GAUTHAM NARAYAN

FUNDAMENTALS OF DATA SCIENCE
WEEK 0

0.0

LOGISTICS

WAYS TO REACH ME

- Slack: #astr596_fds (if you are auditing, let me know and I'll add you)
- Office Hours: MW (3-5 PM) in Astro 129
- GitHub: https://github.com/gnarayan/ast596_2020_Spring (everything course related goes here)
- Email: Good for official-ish messages like "I'm missing class/a test."

WHERE ELSE TO GET HELP

- Slack: #astr596_fds
- Resources in the syllabus!
- Facebook: Python Users in Astronomy
- stackoverflow (can be toxic)

RULES OF ENGAGEMENT

- Being able to explain a concept is the ultimate test of understanding, and it's entirely possible to believe we have an understanding right up to the moment we try to explain it to someone else.
- **Therefore, asking questions during class is providing a service.**
- No feigning surprise: To enable everyone to be comfortable saying "I don't understand", please resist the urge to act surprised when someone admits to not knowing something.
- No well-actually's: A well-actually happens when someone says something that's almost - but not entirely - correct, and you say, "well, actually..." and then give a minor correction. This is especially annoying when the correction has no bearing on the actual conversation.
- No subtle-isms: It goes without saying that overt expressions of racism, sexism, homophobia, transphobia or other biases are unacceptable. We also need to avoid subtle expressions of bias that can make others uncomfortable and that are of no benefit to the learning environment. For example, the expression "It's so easy my grandmother could do it" is a subtle-ism.

SYLLABUS

- Homework (40%) - nothing this week - weekly assignments starting next Tuesday, posted to GitHub. Due before the following Tuesday's class. It's OK if you can't finish it all right after Tuesday - we will cover additional material on Thursday. You can drop one for any reason.
- Exams (30% x 2) - same style as homework, but on all material up to that point.
- Textbooks - links to PDF versions are in syllabus. Others are linked in this slideshow. Dead tree versions are cheaper through Amazon. (http://www.astroml.org/ for course primary resource)

LEARNING OUTCOME

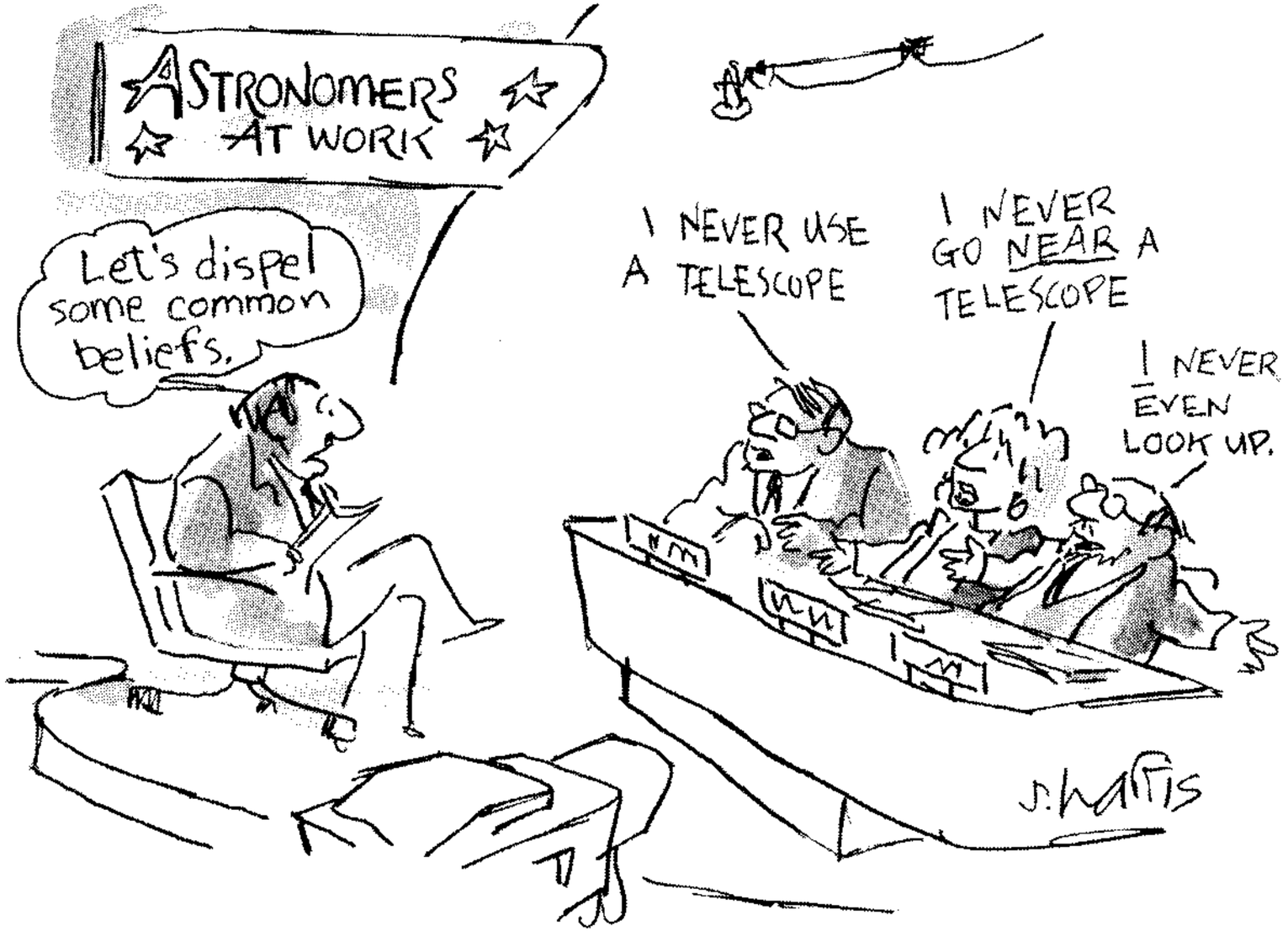
BY THE END OF THE CLASS, YOU SHOULD BE ABLE TO FORMULATE AN APPROPRIATE ANALYSIS PLAN FOR A RESEARCH QUESTION, SELECT GATHER AND PREPARE THE DATA, AND BUILD A MODEL TO DESCRIBE IT.

