



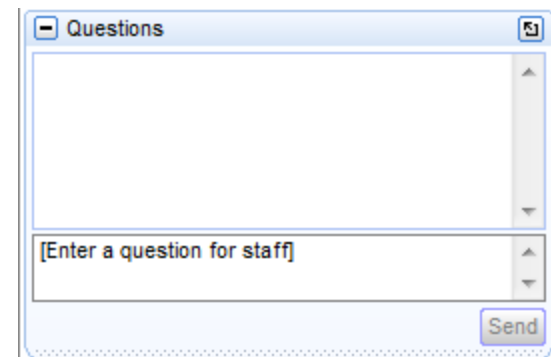
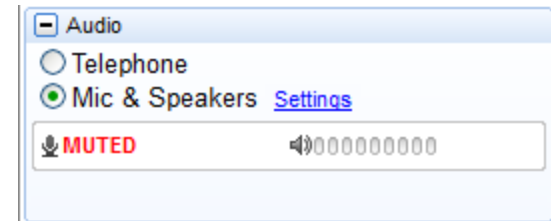
ggplot2

Hadley Wickham
Professor of Statistics
Rice University

Ray DiGiacomo, Jr.
President
Orange County R

GoToWebinar Control Panel

1. Click the Orange button to toggle your panel
2. Select Audio Type
 - Microphone & Speakers
 - Telephone
3. Close other applications
4. Make webinar full screen



Me

- Ray DiGiacomo, Jr.
- President, Lion Data Systems LLC
- President, The Orange County R User Group
(Join us on LinkedIn Groups)
- Board Member, The Data Warehousing Institute
- Living in Southern California
- Contact me at: **info@liondatasystems.com**

Lion Data Systems, LLC

- **Predictive Analytics Consulting**
(for Healthcare and Life Sciences)
- **R Training**
(Internet Based or On-Site)

Upcoming Webinars

Predictive Model Markup Language

With Dr. Alex Guazzelli, Dr. Michael Hahsler and
Dr. Rajarshi Guha

Cost: Free (Sponsored by Orange County R)

Date: January 24, 2012

Visit

liondatasystems.com

to register

R Training Classes

- Course 1: “R for SAS and SPSS Users”
- Course 2: “R for Analytics Beginners”
- Course 3: “Predictive Analytics for Executives”
- Free courses available for qualified customers
- info@liondatasystems.com

Hadley's ggplot2 Book

ggplot2:

Elegant Graphics for Data Analysis

By: Hadley Wickham

ISBN-13: 978-0387981406

Where can I download this webinar's
slides and video?

Check my website in a week or so...

liondatasystems.com

Poll

Start of Presentation

ggplot2

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair
Department of Statistics / Rice University

