Data-driven approach to Type Ia supernovae:

1) Variable selection on the peak luminosity and 2) Clustering in visual analytics

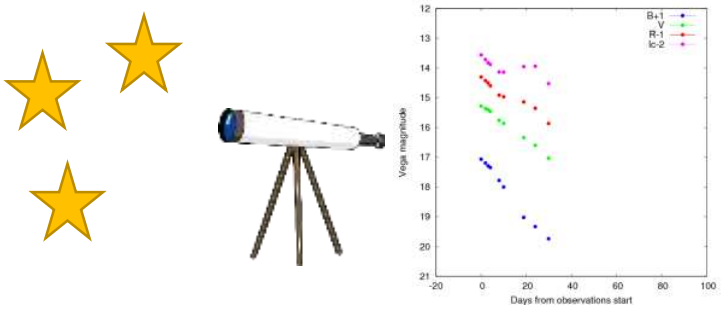
Makoto Uemura (Hiroshima University)

Koji S. Kawabata, Shiro Ikeda, Keiichi Maeda,
Hsiang-Yun Wu, Kazuhiro Watanabe, Shigeo Takahashi, and Issei Fujishiro

Outline

Past

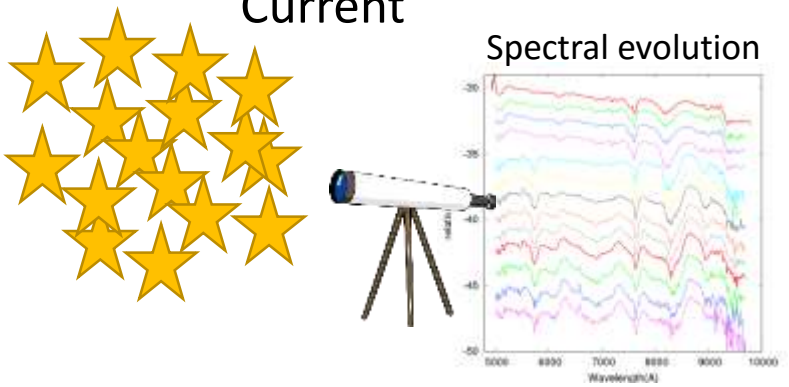
- Magnitude
- Color



Analysis based on the experience of domain experts.

Current

Spectral evolution



Analysis based on the data-driven method is required.

- General introduction about supernovae
- Our recent works
  - Variable selection for the peak luminosity using LASSO
  - Visual analytics for classification

Supernovae

In Large Magellanic Cloud

On Feb. 23, 1987

SuperNova, SN 1987A

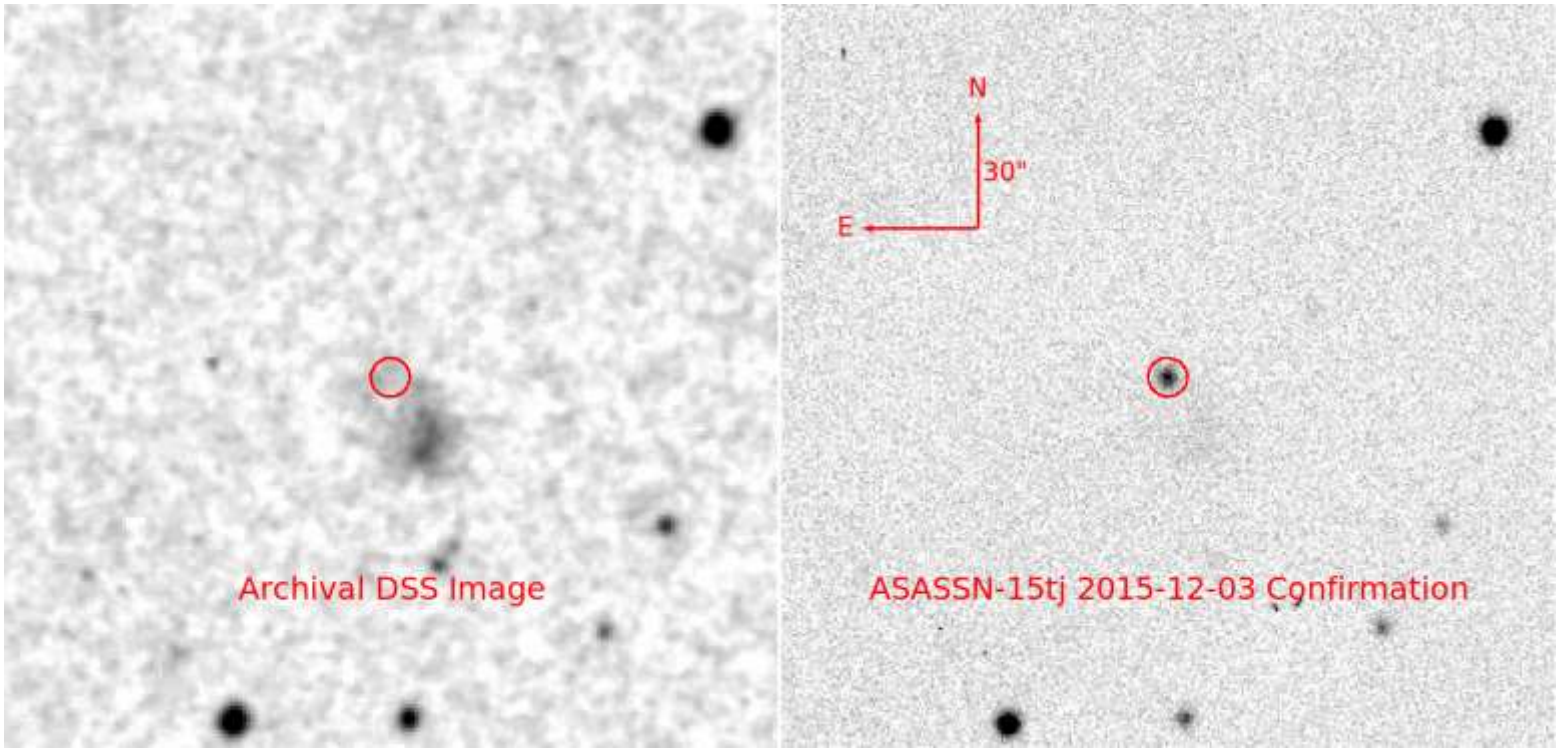
neutrino

Kamiokande

Nobel prize 2002

Supernovae can be brighter than its host galaxy

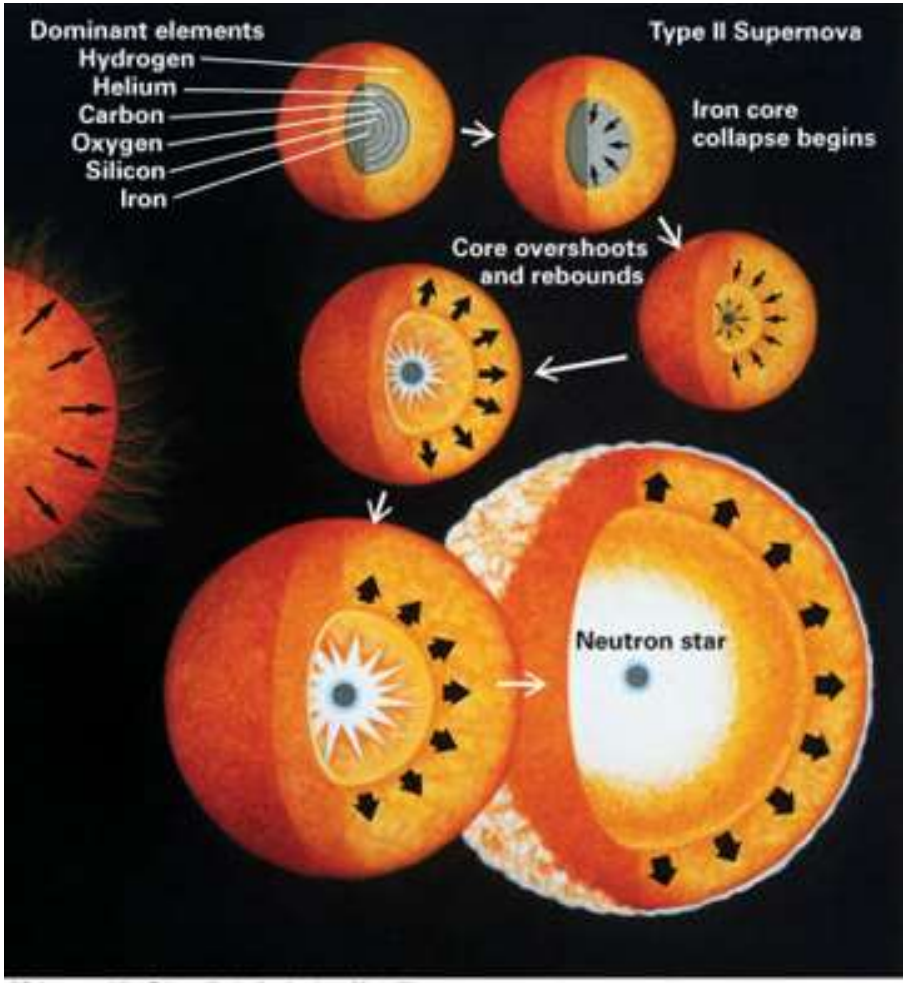
- We can observe supernovae even when its galaxy is too faint to be detected.
- This means that we can see more distant area in the Universe using supernovae.



