Distributed Data Architectures and a little ENCODE

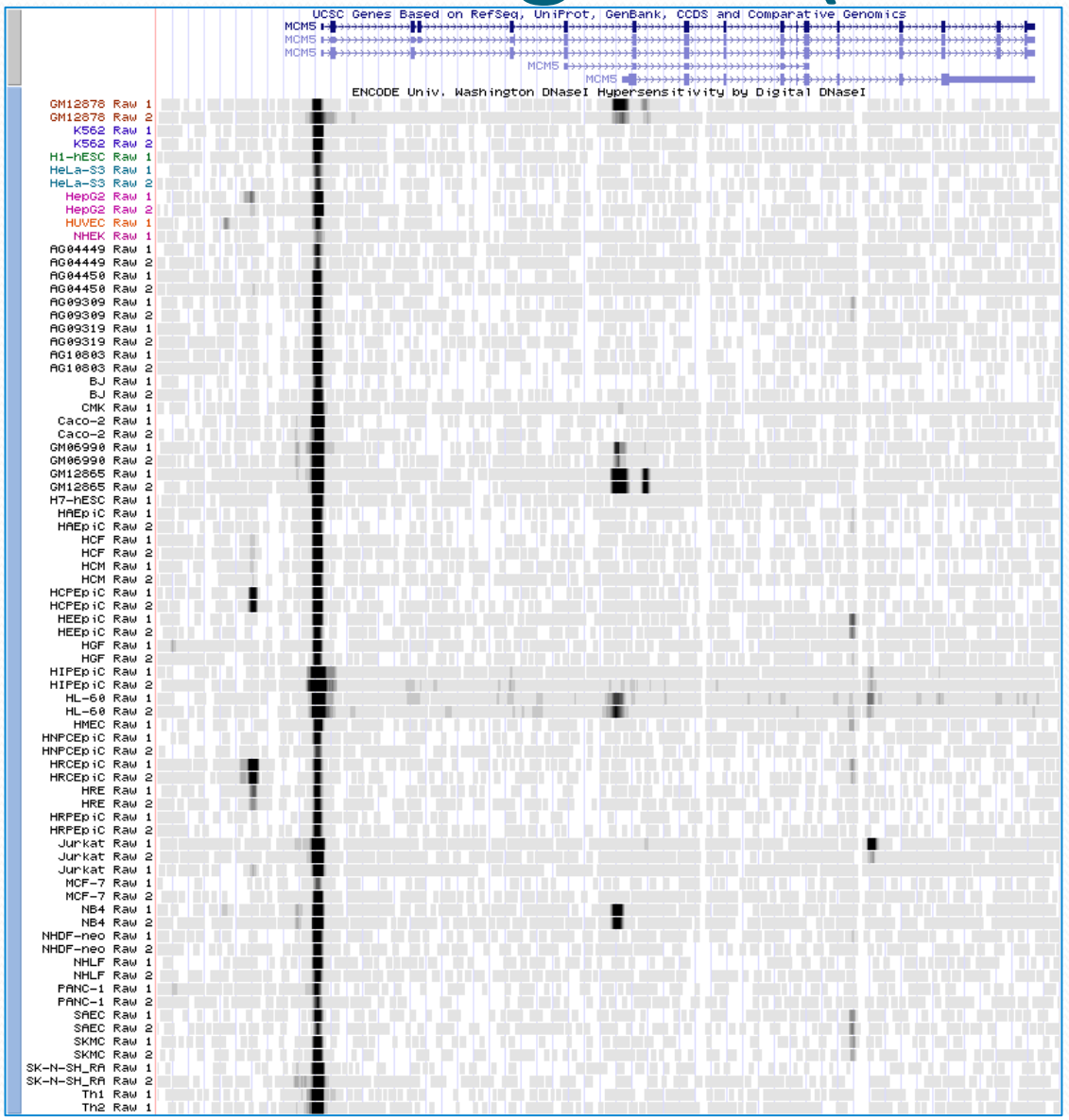
Jim Kent, UC Santa Cruz

A little ENCODE

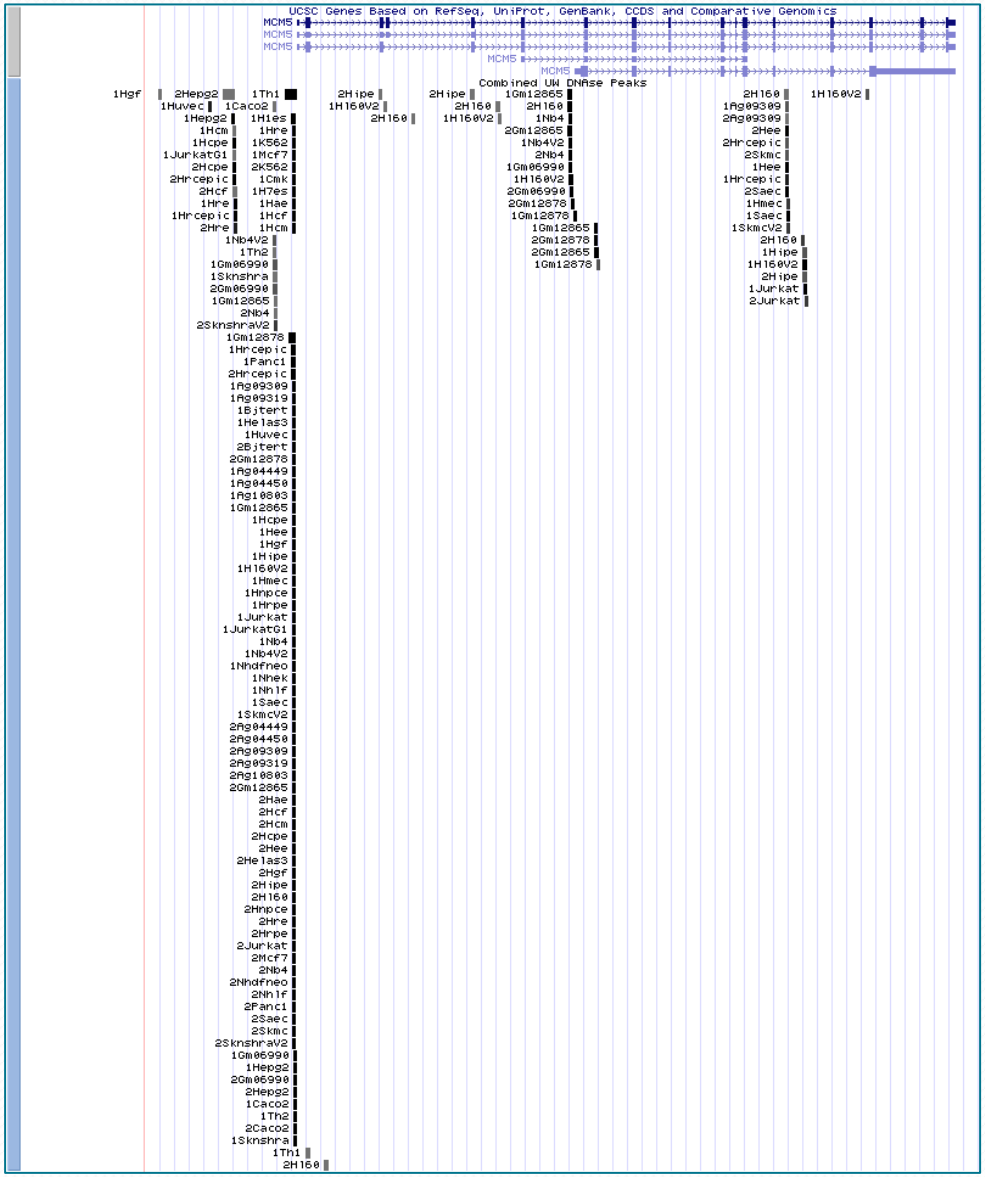
There is a need to do integrated tracks!

- Some work going on at UCSC
- Hope to bring in integrated tracks from analysis working group as well

