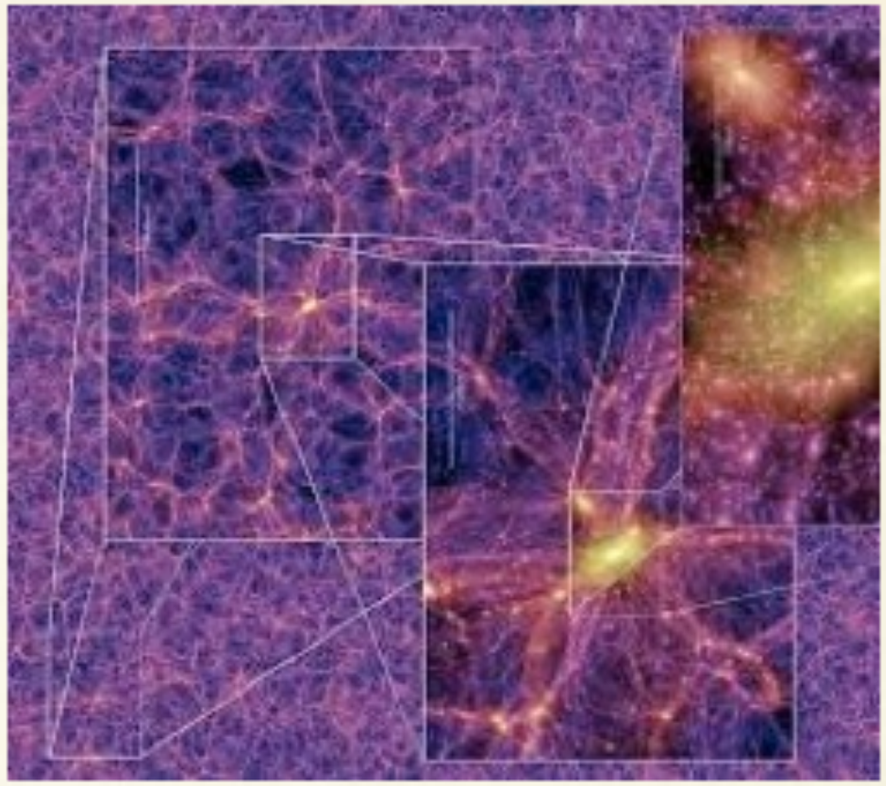
Astronomy 502
Data Mining and Machine Learning in Astronomy

University of Arizona
Fall 2017

Professor Xiaohui Fan

Tel: (520) 626-7558

fan@as.arizona.edu



Plan

- today: introduction and logistics
- next Monday: tools
  - Python notebook
  - astroML
  - github
  - datasets
- next Wednesday: statistics refresher
  - distributions
  - likelihood
  - Bayes' rule
  - frequentist vs. Bayesian views
- From Sept 6 (after labor day): student seminars and guest lectures

DATA NEVER SLEEPS 4.0

How much data is generated every minute? In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like Youtube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the internet every minute? See for yourself below.

2016 every MINUTE of the DAY

PRESENTED BY DOMO

Americans USE **18,264,840** MEGABYTES OF WIRELESS DATA

BUZZFEED USERS VIEW **159,380** PIECES OF CONTENT

SNAPCHAT USERS WATCH **6,944,444** VIDEOS

Netflix SUBSCRIBERS STREAM **86,805** HOURS OF VIDEO

TRA… **69,5…** WORDS

YOUTUBE USERS SHARE **400** HOURS OF NEW VIDEO

FACEBOOK MESSENGER USERS SHARE **216,302** PHOTOS

SIRI ANSWERS **99,206** REQUESTS

Amazon MAKES **$222,283** IN SALES

Tinder USERS SWIPE **972,222** TIMES

3,567,850 TEXT MESSAGES ARE SENT IN THE **U.S.**

Giphy SERVES **569,217** GIFS

TWITTER USERS SEND **9,678** EMOJI-FILLED TWEETS

Dropbox USERS UPLOAD **833,333** NEW FILES

THE WEATHER CHANNEL RECEIVES **13…** FO…

Big Data is growing fast

Annual growth rate **60%** Structured and unstructured data[1]

In social media alone, every 60 seconds **600** new blog posts are published, and **34,000** tweets are sent[2]

The digital universe will grow to **2.7ZB** in 2012, up **48%** from 2011, toward nearly **8ZB** by 2015[3]

The era of surveys...