From “bright supernova” Web site

What are exploding?

- Core-collapse supernovae
  - Death of massive stars

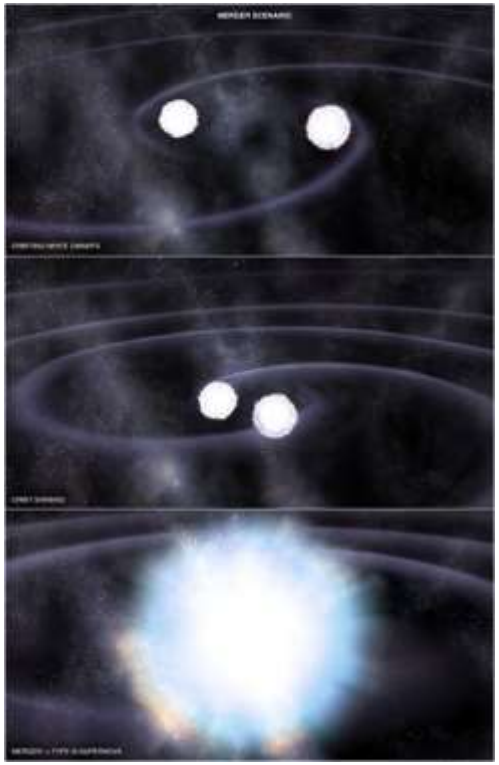


of Science and the Future; illustration by Jane Meredith

What are exploding?

- Thermonuclear supernovae (Type Ia)
  - White dwarf
  - Some accretion → critical mass → thermonuclear runaway reaction

Merger of two white dwarfs



Accretion in a binary system

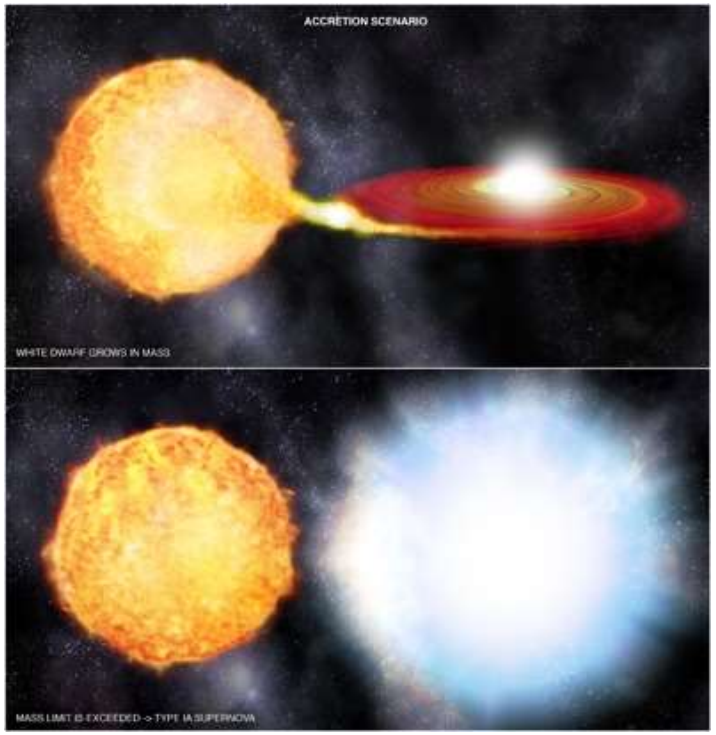


Image: NASA/CXC/M

Supernova as a distance indicator

Distance, *d*

Observed flux, *f*

Luminosity, *L*

- $f = \frac{L}{4\pi d}$

Common luminosity
in the case of supernovae

- We need to take into account the cosmological effect and extinction by the interstellar medium.

Distance of galaxies

- Nobel prize in 2011 → the discovery of the accelerating expansion of the Universe

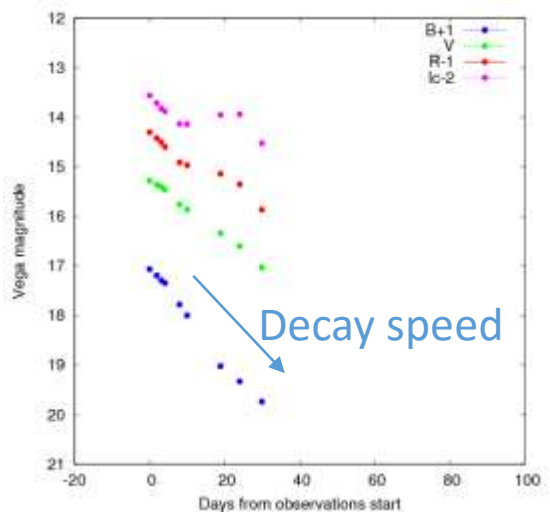
Data we can get

- Photometric data
  - Easier to obtain
- Spectroscopic data
  - Increasing in its volume

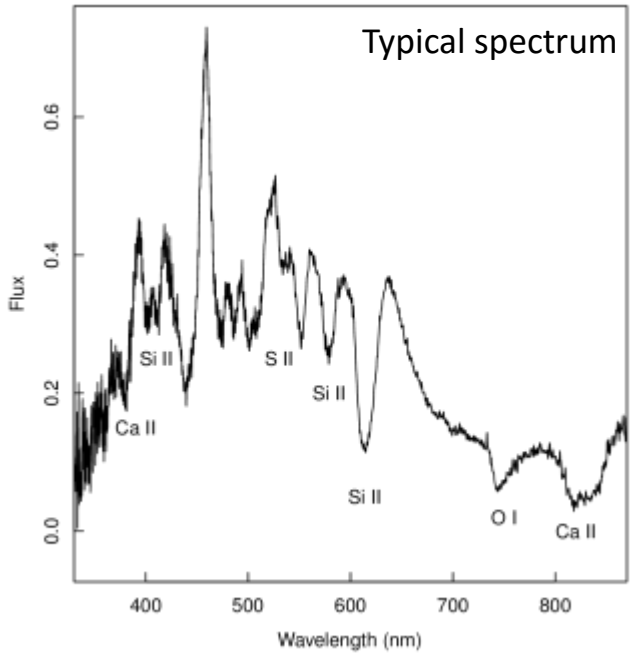
→ data-driven approach

image

Band filter



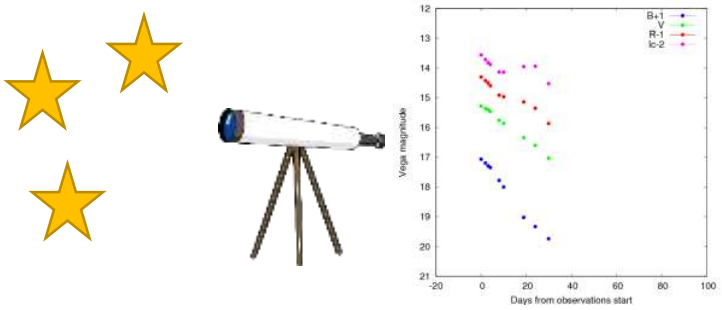
- Magnitude
- Color
- Decay rate



Outline

Past

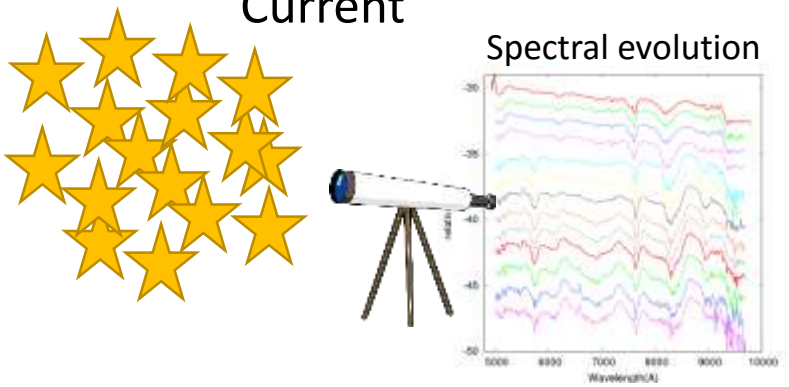
- Magnitude
- Color



Analysis based on the experience of domain experts.