OTHER FAQS

- Is there an attendance policy?: Nope
- Can I work with others?: Yes, you are encouraged to! If you can't find a group to work with, come talk with me.
- Will I get my homework back before the next one is due?: That's the hope, but we've got no graders for class, so occasionally not. If I fall behind, I might make homework group assignments.
- Is there any way to get extra credit?: Not planned, and not very likely
- Do you grade on a curve?: No one will ever look at your GPA, and there's a going to be a homework problem about why this is statistically unsound.
- How do I get an A?: No, also one will ever look at your GPA!

0.1

WHAT IS DATA SCIENCE?



It’s quite likely that you will be the last generation of astronomers to use telescopes directly.
We are moving from:
The classical model - “What data do I have to collect to (dis)prove a hypothesis?”
To the data-driven model - “What theories can I test given the data I already have?”

3x3 arcmin, gri
LSST is 3x3 deg.

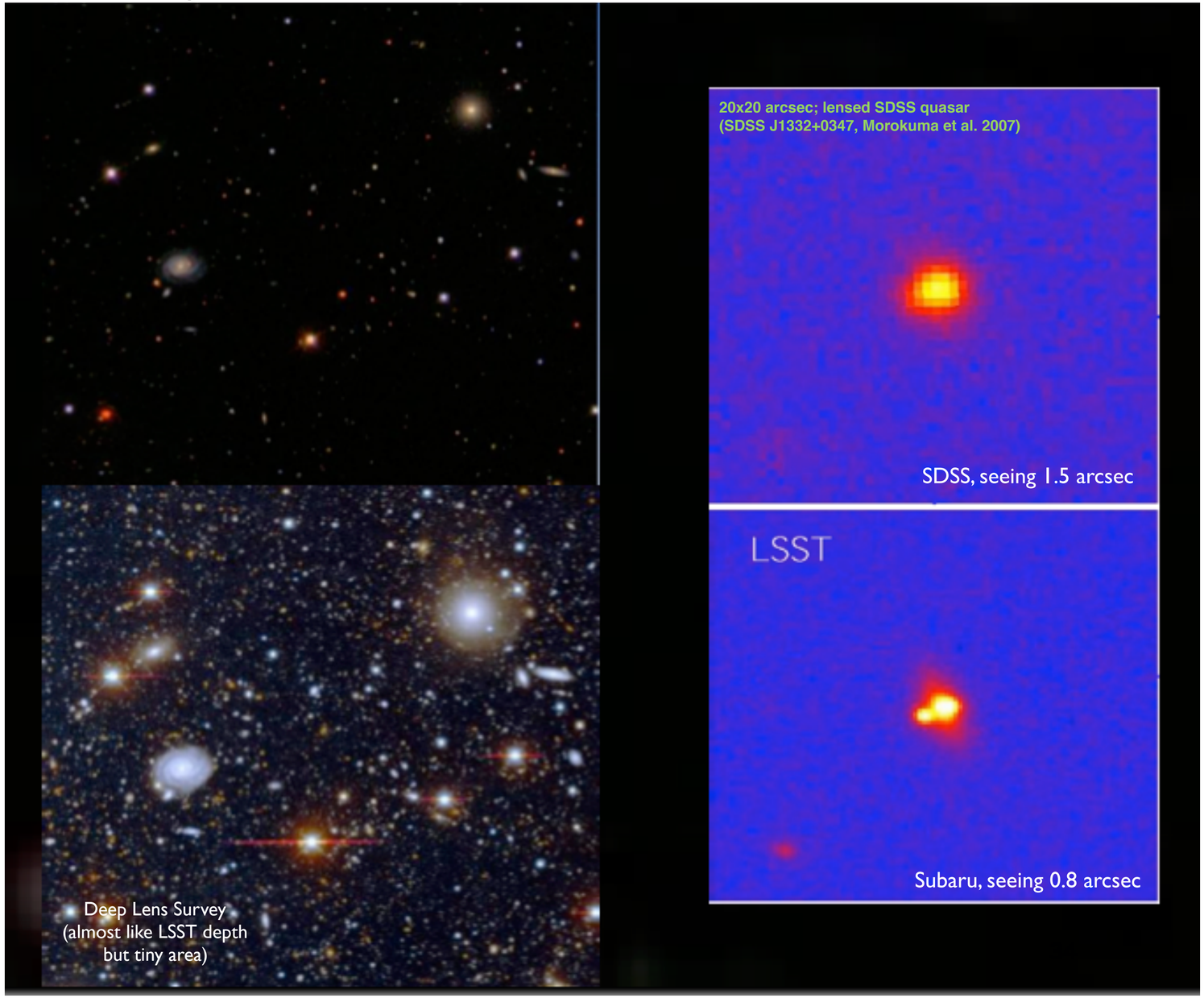
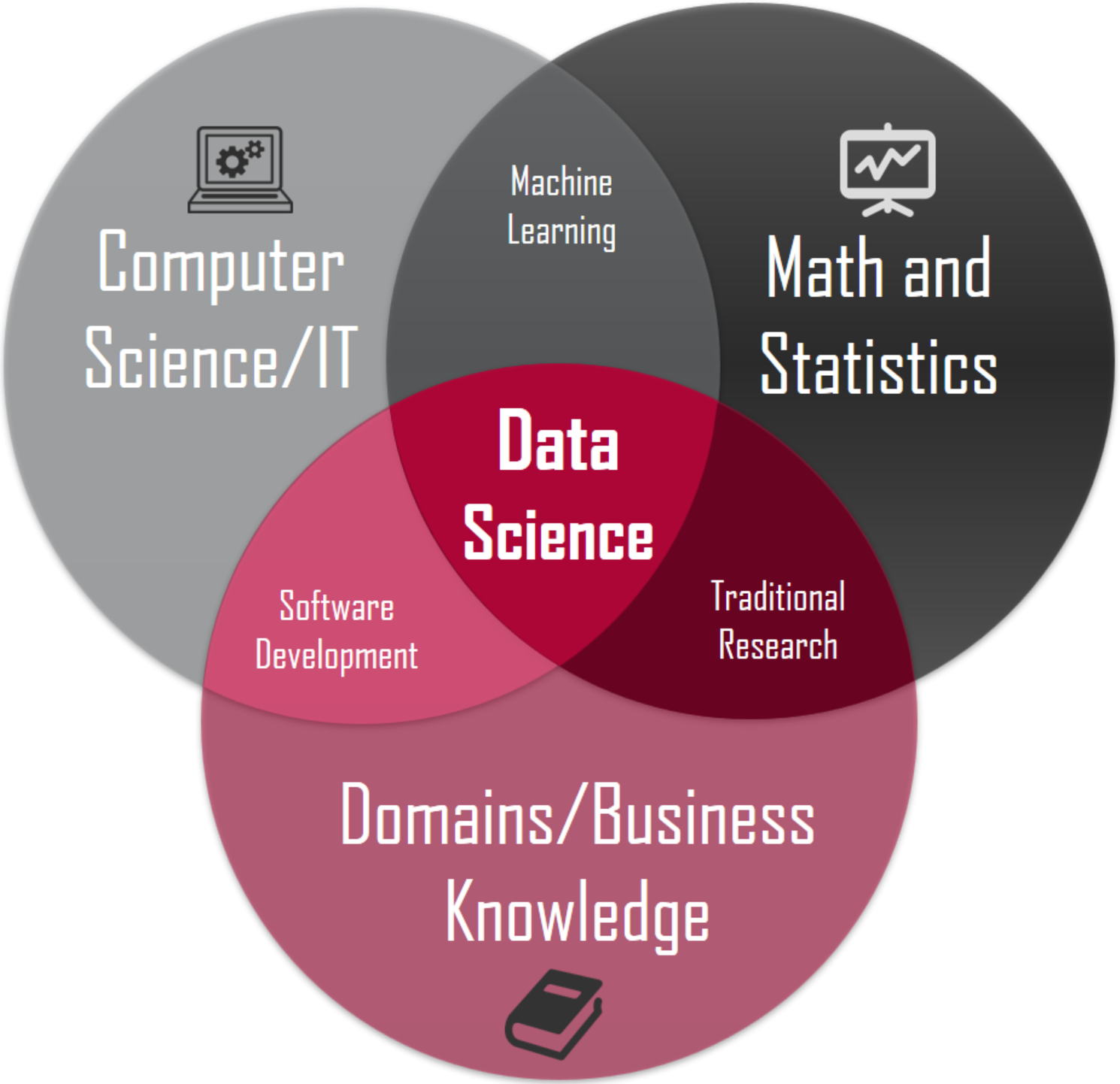