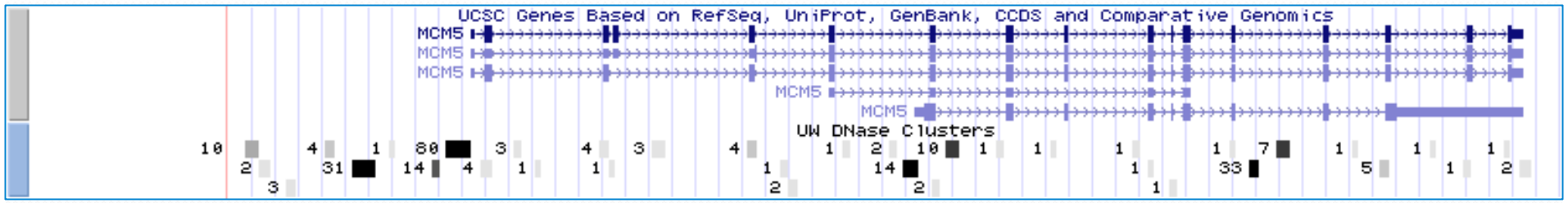
UW DNAse all signals (so far!)



UW DNAse Combined Peaks



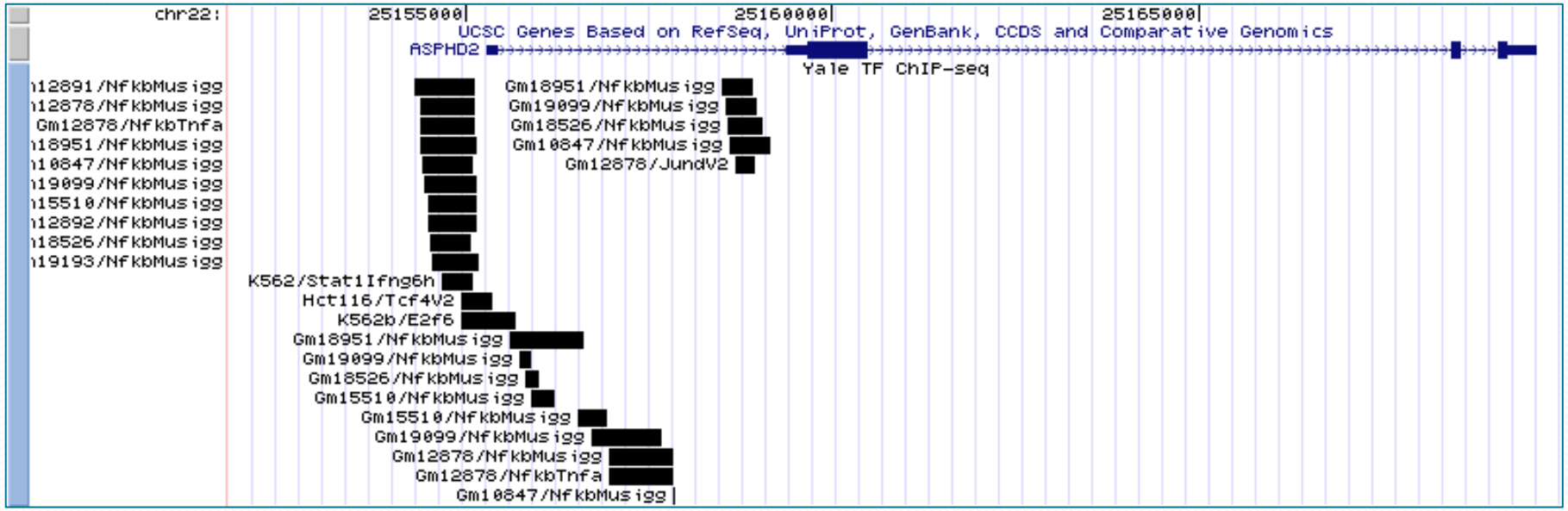
UW DNAse Merged



The number to the left indicates the number of samples (replicates of cell lines) in which the peak is seen.

