O’Reilly Research

Applied Natural Language Processing - SIMS 256

Roger Magoulas
Ben Lorica
O’Reilly Media
roger@oreilly.com
ben@oreilly.com

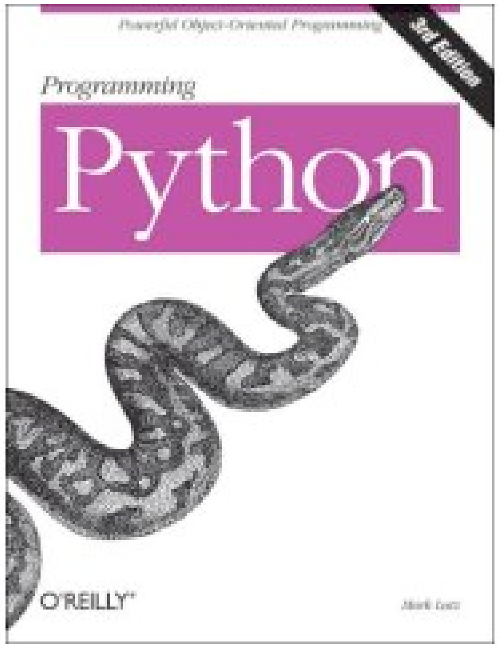
Outline

- Intro / Background
- Disambiguation
- Books
  - Regex
  - SVM
- Jobs
  - Regex - Term Freq
  - Topic Model
- Social Networks
- Book Contents
- Wrap-up

Introduction

O’Reilly Media

- **Changing the world by spreading the knowledge of innovators**
  - The future is here it’s just not evenly distributed (W. Gibson)
- **Third largest independent technical book publisher**
- **Conferences (e.g., Web 2.0, Foo Camp)**



O’Reilly Research

- **Three basic tasks:**
  - Help Editors pick technology book topics
  - Help Retailers stock best selection of books
  - Track technology adoption trends
- **Social Network**
  - O’Reilly has contacts with many technology leaders, from academia, from finance, and, most importantly, from technology entreprenuers
- **Quantitative Analysis**
  - Faint signals from data stores: e.g., book sales, on-line jobs, blogs, mail list servers, futures markets
- **Telling Stories; making sense out of nonsense**

Introduction

Roger Magoulas

- Finance / Computer Science + Haas MBA
- Data Warehouse and Quantitative Analysis Experience
  - audited SIMS 296

Ben Lorica

- Ph.D., UC Santa Barbara - Partial Differential Equations & Probability
- Math Faculty, UC Davis
- Founding Chair of Math and Stats at Cal State Monterey Bay
- Finance, Commerce and Technology Analysis

Why NLP?

- **We use simpler methods when appropriate**
  - **Regex-Based Term Frequency Distribution**
- **Started with desire to categorize > 10,000 books**
  - **Too many books for small team**
  - **Term Frequency Distribution and Regex method too inaccurate**
  - **Need for fast categorization of Retailer inventory**
- **Job and Blog data accelerated need**
  - **Unstructured text to mine for technology trends**
  - **Large data sets**
    - 80mm Jobs
    - 100mm Blogs
    - random samples to manage complexity
    - Fast MPP Database - Greenplum
      - database summary: MySQL, Postgres, XML DBs
- **NLP experience**

Disambiguation

- **Some technology terms are difficult to spot in a technology context:**
  - **Access**
  - **Ruby (Rails)**
  - **Java**
  - **Subversion**
  - **Mercurial**
  - **Python**
  - **c**
- **We know we're looking for technology context:**
  - **In Books, use brand or prefix / suffix words**
    - hand review - needs to be correct
  - **In Jobs / Blogs, multiple key technology mentions**
    - willing to accept errors
    - job metadata, when available, helps

Book Sales

- **Nielsen POS Data - Computer Book 3K**
  - **Weekly Sales**
  - **15K books, 3+ years of data**
- **Exception and Trend Reporting**
  - **Treemap/Dasboard Portal**
  - **Dimensional classifications to make sense of data**
    - But making classifications assignments is tedious
    - Classification Tools
- **Classification Tools**
  - **Based on Book Meta Data: Title, Description, Reviews**
  - **Regex can be good enough**
    - Programming Languages, Databases, Certification
    - Domain mostly known, slow changing and often exposed
  - **SVM for book topic categorization**

SVM Retail Book Classification

- Two large Book retailers have asked us to assist them in stocking their Computer/Technical section.
- We devised a Retail categorization scheme with two levels
  - 19 Shelf Signs
  - 80 Shelf Labels
- We had to classify a few thousand titles into one of these Retail categories.
- Fortunately, we have thousands of titles already classified:
  - Training Set: 13K+ titles already categorized
  - For each title, we have a rich set of text data from Amazon (title, editorial and reader reviews)
  - Some books are difficult to categorize
    - e.g., Beware the Blue E (Firefox)

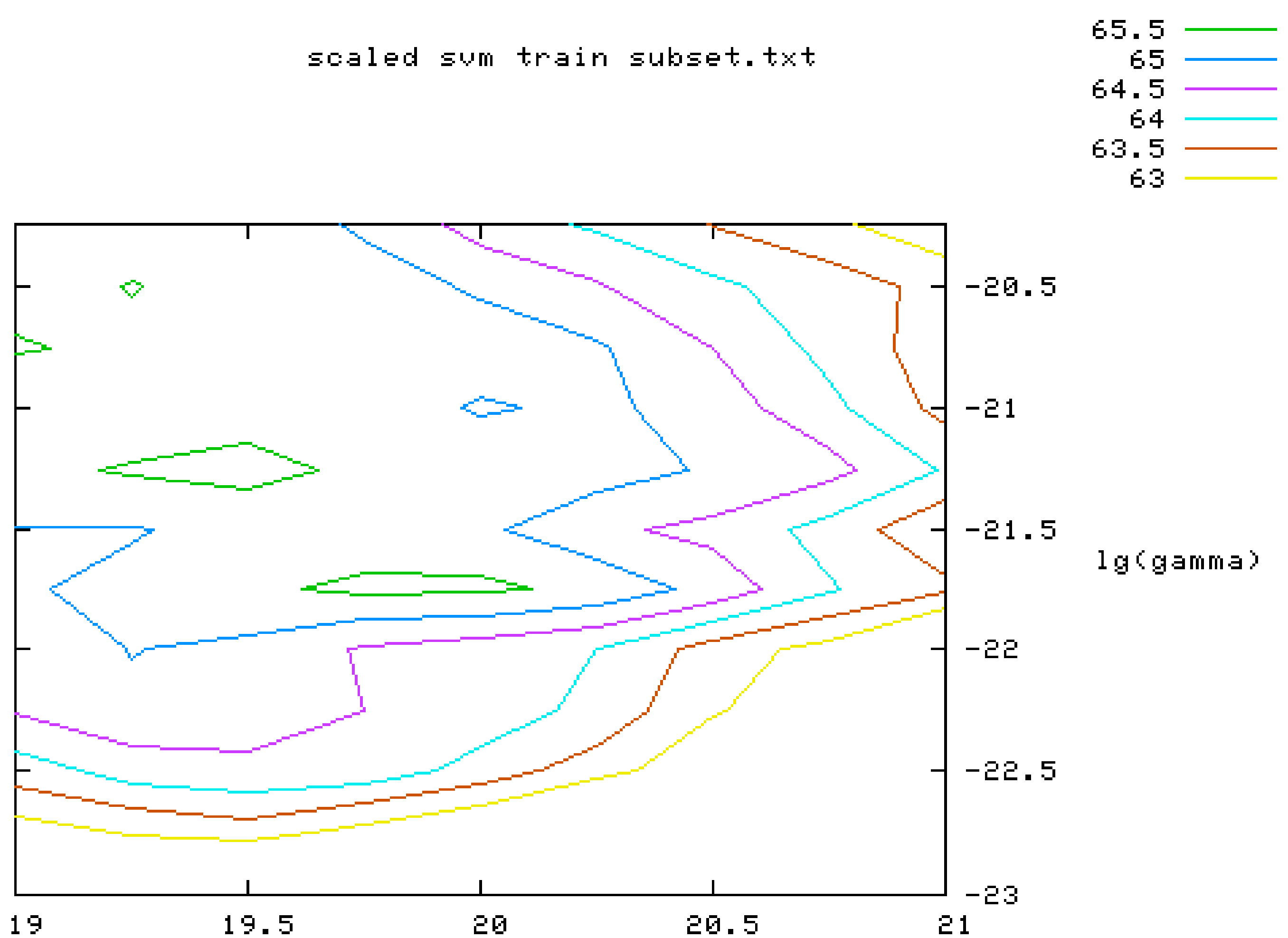
Text Classification: Kernel Methods

- After trying out Naive Bayes and KNN, we have settled on a specific set of Kernel methods.
  - linear and non-linear Support Vector Machines (SVM)
- We currently use the libsvm (C++) implementation.
- Text are parsed, stemmed, and stop words are removed.
- The results for linear SVM serves as a benchmark as we search for optimal Radial Basis Function (RBF) SVM's.
- Some art req'd to set RBF, based on linear SVM results
- Keerthi and Lin (2003):
  - A linear SVM with parameter C, can be approximated by an RBF SVM with parameters

    $$(C_0 , \sigma)$$
- Cross Validation used to check error
  - Check SVM for K groups
  - Mitigates Over-fitting

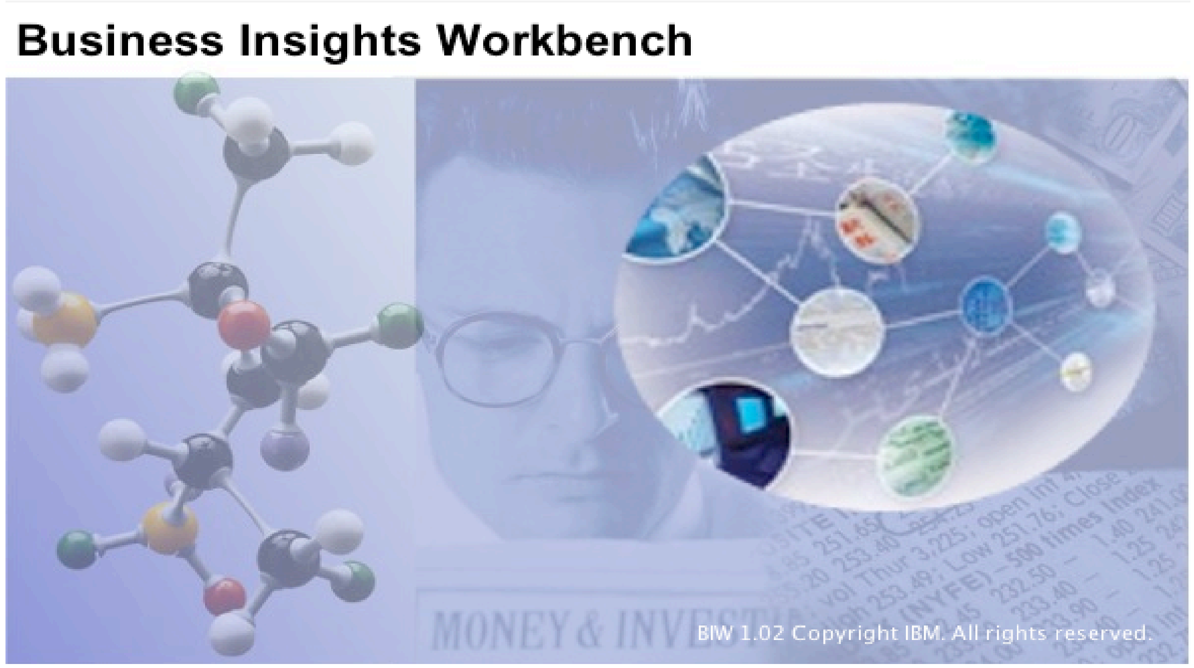
Text Classification: Kernel Methods

- Depending on the project, the classifiers we trained were about 65-75% accurate.
  - Since we care strongly about the results, manually check the results.
- The classifiers speed-up a tedious manual inspection task
- Example results



Creating a Taxonomy - IBM BIW Tool

- **We trialed a tool for Exploring, Understanding, and Analyzing text.**
  - **easy to use Java UI; well-suited for analysts/non-programmers.**
  - **Since UI comes with a lot of features and options, it was difficult to replicate previous work.**
  - **Underlying data can be stored in a RDBMS**
- **The tool also comes with a set of classifiers**
  - **Ideal for building taxonomy and classifying new documents on a regular basis.**
  - **Reduced dimensionality**
    - manual splits
    - meta-data review
- **Ultimately abandoned**
  - **fit w/ ongoing process**
  - **resource constraints**



Book Summary

- **Evaluating Alternatives Categorization Schemes**
- **Integrating Categorization into manual review process**
- **Key Learnings**
  - **Classifying books requires manual review of machine learning results**
  - **Accurate classifications considered a requirement to maintain confidence in analysis and recommendations**
  - **Machine Learning accompanied by Rule Based algorithms for best results**
  - **Careful considerations of categories enhances efficacy of machine learning tools**
    - Machine learning underperforms in poorly defined categories
  - **Challenge to accommodate 800 atomic topic categories with machine learning techniques**
    - preliminary results: 47% accuracy w/ linear SVM
    - about same as Rule-based Regex method
      - requires more maintenance

