Probing Galaxy Properties through Next Generation Data Mining

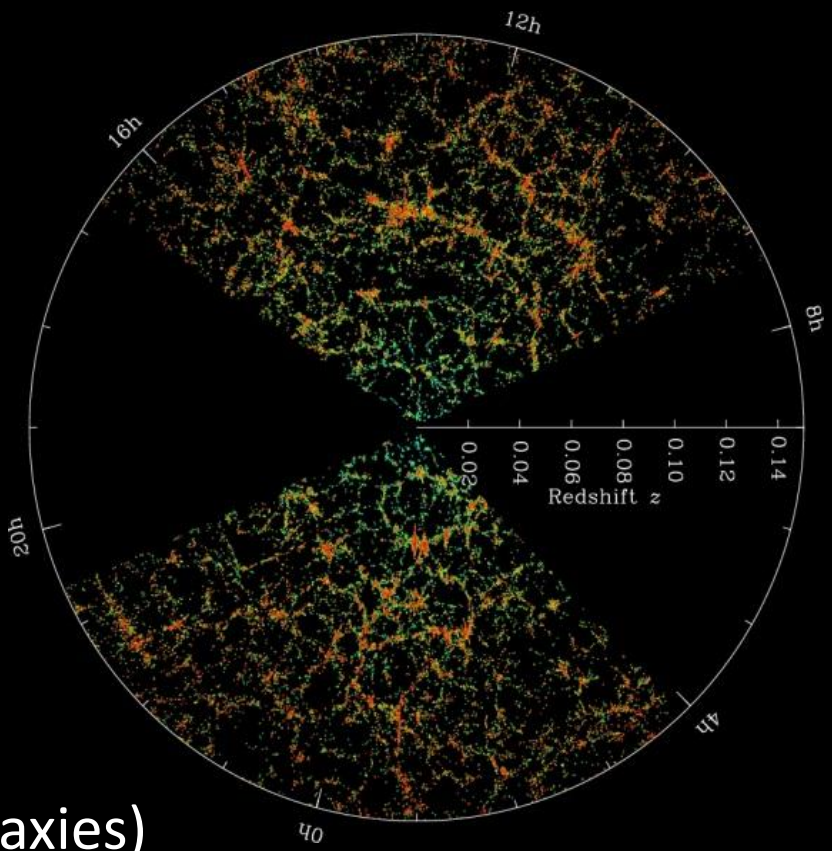
Ching-Wa Yip

JHU

(Michael Mahoney @ Stanford)

Challenges in Probing Galaxies' Properties

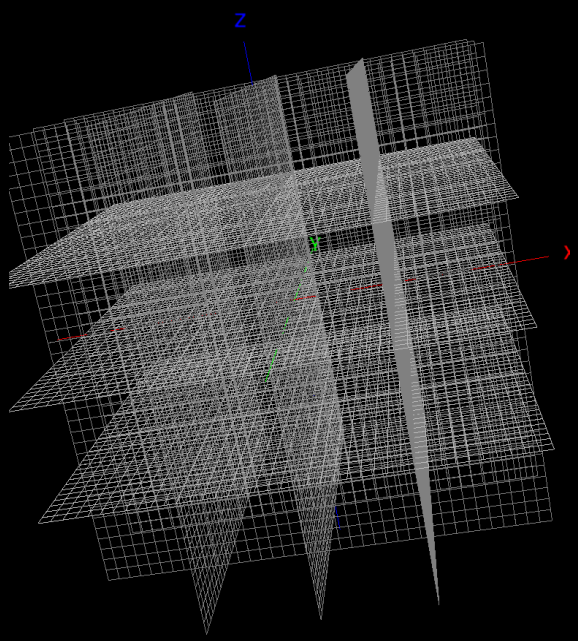
- Many Objects (Big Data)
- Many Parameters
- Noisy Data



SDSS & BOSS (z = 0 – 0.7, 2.5M galaxies)
LAMOST (z = 0 – 0.2, 10M galaxies)
Prime Focus Spectrograph (z = 1 – 2, 200K galaxies)

Probing Galaxies' Properties (Many Parameters)

- Model Fitting of Spectra (SDSS/JHU Catalog: Kauffmann, Heckman, Tremonti, Chen, et al.)
- Empirical Relations



N-Dimensional Model Grid (Yip 2010)

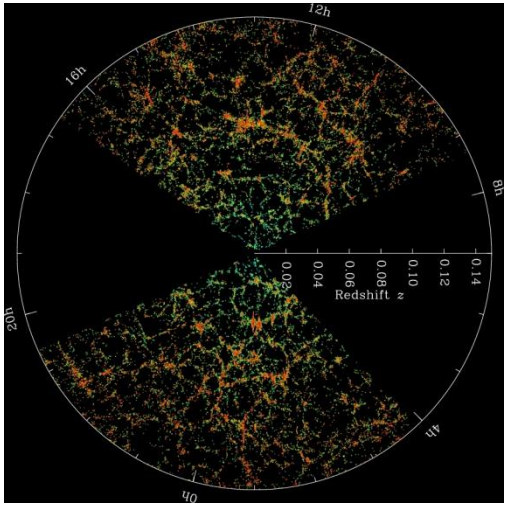
1 mag (g band) effect!

Extinction in SDSS Disks from Inclination Effect (Yip, Szalay, Wyse et al. 2010, 2011)

Probing Galaxies' Properties (Many Data)

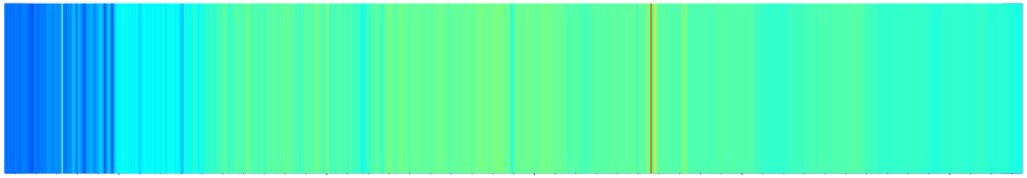
Galaxy Parameter Estimation:

- Stellar Age
- Stellar Metallicity
- Star Formation Rate
- Dust Extinction
- etc.