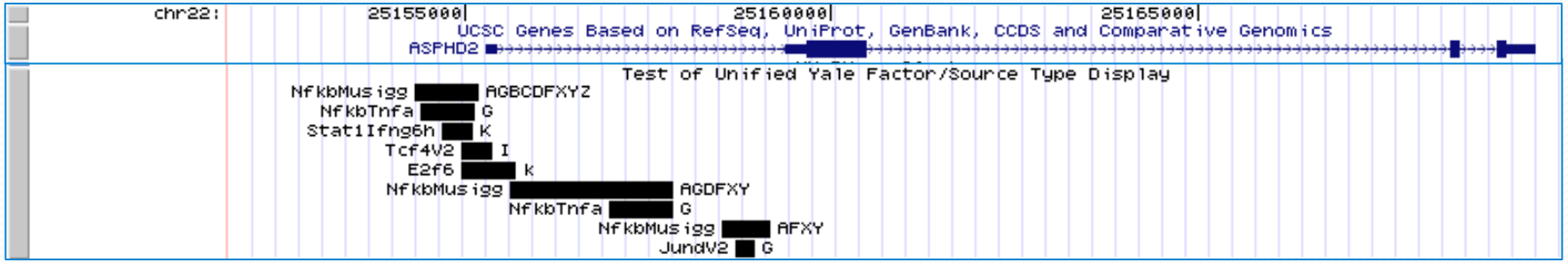
Yale TFBS Raw Signals (no Pol)

Yale TFBS Combined Peaks



Various peak-called “tracks” combined into a single track with label being cell-line/factor. Good, but still get “stacks,” mostly of common factors seen in many cell lines.

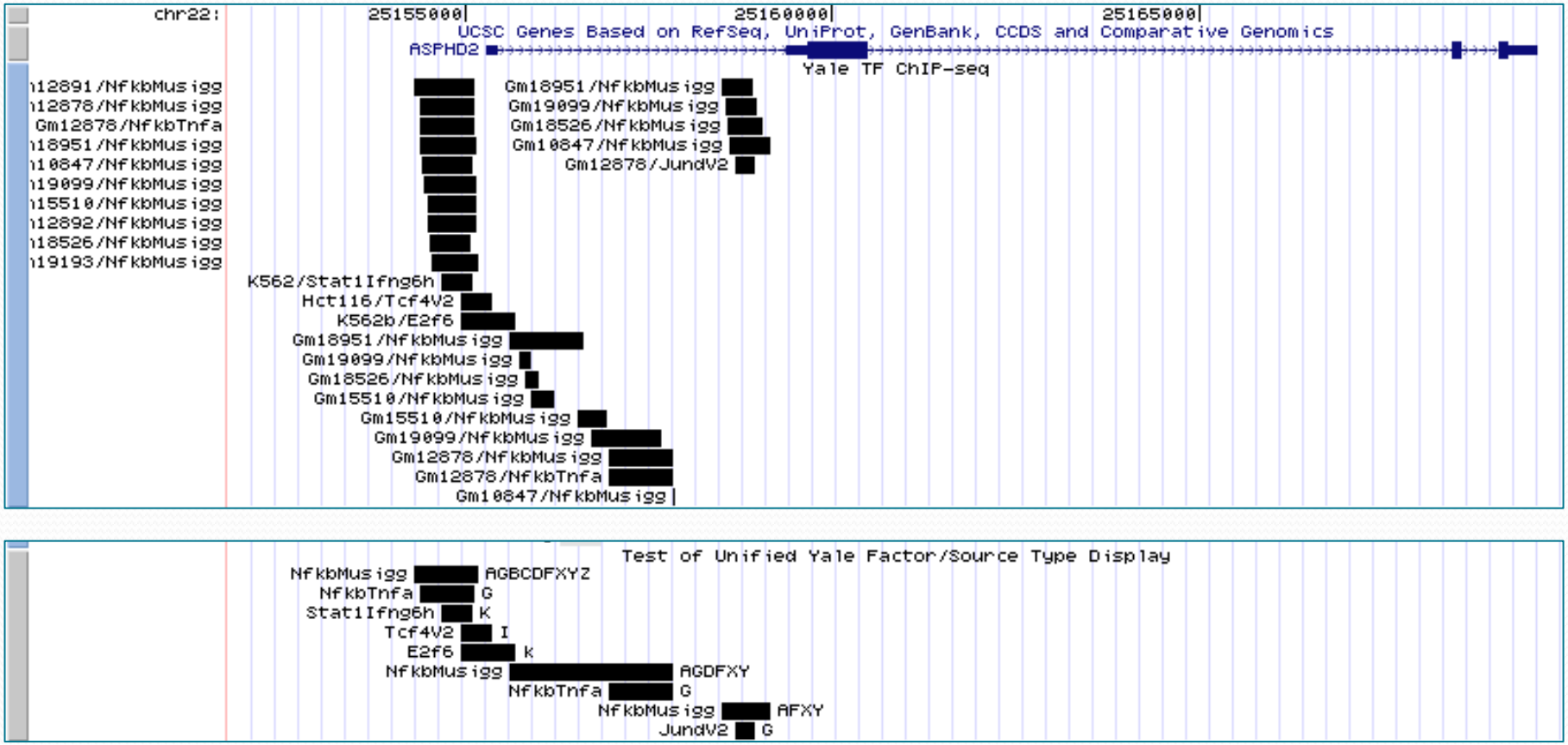
Yale TFBS Merged Peaks



Letters to right indicate which cell lines peak is seen in:

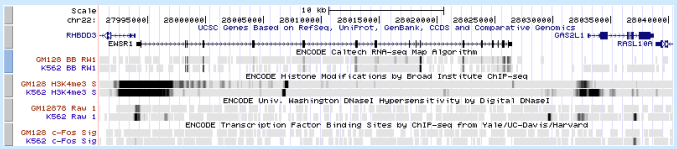
A Gm10847	I Hct116
B Gm12891	K K562
C Gm12892	L Hepg2
D Gm15510	M Mcf7
E Gm18505	N Nt2d1
F Gm18526	X Gm18951
G Gm12878	Y Gm19099
H Helas3	Z Gm19193