Current

Spectral evolution



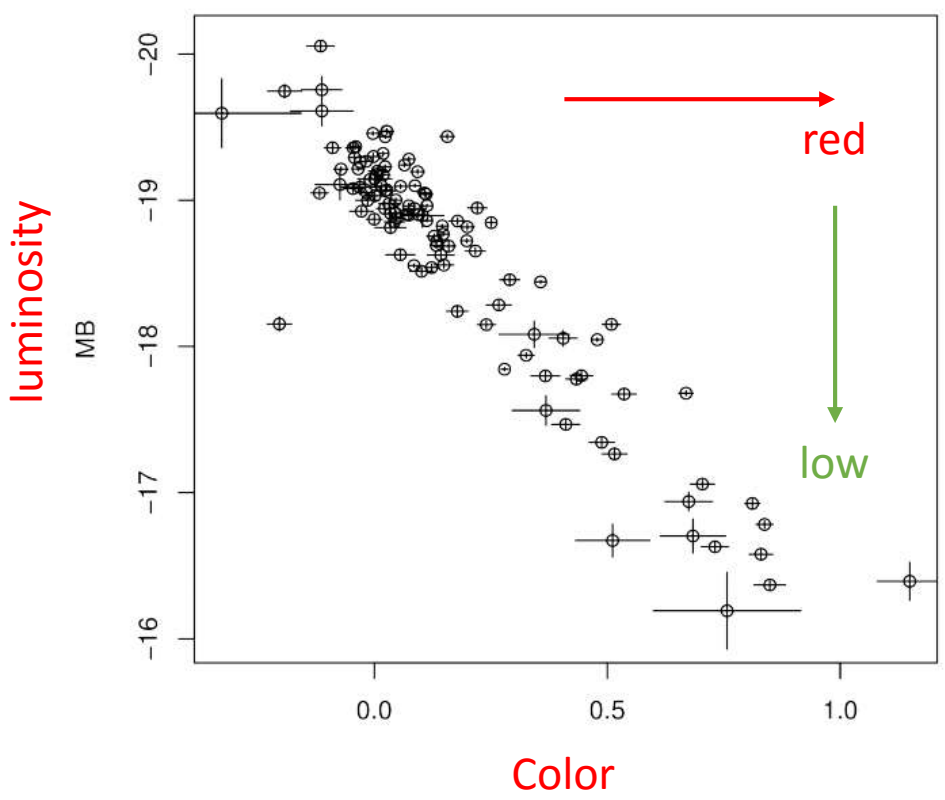
Analysis based on the data-driven method is required.

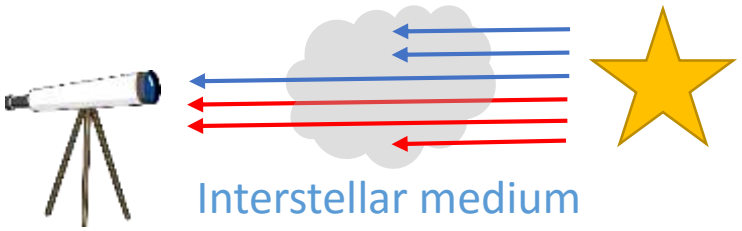
- General introduction about supernovae
- Our recent works
  - Variable selection for the peak luminosity using LASSO
  - Visual analytics for classification

Diversity in the peak luminosity

- Supernovae have a common luminosity, after some corrections.
- Strong correlation with the color

$$M = M_0 + \beta c$$

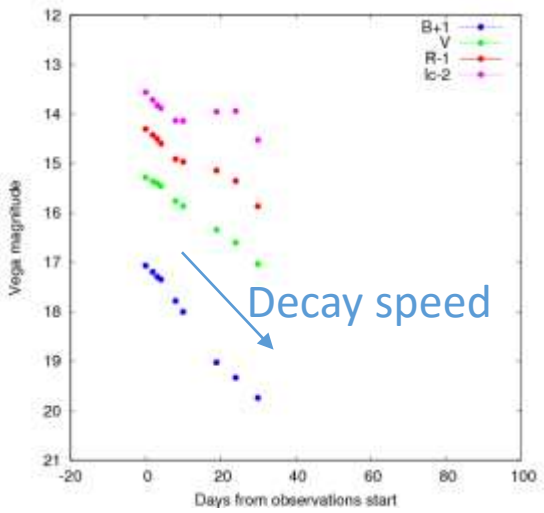


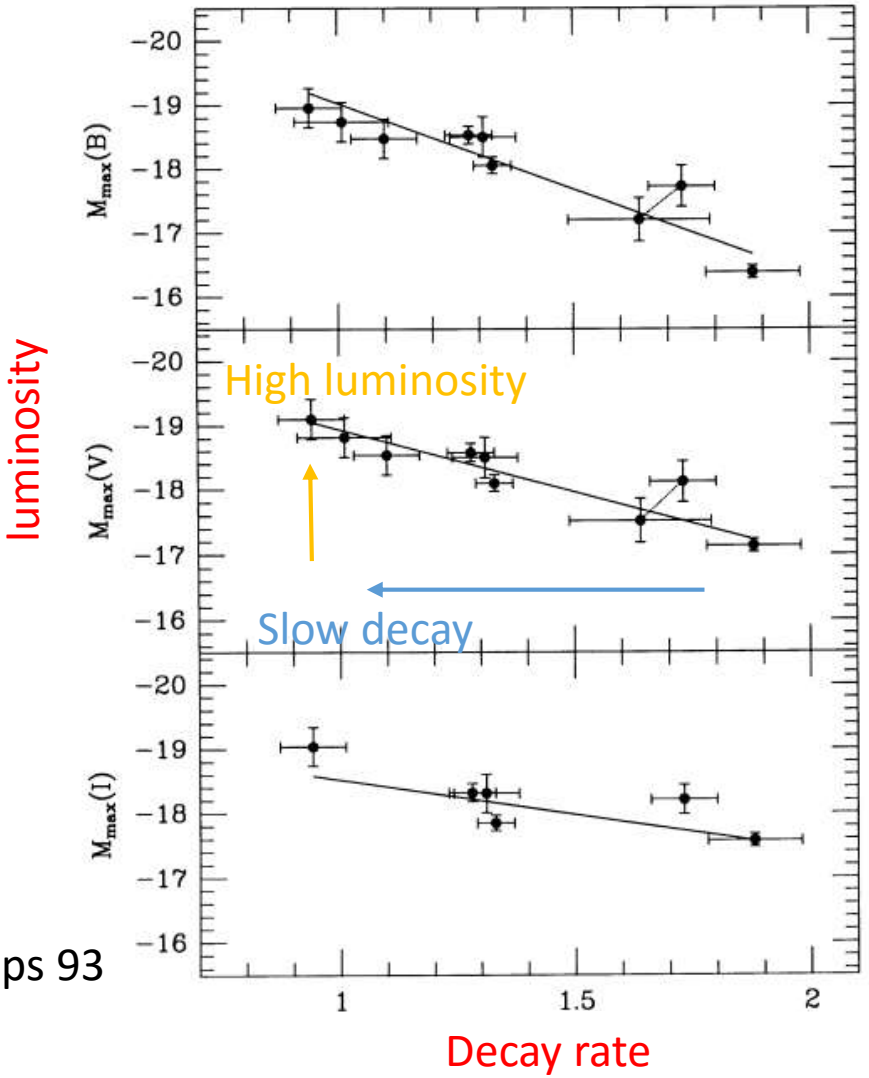


Interstellar extinction
- Flux decreasing
- Color reddening

Diversity in the peak luminosity

- Phillips relation
  - Significant correlation with the decay rate (speed)





Phillips 93

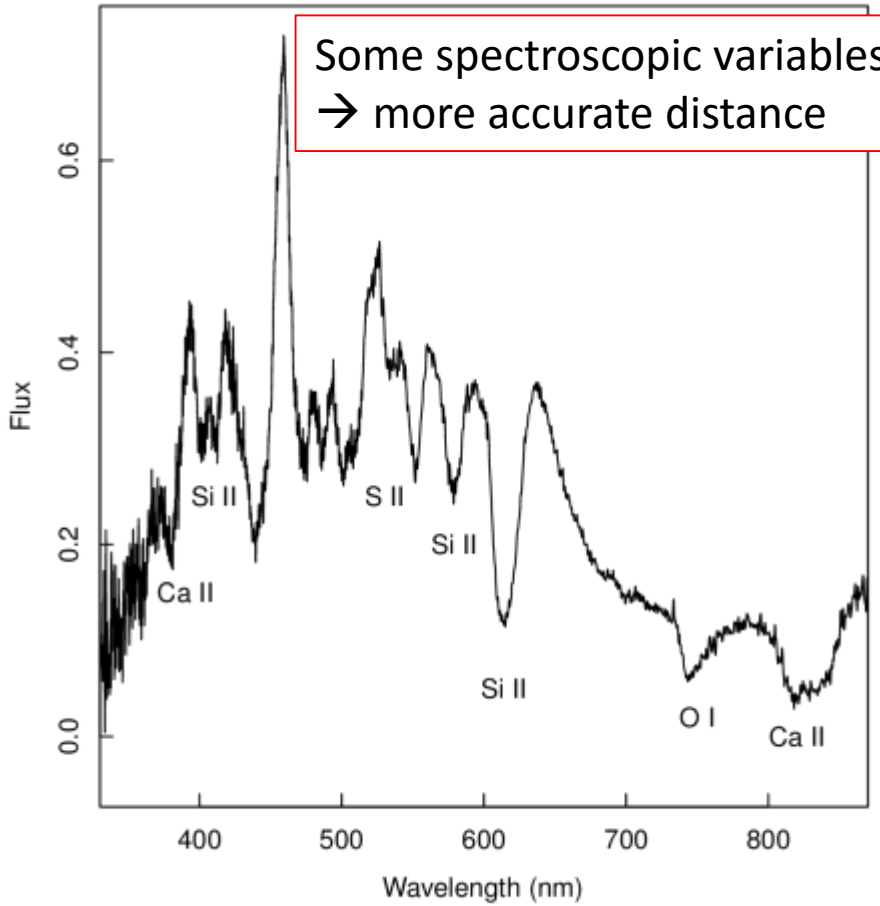
And more…?