- Standard: “What data do I have to collect to (dis)prove a hypothesis”?
- Data-driven: “What theories can I test given the data I already have?”

Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected

Zhang & Zhao (2015): http://datascience.codata.org/articles/10.5334/dsj-2015-011/

Alternative Careers: Leveraging your Astronomy Degree for Data Science

by Ben Cook | Jun 1, 2016 | Career Navigation, Personal Experiences | 0 comments

Big Data in Astronomy

Alongside the recent explosion of "Big Data" into the public consciousness, there has been a similar transition into the age of "Big Astronomy". Astronomers have always been adept at drawing conclusions using advanced statistics and data analysis. Now, with the advent of extremely large simulations like Illustris and surveys like the upcoming LSST, astronomers are increasingly gaining experience in dealing with datasets vastly larger than could ever hope to fit on a single computer.

For early career astronomers looking for advice, I think you can do no better than look at the posts made by Jessica Kirkpatrick, who obtained a PhD in Astronomy and then became a data scientist at Microsoft/Yammer, and I understand she has since taken a position as Director of Data Science at the education start-up InstaEDU.

The term "Data Scientist" is extraordinarily broad. For example, the post "What is a Data Scientist?" describes some of the Data Analyst roles a Data Scientists may play:

- Derive business insight from data.
- Work across all teams within an organization.
- Answer questions using analysis of data.
- Design and perform experiments and tests.
- Create forecasts and models.
- Prioritize which questions and analyses are actionable and valuable.
- Help teams/executives make data-driven decisions.
- Communicate results across the company to technical and non-technical people.

https://astrocompute.wordpress.com/2016/01/29/how-does-an-astronomer-become-a-data-scientist/

LSST: a digital color movie of the Universe...

LSST in one sentence:

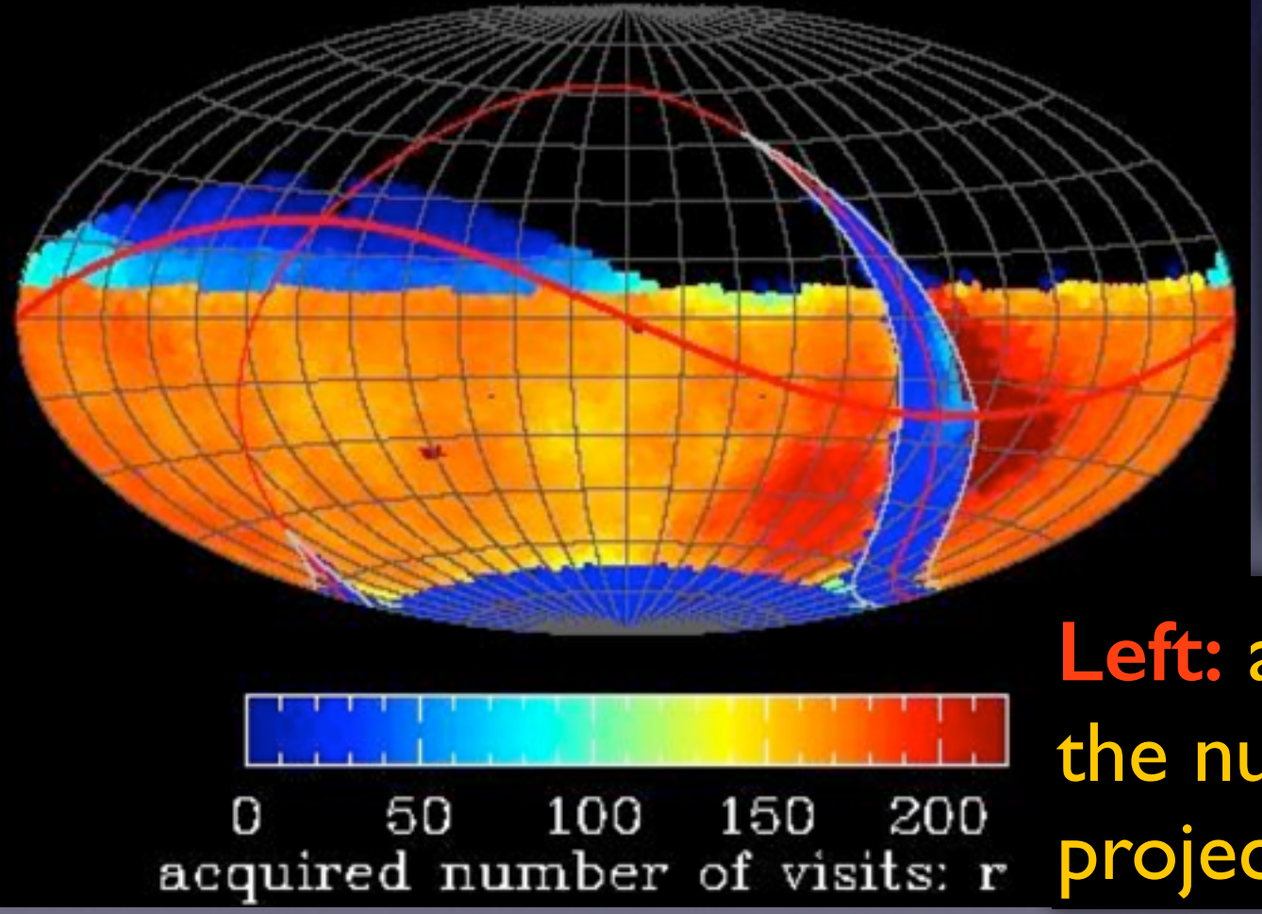
An optical/near-IR survey of half the sky in ugrizy bands to r~27.5 based on ~800 visits over a 10-year period:

More information at www.lsst.org and arXiv:0805.2366

A catalog of 20 billion stars and 20 billion galaxies with exquisite photometry, astrometry and image quality!

Basic idea behind LSST: **a uniform sky survey**

- 90% of time will be spent on a uniform survey: every 3-4 nights, the whole observable sky will be scanned twice per night
- after 10 years, half of the sky will be imaged about 1000 times (in 6 bandpasses, ugrizy): a digital color movie of the sky
- ~100 PB of data: about a billion 16 Mpix images, enabling **measurements for 40 billion objects!**



LSST in three words:
deep wide fast.

Left: a 10-year simulation of LSST survey: the number of visits in the r band (Aitoff projection of eq. coordinates)