Merging Process

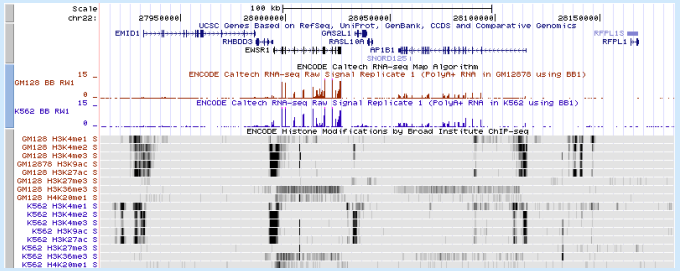


Very simple minded – peaks from same factor in different cell lines that overlap are merged. Extents of merged peak encompass all peaks it is made of.
Would be happy to use something more sophisticated from analysis working group.

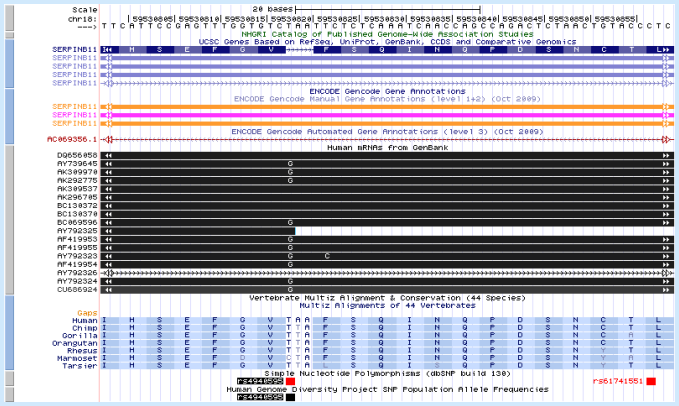
Browser ENCODE Session Gallery



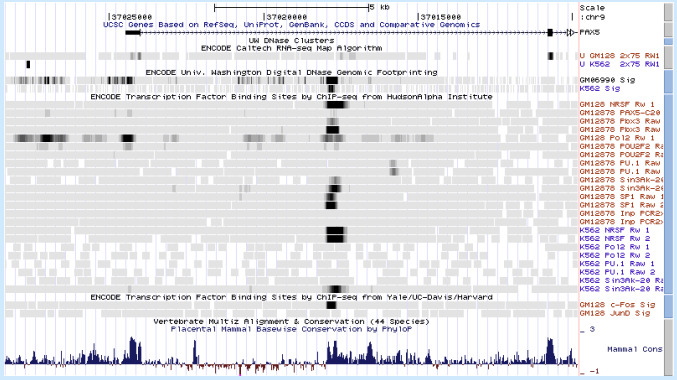
Basic ENCODE data on the tier 1 cell lines GM128, a lymphoid cell line and K562, a myeloid line. The tracks include RNA-seq which shows the level of expression of the gene EWSR1, the histone marks H3K4Me3 which is associated with promoters, DNAse I hypersensitivity which is associated with regulatory regions in general, and ChIP-seq showing levels of occupancy by the transcription factor c-Fos.



Histone ChIP-seq data on the two tier one cell lines. H3K4me1 is associated with enhancers and to an extent with promoters, H3K4me3 is associated with promoters, and H3K36me3 with transcribed regions.



A premature stop codon that is found in the reference genome and about half of people of European descent. The UCSC Genes track is forced to skip the codon.



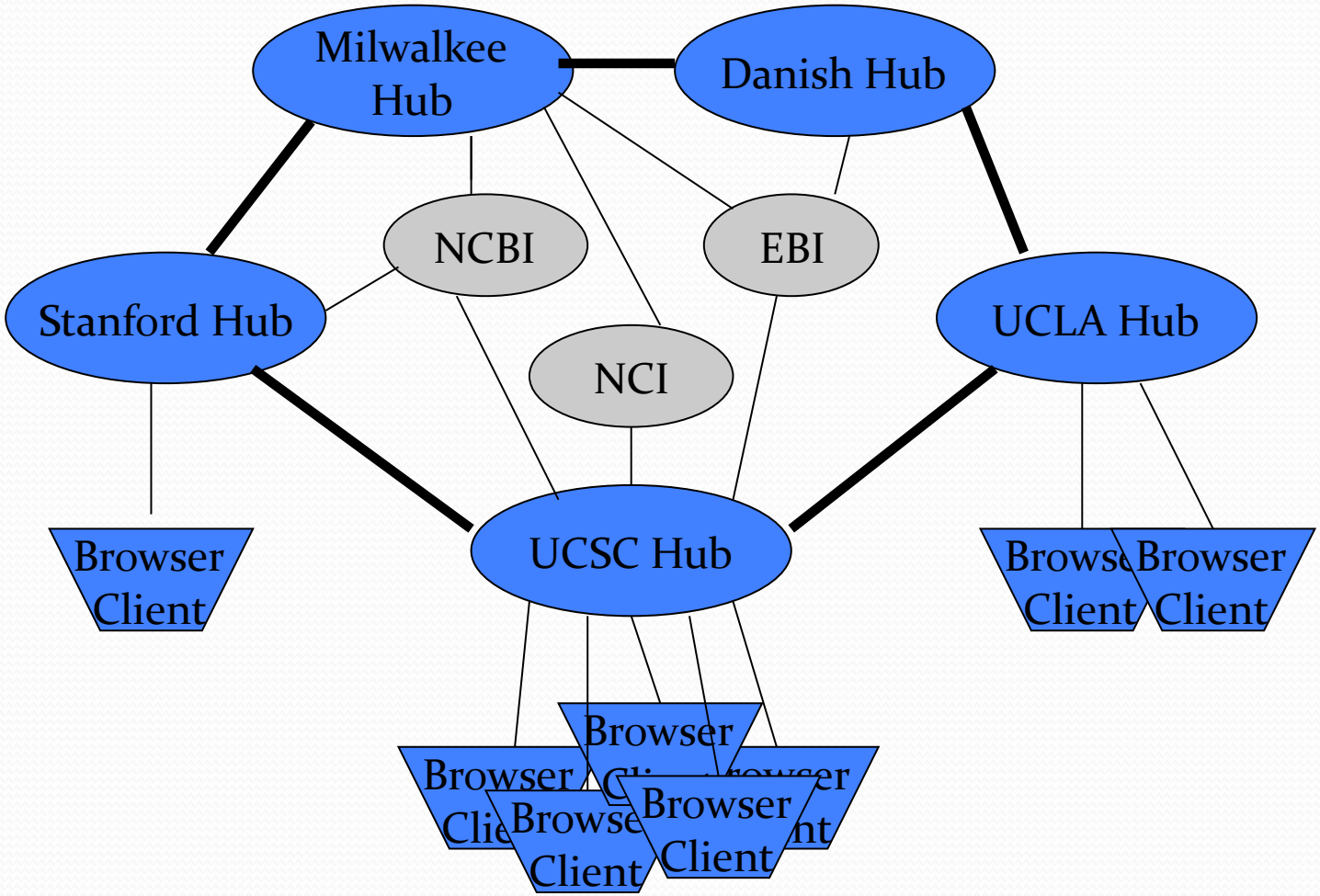
Regulatory elements in the promoter and first intron of the transcription factor PAX5, a gene expressed in GM128 but not K562 cells. The data suggests a complex regulatory circuit with some autoregulation.