November 2011

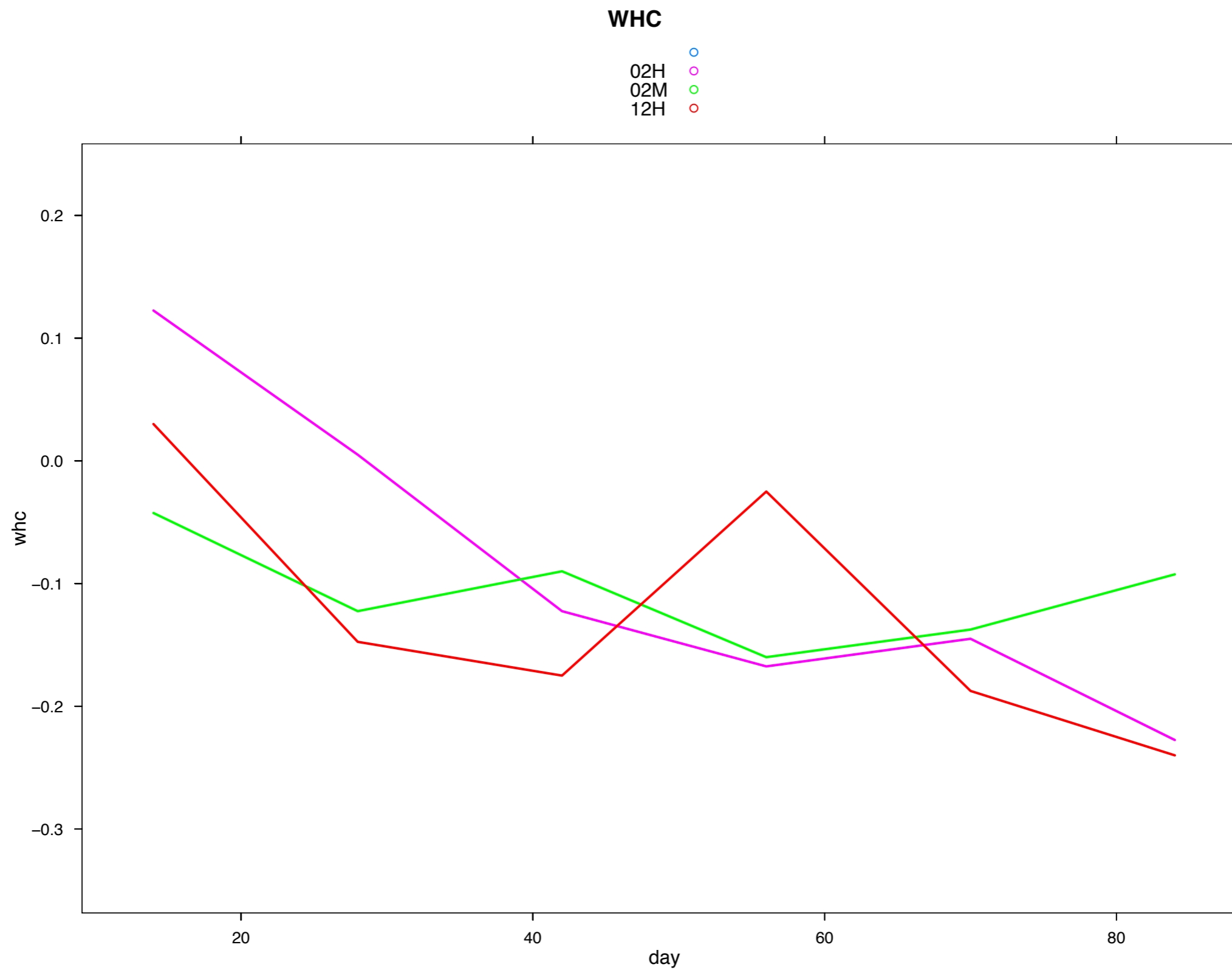


1. Motivation: why ggplot2?

2. Present

3. Sneak peek

Why ggplot2?



Niggles


- Why didn't `panel.line` automatically order the points? Why couldn't you describe histograms by specifying binwidths?
- Why was `xyplot(panel = panel.bwplot)` the same as `bwplot`?
- Why was it so hard to combine different datasets on the same plot?

Statistics and Computing

Leland Wilkinson

**The Grammar
of Graphics**

Second Edition

 Springer

“Nothing is as practical as a good theory”
—*Kurt Lewin*

“[A good model] will bring together in a coherent way things that previously appeared unrelated and which also will provide a basis for dealing systematically with new situations”
—*David Cox*

Milestones

(R2.2.1)

28 Oct **2005** – ggplot development starts

6 Apr **2006** – first release of ggplot

3 Jul **2006** – first ggplot announcement

10 Jun **2007** – first release of ggplot2

7 Nov **2008** – start of ggplot2 mailing list

7 Aug **2009** – ggplot2 book published

Major versions

0.5: ggplot2, + instead of functional style

0.6: documentation, auto legends

0.7: themes

0.8: facet_wrap, free scales

0.9 (Christmas 2011): namespace, roxygen, S3, diaspora

The “hard” stuff

- Implementing grammar is actually pretty easy
- The hard stuff (as always) is in the details:
 - What is a good default scale for colour?
Size? shape?
 - Where should tick marks go?
 - How can the user control the style of the plot?