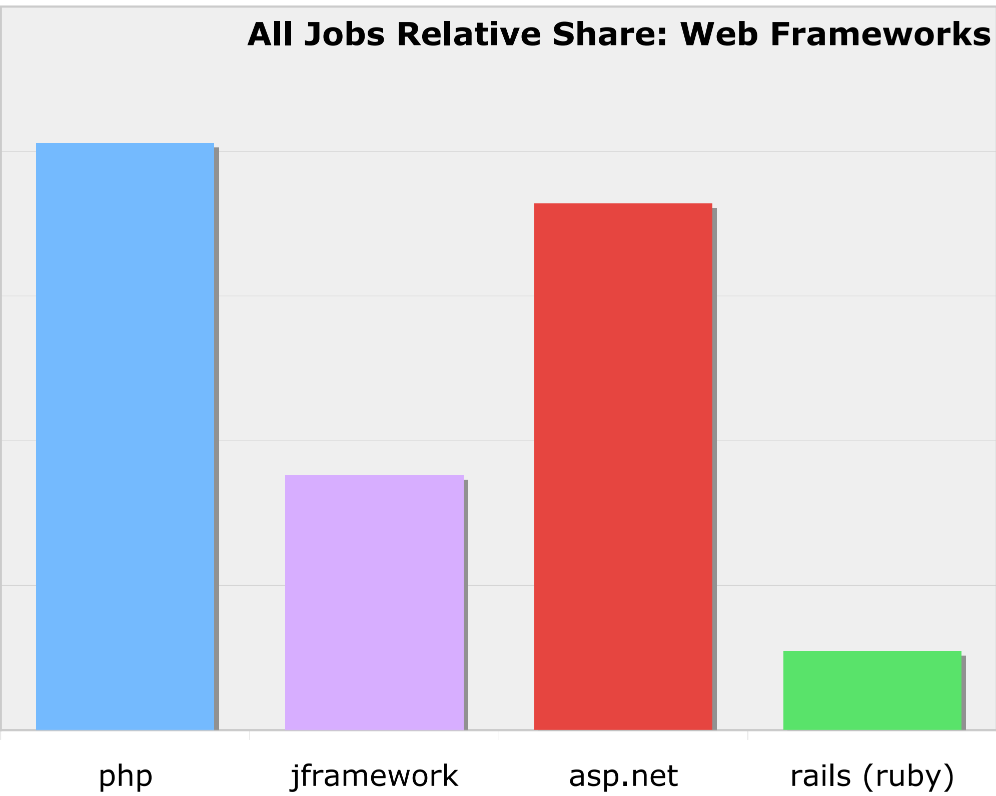
Job Data

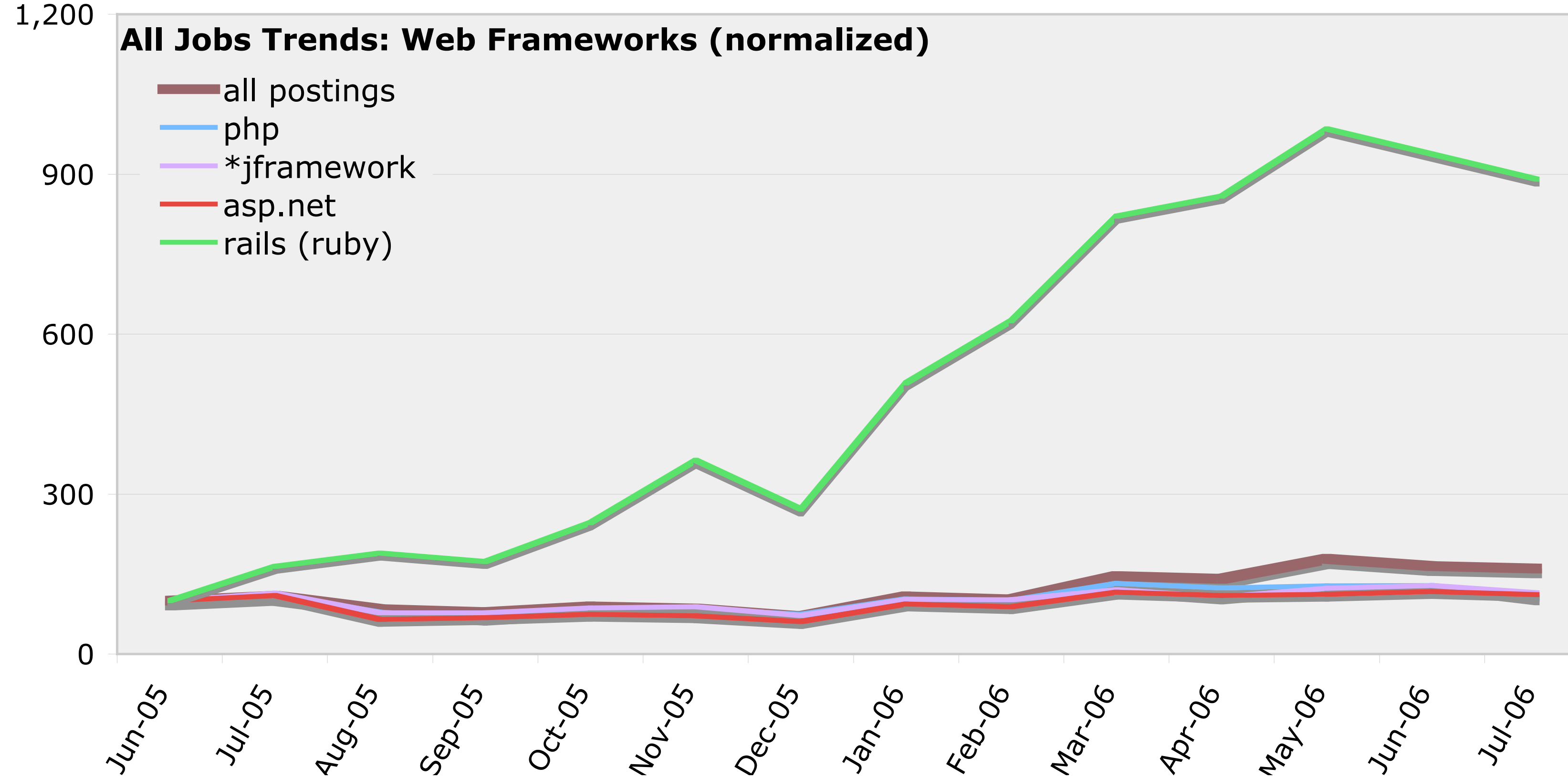
- **80mm on-line job postings**
- **Used for Technology Adoption Trend Analysis**
- **Research Example**
  - **Technology term frequency distribution and trends**
    - Manual analysis
    - via Lucene search
  - **Topic Model**

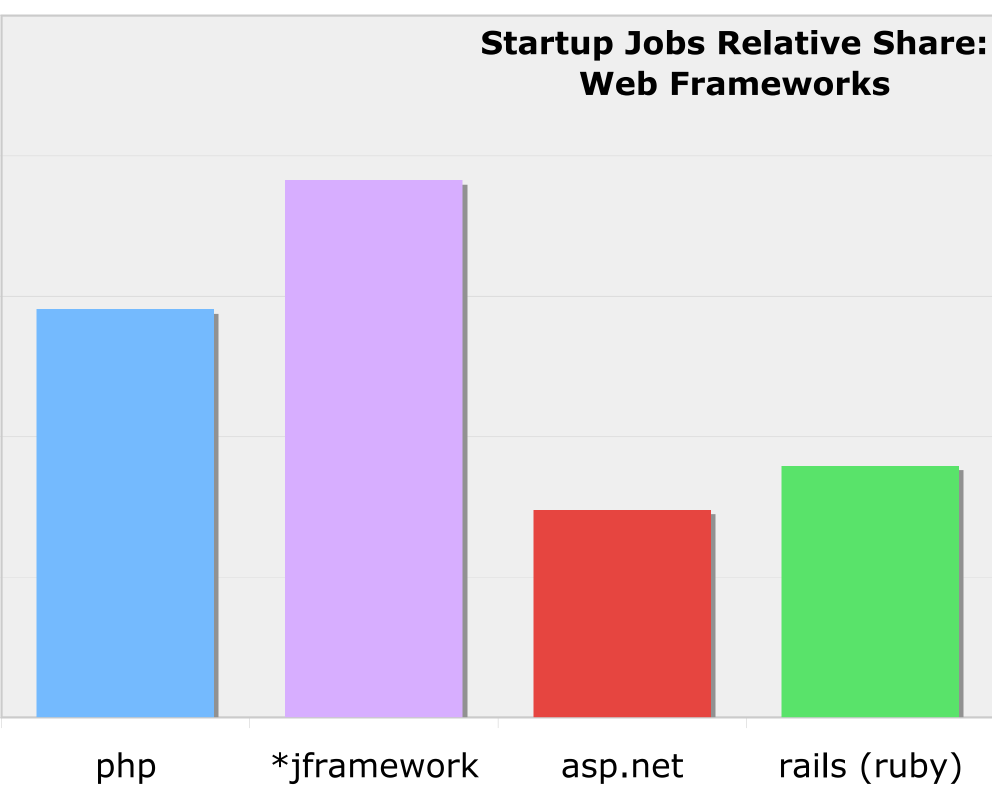
Web Development Frameworks

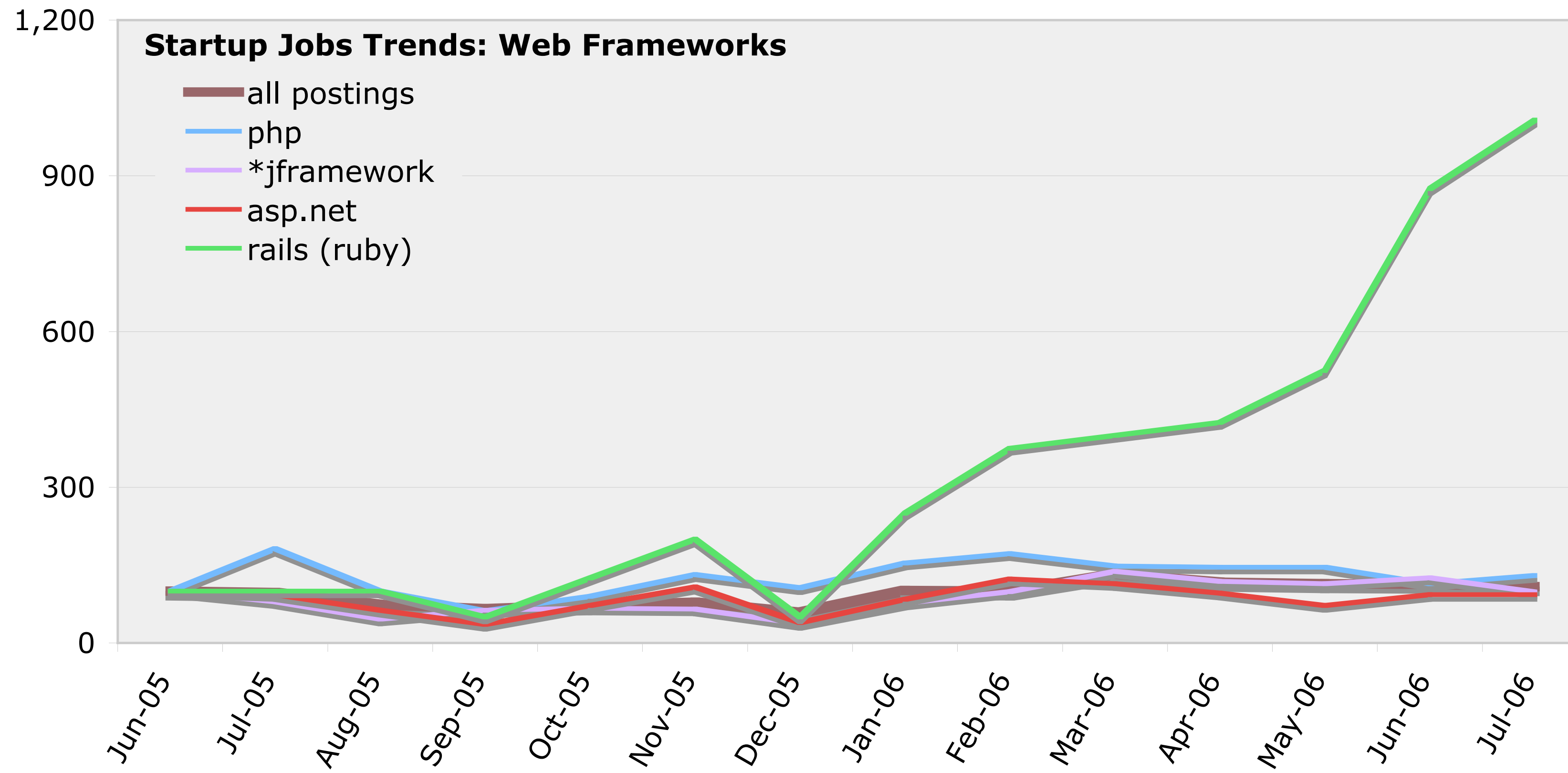
- **Startups: Java Frameworks Share Up; ASP.Net Share Down**
- **Rapid Growth of Ruby**

