우주 관측 대용량 자료와 그 분석을 통해 살펴본 (spatio-temporal) data mining

신민수
한국천문연구원

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

Introduction

한국천문연구원
Korea Astronomy & Space Science Institute

설립목적

천문우주과학의 연구개발을 종합적으로 수행하고 그 성과를 확산하여 천문우주과학의 국가적 발전을 달성

우리는 우주에 대한 근원적 의문에 과학으로 답한다

임무

- 천문학과 우주과학에 대한 연구 및 사업
- 대형 관측시설의 운영 및 기기개발
- 우주환경감시기술 개발사업 수행
- 역 및 표준시의 관리 등 국가 천문업무의 수행
- 대 국민 천문지식 및 정보 보급 사업
- 정부, 민간, 법인, 단체 등과 연구 개발 협력 및 기술용역 수탁·위탁
- 주요 임무 분야의 전문 인력 양성 및 관련 과학기술정책 수립 지원

KASI는 천문우주과학의 발전에 필요한 학술연구와 기술개발을 종합적으로 수행하는
국가 유일의 천문우주과학 전문 연구기관임

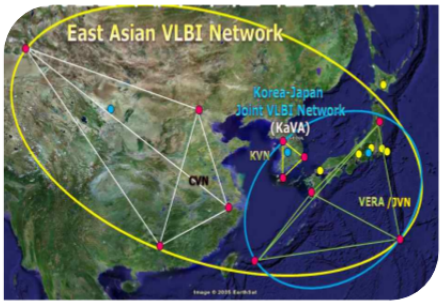
http://www.kasi.re.kr

국내 주요연구시설

- **소백산천문대 건설('74~'78)**
  * 61cm 반사망원경
- **대덕전파천문대 건설('82~'85)**
  * 14m 전파망원경
- **보현산천문대 건설('86~'96)**
  * 1.8m 광학망원경
- **한국우주전파관측망(KVN) 건설('01~'11)**
  - 서울(연세대), 울산(울산대), 제주(탐라대)
  * 21m 전파망원경
- **우주환경예보센터('07~'13)**
  * 태양 및 우주환경연구
- **글로벌데이터 센터('06)**
  * 세계의 GPS 관측자료 수집·제공
- **우주측지용레이저추적시스템('08~'16)**
  * mm 정밀도로 인공위성의 거리 측정
- **동아시아 VLBI 연구센터**
  * KVN 3기 및 일본, 중국, 대만을 포함한 동아시아 허브 역할 수행

국외 주요연구시설(전파망원경)

KVN-VERA (2012)



ALMA(2013, 국제공동운영)

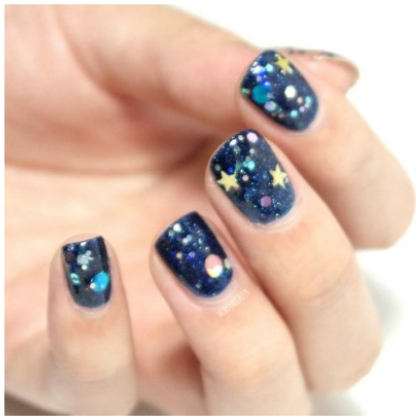
- **KVN-VERA (2012)**
  - KVN 3기, 일본 VERA 4기를 연결한 지름 2,300km급 운영
  - * 한국 KVN(Korea VLBI Network) 전파망원경 : 직경 21m
  - * 일본 VERA(VLBI Exploration of Radio Astrometry) 전파망원경 : 직경 20m
- **ALMA(**ALMA(Atacama Large Millimeter /submillimeter Array)망원경
  - 칠레 아타카마 사막 위치
  - 총 66개의 전파망원경으로 구성
    - * 직경 12m(54기), 7m(12기)
  - 미국, 유럽연합, 일본 등의 컨소시엄으로 구축
  - 천문(연)은 동아시아 지역 파트너 기관으로 합류(일본 국립천문대(NAOJ)와 MOA 체결

국외 주요연구시설(광학망원경)

LSST (Large Synoptic Survey Telescope) project
: 매일 밤 약 천 만 개의 밝기나 위치가 변하는 현상을 검출하여 매 분마다 전세계에 제공.
: 약 육 백 만 개의 소행성 궤도 정보.
: 약 370억개의 천체들 (200억개의 은하와 170억개의 별) 정보.

- **GMT**(Giant Magellan Telescope)
  - 칠레 라스캄파나스에 건설
  - 한국, 호주, 미국이 참여하는 주경 25.4m의 세계 최대급 광학망원경
- **우주물체 전자광학 감시시스템**(Optical Wide-field Patrol)
  - 세계 최초의 인공위성 무인 광학감시 구현(총 5기 구축 예정)
    * 0.5m 광시야 망원경
    ※ 해외관측소 1호기(몽골) 구축 : 2014
    해외관측소 2호기(모로코) 구축 : 2015
- **KMTNet**(Korea Microlensing Telescope Network)
  - 칠레, 호주, 남아공에 위치(총 3기)
  - 지구형 외계행성 탐색을 위한 시스템
    * 지름 1.6m CCD 카메라 3기

에뛰드 은하철도 999 네일 로 아주 쉬운 우주 네일! 해봤어요



안녕!

요즘 아~주 난리인
에뛰드의 은하철도 999 매니큐어!

저도 가장 맘에 드는 2가지 구입해서
아주아주 쉽고 블링블링 예쁜
우주 네일아트 해봤어요!!

Credit:
http://blog.naver.com/minsoo3o/220566508667

BIGBANG

천문학: 우주(즉, 시공간 자체)와 그 곳에서 만들어지는 구조(즉, 천체)에 대한 연구.

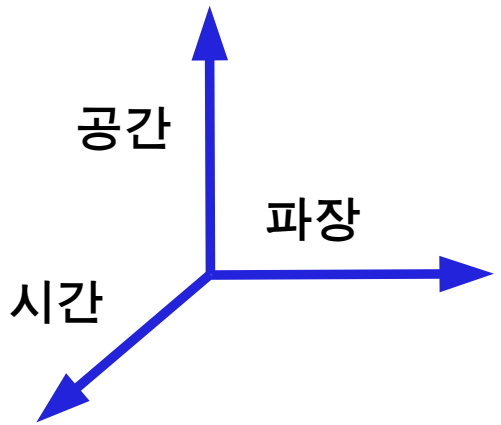
천문학에서 대용량 관측 자료

- 관측 자료가 증가하는 방향은 크게 세 가지.

  - 더 많은 공간을 탐사하는 경우 (더 어두운 천체를 보는 것이나 더 자세하게 관측하는 것도 포함).

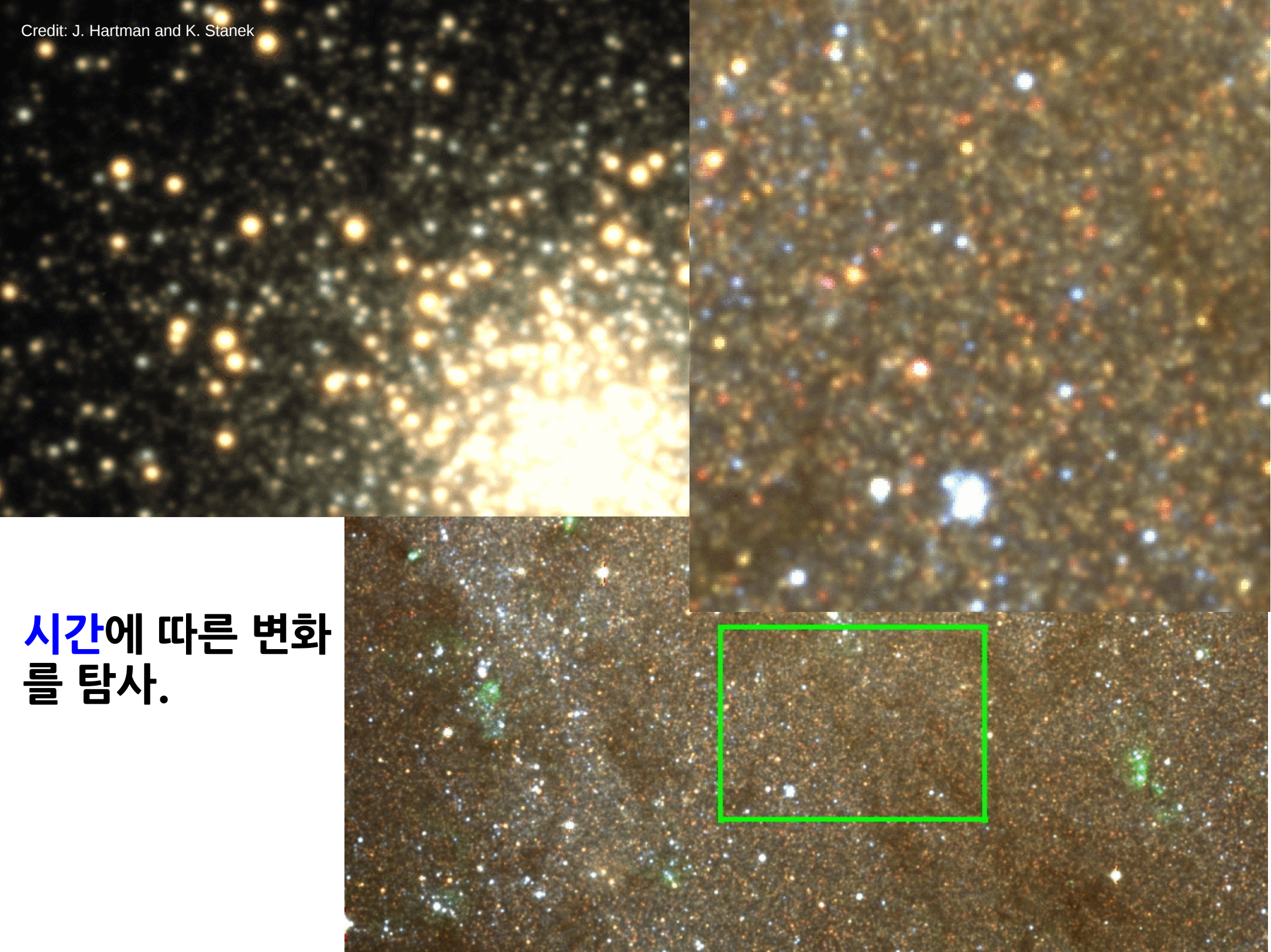
  - 시간에 따른 변화를 탐사하는 경우.

  - 다양한 빛의 파장 대역을 탐사하는 경우.



이 3차원 공간을 완전히 채우는 것이 "다 파장 광역 변화 탐사"

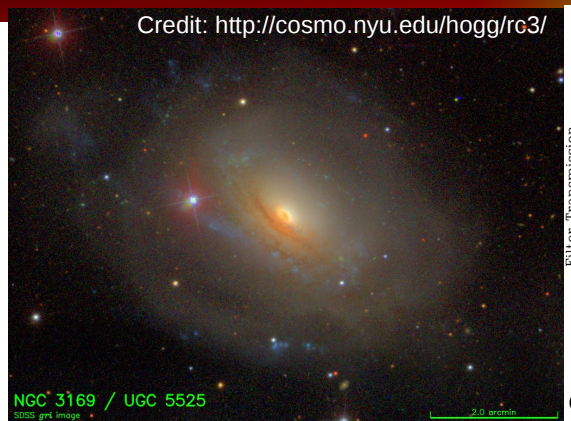
다수 고품질 센서 획득의 용이함.
→ 대용량 관측 자료 획득이 가능함.

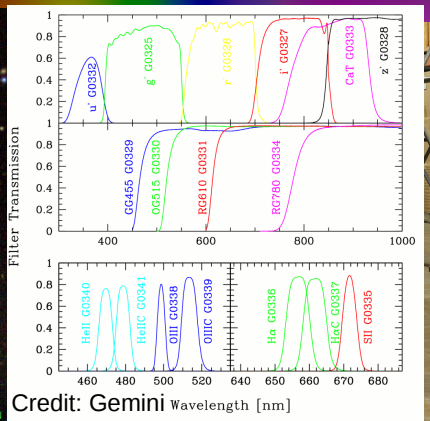


시간에 따른 변화를 탐사.

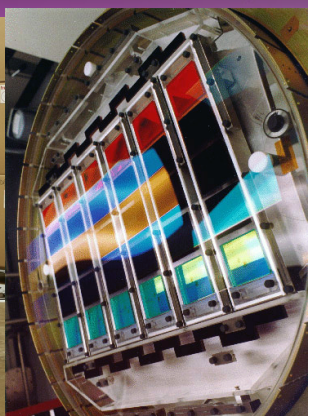
Credit: University of Chicago

다양한 빛의 파장 대역을 탐사.





Credit: Gemini



모든 천문학 자료는 Open Access를 원칙으로 공개.
– 우리의 관측은 인류 전체의 유산.

THE ROYAL SOCIETY

Home Fellows Events Grants, Schemes & Awards Topics & policy Journals Collections About us News

Final report - Science as an open enterprise

21 June 2012

The Science as an open enterprise report highlights the need to grapple with the huge deluge of data created by modern technologies in order to preserve the principle of openness and to exploit data in ways that have the potential to create a second open science revolution.

Exploring massive amounts of data using modern digital technologies has enormous potential for science and its application in public policy and business. The report maps out the changes that are required by scientists, their institutions and those that fund and support science if this potential is to be realised.

Examples

Astronomy and the Virtual Observatory

In the field of astronomy, scientists have for some time already recognised the importance of greater openness in science. Astronomers from around the globe have initiated the Virtual Observatory (VO) project to allow scientists to discover, access, analyse and combine astronomical data archives and make use of novel software tools. The International Virtual Observatory Alliance (IVOA) coordinates various national VO organisations and establishes technical and astronomical standards. The establishment of such standards is vital so that datasets and analysis tools from around the world are interoperable. Metadata are also standardised using the Flexible Image Transport System (FITS) standard and the more recent XML-based IVOA format, IVOTable. It is also an IVOA standard to register datasets in a registry, a sort of web-based Yellow Pages for astronomy databases. These are important to document the existence and location of datasets so that they can be easily found and accessed. IVOA itself collates a registry of registries.

https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf

National Optical Astronomy Observatory

Kitt Peak National Observatory • Cerro Tololo Inter-American Observatory • NOAO Gemini Science Center

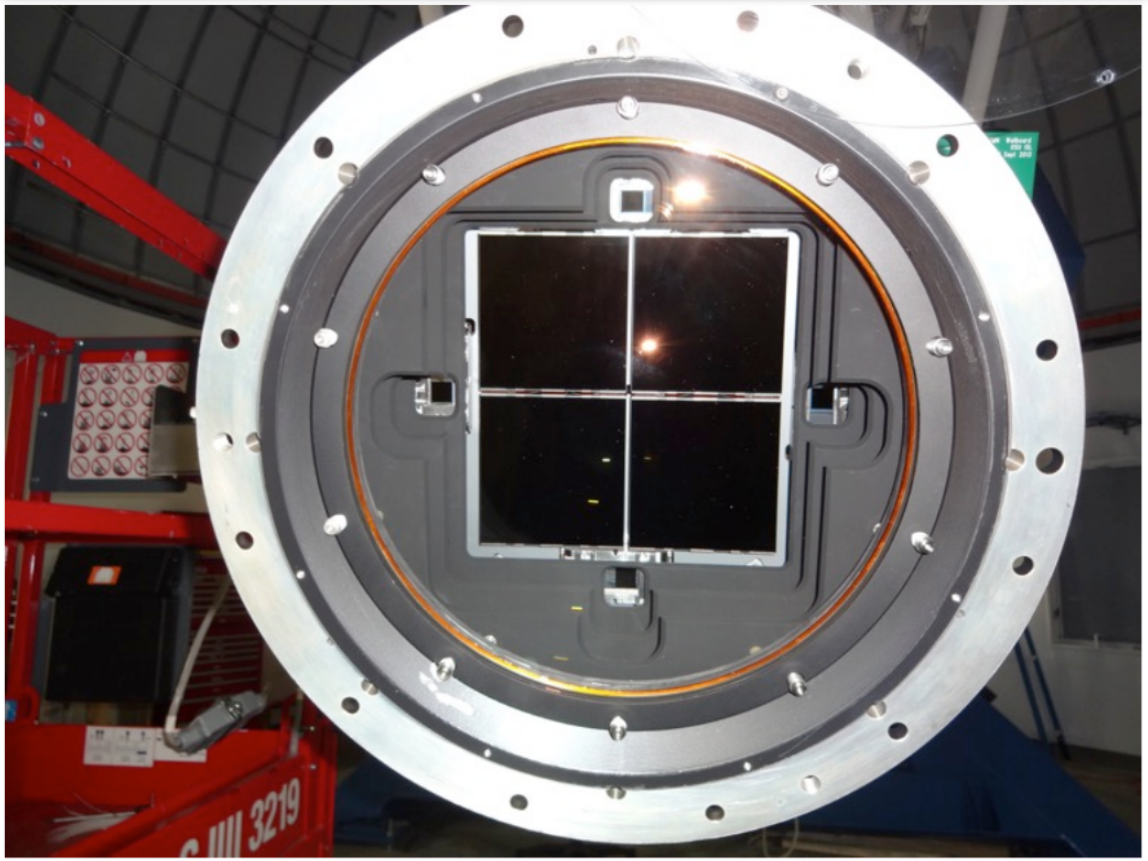
b. Proprietary Period

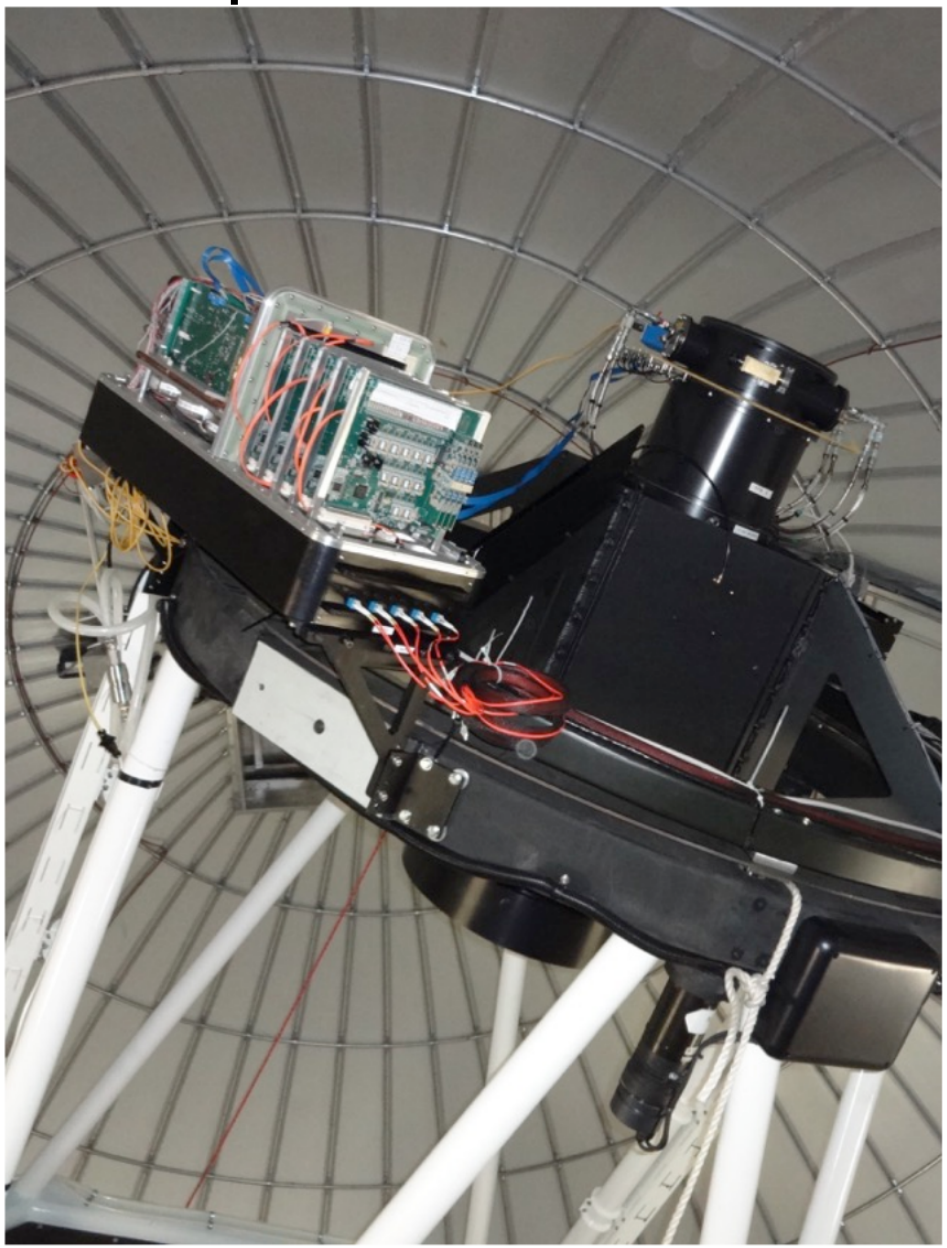
To promote the maximum scientific impact of data collected at NOAO facilities, all data will be retained and archived at NOAO, with the expectation that the data will be made available electronically to the world-wide astronomical community following a nominal period during which the proposing investigators have exclusive scientific rights. This "proprietary period" is subject to the following provisions:

예: KMTNet

http://kmtnet.kasi.re.kr

- 3.4억 화소(9kx9k pixels), 10 마이크론
- 0.40 arcsec/pixel, 2x2도 관측시야
- 60초 노출 I-band 측광정밀도 1% 17.5mag