Image credit: Zeljko Ivezic

BUT THESE SURVEYS ARE MORE COMPLEX

- Ever increasing data volume and complexity- SDSS is ~30 TB;
  LSST will be one SDSS per night, or a total of >100 PB of data (40 billion objects)
- Sophisticated analysis, need for reproducibility - with the increasing data complexity, analysis becomes more complex, too; what do we do in case of disagreement?
- Open-source approach improves efficiency
  - we are not data starved any more!
  - the bottleneck for new results is in human resources and analysis tools
  - nobody has an unlimited budget; collaborate and share!


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

github

reproducibility

https://github.com

Reproducible research means:

all numbers in a data analysis can be recalculated exactly (down to stochastic variables!) using the **code** and **raw data** provided by the analyst.

Claerbout, J. 1990,

Active Documents and Reproducible Results, Stanford Exploration Project Report, 67, 139

allows reproducibility through code distribution

Credit: Federica Bianco

github

version control

https://github.com

the Git software

is a distributed *version control system:*

a version of the files on your local computer is made also available at a central server.

The history of the files is saved remotely so that any version (that was checked in) is retrievable.

allows version control

Credit: Federica Bianco

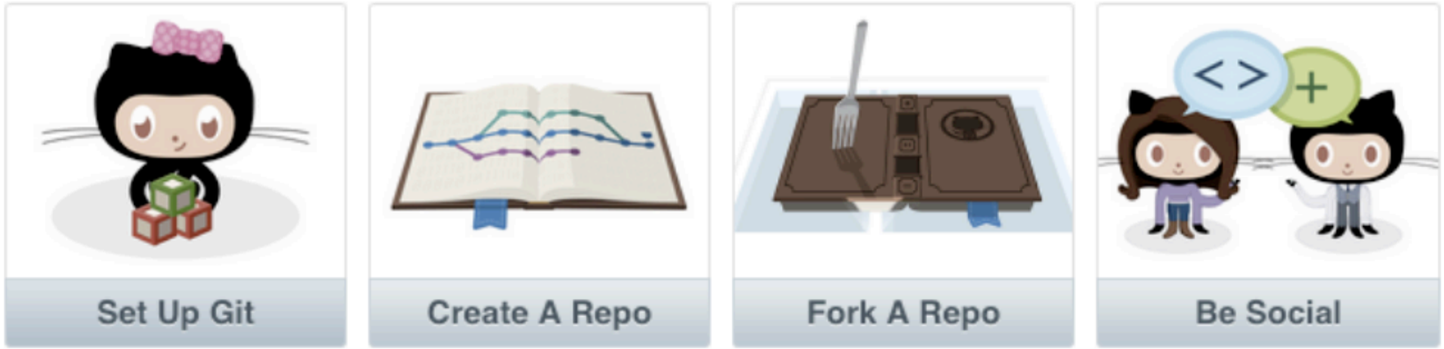
github

collaborative platform

https://github.com

collaboration tool

by fork, fork and pull request, or by working directly as a collaborator



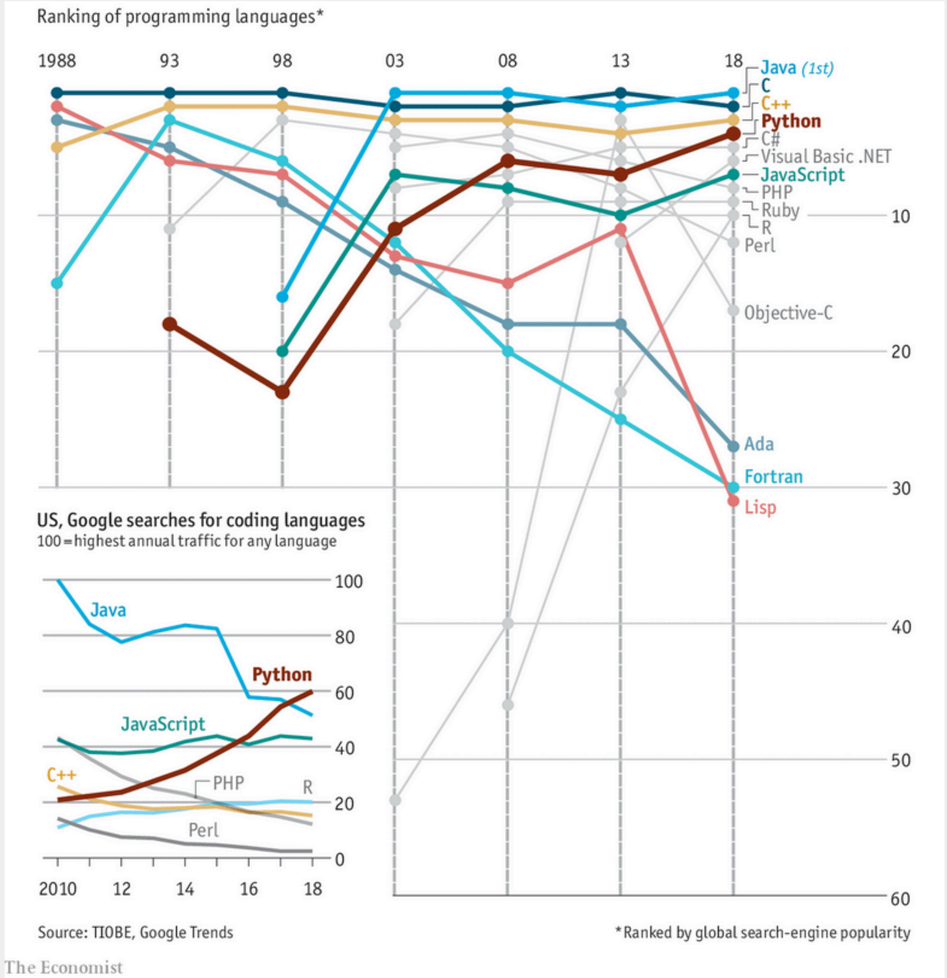
allows effective collaboration

Credit: Federica Bianco

python

- intuitive and readable
- open source
- support **C** integration for performance
- packages designed for science:
  - scipy
  - statsmodels
  - numpy (computation)
  - sklearn (machine learning)

https://www.economist.com/graphic-detail/2018/07/26/python-is-becoming-the-worlds-most-popular-coding-language

