Data Mining In Modern Astronomy Sky Surveys

Ching-Wa Yip
cwyip@pha.jhu.edu; Bloomberg 518

Schedule: TTh 4-6:30pm (Jan 6-24); Bloomberg 274
Office Hours: Wed 2-4pm; And by appointments

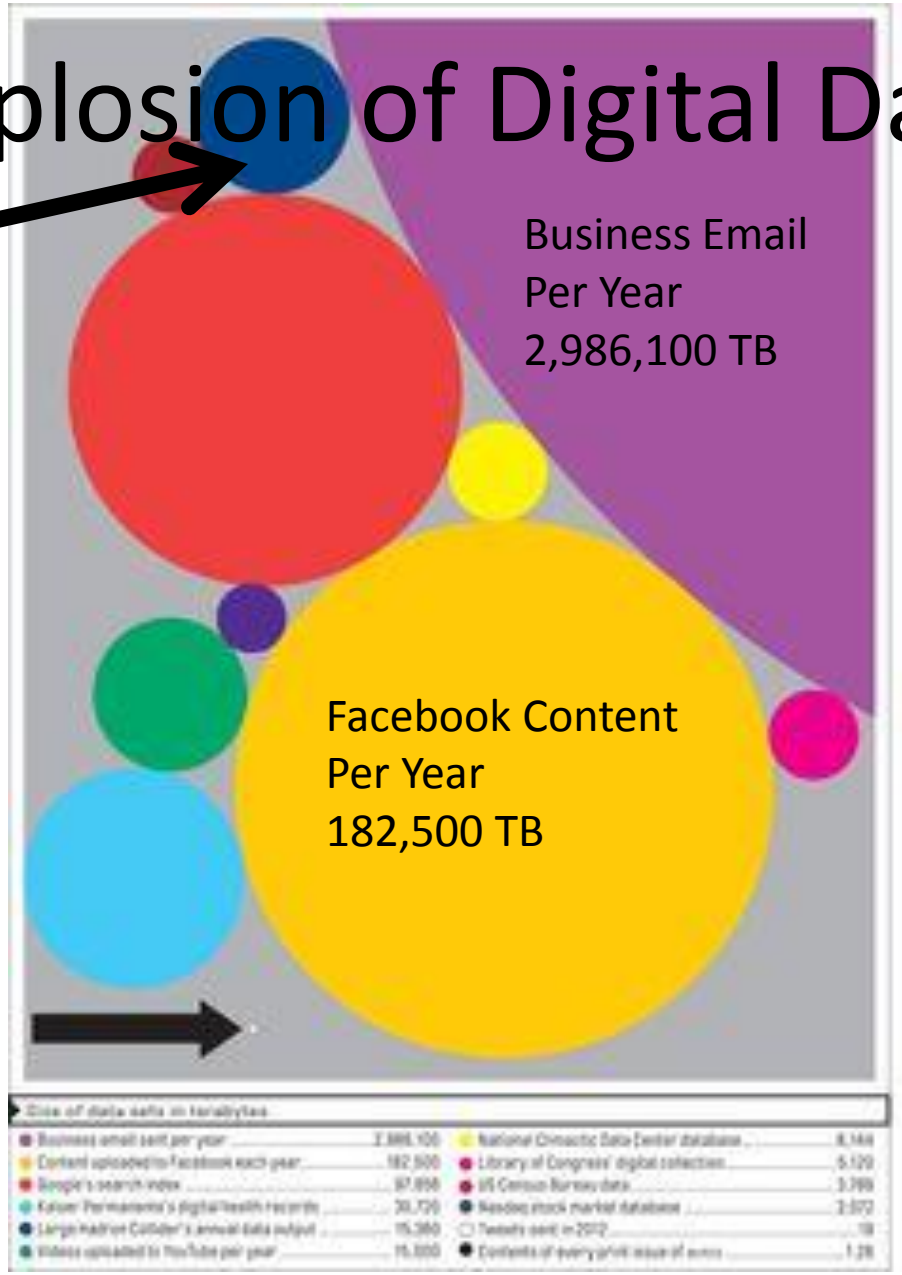
Feb 25, 2010

"How to make sense of all these data? People should be worried about how we train the next generation, not just of scientists, but people in government and industry."

- Alex Szalay @ JHU

Explosion of Digital Data

Large Hadron Collider
15,360 TB



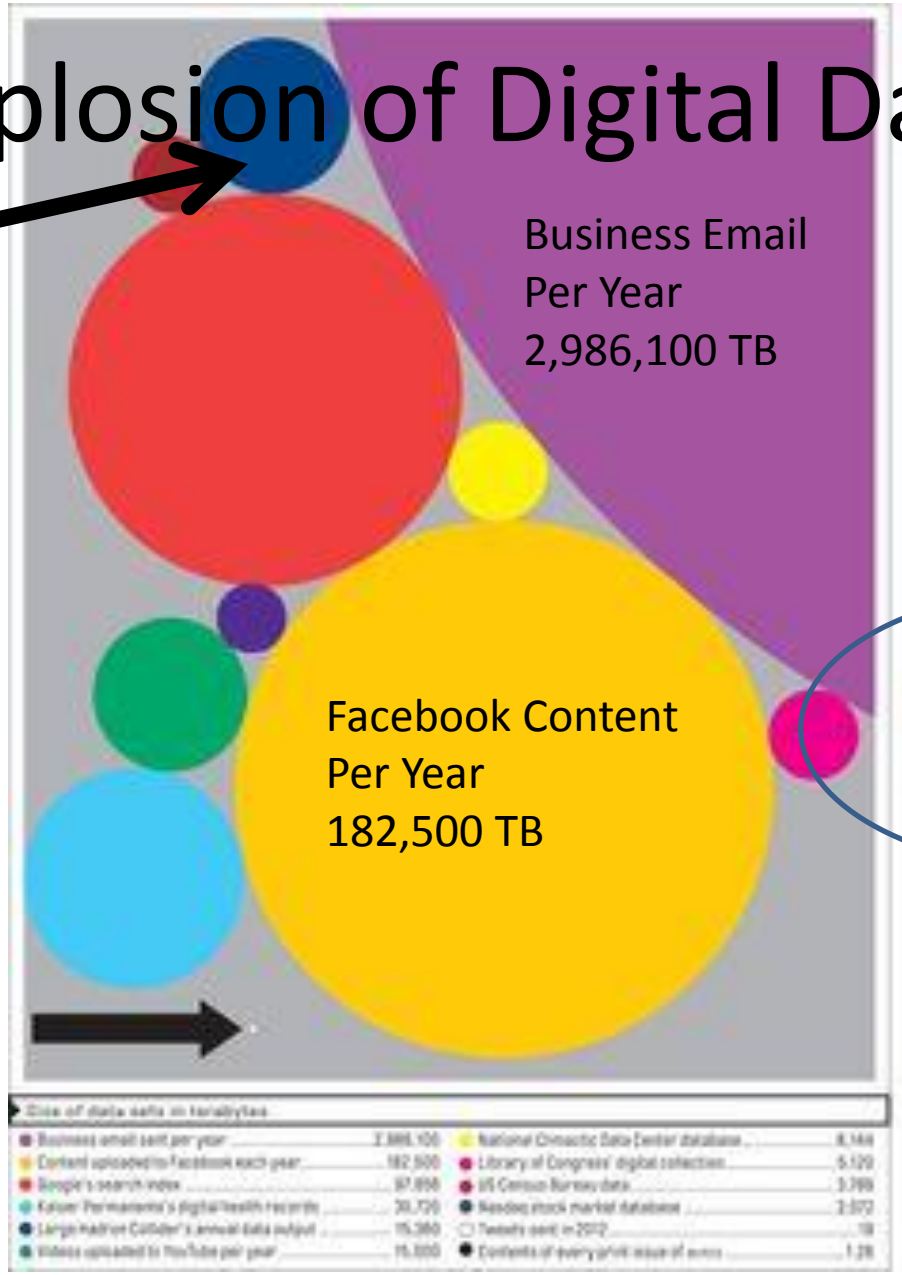
Tweets in 2012
19 TB

1Mega = 1,000,000 = 10^6
1Giga = 10^9
1Tera = 10^{12}
1Peta = 10^{15}
1Exa = 10^{18}
1Zetta = 10^{21}

(WIRED, May 2013)

Explosion of Digital Data

Large Hadron Collider
15,360 TB



Tweets in 2012
19 TB

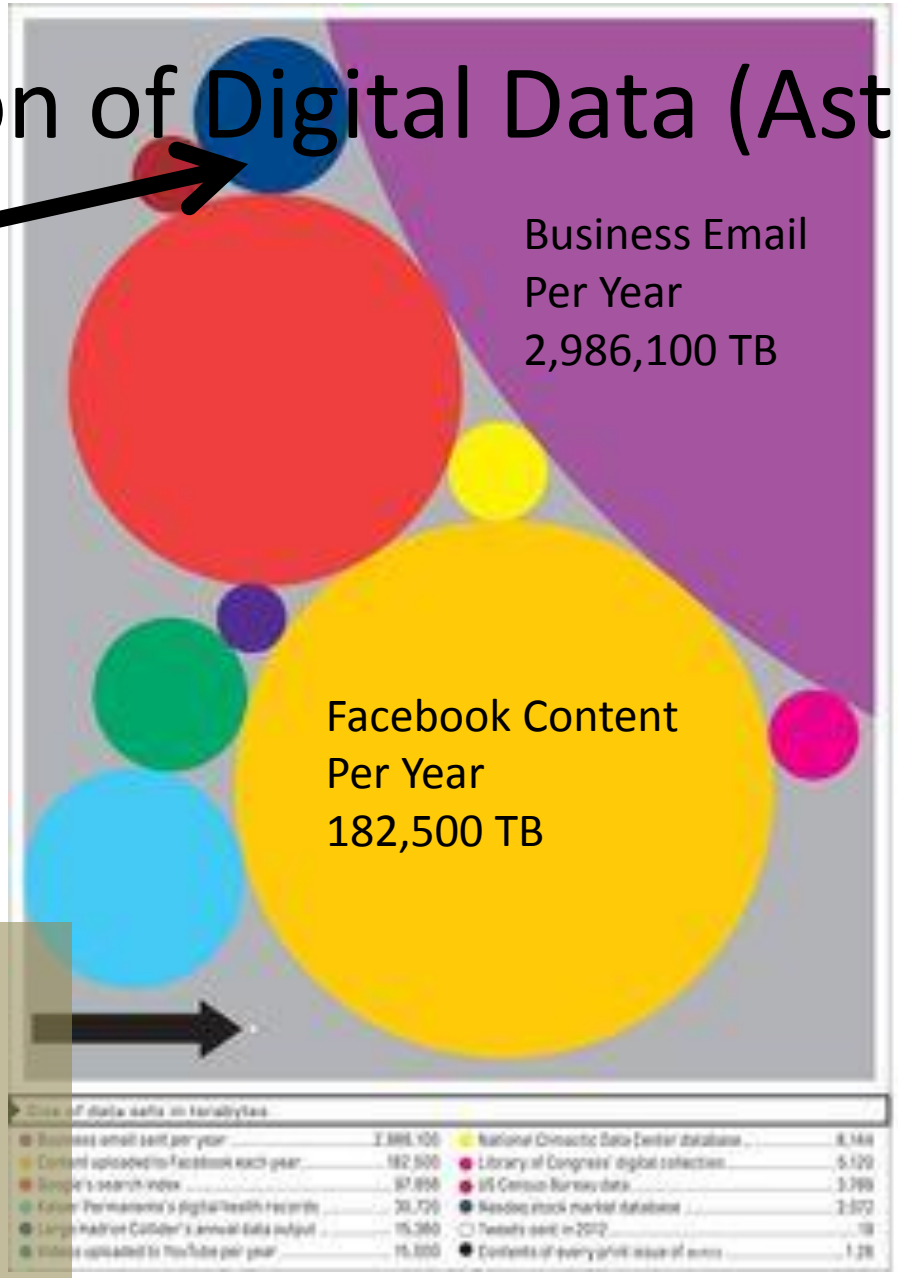
1Mega = 1,000,000 = 10^6
1Giga = 10^9
1Tera = 10^{12}
1Peta = 10^{15}
1Exa = 10^{18}
1Zetta = 10^{21}

2012
2.8 Zettabytes!

(WIRED, May 2013)

Explosion of Digital Data (Astronomy)

Large Hadron Collider
15,360 TB



1Mega = 1,000,000 = 10^6
1Giga = 10^9
1Tera = 10^{12}
1Peta = 10^{15}
1Exa = 10^{18}
1Zetta = 10^{21}

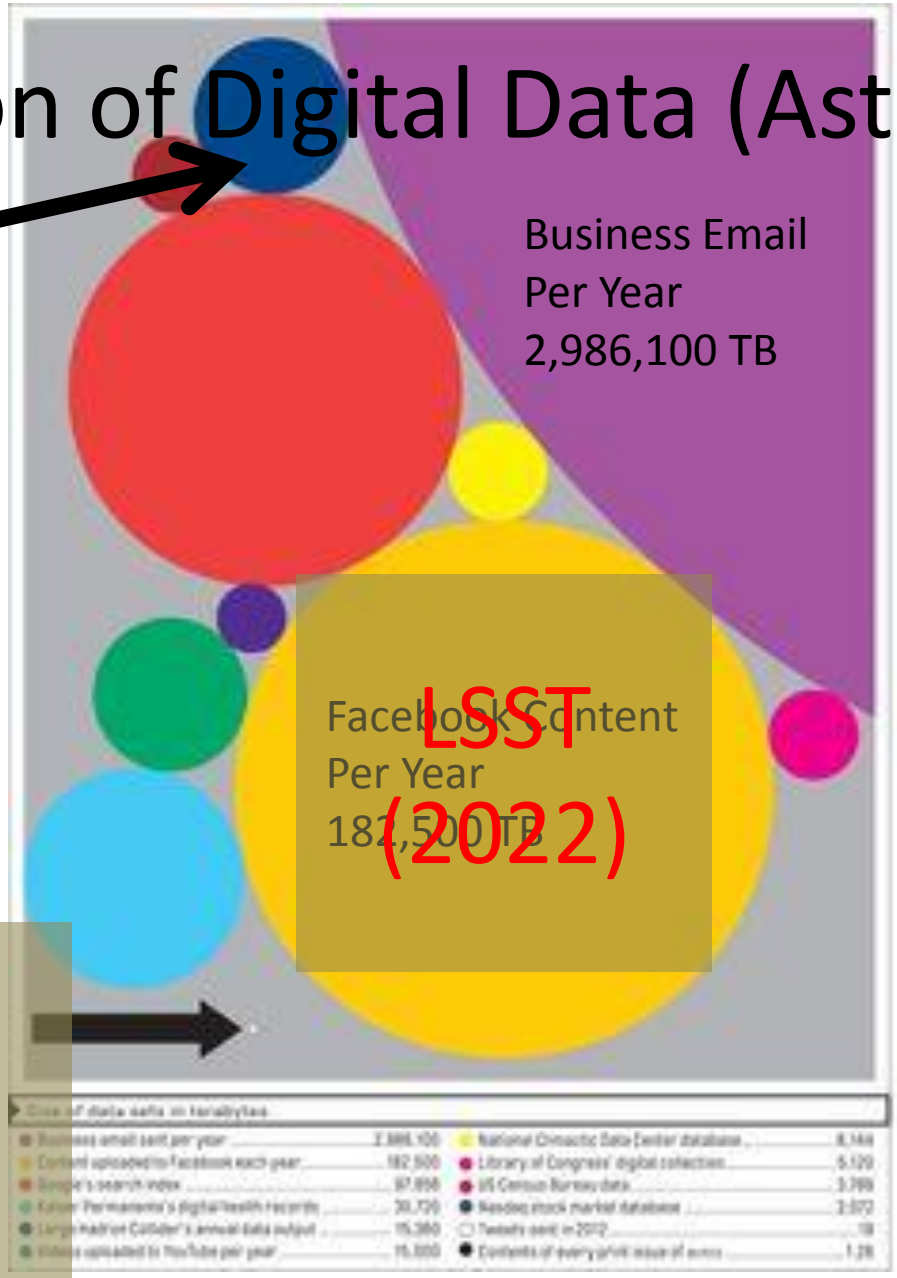
Tweets in 2012
19 TB

SDSS
(now)

(WIRED, May 2013)

Explosion of Digital Data (Astronomy)

Large Hadron Collider
15,360 TB



1Mega = 1,000,000 = 10^6
1Giga = 10^9
1Tera = 10^{12}
1Peta = 10^{15}
1Exa = 10^{18}
1Zetta = 10^{21}

(WIRED, May 2013)

Why the Data Deluge?