Gallery linked to main ENCODE page. Clicking on image takes you to browser at that spot with those tracks configured as shown. Caption explains session to non-specialists.

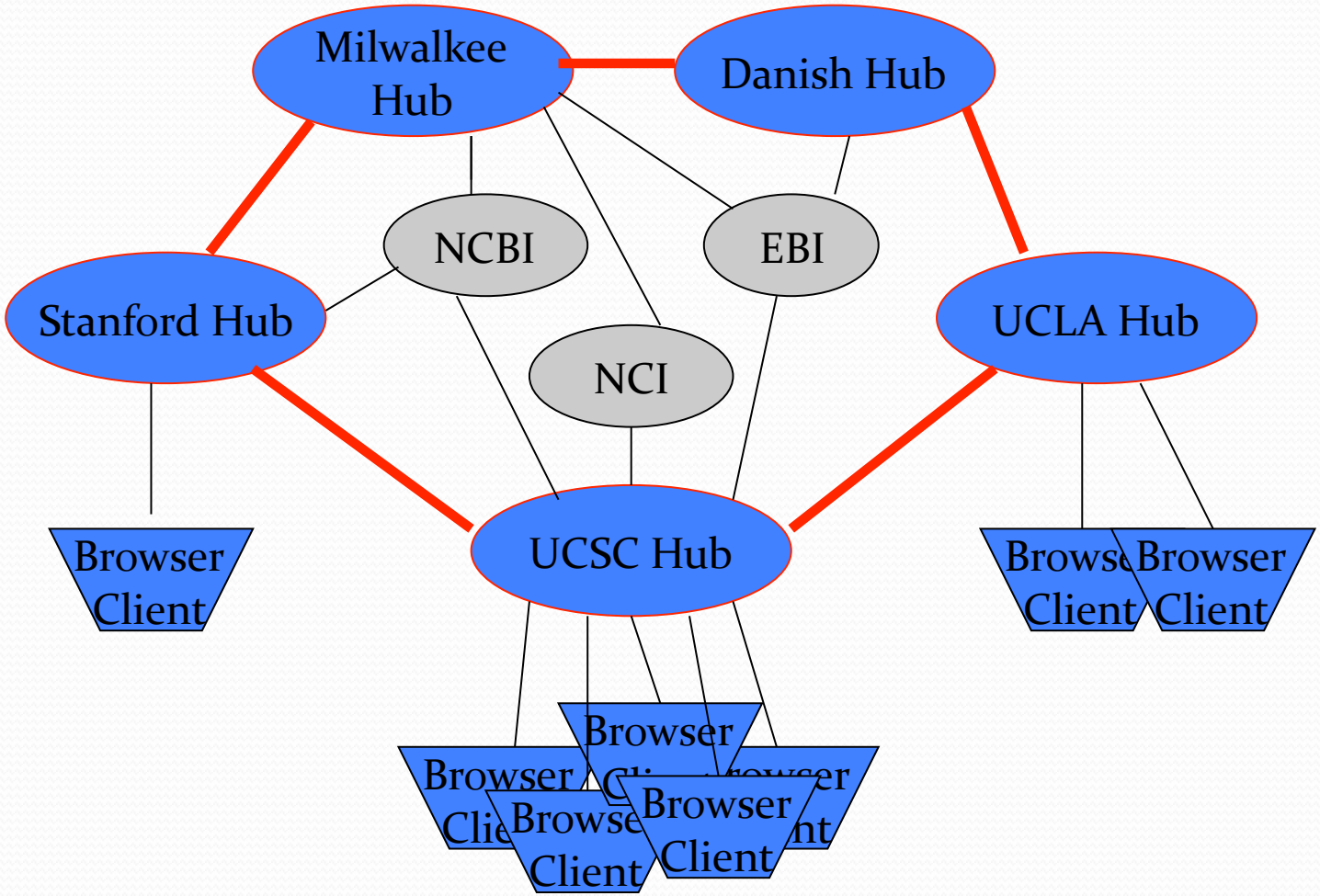
2009 Proposal for Architecture of UCSC Genome Browser