Brute force Bayesian approach:

N = 1 Million	Galaxy Spectra
M = 4,000	Pixels
P = 4	Parameters
T = 20 years	CPU time



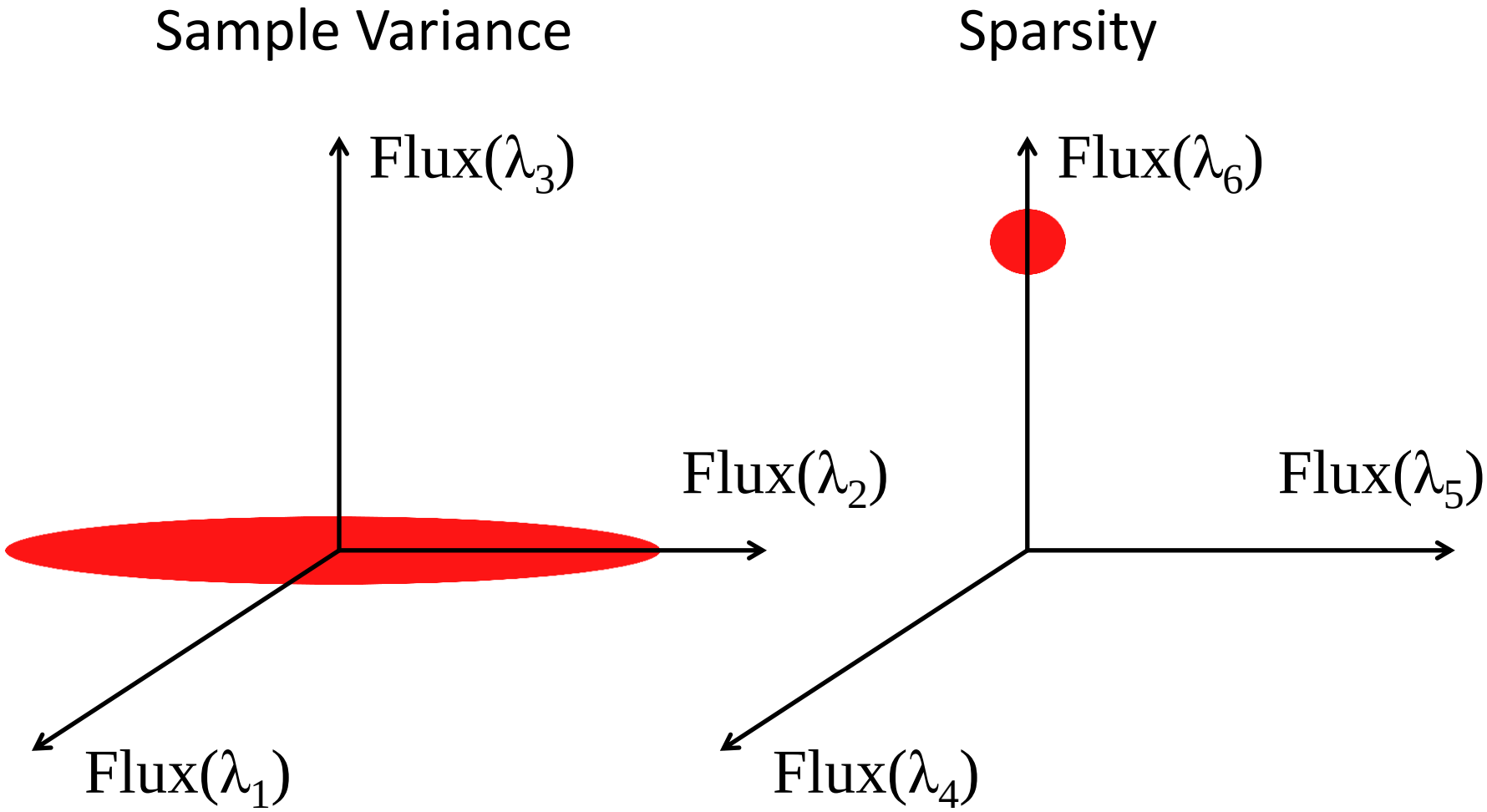
Lick Indices

(Worthey et al. 94; Trager et al. 98)