- We have been transitioning from analog to digital devices.
- Almost all digital circuits use transistors as building blocks.
- Transistor count doubles approximately every 2 years (Moore's Law, 1965).

Intel 80386 (1985)
100,000 transitors

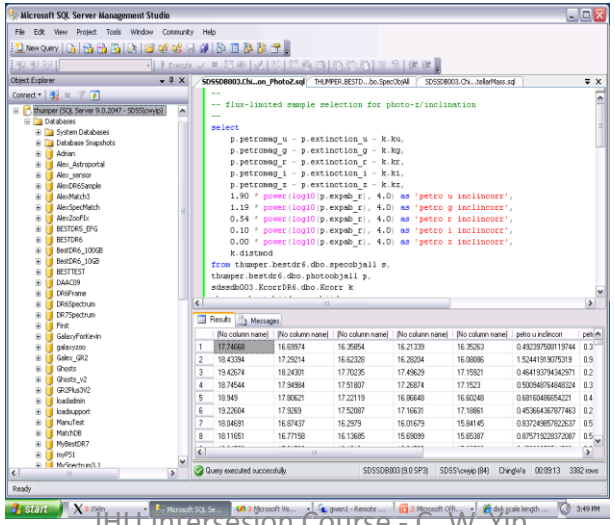
Current Laptops
100,000,000 or more
transitors

From Data to Information

- We don't just want data.
- We want information from the data.

Information ← **Database** ← **Sensors**

Data Analysis
or
Data Mining



Charged-Coupled Device

- CCDs are the state-of-the-art detectors in many areas of observational science.
- First CCD: Boyle & Smith, 1970 at Bell Lab.


Photo: Richard Epworth

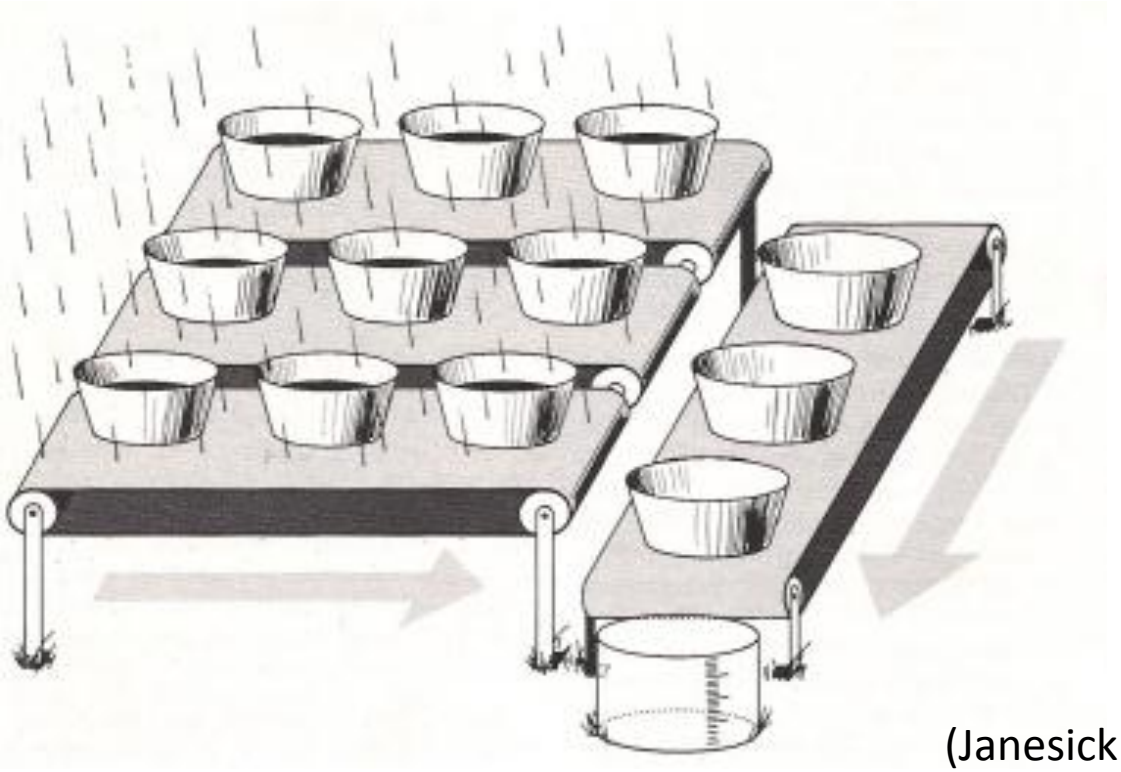

Copyright © National Academy of Engineering


Photo: National Inventors Hall of Fame Foundation/SCANPIX

(Nobel Prize, 2009)

How CCD works?

- A CCD uses pixels (buckets) to collect photons (raindrops) after integration (storm).

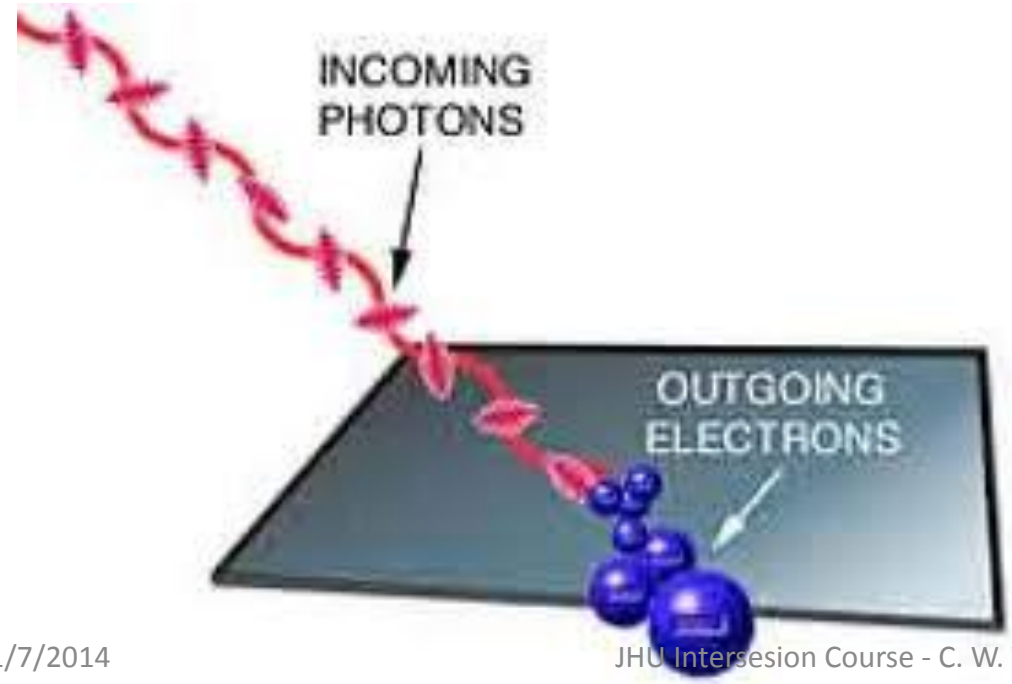


(Janesick & Blouke 1987)

From Photons to Electrons: Photoelectric Effect

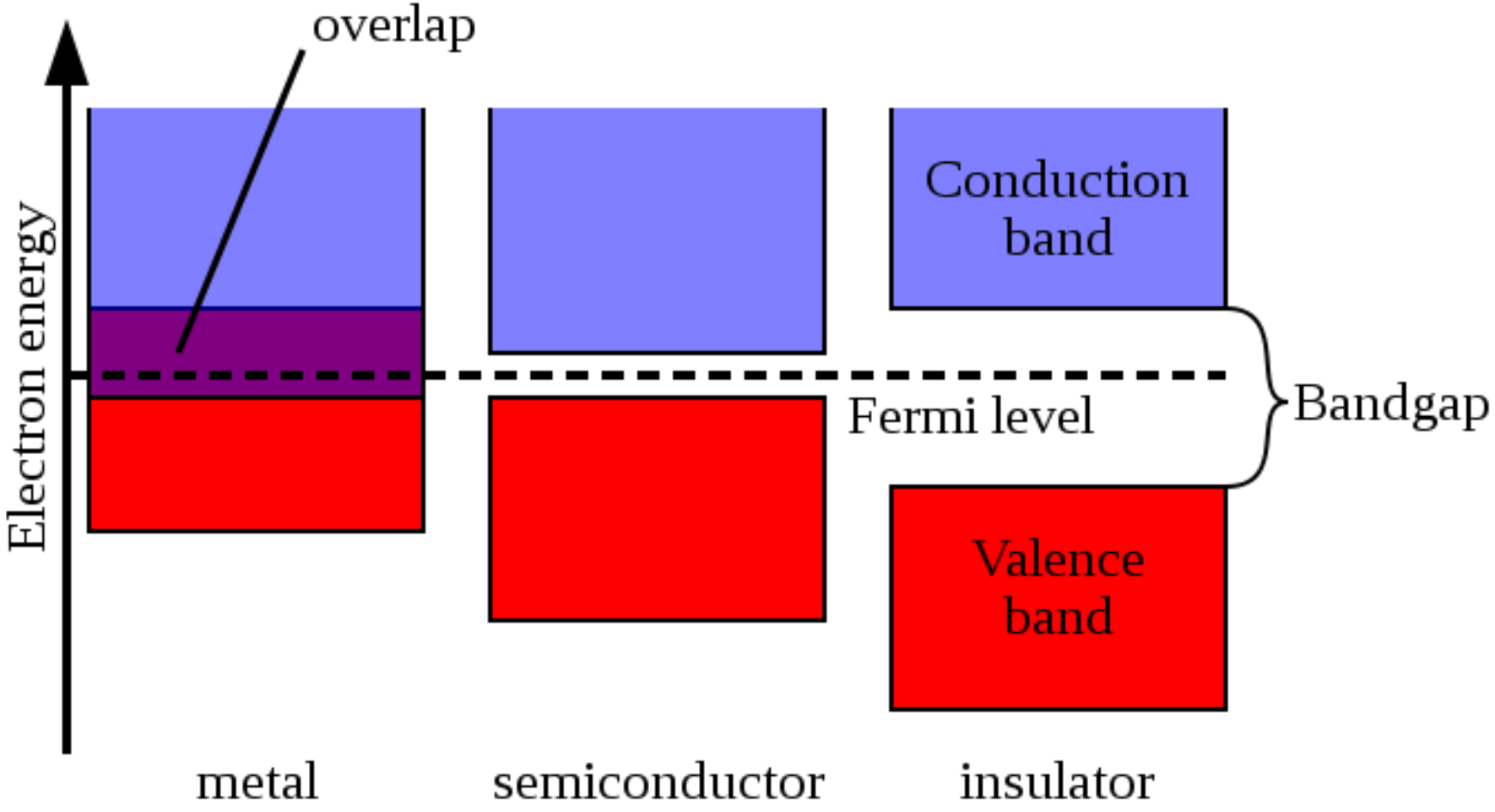
Energy (one photon) = h ν

where:
h = Planck's constant = 6.626 x 10^{-34} m^2 kg s^{-1}
ν = Frequency of the photon, in s^{-1}

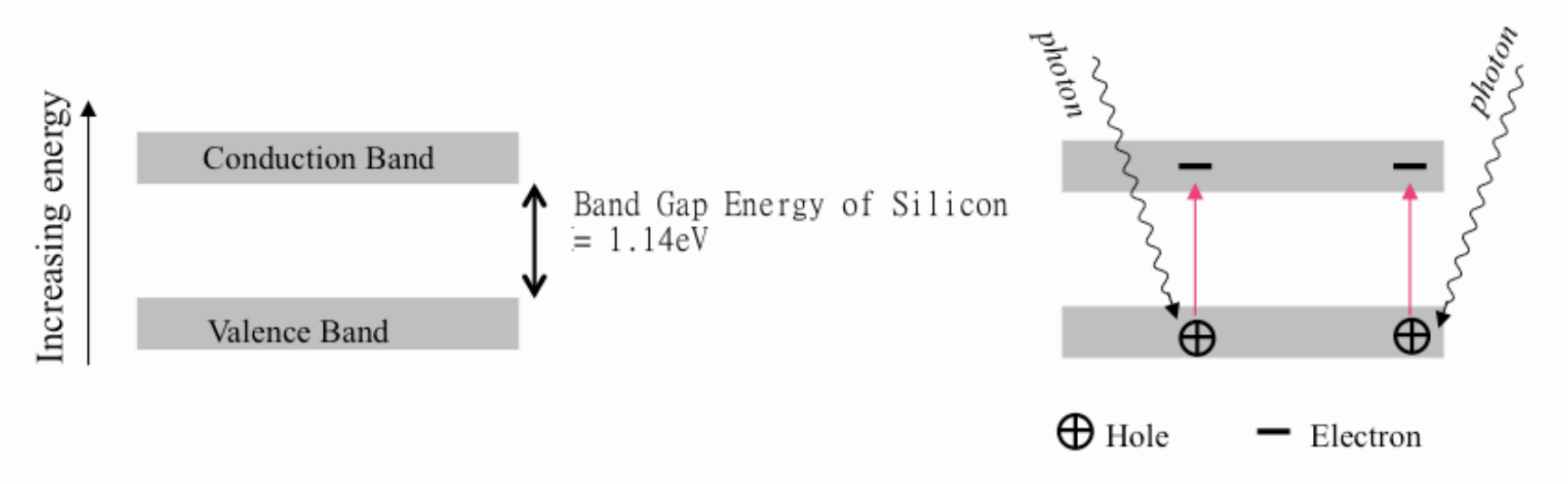


(Nobel Prize, 1921)

Semiconductor: In-between Conductor and Insulator



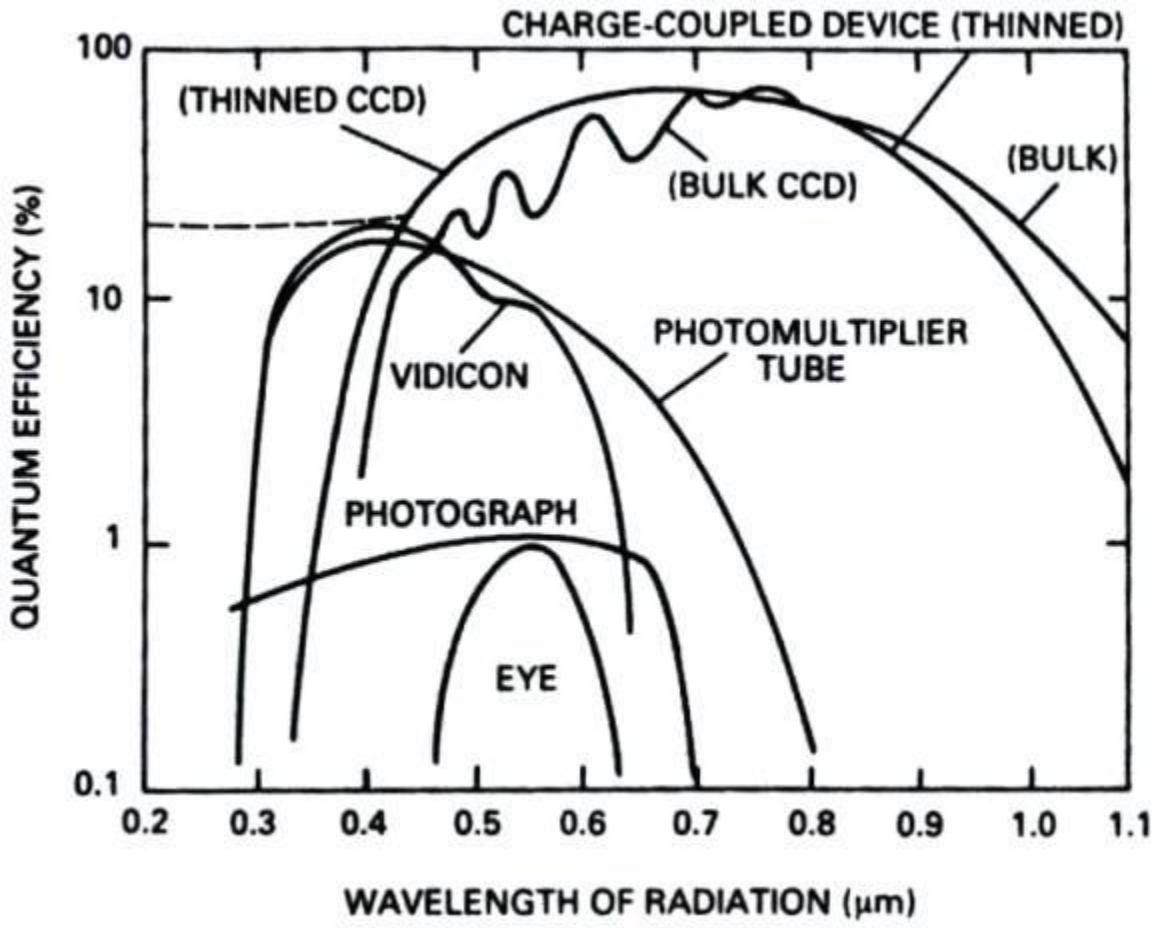
Semi-Conductor Band Gap



Advantages of CCDs over Analog Solutions

- High Quantum Efficiency (QE) over a wide spectral bandpass.
  - QE = 100%, all incoming photons are accounted for in the output [ideal detector]
  - QE = 0%, all incoming photons are missing from the output
- Low Noise.
- Digital output which can be saved to disk or analyzed using computers.