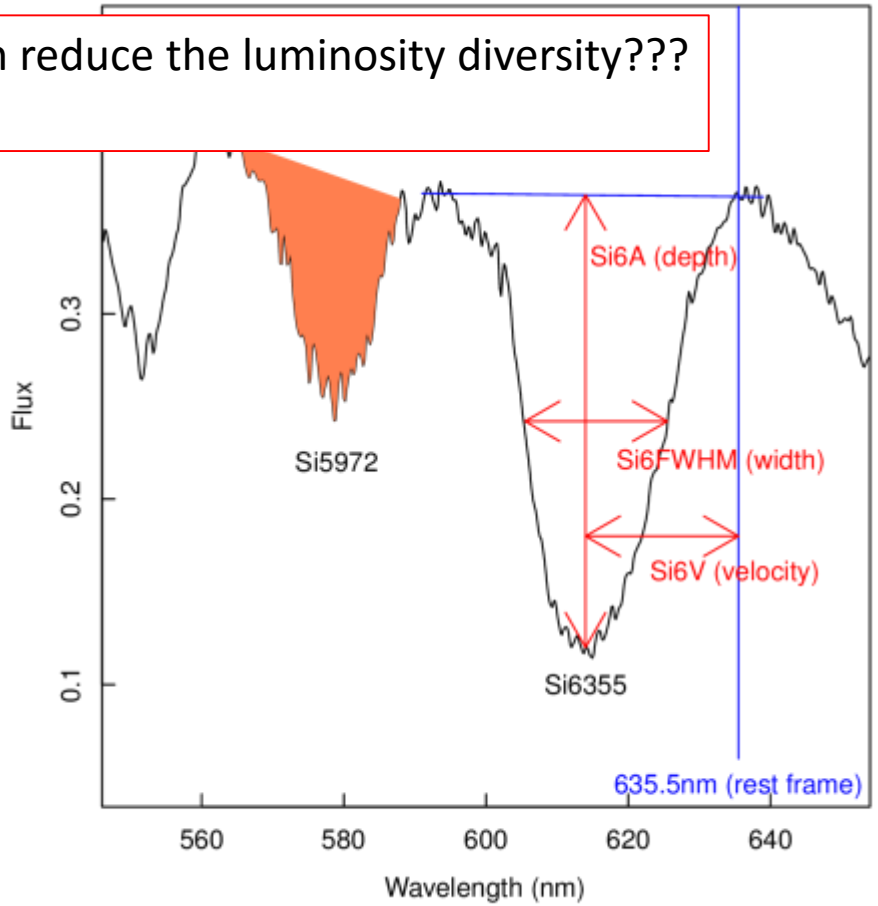
- Searching for the 3rd (or more) parameter

Color Decay rate

$$M = M_0 + \beta_1 c + \beta_2 x_1 + ???$$

Some spectroscopic variables can reduce the luminosity diversity???
→ more accurate distance



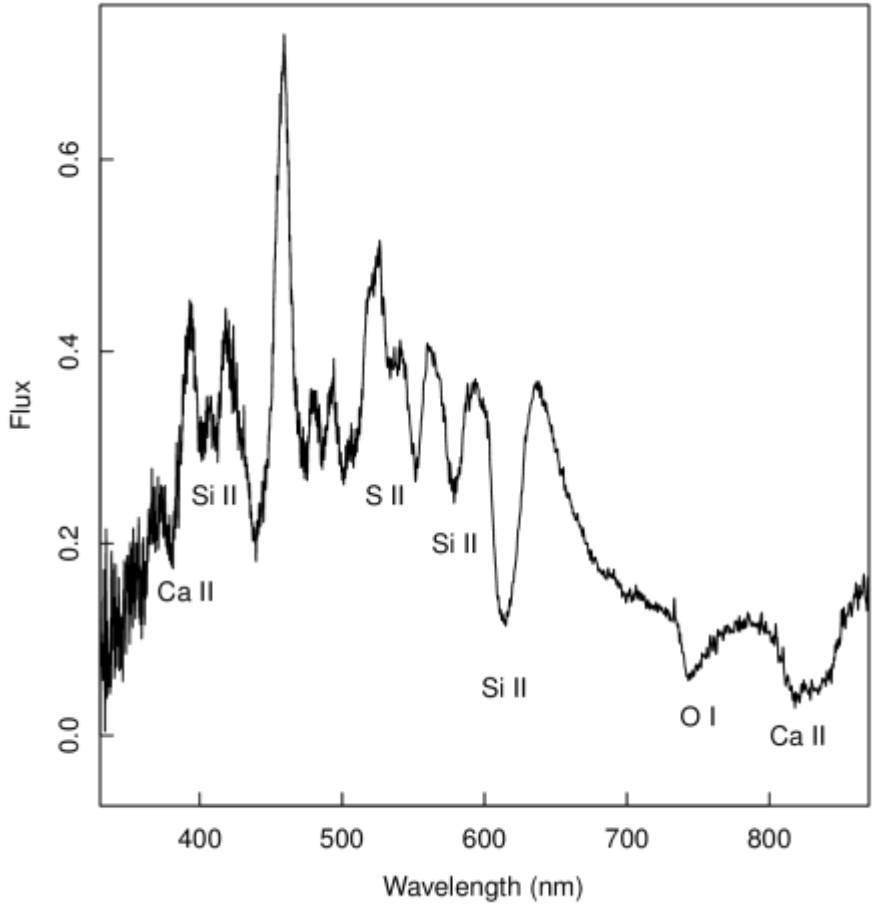


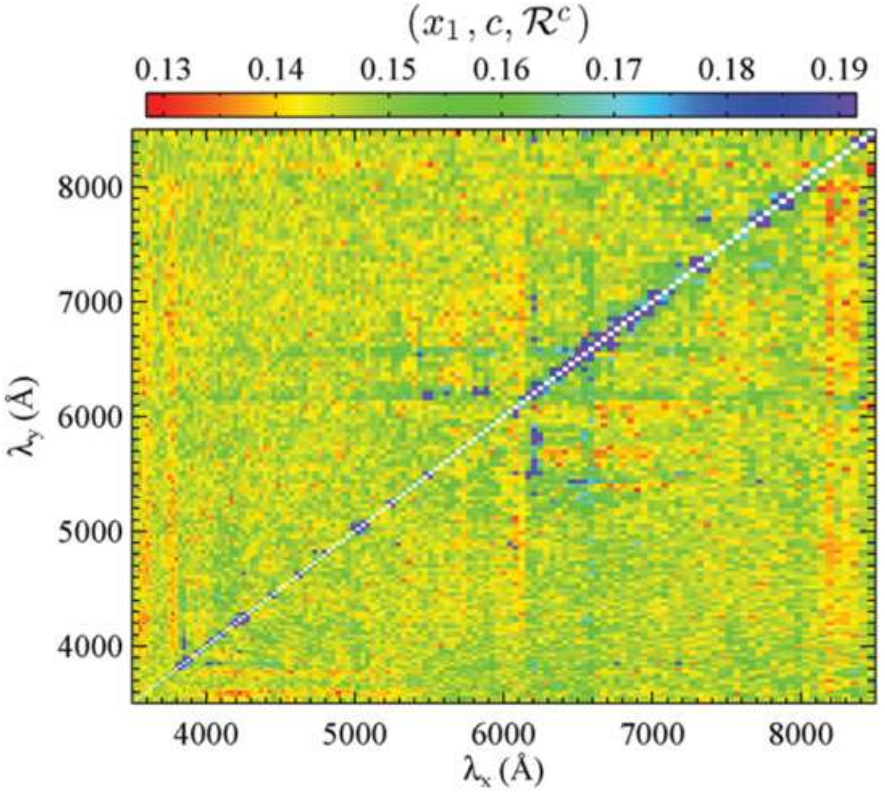
Past studies

- Velocity of Si II 6355 Å (Blondin+11)
- Velocity of Ca II H&K (Foley&Kasen 11)
- Depth of the blue side of S II “W” (Blondin+11)
- Equivalent width (EW) of Si II 4000 Å (Arsenijevic+08, Walker+11, Chotard+11, Nordin+11, Walker+11)
- Equivalent width (EW) of Si II 5972 & 6355 Å (Hachinger+06, Nordin+11)
- EW ratio of silicon, EW(5972)/EW(6355) (Nugent+95, Hachinger+06)
- Flux ratio F(S II “W”)/F(Si II 6355) (Bongard+06)
- Trying all possible combinations of (, or arbitrary) flux ratios as explanatory variables of luminosity.