***jframework = jsp, struts, swing, hibernate, webworks**



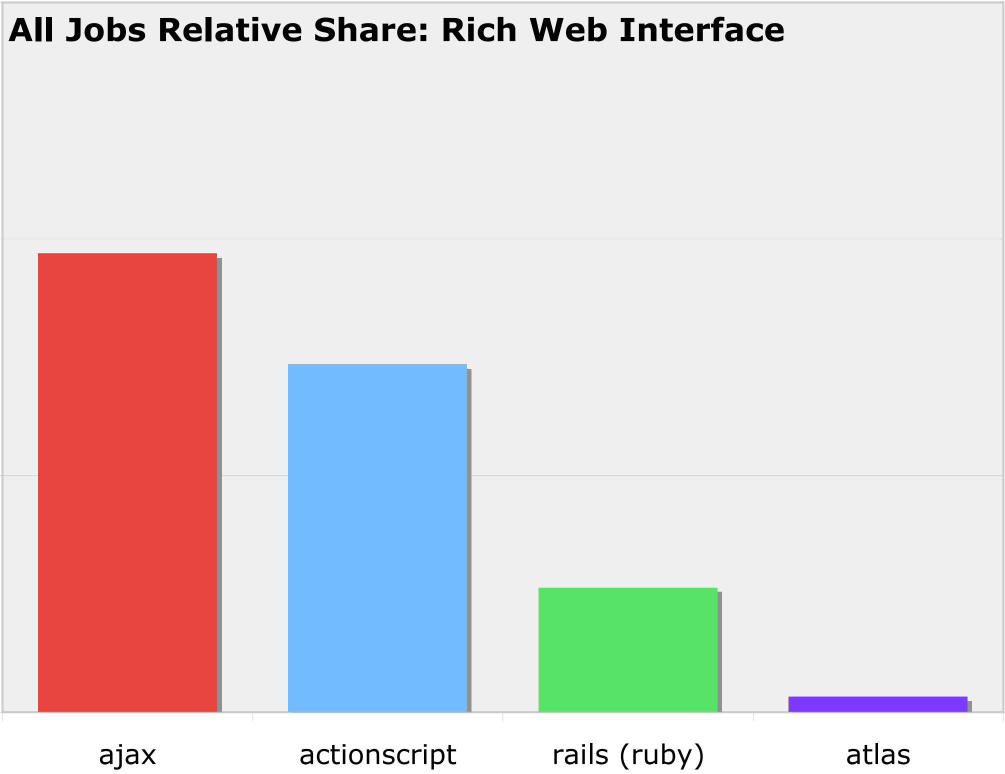


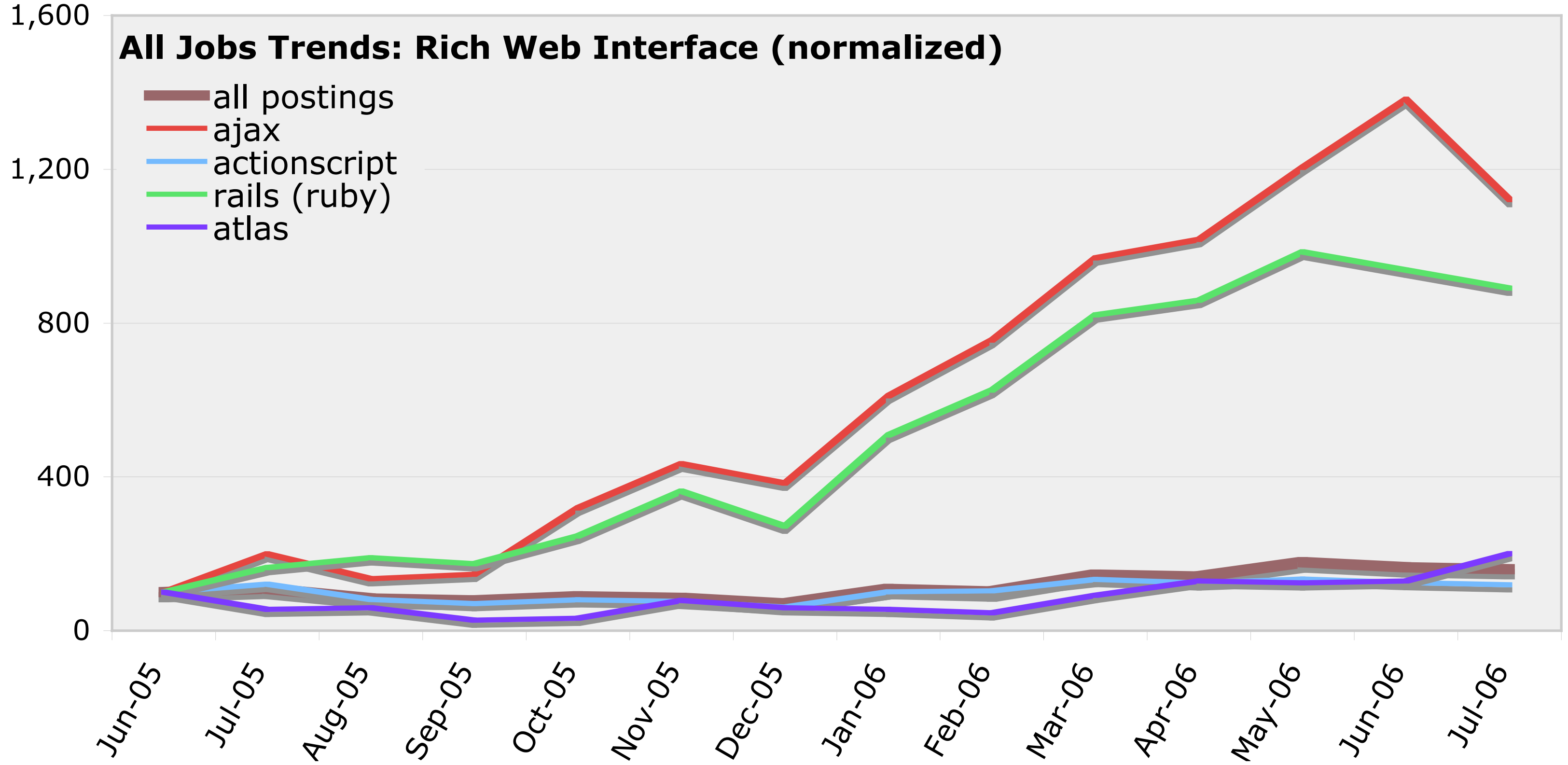


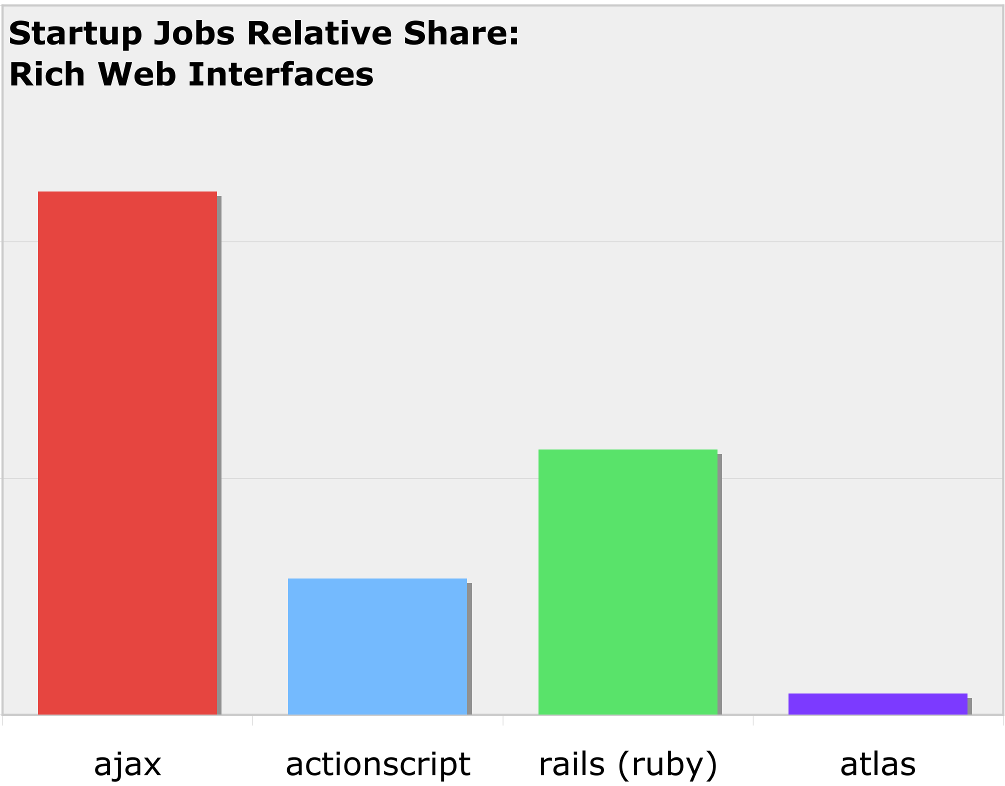


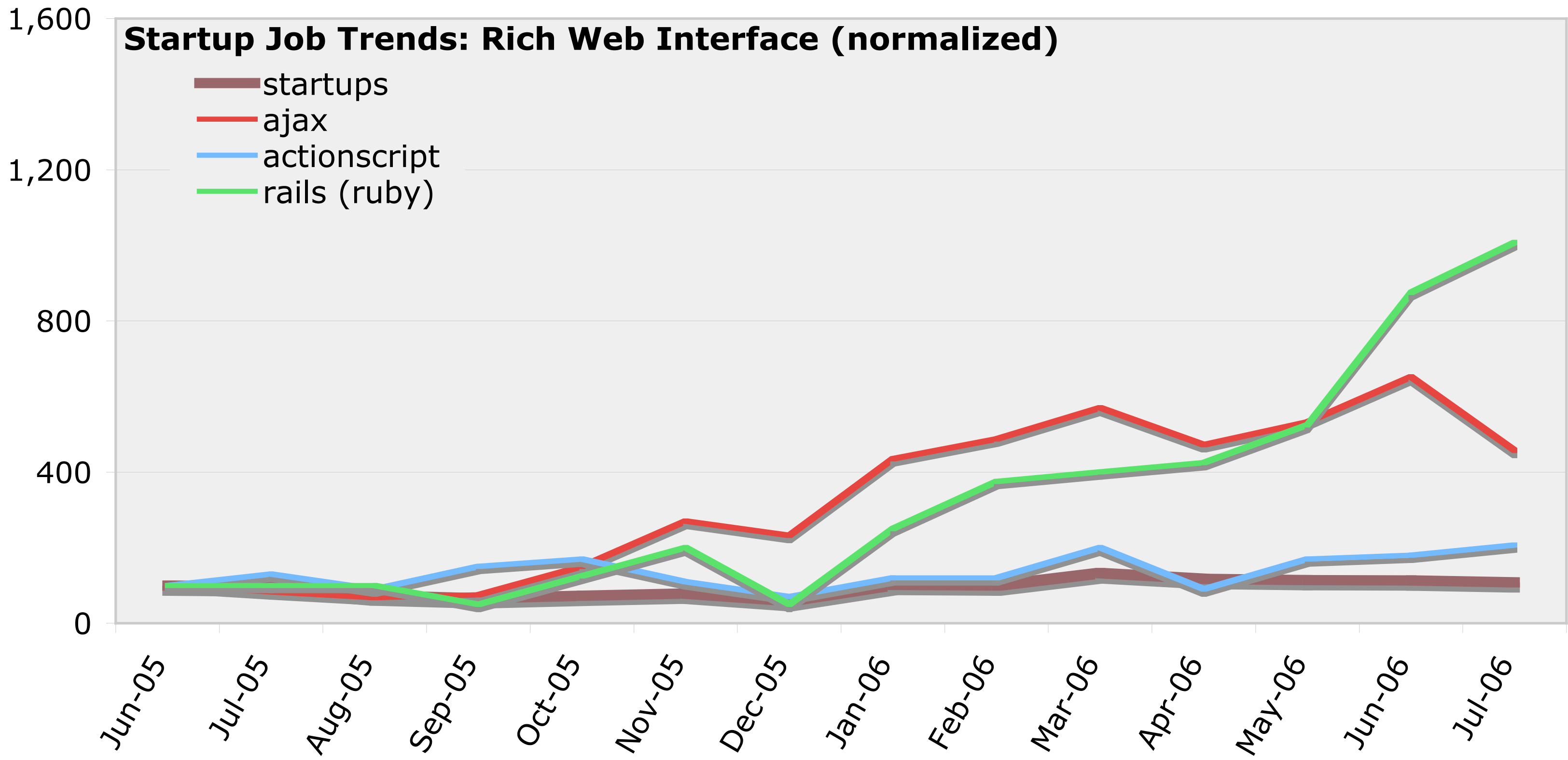
Rich Web Interface Development

- **Startups: AJaX ascendant and appears significantly more frequently**
  - **Rails making inroads among all Jobs and Startups**
    - Too few Atlas mentions to graph Startups trends