CCD Quantum Efficiency Curve



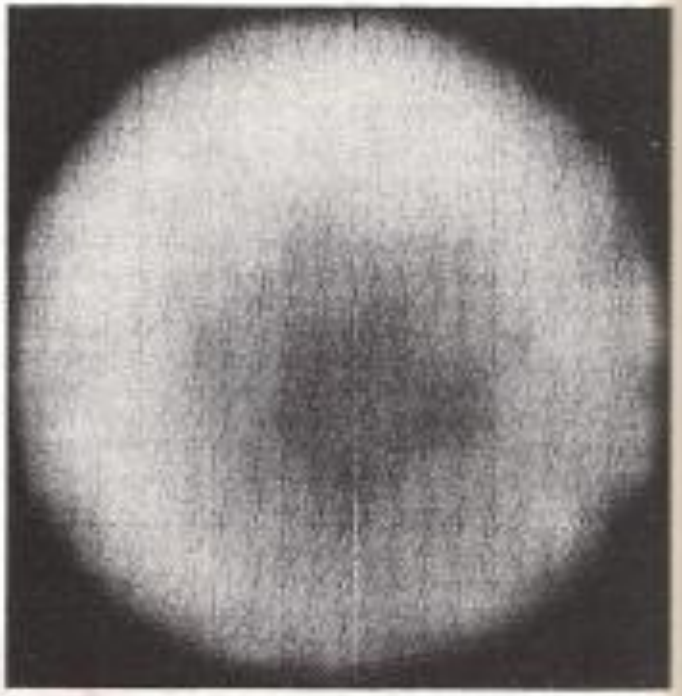
Some Digital Cameras have IR Blocking Filter Removed for Astronomy

Canon 60D

Canon 60Da

First CCD Image in Astronomy, 1975

- The image of planet Uranus was taken by Jet Propulsion Laboratory at 8900Å using Mt. Lemmon 61” telescope.
- Sky & Telescope Article, Janesick & Blouke 1987

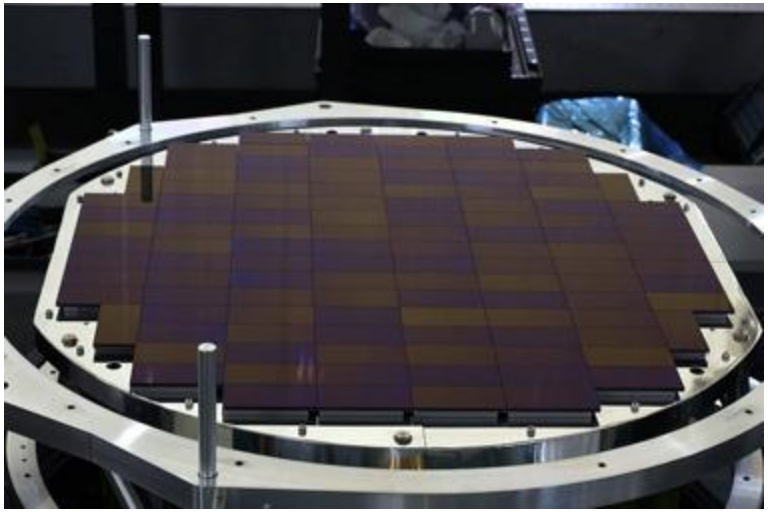


This picture of Uranus is thought to be the first astronomical image made with a charge-coupled device, or CCD. It was obtained in 1975 by scientists from the Jet Propulsion Laboratory and the University of Arizona, using the 61-inch telescope in the Santa Catalina Mountains near Tucson. Recorded at a wavelength of 8900 angstroms in the near infrared, it shows a region of enhanced methane absorption (dark area) near Uranus' south pole.

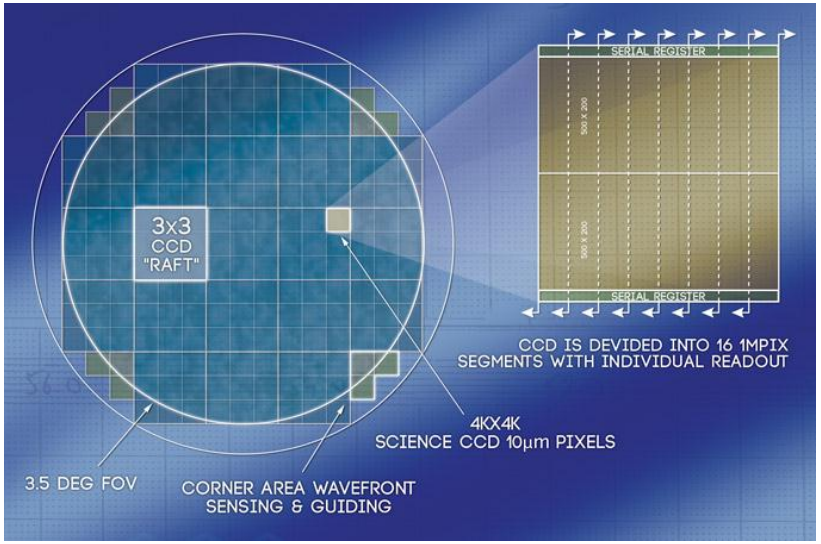
Hubble Space Telescope
2.4m telescope
10 days observation

Hubble Deep Field HST · WFPC2
PRC96-01a · ST ScI OPO · January 15, 1996 · R. Williams (ST ScI), NASA

Big CCDs in Astronomy



870M Pixels
Subaru Hyper Suprime-Cam
(HSC)

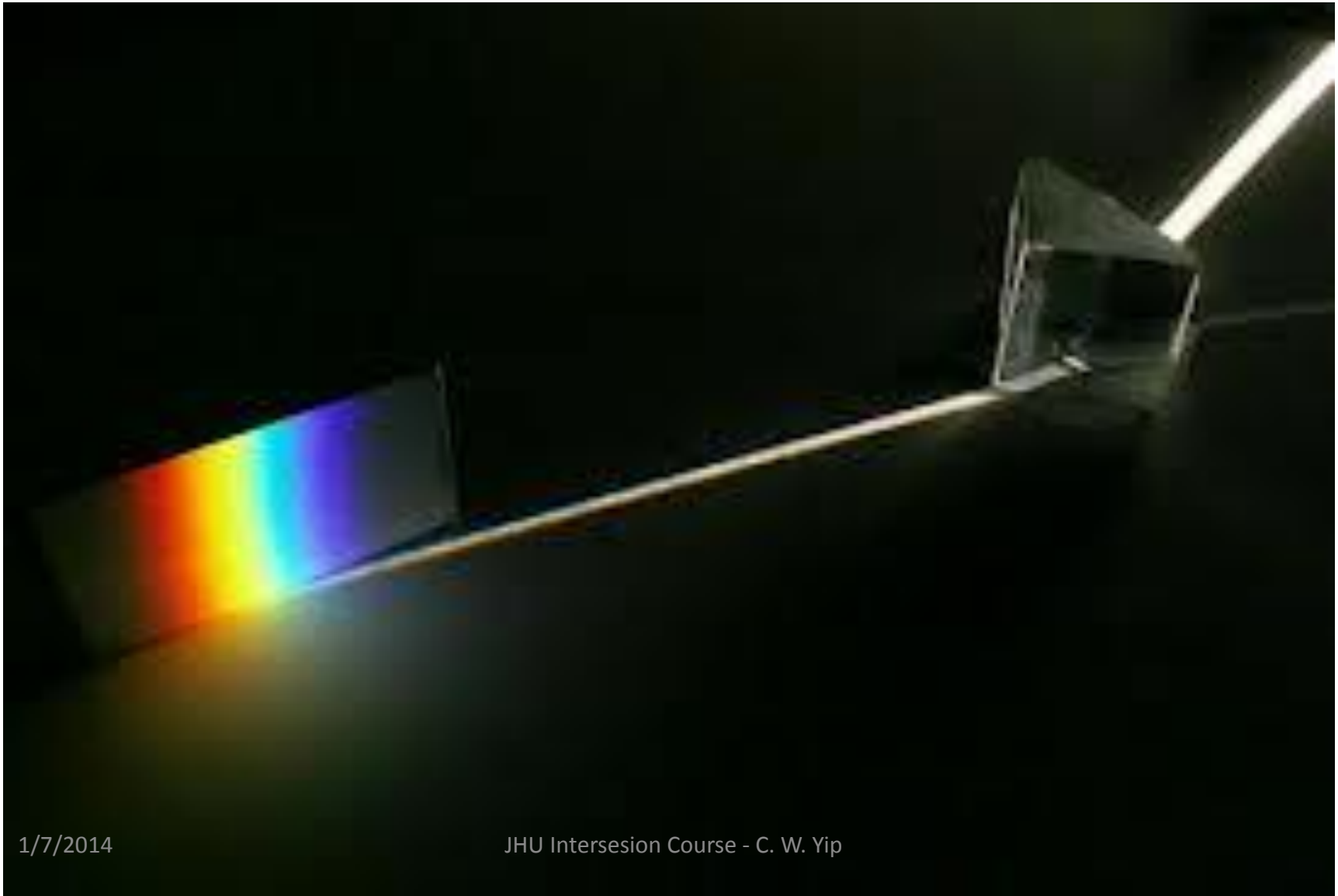


3.2G Pixels
Large Synoptic Survey Telescope
(LSST)

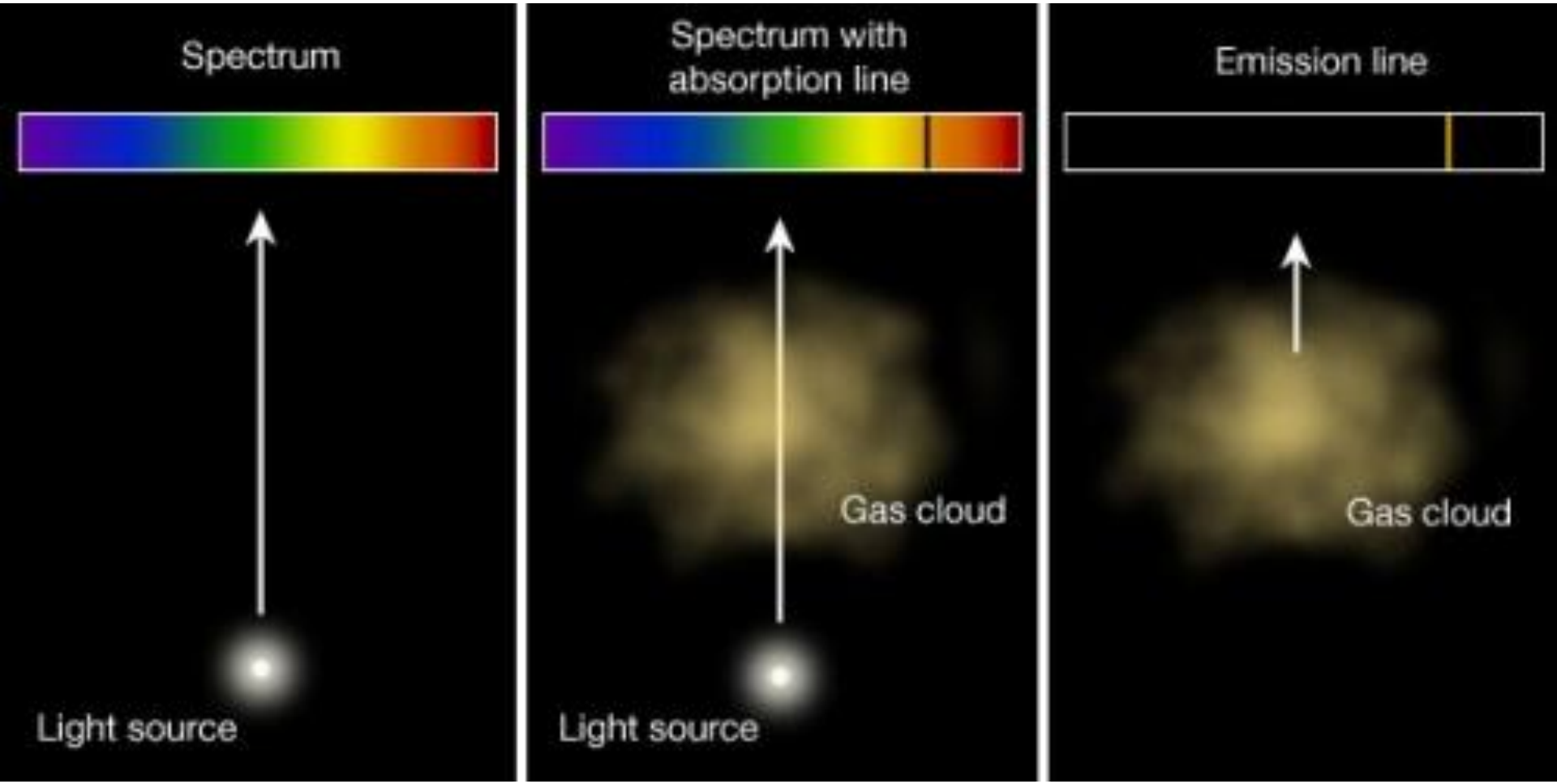
Spectroscopy

- Spectroscopy is the study of light.
- The bulk of astronomical observation is spectroscopy.
- How is it done?
- What are the data?