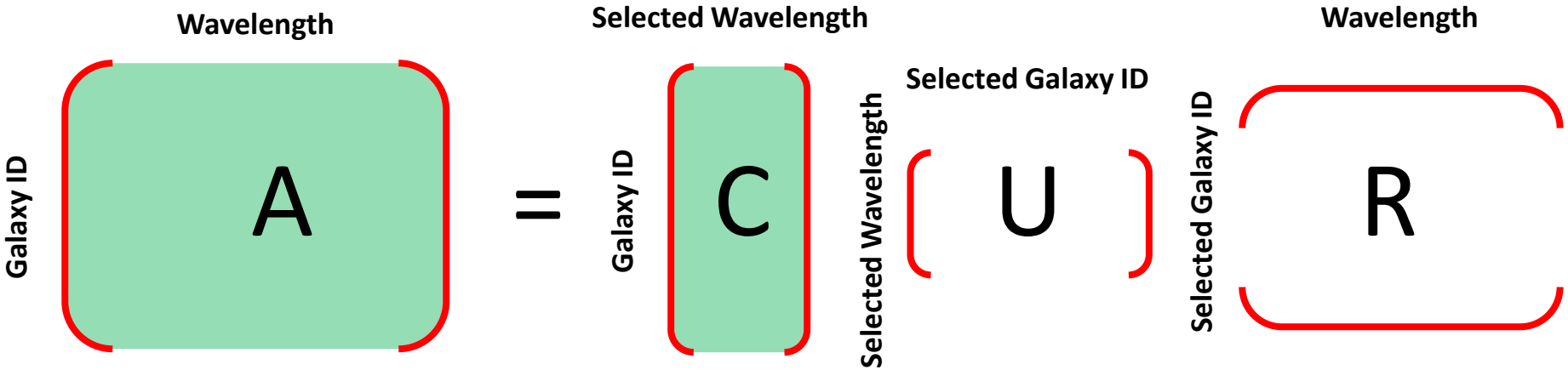
INDEX DEFINITIONS

	Name	Index Bandpass	Pseudocontinua	Units	Measures	Error[1]	Notes
01	CN_1	4143.375-4178.375	4081.375-4118.875 4245.375-4285.375	mag	CN, Fe I	0.021	
02	CN_2	4143.375-4178.375	4085.125-4097.625 4245.375-4285.375	mag	CN, Fe I	0.023	2
03	Ca4227	4223.500-4236.000	4212.250-4221.000 4242.250-4252.250	Å	Ca I, Fe I, Fe II	0.27	2
04	G4300	4282.625-4317.625	4267.625-4283.875 4320.125-4336.375	Å	CH, Fe I	0.39	
05	Fe4383	4370.375-4421.625	4360.375-4371.625 4444.125-4456.625	Å	Fe I, Ti II	0.53	2
06	Ca4455	4453.375-4475.875	4447.125-4455.875 4478.375-4493.375	Å	Ca I, Fe I, Ni I, Ti II, Mn I, V I	0.25	2
07	Fe4531	4515.500-4560.500	4505.500-4515.500 4561.750-4580.500	Å	Fe I, Ti I, Fe II, Ti II	0.42	2
08	Fe4668	4635.250-4721.500	4612.750-4631.500 4744.000-4757.750	Å	Fe I, Ti I, Cr I, Mg I, Ni I, C_2	0.64	2
09	Hβ	4847.875-4876.625	4827.875-4847.875 4876.625-4891.625	Å	Hβ, Fe I	0.22	3
10	Fe5015	4977.750-5054.000	4946.500-4977.750 5054.000-5065.250	Å	Fe I, Ni I, Ti I	0.46	2,3
11	Mg_1	5069.125-5134.125	4895.125-4957.625 5301.125-5366.125	mag	MgH, Fe I, Ni I	0.007	3
12	Mg_2	5154.125-5196.625	4895.125-4957.625 5301.125-5366.125	mag	MgH, Mg *b*, Fe I	0.008	3
13	Mg *b*	5160.125-5192.625	5142.625-5161.375 5191.375-5206.375	Å	Mg *b*	0.23	3
14	Fe5270	5245.650-5285.650	5233.150-5248.150 5285.650-5318.150	Å	Fe I, Ca I	0.28	3
15	Fe5335	5312.125-5352.125	5304.625-5315.875 5353.375-5363.375	Å	Fe I	0.26	3
16	Fe5406	5387.500-5415.000	5376.250-5387.500 5415.000-5425.000	Å	Fe I, Cr I	0.20	2,3
17	Fe5709	5698.375-5722.125	5674.625-5698.375 5724.625-5738.375	Å	Fe I, Ni I, Mg I Cr I, V I	0.18	2
18	Fe5782	5778.375-5798.375	5767.125-5777.125 5799.625-5813.375	Å	Fe I, Cr I Cu I, Mg I	0.20	2
19	Na D	5878.625-5911.125	5862.375-5877.375 5923.875-5949.875	Å	Na I	0.24	
20	TiO_1	5938.375-5995.875	5818.375-5850.875 6040.375-6105.375	mag	TiO	0.007	
21	TiO_2	6191.375-6273.875	6068.375-6143.375 6374.375-6416.875	mag	TiO	0.006	

Some Wavelengths are More Informative: Leverage Score



CUR Matrix Decomposition (Mahoney & Drineas 2009): Best Rank-k Approximation of Data Matrix



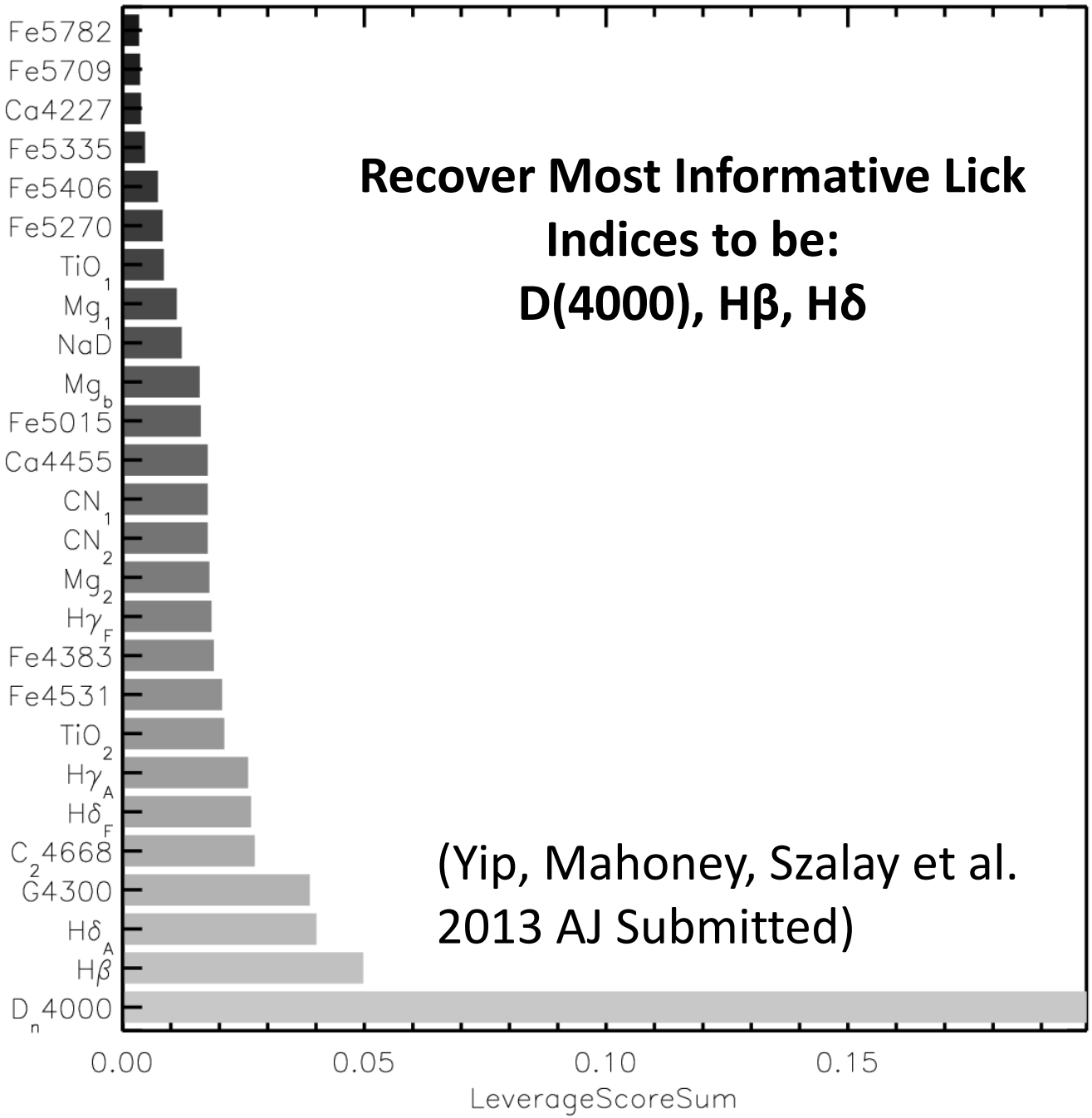
CUR approximates data matrix A:

$$\min \| A - CUR \|_F$$

Identifying Objective Wavelength Regions using CUR Matrix Decomposition

- Simple Stellar Populations (SSP) defined by stellar age (0 – 20 Gyr) and stellar metallicity (Z = 0.0001, 0.0004, 0.004, 0.008, 0.02, 0.05)
  - Bruzual and Charlot 2003
  - Absorption Lines Only, No Emission Lines

