Software Engineering for Data Scientists
Working in Teams

David Beck[1,2], Joseph Hellerstein[1,3], Jake VanderPlas[1,4]

[1]eScience Institute

[2]Chemical Engineering

[3]Computer Science Engineering

[4]Astronomy

University of Washington

March 13, 2017

Agenda

- Software licenses
- Software development overview
- Team process
- Project questions

Software Licenses

Overview of Software Licenses*

A software license is a legal instrument (usually by way of contract law, with or without printed material) governing the use or redistribution of software. Under United States copyright law all software is copyright protected, in source code as also object code form. The only exception is software in the public domain. A typical software license grants the licensee, typically an end-user, permission to use one or more copies of software in ways where such a use would otherwise potentially constitute copyright infringement of the software owner's exclusive rights under copyright law.

*Mark Webbink

Software Licenses*

Rights granted	Public domain	Non-protective FOSS license (e.g. BSD license)	Protective FOSS license (e.g. GPL)
Copyright retained	No	Yes	Yes
Right to perform	Yes	Yes	Yes
Right to display	Yes	Yes	Yes
Right to copy	Yes	Yes	Yes
Right to modify	Yes	Yes	Yes
Right to distribute	Yes	Yes, under same license	Yes, under same license
Right to sublicense	Yes	Yes	No
Example software	SQLite, ImageJ	Apache Webserver, ToyBox	Linux kernel, GIMP

FOSS = Free and Open Source Software

*Mark Webbink

Software Licenses*

Rights granted	Freeware/Shareware/ Freemium	Proprietary license	Trade secret
Copyright retained	Yes	Yes	Yes
Right to perform	Yes	Yes	No
Right to display	Yes	Yes	No
Right to copy	Often	No	No
Right to modify	No	No	No
Right to distribute	Often	No	No
Right to sublicense	No	No	No
Example software	Irfanview, Winamp	Windows, Half-Life 2	Server-side World of Warcraft

The default (no license): No rights are granted.