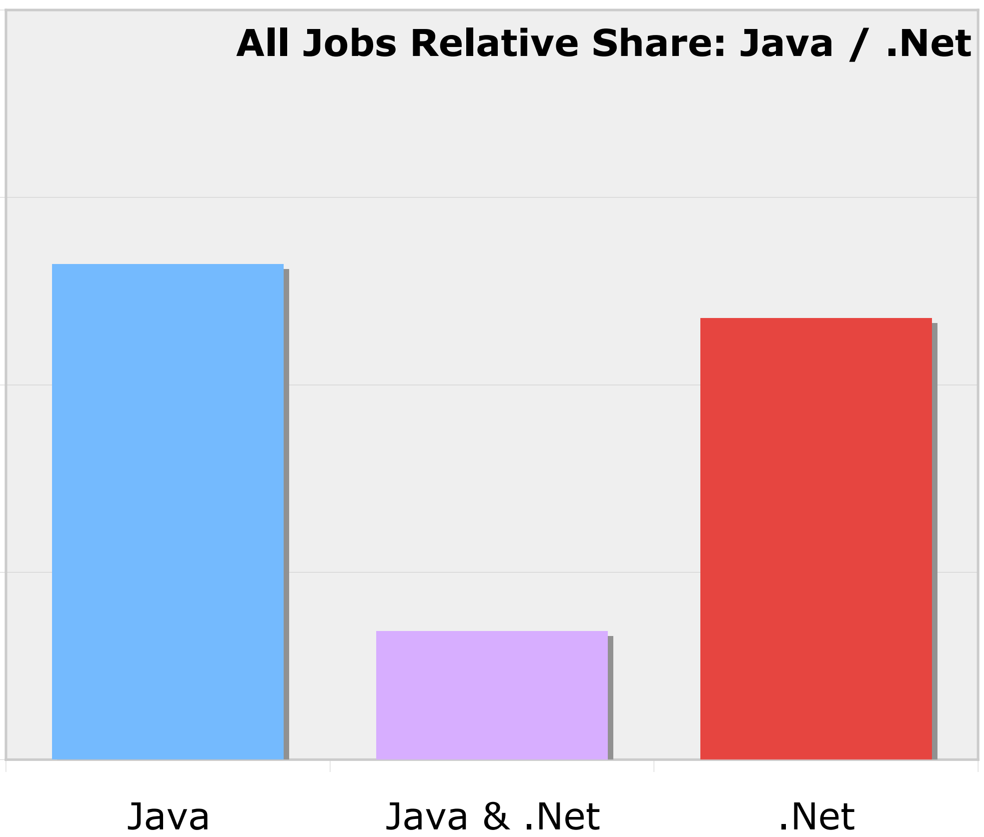


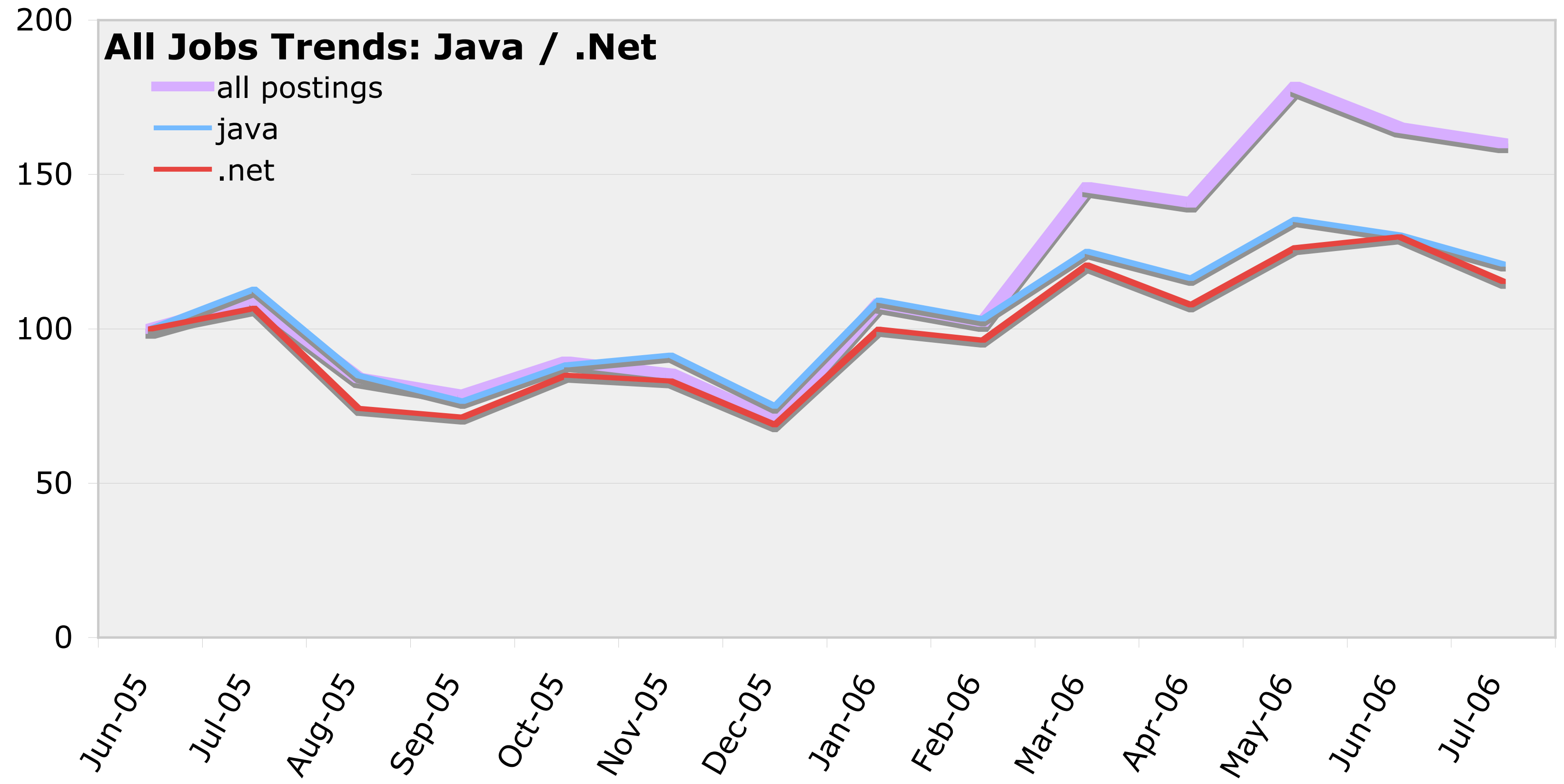


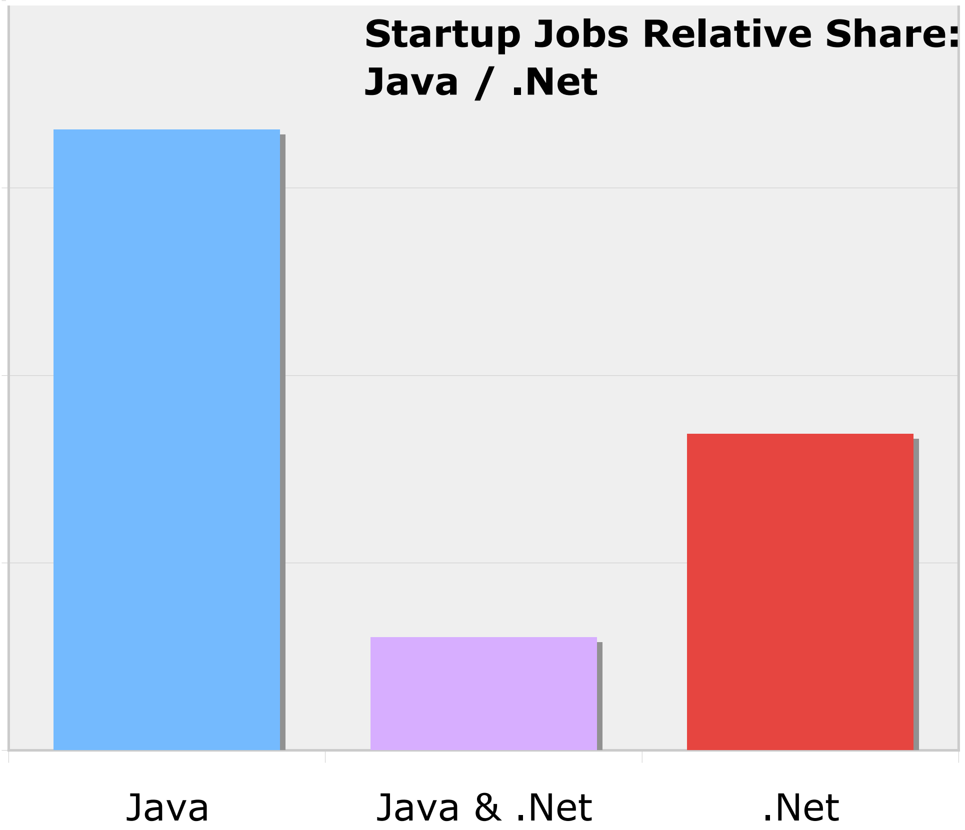


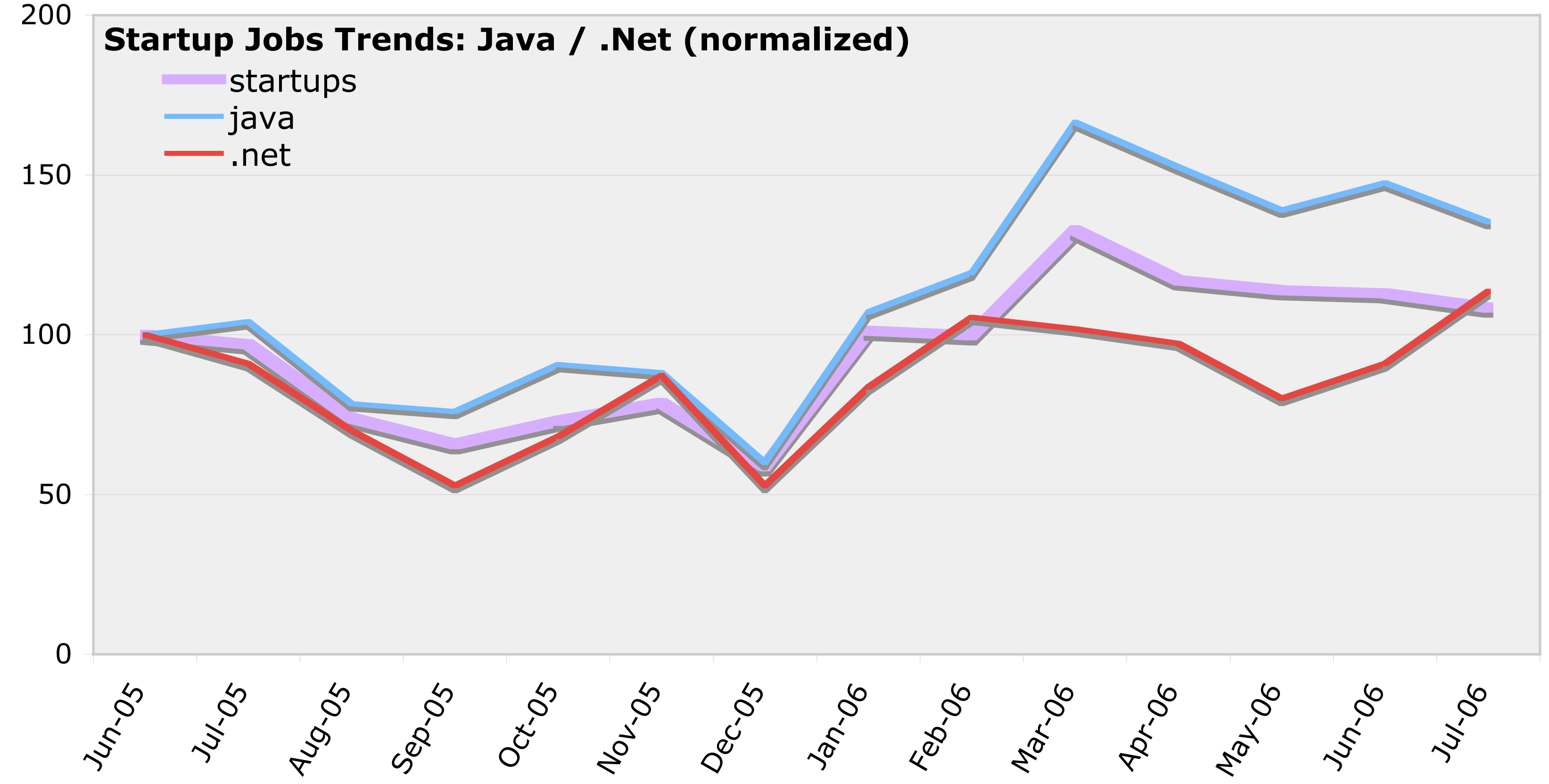
.Net / Java Development

- **Startups: Java share increases and growing faster**