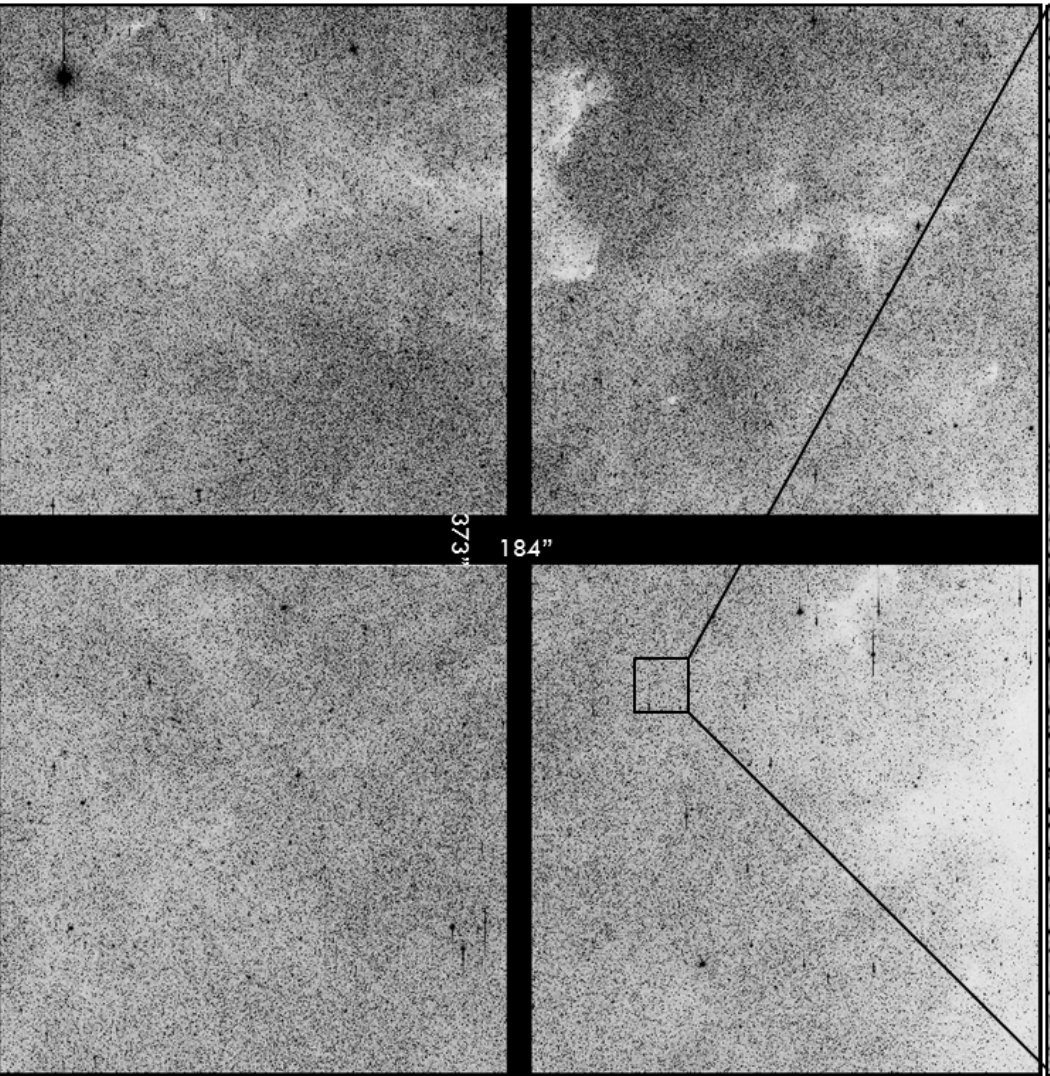


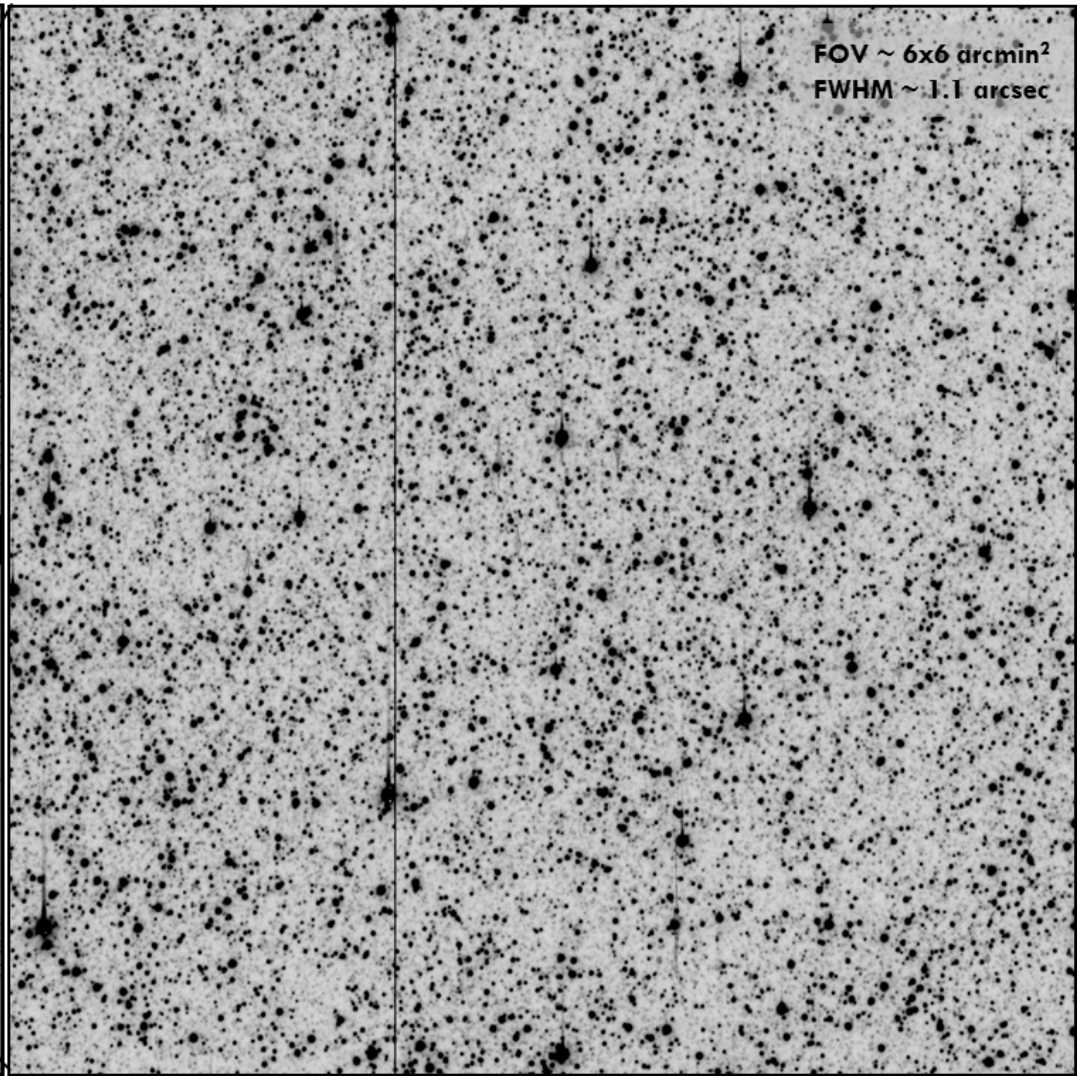


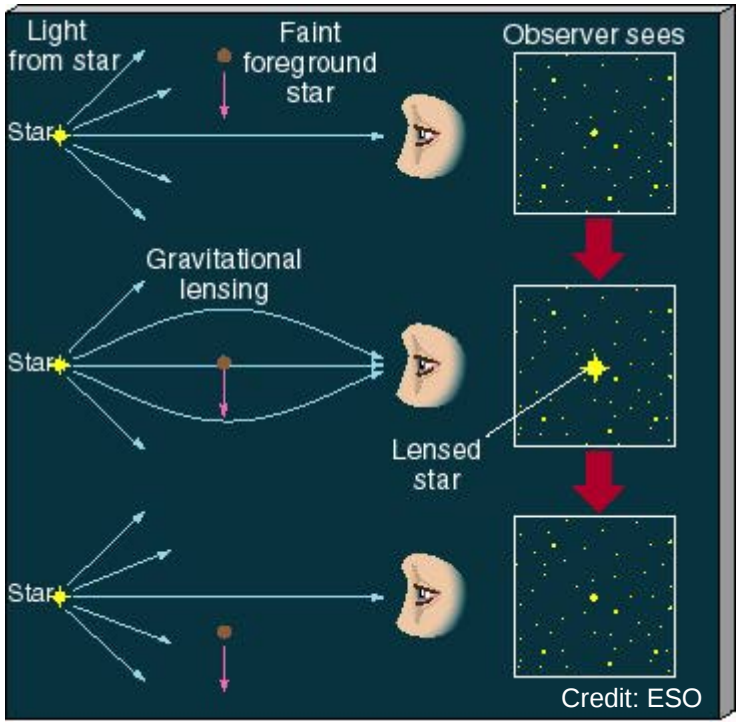
김승리

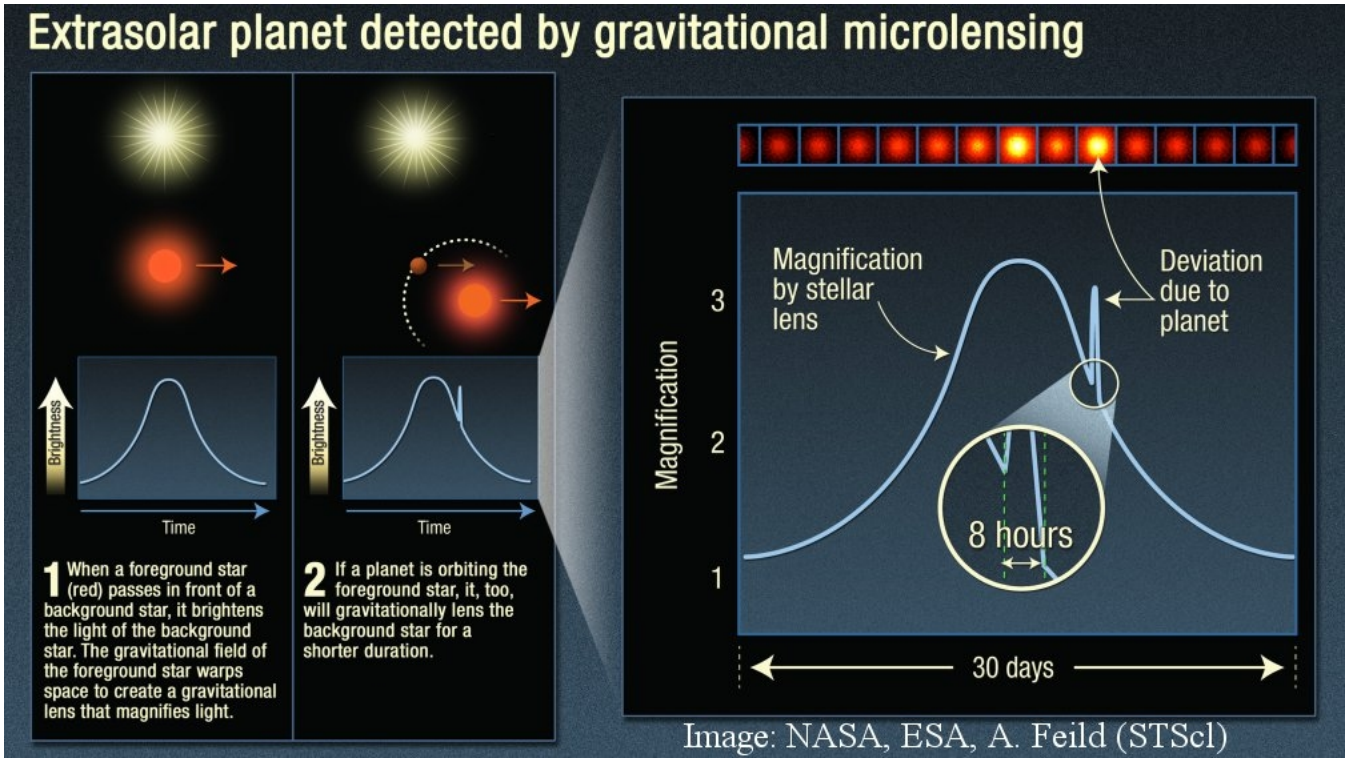
이충욱

– 남반구 3개의 동일 기기.
– 미세 중력 렌즈 현상을 이용한 외계 행성계 발견 및 특성 파악.
– 광시야 관측을 이용한 다양한 과학 연구 (초신성 폭발, 가까운 은하들, 우리 은하에 포섭된 왜소은하들, 태양계의 소행성들).









뉴스 뒤의 뉴스

한국이 만든 관측소, 태양계 밖 행성 2개 첫 발견

박근태 기자 2016-07-29 10:45:03

가 - | 가 + | 🖶 | ✉

한국천문연구원은 지구에서 우리은하 중심 쪽으로 각각 2000광년과 2만7000년 떨어진 곳에서 목성처럼 가스로 가득 찬 외계행성을 찾아냈다고 28일 발표했다. 이충욱 천문연 광학천문본부 변광천체그룹 책임연구원은 "지난해 발견된 2개 중력렌즈 사건을 KMTNet을 활용해 관측한 결과 최종적으로 목성형 외계 행성이 존재한다는 사실을 확인했다"고 말했다.

http://plus.hankyung.com/apps/newsinside.view?aid=201607282745A

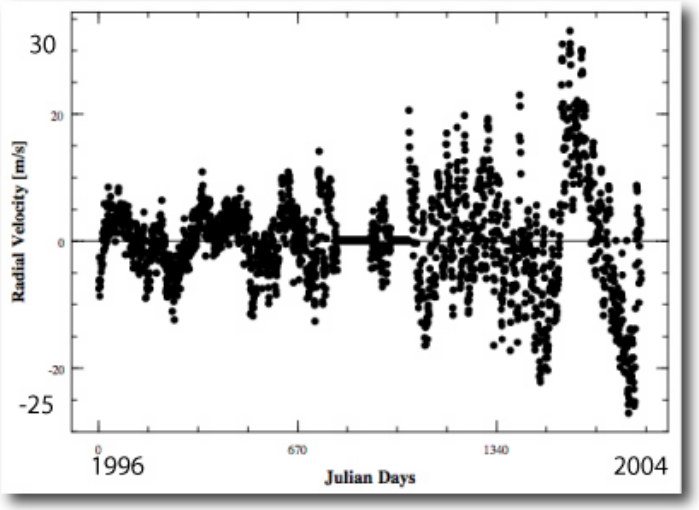
Applications of Statistics and Machine Learning methods

- 인간이 제공하는 자료로부터 **자료의 성질을 학습**하고, 나아가서 **학습 결과를 이용하여 특정 일을 수행**하는 컴퓨터 알고리즘.
- 자료와 수행하고자 하는 일에 따라서 **올바른 알고리즘을 선택하는 것은 인간의 몫** (아직은...!!!).
- 종류들: 1. 문제 (학습 자료의) 유형에 따라서, 2. 학습을 표현하는 방법에 따라서, 3. 학습을 평가하는 방법에 따라서.
- [KEY CONCEPT] LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION.
- "Learning: using a set of observations to uncover an underlying process."

Feature: data representation

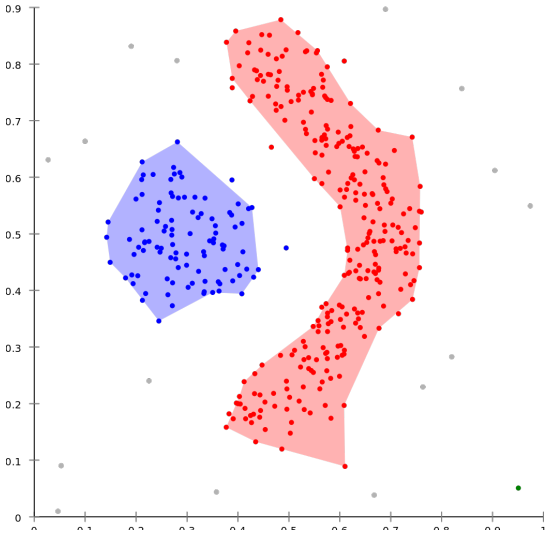
- [KEY CONCEPT] LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION.

- Design matrix = N (row; the number of samples) x D (column; the number of features).



ɹtes = covariates.

자료를 기술하는 값들 추정 및 선택! 컴퓨터가 이해할 수 있는 형태로!

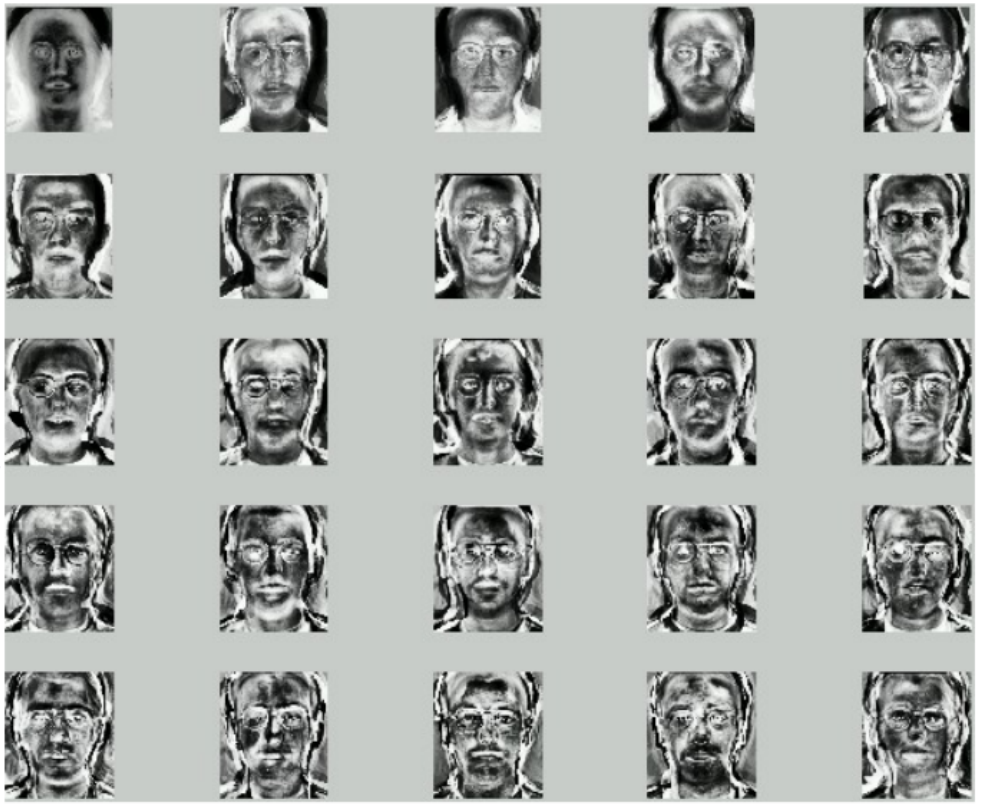


Credit: Mark Craven

Dimensionality reduction example

We can represent a face using all of the pixels in a given image (# features = # pixels)

More effective method: represent each face as a linear combination of *eigenfaces* (# features = 25)



Selection of relevant features and examples

- Relevant features vs. relevant examples: machine learning 방법의 성능과 이용되는 training example 및 feature의 의존성. 각 방법이 이용되는 자료에 대해 다른 의존성을 가짐.
- Sample complexity: the number of training examples needed to reach a desired level of accuracy.
  - over-fitting vs. under-fitting problems.
- Feature-weighting methods vs. explicit feature-selection methods.

Scientists' view on features

- Feature selection and sample selection은 learning data를 가장 잘 아는 과학자들이 가장 중요한 역할을 할 수 있는 부분.
- 크게 두 가지 다른 방향에서 기술: x-space (i.e., raw space) vs. k-space (i.e., basis space).
- Decomposition: PCA, ICA, Wavelet, Fourier analysis 등.
- [주의] 대부분의 ML은 입력 자료의 부정확성 (즉, 자료나 feature의error/uncertainty)를 고려하는 법을 포함하지 않음. 천문학 자료에서 중요한 문제. 그렇다면 과학자가 이런 문제를 feature에 담아 낼 수 있을까? 통계학에서 어떤 도움을 받을 수 있을까?

THE ASTROPHYSICAL JOURNAL, 760:15 (10pp), 2012 November 20 doi:10.1088/0004-637X

STAR–GALAXY CLASSIFICATION IN MULTI-BAND OPTICAL IMAGING

ROSS FADELY[1], DAVID W. HOGG[2,3], AND BETH WILLMAN[1]
[1] Haverford College, Department of Physics and Astronomy, 370 Lancaster Ave., Haverford, PA 19041, USA
[2] Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003, USA
[3] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
Received 2012 June 19; accepted 2012 September 26; published 2012 October 30

Throughout this paper, we restrict our analysis to sources likely to be unresolved in ground-based data (FWHM$_{HST/ACS}$ < 0.2 arcsec). We do so since commonly used morphological classification criteria will easily distinguish quite extended sources, accounting for a majority of galaxies to depths of $r \sim 24$–25. However, galaxies with angular sizes <0.2 arcsec are unlikely to be resolved in surveys with seeing $\gtrsim$0.7 arcsec and so are an appropriate test bed for the type of sources that will rely the most on photometric star–galaxy separation. In total, our sample consists of 7139 stars and 9167 galaxies with apparent magnitudes $22.5 < r < 25$ and is plotted in *ugriz* color–color space in Figure 2. Over this magnitude range, the median S/N in the *r* band ranges from ~50 at $r = 22.5$ to ~15 at $r = 25$, with lower corresponding ranges of 10–7 in the *u*. Of all 18,606 sources with FWHM< 0.2 arcsec in the COSMOS catalog, we identified 2300 AGNs, which we discard from our current analysis.

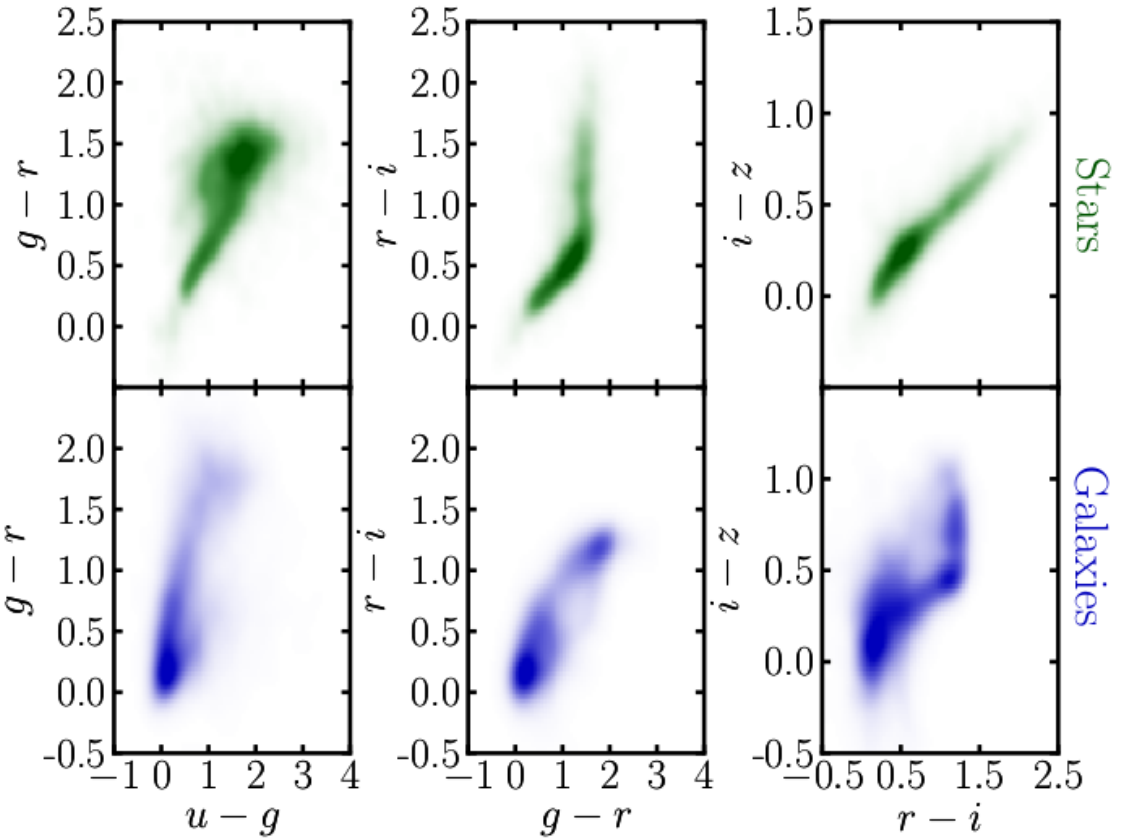
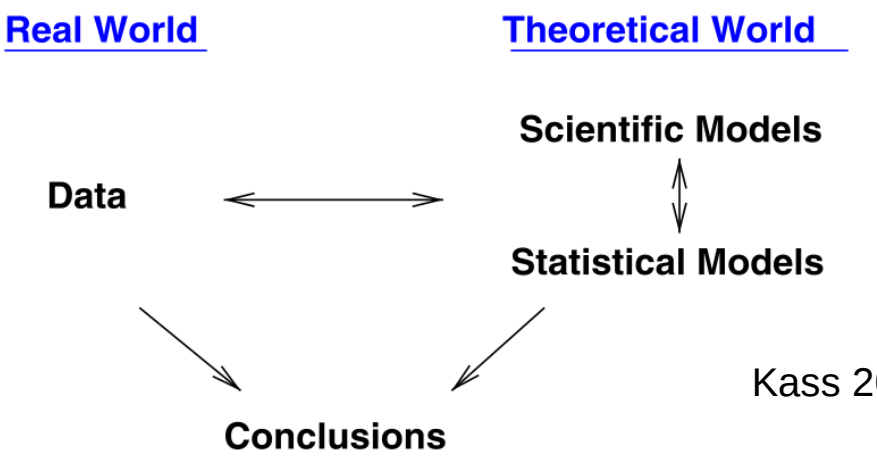


Figure 2. Color–color space distribution of point sources (FWHM < 0.2 arcsec) in the COSMOS catalog. It is clear that stars in the sample follow a tight locus in all slices of color–color space, while galaxies are more generally distributed. Even so, comparison by eye reveals significant overlap between stars and galaxies, particularly for bluer sources.