Jim Kent
Genome Bioinformatics Group
University of California Santa Cruz

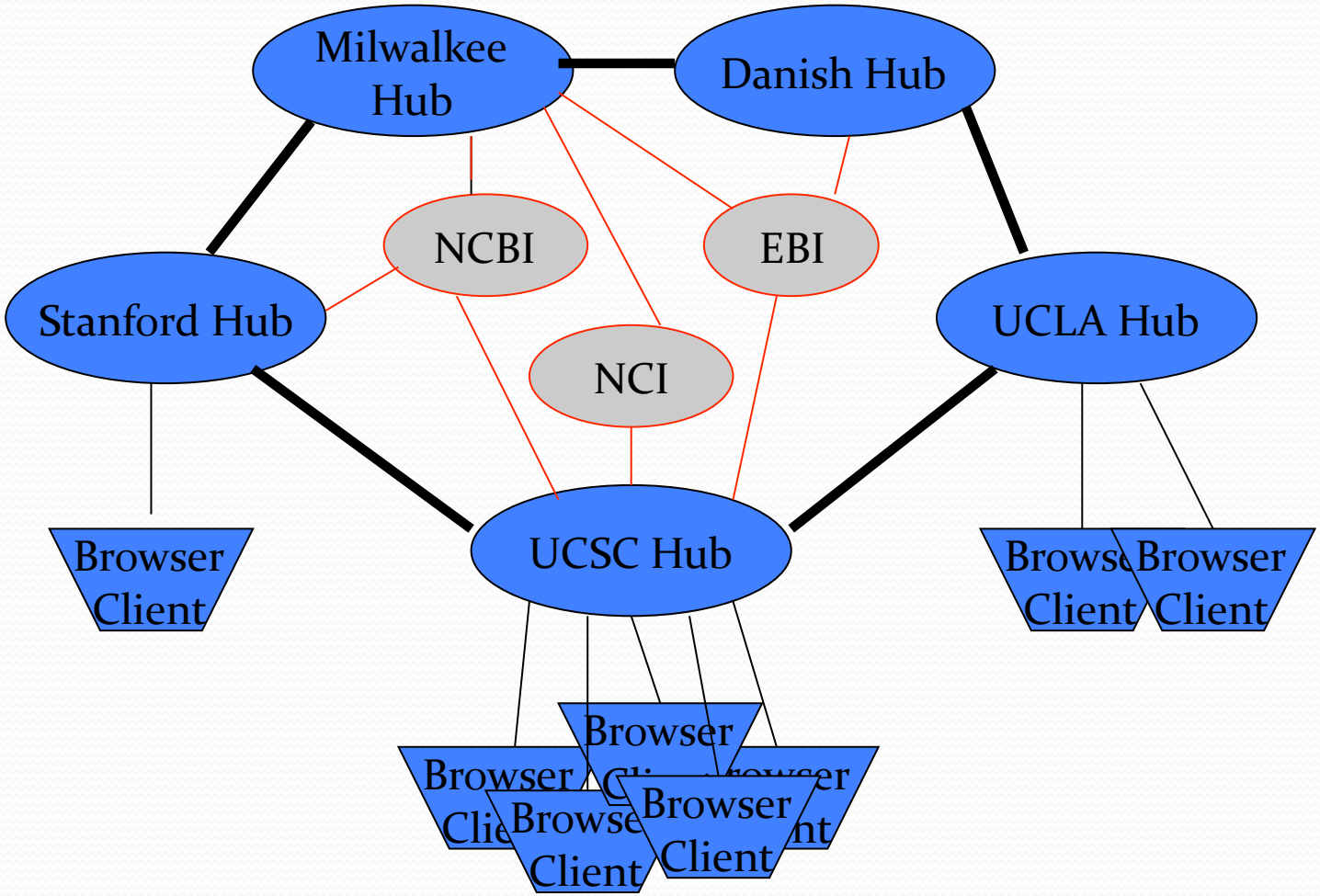
Overall Architecture – JavaScript Clients, C Web Services hubs.



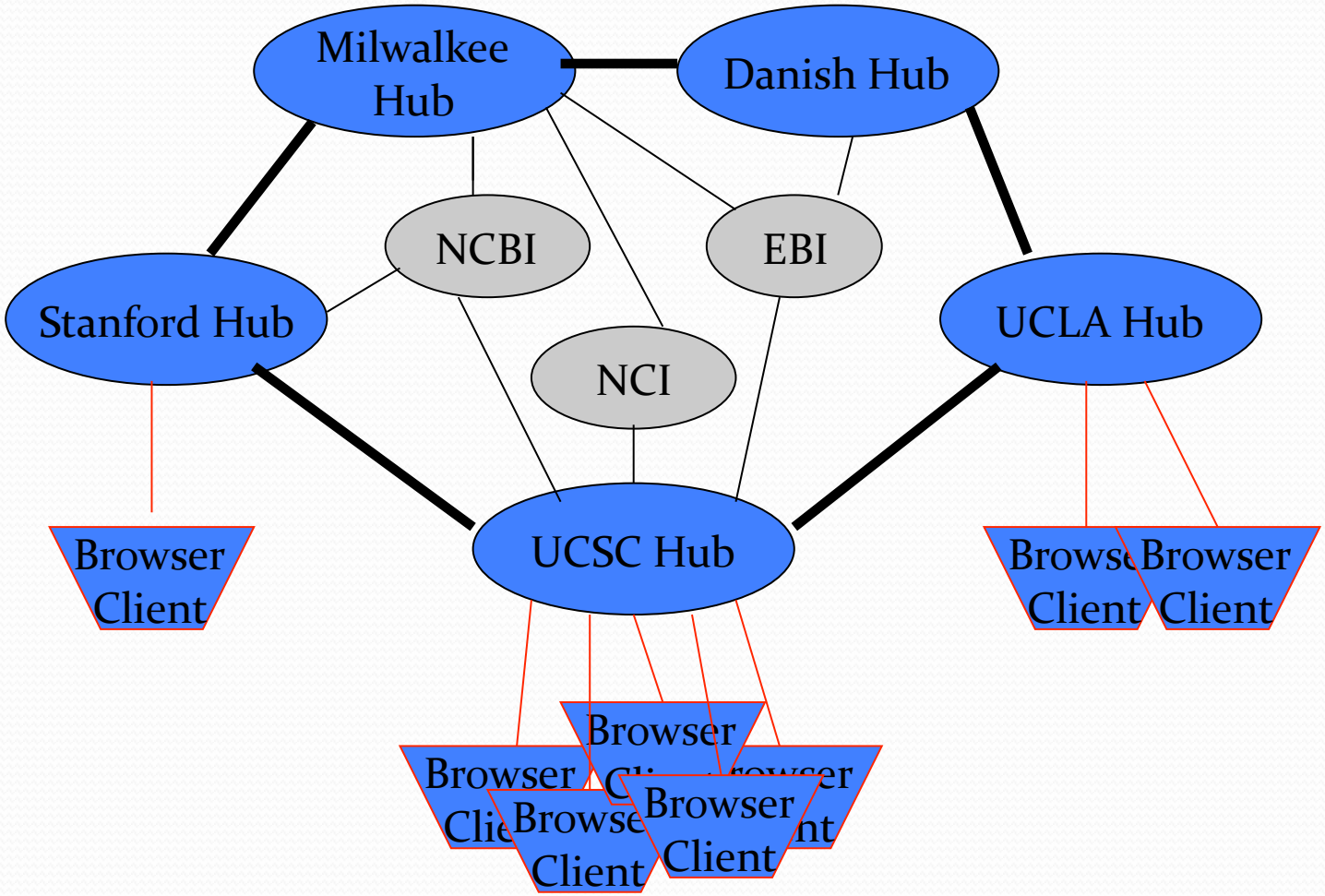
Hubs serve as translators and cache for data.