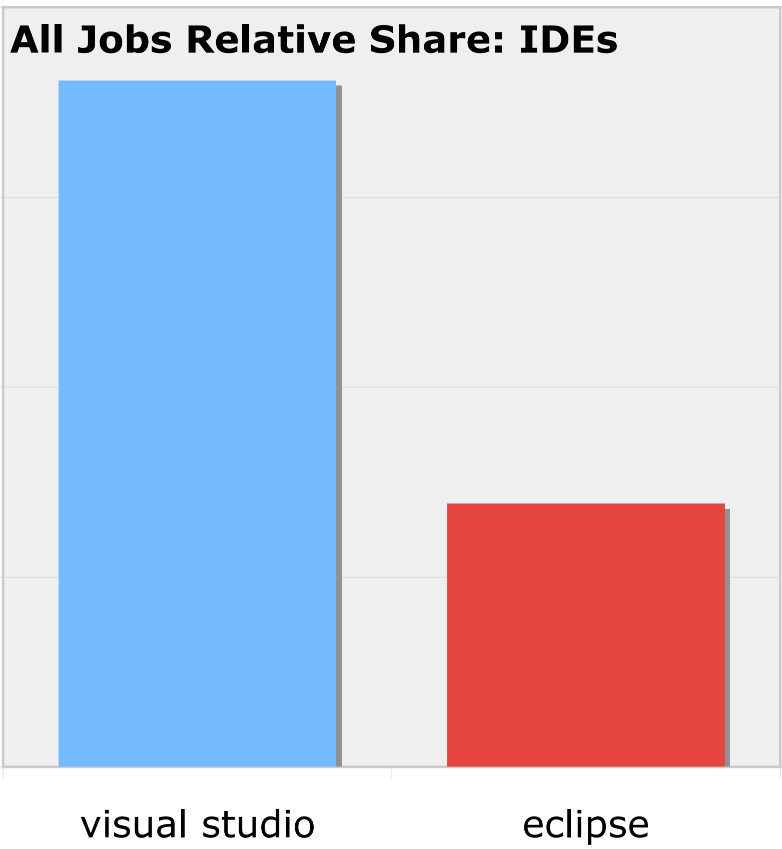


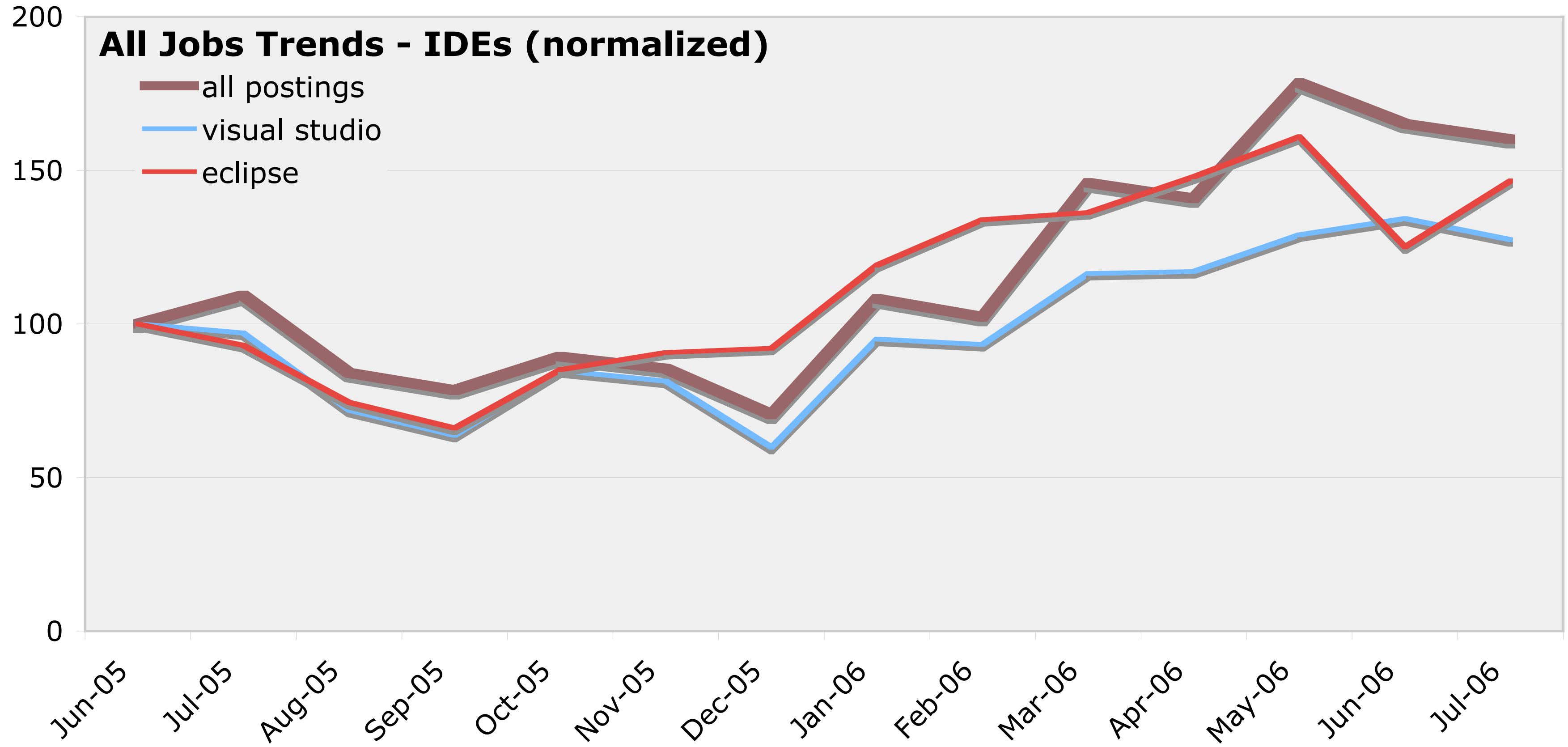


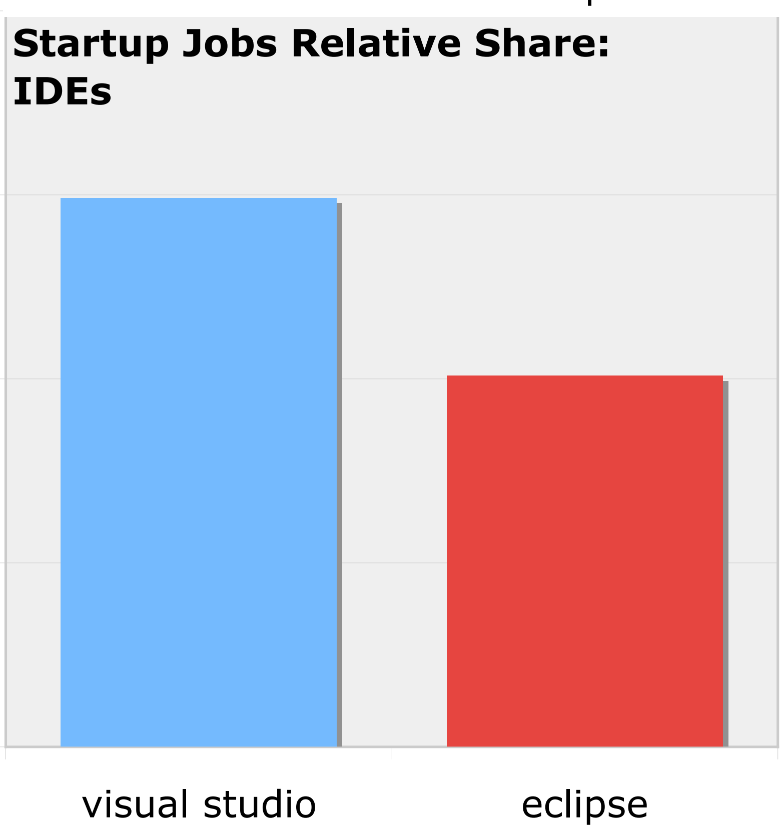


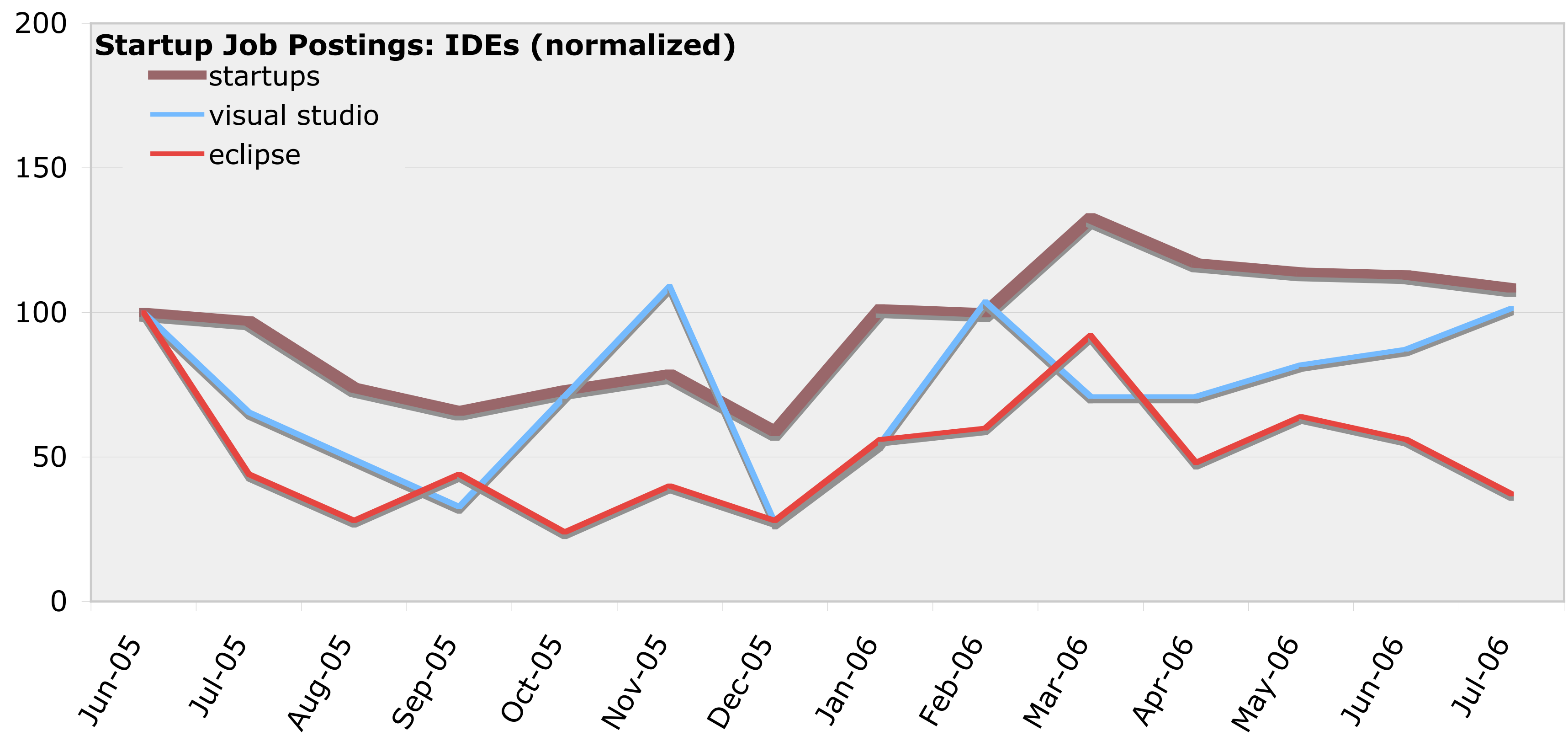
IDE's

- **Startups: Eclipse gains share**
  - **Other IDE's do not appear in startups often enough to chart**