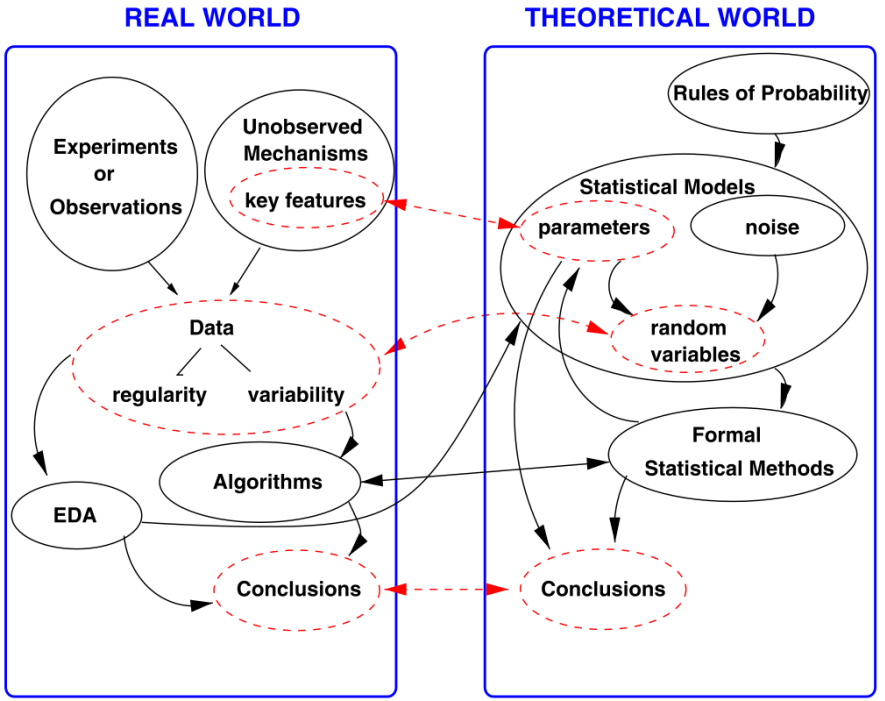
통계적 추론 (statistical inference)?

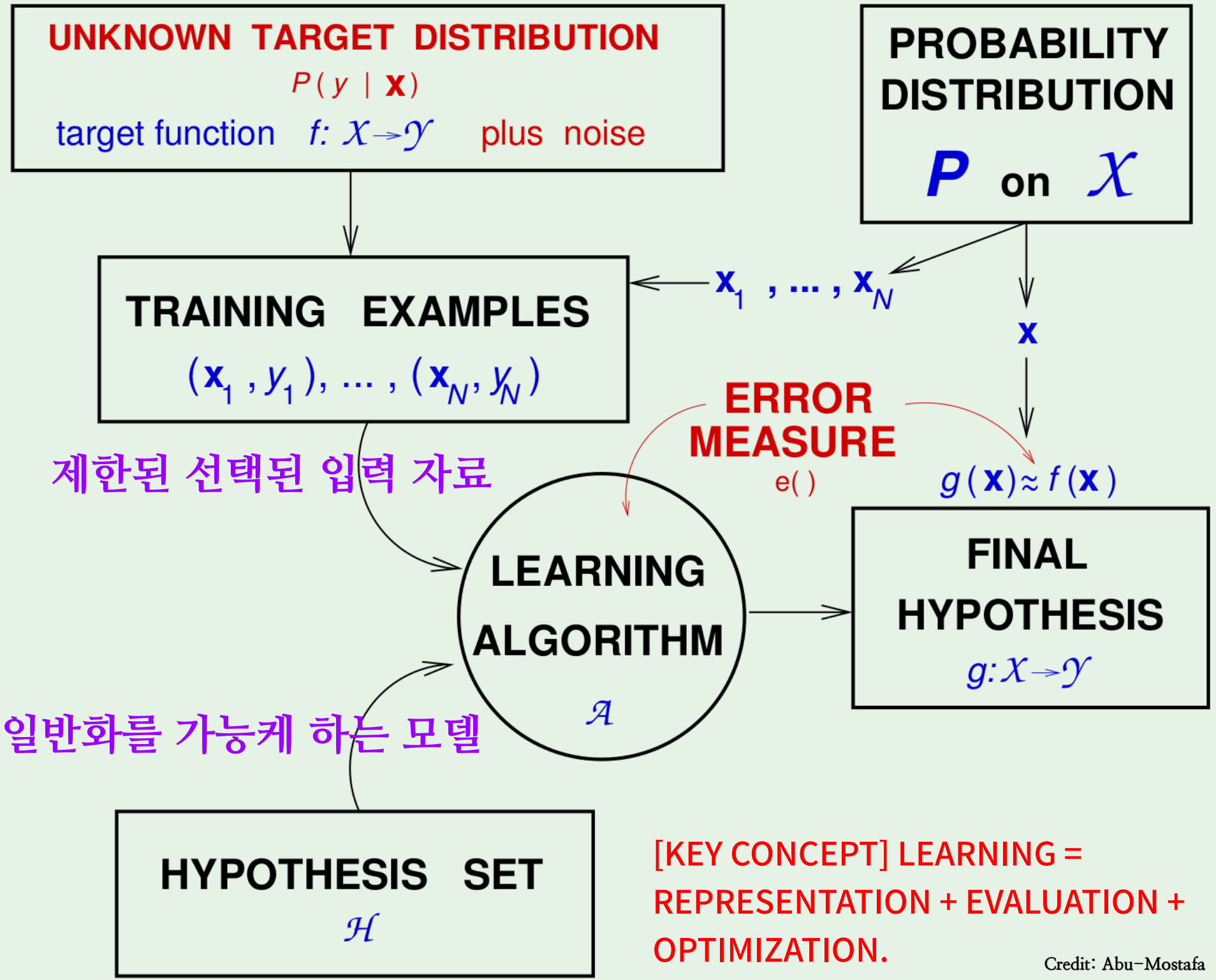
- 사전적 의미: 자료로부터 모집단에 대한 결론이나 과학적 사실을 이끌어 내는 것.

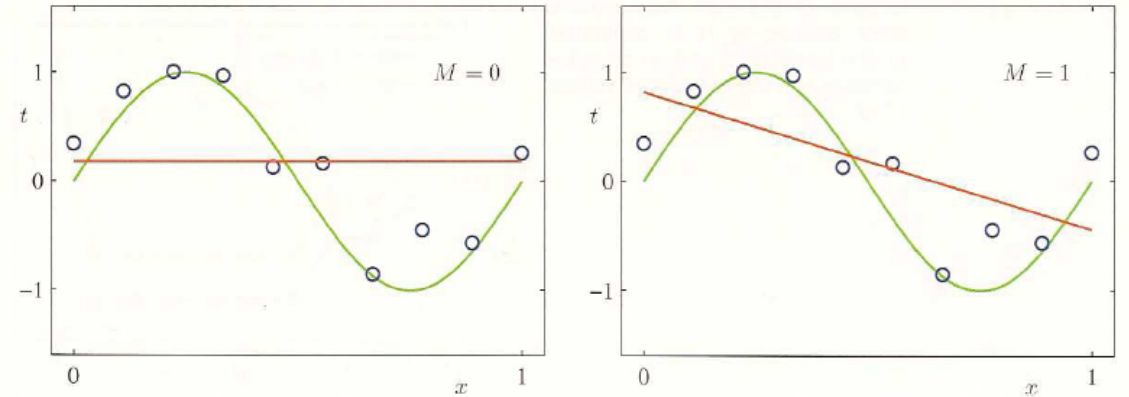
중요: Deductive logic vs. plausible reasoning.

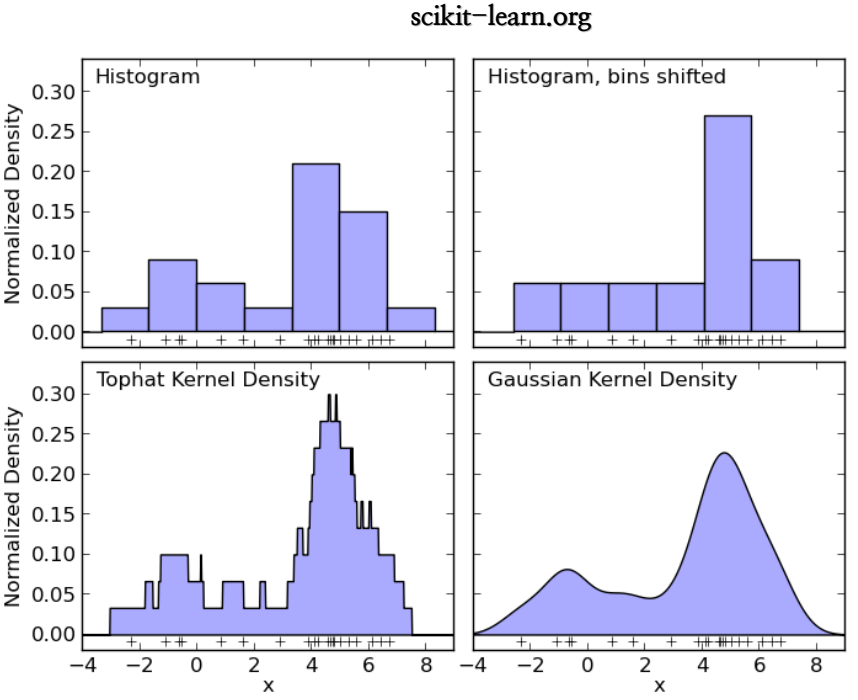


Kass 2011



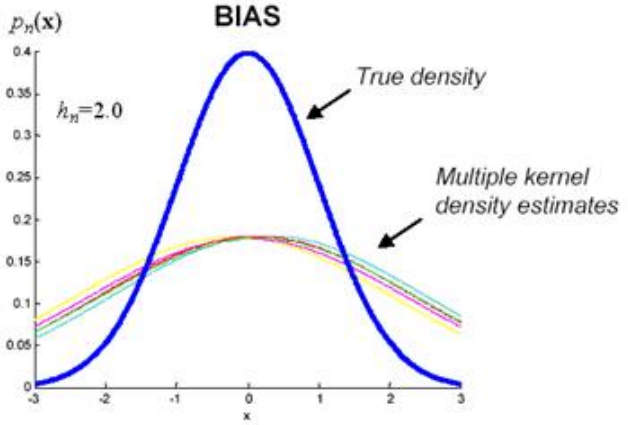

UNKNOWN TARGET DISTRIBUTION
P(y | x)
target function f: X→Y plus noise
PROBABILITY DISTRIBUTION
P on X
TRAINING EXAMPLES
(x1 , y1), ... , (xN, yN)
x1 , ... , xN
x
ERROR MEASURE
e( )
g(x)≈ f (x)
제한된 선택된 입력 자료
LEARNING ALGORITHM
A
FINAL HYPOTHESIS
g: X→Y
일반화를 가능케 하는 모델
HYPOTHESIS SET
H
[KEY CONCEPT] LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION.
Credit: Abu-Mostafa

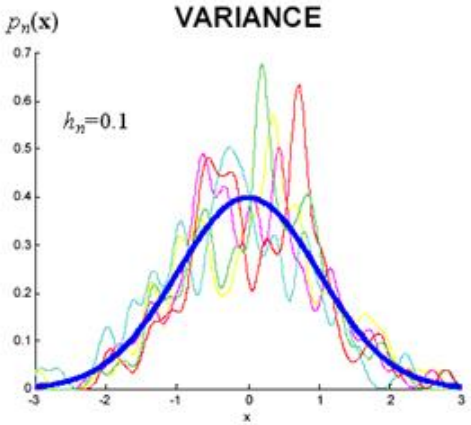




https://www.byclb.com/TR/Tutorials/neural_networks/ch11_1.htm

Bias vs. Variance





Accuracy vs. Precision

논문이나 보고서에 제시된 자료와 통계/과학 모델의 조합이 제시된 통계적 추론을 얼마나 뒷받침하는가? 고려된 통계 모델과 방법들이 과학 모델을 충분히 담아내고 있는가?

자연 관측 (즉, 천문학) 자료의 특성과 분석의 근본 특징들

- 대부분의 기계학습은 **입력 자료의 부정확성 (즉, 자료나 feature의 noise/irreducible error/uncertainty)**를 고려하는 법을 포함하지 않음. 천문학 자료에서 중요한 문제. 모든 측정은 부정확성을 고려해야 한다!

EM Algorithms for Weighted-Data Clusteringwith Application to Audio-Visual Scene Analysis (2016)

- Israel D. Gebru, Xavier Alameda-Pineda, Florence Forbes and Radu Horaud

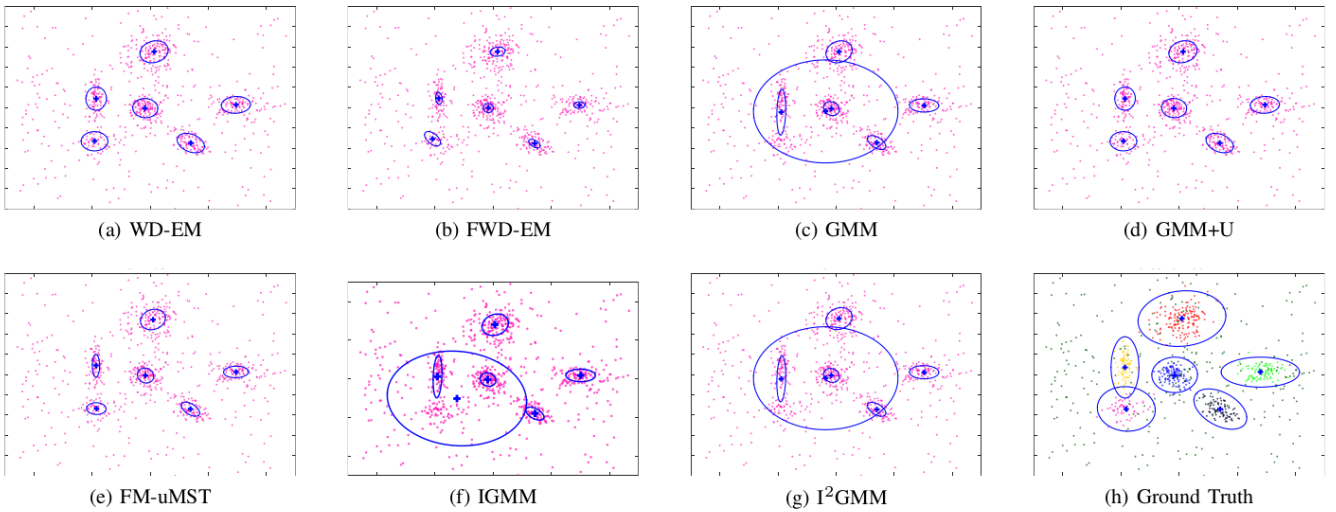
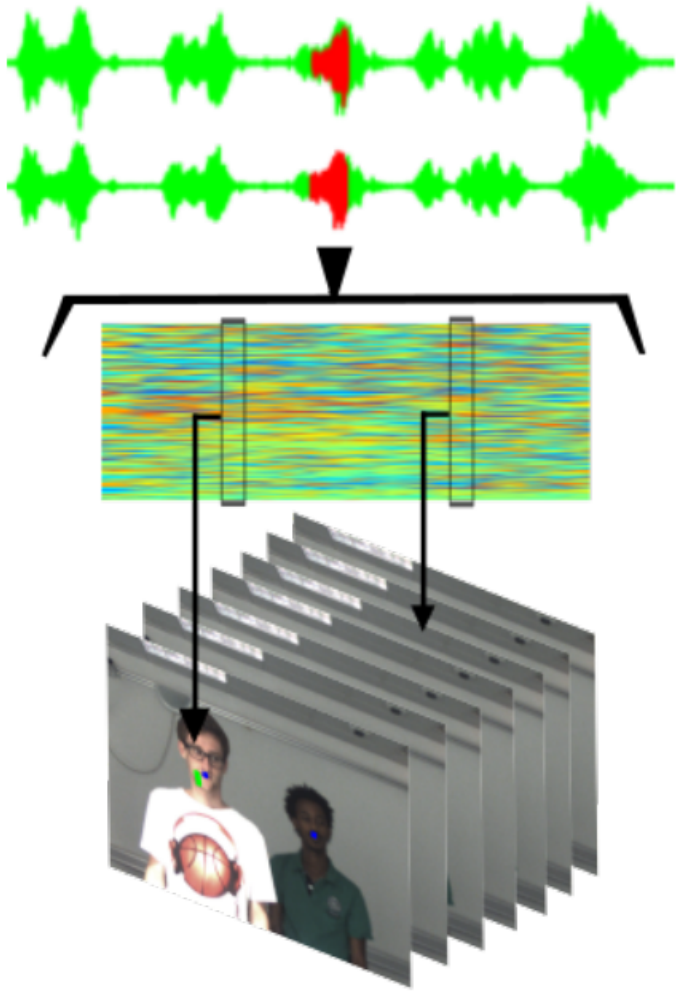


Fig. 2: Results obtained by fitting mixture models to the SIM-Mixed data in the presence of 50% outliers (see Table IV).



g. 3: Audio-visual data acquisition and alignmer
d right-microphone signals. A temporal segmen
tlined in red. *Middle*: Binaural spectrogram that
the outlined segment. This spectrogram is c
binaural vectors, each one being associated wi
me (shown as a vertical rectangle). *Bottom*: v
sociated with a segment. A sound-source directi
OA) is extracted from each binaural vector and r
image plane, hence each green dot in the i
rresponds to a DOA.

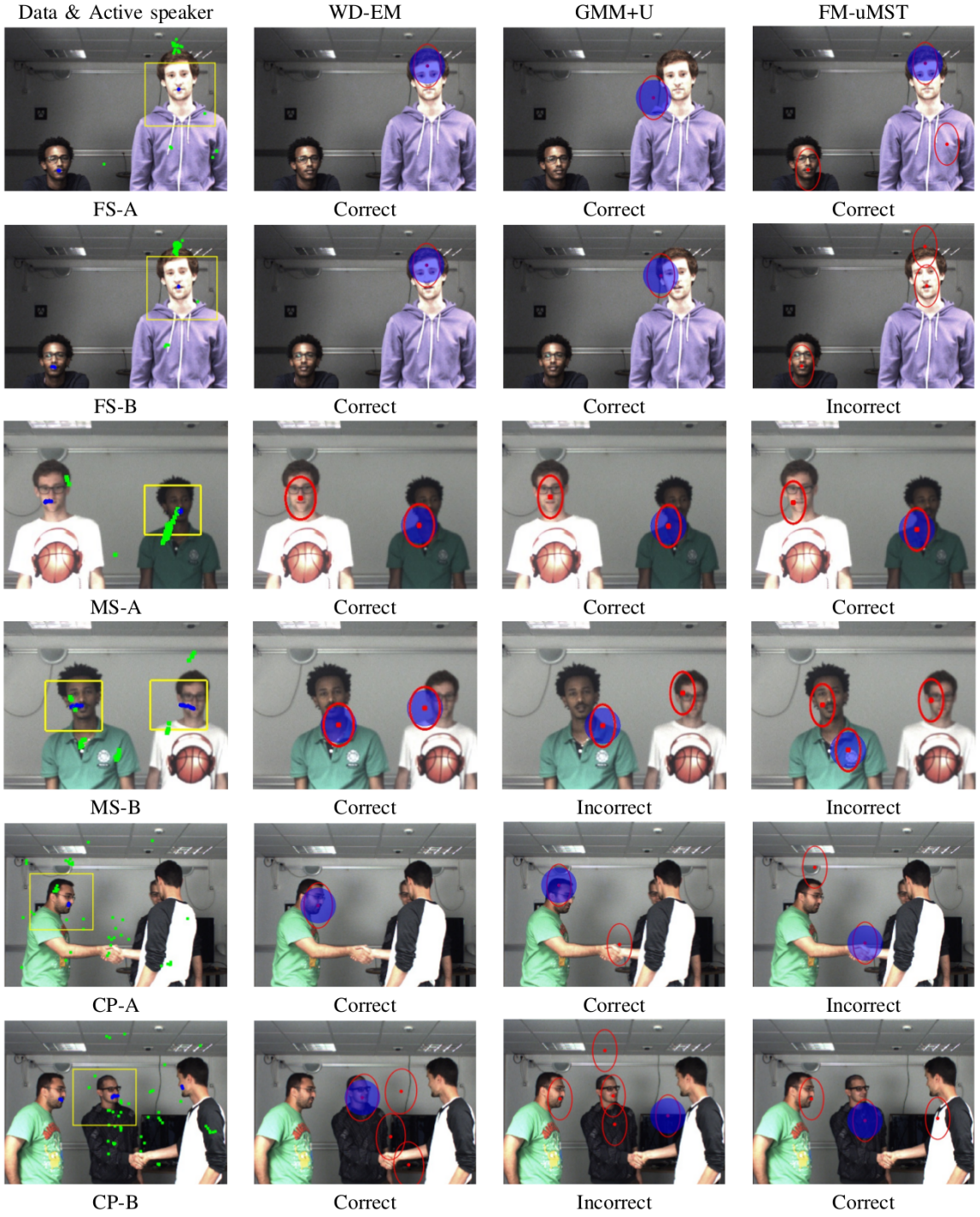


Fig. 4: Results obtained on the *fake speaker* (FS), *moving speaker* (MS) and *cocktail party* (CP) sequences. The first column shows the audio (green) and visual (blue) observations, as well as a yellow bounding box that shows the ground-truth active speaker. The second, third and fourth columns show the mixture components obtained with the WD-EM, GMM+U and FM-uMST methods, respectively. The blue disks mark components that correspond to correct detections of active speakers, namely whenever there is an overlap between a component and the ground-truth bounding box.

자연 관측 (즉, 천문학) 자료의 특성과 분석의 근본 특징들

- Supervised learning에 쓰이는 **training 자료의 질과 양 및 특성**의 문제. 다량의 일관성 있는 training 자료를 확보하기 어려움. 따라서, 다양한 기기로부터의 자료를 결합. Semi-supervised learning 중요!

3.1. Convolutional Neural Network (ConvNet) Configuration

In this work, we mimic human perception with *deep learning* using convolutional neural networks (ConvNets). Although it is clearly beyond the scope of the present paper to provide a complete description of how convolutional neural networks work, we provide a brief introduction below. We refer the interested reader to D15 for more details.

Deep learning is a methodology to automatically learn and extract the most relevant features (or parameters) from raw data for a given classification problem through a set of nonlinear transformations.

Though deep learning architectures have existed since the early 80s (Fukushima 1980), they involve complex technological problems which only allowed their use in massive data sets in the last decade. Several factors have contributed to the rise in their popularity: (i) the availability of much larger training sets with millions of labeled examples[12]; (ii) powerful GPU implementations, making the training of very large models practical; and (iii) improved model regularization algorithms, which helped to reduce computing time.