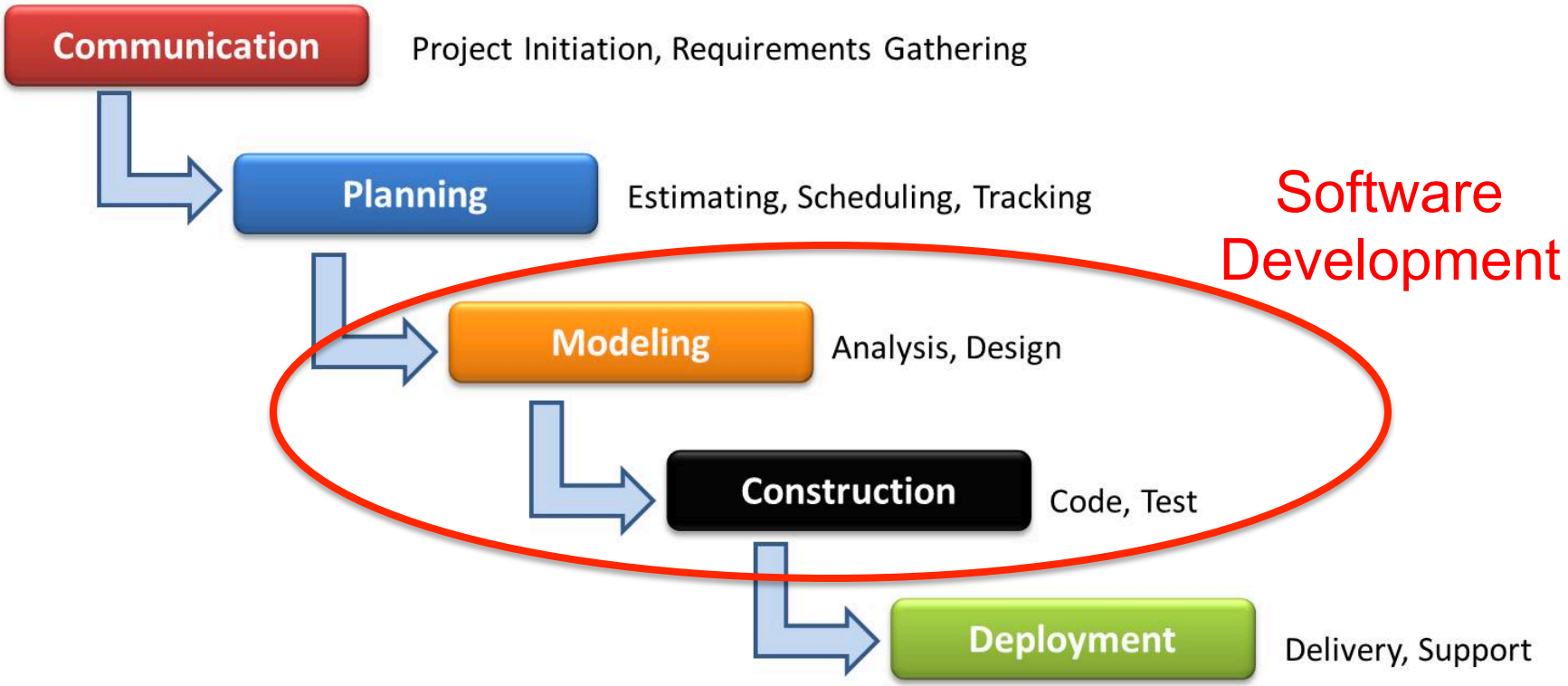
*Mark Webbink

Software Development Overview

Software Development Phases

Beck, Hellerstein & VanderPlas, 2017

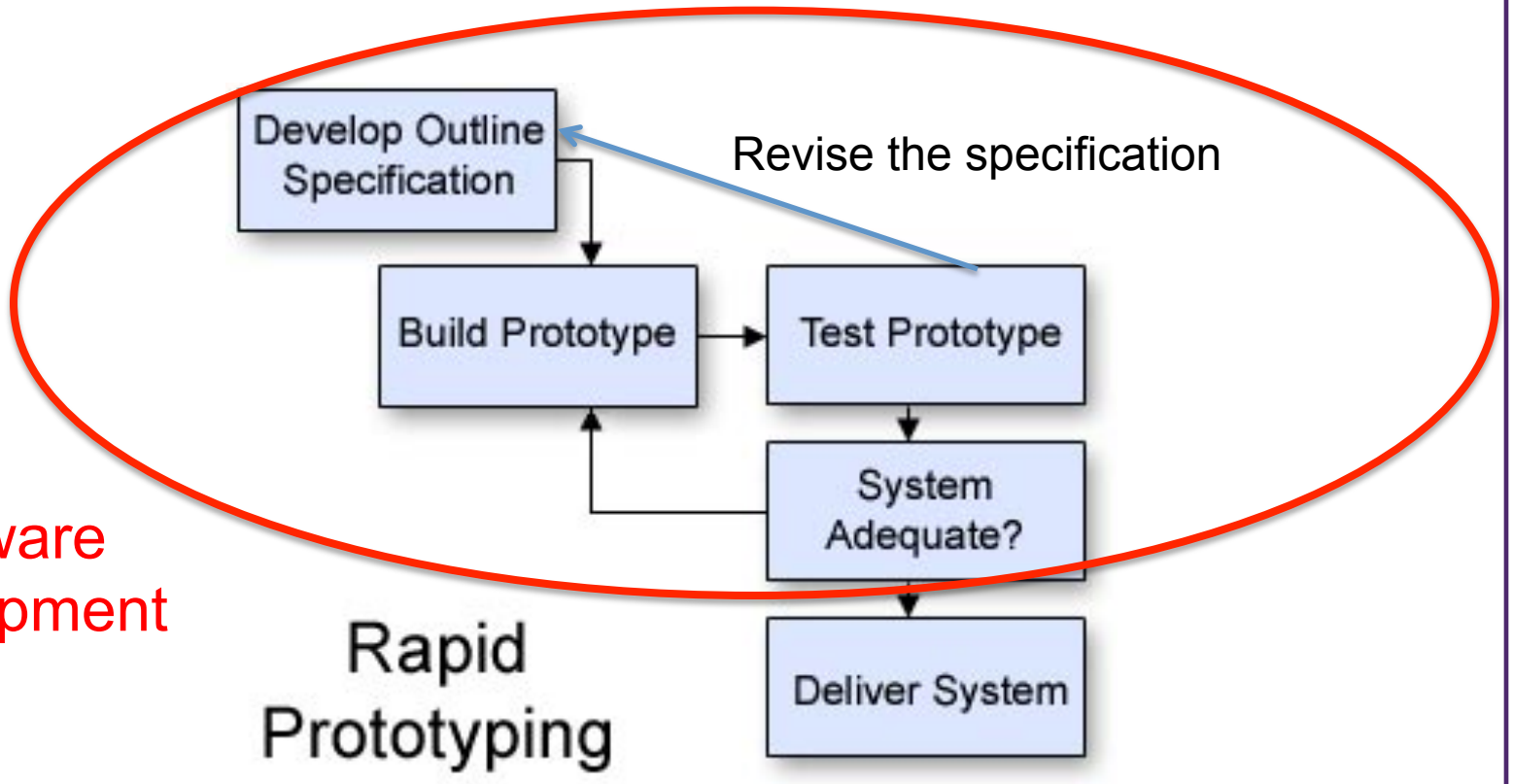
Waterfall Process Model



Why does this work poorly?