Using arbitrary flux ratios

- Silverman et al. 2012
  - ≳17,000 flux ratios
  - The number of samples = 62
  - Choose the flux ratio leading to the smallest error.





Variable selection approach

Candidates of variables

$$\begin{pmatrix} MB_{\mathrm{SN1994S}} \\ MB_{\mathrm{SN1995E}} \\ \vdots \\ MB_{\mathrm{SN2008s1}} \end{pmatrix} = \begin{pmatrix} x1_{\mathrm{SN1994S}} & c_{\mathrm{SN1994S}} & EW_{\mathrm{SiII4000,SN1994S}} & FWHM_{\mathrm{SiII4000,SN1994S}} \\ x1_{\mathrm{SN1995E}} & c_{\mathrm{SN1995E}} & EW_{\mathrm{SiII4000,SN1995E}} & FWHM_{\mathrm{SiII4000,SN1995E}} \\ \vdots & & & \\ x1_{\mathrm{SN2008s1}} & c_{\mathrm{SN2008s1}} & EW_{\mathrm{SiII4000,SN2008s1}} & FWHM_{\mathrm{SiII4000,SN2008s1}} \end{pmatrix}$$

Supernova luminosity

$$\begin{pmatrix} 3535/3512_{\mathrm{SN1994S}} & 3558/3512_{\mathrm{SN1994S}} & \cdots & 8416/8472_{\mathrm{SN1994S}} \\ 3535/3512_{\mathrm{SN1995E}} & 3558/3512_{\mathrm{SN1995E}} & \cdots & 8416/8472_{\mathrm{SN1995E}} \\ \vdots & & & \\ 3535/3512_{\mathrm{SN2008s1}} & 3558/3512_{\mathrm{SN2008s1}} & \cdots & 8416/8472_{\mathrm{SN12008s1}} \end{pmatrix}$$

coefficients

$$\begin{pmatrix} c_{x1} \\ c_c \\ c_{EW} \\ c_{FWHM} \\ c_{3535/3512} \\ c_{3558/3512} \\ \vdots \\ c_{8416/8472} \end{pmatrix}$$

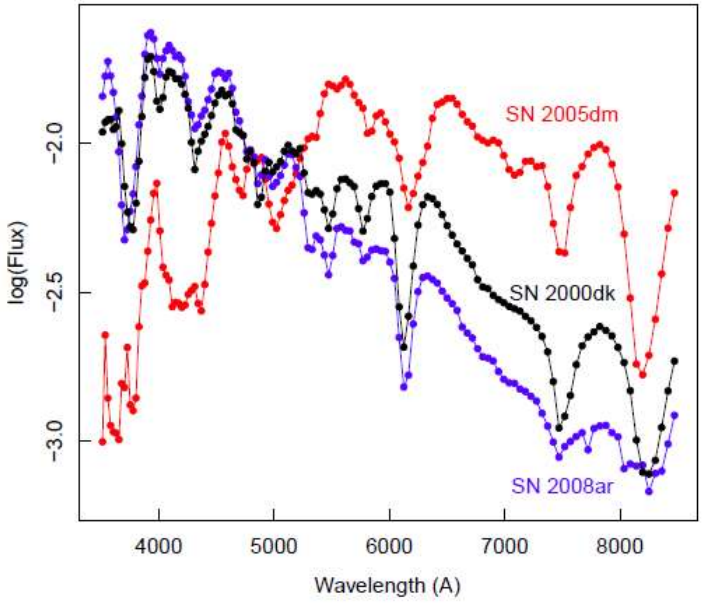
- Our approach
  - Too many candidates → reduce them
  - LASSO + cross-validation

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \qquad \|\boldsymbol{\beta}\|_1 = \sum_i |\beta_i|$$

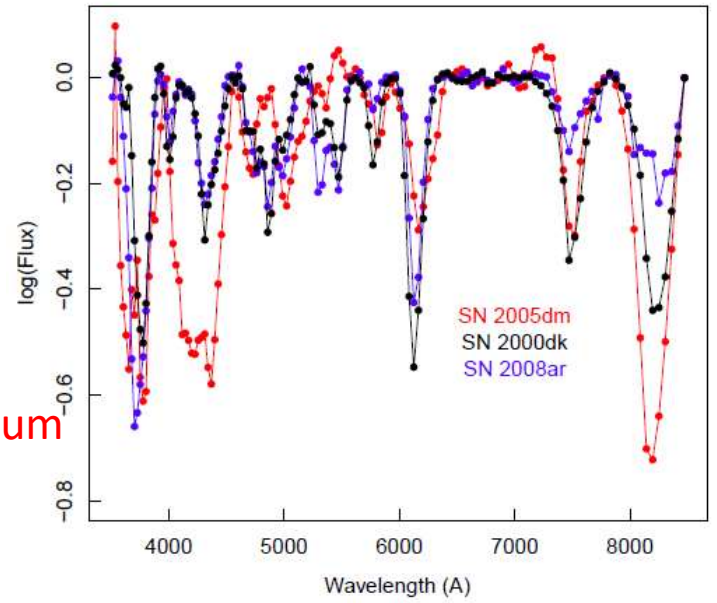
Data

- Berkeley supernova database
  - Filippenko & Silverman
- Our data set is the same as that used in the past study, Silverman+12.
- The number of samples is 78.
- 276 Variables (≪17,000 in arbitrary flux ratios)
  - Two kinds of normalized spectra
  - Log scale ($\log f1 - \log f2 = \log \frac{f1}{f2}$)
  - Other previously proposed variables (color and decay rate, etc..)

Spectra normalized by the total flux
(local color information)



Spectra normalized by the continuum
(absorption line information)



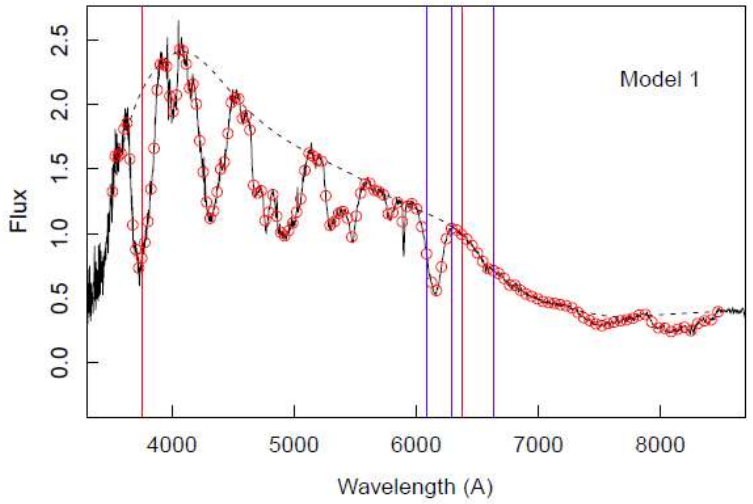
Result 1/3: using all candidates

$$
\begin{aligned}
M_B = {} & M_{B,0} + \beta_1 c + \beta_2 x_1 \\
& + \beta_3 f_{\rm tot}(3512) + \beta_4 f_{\rm tot}(3534) + \cdots + \beta_{136} f_{\rm tot}(8472) \\
& + \beta_{137} f_{\rm cnt}(3512) + \beta_{138} f_{\rm cnt}(3534) + \cdots + \beta_{270} f_{\rm cnt}(8472) \\
& + \beta_{271} \mathcal{R}(3780/4580) + \beta_{272} \mathcal{R}(4610/4260) \\
& + \beta_{273} \mathcal{R}(5690/5360) + \beta_{274} \mathcal{R}(6420/4430) \\
& + \beta_{275} \mathcal{R}(6420/5290) + \beta_{276} \mathcal{R}(6630/4400) + e. \qquad (6)
\end{aligned}
$$

Non-zero elements	coefficients β	p
c	0.376	1.00
$f_{\rm tot}(6373)$	0.100	1.00
x_1	−0.050	0.98
$f_{\rm cnt}(6084)$	−0.034	0.98
$f_{\rm cnt}(6289)$	−0.045	0.95
$f_{\rm cnt}(6631)$	−0.061	0.80
$\mathcal{R}(3780/4580)$	−0.050	0.74
$f_{\rm tot}(3752)$	0.063	0.73

- 78 samples, 276 variables
- Solving using LASSO

- Some spectroscopic variables
  - Independent of the color and decay rate?
  - Just due to high correlations?