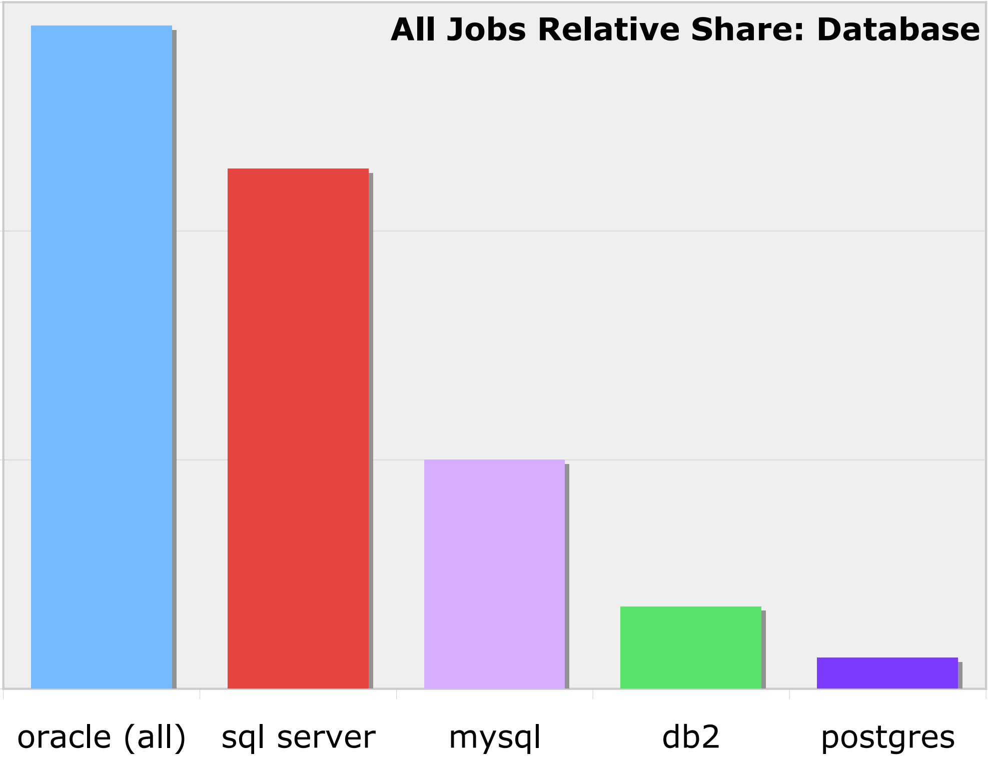


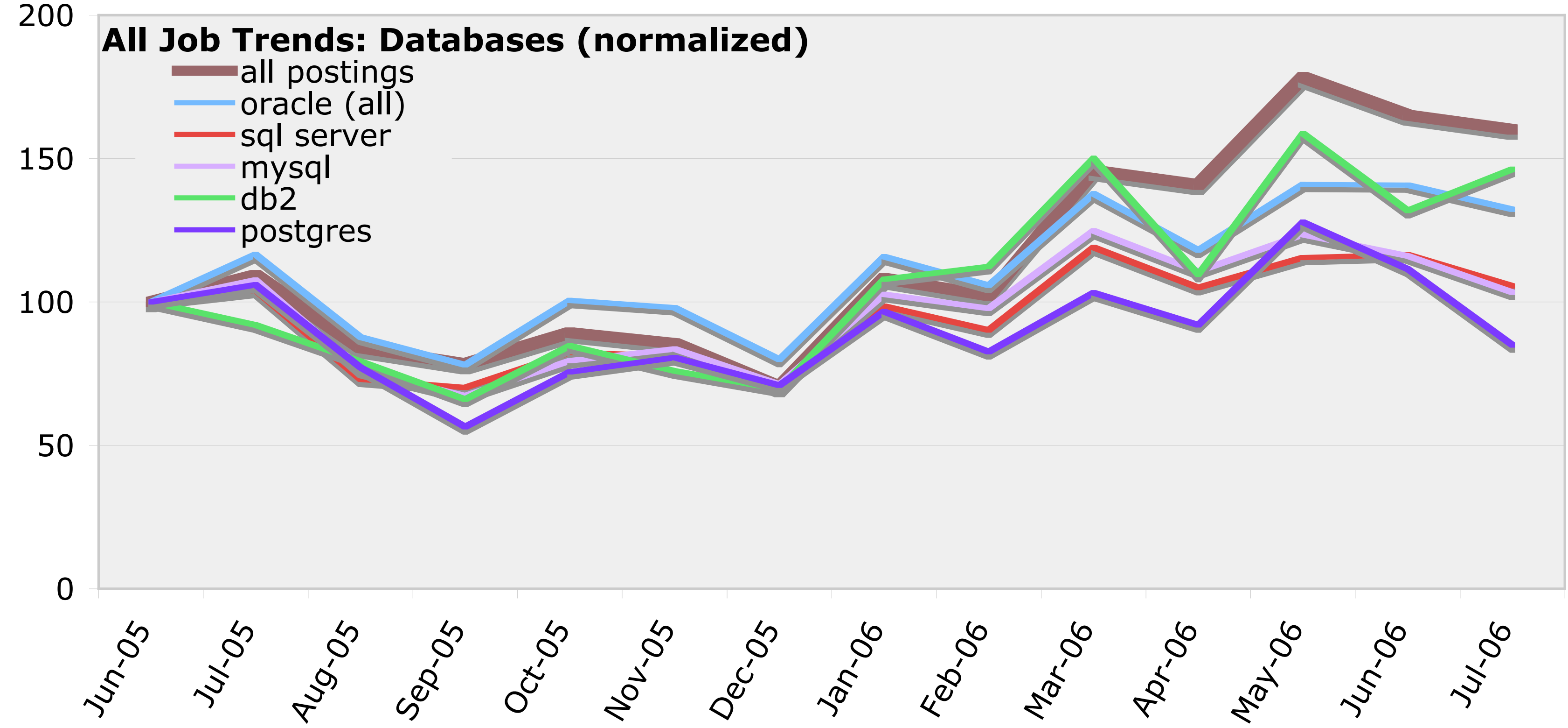


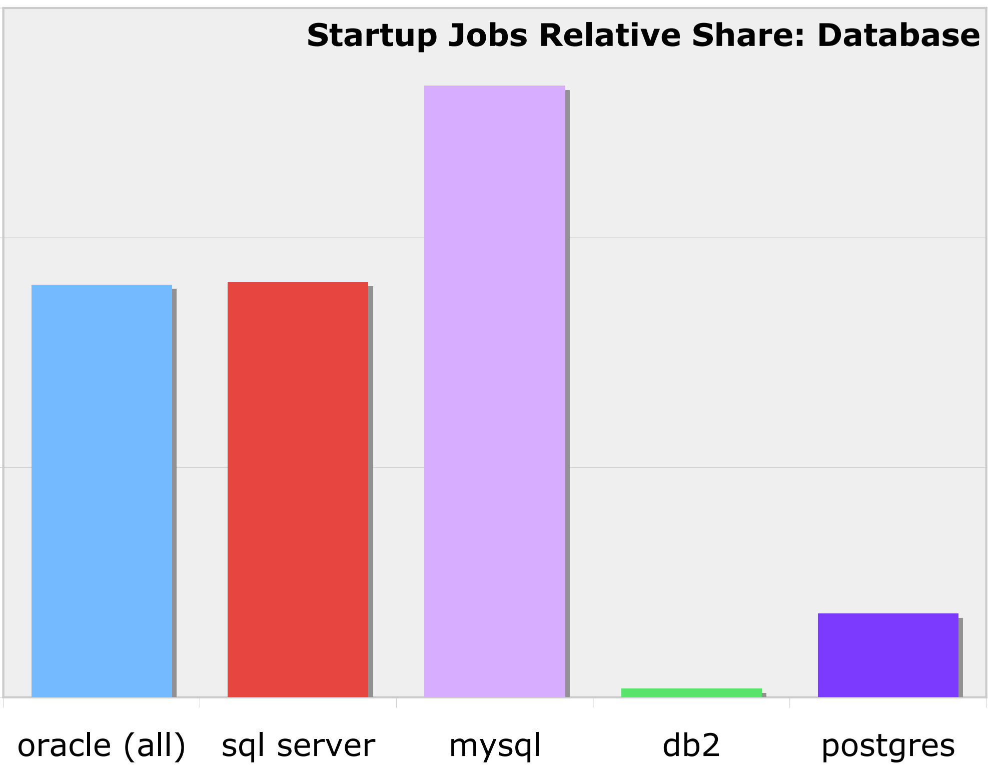


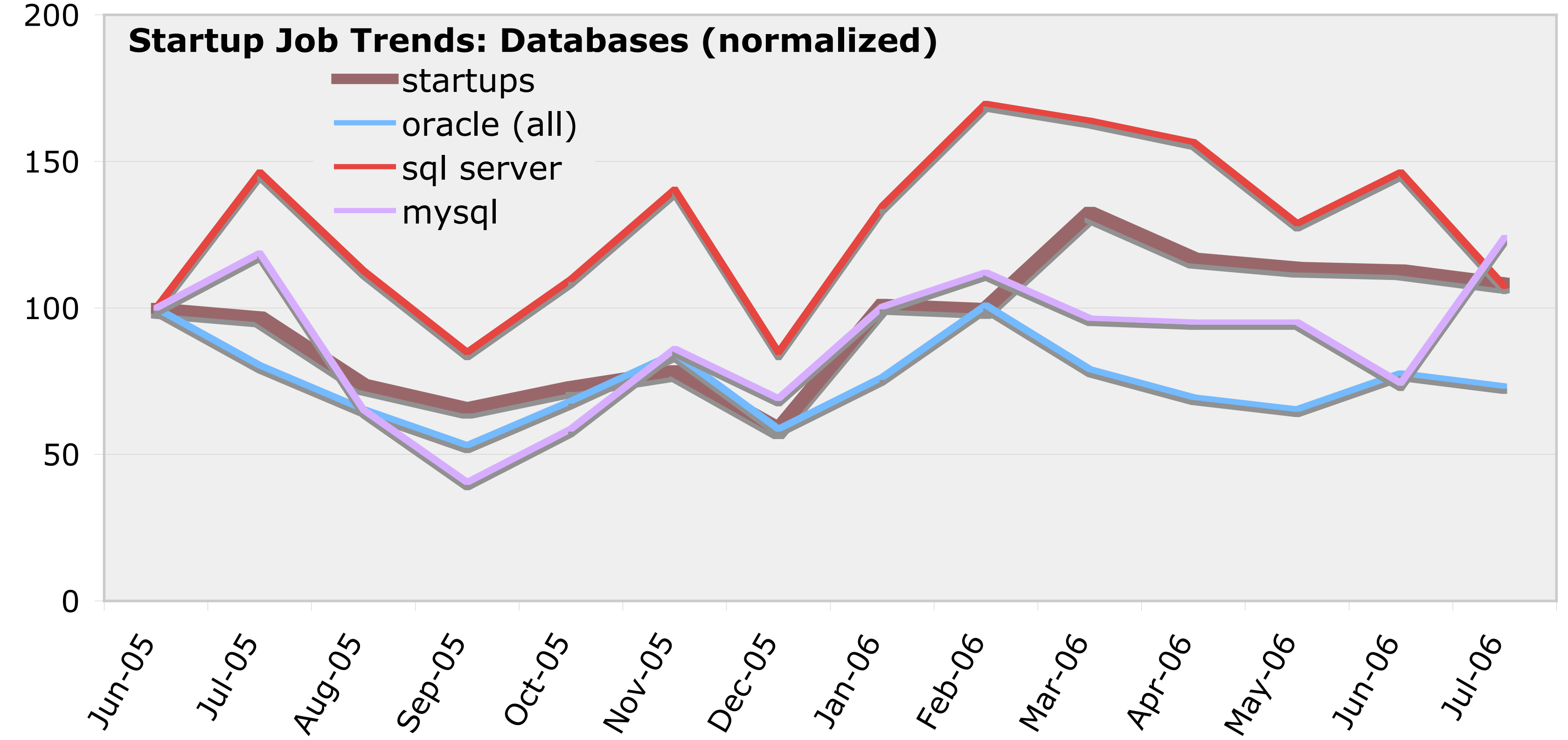
Databases

- Startups: MySQL more popular; Oracle loses more share than SQL Server
  - Oracle (all) includes Oracle applications
  - Too few DB2 Startup jobs, Postgres results too erratic to graph startup trends



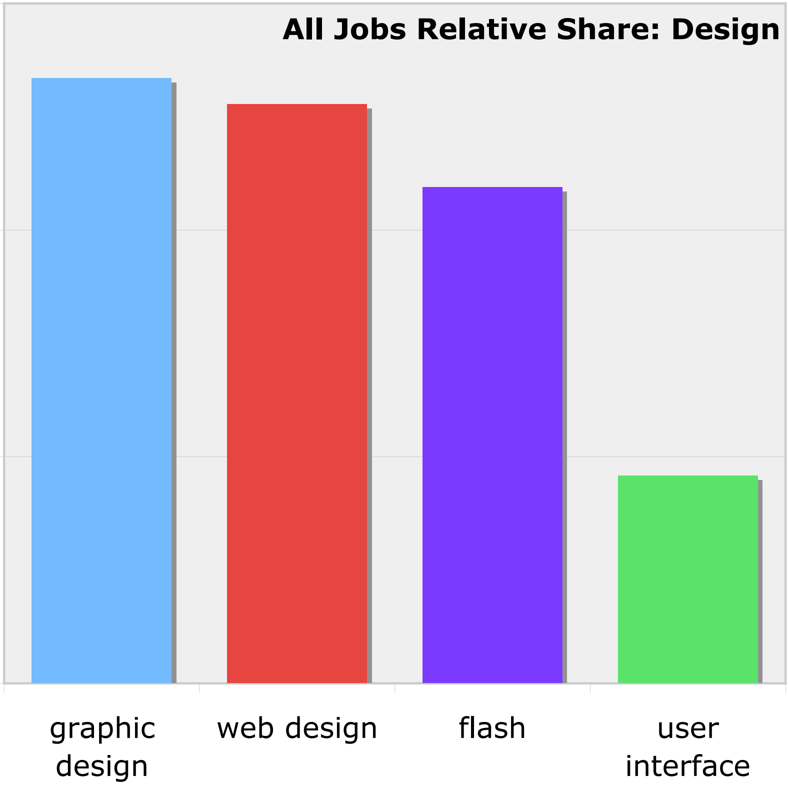


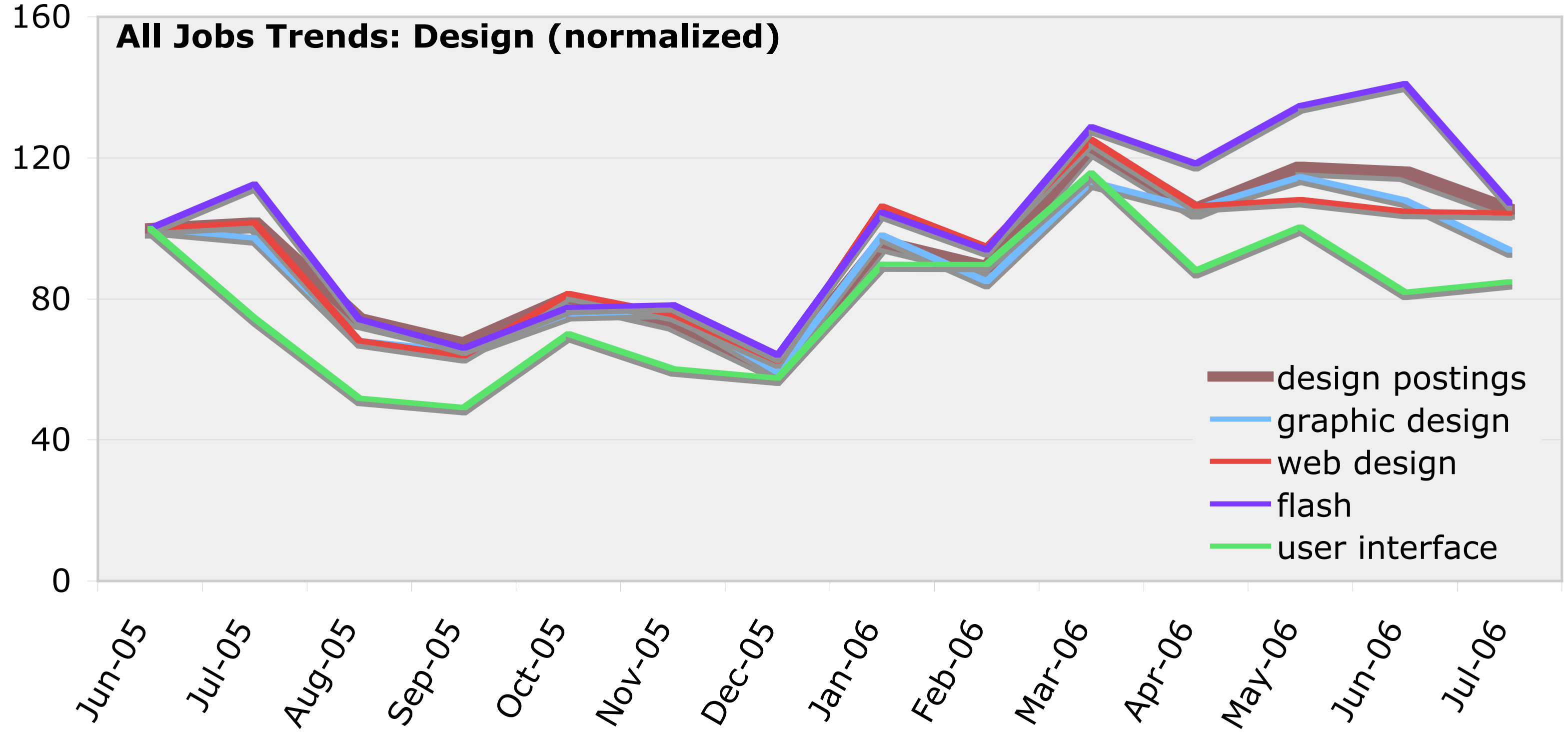


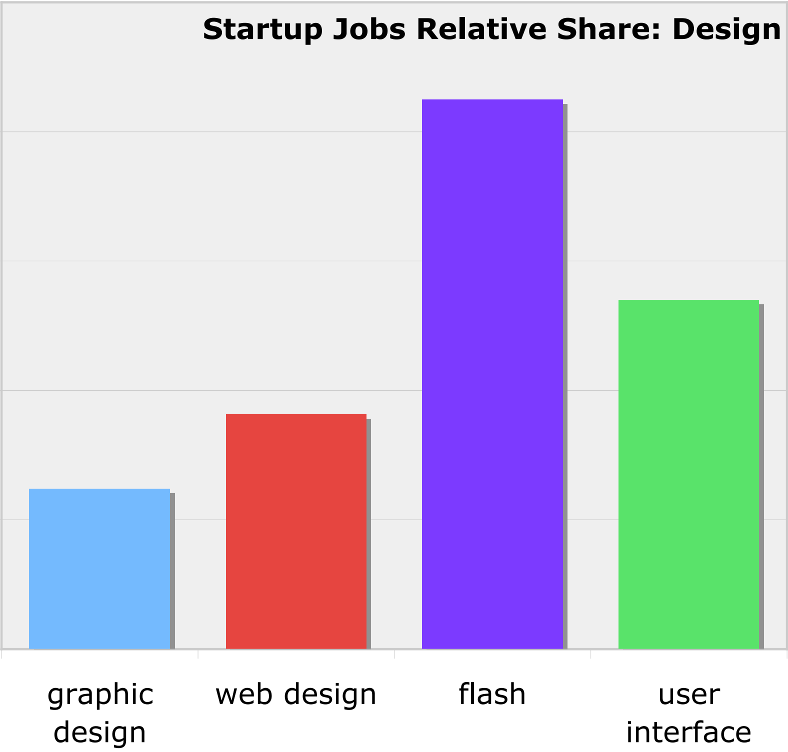


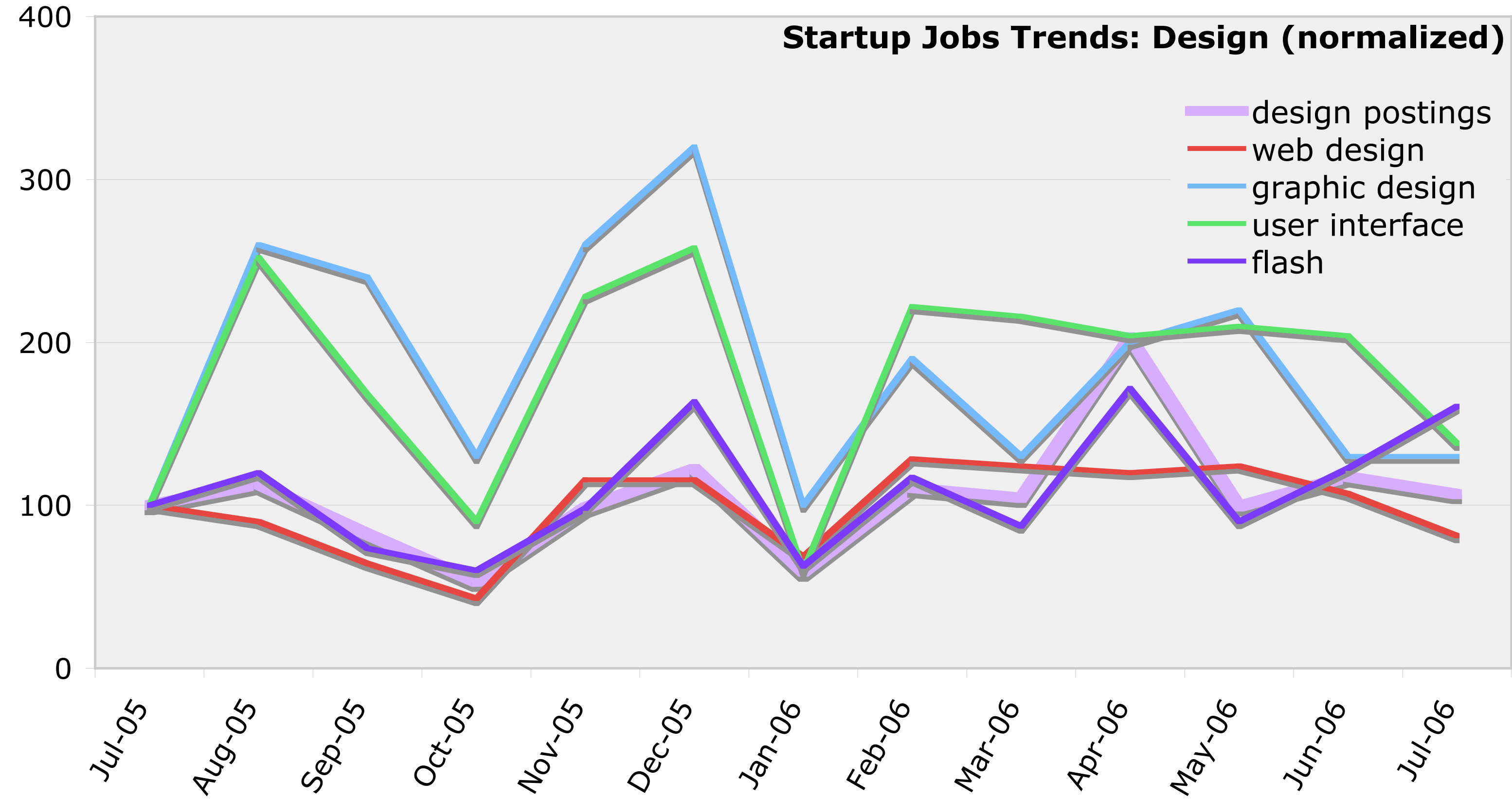
Design Topics

- **Flash most significant technology in design jobs**
- **Flash and User Interface have increased share of Startup Jobs**