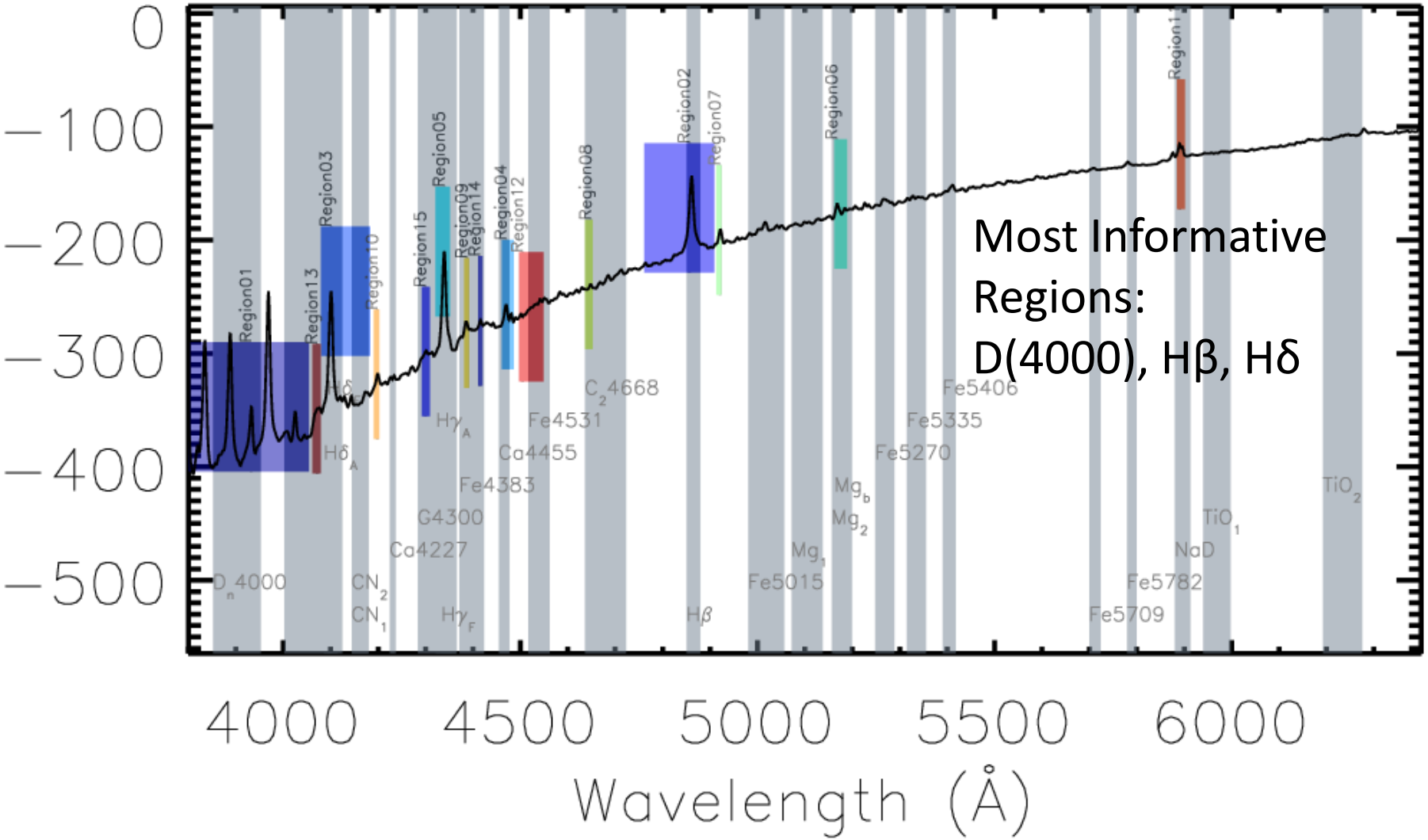
- 2 configurations
  - Lick: 3800-6400 Å with 9 Å resolution
  - SDSS: 3450-8350 Å with 3 Å resolution



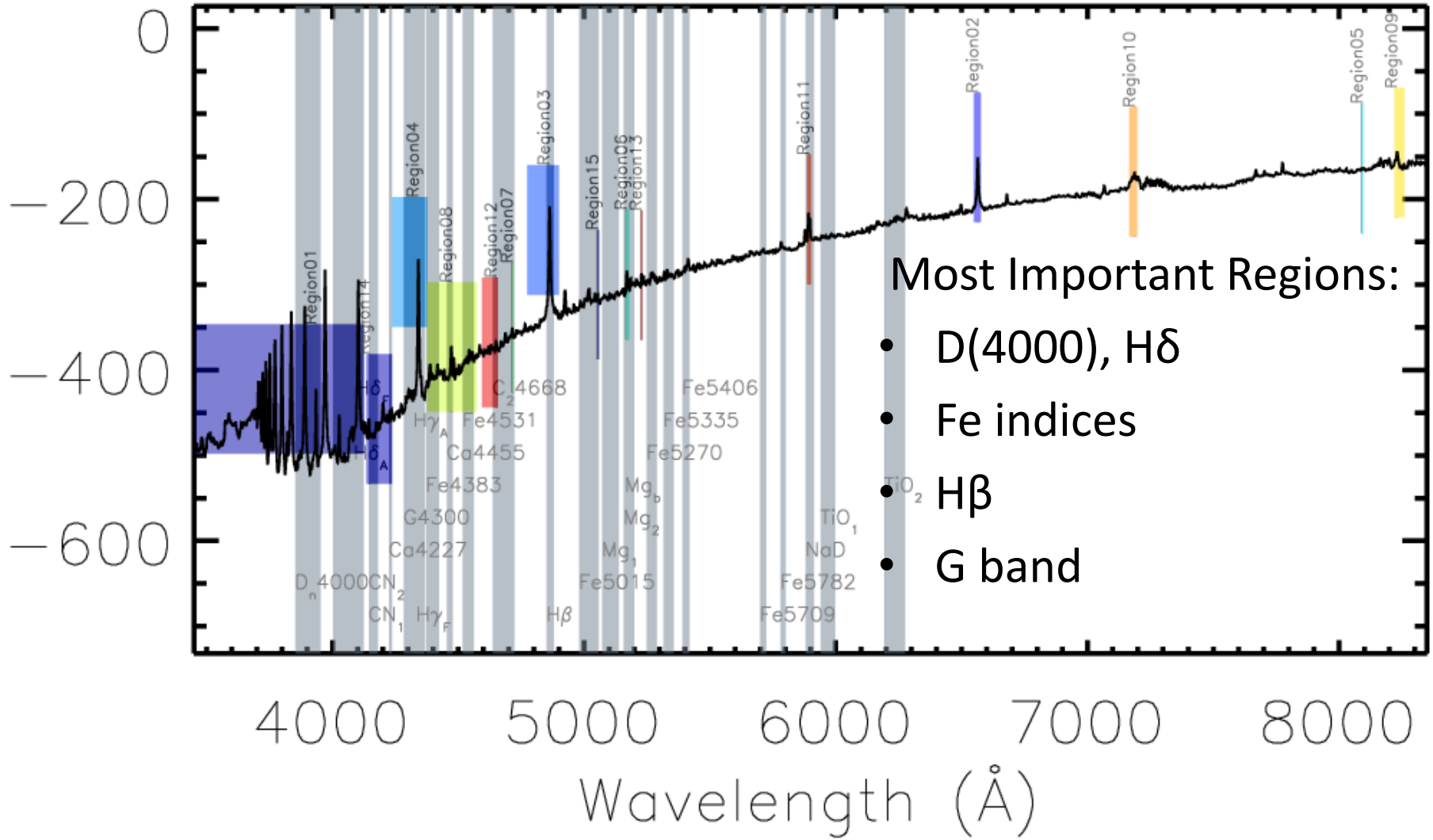
Recover Most Informative Lick Indices to be: D(4000), Hβ, Hδ