Result 2/3: using corrected luminosity

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

$$\begin{aligned} M_B = M_{B,0} &+ \beta_1 c + \beta_2 x_1 \\ &+ \beta_3 f_{\rm tot}(3512) + \beta_4 f_{\rm tot}(3534) + \cdots + \beta_{136} f_{\rm tot}(8472) \\ &+ \beta_{137} f_{\rm cnt}(3512) + \beta_{138} f_{\rm cnt}(3534) + \cdots + \beta_{270} f_{\rm cnt}(8472) \\ &+ \beta_{271} \mathcal{R}(3780/4580) + \beta_{272} \mathcal{R}(4610/4260) \\ &+ \beta_{273} \mathcal{R}(5690/5360) + \beta_{274} \mathcal{R}(6420/4430) \\ &+ \beta_{275} \mathcal{R}(6420/5290) + \beta_{276} \mathcal{R}(6630/4400) + e. \end{aligned} \quad (6)$$

Non-zero elements	coefficients $\boldsymbol{\beta}$	p
x_1	-0.020	0.99

Decay rate

- The decay rate, only
- The other candidates in the last model disappear.

Result 3/3: final result

$$
\begin{aligned}
M_B = M_{B,0} &+ \beta_1 c + \beta_2 x_1 \\
&+ \beta_3 f_{\mathrm{tot}}(3512) + \beta_4 f_{\mathrm{tot}}(3534) + \cdots + \beta_{136} f_{\mathrm{tot}}(8472) \\
&+ \beta_{137} f_{\mathrm{cnt}}(3512) + \beta_{138} f_{\mathrm{cnt}}(3534) + \cdots + \beta_{270} f_{\mathrm{cnt}}(8472) \\
&+ \beta_{271} \mathcal{R}(3780/4580) + \beta_{272} \mathcal{R}(4610/4260) \\
&+ \beta_{273} \mathcal{R}(5690/5360) + \beta_{274} \mathcal{R}(6420/4430) \\
&+ \beta_{275} \mathcal{R}(6420/5290) + \beta_{276} \mathcal{R}(6630/4400) + e. \qquad (6)
\end{aligned}
$$

- No variables have non-zero coefficient.

Short summary

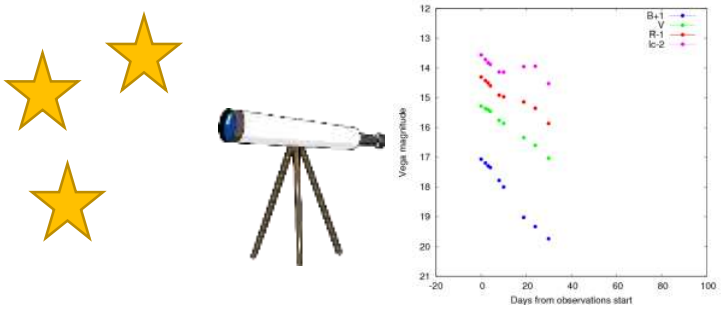
- We estimated the best set of variables for the luminosity of supernovae using LASSO.
- We reduced the number of candidate variables by using the normalized spectra instead of flux ratios.
- Our result supports the classical picture, that is, the color and decay rate is the best set of variables, and does not support to add any other variables.

- Useful framework in future when the data size further increases.

Outline

Past

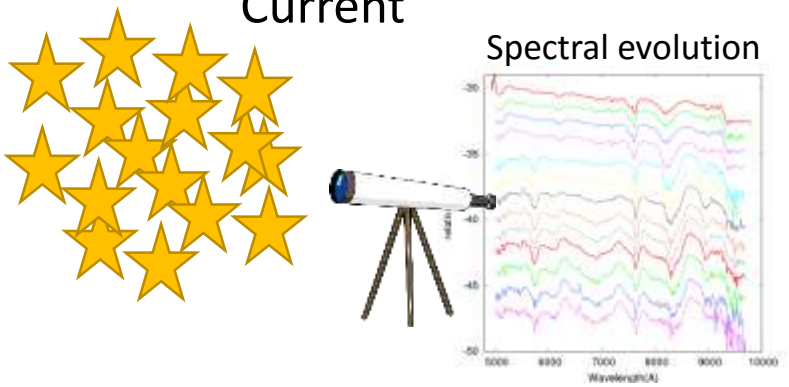
- Magnitude
- Color



Analysis based on the experience of domain experts.

Current

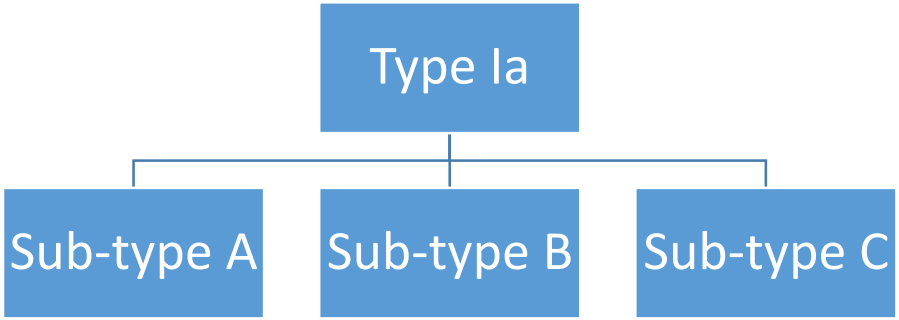
Spectral evolution



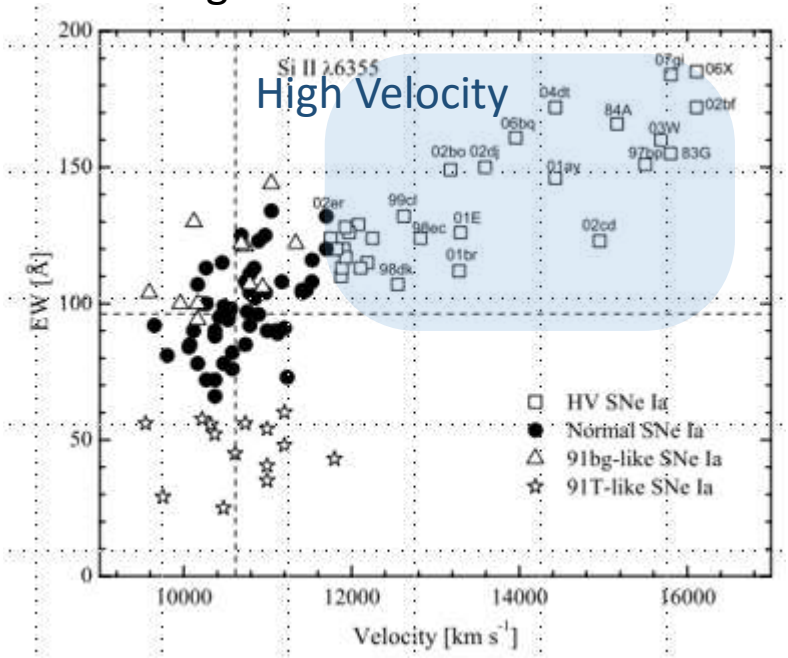
Analysis based on the data-driven method is required.

- General introduction about supernovae
- Our recent works
  - Variable selection for the peak luminosity using LASSO
  - Visual analytics for classification

Classification of supernovae



Wang et al. 2009

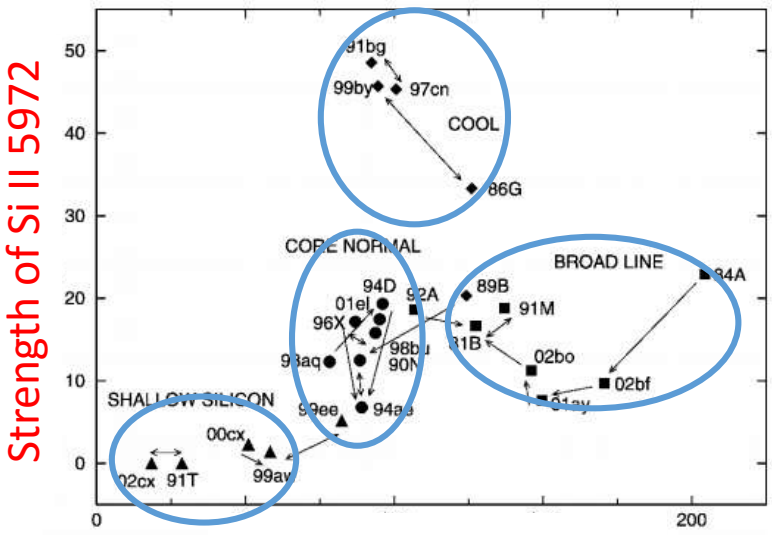


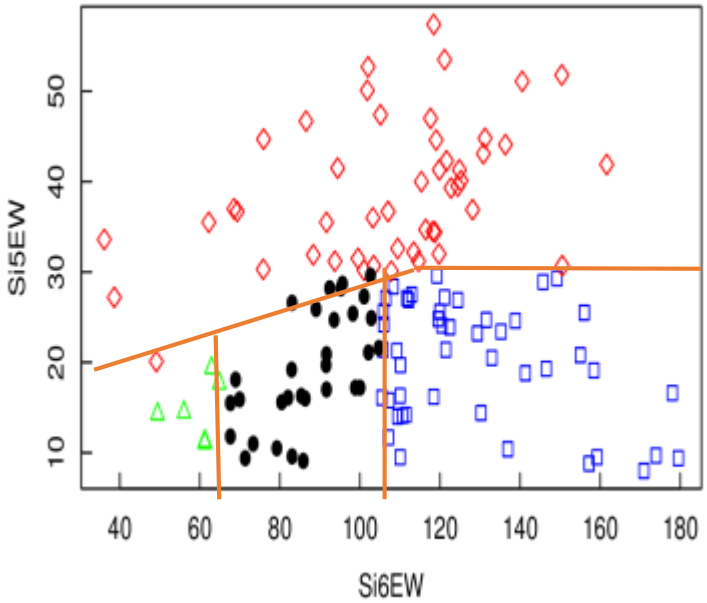
- Wang et al. 2009: The high velocity group has a different color behavior from the ordinary ones.
- Different color correction