[12] ConvNets are particularly sensitive to this since the risk of over-fitting is large given the complexity of the models.

A CATALOG OF VISUAL-LIKE MORPHOLOGIES IN THE 5 CANDELS FIELDS USING DEEP LEARNING (2015)
- M. Huertas-Company et al.

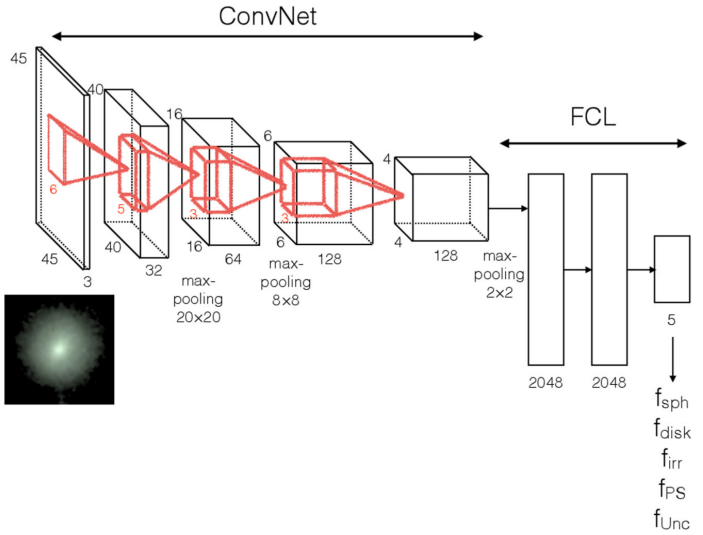


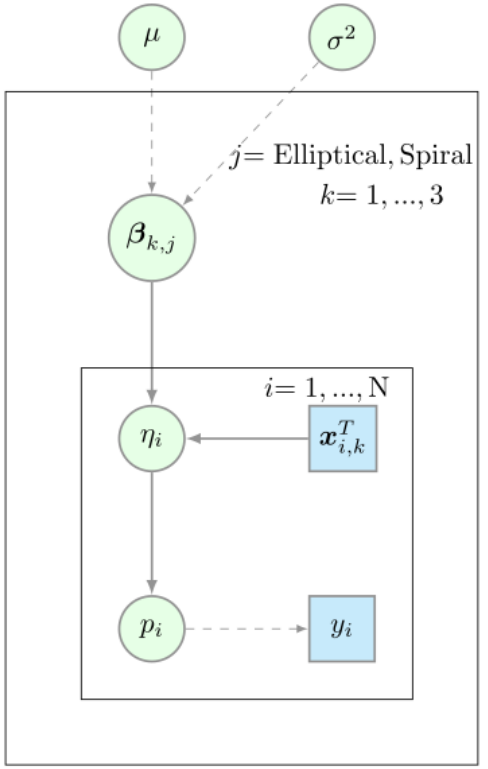
Figure 2. Configuration of the Convolutional Neural Network used in this paper. The Network is based on the one used by Dieleman et al. (2015) on SDSS galaxies. It is made of 5 convolutional layers followed by 2 fully connected perceptron layers. In the convolutional part there are also 3 max-pooling steps of different sizes. The input are *SDSSized* CANDELS galaxies as explained in the text and the output (for this paper) is made of 5 real values corresponding to the fractions defined in the CANDELS classification scheme.

자연 관측 (즉, 천문학) 자료의 특성과 분석의 근본 특징들

- **입력 자료의 불완전성** 문제. 일부 자료의 경우 특정 feature가 존재하지 않거나 추출이 불가능 한 경우.
  - 예를 들어, 지상에서 가시광선으로 측정 가능. 그러나 우주에서 적외선, 자외선 등의 관측이 필요.
  - 여러 필터를 통한 관측 자료는 존재하나, 스펙트럼 관측 자료는 일부에 대해서 존재.
  - **시간에 따른 관측이 균일하게 이루어지지 않음 (irregular time sampling).**
  - **관측의 공간 분포가 균일하지 않은 (gaps in the observation domain).**

자연 관측 (즉, 천문학) 자료의 특성과 분석의 근본 특징들

- 다른 분야와 달리 시간에 따른 변화 예측이 필요한 경우가 흔하지 않음.
- 대신에 **분류에 의존하는 statistical inference** (특히 regression 문제)가 흔함. P(z | discrete type) vs. P(z) (categorically distributed dependent variable) .



$$y_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = \eta_i$$
$$\eta_i = x^T_{i,k}\beta_{k,j}$$
$$x^T_{i,k} = \begin{pmatrix} 1 & (\log M_{200})_1 & \left(\frac{r}{r_{200}}\right)_1 \\ \vdots & \vdots & \vdots \\ 1 & (\log M_{200})_N & \left(\frac{r}{r_{200}}\right)_N \end{pmatrix}$$
$$\beta_{k,j} \sim \text{Normal}(\mu, \sigma^2)$$
$$\mu \sim \text{Normal}(0, 10^3)$$
$$\tau \sim \text{Gamma}(10^{-3}, 10^{-3})$$
$$\sigma^2 = 1/\tau$$
$$j = \text{Elliptical, Spiral}$$
$$k = 1, ..., 3$$
$$i = 1, ..., N$$

Figure 4. A graphical model representing the hierarchy of dependences for a data set of galaxies indexed by the subscript i. The dashed arrows represent stochastic dependences, while straight arrows the systematic ones. Blue square represents input data, and green circles are model parameters.

Is the cluster environment quenching the Seyfert activity in elliptical and spiral galaxies? (2016)
- R. S. de Souza et al.

A Few Useful Things to Know about Machine Learning - Pedro Domingos

1. LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION.

2. IT'S GENERALIZATION THAT COUNTS.

3. DATA ALONE IS NOT ENOUGH.

4. OVERFITTING HAS MANY FACES.

5. INTUITION FAILS IN HIGH DIMENSIONS.

6. THEORETICAL GUARANTEES ARE NOT WHAT THEY SEEM.

7. FEATURE ENGINEERING IS THE KEY.

8. MORE DATA BEATS A CLEVERER ALGORITHM.

9. LEARN MANY MODELS, NOT JUST ONE.

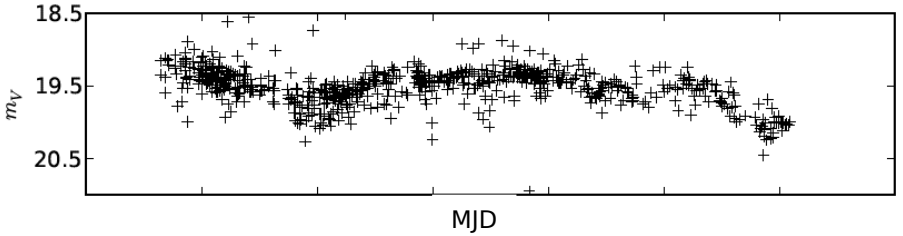
10. SIMPLICITY DOES NOT IMPLY ACCURACY.

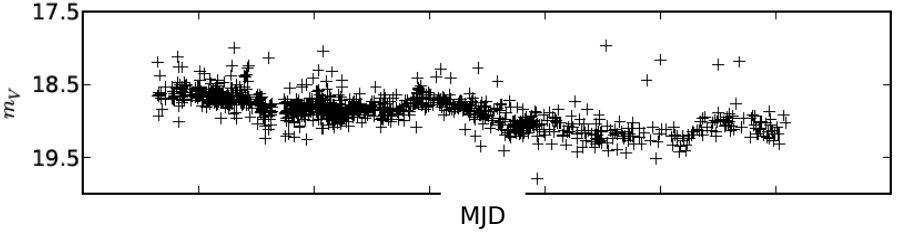
11. REPRESENTABLE DOES NOT IMPLY LEARNABLE.

12. CORRELATION DOES NOT IMPLY CAUSATION.

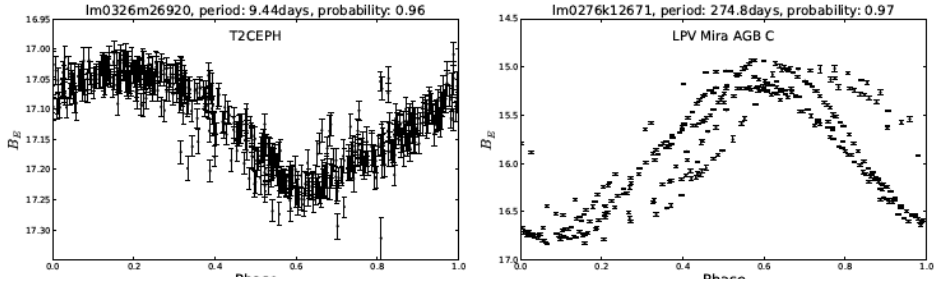
Temporal data and spatio-temporal data

Quasar candidates in MACHO light curves (Kim+ 12).

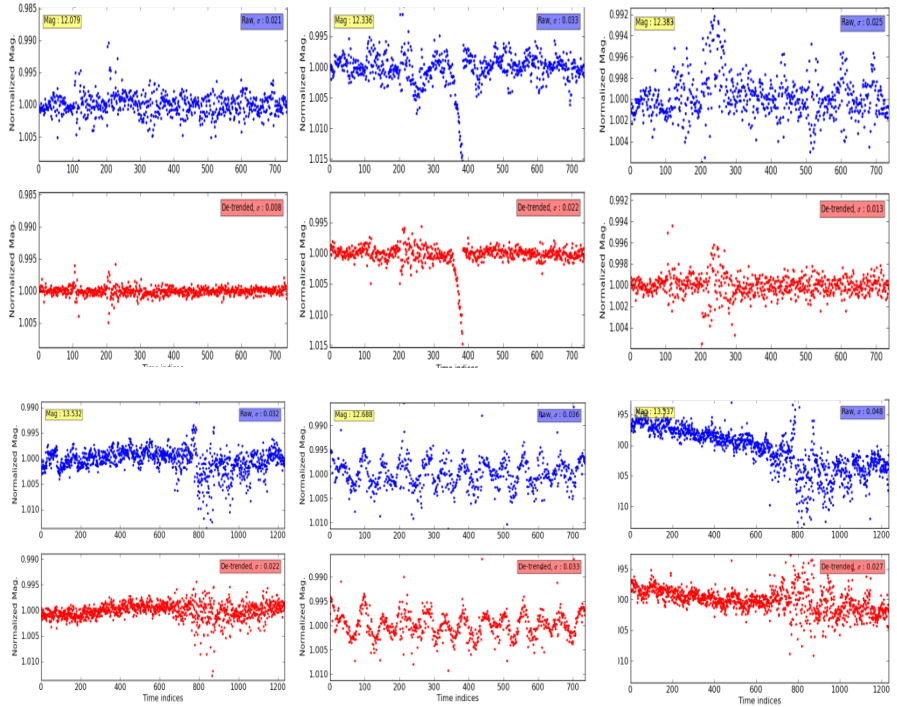




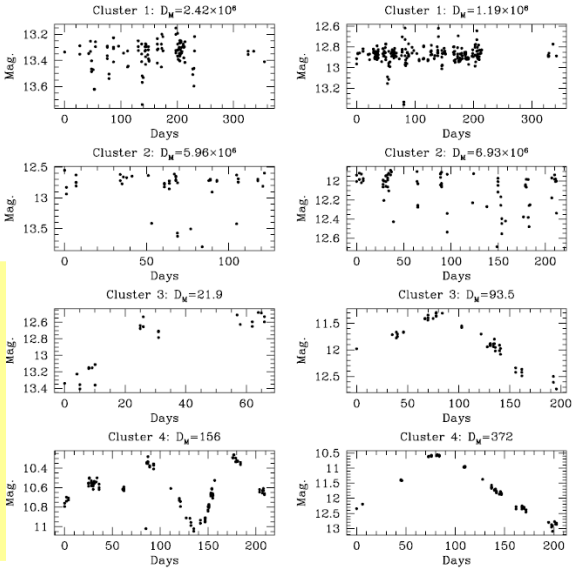
Classification of EROS variable light curves (Kim+ 14).



- Supporting Vector Machine.
- Random Forest.
- Finite vs. Infinite Gaussian Mixture Model.
- Variational approximate Bayesian inference vs. full MCMC inference.
- Ensemble learning.



Identification of trends in light curves (Chang+ 15).

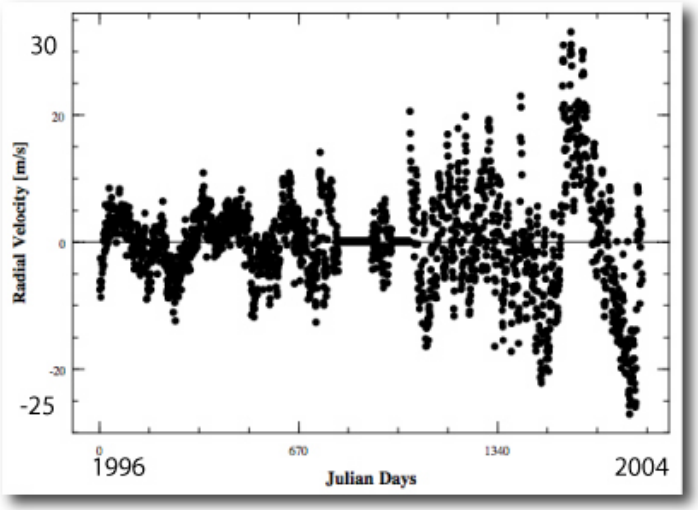


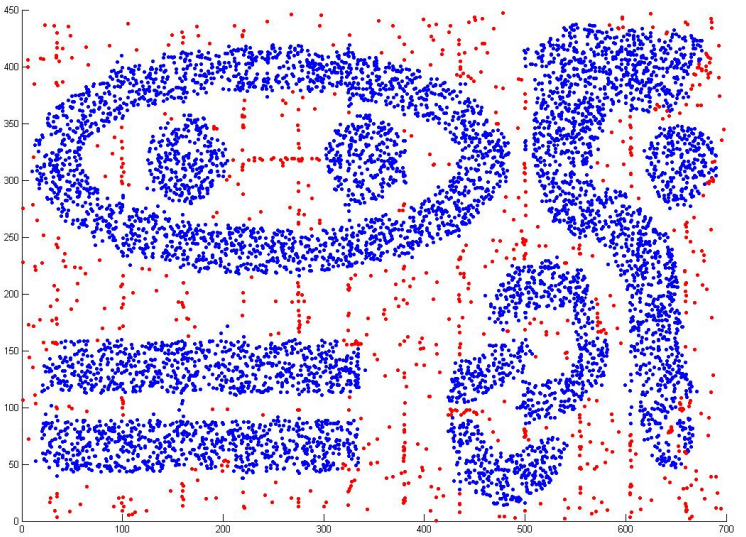
Detection of variable candidates (Shin+ 09, 12, 16).

예: clustering 방법을 이용한 시계열 자료 변화 검출

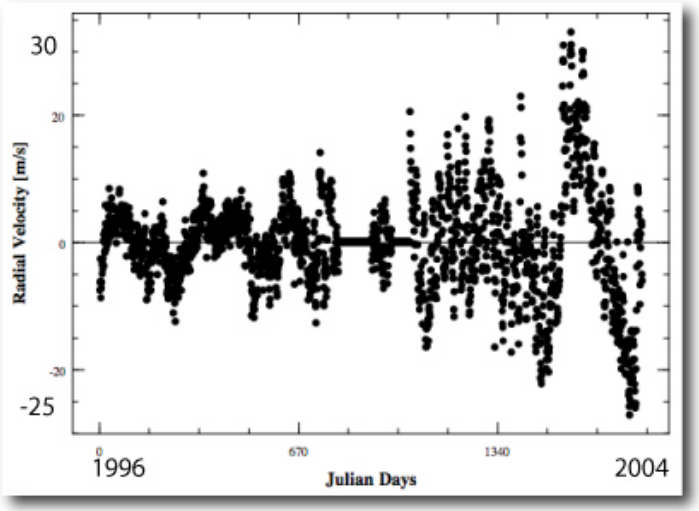
- Outliers are data points that are distant from major groups in clustering results.

- Common systematic patterns and non-variable sources are expected to appear as a major group of input data. *Variable sources are outliers.*





Feature extraction/selection (i.e. dimensional reduction).



Multiple light curves (i.e. measurements as a function of time)

- Time-series have signal, systematic / random noise, and bias together as a function of time and space. Does your time-series follow the Normal distribution? No sampling effects?

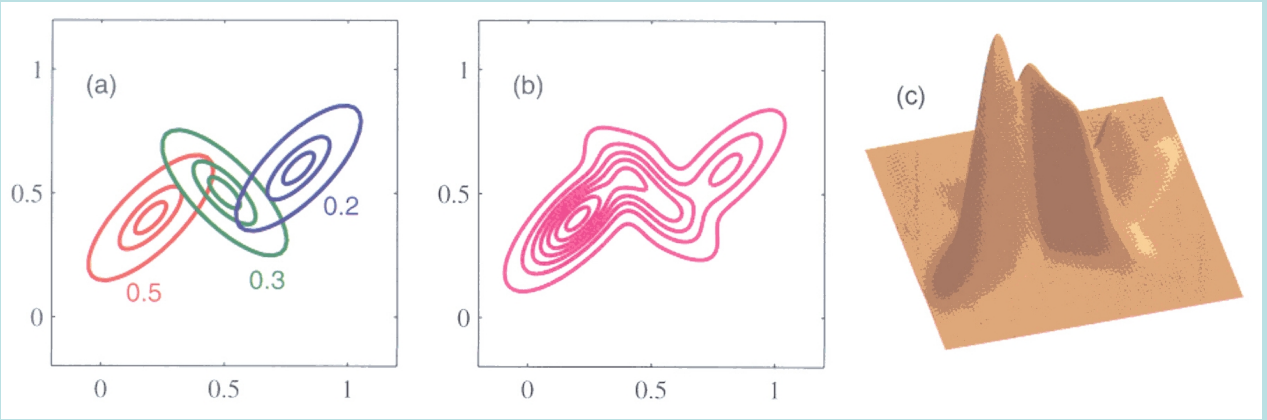
- Two ways to discover variables: defining properties of genuine variable objects vs. defining properties of non-variable objects.

Summarizing the data to a lower dimension (i.e. feature space). Some features are not updated quickly with newly added data.

Feature extraction

σ/ν	Weighted sample standard deviation / median
γ_1	$\frac{\sqrt{N(N-1)}}{N-2}\frac{\sum_{n=1}^{N}(x_n-\mu)^3/N}{\sqrt[3]{\sum_{n=1}^{N}(x_n-\mu)^2/N}}$
γ_2	$\frac{N-1}{(N-2)(N-3)}\left\{(N+1)\left(\frac{\sum_{n=1}^{N}(x_n-\mu)^4/N}{(\sum_{n=1}^{N}(x_n-\mu)^2/N)^2}-3\right)+6\right\}$
Con	$1+\frac{1}{N-2}\sum_{n=1}^{N-2}\begin{cases}1 & \text{if } (x_n-\nu)>2\sigma \text{ and } (x_{n+1}-\nu)>2\sigma \text{ and } (x_{n+2}-\nu)>2\sigma\\ 1 & \text{if } (x_n-\nu)<2\sigma \text{ and } (x_{n+1}-\nu)<2\sigma \text{ and } (x_{n+2}-\nu)<2\sigma\\ 0 & \text{otherwise}\end{cases}$
η	$\frac{\sum_{n=1}^{N-1}(x_{n+1}-x_n)^2/(N-1)}{\sigma^2}$
J	$\sum_{n=1}^{N-1} sign(\delta_n\delta_{n+1})\sqrt{\lvert\delta_n\delta_{n+1}\rvert}$
K	$\frac{1/N\sum_{n=1}^{N}\lvert\delta_n\rvert}{\sqrt{1/N\sum_{n=1}^{N}\delta_n^2}}$
$AoVM$	The maximum value of the analysis of variance (ANOVA) statistic (Schwarzenberg-Czerny 1989

Gaussian Mixture Models (GMMs) to describe the distribution of light curves in the feature space



$$p(\boldsymbol{x}) = \sum_{m=1}^{M} p_m(\boldsymbol{x}) w_m$$

$$p_m(\boldsymbol{x}) = \frac{1}{(2\pi)^{\gamma/2}|\boldsymbol{\Sigma}_m|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_m)\right]$$

Depending on 1) how you infer parameters (the number of components, centers of components, and their dispersions) and their distributions, and 2) cluster given data by using inference of mixture components, there are several different approaches.

Overview: clustering as an unsupervised machine learning method with density estimation

Infinite Gaussian Mixture Model with the Markov Chain Monte Carlo method (non-parametric Bayesian method)

Computationally expensive and slow convergence with stochastic derivation of statistical inference (i.e. MCMC).
Flexible and precise models.

Infinite Gaussian Mixture Model with the variational approximate method (non-parametric Bayesian method)

Finite Gaussian Mixture Model with the expectation-maximization method (parametric maximum likelihood method)

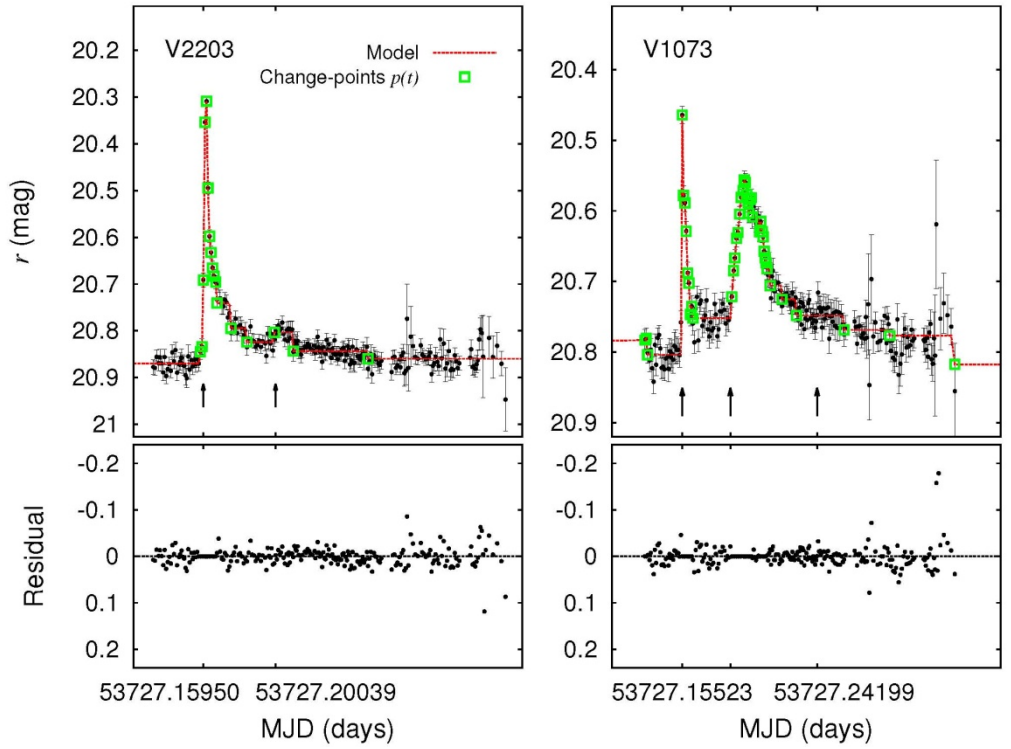
Computationally cheap and fast convergence with a well-known estimation of a convergence rate.
Limitation of models and uncertain statistical inference.

사용되는 용어들에 주목!

예: change-point analysis 방법을 이용한 시계열 자료 변화 검출 feature

It may not always be possible to characterize flux variation purely by detecting and characterizing periodicity.