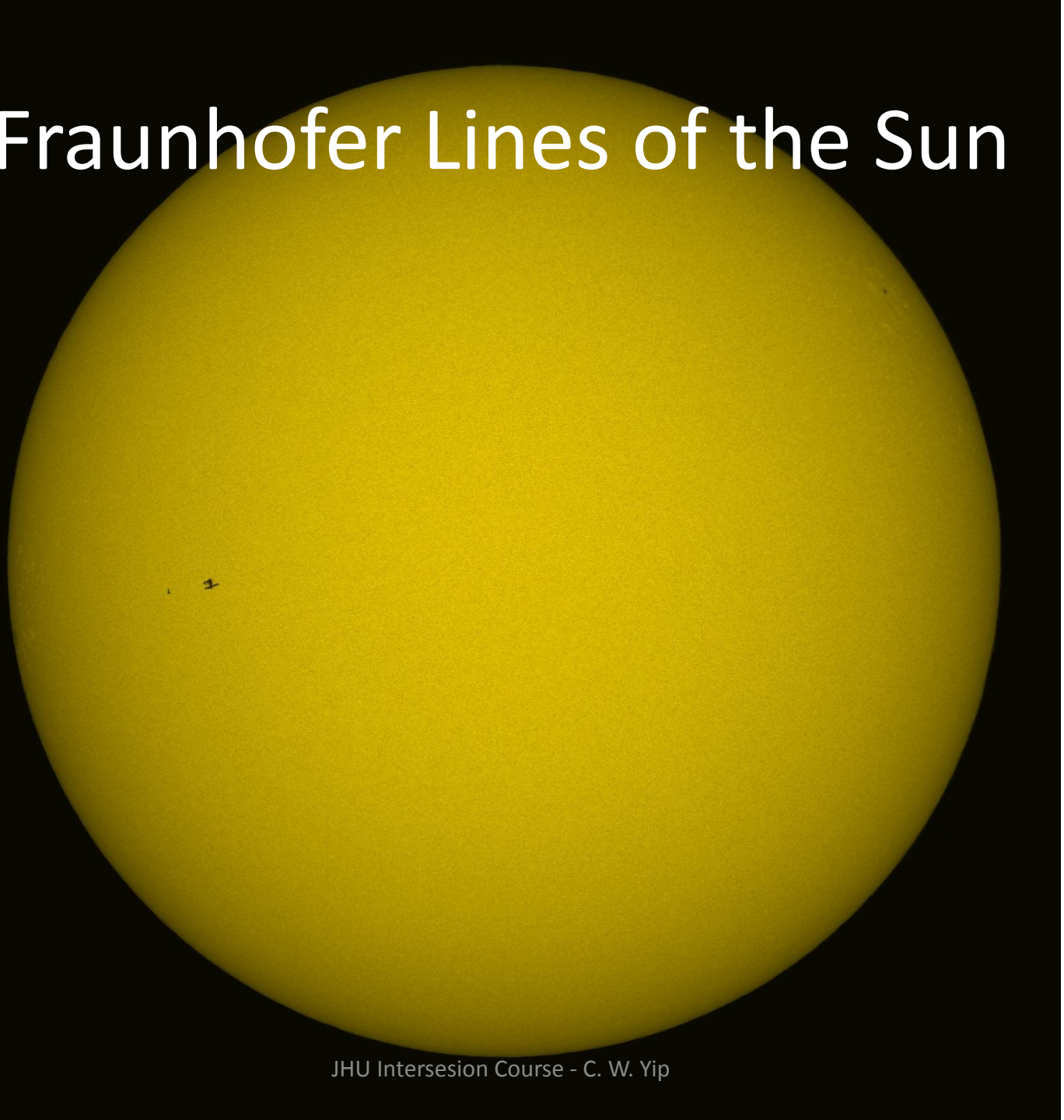
Spectroscopy



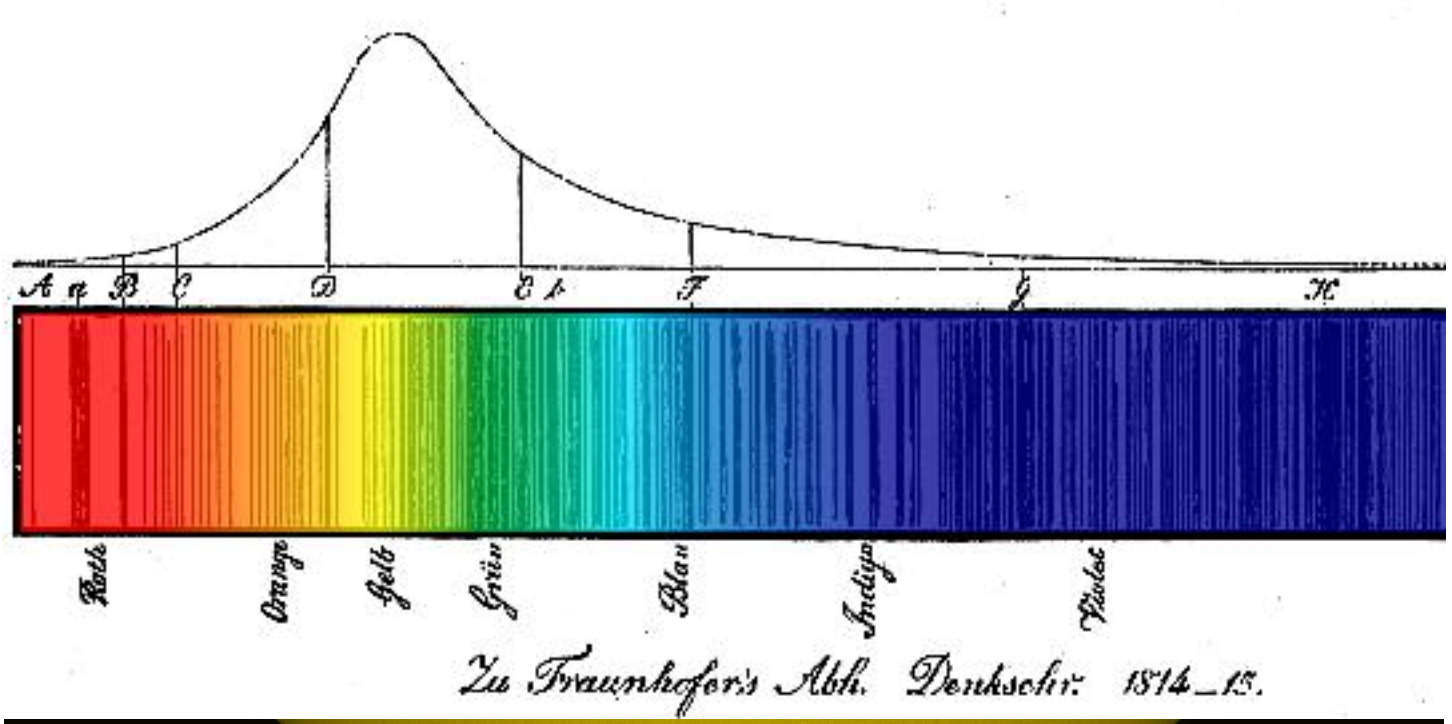
Kirchhoff's Law of Thermal Radiation



Fraunhofer Lines of the Sun



Fraunhofer Lines of the Sun

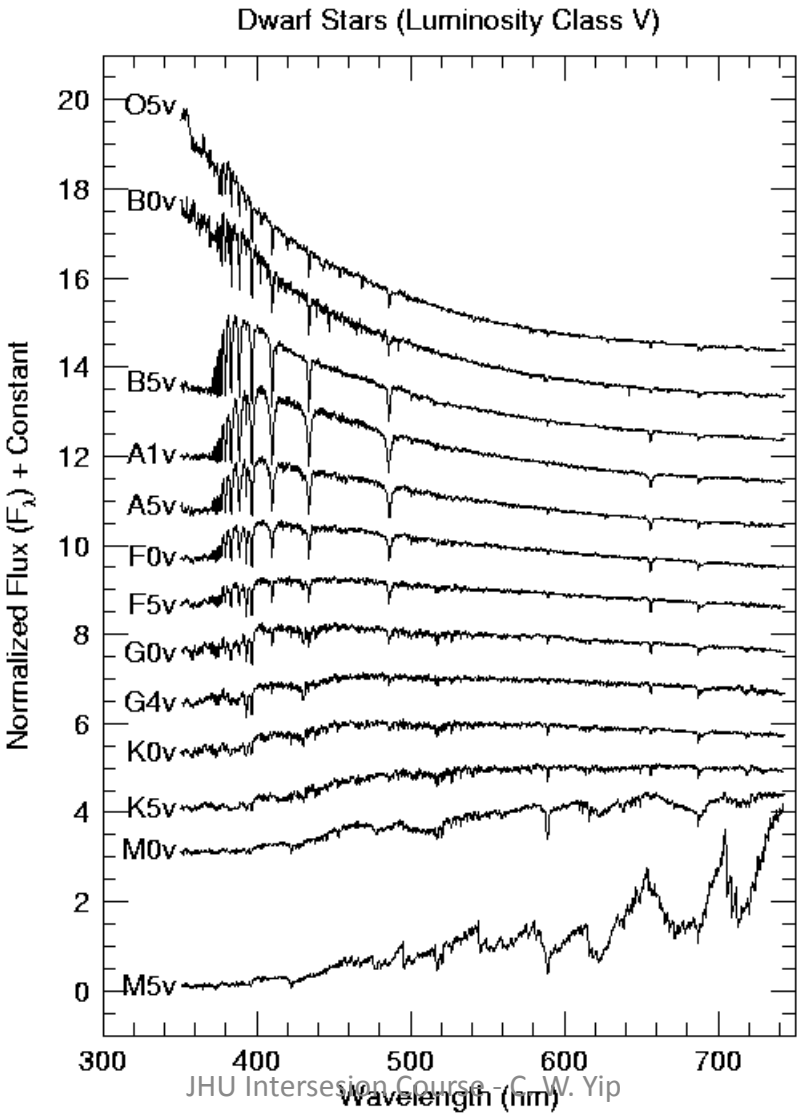


Stellar Spectral Classification

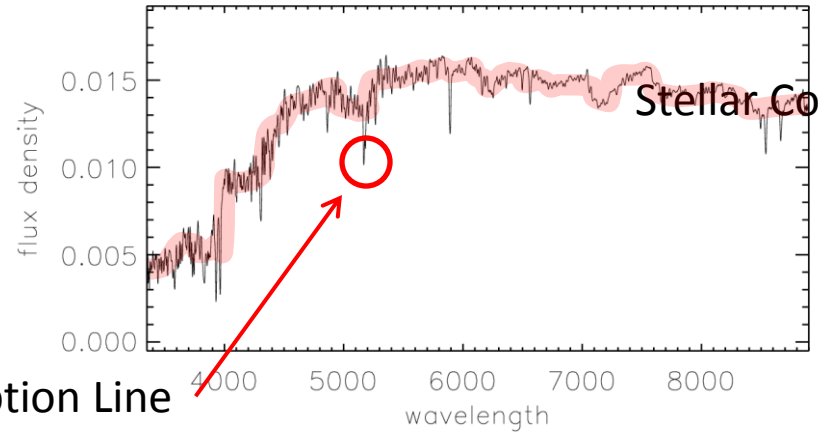
Blue
Hot

Sun is a
G type star.

Red
Cold



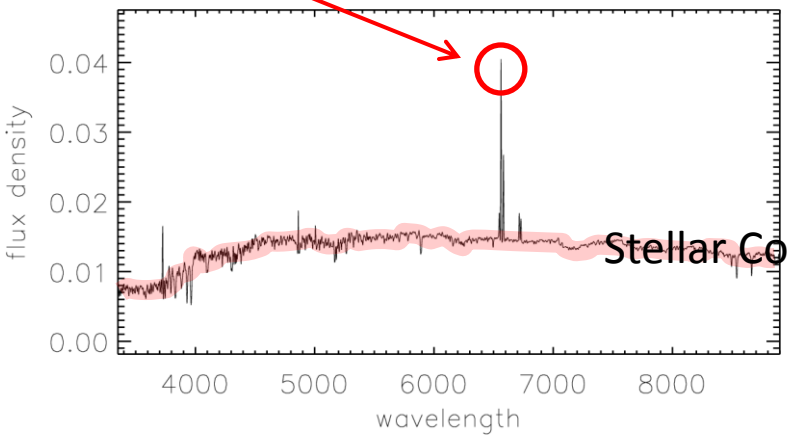
Spectra are the keys to Stars, Gas and Dust in Galaxies



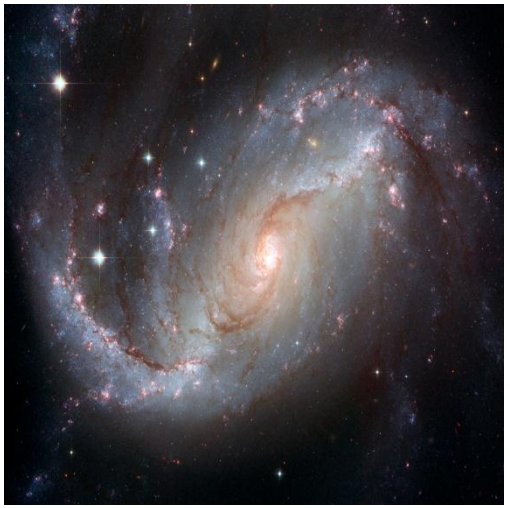
Stellar Continuum

Absorption Line

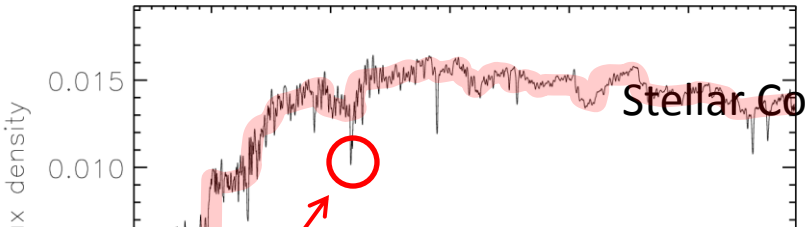
Emission Line



Stellar Continuum



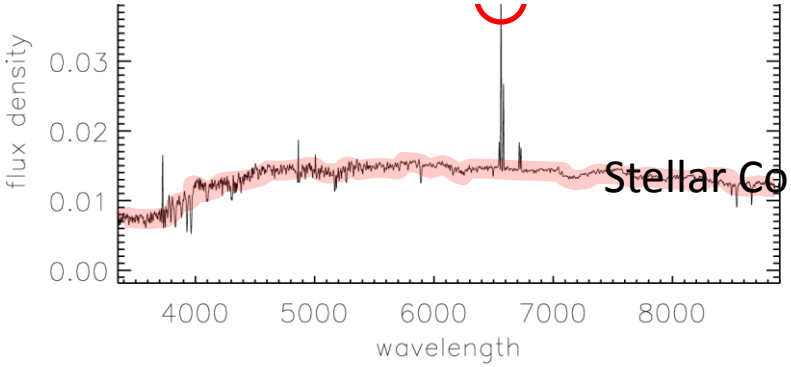
Spectra are the keys to Stars, Gas and Dust in Galaxies