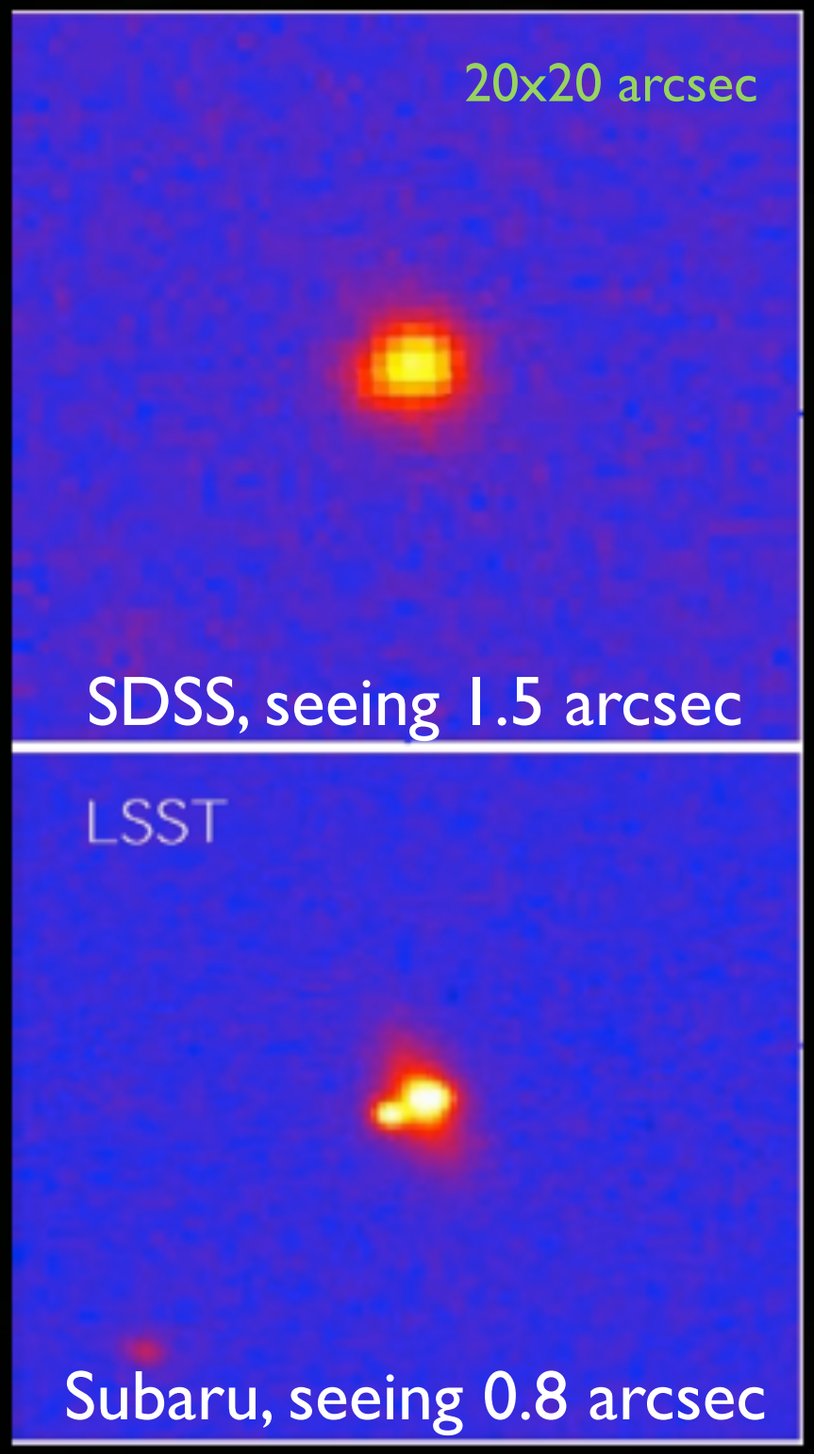
3x3 arcmin, gri

SDSS

Deep Lens Survey (r~26)

(almost) like LSST depth (but tiny area)

20x20 arcsec; lensed SDSS quasar (SDSS J1332+0347, Morokuma et al. 2007)



LSST Science Themes

- dark matter, dark energy, cosmology
  - galaxy large scale structure
  - gravitational lensing
  - supernovae
- Time domaine
  - cosmic explosions
  - variable stars
- The solar system structure
  - asteroids
  - KBOs
- Milky Way Structure
  - stars
  - near-field cosmology