Web Services Interfaces to Other Databases



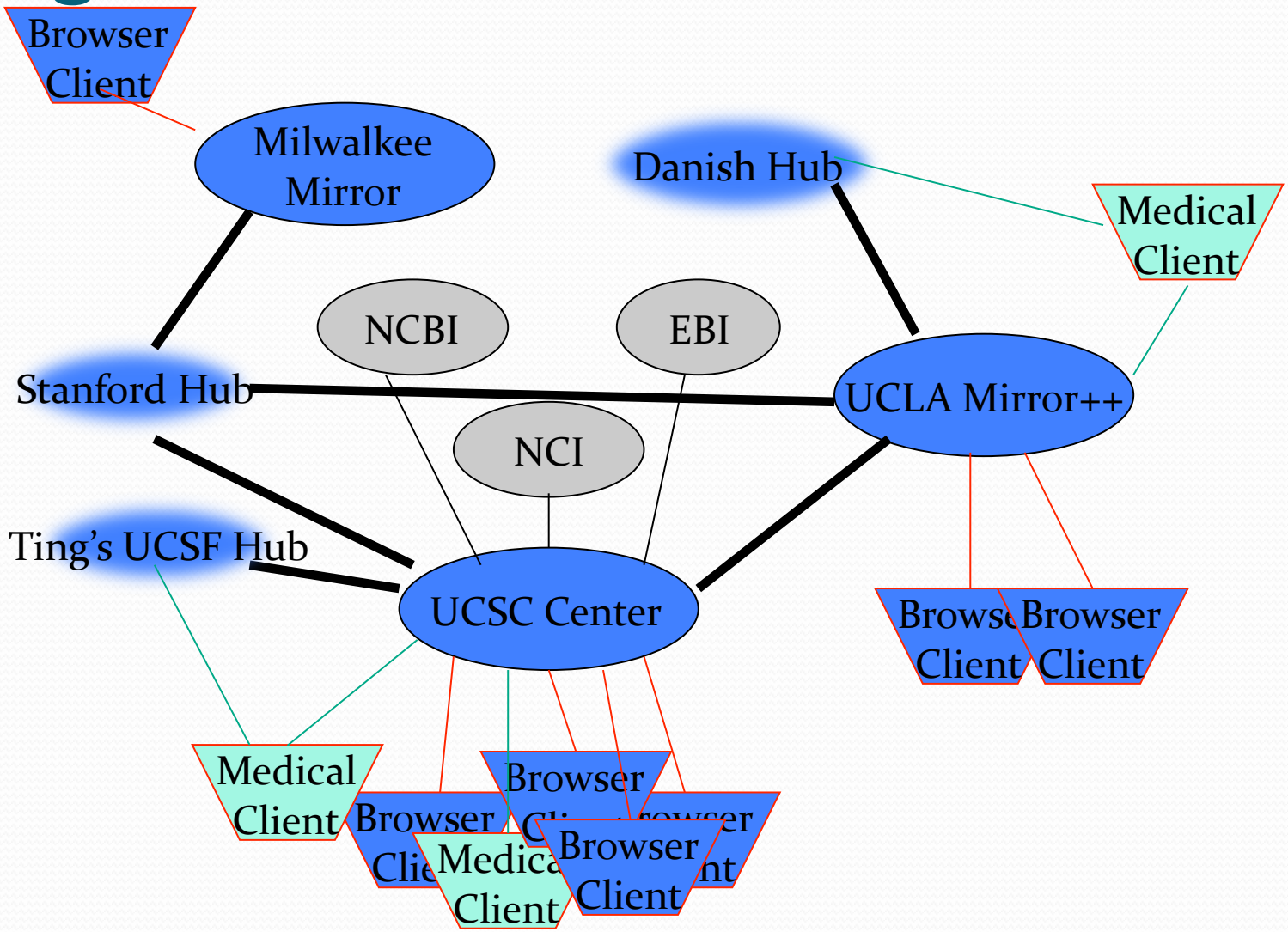
Javascript in web browser talks to just one hub



All good except for the current controversy….

Too much too fast?

Incremental design keeps main in hgTracks.