Stellar Continuum

3 Main Characteristics:

- Stellar Continuum
- Absorption Lines
- Emission Lines

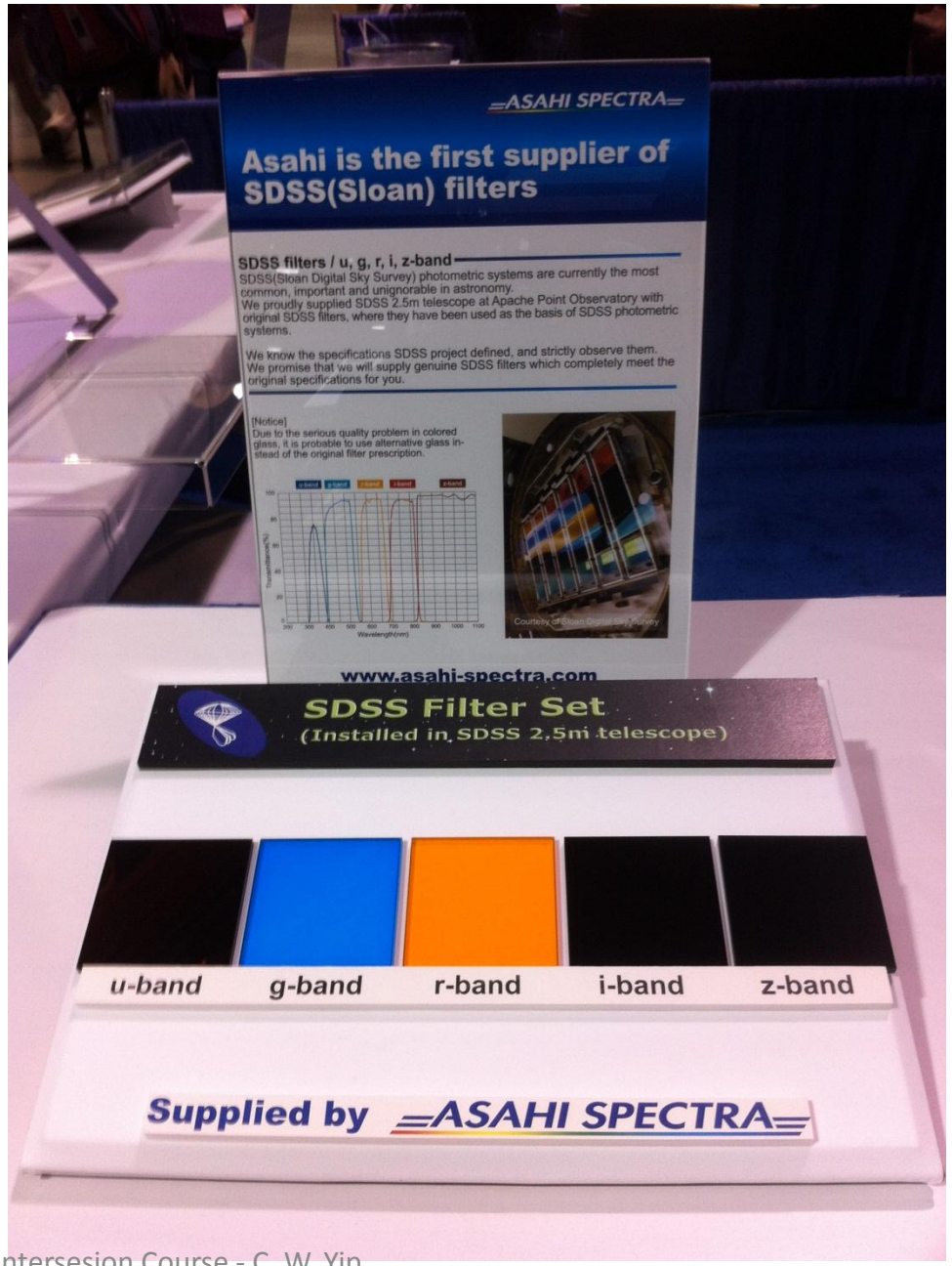


Stellar Continuum

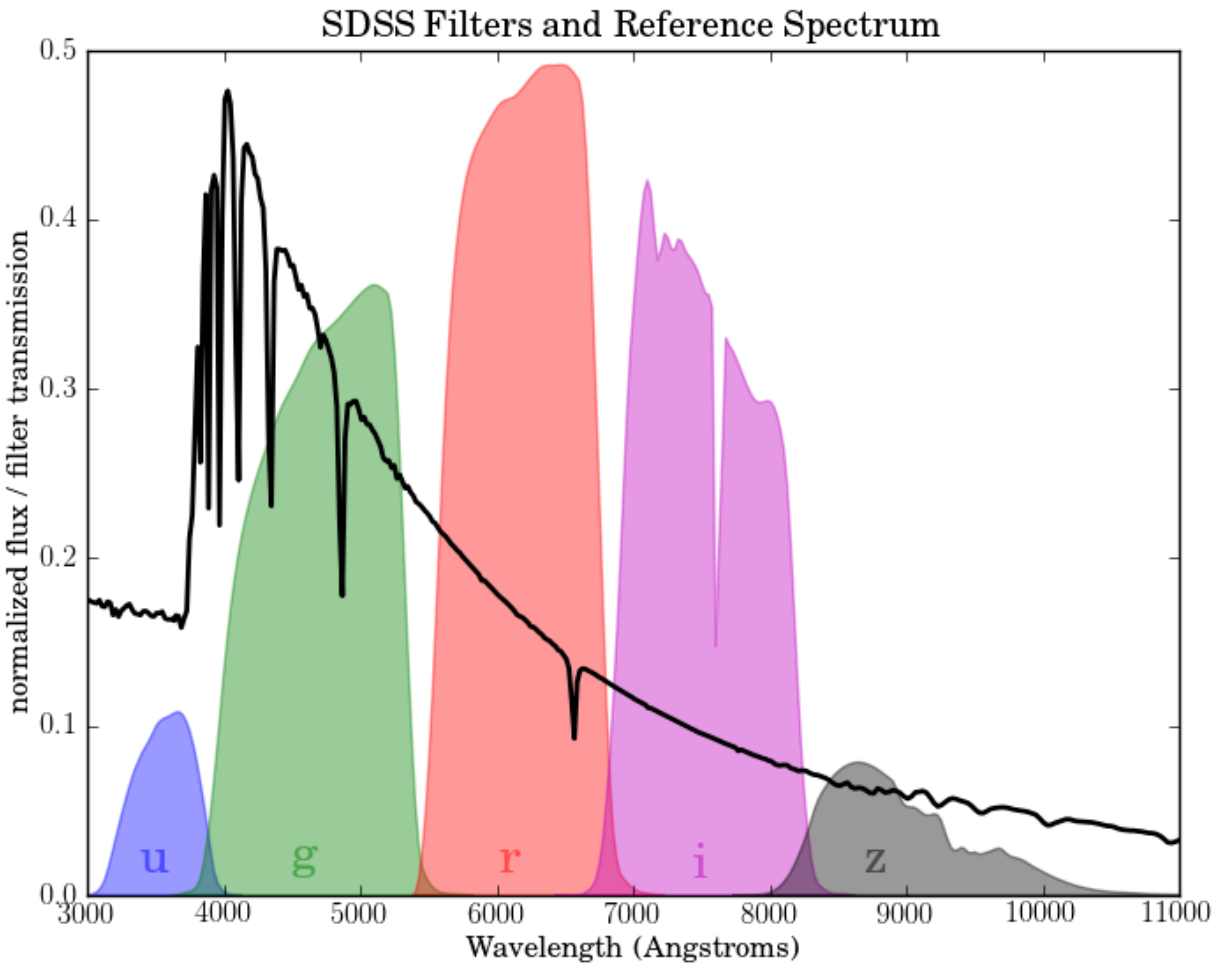
Photometry

- Early sky surveys were mostly photometric.
- How is it done?
- What are the data?