Science questions vs. big data tools

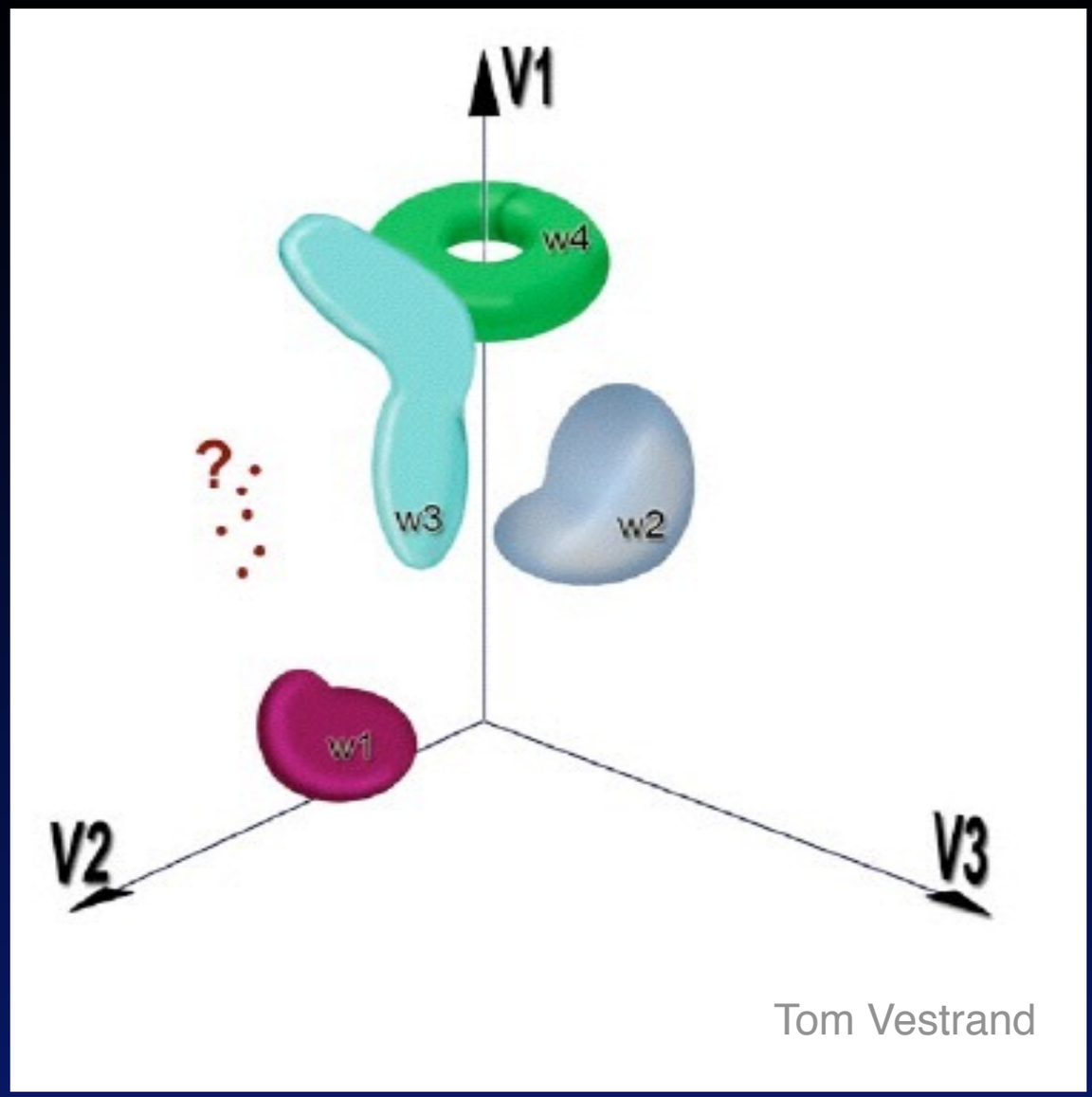
- large scale structure - clustering
- galaxy evolution - density estimates
- cosmological parameters - regression
- spectroscopic/photometric classification - dimension reduction
- rare object detection - classification, outliers
- time domaine - time series analysis
- etc.

Data Challenges in the LSST Era

- large data volume (petabytes)
- large number of objects (billions)
- high dimensionality (thousands)
- unknown statistical distributions (non-Gaussianity)
- Time-series data (irregular sampling)
- Heteroscedastic errors, truncated, censored, and missing data

The bottleneck is not any more data availability but instead our ability to extract useful and reliable information from data.

- ? Characterize the known (clustering)
- ? Assign the new (classification)
- ? Discover the unknown (outlier detection)



Tom Vestrand

Benefits of very large data sets:

- best statistical analysis of “typical” events
- automated search for “rare” events

In this class, we will learn how to do all that.