Job Trends via Lucene Search

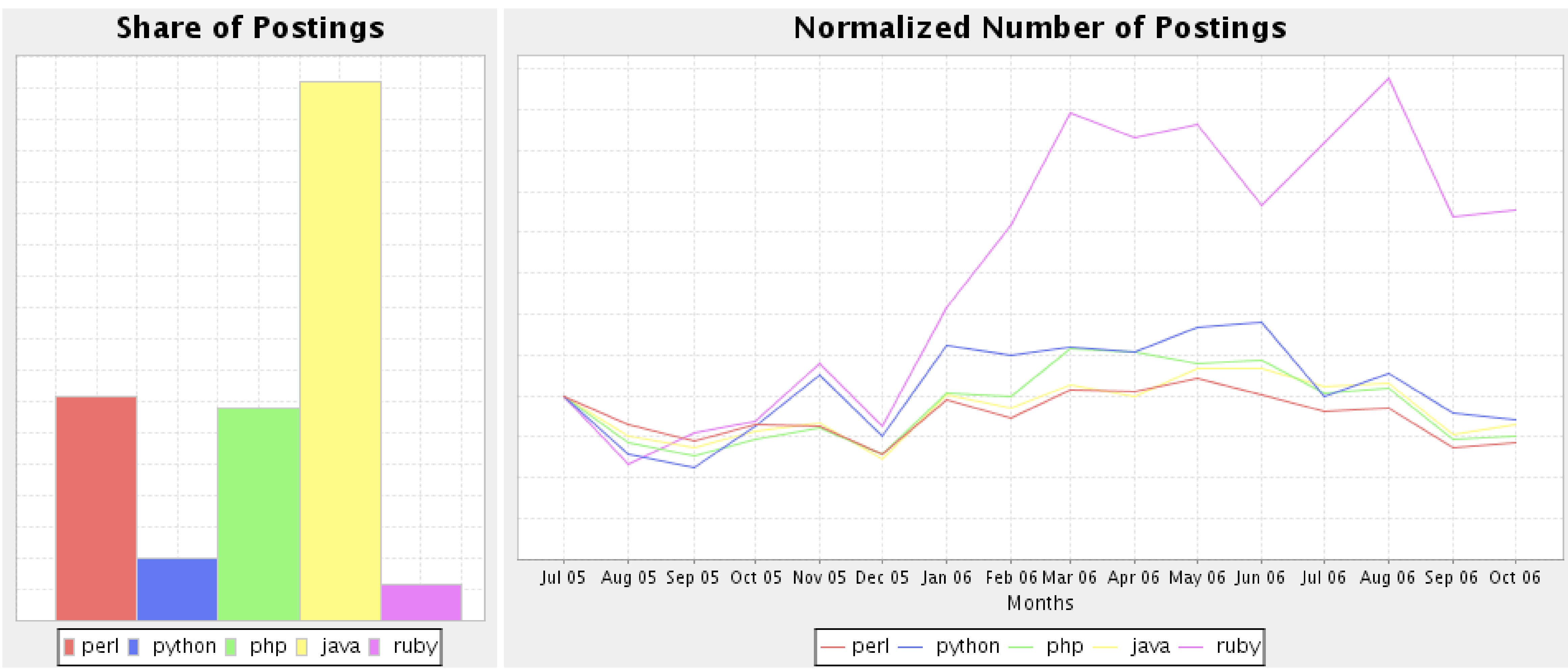
- **Sample of Job Data with Lucene indexes**
- **Results presented as Current Share and Time Series**
- **Disambiguation Issues**

O'REILLY® RESEARCH

Job Search

Search Query: perl, python, php, java, rut Show Graph

Tip: You can compare topics by separating with commas.



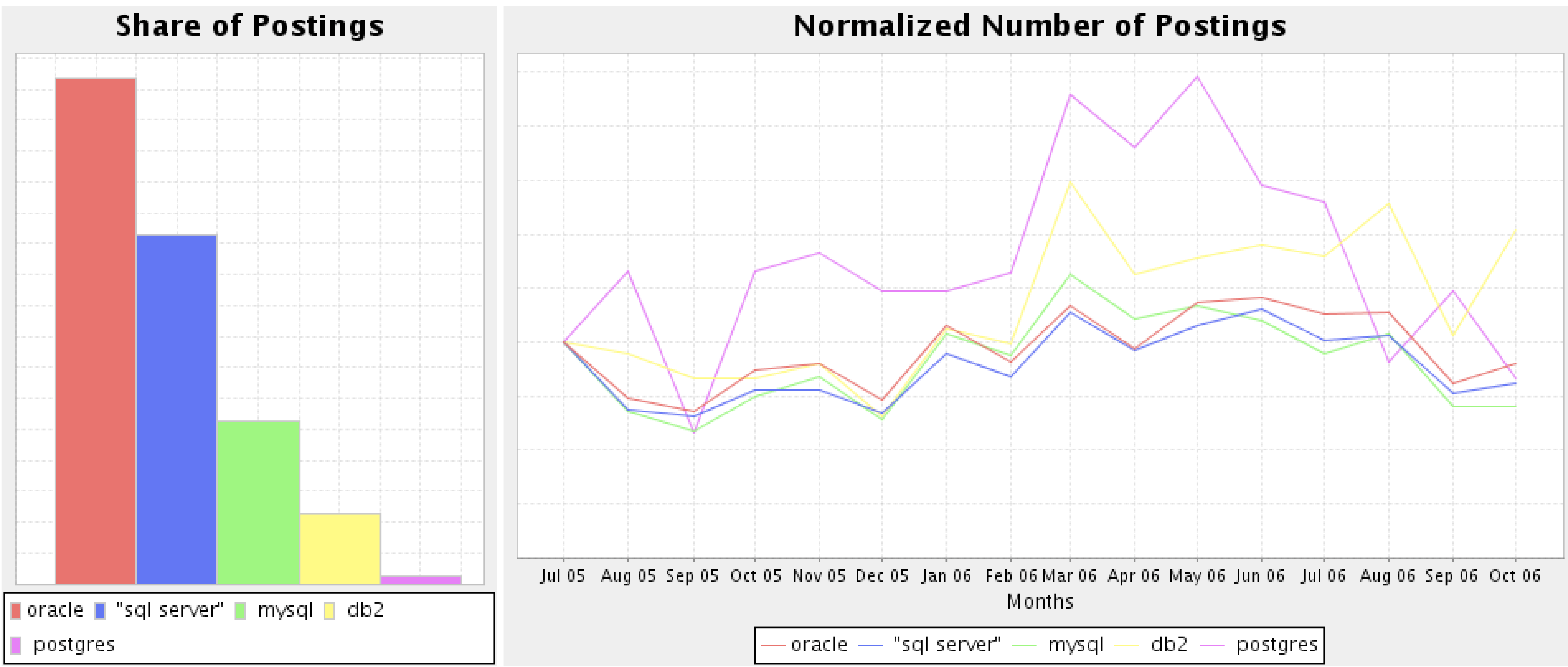
Database Job Trends via Search

O'REILLY® RESEARCH

Job Search

Search Query: oracle, "sql server", mysql, Show Graph

Tip: You can compare topics by separating with commas.



Unsupervised Learning: Topic Models

- **Occasionally we have had the need to analyze a large corpus of unstructured text: e.g. job postings or blogs.**
- **We were interested in a technique that would allow us to identify and "measure" the size of subjects in a corpus.**
- **In a recent project, we primarily used term frequency analysis, and confirmed and complemented some of our findings using a topic model.**
- **A topic is a probability distribution over words.**
  - **Topic models assume that documents are mixtures of topics.**
  - **Probabilistic Generative Process: A new document is generated by first choosing a distribution over topics. Each word in the document is selected by choosing a topic from the topic distribution, then selecting a word from the chosen topic.**

Topic Model

- **Statistical methods are used to invert the Generative Process, and uncover the "latent" topics**
- **Variational Bayes: An approximate procedure introduced by Blei, Ng, and Jordan.**
- **MCMC using a Gibbs Sampler: Griffiths and Steyvers**
- **Collapsed Variational Bayes: Teh, Newman, Welling (2006)**

- **We used MCMC / Gibbs Sampler**
  - **Monte Carlo simulation**
  - **Pick topic count and run until perceived convergence**
  - **Check results and rerun**
    - increase topic count if topics too broad
    - decrease topic count if topics redundant
- **Art and Science**
  - **Knowing how many topics to start with**
  - **Domain knowledge to judge model topic quality**

Topic Model: Topic Size and Trends

- **As an output of the model, each distinct word inside a document gets assigned to an appropriate topic**
  - **The size of a topic is the count of words assigned to it.**
  - **A job posting is possibly a mixture of words from several topics.**

Job Posting #1: Word Java is assigned to the topic "Java & Web Development"
Java Web Developer with Austin-based startup … is seeking a **Java** software engineer with 1-2.5 years professional **Java** experience to lead web development initiatives.

Job Posting #2: Word Java is assigned to the topic "Open Source Web Development"
Senior PHP developer … By building a scalable and distributable content management cluster, developing state-of-the-art server-side **Java** applications, and forging the frontier of website UI using AJAX and PHP, … is positioned to cause a significant stir on the semantic web later this year. And we are looking to hire senior software professionals including an experienced PHP Software Engineer: Requirements: y 2+ years experience using PHP/MySQL

Job Posting #3: Word Java is assigned to the topic "Mobile Apps"
Are you passionate about wireless technology, love mobile devices and are looking to be part of a growing team that delivers cutting edge products to some of the largest players in the telecommunications space? … JOB RESPONSIBILITIES --Design and development of mobile applications for J2ME (**Java**), BREW PalmOS, Windows Mobile and Symbian platforms.

Topic Model

- Text Mining used to gain additional insights and supplement term frequency analysis
- The topic model is a probabilistic model which postulates that a job posting is generated by a mixture of (latent) topics.
  - Startup job postings are generated by first picking topics (from a distribution of topics), then picking words which are prevalent in a topic.
  - Algorithmic technique to identify emerging trends and discover “unknown unknowns” in the data
- Generally, the higher the relative topic size (in parens) for a topic, the more the topic appears in the job postings
  - If the 50 topics in model were equally distributed, topic size (value in parens) would be 2.0%
- Words/technologies associated with a topic are presented in descending order of probability of appearing with the topic
  - The first terms appear more frequently than the later terms
- Descriptive patterns noted in topics and word probabilities