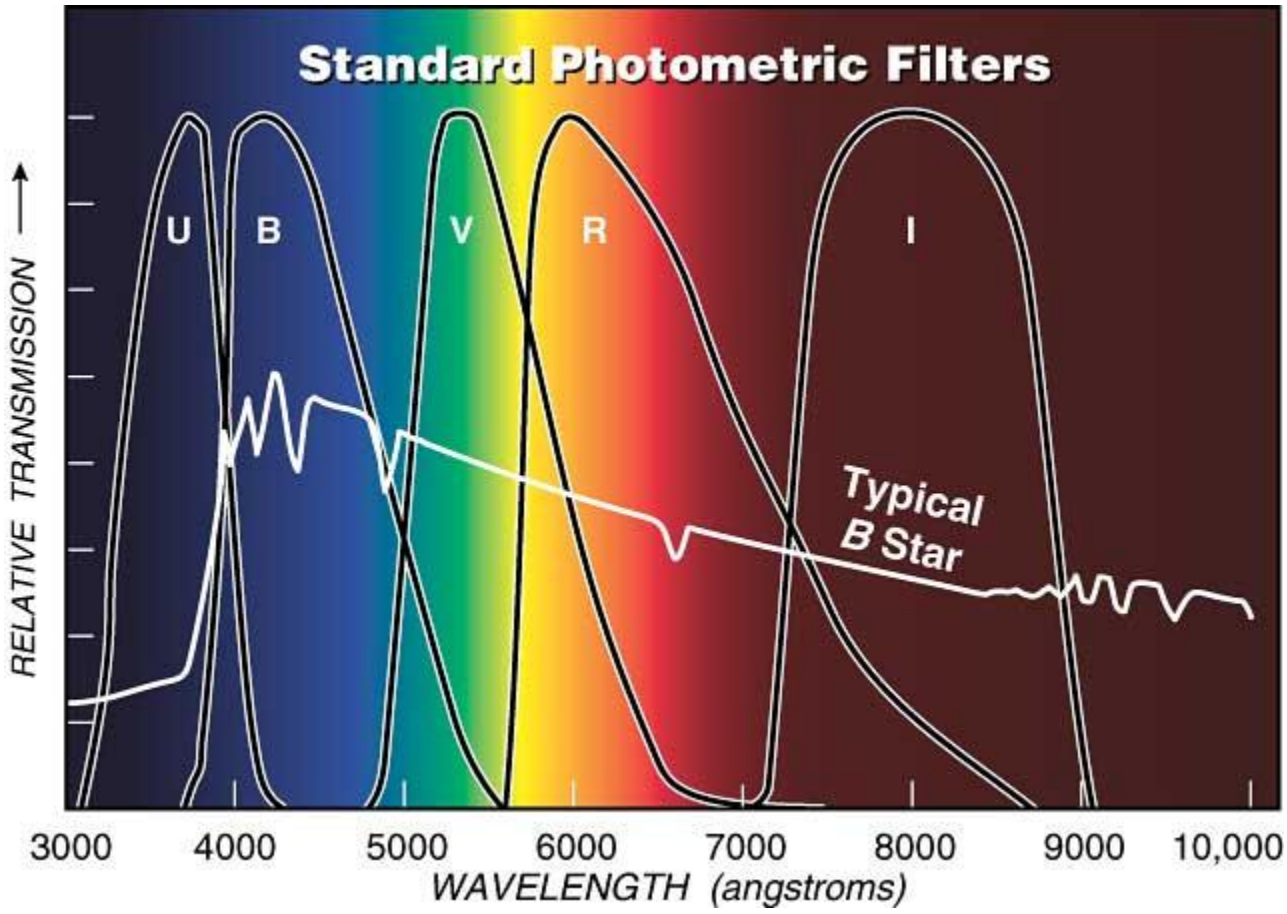
Photometric Filters: SDSS ugriz



Transmission Curves: SDSS ugriz Filters

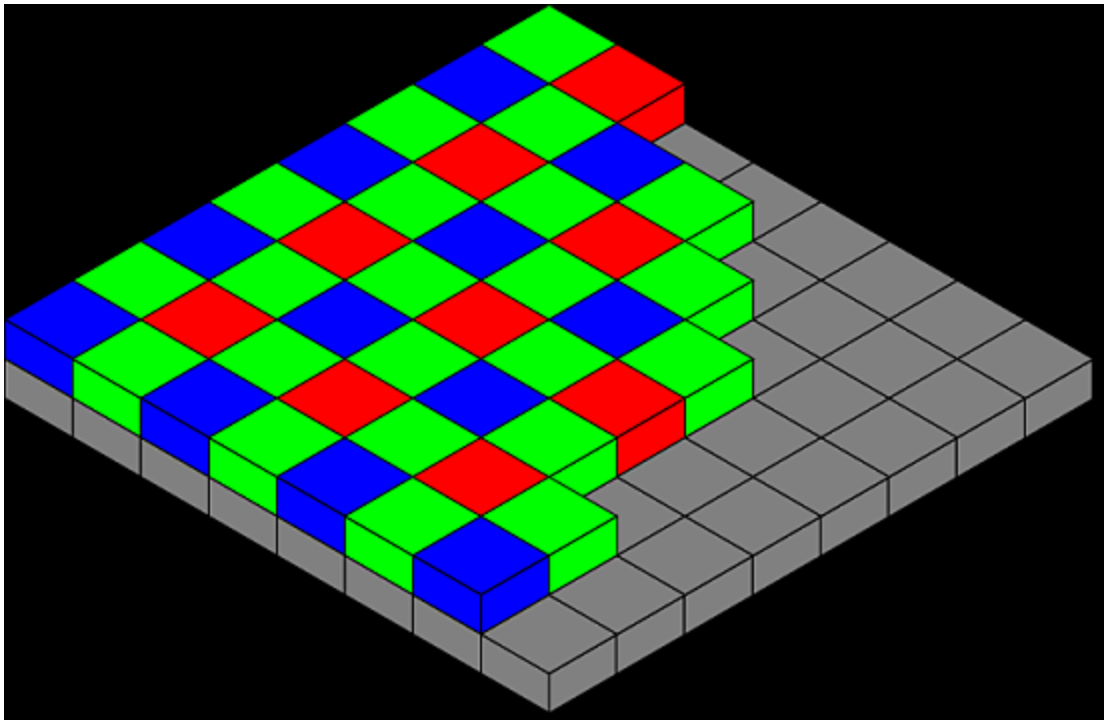


Transmission Curves: Johnson UBV Filters

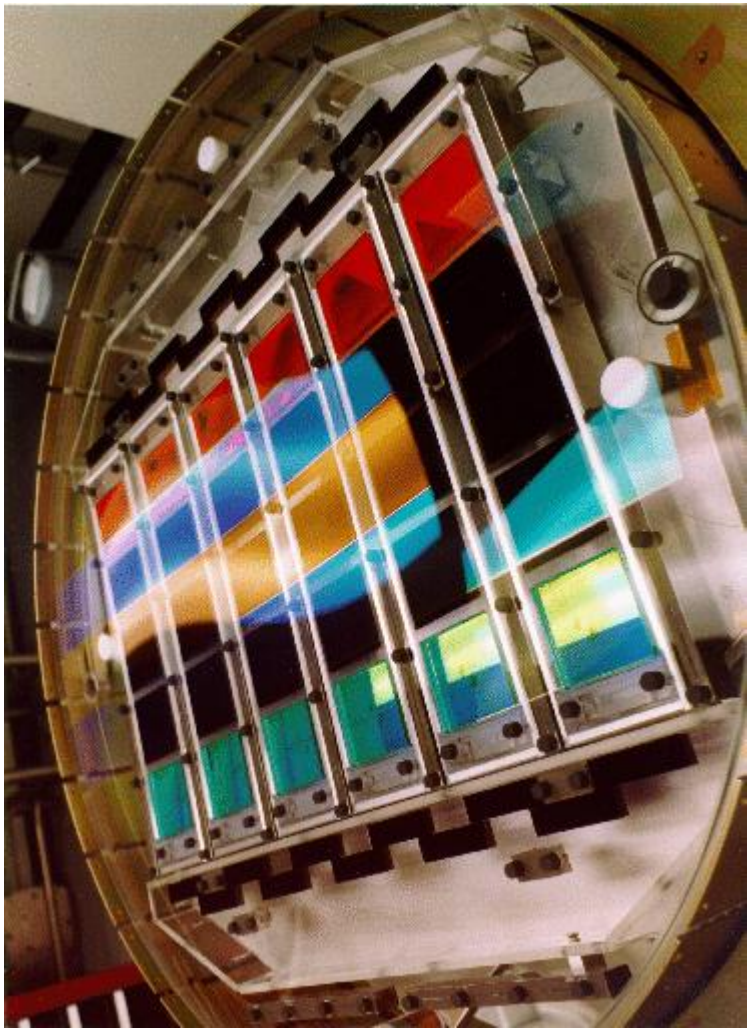


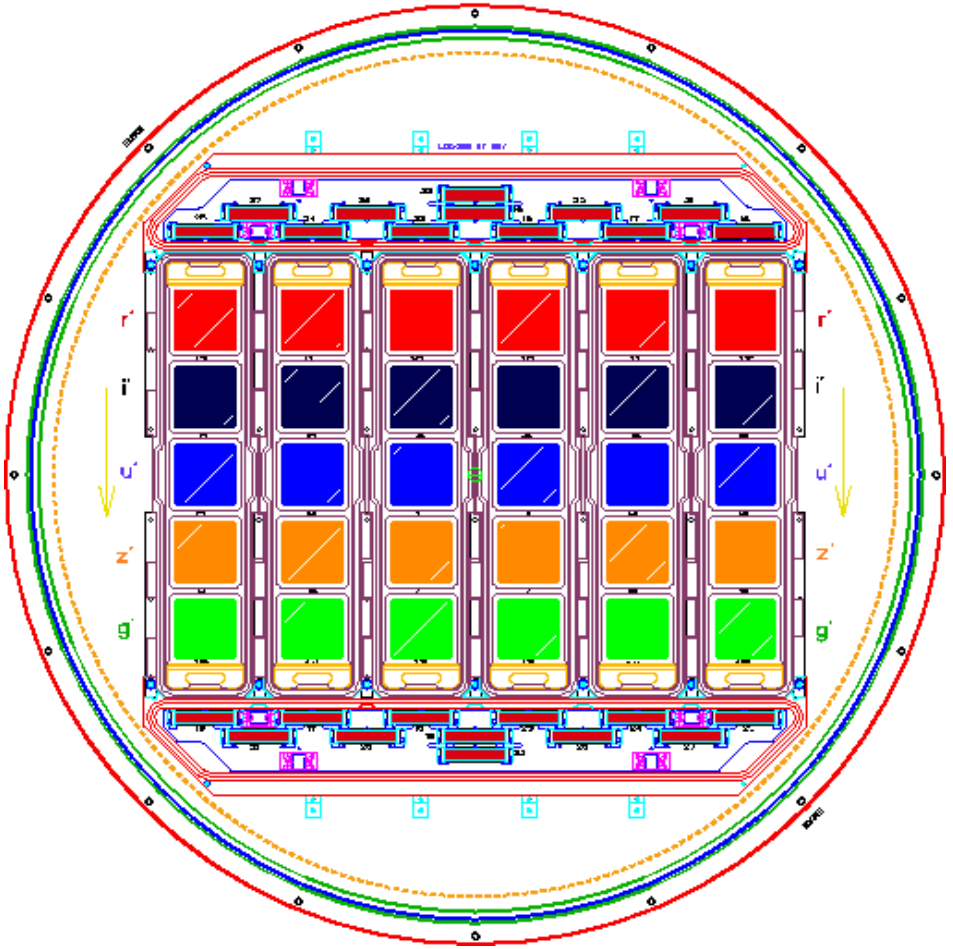
Bayer's Filter Array

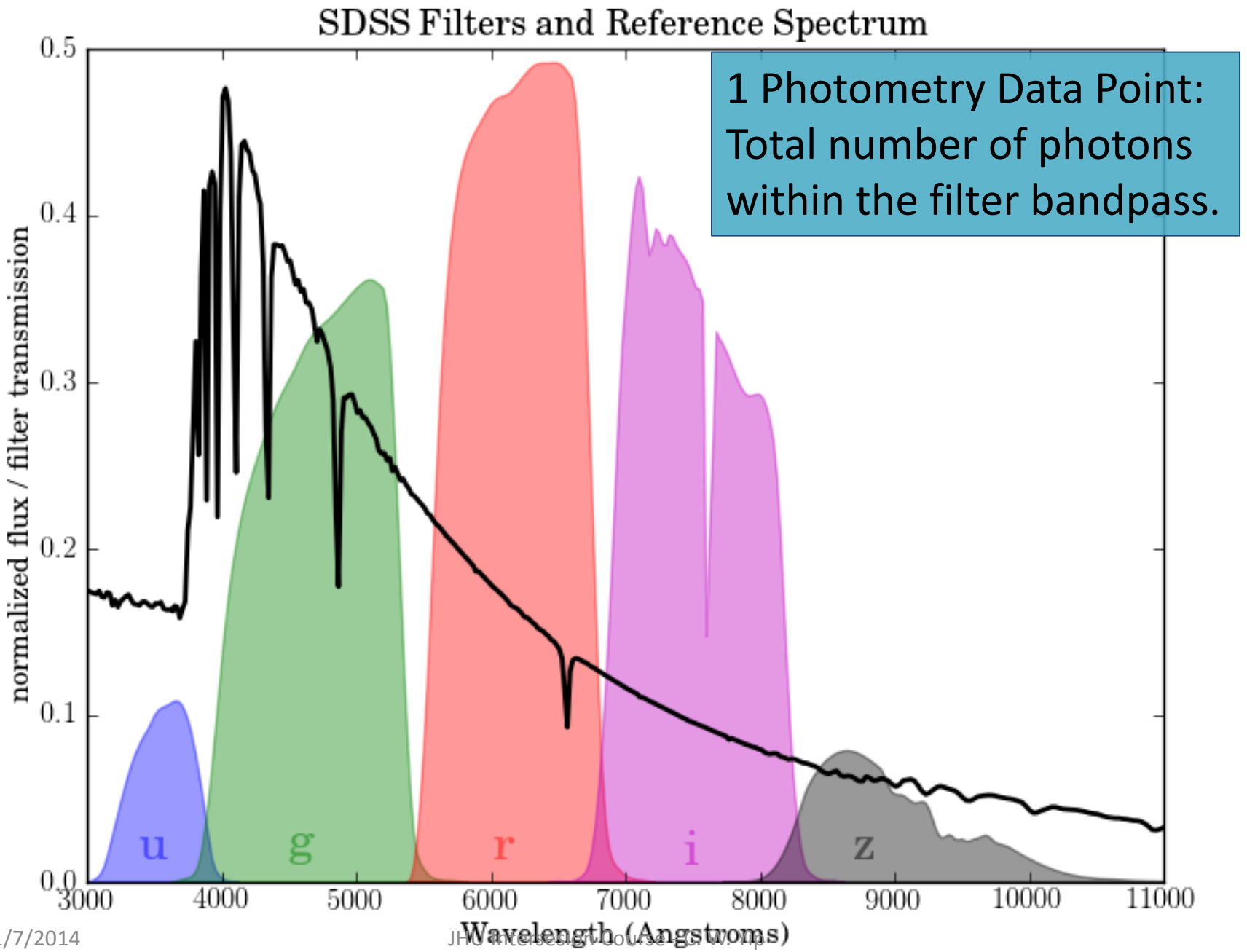
- CCD only counts the number of photons
- How do we get the colors?



SDSS CCD Imaging Camera







1 Photometry Data Point: Total number of photons within the filter bandpass.

Spectroscopic and Photometric Data at a Glance

Spectroscopy:
Number of photons as a function of wavelength.

Photometry:
Total number of photons within the filter bandpass, as a function of filter.

DR6

Explore Home

Search by
ObjId
Ra,dec
5-part SDSS
Plate-MJD-Fiber
SpecObjId

Summary

PhotoObj
PhotoTag
More Observations
Field
Frame
PhotoZ
Neighbors
Finding chart
Navigate
FITS

SpecObj
All Spectra
SpecLine
SpecLineIndex
XCredShift
ELredShift
Spectrum
Plate
FITS

NED search
SIMBAD search
AKARI FIS
AKARI IRC
ADS search

flags	BINNED1 NOPETRO
PrimTarget	TARGET_GALAXY
SecTarget	

getFrames() has failed:
SELECT img , f.a, f.b, f.c, f.d, f.e
f.ra, f.dec, f.fieldID, dbo.fSDSS(f.
FROM dbo.fGetNearbyFrameEq
ON f.fieldID = n.fieldID WHERE z

Exception Message:Out of mem

u	g	r	i	z
19.90	18.16	17.16	16.63	16.20

err_u	err_g	err_r	err_i	err_z
0.09	0.01	0.01	0.01	0.02

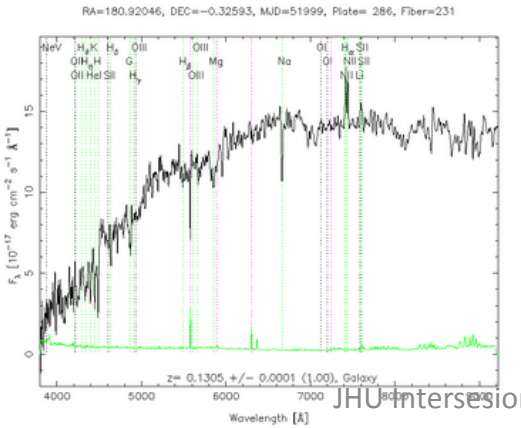
run	rerun	camcol	field	obj	rowc	colc
752	40	3	249	118	1411.3	950.3

fiberMag_r	petroMag_r	devMag_r	expMag_r	psfMag_r	modelMag_r
18.93	17.33	17.16	17.39	18.24	17.16

extinction_r	petroRad_r	parentId	nChild
0.07	6.221	0	0

SpecObjID = 80725178312556544

plate	mjd	fiberId	z	zErr	zConf	specClass	ra	dec	fiberMag_r	objId
286	51999	231	0.131	0.00015	0.999426	GALAXY	180.92046	-0.32593	18.88	587722982814187638



zStatus	XCORR_HIC
zWarning	OK
PrimTarget	TARGET_GALAXY
SecTarget	
eClass	-0.161853
emZ	0.500
emConf	0.11509
xcZ	0.131
xcConf	0.999426

Data Analysis Case Study: Hubble's Law (1929)

100" Mt Wilson Telescope

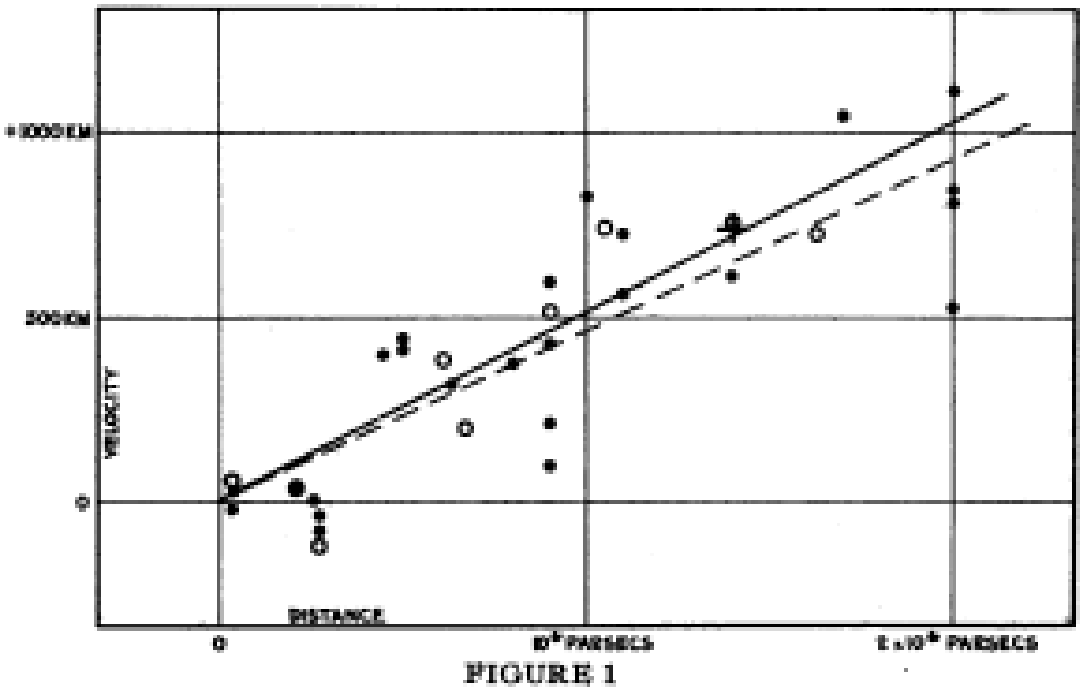
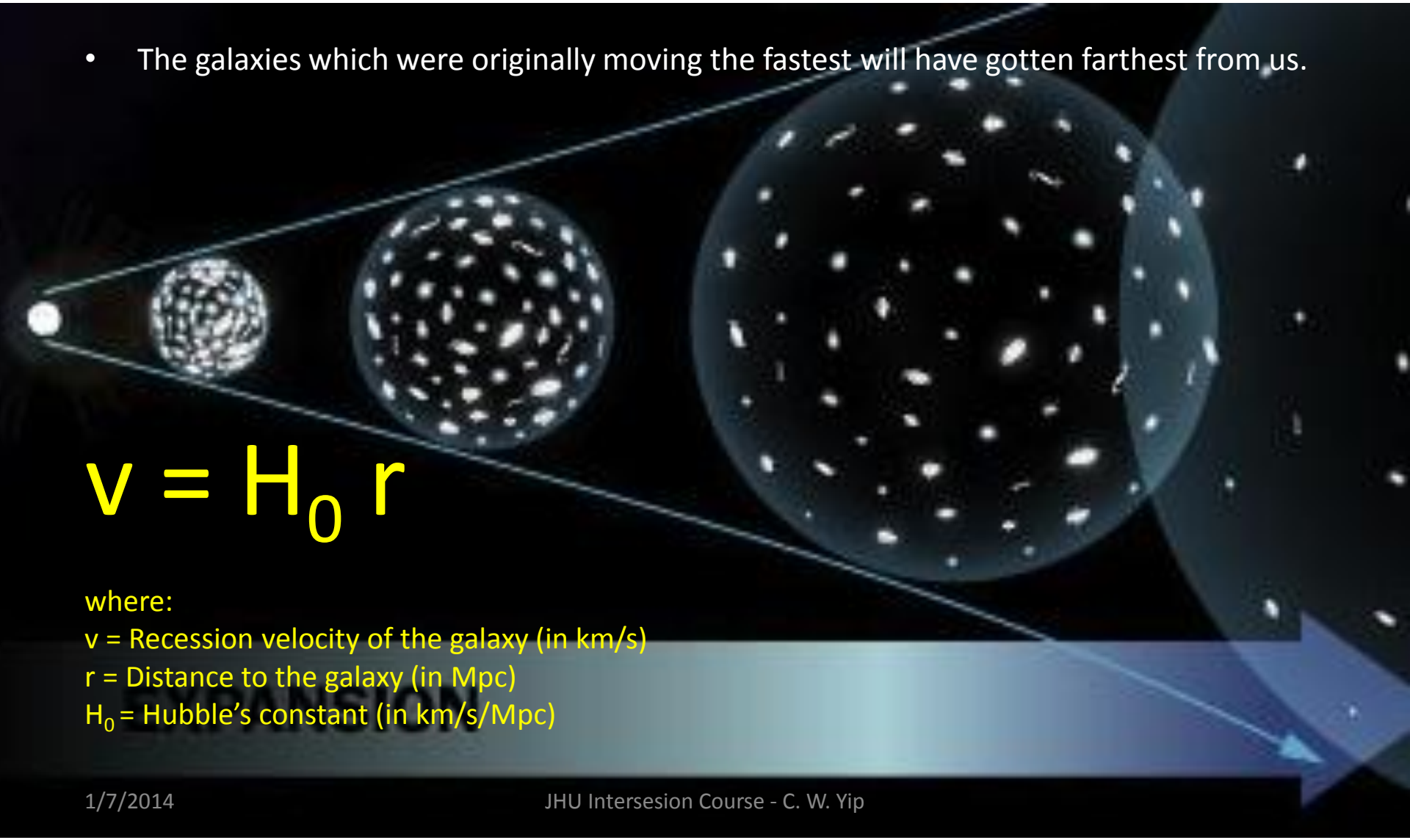


FIGURE 1

Velocity-Distance Relation among Extra-Galactic Nebulae.

Hubble's Law & Universe Expansion

- The galaxies which were originally moving the fastest will have gotten farthest from us.



$$v = H_0 r$$

where:
v = Recession velocity of the galaxy (in km/s)
r = Distance to the galaxy (in Mpc)
H_0 = Hubble's constant (in km/s/Mpc)

Hubble (1929)

- 2D Plot of Variables: v vs. r
- Linearity assumption
- Least-square fitting to get the slope (or, the Hubble's constant)
- Dynamical range
  - Small! (Local Universe)

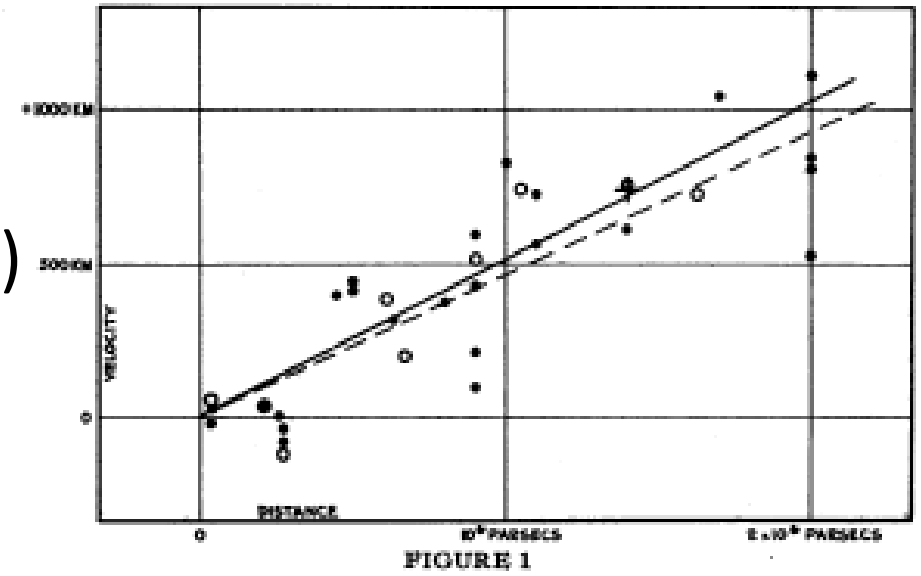
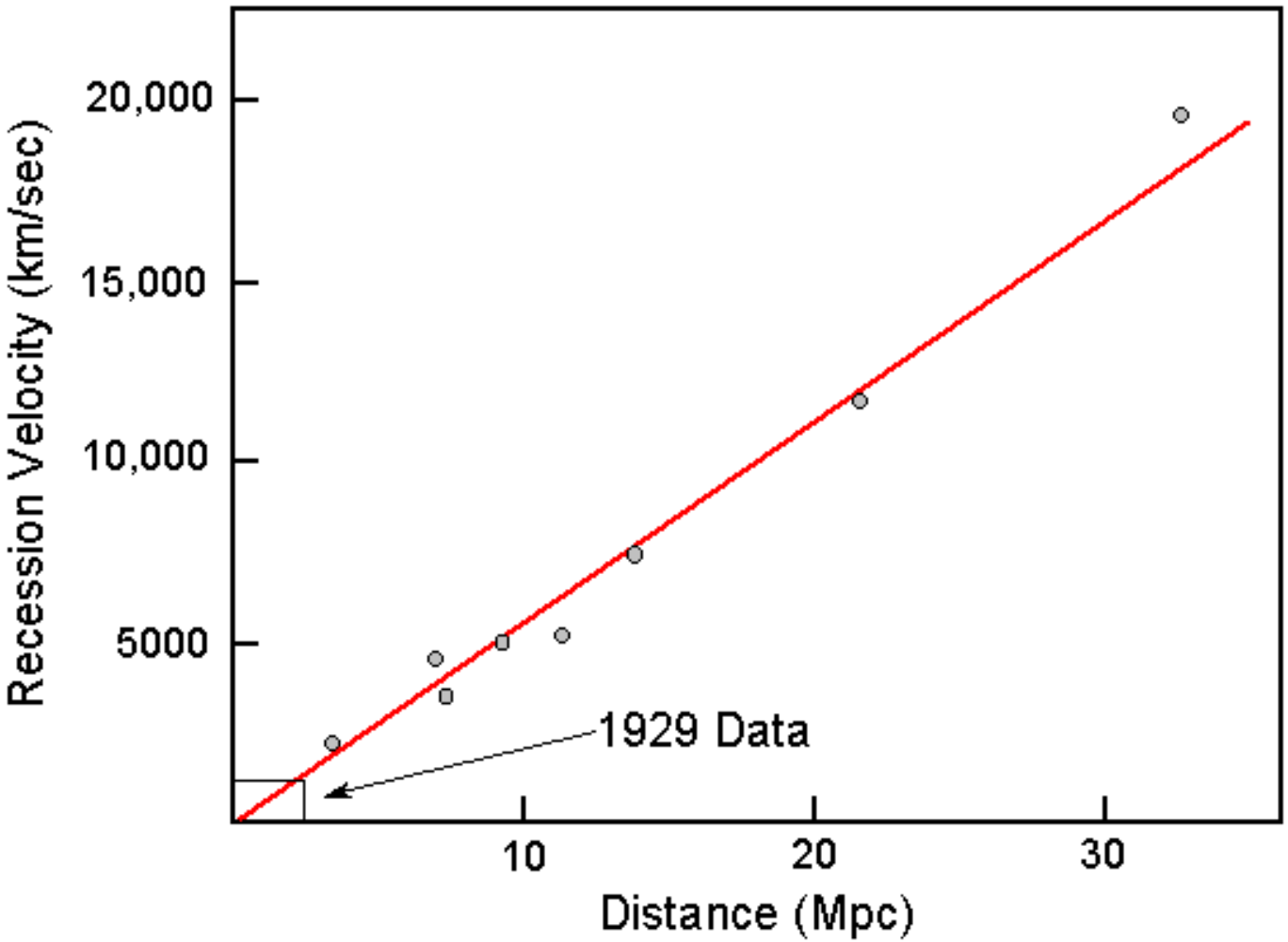