Features of incremental design

- Avoids recasting the “main” flow of control into JavaScript, which would be a radical departure.
- Less worry how to maintain current services and ongoing data integration including ENCODE DCC.
- Sometimes a more incremental approach has a better chance of success.
- Quite a bit of incremental work is already done…

Further Incremental Steps

- Create _file_ based data hubs.
- Put error trapping logic in hgTracks for remote tracks.
- Build web services first just as an alternative interface to the file based data hubs
- After some experience with those web services, expand them to include data in MySQL etc.
- Build robust application-level security module.
- Build up new medical apps that make use of security and web services modules.

File based data hubs

- Extension of custom track mechanism
- A directory full of bed,wig,gff and other genomics files, a trackDb.ra file that describes how to represent the files in the browser, and a genome.ra file that describes the assembly, organism, and the like.
- Imported as a new “group” with a blue-titled section of it’s own in track lists.

Trapping errors from remote sites

- Will be accessing data remotely more and more often.
- Both remote file oriented i/o and remote web services fail much more than local i/o.
- Almost worse than failing remote i/o can hang for long periods.
- Ideally need a system that fetches all remote tracks in parallel.
- At a minimum need a system that wraps a time out and an errCatch around remote track i/o.

Virtues of web services

- Interfacing with web services is easier for Python and JavaScript programs than interfacing with i/o driven by C libraries.
- Web service can be run at a major node in the internet on a server that has enough hard disk to do serious caching, reducing the i/o time over slow links.
- Web service can often reduce the i/o stream by 2x-20x over even a nicely designed, indexed, and summarized data file.
- A web service that served range queries on tracks on a data hub should not be hard to develop.
- Over time it would make sense to expand the web service to include serving things out of a combination of mySQL database plus /gbdb file combination like the genome browser.

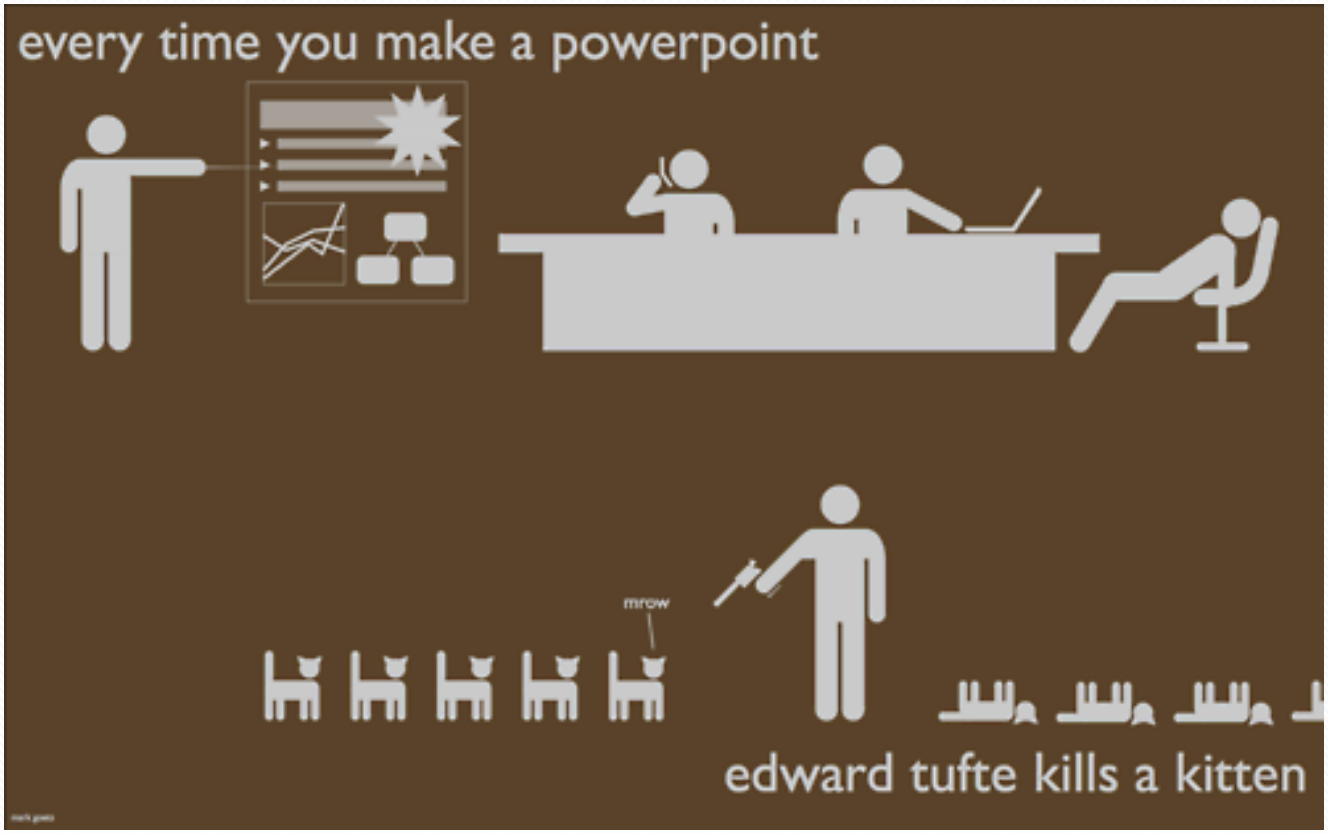
Security at application layer

- Increasingly we need to place layers of encryption and security on the data we serve.
- It's expensive to do security at the machine/virtual host level, and also it is less flexible than doing security at the application level.
- Only problem – security could be time consuming. Is this really what Galt wants to spend the rest of his life doing? Should we outsource some of it to a security expert? Does anyone want to back up Galt on security?
- Alternatives? Discuss!

Some jobs that need volunteers

- sketches for the Ped's copy # track, and possibly for an app around it.
- sketches for the face base app too.
- making track loading and track drawing each to be in multithreaded loops that catch and report errors in the track's space on the graphic.
- Integrate in gl library font that is one pixel higher than current default, and a little easier on the eyes. Possibly bring in other fonts from library.

The End


every time you make a powerpoint
mrow
edward tufte kills a kitten