - NSF

NSF and Internet2 to explore cloud computing to accelerate science frontiers

Cooperative agreement aimed to catalyze the use of cloud computing platforms for scientific computing

Conceptual design using cloud computing to process near-real-time streaming data.
Credit and Larger Version

November 15, 2018

The National Science Foundation (NSF) has announced a new cooperative agreement with Internet2, a nonprofit computer networking consortium, to build partnerships with commercial cloud computing providers and support science applications in new and more effective uses of cloud computing capabilities.

The NSF-funded Internet2 project, Exploring Clouds for Acceleration of Science (E-CAS), will investigate the viability of commercial clouds as an option for leading-edge research computing and computational science supporting a range of academic disciplines. Amazon Web Services and Google Cloud Platform have signed on as the initial cloud computing providers in this endeavor.

"The E-CAS project has the potential to not only demonstrate the effectiveness of current commercial cloud computing services in supporting a range of applications that are important to the science and engineering research communities, but also to enable these communities to leverage the innovative technologies and capabilities to significantly accelerate scientific discoveries," said Manish Parashar, director of NSF's Office of Advanced Cyberinfrastructure.

NSF has pledged $3 million for a two-phased approach to be managed by Internet2 and is expected to produce a deeper understanding of the use of cloud computing in accelerating scientific discoveries.

There are 6 projects chosen to participate in the first phase of the Exploring Clouds for Acceleration of Science (E-CAS) project based on **their need for on-demand, scalable infrastructure, and their innovative use of newer technologies such as hardware accelerators and machine learning platforms**.

- Accelerating Science by Integrating Commercial Cloud Resources in the CIPRES Science Gateway
- **Investigating Heterogeneous Computing at the Large Hadron Collider**
- **Ice Cube computing in the cloud**
- Building Clouds: Worldwide building typology modelling from images
- Deciphering the Brain's Neural Code Through Large-Scale Detailed Simulation of Motor Cortex Circuits
- Development of BioCompute Objects for Integration into Galaxy in a Cloud Computing Environment