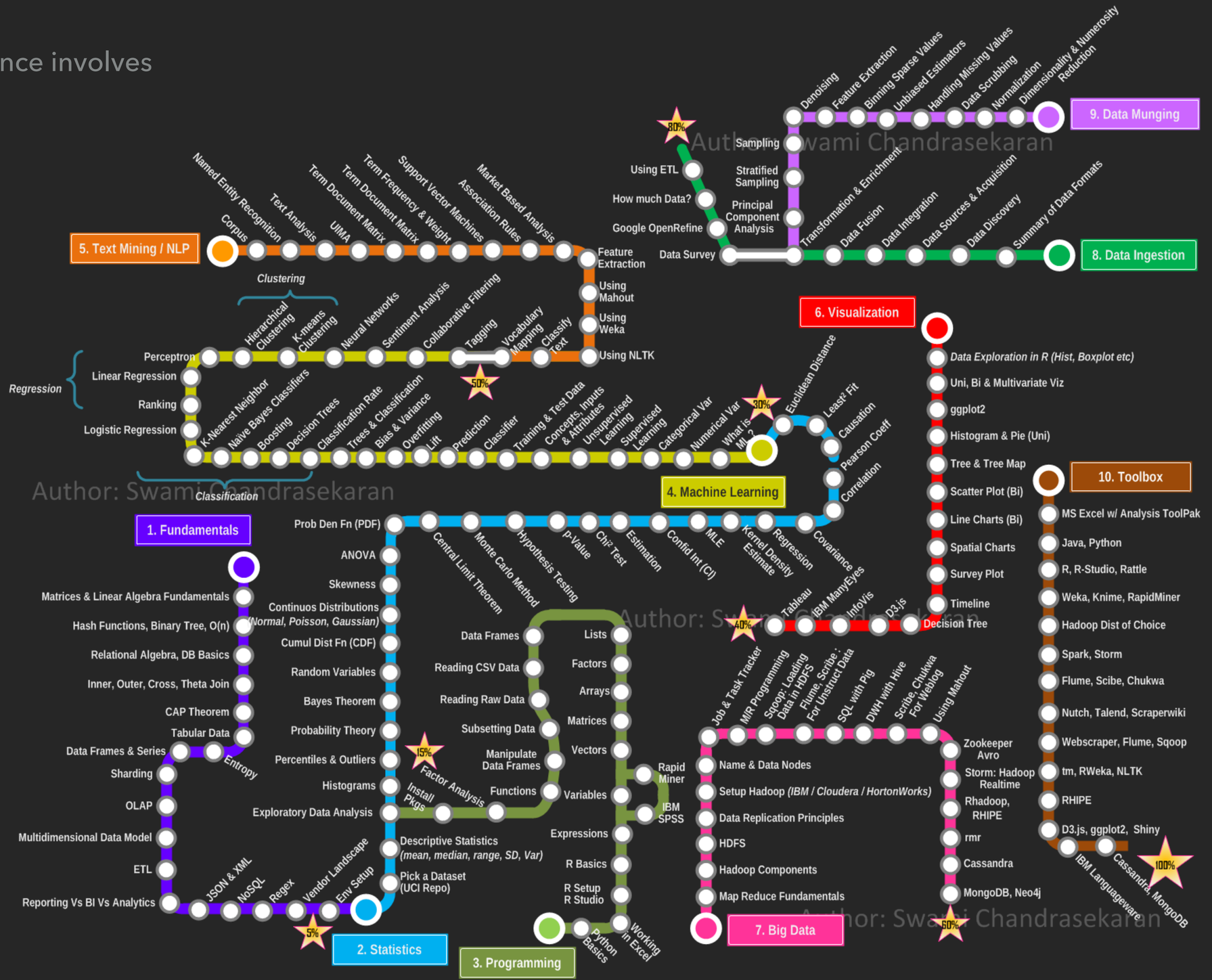


Credit: Federica Bianco

ASTRO DATA ANALYSIS CHALLENGES

- Large data volume (petabytes)
- Large numbers of objects (billions)
- Highly multi-dimensional spaces (thousands)
- Unknown statistical distributions
- Time-series data (irregular sampling)
- Heteroskedastic errors, truncated, censored and missing data
- Unreliable quantities (e.g. unknown systematics and random errors)