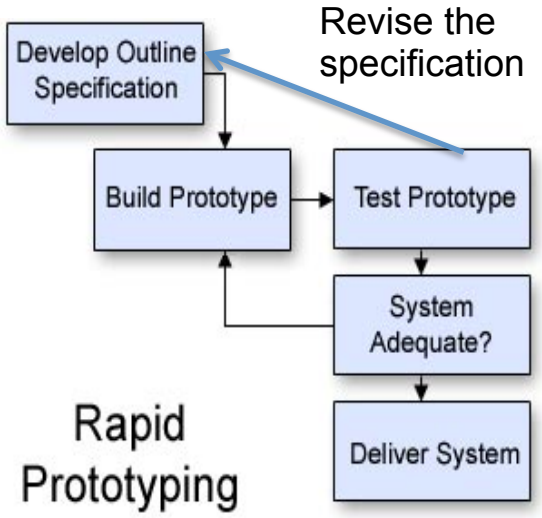
Rapid Prototyping

- Why?
  - Cannot specify all requirements in advance



Team Process

Team Activities



- Reqs gathering (functional spec.)
- Design
  - Technology assessments
  - Write specifications
  - Review specification
- Implementation
  - Code
  - Code review
- Bug prioritization and resolution
- Standups (status update)

Code Review Template

- Why code review?
  - Improve code quality and find bugs

- Background
  - Describe what the application does
  - Describe the role of the code being reviewed
- Comment on
  - Choice of variable and function names
  - Readability of the code
  - How improve reuse and efficiency
  - How use existing python packages

Technology Review Template

- Why technology reviews?
  - Determine if use a package

- Background
  - Requirements that indicate a need for the proposed package
- Discuss
  - How the package works
  - Appeal of using the package
  - Drawbacks of using the package

Example of A Technology Review

Antimony Package for Kinetics Modeling

- Background
  - Need kinetics models to explore certain what-if questions in chemical systems.

Using Antimony

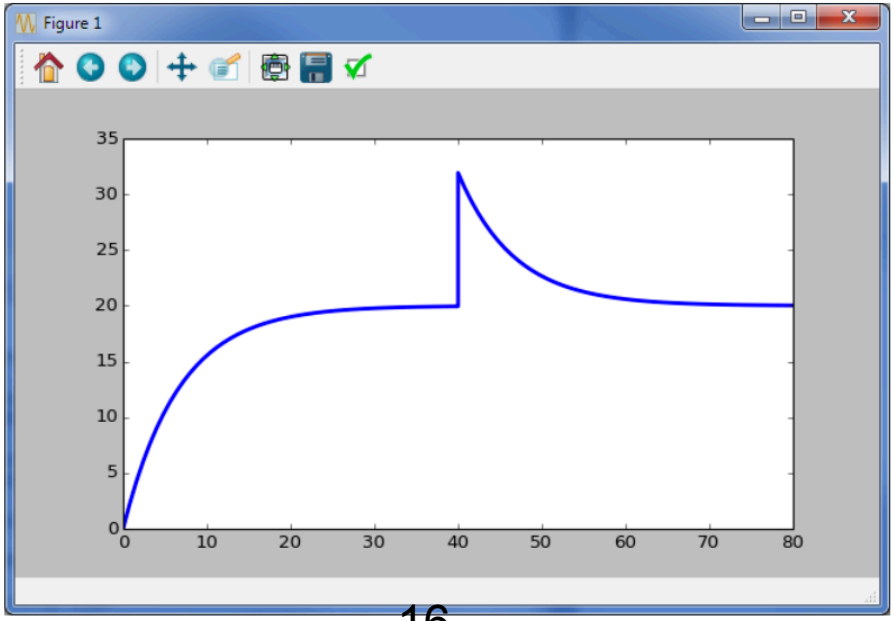
```
import numpy # Required for vstack
import tellurium as te

rr = te.loada ('''
     $Xo -> S1;  k1*Xo;
      S1 -> $X1; k2*S1;

     Xo = 10; k1 = 0.3; k2 = 0.15;
''')

m1 = rr.simulate (0, 40, 50)
```

Kinetics model is a python string



Perturbation of *S1*

Assessment of Antimony

- Appeal
  - Readable kinetics models
  - Can use python with Antimony
  - Exports to and imports from SBML (systems biology modeling language)
- Drawbacks
  - Poor support for the package
  - Scaling may be a problem

Projects

Project Updates

- What is your data?
  - You should have 2 datasets in hand!
- Who are your users?
  - General public? Scientists? Analysts?
- What questions are users trying to answer?
- What are the use cases (user-system interactions) to answer their questions?
- What issues are there ("known unknowns") with building your system?