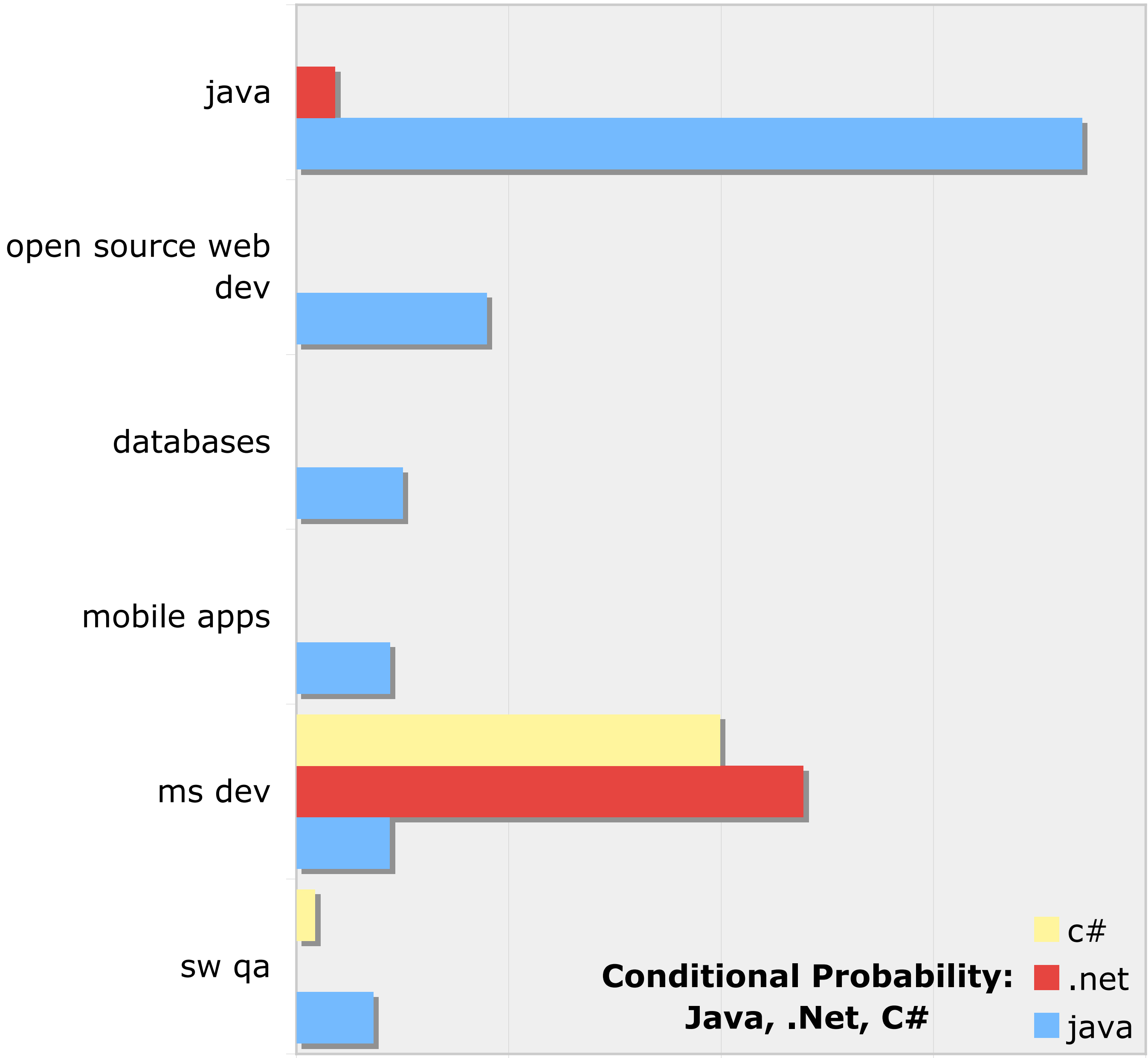
Startup Topics

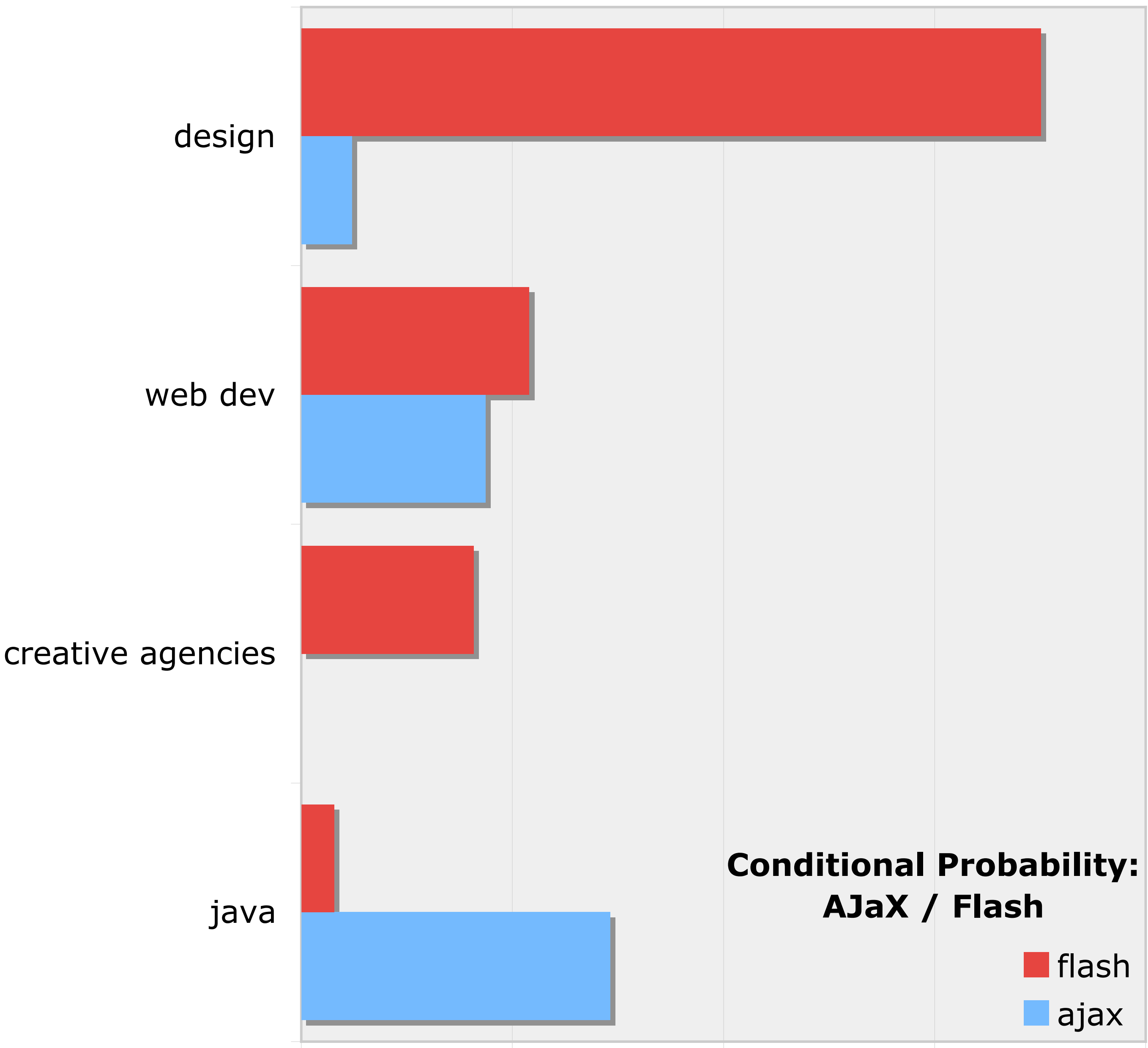
- **Typology that emerges from semantic analysis***
  - **open source web development (3.7%)**
    - php, mysql, linux, html, javascript, xml, java, perl, apache, css, sql, flash, databases, unix, ajax, python, dhtml, c/c++, video, asp, jsp
  - **microsoft development (2.8%)**
    - .net, c#, windows, sql server, asp.net, c++, xml, visual (studio), java, database, sql, vb.net, win32, javascript
  - **java & web development (2.6%)**
    - java, j2ee, javascript, jsp, ajax, struts, xml, hibernate, tomcat, spring, ruby, servlets, eclipse, css, patterns, mysql, jdbc, rails, swing, ant, jboss, agile, dhtml, linux, apache, oracle, database, web 2.0, ejb
  - **design and web design (2.0%)**
    - flash, html, designer, photoshop, css, graphics, illustrator, usability, adobe, layout, javascript, dreamweaver, dhtml, actionscript, xhtml
  - **databases (1.7%)**
    - database, oracle, sql, performance, modeling, tuning, dba, sql server, java, reporting, relational, intelligence, reports, pl/sql, j2ee, unix, xml, mysql
  - **mobile apps (1.7%)**
    - mobile, wireless, video, (palo alto, phoenix), java, j2me, c++, windows, brew
  - **embedded software and devices (1.7%)**
    - c/c++, linux, windows, firmware, components, kernel
  - **enterprise software (0.9%)**
    - enterprise, crm, supply chain, erp, oracle, sap, peoplesoft, siebel, ariba, asp (hosting),

* relative topic size in (parens)
* words in order of declining probability

Startup Topic Model Word Probabilities

- **Shows technology distribution by topic**
  - **no bar, no probability of word in topic**
- **.Net concentrated in Microsoft Development topic**
- **Flash for Design; AJaX for Development**



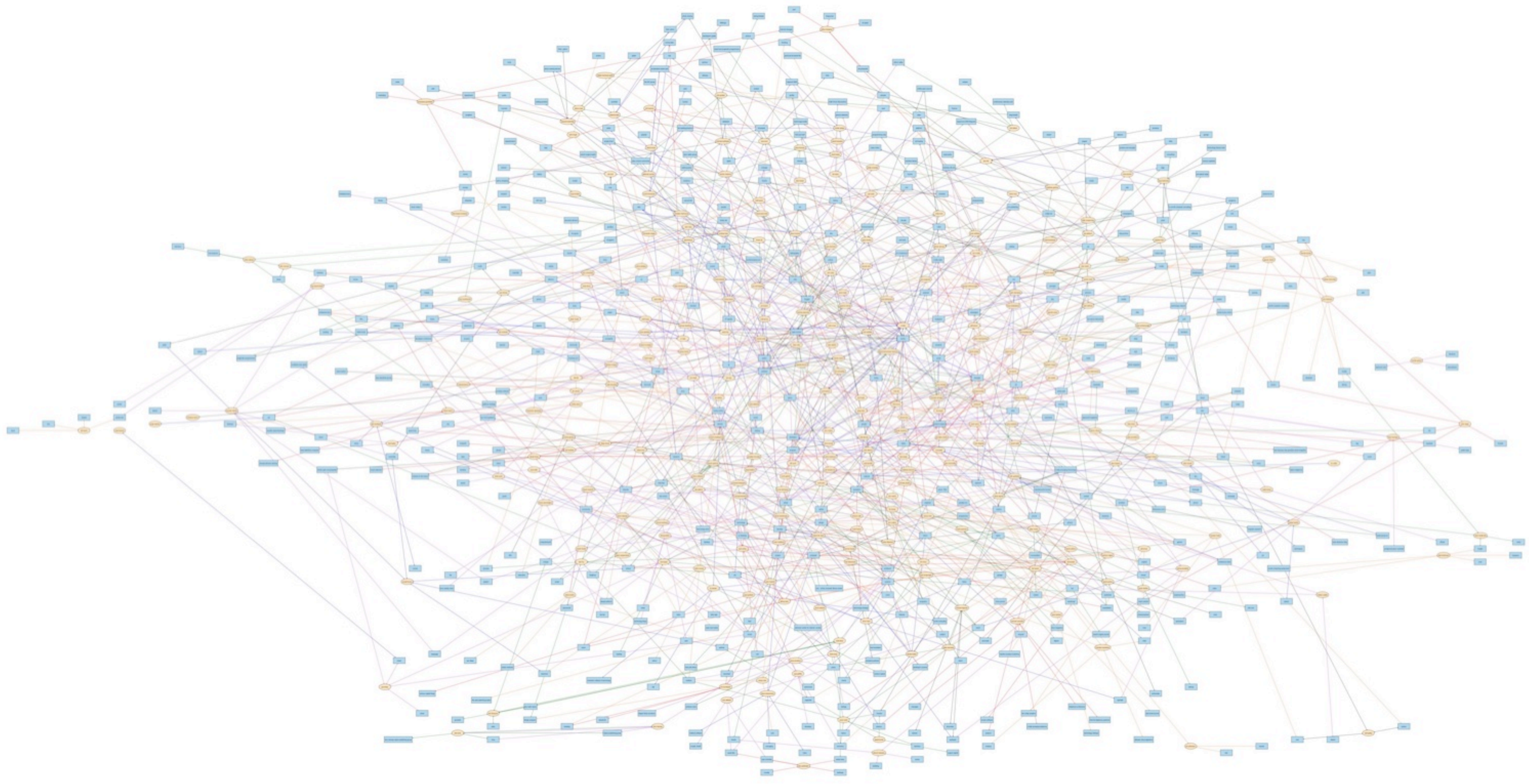


Startup Topic Model Results

- **Combining 'open source web development' and 'java and web development' shows more than double the word occurrence than second ranked 'microsoft developer' topic**
  - **relative rate of 6.3% vs. 2.8%**
  - **startups appear to be requesting open source development frameworks at double the rate of Microsoft frameworks**
- **Silo effect noted for Microsoft technologies**
  - **Microsoft technologies appear in 'microsoft developer topic' but very unlikely appear in other topics**
    - SQL Server appears in 'microsoft developer' and 'databases' topics
    - Windows appears in mobile apps and embedded topics
  - **Java, Javascript, AJaX, Flash appear in multiple topics**
- **Java used significantly for Web Development by Startups**
- **MySQL top database of choice for Web Development**
  - **MySQL appears with less probability at end of in 'database topic'**
- **Flash dominant technology in design topic**
  - **Javascript and Actionscript also appear, but less frequently**

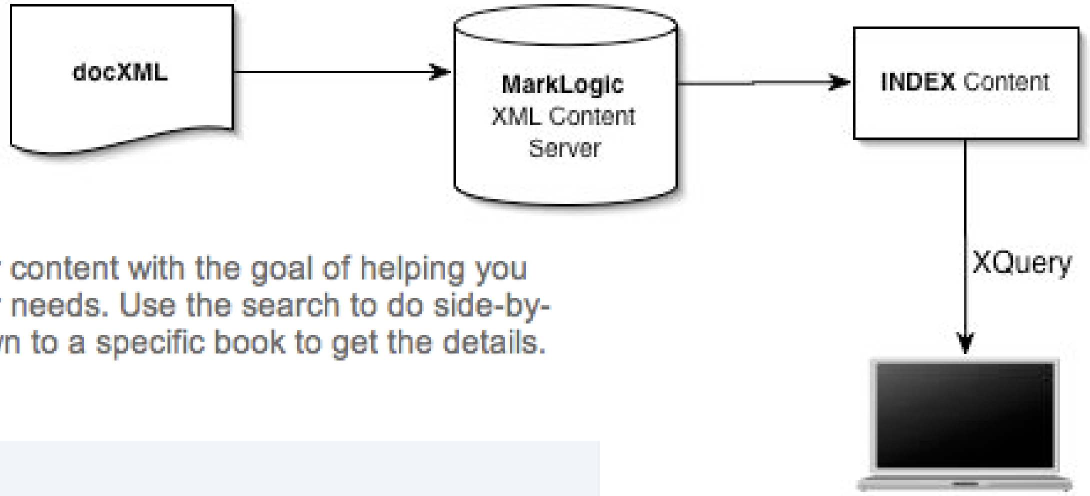
Social Networks: FOO Camp

- **Foo Camp - An experiment in face-to-face Social Networks**
  - **What can we learn from attendees**
    - Collarity search used to seed tags
      - User search behavior clustered to create natural, implicit communities of subject-matter experts
      - Communities and clusters used to generate user tags
    - Compare to use generated Tags
    - The dreaded tag cloud + directed graph
- **Trying to figure how to mine social networks for trends**



Book Content

- **Mark Logic / XQuery**
  - **Indexed Content**



- Content Statistics

Content Statistics
beta

We've collected some statistics on our content with the goal of helping you figure out what content best suits your needs. Use the search to do side-by-side comparisons of books or drill down to a specific book to get the details.

Averages

Page Count:	460
Word Count:	451,166
Words/Page:	979
Chapters:	14
Sections:	124
Reference Sections:	75
Tables:	24
Figures:	92
Sidebars:	15
Paragraphs:	4,109
Tips:	61
Blocks of Code:	183
Lines of Code:	3,732
Index Terms:	1,783

Totals

Page Count:	309,647
Word Count:	303,183,808
Words/Page:	979
Chapters:	9,550
Sections:	83,797
Reference Sections:	50,906
Tables:	16,143
Figures:	62,015
Sidebars:	10,229

Search

Search by book title, isbn, author or use a fielded search like:

tag:replication	**tag:** search in books containing a specific tag
cat:perl	**cat:** allows you to search in a specific category
pubyear:2004	**pubyear:** is used to narrow the search to a specific year
author:hunter	**author:** restricts to books by an author

Try It

Top Tags

addresses applications arrays attributes authentication backups browsers classes clients code commands components configuration configuring controls creating data data types databases datatypes debugging deleting directories DNS documents domains elements email encryption environment variables errors events exceptions files filesystems folders formatting forms functions hardware headers HTML HTTP images installation interfaces Internet Explorer IP addresses Java keyboard shortcuts Linux lists logging memory menus messages methods modules MySQL names networking networks numbers objects operators Oracle packages parameters passwords performance Perl permissions printing processes programs properties queries quick reference regular expressions sample code scripts security servers SQL strings tables tags templates text threads transactions troubleshooting Unix URLs users variables web services web sites windows XML

Summary / Observations

- O'Reilly somewhat unusual in its use of Natural Language Processing / Machine Learning (NLP/ML) are important analysis tools for O'Reilly Research
  - Desire to mine information and trends from structured and unstructured text
- NLP/ML used as recommendation engines to speed up classification
  - 65-75% accurate (SVM)
  - Manual review required
  - Build into taxonomy admin screens
- Combination of supervised and unsupervised NLP/ML techniques will be used to create new taxonomies
- The Web has created large sources of interesting unstructured data
- Organizations housing large volumes of unstructured data are increasingly interested in NLP/ML to help organize and make sense of data, to spot trends, help with search and understand user behavior
- Requires specialized skills to implement
  - Techniques require art and science
- We consider NLP/ML a complement to tagging / folksonomies

References: Kernel Methods

- **SVM vs. other Text Classifiers**
  - **Thorsten Joachims, *Text Classification with Support Vector Machines*. 1997**
  - **Yiming Yang and Xin Liu. *A re-examination of text categorization methods.* In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.**
- **RBF and Linear SVM's**
  - **S. Sathiya Keerthi and Chih-Jen Lin. *Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel*. Neural Computation. 2003;15:1667-1689.**
- **Multi-class SVM's**
  - **Kai-Bo Duan and S. Sathiya Keerthi. *Which Is the Best Multiclass SVM Method? An Empirical Study.* Lecture Notes in Computer Science. Springer. 2005**