Stuff that data science involves



Stuff that data science involves

Stuff we are not covering

Stuff that data science involves

Stuff we are not covering/I'm assuming you know/you'll pickup (hopefully)

Programming

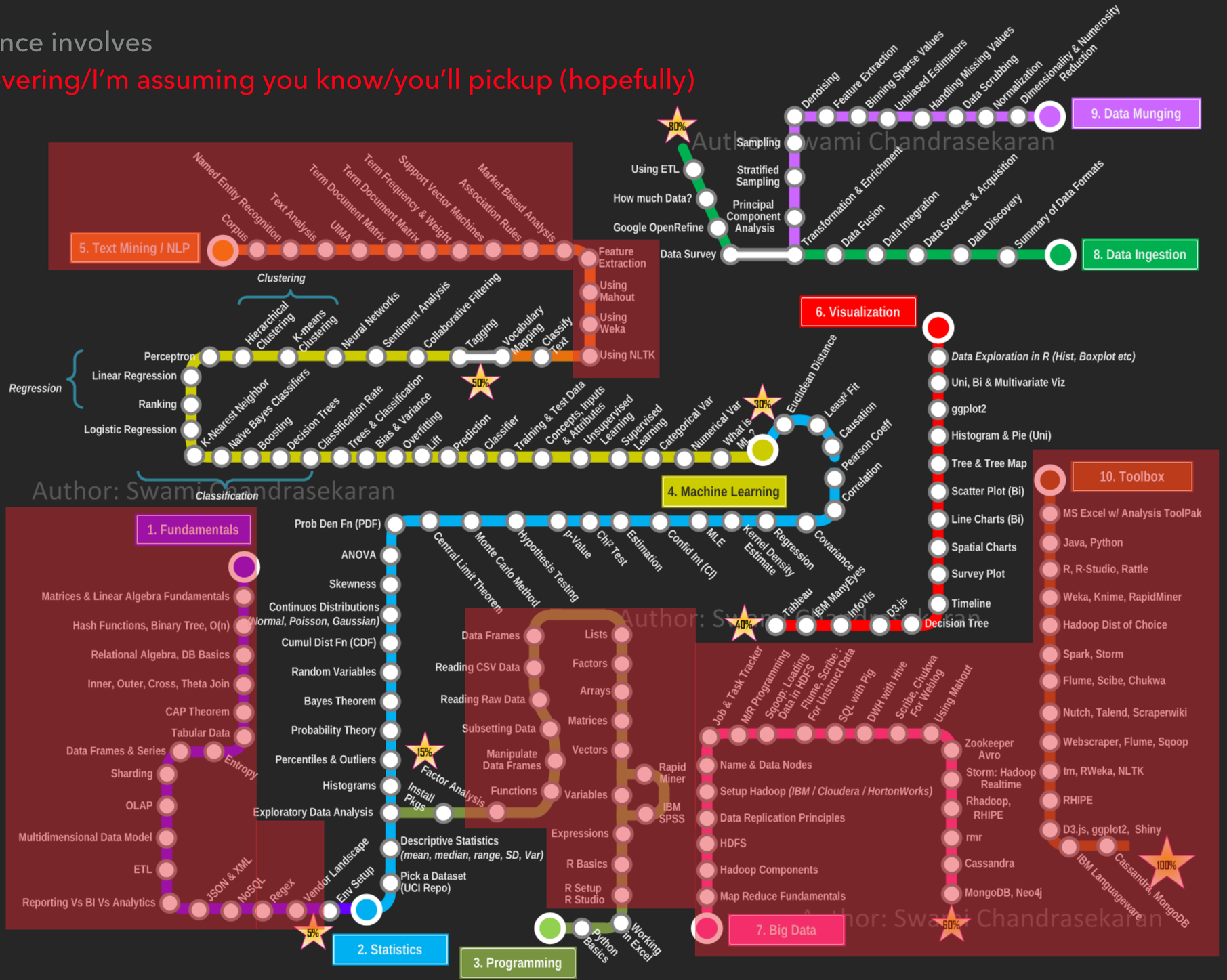
Statistics

Data ingestion

Data munging

Machine learning

Visualization



Python

Probability Distributions Uncertainty MCMC

Regression (linear/template)