Change-point analysis is a method for identifying abrupt variations in the sequential data. It is widely used in the statistics and data mining communities as well as in the field of time domain astronomy (see e.g., Schütz & Holschneider 2011 and reference therein).



ü The location of change-

$$p_t = \arg\max_{t\in[0,n]} |S_t|$$

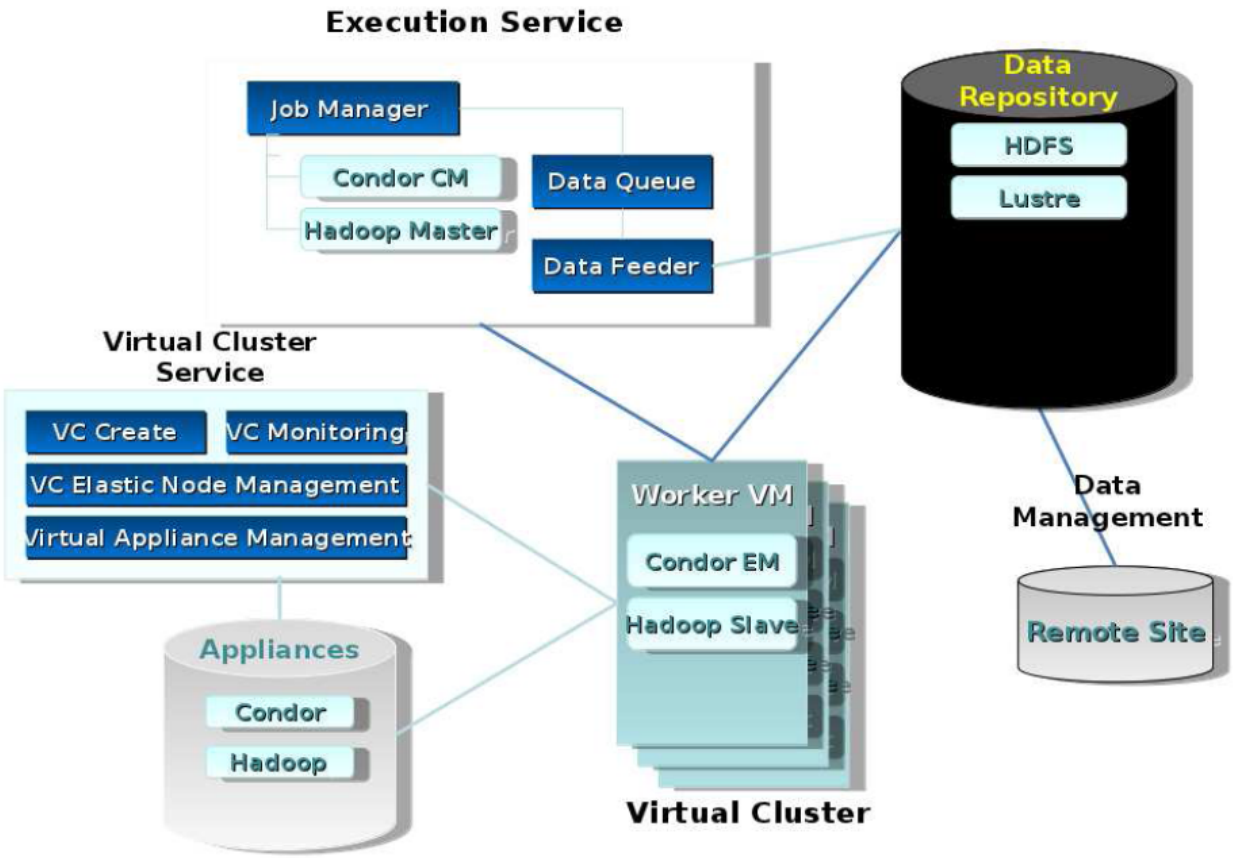
ü The CUSUM values:

$$S_t = \sum_{i=1}^{t} (x_i - \bar{x})$$

Chang+ 2014c

Implementation과 framework는 또 다른 문제...

1. 자료는 일반적으로 database 혹은 file로 존재.
2. 분석 자료의 규모 및 기계 학습/자료 분석 방법에 따른 계산 및 저장 능력의 차이 발생.
3. Computing and storage system에 따른 성능 차이 존재. In most cases, you acquire knowledge from experiments.



Experiments of Hadoop and Cloud computing with the National Institute of Supercomputing and Networking in KISTI.

In Hadoop environment, processes are assigned to nodes that have relevant data in order to exploit data locality.

Astronomical Time Series Data Analysis Leveraging Science Cloud (Hahm+ 12).

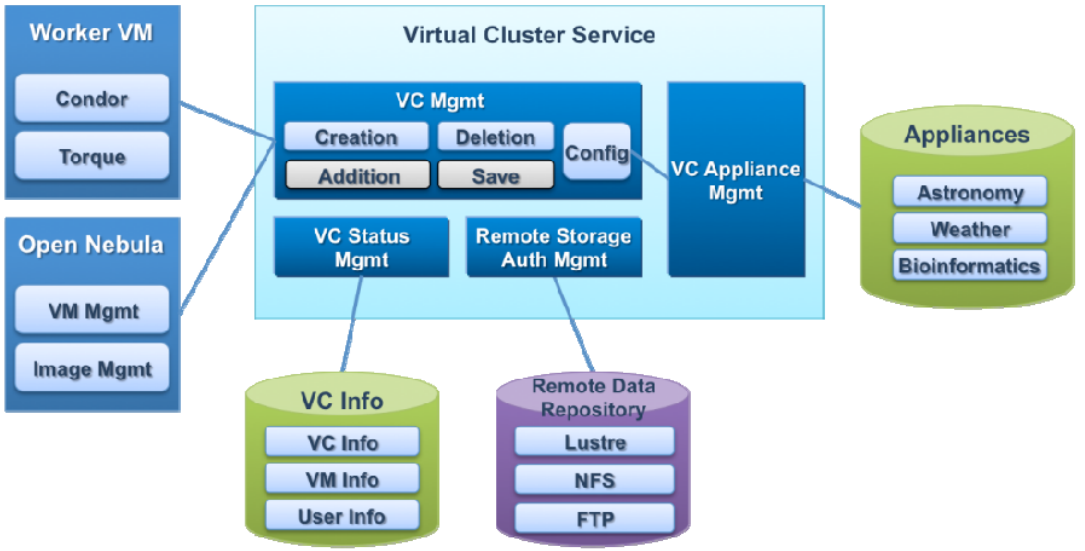


Fig. 3. Virtual Cluster Service Architecture

Experiments of indexing astronomical big data with the LBNL Scientific Data Management Research Group and KISTI.

Efficient Attribute-based Data Access in Astronomy Analysis (Ma+ 12)

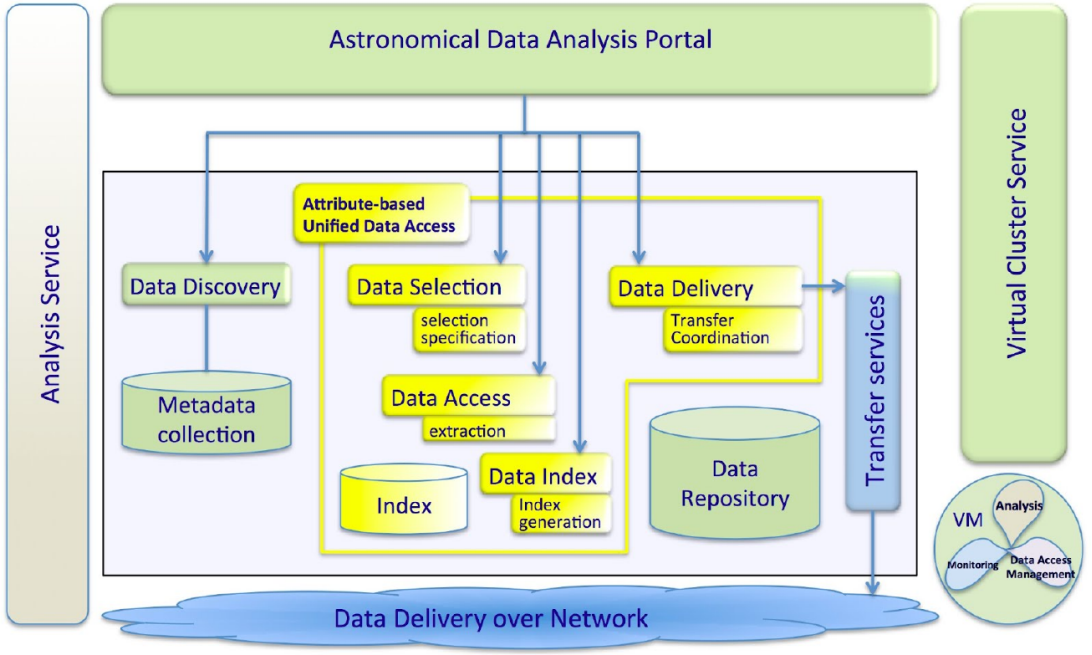
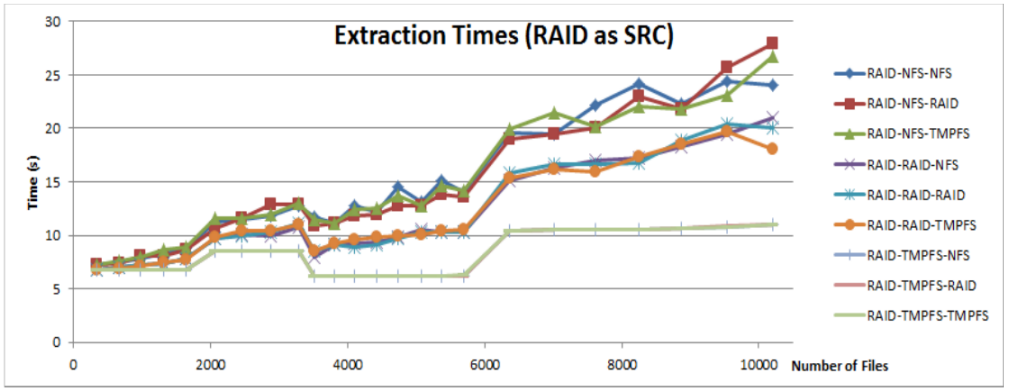
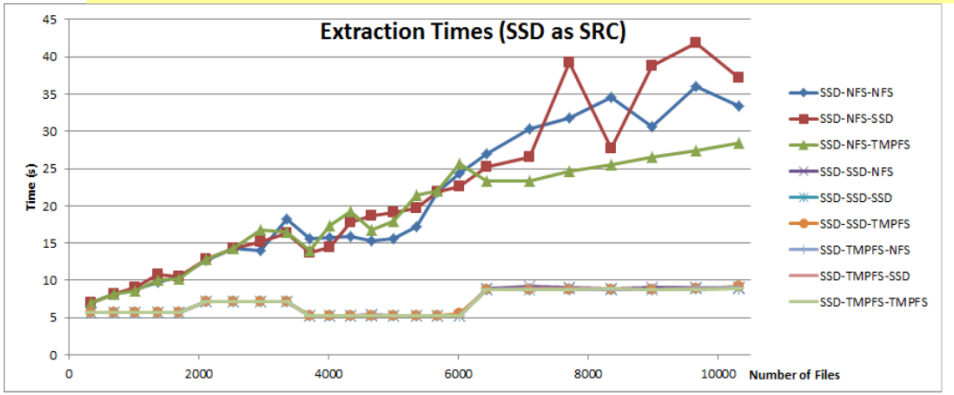


Figure 1: Overall architecture of the Attribute-based Unified Data Access Service





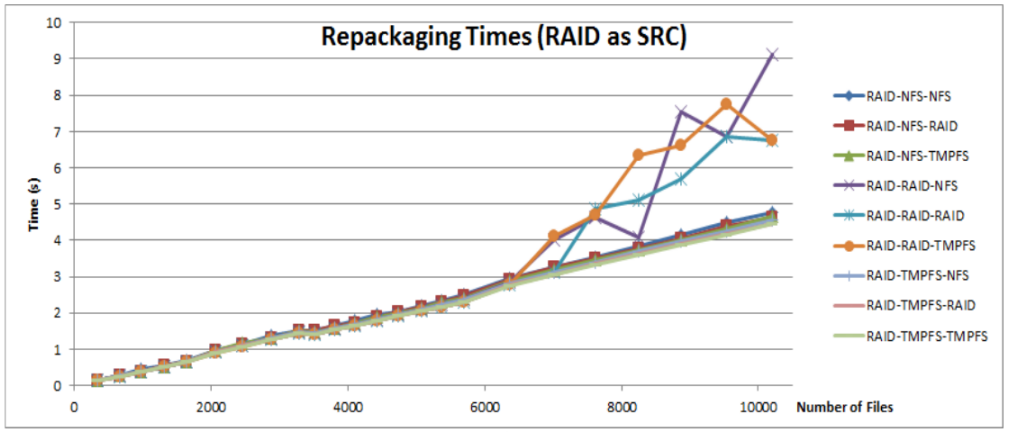


Figure 2: Data access performance (RAID as source)

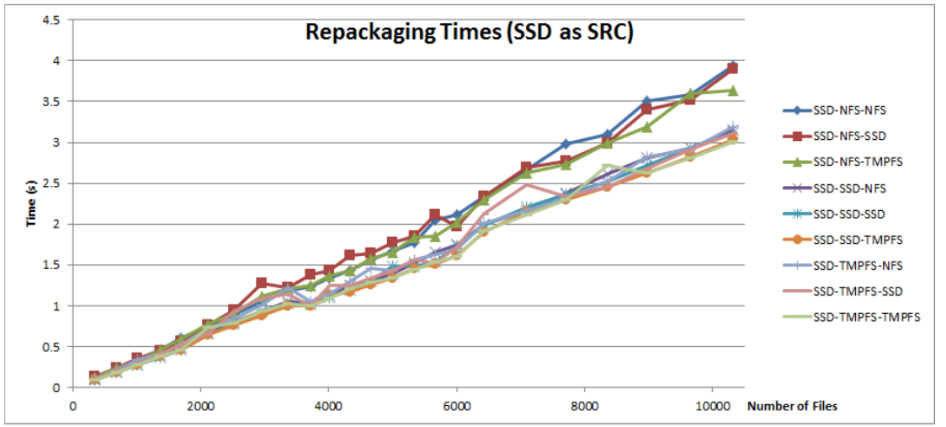
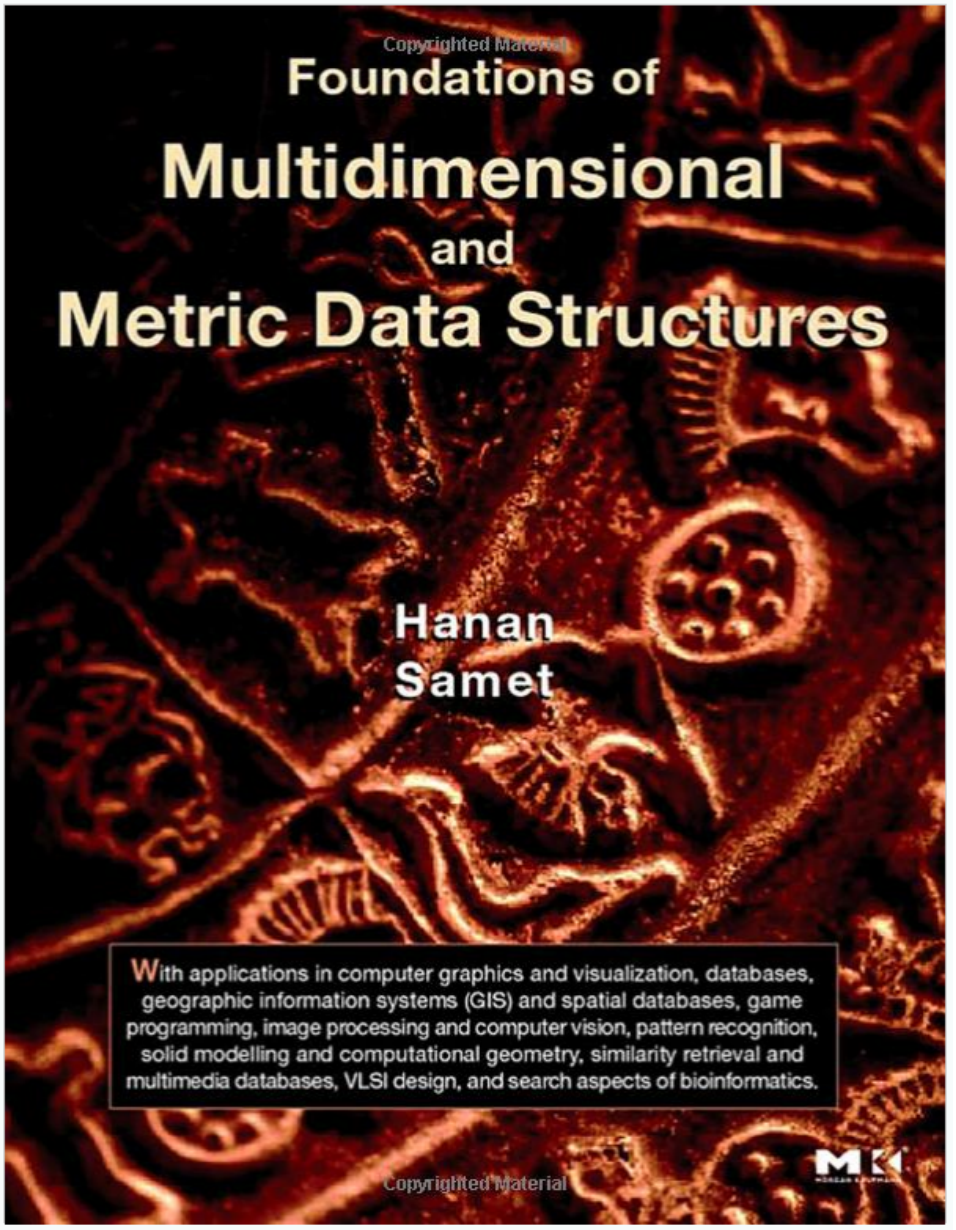
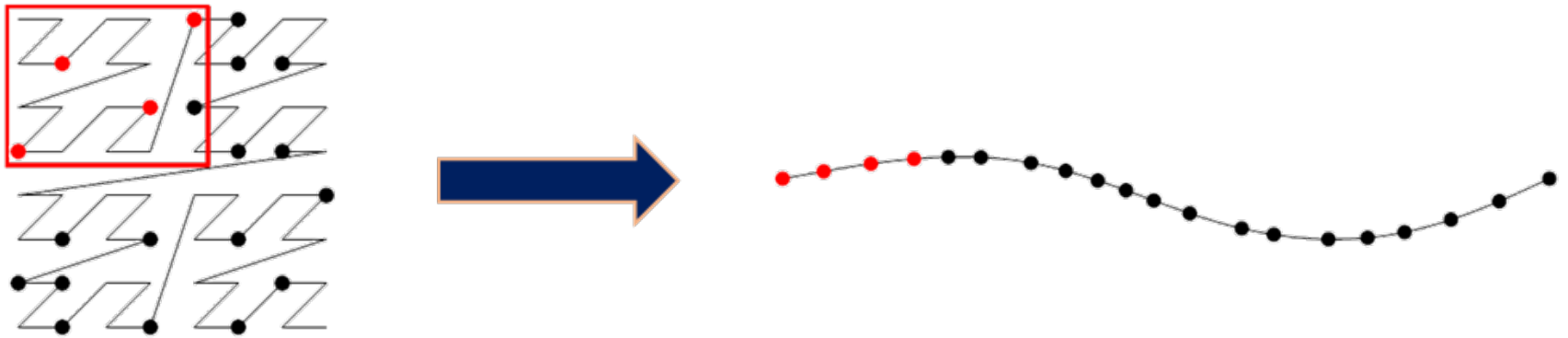


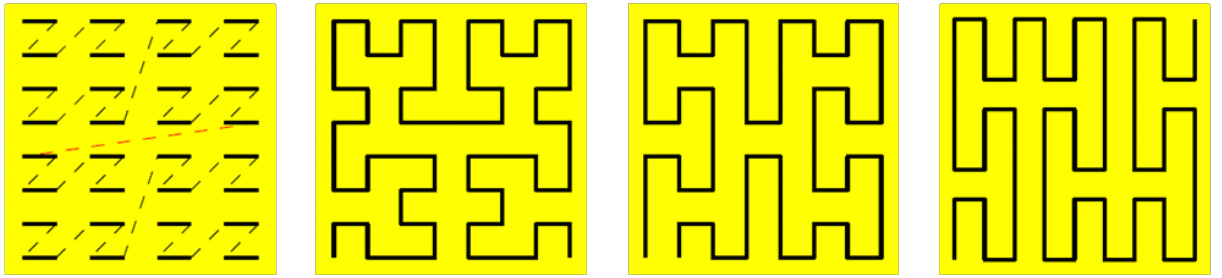
Figure 3: Data access performance (SSD as source)

Big spatio-temporal data

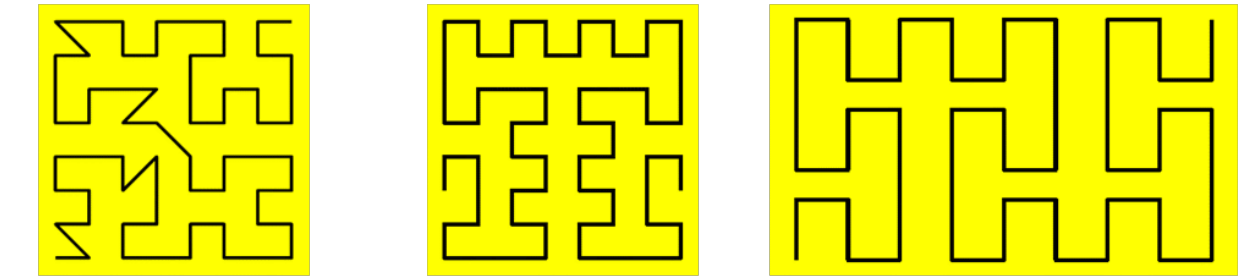
- Key: how to index your data!
- When you deal with multi-dimensional domain data, there are two general options: hash vs. tree.