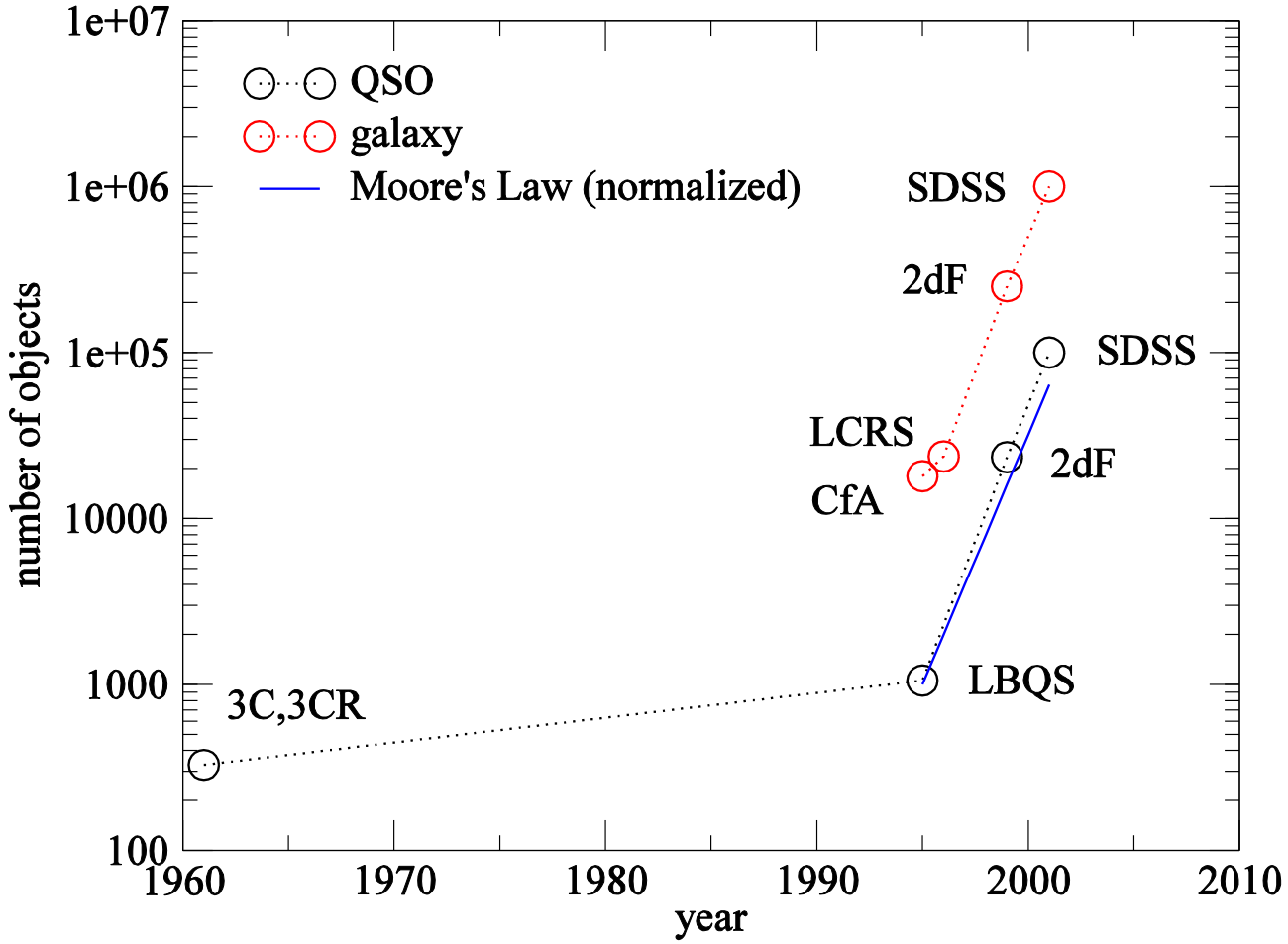
FIGURE 1

Velocity-Distance Relation among Extra-Galactic Nebulae.

Hubble & Humason (1931)



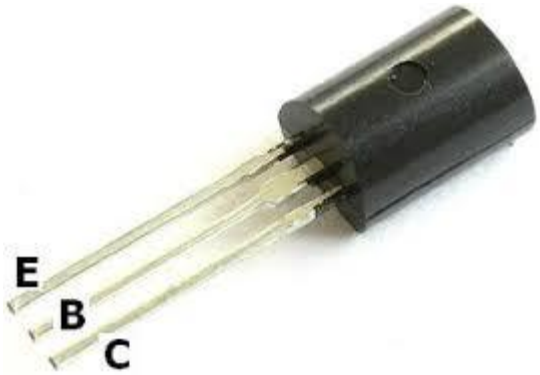
Moore's Law in Astronomy Data

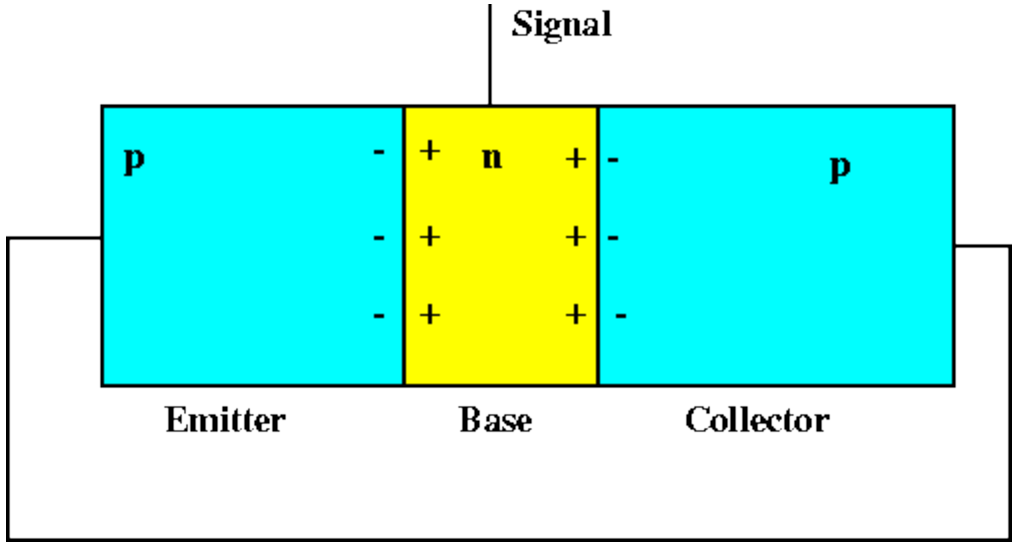


(Yip 2005)

Transistor

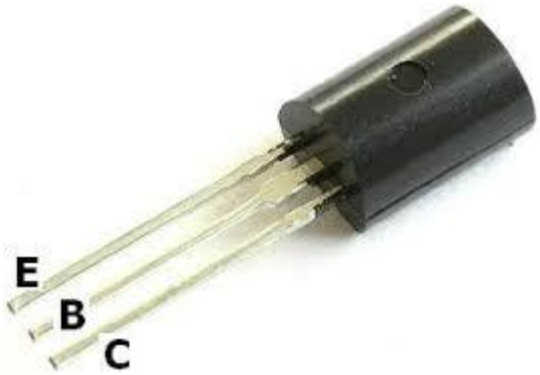
- 2 Main Functions
  - Amplifier
  - Switch (0/1)

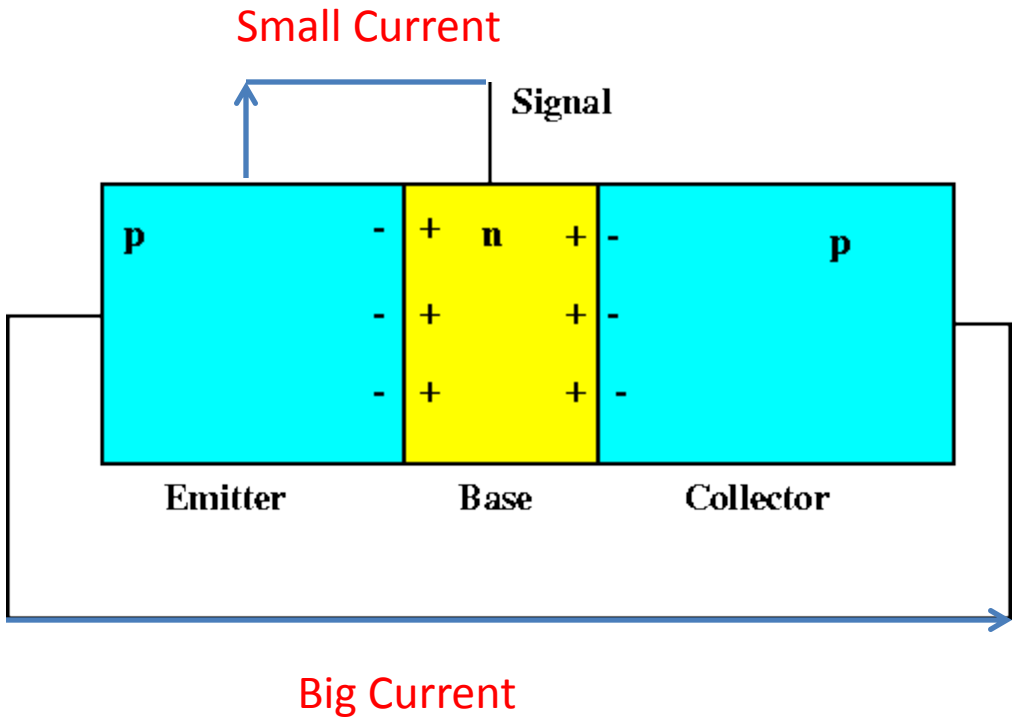




Transistor

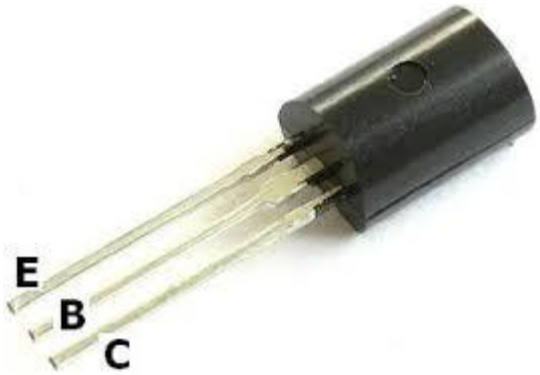
- 2 Main Functions
  - Amplifier
  - Switch (0/1)



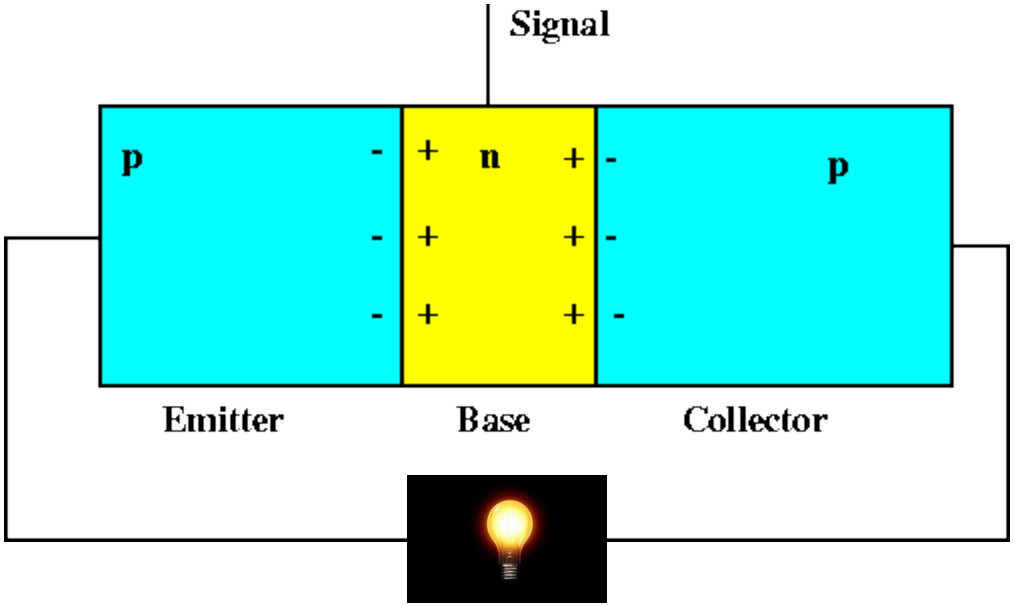


Transistor

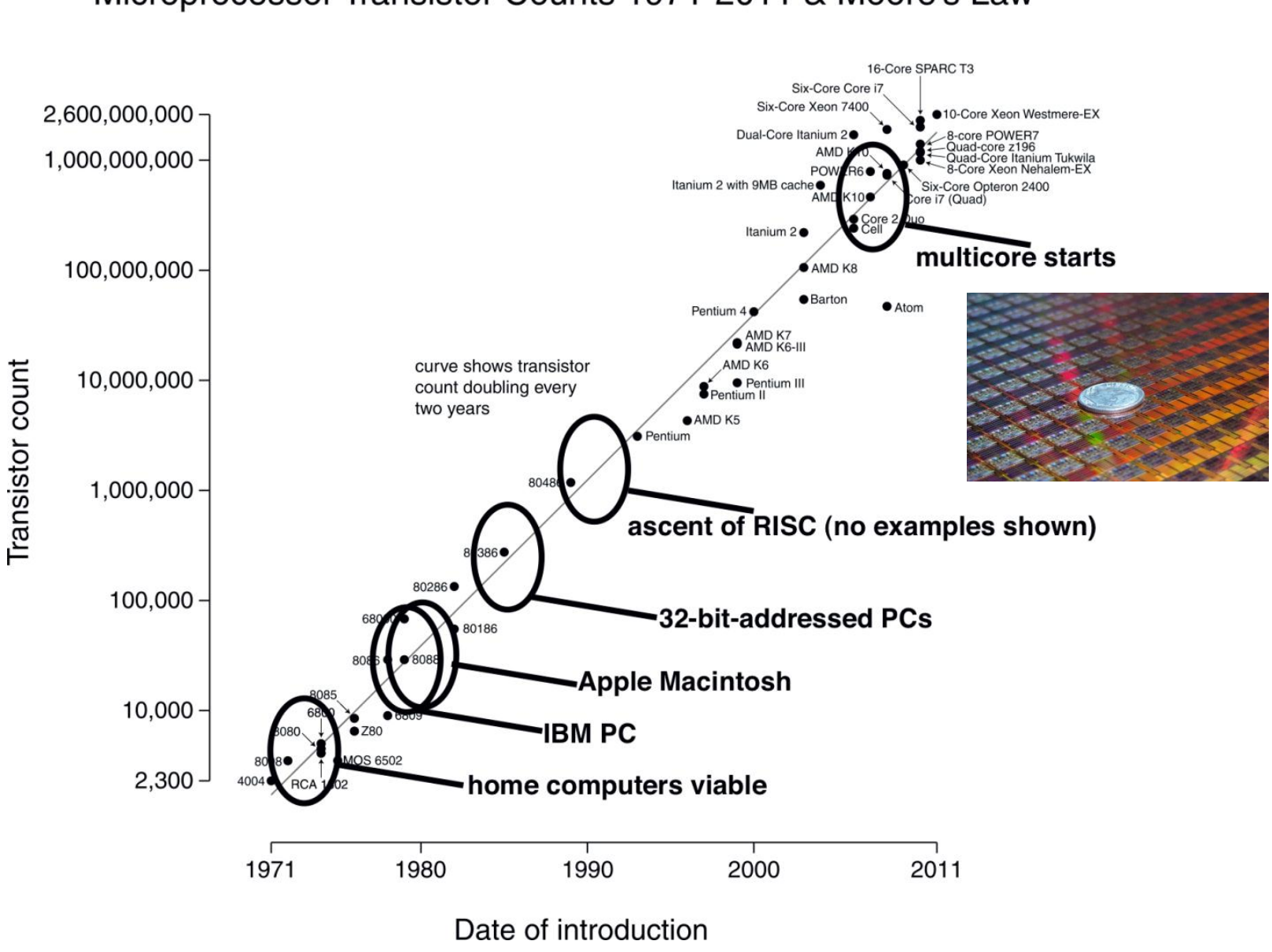
- 2 Main Functions
  - Amplifier
  - Switch (0/1)



Adjust Voltage



Zero Voltage Difference between E & C
=> Electrons flow freely



(http://homes.cs.ru.ac.za/philip/images/Transistor_Count_and_Moores_Law_2011.png)

Properties of Astronomical Objects probed by Sky Surveys

- Direct Observables:
  - Luminosity
  - Color
  - Size
  - Position on the sky
  - Etc.
- Derived Quantities:
  - Distance
  - Age
  - Metallicity (i.e., how much metal are present)
  - Etc. (much longer list)

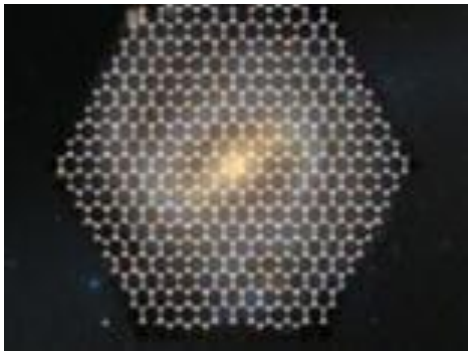
Modern Astronomy Sky Surveys

- We combine large telescopes and CCDs to:
  - Scan across the sky
  - Look deep into the universe
- Many current surveys are:
  - Wide-field
  - Multi-wavelength
  - Probing the time domain

- The instruments can be ground-based or space-borne.