classification

clustering

anomaly detection

CLASS SCHEDULE (subject to revision

- **WEEK 0 (Jan. 23rd)**
  First steps, crash course in python
- **WEEK 1 (Jan. 28th, 30th)**
  Probability distributions, descriptive statistics, the Central Limit theorem and when it doesn't hold, robust statistics, and hypothesis testing (ICVG Ch. 3, FB Ch. 2)
- **WEEK 2 (Feb. 4th, 6th)**
  Statistical inference, frequentist properties such as unbiasedness & the Cramér–Rao bound, consistency, asymptotic limits, mean-squared errors (ICVG Ch. 4, FB Ch. 3)
- **WEEK 3 (Feb. 11th, 13th)**
  Maximum likelihood estimation and applications, ranting about minimizing χ^2 (ICVG Ch. 4)
- **WEEK 4 (Feb. 18th, 20th)**
  Regression & Inference: ordinary least squares, generalized least squares, orthogonal distance regression vs generative modeling of data (ICVG Ch. 8, FB Ch. 7)
- **WEEK 5 (Feb. 25th, 27th)**
  Bayes in practice, sampling and Markov Chain Monte Carlo methods (ICVG Ch. 5)
- **WEEK 6 (Mar. 3rd, 5th)**
  Building models, effective sampling techniques, estimating parameters & uncertainties, posterior predictive checks, other MCMC wizardry (ICVG Ch. 8). Midterm exam posted.
- **WEEK 7 (Mar. 10th, 12th)**
  Visualization (VdP Ch. 4), Midterm exam due. No homework assignment because it's spring break and I'm not that mean.
- **WEEK 8 (Mar. 24th, 26th)**
  Time-series analysis (ICVG Ch. 10, FB Ch. 11), Gaussian processes (ICVG Ch. 8.10, readings from Rasmussen & Williams)
- **WEEK 9 (Mar. 31st, Apr 2nd)**
  Probabilistic Graphical Models (PGMs) & hierarchical Bayes (Readings from Hilbe, de Souza & Ishida)
- **WEEK 10 (Apr. 7th, 9th)**
  The ABCs of not having a likelihood function (Readings from Hilbe, de Souza & Ishida)
- **WEEK 11 (Apr. 14th, 16th)**
  Intro to Machine learning, tree methods (ICVG Ch. 9, VdP Ch. 5)
- **WEEK 12 (Apr. 21st, 23rd)**
  Gaussian mixture models, density estimation, unsupervised clustering techniques, and dimensionality reduction (ICVG Ch. 6, 7, FB Ch. 9 and bits of Ch. 6, VdP Ch. 5)
- **WEEK 13 (Apr. 28th, 30th)**
  Dealing with outliers, imbalanced, and missing data, supervised machine learning techniques
- **WEEK 14 (May 5th)**
  supervised ML continued, putting it all together. Final exam posted on May 7th.
- **May. 14th**
  Final exam due by 1400.

0.2

SETUP.PY

NEW TO PYTHON?

- Mohit Sharma's tutorial from Urban Computing Skills Lab https://sharmamohit.com/work/tutorials/ucsl/
- Federica Bianco's Python Bootcamp: https://github.com/fedhere/PyBOOT
- UW E-Science 2015 seminar on python: https://github.com/uwescience/python-seminar-2015
- Text reference: https://www.southampton.ac.uk/~fangohr/training/python/pdfs/Python-for-Computational-Science-and-Engineering.pdf

WHAT WE'RE GOING TO DO NEXT

- install miniconda
- install python env
- install git if you don't have it
- create SSH keys
- clone repo