(Yip, Mahoney, Szalay et al. 2013 AJ Submitted)

9Å Resolution



3Å Resolution



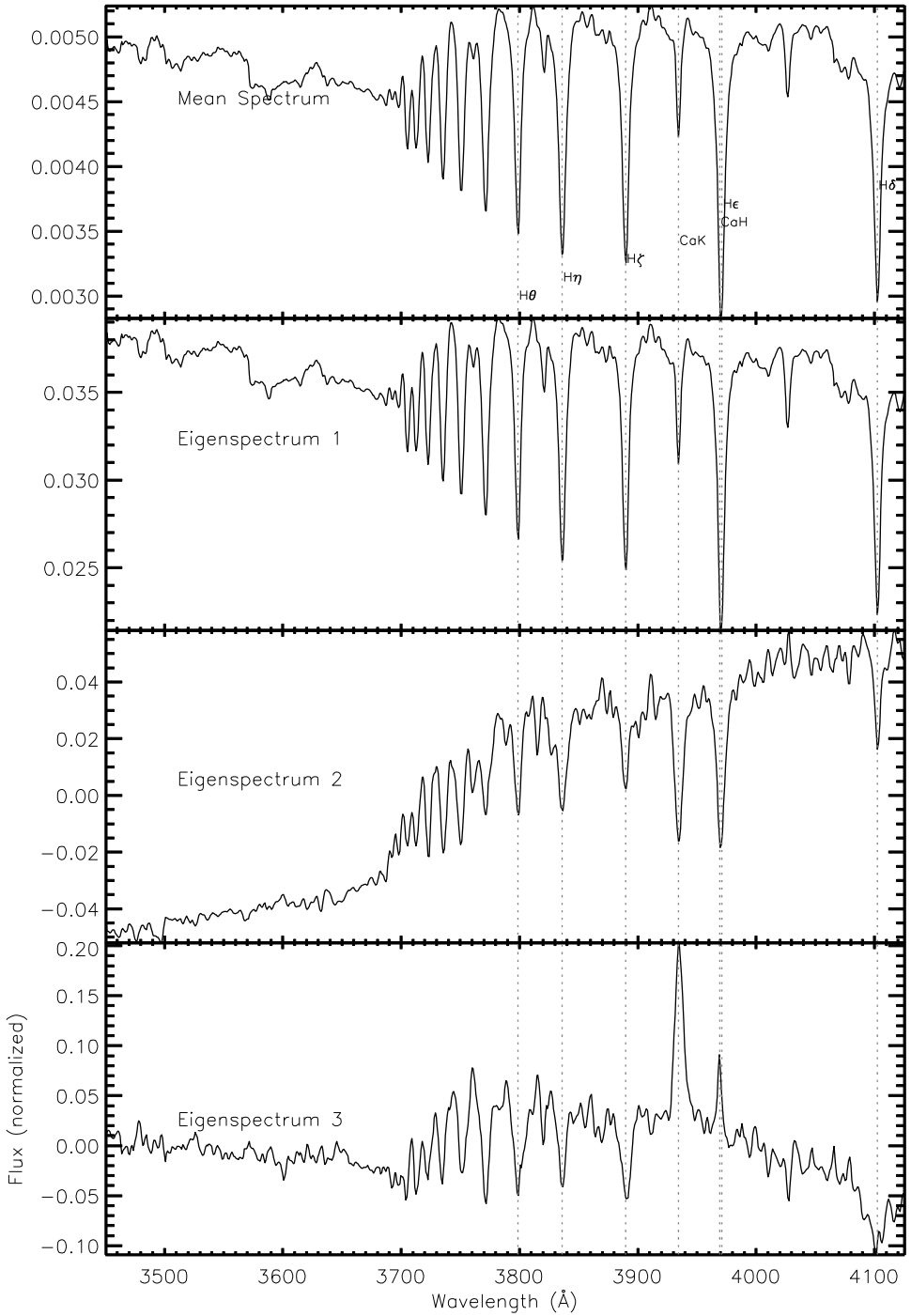
Objective 3Å Absorption Indices

(Yip et al. 2013 AJ submitted)

Table 3: Informative wavelength regions, for wavelength range 3450–8350 Å at 3 Å spectral resolution.