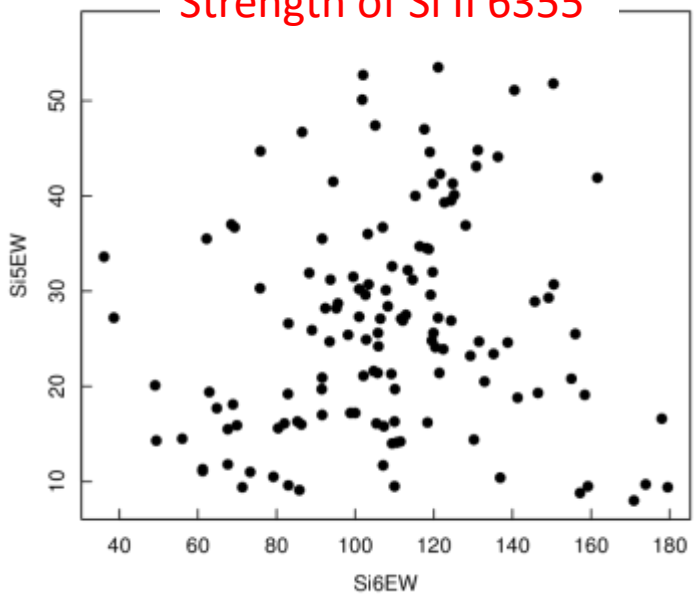
  → more accurate distance

Classification scheme proposed in Branch+06

- Based on spectra.
- A standard classification scheme.







Problem

Velocity | Si II 6355 | color

Line strength | Si II 5972 | luminosity

Line width | Si II 4000 | Decay rate

Line depth | Ca II, etc… | distance

Which are good for classification ?

- We don't know whether the sub-groups really exist.
- Which set of variables is good for classification?

- Difficult to "see" the data structure with eyes because candidates of variables are so many.