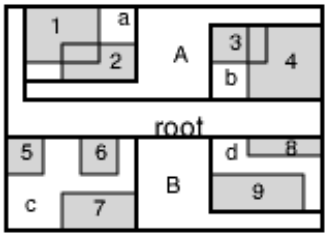
Z-order | Hilbert curve | H-order | Peano's curve

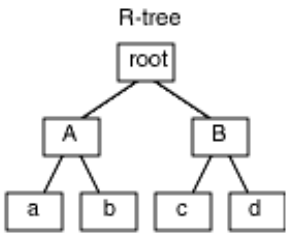


AR^2W^2-curve | $\beta\Omega$-curve | Balanced Peano

https://ngageoint.github.io/geowave/

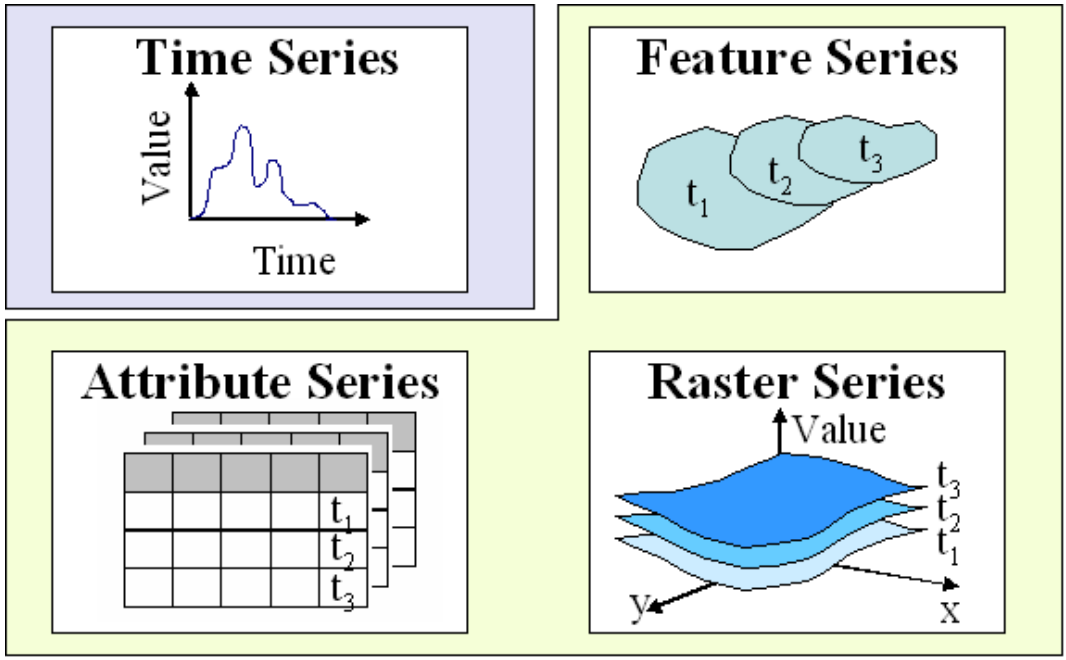
http://doc.oracle.com/



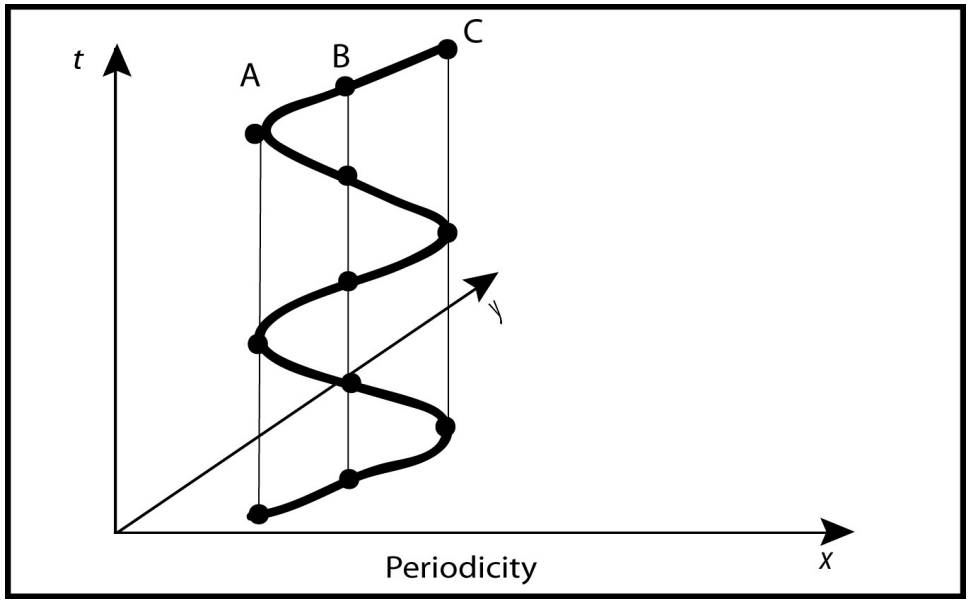


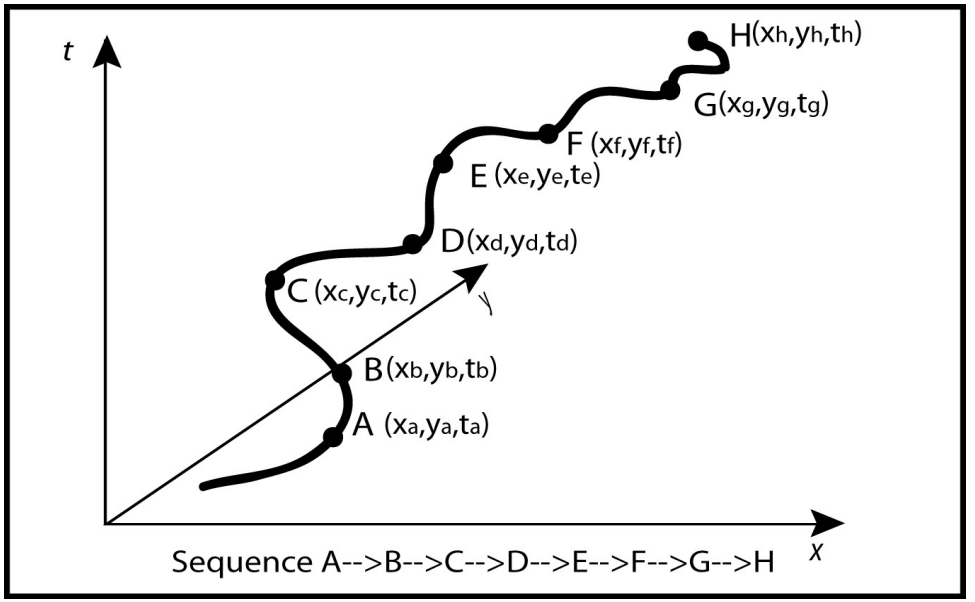
Big spatio-temporal data

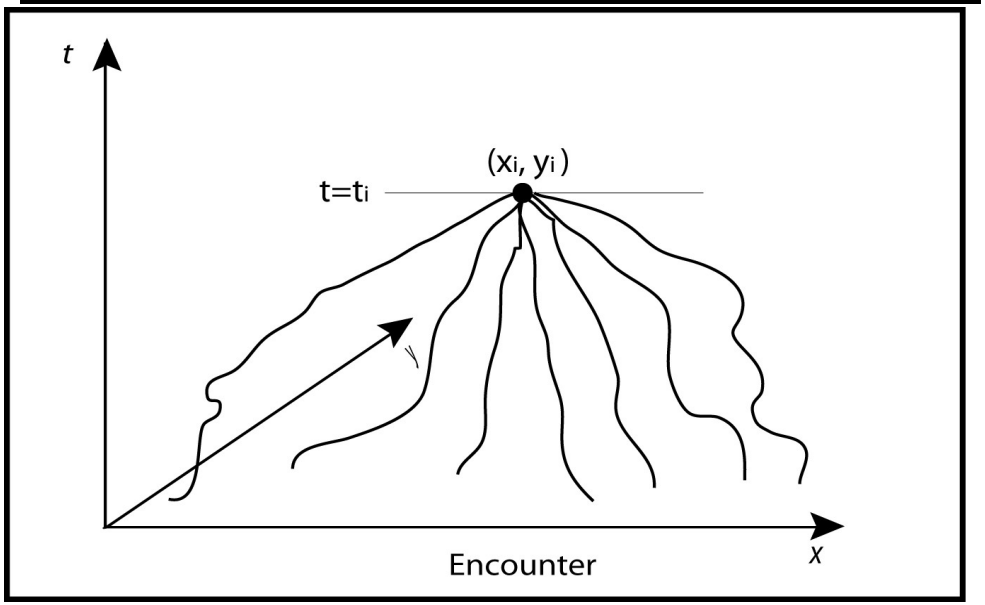
- How can we exploit spatio-temporal (LSST) data efficiently?
  - Fast search and extraction of O(10^{11}) spatio-temporal data.
  - Cheap and scalable system configuration.
  - Easy development and test of science analysis codes.

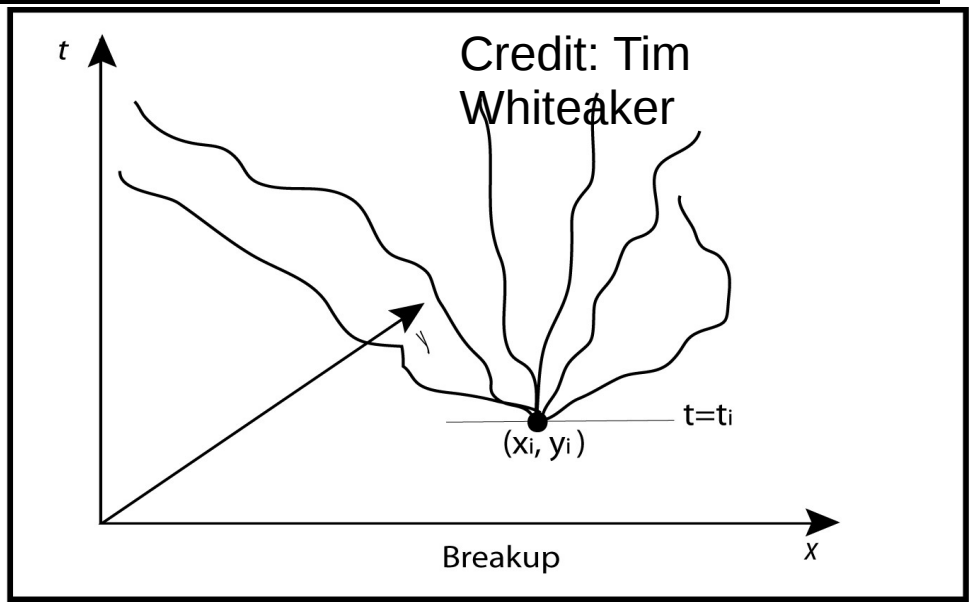


A *GIS(Geographic Information System)* is used to store and process the same kind of data.









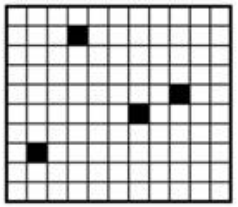
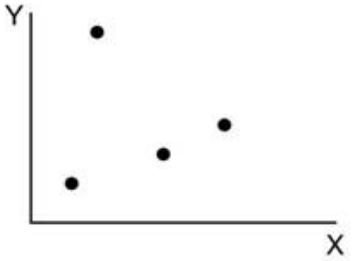
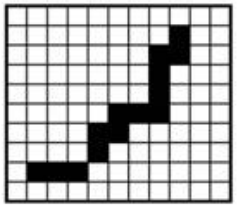
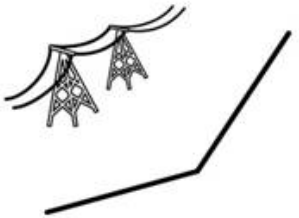
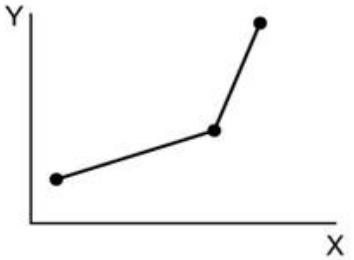
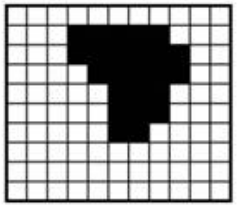
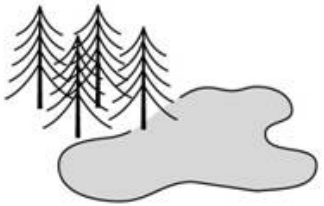
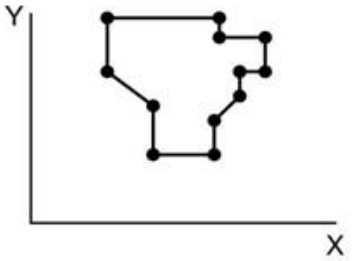
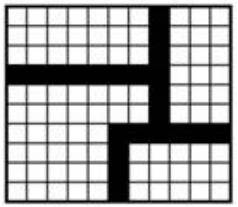
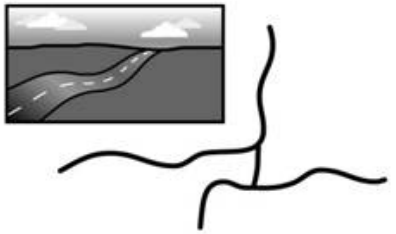
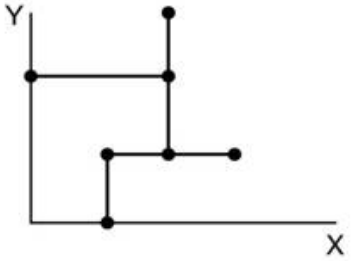
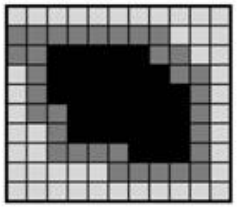
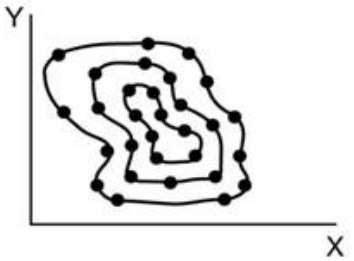
- Speed of searching and analyzing data depends on how you store spatio-temporal data.

- Multi-dimensional data are typically stored in tree structures or hashing (i.e. indexing).

Experiments with big GIS spatio-temporal databases

- Tests of using well-developed open-source GIS systems for big astronomical spatio-temporal data.
- Collaboration with Dr. Seo-Won Chang (Yonsei University) and KISTI Supercomputing Center.
- Geomesa (cf. Geowave).
  - Supporting indexing on both spatial and temporal dimensions.
  - Supporting both vector and raster data.
- Temporal data: InfluxDB.

Vector vs. raster data

The raster view of the world	Happy Valley spatial entities	The vector view of the world
	 Points: hotels	
	 Lines: ski lifts	
	 Areas: forest	
	 Network: roads	
	 Surface: elevation	

Geomesa vs. Geowave

- Both use space and time indexing.
- Easily scalable and deployable.
- Geomesa open-sourced by LocationTech
- Geowave open-sourced by the National Geospatial-Intelligence Agency.
- Both exploits the Apache Hadoop ecosystem, in particular, Apache Accumulo (created by the NSA in 2008).
  - Store in Hadoop filesystem.
  - Columnar storage.
  - Billions of rows.
  - Millions of columns.
  - Very fast read/write.
  - Schema free.

Experiments on Intel Xeon Phi x200 series (2017-)

Software Barrier for Modern HPC

The breakdown in Dennard scaling has led to significant changes in HPC computer architectures that achieve power-efficient performance.

Modifying legacy software for these systems can be a significant barrier to performance

- Example: Still only a subset of HPC codes can efficiently utilize multicore server processors with GPGPU accelerators at scale

The Intel roadmap seeks to address both challenges with x86-based many-core coprocessors and bootable processors that achieve performance with standard programming models

2016
Knights Landing
The processor version of the next generation Intel Xeon Phi product family
- 14 nm process
- Processor & Coprocessor
- Over 3 TF DP Peak[1]
- Up to 72 Cores
- On Package High-Bandwidth Memory
- 3x single-thread performance[2]
- Out-of-order core
- Integrated Intel® Omni-Path

2013
Knights Corner
Intel® Xeon Phi™ x100 product family
- 22 nm process
- Coprocessor
- Over 1 TF DP Peak
- Up to 61 Cores
- Up to 16GB GDDR5

FUTURE
Knights Hill
Next generation of Intel® MIC Architecture Product Line
- 10 nm process
- 2nd Generation Integrated Intel® Omni-Path
- In planning –

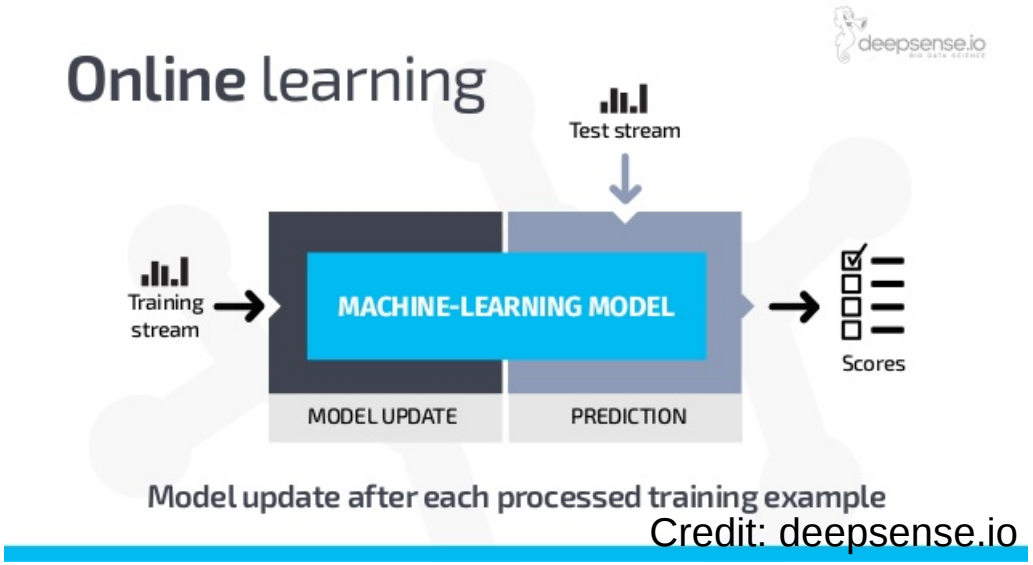
All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. All projections are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

[1]Over 3 Teraflops of peak theoretical double-precision performance is preliminary and based on current expecations of cores, clock frequency and floating point operations per cycle.
[2] Projected peak theoretical single-thread performance relative to 1st Generation Intel® Xeon Phi™ Coprocessor 7120P

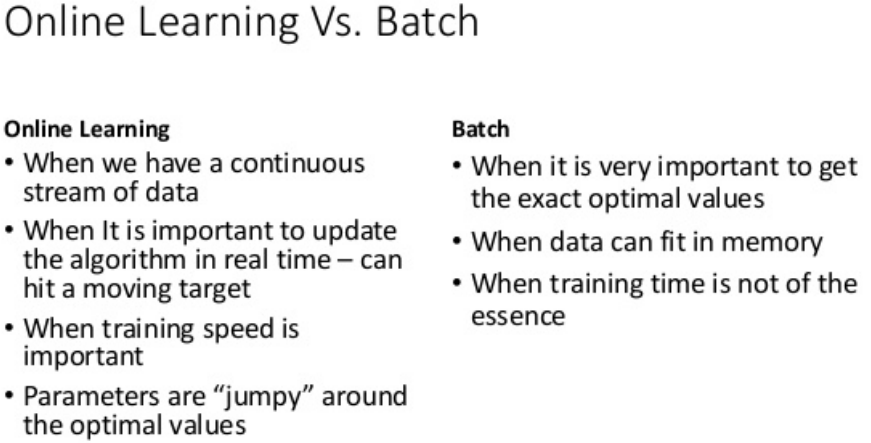
Experiments and code development for time-series analysis in Intel Xeon Phi Knights Corner since 2014.

New demands for streaming processing in astronomy

- **On-line learning** commonly used in industry needs to be adopted in the LSST era with lots of near real-time detection of temporal/spatial variable objects.

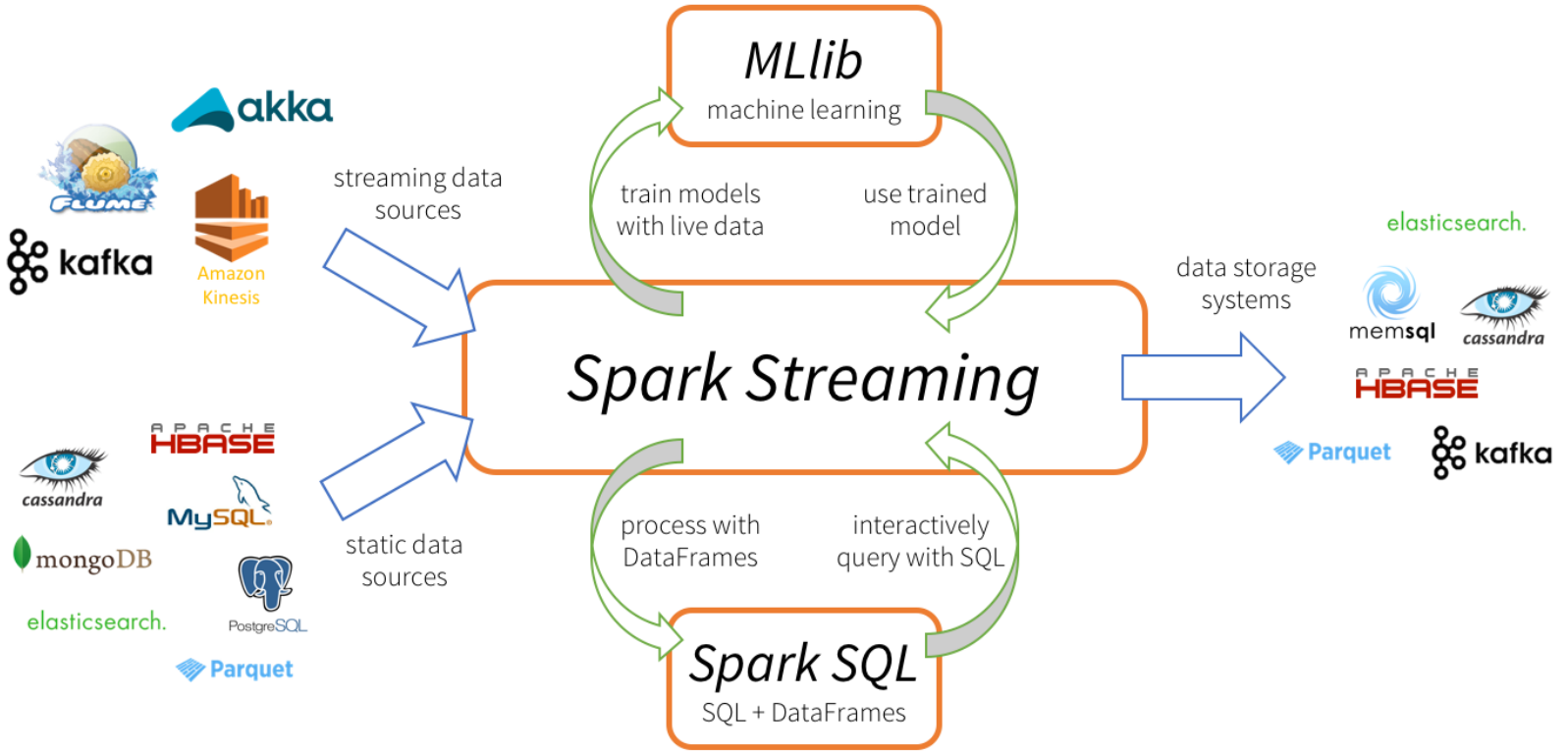


Credit: deepsense.io



Credit: Thomas Jensen@Expedia

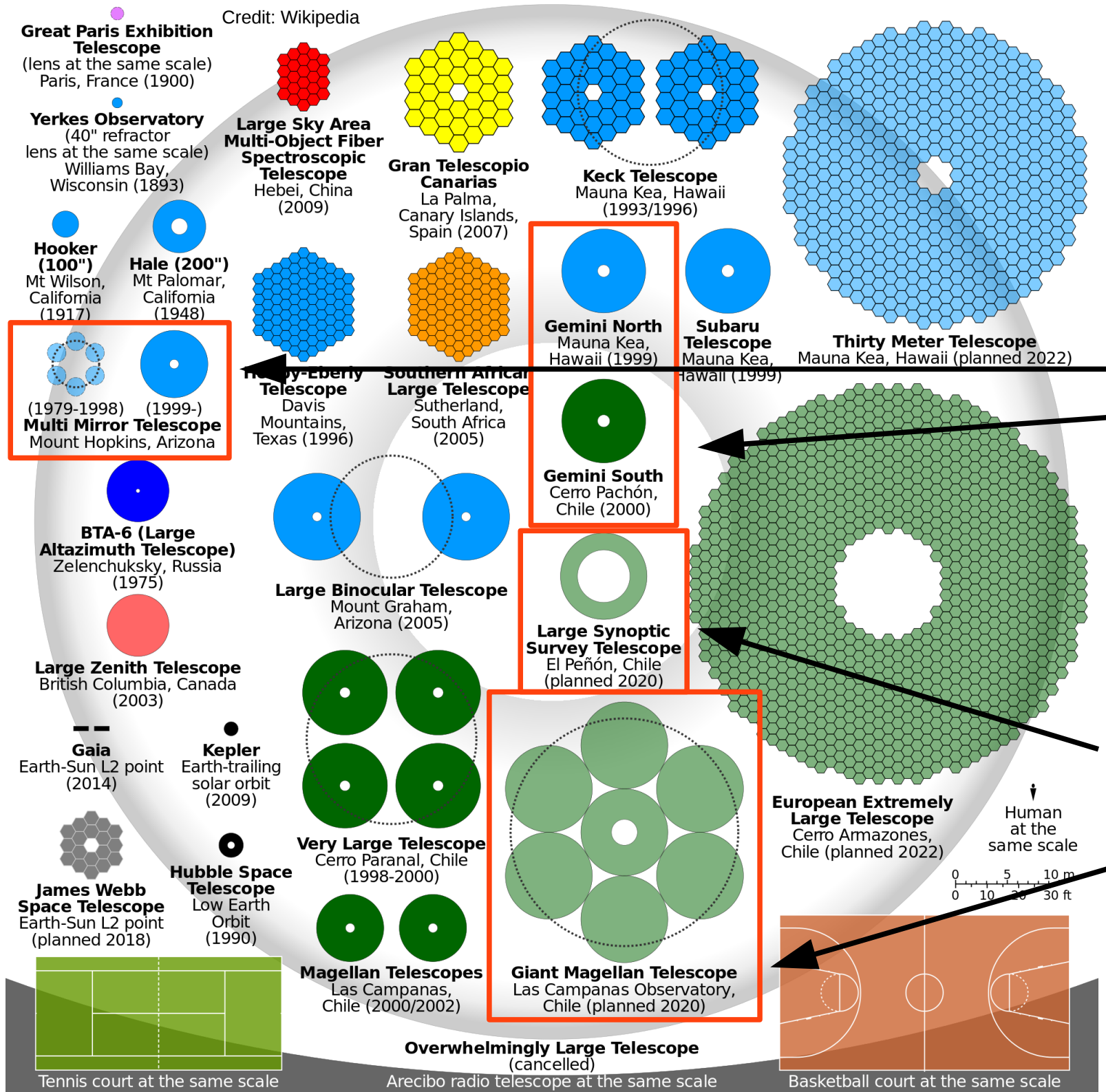
- New computing framework: Apache Spark and Storm.
- Discussion with computer scientists and engineers from Kakao in 2015.
- Experiments using a single node with Apache Storm.
- Key problem: no complex fast algorithms!