Region	λ_S	λ_I	λ_E	$\lambda_E - \lambda_S$	$\lambda_I - \lambda_S$	Symmetric?	$\sum p_\lambda$	Lick Overlaps	1st mode	$u^1 \cdot$ par	2nd mode	$u^2 \cdot$ par
Region01	3450.8	3934.5	4125.6	674.8	483.8	no	0.403	$H\delta_A$;$H\delta_F$;$D_n(4000)$L;$D_n(4000)$R;	AGE	-0.390	METAL	-0.079
Region08	4378.0	4472.7	4574.7	196.8	94.7	no	0.063	Fe4383;Ca4455;Fe4531;	AGE	-0.397	METAL	-0.137
Region03	4773.7	4861.3	4899.5	125.8	87.6	no	0.056	$H\beta$;	AGE	-0.393	METAL	-0.171
Region04	4238.2	4340.9	4377.0	138.8	102.7	no	0.049	G4300;Fe4383;$H\gamma_A$;$H\gamma_F$;	AGE	-0.396	METAL	-0.177
Region14	4135.1	4144.6	4237.2	102.1	9.5	no	0.028	CN_1;CN_2;Ca4227;	AGE	-0.398	METAL	-0.220
Region12	4594.8	4651.2	4656.5	61.7	56.4	no	0.015	$C_2$4668;	AGE	-0.395	METAL	-0.255
Region02	6549.4	6564.5	6576.6	27.2	15.1	no	0.012	···	AGE	-0.371	METAL	0.253
Region09	8219.7	8229.2	8259.5	39.8	9.5	no	0.009	···	AGE	-0.370	METAL	0.290
Region10	7166.1	7187.6	7197.5	31.4	21.5	no	0.008	···	AGE	-0.375	METAL	0.305
Region06	5165.9	5168.2	5179.0	13.1	2.4	no	0.006	Mg_2;Mg_b;	AGE	-0.388	METAL	0.157
Region11	5887.3	5892.8	5902.3	14.9	5.4	no	0.006	NaD;	AGE	-0.374	METAL	0.093
Region19	4914.2	4924.4	4936.9	22.7	10.2	no	0.006	···	AGE	-0.393	METAL	-0.388
Region18	8159.4	8178.2	8182.0	22.6	18.8	no	0.005	···	AGE	-0.373	METAL	0.301
Region17	4681.3	4686.6	4694.2	12.9	5.4	no	0.004	$C_2$4668;	AGE	-0.393	AGE	0.035
Region07	4712.6	4714.8	4720.2	7.6	2.2	no	0.003	$C_2$4668;	AGE	-0.393	METAL	0.428
Region20	5180.2	5183.7	5189.7	9.5	3.6	no	0.003	Mg_2;Mg_b;	AGE	-0.387	METAL	0.164
Region13	5224.5	5226.9	5231.7	7.2	2.4	no	0.003	···	AGE	-0.387	METAL	0.121
Region05	8090.2	8092.1	8097.7	7.5	1.9	no	0.003	···	AGE	-0.364	METAL	-0.121
Region16	7274.1	7277.4	7279.1	5.0	3.3	no	0.002	···	AGE	-0.375	METAL	0.335
Region15	5053.0	5054.1	5056.5	3.5	1.2	no	0.002	Fe5015;	AGE	-0.391	METAL	-0.143

See footnote in Table 2. The extra horizontal line divides regions with *LeverageScoreSum* less than and larger than 0.03.