More milestones

April 2006: how do you put two ggplot figures into the same page?

June 2006: why are the breaks in such bad positions?

July 2006: how do you get rid of the gray background?

Present

Community

Mailing list: 1,900 members, 10,000 messages, 300-500 messages/month

Stackoverflow.com: >750 questions and answers

Code contributions: In last release of ggplot2, 50% of features/fixes contributed by users (mainly by Takahashi Kohske)

Community

- <http://learnr.wordpress.com/>
- <http://www.ling.upenn.edu/~joseff/rstudy/>
- <http://wiki.stdout.org/rcookbook/>
- <https://sites.google.com/site/r4statistics/example-programs/graphics-ggplot2>

Current work

ggplot2 is too complicated - it is hard for me to improve, and hard for new developers to understand.

ggplot2 was my second R package - I have now written around 30. I now know a lot more!

Current work

Break up ggplot2 into simpler pieces:
scales, density vis, spatial vis, ...

Aggressively rewrite to make simpler.

Better development practices: S3 instead of proto; roxygen2; unit testing; namespaces.

New features aren't that exciting, but smooth out many rough edges.

Sneak peek

```
# To get the development version

# install.packages("devtools")
library(devtools)
dev_mode() # don't overwrite your existing install
install_github("ggplot2")
```

Development version

```
hadley — R — 80x24 — 963
> dev_mode()
Dev mode: ON
> library(ggplot2)
>
```

```
hadley — R — 80x24 — 962
> library(ggplot2)
Loading required package: reshape
Loading required package: plyr

Attaching package: 'reshape'

The following object(s) are masked from 'package:plyr':

  rename, round_any

Loading required package: grid
Loading required package: proto
> █
```

Released version

```
library(ggplot2)
ddply
```

```
# Note that plyr, reshape etc aren't automatically
# loaded. This is good development practice -
# it's better to be explicit than implicit.
```

```
# Building up suite of automated tests to ensure  
# that bugs only need to be fixed once  
library(testthat)  
test_package("ggplot2")
```

Will be big improvements to documentation

(Still a work in progress)

?geom_point

?facet_wrap

Scales have been rewritten to be more consistent

and more flexible. Dennis Murphy helping me

to write up the documentation.

?scale_fill_gradient

?continuous_scale

```
# Trying to make ggplot2 faster
system.time(
  print(qplot(carat, price, data = diamonds))
)

# Includes new tools for figuring out what's
# taking all the time
benchplot(qplot(carat, price, data = diamonds))

# Still a lot of work to do!
```

```
# Internally, there has been a big rewrite of  
# the faceting data processing and rendering  
# systems. This lays the foundation for new  
# features, and fixes some annoying long-standing  
# bugs.
```

```
df <- data.frame(x = 1:10, y = 10:1, colour = 1:2)  
qplot(x, y, data = df) + coord_fixed()  
qplot(x, y, data = df) + facet_wrap(~ colour)
```

```
# Kohske Takahashi has contributed substantial
# improvements to the legends

df <- data.frame(x = runif(100), y = runif(100))
df$colour <- with(df, x ^ 2 + y + runif(100))

qplot(x, y, data = df, colour = colour)
qplot(x, y, data = df, colour = colour) +
  guides(colour = guide_legend(nrow = 2, byrow = T))
qplot(x, y, data = df, colour = colour) +
  guides(colour = guide_colorbar())

qplot(x, y, data = df, colour = colour,
      alpha = I(1/4), size = I(30)) +
  guides(colour = guide_legend(
    override.aes = list(alpha = 1, size = 2)))
```

Two new geoms (developed today)

geom_raster

geom_map

High performance special cases of geom_tile

and geom_polygon.

ggplot2 0.9

- Considerably reduces technical debt that has accumulated over the last 5 years
- Fixes some frustrating long-term bugs
- Big improvements to scales and guides

Next year?

- Rewrite of themes system
- Switch geoms, stats and position adjustments to S3 and rewrite
- `geom_histogram` -> `layer_histogram`?
- `big visualisation` -> `big_viz`

- Questions?