Credit: Tathagat Das

홍보- 미래의 대용량 우주 관측 자료

: 대한민국도 참여하는 인류가 획득할 가장 방대한 천문 관측 자료 LSST



더 많은 공간을 탐사 (더 어두운 천체를 보는 것도 포함).

우리나라 천문학자들이 현재 공식적으로 이용하는 대형 광학 망원경.

우리나라 천문학자들이 앞으로 공식적으로 이용하고자 하는 대형 광학 망원경.

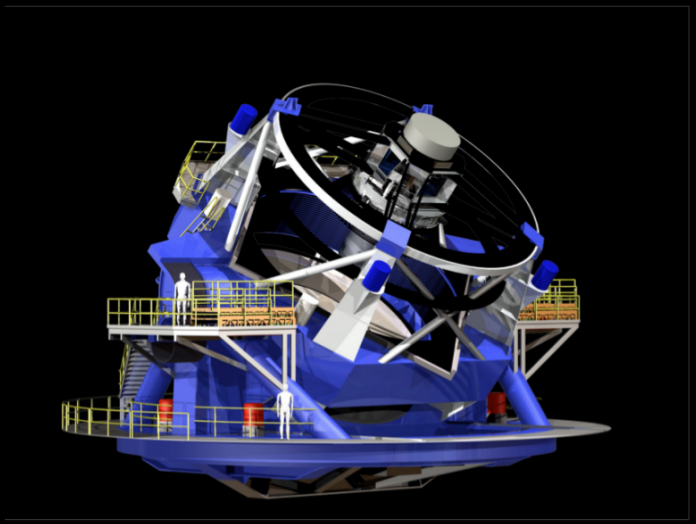
LSST: Large Synoptic Survey Telescope

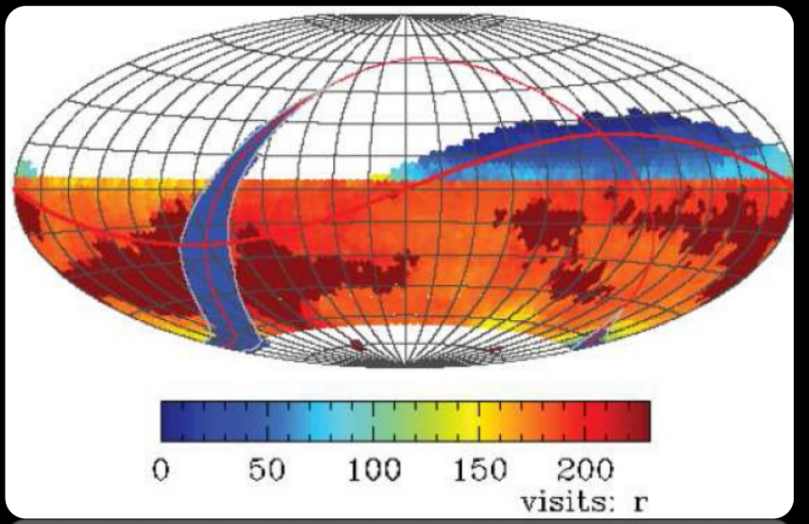
LSST: A Deep, Wide, Fast, Optical Sky Survey

http://lsst.org

8.4m telescope | 18000+ deg^2 | 10mas astrom. | r<24.5 (<27.5@10yr)

ugrizy | 0.5-1% photometry

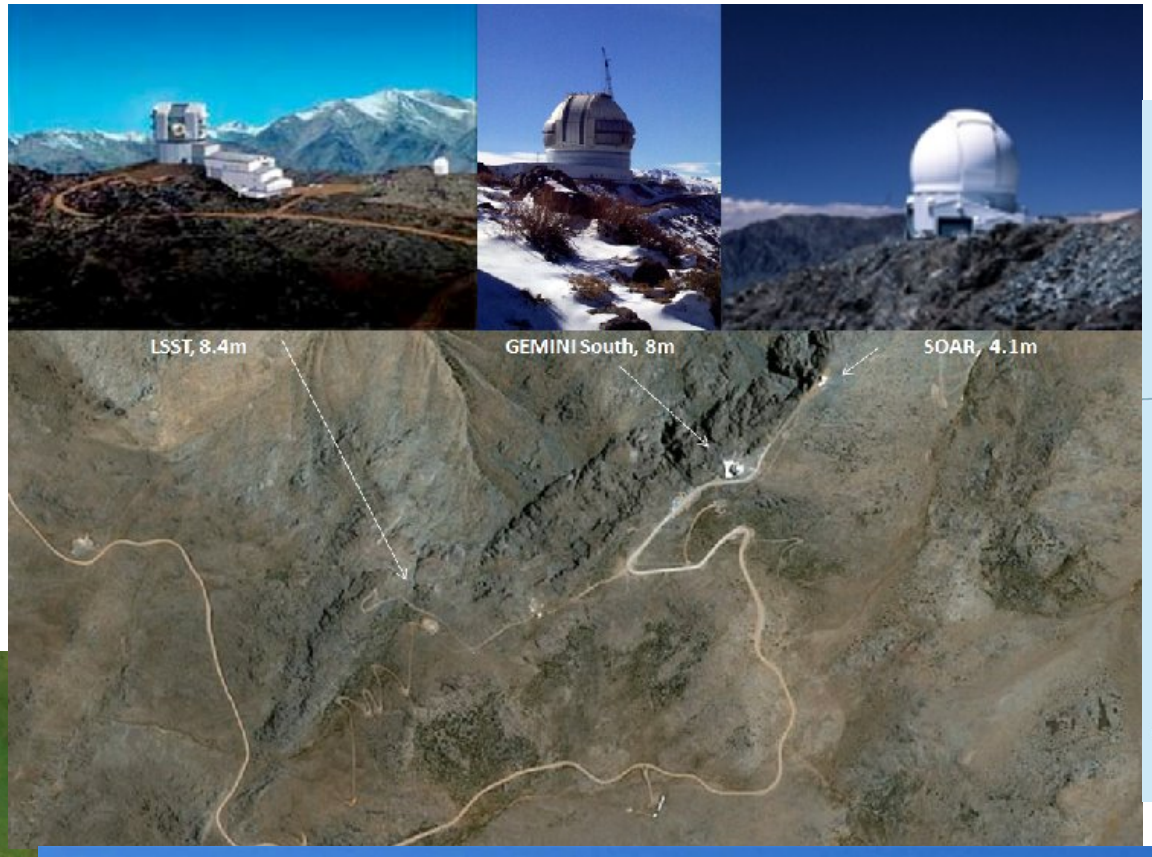




3.2Gpix camera | 30sec exp/4sec rd | 15TB/night | 37 B objects

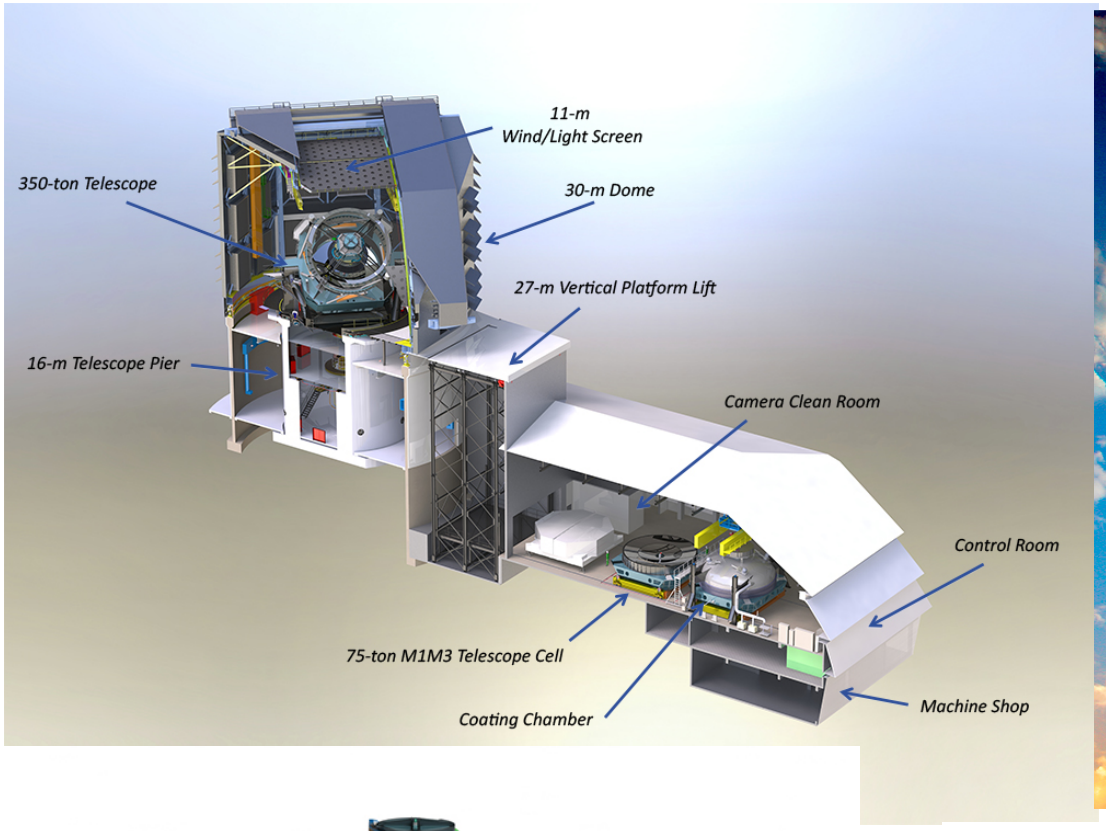
Imaging the visible sky, once every 3 days, for 10 years (825 revisits)

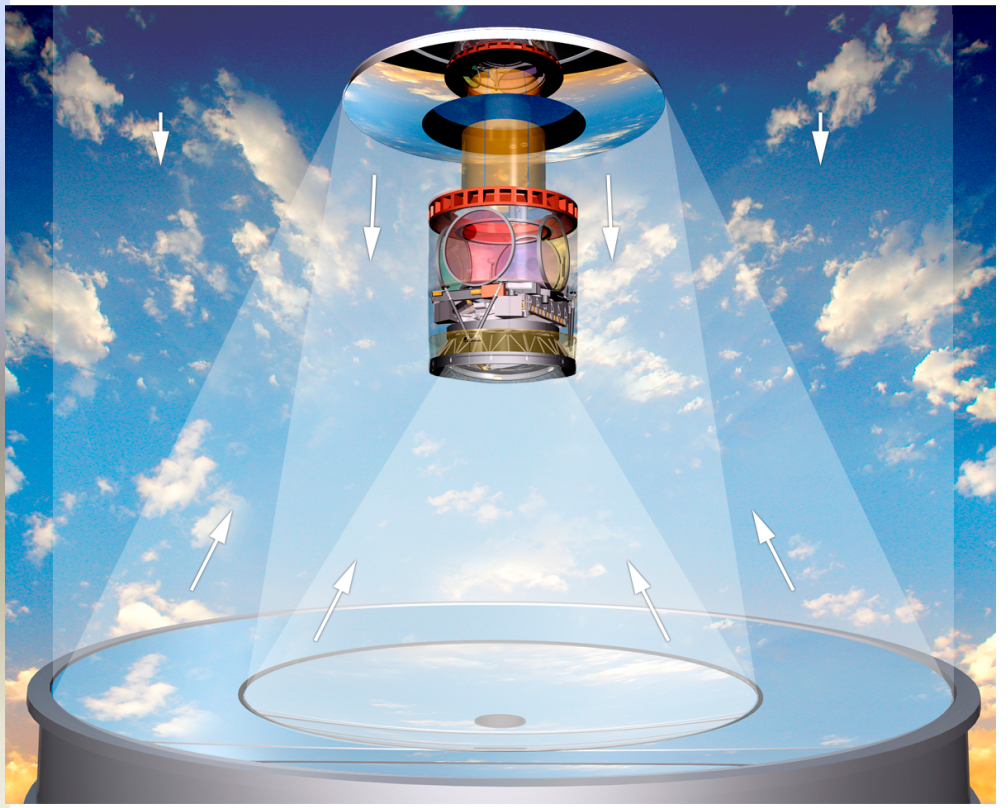
고도 ~ 2647 m
위도 ~ -30도
경도 ~ 서 70도

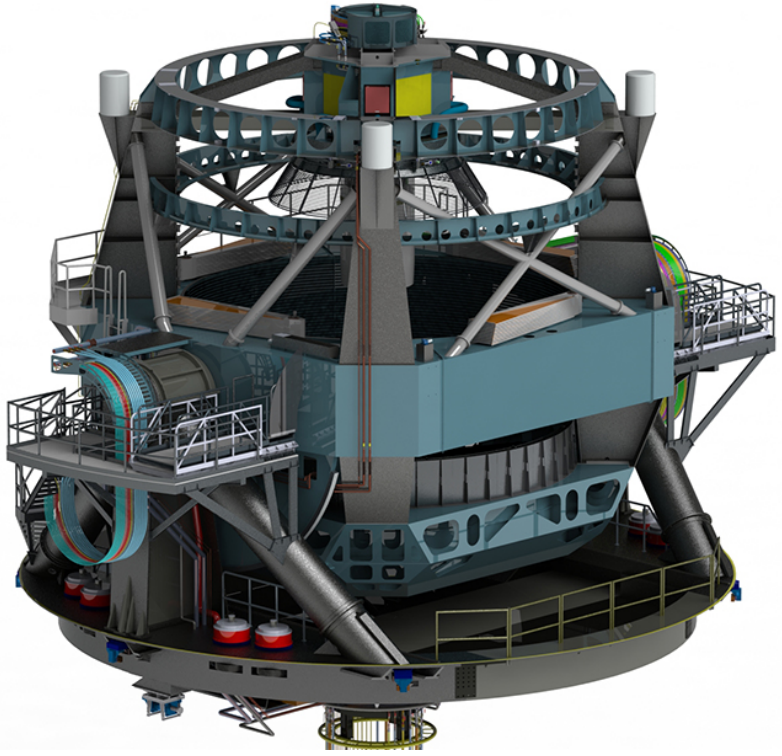


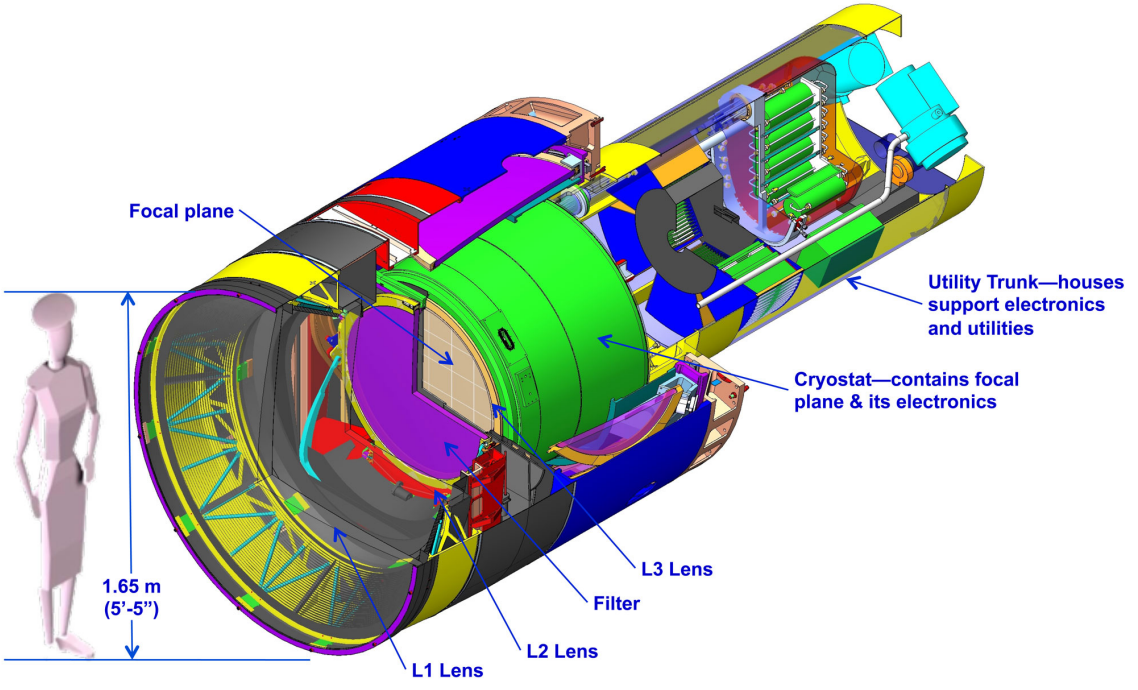
Chile to Host 70 Per cent of World's Astronomical Infrastructure by 2020

World | Indo-Asian News Service | Updated: April 15, 2015 19:04 IST







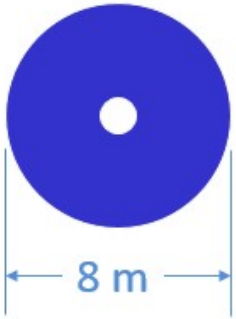


Camera ¾ Section

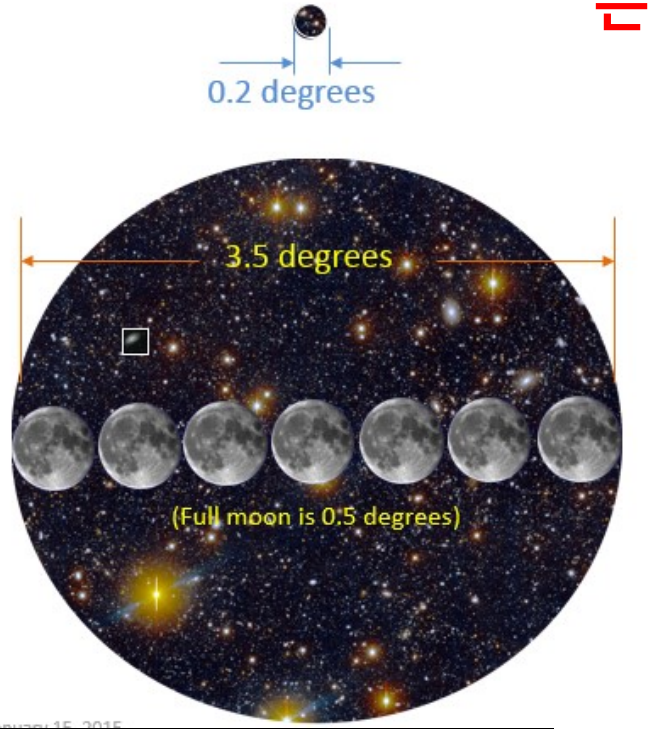
광시야를 가능케 하는 망원경과 그 시야를 한 번에 다 담을 수 있는 거대한 카메라!