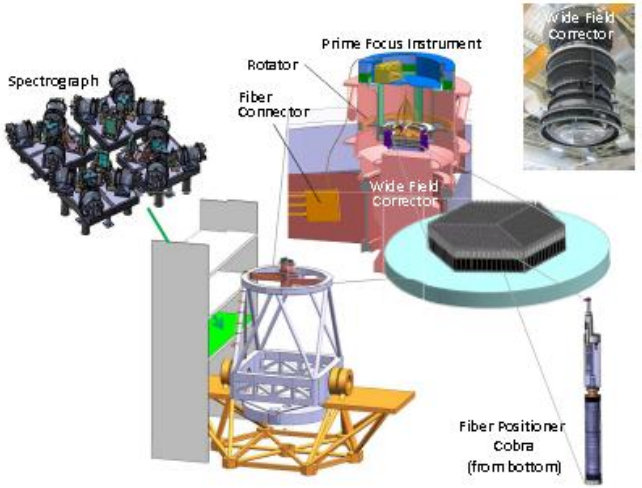
Current & Future Spectroscopic Surveys

- Multiplexing: many objects at a time (LAMOST, MaNGA, PFS)
- 2D In Situ: as a function of projected 2D position on galaxies (Integral Field Units, CALIFA, SAMI, MaNGA)



CALIFA

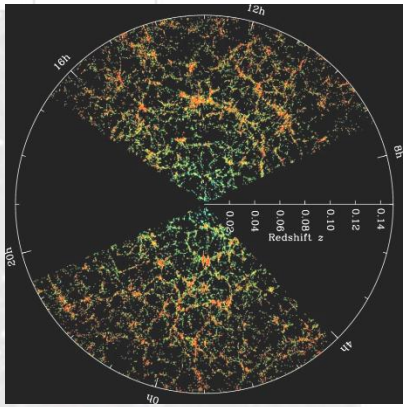
LAMOST



Subaru PFS

Sloan Digital Sky Survey (2000-)

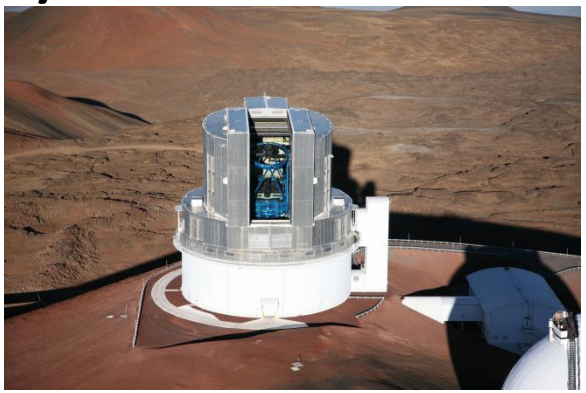
- Photometric + Spectroscopic Surveys
  - 11,000 square degree footprint (DR7)
  - 5.9 $\times 10^8$ u, g, r, i, z photometry
  - 1.6 $\times 10^6$ fiber spectra
- Phases
  - SDSS I (2000-05)
  - SDSS II (2005-08)
  - SDSS III (2008-14)
- Data are public
- Web interfaces for data download & exploration
  - SkyServer, DAS, etc.



(Galaxy Distribution)

Subaru Prime Focus Spectrograph (PFS; 2016-2017)

- High redshift version of SDSS
- 2,400 fiber array, 1.3° FOV
- 0.5 million galaxy spectra ($1.4 < z < 2.2$)
- 140,000 Lyα emitters ($2 < z < 7$)
- 50,000 QSOs ($3 < z < 7$)



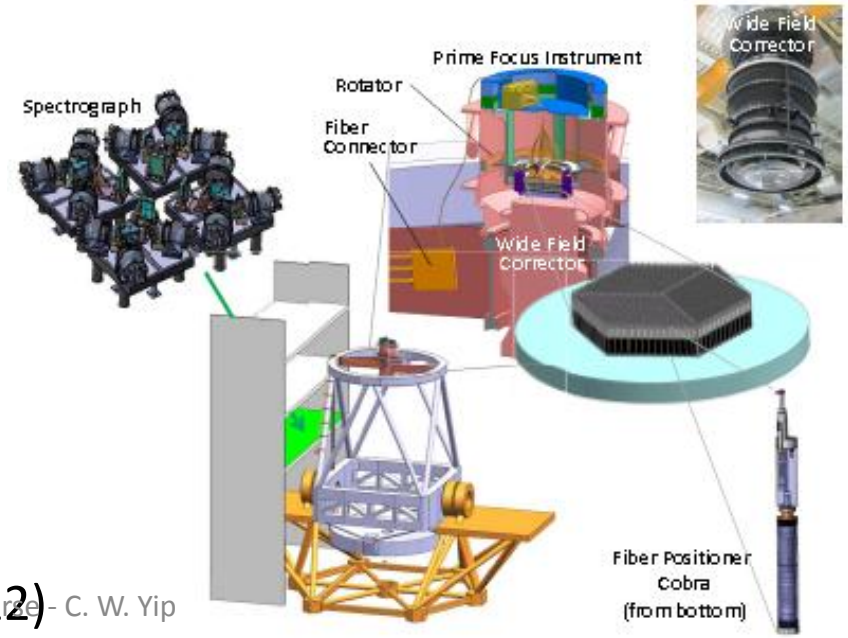
(8m telescope in Hawaii)

Spectral resolution:

3800-6700 Å
R = 1900

6500-10000 Å
R = 2400

9700-13000 Å
R = 3500



(Ellis et al. 2012)

Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; 2011-)

- 4m segmented telescope, 5° FOV (the Moon spans 0.5 °)
- 4,000 fiber spectra into 16 spectrographs
- 10 million fiber spectra, 10x more than SDSS

Spectral resolutions:

medium-low
R = 1000 - 2000

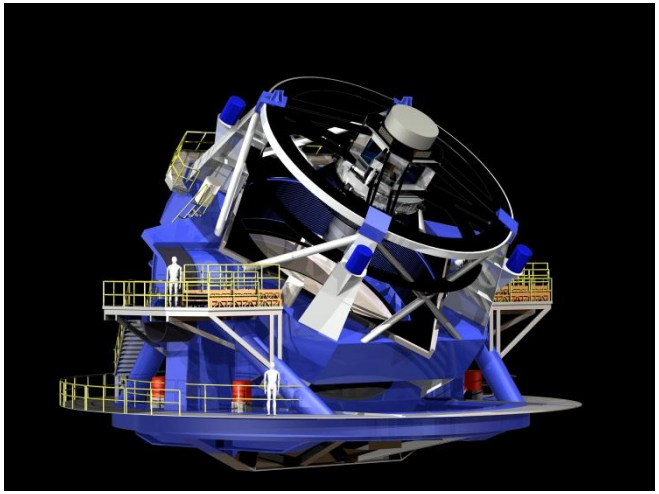
medium
R = 5000 – 10000

(Xinglong Station,
180 km north of Beijing)

Large Synoptic Survey Telescope (LSST; 2022-2032)

- Survey the whole sky every few nights.
- Main science goals: Matter in the Universe; Supernovae explosion; Hazardous near-Earth objects.

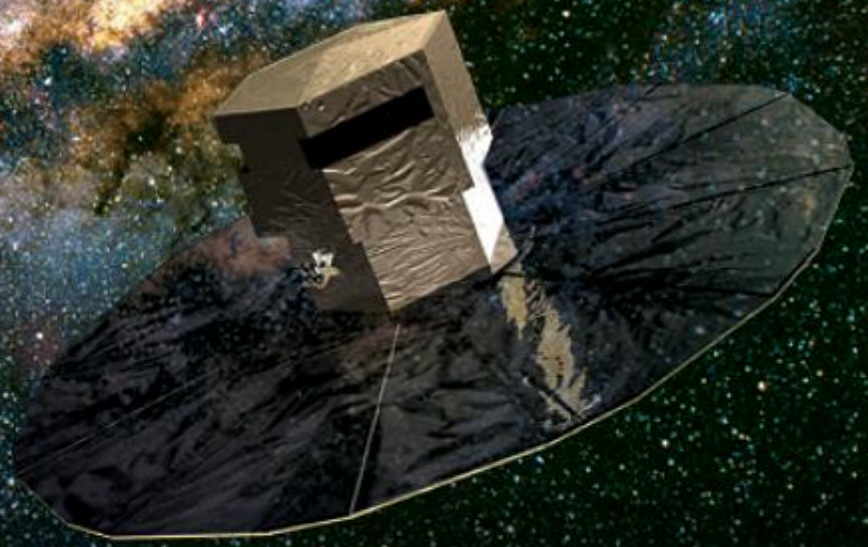
(Simulated LSST image)



(8m telescope in Chile)

GAIA (2013-2018)

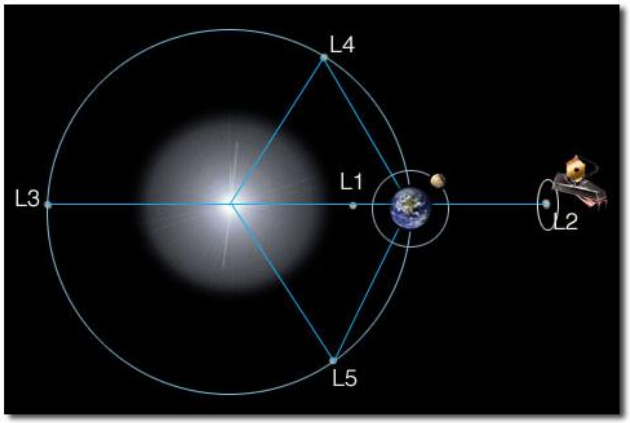
- Map 1 billion (1%) stars in Milky Way.
- Get both stellar positions and velocity (6 dimension)



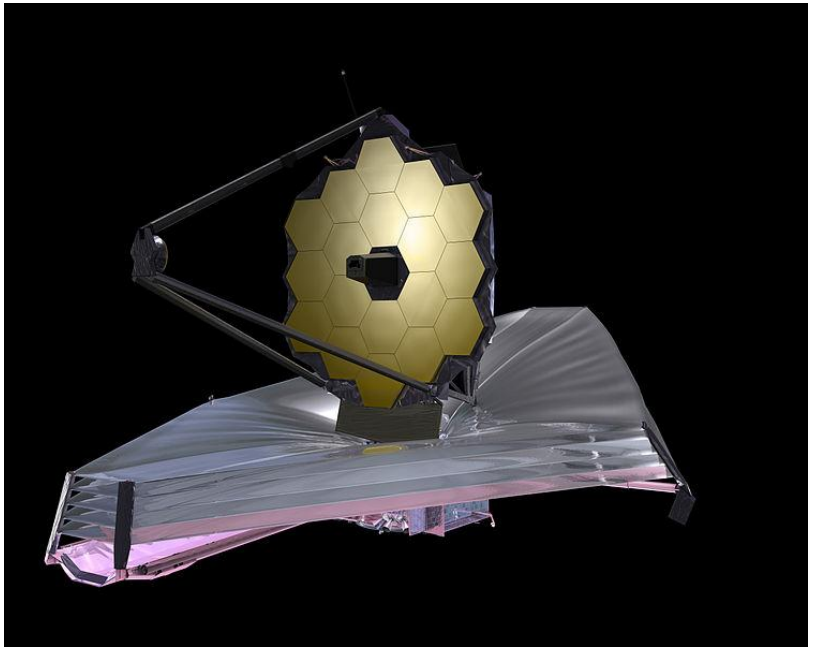
(European Space Agency)

James Webb Space Telescope (JWST; 2018-2028)

- Successor of Hubble Space Telescope.
- Infrared capabilities allows for probing distant universe.
- Many other topics…



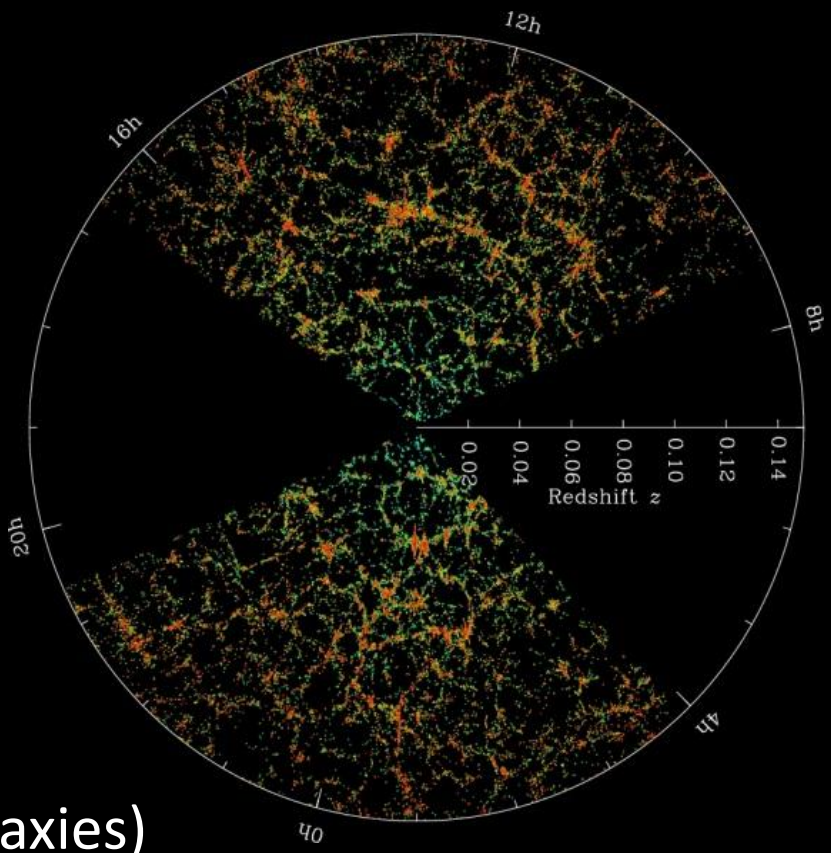
L_2 Lagrangian Point



6.2m Telescope

Challenges in Analyzing Astronomy Data

- 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)

Homework
2014 Jan 7

1. If a CCD has 1024×1024 pixels, express in reasonable units the total number of pixels.
2. GAIA will map 1 billion stars in our Milky Way. If the estimation of distance of each star takes 1 second when 1 computer is used, how much time (in reasonable units) is needed to obtain the distances for all stars in GAIA?
3. Install the statistical software R to your computer (laptop/desktop).
4. If applicable, describe briefly a data analysis exercise/project that you may have done previously (for example: What was the goal? How was the data taken and analyzed? What was the conclusion?). Figures are welcome.