Class format

- introduction (XF)
- student-led seminars on selected topics
- guest lectures
- final student projects

textbook and references

- textbook
  - Statistics, data mining and machine learning in astronomy, by Ivezic et al. (I14)
  - codes, figures etc at astroML.org
  - ebook: http://sabio.library.arizona.edu/record=b7106910~S9
- references
  - Statistics in Theory and Practice by Lupton
  - Practical statistics for astronomers by Wall and Jenkins
  - other references listed in I14
- websites (including notebooks from other classes)
  - UW class: https://github.com/uw-astr-324-s17/astr-324-s17
  - Drexel class: http://www.physics.drexel.edu/~gtr/teaching/physT480
  - LSST data science fellowship program: https://github.com/LSSTC-DSFP/LSSTC-DSFP-Sessions

tools

- computing and coding
  - we will use python (anaconda python 2.7 to be compatible with I14 and astroML)
  - we will use python mostly in the form of Jupyter_Notebooks
- python packages
  - NumPy; SciPy; Scikit-learn; Matplotlib
  - you might need astroPy for some applications later on
- machine learning codes: www.astroML.org
- dataset for exercise: mostly SDSS photometric catalogs and spectra (next Tuesday)
- class website and github depository:
  - http://sancerre.as.arizona.edu/~fan/Home/AST502.html
  - https://github.com/UA-ast502-2017/classnotebooks

Topics I: Statistics

- Chap 3: Probability and statistical distributions (XF, next Wed)
- Chap 4: Classical Statistical Inference:
  - Maximum likelihood estimation
  - confidence estimates
  - hypothesis testing
  - non-parametric modeling
- Chap 5: Bayesian Statistical Inference:
  - Bayesian priors and model selection
  - Markov Chain Monte Carlo (MCMC)

Topic 2: Data Mining (characterizing data)

- Chap 6: Searching for structure in point data
  - density estimation (non-parametric and parametric)
  - clustering and correlation function
- Chap 7: Dimensionality and its reduction
  - Principle Component Analysis
  - manifold learning
  - independent component analysis

Topics 3: Machining Learning (interpreting data)

- Chap 8: Regression and model fitting
  - linear models; non-linear models
- Chap 9: Classification
  - generative models
  - support vector machine; decision trees
- Chap 10: Time Series Analysis
  - Periodic; localized (bursts); stochastic processes

Student Lectures

- I will email the class list of topics and set up poll
- each student should choose three topics of interests to present
- student should see me a week before the lecture to discuss
- basic format:
  - review basic concepts
  - introduce basic computational tools/codes
  - lead exercise demonstrations - could be examples from astroML, or your own data
- lectures should be written in the format of a python notebook (feel free to use UW/Drexel notebooks as a template/starting point)
- notebook needs to be posted on class github before the lecture, and update after the lecture (other students are encouraged to send feedback)

Guest lectures

- NOAO Data Lab: (Stephanie Juneau and Knut Olsen), in late Sept/early Oct
  - 1. Data Lab
  - 2. database and publication
- LSST (Zeljko Ivezic and/or Beth Willman), Nov
- Spectroscopic surveys (Adam Bolton), early Nov
- ARTARES (event alert for LSST, Thomas Matheson), Late Nov

Final Projects

- we have 18 students - divide to 3-4 group projects
- application of data mining and machine learning techniques to:
  - a more complex data of your choice
  - utilizing a number of different approaches, and comparing them in term of accuracy, interpretability, simplicity, speed etc.
    - submit a joint report in the form of a python notebook and make a 30 min presentation at the end of the semester
- topics might be aligned with the individual chapters, or of your own (group) choices
- we will discuss the final project topics in early November.

Grades

- class attendance: 33.3%
- student lecture: 33.3%
- final project: 33.3%

homework

- Please make sure your python installation is update, your jupyter_notebooks work
- please install astroML, scikit-learn and make sure your numPy, matplotlib, sciPY work
- use anaconda. see: http://python4astronomers.github.io/index.html for help.