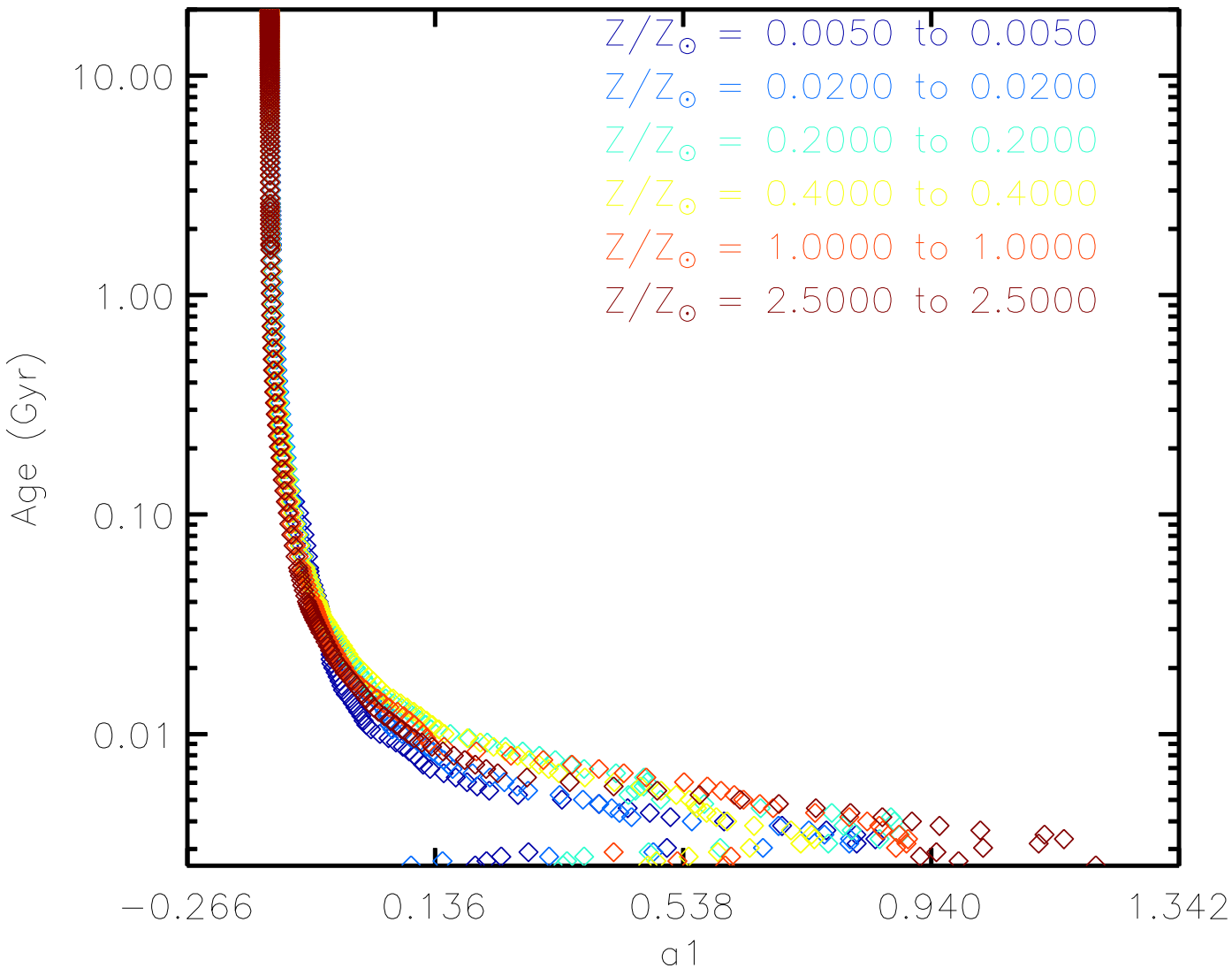


PCA Eigenspectra of 1st Informative Region

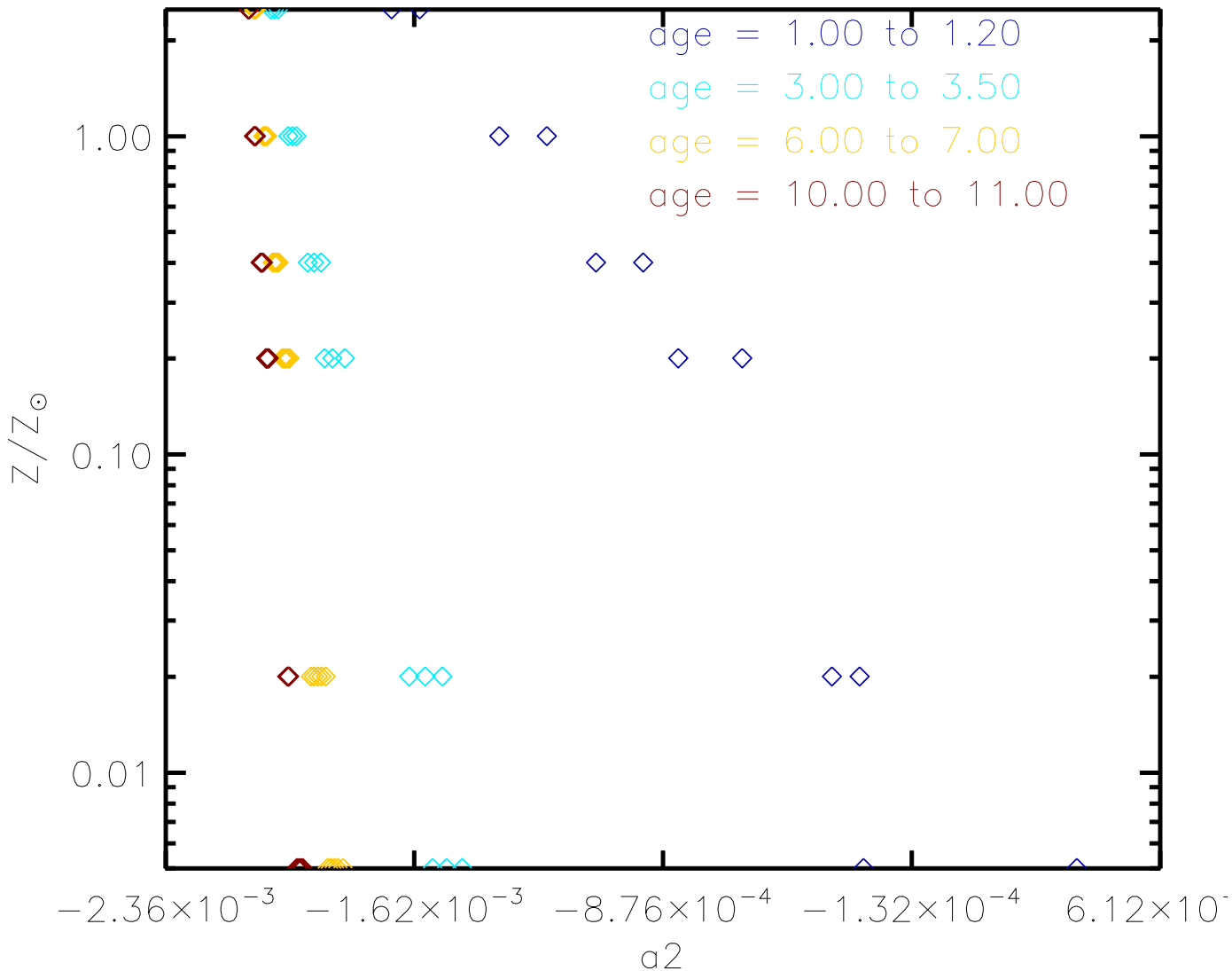
Larger a1 implies:
Younger stellar populations

Larger a2 implies:
Younger, More Metal Rich
OR
Older, Less Metal Rich

Stellar Age Sensitivity of Wavelength Regions

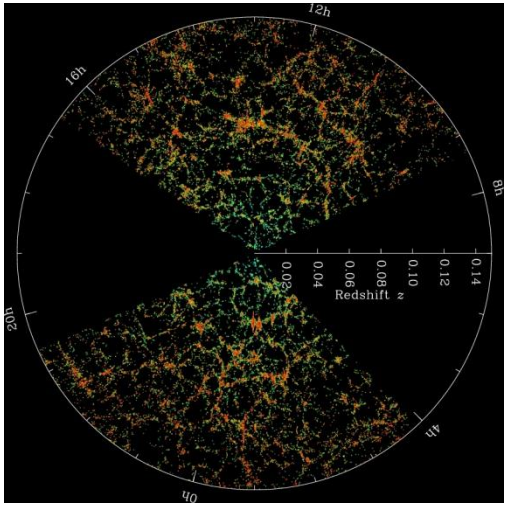


Stellar Metallicity Sensitivity of Wavelength Regions (Seeing Age-Metallicity Degeneracy!)



Galaxy Parameter Estimation:

- Stellar Age
- Stellar Metallicity
- Star Formation Rate
- Dust Extinction
- etc.