8m Class Telescope

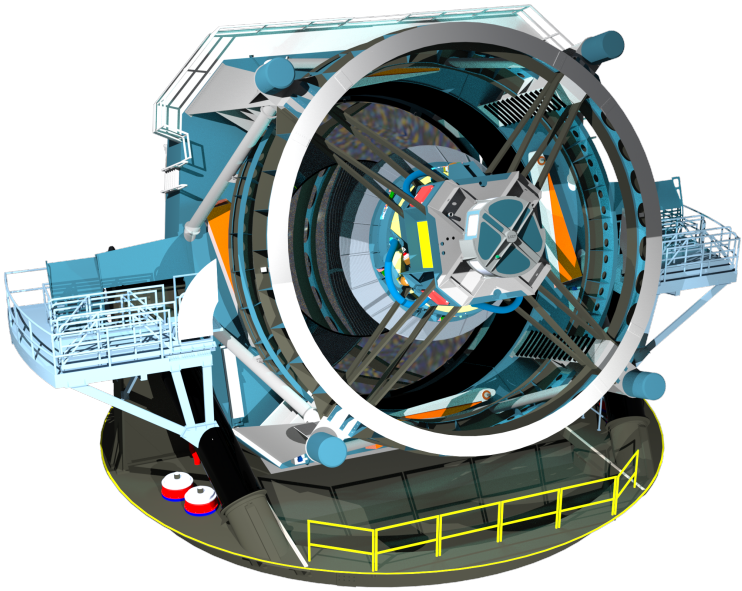
Primary Mirror Diameter

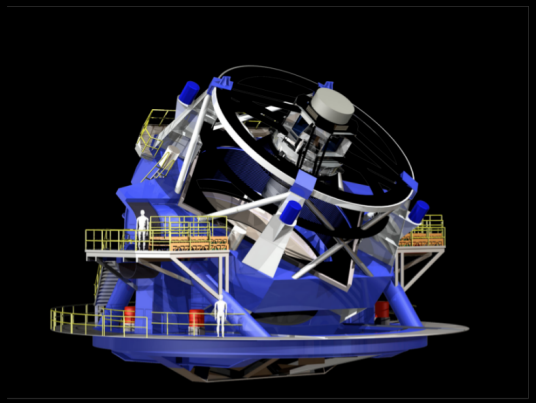


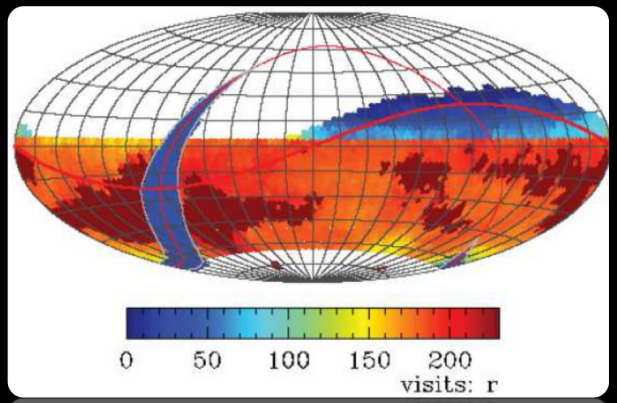
Field of View



LSST

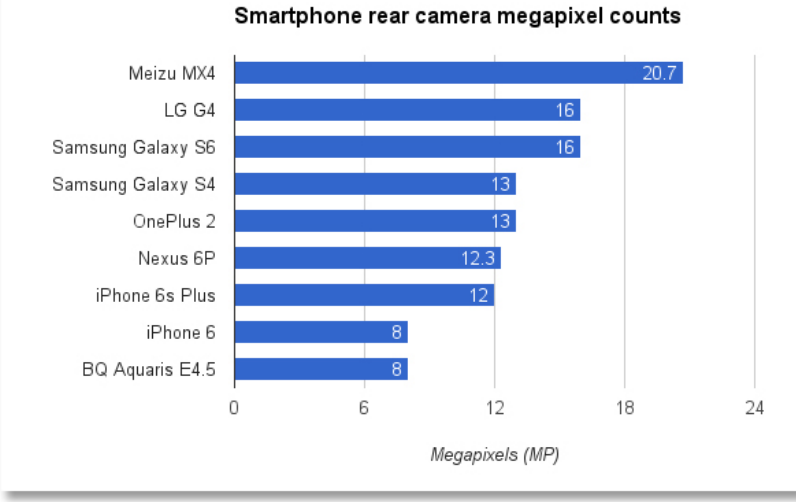






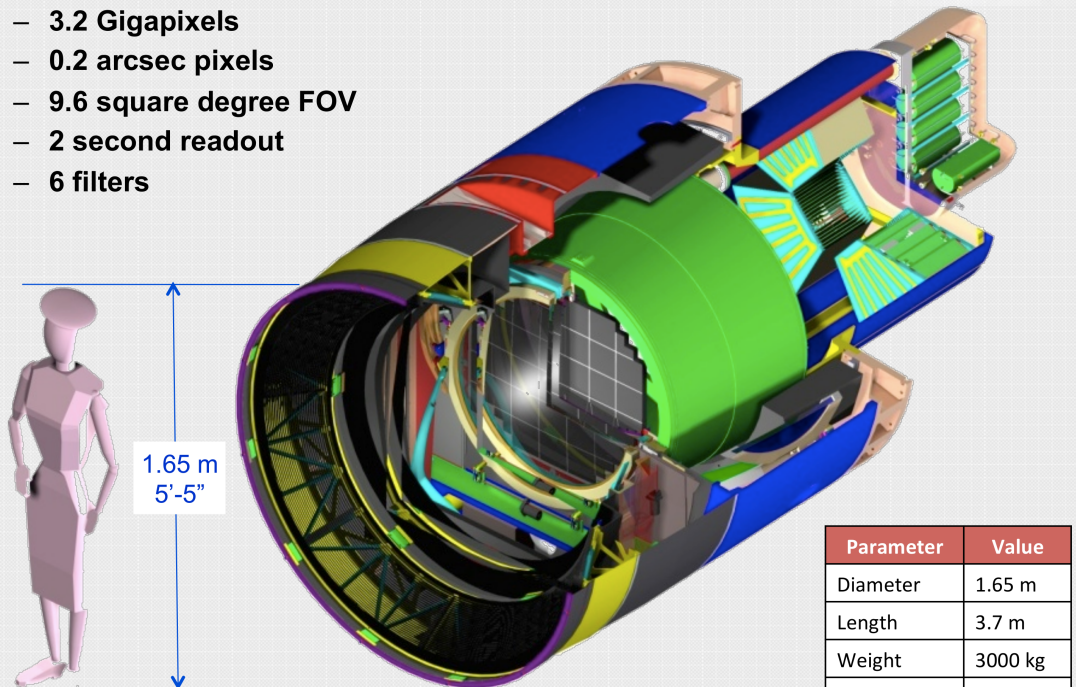
3.2Gpix camera 30sec exp/4sec rd 15TB/night 37 B objects

Imaging the visible sky, once every 3 days, for 10 years (825 revisits)

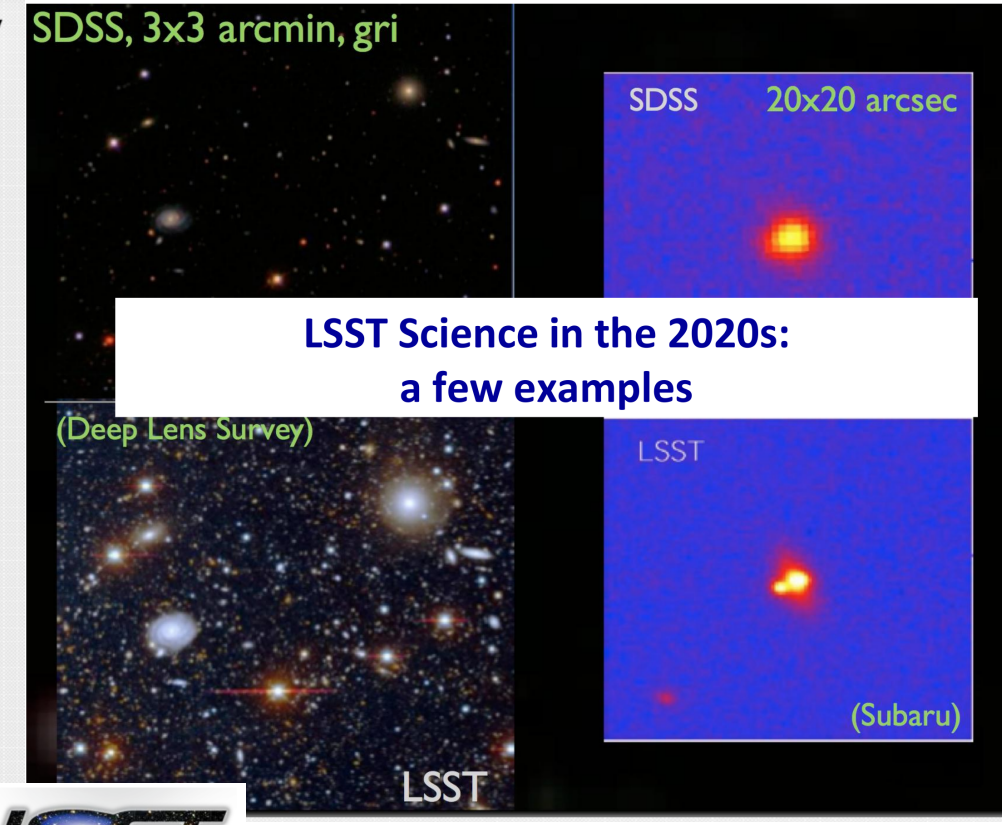


LSST Camera

- 3.2 Gigapixels
- 0.2 arcsec pixels
- 9.6 square degree FOV
- 2 second readout
- 6 filters



Parameter	Value
Diameter	1.65 m
Length	3.7 m
Weight	3000 kg



LSST data volume: 1 night ~ 15 TB …
and in 10 years :

어두운 천체를 자세하게 볼 수 있는 거대한 거울과 같은 영역을 여러 차례에 걸쳐 반복하여 관측!

Number of objects	~37 10^9 (20 10^9 galaxies /17 10^9 stars)
Number of forced measurements	~37 10^9 * 825 ~ 30 10^{12}
Average number of alerts per night	2 10^6 (10 ^{7} including galactic plane)
Number of data collected per 24 hr period	~ 15 TB
Final Raw image	24 PB
Final Disk Storage	0.4 EB (400 PetaBytes)
Final database size	15 PB

- 2020년 시험 관측 시작.
- 2022년 후반기 과학 관측 시작.
- 최소 10년, 기본적으로 15년 관측.

LSST의 4가지 주요 과학 목표

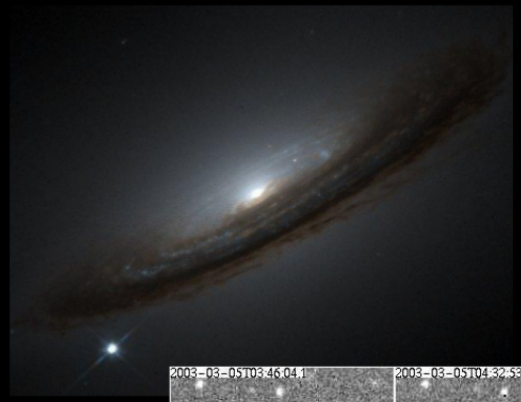
암흑 물질과 암흑 에너지

Multiple investigations into the nature of the dominant components of the universe

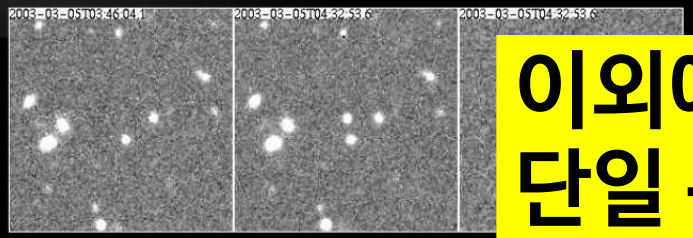
태양계 소천체 파악

Find 82% of hazardous NEOs down to 140 m over 10 yrs & test theories of solar system formation

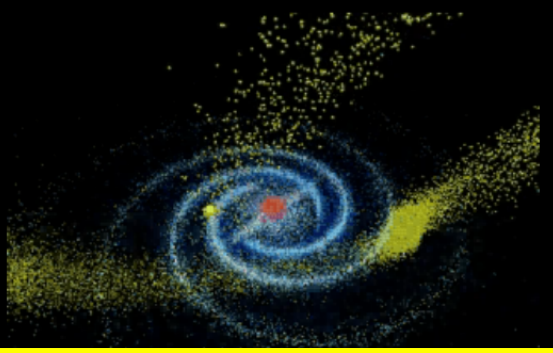
시간에 따른 천체의 변화를 추적



Discovering the transient & unknown on time scales days to years



우리 은하의 별 지도 작성



Map the rich and complex structure of the galaxy in unprecedented detail and extent

이외에도 여러 과학 연구 목표를 위해서 단일 관측 기기를 이용한 관측을 수행.

다양한 목적을 위한 기계학습 활용은 필수.

LSST From the Astronomer's Perspective

매일 밤 약 천 만 개의 밝기나 위치가 변하는 현상을 검출하여 매 분마다 전세계에 제공.

Level 1

- A stream of ~10 million time-domain events per night, detected and transmitted to event distribution networks within 60 seconds of observation.
- A catalog of orbits for ~6 million bodies in the Solar System.

약 육 백 만 개의 소행성 궤도 정보.

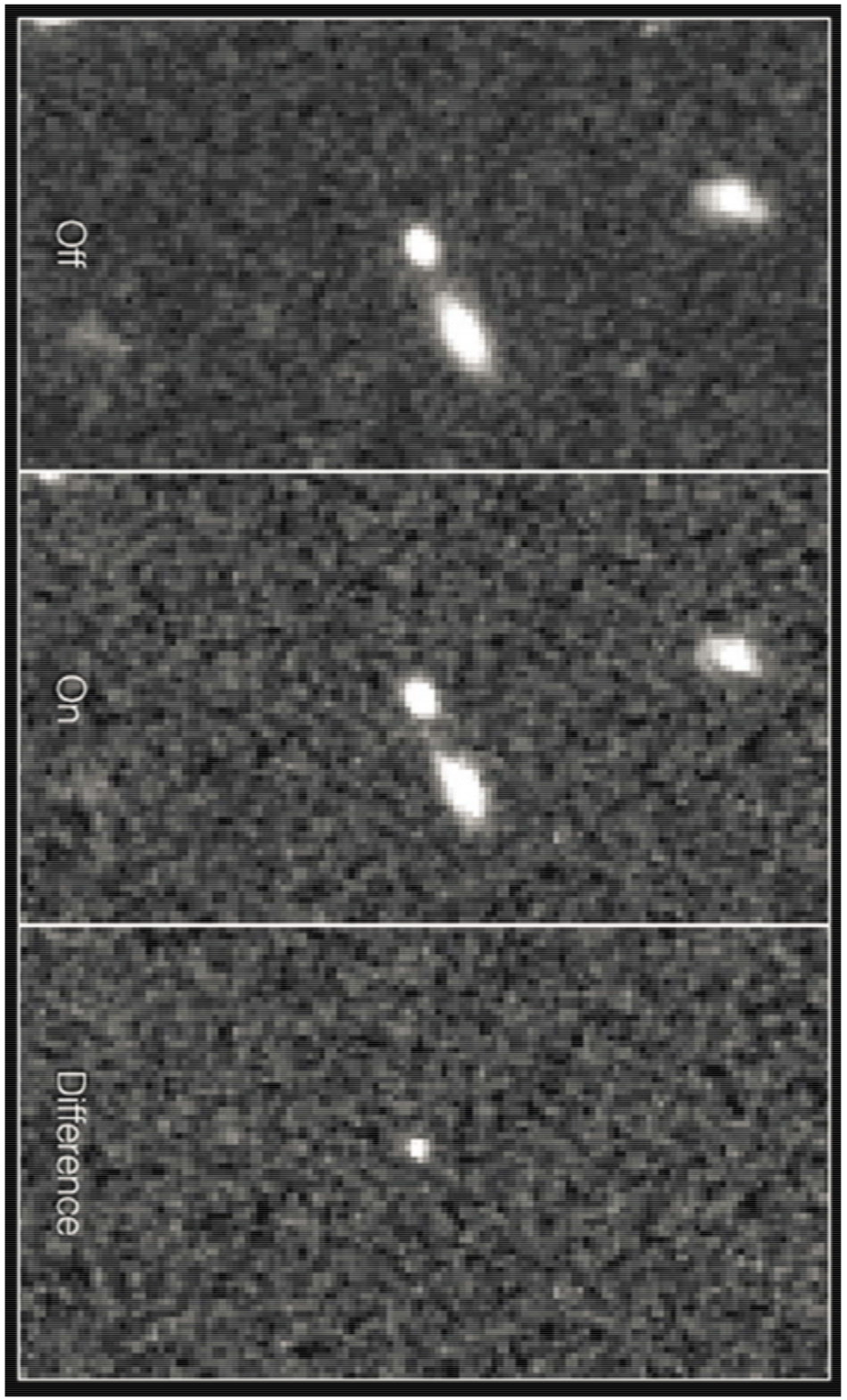
Level 2

- A catalog of ~37 billion objects (20B galaxies, 17B stars), ~7 trillion observations ("sources"), and ~30 trillion measurements ("forced sources"), produced annually, accessible through online databases.
- Deep co-added images.

약 370억개의 천체들 (200억개의 은하와 170억개의 별) 정보.

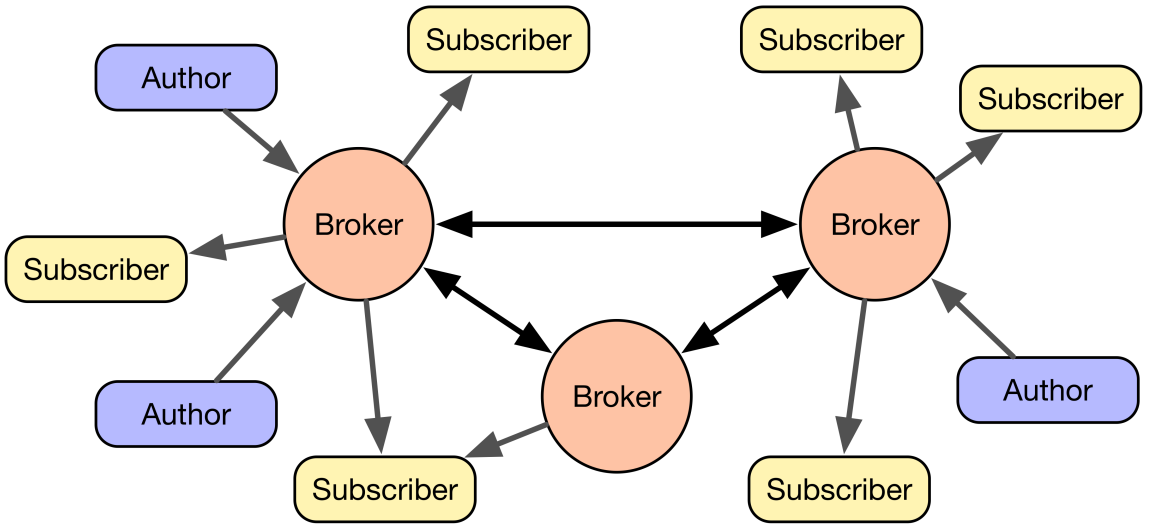
Level 3

- Services and computing resources at the Data Access Centers to enable user-specified custom processing and analysis.
- Software and APIs enabling development of analysis codes.



매일 밤 약 천 만 개의 밝기나 위치가 변하는 현상을 검출하여 매 분마다 전세계에 제공.

- 차등 영상 분석.
- 원래 영상 분석.



- Deep Learning 등의 최신 기계 학습 방법 실험.
- 메시지 손실이 없는 분산 병렬 처리 환경 실험.
- 시간 vs. 정확도의 trade-off 기계학습법 활용 필요.

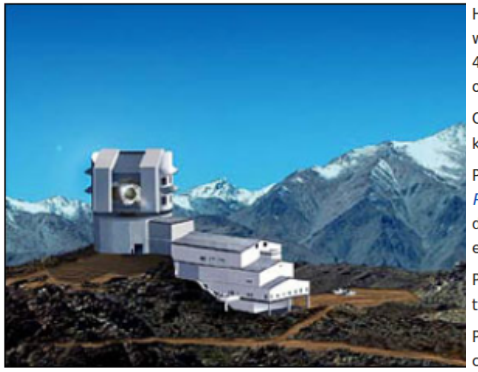
Google is the institutional member of the LSST project!

Rob Pike represents Google in the LSST project.

Google Joins Large Synoptic Survey Telescope Project

January 8, 2007

Google has joined a group of nineteen universities and national labs that are building the Large Synoptic Survey Telescope (LSST).



Artist conception of the Large Synoptic Survey Telescope.

Scheduled to begin operations in 2013, the 8.4-meter LSST will be able to survey the entire visible southern sky deeply in multiple colors every week with its three-billion pixel digital camera, probing the mysteries of Dark Matter and Dark Energy, and opening a movie-like window on objects that change or move rapidly: exploding supernovae, potentially hazardous near-Earth asteroids as small as 100 meters, and distant Kuiper Belt Objects. LSST is a public-private partnership.

LSST and Google share many of the same goals: organizing massive quantities of data and making it useful. Over 30 thousand gigabytes (30TB) of images will be generated every night during the decade-long LSST sky survey. The massive amount of data from LSST must be managed efficiently

Robert Pike (born 1956) is a Canadian software engineer and author. He is best known for his work at Bell Labs, where he was a member of the Unix team and was involved in the creation of the Plan 9 from Bell Labs and Inferno operating systems, as well as the Limbo programming language.

He also co-developed the Blit graphical terminal for Unix; before that he wrote the first window system for Unix in 1981. Pike is the sole inventor named in AT&T's US patent 4,555,775 or "backing store patent" that is part of the X graphic system protocol and one of the first software patents.[1]

Over the years Pike has written many text editors; sam[2] and acme are the most well known and are still in active use and development.

Pike, with Brian Kernighan, is the co-author of *The Practice of Programming* and *The Unix Programming Environment*. With Ken Thompson he is the co-creator of UTF-8. Pike also developed lesser systems such as the vismon program for displaying images of faces of email authors.

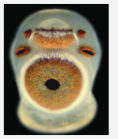
Pike also appeared once on *Late Night with David Letterman*, as a technical assistant to the comedy duo Penn and Teller.[*verification needed*]

Pike is married to Renée French, and currently works for Google, where he is involved in the creation of the programming languages Go and Sawzall.[3]

Rob Pike

Born	1956 (age 58–59)
Nationality	Canadian
Occupation	Software engineer
Employer	Google
Known for	Plan 9, UTF-8, Go
Spouse(s)	Renée French
Website	herpolhode.com/rob/

Rob Pike



Research Area(s)
Data Management
Distributed Systems and Parallel Computing

Co-Authors

Rob Pike is a Distinguished Engineer at Google, Inc. He works on distributed systems, data mining, programming languages, and software development tools. Most recently he has been a co-designer and developer of the Go programming language. Before Google, Rob was a member of the Computing Sciences Research Center at Bell Labs, the lab that developed Unix. While there, he worked on computer graphics, user interfaces, languages, concurrent programming, and distributed systems. He was an architect of the Plan 9 and Inferno operating systems and is the co-author with Brian Kernighan of The Unix Programming Environment and The Practice of Programming. Other details of his life appear on line but vary in veracity.

Google Publications

Data management projects at Google
Wilson Hsieh, Jayant Madhavan, Rob Pike
SIGMOD Conference (2006), pp. 725-726

Interpreting the Data: Parallel Analysis with Sawzall
Rob Pike, Sean Dorward, Robert Griesemer, Sean Quinlan
Scientific Programming Journal, vol. 13 (2005), pp. 277-298

Previous Publications

The Inferno Programming Book: An Introduction to Programming for the Inferno Distributed System
Martin Atkins, Rob Pike, Howard Trickey
John Wiley \& Sons (2005)

Security in Plan 9
Russ Cox, Eric Grosse, Rob Pike, David L. Presotto, Sean Quinlan
USENIX Security Symposium (2002), pp. 3-16

LSST의 우리나라 참여에 예상되는 비용:
약 30명의 천문학자 ~ 최소 약 75억/10년 ~ 최소 약 7.5억/년. 예산 등 마련이 어려운 상황.

Charles and Lisa Simonyi Fund: 20 million,
Bill Gates: 10 million USD!

대중이 참여하여 DO SCINECE!

국내 최초 시도
: 순수과학을 위한 산학연
민간 + 공공 기금 활용.

국내 **LSST** 연구 지원 기부 지원자의 이름으로,

- 온라인을 통한 대중의 참여 기회 마련,
- 체계적인 교육 및 대중 활동 프로그램 운영,
- 과학적 발견의 후원자로 명시.

- 기계 학습의 활용은 대용량 다 파장 광역 변화 탐사 자료 분석에 필수적으로 이용됨.
- 단, 이는 천문학자들이 적절하게 학습 가능한 입력 자료를 설정하고 적절한 학습 방법 및 그 활용이 가능한 경우를 발견하는 능력에 의존함.
- Naver의 다양한 분야와 관련성 존재.
- 미래에는 완전한 인공 지능의 로봇과 같이 하는 천문학 연구의 시기가 올지도? 인간을 대체하지는 못하나, 우주의 비밀을 밝혀내는데 도와주는 역할을 기대.