➔ Visual analytics approach

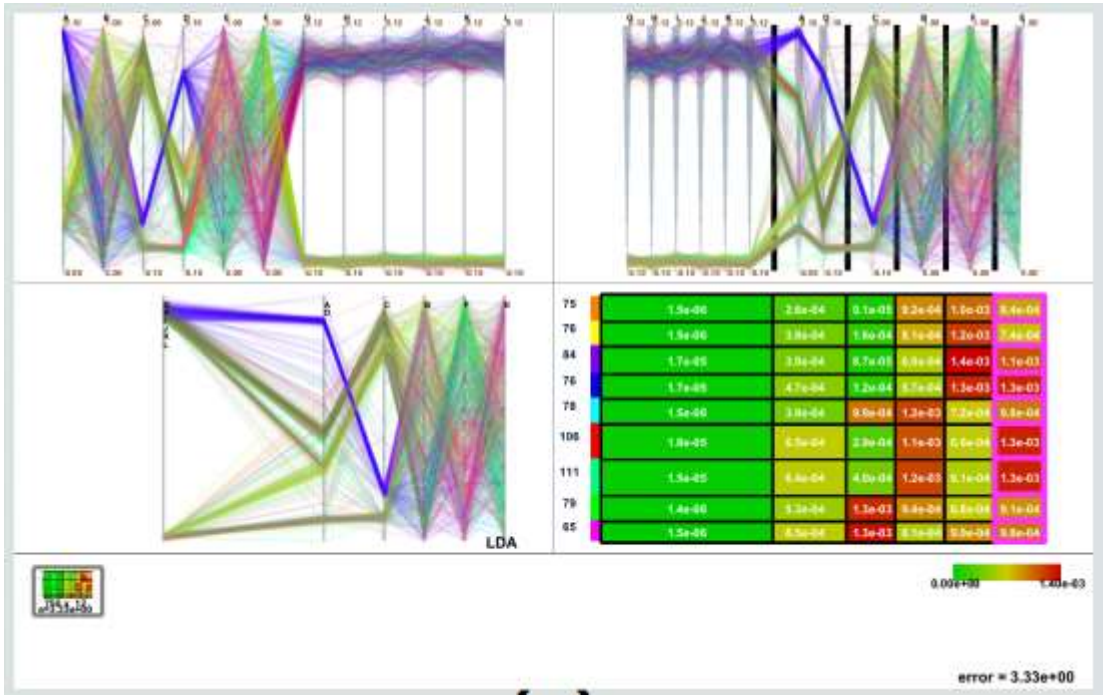
The ABC tool

- Watanabe, et al., IEEE Pacific Visualization Symposium 2015

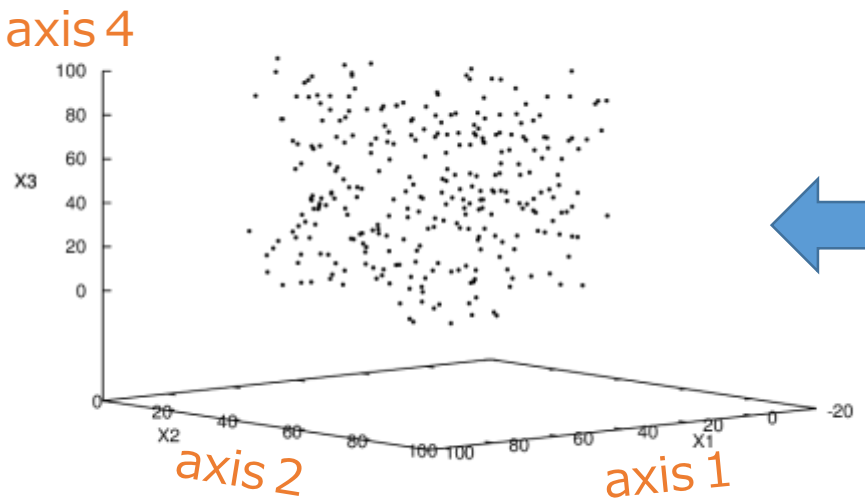
K-means clustering

visualization

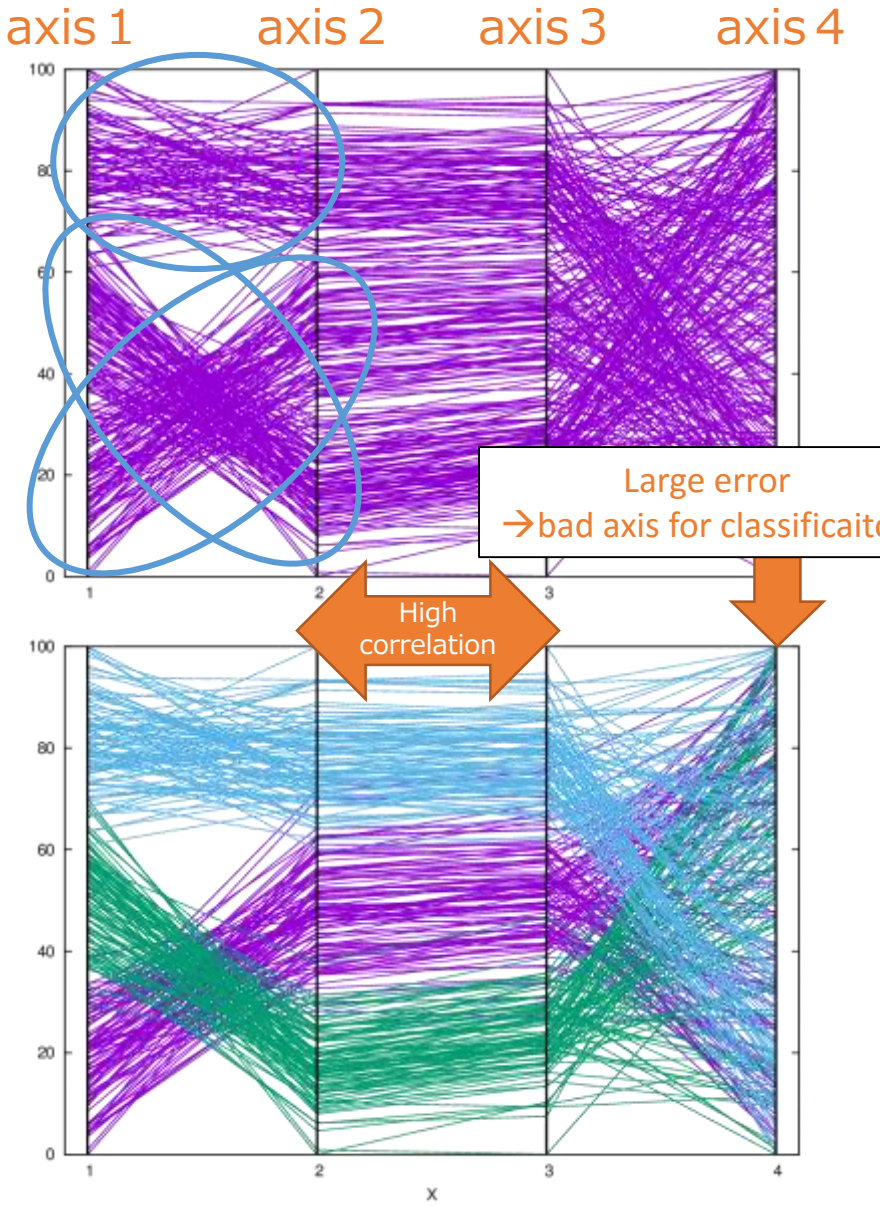
Feed back



Parallel coordinate plot



Scatter plot

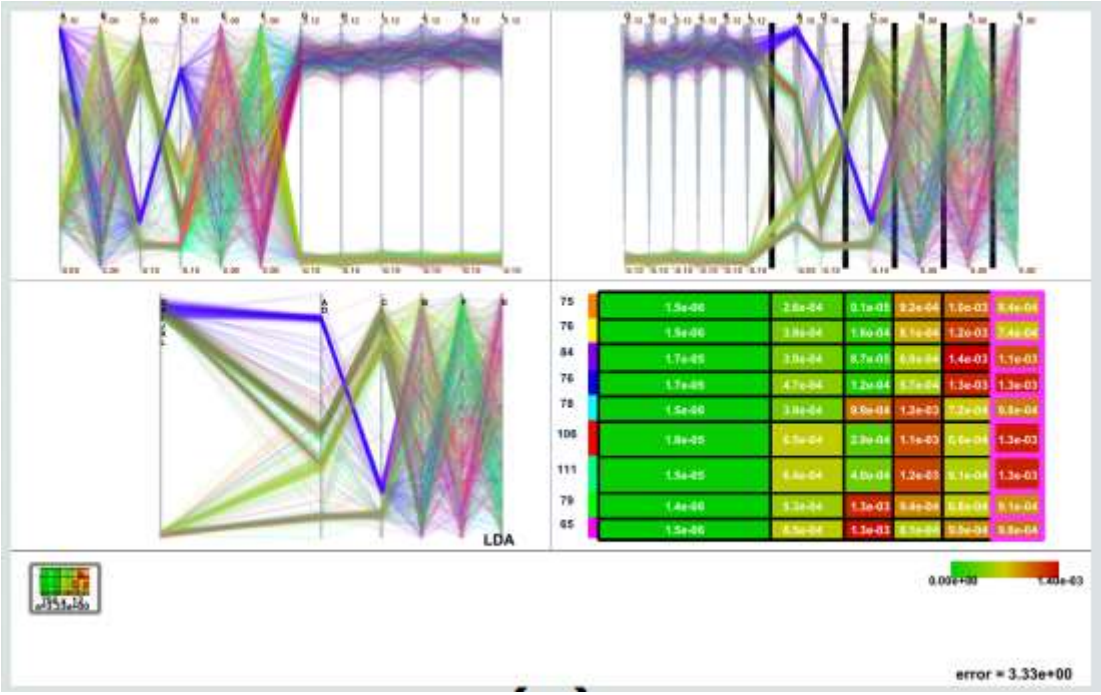


- Even in the same data, some structures can be seen only in the parallel coordinate plot.

Analysis with the ABC tool

- Goal: to find sub-groups and a set of axes for the classification
- Process: delete axes having large errors
- Data: from Berkeley supernova database. 132 samples, 14 variables

  (luminosity, distance, color, strengths of absorption lines, velocity, etc…)



Demo.

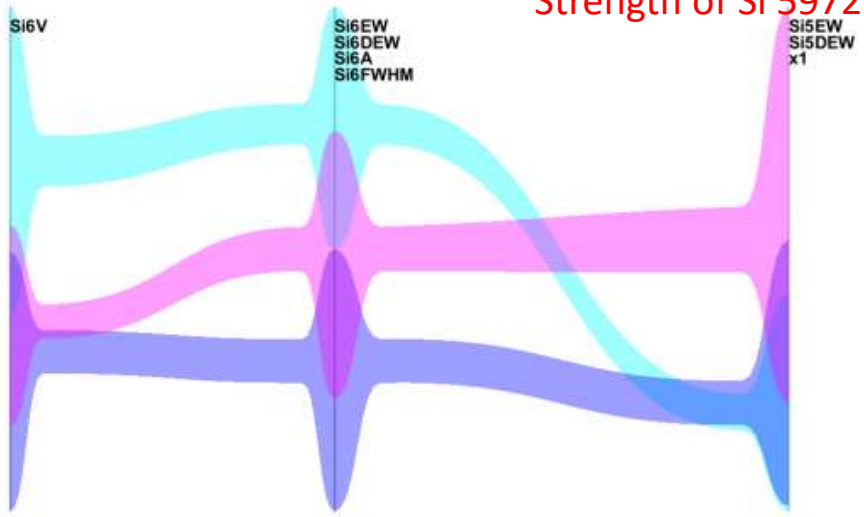
Velocity of Si 6355 Strength of Si 6355

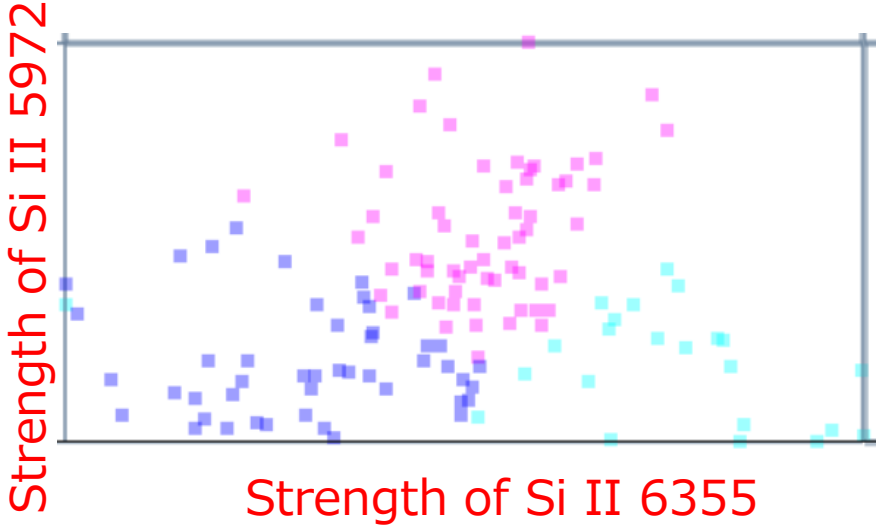
Strength of Si 5972

Result

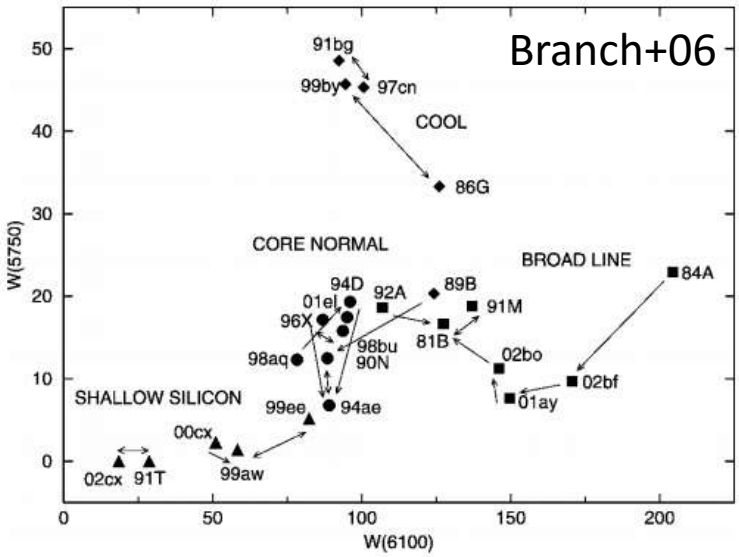
SN Ia

- Expansion velocity ≳12,000 km/s → Si II 6355 ∝ Si II 5972
- Expansion velocity ≲12,000 km/s → Si II 6355 anti-correlation Si II 5972





- Con
  -
  -



Conclusions

- Variable selection for the peak luminosity of supernovae using LASSO
  - Reducing the candidates of explanatory variables
  - Confirming the past classical model.
- Classification of supernovae using visual analytics tool
  - Confirming the past classification scheme.

➔Demonstrating that the data-driven approach provides consistent results to our past understandings

➔In the big-data era near future, those methods would be standard for supernova science.