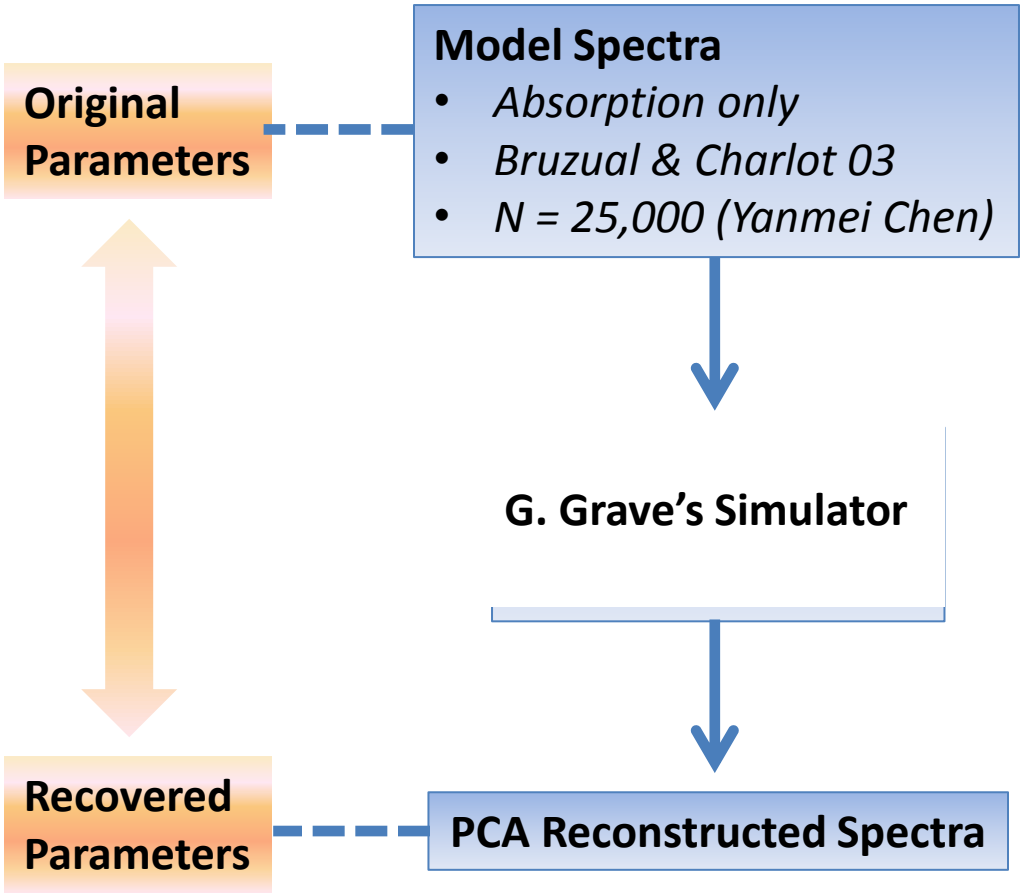
Brute force Bayesian approach:

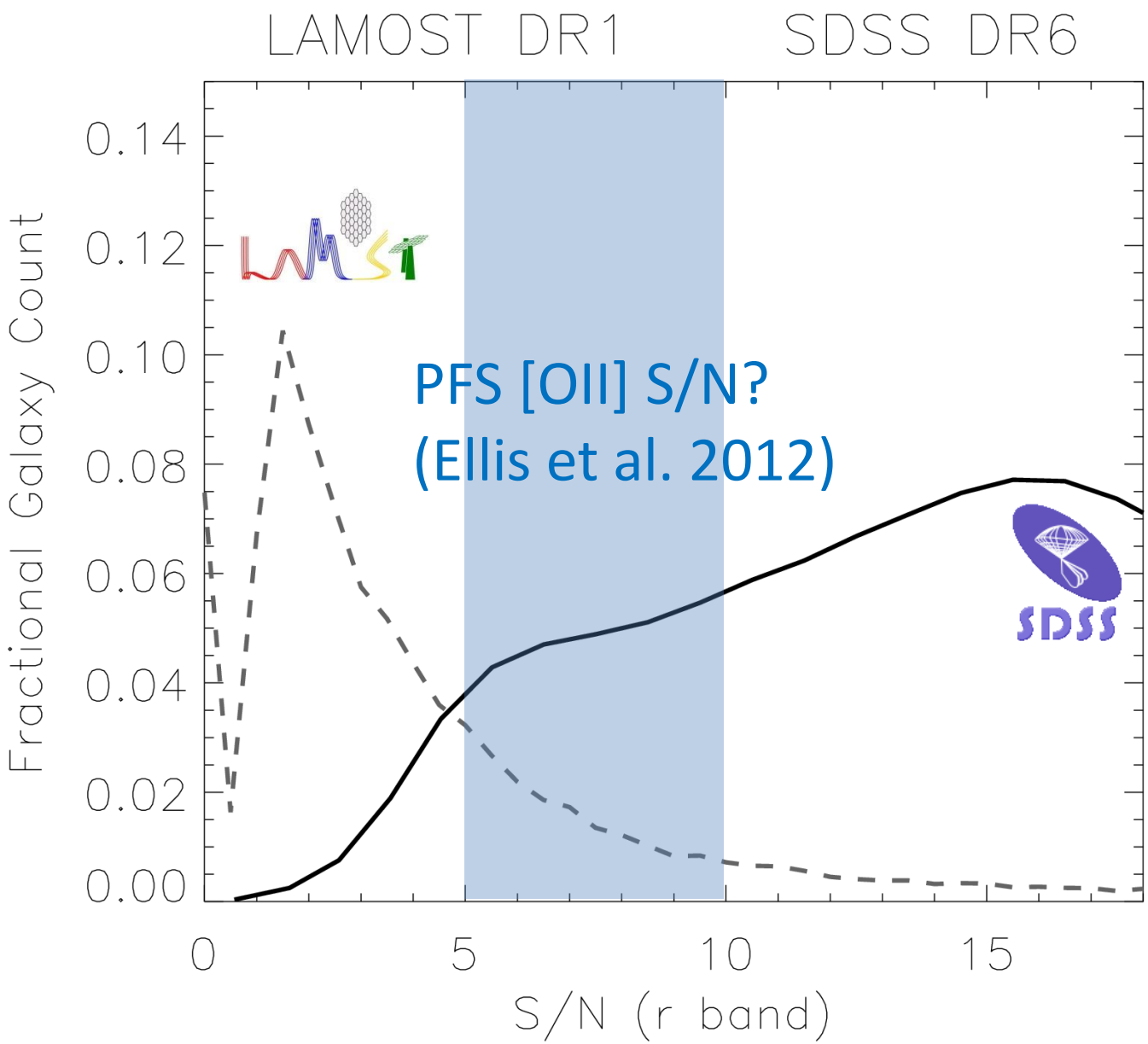
N = 1 Million	Galaxy Spectra
M = 4,000	Pixels
P = 4	Parameters
T = 20 years	CPU time

2 weeks!

Galaxy Parameters from Prime Focus Spectrograph (PFS) Spectra

- 25,000 model spectra simulation
- Realistic star formation history
- 5 model parameters
- 18 derived parameters
- Focus on $D_n(4000)$



(Yip & Heckman, et al.)

Noisy Astronomy Data